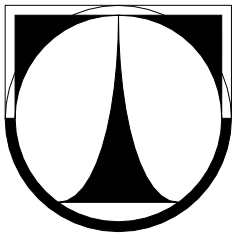
Faculty of Mechatronics and Interdisciplinary Engineering Studies
Technical University of Liberec

Institute of Computer Science
Academy of Sciences of the Czech Republic, Prague



Limiting Accuracy of Iterative Methods

Pavel Jiránek

PhD Thesis November, 2007

Title: Limiting Accuracy of Iterative Methods
Author: Pavel Jiránek
Study programme: P2612 Electrotechnics and Informatics
Field of study: 3901V025 Science Engineering
Department: Faculty of Mechatronics
and Interdisciplinary Engineering Studies
Technical University of Liberec
Hálkova 6, 46117 Liberec
Czech Republic

Institute of Computer Science
Academy of Sciences of the Czech Republic
Pod Vodárenskou věží 2, 18207 Prague 8
Czech Republic
Supervisor: Miroslav Rozložník

Prohlašuji,

že jsem tuto práci vypracoval samostatně a uvedl jsem veškeré prameny, kterých jsem použil.

Datum: _____ Podpis: _____

Abstrakt

Jak je známo, zaokrouhlovací chyby a nepřesné řešení vnitřních úloh mají vliv na numerické chování iteračních metod v aritmetice s konečnou přesností; obecně snižují jejich rychlost konvergence a ovlivňují konečnou přesnost spočteného řešení. V práci se zabýváme analýzou maximální dosažitelné přesnosti některých iteračních metod pro řešení soustav lineárních algebraických rovnic.

Dizertace je rozdělena na dvě části. První z nich obsahuje analýzu limitní přesnosti metod krylovovských podprostorů pro řešení rozsáhlých úloh sedlových bodů. Uvažujeme dva typy segregovaných metod: metodu redukce na Schurův doplněk a metodu projekce na nulový prostor mimodiagonálního bloku. Ukazuje se, že výběr vzorce pro zpětnou substituci má vliv na maximální dosažitelnou přesnost přibližného řešení spočteného v aritmetice s konečnou přesností.

Druhá část je věnována analýze numerického chování některých metod minimálních reziduí, které jsou matematicky ekvivalentní metodě zobecněných minimálních reziduí GMRES. Srovnáváme dva hlavní postupy: jeden, kde přibližné řešení je vypočteno ze soustav s horní trojúhelníkovou maticí, a jeden, kde je přibližné řešení upravováno pomocí jednoduchého rekurentního vzorce. Ukazuje se, že výběr báze má vliv na numerické chování výsledné implementace. Zatímco metody Simpler GMRES a ORTHODIR jsou méně stabilní díky špatné podmíněnosti zvolené báze, báze zkonstruovaná z reziduí může být dobře podmíněná, jestliže jsou normy reziduí dostatečně klesající. Tyto výsledky vedou k nové implementaci, která je podmíněně zpětně stabilní, a v jistém smyslu i vysvětlují experimentálně ověřený fakt, že metoda GCR (ORTHOMIN) dává v praktických aplikacích velmi přesné aproximace řešení.

Klíčová slova. Rozsáhlé lineární soustavy, metody krylovovských podprostorů, úlohy sedlového bodu, metoda redukce na Schurův doplněk, metoda projekce na nulový prostor mimodiagonálního bloku, metody minimálních reziduí, numerická stabilita, analýza zaokrouhlovacích chyb.

Abstract

It is known that inexact solution of inner systems and rounding errors affect the numerical behavior of iterative methods in finite precision arithmetic. In particular, they slow down their convergence rate and have an effect on the ultimate accuracy of the computed solution. Here we focus on the analysis of the maximum attainable accuracy of several iterative methods for solving systems of linear algebraic equations.

The thesis is divided into two parts. The first part is devoted to the analysis of Krylov subspace solvers applied to the large-scale saddle point problems. Two main representatives of segregated solution approaches are analyzed: the Schur complement reduction method and the null-space projection method. We show that the choice of the back-substitution formula can considerably influence the maximum attainable accuracy of approximate solutions computed in finite precision arithmetic.

In the second part we analyze numerical behavior of several minimum residual methods, which are mathematically equivalent to the GMRES method. Two main approaches are compared: the approach, which computes the approximate solution from an upper triangular system, and the approach where the approximate solutions are updated with a simple recursion formula. We show that a different choice of the basis can significantly influence the numerical behavior of resulting implementation. While Simpler GMRES and ORTHODIR are less stable due to ill-conditioning of chosen basis, the residual basis remains well-conditioned when we have a reasonable residual norm decrease. These results lead to a new implementation, which is conditionally backward stable, and in a sense explain an experimentally observed fact that the GCR (ORTHOMIN) method delivers in practical computations very accurate approximate solutions when it converges fast enough without stagnation.

Key words. large-scale linear systems, Krylov subspace methods, saddle point problems, Schur complement reduction, null-space projection method, minimum residual methods, numerical stability, rounding error analysis.

Аннотация

Известно, что неаккуратные решения внутренних проблем и ошибки округления отражаются на вычислительном поведению итерационных методов. Они конкретно затормозят их скорость сходимости и оказывают влияние на финальную аккуратность вычисленного решения. Мы здесь занимаемся анализом максимальной достижимой аккуратности некоторых итерационных методов для решения систем линейных алгебраических уравнений.

Эта диссертация разделена на две части. Первая занимается анализом лимитной аккуратности методов пространств Крылова для решения больших систем седельных точек. Мы рассматриваем два типы сегрегационных методов: методом преобразования на дополнение Шура и методом проекции на ядро недиагонального блока. Мы указываем, что выбор формулы обратной подстановки отражается на максимальной достижимой аккуратности приблизительного решения вычисленного в арифметике с конечной точностью.

Вторая часть содержит анализ вычислительного поведения нескольких методов минимальных невязок, которые математически эквивалентные методу «GMRES». Мы сравниваем два главные методы: один, который определяет приближённое решение из системы с верхней треугольной матрицой, и один, где приближённое решение корректированное с помощью простой рекуррентной формулы. Мы указываем, что выбор базы отражается на вычислительном поведении конечного метода. Пока методы «Simpler GMRES» и «ORTHODIR» менее стабильные за счет плохо обусловленной базы, база невязок может быть хорошо обусловленная, если нормы невязок достаточно снижаются. Эти результаты ведут к новому методу, который условно обратно стабильный, и в определенном смысле объясняют экспериментально удостоверенный факт, что метод «GCR» (также известный как «ORTHOMIN») даёт в практических аппликациях очень аккуратные аппроксимации решения.

Ключевые слова. большие линейные уравнения, методы подпространств Крылова, метод преобразования на дополнение Шура, метод проекции на ядро недиагонального блока, методы минимальных невязок, вычислительная стабильность, анализ ошибок округления.

Acknowledgements

First of all, I would like to thank Miroslav Rozložník for being a patient supervisor. His ideas and our fruitful discussions showed me the way where to go and inspired me to express my own creativity. I am also grateful to him for introducing me to the fascinating world of numerical mathematics and numerical linear algebra.

My thanks also go to Martin Gutknecht, the coauthor of the paper [**62**], Mario Arioli, Julien Langou, and Yvan Notay for their suggestions and discussions which considerably improved the presented results.

Financial support of my work on the thesis was provided in part by the Ministry of Education, Youth and Sports of the Czech Republic under the project 1M0554 "Advanced Remedial Technologies", and by the project 1ET400300415 within the National Program of Research "Information Society".

Finally, I would like to thank my family, especially my parents and sister, for their attention and support throughout the years of my life and studies in Liberec. And most of all, I am grateful to my girlfriend Eva who has been incredibly supportive, understanding and encouraging during the past years and I thank her for love and standing by me during the good and bad times.

Contents

CHAPTER 1

Introduction

Consider a system of linear algebraic equations in the form

$$Ax = b, \tag{1.1}$$

where A is an $N \times N$ nonsingular matrix and b is a right-hand side vector. Usually we assume that A is large and sparse as it is, e.g., when A is a discrete representation of some partial differential operator. We are looking for the solution of (1.1) or for its sufficiently accurate approximation.

The methods for solving (1.1) are usually classified as direct and iterative. Direct methods are mostly based on the successive elimination of unknowns. They factorize the system matrix (with suitably ordered rows or columns), e.g., into the product of lower and upper triangular matrices as in the Gaussian elimination, or to the product of an orthogonal and a triangular matrix as in the QR factorization. The solution of (1.1) can be then found by solving systems with these factors. In general, direct methods are well suited for dense and moderately large systems. However, when solving a large sparse system, the number of newly created non-zero elements in both factors can heavily affect the computational time and storage requirements. In addition, even though direct methods deliver in theory the exact solution, there is no need for such an accuracy in practice due to uncertain data or discretization errors.

Therefore, iterative methods became very popular when solving sparse systems. An iterative method for the solution of (1.1) generates a sequence of approximations x_k so that they ideally converge to the exact solution. The system matrix need not to be explicitly stored. In each iteration we need only to perform a matrix-vector multiplication. Moreover, the approximations converge often monotonously (or almost monotonously) in some fixed norm and so we can stop the iteration process when the approximation is accurate enough. However, the convergence rate of iterative methods can be slow in general (depending on properties of the system) and thus hybrid techniques combining the iterative and

direct approach, such as preconditioned iterations, are widely used to make the process more efficient.

In general, a solution method (no matter if a direct or iterative one) can be interpreted as a solution of a sequence of subproblems which are simpler to solve. In direct methods we can identify following subproblems: the factorization of the system matrix and the solution of systems with computed factors. In each step of an iterative method, we multiply a vector by the system matrix and optionally solve the system with a preconditioner which can be also regarded as the subproblems solved repeatedly in the iteration loop. E.g., the matrix-vector multiplication can involve the solution of an inner system as in the Schur complement reduction method which we will discuss later.

1. The state of the art

From now on we restrict ourselves to iterative methods. In practice, the computations are affected by errors. They are never performed exactly due to rounding errors and some of them are done inexactly with a prescribed level of accuracy, especially when computations with the working accuracy could be a waste of time and resources. E.g., matrix-vector products may involve a solution of inner systems, which (being large and sparse) can be solved inexactly with another iterative method. Preconditioning can be also applied through some iterative process. Usually, a method is called inexact if some involved subproblems are solved only approximately even though we assume exact arithmetic. Rounding errors can also considerably affect the behavior of iterative methods. Since the behavior of inexact iterative methods and "exact" methods in finite precision arithmetic is similar, we will not strictly distinguish between the sources of errors and we will treat them commonly in a unified approach in the following discussion.

When an inexactness is taken into account, there are several important questions which need to be answered. In the following we give a brief overview of the state of art in this field (including results in finite precision arithmetic). Generally the inexactness introduced in an iterative method has two main effects:

- The errors caused by inexact computations are propagated throughout the iterative process. Ideally the error propagation should be restrained so that the local errors are not magnified. There is a limit in the accuracy which cannot be exceeded and it is usually called the maximum attainable (or limiting) accuracy.

- The convergence of an inexact iterative method can be delayed with respect to the convergence of the same method, where all computations are performed exactly. We may ask how many additional iterations should be performed such that the same accuracy is attained as in the ideal (exact) case.

In this thesis we focus on the limiting accuracy of inexact iterative methods. The effects of inexact matrix-vector multiplications in iterative methods (also referred as relaxed methods) on the maximum attainable accuracy were studied simultaneously by van den Eshof and Sleijpen [**97**], and by Simoncini and Szyld [**90**]. Their analysis explains the experimental results of Bourass and Frayssé [**18**] (the report with an extensive experimental basis was published in 2000) who proposed a relaxation strategy for the accuracy of the computed matrix-vector product. They have shown that to achieve the prescribed accuracy of the computed solution we need to compute the matrix-vector product with the accuracy (measured by the backward error) inversely proportional to the actual residual norm. The papers [**97, 90**] provide the theoretical support for this strategy further developed in [**98**]. This topic is closely related to the flexible preconditioning, see, e.g., [**11, 43, 76, 90, 39**]. Here we try to adopt the backward error analysis, widely used in the study of rounding errors, and we analyze the effects of inexact computations on the limiting accuracy of certain iterative methods. The computations are performed in the presence of rounding errors while solutions to certain subproblems are done with more relaxed accuracy. We want to know how the inexactness of these inner systems together with the errors caused by roundoff affect the behavior of the considered algorithms. It appears that some measures of the accuracy are ultimately on the level proportional to the unit roundoff, while other measures depend on the accuracy of inner systems.

The problem of numerical stability of classical iterative methods was addressed in several papers. The first analyzes carried out by Golub [**40**] and Lynn [**69**] provide statistical and non-statistical results for the second order Richardson and SOR method. The statistical error analysis of classical iterative methods was also performed by Arioli and Romani [**5**] clarifying the relation between the conditioning of the preconditioned system matrix and the convergence rate of the iterative method. In [**56**] Higham and Knight give the forward and backward error analysis of a general one-step stationary method. Their analysis among other things shows that the accuracy of the computed solution strongly depends on the oscillations of norms of the iterates which is a common observation not

only in the case of classical iterative methods. Moreover, even though the convergence is driven by the spectral radius of the iteration matrix, the limiting accuracy depends rather on the norm of its powers which can be arbitrarily large in the early stage of the iterative process. This was observed by Hammarling and Wilkinson [**53**]. The stability of classical iterative methods was also analyzed by Woźniakovski in [**107, 108**]. He proved the forward stability of classical methods like Jacobi, Richardson, Gauss-Seidel and SOR (for symmetric systems with the Property A) and Chebyshev method (for symmetric positive definite systems). However, the Chebyshev method appeared to be not normwise backward stable. In [**41**] Golub and Overton discuss the convergence rate of the second order Richardson and Chebyshev method. They consider the inexact solution of inner systems with uniformly bounded relative residuals. The accuracy of the computed solution in the Chebyshev method is further analyzed by Giladi, Golub and Keller [**37**] who show the optimality of the uniform tolerance used in [**41**]. When the system is solved by the classical iterative method in each step we must solve a simpler system induced by the splitting of the system matrix. However, these systems can be also solved iteratively. These methods, referred to as two-stage methods, were addressed, e.g., in [**73, 64, 36**].

One of the most important result in the study of Krylov subspace methods is due to Paige [**77**]. He provides the analysis of the behavior of the symmetric Lanczos algorithm [**65**] in the presence of rounding errors. This algorithm is closely related to the conjugate gradient method by Hestenes and Stiefel [**54**]. It was first studied by Woźniakowski [**109**] and Bollen [**17**]. Woźniakowski shows that this method converges in finite precision arithmetic at least linearly with the convergence rate similar to the steepest descent method. However, his analysis does not reflect the reality very well, since the convergence of the conjugate gradient method cannot be characterized locally but its actual behavior depends on the whole iteration process; see, e.g., [**99, 68**] and the references therein. The new insight into this problem was brought by Greenbaum [**45**] and further developed together with Strakoš [**95, 49**]. It appears that the finite precision Lanczos process as well as the finite precision conjugate gradient method behave as their exact counterparts applied to the matrix of (possibly much) larger dimension with the eigenvalues clustered near the eigenvalues of the original matrix. This issue was further discussed by Notay in [**75**].

The analysis of limiting accuracy of some classes of iterative methods can be performed in rather general setting without referring to any particular method. The

methods based on the coupled two-term recurrences were analyzed by Greenbaum in [**46, 47**]. The papers focus mainly on the conjugate gradient method but the analysis holds for a larger set of methods. In particular, the results of Greenbaum show that the highly irregular convergence behavior (expressed by the oscillations of norms of iterates) observed in the case of non-optimal iterative methods (such as BiCG [**35**] or CGS [**93**]) can have an unfavorable effect on the limiting accuracy of the computed solution. A similar phenomenon is mentioned also by van der Vorst in [**100**], where the loss of accuracy is explained by oscillations of residual norms. On the other hand, such oscilations do not occur (or can be a priori bounded) in the case of optimal methods such as conjugate gradients and conjugate residuals [**94**] applied to symmetric positive definite problems, or in the case of residual minimizing methods (Orthodir [**110**], Orthomin [**102**], GCR [**29**]) for general nonsymmetric systems. The numerical stability of various (equivalent) methods using short recurrences was further studied by Gutknecht and Strakoš in [**52**] and by Sleijpen, van der Vorst and Modersitzki in [**92**]. In [**51**] Gutknecht and Rozložník discuss the effect of residual smoothing on the limiting accuracy.

Finally we survey the results for the finite precision behavior of nonsymmetric Krylov subspace methods with the full-term recurrences such as GMRES [**88**]. The Householder implementation of the underlying Arnoldi process [**6**] is quite straightforward to analyze, see the paper by Drkošová, Greenbaum, Rozložník and Strakoš [**27**], and by Arioli and Fassino [**4**]. This is due to the almost exact orthogonality of the computed Krylov subspace basis. However, when we use the cheaper modified Gram-Schmidt implementation, the orthogonality is gradually lost during the iteration process. The loss of orthogonality however goes hand in hand with the decrease of the backward error of the actual computed solution as observed by Greenbaum, Rozložník and Strakoš in [**48**] and further analyzed by Paige, Rozložník and Strakoš in [**80, 78**]. For more details see [**67**] and the references therein.

2. Organization of the thesis

This thesis is divided into two main parts and is organized as follows. Chapter 3, which is based on the papers [**61, 60**], is devoted to the analysis of inexact methods for solving saddle point problems of the form

$$\begin{pmatrix} A & B \\ B^T & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} f \\ 0 \end{pmatrix}.$$

A brief overview on saddle point problems is presented in Chapter 2. We analyze two segregated methods based on the transformation of the whole indefinite problem to a reduced system with more preferable properties (smaller dimension, positive (semi)definiteness). The reduced system is solved by a suitable iterative method giving the approximations to one of the block components of the solution vector (x or y). The remaining component is computed via some back-substitution formula. We consider three different but mathematically equivalent formulas. In each iteration we have to solve either a nonsingular system with A, or a full rank least squares problem with B. Since such systems are not usually solved exactly, we assume here that they are solved with a prescribed backward error and study the effect on the maximum attainable accuracy of the solution method together with the effects of rounding errors. Such inexact methods have been also considered in many papers but most of them analyzed the delay of convergence; see the references in Chapter 3. Here we provide a qualitative analysis of the maximum attainable accuracy of the computed solution measured by true residuals in the saddle point system, by true residuals in reduced systems and by forward errors of the computed solutions. In addition, we show which residuals (and how) can be affected by the possibly irregular convergence behavior in the case of the nonsymmetric block A. The theoretical results are illustrated on numerical experiments.

Chapter 4, based on the paper [**62**], is devoted to the analysis of several residual minimizing Krylov subspace methods, which are mathematically equivalent to the GMRES method [**88**]. In contrast to GMRES, they, in the nth iteration, build an orthonormal basis of $A\mathcal{K}_n(A, r_0)$ instead of $\mathcal{K}_n(A, r_0)$: $\mathcal{K}_n(A, r_0)$ denotes the nth Krylov subspace generated by the matrix A and the vector r_0. Two approaches are compared: the approach, which computes the approximate solution from an upper triangular system, and the approach, where the approximate solutions are updated step by step with a simple recursion formula. We consider a general basis to generate the orthonormal basis of $A\mathcal{K}_n(A, r_0)$, and it appears that, while Simpler GMRES and ORTHODIR are less stable due to ill-conditioning of the chosen basis, the residual basis can be well-conditioned, when we have a reasonable residual norm decrease. These results lead to a new implementation, which is conditionally backward stable, and to the well known GCR (ORTHOMIN) method, and in a sense explain an experimentally observed fact that GCR (ORTHOMIN) delivers very accurate approximate approximate solutions in practical applications. The theoretical results are illustrated on numerical experiments.

In Chapter 5 we give conclusions and directions of the future work.

3. List of related publications and conference talks

Journal papers.

- P. Jiránek, M. Rozložník. Maximum attainable accuracy of inexact saddle point solvers. Accepted for publication in *SIAM Journal on Matrix Analysis and Applications*, 2007.
- P. Jiránek, M. Rozložník. Limiting accuracy of segregated solution methods for nonsymmetric saddle point problems. Accepted for publication in *Journal of Computational and Applied Mathematics*, 2007.
- P. Jiránek, M. Rozložník, M. H. Gutknecht. How to make Simpler GMRES and GCR more stable. Submitted to *SIAM Journal on Matrix Analysis and Applications*, 2007.

Proceedings contributions.

- P. Jiránek. On a maximum attainable accuracy of some segregated techniques for saddle point problems. *Proceedings of the XI. PhD. Conference*, pages 26–34, Institute of Computer Science, CAS, Matfyzpress, Prague, 2006.
- P. Jiránek, M. Rozložník. On a limiting accuracy of segregated techniques for saddle point problems, *Proceedings of the 3rd International Workshop on Simulation, Modelling and Numerical Analysis SIMONA 2006*, pages 62–69, Liberec, September 2006.

Conference talks.

- P. Jiránek, M. Rozložník. Numerical behavior of iterative methods for saddle point problems. GAMM Annual Meeting 2006, Berlin, March 27–31, 2006.
- P. Jiránek. On a maximum attainable accuracy of some segregated techniques for saddle point solvers. XI. PhD. Conference, Institute of Computer Science, Academy of Sciences of the Czech Republic, Monínec – Sedlec-Prčice, September 18–20, 2006.
- P. Jiránek, M. Rozložník. On a limiting accuracy of segregated techniques for saddle point solvers. Simulation, Modelling and Numerical Analysis SIMONA 2006, Liberec, September 18–20, 2006.
- P. Jiránek, M. Rozložník. Numerical solution of saddle point problems. SNA'07, Seminar on Numerical Analysis, Ostrava, January 22–26, 2007.

- P. Jiránek, M. Rozložník. On the limiting accuracy of segregated saddle point solvers. MAT-TRIAD 2007 – three days full of matrices, Będlewo, Poland, March 22–24, 2007.
- P. Jiránek, M. Rozložník. On the limiting accuracy of segregated saddle point solvers. VIII. vedecká konferencia s medzinárodnou účasťou, Technical University of Košice, Slovakia, May 28–30, 2007.
- P. Jiránek, M. Rozložník. Limiting accuracy of inexact saddle point solvers. 22nd Biennial Conference on Numerical Analysis, University of Dundee, Scotland, UK, June 26–29, 2007.
- P. Jiránek, M. Rozložník, M. H. Gutknecht. On the stability of Simpler GMRES. CEMRACS'07, Lumini, France, Juny 22–August 31, 2007.
- P. Jiránek, M. Rozložník, M. H. Gutknecht. How to make Simpler GMRES and GCR more stable. IMA Conference on Numerical Linear Algebra and Optimisation, University of Birmingham, UK, September 13–15, 2007.

3.1. Posters.

- P. Jiránek, M. Rozložník. Numerical stability of inexact saddle point solvers. ICIAM'07, 6th International Congress on Industrial and Applied Mathematics, Zurich, Switzerland, July 16–20, 2007.

CHAPTER 2

Saddle point problems

The solution of large-scale systems in the saddle point form attracted a lot of attention in recent years. They appear in a large variety of applications and many solution methods were developed so far. The next chapter is devoted to the numerical stability analysis of certain iterative methods for saddle point systems and, before we start, we give a short introduction into this field. For an exhaustive overview we refer to the paper by Benzi, Golub and Liesen [**14**].

We consider the large sparse system of linear algebraic equations in the block form

$$\mathcal{A}\begin{pmatrix} x \\ y \end{pmatrix} \equiv \begin{pmatrix} A & B \\ B^T & -C \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix}, \tag{2.1}$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$ and $C \in \mathbb{R}^{m\times m}$. The solution and right-hand side vectors are partitioned consistently with respect to the partitioning of the system matrix. Let A and B are nonzero matrices and furthermore we assume that the right-hand side is always chosen so that the system is consistent.

The properties of blocks A, B and C may vary depending on the application. In the following section we mention several important examples of problems leading to a saddle point system. Note that the system (2.1) has a symmetric block structure which can be relaxed when solving so called generalized saddle point problems. However, we do not consider this case here.

1. Applications leading to saddle point problems

Saddle point problems arise in a wide selection of problems of computational science and engineering. When A is symmetric positive definite, B has a full column rank and $C = 0$, we have the most common version of the saddle point system

$$\begin{pmatrix} A & B \\ B^T & 0 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix}, \tag{2.2}$$

which appears, e.g., when solving elliptic second order partial differential equations by the mixed finite element method [**24**] or quadratic programming problems with linear constraints [**38, 74**]. The component x of the solution vector (x, y) of (2.2) is the solution of the constrained minimization problem

$$\min_{u \in \mathbb{R}^n} \; J(u) = \frac{1}{2} u^T A u - f^T u \;\; \text{s.t.} \;\; B^T u = g. \tag{2.3}$$

The corresponding Lagrangian is defined as

$$\mathcal{L}(u, v) = J(u) + (B^T u - g)^T v \qquad \forall u \in \mathbb{R}^n, \quad \forall v \in \mathbb{R}^m,$$

where v is the vector of Lagrange multipliers. The vector (x, y) is the saddle point of $\mathcal{L}$,

$$\mathcal{L}(x, v) \leq \mathcal{L}(x, y) \leq \mathcal{L}(u, y).$$

The nonsymmetric block A appears, e.g., when solving linearized Navier-Stokes equation via the sequence of Stokes and Oseen problems. If, in the mixed finite elements, the approximation spaces do not fulfill the LBB condition, the stabilization should be applied leading to the nonzero symetric positive semidefinite matrix C [**24, 32**].

Another important application of saddle point systems is the solution of linear least squares problems. Let B be an $n \times m$ matrix of a full column rank and consider

$$\text{find} \;\; y \;\; \text{s.t.} \;\; \|f - By\| = \min_{v \in \mathbb{R}^m} \|f - Bv\|.$$

It is well-known [**16, 42**] that the solution of this problem is unique and it is characterized by the orthogonality condition $x = f - By \perp R(B) = N(B^T)^\perp$ for the residual vector x (where $R(B)$ and $N(B^T)$ denotes the range and null-space of the matrix B and B^T, respectively). Hence we have $x + By = f$, $B^T x = 0$ leading to the system

$$\begin{pmatrix} I & B \\ B^T & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} f \\ 0 \end{pmatrix}.$$

In general, the system of the form (2.2) (with $g = 0$) corresponds to the weighted least squares problem, where A^{-1}-norm is minimized instead of the Euclidean one (when A is symmetric positive definite).

2. Properties of saddle point matrices

Here we briefly recall the basic properties of saddle point matrices and relate their spectral and nonsingularity properties with respect to the properties of particular blocks. We restrict ourselves to the symmetric case but some results

can be extended to a more general setting. For a more complete discussion, see [**14**].

THEOREM 2.1. *Let A be a symmetric positive definite matrix with eigenvalues contained in the interval $[\underline{\lambda}, \overline{\lambda}]$ and let B be of a full column rank with singular values contained in $[\underline{\sigma}, \overline{\sigma}]$ with $\underline{\lambda} > 0$ and $\underline{\sigma} > 0$ and C is symmetric positive semidefinite. Then*

- *$\mathcal{A}$ has n positive and m negative eigenvalues;*
- *if $C = 0$, the eigenvalues of $\mathcal{A}$ are localized as follows:*

$$\lambda(A) \subset I^- \cup I^+,$$

where

$$I^- \equiv \left[\frac{1}{2}\left(\underline{\lambda} - \sqrt{\underline{\lambda}^2 + 4\overline{\sigma}^2}\right), \frac{1}{2}\left(\overline{\lambda} - \sqrt{\overline{\lambda}^2 + 4\underline{\sigma}^2}\right)\right],$$

$$I^+ \equiv \left[\underline{\lambda}, \frac{1}{2}\left(\overline{\lambda} + \sqrt{\overline{\lambda}^2 + 4\overline{\sigma}^2}\right)\right].$$

PROOF. The saddle point matrix $\mathcal{A}$ can be factorized as follows

$$\mathcal{A} = \begin{pmatrix} I & 0 \\ B^T A^{-1} & I \end{pmatrix} \begin{pmatrix} A & 0 \\ 0 & -B^T A^{-1} B - C \end{pmatrix} \begin{pmatrix} I & A^{-1}B \\ 0 & I \end{pmatrix}.$$

The first statement immediately follows from the Sylvester's law of inertia [**57**], since the Schur complement $-B^T A^{-1} B - C$ is symmetric negative definite. For the proof of the second statement, see [**85**]. □

The matrix $\mathcal{A}$ is indefinite, since it has both positive and negative eigenvalues. Solving highly indefinite matrices (with $n \approx m$) can lead to the slow convergence when using Krylov subspace methods like MINRES [**79**], see [**34**]. A simple modification of the system matrix in the form

$$\hat{\mathcal{A}} \equiv \begin{pmatrix} A & B \\ -B^T & C \end{pmatrix},$$

as observed, e.g., in [**13, 34**], leads to a nonsingular system with a spectrum moved to the right half-plane of the complex plane but, however, for the price of losing the symmetry.

The nonsingularity conditions are summarized in the following theorem (see [**13**]).

THEOREM 2.2. *Let A be symmetric nonnegative real (that is, $\frac{1}{2}(A + A^T)$ is positive semidefinite), B has a full column rank and let C be symmetric positive semidefinite. Then*

- *if $\mathcal{A}$ is nonsingular, then $N(A) \cap N(B^T) = 0$;*
- *if $N(\frac{1}{2}(A + A^T)) \cap N(B^T) = 0$, then $\mathcal{A}$ is nonsingular.*

Here 0 represents the null subspace of $\mathbb{R}^n$. In particular, if A is symmetric positive semidefinite, then $\mathcal{A}$ is nonsingular if and only if $N(A) \cap N(B^T) = 0$.

3. Solution methods

Solution methods for systems of the form (2.1) can be divided into two categories called coupled and segregated methods. Coupled methods solve the system (2.1) as a whole and therefore compute both components x and y of the solution vector at once. They can be both direct, e.g., using LDL^T factorization with 1×1 and 2×2 pivots, and iterative, e.g., using MINRES [**79**] in the symmetric case. On the other hand, segregated methods transform the system (2.1) of the dimension $n+m$ to a reduced system of a smaller dimension solving either for the component x or y. The remaining component is then found by the back-substitution into (2.1). The reduced systems can be also solved either directly or iteratively. They can be hard to compute explicitly, so the iterative approach is more preferable in many cases. Moreover, besides the smaller dimension, the reduced systems can be easier to solve than the whole saddle point system (e.g., the reduced system can be positive (semi)definite). Sometimes the border between coupled and segregated approaches is not sharp, since coupled methods can be treated as segregated and vice versa. Here we review two main representatives of segregated approaches which will be analyzed in the next chapter: the Schur complement reduction method and the null-space projection method. We will not discuss other issues related to the topic and solution methods, especially preconditioning of saddle point problems; see [**14**] for more information.

3.1. The Schur complement reduction method. Assume A is symmetric positive definite, B has a full column rank and C is symmetric positive semidefinite. Then Theorem 2.2 implies that the system (2.1) has a unique solution. It can be regarded as two matrix-vector equations in the form

$$Ax + By = f, \quad B^T x - Cy = g. \tag{2.4}$$

Since A is nonsingular, we can to eliminate x from the first equation, i.e., x can be expressed as

$$x = A^{-1}(f - By), \tag{2.5}$$

and substituted into the second equation. Then we obtain the system

$$Sy = B^T A^{-1} f - g, \quad S \equiv B^T A^{-1} B + C \tag{2.6}$$

with the Schur complement matrix S (which is, more precisely, the negative Schur complement of A in $\mathcal{A}$). The solution of an $(n+m)$-dimensional indefinite problem (2.1) is thus transformed to the solution of two systems of orders m and n with symmetric positive definite matrices. First, the system (2.6) is solved for y. It is not always preferable to compute S directly, since, even though A is sparse, S need not to be. Sometimes the elimination process can be performed such that the sparsity is preserved [**71**]. When (2.6) is solved iteratively, we need to compute the product with S which involves the solution of a system with the matrix A. The iterative method produces the sequence of approximations y_k $(k = 0, 1, 2, \ldots)$ converging ideally to y. When the vector y or an iterate y_k is available, the corresponding approximation to x can be computed by the substitution into (2.5).

One of the most popular methods for solving saddle point systems based on the Schur complement reduction is the Uzawa method [**7**]. The algorithm is as follows: choose y_0, then for $k = 0, 1, 2, \ldots$ do

$$\begin{cases} \text{solve } \ Ax_{k+1} = f - By_k, \\ y_{k+1} = y_k - \alpha(g - B^T x_{k+1} + Cy_k). \end{cases}$$

Here $\alpha > 0$ is a relaxation parameter. Hence we can write the iteration in the form

$$\begin{pmatrix} A & 0 \\ B^T & -\alpha^{-1}I \end{pmatrix} \begin{pmatrix} x_{k+1} \\ y_{k+1} \end{pmatrix} = \begin{pmatrix} 0 & -B \\ 0 & -\alpha^{-1}I - C \end{pmatrix} \begin{pmatrix} x_k \\ y_k \end{pmatrix} + \begin{pmatrix} f \\ g \end{pmatrix}.$$

The direct computation shows that the iteration matrix of the associated stationary method is

$$\begin{pmatrix} A & 0 \\ B^T & -\alpha^{-1}I \end{pmatrix}^{-1} \begin{pmatrix} 0 & -B \\ 0 & -\alpha^{-1}I - C \end{pmatrix} = \begin{pmatrix} 0 & -A^{-1}B \\ 0 & I - \alpha S \end{pmatrix}.$$

Thus the Uzawa method converges if and only if the spectral radius of $I - \alpha S$ is strictly less than one. It is easy to see that the Uzawa method is based on the Schur complement method, since it is nothing but the Richardson iteration applied to the Schur complement system (2.6). On the other hand, the Uzawa method can be regarded as a block Gauss-Seidel method (with a regularization in the block $(2, 2)$) applied to the saddle point system (2.1).

