Technical University of Liberec

Faculty of Mechatronics and Interdisciplinary Engineering Studies

Academy of Sciences of the Czech Republic, Prague

Institute of Computer Science

Limiting Accuracy of Iterative Methods

Pavel Jiránek

Report of the PhD Thesis December, 2007

Technical University of Liberec

Faculty of Mechatronics and Interdisciplinary Engineering Studies

Academy of Sciences of the Czech Republic, Prague

Institute of Computer Science

Limiting Accuracy of Iterative Methods

Pavel Jiránek

Report of the PhD Thesis December, 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
Referees:	Mario Arioli (Rutherford Appleton Laboratory, Oxfordshire, UK) Petr Tichý (Institute of Computer Science, CAS, Prague, CZ)

Typeset by LaTeX.

ISBN: 978-80-7372-285-2

Abstrakt

Jak je známo, zaokrouhlovací chyby a nepřesné řešení vnitřních úloh mají vliv na numerické chování iteračních metod; 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 obsahuje analýzu 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 solutions of inner systems and rounding errors affect the numerical behavior of iterative methods. 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 мимодиагонального 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 пространств Крылова, метод преобразования на дополнение Шура, метод проекции на ядро мимодиагонального блока, методы минимальных невязок, вычислительная стабильность, анализ ошибок округления.

Contents

CHAPTER 1

Introduction

Consider a system of linear algebraic equations in the form

$$Ax = b, \tag{1}$$

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) or for its sufficiently accurate approximation.

The methods for solving (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) 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) 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 [**59**], and by Simoncini and Szyld [**54**]. Their analysis explains the experimental results of Bourass and Frayssé [**7**] (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 [**59, 54**] provide the theoretical support for this strategy further developed in [**60**]. This topic is closely related to the flexible preconditioning, see, e.g., [**4, 21, 46, 54, 18**]. 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 [**19**] and Lynn [**42**] 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 [**2**] clarifying the relation between the conditioning of the preconditioned system matrix and the convergence rate of the iterative method. In [**33**] 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 [**30**]. The stability of classical iterative methods was also analyzed by Woźniakovski in [**67, 68**]. 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 [**20**] 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 [**17**] who show the optimality of the uniform tolerance used in [**20**]. 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 [**44, 37, 16**].

One of the most important result in the study of Krylov subspace methods is due to Paige [**47**]. He provides the analysis of the behavior of the symmetric Lanczos algorithm [**38**] in the presence of rounding errors. This algorithm is closely related to the conjugate gradient method by Hestenes and Stiefel [**31**]. It was first studied by Woźniakowski [**69**] and Bollen [**6**]. 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., [**61, 41**] and the references therein. The new insight into this problem was brought by Greenbaum [23] and further developed together with Strakoš [58, 27]. 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 [45].

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 [24, 25]. 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 [15] or CGS [56]) can have an unfavorable effect on the limiting accuracy of the computed solution. A similar phenomenon is mentioned also by van der Vorst in [62], 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 [57] applied to symmetric positive definite problems, or in the case of residual minimizing methods (Orthodir [70], Orthomin [64], GCR [12]) for general nonsymmetric systems. The numerical stability of various (equivalent) methods using short recurrences was further studied by Gutknecht and Strakoš in [29] and by Sleijpen, van der Vorst and Modersitzki in [55]. In [28] 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 [53]. The Householder implementation of the underlying Arnoldi process [3] is quite straightforward to analyze, see the paper by Drkošová, Greenbaum, Rozložník and Strakoš [11], and by Arioli and Fassino [1]. 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 [26] and further analyzed by Paige, Rozložník and Strakoš in [49, 48]. For more details see [40] 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 [35, 34], 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 [**36**], is devoted to the analysis of several residual minimizing Krylov subspace methods, which are mathematically equivalent to the GMRES method [**53**]. 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

Main results of the thesis

1. Limiting accuracy of segregated saddle point solvers

In this section we summarize the results of the first part of the thesis. Consider the solution of a saddle point system in the block form

$$\begin{pmatrix} A & B \\ B^T & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} f \\ 0 \end{pmatrix}, \tag{2}$$

where the diagonal block $A \in \mathbb{R}^{n\times n}$ is symmetric positive definite, and the off-diagonal block $B \in \mathbb{R}^{n\times m}$ has the full column rank. The solution vector and the right-hand side vector are partitioned consistently with respect to the partitioning of the system matrix. 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 the excellent survey on applications, methods and results on numerical solution of saddle point problems we refer to [**5**] and numerous references therein. Significantly less attention however has been paid so far to the numerical stability aspects. 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 the 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. Here 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.

