Preconditioned Newton methods for ill-posed problems

Dissertation

zur Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultäten

der Georg-August-Universität zu Göttingen

vorgelegt von

Stefan Langer

aus Kassel

Göttingen 2007

D 7
Referent: Prof. Dr. T. Hohage
Korreferent: Prof. Dr. R. Kreß
Tag der mündlichen Prüfung: 21. Juni 2007

Abstract

We consider preconditioned regularized Newton methods tailored to the efficient solution of nonlinear large-scale exponentially ill-posed problems.
In the first part of this thesis we investigate convergence and convergence rates of the iteratively regularized Gauss-Newton method under general source conditions both for an a-priori stopping rule and the discrepancy principle. The source condition determines the smoothness of the true solution of the given problem in an abstract setting. Dealing with large-scale ill-posed problems it is in general not realistic to assume that the regularized Newton equations can be solved exactly in each Newton step. Therefore, our convergence analysis includes the practically relevant case that the regularized Newton equations are solved only approximately and the Newton updates are computed by using these approximations.
In a second part of this thesis we analyze the complexity of the iteratively regularized Gauss-Newton method assuming that the regularized Newton equations are solved by the conjugate gradient method. This analysis includes both mildly and severely ill-posed problems. As a measure of the complexity we count the number of operator evaluations of the Fréchet derivative and its adjoint at some given vectors. Following a common practice for linear ill-posed problems, we express the total complexity of the iteratively regularized Gauss-Newton method in terms of the noise level of the given data.
To reduce the total complexity of these regularized Newton methods we consider spectral preconditioners to accelerate the convergence speed of the inner conjugate gradient iterations. We extend our complexity analysis to these preconditioned regularized Newton methods. This investigation gives us the possibility to compare the total complexity of non preconditioned regularized Newton methods and preconditioned ones. In particular we show the superiority of the latter ones in the case of exponentially ill-posed problems.
Finally, in a third part we discuss the implementation of a preconditioned iteratively regularized Gauss-Newton methods exploiting the close connection of the conjugate gradient method and Lanczos' method as well as the fast decay of the eigenvalues corresponding to the linearized operators in the regularized Newton equations. More precisely, we determine by Lanczos' method approximations to some of the extremal eigenvalues. These are used to construct spectral preconditioners for the following Newton steps. Developing updating techniques to keep the preconditioner efficient while performing Newton's method the total complexity can be significantly reduced compared to the non preconditioned iteratively regularized Gauss-Newton method. Finally, we illustrate in numerical examples from inverse scattering theory the efficiency of the preconditioned regularized Newton methods compared to other regularized Newton methods.

Acknowledgments

After all I wish to thank all those who helped me throughout my studies. First of all, my thank is dedicated to my advisor Prof. Dr. Thorsten Hohage for introducing me into the topic of my thesis. The discussions with him were always helpful, as well as his hints and suggestions when they were needed. Moreover, I gratefully want to thank him for letting me use a C++-class library designed for iterative regularization methods. This tool facilitated the work on the numerical examples a lot. Furthermore, my thank goes to Prof. Dr. Rainer Kreß for acting as the second advisor.
Sincere thanks are given to my office mates Harald Heese and Pedro Serranho for carefully proof-reading parts of this thesis and for the good memories for our joint time in our office. My thanks are extended to Annika Eickhoff-Schachtebeck who also read carefully over parts of this thesis and to my former English teacher Mr. Newton who accurately read over the introduction.
The financial support of the Deutsche Forschungsgemeinschaft Graduiertenkolleg 1023 "Identification in Mathematical Models: Synergy of Stochastic and Numerical Methods" is also gratefully acknowledged.
Finally, I would like to thank my fiancée Antje Packheiser for encouraging me over the last years, especially in the last months while writing this thesis.

Contents

Chapter 0

Introduction

Inverse problems occur in many branches of science and mathematics. Usually these problems involve the determination of some model parameters from observed data, as opposed to the problems arising from physical situations where the model parameters or material properties are known. The latter problems are in general *well-posed*. The mathematical term *well-posed problem* stems from a definition given by Hadamard [28]. He called a problem well-posed, if

a) a solution exists,

b) the solution is unique,

c) the solution depends continuously on the data, in some reasonable topology.

Problems that are not well-posed in the sense of Hadamard are termed *ill-posed*. Inverse problems are typically ill-posed. Of the three conditions for a well-posed problem the condition of stability is most often violated and has our primary interest. This is motivated by the fact that in all applications the data will be measured and therefore perturbed by noise. Typically, inverse problems are classified as linear or nonlinear. Classical examples of linear inverse problems are computerized tomography [67] and heat conduction [16, Chapter 1].
An inherently more difficult family are nonlinear inverse problems. Nonlinear inverse problems appear in a variety of fields such as scattering theory [11] and impedance tomography. During the last decade a variety of problem specific mathematical methods has been developed for solving a given individual ill-posed problem. For example, for the solution of time harmonic *acoustic* inverse scattering problems quite a number of methods have been developed such as the *Kirsch-Kress method* [48, 49, 50], the *Factorization method* [46, 47, 27] and the *Point-source method* [72]. Naturally, the development of such problem specific solution approaches often requires a lot of time and a deep understanding of the mathematical and physical aspects.

Unfortunately, a portability of problem specific solution methods to other problems is often either impossible or a difficult task. For example, although the methods mentioned above exist already for about ten years or even longer, to our knowledge a satisfactory realization of these methods for time harmonic *electromagnetic* inverse scattering problems is still open. Moreover, besides the classical and well known inverse problems due to evolving innovative processes in engineering and business more and more new nonlinear problems arise. Hence, although problem specific methods for nonlinear inverse problems have their advantages, efficient algorithms for solving inverse problems in their general formulation as nonlinear operator equations have proven to become necessary.

It is the topic of this work to develop and analyze a regularized Newton method designed for efficiently solving large scale nonlinear ill-posed problems, in particular nonlinear exponentially ill-posed problems.

Newton's method is one of the most powerful techniques for solving nonlinear equations. Its widespread applications in all areas of mathematics make it one of the most important and best known procedures in this science. Usually it is the first choice to try for solving some given nonlinear equation. Many other good methods designed to solve nonlinear equations often turn out to be variants of Newton's method attempting to preserve its convergence properties without its disadvantages. A motivation of Newton's method is given by the following elementary construction:

We consider the nonlinear equation $f(x) = 0$, where $f : \mathbb{R} \to \mathbb{R}$ is a continuously differentiable function. Let x_n be an approximation to some root x^* of f and $y(x) := f(x_n)+f'(x_n)(x-x_n)$ the tangent on f through $(x_n, f(x_n))$. If $f'(x_n) \neq 0$, then y has exactly one point of intersection with the x-axis, which we examine as new approximation to x^*. Proceeding in this way, which is illustrated in Figure 1, we obtain the algorithm

$$x_{n+1} := x_n - [f'(x_n)]^{-1} f(x_n), \qquad n = 0, 1, 2, \ldots .$$

This idea can be generalized to operator equations

$$F(x) = y, \tag{1}$$

where $F : D(F) \to \mathcal{Y}$ is a nonlinear injective Fréchet differentiable mapping between its domain $D(F) \subset \mathcal{X}$ into $\mathcal{Y}$. Throughout this work $\mathcal{X}$ and $\mathcal{Y}$ are real Hilbert spaces. Substituting F by its linear approximation in each Newton step the least squares problem

$$\|F'[x_n]h + F(x_n) - y\|_{\mathcal{Y}}^2 = \min_{h \in \mathcal{X}} ! \tag{2}$$

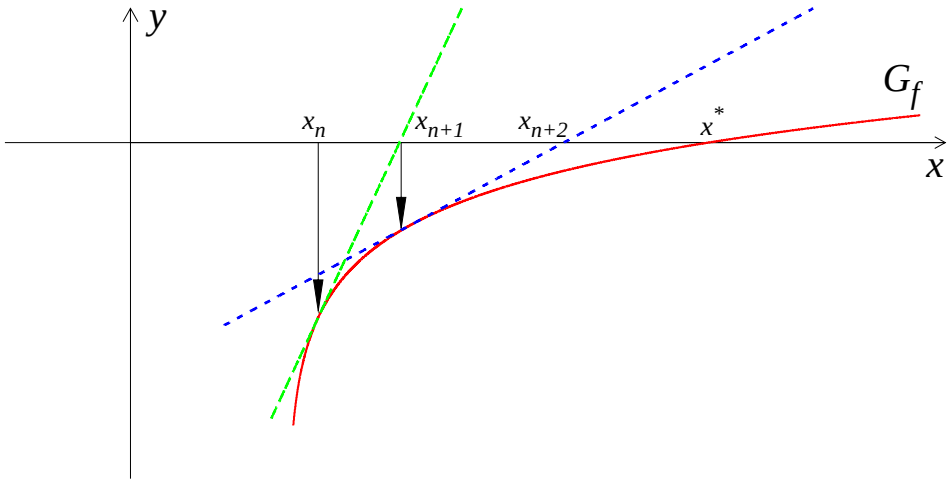

Figure 1: An illustration of Newton's method

needs to be solved. $F'[x_n]$ denotes the Fréchet derivative of F at x_n and the Newton update is given by $h = x_{n+1} - x_n$. This generalized approach is well-known as Gauss-Newton method. If the operator equation (1) is well posed many different local convergence proofs of the Gauss-Newton method have been established to show convergence of quadratic order under some natural conditions on the operator F.

In the case where (1) is ill-posed it is important to study the situation where the right hand side y in (1) is replaced by noisy data y^δ satisfying

$$\|y - y^\delta\|_{\mathcal{Y}} \leq \delta$$

for a known noise level $\delta > 0$. In this case a straightforward implementation of the Gauss-Newton method usually fails and does not lead to a good reconstruction of the solution after several Newton steps. One reason for the failure of the Gauss-Newton approach in this situation is the ill-posedness of the least-squares problem (2), which is inherited from the original operator equation (1). Thus, to perform one Newton step some sort of regularization has to be employed when solving (2). This additional regularization usually complicates the investigation of local convergence of Newton's method. Moreover, different kinds of regularization methods for the linearized equation generate different kinds of regularized Newton methods and each of these methods requires its own convergence analysis. During the last fifteen years many of these methods have been proposed, but often no completely satisfactory convergence proofs could be established so far since often assumptions are made, which could only be proven for a few examples. In Section 2.1 we will discuss some examples of regularized Newton methods.

In this work we consider a regularized Gauss-Newton method where instead of (2), the regularized least-squares problem

$$\|F'[x_n^\delta]h + F(x_n^\delta) - y^\delta\|_{\mathcal{Y}}^2 + \gamma_n\|h + x_n^\delta - x_0\|_{\mathcal{X}}^2 = \min_{h\in\mathcal{X}}! \tag{3}$$

is solved in each Newton step. Here $\gamma_n > 0$ denotes a regularization parameter. This iterative regularization method can be interpreted as a common Newton method, where in each Newton step Tikhonov regularization with initial guess $x_n^\delta - x_0$ is applied to the linearized equation. The problem (3) is well posed, in particular there exists a uniquely defined minimizer $h^\dagger \in \mathcal{X}$ of (3). Moreover, if γ_n is small we expect that the solution of (3) is a stable approximation to the solution of (2). Formulating additional assumptions on the sequence $(\gamma_n)_{n\in\mathbb{N}_0}$ this algorithm is called *iteratively regularized Gauss-Newton method* (IRGNM) and was originally suggested by Bakushinskii [5]. We are going to contribute to the convergence analysis of this method.

When speaking of convergence of iterative regularization methods for ill-posed problems we have to distinguish two different types of convergence. On the one hand, for known exact data y we must ensure that our iterates converge to the true solution of (1). On the other hand if the right hand side of (1) is given by noisy measurements y^δ we have to combine the iterative regularization method with some data-dependent stopping criterion. The most well-known is *Morozov's discrepancy principle* [65]. It states that one should not try to solve the operator equation more accurately than the data noise error. This ensures a stopping of the algorithm before the iterates start to deteriorate. Now a natural requirement is convergence of the final iterates to the true solution $x^\dagger$ of (1) when the noise level δ tends to zero. In this case we are also interested in the convergence rate expressed in terms of the noise level δ of the available data. Unfortunately, it is well known that this convergence can be arbitrarily slow unless the true solution $x^\dagger$ satisfies some smoothness condition. In an abstract setting these smoothness conditions are expressed by so-called source conditions given by

$$x_0 - x^\dagger = f(F'[x^\dagger]^* F'[x^\dagger])w, \qquad w \in \mathcal{X}.$$

Here $\|w\|$ is assumed to be "small" and in a general setting introduced by Mathé and Pereverzev [61] the function $f : [0, \|F'[x^\dagger]\|^2] \to [0, \infty)$ is an increasing and continuous function satisfying $f(0) = 0$. So far mainly Hölder source conditions (see (1.13)) and logarithmic source conditions (see (1.14)) have been discussed in the literature on nonlinear inverse problems and optimal rates of convergence of the IRGNM have been established for both of these types of source conditions (see [9, 37]). In this thesis we will give a proof for optimal rates of convergence under general source conditions for both an a-priori stopping criterion (see Theorem 2.4 and Corollary 2.6) and the discrepancy principle (see Theorem 2.7).

Furthermore, our particular interest is in large-scale problems, where the operator F usually represents a partial differential or integral equation in $\mathbb{R}^3$. Under this condition finding the solution of (3) is a complex task and a straightforward implementation of the IRGNM involving the construction of the derivative matrices representing the Frechét derivatives $F'[x_n^\delta]$, $n = 0, 1, 2, \ldots$ is usually not realizable or at least not realizable in an adequate time period. This is due to several reasons. Setting up the derivative matrix incorporates the evaluation of $F'[x_n^\delta]\varphi_j$ for all basis functions φ_j spanning the approximating subspace of $\mathcal{X}$. For large scale problems the time required by this process is not acceptable. Furthermore, often the number of basis functions φ_j is so large that the derivative matrix would not fit into the fast memory of a workstation and even if we had a decomposition of the matrix such that it would fit into the memory, usage of this matrix would be inefficient.

Therefore, we are restricted to iterative solution methods for solving (3) which just require a "black box" to evaluate $F'[x_n^\delta]h$ and $F'[x_n^\delta]^*\tilde{h}$ for some given vectors $h \in \mathcal{X}$ and $\tilde{h} \in \mathcal{Y}$. Since the least-squares problem (3) can be equivalently reformulated by the linear equation

$$(\gamma_n I + F'[x_n^\delta]^* F'[x_n^\delta])h_n = F'[x_n^\delta]^*(y^\delta - F(x_n^\delta)) + \gamma_n(x_0 - x_n^\delta), \tag{4}$$

with the self-adjoint and strictly coercive operator $\gamma_n I + F'[x_n^\delta]^* F'[x_n^\delta]$, a natural choice to solve this problem is the *conjugate gradient method* (CG-method) coupled with an adequate stopping criterion. This method has become the most widespread way of solving systems of this kind. Moreover, it is possible to construct various efficient preconditioner to speed up its convergence rate (see Section 5.2).
Unfortunately, it is well known that a large condition number of the operator is an indicator of slow convergence of the CG-method (see Theorem 4.3). Since for convergence of the IRGNM it is necessary that the regularization parameter γ_n tends to zero, the condition number of the operator in (4), namely $\gamma_n I + F'[x_n^\delta]^* F'[x_n^\delta]$ explodes when n tends to infinity. Actually, by numerical experience the convergence speed of the CG-method for the problems at hand usually deteriorates, and a large number of steps is required until we obtain a reasonable approximation h_n^{app} to the true solution $h_n^\dagger$ of (4). Hence, it is our goal to investigate the accuracy of the final iterates of the IRGNM when the Newton updates are only computed approximately.

Besides the accuracy of an iterative method its efficiency is an important feature to investigate, especially in the situation of large scale problems. For the IRGNM the main complexity consists in finding in each Newton step the solution of (4). One step of the IRGNM where the linear system is solved by the conjugate gradient method usually requires many evaluations of $F'[x_n^\delta]h$ and $F'[x_n^\delta]^*\tilde{h}$ until some stopping criterion is satisfied. For quite a number of nonlinear inverse problems it can be shown that these evaluations are equivalent to finding the solution of a well-posed integral or differential equation. We will illustrate these correspondences by examples arising in inverse scattering discussed in Chapter 7. For large-scale problems

the corresponding discrete linear systems often involve more than a thousand unknowns. Hence, to perform one step in the CG-algorithm, high-dimensional linear systems need to be set up and solved, which can be rather time consuming. As a consequence we expect that under these conditions already performing one Newton step is a complex task, in particular when the regularization parameter is small.

To summarize the discussion above we are interested in three aspects, which are of particular importance in the investigation of large-scale inverse problems.

a) **Accuracy**: Assume that the systems (4) cannot be solved exactly in each Newton step. Is it possible to formulate reasonable conditions on the additional error $\|h_n^{\text{app}} - h_n^\dagger\|$ such that convergence rates of optimal order for the final iterates of the IRGNM can still be established?

b) **Complexity**: Assume that we measure the total complexity of the IRGNM by counting the total number of operator evaluations of $F'[x_n^\delta]h$ and $F'[x_n^\delta]^*\tilde{h}$ for some given vectors $h \in \mathcal{X}$ and $\tilde{h} \in \mathcal{Y}$ and $F(x_n^\delta)$. Is it possible to give an upper bound on the total number of operator evaluations until some data-dependent stopping-criterion terminates the IRGNM?

c) **Acceleration**: Assume that the linear systems (4) are solved by the CG-method in each Newton step. Is it possible to construct preconditioners significantly reducing the number of CG-steps to compute h_n^{app}? Moreover, can we show superiority of an accelerated IRGNM when compared with a standard IRGNM?

All three questions will be answered in this thesis. Note that when we speak about the standard IRGNM throughout this thesis we consider the IRGNM with inner CG-iteration.

Before we give a detailed overview on the topics discussed in the following chapters, let us take a closer look at the main ideas to accelerate the IRGNM, since this point has not been considered here so far.

To achieve a speed up of the IRGNM a significant reduction of the total number of operator evaluations of $F'[x_n^\delta]^*\tilde{h}$ and $F'[x_n^\delta]^*\tilde{h}$ is necessary. Therefore, when solving the linear systems (4) by the CG-method a reduction of the number of CG-steps until some stopping criterion is satisfied needs to be realized. It is well known that this aim can be achieved by preconditioning techniques.

While for well-posed problems acceleration of iterative solution methods for linear systems by appropriate preconditioning is well-studied, the design and analysis of preconditioners for ill-posed problems is not so well understood. Since the eigenvalue distribution of the operators in ill-posed problems play an important role and is usually known beforehand, this knowledge can be exploited to construct so-called spectral preconditioners especially appropriate for large-scale exponentially

ill-posed problems (see Section 5.2). For example, when linear inverse problems are solved by the CG-method applied to the normal equation, preconditioning techniques based on the eigenvalue distribution of the corresponding linear operator have been proven to be successful [32, 66]. In this case the well known regularizing properties of the CG-method have been exploited. Besides preconditioners based on spectral information Egger & Neubauer constructed preconditioners exploiting the smoothing properties of the operators arising in ill-posed problems [15] yielding a significant reduction of the total complexity.

Based on the article by Hohage [40] our interest in this thesis is devoted to the analysis and improvement of a "frozen" version of the IRGNM where incremental spectral preconditioners are constructed within Newton's method to accelerate the convergence speed of the inner CG-iterations. Similar to the first idea just described above we precondition the original linear system by manipulating the eigenvalue distribution of the operator $\gamma_n I + F'[x_n^\delta]^* F'[x_n^\delta]$ to achieve improved convergence rates in the inner CG-iterations of the IRGNM. Note that we formally deal with well-posed linear systems given by (4). Still, if the regularization parameter γ_n is small these systems will be ill-conditioned.
Let us briefly review the idea of the preconditioned IRGNM as it was suggested in [40] such that we are in a position to explain our improvements. Assuming that the eigenvalues of the compact operator $F'[x_n^\delta]^* F'[x_n^\delta]$ have an exponential decay, the linear operator $\gamma_n I + F'[x_n^\delta]^* F'[x_n^\delta]$ has a cluster of eigenvalues in a neighborhood of γ_n, whereas only a few eigenvalues are far away from this limit point. Solving the arising linear systems (4) by the CG-method we can exploit its close connection to Lanczos' method, which computes Ritz values and Ritz vectors approximating eigenpairs of $F'[x_n^\delta]^* F'[x_n^\delta]$. In particular, Lanczos' method has a tendency to approximate those eigenvalues with their corresponding eigenvectors, which are not in a neighborhood of γ_n. Since these eigenvalues are well separated usually the approximations are of high quality.
Assume we have exact knowledge of the k_n largest eigenvalues $\lambda_1 \geq \ldots \geq \lambda_{k_n}$ of $F'[x_n^\delta]^* F'[x_n^\delta]$ with their corresponding eigenvectors φ_j, $j = 1, \ldots, k_n$. To reduce the complexity for the inner CG-iterations in the following Newton steps we set up a spectral preconditioner defined by

$$M_n x := \gamma_n x + \sum_{j=1}^{k_n} \lambda_j \langle x, \varphi_j \rangle \varphi_j \tag{5}$$

and solve instead of (4) the mathematically equivalent linear systems

$$M_s^{-1}(\gamma_s I + F'[x_*]^* F'[x_*]) h_s = M_s^{-1} \left(F'[x_*]^* (y^\delta - F(x_s^\delta)) + \gamma_n (x_0 - x_s^\delta) \right), \tag{6}$$

where $x_* := x_n^\delta$ is kept fixed and $s > n$. Note that the k_n known largest eigenvalues of $\gamma_s I + F'[x_*]^* F'[x_*]$ are shifted by the preconditioner M_s to one, whereas the

rest of the spectrum is amplified by the factor $1/\gamma_s$ (see Theorem 4.6). Hence, the standard error estimate for the CG-method (see Theorem 4.3) indicates an improved convergence rate of the CG-method applied to (6) when compared with the non-preconditioned case. In [40] it was shown that this idea leads to a significant reduction of the total complexity when applied to nonlinear exponentially ill-posed problems. Moreover, the final iterates of this frozen IRGNM and the standard one were comparable for the examples presented.

Several reasons yielding an undesirable increase of the total complexity of the frozen IRGNM, have not been considered in [40]. We just mention here two reasons, further ones are pointed out in Section 6.1:

- Lanczos' method has approximated just a few of the largest eigenvalues,
- the linear operator $F'[x_n^\delta]^* F'[x_n^\delta]$ has multiple eigenvalues.

Since it is well known that Lanczos' method approximates at most one of each multiple eigenvalue (see Theorem 3.10) it is clear that a preconditioner given by M_n is unrealistic in practice and serves therefore only as a motivation.
Even more important to ensure efficiency of the preconditioner M_n it is essential to investigate the behavior of the eigenvalues of the preconditioned operator given only approximations to the eigenpairs. We will show in Chapter 5 that the behavior of the eigenvalues is rather sensitive to errors in the eigenelements used to construct M_n, in particular if the targeted eigenvalues are small or clustered. Unfortunately, the widest part of the spectrum of $\gamma_n I + F'[x_n^\delta]^* F'[x_n^\delta]$ satisfies this condition, in particular if the regularization parameter is small. As a consequence one has to be rather careful which approximations computed by Lanczos' method are chosen. To this end we use a-posteriori bounds to select approximations of high quality (see Theorem 3.14). Still, confirmed by the theory and supported by numerical examples preconditioners of the form (5) have their limits if the eigenvalues are too small compared with the errors in the approximations.

To improve the algorithm suggested in [40] we propose to update the preconditioner while performing the frozen IRGNM. Obviously, further spectral information of $F'[x_*]^* F'[x_*]$ is required to make the preconditioner more efficient. To this end, we apply Lanczos' method after having solved the preconditioned equation (6). This yields approximations to eigenpairs of the preconditioned operator $M_s^{-1}(\gamma_s I + F'[x_*]^* F'[x_*])$. By elementary computations these approximations can be used to compute approximations to eigenpairs of $F'[x_*]^* F'[x_*]$ (see Lemma 6.2). Adding this additional spectral information to the preconditioner reduces the total complexity of the frozen IRGNM significantly once again. Besides this idea we have developed another procedure to update the preconditioner, which is based on the approximation properties of the preconditioner to the operator $\gamma_s I + F'[x_*]^* F'[x_*]$. Both algorithms are presented in detail in Chapter 6.

Finally, the work at hand is organized as follows:

It is roughly divided into three parts. Chapters 1 and 2 deal with the theoretical proof of convergence and convergence rates of optimal order for the IRGNM. The fundamentals of the CG-method and Lanczos' method are described in Chapter 3, which will be used to analyze the complexity of the IRGNM and its preconditioned version in Chapter 4. The last part is dedicated to the derivation of the preconditioned IRGNM and numerical examples. These topics can be found in Chapters 5, 6 and 7. More precisely:

In **Chapter 1** we review the basic concepts of the theory of linear regularization methods for ill-posed problem. In particular we recall the concepts of source sets defined by general index functions to investigate the best possible accuracy to recover the solution of a linear ill-posed problem given only noisy data y^δ. This analysis leads to the definition of optimal rates of convergence for ill-posed problems (see Definition 1.6). Subsequently, we show that for linear ill-posed problems regularization methods with a-priori parameter choice rule can be constructed, which yield to optimal rates of convergence (see Theorem 1.11). In particular the link between the qualification of a regularization method and the index function determining the source set is explained and used to prove this assertion. Finally we consider some type of IRGNM in combination with the discrepancy principle when applied to a linear ill-posed problem. In addition we prove optimal rates of convergence for this regularization method (see Section 1.3). This proof serves as illustration for the main ideas for the inherently more difficult nonlinear case.