3.2. The null-space projection method. The Schur complement reduction relies on the effective solution of systems with the matrix A. Sometimes the application of A^{-1} is hard to compute in which case the null-space projection method can be the method of choice. Assume here that A is symmetric positive

definite on $N(B^T)$, B has a full column rank and $C = 0$. The system (2.2) is thus by Theorem 2.2 uniquely solvable and can be expressed as two matrix-vector equations

$$Ax + By = f, \quad B^T x = g. \tag{2.7}$$

Let x_0 be a particular solution of the second equation and $Z \in \mathbb{R}^{n\times(n-m)}$ be a matrix containing a basis of the null-space of B^T. Every such solution lies in the affine space $x_0 + N(B^T)$ and hence has the form $x = x_0 + Zx_Z$, where $x_Z \in \mathbb{R}^{n-m}$ are the coordinates of $x - x_0$ in the null-space basis Z. Substitution into the first equation of (2.4) and premultiplying by Z^T gives the symmetric positive definite system

$$Z^T AZ x_Z = Z^T(f - Ax_0) \tag{2.8}$$

that is, the reduced system of the order $n - m$ for the components of $x - x_0$ in the basis of $N(B^T)$. The system $Z^T AZ$ can be solved directly or iteratively. When we have Z explicitly available (e.g., by the sparse QR factorization) both approaches can be applied. However, when using an iterative method, it can be implemented so that the matrix Z is kept only implicitly [**44**]. We can view the solution of (2.8) as the solution of a projected system

$$(I - \Pi)A(I - \Pi)x_1 = (I - \Pi)f, \tag{2.9}$$

where $x_1 = Zx_Z$ and Π is the orthogonal projector onto $R(B)$. The solution component y can be then found via the solution of the least squares problem

$$\|f - Ax - By\| = \min_{v\in\mathbb{R}^m} \|f - Ax - Bv\|. \tag{2.10}$$

When (2.8) or (2.9) is solved iteratively producing the sequence of approximations x_k ($k = 0, 1, 2, \ldots$), solving (2.10) gives an approximation y_k to y with x replaced by x_k.

CHAPTER 3

Limiting accuracy of segregated saddle point solvers

We want to solve a saddle point system which is in fact the symmetric indefinite system with 2×2 block structure

$$\begin{pmatrix} A & B \\ B^T & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} f \\ 0 \end{pmatrix}, \tag{3.1}$$

where the diagonal $n \times n$ block A is symmetric positive definite and the $n \times m$ off-diagonal block B has full column rank. Saddle point problems have recently attracted a lot of attention and appear to be a time-critical component in the solution of large-scale problems in many applications of computational science and engineering. A large amount of work has been devoted to a wide selection of solution techniques varying from the fully direct approach, through the use of iterative stationary or Krylov subspace methods, up to the combination of direct and iterative techniques including preconditioned iterative schemes. For an excellent survey on applications, methods, and results on numerical solution of saddle point problems, we refer to [**14**] and numerous references therein (relevant references will be given later in the text). Significantly less attention, however, has been paid so far to the numerical stability aspects. Here we concentrate on the numerical behavior of schemes which compute separately the unknown vectors x and y: one of them is first obtained from a reduced system of a smaller dimension, and, once it has been computed, the other unknown is obtained by back-substitution solving exactly or inexactly another reduced problem. The main representatives of such a segregated approach are the Schur complement reduction method and the null-space projection method. We analyze such algorithms which can be interpreted as iterations for the reduced system but compute the approximate solutions x_k and y_k to both unknown vectors x and y simultaneously.

The Schur complement reduction method uses the block factorization in the form

$$\begin{pmatrix} A & B \\ B^T & 0 \end{pmatrix} = \begin{pmatrix} I & 0 \\ B^T A^{-1} & I \end{pmatrix} \begin{pmatrix} A & B \\ 0 & -B^T A^{-1} B \end{pmatrix},$$

where the matrix $-B^T A^{-1} B$ is the Schur complement of A in (3.1). Such decomposition leads to solving the resulting block triangular system

$$\begin{pmatrix} A & B \\ 0 & -B^T A^{-1} B \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} f \\ -B^T A^{-1} f \end{pmatrix}, \tag{3.2}$$

which is nothing but a block Gaussian elimination applied to the original system (3.1). The block triangular system (3.2) is solved by computing the unknown y from the symmetric positive definite Schur complement system

$$B^T A^{-1} B y = B^T A^{-1} f \tag{3.3}$$

of order m and then by computing the unknown x from a system of order n with the symmetric positive definite matrix A. This approach leads to the explicit formula for the unknown vector $x = A^{-1}(f - By)$. The null-space projection method is based on the projection of the first block equation in (3.1) onto the null-space $N(B^T)$ and onto its orthogonal complement $R(B)$, respectively. According to the second block equation of (3.1) the unknown x belongs to $N(B^T)$ and therefore we get the block triangular system

$$\begin{pmatrix} (I - \Pi) A (I - \Pi) & 0 \\ B^T A & B^T B \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} (I - \Pi) f \\ B^T f \end{pmatrix}, \tag{3.4}$$

where $\Pi \equiv B(B^T B)^{-1} B^T$ denotes the orthogonal projector onto $R(B)$. This triangular system is solved by forward substitution, where we first compute the unknown x from the projected system

$$(I - \Pi) A (I - \Pi) x = (I - \Pi) f \tag{3.5}$$

of order n with the symmetric positive semi-definite matrix $(I - \Pi)A(I - \Pi)$. Once it has been computed, the unknown y is obtained as $y = B^{\dagger}(f - Ax)$ by solving the least squares problem

$$\|f - Ax - By\| = \min_{v \in \mathbb{R}^m} \|f - Ax - Bv\|, \tag{3.6}$$

where $B^{\dagger}$ denotes the Moore–Penrose pseudoinverse of B. The success of algorithms for solving the block triangular systems (3.2) or (3.4) depends on the availability of good approximations to the inverse of the block A or to the

pseudoinverse of B, respectively. More precisely, one looks for a cheap approximate solution to the inner systems with the matrix A and/or to the associated least squares problems with the matrix B. Numerous inexact schemes have been used and analyzed, see, e.g., the analysis of inexact Uzawa algorithms [**31, 22, 23, 12, 112**], inexact null-space methods [**89, 105, 111**], multilevel or multigrid methods [**21, 20, 111**], domain decomposition methods [**19**], two-stage iterative processes [**73, 36**] or inner-outer iterations [**43**]. These papers contain mainly the analysis of a convergence delay caused by the inexact solution of inner systems or least squares problems.

We concentrate on the question of what is the best accuracy we can get from inexact schemes solving either (3.2) or (3.4) when implemented in finite precision arithmetic. The fact that the inner solution tolerance strongly influences the accuracy of computed iterates is known and was studied in several contexts. The general framework for understanding inexact Krylov subspace methods has been developed in [**90**] and [**97**]. Assuming exact arithmetic, Simoncini and Szyld [**90**] and van den Eshof and Sleijpen [**97**] investigated the effect of an approximately computed matrix-vector product in every iteration on the ultimate accuracy of several solvers and explained the success of relaxation strategies for the inner accuracy tolerance from [**18, 19, 39**]. The developed theory strongly exploits the particular properties of an iterative method used for solving the associated system. In the context of saddle point problems, this requires a deep analysis of the outer iteration scheme for solving the reduced Schur complement or projected system (in particular, we refer to [**90**, Section 8]).

The effects of rounding errors in the Schur complement reduction method and the null-space projection method have been studied, e.g., in [**2, 3, 26, 70**], where the maximum attainable accuracy of computed approximate solutions by means of residuals and errors is estimated depending on the user tolerance specified in the outer iteration. We analyze the influence of the inexact solution of inner systems/least squares problems on the same quantities. Our approach is based on a standard backward analysis which allows us to take into account both the inexactness of the inner iteration loops as well as the accompanying rounding errors that occur in finite precision arithmetic.

The theory developed for the outer iteration process is similar to the analysis of Greenbaum in [**47, 46**] who estimated the gap between the true and recursively updated residual for a general class of iterative methods using coupled two-term recursions. The difference here is that every computed approximate solution of inner problem is interpreted as an exact solution of a perturbed problem induced

by the actual stopping criterion, while the theory of [**47**] considered only the rounding errors associated with a fixed matrix-vector multiplication. In contrast to the theory of inexact Krylov methods [**90, 97**], the bounds for the true residual in the outer iteration loop are obtained without specifying the solver used for solving the Schur complement or the projected Hessian system. It appears that the maximum attainable accuracy level in the outer process is mainly given by the inexactness of solving the inner problems and it is not further magnified by the associated rounding errors. These results are thus similar to ones which can be obtained in exact arithmetic.

The situation is different when looking at the numerical behavior of residuals associated with the original saddle point system, which describe how accurately the block equations (3.1) are satisfied. It is shown that the attainable accuracy of computed approximate solutions then depends significantly on the back-substitution formula used for computing the remaining unknowns. Our results show that, independent of the fact that the inner systems are solved inexactly, some back-substitution schemes lead ultimately to residuals on the roundoff unit level. Indeed, our results confirm that depending which back-substitution formula is used the computed iterates may satisfy either the first or the second block equation to the working accuracy. We believe that such results cannot be obtained using the exact arithmetic considerations and are of importance in applications requiring accurate approximations (see e.g. [**44, 38, 24**]). On the other hand, we agree that in many applications the saddle point system comes from a discretization of certain partial differential equations and much lower accuracy is sufficient. In any case, we give a theoretical explanation for the behavior which was probably observed or is already implicitly known. However, we have not found any explicit references to this issue. The implementations that we point out as optimal are actually those which are widely used and suggested in applications.

The chapter is organized as follows. Sections 1 and 2 are devoted to the rounding error analysis of the Schur complement reduction method and the null-space projection method, respectively. Each section is divided into five subsections. In subsections 1.1 and 2.1 we analyze the influence of inexact solution of inner systems or least squares on the maximum attainable accuracy in the outer iteration process for solving (3.2) or (3.4), and we estimate the ultimate norms of the true residuals $-B^TA^{-1}f + B^TA^{-1}B\bar{y}_k$ and $(I-\Pi)f - (I-\Pi)A(I-\Pi)\bar{x}_k$. In the consequent three subsections of Sections 1 and 2, we give bounds for the ultimate norm of the true residuals $f - A\bar{x}_k - B\bar{y}_k$ and $-B^T\bar{x}_k$. As we will see

in subsections 1.2–1.4 and 2.2–2.4, the limiting accuracy of these residuals may significantly differ for various back-substitution formulas for computing x_k or y_k, respectively. Subsections 1.5 and 2.5 contain forward analysis with the bounds for the errors $x - \bar{x}_k$ and $y - \bar{y}_k$. Throughout this chapter our theoretical results are illustrated on the model example taken from [**83**]: we put $n = 100$, $m = 20$, and

$$A = \text{tridiag}(1, 4, 1) \in \mathbb{R}^{n \times n}, \quad B = \text{rand}(n, m), \quad f = \text{rand}(n, 1).$$

The spectrum of A and singular values of B lie in the interval $[2.001, 5.999]$ and $[2.173, 7.170]$, respectively. Therefore the conditioning of A or B does not play an important role in our experiments. For further discussion, we refer to subsections 1.5 and 2.5.

For distinction, we denote quantities computed in finite precision arithmetic by bars. We assume that the usual rules of a well-designed floating-point arithmetic hold, and use occasionally the notation $\text{fl}(\cdot)$ for a computed result of an expression. The roundoff unit is denoted by u. In particular, for a matrix-vector multiplication the bound $\|\text{fl}(Ax) - Ax\| \leq O(u)\|A\|\|x\|$ is used and $\|x\|$ denotes the 2-norm of the vector x; for a general matrix A we make use of the spectral norm $\|A\|$ and the corresponding condition number $\kappa(A) = \|A\|/\sigma_{min}(A)$, where $\sigma_{min}(A)$ is the minimal singular value of A. For a symmetric positive definite matrix A, $\|x\|_A$ denotes the A-norm of the vector x. Finally, we apply the O-notation when suitable.

1. Schur complement reduction method

In this section we will discuss algorithms which compute simultaneously approximations x_k and y_k to the unknowns x and y and ideally fulfill the first block equation of (3.1)

$$Ax_k + By_k = f. \tag{3.7}$$

Our goal here is not to survey all existing schemes based on (3.7) but to analyze the numerical behavior of three implementations which use different back-substitution formulas for computing the approximate solution x_k. More precisely, without specifying any particular method, we assume that we have computed the approximate solution y_{k+1} and the residual vector $r_{k+1}^{(y)}$ using the recursions

$$y_{k+1} = y_k + \alpha_k p_k^{(y)}, \tag{3.8}$$

$$r_{k+1}^{(y)} = r_k^{(y)} + \alpha_k B^T A^{-1} B p_k^{(y)} \tag{3.9}$$

with $r_0^{(y)} = -B^T A^{-1}(f - By_0)$. We will distinguish between the following three mathematically equivalent formulas:

$$x_{k+1} = x_k + \alpha_k(-A^{-1}Bp_k^{(y)}), \tag{3.10}$$

$$x_{k+1} = A^{-1}(f - By_{k+1}), \tag{3.11}$$

$$x_{k+1} = x_k + A^{-1}(f - Ax_k - By_{k+1}). \tag{3.12}$$

The resulting schemes are summarized in Figure 3.1. These schemes have been used and studied in the context of many applications, including various classical Uzawa algorithms, two-level pressure correction approach, or inner-outer iteration method for solving (3.1); see, e.g., the schemes with (3.10) in [**82, 10**], (3.11) in [**31**], or (3.12) in [**22, 23, 12, 112**], respectively. Because the solves with matrix A in formulas (3.10)–(3.12) are expensive, these systems are in practice solved only approximately. Our analysis is based on the assumption that every solution of a symmetric positive definite system with the matrix A is replaced by an approximate solution produced by an arbitrary method. The resulting vector is then interpreted as an exact solution of the system with the same right-hand side vector but with a perturbed matrix $A + \Delta A$. We always require that the relative norm of the perturbation is bounded as $\|\Delta A\| \leq \tau\|A\|$, where τ represents a backward error associated with the computed solution vector. We will always assume that the perturbation ΔA does not exceed the limitation given by the distance of A to the nearest singular matrix and put restriction in the form $\tau\kappa(A) \ll 1$. It follows then from the standard perturbation analysis (see, e.g., [**55, 16**]) that

$$\|(A + \Delta A)^{-1} - A^{-1}\| \leq \frac{\tau\kappa(A)}{1 - \tau\kappa(A)}\|A^{-1}\|.$$

Note that if $\tau = O(u)$, then we have a backward stable method for solving the positive definite system with A. In our numerical experiments, we solve the systems with A inexactly using the conjugate gradient method or with the Cholesky factorization as indicated by the notation $\tau = O(u)$.

1.1. The attainable accuracy in the Schur complement system. In this subsection we look at the ultimate accuracy in the outer iteration process by means of the true residual $-B^T A^{-1} f + B^T A^{-1} B\bar{y}_k$. It is clear that if we perturb the Schur complement system $-B^T A^{-1} By = -B^T A^{-1} f$ to $-B^T(A + \Delta A)^{-1} B\hat{y} = -B^T A^{-1} f$, where $\|\Delta A\| \leq \tau\|A\|$, then the residual associated with

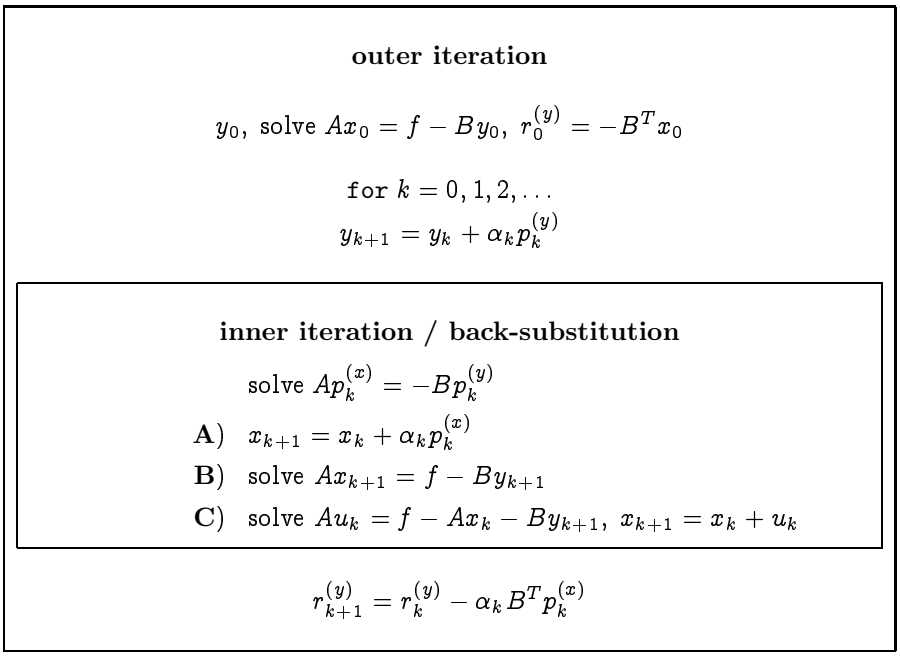


FIGURE 3.1. Schur complement reduction: Three different schemes for computing the approximate solution x_{k+1} (called in the text the updated approximate solution (A), the approximate solution computed by a direct substitution (B), and the approximate solution computed by a corrected direct substitution (C), respectively).

$\hat{y}$ can be bounded as

$$\| - B^T A^{-1} f + B^T A^{-1} B\hat{y}\| \leq \frac{\tau\kappa(A)}{1 - \tau\kappa(A)} \|A^{-1}\| \|B\|^2 \|\hat{y}\|. \tag{3.13}$$

We see from (3.13) that there is a limitation to the accuracy of the residual obtained directly from $\hat{y}$ and its bound is proportional to τ. Note that these considerations were made assuming exact arithmetic. The effects of rounding errors on the same quantity have been studied by Greenbaum [**47**], who considered a general class of methods for solving the fixed system of linear equations using

two-term recursions given by (3.8) and (3.9). Using a similar approach we can extend these results and formulate the following theorem.

THEOREM 3.1. *The gap between the true residual* $-B^TA^{-1}f+B^TA^{-1}B\bar{y}_k$ *and the updated residual* $\bar{r}_k^{(y)}$ *can be bounded as*

$$\begin{aligned}&\| - B^TA^{-1}f + B^TA^{-1}B\bar{y}_k - \bar{r}_k^{(y)}\| \\ &\leq \frac{[(2k+1)\tau+O(u)]\kappa(A)}{1-\tau\kappa(A)}\|A^{-1}\|\|B\|(\|f\|+\|B\|\bar{Y}_k),\end{aligned}$$

where $\bar{Y}_k$ *is defined as a maximum norm over all computed approximate solutions* $\bar{Y}_k \equiv \max_{i=0,\ldots,k}\|\bar{y}_i\|$.

PROOF. The initial residual $\bar{r}_0^{(y)}$ is computed as $\bar{r}_0^{(y)} = -\mathrm{fl}(B^T\bar{x}_0)$, where $(A+\Delta A_0)\bar{x}_0 = \mathrm{fl}(f-By_0)$, $\|\Delta A_0\| \leq \tau\|A\|$. It is easy to see that the statement holds for $k=0$. The computed approximate solution $\bar{y}_{k+1}$ and the residual $\bar{r}_{k+1}^{(y)}$ satisfy

$$\begin{aligned}\bar{y}_{k+1} &= \bar{y}_k + \bar{\alpha}_k\bar{p}_k^{(y)} + \Delta y_{k+1}, \\ \|\Delta y_{k+1}\| &\leq u\|\bar{y}_k\| + (2u+u^2)\|\bar{\alpha}_k\bar{p}_k^{(y)}\|,\end{aligned} \tag{3.14}$$

$$\begin{aligned}\bar{r}_{k+1}^{(y)} &= \bar{r}_k^{(y)} - \bar{\alpha}_k B^T\bar{p}_k^{(x)} + \Delta r_{k+1}^{(y)}, \\ \|\Delta r_{k+1}^{(y)}\| &\leq u\|\bar{r}_k^{(y)}\| + O(u)\|B\|\|\bar{\alpha}_k\bar{p}_k^{(x)}\|,\end{aligned} \tag{3.15}$$

where $\bar{p}_k^{(x)}$ is the exact solution of the perturbed system

$$(A+\Delta A_k)\bar{p}_k^{(x)} = -\mathrm{fl}(B\bar{p}_k^{(y)}),\ \|\Delta A_k\| \leq \tau\|A\|. \tag{3.16}$$

Multiplying (3.14) by $B^TA^{-1}B$, substituting (3.16) into the recurrence (3.15), and subtracting these two equations we get the recurrence

$$\begin{aligned}-B^TA^{-1}f + B^TA^{-1}B\bar{y}_{k+1} - \bar{r}_{k+1}^{(y)} &= -B^TA^{-1}f + B^TA^{-1}B\bar{y}_k - \bar{r}_k^{(y)} \\ -\bar{\alpha}_k(B^T\bar{p}_k^{(x)} + B^TA^{-1}B\bar{p}_k^{(y)}) &+ B^TA^{-1}B\Delta y_k - \Delta r_k^{(y)}.\end{aligned}$$

The norm of the vector $\bar{\alpha}_k\bar{p}_k^{(y)}$ can be bounded as $\|\bar{\alpha}_k\bar{p}_k^{(y)}\| \leq \|\bar{y}_{k+1}\| + \|\bar{y}_k\| + \|\Delta y_{k+1}\|$. This bound in combination with (3.14) gives $\|\Delta y_{k+1}\| \leq O(u)\bar{Y}_{k+1}$ and $\|\bar{\alpha}_k\bar{p}_k^{(y)}\| \leq 3\bar{Y}_{k+1}$ which also implies

$$\|\bar{\alpha}_k\bar{p}_k^{(x)}\| \leq \frac{3\|A^{-1}\|}{1-\tau\kappa(A)}\|B\|\bar{Y}_{k+1}. \tag{3.17}$$

Using (3.16), the bound on $\|\bar{\alpha}_k \bar{p}_k^{(y)}\|$, and some elementary manipulation, we can estimate the term $\bar{\alpha}_k(B^T \bar{p}_k^{(x)} + B^T A^{-1} B \bar{p}_k^{(y)})$

$$\|\bar{\alpha}_k(B^T \bar{p}_k^{(x)} + B^T A^{-1} B \bar{p}_k^{(y)})\| \leq \|\bar{\alpha}_k B^T[(A + \Delta A_k)^{-1} - A^{-1}] \mathrm{fl}(B \bar{p}_k^{(y)})\|$$
$$+ \|\bar{\alpha}_k B^T A^{-1}[\mathrm{fl}(B \bar{p}_k^{(y)}) - B \bar{p}_k^{(y)}]\| \leq \frac{[\tau + O(u)]\kappa(A)}{1 - \tau\kappa(A)} \|A^{-1}\| \|B\|^2 \bar{Y}_{k+1}.$$

Considering (3.15), (3.17), and the induction assumption on the gap between $-B^T A^{-1} f + B^T A^{-1} B \bar{y}_k$ and $\bar{r}_k^{(y)}$ (similar to the one used in [**47**]), we obtain the bound for the error vector $\Delta r_{k+1}^{(y)}$ in the form

$$\|\Delta r_{k+1}^{(y)}\| \leq \frac{O(u)\kappa(A)}{1 - \tau\kappa(A)} \|A^{-1}\| \|B\| (\|f\| + \|B\| \bar{Y}_{k+1})$$

which proves the statement of the theorem. □

It is a well-known fact that the residual $\bar{r}_k^{(y)}$ computed recursively via (3.9) usually converges far below $O(u)$. Using this assumption we can obtain from the estimate for the gap $-B^T A^{-1} f + B^T A^{-1} B \bar{y}_k - \bar{r}_k^{(y)}$ the estimate for the maximum attainable accuracy of the true residual $-B^T A^{-1} f + B^T A^{-1} B \bar{y}_k$ itself. Summarizing, while the updated residual $\bar{r}_k^{(y)}$ converges to zero the true residual stagnates at the level proportional to τ. This is also illustrated in our numerical example, where the Schur complement system $-B^T A^{-1} B y = -B^T A^{-1} f$ is solved using the steepest descent method with the initial approximation y_0 set to zero. In Figure 3.2 we show the relative norms of the true residual $-B^T A^{-1} f + B^T A^{-1} B \bar{y}_k$ (solid lines) and the updated residual $\bar{r}_k^{(y)}$ (dashed lines).

Similar to Greenbaum [**47**], we have shown that the gap between the true and updated residual is proportional to the maximum norm of approximate solutions computed during the whole iteration process. Since the Schur complement system is symmetric negative definite, the norm of the error or residual converges monotonically for the most iterative methods like the steepest descent, the conjugate gradient, conjugate residual method, or other error/residual minimizing methods or at least becomes orders of magnitude smaller than initial error/residual without exceeding this limit. In such cases, the quantity $\bar{Y}_k$ does not play an important role in the bound, and it can usually be replaced by $\|y_0\|$ or a small multiple of $\|y\|$. The situation is more complicated when A is nonsingular and nonsymmetric; see [**60**].

As we already noted, the main difference with respect to the analysis of Greenbaum is that the floating-point multiplication with the fixed A^{-1} is replaced by the step-dependent inexact solution of the system with A such that it can be interpreted as the exact application of the matrix $(A + \Delta A_k)^{-1}$, where the perturbation matrix ΔA_k changes at every step k. This concept is very similar to the notion of inexact Krylov subspace methods (see [**90**] or [**97**]), which, on the other hand, does not take into account the effects of rounding errors. The theory of Greenbaum [**47**] could be directly applied only if we have at each iteration $\|\mathrm{fl}(B^T A^{-1} Bx) - B^T A^{-1} Bx\| \leq O(u)\|A^{-1}\|\|B\|^2\|x\|$. Since in our idealized case $\mathrm{fl}(B^T A^{-1} Bx) = B^T (A + \Delta A_k)^{-1} Bx$ with $\|\Delta A_k\| \leq \tau \|A\|$, we have only

$$\|\mathrm{fl}(B^T A^{-1} Bx) - B^T A^{-1} Bx\| \leq \frac{\tau\kappa(A)}{1 - \tau\kappa(A)} \|A^{-1}\|\|B\|^2\|x\|.$$

This bound could be improved if we make a restriction and use a variable tolerance for inner systems. If we require that every inner system is solved so that the relative residual of its computed solution needs the tolerance τ, then every inexact application of the matrix $B^T A^{-1} B$ would satisfy the inequality

$$\|\mathrm{fl}(B^T A^{-1} Bx) - B^T A^{-1} Bx\| \leq \tau \|A^{-1}\|\|B\|^2\|x\|. \tag{3.18}$$

Then the whole outer process (3.8) and (3.9) together with (3.18) could be interpreted as a floating-point iteration with the roundoff unit equal to τ. The computation in this "extended" arithmetic would lead to

$$\| - B^T A^{-1} f + B^T A^{-1} B\bar{y}_k - \bar{r}_k^{(y)}\| \leq \frac{O(\tau)}{1 - \tau\kappa(A)} \|A^{-1}\|\|B\|^2(\|y\| + \bar{Y}_k).$$

A thorough rounding analysis of the block LU factorization has been given in [**26**] and further developed in the saddle point context in [**70**]. The approach was quite converse to the one used here. It is assumed that all inner systems are solved in a backward stable way and the accuracy of computed approximate solutions is estimated in terms of the user prescribed tolerance for the outer Schur complement system. Roughly speaking, the higher stopping tolerance η leads to the higher attainable accuracy of the true residuals $f - A\bar{x}_k - B\bar{y}_k$ and $-B^T\bar{x}_k$. This level is magnified by the quantities that play a similar role as the growth factor in the Gaussian elimination with partial pivoting (see, e.g., [**55**]). On the other hand, the parameter η giving the threshold for the backward error cannot be infinitely small. Theorem 3.1 actually gives its lower bound. Dividing the right-hand side by $\|A^{-1}\|B\|^2\|\bar{y}\|$ we end up with $\eta \geq O(u)\kappa(A)/(1 - O(u)\kappa(A))$.

In the following we will estimate the residuals $f - A\bar{x}_k - B\bar{y}_k$ and $-B^T\bar{x}_k$. We will show that these quantities depend on the actual implementation of the back-substitution formula for x_k and distinguish between three schemes (3.10), (3.11) and (3.12). No matter how we compute the approximations $\bar{x}_k$ and $\bar{y}_k$ it holds that

$$-B^TA^{-1}f + B^TA^{-1}B\bar{y}_k = -B^T\bar{x}_k - B^TA^{-1}(f - A\bar{x}_k - B\bar{y}_k), \qquad (3.19)$$

which gives the relation between the residual $-B^TA^{-1}f + B^TA^{-1}B\bar{y}_k$ in the Schur complement system and the residuals $f - A\bar{x}_k - B\bar{y}_k$ and $-B^T\bar{x}_k$ associated with the saddle point system (3.1). According to Theorem 3.1, $\| - B^TA^{-1}f + B^TA^{-1}B\bar{y}_k\|$ is ultimately $O(\tau)$. Then it is clear from (3.19) that both $f - A\bar{x}_k - B\bar{y}_k$ and $-B^T\bar{x}_k$ cannot be proportional to the roundoff unit u. We will show that, depending on the chosen back-substitution scheme, we can ensure either that $f - A\bar{x}_k - B\bar{y}_k = O(\tau)$ with $-B^T\bar{x}_k = O(u)$ (scheme A (3.10)), or that $f - A\bar{x}_k - B\bar{y}_k = O(u)$ with $-B^T\bar{x}_k = O(\tau)$ (scheme C (3.12)), while the most straightforward scheme B (3.11) leads to both $f - A\bar{x}_k - B\bar{y}_k = O(\tau)$ and $-B^T\bar{x}_k = O(\tau)$.

1.2. Scheme A: The updated approximate solution. In this subsection we analyze the generic update (3.10). It is clear that this scheme requires only one system solve with A per iteration. Indeed, we compute only the direction vector $p_k^{(x)} = -A^{-1}Bp_k^{(y)}$, which appears in the recurrence $r_{k+1}^{(y)} = r_k^{(y)} - \alpha_k B^T p_k^{(x)}$ anyway. As we will see, in finite precision arithmetic this algorithm guarantees that $-B^T\bar{x}_k$ will ultimately reach $O(u)$. This happens despite the fact that the systems with the matrix block A are computed inexactly with the parameter τ frequently much larger than $O(u)$.

THEOREM 3.2. *The true residual $f - A\bar{x}_k - B\bar{y}_k$ satisfies the bound*

$$\|f - A\bar{x}_k - B\bar{y}_k\| \leq O(u)(\|f\| + \|B\|\bar{Y}_k) + [(k+1)\tau + O(u)]\|A\|\bar{X}_k. \qquad (3.20)$$

The gap between the residuals $-B^T\bar{x}_k$ and $\bar{r}_k^{(y)}$ can be estimated as

$$\| - B^T\bar{x}_k - \bar{r}_k^{(y)}\| \leq O(u)\|A^{-1}\|\|B\|(\|f\| + \|A\|\bar{X}_k + \|B\|\bar{Y}_k),$$

where $\bar{X}_k$ is now defined as a maximum norm over all computed approximate solutions $\bar{X}_k \equiv \max_{i=0,\ldots,k}\|\bar{x}_i\|$.