We concentrate on the question what is the best accuracy we can get from the Schur complement reduction method and the null-space projection method when inner systems are solved with a prescribed accuracy 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 [**54**] and [**59**]. Assuming exact arithmetic, the authors of [**54**] and [**59**] 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 [**7, 8, 18**]. 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.

The theory developed here for the outer iteration process is similar to the analysis of Greenbaum in [**25, 24**] 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 [**25**] considered only the rounding errors associated with a fixed matrix-vector multiplication. In contrast to the theory of inexact Krylov methods [**54, 59**] 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 are the two block equations of (2) 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 independently of the fact that the inner systems are solved inexactly some back-substitution schemes lead ultimately to residuals on the roundoff unit level.

1.1. Schur complement reduction method. The Schur complement reduction method uses the equivalent formulation of (2) in the form

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

which is nothing but a block Gaussian elimination applied to (2). This block triangular system is solved by computing the unknown y from the symmetric positive definite Schur complement system

$$B^TA^{-1}By = B^TA^{-1}f \tag{3}$$

and then by computing the unknown x from a system

$$Ax = f - By. \tag{4}$$

Here we discuss algorithms which compute simultaneously approximations y_k and x_k solving iteratively the Schur complement system (3) and ideally fulfill the first block equation of (2), i.e., they satisfy

$$Ax_k + By_k = f.$$

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

$$y_{k+1} = y_k + \alpha_k p_k^{(y)}, \tag{5}$$

$$r_{k+1}^{(y)} = r_k^{(y)} + \alpha_k B^T A^{-1} B p_k^{(y)} \tag{6}$$

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

$$x_{k+1} = x_k + \alpha_k(-A^{-1}Bp_k^{(y)}), \tag{7}$$

$$x_{k+1} = A^{-1}(f - By_{k+1}), \tag{8}$$

$$x_{k+1} = x_k + A^{-1}(f - Ax_k - By_{k+1}). \tag{9}$$

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 (2). Because the solves with matrix A in (7)-(9) 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 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, and we 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 $\tau\kappa(A) \ll 1$.

Using (5) and (6), we can estimate the gap between the true residual in the outer iteration, i.e., the residual in the Schur complement system (3), and the updated residual $r_k^{(y)}$ as

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

where $\bar{Y}_k$ is defined as a maximum norm over all computed approximate solutions $\bar{Y}_k \equiv \max_{i=0,\ldots,k} \|\bar{y}_i\|$. While the updated residual $\bar{r}_k^{(y)}$ converges to zero, the true residual stagnates at the level proportional to τ. On the other hand, the accuracy measured by the residuals $f - A\bar{x}_k - B\bar{y}_k$ and $-B^T\bar{x}_k$ in (2) depends on the particular choice of the back-substitution formula. No matter how we compute the approximations $\bar{x}_k$ and $\bar{y}_k$, we have

$$-B^T A^{-1} f + B^T A^{-1} B\bar{y}_k = -B^T \bar{x}_k - B^T A^{-1}(f - A\bar{x}_k - B\bar{y}_k) \tag{10}$$

which gives the mutual relation between the residual $-B^T A^{-1} f + B^T A^{-1} B\bar{y}_k$ in the Schur complement system (3) and the residuals $f - A\bar{x}_k - B\bar{y}_k$ and $-B^T\bar{x}_k$ associated with the saddle point system (2). Since $\| - B^T A^{-1} f + B^T A^{-1} B\bar{y}_k\|$ is ultimately $O(\tau)$, it is clear from (10) 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.