Chapter 2 is dedicated to the analysis of the IRGNM applied to some general nonlinear ill-posed operator equation in Hilbert spaces. The smoothness of the true solution is expressed by a source condition defined by some general index function. Optimal rates of convergence under these assumptions will be proven for both an a-priori stopping rule and the discrepancy principle (see Corollary 2.6 and Theorem 2.7). The proof includes the important case that (3) cannot be solved exactly in each Newton step. Furthermore, we formulate reasonable conditions on the difference $\|h_n^{\text{app}} - h_n^\dagger\|$ (see (2.24) and (2.32)) such that convergence and optimal rates of convergence for the IRGNM are not destroyed by this additional error. It can be shown that these conditions can be satisfied if (3) is solved by the CG-method coupled with an adequate stopping criterion (see Theorem 4.4). Besides the IRGNM in Section 2.1 other iterative regularization methods, which have been suggested in the literature are reviewed and briefly discussed.

In **Chapter 3** we develop the fundamentals for both a theoretical complexity analysis with inner CG-iteration and an efficient realization of the IRGNM. While writing this thesis it has turned out that none of the textbooks at hand presenting the CG-method and Lanczos' method had an illustration, which fitted into our framework. To this end, we reformulated the CG-method in a general Hilbert space setting for an

arbitrary inner product and some bounded linear operator, which is self-adjoint and strictly coercive with respect to this inner product. Our formulation allows an easy incorporation of a preconditioner into the algorithm. Moreover, we show in a short and precise way the connection of Lanczos' and the CG-method (see Section 3.5). Sections 3.6 and 3.7 are devoted to present error bounds for the approximations computed by Lanczos' method. To determine computable a-posteriori bounds we use the relation of Lanczos' method to the Rayleigh-Ritz method. A purely theoretical error bound shedding light on convergence rates of Lanczos' method is discussed in Theorem 3.15. The result formulated there is known in the literature as Kaniel-Paige convergence theory for Lanczos' method.

Chapter 4 deals with the investigation of the complexity of the IRGNM and its preconditioned version. Moreover, the complexity analysis presented includes both mildly and exponentially ill-posed problems. We exploit the close connection between the iteration error of the CG-method and polynomials (see (4.4)) to derive convergence rates for the CG-method. In particular, we consider polynomials tailored for eigenvalue distributions corresponding to ill-posed problems leading to improved convergence rates (see Theorem 4.13). Splitting the spectrum into the eigenvalues, which lie in a neighborhood of γ_n and the rest, we will prove upper bounds on the total number of CG-steps, which are necessary to satisfy some reasonable stopping criterion. These upper bounds depend for the most part on the degree of ill-posedness of the original problem and the Newton step (see Theorem 4.19). Finally, a simple summation over all Newton steps required to reach the stopping criterion for the outer Newton iteration yields to the total complexity of the IRGNM and its frozen version. Moreover, by results of Chapter 2 the stopping index of the IRGNM can be expressed in terms of the noise level δ. As a consequence we can express the total complexity of the IRGNM in terms of δ (see Theorems 4.20 and 4.21). The complexity analysis confirms quantitatively the superiority of the preconditioned IRGNM when compared with a standard IRGNM.

In **Chapter 5** we switch to the practically relevant case of discrete systems. Our major interest in this chapter is the investigation of the efficiency of preconditioners of the form (5), since they are especially adequate for large-scale ill-posed problems (see Section 5.2). To this end, we carry out a first order analysis to sniff out the dependency of the eigenvalues on the preconditioned operator given only approximate eigenelements for constructing a spectral preconditioner. This analysis motivates the definition of a condition number for the targeted eigenvalues. Furthermore, an upper bound on this condition number is computed (see Definition 5.5 and Corollary 5.7) implying that preconditioners of the form (5) are extremely sensitive to errors in approximations to small and clustered eigenvalues. In Section 5.4 we interpret this result for the problem at hand.

In **Chapter 6** we derive a realization of the preconditioned IRGNM. To this end we have summarized in Section 6.1 the key points, which need to be considered for

an implementation. All simplifying assumptions for a theoretical analysis of the algorithm are not taken into account any more. Hence, a subsequent discussion is put up where additional difficulties arising in practice and suggestions for their solutions are presented. Moreover, in Section 6.2 we present an iterated Lanczos' method to construct incremental spectral preconditioners of the form (5) significantly reducing the complexity required for solving the linear systems (4). More precisely, two different types of iterated Lanczos' methods are studied (see Algorithm 6.5 and 6.6). Finally we incorporate these methods into the frozen IRGNM (see Algorithm 6.8) eliminating the drawbacks of the algorithm suggested in [40].

Numerical examples confirming the superiority of the Algorithm are presented in **Chapter 7**. In particular we consider inverse inhomogeneous medium scattering problems for time-harmonic acoustic and electromagnetic waves in three space dimensions. The problem is to determine the refractive index of an inhomogeneity from far-field measurements. In this chapter we restrict ourselves to a presentation of the main features, which are necessary for an application of our algorithms. In particular we point out how these inverse problems can be described by an operator equation (1) and how the Fréchet derivatives and their adjoints at a point $h \in \mathcal{X}$ and $\tilde{h} \in \mathcal{Y}$, that is $F'[x_n^\delta]h$ and $F'[x_n^\delta]^*\tilde{h}$ can be evaluated without setting up the matrix representing the Fréchet derivative.

Finally, we discuss our results and conclude this thesis with an outlook in **Chapter 8**.

Chapter 1

Linear inverse problems under general source conditions

To construct a stable approximation to the solution of an ill-posed problem given only noisy data many different regularization methods have been established. Whereas for several regularization methods for linear ill-posed problems optimal rates of convergence under general source conditions have been proven, so far such optimal convergence rates for regularization methods for nonlinear ill-posed problems have not been shown. Of particular interest for nonlinear problems are iterative regularization methods of Newton type, as considered in the introduction. Since in each step of such a method a linearized equation is solved, an analysis requires a deep knowledge of regularization for linear problems.

To this end we will review in this chapter the main results of the linear theory. Our exposition mainly follows the articles by Mathé and Pereverzev [61] and Hohage [39]. In particular we will formulate the main definitions and results concerning linear regularization methods under general source conditions.

The chapter is organized as follows: In Section 1.1 we describe and motivate the important definition of optimality. Section 1.2 deals with linear regularization methods, in particular the interplay between their qualification and the index function determining the source set. Moreover, motivated by the IRGNM for nonlinear ill-posed problems we will discuss in Section 1.3 a corresponding iterative regularization method for approximating the solution of a linear ill-posed problem where we stop the iteration by the discrepancy principle. Optimal rates of convergence of this method will be proven. This method together with the proofs serves as an illustration of the inherently more difficult nonlinear case presented in the next chapter.

1.1 Optimality

We consider in this chapter a linear, ill-posed operator equation

$$Ax = y, \qquad y \in R(A), \tag{1.1}$$

where the bounded operator $A : \mathcal{X} \to \mathcal{Y}$ acts between Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$ and $R(A)$ is not closed. Naturally, in applications the right hand side y of (1.1) is given by measured data and is perturbed by noise. So, we assume that instead of y only noisy data $y^\delta \in \mathcal{Y}$ satisfying

$$\|y^\delta - y\| \leq \delta \tag{1.2}$$

are available. The nonnegative noise level δ is assumed to be known. Notice that in general $y^\delta \notin R(A)$.

It is well known that equation (1.1) has a unique solution $x^\dagger \in \mathcal{X}$, which has minimal norm among all solutions of (1.1). $x^\dagger$ is given by $x^\dagger = A^\dagger y$ where $A^\dagger$ denotes the Moore-Penrose generalized inverse of A.

Since (1.1) is ill-posed the generalized inverse $A^\dagger$ is unbounded. Due to our assumption that instead of the exact right hand side y only noisy data y^δ are available, in general $A^\dagger y^\delta$ is not a good approximation to $x^\dagger$. So, in order to obtain a stable approximation to $x^\dagger$, the unbounded operator $A^\dagger$ has to be approximated by a continuous operator. Any (possibly nonlinear) numerical method to approximately recover $x^\dagger$ from noisy data y^δ is described by an arbitrary mapping $R : \mathcal{Y} \to \mathcal{X}$. We consider here numerical methods with the following regularizing properties:

Definition 1.1 *Let $A : \mathcal{X} \to \mathcal{Y}$ be a bounded linear operator between the Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$, $\alpha_0 \in (0, \infty]$ and $y \in D(A^\dagger)$. The family $\{R_\alpha\}$ of continuous (not necessarily linear) operators*

$$R_\alpha : \mathcal{Y} \to \mathcal{X}$$

together with some parameter choice rule *$\alpha : \mathbb{R}^+ \times \mathcal{Y} \to (0, \alpha_0)$ satisfying*

$$\lim_{\delta \to 0} \sup\{\alpha(\delta, y^\delta) : y^\delta \in \mathcal{Y}, \|y^\delta - y\| \leq \delta\} = 0 \tag{1.3}$$

is called a regularization method *for A if*

$$\lim_{\delta \to 0} \sup\{\|R_{\alpha(\delta, y^\delta)} y^\delta - A^\dagger y\| : y^\delta \in \mathcal{Y}, \|y^\delta - y\| \leq \delta\} = 0 \tag{1.4}$$

holds. If α depends only on the noise level δ, we call it an a-priori *parameter choice rule, otherwise an* a-posteriori *parameter choice rule.*

Naturally, we want to investigate the behavior of the error of the approximate solution $Ry^\delta := R_{\alpha(\delta, y^\delta)} y^\delta$ to (1.1) obtained by a regularization method (R_α, α) for

given observations y^δ as the noise level δ tends to 0. To this end we define the *worst case error* over a class $M \subset \mathcal{X}$ of problem instances by

$$err(M, R, \delta) := \sup \left\{ \|Ry^\delta - x^\dagger\| : x^\dagger \in M, \|Ax^\dagger - y^\delta\| \leq \delta \right\}$$

and the *best possible accuracy* by minimizing over all numerical methods, i.e.

$$err(M, \delta) := \inf_{R:\mathcal{Y}\to\mathcal{X}} err(M, R, \delta). \tag{1.5}$$

Unfortunately, it is well known that if $M = \mathcal{X}$ the error $err(M, \delta)$ may converge to 0 arbitrarily slow for $\delta \to 0$. (cf. for example [16, Proposition 3.11, Remark 3.12]). Convergence rates in terms of δ can thus be established only on subsets of $\mathcal{X}$.

Throughout this chapter we are interested in the asymptotic behavior of $err(M, \delta)$ as $\delta \to 0$ when the class of problem instances $M_f(\rho) \subset \mathcal{X}$ is given by

$$M_f(\rho) := \{x \in \mathcal{X} : x = f(A^*A)w, \|w\| \leq \rho\}. \tag{1.6}$$

$M_f(\rho)$ is called a *source set* and $f : [0, \|A\|^2] \to [0, \infty)$ is an index function.

Definition 1.2 *A function $f : [0, \|A\|^2] \to [0, \infty)$ is called an* index function*, if it is increasing, continuous and satisfies $f(0) = 0$.*

In the case where the subset $M \subset \mathcal{X}$ is given by (1.6) it can be shown (cf. [58, 63]) that the infimum in (1.5) is actually attained and that

$$err(M_f(\rho), \delta) = \sup\{\|x\| : x \in M_f(\rho), \|Ax\| \leq \delta\}. \tag{1.7}$$

Furthermore, it is known (see Engl & Hanke & Neubauer [16] for linear and Bakushinskii & Kokurin [7] for nonlinear inverse problems) that a so-called *source condition* $x^\dagger \in M_f(\rho)$, that is

$$x^\dagger = f(A^*A)w, \qquad \|w\| \leq \rho, \tag{1.8}$$

is also almost necessary to prove rates of convergence. As the operator A is usually smoothing conditions in the form of (1.8) can often be interpreted as abstract smoothness conditions for some given index function f (see for example [16, 75]). The behavior of f near 0 determines how much smoothness of $x^\dagger$ is required compared to the smoothing properties of A^*A.

In many cases there exists an explicit formula for the right hand side of (1.7) (see for example Ivanov & Korolyuk [43]). Following Tautenhahn [81] to derive a formula for the right hand side of (1.7) we have to impose a further condition on the index function f.

Assumption 1.3 *Let $f \in C[0, \|A\|^2]$ be a strictly monotonically increasing index function for which the function $\Phi : [0, f(\|A\|^2)^2] \to [0, \|A\|^2 f(\|A\|^2)^2]$ defined by*

$$\Phi(t) := t(f \cdot f)^{-1}(t) \tag{1.9}$$

is convex and twice differentiable.

Under this assumption the following stability result holds true.

Lemma 1.4 *Assume that the index function f satisfies Assumption 1.3 and that $x \in M_f(\rho)$. Then x satisfies the stability estimate*

$$\|x\|^2 \leq \rho^2 \Phi^{-1}\left(\frac{\|Ax\|^2}{\rho^2}\right) = \rho^2 f^2\left(u^{-1}\left(\frac{\|Ax\|}{\rho}\right)\right), \tag{1.10}$$

where the function u is defined by

$$u(\lambda) := \sqrt{\lambda} f(\lambda). \tag{1.11}$$

Consequently,

$$\sup\{\|x\| : x \in M_f(\rho), \|Ax\| \leq \delta\} \leq \rho f(u^{-1}(\delta/\rho)),$$

for $\delta \leq \rho \|A\| f(\|A\|^2)$.

Proof: Due to the assumptions on f and Φ the function Φ is invertible and an application of Jensen's inequality gives us the estimate in (1.10) (see Mair [59, Theorem 2.10]). The equality in (1.10) is a consequence of the identity $\Phi^{-1}(t^2) = f^2(u^{-1}(t))$, which follows from

$$\Phi(f^2(u^{-1}(t))) = f^2(\xi)(f \cdot f)^{-1}(f^2(\xi)) = f^2(\xi)\xi = [u(\xi)]^2 = t^2$$

with $\xi = u^{-1}(t)$.

□

An application of Lemma 1.4 now yields the following result:

Proposition 1.5 *Assume that the index function f satisfies Assumption 1.3, and let $\tilde{\Phi} : [0, f(\|A\|^2)^2] \to [0, \|A\|^2 f(\|A\|^2)^2]$ be the largest convex function satisfying (1.9) for all $t \in \{f(\lambda)^2 : \lambda \in \sigma(A^*A) \cup \{0\}\}$. Then*

$$\sup\{\|x\| : x \in M_f(\rho), \|Ax\| \leq \delta\} = \rho \sqrt{\tilde{\Phi}^{-1}(\delta^2/\rho^2)} \tag{1.12}$$

for $\delta \leq \rho \|A\| f(\|A\|^2)$, and

$$\sup\{\|x\| : x \in M_f(\rho), \|Ax\| \leq \delta\} = \rho f(\|A\|^2)$$

for $\delta > \rho \|A\| f(\|A\|^2)$.

Proof: See Hohage [39, Proposition 2].

□

Proposition 1.5 together with (1.7) answers the question, what the *best possible accuracy* over all numerical methods to recover $x^\dagger$ is as the noise level δ tends to 0 provided that Assumption 1.3 holds. Motivated by this discussion we recall the following definition (see Hohage [39, Definition 3]).

Definition 1.6 *Let (R_α, α) be a regularization method for (1.1), and let Assumption 1.3 be satisfied. Convergence on the source sets $M_f(\rho)$ is said to be*

- *optimal if*
$$err(M_f(\rho), R_\alpha, \delta) \le \rho f(u^{-1}(\delta/\rho)),$$
- *asymptotically optimal if*
$$err(M_f(\rho), R_\alpha, \delta) \ = \rho f(u^{-1}(\delta/\rho))(1+o(1)), \qquad \delta \to 0,$$
- *of optimal order if there is a constant $C \ge 1$ such that*
$$err(M_f(\rho), R_\alpha, \delta) \le C\rho f(u^{-1}(\delta/\rho))$$
for δ/ρ sufficiently small.

So far two classes of index functions have been discussed with major interest in the literature. The first class leading to *Hölder type source conditions* is given by

$$f(t) := t^\nu, \qquad 0 < \nu \le 1. \tag{1.13}$$

So-called *logarithmic source conditions* are described by the functions

$$f(t) := \begin{cases} (-\ln t)^{-p}, & 0 < t \le \exp(-1), \\ 0, & t = 0, \end{cases} \qquad p > 0. \tag{1.14}$$

The former conditions are usually appropriate for mildly ill-posed problems, i.e. finitely smoothing operators A whereas the latter conditions (where the scaling condition $\|A\|^2 \le \exp(-1)$ must be imposed) lead to natural smoothness conditions in terms of Sobolev spaces for a number of exponentially ill-posed problems. A generalization of the latter functions were discussed by Mathé & Pereverzev [61]. For Hölder type source conditions it can be shown by direct computations that the corresponding functions Φ defined by (1.9) are convex and twice differentiable. For logarithmic source conditions a proof of these properties can be found in [59].

Another class of index functions, which have been considered by Hähner and Hohage [29], are given by

$$f(t) := \begin{cases} \exp\left(-\frac{1}{2}(-\ln t)^\theta\right), & 0 < t \le \exp(-1), \\ 0, & t = 0, \end{cases} \qquad 0 < \theta < 1. \tag{1.15}$$

The corresponding source conditions are stronger than logarithmic, but weaker than Hölder source conditions. The functions Φ defined in (1.9) and their second derivatives in this case are given by

$$\begin{aligned}\Phi(t) &= t\exp\left(-(-\ln t)^{1/\theta}\right), \qquad 0 < t \leq \exp(-1),\\ \Phi''(t) &= \exp\left((-\ln t)^{1/\theta}\right)\left(\frac{(-\ln t)^{1/\theta-2}}{\theta t}\right)\left(1-\ln t+\frac{(-\ln t)^{1/\theta}-1}{\theta}\right).\end{aligned}$$

It is obvious that (1.15) defines an index function and that $\Phi''(t) > 0$ for $0 < \theta < 1$ and $0 < t < \exp(-1)$.
But, to the author's knowledge so far there exist only examples where source conditions given by the index functions (1.13) and (1.14) could be interpreted as abstract smoothness conditions.

1.2 Linear regularization methods

We now consider a class of regularization methods based on spectral theory for self-adjoint linear operators. More precisely, we analyze regularization methods (R_α, α) of the form

$$R_\alpha y^\delta := g_\alpha(A^*A)A^*y^\delta \tag{1.16}$$

with some functions $g_\alpha \in C[0, \|A\|^2]$ depending on some regularization parameter $\alpha > 0$. (1.16) has to be understood in the sense of the functional calculus. For an introduction to spectral theory for selfadjoint operators we refer to [16] and [36]. The function g_α is also called a filter. Corresponding to g_α we define the function

$$r_\alpha(t) := 1 - tg_\alpha(t), \qquad t \in [0, \|A\|^2]. \tag{1.17}$$

Now we will study the connection between the qualification of a regularization method specified by the function g_α and properties of an index function f. To this end we recall a definition given by Mathé and Pereverzev [61].

Definition 1.7 *A family $\{g_\alpha\}$, $0 < \alpha \leq \|A\|^2$ is called* regularization*, if there are constants C_r and C_g for which*

$$\sup_{0<t\leq\|A\|^2} |r_\alpha(t)| \leq C_r, \qquad 0 < \alpha \leq \|A\|^2, \tag{1.18}$$

and

$$\sup_{0<t\leq\|A\|^2} \sqrt{t}|g_\alpha(t)| \leq \frac{C_g}{\sqrt{\alpha}}, \qquad 0 < \alpha \leq \|A\|^2. \tag{1.19}$$

The regularization is said to have qualification ξ*, if*

$$\sup_{0<t\leq\|A\|^2} |r_\alpha(t)|\xi(t) \leq C_r\xi(\alpha), \qquad 0 < \alpha \leq \|A\|^2,$$

for an increasing function $\xi : (0, \|A\|^2) \to \mathbb{R}_+$.

In the following theorem we show the connection between Definition 1.1 and Definition 1.7. The assertion can be also found for example in [16]. To shorten the notation we denote the reconstructions for exact and noisy data by $x_\alpha := R_\alpha y$ and $x_\alpha^\delta := R_\alpha y^\delta$. Hence, the reconstruction error for exact data is given by

$$x^\dagger - x_\alpha = (I - g_\alpha(A^*A)A^*A)x^\dagger = r_\alpha(A^*A)x^\dagger. \tag{1.20}$$

Theorem 1.8 *Assume that the family $\{g_\alpha\}$ is a regularization, which additionally satisfies*

$$\lim_{\alpha\to 0} r_\alpha(t) = \begin{cases} 0, & t > 0, \\ 1, & t = 0. \end{cases} \tag{1.21}$$

Then the operators R_α defined by (1.16) converge pointwise to $A^\dagger$ on $D(A^\dagger)$ as $\alpha \to 0$. If α is a parameter choice rule satisfying

$$\alpha(\delta, y^\delta) \to 0, \qquad \text{and} \qquad \delta/\sqrt{\alpha(\delta, y^\delta)} \to 0 \quad \text{as} \quad \delta \to 0, \tag{1.22}$$

then (R_α, α) is a regularization method.

Proof: Let $y \in D(A^\dagger)$. Using (1.20) and condition (1.18), it follows by an application of the functional calculus that

$$\lim_{\alpha\to 0} r_\alpha(A^*A)x^\dagger = r_0(A^*A)x^\dagger,$$

where r_0 denotes the limit function defined by the right hand side of (1.21). Since r_0 is real valued and $r_0^2 = r_0$, the operator $r_0(A^*A)$ is an orthogonal projection. Moreover, $R(r_0(A^*A)) \subset N(A^*A)$ since $tr_0(t) = 0$ for all t. Hence,

$$\|r_0(A^*A)x^\dagger\|^2 = \left\langle r_0(A^*A)x^\dagger, x^\dagger \right\rangle = 0 \qquad \text{as} \qquad x^\dagger \in N(A)^\perp = N(A^*A)^\perp.$$

This yields

$$\lim_{\alpha\to 0} \|R_\alpha y - A^\dagger y\|^2 = \lim_{\alpha\to 0} \|r_\alpha(A^*A)x^\dagger\|^2 = 0. \tag{1.23}$$

Now, by the isometry of the functional calculus and (1.19) we obtain for all $z \in \mathcal{Y}$

$$\|R_\alpha z\| = \|A^* g_\alpha(AA^*)z\| = \|(AA^*)^{1/2} g_\alpha(AA^*)z\| \le \|\sqrt{t} g_\alpha(t)\|_\infty \|z\| \le \frac{C_g}{\sqrt{\alpha}}\|z\|.$$

We now split the total error into the approximation and the data noise error,

$$\|x^\dagger - x_\alpha^\delta\| \le \|x^\dagger - x_\alpha\| + \|x_\alpha - x_\alpha^\delta\|.$$

Due to the first assumption in (1.22) and (1.23) we observe that the reconstruction error $\|x^\dagger - x_\alpha\| \to 0$ as $\delta \to 0$. The data noise error

$$\|x_\alpha - x_\alpha^\delta\| = \|R_\alpha(y - y^\delta)\| \le C_g \frac{\delta}{\sqrt{\alpha}}$$

tends to zero by the second assumption in (1.22).

□

In a classical setting the qualification $p \in [0, \infty]$ of a regularization $\{g_\alpha\}$ is defined by the inequality

$$\sup_{0<t\leq\|A\|^2} t^q |r_\alpha(t)| \leq C_q \alpha^q, \qquad \text{for every} \quad 0 \leq q \leq p,$$

and some constant $C_q > 0$. In this case, we call this *classical qualification of order* p. That is classical qualifications are special cases of the general Definition 1.7 by using polynomials of prescribed degree.

For example, Tikhonov regularization given by the functions

$$g_\alpha(t) = \frac{1}{\alpha + t}$$

has qualification $\xi(t) = t$ in the sense of Definition 1.7, since

$$|r_\alpha(t)|t = \frac{\alpha t}{\alpha + t} \leq \alpha.$$

In the classical sense Tikhonov regularization has qualification order 1 and one can show that this is the maximal qualification order of Tikhonov regularization. Following [61] we now turn to study the connection between the qualification ξ of a regularization and an index function f.

Definition 1.9 *The qualification* ξ covers *an index function* f*, if there is a constant* $c > 0$ *such that*

$$c\frac{\xi(\alpha)}{f(\alpha)} \leq \inf_{\alpha\leq t\leq\|A\|^2} \frac{\xi(t)}{f(t)}, \qquad 0 < \alpha \leq \|A\|^2.$$

Theorem 1.11 below illuminates the correspondence between the qualification of a regularization method and an index function f representing the smoothing properties of the operator A^*A. The next lemma serves as a preparation.

Lemma 1.10 *Let* f *be a non-decreasing index function and let* $\{g_\alpha\}$ *be a regularization with qualification* ξ *that covers* f*. Then*

$$\sup_{0<t\leq\|A\|^2} |r_\alpha(t)| f(t) \leq \frac{C_r}{c} f(\alpha), \qquad 0 < \alpha \leq \|A\|^2.$$

In particular, for Tikhonov regularization we have that $C_r = 1$*.*

Proof: See [61, Proposition 3].

□

Theorem 1.11 *Let f be an index function which satisfies Assumption 1.3 and $x^\dagger \in M_f(\rho)$. If the regularization parameter α is chosen to satisfy $u(\alpha) = \delta/\rho$, where u is given by (1.11), and the regularization $\{g_\alpha\}$ covers f with constant c, then the convergence $\|x_\alpha^\delta - x^\dagger\| \to 0$ is of optimal order as δ/ρ tends to 0.*

Proof: By splitting the error into the approximation and the data noise error we can estimate using (1.8), (1.18), (1.19), (1.20) and Lemma 1.10

$$\begin{aligned}
\|x^\dagger - x_\alpha^\delta\| &\leq \|r_\alpha(A^*A)f(A^*A)w\| + \|g_\alpha(A^*A)A^*(y-y^\delta)\| \\
&\leq \rho \sup_{0<t\leq\|A\|^2} |r_\alpha(t)|\, f(t) + \delta \sup_{0<t\leq\|A\|^2} \left|\sqrt{t} g_\alpha(t)\right| \\
&\leq \rho \frac{C_r}{c} f(\alpha) + \delta \frac{C_g}{\sqrt{\alpha}} \\
&= \rho \left(\frac{C_r}{c} + C_g\right) f(\alpha).
\end{aligned}$$

Since $\alpha = u^{-1}(\delta/\rho)$, the assertion follows.