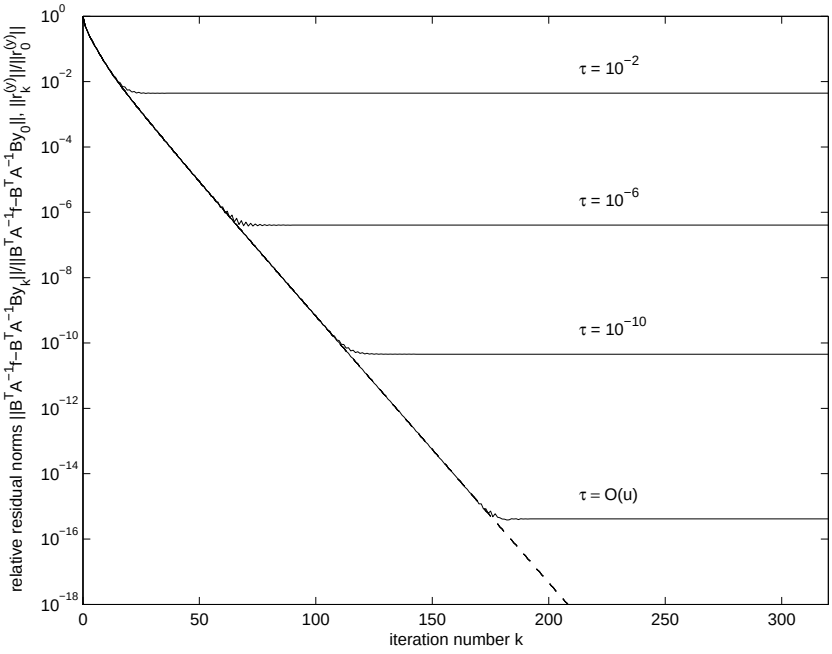


FIGURE 3.2. Schur complement reduction method: The relative norms of the true residual $-B^TA^{-1}f+B^TA^{-1}\bar{y}_k$ (solid lines) and the updated residual $\bar{r}_k^{(y)}$ (dashed lines) – the updated solution scheme (3.10).

PROOF. The computed approximate solution $\bar{x}_{k+1}$ satisfies

$$\begin{aligned} \bar{x}_{k+1} &= \bar{x}_k + \bar{\alpha}_k\bar{p}_k^{(x)} + \Delta x_{k+1}, \\ \|\Delta x_{k+1}\| &\leq u\|\bar{x}_k\| + (2u+u^2)\|\bar{\alpha}_k\bar{p}_k^{(x)}\|. \end{aligned} \tag{3.21}$$

Substituting recurrently (3.21) and (3.14) into the residual

$$\begin{aligned} f - A\bar{x}_{k+1} - B\bar{y}_{k+1} = f - A\bar{x}_k - B\bar{y}_k - \bar{\alpha}_k(A\bar{p}_k^{(x)} + B\bar{p}_k^{(y)}) \\ - A\Delta x_{k+1} - B\Delta y_{k+1}, \end{aligned}$$

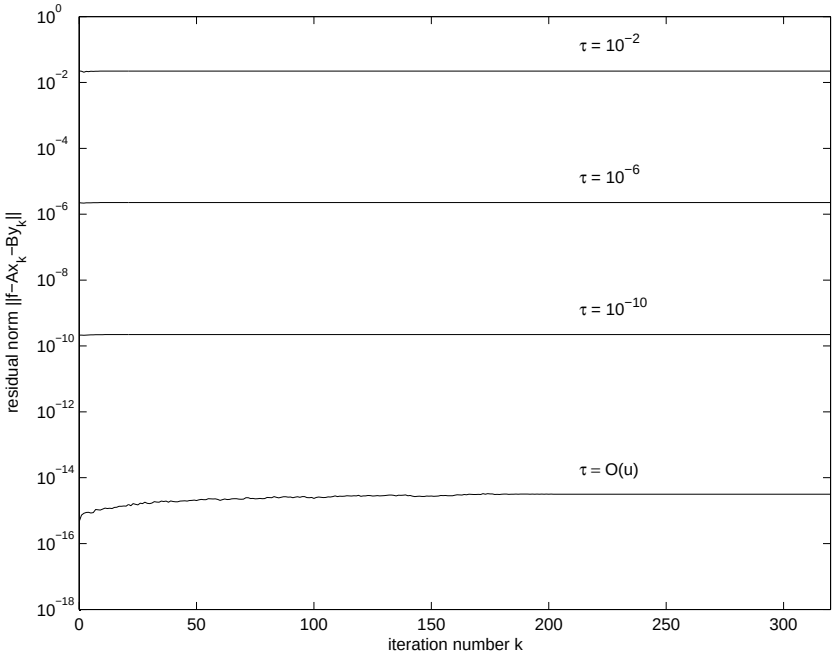


FIGURE 3.3. Schur complement reduction method: The norms of the true residual $f - A\bar{x}_k - B\bar{y}_k$ – the updated solution scheme (3.10).

we obtain the following bound:

$$\|f - A\bar{x}_k - B\bar{y}_k\| \leq \|f - A\bar{x}_0 - By_0\|$$
$$+ \sum_{i=0}^{k-1} \left(\|\bar{\alpha}_i(A\bar{p}_i^{(x)} + B\bar{p}_i^{(y)})\| + \|A\|\|\Delta x_{i+1}\| + \|B\|\|\Delta y_{i+1}\| \right).$$

Here we, in fact, reformulate the main result of Greenbaum [**47**, Theorem 2.2] and heavily use the fact that the vectors $\bar{p}_k^{(x)}$ satisfy the perturbed system (3.16). From Theorem 3.1 we have bounds $\|\Delta y_{k+1}\| \leq O(u)\bar{Y}_{k+1}$ and $\|\bar{\alpha}_k\bar{p}_k^{(y)}\| \leq 3\bar{Y}_{k+1}$ which also imply the bound (3.17). Using all of these results we get

$$\|\bar{\alpha}_k(A\bar{p}_k^{(x)} + B\bar{p}_k^{(y)})\| \leq \|\bar{\alpha}_k[fl(B\bar{p}_k^{(y)}) - B\bar{p}_k^{(y)}]\| + \|\Delta A_k\|\|\bar{\alpha}_k\bar{p}_k^{(x)}\|.$$

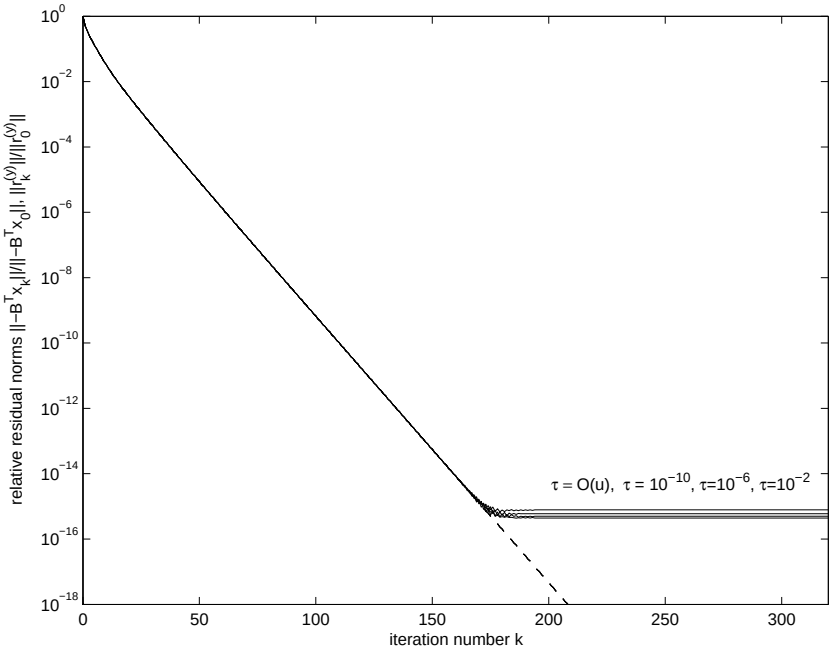


FIGURE 3.4. Schur complement reduction method: The relative norms of the true residual $-B^T\bar{x}_k$ (solid lines) and the recursively computed residual $\bar{r}_k^{(y)}$ (dashed lines) – the updated solution scheme (3.10).

Further we use $\|\Delta x_{k+1}\| \leq O(u)\bar{X}_{k+1}$ and $\|\bar{\alpha}_k\bar{p}_k^{(x)}\| \leq 3\bar{X}_{k+1}$. Summarizing, we get the first result. The gap between $-B^T\bar{x}_{k+1}$ and $\bar{r}_{k+1}^{(y)}$ is equal to

$$-B^T\bar{x}_{k+1} - \bar{r}_{k+1}^{(y)} = -B^T\bar{x}_k - \bar{r}_k^{(y)} - B^T\Delta x_{k+1} - \Delta r_{k+1}^{(y)}$$

and it leads to the expansion containing just the local errors Δx_{i+1}, Δy_{i+1} and the initial gap $-B^T\bar{x}_0 - \bar{r}_0^{(y)}$

$$-B^T\bar{x}_k - \bar{r}_k^{(y)} = -B^T\bar{x}_0 - \bar{r}_0^{(y)} - \sum_{i=0}^{k-1} B^T\Delta x_{i+1} - \sum_{i=0}^{k-1}\Delta r_{k+1}^{(y)}.$$

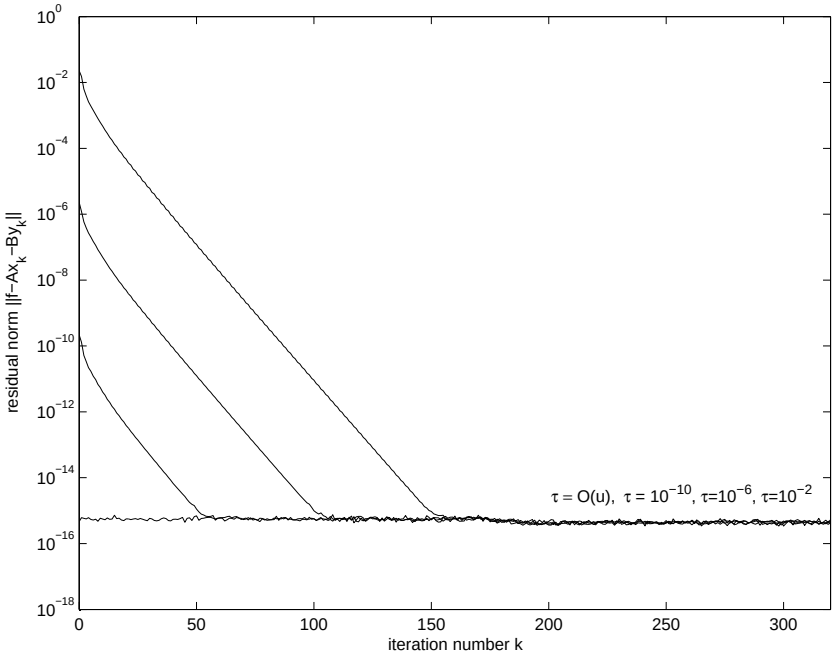


FIGURE 3.5. Schur complement reduction method: The norms of the true residual $f - A\bar{x}_k - B\bar{y}_k$ – the corrected direct substitution scheme (3.12).

Taking norms, considering the bounds on $\|\Delta x_{k+1}\|$, $\|\Delta y_{k+1}\|$, (3.15), and the relation $\bar{r}_0^{(y)} = -\mathrm{fl}(B^T\bar{x}_0)$, we get the second result. □

COROLLARY 3.3. *The true residual $f - A\bar{x}_k - B\bar{y}_k$ satisfies the bound*

$$\|f - A\bar{x}_k - B\bar{y}_k\| \leq \frac{O(\tau)\kappa(A)}{1 - \tau\kappa(A)}(\|f\| + \|B\|\bar{Y}_k).$$

The gap between the residuals $-B^T\bar{x}_k$ and $\bar{r}_k^{(y)}$ can be estimated as

$$\| - B^T\bar{x}_k - \bar{r}_k^{(y)}\| \leq \frac{O(u)\kappa(A)}{1 - \tau\kappa(A)}\|A^{-1}\|\|B\|(\|f\| + \|B\|\bar{Y}_k).$$

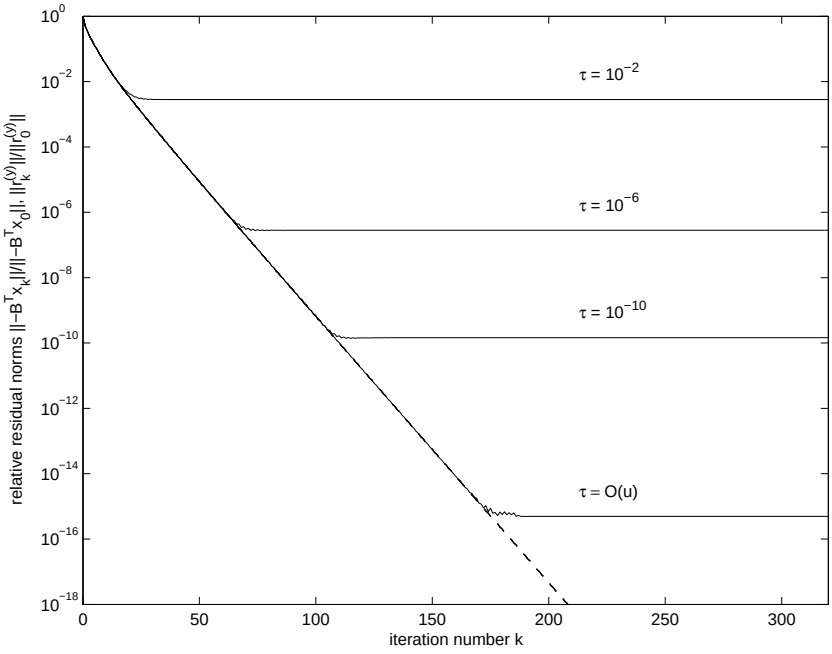


FIGURE 3.6. Schur complement reduction method: The relative norms of the true residual $-B^T\bar{x}_k$ (solid lines) and the recursively computed residual $\bar{r}_k^{(y)}$ (dashed lines) – the direct substitution scheme (3.11).

As we will see in the next subsection, the bound for the gap $-B^T\bar{x}_k - \bar{r}_k^{(y)}$ is considerably better than for the scheme (3.11). In contrast to (3.24), it does not depend on τ. Provided that $\bar{r}_k^{(y)}$ converges to zero, the true residual $-B^T\bar{x}_k$ will stagnate at the level proportional to u and the second block equation of (3.1) will be satisfied to working accuracy.

Figs. 3.3 and 3.4 show the norms of the true residual $f - A\bar{x}_k - B\bar{y}_k$ and $-B^T\bar{x}_k$ (solid lines), respectively, including the norms of the updated residual $\bar{r}_k^{(y)}$ (dashed lines). The numerical results are in good agreement with Theorem 3.2. The residual $f - A\bar{x}_k - B\bar{y}_k$ is growing slightly due to the accumulation of

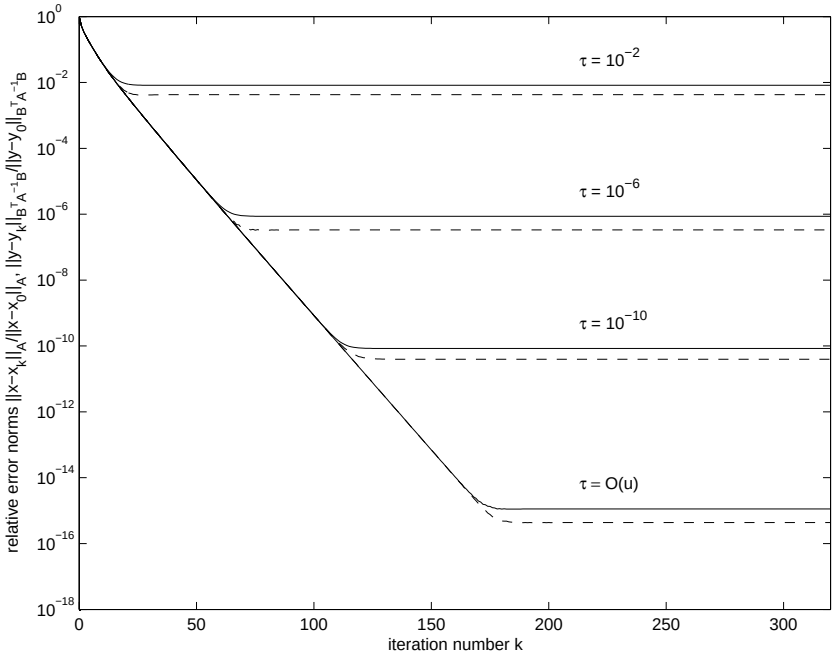


FIGURE 3.7. Schur complement reduction method: The relative error norms $\|x - \bar{x}_k\|_A / \|x - \bar{x}_0\|_A$ (solid lines) and $\|y - \bar{y}_k\|_{B^T A^{-1} B} / \|y - y_0\|_{B^T A^{-1} B}$ (dashed lines) – the updated solution scheme (3.10).

errors in inner systems $Ap_k^{(x)} = -Bp_k^{(y)}$ but it essentially remains on the level proportional to τ. The residual $-B^T \bar{x}_k$ ultimately stagnates at $O(u)$. The formula (3.10) is suitable whenever the second block equation of (3.1) must be satisfied accurately, no matter how small or big the inner tolerance τ is.

1.3. Scheme B: The approximate solution computed by a direct substitution. In this subsection we assume that x_k is computed by the direct substitution (3.11). The computed $\bar{x}_k$ then satisfies the equality

$$(A + \Delta A_k)\bar{x}_k = \mathrm{fl}(f - B\bar{y}_k), \ \|\Delta A_k\| \leq \tau \|A\|. \tag{3.22}$$

The perturbation matrices ΔA_k are different from those defined in Subsection 1.1, but for simplicity we will keep the same notation. In the following we will show that the residual $\bar{r}_k^{(y)}$ is a good approximation for the residual $-B^T\bar{x}_k$, provided that they are above the level given by the bound for $-B^T\bar{x}_k - \bar{r}_k^{(y)}$. This quantity is now, however, proportional to τ.

THEOREM 3.4. *The true residual $f - A\bar{x}_k - B\bar{y}_k$ satisfies the bound*

$$\|f - A\bar{x}_k - B\bar{y}_k\| \leq O(u)(\|f\| + \|B\|\|\bar{y}_k\|) + \tau\|A\|\|\bar{x}_k\|. \tag{3.23}$$

The gap between the residuals $-B^T\bar{x}_k$ and $\bar{r}_k^{(y)}$ can be bounded as follows:

$$\begin{aligned} \| - B^T\bar{x}_k - \bar{r}_k^{(y)}\| \leq\ & O(u)\|A^{-1}\|\|B\|(\|f\| + \|B\|\bar{Y}_k) \\ & + [(k+3)\tau + O(u)]\kappa(A)\|B\|\bar{X}_k, \end{aligned} \tag{3.24}$$

where $\bar{X}_k$ is defined as $\bar{X}_k \equiv \max_{i=0,\ldots,k-1}\{\|\bar{x}_0\|, \|\bar{x}_k\|, \|\bar{\alpha}_i\bar{p}_i^{(x)}\|\}$.

PROOF. The first result follows from (3.22) and the relation for the true residual

$$f - A\bar{x}_k - B\bar{y}_k = f - B\bar{y}_k - \mathrm{fl}(f - B\bar{y}_k) - \Delta A_k\bar{x}_k.$$

For the gap between $-B^T\bar{x}_k$ and $\bar{r}_k^{(y)}$ we have the identity

$$\begin{aligned} -B^T\bar{x}_k - \bar{r}_k^{(y)} =\ & -B^TA^{-1}f + B^TA^{-1}B\bar{y}_k - \bar{r}_k^{(y)} + B^TA^{-1}\Delta A_k\bar{x}_k \\ & + B^TA^{-1}[\mathrm{fl}(f - B\bar{y}_k) - (f - B\bar{y}_k)]. \end{aligned} \tag{3.25}$$

The statement of Theorem 3.1 together with (3.25) gives the second result (3.24). □

COROLLARY 3.5. *The true residual $f - A\bar{x}_k - B\bar{y}_k$ satisfies the bound*

$$\|f - A\bar{x}_k - B\bar{y}_k\| \leq \frac{O(\tau)\kappa(A)}{1 - \tau\kappa(A)}(\|f\| + \|B\|\|\bar{y}_k\|).$$

The gap between the residuals $-B^T\bar{x}_k$ and $\bar{r}_k^{(y)}$ can be bounded as follows

$$\| - B^T\bar{x}_k - \bar{r}_k^{(y)}\| \leq \frac{O(\tau)\kappa(A)}{1 - \tau\kappa(A)}\|A^{-1}\|\|B\|(\|f\| + \|B\|\bar{Y}_k). \tag{3.26}$$

Indeed while the residual $\bar{r}_k^{(y)}$ converges ultimately below $O(u)$, the residual $-B^T\bar{x}_k$ will remain proportional to τ. The norm of $f - A\bar{x}_k - B\bar{y}_k$ is unconditionally bounded by the term proportional to τ dominating other terms in (3.23).

Figure 3.6 shows the norms of $-B^T\bar{x}_k$ (solid lines) and $\bar{r}_k^{(y)}$ (dashed lines). The residual $f - A\bar{x}_k - B\bar{y}_k$ behaves similarly to that of the scheme (3.10) shown in plot 3.3. The residual $f - A\bar{x}_k - B\bar{y}_k$ remains almost constant since it is nothing but the residual of the system $Ax_k = f - By_k$ solved in each iteration with the uniform accuracy.

1.4. Scheme C: The approximate solution computed with a corrected direct substitution. The third back-substitution formula (3.12) can be derived by a correction of the scheme (3.11) and requires two system solves with A. In this subsection we show that its numerical behavior is very similar to the behavior of classical nonstationary iterative methods described and analyzed by Higham [**55**]. We prove that under certain conditions the true residual $f - A\bar{x}_k - B\bar{y}_k$ ultimately converges to the level proportional to u, which is significantly smaller than those for the previous two schemes.

THEOREM 3.6. *Assuming for sufficiently large k with $\|\bar{y}_{k+1} - \bar{y}_k\| \le O(u)\bar{Y}_{k+1}$, there exists a step k_0 such that the true residual $f - A\bar{x}_k - B\bar{y}_k$ is bounded by*

$$\|f - A\bar{x}_k - B\bar{y}_k\| \le O(u)(\|f\| + \|A\|\bar{X}_k + \|B\|\bar{Y}_k) \tag{3.27}$$

for all steps $k \ge k_0$. The gap between $-B^T\bar{x}_k$ and $\bar{r}_k^{(y)}$ can be estimated as follows:

$$\begin{aligned}\| - B^T\bar{x}_k - \bar{r}_k^{(y)}\| \le\ & O(u)\|A^{-1}\|\|B\|(\|f\| + \|B\|\bar{Y}_k) \\ & + [(k+3)\tau + O(u)]\kappa(A)\|B\|\bar{X}_k.\end{aligned}$$

The quantity $\bar{X}_k$ is here defined as $\bar{X}_k \equiv \max_{i=0,\ldots,k-1}\{\|\bar{x}_0\|, \|\bar{x}_k\|, \|\bar{\alpha}_i\bar{p}_i^{(x)}\|\}$.

PROOF. The computed approximate solution $\bar{x}_{k+1}$ satisfies

$$\bar{x}_{k+1} = \bar{x}_k + \bar{u}_k + \Delta x_{k+1},\ \|\Delta x_{k+1}\| \le u(\|\bar{x}_k\| + \|\bar{u}_k\|), \tag{3.28}$$

where the vector $\bar{u}_k$ is the exact solution of the system

$$(A + \Delta A_{k+1})\bar{u}_k = \mathrm{fl}(f - A\bar{x}_k - B\bar{y}_{k+1}),\ \|\Delta A_{k+1}\| \le \tau\|A\|. \tag{3.29}$$

The residual $f - A\bar{x}_{k+1} - B\bar{y}_{k+1}$ can be expressed using (3.28) and (3.29) as

$$\begin{aligned}f - A\bar{x}_{k+1} - B\bar{y}_{k+1} &= \Delta A_{k+1}\bar{u}_k - A\Delta x_{k+1} \\ &\quad + \mathrm{fl}(f - A\bar{x}_k - B\bar{y}_{k+1}) - (f - A\bar{x}_k - B\bar{y}_{k+1}) \\ &= G_{k+1}(f - A\bar{x}_k - B\bar{y}_k) - G_{k+1}B(\bar{\alpha}_k\bar{p}_k^{(y)}) + h_{k+1},\end{aligned} \tag{3.30}$$

where $G_{k+1} \equiv \Delta A_{k+1}(A + \Delta A_{k+1})^{-1}$ and $h_{k+1} \equiv (I + G_{k+1})[\mathrm{fl}(f - A\bar{x}_k - B\bar{y}_{k+1}) - (f - A\bar{x}_k - B\bar{y}_{k+1})] - A\Delta x_{k+1} - G_{k+1}B\Delta y_{k+1}$. From a recursive use of the formula (3.30) we obtain

$$\begin{aligned} f - A\bar{x}_k - B\bar{y}_k = {} & G_k \cdots G_1(f - A\bar{x}_0 - By_0) \\ & - \sum_{i=0}^{k-1} G_k \cdots G_{i+2}(G_{i+1}B\bar{\alpha}_i\bar{p}_i^{(y)} - h_{i+1}). \end{aligned}$$

Taking norms, using the relation $\|\bar{\alpha}_i\bar{p}_i^{(y)}\| \leq \|\bar{y}_{i+1} - \bar{y}_i\| + \|\Delta y_{i+1}\|$ and $\|\Delta A_i\| \leq \tau\|A\|$ we obtain the uniform bound $\|G_i\| \leq \tau\kappa(A)[1 - \tau\kappa(A)]^{-1} < 1$. This leads to the inequality

$$\begin{aligned} \|f - A\bar{x}_k - B\bar{y}_k\| \leq {} & \left(\frac{\tau\kappa(A)}{1 - \tau\kappa(A)}\right)^k \|f - A\bar{x}_0 - By_0\| \\ & + \sum_{i=0}^{k-1} \left(\frac{\tau\kappa(A)}{1 - \tau\kappa(A)}\right)^{k-i} \|B\|\|\bar{y}_{i+1} - \bar{y}_i\| \\ & + k \max_{i=0,\ldots,k-1} \|h_{i+1}\| + k \max_{i=0,\ldots,k-1} \|B\|\|\Delta y_{i+1}\|. \end{aligned} \tag{3.31}$$

For the vector h_{k+1} it further follows that

$$\|h_{k+1}\| \leq O(u)[\|f\| + \|A\|(\|\bar{x}_{k+1}\| + \|\bar{x}_k\|) + \|B\|\bar{Y}_{k+1}].$$

It is easy to see that for sufficiently large k the first term on the right-hand side of (3.31) will decrease far below $O(u)$, while the second term will be at most $O(u)\|B\|\bar{Y}_{k+1}$ for all steps k starting from some index k_0. Summarizing, for sufficiently large $k \geq k_0$ we have the bound

$$\|f - A\bar{x}_k - B\bar{y}_k\| \leq O(u)[\|f\| + \|A\|(\|\bar{x}_{k+1}\| + \|\bar{x}_k\|) + \|B\|\bar{Y}_k].$$

The second statement can be proved considering

$$\begin{aligned} -B^T\bar{x}_{k+1} - \bar{r}_{k+1}^{(y)} = {} & -B^TA^{-1}f + B^TA^{-1}B\bar{y}_{k+1} - \bar{r}_{k+1}^{(y)} \\ & - B^T[(A + \Delta A_{k+1})^{-1} - A^{-1}]\mathrm{fl}(f - A\bar{x}_k - B\bar{y}_{k+1}) \\ & - B^TA^{-1}[\mathrm{fl}(f - A\bar{x}_k - B\bar{y}_{k+1}) - (f - A\bar{x}_k - B\bar{y}_{k+1})]. \end{aligned}$$

The first term on the right-hand side can be estimated using Theorem 3.1. Based on (3.29) we have

$$\|[(A + \Delta A_{k+1})^{-1} - A^{-1}]\mathrm{fl}(f - A\bar{x}_k - B\bar{y}_{k+1})\| \leq \frac{\tau\kappa(A)}{1 - \tau\kappa(A)}\|\bar{u}_k\|$$

which together with the bound on $\|\bar{u}_k\|$ completes the proof. □

COROLLARY 3.7. *Assuming for sufficiently large k with $\|\bar{y}_{k+1}-\bar{y}_k\| \le O(u)\bar{Y}_{k+1}$, there exists a step k_0 such that the true residual $f - A\bar{x}_k - B\bar{y}_k$ is bounded by*

$$\|f - A\bar{x}_k - B\bar{y}_k\| \le \frac{O(u)\kappa(A)}{1-\tau\kappa(A)}(\|f\| + \|B\|\bar{Y}_k^{(k_0)})$$

for all steps $k \ge k_0$. The quantity $\bar{Y}_k^{(k_0)}$ is defined as $\bar{Y}_k^{(k_0)} \equiv \max_{i=k_0,\ldots,k} \|\bar{y}_i\|$. The gap between $-B^T\bar{x}_k$ and $\bar{r}_k^{(y)}$ can be estimated as follows

$$\| - B^T\bar{x}_k - \bar{r}_k^{(y)}\| \le \frac{O(u)\kappa(A)}{1-\tau\kappa(A)}\|A^{-1}\|\|B\|(\|f\| + \|B\|\bar{Y}_k).$$

In Theorem 3.6, we assume that $\bar{y}_k$ ultimately stagnate so that $\|\bar{y}_{k+1} - \bar{y}_k\| \le O(u)\bar{Y}_{k+1}$ for sufficiently large $k \ge k_0$. It appears that this condition does not represent a serious restriction. Using (3.14) we have $\|\bar{y}_{k+1} - \bar{y}_k\| \le \|\bar{\alpha}_k\bar{p}_k^{(y)}\| + O(u)\bar{Y}_{k+1}$. We will show that the norm of $\bar{\alpha}_k\bar{p}_k^{(y)}$ is much smaller than u for large k, i.e., we can absorb it into the term $O(u)\bar{Y}_{k+1}$. Denoting $\hat{S}_k \equiv B^T(A + \Delta A_k)^{-1}B$, using (3.15) and (3.16) we have the bound

$$\|\bar{\alpha}_k\bar{p}_k^{(y)}\| \le 2\|\hat{S}_k^{-1}\|(\|\bar{r}_{k+1}^{(y)}\| + \|\bar{r}_k^{(y)}\|) + O(u)\|\hat{S}_k^{-1}\|\|(A+\Delta A_k)^{-1}\|\|B\|^2\|\bar{\alpha}_k\bar{p}_k^{(y)}\|.$$

Provided that $O(u)\|\hat{S}_k^{-1}\|\|(A+\Delta A_k)^{-1}\|\|B\|^2 < 1$, we obtain

$$\|\bar{\alpha}_k\bar{p}_k^{(y)}\| \le \frac{2\|\hat{S}_k^{-1}\|(\|\bar{r}_{k+1}^{(y)}\| + \|\bar{r}_k^{(y)}\|)}{1 - O(u)\|\hat{S}_k^{-1}\|\|(A+\Delta A_k)^{-1}\|\|B\|^2}.$$

Since the norms of updated residuals decrease far below the roundoff unit, the assumption on $\|\bar{y}_{k+1} - \bar{y}_k\|$ will be true for sufficiently large k. Note that $O(u)\|\hat{S}_k^{-1}\|\|(A+\Delta A_k)^{-1}\|\|B\|^2 < 1$ is nothing but the restricted assumption of numerical nonsingularity of the Schur complement matrix $B^TA^{-1}B$.

The bound (3.27) is significantly better than its counterparts (3.20) and (3.23). Theorem 3.6 describes that the residual $f - A\bar{x}_k - B\bar{y}_k$ will ultimately reach the roundoff unit level provided that the matrix $G_kG_{k-1}\cdots G_1$ converges to zero for $k \to \infty$. As soon as iterates $\bar{y}_k$ start to stagnate at their limiting accuracy level, the rate of convergence of this nonstationary iteration process is bounded by the factor $\tau\kappa(A)[1-\tau\kappa(A)]^{-1}$. The behavior of $-B^T\bar{x}_k$ is similar to that of scheme (3.11). Indeed, when $\bar{r}_k^{(y)}$ converges ultimately below $O(u)$, the residual $-B^T\bar{x}_k$ remains proportional to τ. Figure 3.5 shows the norms of the residual $f - A\bar{x}_k - B\bar{y}_k$. The plot for $-B^T\bar{x}_k$ (not reported) is similar to the plot (d) for the scheme (3.11). It is clear that in our well-conditioned case the stationary

method converges very fast and the rate of decrease of $f - A\bar{x}_k - B\bar{y}_k$ is essentially comparable to the convergence rate of the outer iteration.