In the update scheme (7), the true residual $f - A\bar{x}_k - B\bar{y}_k$ satisfies the bound

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

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

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

Hence this scheme guarantees that the residual $-B^T\bar{x}_k$ will ultimately reach the level of $O(u)$ independently on the fact that the inner systems are solved with the relaxed accuracy given by the parameter τ.

In the direct substitution scheme (8), the true residual $f - A\bar{x}_k - B\bar{y}_k$ satisfies the bound

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

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

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

In this most straightforward scheme, both residuals thus stagnate ultimately on the level of $O(\tau)$.

In the corrected direct substitution scheme (9), the true residual $f - A\bar{x}_k - B\bar{y}_k$ satisfies

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

for all steps k starting from some k_0, where $\bar{Y}_k^{k_0} \equiv \max_{i=k_0,\dots,k}\|\bar{y}_i\|$. The gap between $-B^T\bar{x}_k$ and $\bar{r}_k^{(y)}$ can be estimated as follows

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

The convergence of $\|f - A\bar{x}_k - B\bar{y}_k\|$ is driven by the stationary iteration with the norm of the iteration matrix bounded by $O(\tau)\kappa(A)/(1 - \tau\kappa(A))$ and after some initial stage the residual converges ultimately to the level of $O(u)$. However, the second block equation of (2) is satisfied to the accuracy given by τ.

Independently of the chosen back-substitution formula, the ultimate levels of error norms $\|x - \bar{x}_k\|$ and $\|y - \bar{y}_k\|$ are $O(\tau)$ as indicated by the estimates

$$\|x - \bar{x}_k\| \leq \gamma_1\|f - A\bar{x}_k - B\bar{y}_k\| + \gamma_2\| - B^T\bar{x}_k\|,$$
$$\|y - \bar{y}_k\| \leq \gamma_2\|f - A\bar{x}_k - B\bar{y}_k\| + \gamma_3\| - B^T\bar{x}_k\|,$$

where $\gamma_1 \equiv \sigma_{min}^{-1}(A)$, $\gamma_2 \equiv \sigma_{min}^{-1}(B)$ and $\gamma_3 \equiv \sigma_{min}^{-1}(B^TA^{-1}B)$ are constants independent of the iteration step k, and depend on the conditioning of the blocks A and B. In practice, these blocks can be ill-conditioned and in such cases the constants γ_1, γ_2 and γ_3 may play an important role.

1.2. Null-space projection method. The null-space projection method is based on the projection of the first block equation in (2) onto the null-space $N(B^T)$ and onto its orthogonal complement, the range $R(B)$. Denoting by Π the orthogonal projector onto $R(B)$, we first compute the unknown $x \in N(B^T)$ from the projected system

$$(I - \Pi)A(I - \Pi)x = (I - \Pi)f \tag{11}$$

with the symmetric positive semi-definite matrix $(I-\Pi)A(I-\Pi)$, and then the unknown y is obtained as $y = B^{\dagger}(f - Ax)$ by solving the least squares problem

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

We discuss algorithms which compute simultaneously approximations x_k and y_k by solving iteratively the projected system (11) and minimize the residual norm $f - Ax_k - By_k$, i.e., y_{k+1} is given by $y_{k+1} = B^{\dagger}(f - Ax_{k+1})$. We assume that the approximate solution x_{k+1} and the residual vector $r_{k+1}^{(x)}$ are computed using

$$x_{k+1} = x_k + \alpha_k p_k^{(x)}, \tag{12}$$

$$r_{k+1}^{(x)} = r_k^{(x)} - \alpha_k A p_k^{(x)} - B p_k^{(y)}, \tag{13}$$

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 [**22**] (see also [**10, 9**]) and it is used to reduce the roundoff errors in the projection onto $N(B^T)$. Again we distinguish between three back-substitution formulas

$$y_{k+1} = y_k + p_k^{(y)},\ p_k^{(y)} = B^{\dagger}(r_k^{(x)} - \alpha_k A p_k^{(x)}), \tag{14}$$

$$y_{k+1} = B^{\dagger}(f - Ax_{k+1}), \tag{15}$$

$$y_{k+1} = y_k + B^{\dagger}(f - Ax_{k+1} - By_k). \tag{16}$$

The pseudoinverse $B^{\dagger}$ in (14)-(16) is applied by solving the least squares with the matrix B. These problems are solved inexactly. In our considerations we assume that the computed solution $\bar{v}$ of a 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 it 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. We assume $\tau\kappa(B) \ll 1$ which guarantees $B + \Delta B$ to have a full column rank.