□

Theorem 1.11 shows that for a source set (1.6) defined by an arbitrary index function f satisfying Assumption 1.3 regularization methods with a-priori parameter choice rule can be constructed leading to convergence of optimal order. In the next section we will show that convergence of optimal order can also be obtained if we use the discrepancy principle to determine the regularization parameter.
We want to close this section with a corollary, which we will need later in the chapter followed.

Corollary 1.12 *Assume that $\{g_\alpha\}$ has qualification $t \mapsto \sqrt{t}f(t)$ for an index function $f : [0, \|A\|^2] \to [0, \infty)$. Then $\{g_\alpha\}$ has qualification $t \mapsto f(t)$.*

Proof: Since $\{g_\alpha\}$ has qualification $t \mapsto \sqrt{t}f(t)$ the estimate

$$\sup_{0<t\leq\|A\|^2|} |r_\alpha(t)|\sqrt{t}f(t) \leq C_r\sqrt{\alpha}f(\alpha), \qquad 0 < \alpha \leq \|A\|^2,$$

holds. The equality

$$\frac{\sqrt{\alpha}f(\alpha)}{f(\alpha)} = \inf_{\alpha\leq t\leq\|A\|^2} \frac{\sqrt{t}f(t)}{f(t)}$$

shows that the mapping $t \mapsto \sqrt{t}f(t)$ covers f with constant $c = 1$. An application of Lemma 1.10 now yields

$$\sup_{0<t\leq\|A\|^2} |r_\alpha(t)|f(t) \leq C_r f(\alpha), \qquad 0 < \alpha \leq \|A\|^2,$$

which proves the assertion.

□

1.3 Discrepancy principle for linear problems

Before we analyze convergence rates of the IRGNM for nonlinear ill-posed problems in chapter 2, we close this chapter by studying a corresponding iterative regularization method for the special case of the linear ill-posed operator equation (1.1) where the right hand side y is replaced by noisy data y^δ satisfying (1.2). We assume that the true solution of (1.1) satisfies a source condition, that is $x^\dagger \in M_f(\rho)$ (see (1.6)) for a given index function f and a given bound $\rho > 0$. Motivated by a regularized Newton method as presented in the introduction we consider the Tikhonov-regularized solution of (1.1) defined by

$$x_{n+1}^\delta = (\gamma_n I + A^*A)^{-1} A^* y^\delta. \tag{1.24}$$

The iterates (1.24) correspond to the iterates of the IRGNM applied to (1.1) with initial guess $x_0 = 0$. Here (γ_n) is a fixed sequence satisfying

$$\lim_{n\to\infty} \gamma_n = 0 \qquad \text{and} \qquad 1 \le \frac{\gamma_n}{\gamma_{n+1}} \le \gamma \tag{1.25}$$

for some $\gamma > 1$.
Dealing with ill-posed problems the choice of some data-dependent stopping rule is an important issue. On the one hand the iteration should not stop too early. In this case a better reconstruction out of noisy data y^δ can be computed. On the other hand the stopping index should not be too large, since typically the iterations deteriorate quite rapidly. We consider as stopping rule the well-known Morozov discrepancy principle, i.e. we stop the iteration at the first index N, for which the residual $\|Ax_N^\delta - y^\delta\|$ satisfies

$$\|Ax_N^\delta - y^\delta\| \le \tau\delta < \|Ax_n^\delta - y^\delta\|, \qquad 0 \le n < N, \tag{1.26}$$

for a fixed parameter $\tau > 1$. In the last years also Lepskij-type stopping rules have been considered (see [8]).
Our aim is to show that in the linear case the discrepancy principle yields optimal rates of convergence for a certain class of index functions. This result was originally published in [62]. We will prove it here in a different way based on Assumption 1.3 and Lemma 1.4, that is for a class of index functions that guarantees inequality (1.10). Our intention is to illustrate in the linear case the main idea to prove convergence rates of the IRGNM in the nonlinear case, which will be treated in the next chapter.

To prove optimal rates of convergence we first have to formulate some additional assumptions on the index function f. To shorten the notation we make the definitions

$$g_n(\lambda) := \frac{1}{\gamma_n + \lambda}, \qquad \text{and} \qquad r_n(\lambda) := 1 - \lambda g_n(\lambda). \tag{1.27}$$

Note that $x_{n+1}^\delta = g_n(A^*A)A^*y^\delta$. So g_n denotes the filter corresponding to Tikhonov regularization. To formulate our convergence result for general source conditions as presented in Sections 1.1 and 1.2, we assume that the regularization additionally satisfies the inequality

$$\sup_{0<\lambda\le\|F'[x^\dagger]\|^2} |r_n(\lambda)|\sqrt{\lambda}f(\lambda) \;\le\; c_f\sqrt{\gamma_n}f(\gamma_n), \qquad n\in\mathbb{N}_0, \tag{1.28a}$$

that is it has the qualification $t\mapsto\sqrt{t}f(t)$. We further assume that

$$\frac{f(\gamma\lambda)}{f(\lambda)} \;\le\; C_f \qquad \text{for all} \qquad \lambda\in(0,\|A\|^2/\gamma]. \tag{1.28b}$$

The class of index function determined by (1.28) and (1.28b) corresponds to the index function class $\mathcal{F}_{c_f,C_f}$ defined in [61]. Note, as in the proof of Theorem 1.11 with the help of (1.16), (1.20) and (1.8) we can decompose the total error $x^\dagger - x_n^\delta$ into the approximation error e_{n+1}^{app} and the data noise error e_{n+1}^{noi}, more precisely

$$x^\dagger - x_n^\delta = e_{n+1}^{\text{app}} + e_{n+1}^{\text{noi}},$$

where

$$e_{n+1}^{\text{app}} \;:=\; r_n(A^*A)f(A^*A)w, \tag{1.29a}$$
$$e_{n+1}^{\text{noi}} \;:=\; g_n(A^*A)A^*(y^\delta - y). \tag{1.29b}$$

We now state the main theorem establishing optimal rates of convergence for the final iterates x_N^δ produced by the sequence (1.24), where the stopping index N is determined by the discrepancy principle. A similar result can be found in [62].

Theorem 1.13 *Assume that $Ax^\dagger = y$, $x^\dagger\in M_f(\rho)$, and that (1.2), (1.25), (1.28a), (1.28b) and Assumption 1.3 hold. Let x_n^δ be defined by (1.24), and let $N\ge 1$ be determined by the discrepancy principle (1.26). ($N\ge 1$ implies that γ_0 must be sufficiently large.) Then*

$$\|x_N^\delta - x^\dagger\| \le C\rho f\left(u^{-1}\left(\frac{\delta}{\rho}\right)\right)$$

for $\delta/\rho\le\|A\|$, i.e. the convergence $\|x_N^\delta - x^\dagger\|\to 0$ as the noise level δ tends to 0 is of optimal order in the sense of Definition 1.6.

Proof: It is our goal to prove for the error components e_{n+1}^{app} and e_{n+1}^{noi} the predicted behavior as $\delta/\rho\to 0$ separately. To prove this behavior for e_{n+1}^{app} the observation

$$r_n(A^*A)x^\dagger = f(A^*A)r_n(A^*A)w \in M_f(\rho)$$

is crucial, since it implies that e_{n+1}^{app} satisfies a source condition. Hence, we can apply inequality (1.10) to obtain

$$\|e_{n+1}^{\mathrm{app}}\| \leq \rho f\left(u^{-1}\left(\frac{\|Ae_{n+1}^{\mathrm{app}}\|}{\rho}\right)\right). \tag{1.30}$$

By (1.29b) and the definition of r we have the estimate

$$\begin{aligned}\|Ae_{n+1}^{\mathrm{noi}} - (y^\delta - y)\| &= \|Ag_n(A^*A)A^*(y^\delta - y) - (y^\delta - y)\| \\ &= \|r_n(AA^*)(y^\delta - y)\| \leq \delta,\end{aligned}$$

which in the case $N = n + 1$ leads to

$$\begin{aligned}\|Ae_N^{\mathrm{app}}\| &\leq \|y^\delta - Ax_N^\delta\| + \|Ax_N^\delta - y^\delta + Ae_N^{\mathrm{app}}\| \\ &= \|y^\delta - Ax_N^\delta\| + \|Ag_{N-1}(A^*A)A^*y^\delta - y^\delta + A(I - A^*Ag_{N-1}(A^*A))x^\dagger\| \\ &= \|y^\delta - Ax_N^\delta\| + \|Ag_{N-1}(A^*A)A^*y^\delta - y^\delta + y - Ag_{N-1}(A^*A))A^*y\| \\ &= \|y^\delta - Ax_N^\delta\| + \|Ae_N^{\mathrm{noi}} - (y^\delta - y)\| \\ &\leq \|y^\delta - Ax_N^\delta\| + \delta \leq (\tau + 1)\delta,\end{aligned}$$

where we have used (1.29a), (1.29b), (1.24) and (1.26). Hence, inserting this into (1.30) it follows by an application of the inequality

$$f\left(u^{-1}\left(t\frac{\delta}{\rho}\right)\right) \leq tf\left(u^{-1}\left(\frac{\delta}{\rho}\right)\right), \qquad t \geq 1, \tag{1.31}$$

which is due to the concavity of $f \circ u^{-1}$, that

$$\|e_N^{\mathrm{app}}\| \leq \rho f\left(u^{-1}\left(\frac{(\tau+1)\delta}{\rho}\right)\right) \leq (\tau+1)f\left(u^{-1}\left(\frac{\delta}{\rho}\right)\right). \tag{1.32}$$

We continue by estimating $\|e_N^{\mathrm{noi}}\|$. Since, by (1.2) and (1.29b)

$$\|e_N^{\mathrm{noi}}\| \leq \|g_{N-1}(A^*A)A^*\|\delta = \|(\gamma_{N-1}I + A^*A)^{-1}A^*\|\delta \leq \frac{\delta}{2\sqrt{\gamma_{N-1}}},$$

in order to prove the assertion we need to show that

$$\frac{\delta}{\sqrt{\gamma_{N-1}}} \leq C\rho f\left(u^{-1}\left(\frac{\delta}{\rho}\right)\right)$$

with a constant $C > 0$. Using the triangle inequality and (1.26) it follows

$$\begin{aligned}\|Ae_{N-1}^{\mathrm{app}}\| &\geq \|y^\delta - Ax_{N-1}^\delta\| - \|Ae_{N-1}^{\mathrm{noi}} - (y^\delta - y)\| \\ &> \tau\delta - \delta = (\tau - 1)\delta.\end{aligned}$$

Therefore, (1.25), (1.29a) and assumptions (1.28a) and (1.28b) imply

$$\begin{aligned}\delta &< \frac{\|Ae_{N-1}^{\text{app}}\|}{(\tau-1)} \\ &= \frac{\|(A^*A)^{1/2} r_{N-2}(A^*A) f(A^*A) w\|}{(\tau-1)} \\ &\leq \frac{c_f}{(\tau-1)} u(\gamma_{N-2})\rho \leq \frac{c_f C_f \gamma}{(\tau-1)} u(\gamma_{N-1})\rho, \end{aligned} \tag{1.33}$$

which yields

$$u^{-1}\left(\left(\frac{\tau-1}{c_f C_f \gamma}\right)\left(\frac{\delta}{\rho}\right)\right) \leq \gamma_{N-1}. \tag{1.34}$$

Therefore,

$$\begin{aligned}\frac{\delta}{\sqrt{\gamma_{N-1}}} &= \rho\left(\frac{c_f C_f \gamma}{\tau-1}\right) \frac{u\left(u^{-1}\left(\left(\frac{\tau-1}{c_f C_f \gamma}\right)\left(\frac{\delta}{\rho}\right)\right)\right)}{\sqrt{\gamma_{N-1}}} \\ &= \rho\left(\frac{c_f C_f \gamma}{\tau-1}\right) \sqrt{\frac{u^{-1}\left(\left(\frac{\tau-1}{c_f C_f \gamma}\right)\left(\frac{\delta}{\rho}\right)\right)}{\gamma_{N-1}}} f\left(u^{-1}\left(\left(\frac{\tau-1}{c_f C_f \gamma}\right)\left(\frac{\delta}{\rho}\right)\right)\right) \\ &\leq \rho\left(\frac{c_f C_f \gamma}{\tau-1}\right) f\left(u^{-1}\left(\max\left\{\frac{\tau-1}{c_f C_f \gamma}, 1\right\}\left(\frac{\delta}{\rho}\right)\right)\right) \\ &\leq \rho\left(\frac{c_f C_f \gamma}{\tau-1}\right) \max\left\{\frac{\tau-1}{c_f C_f \gamma}, 1\right\} f\left(u^{-1}\left(\frac{\delta}{\rho}\right)\right) \\ &= \max\left\{1, \frac{c_f C_f \gamma}{\tau-1}\right\} \rho f\left(u^{-1}\left(\frac{\delta}{\rho}\right)\right). \end{aligned}$$

In the second line we have used the definition of u (see (1.11)), in the third line inequality (1.34) and the monotonicity of $f \circ u^{-1}$, and in the fourth line inequality (1.31). Altogether we have proved

$$\|x_N^\delta - x^\dagger\| \leq \left(\tau + 1 + \frac{1}{2}\max\left\{1, \frac{c_f C_f \gamma}{\tau-1}\right\}\right) \rho f\left(u^{-1}\left(\frac{\delta}{\rho}\right)\right),$$

which shows the assertion.

□

Remark 1.14 *The assertion of Theorem 1.13 is not restricted to Tikhonov regularization. The result remains true for any regularization $\{g_\alpha\}$ satisfying $\|r_\alpha\|_\infty \leq 1$, since we can always split the total error into the approximation error and the data noise error.*

We will discuss the main points of the proof of Theorem 1.13 at the beginning of Section 2.3 since it includes the main ideas to treat the nonlinear case, which is inherently more difficult.

We can conclude from inequality (1.33) an upper bound for the total number of steps until the stopping criterion is satisfied given noisy data y^δ with noise level $\delta > 0$.

Corollary 1.15 *Let the assumptions of Theorem 1.13 hold. If $\delta > 0$ and the regularization parameters γ_n are chosen by*

$$\gamma_n = \gamma_0 \gamma^{-n}, \qquad n = 0, 1, 2, \ldots,$$

then the stopping index is finite and we have $N = O(-\ln(u^{-1}(\delta/\rho)))$, $\delta/\rho \to 0$. The function u is given by (1.11).

Proof: From inequality (1.33) we conclude

$$\frac{\delta}{\rho} \leq C u(\gamma_{N-1}) \tag{1.35}$$

with some constant $C > 0$ and for $\delta/\rho > 0$. Hence, the stopping index is finite with

$$N = \mathrm{O}\left(-\ln\left(u^{-1}\left(\frac{\delta}{\rho}\right)\right)\right)$$

for the choice $\gamma_n = \gamma_0 \gamma^{-n}$.

□

Chapter 2

Convergence analysis of an inexact iteratively regularized Gauss-Newton method under general source conditions

In Chapter 1 we presented some of the main results of the classical theory for linear ill-posed problems dealing with error estimates for approximate solutions for a certain data noise level $\delta > 0$. In particular we have reviewed the convergence of linear regularization methods under source conditions with general index functions as studied in a series of articles starting with the work of Mathé & Pereverzev [61]. From this point of view the theory dealing with linear ill-posed problems is rather complete.

For nonlinear ill-posed problems iterative regularization methods have been established to construct a stable approximation to the true solution given only noisy data. In the past a lot of iterative regularization methods have been investigated under either Hölder source conditions or logarithmic source conditions. For these source conditions often convergence rate results could be shown. But so far no convergence rate results have been proven for these methods under general source conditions, that is the source sets are determined by an index function.

It is the topic of this chapter to obtain convergence and convergence rate results for the iteratively regularized Gauss-Newton method (IRGNM) under such general source conditions where the stopping index is determined by the discrepancy principle. Moreover, we will prove under these conditions and for an a-priori stopping rule convergence rates of optimal order. Furthermore, our proof involves the realistic assumption that in each Newton step the linearized equation cannot be solved exactly, an important issue which has also not been considered so far. The approximate solution of the linearized equation will be the topic of the following chapters of this thesis.

This chapter is organized as follows: in Section 2.1 we give a precise description of the abstract mathematical setting suited for this topic. Moreover, we give a short outlook and characterization of some other methods to solve nonlinear ill-posed problems. The Sections 2.2 and 2.3 deal with the rather technical proof of convergence and convergence rates for the IRGNM. In Section 2.3 we will show that the discrepancy principle leads to order optimal rates of convergence under the usual conditions concerning the finite qualification of Tikhonov regularization. This result can be interpreted as generalization of Theorem 1.13.

2.1 Iterative regularization methods for nonlinear ill-posed problems

Let us consider a nonlinear, ill-posed operator equation

$$F(x) = y, \tag{2.1}$$

where the operator $F : D(F) \to \mathcal{Y}$ is injective and continuously Fréchet differentiable on its domain $D(F) \subset \mathcal{X}$. We assume that there exists an $x^\dagger \in D(F)$ with

$$F(x^\dagger) = y \tag{2.2}$$

and that only noisy data $y^\delta \in \mathcal{Y}$ are available satisfying

$$\|y^\delta - y\| \leq \delta \tag{2.3}$$

with known noise level $\delta > 0$.

Definition 2.1 *An iterative method $x^\delta_{n+1} := \Phi(x^\delta_n, \dots, x^\delta_1, x_0, y^\delta)$ together with a stopping rule $N = N(\delta, y^\delta)$ is called an* iterative regularization method *for F if for all $x^\dagger \in D(F)$, all y^δ satisfying (2.3), and all initial guesses x_0 sufficiently close to $x^\dagger$ the following conditions hold:*

a) x^δ_n is well defined for $n = 1, \dots, N$ and $N < \infty$ for $\delta > 0$.

b) For exact data ($\delta = 0$) either $N < \infty$ and $x^\delta_N = x^\dagger$ or $N = \infty$ and

$$\|x_n - x^\dagger\| \to 0, \qquad n \to \infty.$$

c) The following regularization property holds:

$$\sup_{\|y^\delta - y\| \leq \delta} \|x^\delta_N - x^\dagger\| \to 0, \qquad \delta \to 0.$$

As in the linear case for a given arbitrary iterative method Φ together with a stopping rule, convergence of $\|x_N^\delta - x^\dagger\| \to 0$ as the noise level $\delta \to 0$ may be arbitrarily slow unless a *source condition* is satisfied. For nonlinear problems these conditions have the form

$$x_0 - x^\dagger = f(F'[x^\dagger]^* F'[x^\dagger])w, \tag{2.4}$$

where $f : [0, \|F'[x^\dagger]\|^2] \to [0, \infty)$ is again considered to be an index function, and $w \in \mathcal{X}$ is 'small', i.e. $\|w\| \leq \rho$ for a $\rho > 0$. Analogous to the linear case, $F'[x^\dagger]$ is usually smoothing and (2.4) can be often interpreted as an abstract smoothness condition (see for example [37, 44]).

We want to investigate the behavior of the error, which is committed by approximating the solution of (2.1) by a given iterative regularization method given noisy data y^δ as the noise level δ tends to 0. To this end by a similar approach as in the linear case we define the *worst case error*

$$err(N_f(\rho), (\Phi, N), \delta, x_0) := \sup\left\{\|x_{N(\delta, y^\delta)}^\delta - x^\dagger\| : x_0 - x^\dagger \in N_f(\rho), \|F(x^\dagger) - y^\delta\| \leq \delta\right\},$$

where the *source set* is given by

$$N_f(\rho) := \left\{x_0 - x \in \mathcal{X} : x_0 - x = f(F'[x^\dagger]^* F'[x^\dagger])w, \|w\| \leq \rho\right\}.$$

Unfortunately, in the nonlinear case no explicit characterization of the best possible accuracy is known in general. However, since for a linear problem the best possible accuracy is given by (1.12), we cannot expect better accuracies for nonlinear problems as it should subsume the linear case. Hence, when speaking of convergence of optimal order for an iterative regularization method for a nonlinear ill-posed problem this concept is understood as in the linear case. In [5] the necessity of the source condition (2.4) to obtain certain rates of convergence has been discussed even for nonlinear problems.

Corresponding to the iterative method (1.24) for a linear ill-posed problem also for nonlinear ill-posed problems the choice of some data-dependent stopping rule is an important issue, in particular for iterative regularization methods. Convergence results for some iterative regularization methods have been established for both a-priori and a-posteriori stopping rules (see Hohage [38] and Kaltenbacher [44]). Therefore, we also consider both an a-priori and an a-posteriori stopping criterion. In the a-priori case we stop the iteration at the first index N for which the condition

$$\eta\sqrt{\gamma_N} f(\gamma_N) < \delta \leq \eta\sqrt{\gamma_n} f(\gamma_n), \qquad 0 \leq n < N, \tag{2.5a}$$

is satisfied. Here η is a sufficiently small constant. As a-posteriori stopping rule we consider again Morozov's discrepancy principle, i.e. we stop the iteration at the first index N, for which the residual $\|F(x_N^\delta) - y^\delta\|$ satisfies

$$\|F(x_N^\delta) - y^\delta\| \leq \tau\delta < \|F(x_n^\delta) - y^\delta\|, \qquad 0 \leq n < N, \tag{2.5b}$$

for a fixed parameter $\tau > 1$. Note that the a-priori stopping criterion (2.5a) in general is not realistic since usually the index function f is unknown.

Before we characterize the IRGNM, let us briefly recall some other examples of regularization methods for solving nonlinear ill-posed problems.

A straightforward generalization of linear Tikhonov regularization leads to the minimization problem

$$\|F(x) - y^\delta\|^2 + \gamma_{Tikh}\|x - x_0\|^2 = \min!$$

over $x \in D(F)$. Here $\gamma_{Tikh} > 0$ denotes the regularization parameter and x_0 is some initial guess for $x^\dagger$. Neubauer [68] has proved convergence of optimal order for nonlinear Tikhonov regularization under a Lipschitz-condition on the Fréchet derivative of F and Hölder source conditions for $\nu \in [1/2, 1]$. Details of this method can be found in [16, chapter 10].

One can also apply Landweber iteration to a nonlinear, ill-posed problem. The iterations are defined by the formula

$$x_{n+1}^\delta := x_n^\delta + \mu F'[x_n^\delta]^* \left(y^\delta - F(x_n^\delta)\right),$$

where μ is a scaling parameter that has to be chosen such that $\|F'[x]\| \leq 1/\mu$ for all x in a neighborhood of $x^\dagger$. In [33] Hanke, Neubauer & Scherzer have proved that Landweber iteration together with the discrepancy principle as stopping rule is an iterative regularization method under certain conditions. It is well known that the convergence of Landweber iteration is very slow. However, Egger & Neubauer [15] have shown that the number of steps Landweber iteration requires to match an appropriate stopping criterion can be significantly reduced by considering the modified Landweber iteration

$$x_{n+1}^\delta := x_n^\delta + L^{-2s} F'[x_n^\delta]^* \left(y^\delta - F(x_n^\delta)\right),$$

where L is a densely defined, unbounded, self-adjoint, and strictly positive operator in $\mathcal{X}$. By choosing $s \leq 0$ the operator L^{-2s} acts as a preconditioner for the smoothing operator $F'[x_n^\delta]^*$ yielding a significant reduction in the number of iterations needed to satisfy the stopping criterion.

To iteratively compute an approximation to $x^\dagger$ it is also popular to apply Newton's method to the nonlinear equation $F(x) = y^\delta$, i.e. to replace the nonlinear operator equation in the n-th step by the linearized equation

$$F'[x_n^\delta]h_n = y^\delta - F(x_n^\delta), \qquad n = 0, 1, 2, \ldots, \tag{2.6}$$

where h_n denotes the update $h_n = x_{n+1}^\delta - x_n^\delta$. Since in general the linearized equation inherits the ill-posedness of the equation $F(x) = y$, some sort of regularization has to be employed. In principle any regularization method for linear ill-posed problems

can be used to compute a stable solution to the linearized equation. Tikhonov regularization with regularization parameter $\gamma_n > 0$ leads to the iteration formula

$$x_{n+1}^\delta = x_n^\delta + \left(\gamma_n I + F'[x_n^\delta]^* F'[x_n^\delta]\right)^{-1} F'[x_n^\delta]^* \left(y^\delta - F(x_n^\delta)\right). \tag{2.7}$$

This method is known as the Levenberg-Marquardt algorithm. For a detailed analysis of this algorithm we refer to Hanke [30].

An alternative approach is to apply an iterative method such as Landweber iteration, ν-method or CGNE directly to (2.6) and use the regularizing properties of such methods with early stopping (see Bakushinskii [6], Kaltenbacher [45], Rieder [74, 76], Hanke [31]). Choosing for example CGNE means that in each Newton step the *normal equation*

$$F'[x_n^\delta]^* F'[x_n^\delta] h_n = F'[x_n^\delta]^* (y^\delta - F(x_n^\delta))$$

is solved using the conjugate gradient method with an appropriate stopping rule. This exploits the well known regularizing properties of the conjugate gradient method (see [16, chapter 7]). This algorithm is called *Newton-CG method*.

However, for linear regularization methods the number of inner iterations typically grows exponentially with the Newton step. For the Newton-CG method no convergence rate results are available so far for weak source conditions (e.g. logarithmic and Hölder with small ν), and experimentally one often observes a slow-down of the convergence after some good progress in the initial phase.