1.5. Forward error analysis. In this subsection we estimate the maximum attainable accuracy in terms of the errors $x - \bar{x}_k$ and $y - \bar{y}_k$. First we formulate the bounds in the 2-norm, then in the A-norm of the error $x - \bar{x}_k$, and then in the $B^T A^{-1} B$-norm of the error $y - \bar{y}_k$. The errors $x - \bar{x}_k$ and $y - \bar{y}_k$, and the residuals $f - A\bar{x}_k - B\bar{y}_k$ and $-B^T\bar{x}_k$ satisfy

$$\begin{pmatrix} A & B \\ B^T & 0 \end{pmatrix} \begin{pmatrix} x - \bar{x}_k \\ y - \bar{y}_k \end{pmatrix} = \begin{pmatrix} f - A\bar{x}_k - B\bar{y}_k \\ -B^T\bar{x}_k \end{pmatrix}. \tag{3.32}$$

We have the explicit expression for the inverse of the saddle point matrix

$$\begin{pmatrix} A & B \\ B^T & 0 \end{pmatrix}^{-1} = \begin{pmatrix} (I - \Pi)A^{-1} & -\Pi B(B^T B)^{-1} \\ -(B^T B)^{-1} B^T \Pi^T & -(B^T A^{-1} B)^{-1} \end{pmatrix},$$

where $\Pi \equiv A^{-1}B(B^T A^{-1} B)^{-1} B^T$ represents the oblique projector onto $R(B)$ along $N(B^T)$. Considering (3.32), the inequalities

$$\|(I - \Pi)A^{-1}\| = \|A^{-1/2}(I - A^{-1/2}B(B^T A^{-1} B)^{-1} B^T A^{-1/2})A^{-1/2}\| \leq \sigma_{min}^{-1}(A)$$

and

$$\begin{aligned} \|\Pi B^T (B^T B)^{-1}\| &= \|A^{-1/2}(A^{-1/2}B(B^T A^{-1} B)^{-1} A^{-1/2})A^{1/2} B(B^T B)^{-1}\| \\ &\leq \kappa^{1/2}(A)\sigma_{min}^{-1}(B), \end{aligned}$$

(note that $A^{-1/2}B(B^T A^{-1} B)B^T A^{-1/2}$ is the orthogonal projector onto the range of $R(A^{-1/2}B)$, we obtain the bounds

$$\|x - \bar{x}_k\| \leq \gamma_1 \|f - A\bar{x}_k - B\bar{y}_k\| + \gamma_2 \|- B^T\bar{x}_k\|, \tag{3.33}$$

$$\|y - \bar{y}_k\| \leq \gamma_2 \|f - A\bar{x}_k - B\bar{y}_k\| + \gamma_3 \|- B^T\bar{x}_k\|, \tag{3.34}$$

where $\gamma_1 \equiv \sigma_{min}^{-1}(A)$, $\gamma_2 \equiv \kappa^{1/2}(A)\sigma_{min}^{-1}(B)$, and $\gamma_3 \equiv \sigma_{min}^{-1}(B^T A^{-1} B)$ are constants independent of the iteration step k. It is clear from (3.33), (3.34), and Theorems 3.2, 3.4 and 3.6 that $\|x - \bar{x}_k\|$ and $\|y - \bar{y}_k\|$ will be $O(\tau)$ for all back-substitution schemes. In contrast to our numerical example, the saddle point systems that arise in practice can be ill-conditioned. In such cases the constants γ_1, γ_2, and γ_3 may play an important role.

In exact arithmetic we have $\|x - x_k\|_A = \|y - y_k\|_{B^T A^{-1} B}$. Since in finite precision arithmetic the residual $f - A\bar{x}_k - B\bar{y}_k$ is no longer zero, instead of this identity we get

$$\left| \|x - \bar{x}_k\|_A - \|y - \bar{y}_k\|_{B^T A^{-1} B} \right| \leq \gamma_1^{1/2} \|f - A\bar{x}_k - B\bar{y}_k\|. \tag{3.35}$$

We can also formulate the proposition, which gives bounds for the errors in terms of the residuals $f - A\bar{x}_k - B\bar{y}_k$ and $-B^T A^{-1} f + B^T A^{-1} B\bar{y}_k$.

THEOREM 3.8. *The A-norm of the error $x - \bar{x}_k$ and the $B^T A^{-1} B$-norm of the error $y - \bar{y}_k$ can be bounded as*

$$\|x - \bar{x}_k\|_A \leq \gamma_1^{1/2} \|f - A\bar{x}_k - B\bar{y}_k\| + \gamma_3^{1/2} \| - B^T A^{-1} f + B^T A^{-1} B\bar{y}_k\|, \quad (3.36)$$

$$\|y - \bar{y}_k\|_{B^T A^{-1} B} \leq \gamma_3^{1/2} \| - B^T A^{-1} f + B^T A^{-1} B\bar{y}_k\|. \quad (3.37)$$

PROOF. It follows from (3.35) that

$$\begin{aligned} \|x - \bar{x}_k\|_A &\leq \|y - \bar{y}_k\|_{B^T A^{-1} B} + |\|x - \bar{x}_k\|_A - \|y - \bar{y}_k\|_{B^T A^{-1} B}| \\ &\leq \|y - \bar{y}_k\|_{B^T A^{-1} B} + \sigma_{min}^{-1/2}(A) \|f - A\bar{x}_k - B\bar{y}_k\|. \end{aligned} \quad (3.38)$$

For the $B^T A^{-1} B$-norm of the error $y - \bar{y}_k$ we have

$$\|y - \bar{y}_k\|_{B^T A^{-1} B} = \|B^T A^{-1} f - B^T A^{-1} B\bar{y}_k\|_{(B^T A^{-1} B)^{-1}}, \quad (3.39)$$

which completes the proof. □

The first term on the right-hand side of (3.36) should be zero in exact arithmetic and it describes how well the computed $\bar{x}_k$ and $\bar{y}_k$ satisfy (3.7). The second term is related to the Schur complement residual which in exact arithmetic should converge to zero. The recursively computed residual $\bar{r}_k^{(y)}$ is a good approximation to $-B^T A^{-1} f + B^T A^{-1} B\bar{y}_k$, provided they are above the level given by Theorem 3.1. Therefore its norm represents an easily computable quantity for the second term on the right-hand side of (3.36). The residual $f - A\bar{x}_k - B\bar{y}_k$ depends on the computed $\bar{x}_k$ and we distinguish between three schemes with (3.10), (3.11) or (3.12), respectively. We can see that, no matter which implementation we use, $-B^T A^{-1} f + B^T A^{-1} B\bar{y}_k$ is a dominating quantity in (3.36). Therefore, $\|x - \bar{x}_k\|_A$ can be thus well approximated during the convergence by the quantity $\gamma_3^{1/2} \|\bar{r}_k^{(y)}\|$ or its estimate. Similar can be said also for $\|y - \bar{y}_k\|_{B^T A^{-1} B}$, see (3.37).

The errors $x - \bar{x}_k$ and $y - \bar{y}_k$ can be estimated with more sophisticated but easily computable bounds (without explicit use of residuals and conditioning). As an example we refer to the rounding error analysis of the conjugate gradient method and various mathematically equivalent formulas for estimating $\|x - \bar{x}_k\|_A$ [**96**]. It appears that although many existing bounds were developed using exact arithmetic considerations, they estimate successfully the energy error using computed quantities which can be orders of magnitude different from their exact precision

counterparts. Therefore despite that we assume that A^{-1} is performed inexactly, it is feasible to estimate the $B^TA^{-1}B$-norm of the error $y - \bar{y}_k$.

In Figure 3.7 we report the relative error norms $\|x - \bar{x}_k\|_A / \|x - \bar{x}_0\|_A$ and $\|y - \bar{y}_k\|_{B^TA^{-1}B} / \|y - y_0\|_{B^TA^{-1}B}$. The inverse of A in the computation of $B^TA^{-1}B$-norm is computed by a direct solver. In agreement with (3.36) and (3.37) and Theorems 3.2, 3.4 and 3.6 (see also Figures 3.3-3.6), the relative A-norm of the error $x - \bar{x}_k$ and also the relative $B^TA^{-1}B$-norm of the error $y - \bar{y}_k$ begin to stagnate at the level proportional to τ. Since the behavior of these quantities for all implementations is similar, we present only the results for the scheme (3.11). The slight difference is visible only in the gap between both error norms given by the estimate (3.35).

2. Null-space projection method

In this section we deal with algorithms which compute approximations x_k and y_k such that x_k satisfies $B^Tx_k = 0$ and y_k solves the least squares problem minimizing the residual $f - Ax_k - By_k$, i.e.,

$$\|f - Ax_k - By_k\| = \min_{v \in \mathbb{R}^m} \|f - Ax_k - Bv\|. \tag{3.40}$$

We will denote (3.40) by $By_k \approx f - Ax_k$ and assume that the approximate solution x_{k+1} and the residual vector $r_{k+1}^{(x)}$ are computed using

$$x_{k+1} = x_k + \alpha_k p_k^{(x)}, \tag{3.41}$$

$$r_{k+1}^{(x)} = r_k^{(x)} - \alpha_k A p_k^{(x)} - B p_k^{(y)}, \tag{3.42}$$

where $r_0^{(x)} = B^\dagger(f - Ax_0)$. The vectors x_0 and $p_k^{(x)}$ belong to $N(B^T)$ and $p_k^{(y)}$ solves the problem $Bp_k^{(y)} \approx r_k^{(x)} - \alpha_k A p_k^{(x)}$ minimizing the residual

$$\|r_k^{(x)} - \alpha_k A p_k^{(x)} - B p_k^{(y)}\| = \min_{p \in \mathbb{R}^m} \|r_k^{(x)} - \alpha_k A p_k^{(x)} - Bp\|.$$

This residual update strategy was proposed in [**44**] (see also [**21, 20**]) and is used to reduce the roundoff errors in the projection onto $N(B^T)$. Note that the vectors $p_k^{(y)}$ can be, with no additional cost, used as direction vectors for computing the approximate solution y_{k+1}. Again we will distinguish between

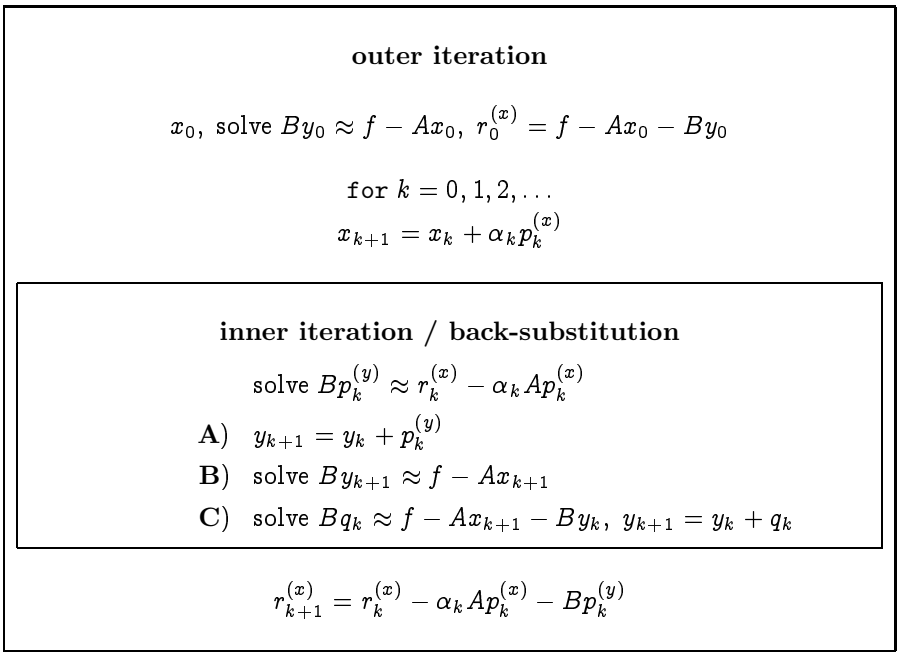


FIGURE 3.8. Null-space projection method: Three different schemes for computing the approximate solution y_{k+1} (called in the text the updated approximate solution (A), the approximate solution computed by a direct substitution (B), the approximate solution computed by a corrected direct substitution (C), respectively).

three back-substitution formulas (the schemes are described in Figure 3.8)

$$y_{k+1} = y_k + p_k^{(y)},\ p_k^{(y)} = B^\dagger(r_k^{(x)} - \alpha_k A p_k^{(x)}), \tag{3.43}$$

$$y_{k+1} = B^\dagger(f - Ax_{k+1}), \tag{3.44}$$

$$y_{k+1} = y_k + B^\dagger(f - Ax_{k+1} - By_k). \tag{3.45}$$

The pseudoinverse $B^\dagger$ in (3.43)–(3.45) is applied by solving the least squares with the matrix B. These problems are solved inexactly. In our considerations we will

assume that the computed solution $\bar{v}$ of the least squares problem $Bv \approx c$ is an exact solution of a perturbed problem $(B+\Delta B)\bar{v} \approx c+\Delta c$ with $\|\Delta B\|/\|B\| \leq \tau$ and $\|\Delta c\|/\|c\| \leq \tau$. The parameter τ again represents the measure for inexact solution of the least squares with B and actually describes the backward error. This can be achieved in many different ways considering the inner iteration loop solving the associated system of normal equations, the augmented system formulation, or solving it directly. Similar inexact schemes have been considered for solving quadratic programming problems [**2, 3**], multigrid methods [**20, 21**] or constraint preconditioners [**63, 83, 89**]. We assume $\tau\kappa(B) \ll 1$ which guarantees $B+\Delta B$ to have a full column rank. This allows the use of the perturbation theory (see [**104**] or [**55**, Lemma 19.8]), in particular the inequalities

$$\|(B+\Delta B)^\dagger\| \leq \frac{\|B^\dagger\|}{1-\tau\kappa(B)}, \ \|BB^\dagger - B(B+\Delta B)^\dagger\| \leq \frac{2\tau\kappa(B)}{1-\tau\kappa(B)}.$$

Note that if $\tau = O(u)$, then we have a backward stable method for solving the least squares problem with B. In our experiments we applied the CGLS method [**16**] with the stopping criterion based on the corresponding backward error. Notation $\tau = O(u)$ stands for the Householder QR factorization.

2.1. The attainable accuracy in the projected system. In this subsection we look at the accuracy in the outer iteration for solving the projected system $(I-\Pi)A(I-\Pi)x = (I-\Pi)f$. We can consider the perturbed system

$$(I-\hat{\Pi})A(I-\hat{\Pi})\hat{x} = (I-\hat{\Pi})f, \tag{3.46}$$

where $\hat{\Pi} = (B+\Delta B)(B+\Delta B)^\dagger$ such that $\|\Delta B\| \leq \tau\|B\|$. The residual associated with the solution of (3.46) can be written as

$$(I-\Pi)f - (I-\Pi)A(I-\Pi)\hat{x} = (\hat{\Pi}-\Pi)f + (I-\hat{\Pi})A(\Pi-\hat{\Pi})\hat{x} + (\Pi-\hat{\Pi})A(I-\Pi)\hat{x}$$

and due to $\|\hat{\Pi}-\Pi\| \leq \|\Delta B\|\min\{\|B^\dagger\|, \|(B+\Delta B)^\dagger\|\}$ [**55**, Lemma 19.8] we have

$$\|(I-\Pi)f - (I-\Pi)A(I-\Pi)\hat{x}\| \leq \frac{2\tau\kappa(B)}{1-\tau\kappa(B)}(\|f\| + \|A\|\|\hat{x}\|).$$

Indeed, even if we assume exact arithmetic, the residual obtained directly from $\hat{x}$ is proportional to the parameter τ. In addition, we ideally have $(B+\Delta B)^T\hat{x} = 0$ which implies $\| - B^T\hat{x}\| \leq \tau\|B\|\|\hat{x}\|$. Therefore we can expect that also the residual $-B^T\bar{x}_k$ associated with the computed approximate solution $\bar{x}_k$ will be proportional to τ. Such analysis is dependent on the choice of a particular method with the recurrences (3.41) and (3.42), and therefore we do not give it here. In

accordance with [**47**] it seems reasonable that the bound for $-B^T\bar{x}_k$ is proportional to the factor $\bar{X}_k$. Moreover, the error in the projection of an arbitrary vector is represented in the bounds by $\tau\kappa(B)/[1-\tau\kappa(B)]$. Therefore $-B^T\bar{x}_k$ and $\Pi\bar{x}_k$ can be expected to have the form

$$\|-B^T\bar{x}_k\| \le \frac{O(\tau)\|B\|}{1-\tau\kappa(B)}\bar{X}_k, \quad \|\Pi\bar{x}_k\| \le \frac{O(\tau)\kappa(B)}{1-\tau\kappa(B)}\bar{X}_k. \tag{3.47}$$

Theorem 3.9 shows that the true residual $(I-\Pi)f-(I-\Pi)A(I-\Pi)\bar{x}_k$ is ultimately proportional to τ, while its projection onto $N(B^T)$ will finally reach the level $O(u)$ provided that the updated residual $\bar{r}_k^{(x)}$ converges far below that level.

THEOREM 3.9. *The gap between the true residual* $(I-\Pi)f-(I-\Pi)A(I-\Pi)\bar{x}_k$ *and the projection of the updated residual* $(I-\Pi)\bar{r}_k^{(x)}$ *can be bounded by*

$$\|(I-\Pi)f-(I-\Pi)A(I-\Pi)\bar{x}_k-(I-\Pi)\bar{r}_k^{(x)}\| \le \frac{O(\tau)\kappa(B)}{1-\tau\kappa(B)}(\|f\|+\|A\|\bar{X}_k),$$

where $\bar{X}_k \equiv \max_{i=0,\ldots,k}\|\bar{x}_i\|$.

PROOF. The computed approximation $\bar{x}_{k+1}$ satisfies the relations

$$\bar{x}_{k+1} = \bar{x}_k + \bar{\alpha}_k\bar{p}_k^{(x)} + \Delta x_{k+1}, \ \|\Delta x_{k+1}\| \le u\|\bar{x}_k\| + (2u+u^2)\|\bar{\alpha}_k\bar{p}_k^{(x)}\|. \tag{3.48}$$

The inequality $\|\bar{\alpha}_k\bar{p}_k^{(x)}\| \le \|\bar{x}_{k+1}\| + \|\bar{x}_k\| + \|\Delta x_{k+1}\|$ gives $\|\bar{\alpha}_k\bar{p}_k^{(x)}\| \le 3\bar{X}_{k+1}$ and $\|\Delta x_{k+1}\| \le O(u)\bar{X}_{k+1}$. The vectors $\bar{y}_0$ and $\bar{p}_k^{(y)}$ satisfy $(B+\Delta B_0)\bar{y}_0 \approx \mathrm{fl}(f-Ax_0)+\Delta c_0$ with $\|\Delta B_0\| \le \tau\|B\|$, $\|\Delta c_0\| \le \tau\|\mathrm{fl}(f-Ax_0)\|$ and

$$(B+\Delta B_k)\bar{p}_k^{(y)} \approx \mathrm{fl}(\bar{r}_k^{(x)} - \bar{\alpha}_k A\bar{p}_k^{(x)}) + \Delta c_k, \tag{3.49}$$

$$\|\Delta B_k\| \le \tau\|B\|, \ \|\Delta c_k\| \le \tau\|\mathrm{fl}(\bar{r}_k^{(x)} - \bar{\alpha}_k A\bar{p}_k^{(x)})\|. \tag{3.50}$$

For updated residuals we have $\bar{r}_0^{(x)} = \mathrm{fl}(f-Ax_0-B\bar{y}_0)$ and

$$\bar{r}_{k+1}^{(x)} = \bar{r}_k^{(x)} - \bar{\alpha}_k A\bar{p}_k^{(x)} - B\bar{p}_k^{(y)} + \Delta r_{k+1}^{(x)}, \tag{3.51}$$

$$\|\Delta r_{k+1}^{(x)}\| \le O(u)(\|\bar{r}_k^{(x)}\| + \|A\|\|\bar{\alpha}_k\bar{p}_k^{(x)}\| + \|B\|\|\bar{p}_k^{(y)}\|). \tag{3.52}$$

The recursive use of (3.48) and (3.51) leads to the expression for the gap between the projections of $f - A\bar{x}_k$ and $\bar{r}_k^{(x)}$

$$(I-\Pi)(f - A\bar{x}_k - \bar{r}_k^{(x)}) = (I-\Pi)(f - A\bar{x}_0 - \bar{r}_0^{(x)}) - \sum_{i=0}^{k-1}(I-\Pi)(A\Delta x_{i+1} + \Delta r_{i+1}^{(x)}).$$

Taking norms and corresponding bounds we get, after some manipulation, the following:

$$\|(I-\Pi)(f - A\bar{x}_k - \bar{r}_k^{(x)})\| \leq \frac{O(u)\kappa(B)}{1-\tau\kappa(B)}\left(\|f\| + \|A\|\bar{X}_k\right). \tag{3.53}$$

Here we have used that $\|\bar{r}_k^{(x)}\| \leq \|\bar{r}_0^{(x)}\|$ for $k = 0, 1, \ldots$ which seems reasonable when solving the positive semi-definite problem. For the gap between $(I-\Pi)f - (I-\Pi)A(I-\Pi)\bar{x}_k$ and $(I-\Pi)\bar{r}_k^{(x)}$, we can write

$$\|(I-\Pi)f - (I-\Pi)A(I-\Pi)\bar{x}_k - (I-\Pi)\bar{r}_k^{(x)}\| \leq \|(I-\Pi)(f - A\bar{x}_k - \bar{r}_k^{(x)})\| + \|(I-\Pi)A\Pi\bar{x}_k\|.$$

Considering (3.53) and (3.47) we can conclude the proof. □

In Figure 3.9 we report the relative norms of the true residual $(I-\Pi)f - (I-\Pi)A(I-\Pi)\bar{x}_k$ (solid lines) and the updated residual $\bar{r}_k^{(x)}$ (dashed lines). The numerical results confirm that the residual $f - A\bar{x}_k$ is within $N(B^T)$ approximated by $\bar{r}_k^{(x)}$ to the working precision u. However, this is not true for the residual $(I-\Pi)f - (I-\Pi)A(I-\Pi)\bar{x}_k$ which is ultimately $O(\tau)$ as it follows from Theorem 3.9. The residual $-B^T\bar{x}_k$ obviously does not depend on the back-substitution scheme; see Figure 3.10.

In contrast to the Schur complement reduction method, the inexactness is connected with the matrix B instead of A. In practice, the sequential application of the matrix $(I-\Pi)A(I-\Pi)$ does not represent a symmetric operator. This is also reflected in the fact that we assume a general framework for computing the vector x_k and analyze another projection of residuals $f - A\bar{x}_k - B\bar{y}_k$ and $\bar{r}_k^{(x)}$. Ideally at every iteration step we apply the matrix-vector product with the matrix $(I-\hat{\Pi})A(I-\hat{\Pi})$, where $\hat{\Pi}$ represents the orthogonal projector $\hat{\Pi} = (B+\Delta B)(B+\Delta B)^\dagger$ with $\|\Delta B\| \leq \tau\|B\|$. A question similar to one in subsection 1.1 arises whether we can apply the results of [**47**] directly to the system

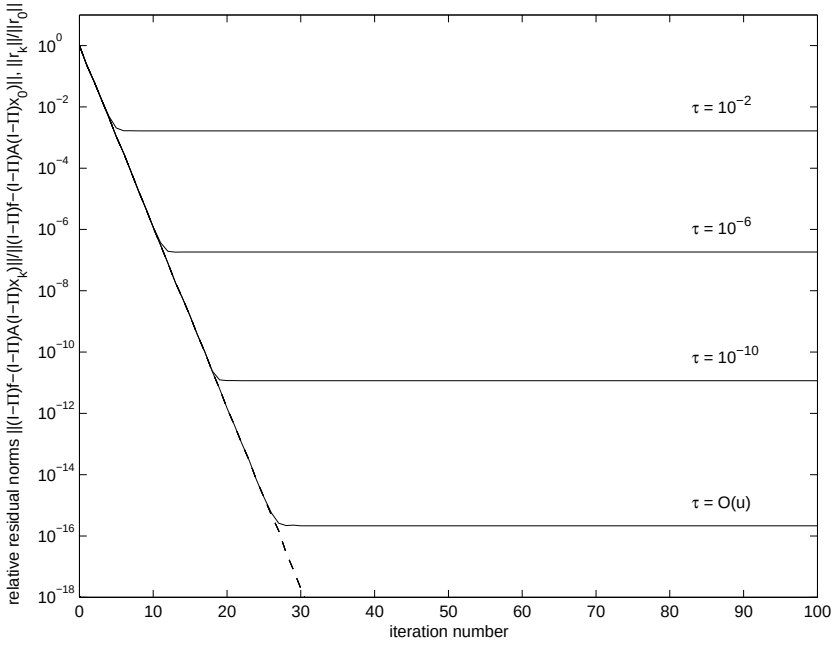


FIGURE 3.9. Null-space projection method: the relative norms of the true residual $(I-\Pi)f-(I-\Pi)A(I-\Pi)\bar{x}_k$ of the projected system (solid lines) and the updated residual $\bar{r}_k^{(x)}$ (dashed lines) – the updated solution scheme (3.43).

$(I-\hat{\Pi})A(I-\hat{\Pi})\hat{x}=(I-\hat{\Pi})f$. Theorem 3.9 shows that in finite precision arithmetic the residual $(I-\Pi)f-(I-\Pi)A(I-\Pi)\bar{x}_k$ will remain proportional to the parameter τ. The theory of Greenbaum can be directly applied only if the multiplication by $(I-\Pi)A(I-\Pi)$ satisfies $\|\mathrm{fl}[(I-\Pi)A(I-\Pi)x]-(I-\Pi)A(I-\Pi)x\|\leq O(u)\|(I-\Pi)A(I-\Pi)\|\|x\|$ which is obviously not the case here. In the idealized case we have $\mathrm{fl}[(I-\Pi)A(I-\Pi)x]=(I-\hat{\Pi})A(I-\hat{\Pi})x$ and hence

$$\|\mathrm{fl}[(I-\Pi)A(I-\Pi)x]-(I-\Pi)A(I-\Pi)x\|\leq\frac{O(\tau)\kappa(B)}{1-\tau\kappa(B)}\|A\|\|x\|.$$

If we could improve this bound to satisfy $\|\mathrm{fl}[(I-\Pi)A(I-\Pi)x]-(I-\Pi)A(I-\Pi)x\|\leq\tau\|A\|\|x\|$, the outer iteration process could be viewed as an iteration in finite precision arithmetic with the roundoff unit equal to τ and the theory of

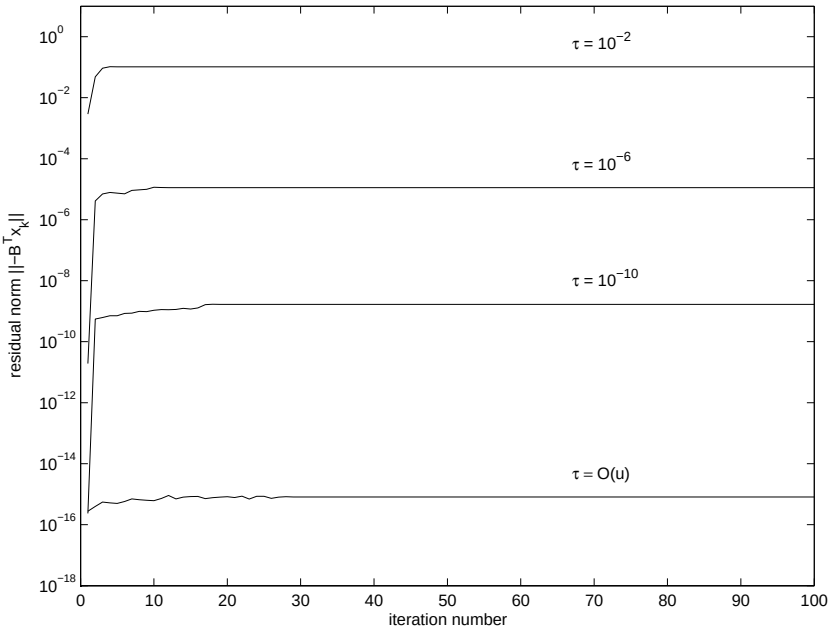


FIGURE 3.10. Null-space projection method: The norms of the residual $-B^T\bar{x}_k$ – the updated solution scheme (3.43).

Greenbaum would lead to the estimate

$$\|(I-\Pi)f-(I-\Pi)A(I-\Pi)\bar{x}_k-\bar{r}_k^{(x)}\| \leq \frac{O(\tau)}{1-\tau\kappa(B)}\|A\|(\|x\|+\bar{X}_k).$$

The numerical behavior of the null-space projection method was studied also in [**2, 3**], where the inner least squares are solved by the QR or LU factorization with $\tau = O(u)$ and the projected system is solved inexactly with the parameter η. Our Theorem 3.9 thus gives an answer to the question of how small can the parameter η be in the outer iteration. Roughly speaking, when using the error or residual minimizing method for solving the projected Hessian system the backward error associated with the iterate $\bar{x}_k$ cannot be smaller than $O(u)\kappa(B)/[1-O(u)\kappa(B)]$.

It is clear that no matter how we compute $\bar{x}_k$ and $\bar{y}_k$ we have the following relation between $(I-\Pi)f-(I-\Pi)A(I-\Pi)\bar{x}_k$, $f-A\bar{x}_k-B\bar{y}_k$ and $-B^T\bar{x}_k$:

$$(I-\Pi)f-(I-\Pi)A(I-\Pi)\bar{x}_k=(I-\Pi)(f-A\bar{x}_k-B\bar{y}_k)+(I-\Pi)A\Pi\bar{x}_k. \quad (3.54)$$

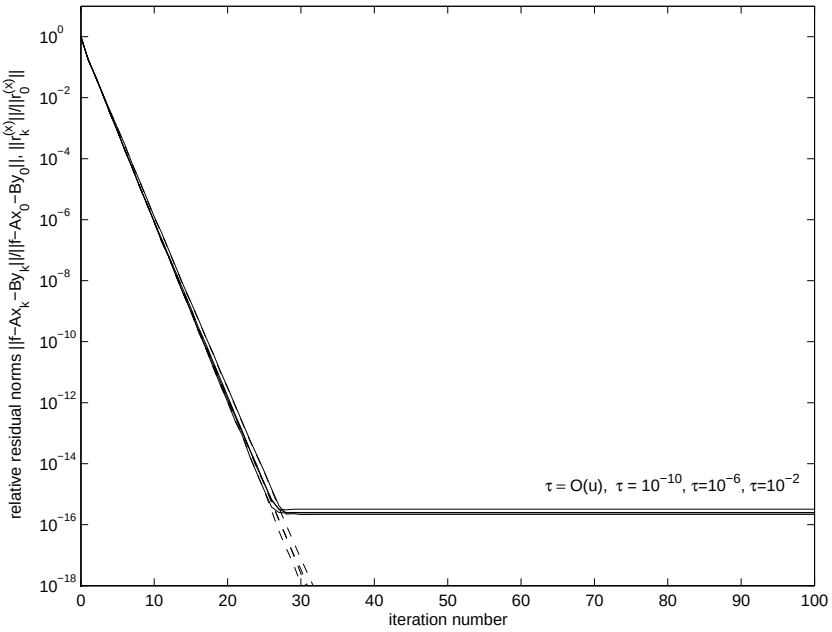


FIGURE 3.11. Null-space projection method: The relative norms of the true residual $f - A\bar{x}_k - B\bar{y}_k$ and the updated residual $\bar{r}_k^{(x)}$ (for the updated solution scheme (3.43)).

Owing to (3.47), $\Pi\bar{x}_k$ (and thus also $-B^T\bar{x}_k$) is $O(\tau)$. From Theorem 3.9 we have that $\|(I-\Pi)f-(I-\Pi)A(I-\Pi)\bar{x}_k\|$ is ultimately $O(\tau)$. Since $(I-\Pi)(f-A\bar{x}_k) = (I-\Pi)(f-A\bar{x}_k-B\bar{y}_k)$ for any $\bar{y}_k$ it also follows from Theorem 3.9 that the projection of $f-A\bar{x}_k-B\bar{y}_k$ onto $N(B^T)$ will ultimately reach $O(u)$. It is not clear from (3.54) whether the whole residual $f-A\bar{x}_k-B\bar{y}_k$ will be ultimately $O(\tau)$ or $O(u)$. It strongly depends on the back-substitution scheme used for computing the approximate solutions y_{k+1}. The following subsections show that the residual $f-A\bar{x}_k-B\bar{y}_k$ for the schemes with (3.43) (scheme A) and with (3.45) (scheme C) will finally reach $O(u)$, while the scheme B using (3.44) leads to the accuracy that is proportional only to τ.

2.2. Scheme A: The updated approximate solution. In this subsection we analyze the generic scheme with the update (3.43). This implementation

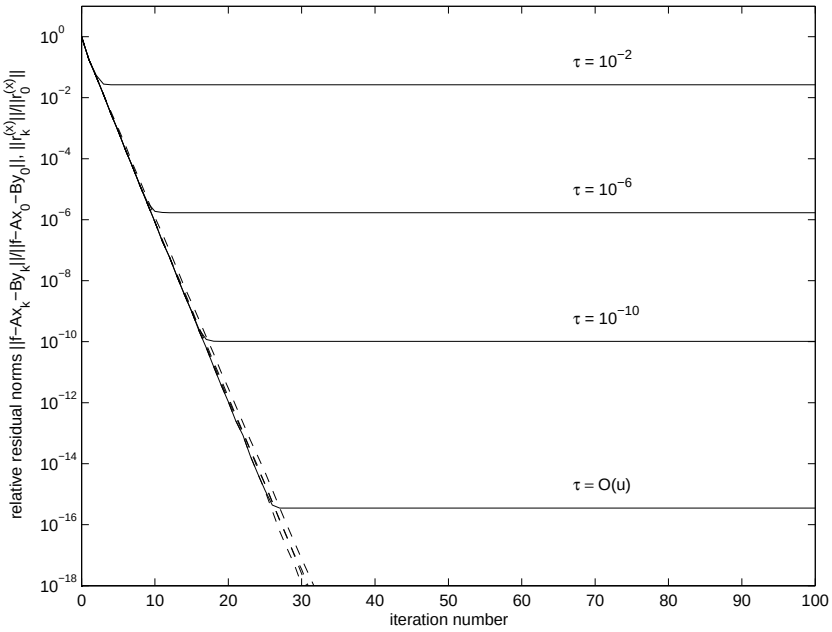


FIGURE 3.12. Null-space projection method: The relative norms of the true residual $f - A\bar{x}_k - B\bar{y}_k$ and the updated residual $\bar{r}_k^{(x)}$ (for the direct substitution scheme (3.44)).

does not require any additional solution of a least squares problem with the matrix B. Indeed, the computed direction vector $p_k^{(y)}$ is used to update both the iterate y_k and the residual $\bar{r}_k^{(x)}$. As we will see, this algorithm computes the residual $f - A\bar{x}_k - B\bar{y}_k$ which will ultimately reach the level of roundoff unit u independently on the fact that the inner least squares are solved with the accuracy determined by the parameter τ.