Using (12) and (13), we can estimate the gap between the true residual in the outer iteration, i.e., in the projected system (11), and the updated residual $\bar{r}_k^{(x)}$ as

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

where $\bar{X}_k \equiv \max_{i=0,\dots,k}\|\bar{x}_i\|$. While the updated residual $\bar{r}_k^{(x)}$ converges to zero, the true residual stagnates at the level proportional to τ independently on the back-substitution formula. Moreover, 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 (12) and (13) and we do not give it here. In accordance with [**25**] it seems reasonable that the bound for $-B^T\bar{x}_k$ is proportional to the factor $\bar{X}_k$, i.e.,

$$\|-B^T\bar{x}_k\| \leq \frac{O(\tau)\|B\|}{1-\tau\kappa(B)}\bar{X}_k.$$

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 $-B^T\bar{x}_k$

$$(I-\Pi)f-(I-\Pi)A(I-\Pi)\bar{x}_k=(I-\Pi)(f-A\bar{x}_k-B\bar{y}_k)+(I-\Pi)A\Pi\bar{x}_k. \tag{17}$$

Owing to our assumption, the norm of $-B^T\bar{x}_k$ is finally on the level of $O(\tau)$. We have that $\|(I-\Pi)f-(I-\Pi)A(I-\Pi)\bar{x}_k\|$ is ultimately $O(\tau)$ and, on the other hand, the norm of the projection of $f-A\bar{x}_k-B\bar{y}_k$ onto $N(B^T)$ reaches the level of $O(u)$. It is not clear from (17) 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.

In the updated scheme (14), the gap between the residuals $f-A\bar{x}_k-B\bar{y}_k$ and $\bar{r}_k^{(x)}$ can be bounded as

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

Thus using the simple update formula makes the first component of the residual in (2) stagnating ultimately on the level proportional to unit roundoff.

In the direct substitution scheme (15), the gap between the residuals $f-A\bar{x}_k-B\bar{y}_k$ and $\bar{r}_k^{(x)}$ can be bounded as

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

Comparing this scheme with the generic update formula, the first component of the residual in (2) ultimately stagnates on the level proportional to the parameter τ.

In the corrected direct substitution scheme (16), the gap between the residuals $f - A\bar{x}_k - B\bar{y}_k$ and $\bar{r}_k^{(x)}$ can be bounded as

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

for all k large enough. This scheme gives a similar accuracy as the generic update but it costs one additional solution of the least squares problem with B.

For the error norms, we have the same results as in the case of the Schur complement method – they do not depend on the back-substitution scheme and ultimately stagnate on the level proportional to τ.

2. Numerical stability of some residual minimizing iterative methods

In this section we summarize the results of the second part of the thesis. We consider certain Krylov subspace methods for solving a system of linear algebraic equations

$$Ax = b, \qquad A \in \mathbb{R}^{N \times N}, \qquad b \in \mathbb{R}^N, \tag{18}$$

where A is a large and sparse nonsingular matrix that is, in general, nonsymmetric. A Krylov subspace method builds 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 [**25**], Meurant [**43**], and Saad [**52**]. In particular, due to their smooth convergence behavior, minimum residual methods satisfying

$$\|r_n\| = \min_{\tilde{x} \in x_0 + \mathcal{K}_n(A, r_0)} \|b - A\tilde{x}\|, \qquad r_n \equiv b - Ax_n, \tag{19}$$