Another class of iterative regularization methods, which we will consider in the rest of this chapter, is given by applying Tikhonov regularization with initial guess $x_0 - x_n^\delta$ to the Newton equation (2.6). This idea leads for $n = 0, 1, 2, \ldots$ to the regularized equations

$$(\gamma_n I + F'[x_n^\delta]^* F'[x_n^\delta]) h_n = F'[x_n^\delta]^* (y^\delta - F(x_n^\delta)) + \gamma_n (x_0 - x_n^\delta), \tag{2.8}$$

where γ_n is the regularization parameter. This algorithm is called iteratively regularized Gauss-Newton method (IRGNM) and was first studied by Bakushinskii [5]. In our case (γ_n) is a fixed sequence satisfying (1.25). As mentioned above, we consider the iteration in combination with the stopping rule (2.5a) or (2.5b). For both stopping criteria we will establish convergence of optimal order under general source conditions.

2.2 Convergence of the IRGNM with a-priori stopping rule

So far the convergence of the IRGNM has been studied under Hölder type source conditions (1.13) (see Bakushinskii [5] and Kaltenbacher, Neubauer & Scherzer

[9]) and logarithmic source conditions (1.14) (see Hohage [37]). Prior to [61] the IRGNM under general source conditions has been investigated by Deuflhard, Engl & Scherzer [14], but no rates of convergence as the noise level δ tends to 0 have been established.

Moreover, up to the present convergence proofs for the iteratively regularized Gauss-Newton method have assumed that the linear equations (2.8) are solved exactly in each Newton step (see [5, 9, 37]). For large scale problems this is unrealistic. One usually computes just an approximation

$$h_n^{\text{app}} \approx (\gamma_n I + F'[x_n^\delta]^* F'[x_n^\delta])^{-1} \left(F'[x_n^\delta]^*(y^\delta - F(x_n^\delta)) + \gamma_n(x_0 - x_n^\delta) \right) \tag{2.9}$$

in each Newton step. It is our goal to formulate conditions under which this additional error does not impair the rate of convergence.

We will show in section 4.2 that the conjugate gradient method applied to (2.8) satisfies the assumptions of our convergence analysis for an appropriate stopping criterion. The CG-method has been shown to be an efficient choice for large scale, exponentially ill-posed problems, especially in combination with a preconditioner (see [40]), but in principle our convergence analysis applies to any iterative method.

To prove convergence and convergence rate results for the IRGNM the index function f and the operator F have to satisfy additional requirements formulated in the following. To this end recall the definition (1.27) of the functions g_n and r_n, $n \in \mathbb{N}_0$. To shorten the notation we make the definitions

$$A := F'[x^\dagger] \qquad \text{and} \qquad A_n := F'[x_n^\delta].$$

As in the linear we assume that the index function f satisfies the additional assumptions (1.28a) and (1.28b) and furthermore, w.l.o.g. that

$$f(\lambda) \le 1 \qquad \text{for all} \qquad \lambda \in (0, \|A\|^2]. \tag{2.10a}$$

Moreover, we assume that the scaling condition $\|A\|^2 \le 1$ is satisfied and that the initial regularization parameter satisfies $\gamma_0 \le \|A\|^2$. For example, for logarithmic source conditions usually the condition $\|A\|^2 \le 1/\exp(-1)$ must be imposed. An application of Corollary 1.12 shows that (1.28a) implies

$$\sup_{0<\lambda\le\|F'[x^\dagger]\|^2} |r_n(\lambda)| f(\lambda) \le c_f f(\gamma_n), \qquad n \in \mathbb{N}_0. \tag{2.10b}$$

We will discuss the classes of index functions defined in (1.13) and (1.14) at the end of this chapter.
As in [9, 37, 44] our analysis relies heavily on a local factorization of the operator F. We assume that for all $\bar{x}, x \in B(x^\dagger, E) := \{x : \|x - x^\dagger\| \le E\}$, $E > 0$, there

exist linear operators $R(\bar{x}, x) \in L(\mathcal{Y}, \mathcal{Y})$ and $Q(\bar{x}, x) \in L(\mathcal{X}, \mathcal{Y})$ such that

$$\begin{aligned} F'[\bar{x}] &= R(\bar{x}, x)F'[x] + Q(\bar{x}, x) && \text{(2.11a)} \\ \|I - R(\bar{x}, x)\| &\leq C_R && \text{(2.11b)} \\ \|Q(\bar{x}, x)\| &\leq C_Q\|F'[x^\dagger](\bar{x} - x)\| && \text{(2.11c)} \end{aligned}$$

for all $\bar{x}, x \in B(x^\dagger, E)$.

By x_{n+1}^δ we denote the computed new iterate and by $x_{n+1}^{\delta,\text{exc}}$ the new iterate for the exact update, i.e.

$$x_{n+1}^\delta = x_n^\delta + h_n^{\text{app}}, \qquad x_{n+1}^{\delta,\text{exc}} = x_n^\delta + h_n. \tag{2.12}$$

Hence, the computed new iterate can be written as

$$x_{n+1}^\delta = x_n^\delta + h_n^{\text{app}} = x_{n+1}^{\delta,\text{exc}} + (h_n^{\text{app}} - h_n).$$

A straightforward computation shows that the total error $e_n := x_n^\delta - x^\dagger$ for the iteratively regularized Gauss-Newton method can be decomposed into

$$\begin{aligned} e_{n+1}^{\text{app}} &:= r_n(A^*A)f(A^*A)w, && \text{(2.13a)} \\ e_{n+1}^{\text{noi}} &:= g_n(A_n^*A_n)A_n^*(y^\delta - y), && \text{(2.13b)} \\ e_{n+1}^{\text{nl}} &:= \left(r_n(A_n^*A_n) - r_n(A^*A)\right) f(A^*A)w, && \text{(2.13c)} \\ e_{n+1}^{\text{tay}} &:= g_n(A_n^*A_n)A_n^*(F(x^\dagger) - F(x_n^\delta) + A_n e_n), && \text{(2.13d)} \\ e_{n+1}^{\text{ls}} &:= h_n^{\text{app}} - h_n. && \text{(2.13e)} \end{aligned}$$

Here e_n^{app} is the linear approximation error, e_n^{noi} is the propagated data noise error, e_n^{tay} involves the Taylor remainder, e_n^{nl} describes the nonlinearity effect that $A_n \neq A$ in general and e_n^{ls} is the error caused by the approximate solution of the linear system.

In the following lemma we prove important estimates for the error components (2.13a)–(2.13d) and for their images under A. These estimates are already implicitly contained in [9] for the special case $f(t) = t^\nu$.

Lemma 2.2 *Assume that (1.28), (2.2), (2.3), (2.4), (2.10) – (2.12) and $\|e_n\| \leq E$ hold. Then the following estimates hold for the error components defined above*

$$\begin{aligned} \|e_{n+1}^{\text{app}}\| &\leq c_f f(\gamma_n)\rho && \text{(2.14a)} \\ \|e_{n+1}^{\text{noi}}\| &\leq \frac{1}{2\sqrt{\gamma_n}}\delta && \text{(2.14b)} \\ \|e_{n+1}^{\text{nl}}\| &\leq C_R\frac{\|Ae_{n+1}^{\text{app}}\|}{\sqrt{\gamma_n}} + \frac{3c_f}{2}C_Q\frac{\|Ae_n\|}{\sqrt{\gamma_n}}f(\gamma_n)\rho && \text{(2.14c)} \\ \|e_{n+1}^{\text{tay}}\| &\leq \frac{1}{2\sqrt{\gamma_n}}\left(2C_R + \frac{3}{2}C_Q\|e_n\|\right)\|Ae_n\| && \text{(2.14d)} \end{aligned}$$

and for their images under A

$$\begin{aligned}
\|Ae_{n+1}^{\mathrm{app}}\| &\leq c_f\sqrt{\gamma_n}f(\gamma_n)\rho && \text{(2.15a)}\\
\|Ae_{n+1}^{\mathrm{noi}}\| &\leq \left(C_R+1+C_Q\frac{\|Ae_n\|}{2\sqrt{\gamma_n}}\right)\delta && \text{(2.15b)}\\
\|Ae_{n+1}^{\mathrm{nl}}\| &\leq \left[2C_R\|Ae_{n+1}^{\mathrm{app}}\|+C_Q\|e_{n+1}^{\mathrm{app}}\|\|Ae_n\|+\frac{\|Ae_{n+1}^{\mathrm{app}}\|}{2\sqrt{\gamma_n}}C_Q\|Ae_n\|\right]\\
&\quad\cdot(C_R+1)+C_Q\|Ae_n\|\|e_{n+1}^{\mathrm{nl}}\| && \text{(2.15c)}\\
\|Ae_{n+1}^{\mathrm{tay}}\| &\leq \left(C_R+1+C_Q\frac{\|Ae_n\|}{2\sqrt{\gamma_n}}\right)\left(2C_R+\frac{3}{2}\|e_n\|C_Q\right)\|Ae_n\|. && \text{(2.15d)}
\end{aligned}$$

Proof: (2.14a) follows from (2.10b) and the isometry of the functional calculus. The proof of (2.15a) is analogous and uses (1.28a) and the identity $\|Az\| = \|(A^*A)^{1/2}z\|$, which holds for all $z \in \mathcal{X}$.
The proofs of the other estimates formulated in the lemma are rather lengthy, but the calculations are straightforward. We will make frequent use of the conditions (2.11) postulated on the nonlinear operator F. To shorten the notation we define the operators $T := (\gamma_n I + A^*A)$ and $T_n := (\gamma_n I + A_n^*A_n)$. Note that the operator T corresponds to the filter function g_n defined above. Now recall the important estimates

$$\begin{aligned}
\|T_n^{-1}A_n^*\| &\leq \frac{1}{2\sqrt{\gamma_n}}, \qquad \|A_nT_n^{-1}\| \leq \frac{1}{2\sqrt{\gamma_n}}, && \text{(2.16)}\\
\|T_n^{-1}\| &\leq \frac{1}{\gamma_n}, \qquad \|A_nT_n^{-1}A_n^*\| \leq 1. && \text{(2.17)}
\end{aligned}$$

Then (2.14b) follows directly from (2.16), and (2.15b) is a consequence of

$$\begin{aligned}
\|Ae_{n+1}^{\mathrm{noi}}\| &\leq \|R(x^\dagger,x_n^\delta)A_nT_n^{-1}A_n^*(y^\delta-y)\|+\|Q(x^\dagger,x_n^\delta)T_n^{-1}A_n^*(y^\delta-y)\|\\
&\leq \left((C_R+1)+C_Q\frac{\|Ae_n\|}{2\sqrt{\gamma_n}}\right)\delta,
\end{aligned}$$

where we have used (2.11). For the following note the important equality

$$e_{n+1}^{\mathrm{app}} = \gamma_nT^{-1}f(A^*A)w.$$

To show (2.14c) we estimate

$$\begin{aligned}
\|e_{n+1}^{\mathrm{nl}}\| &= \|\gamma_n T_n^{-1}(A^*A - A_n^*A_n)T^{-1}f(A^*A)w\| \\
&= \|\gamma_n T_n^{-1}[A_n^*(R(x^\dagger, x_n^\delta)^* - R(x_n^\delta, x^\dagger))A \\
&\quad + Q(x^\dagger, x_n^\delta)^*A - A_n^*Q(x_n^\delta, x^\dagger)]T^{-1}f(A^*A)w\| \\
&\le \|T_n^{-1}A_n^*\|\|R(x^\dagger, x_n^\delta)^* - R(x_n^\delta, x^\dagger)\|\|\gamma_n AT^{-1}f(A^*A)w\| \\
&\quad + \|\gamma_n T_n^{-1}Q(x^\dagger, x_n^\delta)^*AT^{-1}f(A^*A)w\| \\
&\quad + \|\gamma_n T_n^{-1}A_n^*Q(x_n^\delta, x^\dagger)T^{-1}f(A^*A)w\| \\
&\le \frac{1}{2\sqrt{\gamma_n}}2C_R\|Ae_{n+1}^{\mathrm{app}}\| + \frac{1}{\gamma_n}C_Q\|Ae_n\|\|Ae_{n+1}^{\mathrm{app}}\| \\
&\quad + \frac{1}{2\sqrt{\gamma_n}}C_Q\|Ae_n\|\|e_{n+1}^{\mathrm{app}}\| \\
&\le C_R\frac{\|Ae_{n+1}^{\mathrm{app}}\|}{\sqrt{\gamma_n}} + \frac{3c_f}{2}C_Q\frac{\|Ae_n\|}{\sqrt{\gamma_n}}f(\gamma_n)\rho.
\end{aligned}$$

(2.15c) follows from

$$\begin{aligned}
\|Ae_{n+1}^{\mathrm{nl}}\| &= \|\gamma_n AT_n^{-1}(A^*A - A_n^*A_n)T^{-1}f(A^*A)w\| \\
&\le \|\gamma_n R(x^\dagger, x_n^\delta)A_nT_n^{-1}(A^*A - A_n^*A_n)T^{-1}f(A^*A)w\| \\
&\quad + \|\gamma_n Q(x^\dagger, x_n^\delta)T_n^{-1}(A^*A - A_n^*A_n)T^{-1}f(A^*A)w\| \\
&\le \|\gamma_n R(x^\dagger, x_n^\delta)A_nT_n^{-1}[A_n^*(R(x^\dagger, x_n^\delta)^* - R(x_n^\delta, x^\dagger))A \\
&\quad + Q(x^\dagger, x_n^\delta)^*A - A_n^*Q(x_n^\delta, x^\dagger)]T^{-1}f(A^*A)w\| + C_Q\|Ae_n\|\|e_{n+1}^{\mathrm{nl}}\| \\
&\le (C_R + 1)\left[2C_R\|Ae_{n+1}^{\mathrm{app}}\| + \frac{\|Ae_{n+1}^{\mathrm{app}}\|}{2\sqrt{\gamma_n}}C_Q\|Ae_n\| + C_Q\|Ae_n\|\|e_{n+1}^{\mathrm{app}}\|\right] \\
&\quad + C_Q\|Ae_n\|\|e_{n+1}^{\mathrm{nl}}\|.
\end{aligned}$$

To show (2.14d) we estimate

$$\begin{aligned}
\left\|\int_0^1 \left(A_n - F'[x^\dagger + te_n]\right)e_n\,\mathrm{d}t\right\| &= \left\|\int_0^1 \big(R(x_n^\delta, x^\dagger)A + Q(x_n^\delta, x^\dagger)\right. \\
&\quad \left. - R(x^\dagger + te_n, x^\dagger)A - Q(x^\dagger + te_n, x^\dagger)\big)e_n\,\mathrm{d}t\right\| \\
&\le \int_0^1 \left\|\big(R(x_n^\delta, x^\dagger) - R(x^\dagger + te_n, x^\dagger)\big)Ae_n\right\|\,\mathrm{d}t \\
&\quad + \int_0^1 \left\|\big(Q(x_n^\delta, x^\dagger) - Q(x^\dagger + te_n, x^\dagger)\big)e_n\right\|\,\mathrm{d}t \\
&\le \left(2C_R + \frac{3}{2}C_Q\|e_n\|\right)\|Ae_n\|.
\end{aligned}$$

This and (2.16) proves (2.14d).

(2.15d) follows from

$$\begin{aligned}\|Ae_{n+1}^{\mathrm{tay}}\| &= \|AT_n^{-1}A_n^*(F(x^\dagger)-F(x_n^\delta)+A_ne_n)\| \\ &\le \|(R(x^\dagger,x_n^\delta))A_nT_n^{-1}A_n^*(F(x^\dagger)-F(x_n^\delta)+A_ne_n)\| \\ &\quad +\|Q(x^\dagger,x_n^\delta)T_n^{-1}A_n^*(F(x^\dagger)-F(x_n^\delta)+A_ne_n)\| \\ &\le \left(C_R+1+C_Q\frac{\|Ae_n\|}{2\sqrt{\gamma_n}}\right)\left(2C_R+\frac{3}{2}C_Q\|e_n\|\right)\|Ae_n\|.\end{aligned}$$

□

Lemma 2.3 *Assume that (1.28), (2.2), (2.3), (2.4), (2.10) – (2.12), (2.5b) and* $\|e_n\| \le E$ *are satisfied for* $0 \le n < N$ *and for some sufficiently large* $\tau > 1$*. Then the inequalities*

$$\|e_{n+1}\| \le \|e_{n+1}^{\mathrm{ls}}\| + (c_f + C_Rc_f)\rho f(\gamma_n) + \frac{\overline{c}}{C_Q}\frac{\|Ae_n\|}{\sqrt{\gamma_n}}, \tag{2.18}$$

$$\|Ae_{n+1}\| \le \|Ae_{n+1}^{\mathrm{ls}}\| + \overline{a}\|Ae_{n+1}^{\mathrm{app}}\| + \overline{b}\|Ae_n\| + \frac{\overline{c}}{\sqrt{\gamma_n}}\|Ae_n\|^2, \tag{2.19}$$

$$\underline{a}\|Ae_{n+1}^{\mathrm{app}}\| \le \|Ae_{n+1}^{\mathrm{ls}}\| + \|Ae_{n+1}\| + \overline{b}\|Ae_n\| + \frac{\overline{c}}{\sqrt{\gamma_n}}\|Ae_n\|^2 \tag{2.20}$$

with constants

$$\begin{aligned}\overline{a} &:= 1+2C_R(C_R+1), \qquad \underline{a} := 1-2C_R(C_R+1), \\ \overline{b} &:= (C_R+1)\left(\frac{1+C_R+\frac{1}{2}EC_Q}{\tau-1}+\left(2C_R+\frac{3}{2}EC_Q\right)+C_Q\frac{3c_f}{2}\rho\right)+C_QC_Rc_f\rho, \\ \overline{c} &:= C_Q\left(\frac{1+C_R+\frac{1}{2}EC_Q}{2(\tau-1)}+\frac{3c_f}{2}C_Q\rho+\frac{1}{2}\left(2C_R+\frac{3}{2}EC_Q\right)\right)\end{aligned}$$

hold. (Note that $\underline{a} > 0$ *if* C_R *is sufficiently small.) In the case where we replace the discrepancy principle (2.5b) by the a-priori stopping rule (2.5a) the inequalities*

$$\|e_{n+1}\| \le \hat{a}f(\gamma_n) + \hat{b}\frac{\|Ae_n\|}{\sqrt{\gamma_n}}, \tag{2.21}$$

$$\|Ae_{n+1}\| \le \|Ae_{n+1}^{\mathrm{ls}}\| + \tilde{a}\sqrt{\gamma_n}f(\gamma_n) + \tilde{b}\|Ae_n\| + \frac{\tilde{c}}{\sqrt{\gamma_n}}\|Ae_n\|^2 \tag{2.22}$$

are satisfied where the constants $\hat{a}$, $\hat{b}$, $\tilde{a}$, $\tilde{b}$ *and* $\tilde{c}$ *are given by*

$$\begin{aligned}\hat{a} &:= c_f\rho+\frac{1}{2}\eta+C_Rc_f\rho, \qquad \hat{b} := \frac{3c_f}{2}C_Q\rho+\frac{1}{2}\left(2C_R+\frac{3}{2}C_QE\right), \\ \tilde{a} &:= c_f\rho+(C_R+1)(\eta+2C_Rc_f\rho), \\ \tilde{b} &:= \frac{C_Q}{2}\eta+(C_R+1)\left(2C_R+\frac{3}{2}EC_Q+C_Q\frac{3c_f}{2}\rho\right)+C_QC_Rc_f\rho, \\ \tilde{c} &:= C_Q\left(\frac{3c_f}{2}C_Q\rho+C_R+\frac{3}{2}EC_Q\right).\end{aligned}$$

Proof: Notice that from (2.11) and (2.5b) we obtain

$$\begin{aligned}
\tau\delta &\leq \|F(x_n^\delta) - F(x^\dagger) + y - y^\delta\| \\
&\leq \left\| \int_0^1 F'[x^\dagger + te_n]e_n \,\mathrm{d}t \right\| + \delta \\
&= \left\| \int_0^1 \left(R(x^\dagger + te_n, x^\dagger)A + Q(x^\dagger + te_n, x^\dagger) \right) e_n \,\mathrm{d}t \right\| + \delta \\
&\leq \left(C_R + 1 + \frac{1}{2}\|e_n\|C_Q \right) \|Ae_n\| + \delta
\end{aligned}$$

and thus

$$\delta \leq \frac{1}{\tau - 1} \left(C_R + 1 + \frac{1}{2} E C_Q \right) \|Ae_n\|. \tag{2.23}$$

Then the sum of the estimates (2.14) together with (2.23) leads to inequality (2.18), where we have also used (2.10a). Analogously the sum of the estimates (2.15) together with the estimates (2.14) and (2.23) leads to inequality (2.19). By an application of (2.5a) as above the sum of the estimates (2.14) and (2.15) yields (2.21) and (2.22). To show (2.20) we use the equality

$$Ae_{n+1}^{\mathrm{app}} + Ae_{n+1}^{\mathrm{nl}} = Ae_{n+1} - Ae_{n+1}^{\mathrm{noi}} - Ae_{n+1}^{\mathrm{tay}} - Ae_{n+1}^{\mathrm{ls}}.$$

Writing

$$\begin{aligned}
Ae_{n+1}^{\mathrm{nl}} &= \gamma_n R(x^\dagger, x_n^\delta) A_n T_n^{-1} \left[A_n^*(R(x^\dagger, x_n^\delta)^* - R(x_n^\delta, x^\dagger))A \right] T^{-1} f(A^*A)w \\
&\quad + \gamma_n R(x^\dagger, x_n^\delta) A_n T_n^{-1} \left[Q(x^\dagger, x_n^\delta)^* A - A_n^* Q(x_n^\delta, x^\dagger) \right] T^{-1} f(A^*A)w \\
&\quad + Q(x^\dagger, x_n^\delta) e_{n+1}^{\mathrm{nl}}
\end{aligned}$$

we get

$$\begin{aligned}
&Ae_{n+1}^{\mathrm{app}} + \gamma_n R(x^\dagger, x_n^\delta) A_n T_n^{-1} \left[A_n^*(R(x^\dagger, x_n^\delta)^* - R(x_n^\delta, x^\dagger))A \right] T^{-1} f(A^*A)w \\
&\quad = -\gamma_n R(x^\dagger, x_n^\delta) A_n T_n^{-1} \left[Q(x^\dagger, x_n^\delta)^* A - A_n^* Q(x_n^\delta, x^\dagger) \right] T^{-1} f(A^*A)w \\
&\qquad\qquad - Q(x^\dagger, x_n^\delta) e_{n+1}^{\mathrm{nl}} + Ae_{n+1} - Ae_{n+1}^{\mathrm{noi}} - Ae_{n+1}^{\mathrm{tay}} - Ae_{n+1}^{\mathrm{ls}}.
\end{aligned}$$

Now the assertion follows by estimating, and by using the second triangle inequality on the left hand side.

□

With these lemmas we can prove the following convergence result. A similar result has been shown in [14] for the special case $C_{\mathrm{ls}} = 0$ and under a different nonlinearity condition on the operator F.

Proposition 2.4 *Let (1.28), (2.2), (2.3), (2.4), (2.10) and (2.11) hold. Assume that the error e_{n+1}^{ls} and its image $F'[x_n^\delta]e_{n+1}^{\mathrm{ls}}$ satisfy*

$$\begin{aligned} \|e_{n+1}^{\mathrm{ls}}\| &\leq C_{\mathrm{ls}} f(\gamma_n), \qquad 0 \leq n < N, && \text{(2.24a)} \\ \|F'[x_n^\delta]e_{n+1}^{\mathrm{ls}}\| &\leq C_{\mathrm{ls}}\sqrt{\gamma_n} f(\gamma_n), \qquad 0 \leq n < N, && \text{(2.24b)} \end{aligned}$$

and that $C_R, C_Q, C_{\mathrm{ls}}, \gamma, 1/\gamma_0$ and ρ are sufficiently small. Then there exists $E > 0$ such that the inexact Gauss-Newton iterates x_n^δ, $0 \leq n \leq N$, given by (2.12) are well defined for every $x_0 \in D(F)$ satisfying

$$\|x_0 - x^\dagger\| \leq E \tag{2.25}$$

if the stopping index $N = N(\delta, y^\delta)$ is determined by either (2.5a) or (2.5b). Moreover,

$$\|x_n^\delta - x^\dagger\| = O(f(\gamma_n)) \qquad \text{for} \qquad 1 \leq n \leq N \tag{2.26}$$

and for exact data, that is $\delta = 0$, either $N < \infty$ and $x_N^\delta = x^\dagger$ or $N = \infty$ and $\|x_n^\delta - x^\dagger\| \to 0$, $n \to \infty$. Conditions specifying "sufficiently small" are given in the proof.