THEOREM 3.10. *The gap between the residuals* $f - A\bar{x}_k - B\bar{y}_k$ *and* $\bar{r}_k^{(x)}$ *can be bounded as follows:*

$$\|f - A\bar{x}_k - B\bar{y}_k - \bar{r}_k^{(x)}\| \leq O(u)(\|f\| + \|A\|\bar{X}_k + \|B\|\bar{Y}_k),$$

where $\bar{Y}_k \equiv \max_{i=0,\ldots,k} \|\bar{y}_i\|$. *The statement of the theorem remains true if we replace* $\bar{Y}_k$ *by* $\max\{\|y_0\|, \|p_i^{(y)}\|,\ i = 0, 1, \ldots, k-1\}$.

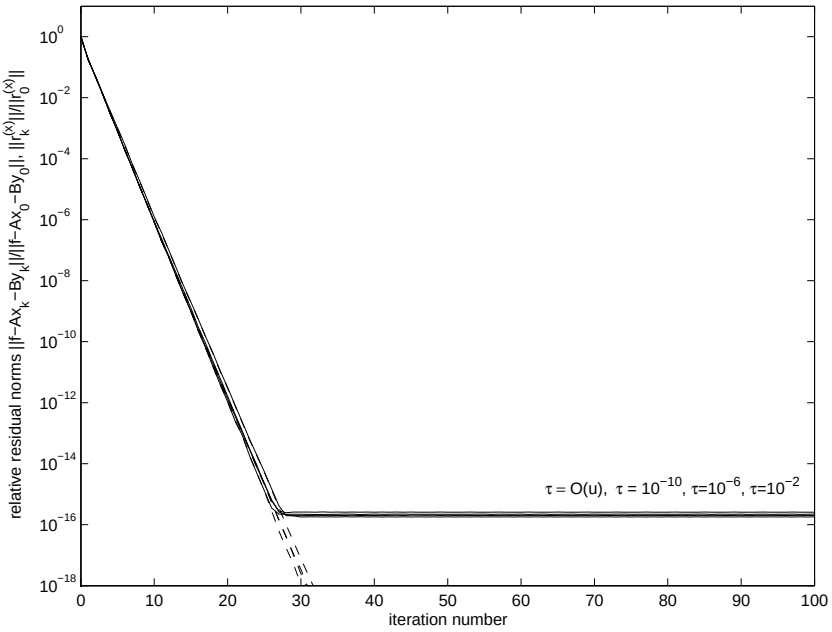


FIGURE 3.13. Null-space projection method: The relative norms of the true residual $f - A\bar{x}_k - B\bar{y}_k$ and the updated residual $\bar{r}_k^{(x)}$ (for the corrected direct substitution scheme (3.45)).

PROOF. The vector $\bar{x}_{k+1}$ satisfies (3.48) with $\|\Delta x_{k+1}\| \leq O(u)\bar{X}_{k+1}$ and similarly for $\bar{y}_{k+1}$ we have

$$\bar{y}_{k+1} = \bar{y}_k + \bar{p}_k^{(y)} + \Delta y_{k+1},\ \|\Delta y_{k+1}\| \leq u\|\bar{y}_k\| + (2u + u^2)\|\bar{p}_k^{(y)}\|$$

with $\|\Delta y_{k+1}\| \leq O(u)\bar{Y}_{k+1}$. The residual $\bar{r}_{k+1}^{(x)}$ satisfies (3.51) and thus $\|\Delta r_{k+1}^{(x)}\| \leq O(u)(\|\bar{r}_k^{(x)}\| + \|A\|\bar{X}_{k+1} + \|B\|\bar{Y}_{k+1})$. Using the above relations we obtain the recursive formula

$$f - A\bar{x}_{k+1} - B\bar{y}_{k+1} - \bar{r}_{k+1}^{(x)} = f - A\bar{x}_k - B\bar{y}_k - \bar{r}_k^{(x)} - A\Delta x_{k+1} - B\Delta y_{k+1} - \Delta r_{k+1}^{(x)}.$$

Taking the norms we get, after some manipulation, the following:

$$\|f - A\bar{x}_k - B\bar{y}_k - \bar{r}_k^{(x)}\| \leq O(u)\left(\|f\| + \|A\|\bar{X}_k + \|B\|\bar{Y}_k + \sum_{i=0}^{k-1}\|\bar{r}_i^{(x)}\|\right).$$

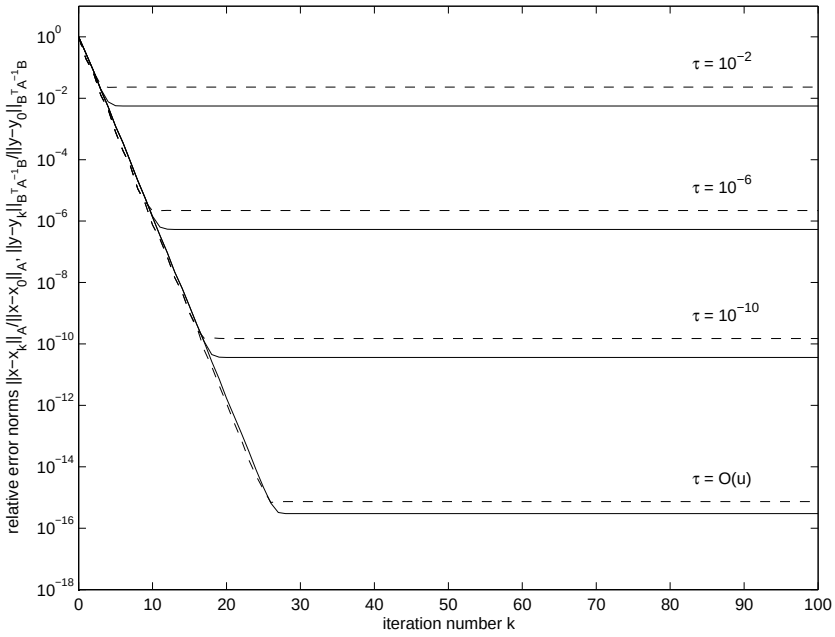


FIGURE 3.14. Null-space projection method: the relative norms of the errors $\|x - \bar{x}_k\|_A / \|x - x_0\|_A$ (solid lines) and $\|y - \bar{y}_k\|_{B^T A^{-1} B} / \|y - \bar{y}_0\|_{B^T A^{-1} B}$ (dashed lines) – the update solution scheme (3.43).

The statement can now be proved by induction on k. □

COROLLARY 3.11. *The gap between the residuals $f - A\bar{x}_k - B\bar{y}_k$ and $\bar{r}_k^{(x)}$ can be bounded as follows:*

$$\|f - A\bar{x}_k - B\bar{y}_k - \bar{r}_k^{(x)}\| \leq \frac{O(u)\kappa(B)}{1 - \tau\kappa(B)}(\|f\| + \|A\|\bar{X}_k).$$

We have shown that $\bar{r}_k^{(x)}$ is a good approximation to $f - A\bar{x}_k - B\bar{y}_k$ independently of the fact that $\bar{p}_k^{(y)}$ are computed inexactly. Note that Theorem 3.9 can be derived using Theorem 3.10 due to $\|(I - \Pi)(f - A\bar{x}_k - \bar{r}_k^{(x)})\| = \|(I - \Pi)(f - A\bar{x}_k - B\bar{y}_k - \bar{r}_k^{(x)})\| \leq \|f - A\bar{x}_k - B\bar{y}_k - \bar{r}_k^{(x)}\|$. In Figure 3.11 we show the relative

norms of $f - A\bar{x}_k - B\bar{y}_k$ (solid lines) and $\bar{r}_k^{(x)}$ (dashed lines). The results of our numerical experiment are in a good agreement with Theorem 3.10.

2.3. Scheme B: The approximate solution computed by a direct substitution. In this subsection we analyze the scheme (3.44), which uses the directly computed right-hand side vector $f - Ax_k$. The computed $\bar{y}_k$ is then a solution of the perturbed problem

$$(B + \Delta B_k)\bar{y}_k \approx \mathrm{fl}(f - A\bar{x}_k) + \Delta c_k \tag{3.55}$$

with $\|\Delta B_k\| \leq \tau\|B\|$ and $\|\Delta c_k\| \leq \tau\|\mathrm{fl}(f - A\bar{x}_k)\|$. We will show that $(I - \Pi)\bar{r}_k^{(x)}$ is a good approximation of $f - A\bar{x}_k - B\bar{y}_k$ provided that both are above their level of maximum attainable accuracy.

THEOREM 3.12. *The gap between the residuals $f - A\bar{x}_k - B\bar{y}_k$ and $(I - \Pi)\bar{r}_k^{(x)}$ can be bounded by*

$$\|f - A\bar{x}_k - B\bar{y}_k - (I - \Pi)\bar{r}_k^{(x)}\| \leq \frac{5\tau\kappa(B)}{1 - \tau\kappa(B)}(\|f\| + \|A\|\|\bar{x}_k\|) + O(u)(\|f\| + \|A\|\bar{X}_k + \|B\|\bar{Y}_k).$$

PROOF. Considering (3.55) it follows for the true residual that

$$\begin{aligned} f - A\bar{x}_k - B\bar{y}_k &= f - A\bar{x}_k - B(B + \Delta B_k)^\dagger[\mathrm{fl}(f - A\bar{x}_k) + \Delta c_k] \\ &= (I - \Pi)(f - A\bar{x}_k) + B[B^\dagger - (B + \Delta B_k)^\dagger]\mathrm{fl}(f - A\bar{x}_k) \\ &\quad + BB^\dagger[\mathrm{fl}(f - A\bar{x}_k) - (f - A\bar{x}_k)] - B(B + \Delta B_k)^\dagger\Delta c_k. \end{aligned}$$

Taking (3.55), the bounds on $B[B^\dagger - (B + \Delta B_k)^\dagger]$, $(B + \Delta B_k)^\dagger$ and Theorem 3.9 we get the desired result. □

COROLLARY 3.13. *The gap between the residuals $f - A\bar{x}_k - B\bar{y}_k$ and $(I - \Pi)\bar{r}_k^{(x)}$ can be bounded by*

$$\|f - A\bar{x}_k - B\bar{y}_k - (I - \Pi)\bar{r}_k^{(x)}\| \leq \frac{O(\tau)\kappa(B)}{1 - \tau\kappa(B)}(\|f\| + \|A\|\|\bar{x}_k\|) + \frac{O(u)\kappa(B)}{1 - \tau\kappa(B)}(\|f\| + \|A\|\bar{X}_k).$$

When using the formula (3.44) the residual $f - A\bar{x}_k - B\bar{y}_k$ will not decrease below a level proportional to τ, while $(I - \Pi)\bar{r}_k^{(x)}$ converges beyond the level $O(u)$. This result is illustrated by our numerical experiment. In Figure 3.12 we plotted the relative norms of $f - A\bar{x}_k - B\bar{y}_k$ (solid lines) and $\bar{r}_k^{(x)}$ (dashed lines).

2.4. Scheme C: The approximate solution computed with a corrected direct substitution. In this subsection we analyze the scheme (3.45) requiring a solution of two least squares problems with B. We show that its behavior is similar to the algorithm using the update (3.43). We prove that under certain assumptions the true residual $f - A\bar{x}_k - B\bar{y}_k$ converges ultimately to the $O(u)$ level. The difference is that while Theorem 3.10 holds without any additional conditions, here we have a situation analogous to the behavior of nonstationary iterative methods (see [**55**, Chapter 16]).

THEOREM 3.14. *Provided that for sufficiently large step k the computed vector $\bar{x}_k$ stagnates, i.e., we have $\|\bar{x}_{k+1} - \bar{x}_k\| \leq O(u)\bar{X}_{k+1}$, there exists some iteration step k_0 such that*

$$\|f - A\bar{x}_k - B\bar{y}_k - (I - \Pi)\bar{r}_k^{(x)}\| \leq O(u)(\|f\| + \|A\|\bar{X}_k + \|B\|\bar{Y}_k)$$

holds for all $k \geq k_0$.

PROOF. The vector $\bar{y}_{k+1}$ satisfies $\bar{y}_{k+1} = \bar{y}_k + \bar{q}_k^{(y)} + \Delta y_{k+1}$ and $\|\Delta y_{k+1}\| \leq O(u)\bar{Y}_{k+1}$, where $\bar{q}_k^{(y)}$ is the solution of the problem $(B + \Delta B_k)\bar{q}_k^{(y)} \approx \mathrm{fl}(f - A\bar{x}_{k+1} - B\bar{y}_k) + \Delta c_k$ with $\|\Delta B_k\| \leq \tau\|B\|$ and $\|\Delta c_k\| \leq \tau\|\mathrm{fl}(f - A\bar{x}_{k+1} - B\bar{y}_k)\|$. For $f - A\bar{x}_{k+1} - B\bar{y}_{k+1}$ we can then write

$$\begin{aligned} f - A\bar{x}_{k+1} - B\bar{y}_{k+1} = (I - \Pi)(f - A\bar{x}_{k+1}) + G_k(f - A\bar{x}_{k+1} - B\bar{y}_k) \\ - B(B + \Delta B_k)^\dagger \Delta c_k + h_k, \end{aligned}$$

where $G_k = B[B^\dagger - (B + \Delta B_k)^\dagger]$ and $h_k = -B(B + \Delta B_k)^\dagger[\mathrm{fl}(f - A\bar{x}_{k+1} - B\bar{y}_k) - (f - A\bar{x}_{k+1} - B\bar{y}_k)] - B\Delta y_{k+1}$. Projecting $f - A\bar{x}_{k+1} - B\bar{y}_{k+1}$ onto $R(B)$ and taking norms, we obtain

$$\begin{aligned} \|\Pi(f - A\bar{x}_{k+1} - B\bar{y}_{k+1})\| \leq \left[\|G_k\| + \tau\|B(B + \Delta B_k)^\dagger\|\right] \|f - A\bar{x}_{k+1} - B\bar{y}_k\| \\ + \tau\|B(B + \Delta B_k)^\dagger\|\|\mathrm{fl}(f - A\bar{x}_{k+1} - B\bar{y}_k) - (f - A\bar{x}_{k+1} - B\bar{y}_k)\| + \|h_k\|. \end{aligned}$$

The term $\|f - A\bar{x}_{k+1} - B\bar{y}_k\|$ can be further bounded by

$$\begin{aligned} \|f - A\bar{x}_{k+1} - B\bar{y}_k\| \leq \|(I - \Pi)(f - A\bar{x}_{k+1})\| + \|\Pi(f - A\bar{x}_k - B\bar{y}_k)\| \\ + \|A(\bar{x}_{k+1} - \bar{x}_k)\| \end{aligned}$$

which together with the bound on $\|G_k\|$, $\|h_k\| \leq O(u)(\|f\|+\|A\|\bar{X}_{k+1}+\|B\|\bar{Y}_{k+1})$, and $\tau\|B(B+\Delta B_k)^\dagger\| \leq \tau\kappa(B)[1-\tau\kappa(B)]^{-1} < 1$ leads to

$$\begin{aligned}\|\Pi(f - A\bar{x}_{k+1} - B\bar{y}_{k+1})\| \leq \frac{3\tau\kappa(B)}{1-\tau\kappa(B)} \,[&\|\Pi(f - A\bar{x}_k - B\bar{y}_k)\| \\ &+\|(I-\Pi)(f - A\bar{x}_{k+1})\| + \|A\|\|\bar{x}_{k+1} - \bar{x}_k\|] \\ &+O(u)(\|f\| + \|A\|\bar{X}_{k+1} + \|B\|\bar{Y}_{k+1}).\end{aligned}$$

After the recursive use of the previous inequality we obtain

$$\begin{aligned}\|\Pi(f - A\bar{x}_k - B\bar{y}_k)\| \leq &\left(\frac{3\tau\kappa(B)}{1-\tau\kappa(B)}\right)^k \|f - A\bar{x}_0 - B\bar{y}_0\| \\ &+ \sum_{i=0}^{k-1}\left(\frac{3\tau\kappa(B)}{1-\tau\kappa(B)}\right)^{k-i} [\|(I-\Pi)(f - A\bar{x}_{i+1})\| + \|A\|\|\bar{x}_{i+1} - \bar{x}_i\|] \\ &+ O(u)(\|f\| + \|A\|\bar{X}_k + \|B\|\bar{Y}_k).\end{aligned} \tag{3.56}$$

Under the assumption on the stagnation of iterates there exist some index k_0 such that the second term on the right-hand side of (3.56) will be of order $O(u)(\|f\| + \|A\|\bar{X}_k + \|B\|\bar{Y}_k)$ for all iteration steps $k \geq k_0$. Finally, from Theorem 3.10 we have $\|(I-\Pi)(f - A\bar{x}_k) - (I-\Pi)\bar{r}_k^{(x)}\| \leq O(u)(\|f\| + \|A\|\bar{X}_k + \|B\|\bar{Y}_k)$. □

COROLLARY 3.15. *Provided that for sufficiently large step k the computed vector $\bar{x}_k$ stagnates, i.e., we have $\|\bar{x}_{k+1} - \bar{x}_k\| \leq O(u)\bar{X}_{k+1}$, there exists some iteration step k_0 such that*

$$\|f - A\bar{x}_k - B\bar{y}_k - (I-\Pi)\bar{r}_k^{(x)}\| \leq \frac{O(u)\kappa(B)}{1-\tau\kappa(B)}(\|f\| + \|A\|\bar{X}_k)$$

holds for all $k \geq k_0$.

Theorem 3.14 shows that $f - A\bar{x}_k - B\bar{y}_k$ will ultimately reach the $O(u)$ level. As soon as the approximate solutions $\bar{x}_k$ stagnate with $\|\bar{x}_{k+1} - \bar{x}_k\| \leq O(u)\bar{X}_{k+1}$, the rate of convergence of this process is roughly given by the factor $3\tau\kappa(B)[1-\tau\kappa(B)]^{-1}$. Note that similar to subsection 1.4 the assumption on the stagnation is not restrictive. The numerical results on a model example are shown in Figure 3.13, which reports the relative norms of $f - A\bar{x}_k - B\bar{y}_k$ (solid lines) and $\bar{r}_k^{(x)}$ (dashed lines), and are in a good agreement with Theorem 3.14.

2.5. Forward error analysis. In this subsection we look at the maximum attainable accuracy measured by errors $x - \bar{x}_k$ and $y - \bar{y}_k$. The analysis is very similar to the Schur complement reduction method and therefore we focus only on issues particular to the null-space projection method. We recall that relation (3.32) gives the universal bounds (3.33), (3.34), and (3.35). Independent of the back-substitution scheme used for computing $\bar{y}_k$, the terms $\gamma_2\| - B^T\bar{x}_k\|$ and $\gamma_3\| - B^T\bar{x}_k\|$ on the right-hand side of (3.33) and (3.34), respectively, are always proportional to τ. The terms with $f - A\bar{x}_k - B\bar{y}_k$ depend on the back-substitution formula and their final magnitude will be at most $O(\tau)$, leading to similar conclusions on errors as in subsection 1.5. The estimate for $\|x - \bar{x}_k\|_A$ is given in the following theorem.

THEOREM 3.16. *The A-norm of the error $x - \bar{x}_k$ can be bounded as*

$$\|x - \bar{x}_k\|_A \leq \delta_1\| - B^T\bar{x}_k\| + \delta_2\|(I - \Pi)(f - A\bar{x}_k)\|, \tag{3.57}$$

where $\delta_1 \equiv \|A\|^{1/2}/\sigma_{min}(B)$ and $\delta_2 \equiv \sigma_{min}^{-1/2}(A)$ are constants independent of the iteration step k.

PROOF. Since $(I - \Pi)A(x - \bar{x}_k) = (I - \Pi)(f - A\bar{x}_k)$, $B^T x = 0$ and using $\|B(B^TB)^{-1}\| = \sigma_{min}^{-1}(B)$, $\|x - \bar{x}_k\|_A^2$ can be written as

$$\begin{aligned}\|x - \bar{x}_k\|_A^2 &= (\Pi(x - \bar{x}_k), A(x - \bar{x}_k)) + ((I - \Pi)A(x - \bar{x}_k), x - \bar{x}_k) \\ &\leq \|A^{1/2}\|\|x - \bar{x}_k\|_A(\|B(B^TB)^{-1}\|\|B^T(x - \bar{x}_k)\| \\ &\quad + \|(I - \Pi)(f - A\bar{x}_k)\|).\end{aligned}$$

Dividing both sides by $\|x - \bar{x}_k\|_A$ gives the statement (3.57). □

The first term on the right-hand side of (3.57) should be zero in exact arithmetic. The computed $\bar{x}_k$, however, does not fulfill $-B^T\bar{x}_k = 0$ and its departure from $N(B^T)$ was discussed in (3.47). The second term converges to zero in exact arithmetic and it is related to the projected residual $(I-\Pi)(f-A\bar{x}_k)$, see Theorem 3.53. The result for $y-\bar{y}_k$ can be obtained from (3.57) using (3.35). Provided that $\bar{r}_k^{(x)}$ is larger than $O(\tau)$, $\|x - \bar{x}_k\|_A$ is then well approximated by $\delta_2\|(I - \Pi)\bar{r}_k^{(x)}\|$.

3. Numerical experiments in the nonsymmetric case

In this section we consider a nonsymmetric block A in the system (3.1). Hence the difference here is that we apply a nonsymmetric iterative method to solve the Schur complement system $B^TA^{-1}By = B^TA^{-1}f$ and the projected system

$(I-\Pi)A(I-\Pi)x = (I-\Pi)f$. We demonstrate the theoretical results of Sections 1 and 2 on a simple numerical example of a nonsymmetric system (3.1) with

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

Since $\kappa(A) = \|A\|\|A^{-1}\| = 2.00 \cdot 32.15 = 64.27$ and $\kappa(B) = \|B\|\|B^{\dagger}\| = 7.39 \cdot 0.75 = 5.55$, the conditioning of matrices A and B has not a significant effect on the behavior of considered schemes. For each test we set $y_0 = 0$ and $x_0 = 0$ for the Schur complement reduction method and for the null-space projection method, respectively.

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 $\bar{x}_k$ and $\bar{y}_k$ stagnate on the level proportional to the maximum norms (measured either by $\bar{X}_k$ or $\bar{Y}_k$) 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 [**54, 25**] 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 [**54**] 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 [**88**] 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 $\bar{X}_k$ and $\bar{Y}_k$ in CGNE and GMRES applied either to the Schur complement system or to the projected system cannot be significantly larger than the size of the initial approximations x_0, y_0 and unknowns x and y. Depending on the actual backsubstitution formula the maximum attainable accuracy level is then proportional either to roundoff unit u or to the parameter τ, and the quantities $\bar{Y}_k$ and $\bar{X}_k$ do not play an important role in our bounds.

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 [**35**] or CGS [**93**]. 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 near the true solution. For this reason one cannot give an a priori bound on $\bar{X}_k$ and $\bar{Y}_k$, and indeed the algorithms for solving the Schur complement system and the projected system 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 Figure 3.15 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. Similarly in Figure 3.16 we report the results for the null-space projection method, where the inner systems were solved using the Householder QR factorization of the matrix B. We have plotted the true residual $B^T A^{-1} f - B^T A^{-1} B\bar{y}_k$ and $(I - \Pi)(f - A\bar{x}_k)$ and the updated residuals $\bar{r}_k^{(y)}$ and $\bar{r}_k^{(x)}$, 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), true residuals in the Schur complement system and in the projected system behave as indicated by the estimates of Theorem 3.1 and 3.9. It is clear from Figures 3.15 and 3.16 that for the error norm minimizing CGNE and the residual minimizing GMRES is the maximum attainable accuracy level proportional to the unit roundoff. The quantities $\bar{Y}_k$ and $\bar{X}_k$ 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 Tab. 3.1). Indeed, the results confirm that the final residuals reach the levels which are roughly equal to $O(u)\bar{Y}_k$ or $O(u)\bar{X}_k$ instead of $O(u)$. Note that the matrices A and B

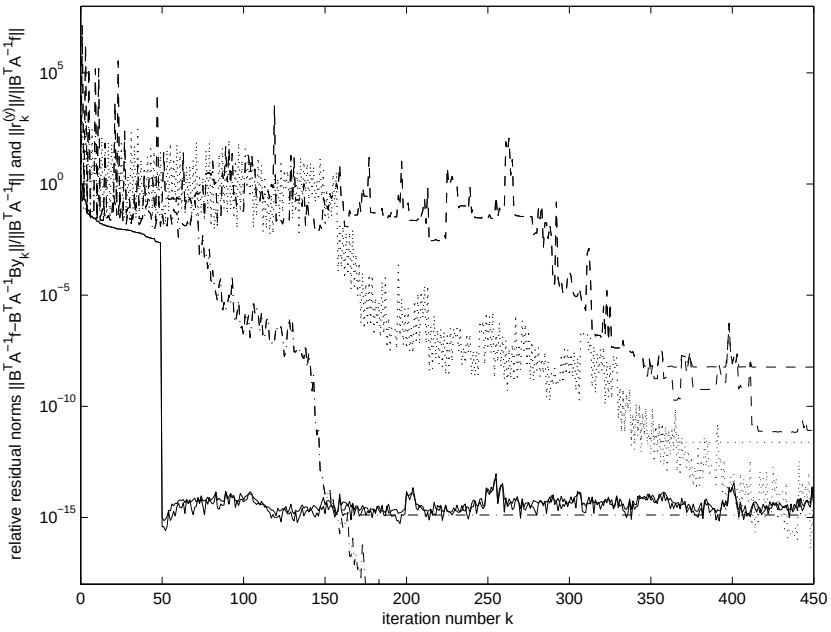


FIGURE 3.15. Relative norms of the residual $B^T A^{-1} f - B^T A^{-1} B \bar{y}_k$ in the Schur complement reduction method with respect to the iteration number for GMRES (solid lines), CGNE (dash-dotted lines), Bi-CG (dotted lines) and CGS (dashed lines) with a direct solver used for the solution of inner systems.

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 Figures 3.17 and 3.18 we report the norms of the residual $f - A\bar{x}_k - B\bar{y}_k$ in the Schur complement reduction method where the system (3.3) 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 $f - A\bar{x}_k - B\bar{y}_k$ 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 Figures 3.17 and 3.18 we can see the difference between the final accuracy levels of the norm of $f - A\bar{x}_k - B\bar{y}_k$ for the generic

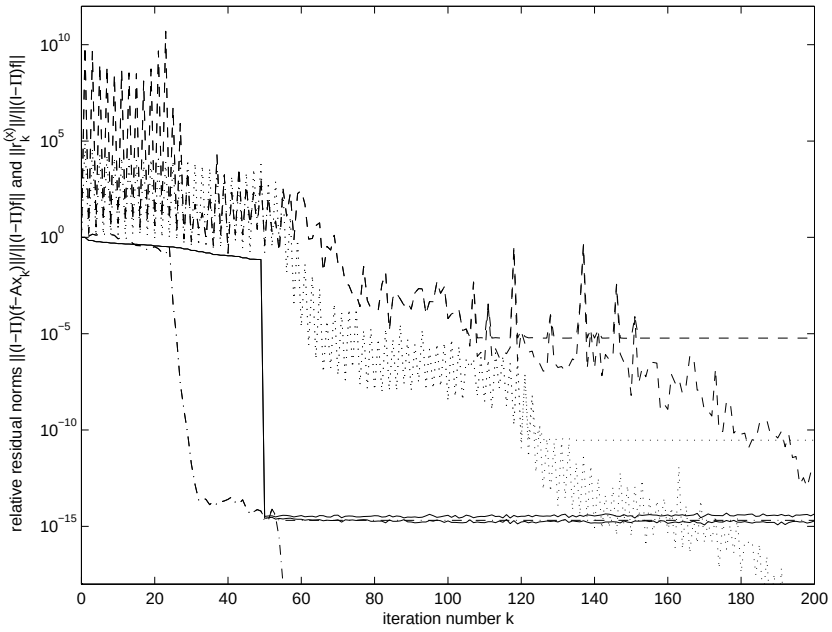


FIGURE 3.16. Relative norms of the residual $B^T A^{-1} f - B^T A^{-1} B \bar{y}_k$ in the null-space projection method with respect to the iteration number for GMRES (solid lines), CGNE (dash-dotted lines), Bi-CG (dotted lines) and CGS (dashed lines) with a direct solver used for the solution of inner systems.

update (3.10) and for the direct substitution (3.11) (see Corollary 3.3 and 3.5). In the first case, where the ultimate accuracy level depends on the maximum norm of the iterates $\bar{Y}_k$, the residual is essentially growing due to the accumulation of the residuals in inner systems. On the other hand, for the direct substitution (3.11) the maximum attainable accuracy of the first equation in (3.1) is bounded by the norm of the actual iterate $\bar{y}_k$. The norms of $f - A\bar{x}_k - B\bar{y}_k$ are somewhat oscillating which reflects the jumps of $\|\bar{y}_k\|$ in the initial phase of the iteration process. When the norms of $\bar{y}_k$ begin to stagnate, the norms of $f - A\bar{x}_k - B\bar{y}_k$ do so but on much smaller level than for the generic update (3.10). This difference between the accuracy levels is even more significant for the CGS method which exhibits much larger oscillations of the iterates. Note that both for Bi-CG and

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

TABLE 3.1. Quantities $\bar{Y}_k$ and $\bar{X}_k$ in the Schur complement method and in the null-space projection method, respectively, for GMRES, CGNE, BiCG and CGS.

CGS the residual norms for the corrected direct substitution converge to the unit roundoff level and it is not affected by the oscillations in the initial phase (see Corollary 3.7).

In Figures 3.19 and 3.20 we report the norms of the residual $f - A\bar{x}_k - B\bar{y}_k$ for the null-space projection method where the projected system 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 $f - A\bar{x}_k - B\bar{y}_k$ 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 (3.44) 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 is proportional to the unit roundoff u and to the quantity $\bar{X}_k$, while the second term is proportional to τ and to the norm of the actual iterate $\bar{x}_k$ (see Corollary 3.11 and 3.13). 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 $f - A\bar{x}_k - B\bar{y}_k$ does depend also on τ. This can be seen in Figures 3.19 and 3.20. The final level of the residual $f - A\bar{x}_k - B\bar{y}_k$ 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

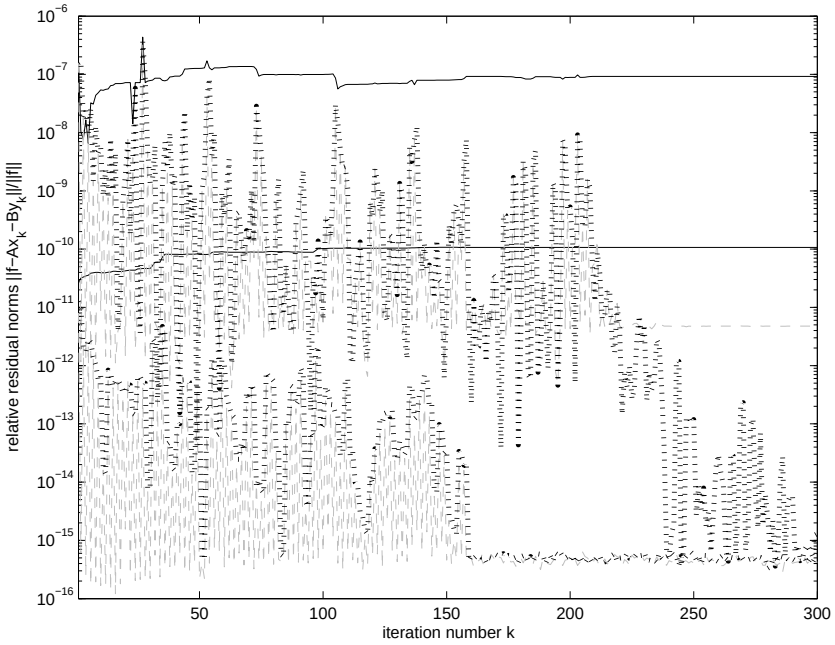


FIGURE 3.17. Schur complement reduction method: Relative norms of the residual $f - A\bar{x}_k - B\bar{y}_k$ for the Bi-CG method using the generic update (solid lines), the direct substitution (dashed lines) and the corrected direct substitution (dotted lines) with the inner systems solved either by a direct solver or by an iterative method where $\tau = 10^{-12}$.

errors (on the right). For other two back-substitution formulas the norms of $f - A\bar{x}_k - B\bar{y}_k$ 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 scheme (3.45) converge to the level of unit roundoff affected however by the oscillations of the iterates (see Corollary 3.15).