are widely used, e.g., the GMRES algorithm of Saad and Schultz [**53**]. In [**11, 26, 48**] 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 [**65**], who called it the Simpler GMRES method. The minimum residual property (19) 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 $\langle \cdot, \cdot \rangle$. We propose a generalization of the Simpler GMRES method that makes use of 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 $\mathcal{K}_n(A, r_0)$. We may assume that the columns z_k of Z_{n-1} have unit length and need not be mutually orthogonal. The orthonormal basis V_n of $A\mathcal{K}_n(A, r_0)$ is obtained from the QR factorization of the image of $[q_1, Z_{n-1}]$:

$$A[q_1, Z_{n-1}] = V_n U_n. \tag{20}$$

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 $r_n = (I - V_n V_n^T)r_0$. To compute it we apply the modified Gram-Schmidt method, which leads to the recursion

$$r_n = r_{n-1} - \alpha_n v_n, \qquad \alpha_n \equiv \langle r_{n-1}, v_n \rangle. \tag{21}$$

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

$$x_n = x_0 + [q_1, Z_{n-1}]t_n. \tag{22}$$

Due to the minimum residual property, we have $r_n \perp \mathcal{R}(V_n)$, and thus simply

$$U_n t_n = V_n^T r_0 = [\alpha_1, \ldots, \alpha_n]^T. \tag{23}$$

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".

Recursion (21) 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 (21) by A^{-1} yields

$$x_n = x_{n-1} + \alpha_n p_n, \tag{24}$$

Now, note that left-multiplying (20) by A^{-1} yields

$$[q_1, Z_{n-1}] = P_n U_n. \tag{25}$$

If U_n is known from (20), a recursion for p_n can be extracted from this formula. 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 it is a refined version of the residual norm minimizing ORTHODIR algorithm [**14**, **70**]. 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 [**64**, **70**] (or the GCR method of Elman [**13**, **12**]) and is identical to the GMRESR method without preconditioning.

2.1. The maximum attainable accuracy. We analyze the numerical stability of the simpler and update approaches, and assume that only the computations performed in (20), (23) and (25) are affected by rounding errors and that the computed Q-factor in the QR factorization (20) 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 (20) satisfy

$$A[q_1, Z_{n-1}] = V_n U_n + F_n, \qquad \|F_n\| \leq cu\|A\|\|[q_1, Z_{n-1}]\|, \tag{26}$$

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 do not distinguish between V_n and $\hat{V}_n$ and assume that V_n is exactly orthonormal. In the simpler approach, we have from [**66, 32**] for the computed solution $\hat{t}_n$ of (23) that

$$(U_n + \Delta U_n)\hat{t}_n = D_n e, \qquad |\Delta U_n| \leq cu|U_n|, \tag{27}$$

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

$$\hat{x}_n = x_0 + [q_1, Z_{n-1}]\hat{t}_n. \tag{28}$$

In accordance with (26) we assume in the update approach that in finite precision arithmetic the computed direction vectors satisfy

$$[q_1, Z_{n-1}] = P_n U_n + G_n, \qquad ||G_n|| \leq cu||P_n||||U_n||. \tag{29}$$

As in (24) we compute then the approximate solution $\hat{x}_n$ as

$$\hat{x}_n = \hat{x}_{n-1} + \alpha_n p_n. \tag{30}$$

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 (21) describing the projection of the previous residual; see [**25, 29**]. In fact, once the true residual becomes negligible compared to the true one, 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 $\hat{x}_n$. In the simpler approach, the gap between the true residual $b - A\hat{x}_n$ and the updated residual r_n satisfies

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

while in the update approach, we have

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

provided that $1 - cu\kappa(A)\kappa([q_1, Z_{n-1}]) > 0$. The bound on the ultimate backward error for the update approach is worse that the one for the simpler approach. 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. Such a difference is hard to be seen in practice, but a model example can be constructed, where this difference is clearly visible.

In contrast to the difference in the attainable accuracy measured by the backward errors, it appears that the update approach leads to an approximate solution on essentially the same accuracy level in the error as the simpler approach, as indicated by the estimate

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

which holds for both approaches. A similar phenomenon was also observed by Sleijpen, van der Vorst and Modersitzki [**55**] in the symmetric case for GMRES and MINRES.