Proof: We will use an induction argument to prove for θ_n and C_θ^i, $i = 1, 2$, defined by

$$\theta_n := \frac{\|Ae_n\|}{u(\gamma_n)}, \qquad C_\theta^i := \max\left\{\theta_0, \frac{2a_i}{1 - b_i + \sqrt{(1-b_i)^2 - 4a_i c_i}}\right\}$$

with constants

$$\begin{aligned} a_1 &:= \sqrt{\gamma} C_f(\tilde{a} + (C_R + 1)C_{\mathrm{ls}}), \\ b_1 &:= \sqrt{\gamma} C_f(\tilde{b} + C_Q C_{\mathrm{ls}}), \\ c_1 &:= \sqrt{\gamma} C_f \tilde{c}, \\ a_2 &:= \sqrt{\gamma} C_f(c_f \rho \overline{a} + C_{\mathrm{ls}}(C_R + 1)), \\ b_2 &:= \sqrt{\gamma} C_f(\overline{b} + C_Q C_{\mathrm{ls}}), \\ c_2 &:= \sqrt{\gamma} C_f \overline{c} \end{aligned}$$

that for $0 \leq n \leq N$ the estimates

$$\begin{aligned} \theta_n &\leq C_\theta^i, \qquad i = 1, 2, && \text{(2.27a)} \\ \|e_n\| &\leq E. && \text{(2.27b)} \end{aligned}$$

hold true. The case $i = 1$ corresponds to the case where the stopping criterion is determined by (2.5a), the case $i = 2$ where the iteration is stopped by (2.5b).
Notice that (2.27b) implies $x_n^\delta \in B(x^\dagger, E)$. Hence, if (2.27) is true for some $n \in \{0, 1, \ldots, N-1\}$, the estimate (2.19) or (2.22) holds. From (2.24a) and (2.24b) using (2.11) and (2.10a) we get

$$\begin{aligned} \|Ae_{n+1}^{\mathrm{ls}}\| &\leq \|R(x^\dagger, x_n^\delta)F'[x_n^\delta]e_{n+1}^{\mathrm{ls}}\| + \|Q(x^\dagger, x_n^\delta)\|\|e_{n+1}^{\mathrm{ls}}\| \\ &\leq (C_R + 1)C_{\mathrm{ls}}\sqrt{\gamma_n} f(\gamma_n) + C_Q C_{\mathrm{ls}}\|Ae_n\|. && \text{(2.28)} \end{aligned}$$

From (1.25) and (1.28b) the estimates

$$\frac{\sqrt{\gamma_n}}{\sqrt{\gamma_{n+1}}} \le \sqrt{\gamma} \qquad \text{and} \qquad \frac{f(\gamma_n)}{f(\gamma_{n+1})} \le \frac{f(\gamma\gamma_{n+1})}{f(\gamma_{n+1})} \le C_f$$

follow, yielding

$$\begin{aligned} \frac{\sqrt{\gamma_n} f(\gamma_n)}{\sqrt{\gamma_{n+1}} f(\gamma_{n+1})} &\le \sqrt{\gamma} C_f, \\ \frac{\|Ae_n\|}{\sqrt{\gamma_{n+1}} f(\gamma_{n+1})} = \frac{\sqrt{\gamma_n} f(\gamma_n)}{\sqrt{\gamma_{n+1}} f(\gamma_{n+1})} \frac{\|Ae_n\|}{\sqrt{\gamma_n} f(\gamma_n)} &\le \sqrt{\gamma} C_f \theta_n, \\ \frac{\|Ae_n\|^2}{\sqrt{\gamma_n}\sqrt{\gamma_{n+1}} f(\gamma_{n+1})} \le \frac{\sqrt{\gamma_n} f(\gamma_n)}{\sqrt{\gamma_{n+1}} f(\gamma_{n+1})} \frac{\|Ae_n\|^2}{\gamma_n f(\gamma_n)^2} &\le \sqrt{\gamma} C_f \theta_n^2. \end{aligned}$$

Using these estimates together with (2.10a), (2.15a) and (2.28) we derive from (2.19) and (2.22) the recursive estimates

$$\theta_{n+1} \le a_i + b_i \theta_n + c_i \theta_n^2, \qquad i = 1, 2.$$

Let for $i = 1, 2$ t_1^i and t_2^i be the solutions to $a_i + b_i t + c_i t^2 = t$, i.e.

$$t_1^i = \frac{2a_i}{1 - b_i + \sqrt{(1-b_i)^2 - 4a_i c_i}}, \qquad t_2^i = \frac{1 - b_i + \sqrt{(1-b_i)^2 - 4a_i c_i}}{2c_i},$$

and assume that the constants $C_R, C_Q, C_{\mathrm{ls}}, \gamma, 1/\gamma_0$ and ρ are sufficiently small such that the smallness conditions

$$b_i + 2\sqrt{a_i c_i} < 1 \tag{2.29a}$$

$$\theta_0 \le \frac{1 - b_i + \sqrt{(1-b_i)^2 - 4a_i c_i}}{2c_i} \tag{2.29b}$$

hold. Now we can show (2.27a) for both $i = 1$ and $i = 2$. For $n = 0$ (2.27a) is true by the definition of C_θ and (2.27b) by virtue of (2.25). Assume that (2.27) is true for $n = k$, $k < N$. Then the assumptions of Lemma 2.3 are satisfied, and therefore the estimate

$$\theta_{k+1} \le a_i + b_i \theta_k + c \theta_k^2$$

is true. By virtue of assumption (2.29a) we have $t_1^i, t_2^i \in \mathbb{R}$ and $t_1^i < t_2^i$. By the induction hypothesis (2.27a) either $0 \le \theta_k \le t_1^i$ or $t_1^i < \theta_k \le \theta_0$. In the first case, the non-negativity of a_i, b_i, and c_i implies

$$\theta_{k+1} \le a_i + b_i \theta_k + c_i \theta_k^2 \le a + b t_1^i + c_i (t_1^i)^2 = t_1^i,$$

and in the second case we use assumption (2.29b) and the fact that

$$a_i + (b_i - 1)t + c_i t^2 \le 0, \quad t_1^i \le t \le t_2^i,$$

to show that

$$\theta_{k+1} \le a_i + b_i\theta_k + c_i\theta_k^2 \le \theta_k \le \theta_0.$$

Thus, in both cases (2.27a) is true for $n = k+1$ and $i = 1, 2$.

To prove (2.27b) consider first the case $i = 1$. The estimate (2.21) together with (2.27a) yields

$$\|e_{n+1}\| \le \left(\hat{a}f(\gamma_n) + \hat{b}\frac{\|Ae_n\|}{u(\gamma_n)}\right) f(\gamma_n) \le (\hat{a} + \hat{b}C_\theta^1)f(\gamma_n), \qquad 0 \le n < N.$$

Hence, under the additional smallness assumption

$$\hat{a} + \hat{b}C_\theta^1 \le E \tag{2.29c}$$

(2.27b) holds in the case $i = 1$. In the case $i = 2$ we replace the latter smallness assumption (2.29c) by

$$C_{\mathrm{ls}} + (c_f + C_R c_f)\rho + \frac{\overline{c}}{C_Q}C_\theta^2 \le E. \tag{2.29d}$$

Using (2.18), assumptions (2.24a) and (2.29d), and the induction hypothesis we get for $0 \le n < N$ the estimate

$$\begin{aligned} \|e_{n+1}\| &\le (C_{\mathrm{ls}} + (c_f + C_R c_f)\rho)f(\gamma_n) + \frac{\overline{c}}{C_Q}\frac{\|Ae_n\|f(\gamma_n)}{\sqrt{\gamma_n}f(\gamma_n)} \\ &\le Ef(\gamma_n). \end{aligned} \tag{2.30}$$

This proves (2.27b) in the case $i = 2$.

Hence, if the stopping index is determined by (2.5a) or by (2.5b) the iterates x_n^δ, $0 \le n \le N$ are well defined. Furthermore, note that both constants $\hat{a} + \hat{b}C_\theta^1$ and $C_{\mathrm{ls}} + (c_f + C_R c_f)\rho + \overline{c}C_\theta/C_Q$ do not depend on δ or y^δ. This proves (2.26) and shows convergence for the noise-free case $\delta = 0$, since $\gamma_n \searrow 0$, $n \to \infty$, that is either $N < \infty$ and $x_N^\delta = x^\dagger$ or $N = \infty$ and $\|x_n^\delta - x^\dagger\| \to 0$, $n \to \infty$.

□

Proposition 2.4 gives us a possibility to compute an upper bound for the total number of steps until the stopping criterion is satisfied if only noisy data are given.

Corollary 2.5 *Let the assumptions of Proposition 2.4 hold. If $\delta > 0$ and the regularization parameters γ_n are chosen by*

$$\gamma_n = \gamma_0\gamma^{-n}, \qquad n = 0, 1, 2, \ldots,$$

then the stopping index is finite. If the iteration is stopped by either (2.5a) or (2.5b) we have $N = O(-\ln(u^{-1}(\delta)))$, $\delta \to 0$. The function u is given by (1.11).

Proof: Let us first consider the case where we stop the iteration by (2.5b). An application of (2.23) and (2.27) yields

$$\delta \leq \left(\frac{1 + C_R + \frac{E}{2} C_Q}{\tau - 1} \right) \frac{\|Ae_{N-1}\|}{u(\gamma_{N-1})} u(\gamma_{N-1}) \leq \left(\frac{1 + C_R + \frac{E}{2} C_Q}{\tau - 1} \right) C_\theta^2 u(\gamma_{N-1}) \quad (2.31)$$

for $\delta > 0$. Hence, the stopping index is finite with $N = \mathrm{O}(-\ln(u^{-1}(\delta)))$ for the choice $\gamma_n = \gamma_0 \gamma^{-n}$. In the case of (2.5a) we have the estimate

$$\delta \leq \eta \sqrt{\gamma_{N-1}} f \gamma_{N-1} = \eta u(\gamma_0 \gamma^{-N+1}).$$

Thus, $N = O(-\ln(u^{-1}(\delta)))$.

□

We are now in a position to prove convergence rates of optimal order for the IRGNM under general source conditions where the stopping criterion is determined by the a-priori choice (2.5a).

Corollary 2.6 *Let the assumptions of Proposition 2.4 hold and assume that the stopping index of the IRGNM is determined by (2.5a). Then the optimal convergence rate*

$$\|x_N^\delta - x^\dagger\| = O(f(u^{-1}(\delta))), \qquad \delta \to 0,$$

holds true.

Proof: The stopping criterion (2.5a) implies that $\eta u(\gamma_N) < \delta$, that is $\gamma_N = O(u^{-1}(\delta))$, $\delta \to 0$. Inserting this into (2.26) the assertion follows.

□

As mentioned above, usually the index function f is unknown and therefore the a-priori stopping criterion (2.5a) is in practice not realizable. Therefore, it is our goal in the section followed to establish convergence rates of optimal order for the IRGNM in combination with the discrepancy principle.

2.3 Convergence of the IRGNM for the discrepancy principle

Before we dicuss the proof to establish convergence rates of optimal order of the IRGNM in combination with (2.5b) we want to illustrate its main ideas. To this end recall the proof of Theorem 1.13. We want to emphasize the main points of it and dicuss its relation to the nonlinear case.

To prove the convergence rate of the approximation error e_N^{app} we showed that the error behavior of Ae_N^{app} could be expressed in terms of δ. Combining this result with inequality (1.10) yielded the predicted convergence rate. In the nonlinear case we

will proceed analogously. As expected, it will turn out that in this situation it is a technical task to find an expression for Ae_N^{app} in terms of δ. The inequalities (2.19) and (2.20) will play an essential role for this task. Subsequently, it will be proven that the remaining error components (2.14b) – (2.14d) behave like $O(\delta/\sqrt{\gamma_{N-1}})$, which is already clear for e_N^{noi} due to (2.14b). Moreover, the behavior $O(\delta/\sqrt{\gamma_{N-1}})$ already occurred in the proof of Theorem 1.13 and in the nonlinear case it can be shown similar to the linear case that it implies convergence of optimal oder. Finally we are left with e_N^{ls}, the error we commit by not solving the linear systems (2.8) exactly. We will formulate reasonable assumptions for this error incorporating the inequalities (2.24) guaranteeing optimal rates of convergence.
Concluding, we want to point out that the proof of the linear case presented above includes all the important ideas of the nonlinear case, which is inherently more technical.

Since we assume in this section that the index function f is unknown, we replace the error bounds (2.24) on e_{n+1}^{ls} by the strongest possible bounds

$$\begin{aligned} \|e_{n+1}^{\text{ls}}\| &\leq C_{\text{ls}}\sqrt{\gamma_n}, \qquad 0 \leq n < N, && \text{(2.32a)} \\ \|F'[x_n^\delta]e_{n+1}^{\text{ls}}\| &\leq C_{\text{ls}}\gamma_n, \qquad 0 \leq n < N, && \text{(2.32b)} \end{aligned}$$

considering that f satisfies (1.28a) and that the classical qualification order of Tikhonov regularization is 1.

The following main result of this chapter shows that optimal rates of convergence are achieved by the IRGNM where the source condition is given by an index function and the iteration is stopped by the discrepancy principle.

Theorem 2.7 *Let the assumptions of Proposition 2.4 hold and let f satisfy Assumption 1.3. Assume furthermore that the inequalities (2.32) and the smallness conditions*

$$\begin{aligned} q\gamma &< 1 && \text{(2.33a)} \\ \frac{\overline{a}}{1-q\gamma} + q\left(1+\frac{\|A\|^2}{\gamma_0}\right) + \frac{(\|A\|^2+\gamma_0)}{\|Ae_0\|}\frac{(C_R+1)C_{\text{ls}}}{1-\gamma q} &< 2 && \text{(2.33b)} \end{aligned}$$

are satisfied. Here the constant q is defined by $q := C_Q C_{\text{ls}} + \overline{b} + C_Q E$ with the notation of Lemma 2.3. Then the final iterates x_N^δ satisfy the order optimal estimate

$$\|x_N^\delta - x^\dagger\| = O(f(u^{-1}(\delta))), \qquad \delta \to 0, \tag{2.34}$$

if the stopping index N is determined by (2.5b).

Proof: Dealing only with the discrepancy principle (2.5b) we do not need to distinguish two different cases in this proof. Therefore, with the notation of Proposition 2.4 we use throughout this proof the definition $C_\theta := C_\theta^2$.

Due to (2.32b) we get as in (2.28) the estimate

$$\|Ae_{n+1}^{\mathrm{ls}}\| \le (C_R+1)C_{\mathrm{ls}}\gamma_n + C_Q C_{\mathrm{ls}}\|Ae_n\|.$$

Then by (2.19) and (2.27a)

$$\|Ae_{n+1}\| \le (C_R+1)C_{\mathrm{ls}}\gamma_n + \overline{a}\|Ae_{n+1}^{\mathrm{app}}\| + (C_Q C_{\mathrm{ls}} + \overline{b} + \overline{c}C_\theta)\|Ae_n\|$$

holds. Since (2.29d) implies $\overline{c}C_\theta \le C_Q E$ and hence $C_Q C_{\mathrm{ls}} + \overline{b} + \overline{c}C_\theta \le q$, it follows by induction that

$$\|Ae_{n+1}\| \le (C_R+1)C_{\mathrm{ls}}\sum_{k=0}^{n} q^{n-k}\gamma_k + \overline{a}\sum_{k=0}^{n}\|Ae_{k+1}^{\mathrm{app}}\|q^{n-k} + \|Ae_0\|q^{n+1}.$$

The inequality

$$r_k(\lambda) = \left(\frac{\gamma_k}{\gamma_k+\lambda}\right) \le \left(\frac{\gamma_k}{\gamma_{k+1}}\right)\left(\frac{\gamma_{k+1}}{\gamma_{k+1}+\lambda}\right) \le \gamma r_{k+1}(\lambda), \qquad \lambda \ge 0,$$

together with the isometry of the functional calculus implies that

$$\begin{aligned}\|Ae_{k+1}^{\mathrm{app}}\| &= \|Ar_{k+1}(A^*A)f(A^*A)w\| \\ &\le \gamma\|Ar_{k+2}(A^*A)f(A^*A)w\| = \gamma\|Ae_{k+2}^{\mathrm{app}}\|. \end{aligned} \tag{2.35}$$

Analogously the inequality

$$\sqrt{\lambda} \le \left(\frac{\gamma_n + \|A\|^2}{\gamma_n}\right)\left(\frac{\gamma_n}{\gamma_n+\lambda}\right)\sqrt{\lambda} \le \left(1+\frac{\|A\|^2}{\gamma_0}\right)\gamma^n\sqrt{\lambda}r_n(\lambda), \qquad 0 \le \lambda \le \|A\|^2,$$

implies that

$$\|Ae_0\| \le \left(1+\frac{\|A\|^2}{\gamma_0}\right)\gamma^n\|Ae_{n+1}^{\mathrm{app}}\|.$$

Using assumption (2.33a) we obtain

$$\sum_{k=0}^{n} q^{n-k}\|Ae_{k+1}^{\mathrm{app}}\| \le \sum_{k=0}^{n} q^{n-k}\gamma^{n-k}\|Ae_{n+1}^{\mathrm{app}}\| \le \frac{1}{1-q\gamma}\|Ae_{n+1}^{\mathrm{app}}\|.$$

Combining the last inequalities, we have shown that

$$\|Ae_{n+1}\| \le (C_R+1)C_{\mathrm{ls}}\sum_{k=0}^{n} q^{n-k}\gamma_k + \left(\frac{\overline{a}}{1-q\gamma} + q\left(1+\frac{\|A\|^2}{\gamma_0}\right)\right)\|Ae_{n+1}^{\mathrm{app}}\|. \tag{2.36}$$

Now it follows from (2.20), (2.27a), (2.28) and (2.32b), assumption (2.29d) and (2.33a), $q < 1$ (because $\gamma > 1$), $\overline{a} + \underline{a} = 2$ and the last inequality for $n = N - 2$ that

$$\begin{aligned}
\|Ae_N\| &\geq \underline{a}\|Ae_N^{\text{app}}\| - (\overline{b} + C_Q C_{\text{ls}})\|Ae_{N-1}\| - \frac{\overline{c}}{\sqrt{\gamma_{N-1}}}\|Ae_{N-1}\|^2 \\
&\quad -(C_R + 1)C_{\text{ls}}\gamma_{N-1} \\
&\geq \underline{a}\|Ae_N^{\text{app}}\| - (\overline{b} + C_Q C_{\text{ls}} + \overline{c}C_\theta)\|Ae_{N-1}\| - (C_R + 1)C_{\text{ls}}\gamma_{N-1} \\
&\geq \underline{a}\|Ae_N^{\text{app}}\| - q\left(\frac{\overline{a}}{1 - q\gamma} + q\left(1 + \frac{\|A\|^2}{\gamma_0}\right)\right)\gamma\|Ae_N^{\text{app}}\| \\
&\quad -(C_R + 1)C_{\text{ls}}\left(q\sum_{k=0}^{N-2} q^{N-2-k}\gamma_k + \gamma_{N-1}\right) \\
&\geq \left(\underline{a} - \frac{q\gamma\overline{a}}{1 - q\gamma} - q^2\gamma\left(1 + \frac{\|A\|^2}{\gamma_0}\right)\right)\|Ae_N^{\text{app}}\| \\
&\quad -(C_R + 1)C_{\text{ls}}\sum_{k=0}^{N-1} q^{N-1-k}\gamma_k \\
&= \left(2 - \frac{\overline{a}}{1 - q\gamma} - q^2\gamma\left(1 + \frac{\|A\|^2}{\gamma_0}\right)\right)\|Ae_N^{\text{app}}\| \qquad (2.37) \\
&\quad -(C_R + 1)C_{\text{ls}}\sum_{k=0}^{N-1} q^{N-1-k}\gamma_k.
\end{aligned}$$

Furthermore, as above the inequality $\sqrt{\lambda}r_{N-1}(\lambda) \geq \sqrt{\lambda}\left(\frac{\gamma_{N-1}}{\gamma_0 + \|A\|^2}\right)$ for $0 \leq \lambda \leq \|A\|^2$ implies

$$\|Ae_N^{\text{app}}\| \geq \frac{\gamma_{N-1}}{\|A\|^2 + \gamma_0}\|Ae_0\|. \qquad (2.38)$$

From (2.5b) we get

$$\begin{aligned}
\tau\delta &\geq \|F(x_N^\delta) - F(x^\dagger) + y - y^\delta\| \\
&\geq \left\|\int_0^1 F'[x^\dagger + te_N]e_N \,\mathrm{d}t\right\| - \delta \\
&= \left\|\int_0^1 \left(R(x^\dagger + te_N, x^\dagger)A + Q(x^\dagger + te_N, x^\dagger)\right)e_N \,\mathrm{d}t\right\| - \delta \\
&\geq \left(1 - C_R - \frac{1}{2}EC_Q\right)\|Ae_N\| - \delta,
\end{aligned}$$

and thus

$$\delta \geq \frac{1 - C_R - \frac{E}{2}C_Q}{\tau + 1}\|Ae_N\|. \qquad (2.39)$$

It follows from the condition (2.33a), $\gamma > 1$, and the definition of $\overline{b}$ that

$$1 > q = C_Q C_{\text{ls}} + \overline{b} + C_Q E > C_R + \frac{E}{2}C_Q. \qquad (2.40)$$

Using the inequality $\gamma_k/\gamma_{N-1} \le \gamma^{N-1-k}$ we get

$$\gamma_k q^{N-1-k} \le \gamma_{N-1}(\gamma q)^{N-1-k}$$

and thus

$$\sum_{k=0}^{N-1} q^{N-1-k}\gamma_k \le \gamma_{N-1}\sum_{k=0}^{N-1}(\gamma q)^{N-1-k} \le \frac{\gamma_{N-1}}{1-\gamma q}. \tag{2.41}$$

Then we can estimate using (2.37), (2.39), (2.41) and assumption (2.33b)

$$\begin{aligned}\delta \;\ge\; & \frac{1 - C_R - \frac{E}{2}C_Q}{\tau+1}\left[\left(2 - \frac{\overline{a}}{1-q\gamma} - q^2\gamma\left(1 + \frac{\|A\|^2}{\gamma_0}\right)\right)\|Ae_N^{\mathrm{app}}\| \right.\\ & \left. - (C_R+1)C_{\mathrm{ls}}\gamma_{N-1}\frac{1}{1-\gamma q}\right].\end{aligned}$$

This, (2.38) and (2.40) imply

$$C_1\gamma_{N-1} \;\le\; \delta, \tag{2.42}$$

$$C_2\|Ae_N^{\mathrm{app}}\| \;\le\; C_3\gamma_{N-1} + \delta, \tag{2.43}$$

with the constants

$$\begin{aligned}C_1 &:= \left(\frac{1-q}{\tau+1}\right)\left[\left(2 - \frac{\overline{a}}{1-q\gamma} - q\left(1 + \frac{\|A\|^2}{\gamma_0}\right)\right)\frac{\|Ae_0\|}{\|A\|^2+\gamma_0} - \frac{(C_R+1)C_{\mathrm{ls}}}{1-\gamma q}\right],\\ C_2 &:= \left(\frac{1-q}{\tau+1}\right)\left(2 - \frac{\overline{a}}{1-q\gamma} - q^2\gamma\left(1 + \frac{\|A\|^2}{\gamma_0}\right)\right),\\ C_3 &:= \left(\frac{1-q}{\tau+1}\right)\frac{(C_R+1)C_{\mathrm{ls}}}{1-\gamma q},\end{aligned}$$

independent of δ and y^δ. (2.33b) implies $C_1 > 0$ and $C_2 > 0$, and so using (2.42) and (2.43) we conclude

$$\|Ae_N^{\mathrm{app}}\| \le C_4\delta \qquad \text{with} \qquad C_4 := \left(\frac{\frac{C_3}{C_1}+1}{C_2}\right). \tag{2.44}$$

Now we can apply Lemma 1.4 with w replaced by $C_4^{-1}r_{N-1}(A^*A)w$ and (2.44) to obtain

$$\frac{C_4}{\rho}\left\|\left(\frac{\rho}{C_4}\right)e_N^{\mathrm{app}}\right\| \le C_4 f\left(u^{-1}\left(\left(\frac{\rho}{C_4}\right)\left(\frac{\|Ae_N^{\mathrm{app}}\|}{\rho}\right)\right)\right) \le C_4 f\left(u^{-1}(\delta)\right).$$

From (2.31) we have that

$$\delta \le C_5 u(\gamma_{N-1}), \quad \text{where} \quad C_5 := \left(\frac{1 + C_R + \frac{E}{2}C_Q}{\tau - 1}\right)C_\theta.$$

To obtain an estimate for $\sqrt{\gamma_{N-1}}$ in terms of δ we estimate

$$\begin{aligned}\sqrt{\gamma_{N-1}} &= C_5 \frac{\sqrt{\gamma_{N-1}}}{\delta} u\left(u^{-1}\left(\frac{\delta}{C_5}\right)\right) \\ &\leq \frac{C_5}{C_1}\sqrt{\frac{u^{-1}\left(\frac{\delta}{C_5}\right)}{\gamma_{N-1}}} f\left(u^{-1}\left(\max\left\{1, \frac{1}{C_5}\right\}\delta\right)\right) \\ &\leq \frac{C_5}{C_1}\max\left\{1, \frac{1}{C_5}\right\} f\left(u^{-1}(\delta)\right).\end{aligned}$$

In the second line we have used the definition of u, (2.42) and the monotonicity of $f \circ u^{-1}$, and in the last line the inequality $f(u^{-1}(t\delta)) \leq t f(u^{-1}(\delta)$, $t \geq 1$, which follows from concavity of $f \circ u^{-1}$ (see Assumption 1.3). Then from the last estimate and assumption (2.32a) we obtain that

$$\|e_N^{\mathrm{ls}}\| \leq \frac{C_{\mathrm{ls}}}{C_1}\max\{1, C_5\} f\left(u^{-1}(\delta)\right).$$

Now it remains to be shown that the error components in (2.14b) – (2.14d) are of order $O(f(u^{-1}(\delta)))$. To estimate the right hand side of (2.14c) and (2.14d) we combine (2.35), (2.36) again for the case $n = N-2$, (2.41), (2.42) and (2.43) to conclude $\|Ae_{N-1}\| = O(\delta)$. Then an application of $\|e_N\| \leq E$, (2.44) and (2.10a) together with the last result shows that $\|e_N^{\mathrm{nl}}\|$ and $\|e_N^{\mathrm{tay}}\|$ are of order $O\left(\delta/\sqrt{\gamma_{N-1}}\right)$. For $\|e_N^{\mathrm{noi}}\|$ this already follows from (2.14b). Now applying a similar idea as above, we have

$$\frac{\delta}{\sqrt{\gamma_{N-1}}} \leq \max\{1, C_5\} f\left(u^{-1}(\delta)\right).$$

So, altogether we have proven

$$\|e_N\| = \|x_N^\delta - x^\dagger\| = O(f(u^{-1}(\delta))), \qquad \delta \to 0.$$

□

It is worthwhile to note that the convergence theorems of the IRGNM given here comprise the results formulated in [9] and [37], where the additional error term e_n^{ls} was not considered and the theorems were formulated for either Hölder source conditions or logarithmic source conditions. To get the results stated there one has to check that the conditions (1.28) are satisfied for the functions defined in (1.13) and (1.14). The proofs for this can be found in [9] and [37].