4. Backward error estimate for the Schur complement reduction

We can also interpret the solution computed by an inexact method as the exact solution of a perturbed problem. It seems quite reasonable to use the local

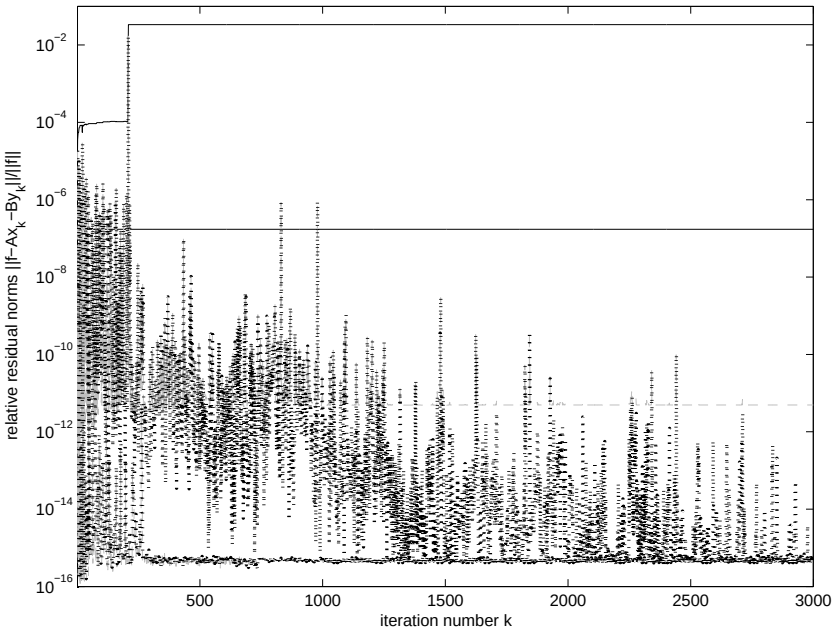


FIGURE 3.18. Schur complement reduction method: Relative norms of the residual $f - A\bar{x}_k - B\bar{y}_k$ for the CGS method using the generic update (solid lines), the direct substitution (dashed lines) and the corrected direct substitution (dotted lines) with the inner systems solved either by a direct solver or by an iterative method where $\tau = 10^{-12}$.

backward errors of inner systems to give an estimate on the global backward error associated with the original saddle point system. In this section we try to illustrate these ideas to the case of the scheme A of the Schur complement reduction (see subsection 1.2). Instead of the system (3.1) we consider the generalized saddle point system

$$\begin{pmatrix} A & B \\ B^T & -C \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} f \\ y \end{pmatrix}, \tag{3.58}$$

where A, B and f are as in the previous sections and C is an $m \times m$ matrix (often symmetric positive semidefinite in applications).

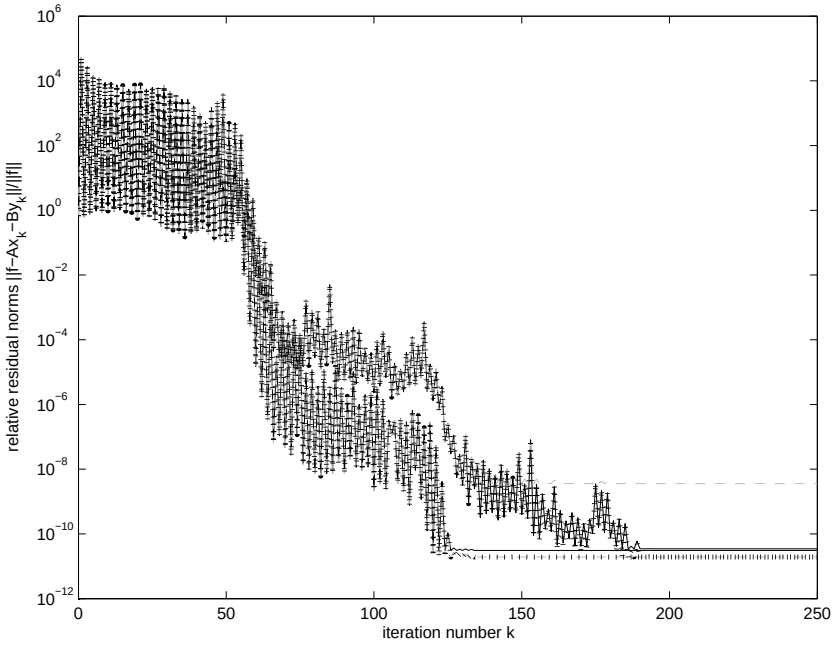


FIGURE 3.19. Null-space projection method: Relative norms of the residual $f - A\bar{x}_k - B\bar{y}_k$ for the Bi-CG method using the generic update (solid lines), the direct substitution (dashed lines) and the corrected direct substitution (dotted lines) with the inner systems solved either by a direct solver or by an iterative method where $\tau = 10^{-9}$.

Assume that the initial approximation $\bar{x}_0$ satisfies

$$A\bar{x}_0 = \mathrm{fl}(f - By_0) + s_0^{(x)}, \ \|s_0^{(x)}\| \leq \tau_0^{(x)}\|A\|\|\bar{x}_0\|, \tag{3.59}$$

where $s_0^{(x)}$ is the residual. Note that the condition on $\|s_0^{(x)}\|$ is equivalent to that used in Section 1.2. Similarly let the computed direction vectors $p_i^{(x)}$ satisfy

$$A\bar{p}_i^{(x)} = \mathrm{fl}(-B\bar{p}_i^{(y)}) + s_i^{(p)}, \ \|s_i^{(p)}\| \leq \tau_i^{(p)}\|A\|\|\bar{p}_i^{(x)}\|. \tag{3.60}$$

The vector $s_i^{(p)}$ is the corresponding residual. Based on these considerations we can formulate the following theorem which states that the computed iterates $\bar{x}_k$

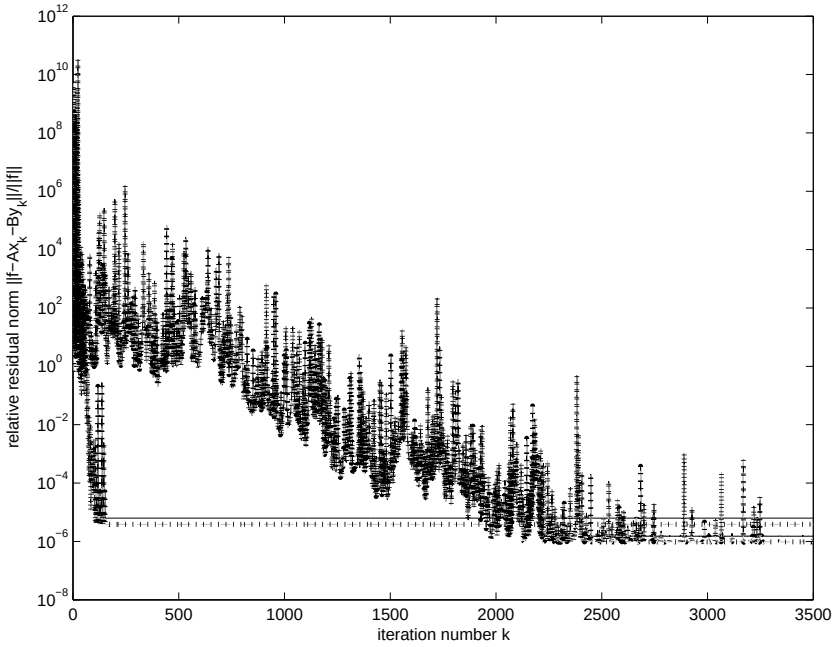


FIGURE 3.20. Null-space projection method: Relative norms of the residual $f - A\bar{x}_k - B\bar{y}_k$ for the CGS method using the generic update (solid lines), the direct substitution (dashed lines) and the corrected direct substitution (dotted lines) with the inner systems solved either by a direct solver or by an iterative method where $\tau = 10^{-9}$.

and $\bar{y}_k$ satisfy a perturbed equation $f - (A + \Delta A)x - By = 0$. In addition, we give a bound on the norm of the difference $g - B^T\bar{x}_k + C\bar{y}_k - \bar{r}_k^{(y)}$.

THEOREM 3.17. *The iterates computed with the algorithm of the Schur complement reduction method using the back-substitution formula (3.10) satisfy the inequality*

$$\begin{aligned} &\|f - (A + \Delta A^{(k)})\bar{x}_k - B\bar{y}_k\| \\ &\leq u\|f\| + 5ku\|A\|\bar{X}_k + (1 + c + (5 + 2c)k)u\|B\|\bar{Y}_k. \end{aligned} \tag{3.61}$$

where the perturbation matrix $\Delta A^{(k)}$ *is given by*

$$\Delta A^{(k)} \equiv \left(-s_0^{(x)} - \sum_{i=0}^{k-1} \alpha_i s_i^{(p)} \right) \frac{\bar{x}_k^T}{\|\bar{x}_k\|^2}$$

with

$$\|\Delta A^{(k)}\| \leq \gamma_k \|A\|, \ \gamma_k \equiv \frac{\tau_0^{(x)} \|\bar{x}_0\| + \sum_{i=0}^{k-1} \tau_i^{(p)} \|\bar{\alpha}_i \bar{p}_i^{(x)}\|}{\|\bar{x}_k\|}$$

and

$$\bar{X}_k \equiv \max\{\|\bar{x}_i\| \mid i = 0, 1, \ldots, k\}, \ \bar{Y}_k \equiv \max\{\|\bar{y}_i\| \mid i = 0, 1, \ldots, k\}.$$

The norm of the gap between the true residual $g - B^T \bar{x}_k + C\bar{y}_k$ *and the updated one* $\bar{r}_k^{(y)}$ *can be bounded as follows*

$$\|g - B^T \bar{x}_k + C\bar{y}_k - \bar{r}_k^{(y)}\| \leq u\|g\| + (3 + c + (12 + 2c)k)u(\|B\|\bar{X}_k + \|C\|\bar{Y}_k). \quad (3.62)$$

PROOF. The computed iterates $\bar{x}_i$ and $\bar{y}_i$ $(i = 0, 1, \ldots)$ satisfy

$$\bar{x}_{i+1} = \bar{x}_i + \bar{\alpha}_i \bar{p}_i^{(x)} + \Delta x_{i+1}, \quad \|\Delta x_{i+1}\| \leq u\|\bar{x}_i\| + 2u\|\bar{\alpha}_i \bar{p}_i^{(x)}\| + O(u^2), \quad (3.63)$$

$$\bar{y}_{i+1} = \bar{y}_i + \bar{\alpha}_i \bar{p}_i^{(y)} + \Delta y_{i+1}, \quad \|\Delta y_{i+1}\| \leq u\|\bar{y}_i\| + 2u\|\bar{\alpha}_i \bar{p}_i^{(y)}\| + O(u^2). \quad (3.64)$$

Since $\|\bar{\alpha}_i \bar{p}_i^{(x)}\| \leq \|\bar{x}_{i+1}\| + \|\bar{x}_i\| + \|\Delta x_{i+1}\|$, we obtain

$$\|\bar{\alpha}_i \bar{p}_i^{(x)}\| \leq (1+2u)\|\bar{x}_{i+1}\| + (1+3u)\|\bar{x}_i\| + O(u^2) \leq (2+5u)\bar{X}_{i+1} + O(u^2) \quad (3.65)$$

and hence the inequality (3.63) becomes

$$\|\Delta x_{i+1}\| \leq 2u\|\bar{x}_{i+1}\| + 3u\|\bar{x}_i\| + O(u^2) \leq 5u\bar{X}_{i+1} + O(u^2). \quad (3.66)$$

Similarly

$$\|\bar{\alpha}_i \bar{p}_i^{(y)}\| \leq (1+2u)\|\bar{y}_{i+1}\| + (1+3u)\|\bar{y}_i\| + O(u^2) \leq (2+5u)\bar{Y}_{i+1} + O(u^2) \quad (3.67)$$

and hence the inequality (3.64) becomes

$$\|\Delta y_{i+1}\| \leq 2u\|\bar{y}_{i+1}\| + 3u\|\bar{y}_i\| + O(u^2) \leq 5u\bar{Y}_{i+1} + O(u^2). \quad (3.68)$$

The computed updated residual satisfies

$$\bar{r}_{i+1}^{(y)} = \bar{r}_i^{(y)} - \bar{\alpha}_i B^T \bar{p}_i^{(x)} + \bar{\alpha}_i C \bar{p}_i^{(y)} + \Delta r_{i+1}^{(y)} \quad (3.69)$$

with

$$\|\Delta r_{i+1}^{(y)}\| \leq u\|\bar{r}_i^{(y)}\| + (3+c)u(\|B\|\|\bar{\alpha}_i \bar{p}_i^{(x)}\| + \|C\|\|\bar{\alpha}_i \bar{p}_i^{(y)}\|) + O(u^2).$$

Using (3.65) and (3.67) we get

$$\|\Delta r_{i+1}^{(y)}\| \leq u\|\bar{r}_i^{(y)}\| + (6+2c)u(\|B\|\bar{X}_{i+1} + \|C\|\bar{Y}_{i+1}) + O(u^2). \tag{3.70}$$

To obtain the first statement (3.61), we start with

$$\begin{aligned} f - A\bar{x}_{i+1} - B\bar{y}_{i+1} &= f - A\bar{x}_i - B\bar{y}_i - \bar{\alpha}_i A\bar{p}_i^{(x)} - \bar{\alpha}_i B\bar{p}_i^{(y)} - A\Delta x_{i+1} - B\Delta y_{i+1} \\ &= f - A\bar{x}_i - B\bar{y}_i - \bar{\alpha}_i s_i^{(p)} + \bar{\alpha}_i(\mathrm{fl}(B\bar{p}_i^{(y)}) - B\bar{p}_i^{(y)}) - A\Delta x_{i+1} - B\Delta y_{i+1} \end{aligned}$$

which gives

$$\begin{aligned} f - A\bar{x}_k - B\bar{y}_k = &-s_0^{(x)} - \sum_{i=0}^{k-1} \bar{\alpha}_i s_i^{(p)} \\ &- (\mathrm{fl}(f - By_0) - (f - By_0)) \\ &+ \sum_{i=0}^{k-1} \left(\bar{\alpha}_i(\mathrm{fl}(B\bar{p}_i^{(y)}) - B\bar{p}_i^{(y)}) - A\Delta x_{i+1} - B\Delta y_{i+1} \right) \end{aligned}$$

using (3.59) and (3.60). Now (3.61) follows by taking norms and using (3.66), (3.68) and the definition of $\Delta A^{(k)}$. The second statement (3.62) follows from

$$g - B^T\bar{x}_{i+1} + C\bar{y}_{i+1} - \bar{r}_{i+1}^{(y)} = g - B^T\bar{x}_i + C\bar{y}_i - \bar{r}_i^{(y)} - B^T\Delta x_{i+1} + C\Delta y_{i+1} - \Delta r_{i+1}^{(y)}.$$

The recursive use of this identity gives

$$\begin{aligned} g - B^T\bar{x}_k + C\bar{y}_k - \bar{r}_k^{(y)} = &\; g - B^T\bar{x}_0 + Cy_0 - \bar{r}_0^{(y)} \\ &+ \sum_{i=0}^{k-1} (-B^T\Delta x_{i+1} + C\Delta y_{i+1} - \Delta r_{i+1}^{(y)}). \end{aligned} \tag{3.71}$$

It can be easily shown by induction that $\bar{r}_i^{(y)} = g - B^T\bar{x}_i + C\bar{y}_i + O(u)$ and hence (3.70) becomes $\|\Delta r_{i+1}^{(y)}\| \leq (7+2c)u(\|B\|\bar{X}_{i+1} + \|C\|\bar{Y}_{i+1}) + O(u^2)$ and taking a norm on both sides of (3.71) proves the desired result. □

The theorem shows that the computed iterates $\bar{x}_k$ and $\bar{y}_k$ are the block components of the exact solution vector of the perturbed saddle point problem

$$\begin{pmatrix} A + \Delta A^{(k)} & B \\ B^T & -C \end{pmatrix} \begin{pmatrix} \bar{x}_k \\ \bar{y}_k \end{pmatrix} = \begin{pmatrix} f + \Delta f_k \\ g + \Delta g_k \end{pmatrix}, \tag{3.72}$$

where

$$\begin{aligned} \|\Delta f_k\| &\leq O(u)(\|f\| + \|A\|\bar{X}_k + \|B\|\bar{Y}_k), \\ \|\Delta g_k\| &\leq O(u)(\|g\| + \|B\|\bar{X}_k + \|C\|\bar{Y}_k) + \|\bar{r}_k^{(y)}\|. \end{aligned}$$

When the norm of $\bar{r}_k^{(y)}$ drops below the level of unit roundoff the iterates $\bar{x}_k$ and $\bar{y}_k$ satisfy the system (3.72), where the inexactness of inner systems is concentrated mainly in the perturbed matrix $A + \Delta A^{(k)}$, while the right-hand side is affected only by an $O(u)$ perturbation. The inner backward errors $\tau_0^{(x)}$ and $\tau_i^{(p)}$ should be small enough to ensure that the perturbed matrix $A + \Delta A^{(k)}$ is nonsingular which gives an upper bound on γ_k. However this quantity depends on terms known the step k of the iteration process and it is not clear at the moment how to choose a priori the inner tolerances $\tau_0^{(x)}$ and $\tau_i^{(p)}$ to ensure that the condition $\gamma_k < 1/\kappa(A)$ will hold. See [**90, 39**] for similar issues related to GMRES and FOM, and [**2, 3**] for the backward error analysis when sparse elimination techniques combined with iterative methods are applied to the solution of saddle point problems arising in sparse quadratic programming problems.

CHAPTER 4

Numerical stability of some residual minimizing Krylov subspace methods

In this chapter we consider certain methods for solving a system of linear algebraic equations

$$Ax = b, \qquad A \in \mathbb{R}^{N\times N}, \qquad b \in \mathbb{R}^N, \tag{4.1}$$

where A is a large and sparse nonsingular matrix that is, in general, nonsymmetric. For solving such systems, Krylov subspace methods are very popular. They build a sequence of iterates x_n $(n = 0, 1, 2, \ldots)$ such that $x_n \in x_0 + \mathcal{K}_n(A, r_0)$, where $\mathcal{K}_n(A, r_0) \equiv \text{span}\{r_0, Ar_0, \ldots, A^{n-1}r_0\}$ is the nth Krylov subspace generated by the matrix A from the residual $r_0 \equiv b - Ax_0$ that corresponds to the initial guess x_0. Many approaches for defining such approximations x_n have been proposed, see, e.g., the books by Greenbaum [**47**], Meurant [**72**], and Saad [**87**]. In particular, due to their smooth convergence behavior, minimum residual methods satisfying

$$\|r_n\| = \min_{\tilde{x}\in x_0+\mathcal{K}_n(A,r_0)} \|b - A\tilde{x}\|, \qquad r_n \equiv b - Ax_n, \tag{4.2}$$

are widely used, e.g., the GMRES algorithm of Saad and Schultz [**88**].

The classical implementation of GMRES makes use of a nested sequence of orthonormal bases of the Krylov subspaces $\mathcal{K}_n(A, r_0)$. These bases are generated by an Arnoldi process [**6**]. With the notation $\rho_0 \equiv \|r_0\|$, $q_1 \equiv \rho_0^{-1} r_0$, $Q_n \equiv [q_1, \ldots, q_n]$, where the columns of Q_n form this orthonormal basis of $\mathcal{K}_n(A, r_0)$, and with an $(n + 1) \times n$ upper Hessenberg matrix $H_{n+1,n}$, its result can be cast in matrix form as

$$[q_1, AQ_n] = Q_{n+1}[e_1, H_{n+1,n}].$$

This can be viewed as the QR factorization of the matrix $[q_1, AQ_n]$. Ultimately, an approximate solution x_n satisfying the minimum residual property (4.2) is constructed in the form $x_n = x_0 + Q_n y_n$, but x_n is not needed at every step.

From the relation

$$\|r_n\| = \|r_0 - AQ_n y_n\| = \|\rho_0 e_1 - H_{n+1,n} y_n\|$$

it follows that y_n is the solution of the $(n+1) \times n$ least squares problem $H_{n+1,n} y_n \approx \rho_0 e_1$, and that $\|r_n\|$ equals the norm of its residual $\rho_0 e_1 - H_{n+1,n} y_n \in \mathbb{R}^{n+1}$. This problem can be solved via the recursive QR factorization of $H_{n+1,n}$, updated by applying n Givens rotations and determining a new one in the nth step. Once the norm of the residual is small enough — which can be seen without explicitly solving the least squares problem — the triangular system with the computed R-factor is solved, and the approximate solution x_n is computed. In [**27, 48, 78**] it was shown that this "classical" version of the GMRES method is backward stable provided that the Arnoldi process is implemented using the modified Gram-Schmidt algorithm or Householder reflections.

Here we deal with a different approach proposed by Walker and Zhou [**103**], who called it the Simpler GMRES method. To derive it, we recall that the minimum residual property (4.2) is equivalent to the orthogonality condition

$$r_n \perp A\mathcal{K}_n(A, r_0),$$

where $\perp$ is the orthogonality relation induced by the standard Euclidean inner product $\langle \cdot, \cdot \rangle$. Instead of building an orthonormal basis of $\mathcal{K}_n(A, r_0)$ we look for an orthonormal basis $V_n \equiv [v_1, \ldots, v_n]$ of $A\mathcal{K}_n(A, r_0)$. As proposed by Walker and Zhou, we could construct it again by an Arnoldi process. This leads to the QR factorization

$$A[q_1, V_{n-1}] = V_n U_n, \tag{4.3}$$

where U_n is an $n \times n$ upper triangular matrix. We propose a generalization that consists in allowing to replace this Arnoldi process. Instead of using the image Av_{n-1} of the last constructed orthonormal basis vectors to extend the basis we consider any nested sequence of matrices $Z_{n-1} \equiv [z_1, \ldots, z_{n-1}]$ such that the columns of $[q_1, Z_{n-1}]$ form a basis of $\mathcal{K}_n(A, r_0)$, and we make use of Az_{n-1} to extend the basis. We may assume that the columns z_k of Z_{n-1} have unit length (and we will do so in the error analysis), but they need not be mutually orthogonal. The orthonormal basis V_n of $A\mathcal{K}_n(A, r_0)$ is thus obtained from the QR factorization of the image of $[q_1, Z_{n-1}]$:

$$A[q_1, Z_{n-1}] = V_n U_n. \tag{4.4}$$

Since $r_n \in r_0 + A\mathcal{K}_n(A, r_0) = r_0 + \mathcal{R}(V_n)$ and $r_n \perp \mathcal{R}(V_n)$, we can obtain the residual from $r_n = (I - V_n V_n^T) r_0$. Note that r_n is just the orthogonal projection

of r_0 onto the orthogonal complement of $\mathcal{R}(V_n)$. To compute it we apply the modified Gram-Schmidt method, which leads to the recursion

$$r_n = r_{n-1} - \alpha_n v_n, \qquad \alpha_n \equiv \langle r_{n-1}, v_n \rangle. \tag{4.5}$$

This recursion can be cast into a matrix relation too. Let $R_{n+1} \equiv [r_0, \ldots, r_n]$, let $D_n \equiv \mathrm{diag}(\alpha_1, \ldots, \alpha_n)$, and let $L_{n+1,n} \in \mathbb{R}^{(n+1)\times n}$ be the bidiagonal matrix with ones on the main diagonal and minus ones on the first subdiagonal; then (4.5) can be written as

$$R_{n+1} L_{n+1,n} = V_n D_n. \tag{4.6}$$

Since the columns of $[q_1, Z_{n-1}]$ are a basis of $\mathcal{K}_n(A, r_0)$, we can represent x_n in the form

$$x_n = x_0 + [q_1, Z_{n-1}] t_n, \tag{4.7}$$

so that $r_n = r_0 - A[q_1, Z_{n-1}]t_n = r_0 - V_n U_n t_n$. Due to the minimum residual property, we have $r_n \perp \mathcal{R}(V_n)$, and thus simply

$$U_n t_n = V_n^T r_0 = [\alpha_1, \ldots, \alpha_n]^T. \tag{4.8}$$

Hence, once the residual norm is small enough, we can solve this triangular system and compute $x_n = x_0 + [q_1, Z_{n-1}]t_n$. We call this general approach the *simpler approach*. It includes, as a special case, Simpler GMRES, where $Z_{n-1} \equiv V_{n-1}$. We will also be interested in the case of the residual basis $[q_1, Z_{n-1}] = [\frac{r_0}{\|r_0\|}, \ldots, \frac{r_{n-1}}{\|r_{n-1}\|}]$, which we will call SGMRES/RB, where "RB" refers to "residual basis" (this method has been recently derived and implemented also by Yvan Notay).

Recursion (4.5) reveals the connection between the simpler approach and yet another minimum residual approach. Let us set $p_n \equiv A^{-1} v_n$, $P_n \equiv [p_1, \ldots, p_n]$. Then, left-multiplying (4.5) by A^{-1} yields

$$x_n = x_{n-1} + \alpha_n p_n, \qquad \alpha_n = \langle r_{n-1}, A p_n \rangle, \tag{4.9}$$

or, in matrix form,

$$X_{n+1} L_{n+1,n} = -P_n D_n$$

with $X_{n+1} \equiv [x_0, \ldots, x_n]$. This shows that $p_n \in \mathcal{K}_n(A, r_0)$ is a direction vector: it has the direction in which one moves from x_{n-1} to x_n. The step length α_n can be determined from one of the formulas on the right-hand side of (4.5) or (4.9). Recall that it follows from the condition $\langle r_{n-1}, v_n \rangle = 0$, which enforces the minimization of $\|r_n\|$ on the line $\alpha \mapsto r_{n-1} - \alpha v_n$. So, instead of computing the coordinates t_n of $x_n - x_0$ with respect to the columns of $[q_1, Z_{n-1}]$ first, we can directly update x_n from (4.9). However, this requires that we construct the

direction vector p_n (or a scalar multiple of it). Now, note that left-multiplying (4.4) by A^{-1} yields

$$[q_1, Z_{n-1}] = P_n U_n. \tag{4.10}$$

If U_n is known from (4.4), a recursion for p_n can be extracted from this formula. Note that it has the same recurrence coefficients (stored in the columns of U_n) that are used in the Gram-Schmidt process in (4.4); so the two recursions can be run in the same loop. The obvious disadvantages of this approach is that we have to store both all the direction vectors p_n and all the original orthonormal basis vectors $v_n = Ap_n$. Moreover, any roundoff errors in U_n may have a strong effect on P_n. However, as we will see, this is the price we have to pay if we want to apply the simple and convenient 2-term update formulas (4.5) and (4.9) and spend only one matrix-vector (MV) product per step, namely Az_{n-1} in (4.4) (or Av_{n-1} in (4.3) if $Z_{n-1} \equiv V_{n-1}$). The case $Z_{n-1} \equiv V_{n-1}$ of this method was proposed in [**84**] under the name A^TA–variant of GMRES. We will use here the terminology *update approach* for this case and, more exactly, refined ORTHODIR for the particular case with $Z_{n-1} \equiv V_{n-1}$, since, as we will see, it is a refined version of the residual norm minimizing ORTHODIR algorithm [**33, 110**]. Likewise the case with $Z_{n-1} = [\frac{r_1}{\|r_1\|}, \ldots, \frac{r_{n-1}}{\|r_{n-1}\|}]$, which can be viewed as a refined version of the ORTHOMIN algorithm [**102, 110**] (or the GCR method of Elman [**30, 29**], and is identical to the GMRESR method [**101**] of van der Vorst and Vuik with the choice $u_n^{(0)} = r_n$), will be referred to as refined ORTHOMIN (see our comments below).

The refined ORTHODIR and ORTHOMIN algorithms with residual norm minimization started from the fact that the direction vectors p_n of the minimum residual method characterized by (4.2) are A^TA–orthonormal to each other: since $V_n = AP_n$, we have $P_n^TA^TAP_n = V_n^TV_n = I$. Because directions are only determined up to a scalar multiple, we might give up the normalization of V_n and choose instead $P_n^TA^TAP_n = V_n^TV_n$ to be a nonsingular diagonal matrix. So, in analogy to (4.4), we can directly compute the columns of $P_n = [p_1, \ldots, p_n]$ and U_n from (4.10), and complement this by the explicit successive evaluation of $V_n = AP_n$ (which, at the same time, serves for extending the Krylov subspace). Again, we can view (4.10) as either an Arnoldi process for an A^TA-orthogonal basis if we choose $Z_{n-1} \equiv AP_{n-1}$, or as a Gram-Schmidt implementation of a QR decomposition of $[q_1, Z_{n-1}]$ with respect to the A^TA–inner product if Z_{n-1} originates elsewhere. The case where $Z_{n-1} \equiv AP_{n-1}$, $q_1 \equiv r_0$, and U_n is unit triangular corresponds to the original ORTHODIR algorithm [**33, 110**]; the case where $Z_{n-1} \equiv [r_1, \ldots, r_{n-1}]$, $q_1 \equiv r_0$, and U_n is unit triangular yields a version

of the ORTHOMIN algorithm as proposed by Young and Jea [**110**], which was called GCR by Elman [**30**]. Despite the popularity of the name GCR we will mostly use the older name ORTHOMIN here, which also underlines the analogy to ORTHODIR. Details can also be found in [**8**] (choosing $B = A^T A$ and $C = I$ there). The cases with short-term recurrences have been treated in detail in [**59**] and [**9**].

However, what we have concealed in these descriptions is that we need a second matrix-vector product, namely Av_{n-1} in ORTHODIR and Ar_n in ORTHOMIN, to compute the coefficients of the orthogonal projection (i.e., of the Gram-Schmidt algorithm). Due to the $A^T A$–orthogonality, in ORTHODIR the relevant projection of Ap_{n-1} is $p_n = (I - P_{n-1}(AP_{n-1})^T A)Ap_{n-1}$, which with $V_{n-1} = AP_{n-1}$ may be written as $p_n = (I - P_{n-1}V_{n-1}^T A)v_{n-1}$. The new vector v_n could be instead of $v_n = (I - V_{n-1}V_{n-1}^T)Av_{n-1}$ computed directly as $v_n = Ap_n$, which requires an extra MV. An analogue consideration holds for ORTHOMIN. So, in this form, these algorithms are not competitive. Some remarks on their stability were drawn in [**47**]; we will not cover these implementations here.

The well-known remedy suggested by Vinsome [**102**] and Eisenstadt, Elman, and Schultz [**29**] consists in computing and storing both P_n and V_n. This is achieved by computing V_n with either the Arnoldi process (4.3) or with another QR decomposition of $A[r_0, r_1, \ldots, r_{n-1}]$ analogous to (4.4). But this means that up to the scaling of the bases P_n, V_n, and Z_n we return to the refined ORTHODIR and refined ORTHOMIN algorithms discussed above. The remaining difference between Vinsome's ORTHOMIN and our refined ORTHOMIN is that we normalize the residuals before orthogonalizing them, and that we use normalized direction vectors. The analog is true for the difference between the usual implementation of ORTHODIR and our refined ORTHODIR. The importance of normalizing the residuals before the orthogonalization will be seen later.

The sections of this chapter are organized as follows. In Section 1 we analyze first the maximum attainable accuracy of the simpler approach based on (4.3) or (4.4) for v_n and (4.7), (4.8) for x_n. Then we turn to the update approach based on (4.3) or (4.4) for v_n, (4.10) for p_n, and (4.9), (4.5) for x_n and r_n. To keep the text readable, we assume rounding errors only in selected, most relevant parts of the computation. The bounds presented in Theorems 4.1 and 4.2 show that the conditioning of the matrix $[q_1, Z_{n-1}]$ plays an important role in the numerical stability of these schemes. Both theorems give bounds on the maximum attainable accuracy measured by the normwise backward error. While for the simpler approach this quantity does not depend on the conditioning of

A, the bound for the update approach is proportional to $\kappa(A)$ (as we will show in our constructed numerical example, the bound is attainable). However, the dependence on $\kappa(A)$ is usually an overestimate; in practice, both the simpler and update approaches behave almost equally for the same choice of the basis. This is especially true for the relative errors of the computed approximate solutions, where we give essentially the same upper bound. The situation is completely analogous to results for the GMRES method [**88**] and the MINRES method [**79**] given by Sleijpen, van der Vorst and Modersitzki in [**92**].