2.2. Choice of the basis. First, we choose $Z_{n-1} = V_{n-1}$, which leads to the Simpler GMRES method of Walker and Zhou [**65**] and to the refined version of ORTHODIR by Young and Jea [**70**], respectively. Hence, we choose $\{q_1, v_1, \dots, v_{n-1}\}$ as a basis of $\mathcal{K}_n(A, r_0)$. If $r_0 \notin A\mathcal{K}_n(A, r_0)$, these vectors are linearly independent and hence form a basis. Note that if $r_0 \in A\mathcal{K}_n(A, r_0)$, then the condition (19) yields $x_n = A^{-1}b$, $r_n = 0$, and any implementation of the minimum residual method will terminate. As observed by Liesen, Rozložník and Strakoš [**39**], this choice of the basis is not very suitable from the numerical stability point of view. This shortcoming is reflected by the unbounded growth of the condition number of $[q_1, V_{n-1}]$ expressed by the two-sided inequalities

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

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. Second, we choose $Z_{n-1} = [\frac{r_1}{\|r_1\|}, \dots, \frac{r_{n-1}}{\|r_{n-1}\|}]$, which leads to SGMRES/RB (which we propose as a more stable counterpart of Simpler GMRES) and to the refined version of ORTHOMIN by Vinsome [**64**] known also under the name GCR [**13, 12**]. We have $[q_1, Z_{n-1}] = R_n B_n^{-1}$, where $B_n \equiv \mathrm{diag}(\|r_0\|, \dots, \|r_{n-1}\|)$, i.e., we choose scaled residuals $r_0, \dots, r_{n-1}$ as the basis of $\mathcal{K}_n(A, r_0)$. The linear independence of the residual is guaranteed by the strictly monotonous convergence of their 2-norms and the condition that the exact solution was not reached yet, i.e., $r_0 \notin A\mathcal{K}_n(A, r_0)$. Moreover, when the minimum residual method does not stagnate, the residuals form a well-conditioned basis, as indicated by the estimate

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

We define the quantity γ_n 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.

CHAPTER 3

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 the first part 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 the second part 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 of segregated saddle point solvers 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 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 [**33**] 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 $r_k = b - Ax_k$. The final level of the residual $f - A\bar{x}_k - B\bar{y}_k$ in the Schur complement reduction method with the corrected direct substitution 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., [**50**]).

Backward error analysis of segregated methods. At the end of the first part of the thesis, 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 the analysis of the minimal residual Krylov subspace methods, 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 [**63**]) is more preferable than, e.g., preconditioned GMRES (with a fixed preconditioner) or flexible GMRES [**51**], 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] M. Arioli and C. Fassino. Roundoff error analysis of algorithms based on Krylov subspace methods. *BIT*, 36(2):189–206, 1996.

[2] 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.

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

[4] 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.

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

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

[7] 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.

[8] 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.

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

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

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

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

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

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

[15] 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.

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

[17] 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.

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

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

[20] 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.

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

[22] 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.

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

[24] 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.

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

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

[27] 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.

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

[29] 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.

[30] 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.

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

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

[33] 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.

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

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

[36] P. Jiránek, M. Rozložník, and M. H. Gutknecht. How to make Simpler GMRES and GCR more stable. *SIAM J. Matrix Anal. Appl.*, 2007. submitted.

[37] 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.

[38] 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.

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

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

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

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

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

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

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

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

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

[48] 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.

[49] 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.

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

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

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

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

[54] 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.

[55] 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.

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

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

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

[59] 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.

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

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

[62] 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. Statist. Comput.*, 13:631–644, 1992.

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

[64] 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.

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

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

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

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

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

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

Název:	Limiting Accuracy of Iterative Methods
Autor:	Pavel Jiránek
Vydavatel:	Technická univerzita v Liberci
Schváleno:	Rektorátem TU v Liberci dne 5. 12. 2007, čj. RE 163/07
Vyšlo:	v prosinci 2007
Počet stran:	36
Vydání:	první
Tiskárna:	Vysokoškolský podnik Liberec, s.r.o., Hálkova 6, Liberec
Číslo publikace:	55-119-07

Tato publikace neprošla redakční ani jazykovou úpravou.

ISBN: 978-80-7372-285-2