Corollary 2.8 *Let the assumptions of Theorem 2.7 be satisfied.*

a) If the source condition is defined via (1.13), then the stopping index of the IRGNM satisfies $N = O(-\ln(\delta^{2/(1+2\nu)}))$ and the optimal convergence rate

$$\|x_N^\delta - x^\dagger\| = O(\delta^{2\nu/(2\nu+1)}), \qquad \delta \to 0,$$

holds true.

b) If the source condition is defined via (1.14), then the stopping index of the IRGNM satisfies $N = O(-\ln(\delta))$ and the optimal convergence rate

$$\|x_N^\delta - x^\dagger\| = O((-\ln(\delta))^{-p}), \qquad \delta \to 0,$$

holds true.

Proof: Consider first the case where the index function is given by (1.13). Then the function u is given by $u(t) = t^{1/2+\nu}$, hence $u^{-1}(t) = t^{2/(1+2\nu)}$. Now using Corollary 2.5 we conclude

$$N = O(-\ln(u^{-1}(\delta))) = O(-\ln(\delta^{2/(1+2\nu)})), \qquad \delta \to 0,$$

and Theorem 2.7 yields

$$\|x_N^\delta - x^\dagger\| = O(f(\delta^{2/(1+2\nu)})) = O(\delta^{2\nu/(1+2\nu)}), \qquad \delta \to 0.$$

In the case where f is given by (1.14) we make use of $f(u^{-1}(t)) = f(t)(1 + o(1))$ if $t \to 0$ (see [59]). Hence,

$$(-\ln(u^{-1}(\delta)))^{-p} = (-\ln(\delta))^{-p}(1 + o(1)), \qquad \delta \to 0.$$

So, by Corollary 2.5

$$N = O\left(\left[(-\ln(u^{-1}(\delta)))^{-p}\right]^p\right) = O\left(-\ln(\delta)\right), \qquad \delta \to 0,$$

and by Theorem 2.7

$$\|x_N^\delta - x^\dagger\| = O((-\ln(u^{-1}(\delta)))^{-p}) = O((-\ln(\delta))^{-p}), \qquad \delta \to 0.$$

□

2.4 Remarks on the nonlinearity conditions

Unfortunately, for many interesting examples the nonlinearity conditions (2.11) could not be proven so far. In particular for the inverse scattering problems considered in Chapter 7 these conditions are an open problem. This is the main reason why the local convergence proof of the IRGNM presented in this chapter is not satisfactory and still open for these examples.

Therefore, in the following we want to take a closer look at the the nonlinearity conditions (2.11). To this end we assume that $x, \bar{x} \in B(x^\dagger, E)$. First note that if $x = \bar{x}$ the conditions (2.11) are obviously satisfied with $R = I$ and $Q = 0$ for any constants C_R and C_Q. Hence, in the following we can assume that $x \neq \bar{x}$ and for

simplicity that the corresponding Fréchet derivatives $F'[\bar{x}]$ and $F'[x]$ are compact operators with singular systems

$$\begin{aligned}\{(\sigma_j; v_j, u_j) : j \in \mathbb{N}\} &\subset (0,\infty) \times \mathcal{X} \times \mathcal{Y},\\ \{(\tilde{\sigma}_j; \tilde{v}_j, \tilde{u}_j) : j \in \mathbb{N}\} &\subset (0,\infty) \times \mathcal{X} \times \mathcal{Y}.\end{aligned}$$

That is, for each $\varphi \in \mathcal{X}$ we have the representations

$$\begin{aligned}F'[\bar{x}]\varphi &= \sum_{j=1}^{\infty} \sigma_j \langle \varphi, v_j \rangle_{\mathcal{X}} u_j,\\ F'[x]\varphi &= \sum_{j=1}^{\infty} \tilde{\sigma}_j \langle \varphi, \tilde{v}_j \rangle_{\mathcal{X}} \tilde{u}_j.\end{aligned}$$

Without loss of generality we can assume that the singular values are in nonincreasing order. To shorten the notation we set $R := R(\bar{x}, x)$ and $Q := Q(\bar{x}, x)$ and define for some threshold integer $m \in \mathbb{N}$ and for all $\psi \in \mathcal{Y}$ and $\varphi \in \mathcal{X}$ the linear operators

$$\begin{aligned}R_1\psi &:= \sum_{j=1}^{\infty} \tilde{\sigma}_j \sum_{i=1}^{m} \frac{1}{\sigma_i} \langle \psi, u_i \rangle_{\mathcal{Y}} \langle v_i, \tilde{v}_j \rangle_{\mathcal{X}} \tilde{u}_j,\\ R_2\psi &:= \sum_{j=m+1}^{\infty} \langle \psi, \tilde{u}_j \rangle_{\mathcal{Y}} \tilde{u}_j,\\ Q_1\varphi &:= \sum_{j=1}^{\infty} \tilde{\sigma}_j \left\langle \sum_{k=m+1}^{\infty} \langle \varphi, v_k \rangle v_k, \tilde{v}_j \right\rangle_{\mathcal{X}} \tilde{u}_j,\\ Q_2\varphi &:= \sum_{j=1}^{\infty} \sigma_j \langle \varphi, v_j \rangle_{\mathcal{X}} \sum_{k=m+1}^{\infty} \langle u_j, \tilde{u}_k \rangle_{\mathcal{Y}} \tilde{u}_k,\end{aligned}$$

and

$$\begin{aligned}R &:= R_1 + R_2, &(2.45)\\ Q &:= Q_1 - Q_2. &(2.46)\end{aligned}$$

We first show that the operators R and Q are constructed such that (2.11a) is satisfied. To this end we compute

$$\begin{aligned}R_2F'[\bar{x}]\varphi &= \sum_{j=1}^{\infty} \sigma_j \langle \varphi, v_j \rangle_{\mathcal{X}} R_2 u_j\\ &= \sum_{j=1}^{\infty} \sigma_j \langle \varphi, v_j \rangle_{\mathcal{X}} \sum_{k=m+1}^{\infty} \langle u_j, \tilde{u}_k \rangle_{\mathcal{Y}} \tilde{u}_k\\ &= Q_2\varphi.\end{aligned}$$

Hence, we have

$$RF'[\bar{x}] + Q = R_1F'[\bar{x}] + R_2F'[\bar{x}] + Q_1 - Q_2 = R_1F'[\bar{x}] + Q_1.$$

Now using the equalities

$$\begin{aligned} R_1u_\ell &= \begin{cases} \sum_{j=1}^{\infty} \frac{\tilde{\sigma}_j}{\sigma_\ell} \langle v_\ell, \tilde{v}_j \rangle_{\mathcal{X}} \tilde{u}_j, & \ell = 1, \dots, m, \\ 0 & i > m, \end{cases} \\ Q_1v_\ell &= \begin{cases} 0, & \ell = 1, \dots, m, \\ \sum_{j=1}^{\infty} \tilde{\sigma}_j \langle v_\ell, \tilde{v}_j \rangle_{\mathcal{X}} \tilde{u}_j & \ell > m, \end{cases} \end{aligned}$$

we have for all $\varphi \in \mathcal{X}$ the equality

$$\begin{aligned} & (R_1F'[\bar{x}] + Q_1)\varphi \\ = \; & R_1 \sum_{\ell=1}^{\infty} \sigma_\ell \langle \varphi, v_\ell \rangle_{\mathcal{X}} u_\ell + Q_1 \sum_{\ell=1}^{\infty} \langle \varphi, v_\ell \rangle_{\mathcal{X}} v_\ell \\ = \; & \sum_{\ell=1}^{m} \sigma_\ell \langle \varphi, v_\ell \rangle_{\mathcal{X}} \sum_{j=1}^{\infty} \frac{\tilde{\sigma}_j}{\sigma_\ell} \langle v_\ell, \tilde{v}_j \rangle_{\mathcal{X}} \tilde{u}_j + \sum_{\ell=m+1}^{\infty} \langle \varphi, v_\ell \rangle_{\mathcal{X}} \sum_{j=1}^{\infty} \tilde{\sigma}_j \langle v_\ell, \tilde{v}_j \rangle_{\mathcal{X}} \tilde{u}_j \\ = \; & \sum_{j=1}^{\infty} \tilde{\sigma}_j \left(\left\langle \sum_{\ell=1}^{m} \langle \varphi, v_\ell \rangle_{\mathcal{X}} v_\ell, \tilde{v}_j \right\rangle_{\mathcal{X}} + \left\langle \sum_{\ell=m+1}^{\infty} \langle \varphi, v_\ell \rangle_{\mathcal{X}} v_\ell, \tilde{v}_j \right\rangle_{\mathcal{X}} \right) \tilde{u}_j \\ = \; & \sum_{j=1}^{\infty} \tilde{\sigma}_j \langle \varphi, \tilde{v}_j \rangle_{\mathcal{X}} \tilde{u}_j \\ = \; & F'[x]\varphi. \end{aligned}$$

Therefore, for the operators R and Q defined through (2.45) and (2.46) condition (2.11a) is satisfied.

For the following discussion we assume that the Fréchet derivatives $F'[x]$ and $F'[\bar{x}]$ belong to the class of so-called Hilbert-Schmidt operators, that is

$$\sum_{j=1}^{\infty} \sigma_j^2 < \infty \qquad \text{and} \qquad \sum_{j=1}^{\infty} \tilde{\sigma}_j^2 < \infty. \tag{2.47}$$

Naturally, condition (2.47) does not include each mildly ill-posed problem. On the other hand (2.47) is satisfied for exponentially ill-posed problems, which have our main interest.

Proposition 2.9 *Assume that (2.47) is satisfied. Then the linear operators R and Q defined through (2.45) and (2.46) are bounded.*

Proof: It follows by a direct computation that

$$\begin{aligned}\|R_1\psi\|^2 &\leq \left(\sum_{j=1}^{\infty}\tilde{\sigma}_j^2\right)\left(\sum_{i=1}^{m}\frac{1}{\sigma_i^2}\right)\|\psi\|^2, \qquad \|R_2\|\leq 1,\\ \|Q_1\varphi\|^2 &\leq \left(\sum_{j=1}^{\infty}\tilde{\sigma}_j^2\right)\|\varphi\|^2, \qquad \|Q_2\|\leq\|F'[\bar{x}]\|.\end{aligned}$$

These estimates show the boundedness of R_i and Q_i, $i=1,2$. Hence, the linear operators R and Q are bounded by their definition (2.45) and (2.46).

□

It is our intention to give some heuristic arguments that the threshold parameter m can possibly be chosen in such a way that the conditions (2.11b) and (2.11c) are satisfied. To this end we proof explicit bounds for $\|I-R\|$ and $\|Q\|$ in the next two theorems.

Theorem 2.10 *Assume that (2.47) is satisfied and that $\psi\in R(F'[x])$. Then we have the estimate*

$$\begin{aligned}\|R\psi-\psi\|_{\mathcal{Y}} \leq &\left\{\left(\sum_{j=1}^{m}\left\|\frac{\tilde{\sigma}_j}{\sigma_j}\langle v_j,\tilde{v}_j\rangle_{\mathcal{X}}\,u_j-\tilde{u}_j\right\|_{\mathcal{Y}}^2\right)^{1/2}\right.\\ &+\frac{\tilde{\sigma}_1}{\sigma_m}\left[\sum_{j=1}^{m}\left(\sum_{i=1,i\neq j}^{m}\left|\langle v_i,\tilde{v}_j\rangle_{\mathcal{X}}\right|\right)^2\right]^{1/2}\\ &\left.+\frac{1}{\sigma_m}\left[\sum_{j=m+1}^{\infty}\tilde{\sigma}_j^2\left(\sum_{i=1}^{m}\left|\langle v_i,\tilde{v}_j\rangle_{\mathcal{X}}\right|\right)^2\right]^{1/2}\right\}\|\psi\|_{\mathcal{Y}}. \qquad (2.48)\end{aligned}$$

Proof: Since $\psi\in R(F'[x])$ we can represent it through $\psi=\sum_{j=1}^{\infty}\langle\psi,\tilde{u}_j\rangle_{\mathcal{Y}}\,\tilde{u}_j$. Now using the definition of R and the triangle inequality we can estimate

$$\begin{aligned}\|R\psi-\psi\|_{\mathcal{Y}} &= \left\|\sum_{j=1}^{\infty}\tilde{\sigma}_j\sum_{i=1}^{m}\frac{1}{\sigma_i}\langle\psi,u_i\rangle_{\mathcal{Y}}\langle v_i,\tilde{v}_j\rangle_{\mathcal{X}}\,\tilde{u}_j-\sum_{j=1}^{m}\langle\psi,\tilde{u}_j\rangle_{\mathcal{Y}}\,\tilde{u}_j\right\|_{\mathcal{Y}}\\ &\leq \left\|\sum_{j=1}^{m}\left\langle\psi,\frac{\tilde{\sigma}_j}{\sigma_j}\langle v_j,\tilde{v}_j\rangle_{\mathcal{X}}\,u_j-\tilde{u}_j\right\rangle_{\mathcal{Y}}\tilde{u}_j\right\|_{\mathcal{Y}}\\ &\quad+\left\|\sum_{j=1}^{m}\tilde{\sigma}_j\sum_{i=1,i\neq j}^{m}\frac{1}{\sigma_i}\langle\psi,u_i\rangle_{\mathcal{Y}}\langle v_i,\tilde{v}_j\rangle_{\mathcal{X}}\,\tilde{u}_j\right\|_{\mathcal{Y}}\\ &\quad+\left\|\sum_{j=m+1}^{\infty}\tilde{\sigma}_j\sum_{i=1}^{m}\frac{1}{\sigma_i}\langle\psi,u_i\rangle_{\mathcal{Y}}\langle v_i,\tilde{v}_j\rangle_{\mathcal{X}}\,\tilde{u}_j\right\|_{\mathcal{Y}}. \qquad (2.49)\end{aligned}$$

The first term on the right hand side of (2.49) can be estimated by

$$\begin{aligned}\left\|\sum_{j=1}^{m}\left\langle\psi,\frac{\tilde\sigma_j}{\sigma_j}\langle v_j,\tilde v_j\rangle_{\mathcal X}u_j-\tilde u_j\right\rangle_{\mathcal Y}\tilde u_j\right\|_{\mathcal Y} &\le \left(\sum_{j=1}^{m}\left|\left\langle\psi,\frac{\tilde\sigma_j}{\sigma_j}\langle v_j,\tilde v_j\rangle_{\mathcal X}u_j-\tilde u_j\right\rangle_{\mathcal Y}\right|^2\right)^{1/2}\\ &\le \left(\sum_{j=1}^{m}\left\|\frac{\tilde\sigma_j}{\sigma_j}\langle v_j,\tilde v_j\rangle_{\mathcal X}u_j-\tilde u_j\right\|_{\mathcal Y}^2\right)^{1/2}\|\psi\|_{\mathcal Y},\end{aligned}$$

and the second term on the right hand side of (2.49) by

$$\begin{aligned}\left\|\sum_{j=1}^{m}\tilde\sigma_j\sum_{i=1,i\neq j}^{m}\frac{1}{\sigma_i}\langle\psi,u_i\rangle_{\mathcal Y}\langle v_i,\tilde v_j\rangle_{\mathcal X}\tilde u_j\right\|_{\mathcal Y} &\le \left[\sum_{j=1}^{m}\left(\sum_{i=1,i\neq j}^{m}\frac{\tilde\sigma_j}{\sigma_i}\langle\psi,u_i\rangle_{\mathcal Y}\langle v_i,\tilde v_j\rangle_{\mathcal X}\right)^2\right]^{1/2}\\ &\le \frac{\tilde\sigma_1}{\sigma_m}\left[\sum_{j=1}^{m}\left(\sum_{i=1,i\neq j}^{m}\left|\langle v_i,\tilde v_j\rangle_{\mathcal X}\right|\right)^2\right]^{1/2}\|\psi\|_{\mathcal Y}.\end{aligned}$$

Finally, we can estimate the last term on the right hand side of (2.49) by

$$\begin{aligned}\left\|\sum_{j=m+1}^{\infty}\tilde\sigma_j\sum_{i=1}^{m}\frac{1}{\sigma_i}\langle\psi,u_i\rangle_{\mathcal Y}\langle v_i,\tilde v_j\rangle_{\mathcal X}\tilde u_j\right\|_{\mathcal Y} &\le \left[\sum_{j=m+1}^{\infty}\left(\sum_{i=1}^{m}\frac{\tilde\sigma_j}{\sigma_i}\langle\psi,u_i\rangle_{\mathcal Y}\langle v_i,\tilde v_j\rangle_{\mathcal X}\right)^2\right]^{1/2}\\ &\le \frac{1}{\sigma_m}\left[\sum_{j=m+1}^{\infty}\tilde\sigma_j^2\left(\sum_{i=1}^{m}\left|\langle v_i,\tilde v_j\rangle_{\mathcal X}\right|\right)^2\right]^{1/2}\|\psi\|_{\mathcal Y}.\end{aligned}$$

Hence, altogether we have proven (2.48).

□

Theorem 2.11 *Assume that (2.47) is satisfied. Then we have the estimate*

$$\begin{aligned}\|Q\varphi\|_{\mathcal Y} \le &\left[\left\{\sum_{j=1}^{m}\tilde\sigma_j^2\left(\sum_{k=m+1}^{\infty}|\langle v_k,\tilde v_j\rangle_{\mathcal X}|^2\right)+\sum_{j=m+1}^{\infty}\tilde\sigma_j^2\right\}^{1/2}\right.\\ &\left.\left\{\sum_{j=1}^{m}\sigma_j^2\left(\sum_{k=m+1}^{\infty}\langle u_j,\tilde u_k\rangle_{\mathcal Y}\right)^2+\sum_{j=m+1}^{\infty}\sigma_j^2\right\}^{1/2}\right]\|\varphi\|_{\mathcal X}. \qquad (2.50)\end{aligned}$$

Proof: Obviously, by the definition (2.46) of Q for all $\varphi\in\mathcal X$ the inequality

$$\|Q\varphi\|_{\mathcal Y}\le\|Q_1\varphi\|_{\mathcal Y}+\|Q_2\varphi\|_{\mathcal Y}.$$

holds. Then we can estimate

$$\begin{aligned}\|Q_1\varphi\|_{\mathcal{Y}}^2 &= \left\|\sum_{j=1}^{\infty}\tilde{\sigma}_j\left\langle\sum_{k=m+1}^{\infty}\langle\varphi,v_k\rangle_{\mathcal{X}}\,v_k,\tilde{v}_j\right\rangle_{\mathcal{X}}\tilde{u}_j\right\|_{\mathcal{Y}}^2\\ &\le \sum_{j=1}^{m}\tilde{\sigma}_j^2\left(\sum_{k=m+1}^{\infty}\langle\varphi,v_k\rangle_{\mathcal{X}}\langle v_k,\tilde{v}_j\rangle_{\mathcal{X}}\right)^2\\ &\quad+\sum_{j=m+1}^{\infty}\tilde{\sigma}_j^2\left(\sum_{k=m+1}^{\infty}\langle\varphi,v_k\rangle_{\mathcal{X}}\langle v_k,\tilde{v}_j\rangle_{\mathcal{X}}\right)^2. \qquad (2.51)\end{aligned}$$

The first term on the right hand side of (2.51) can be estimated by

$$\sum_{j=1}^{m}\tilde{\sigma}_j^2\left(\sum_{k=m+1}^{\infty}\langle\varphi,v_k\rangle_{\mathcal{X}}\langle v_k,\tilde{v}_j\rangle_{\mathcal{X}}\right)^2 \le \sum_{j=1}^{m}\tilde{\sigma}_j^2\left(\sum_{k=m+1}^{\infty}|\langle v_k,\tilde{v}_j\rangle|_{\mathcal{X}}^2\right)\|\varphi\|_{\mathcal{X}}^2, \quad (2.52)$$

and the second term by

$$\sum_{j=m+1}^{\infty}\tilde{\sigma}_j^2\left(\sum_{k=m+1}^{\infty}\langle\varphi,v_k\rangle_{\mathcal{X}}\langle v_k,\tilde{v}_j\rangle_{\mathcal{X}}\right)^2 \le \left(\sum_{j=m+1}^{\infty}\tilde{\sigma}_j^2\right)\|\varphi\|_{\mathcal{X}}^2. \qquad (2.53)$$

Similar, we can estimate $\|Q_2\varphi\|$ by

$$\begin{aligned}\|Q_2\varphi\|_{\mathcal{Y}}^2 &\le \sum_{j=1}^{m}\sigma_j^2|\langle\varphi,v_j\rangle_{\mathcal{X}}|^2\left(\sum_{k=m+1}^{\infty}\langle u_j,\tilde{u}_k\rangle_{\mathcal{Y}}\right)^2\\ &\quad+\sum_{j=m+1}^{\infty}\sigma_j^2|\langle\varphi,v_j\rangle_{\mathcal{X}}|^2\left(\sum_{k=m+1}^{\infty}\langle u_j,\tilde{u}_k\rangle_{\mathcal{Y}}\right)^2\\ &\le \left\{\sum_{j=1}^{m}\sigma_j^2\left(\sum_{k=m+1}^{\infty}\langle u_j,\tilde{u}_k\rangle_{\mathcal{Y}}\right)^2+\sum_{j=m+1}^{\infty}\sigma_j^2\right\}\|\varphi\|_{\mathcal{X}}^2. \qquad (2.54)\end{aligned}$$

The sum of the estimates (2.52), (2.53) and (2.54) yields (2.50).

□

In the following we want to establish a connection between the conditions (2.11b) and (2.11c) and the inequalities (2.48) and (2.50). To this end we give heuristic arguments that if $x, \bar{x}$ lie in a small neighborhood of $x^\dagger$ possibly $\|R-I\|$ and $\|Q\|$ are so small such that (2.11b) and (2.11c) could be satisfied.
By inequality (2.48) to ensure that $\|R-I\|$ is small it is required that the threshold

parameter m can be chosen such that the terms

$$\left\| \frac{\tilde{\sigma}_j}{\sigma_j} \langle v_j, \tilde{v}_j \rangle \, u_j - \tilde{u}_j \right\|_{\mathcal{Y}}, \qquad j = 1, \dots, m, \tag{2.55a}$$

$$\left| \langle v_i, \tilde{v}_j \rangle_{\mathcal{X}} \right|, \qquad i, j = 1, \dots, m, \quad i \neq j, \tag{2.55b}$$

$$| \langle v_i, \tilde{v}_j \rangle_{\mathcal{X}} |, \qquad i = 1, \dots, m, \quad j = m+1, m+2, \dots, \tag{2.55c}$$

$$\sum_{j=m+1}^{\infty} \tilde{\sigma}_j^2 \tag{2.55d}$$

are sufficiently small. But this in fact could be satisfied since the singular values and singular vectors of a linear operator depend continuously on small perturbations. Hence, if the distance $\|x - \bar{x}\|$ is sufficiently small and we interpret $F'[\bar{x}]$ as a small perturbation of $F'[x]$, then by continuity arguments the threshold parameter m can possibly be chosen such that the terms (2.55a) – (2.55d) are sufficiently small such that (2.11b) holds.
To argue that $\|Q\|$ is sufficiently small such that (2.11c) is possibly satisfied using inequality (2.50) we need to take care about the terms

$$| \langle v_k, \tilde{v}_j \rangle_{\mathcal{X}} |, \qquad j = 1, \dots, m, \quad k = m+1, m+2, \dots, \tag{2.56a}$$

$$\langle u_j, \tilde{u}_k \rangle_{\mathcal{Y}}, \qquad j = 1, \dots, m, \quad k = m+1, m+2, \dots, \tag{2.56b}$$

$$\sum_{j=m+1}^{\infty} \sigma_j^2 \tag{2.56c}$$

and (2.55d). Using again continuity arguments the threshold parameter m can possibly be chosen such that the terms (2.56c) – (2.56a) are sufficiently small such that (2.11c) holds.

Note that our discussion does not prove the nonlinearity conditions (2.11), since for some given nonlinear ill-posed problem it is open if in general the threshold integer m can be chosen such that the terms (2.55a) – (2.55d) and (2.56a) – (2.56c) can be simultaneously sufficiently small for $x, \bar{x}$ in a sufficiently small neighborhood of $x^\dagger$. Furthermore, naturally the choice of m depends on x and $\bar{x}$.
However, our discussion sheds some light on the nonlinearity conditions from the direction where the linear operators are represented by their singular value decomposition. It possibly serves as a further step to either prove these conditions in a general way or it maybe gives some hints to disprove these conditions for a certain given problem.

Chapter 3

Conjugate gradient and Lanczos' method

The main complexity of the IRGNM consists in solving the linearized and regularized linear systems (2.8). In applications the operators F, $F'[x]$ and $F'[x]^*$ occuring in these equations usually represent operators corresponding to some differential or integral equation. For small-scale problems it has often been pointed out that setting up the matrix representing the operators $F'[x]$ and $F'[x]^*$ is an appropriate way of realizing the IRGNM. Throughout this work we are concerned with the development of an efficient numerical solver for large-scale problems, where the computation of the system matrix is inefficient for several reasons already discussed in the introduction. Besides the numerical aspects our aim is to establish a complexity analysis of the IRGNM where the linear systems are solved by the conjugate gradient method. This analysis shall include certain types of preconditioners.
The goal of this chapter is to summarize the known foundations for an efficient implementation of the IRGNM, that is an efficient numerical solver for the linear systems (2.8). Moreover, the tools designed in this chapter are fundamental for an analysis of the complexity and an efficient realization of the IRGNM discussed in the following chapters.