In Section 2 we derive particular results for two choices of the basis $[q_1, Z_{n-1}]$. First for $[q_1, Z_{n-1}] = [q_1, V_{n-1}]$ leading to Simpler GMRES by Walker and Zhou [**103**] and to refined ORTHODIR. Then for $[q_1, Z_{n-1}] = [\frac{r_0}{\|r_0\|}, \ldots, \frac{r_{n-1}}{\|r_{n-1}\|}]$, which leads to SGMRES/RB and refined ORTHOMIN, respectively. It appears that the two choices lead to truly different behavior in the condition number of U_n, which governs the stability of the considered schemes. Since all these methods converge in a finite number of iterations, we fix the iteration index n such that $r_0 \notin AK_{n-1}(A, r_0)$, that is, the exact solution has not yet been reached. Based on this we give conditions on the linear independence of the basis $[q_1, Z_{n-1}]$. It is known that $[r_0, \ldots, r_{n-1}]$ can be rank deficient when the GMRES method stagnates (the breakdown occurs in ORTHOMIN and hence also in SGMRES/RB), while this does not happen for $[q_1, V_{n-1}]$ (Simpler GMRES and ORTHODIR are breakdown-free). On the other hand, we show that while the choice $[q_1, Z_{n-1}] = [q_1, V_{n-1}]$ leads to inherently less numerically stable schemes, the second selection $[q_1, Z_{n-1}] = [\frac{r_0}{\|r_0\|}, \ldots, \frac{r_{n-1}}{\|r_{n-1}\|}]$ gives rise to conditionally stable implementations provided we have some reasonable residual decrease. In particular, we show that the SGMRES/RB implementation is conditionally backward stable. Our theoretical results are illustrated by selected numerical experiments.

Throughout the text, we denote by $\|\cdot\|$ the Euclidean vector norm and the induced matrix norm, and by $\|\cdot\|_F$ the Frobenius norm. Moreover, for $B \in \mathbb{R}^{N\times n}$ $(N \geq n)$ of rank n, $\sigma_1(B) \geq \sigma_n(B) > 0$ are the extremal singular values of B, and $\kappa(B) = \sigma_1(B)/\sigma_n(B)$ is the spectral condition number. By I we denote the unit matrix of a suitable dimension, by e_k $(k = 1, 2, \ldots)$ its kth column, and we let $e \equiv [1, \ldots, 1]^T$. We assume the standard model of finite precision arithmetic with the unit roundoff u (see Higham [**55**] for details). In our bounds, instead of distinguishing between several constants (which are in fact polynomials in N and n that can differ from place to place), we use a generic constant c.

1. Maximum attainable accuracy of simpler and update approaches

In this section we analyze the numerical stability of the simpler and update approaches formulated in the previous section. In order to make our analysis readable, we assume that only the computations performed in (4.4), (4.8) and (4.10) are affected by rounding errors and that the computed Q-factor in the QR factorization (4.4) is close to an orthonormal matrix and has beed computed in a backward stable way. Hence we assume that the computed (orthogonal) factor V_n and the upper triangular factor U_n in the QR factorization (4.4) satisfy

$$A[q_1, Z_{n-1}] = V_n U_n + F_n, \qquad \|F_n\| \leq cu\|A\|\|[q_1, Z_{n-1}]\|, \tag{4.11}$$

and $\|V_n - \hat{V}_n\| \leq cu$, where $\hat{V}_n$ is the nearest orthonormal matrix satisfying $\hat{V}_n^T \hat{V}_n = I$. For simplicity, we will not distinguish between V_n and $\hat{V}_n$ and assume that V_n is exactly orthonormal. For details we refer to [**15, 55**]. From [**106, 55**] we have for the computed solution $\hat{t}_n$ of (4.8) that

$$(U_n + \Delta U_n)\hat{t}_n = D_n e, \qquad |\Delta U_n| \leq cu|U_n|, \tag{4.12}$$

where the absolute value and inequalities are understood component-wise. The approximation $\hat{x}_n$ to x is then computed as

$$\hat{x}_n = x_0 + [q_1, Z_{n-1}]\hat{t}_n. \tag{4.13}$$

The crucial quantity for the analysis of the maximum attainable accuracy is the gap between the true residual $b - A\hat{x}_n$ of the computed approximation and the updated residual r_n obtained from the update formula (4.5) describing the projection of the previous residual; see [**47, 52**]. In fact, once the true residual becomes negligible compared to the true one (and in the algorithms considered here it ultimately will), the gap equals the true residual divided by $\|A\|\|\hat{x}_n\|$, which therefore can be thought of as the backward error of the ultimate approximate solution $\hat{x}_n$ (after suitable normalization). Here is our basic result on this gap for the simpler approach.

THEOREM 4.1. *In the simpler approach, the gap between the true residual $b - A\hat{x}_n$ and the updated residual r_n satisfies*

$$\frac{\|b - A\hat{x}_n - r_n\|}{\|A\|\|\hat{x}_n\|} \leq cu\kappa([q_1, Z_{n-1}]) \left(1 + \frac{\|x_0\|}{\|\hat{x}_n\|}\right).$$

PROOF. From (4.13) we have $b - A\hat{x}_n = r_0 - A[q_1, Z_{n-1}]\hat{t}_n = r_0 - (V_n U_n + F_n)(U_n + \Delta U_n)^{-1} D_n e$, and (4.5) gives $r_n = r_0 - V_n D_n e$. Using the identity $I - U_n(U_n + \Delta U_n)^{-1} = \Delta U_n (U_n + \Delta U_n)^{-1}$ and the relation $[q_1, Z_{n-1}](U_n +$

$\Delta U_n)^{-1} D_n e = [q_1, Z_{n-1}]\hat{t}_n = \hat{x}_n - x_0$ we can express the gap between $b - A\hat{x}_n$ and r_n as

$$\begin{aligned} b - A\hat{x}_n - r_n &= (V_n - (V_n U_n + F_n)(U_n + \Delta U_n)^{-1}) D_n e \\ &= (V_n \Delta U_n + F_n)(U_n + \Delta U_n)^{-1} D_n e \\ &= (V_n \Delta U_n + F_n)[q_1, Z_{n-1}]^{\dagger}[q_1, Z_{n-1}](U_n + \Delta U_n)^{-1} D_n e \\ &= (V_n \Delta U_n + F_n)[q_1, Z_{n-1}]^{\dagger}(\hat{x}_n - x_0). \end{aligned} \tag{4.14}$$

Taking the norm, considering (4.11), and noting that the terms involving $V_n \Delta U_n$ and F_n can be subsumed into the generic constant c, we get

$$\|b - A\hat{x}_n - r_n\| \leq cu\|A\|\|[q_1, Z_{n-1}]\|\|[q_1, Z_{n-1}]^{\dagger}\|(\|\hat{x}_n\| + \|x_0\|). \tag{4.15}$$

Division by $\|A\|\|\hat{x}_n\|$ concludes the proof. □

In the following we analyze the maximum attainable accuracy of the update approach. In accordance with (4.11) we assume that in finite precision arithmetic the computed direction vectors satisfy

$$[q_1, Z_{n-1}] = P_n U_n + G_n, \qquad \|G_n\| \leq cu\|P_n\|\|U_n\|. \tag{4.16}$$

Note that the norm of the matrix G_n cannot be bounded by $cu\|A\|\|[q_1, Z_{n-1}]\|$ as it is in the case of the QR factorization (4.11). As in (4.9) we compute then the approximate solution $\hat{x}_n$ as

$$\hat{x}_n = \hat{x}_{n-1} + \alpha_n p_n. \tag{4.17}$$

THEOREM 4.2. *In the update approach, the gap between the true residual $b - A\hat{x}_n$ and the updated residual r_n satisfies*

$$\frac{\|b - A\hat{x}_n - r_n\|}{\|A\|\|\hat{x}_n\|} \leq cu\kappa(A)\kappa([q_1, Z_{n-1}]) \left(1 + \frac{\|x_0\|}{\|\hat{x}_n\|}\right),$$

provided that $\eta_n \equiv 1 - cu\kappa(A)\kappa([q_1, Z_{n-1}]) > 0$.

PROOF. Since $\hat{x}_n = x_0 + P_n D_n e = x_0 + ([q_1, Z_{n-1}] - G_n)U_n^{-1} D_n e$ and $r_n = r_0 - V_n D_n e$, we have that

$$\begin{aligned} b - A\hat{x}_n - r_n &= (V_n - A[q_1, Z_{n-1}]U_n^{-1}) D_n e + AG_n U_n^{-1} D_n e \\ &= (-F_n + AG_n) U_n^{-1} D_n e \end{aligned} \tag{4.18}$$

due to (4.4). From (4.4) and (4.16), we get $P_n = A^{-1}V_n + (A^{-1}F_n - G_n)U_n^{-1}$. Taking a norm we obtain $\|P_n\| \leq \|A^{-1}\| + cu\kappa(A)\|U_n^{-1}\| + cu\|P_n\|\kappa(U_n)$. The

norm of the residual matrix G_n in (4.16) can hence be estimated as

$$\|G_n\| \leq cu\kappa(A)\|[q_1, Z_{n-1}]\|. \tag{4.19}$$

Owing to (4.17), we have the identity $U_n^{-1}D_ne = U_n^{-1}P_n^{\dagger}P_nD_ne = U_n^{-1}P_n^{\dagger}(\hat{x}_n - x_0)$, and $\|U_n^{-1}P_n^{\dagger}\| \leq \eta_n^{-1}\|[q_1, Z_{n-1}]^{\dagger}\|$ following from (4.16). Thus we obtain

$$\|U_n^{-1}D_ne\| \leq \eta_n^{-1}\|[q_1, Z_{n-1}]^{\dagger}\|(\|\hat{x}_n\| + \|x_0\|), \tag{4.20}$$

which together with (4.18), (4.19), and (4.11) proves the statement of the theorem. □

The bound on the ultimate backward error given in Theorem 4.2 is worse that the one of Theorem 4.1. We see that for the simpler approach the normwise backward error is on the order of the roundoff unit, whereas for the update approach we have an upper bound proportional to the condition number of A. In terms of the residual norms, this leads to the bounds involving $cu\kappa(A)\kappa([q_1, Z_{n-1}])$ and $cu\kappa^2(A)\kappa([q_1, Z_{n-1}])$ terms for the simpler and update approach, respectively.

From Theorems 4.1 and 4.2, we can also estimate the ultimate level of the relative 2-norm of the error of both the simpler and update approach. However, as shown below, it appears that the update approach leads to the approximate solution with essentially the same accuracy level in the error as the simpler approach. Similar phenomenon was also observed by Sleijpen, van der Vorst and Modersitzki [**92**] in the symmetric case for GMRES and MINRES.

COROLLARY 4.3. *The gap between the computed approximate solutions $\hat{x}_n$ and exact approximations x_n in both the simpler ($x_n = x_0 + [q_1, Z_{n-1}]t_n$) and update ($x_n = x_{n-1} + \alpha_n p_n$) approaches can be bounded by*

$$\frac{\|x_n - \hat{x}_n\|}{\|x\|} \leq cu\kappa(A)\kappa([q_1, Z_{n-1}])\frac{\|\hat{x}_n\| + \|x_0\|}{\|x\|}, \tag{4.21}$$

provided that $\eta_n \equiv 1 - cu\kappa(A)\kappa([q_1, Z_{n-1}]) > 0$.

PROOF. For the simpler approach, the result follows directly from Theorem 4.1. For the update approach, using (4.18) we have

$$x_n - \hat{x}_n = x - \hat{x}_n - A^{-1}r_n = (-A^{-1}F_n + G_n)U_n^{-1}D_ne$$

and the statement now follows from (4.11), (4.19) and (4.20). □

The bound (4.21) from Corollary 4.3 depends on the quantity $(\|\hat{x}_n\| + \|x_0\|)/\|x\|$ (or more precisely on $\|\hat{x}_n - x_0\|/\|x\|$), which is, however, strongly influenced by the conditioning of the upper triangular matrix U_n. As shown in Section 2,

the matrix U_n can be ill-conditioned for a particular case $[q_1, Z_{n-1}] = [q_1, V_{n-1}]$ leading thus to inherently less numerically stable schemes, whereas the schemes with $[q_1, Z_{n-1}] = [\frac{r_0}{\|r_0\|}, \ldots, \frac{r_{n-1}}{\|r_{n-1}\|}]$ under some assumptions give rise to the well-conditioned triangular matrix U_n. In the following lemma we give bounds on $\|\hat{x}_n - x_0\|$ in terms of the singular values of the matrix U_n.

LEMMA 4.4. *In the simpler approach, we have*

$$\|\hat{x}_n - x_0\| \leq \|[q_1, Z_{n-1}]\|\|\hat{t}_n\| \leq \|[q_1, Z_{n-1}]\|\|(U_n + \Delta U_n)^{-1} D_n e\|,$$

and in the update approach,

$$\|\hat{x}_n - x_0\| \leq \|P_n D_n e\| \leq (1 + cu\kappa(A))\|[q_1, Z_{n-1}]\|\|U_n^{-1} D_n e\|.$$

The norms of $(U_n + \Delta U_n)^{-1} D_n e$ and $U_n^{-1} D_n e$ satisfy

$$\left.\begin{array}{l} \|(U_n + \Delta U_n)^{-1} D_n e\| \\ \|U_n^{-1} D_n e\| \end{array}\right\} \leq \sqrt{2} \sum_{k=1}^{n} \frac{\|r_{k-1}\|}{\sigma_k(U_k)} \leq \sqrt{2}\|A^{-1}\| \sum_{k=1}^{n} \frac{\eta_k^{-1}\|r_{k-1}\|}{\sigma_k([q_1, Z_{k-1}])}, \tag{4.22}$$

provided that $\eta_k \equiv 1 - cu\kappa(A)\kappa([q_1, Z_{k-1}]) > 0$ for all $k = 1, \ldots, n$.

PROOF. Since $e_k^T D_n e_k = \alpha_k$ and $|\alpha_k| = \sqrt{\|r_{k-1}\|^2 - \|r_k\|^2} \leq \sqrt{2}\|r_{k-1}\|$, we have

$$\|(U_n + \Delta U_n)^{-1} D_n e\| \leq \sum_{k=1}^{n} \|(U_n + \Delta U_n)^{-1} D_n e_k\| \leq \sqrt{2} \sum_{k=1}^{n} \frac{\|r_{k-1}\|}{\sigma_k([U_n + \Delta U_n]_{1:k,1:k})}, \tag{4.23}$$

where $[U_n + \Delta U_n]_{1:k,1:k}$ denotes the principal $k \times k$ submatrix of $U_n + \Delta U_n$. Owing to (4.12), we can estimate the perturbation of $[U_n]_{1:k,1:k} = U_k$ as $\|[\Delta U_n]_{1:k,1:k}\| \leq cu\|U_k\|$. Perturbation theory of singular values shows that

$$\begin{aligned} \sigma_k([U_n + \Delta U_n]_{1:k,1:k}) &\geq \sigma_k(U_k) - cu\|U_k\| \\ &\geq \sigma_k(A[q_1, Z_{k-1}]) - cu\|A\|\|[q_1, Z_{k-1}]\| \\ &\geq \sigma_N(A)\sigma_k([q_1, Z_{k-1}]) - cu\|A\|\|[q_1, Z_{k-1}]\|, \end{aligned} \tag{4.24}$$

which, together with (4.23), concludes the proof of the first inequality. The second inequality is proved analogously. □

The first estimate given in (4.22), which involves the minimal singular values of U_k $(k = 1, \ldots, n)$, is quite sharp. However, the second estimate relating the minimal singular values of U_k to those of $[q_1, Z_{k-1}]$ can be a large overestimate, as also observed in our numerical experiments in Section 2. Using Lemma 4.4 we can give the following estimates for the gap between the true and updated residuals in the simpler and update approaches.

COROLLARY 4.5. *In the simpler approach, the gap between the true residual* $\|b - A\hat{x}_n\|$ *and the updated residual* r_n *satisfies*

$$\|b - A\hat{x}_n\| \leq cu\kappa(A)\|[q_1, Z_{n-1}]\| \sum_{k=1}^{n} \frac{\eta_k^{-1}\|r_{k-1}\|}{\sigma_k([q_1, Z_{k-1}])}.$$

In the update approach, the same quantity can be estimated as

$$\|b - A\hat{x}_n\| \leq cu\kappa^2(A)\|[q_1, Z_{n-1}]\| \sum_{k=1}^{n} \frac{\eta_k^{-1}\|r_{k-1}\|}{\sigma_k([q_1, Z_{k-1}])}.$$

Theorems 4.1 and 4.2 indicate that as soon as the backward error of the approximate solution in the simpler approach gets below $cu\kappa(A)\kappa([q_1, Z_{n-1}])$, then the difference between the backward errors in the simpler and update approaches may become visible and can be expected to be up to the order of $\kappa(A)$. Based on our experience it is difficult to find an example where this difference is significant. Similarly to Sleijpen, van der Vorst and Modersitzki [**92**], we use here a model example, where $A = G_1 D G_2^T \in \mathbb{R}^{100\times 100}$ with $D = \mathrm{diag}(10^{-8}, 2\cdot 10^{-8}, 3, 4, \ldots, 100)$ and with G_1 and G_2 being Givens rotations over an angle of $\frac{\pi}{4}$ in the $(1, 10)$-plane and the $(1, 100)$-plane, respectively; finally, $b = e$. The numerical experiments were performed in MATLAB using double precision arithmetic ($u \approx 10^{-16}$), and the zero vector was chosen as the initial guess x_0. In Figure 4.1 we have plotted the normwise backward errors $\|b - A\hat{x}_n\|/(\|A\|\|\hat{x}_n\|)$ (solid lines), relative 2-norms of the residuals $\|b - A\hat{x}_n\|/\|b\|$ (dashed lines) and the relative 2-norms of the errors $\|x - \hat{x}_n\|/\|x\|$ (dash-dotted lines) for Simpler GMRES and refined ORTHODIR, respectively. The same quantities for SGMRES/RB and refined ORTHOMIN are reported in Figure 4.2. We see that the actual backward errors and relative residual norms are close until where they stagnate: for refined ORTHODIR and refined ORTHOMIN this happens approximately at a level close to $u\kappa(A)$ for the backward errors and $u\kappa^2(A)$ for the residuals, while for Simpler GMRES and SGMRES/RB we have stagnation on the roundoff unit level. In contrast, the 2-norms of the errors stagnate on the $u\kappa(A)$ level in all considered schemes.

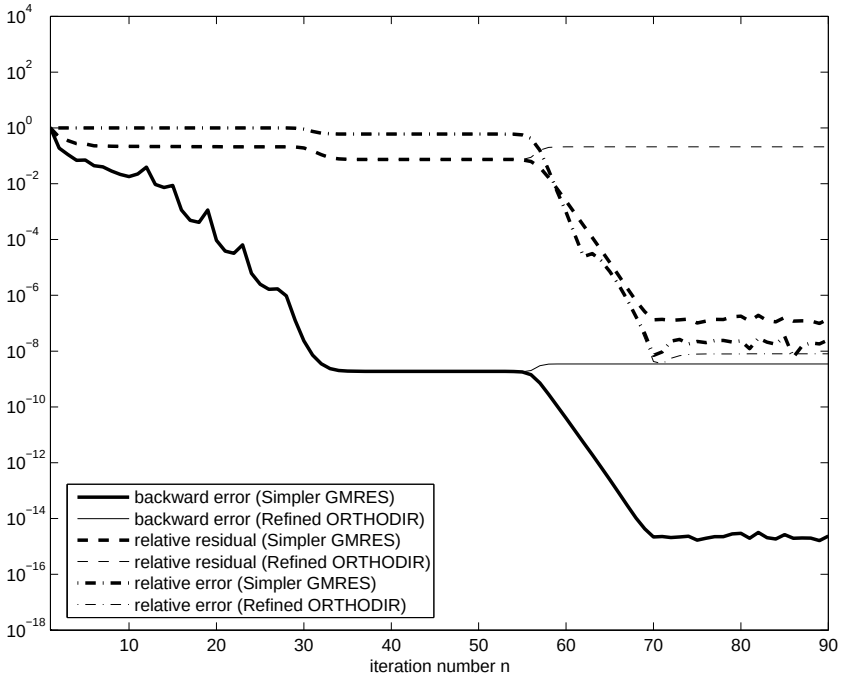


FIGURE 4.1. The test problem solved by Simpler GMRES and refined ORTHODIR.

2. Choice of basis and numerical stability

In this section we discuss the two main particular choices for the matrix Z_{n-1} leading to different algorithms for the simpler and update schemes. For the sake of simplicity, we assume exact arithmetic here. First, we choose $Z_{n-1} = V_{n-1}$, which leads to the Simpler GMRES method of Walker and Zhou [**103**] and to the refined version of ORTHODIR by Young and Jea [**110**], respectively. Hence, we choose $\{q_1, v_1, \ldots, v_{n-1}\}$ as a basis of $\mathcal{K}_n(A, r_0)$. To be sure that such a choice is adequate, we state the following simple lemma.

LEMMA 4.6. *Let $v_1, \ldots, v_{n-1}$ be an orthonormal basis of $A\mathcal{K}_{n-1}(A, r_0)$ and let $r_0 \notin A\mathcal{K}_{n-1}(A, r_0)$. Then the vectors $q_1, v_1, \ldots, v_{n-1}$ form a basis of $\mathcal{K}_n(A, r_0)$.*

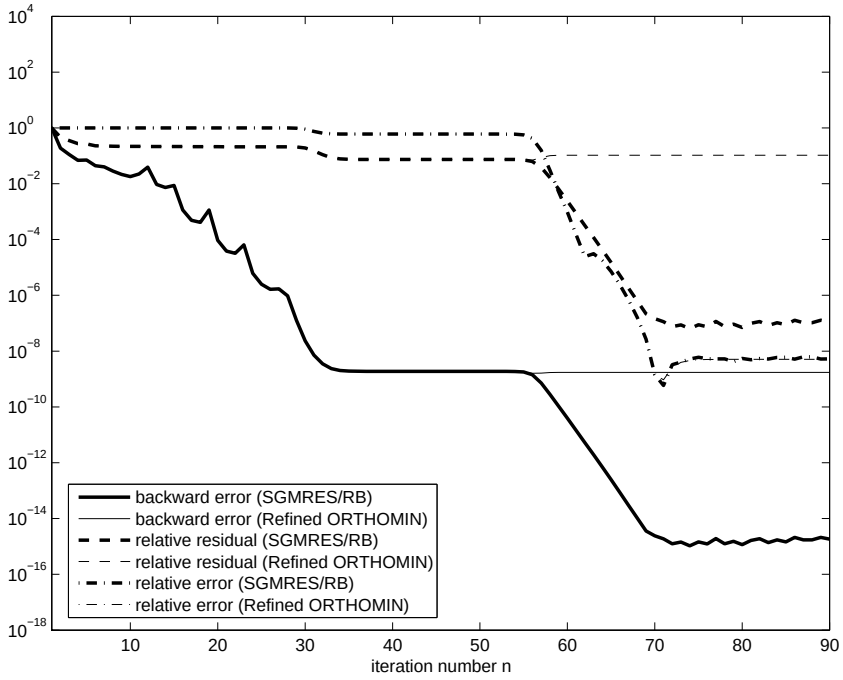


FIGURE 4.2. The test problem solved by SGMRES/RB and refined ORTHOMIN.

PROOF. It follows from the assumption $r_0 \notin A\mathcal{K}_{n-1}(A, r_0)$ implying that $q_1 \notin A\mathcal{K}_{n-1}(A, r_0) = \text{span}\{v_1, \ldots, v_{n-1}\}$. □

Note that if $r_0 \in A\mathcal{K}_n(A, r_0)$, then the condition (4.2) yields $x_n = A^{-1}b$, $r_n = 0$, and any implementation of a minimum residual method will terminate. Lemma 4.6 ensures that it makes sense to build an orthonormal basis V_n of $A\mathcal{K}_n(A, r_0)$ by the successive orthogonalization of the columns of the matrix $A[q_1, V_{n-1}]$ via (4.4). It reflects the fact that, for any initial residual r_0, both Simpler GMRES and ORTHODIR converge (in exact arithmetic) to the exact solution; see [**110**]. However, as observed by Liesen, Rozložník and Strakoš [**66**], this choice of the basis is not very suitable from the stability point of view. This shortcoming is reflected by the unbounded growth of the condition number of $[q_1, V_{n-1}]$ discussed next. The upper bound was also derived in the paper [**103**].

THEOREM 4.7. *Let $r_0 \notin AK_{n-1}(A, r_0)$. Then the condition number of $[q_1, V_{n-1}]$ satisfies*

$$\frac{\|r_0\|}{\|r_{n-1}\|} \leq \kappa([q_1, V_{n-1}]) \leq 2\frac{\|r_0\|}{\|r_{n-1}\|}.$$

PROOF. Since $r_{n-1} = (I - V_{n-1}V_{n-1}^T)r_0$, it is easy to see that r_{n-1} is the residual of the least squares problem $V_{n-1}y \approx r_0$. The statement follows from Theorem 3.2 of [**66**]. □

The conditioning of $[q_1, V_{n-1}]$ is thus related to the convergence of the method; in particular, it is inversely proportional to the actual relative norm of the residual. Hence, if the residual is small enough, Simpler GMRES and refined ORTHODIR behave unstably. In practice, this difficulty can be counteracted by frequent restarts.

Now we turn to the second choice, $Z_{n-1} = [\frac{r_1}{\|r_1\|}, \ldots, \frac{r_{n-1}}{\|r_{n-1}\|}]$, which leads to SGMRES/RB (which we propose here as a more stable counterpart of Simpler GMRES) and to the refined version of ORTHOMIN by Vinsome [**102**] known also under the name GCR; see Eisenstat, Elman and Schultz [**30, 29**]. We have $[q_1, Z_{n-1}] = R_n B_n^{-1}$, where $B_n \equiv \text{diag}(\|r_0\|, \ldots, \|r_{n-1}\|)$, i.e., we choose scaled residuals $r_0, \ldots, r_{n-1}$ as the basis of $\mathcal{K}_n(A, r_0)$. To be sure that such a choice is adequate, we state the following result.

LEMMA 4.8. *Let $v_1, \ldots, v_{n-1}$ be an orthonormal basis of $AK_{n-1}(A, r_0)$ and let $r_0 \notin AK_{n-1}(A, r_0)$ and $r_k = (I - V_k V_k^T)r_0$, where $V_k \equiv [v_1, \ldots, v_k]$, $k = 1, 2, \ldots, n-1$. Then the following statements are equivalent:*

(1) *$\|r_k\| < \|r_{k-1}\|$ for all $k = 1, \ldots, n-1$,*
(2) *$r_0, \ldots, r_{n-1}$ are linearly independent.*

PROOF. Since $r_0 \notin AK_{n-1}(A, r_0) = \mathcal{R}(V_{n-1})$, $r_k \neq 0$ for all $k = 0, 1, \ldots, n-1$. It is clear that $\|r_k\| < \|r_{k-1}\|$ if and only if $\langle r_{k-1}, v_k \rangle \neq 0$. If that holds for all $k = 1, \ldots, n-1$ the diagonal matrix D_{n-1} is nonsingular. Using the relation (4.6) we find that $R_n[L_{n,n-1}, e_n] = [V_{n-1}D_{n-1}, r_{n-1}]$. Since $r_{n-1} \perp V_{n-1}$, the matrix $[V_{n-1}D_{n-1}, r_{n-1}]$ has orthogonal nonzero columns, and hence its rank equals n. Moreover, $\text{rank}([L_{n,n-1}, e_n]) = n$ and thus $\text{rank}(R_n) = n$, i.e., $r_0, \ldots, r_{n-1}$ are linearly independent. Conversely, from the same matrix relation we find that if $r_0, \ldots, r_{n-1}$ are linearly independent, then $\text{rank}([V_{n-1}D_{n-1}, r_{n-1}]) = n$, and hence D_{n-1} is nonsingular, which proves that $\|r_k\| < \|r_{k-1}\|$ for all $k = 1, \ldots, n-1$. □

Therefore if the method does not stagnate, i.e., if the 2-norms of the residuals $r_0, \ldots, r_{n-1}$ are strictly monotonously decreasing, then $r_0, \ldots, r_{n-1}$ are linearly independent. In this case, we can build an orthonormal basis V_n of $A\mathcal{K}_n(A, r_0)$ by the successive orthogonalization of the columns of $AR_nB_n^{-1}$ via (4.4). If $r_0 \in A\mathcal{K}_{n-1}(A, r_0)$, we have an exact solution of (4.1), and the method stops with $x_{n-1} = A^{-1}b$.

Several conditions for the non-stagnation of the minimum residual method have been given in the literature. For example, Eisenstat, Elman and Schultz [**29, 30**] show that GCR (and hence any minimum residual method) does not stagnate if the symmetric part of A is positive definite, i.e., if the origin is not contained in the field of values of A. See also Greenbaum and Strakoš [**50**] for a different proof, and Eiermann and Ernst [**28**]. Several other conditions can be found in Simoncini and Szyld [**91**] and the references therein. If stagnation occurs, the residuals are no longer linearly independent, and thus the method prematurely breaks down. In particular, if $0 \in \mathcal{F}(A)$, choosing x_0 such that $r_0 \in \mathcal{F}(A)$ leads to a breakdown in the first step. This was first pointed out by Young and Jea [**110**] with a simple 2×2 example.

However, as shown in the following theorem, when the minimum residual method does not stagnate, the columns of $R_nB_n^{-1}$ are a reasonable choice for the basis of $\mathcal{K}_n(A, r_0)$.

THEOREM 4.9. *If $r_0 \notin A\mathcal{K}_{n-1}(A, r_0)$, the condition number of $R_nB_n^{-1}$ satisfies*

$$1 \leq \kappa(R_nB_n^{-1}) \leq \sqrt{n}\,\gamma_n, \qquad \gamma_n \equiv \sqrt{1 + \sum_{k=1}^{n-1} \frac{\|r_{k-1}\|^2 + \|r_k\|^2}{\|r_{k-1}\|^2 - \|r_k\|^2}}. \tag{4.25}$$

PROOF. From (4.6) it follows that

$$R_nB_n^{-1}[Q_{n,n-1}, e_n] = [V_{n-1}, \tfrac{r_{n-1}}{\|r_{n-1}\|}], \qquad Q_{n,n-1} \equiv B_nL_{n,n-1}D_{n-1}^{-1}.$$

Since $[V_{n-1}, \frac{r_{n-1}}{\|r_{n-1}\|}]$ is an orthonormal matrix, we have from Theorem 3.3.16 of [**58**]

$$\begin{aligned} 1 &= \sigma_n([V_{n-1}, \tfrac{r_{n-1}}{\|r_{n-1}\|}]) \leq \sigma_n(R_nB_n^{-1})\|[Q_{n,n-1}, e_n]\| \\ &\leq \sigma_n(R_nB_n^{-1})\|[Q_{n,n-1}, e_n]\|_F. \end{aligned}$$

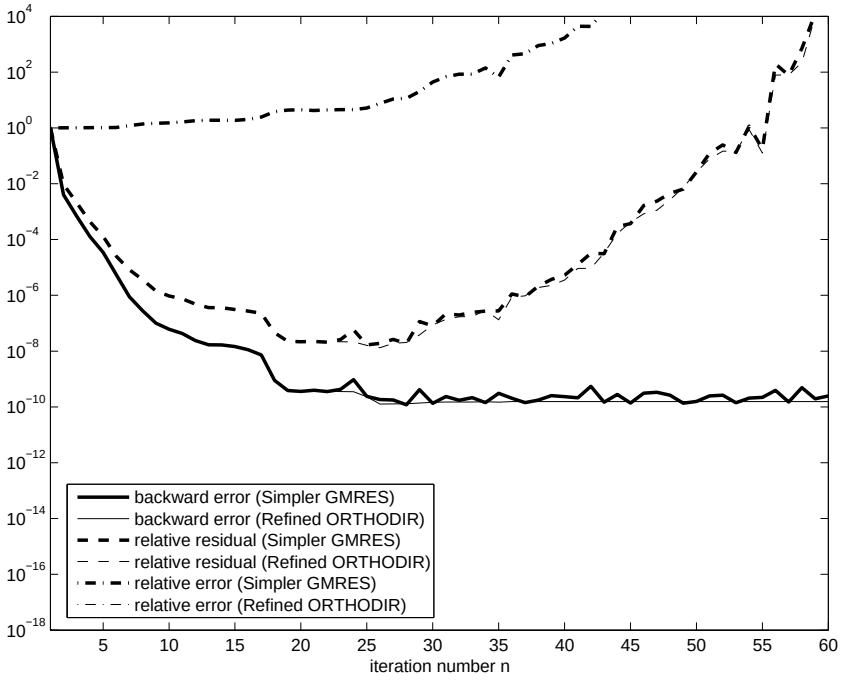


FIGURE 4.3. The test problem FS1836 solved by Simpler GMRES and refined ORTHODIR.

The value of $\|[Q_{n,n-1}, e_n]\|_F$ can be directly computed as

$$\|[Q_{n,n-1}, e_n]\|_F = \sqrt{1 + \sum_{k=1}^{n-1} \frac{\|r_{k-1}\|^2 + \|r_k\|^2}{\|r_{k-1}\|^2 - \|r_k\|^2}},$$

since $\alpha_k^2 = \|r_{k-1}\|^2 - \|r_k\|^2$. The statement follows using $\|R_n B_n^{-1}\| \leq \sqrt{n}$. □

We define the quantity γ_n in (4.25) as the *stagnation factor*. The conditioning of $R_n B_n^{-1}$ is thus related to the convergence of the method, but in contrast to the conditioning of $[q_1, V_{n-1}]$, it is related to the intermediate decrease of the residual norms, not to the residual decrease with respect to the initial residual.

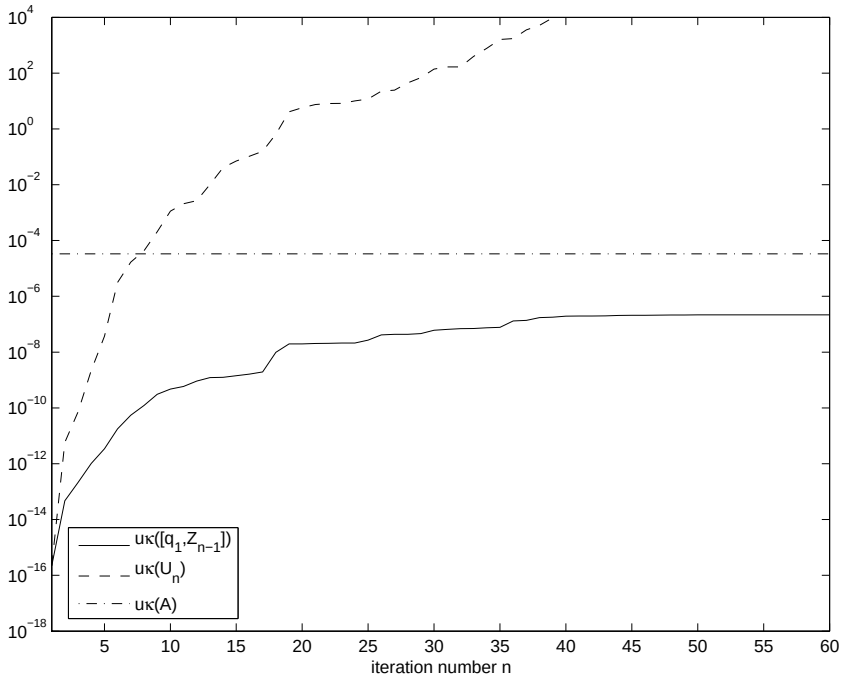


FIGURE 4.4. The test problem FS1836 solved by Simpler GMRES and refined ORTHODIR.

We illustrate our theoretical results by a numerical example using the matrix FS1836 ($\|A\| \approx 1.18 \cdot 10^9$, $\|A^{-1}\| \approx 1.47 \cdot 10^2$) obtained from the Matrix Market [**1**] with the right-hand side $b = Ae$ (see also the experiments in [**66**], where the relative residual norms were reported). In Figures 4.3 and 4.5, we show the normwise backward error $\|b - A\hat{x}_n\|/(\|A\|\|\hat{x}_n\|)$ (solid lines), relative 2-norms of the residuals $\|b - A\hat{x}_n\|/\|b\|$ (dashed lines) and relative 2-norms of the error $\|x - \hat{x}_n\|/\|x\|$ (dotted lines with circles and crosses) for the choice $[q_1, Z_{n-1}] = [q_1, V_{n-1}]$ that corresponds to Simpler GMRES and refined ORTHODIR, and for $[q_1, Z_{n-1}] = [\frac{r_0}{\|r_0\|}, \ldots, \frac{r_{n-1}}{\|r_{n-1}\|}]$ corresponding to SGMRES/RB and refined ORTHOMIN, respectively. The quantities $u\kappa([q_1, Z_{n-1}])$, $u\kappa(U_n)$ and $u\kappa(A)$ are depicted by solid, dashed and dash-dotted lines in Figures 4.4 and 4.6. We see that the backward errors, the residual norms, and the error norms are

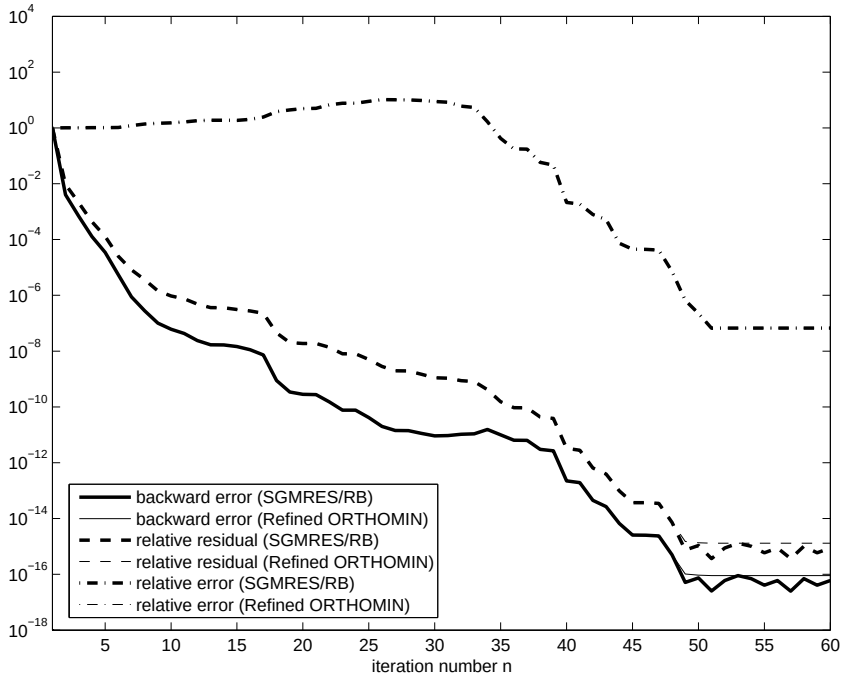


FIGURE 4.5. The test problem FS1836 solved by SGMRES/RB and refined ORTHOMIN.

almost identical for the simpler and update approaches. This can be observed in most cases leading to practically negligible difference between Simpler GMRES and refined ORTHODIR, and SGMRES/RB and refined ORTHOMIN, respectively. Figure 4.3 illustrates our theoretical considerations and shows that, after some initial reduction, the backward error (or residual norm) of Simpler GMRES and refined ORTHODIR may stagnate on a significantly higher level than the backward error (or residual norm) of SGMRES/RB or refined ORTHOMIN, which stagnates on a level proportional to the roundoff unit, as shown in Figure 4.5. Due to Theorem 4.7, after some initial phase, the norms of errors (as well as residuals) start to diverge in Simpler GMRES and refined ORTHODIR, while for SGMRES/RB and refined ORTHOMIN we have a stagnation on a level approximately proportional to $u\kappa(A)$. The difference is clearly caused by the choice of

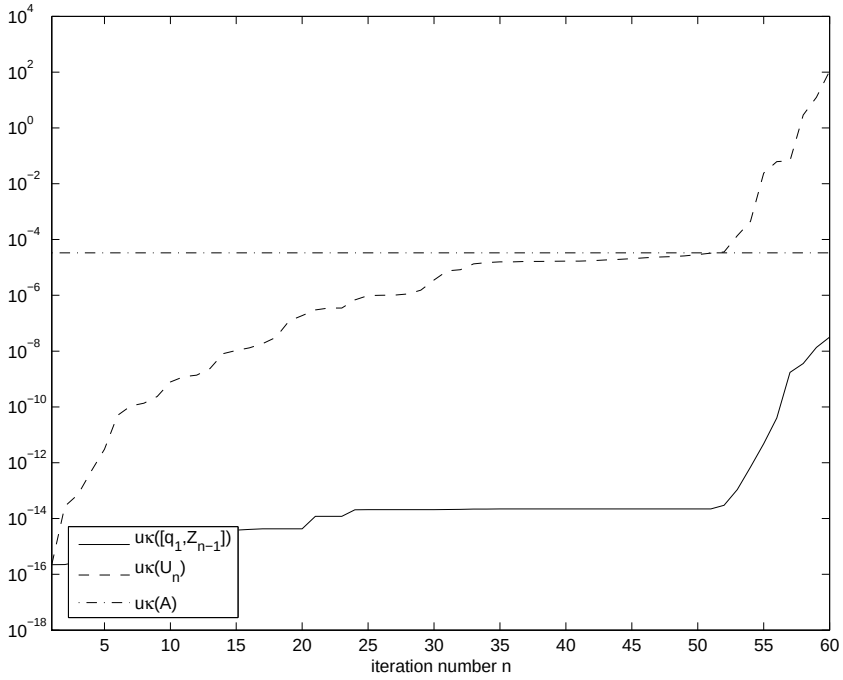


FIGURE 4.6. The test problem FS1836 solved by SGMRES/RB and refined ORTHOMIN.

the basis $[q_1, Z_{n-1}]$, which has an effect on the conditioning of the matrix U_n. We see that $[q_1, Z_{n-1}] = [\frac{r_0}{\|r_0\|}, \ldots, \frac{r_{n-1}}{\|r_{n-1}\|}]$ remains well-conditioned up to the very end of the iteration process, while the conditioning of $[q_1, V_{n-1}]$ is linked to the convergence of Simpler GMRES and may lead to a very ill-conditioned triangular matrix U_n. Consequently the approximate solution $\hat{x}_n$ computed from (4.8) becomes inaccurate and its error starts to diverge. Since the stagnation factor $\gamma_n \approx 55.8$ (for $n = 50$), the matrix U_n remains well-conditioned, and this problem does not occur in the SGMRES/RB method.

CHAPTER 5

Conclusions and open questions

In this thesis we studied the numerical behavior of several iterative methods for the solution of systems of linear algebraic equations. In Chapter 3 we looked at the numerical behavior of certain inexact saddle point solvers. In particular, for several mathematically equivalent implementations, we studied the influence of inexact solution of inner systems and estimate their maximum attainable accuracy. When considering the outer iteration process, our analysis lead to results similar to ones which can be obtained assuming exact arithmetic. The situation was different, when we looked at the residuals in the saddle point system. We showed that some implementations lead ultimately to residuals on the level of roundoff unit independently on the fact that the inner systems were solved inexactly. Indeed, our results confirm that the generic and actually the cheapest implementations deliver the approximate solutions, which satisfy either the second or the first block equation to the working accuracy. In addition, the implementations with corrected direct substitution are also very attractive. We gave a theoretical explanation for the behavior which was probably observed or is already tacitly known. The implementations that we point out as optimal are actually those, which are widely used and suggested in applications. It appears that, when measured in terms of the errors, the maximum attainable accuracy level is similar for all considered implementations and it is proportional to the backward error tolerance of inner systems.

In Chapter 4 we studied the numerical behavior of several minimum residual methods mathematically equivalent to GMRES. Two general formulations were analyzed: the simpler approach that does not require an upper Hessenberg factorization and the update approach which is based on generating a sequence of appropriately computed direction vectors. It was shown that for the simpler approach our analysis leads to an upper bound for the backward error proportional to the roundoff unit, whereas for the update approach the same quantity can be bounded by a term proportional to the condition number of A. Although our

analysis suggests that there maybe a difference between both approaches up to the order of $\kappa(A)$, in practice they behave very similarly and it is very difficult to find an example with a significant difference in the limiting accuracy. Moreover, when looking at the errors, we note that both approaches lead essentially to the same accuracy of the computed approximate solutions.

We indicated that the choice of the basis $[q_1, Z_{n-1}]$ is the most important issue for the stability of the considered schemes. Our analysis supports the well-known fact that, even when implemented with the best possible orthogonalization techniques, Simpler GMRES and ORTHODIR are inherently less stable due to the choice $[q_1, Z_{n-1}] = [q_1, V_{n-1}]$. The situation becomes significantly better, when we use the residual basis $[q_1, Z_{n-1}] = [\frac{r_0}{\|r_0\|}, \ldots, \frac{r_{n-1}}{\|r_{n-1}\|}]$. This choice leads to the popular GCR, ORTHOMIN and GMRESR methods, which are widely used in applications. Assuming some reasonable residual decrease (which happens almost always in finite precision arithmetic), we showed that this scheme is quite efficient and proposed a conditionally backward stable variant (called SGMRES/RB here). Our theoretical results in a sense justify the use of the GCR method in practical computations.

There are several open problems connected to the topic of this thesis.

Various stopping criteria for inner systems. The analysis in Chapter 3 is based on the backward error stopping criterion in inner systems. It could be interesting to compare other stopping criteria based, e.g., on the relative residuals or estimates of energy errors in the Schur complement method. The relation between the A-norm of $x - x_k$ and the $B^T A^{-1} B$-norm of $y - y_k$ can lead to a stopping criterion based on the energy norm of $x - x_k$. However, it is not completely clear how to do this, when the systems with A are not solved exactly.

Corrected substitution in stationary iterative methods. We saw in Chapter 3 that for the Schur complement reduction and null-space projection methods, it is more preferable to update the approximation x_{k+1} using the corrected direct substitution than to compute it directly. Analogous results hold also for stationary iterative methods. Consider the system $Ax = b$ with a nonsingular matrix A and its splitting $A = M - N$, where M is also nonsingular. A stationary iterative method then generates the approximations to x satisfying $Mx_{k+1} = Nx_k + b$ starting from some x_0. Higham and Knight [**56**] analyzed this implementation in finite precision arithmetic, and they showed that the limiting accuracy depends on the maximum relative norm of the approximate solutions $\bar{x}_i$ $(i = 0, \ldots, k)$. However, it is much more beneficial, in such a case,

rather than compute $x_{k+1} = M^{-1}(Nx_k + b)$, to use the "corrected" formula $x_{k+1} = x_k + M^{-1}r_k$, where $r_k = b - Ax_k$. We saw in Section 1.4 of Chapter 3 that the final level of the residual $f - A\bar{x}_k - B\bar{y}_k$ does not depend on the maximum norm of the iterates during the whole iteration process but only on those in a few last iterations. The similar observation can be made also in the case of the "corrected" implementation of the stationary iteration, and the idea can be also extended to two-stage iterative methods, e.g., when applying the SIMPLE method for the solution of fluid flow problems (see, e.g., [**81**]).

Backward error analysis of segregated methods. In Section 4 of Chapter 3 we interpret the inexact solution computed with the Schur complement reduction method (using the generic update) as an exact solution of the saddle point problem with a perturbed upper-left matrix block. The similar backward error analysis should be performed also for other implementations of the Schur complement reduction method and for the null-space projection method. Moreover, the analysis of the null-space projection should consider also a particular projection method for computing the direction vectors.

Preconditioned residual basis. In Chapter 4, we did not consider the issue of preconditioning or, we assume, that the system $Ax = b$ is already preconditioned. It does not make much sense to precondition the methods using the basis $[q_1, V_{n-1}]$ such as Simpler GMRES or ORTHODIR due to their inherent instability. One can restart the method to overcome this problem, but note that the restart is necessary when the method becomes unstable, i.e., when it converges fast! It seems reasonable to use (fixed or flexible) preconditioning in the case of the residual basis (the preconditioned SGMRES/RB and GCR). It is sometimes observed that the preconditioned residual basis of GCR (i.e., GMRESR [**101**]) is more preferable than, e.g., preconditioned GMRES (with a fixed preconditioner) or flexible GMRES [**86**], which use the preconditioned orthonormal basis of $\mathcal{K}_n(A, r_0)$. Moreover, faster convergence could be observed when using preconditioned residuals. This issue needs to be analyzed further.

Bibliography

[1] Matrix Market. URL: `http://math.nist.gov/MatrixMarket`.

[2] M. Arioli. The use of QR factorization in sparse quadratic programming and backward error issues. *SIAM J. Matrix Anal. Appl.*, 21(3):825–839, 2000.

[3] M. Arioli and L. Baldini. A backward error analysis of a null space algorithm in sparse quadratic programming. *SIAM J. Matrix Anal. Appl.*, 23(2):425–442, 2001.

[4] M. Arioli and C. Fassino. Roundoff error analysis of algorithms based on Krylov subspace methods. *BIT*, 36(2):189–206, 1996.

[5] M. Arioli and F. Romani. Stability, convergence, and conditioning of stationary iterative methods of the form $x^{(i+1)} = Px^{(i)} + q$ for the solution of linear systems. *IMA J. Numer. Anal.*, 12:21–30, 1992.

[6] W. E. Arnoldi. The principle of minimized iterations in the solution of the matrix eigenvalue problem. *Quart. Appl. Math.*, 9:17–29, 1951.

[7] K. Arrow, L. Hurwicz, and H. Uzawa. *Studies in Nonlinear Programming*. Stanford University Press, Stanford, CA, 1958.

[8] S. F. Ashby and M. H. Gutknecht. A matrix analysis of conjugate gradient algorithms. In M. Natori and T. Nodera, editors, *Advances in Numerical Methods for Large Sparse Sets of Linear Systems, Parallel Processing for Scientific Computing*, volume 9, pages 32–47, Yokohama, Japan, 1993. Keio University.

[9] S. F. Ashby, T. A. Manteuffel, and P. E. Saylor. A taxonomy for conjugate gradient methods. *SIAM J. Numer. Anal.*, 27(6):1542–1568, 1990.

[10] J. Atanga and D. Silvester. Iterative methods for stabilized mixed velocity-pressure finite elements. *Internat. J. Numer. Methods Fluids*, 14:71–81, 1992.

[11] O. Axelsson and P. S. Vassilevski. A black box generalized conjugate gradient solver with inner iterations and variable-step preconditioning. *SIAM J. Matrix Anal. Appl.*, 12(4):625–644, 1991.

[12] C. Bacuta. A unified approach for Uzawa algorithms. *SIAM J. Numer. Anal.*, 44(6):2633–2649, 2006.

[13] M. Benzi and G. H. Golub. A preconditioner for generalized saddle point problems. *SIAM J. Matrix Anal. Appl.*, 26:20–41, 2004.

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

[15] A. Björck. Solving linear least squares problems by Gram–Schmidt orthogonalization. *BIT*, 7:1–21, 1967.

[16] A. Björck. *Numerical Methods for Least Squares Problems*. SIAM, Philadelphia, 1996.

[17] A. M. Bollen. Numerical stability of descent methods for solving linear equations. *Numer. Math.*, 43:361–377, 1984.

[18] A. Bouras and V. Frayssé. Inexact matrix-vector products in Krylov methods for solving linear systems: a relaxation strategy. *SIAM J. Matrix Anal. Appl.*, 26(3):660–678, 2005.

[19] A. Bouras, V. Frayssé, and L. Giraud. A relaxation strategy for inner-outer linear solvers in domain decomposition methods. Technical Report TR/PA/00/17, CERFACS, France, 2000.

[20] D. Braess, P. Deuflhard, and K. Lipnikov. A subspace cascadic multigrid method for mortar elements. *Computing*, 69(3):205–225, 2002.

[21] D. Braess and R. Sarazin. An efficient smoother for the Stokes problem. *Appl. Numer. Math.*, 23(1):3–19, 1997.

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

[23] J. H. Bramble, J. E. Pasciak, and A. T. Vassilev. Inexact Uzawa algorithms for nonsymmetric saddle point problems. *Math. Comp.*, 69:667–689, 2000.

[24] F. Brezzi and M. Fortin. *Mixed and Hybrid Finite Element Methods*, volume 15 of *Springer Series in Computational Mathematics*. Springer-Verlag, New York, 1991.

[25] E. J. Craig. The N-step iteration procedures. *J. Math. Physics*, 34:64–73, 1955.

[26] J. W. Demmel, N. J. Higham, and R. S. Schreiber. Stability of the block LU factorization. *Numer. Linear Algebra Appl.*, 2(2):173–190, 1995.

[27] J. Drkošová, A. Greenbaum, M. Rozložník, and Z. Strakoš. Numerical stability of GMRES. *BIT*, 35(3):309–330, 1995.

[28] M. Eiermann and O. Ernst. Geometric aspects of the theory of Krylov subspace methods. *Acta Numerica*, pages 251–312, 2001.

[29] S. C. Eisenstat, H. C. Elman, and M. H. Schultz. Variational iterative methods for nonsymmetric systems of linear equations. *SIAM J. Numer. Anal.*, 20(2):345–357, 1983.

[30] H. C. Elman. *Iterative methods for large sparse nonsymmetric systems of linear equations*. PhD thesis, New Haven, 1982.

[31] H. C. Elman and G. H. Golub. Inexact and preconditioned Uzawa algorithms for saddle point problems. *SIAM J. Numer. Anal.*, 31(6):1645–1661, 1994.

[32] H. C. Elman, D. J. Silvester, and A. J. Wathen. *Finite Elements and Fast Iterative Solvers: with Applications in Incompressible Fluid Dynamics*. Oxford University Press, New York, 2005.

[33] D. K. Faddeev and V. N. Faddeeva. *Computational Methods of Linear Algebra*. Fizmatgiz, Moskow, 1960. in russian.

[34] B. Fischer, A. Ramage, D. J. Silvester, and A. J. Wathen. Minimum residual methods for augmented systems. *BIT*, 38:527–543, 1998.

[35] R. Fletcher. Conjugate gradient methods for indefinite systems. In G. A. Watson, editor, *Proceedings of the Dundee Biennial Conference on Numerical Analysis*, pages 73–89, New York, 1975. Springer-Verlag.

[36] A. Frommer and D. B. Szyld. H-Splittings and two-stage iterative methods. *Numer. Math.*, 63:345–356, 1992.

[37] E. Giladi, G. H. Golub, and J. B. Keller. Inner and outer iterations for the Chebyshev algorithm. *SIAM J. Numer. Anal.*, 35:300–319, 1998.

[38] P. E. Gill, W. Murray, and M. H. Wright. *Practical Optimization*. Academic Press Inc., London, 1981.

[39] L. Giraud, S. Gratton, and J. Langou. Convergence in backward error of relaxed GMRES. *SIAM J. Sci. Comput.*, 29(2):710–728, 2007.

[40] G. H. Golub. Bounds for the round-off errors in the Richardson second order method. *BIT*, 2:212–223, 1962.

[41] G. H. Golub and M. L. Overton. The convergence of inexact Chebyshev and Richardson iterative methods for solving linear systems. *Numer. Math.*, 53(5):571–593, 1988.

[42] G. H. Golub and C. F. Van Loan. *Matrix Computations*. The John Hopkins University Press, 3rd edition, 1996.

[43] G. H. Golub and Q. Ye. Inexact preconditioned conjugate gradient method with inner-outer iteration. *SIAM J. Sci. Comput.*, 21(4):1305–1320, 1999.

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

[45] A. Greenbaum. Behavior of slightly perturbed Lanczos and conjugate-gradient recurrences. *Linear Algebra Appl.*, 113:7–63, 1989.

[46] A. Greenbaum. Accuracy of computed solutions from conjugate-gradient-like methods. In M. Natori and T. Nodera, editors, *Advances in Numerical Methods for Large Sparse Sets of Linear Systems*, volume 10, pages 126–138, Keio University, Yokohama, Japan, 1994.

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

[48] A. Greenbaum, M. Rozložník, and Z. Strakoš. Numerical behaviour of the modified Gram-Schmidt GMRES implementation. *BIT*, 37(3):706–719, 1997.

[49] A. Greenbaum and Z. Strakoš. Predicting the behaviour of finite precision Lanczos and conjugate gradient computations. *SIAM J. Matrix Anal. Appl.*, 13:121–137, 1992.

[50] A. Greenbaum and Z. Strakoš. Matrices that generate the same Krylov residual spaces. In G. H. Golub, A. Greenbaum, and M. Luskin, editors, *Recent Advances in Iterative Methods*, pages 95–119, New York, 1994. Springer-Verlag.

[51] M. H. Gutknecht and M. Rozložník. Residual smoothing techniques: do they improve the limiting accuracy of iterative solvers? *BIT*, 41(1):86–114, 2001.

[52] M. H. Gutknecht and Z. Strakoš. Accuracy of two three-term and three two-term recurrences for Krylov space solvers. *SIAM J. Matrix Anal. Appl.*, 22(1):213–229, 2000.

[53] S. J. Hammarling and J. H. Wilkinson. The practical behaviour of linear iterative methods with particular reference to S.O.R. Technical Report NAC 69, National Physical Laboratory, England, Sept. 1976.

[54] M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *J Res. Natl. Bur. Stand.*, 49:409–436, 1952.

[55] N. J. Higham. *Accuracy and Stability of Numerical Algorithms*. SIAM, Philadelphia, 1996.

[56] N. J. Higham and P. A. Knight. Componentwise error analysis for stationary iterative methods. In C. D. Meyer and R. J. Plemmons, editors, *Linear Algebra, Markov Chains, and Queueing Models*, volume 48 of *IMA Volumes in Mathematics and Its Applications*, pages 29–46, 1993.

[57] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, New York, 1985.

[58] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, new edition, 1994.

[59] K. C. Jea and D. M. Young. On the simplification of generalized conjugate-gradient methods for nonsymmetrizable linear systems. *Linear Algebra Appl.*, 52:399–417, 1983.

[60] P. Jiránek and M. Rozložník. Limiting accuracy of segregated solution methods for nonsymmetric saddle point problems. *J. Comput. Appl. Math.*, 2007. to appear.

[61] P. Jiránek and M. Rozložník. Maximum attainable accuracy of inexact saddle point solvers. *SIAM J. Matrix Anal. Appl.*, 2007. to appear.

[62] P. Jiránek, M. Rozložník, and M. H. Gutknecht. How to make Simpler GMRES and GCR more stable. 2007. in preparation.

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

[64] P. J. Lanczkron, D. J. Rose, and D. B. Szyld. Convergence of nested classical iterative methods for linear systems. *Numer. Math.*, 58:685–702, 1991.

[65] C. Lanczos. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *J. Res. Natl. Bur. Stand.*, 45:255–281, 1950.

[66] J. Liesen, M. Rozložník, and Z. Strakoš. Least squares residuals and minimal residual methods. *SIAM J. Sci. Comp.*, 23(5):1503–1525, 2002.

[67] J. Liesen and Z. Strakoš. On numerical stability in large scale linear algebraic computations. *Z. Angew. Math. Mech.*, 85:307–325, 2005.

[68] J. Liesen and P. Tichý. Convergence analysis of Krylov subspace methods. *GAMM Mitt. Ges. Angew. Math. Mech.*, 27(2):153–173 (2005), 2004.

[69] M. S. Lynn. On the round-off error in the method of successive overrelaxation. *Math. Comp.*, 18(85):36–49, 1964.

[70] J. Maryška, M. Rozložník, and M. Tůma. Schur complement reduction in the mixed-hybrid approximation of Darcy's law: rounding error analysis. *J. Comput. Appl. Math.*, 117:159–173, 2000.

[71] J. Maryška, M. Rozložník, and M. Tůma. Schur complement systems in the mixed-hybrid finite element approximation of the potential fluid flow problem. *SIAM J. Sci. Comput.*, 22:704–723, 2000.

[72] G. Meurant. *Computer Solution of Large Linear Systems*. North Holland, 1999.

[73] N. K. Nichols. On the convergence of two-stage iterative processes for solving linear equations. *SIAM J. Numer. Anal.*, 10(3):460–469, 1973.

[74] J. Nocedal and S. Wright. *Numerical Optimization*. Springer, 1999.

[75] Y. Notay. On the convergence rate of the conjugate gradients in presence of rounding errors. *Numer. Math.*, 65:301–317, 1993.

[76] Y. Notay. Flexible conjugate gradients. *SIAM J. Sci. Comp.*, 22(4):1444–1460, 2000.

[77] C. C. Paige. Error analysis of the Lanczos algorithm for tridiagonalizing a symmetric matrix. *J. Inst. Maths. Applics*, 18:341–349, 1976.

[78] C. C. Paige, M. Rozložník, and Z. Strakoš. Modified Gram-Schmidt (MGS), least squares, and backward stability of MGS-GMRES. *SIAM J. Matrix Anal. Appl.*, 28(1):264–284, 2006.

[79] C. C. Paige and M. A. Saunders. Solution of sparse indefinite systems of linear equations. *SIAM J. Numer. Anal.*, 12:617–629, 1975.

[80] C. C. Paige and Z. Strakoš. Residual and backward error bounds in minimum residual Krylov subspace methods. *SIAM J. Sci. Comput.*, 23(6):1899–1924, 2002.

[81] S. V. Parankar. *Numerical Heat Transfer and Fluid Flow*. McGraw-Hill, 1980.

[82] A. Ramage and A. J. Wathen. Iterative solution techniques for the Stokes and Navier-Stokes equations. *Internat. J. Numer. Methods Fluids*, 19(1):67–83, 1994.

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

[84] M. Rozložník and Z. Strakoš. Variants of residual minimizing Krylov subspace methods. In I. Marek, editor, *Proceedings of the 6th Summer School Software and Algorithms of Numerical Mathematics*, pages 208–225, 1995.

[85] T. Rusten and R. Winther. A preconditioned iterative method for saddle-point problems. *SIAM J. Matrix Anal. Appl.*, 13:887–904, 1992.

[86] Y. Saad. Flexible inner-outer preconditioned GMRES algorithm. *SIAM J. Sci. Comput.*, 14(2):461–469, 1993.

[87] Y. Saad. *Iterative Methods for Sparse Linear Systems*. SIAM, 2nd edition, 2003.

[88] Y. Saad and M. H. Schultz. GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Stat. Comp.*, 7(3):856–869, 1986.

[89] V. Simoncini and I. Perugia. Block-diagonal and indefinite symmetric preconditioners for mixed finite element formulations. *Numer. Linear Algebra Appl.*, 7(8):585–616, 2000.

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

[91] V. Simoncini and D. B. Szyld. New conditions for non-stagnation of minimal residual methods. Technical Report 07-4-17, Apr. 2007.

[92] G. L. G. Sleijpen, H. A. van der Vorst, and J. Modersitzki. Differences in the effects of rounding errors in Krylov solvers for symmetric indefinite linear systems. *SIAM J. Matrix Anal. Appl.*, 22(3):726–751, 2000.

[93] P. Sonneveld. CGS, A fast Lanczos-type solver for nonsymmetric linear systems. *SIAM J. Sci. Stat. Comp.*, 10:36–52, 1989.

[94] E. Stiefel. Relaxationsmethoden bester Strategie zur Lösung linearer Gleichungssysteme. *Comm. Math. Helv.*, 29:157–179, 1955.

[95] Z. Strakoš. On the real convergence rate of the conjugate gradient method. *Linear Algebra Appl.*, 154–156:535–549, 1991.

[96] Z. Strakoš and P. Tichý. On error estimation in the conjugate gradient method and why it works in finite precision computations. *Electron. Trans. Numer. Anal.*, 13:56–80, 2002.

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

[98] J. van den Eshof, G. L. G. Sleijpen, and M. B. van Gijzen. Relaxation strategies for nested Krylov methods. *J. Comp. Appl. Math.*, 177(2):125–153, 2005.

[99] A. van der Sluis and H. A. van der Vorst. The rate of convergence of conjugate gradients. *Numer. Math.*, 48:543–560, 1986.

[100] H. A. van der Vorst. Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of non-symmetric linear systems. *SIAM J. Sci. Stat. Comput.*, 13:631–644, 1992.

[101] H. A. van der Vorst and C. Vuik. GMRESR: a family of nested GMRES methods. *Numer. Linear Algebra Appl.*, 1(4):369–386, 1994.

[102] P. K. W. Vinsome. Orthomin, an iterative method for solving sparse sets of simultaneous linear equations. In *Proceedings Fourth Symposium on Reservoir Simulation*, SPE of AIME, Los Angeles, Feb. 1976.

[103] H. F. Walker and L. Zhou. A simpler GMRES. *Numer. Linear Algebra Appl.*, 1(6):571–581, 1994.

[104] P. A. Wedin. Perturbation theory for pseudo-inverses. *BIT*, 13(2):217–232, 1973.

[105] C. Wieners and B. I. Wohlmuth. Duality estimates and multigrid analysis for saddle point problems arising from mortar discretizations. *SIAM J. Sci. Comp.*, 24(6):2163–2184, 2003.

[106] J. H. Wilkinson. *Rounding Errors in Algebraic Processes*. Prentice Hall, Inc., New Jersey, 1963.

[107] H. Woźniakowski. Numerical stability of the Chebyshev method for the solution of large linear systems. *Numer. Math.*, 28:191–209, 1977.

[108] H. Woźniakowski. Round-off error analysis of iterations for large linear systems. *Numer. Math.*, 30:301–314, 1978.

[109] H. Woźniakowski. Roundoff-error analysis of a new class of conjugate-gradient algorithms. *Linear Algebra Appl.*, 29:507–529, 1980.

[110] D. M. Young and K. C. Jea. Generalized conjugate gradient acceleration of nonsymmetrizable iterative methods. *Linear Algebra Appl.*, 34:159–194, 1980.

[111] W. Zulehner. A class of smoothers for saddle point problems. *Computing*, 65:227–246, 2000.

[112] W. Zulehner. Analysis of iterative methods for saddle point problems: a unified approach. *Math. Comp.*, 71(238):479–505, 2002.

Pavel Jiránek

Limiting Accuracy of Iterative Methods

PhD Thesis

Faculty of Mechatronics and Interdisciplinary Engineering Studies
Technical University of Liberec
Czech Republic

Institute of Computer Science
Academy of Sciences of the Czech Republic
Czech Republic