3.1 Introduction and notation

The conjugate gradient method (CG-method) has become the most widespread way of solving symmetric positive definite linear algebraic systems since it was first presented by Hestenes and Stiefel [35]. Since the operator $\gamma_n I + F'[x_n^\delta]^* F'[x_n^\delta]$ is bounded self-adjoint and strictly coercive with respect to the inner product $\langle \cdot\,,\,\cdot \rangle_{\mathcal{X}}$, the CG-method is a natural choice to solve the linear systems (2.8). Moreover, it is well known that efficient preconditioners can be constructed to speed up the convergence of the CG-method. As in the last chapters, $(\mathcal{X}, \langle \cdot\,,\,\cdot \rangle_{\mathcal{X}})$ and $\big(\mathcal{Y}, \langle \cdot\,,\,\cdot \rangle_{\mathcal{Y}}\big)$ denote real Hilbert spaces with inner products $\langle \cdot\,,\,\cdot \rangle_{\mathcal{X}}$ and $\langle \cdot\,,\,\cdot \rangle_{\mathcal{Y}}$.

About fifty years ago superlinear convergence of the CG-method was already proved by Hayes [34] (see also Winther [85]). Closely related to the CG-method is Lanczos' method originally introduced in [55]. This method deals with the problem of approximating extremal eigenvalues and corresponding eigenvectors. It can be interpreted as a generalization of the Rayleigh quotient iteration. The quality of the approximations strongly depends on the eigenvalue distribution of the linear operator. Both methods will be introduced and explained in this chapter.

Although the CG-method as well as Lanczos' method are described in nearly every modern textbook on numerical linear algebra it turns out to be necessary to introduce both methods once again in this work. The reasons are the following:

- Usually the CG-method described in textbooks is restricted to finite dimensional linear systems with respect to the Euclidean inner product in $\mathbb{R}^n$. Both assumptions are too restrictive for our goals.
- To our purposes it is sufficient to have a short and precise description of the connection of the CG-method and Lanczos' method.
- The latter point needs to include the case when the CG-method is preconditioned.

Hence, although there are many textbooks and articles concerning with the CG-method and Lanczos' method, no description was suited for the problem at hand. The main goal of this chapter is to introduce these algorithms in a way suitable to the rest of this work.

Our introduction of the CG-method and Lanczos' method is based on the textbooks of Axelsson [2], Demmel [12], Golub & van Loan [24], Saad [77], van der Vorst [84], Engl, Hanke & Neubauer [16], and Rieder [75].

To shorten the notation we define the operator

$$G_n := \begin{pmatrix} F'[x_n^\delta] \\ \sqrt{\gamma_n} I \end{pmatrix} \in L(\mathcal{X}, \mathcal{Y} \times \mathcal{X}), \qquad x_n^\delta \in D(F), \tag{3.1}$$

and the right hand side vector $g_n^\delta := \left(y^\delta - F(x_n^\delta), \sqrt{\gamma_n} b_n\right)^T$. The choice $b_n = x_0 - x_n^\delta$ corresponds to the IRGNM and $b_n = 0$ to the Levenberg-Marquardt algorithm.
Furthermore, we assume that the linear and bounded operator $F'[x_n^\delta]^* F'[x_n^\delta] : \mathcal{X} \to \mathcal{X}$, which is self-adjoint with respect to $\langle \cdot\,, \cdot \rangle_{\mathcal{X}}$ is compact. The nonnegative eigenvalues λ_j, $j \in \mathbb{N}$, are enumerated in nonincreasing order with multiplicity, and the corresponding orthonormal eigenvectors are denoted by φ_j.
Then, by virtue of (2.7) resp. (2.8) we need to solve in each step of these algorithms the normal equation

$$G_n^* G_n h_n = G_n^* g_n^\delta, \qquad n = 0, 1, 2, \ldots \tag{3.2}$$

with the bounded and strictly coercive operator $G_n^* G_n : \mathcal{X} \to \mathcal{X}$. By the Theorem of Lax-Milgram (see Kress [52, Theorem 13.26]) the operator $G_n^* G_n$ has a bounded inverse with

$$\|(G_n^* G_n)^{-1}\| \leq 1/\gamma_n. \tag{3.3}$$

Still, for a derivation of the CG-method the exact form of the linear system we are concerned with is not important. To formulate our results in a more general setting, we skip the index and replace the operator $G_n^* G_n$ by a bounded and strictly coercive operator $G : \mathcal{X} \to \mathcal{X}$ which is self-adjoint with respect to some other inner product $(\,\cdot\,,\,\cdot\,)$ on $\mathcal{X}$, i.e. we consider the linear operator equation

$$Gh = g \tag{3.4}$$

with right hand side $g \in \mathcal{X}$. The uniquely determined solution of (3.4) is denoted by $h^\dagger$.

3.2 The standard conjugate gradient method

The standard form of the CG-method can be seen as a method to find the minimizer of the functional

$$J(h) = \frac{1}{2}(r, G^{-1}r), \tag{3.5}$$

where the residual r is defined by $r := g - Gh$. Obviously, the minimizer of J is the solution $h^\dagger = G^{-1}g$, since G^{-1} is also bounded strictly coercive and self-adjoint with respect to the inner product $(\,\cdot\,,\,\cdot\,)$. To find the minimizer of J, the CG-method starts with an initial guess h^0 to the minimizer $h^\dagger$ and constructs at each stage a new search direction p^k, which will be *conjugately orthogonal* to the previous search directions, that is:

Definition 3.1 *Two vectors $p, q \in \mathcal{X} \setminus \{0\}$ are said to be* conjugately orthogonal *(with respect to G), if $(p, Gq) = 0$.*

Then the method computes the local minimizer along this search direction p^k, i.e. given h^{k-1}, the approximation to the solution $h^\dagger$ at stage $k-1$, we compute α_k such that

$$J(h^{k-1} + \alpha_k p^k), \qquad -\infty < \alpha_k < \infty,$$

is minimized by α_k and then let

$$h^k := h^{k-1} + \alpha_k p^k \tag{3.6}$$

be the new approximation.

Let $r^k := g - Gh^k$. To compute α_k consider the function

$$m(\alpha) := J(h^{k-1} + \alpha p^k) - J(h^{k-1}).$$

Rewriting m as

$$\begin{aligned} m(\alpha) &= \frac{1}{2}(g - G(h^{k-1} + \alpha p^k), G^{-1}(g - G(h^{k-1} + \alpha p^k))) \\ &\quad -\frac{1}{2}(g - Gh^{k-1}, G^{-1}(g - Gh^{k-1})) \\ &= \frac{1}{2}(r^{k-1} - \alpha G p^k, h^\dagger - h^{k-1} - \alpha p^k) - \frac{1}{2}(r^{k-1}, h^\dagger - h^{k-1}) \\ &= \frac{1}{2}\alpha^2(p^k, Gp^k) - \alpha(r^{k-1}, p^k), \end{aligned}$$

we see that it is quadratic in α and takes its smallest value when

$$(r^{k-1}, p^k) - \alpha(p^k, Gp^k) = 0,$$

i.e. in each stage of the iteration we determine

$$\alpha_k = \frac{(r^{k-1}, p^k)}{(p^k, Gp^k)} \tag{3.7}$$

provided that $p^k \neq 0$. The equality

$$r^k = g - Gh^k = g - Gh^{k-1} - \alpha_k G p^k = r^{k-1} - \alpha_k G p^k \tag{3.8}$$

yields

$$(r^k, p^k) = (r^{k-1}, p^k) - \alpha_k(p^k, Gp^k) = 0, \tag{3.9}$$

that is the search direction is orthogonal to the new residual. Now, for the next iteration step a new search direction is required, which will be computed under the side condition

$$(p^{k+1}, Gp^j) = 0, \qquad 1 \leq j \leq k, \tag{3.10}$$

that is the search directions become mutually conjugate orthogonal with respect to G and the inner product $(\,\cdot\,,\,\cdot\,)$. This still holds for many sets of search directions. So we restrict ourselves to the following requirement.

Let the vectors p^j, $1 \leq j \leq k$, satisfy (3.10) and assume furthermore that

$$(r^k, p^j) = 0, \qquad 1 \leq j \leq k, \tag{3.11}$$

where $k \geq 1$. Now, by (3.8)

$$(r^{k+1}, p^j) = (r^k, p^j) - \alpha_{k+1}(Gp^{k+1}, p^j),$$

and so using (3.10) and (3.11) together with $(r^{k+1}, p^{k+1}) = 0$ we see that

$$(r^{k+1}, p^j) = 0, \qquad 1 \leq j \leq k+1. \tag{3.12}$$

Hence, by induction, it follows that when the search directions are conjugately orthogonal, the residuals become orthogonal to the previous search directions. As we shall see, this property implies that the method computes the best approximation $h^k = h^{k-1} + p$ of all vectors $p \in \text{span}\{p^1, \ldots, p^k\}$. It remains, therefore, to compute the search directions in an efficient way to make them mutually conjugately orthogonal with respect to G. To this end, let

$$p^{k+1} = r^k + \beta_k p^k, \qquad k = 1, 2, \ldots, \tag{3.13}$$

where initially $p^1 = r^0$ and $\beta_1, \beta_2, \ldots$ need to be determined. The relation (3.10) directly yields

$$\beta_k = -\frac{(r^k, Gp^k)}{(p^k, Gp^k)}. \tag{3.14}$$

Now it is left to show that p^{k+1} given by (3.13) satisfies (3.10). We will prove this property and summarize many other important features of the CG-algorithm defined by (3.6), (3.7), (3.8), (3.13) and (3.14) in the following theorem.

Theorem 3.2 *Let the linear operator $G : \mathcal{X} \to \mathcal{X}$ be bounded self-adjoint and strictly coercive with respect to the inner product $(\cdot\,,\cdot)$. Let $h^0 \in \mathcal{X}$ be an arbitrary vector, $r^0 = g - Gh^0$, $p^1 = r^0$ and assume that $r^k \neq 0$ for all $k = 0, \ldots, m$. Then the iterates*

$$\begin{aligned} h^k &= h^{k-1} + \alpha_k p^k, \\ r^k &= r^{k-1} - \alpha_k G p^k \\ p^{k+1} &= r^k + \beta_k p^k, \end{aligned}$$

where α_k is computed by (3.7) and β_k by (3.14), are well defined for all $k = 1, \ldots, m+1$ and the following assertions hold true:

a) The following orthogonality properties hold for $m \geq 1$:

$$(r^m, p^j) = 0, \qquad 1 \leq j \leq m, \tag{3.15a}$$

$$(r^m, r^j) = 0, \qquad 0 \leq j \leq m-1, \tag{3.15b}$$

$$(p^{m+1}, Gp^j) = 0, \qquad 1 \leq j \leq m. \tag{3.15c}$$

b) For all $k = 1, \ldots, m$ we have:

$$\text{span}\{r^0, \ldots, r^{k-1}\} = \text{span}\{p^1, \ldots, p^k\} = \text{span}\{r^0, Gr^0, \ldots, G^{k-1}r^0\}. \tag{3.16}$$

c)

$$h^k \in h^0 + \mathcal{K}_k(G, r^0), \qquad k = 1, \ldots, m$$

where $\mathcal{K}_k(G, r^0)$ denotes the Krylov subspace

$$\mathcal{K}_k(G, r^0) := \text{span}\{r^0, Gr^0, \ldots, G^{k-1}r^0\}.$$

d)

$$\inf_{u\in\mathcal{S}_k(G,r^0)}(r^0+u, G^{-1}(r^0+u)) = (r^k, G^{-1}r^k),$$

where $\mathcal{S}_k(G,r^0) := \mathrm{span}\{Gr^0, G^2r^0, \ldots, G^kr^0\}$.

e) If the inner product is given by $(u,v) := \langle u, v\rangle_{\mathcal{X}}$*, then the conjugate gradient method satisfies*

$$\left\langle e^k, Ge^k\right\rangle_{\mathcal{X}} = \inf_{v\in\mathcal{K}_k(G,r^0)}\left\langle e^0+v, G(e^0+v)\right\rangle_{\mathcal{X}}.$$

Here $e^k := h^\dagger - h^k$ *denotes the iteration error.*

f) If the inner product is given by $(u,v) := \langle u, Gv\rangle_{\mathcal{X}}$*, then the conjugate gradient method gives the best least square residual solution, that is*

$$\|r^k\|^2_{\mathcal{X}} = \inf_{v\in\mathcal{K}_k(G,r^0)}\left\langle r^0+Gv, r^0+Gv\right\rangle_{\mathcal{X}}.$$

Proof: To prove that the iterates are well defined for $k = 1,\ldots,m+1$ assume that $p^{k+1} = 0$ for some $k \le m$, and let k be the smallest number such that this condition is satisfied. By (3.13) we have

$$0 = (r^k, p^k) = -\beta_k(p^k, p^k).$$

By the choice of k, $\beta_k = 0$, implying $r^k = 0$ due to (3.13), which contradicts the assumption $r^k \neq 0$, $k = 0,\ldots,m$.

Assertion a) and b) are proven by induction. So, let $m = 1$. Then (3.15a) is a consequence of (3.9) and (3.15b) follows from the definition $p^1 = r^0$. Using the definition of β_k and (3.13) we conclude

$$(p^{k+1}, Gp^k) \;=\; (r^k, Gp^k) + \beta_k(p^k, Gp^k) = 0, \qquad k = 1,\ldots,m. \tag{3.17}$$

Let the orthogonality relations now be satisfied for some $n \in \{1,\ldots,m\}$. The equality

$$(r^{n+1}, p^j) = (r^n, p^j) - \alpha_k(p^{n+1}, Gp^j) = 0, \qquad 1 \le j \le n,$$

together with (3.9) proves (3.15a). (3.15b) follows from

$$(r^{n+1}, r^j) = (r^{n+1}, p^{j+1}) - \beta_j(r^{n+1}, p^j) = 0, \qquad 1 \le j \le n,$$

and $(r^{n+1}, r^0) = (r^{n+1}, p^1) = 0$. To show (3.15c) we combine (3.17) and

$$(p^{n+2}, Gp^j) = (r^{n+1}, Gp^j) - \beta_{n+1}(p^{n+1}, Gp^j) = \frac{1}{\alpha_j}(r^{n+1}, r^{j-1} - r^j) = 0, \quad 1 \le j \le n,$$

which is true by the induction assumption.

Assertion b) is clear for $k = 1$. Now using the induction assumption and (3.13) we have $p^k \in \text{span}\{r^0, \ldots, r^{k-1}\}$. Then the linear independence of the vectors $p^1, \ldots, p^k$ and $r^0, \ldots, r^{k-1}$, which follows from (3.15b) resp. (3.15c), yields

$$\text{span}\{r^0, \ldots, r^{k-1}\} = \text{span}\{p^1, \ldots, p^k\}, \qquad k = 1, \ldots m.$$

To show the other equality again by the induction assumption it follows that

$$p^{k-1} \in \text{span}\{r^0, \ldots, r^{k-2}\} = \text{span}\{r^0, Gr^0, \ldots, G^{k-2}r^0\}$$

and so by (3.8) $r^{k-1} = r^{k-2} - \alpha_{k-1}Gp^{k-1} \in \text{span}\{r^0, Gr^0, \ldots, G^{k-1}r^0\}$, that is

$$\text{span}\{r^0, \ldots, r^{k-1}\} \subset \{r^0, Gr^0, \ldots, G^{k-1}r^0\}.$$

Again by the linear independence of $r^0, \ldots, r^{k-1}$ we obtain the latter equality in (3.16).

Assertion c) follows by induction from the equality

$$h^k = h^{k-1} + \alpha_k p^k = h^{k-2} + \alpha_{k-1}p^{k-1} + \alpha_k p^k = \ldots = h^0 + \sum_{j=1}^{k} \alpha_j p^j.$$

together with (3.16).

To prove d) note that (3.15b) together with (3.16) yields

$$(r^k, u) = 0 \qquad \text{for all} \qquad u \in \text{span}\{r^0, Gr^0, \ldots, G^{k-1}r^0\}.$$

Hence, we see that $(r^k, G^{-1}Gr^j) = (r^k, G^{-1}u) = 0$ for all $u \in \mathcal{S}_k(G, r^0)$. Letting $u^k := r^k - r^0$, this can be written in the form

$$(r^0 + u^k, G^{-1}u) = 0 \qquad \text{for all} \qquad u \in \mathcal{S}_k(G, r^0).$$

This orthogonality property shows that $(r^0 + u, G^{-1}(r^0 + u))$ is smallest among all $u \in \mathcal{S}_k(G, r^0)$, if and only if $u = u^k$. To see this, note that for any other $u = u^k + \tilde{u}$ with $\tilde{u} \in \mathcal{S}_k$, we have

$$\begin{aligned} (r^0 + u, G^{-1}(r^0 + u)) &= (r^0 + u^k + \tilde{u}, G^{-1}(r^0 + u^k + \tilde{u})) \\ &= (r^0 + u^k, G^{-1}(r^0 + u^k)) + (\tilde{u}, G^{-1}\tilde{u}) \\ &\geq (r^0 + u^k, G^{-1}(r^0 + u^k)). \end{aligned}$$

That is, $r^k = r^0 + u^k$ is the minimizer of the functional $J(h) = (r, G^{-1}r)$ on the subspace $\mathcal{S}_k(G, r^0)$.

Assertion e) follows from the computation

$$\begin{aligned}
\left\langle r^k, G^{-1}r^k\right\rangle_{\mathcal{X}} &= \inf_{u\in\mathcal{S}_k(G,r^0)} \left\langle r^0+u, G^{-1}(r^0+u)\right\rangle_{\mathcal{X}} \\
&= \inf_{u\in\mathcal{S}_k(G,r^0)} \left\langle g-Gh^0+u, h^\dagger-h^0+G^{-1}u\right\rangle_{\mathcal{X}} \\
&= \inf_{u\in\mathcal{S}_k(G,r^0)} \left\langle g-Gh^0+u, G^{-1}G(h^\dagger-h^0+G^{-1}u)\right\rangle_{\mathcal{X}} \\
&= \inf_{u\in\mathcal{S}_k(G,r^0)} \left\langle h^\dagger-h^0+G^{-1}u, G(h^\dagger-h^0+G^{-1}u)\right\rangle_{\mathcal{X}} \\
&= \inf_{v\in\mathcal{K}_k(G,r^0)} \left\langle h^\dagger-h^0+v, G(h^\dagger-h^0+v)\right\rangle_{\mathcal{X}} \\
&= \inf_{v\in\mathcal{K}_k(G,r^0)} \left\langle e^0+v, G(e^0+v)\right\rangle_{\mathcal{X}}
\end{aligned}$$

and

$$\left\langle r^k, G^{-1}r^k\right\rangle_{\mathcal{X}} = \left\langle G^{-1}(g-Gh^k), G(h^\dagger-h^k\right\rangle_{\mathcal{X}} = \left\langle e^k, Ge^k\right\rangle.$$

The proof of f) is analogous to the proof of e).

□

It is left to investigate the important situation $p^{k+1}=0$ for some k, since α_k is not defined in this case. As in the beginning of the proof of the last theorem this assumption yields $r^k=0$ and so $0=g-Gh^k$, that is $h^k=h^\dagger$ is the exact solution of (3.4). Hence, a zero search direction can be produced only after the solution of (3.4) has already been found, at which stage we stop the iteration.

Altogether we can conclude that the CG-iterates are well defined until the true solution $h^\dagger$ has been found. If this is the case, the algorithm produces a zero search direction p^k, which implies that the residual vanishes.

In the computationally relevant case $\dim(\mathcal{X})<\infty$, the particular choice of $\beta_1,\beta_2,\ldots$ making the set of search directions conjugately orthogonal, the algorithm must stop with $m\le\dim(\mathcal{X})$, because we can generate at most $\dim(\mathcal{X})$ such mutually orthogonal vectors. Hence, the CG-method can be considered as a direct solution method. This holds at least in the absence of rounding errors, which usually lead to a loss in the orthogonality of the residual vectors. We will discuss this handicap in Chapter 6.

However, usually m is a large number or in the case when we deal for theoretical purposes with an infinite dimensional Hilbert space $\mathcal{X}$, generally $h^j\neq 0$ for all $j\in\mathbb{N}$. Accordingly, we define the ultimate termination index

$$\sup\{j\in\mathbb{N} : Gh^j-g\neq 0\}.$$

But primarily we are interested in the case where the residual r^j is sufficiently small after a few steps. Then the method is used as an iterative method, coupled with some stopping criterion.

3.3 Preconditioned conjugate gradient method

We will show in Section 4.1 that the speed of convergence of the CG-method strongly depends on the eigenvalue distribution of the operator G. So, to improve convergence rates of the CG-method, one possible way is to manipulate the eigenvalue distribution of the operator G. This, for instance, can be done by multiplying the operator equation (3.4) from the left by a boundedly invertible operator M. More precisely, the operator M should satisfy the following conditions:

- $M \in L(\mathcal{X}, \mathcal{X})$ and M is self-adjoint.
- $M \approx G$, i.e. M should be a good approximation to G.
- The storage requirements for M should be acceptable.
- The system $Mz = c$ must be efficiently solvable.

We call such an operator M a preconditioner for G.

Assuming that a preconditioner M for G is available, instead of solving the linear system (3.4) we replace it by the mathematically equivalent system

$$M^{-1}Gh = M^{-1}g. \tag{3.18}$$

Obviously, in general the operator $M^{-1}G$ is not self-adjoint with respect to $(\cdot\,,\cdot)$. But since we have formulated the CG-method for an arbitrary inner product on $\mathcal{X}$ we have freedom in the choice of this inner product. So we can choose it suitable for the preconditioned equation (3.18), that is we define the inner product via the preconditioner M:

Lemma 3.3 *Assume that $M : \mathcal{X} \to \mathcal{X}$ is a linear bounded strictly coercive and self-adjoint operator with respect to $(\cdot\,,\cdot)$.*

a) The mapping

$$(x, y)_M := (x, My) \tag{3.19}$$

defines an inner product on $\mathcal{X}$.

b) The linear operator $M^{-1}G$ is selfadjoint and strictly coercive with respect to $(\cdot\,,\cdot)_M$.

Proof: Both assertions follow by a straightforward computation.

□

Note that the residuals r^k are replaced in the preconditioned CG-method by the so-called pseudo residuals

$$z^k = M^{-1}r^k. \tag{3.20}$$

This statement is to be interpreted as solving the system $Mz^k = r^k$ if no explicit formula for M^{-1} is available. This underlines the last requirement for a preconditioner formulated above, since system (3.18) should be more efficiently solvable than system (3.4). By Theorem 3.2e) assuming that G is self-adjoint with respect to $\langle \cdot\,,\,\cdot \rangle_{\mathcal{X}}$ we have that

$$\left\langle e^0 + v, G(e^0 + v) \right\rangle_{\mathcal{X}}$$

is minimized over the Krylov subspace $\mathcal{K}_k(G, r^0)$. In the preconditioned case this Krylov subspace is replaced by $\mathcal{K}(M^{-1}G, z^0)$. With a good choice of the preconditioner, this Krylov subspace may generate vectors to minimize this functional much faster than for the unpreconditioned subspace. To formulate this result more precisely we define Π_k to be the set of all polynomials with degree at most k. With Π_k^1 we denote the set of polynomials $p \in \Pi_k$ satisfying $p(0) = 1$. The following theorem holds:

Theorem 3.4 *The k-th residual and the k-th iterate of the preconditioned CG-method satisfy*

$$\begin{aligned} h^k &= h^0 + q_{k-1}(M^{-1}G)z^0, && (3.21)\\ e^k &= p_k(M^{-1}G)e^0, && (3.22)\\ z^k &= p_k(M^{-1}G)z^0, && (3.23) \end{aligned}$$

where $q_k \in \Pi_{k-1}$ and $p_k \in \Pi_k^1$ is given by $p_k(t) = 1 - tq_{k-1}(t)$. If the inner product is given by (3.19), then the preconditioned CG-method satisfies

$$\left\langle e^k, Ge^k \right\rangle_{\mathcal{X}} = \inf_{v \in \mathcal{K}_k(M^{-1}G, z^0)} \left\langle e^0 + v, G(e^0 + v) \right\rangle_{\mathcal{X}}. \qquad (3.24)$$

Recall that $e^k := h^\dagger - h^k$ denotes the iteration error.

Proof: (3.21) follows from Theorem 3.2c) and (3.20). The computations

$$\begin{aligned} e^k &= h^\dagger - h^0 + q_{k-1}(M^{-1}G)M^{-1}G(h^\dagger - h^0) = p_k(M^{-1}G)e^0,\\ z^k &= M^{-1}Ge^k = p_k(M^{-1}G)M^{-1}Ge^0 = p_k(M^{-1}G)z^0 \end{aligned}$$

prove (3.22) and (3.23). Using assertion d) of Theorem 3.2 we have

$$\begin{aligned} (z^k, (M^{-1}G)^{-1}z^k)_M &= \inf_{u \in \mathcal{S}_k(M^{-1}G, z^0)} (z^0 + u, G^{-1}M(z^0 + u))_M\\ &= \inf_{u \in \mathcal{S}_k(M^{-1}G, z^0)} \left\langle r^0 + Mu, G^{-1}(r^0 + Mu) \right\rangle_{\mathcal{X}}\\ &= \inf_{u \in \mathcal{S}_k(M^{-1}G, z^0)} \left\langle e^0 + G^{-1}Mu, G(e^0 + G^{-1}Mu) \right\rangle_{\mathcal{X}}\\ &= \inf_{u \in \mathcal{K}_k(M^{-1}G, z^0)} \left\langle e^0 + v, G(e^0 + v) \right\rangle_{\mathcal{X}}. \end{aligned}$$

Since the functional J from (3.5) with the inner product (3.19) takes the form

$$\begin{aligned} J(h) &= \frac{1}{2}(z, (M^{-1}G)^{-1}z)_M \\ &= \frac{1}{2}\left\langle M^{-1}r, MG^{-1}MM^{-1}r\right\rangle_{\mathcal{X}} \\ &= \frac{1}{2}\left\langle r, G^{-1}r\right\rangle_{\mathcal{X}}, \end{aligned}$$

where $r = g - Gh$ and $z = M^{-1}r$, we see that the preconditioned CG-method minimizes the functional, $\langle r, G^{-1}r\rangle_{\mathcal{X}}$ which coincides with the functional of Theorem 3.2e). This proves (3.24).

□

Unfortunately, it is in general not an easy task to construct an efficient preconditioner, since often detailed knowledge about the operator G needs to be at hand. We will discuss this topic in more detail in Chapters 4, 5 and 6.

3.4 Computational considerations

Various identities allow a number of different formulations of the CG-algorithm. For example, the residuals can be computed directly by $r^k = g - Gh^k$ instead of using the recursion (3.8). Of course, we are interested in an efficient formulation of the CG-method.

Therefore, instead of computing α_k and β_k by (3.7) and (3.14), these formulas should be replaced by

$$\alpha_k = \frac{(r^{k-1}, r^{k-1})}{(p^k, Gp^k)} \qquad \text{and} \qquad \beta_k = \frac{(r^k, r^k)}{(r^{k-1}, r^{k-1})},$$

which are more stable and follow from (3.8) and (3.13) together with (3.15a). Moreover, in our case we are interested in a formulation of the CG-method suited for our problem at hand. We need to solve in each Newton step the normal equation $G_n^* G_n h_n = G_n^* g_n^\delta$, where the inner product is given by $(\,\cdot\,,\,\cdot\,) = \langle\,\cdot\,,\,\cdot\,\rangle_{\mathcal{X}}$. Hence, the formulas for α_k and β_k take the form

$$\alpha_k = \frac{\left\langle r^{k-1}, r^{k-1}\right\rangle_{\mathcal{X}}}{\langle p^k, G_n^* G_n p^k\rangle_{\mathcal{X}}} = \frac{\|r^{k-1}\|_{\mathcal{X}}^2}{\|G_n p^k\|_{\mathcal{X}\times\mathcal{Y}}^2}, \qquad \beta_k = \frac{\left\langle r^k, r^k\right\rangle_{\mathcal{X}}}{\left\langle r^{k-1}, r^{k-1}\right\rangle_{\mathcal{X}}}.$$

Introducing auxiliary vectors

$$d^0 := g_n^\delta, \qquad d^k := d^{k-1} - \alpha_k G_n p^k, \qquad k = 1, 2, \ldots, \tag{3.25}$$

we can show by induction that

$$G_n^* d^k = G_n^* d^{k-1} - \alpha_k G_n^* G_n p^k = r^{k-1} - \alpha_k G_n^* G_n p^k = r^k.$$

Finally, the CG-iteration applied to the system (3.2) can be coded as follows:

Algorithm 3.5 (Conjugate Gradient Algorithm)

$h_n^0 = 0;\ d^0 = g_n^\delta;\ r^0 = G_n^* d^0;\ p^1 = r^0;\ k = 0;$

while $\|r^k\|_{\mathcal{X}} > \varepsilon C(\gamma_n, k)$

$k = k + 1;$
$q^k = G_n p^k;$
$\alpha_k = \|r^{k-1}\|_{\mathcal{X}}^2 / \|q^k\|_{\mathcal{X}\times\mathcal{Y}}^2;$
$h_n^k = h_n^{k-1} + \alpha_k p^k;$
$d^k = d^{k-1} - \alpha_k q^k;$
$r^k = G_n^* d^k;$
$\beta_k = \|r^k\|_{\mathcal{X}}^2 / \|r^{k-1}\|_{\mathcal{X}}^2;$
$p^{k+1} = r^k + \beta_k p^k.$

Alternatively, we consider the preconditioned normal equation

$$M_n^{-1} G_n^* G_n h_n = M_n^{-1} G_n^* g_n^\delta, \tag{3.26}$$

where M_n denotes a preconditioner for the operator $G_n^* G_n$. In this case the choice for the inner product is $(x, y)_{M_n} := \langle x, M_n y\rangle_{\mathcal{X}}$ and the coefficients α_k and β_k can be efficiently computed by

$$\begin{aligned}
\alpha_k &= \frac{(z^{k-1}, z^{k-1})_{M_n}}{(p^k, M_n^{-1} G_n^* G_n)_{M_n}} = \frac{\langle z^{k-1}, r^{k-1}\rangle_{\mathcal{X}}}{\langle G_n p^k, G_n p^k\rangle_{\mathcal{X}\times\mathcal{Y}}} = \frac{\langle z^{k-1}, r^{k-1}\rangle_{\mathcal{X}}}{\|G_n p^k\|_{\mathcal{X}\times\mathcal{Y}}^2}, \\
\beta_k &= \frac{(z^k, z^k)_{M_n}}{(z^{k-1}, z^{k-1})_{M_n}} = \frac{\langle z^k, r^k\rangle_{\mathcal{X}}}{\langle z^{k-1}, r^{k-1}\rangle_{\mathcal{X}}}.
\end{aligned}$$

The preconditioned CG-method applied to (3.26) takes the form:

Algorithm 3.6 (Preconditioned conjugate gradient algorithm)

$h_n^0 = 0;\ d^0 = g_n^\delta;\ r^0 = G_n^* d^0;\ p^1 = z^0 = M_n^{-1} r^0;\ k = 0;$

while $\|r^k\|_{\mathcal{X}} > \varepsilon C(\gamma_n, k)$

$k = k + 1;$
$q^k = G_n p^k;$
$\alpha_k = \langle r^{k-1}, z^{k-1}\rangle_{\mathcal{X}} / \|q^k\|_{\mathcal{X}\times\mathcal{Y}}^2;$
$h_n^k = h_n^{k-1} + \alpha_k p^k;$
$d^k = d^{k-1} - \alpha_k q^k;$
$r^k = G_n^* d^k;$
$z^k = M_n^{-1} r^k;$

$$\beta_k = \langle r^k, z^k \rangle_{\mathcal{X}} / \langle r^{k-1}, z^{k-1} \rangle_{\mathcal{X}};$$
$$p^{k+1} = z^k + \beta_k p^k.$$

Although the operator equations (3.2) and (3.26) are formally well posed, in practice the iterates of the CG-method start to deteriorate after some CG-steps, in particular for "small" regularization parameters γ_n. The main reasons are the tendency towards increasing round-off error and loss of orthogonality of the residual vectors in the method (see [71, Chapter 13]). On the other hand, one frequently has a sufficiently accurate solution after a small number of steps. Therefore, the stopping criterion $\|r^k\|_{\mathcal{X}} > \varepsilon C(\gamma_n, k)$ possibly stops the iteration after a relatively small number of steps, if ε and $C(\gamma_n, k)$ are chosen in a proper way. Moreover, for our class of problems it can be observed that in the steps before the iterates start to deteriorate there is no significant upgrade in the iterates. Since any CG-step involves the evaluation of $F'[x]$ and $F'[x]^*$ to some given vectors it is possible to significantly reduce the total complexity of the IRGNM by the choice of a proper termination criterion.

Note that we have formulated Algorithms 3.5 and 3.6 with initial guess $h_n^0 = 0$. This is due to the fact that in general we do not have any information on an approximation to the minimizer of (3.5) beforehand. Naturally, if an initial guess is available it is recommended to use it. This possibly also reduces the number of steps until the stopping criterion is satisfied.

3.5 Lanczos' method

In this section we discuss Lanczos' method, which is an iterative method for approximating some of the eigenvalues and eigenvectors of a linear operator. The method's idea is to approximate this operator by a "small" matrix representing the operator on a low-dimensional Krylov subspace. Now, the eigenvalues of this matrix can be interpreted as approximations to the eigenvalues of the linear operator. The corresponding eigenvectors of the matrix can be used to compute approximations to the eigenvectors of the operator. Lanczos' method is closely related to the CG-method, and we restrict ourselves here to a formulation of this method which is based on quantities occurring in the CG-method. Moreover, for our purposes it turns out to be necessary to generalize some of the results that are presented in the books of Demmel [12], Golub & van Loan [24], Axelsson [2], Saad [77] and Fischer [17] for the finite dimensional case.

For the rest of this chapter we restrict ourselves to a description of Lanczos method when applied to the preconditioned linear system (3.26). The case for the system (3.2) is a special case of (3.26) with $M_n = I$. Then the inner product $(\cdot\,,\cdot)_{M_n}$ simplifies to $\langle \cdot\,,\cdot \rangle_{\mathcal{X}}$ and instead of considering the pseudo-residuals z^k we have to

consider the residuals r^k. To shorten the notation we skip for the rest of the chapter the index n and use the notation

$$\|x\|_{\mathcal{X}_M} := \sqrt{(x,x)_M}.$$

With $\mathcal{X}_M$ we denote the Hilbert space $\mathcal{X}$ equipped with the inner product $(\,\cdot\,,\,\cdot\,)_M$.

One important property of the CG-method is the construction of an orthogonal basis of the Krylov subspace $\mathcal{K}_k(M^{-1}G^*G, z^0)$. Lanczos' method takes advantage of this basis by approximating the operator $M^{-1}G^*G$ on this subspace. This approximation can be used to compute approximations to some of the eigenvalues and corresponding eigenvectors of $M^{-1}G^*G$. We start the description of Lanczos' method suited for the system (3.26) with the following two lemmas.

Lemma 3.7 *Let $\{u_j : j \in \mathbb{N}\}$ be a set of vectors in a Hilbert space $\mathcal{X}$ which are orthonormal with respect to some inner product $(\,\cdot\,,\,\cdot\,)$ on $\mathcal{X}$. Then the linear operator*

$$\begin{aligned} U_k : \mathbb{R}^k &\to \mathcal{X}, \\ (\xi_1, \ldots, \xi_k) &\mapsto \sum_{j=1}^{k} \xi_j u_j \end{aligned}$$

is bounded and isometric. Here we consider $\mathbb{R}^k$ equipped with the Euclidean inner product. Furthermore, the linear operator

$$\begin{aligned} U : l^2(\mathbb{N}) &\to \mathcal{X}, \\ \xi &\mapsto \sum_{j=1}^{\infty} \xi_j u_j, \qquad \xi = (\xi_j) \in l^2(\mathbb{N}), \end{aligned}$$

is bounded and isometric. The adjoint operators are given by

$$\begin{aligned} U_k^* : u &\mapsto ((u,u_1), \ldots, (u,u_k)), \\ U^* : u &\mapsto ((u,u_1), (u,u_2), \ldots). \end{aligned}$$

Moreover, the equality $U_k^* U_k = I_{\mathbb{R}^k}$ holds.

Proof: It is clear that U_k and U are linear and that $\|U_k\| \leq 1$ and $\|U\| \leq 1$. The isometry of U_k follows from the computation

$$(U_k\xi, U_k\xi) = \left(\sum_{j=1}^{k} \xi_j u_j, \sum_{j=1}^{k} \xi_j u_j \right) = \sum_{j=1}^{k} \xi_j^2 = (\xi,\xi)_{l^2(k)}.$$

To show that U_k^* is the adjoint operator we compute

$$(u, U_k\xi) = \left(u, \sum_{j=1}^{k} \xi_j u_j\right) = \sum_{j=1}^{k} \xi_j (u, u_j) = (U_k^* u, \xi)_{l^2(k)}.$$

Analogously we can prove the properties of U. To show that $U_k^* U_k = I_{\mathbb{R}^k}$ we compute

$$U_k^* U_k \xi = \sum_{j=1}^{k} \xi_j U_k^* u_j = \xi,$$

which is true for all $\xi \in \mathbb{R}^k$.

□

With the help of Lemma 3.7 we now define isometric mappings constructed out of quantities occuring in the CG-algorithm.

Corollary 3.8 *Let $\tilde{z}^j := z^j/\|z^j\|_{\mathcal{X}_M}$, $j = 0, \ldots, k-1$ and $\tilde{q}^j := q^j/\|q^j\|_{\mathcal{X}\times\mathcal{Y}}$, $j = 1, \ldots, k$, where z^j and q^j are determined by Algorithm 3.6. The linear operators $Z_k : \mathbb{R}^k \to \mathcal{X}_M$ and $Q_k : \mathbb{R}^k \to \mathcal{X} \times \mathcal{Y}$ given by*

$$\begin{aligned} Z_k\xi &:= \sum_{j=0}^{k-1} \xi_j \tilde{z}^j, \qquad \xi = (\xi_0, \ldots, \xi_{k-1}) \in \mathbb{R}^k, \\ Q_k\eta &:= \sum_{j=1}^{k} \eta_j \tilde{q}^j, \qquad \eta = (\eta_1, \ldots, \eta_k) \in \mathbb{R}^k \end{aligned}$$

are isometric and the adjoint operator of Z_k is given by

$$Z_k^* : z \;\mapsto\; ((z, \tilde{z}^0)_M, \ldots, (z, \tilde{z}^{k-1})_M)^T.$$

Moreover, $Z_k^* Z_k = I_{\mathbb{R}^k}$.

Proof: By the definition of $\tilde{z}^j$ and (3.15b) the set $\{\tilde{z}^j : j = 0, \ldots, k-1\}$ is an orthonormal system in $\mathcal{X}_M$. Moreover, using (3.15c) and the relation $q^j = Gp^j$ we compute for $i \neq \ell$

$$\langle q^i, q^\ell \rangle_{\mathcal{X}\times\mathcal{Y}} = \langle p^i, G^*Gp^\ell \rangle_{\mathcal{X}} = 0.$$

Hence, the set $\{\tilde{q}^j : j = 1, \ldots, k\}$ is an orthonormal system in $\mathcal{X}\times\mathcal{Y}$. The assertions now follow from Lemma 3.7.

□

It is our aim to derive formulas which show that the linear operator $M^{-1}G^*G$ can be approximately represented on the subspace $\text{span}\{\tilde{z}^0, \ldots, \tilde{z}^{k-1}\}$ by a symmetric and positive definite tridiagonal matrix. Several formulas used in the following arise from Algorithm 3.6.

Multiplying the identity $z^j = p^{j+1} - \beta_j p^j$ from the left by $\|z_j\|_{\mathcal{X}_M}^{-1} G$ and using the equalities

$$\begin{aligned}\frac{\|q^{j+1}\|_{\mathcal{X}\times\mathcal{Y}}}{\|z^j\|_{\mathcal{X}_M}} &= \frac{\|q^{j+1}\|_{\mathcal{X}\times\mathcal{Y}}}{\langle r^j, z^j\rangle_{\mathcal{X}}^{1/2}} = \frac{1}{\sqrt{\alpha_{j+1}}},\\ \frac{\|q^j\|_{\mathcal{X}\times\mathcal{Y}}}{\|z^j\|_{\mathcal{X}_M}} &= \frac{\langle r^{j-1}, z^{j-1}\rangle_{\mathcal{X}}^{1/2}}{\sqrt{\alpha_j}} \frac{1}{\langle r^j, z^j\rangle_{\mathcal{X}}^{1/2}} = \frac{1}{\sqrt{\alpha_j\beta_j}},\end{aligned}$$

yields for all $j = 1, \ldots, k-1$

$$\begin{aligned}G\tilde{z}^j &= \frac{1}{\|z^j\|_{\mathcal{X}_M}} G(p^{j+1} - \beta_j p^j) = \frac{1}{\|z^j\|_{\mathcal{X}_M}}(q^{j+1} - \beta_j q^j)\\ &= \frac{1}{\|z^j\|_{\mathcal{X}_M}}(\|q^{j+1}\|_{\mathcal{X}\times\mathcal{Y}}\tilde{q}^{j+1} - \beta_j\|q^j\|_{\mathcal{X}\times\mathcal{Y}}\tilde{q}^j)\\ &= \frac{1}{\sqrt{\alpha_{j+1}}}\tilde{q}^{j+1} - \sqrt{\frac{\beta_j}{\alpha_j}}\tilde{q}^j. \qquad (3.27)\end{aligned}$$

For $k = 0$ we have

$$G\tilde{z}^0 = \frac{1}{\|z^0\|_{\mathcal{X}_M}} Gp^1 = \frac{\|q^1\|}{\langle r^0, z^0\rangle^{1/2}}\tilde{q}^1 = \frac{1}{\sqrt{\alpha_1}}\tilde{q}^1. \qquad (3.28)$$

Analogously, the identity $\alpha_j q^j = d^{j-1} - d^j$ multiplied from the left by $(\|q^j\|_{\mathcal{X}\times\mathcal{Y}}\alpha_j)^{-1} G^*$ together with the equalities

$$\begin{aligned}\frac{\|z^{j-1}\|_{\mathcal{X}_M}}{\|q^j\|_{\mathcal{X}\times\mathcal{Y}}\alpha_j} &= \frac{\langle r^{j-1}, z^{j-1}\rangle_{\mathcal{X}}^{1/2}}{\|q^j\|_{\mathcal{X}\times\mathcal{Y}}\alpha_j} = \frac{1}{\sqrt{\alpha_j}},\\ \frac{\|z^j\|_{\mathcal{X}_M}}{\|q^j\|_{\mathcal{X}\times\mathcal{Y}}\alpha_j} &= \frac{\langle r^j, z^j\rangle_{\mathcal{X}}^{1/2}}{\|q^j\|_{\mathcal{X}\times\mathcal{Y}}\alpha_j} \frac{\langle r^{j-1}, z^{j-1}\rangle_{\mathcal{X}}^{1/2}}{\langle r^{j-1}, z^{j-1}\rangle_{\mathcal{X}}^{1/2}} = \sqrt{\frac{\beta_j}{\alpha_j}},\end{aligned}$$

yields for all $j = 1, \ldots, k$

$$\begin{aligned}M^{-1}G^*\tilde{q}_j &= \frac{1}{\|q^j\|_{\mathcal{X}\times\mathcal{Y}}\alpha_j} M^{-1}G^*(d^{j-1} - d^j) = \frac{1}{\|q^j\|_{\mathcal{X}\times\mathcal{Y}}\alpha_j}(z^{j-1} - z^j)\\ &= \frac{1}{\|q^j\|\alpha_j}\left(\|z^{j-1}\|_{\mathcal{X}_M}\tilde{z}^{j-1} - \|z^j\|_{\mathcal{X}_M}\tilde{z}^j\right)\\ &= \frac{1}{\sqrt{\alpha_j}}\tilde{z}^{j-1} - \sqrt{\frac{\beta_j}{\alpha_j}}\tilde{z}^j. \qquad (3.29)\end{aligned}$$

Putting (3.27), (3.28) and (3.29) together, we have proven the formulas

$$M^{-1}G^*G\tilde{z}^0 = \frac{1}{\alpha_1}\tilde{z}^0 - \frac{\sqrt{\beta_1}}{\alpha_1}\tilde{z}^1, \qquad (3.30a)$$

$$M^{-1}G^*G\tilde{z}^j = -\frac{\sqrt{\beta_j}}{\alpha_j}\tilde{z}^{j-1} + \left(\frac{1}{\alpha_{j+1}} + \frac{\beta_j}{\alpha_j}\right)\tilde{z}^j - \frac{\sqrt{\beta_{j+1}}}{\alpha_{j+1}}\tilde{z}^{j+1}, \qquad (3.30b)$$

which hold for all $j = 1, \ldots, k-1$. Using the definition

$$T_k := \begin{pmatrix} \frac{1}{\alpha_1} & -\frac{\sqrt{\beta_1}}{\alpha_1} & & & \\ -\frac{\sqrt{\beta_1}}{\alpha_1} & \frac{1}{\alpha_2} + \frac{\beta_1}{\alpha_1} & -\frac{\sqrt{\beta_2}}{\alpha_2} & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & -\frac{\sqrt{\beta_{k-1}}}{\alpha_{k-1}} \\ & & & -\frac{\sqrt{\beta_{k-1}}}{\alpha_{k-1}} & \frac{1}{\alpha_k} + \frac{\beta_{k-1}}{\alpha_{k-1}} \end{pmatrix} \tag{3.31}$$

we can express (3.30) in matrix notation by

$$M^{-1}G^*GZ_k = Z_kT_k - \left(0, \ldots, 0, \frac{\sqrt{\beta_k}}{\alpha_k} z^k\right). \tag{3.32}$$

Multiplying equation (3.32) with Z_k^* from the left yields

$$Z_k^* M^{-1} G^* G Z_k = T_k, \tag{3.33}$$

which shows that T_k is positive definite. Now, let us denote by $\theta_1 \geq \ldots \geq \theta_k > 0$ and $v_1, \ldots, v_k$ the eigenvalues with corresponding eigenvectors of the symmetric and positive definite matrix $T_k \in \mathbb{R}^{k \times k}$. Then it follows from (3.33) that the equation

$$Z_k^* M^{-1} G^* G Z_k v_j = \theta_j v_j, \qquad j = 1, \ldots, k$$

holds. Moreover, in the case when $z^k = 0$, that is the preconditioned CG-method terminates with the exact solution of (3.26), we have

$$M^{-1}G^*GZ_kv_j = \theta_j Z_k v_j, \qquad j = 1, \ldots, k,$$

that is $\theta_1 \geq \ldots \geq \theta_k$ are exact eigenvalues and $Z_kv_1, \ldots, Z_kv_k$ are corresponding eigenvectors of $M^{-1}G^*G$.

However, the assumption $z^k = 0$ is not realistic. As already mentioned, we use the CG-method as an iterative method coupled with some stopping criterion. Hence, in the expected case $z^k \neq 0$ we need a justification to interpret the values $\theta_1 \geq \ldots \geq \theta_k$ and vectors $Z_kv_1, \ldots, Z_kv_k$ as approximations to some eigenvalues with corresponding eigenvectors of $M^{-1}G^*G$. This will be the topic of the next section. Moreover, we will prove that these approximations can be considered as optimal in the sense that they satisfy a minimization problem.

Remark 3.9 *There is a close connection between the CG-method and orthogonal polynomials (see for example* [17]*). With the help of Theorems 3.2 and 3.4 it can be shown that the mapping*

$$(p, q)_\pi := \left(p(M^{-1}G^*G)z^0), q(M^{-1}G^*G)z^0)\right)_M. \tag{3.34}$$

defines an inner product on $\Pi = \bigcup_{m\in\mathbb{N}_0} \Pi_m$ *and that the polynomials* p_k *arising in Theorem 3.4 are orthogonal with respect to this inner product, that is* $(p_i, p_j)_\pi = 0$ *if* $i \neq j$*. The matrices* T_k *derived above just collect the three-term recurrence coefficients of the orthogonal polynomial* p_k*,* $k = 1, 2, \ldots$*, given by (3.23).*

As a consequence of Remark 3.9 together with properties of orthogonal polynomials the following theorem can be proven.

Theorem 3.10 *Let us denote the eigenvalues of the operator* $M^{-1}G^*G$*, which is self-adjoint with respect to* $(\cdot\,,\cdot)_M$*, by* $\tilde{\mu}_1 \geq \tilde{\mu}_2 \geq \ldots$ *and by* $\theta_j^{(k)}$*,* $j = 1, \ldots, k$*, the eigenvalues of the matrices* T_k*,* $k = 1, 2, \ldots$*. Then the following assertions hold true:*

a) The eigenvalues of T_k *are all real, simple and*

$$\inf_{s\in\mathbb{N}} \tilde{\mu}_s < \theta_1^{(k)} < \theta_2^{(k)} < \ldots < \theta_k^{(k)} < \tilde{\mu}_1.$$

In particular, if $M = I$ *we have that*

$$\gamma_n < \theta_1^{(k)} < \theta_2^{(k)} < \ldots < \theta_k^{(k)} < \gamma_n + \lambda_1,$$

where γ_n *denotes the regularization parameter and* λ_1 *the largest eigenvalue of the compact operator* $F'[x_n^\delta]^*F'[x_n^\delta]$ *(see Section 3.1).*

b) The eigenvalues of T_k *and* T_{k+1} *separate each other, i.e.*

$$\theta_1^{(k+1)} < \theta_1^{(k)} < \theta_2^{(k+1)} < \ldots < \theta_k^{(k)} < \theta_{k+1}^{(k+1)}.$$

c) The polynomial p_k *defined through (3.23) is given by*

$$p_k(t) = \Pi_{j=1}^k (1 - t/\theta_j^{(k)}).$$

Proof: See [17, Chapter 2] and [80].

□

Note that a direct consequence of Theorem 3.10 is that Lanczos' method approximates at most one of a multiple eigenvalue with a corresponding eigenvector. In our case this is an important drawback for the construction of a preconditioner (see Chapter 6). On the other hand, if there are multiple eigenvalues we expect faster convergence of the CG-method, since this method performs at most one step for a multiple eigenvalue. Hence, Theorem 3.10 indicates that there is a correspondence between the eigenvalue distribution and the convergence behavior of the CG-method.

3.6 The Rayleigh-Ritz Method

Instead of considering the special operator Z_k defined by the pseudo residuals of the preconditioned CG-method as in the last section, we generalize this idea. Therefore assume that $\{u_j : j \in \mathbb{N}\}$ is an orthonormal system in $\mathcal{X}_M$. We now approximate eigenvalues of $M^{-1}G^*G$ on the subspace $\text{span}\{u_1, \ldots, u_k\} \subset \mathcal{X}_M$. To this end recall the isometric operator U_k defined in Lemma 3.7, which we consider here as a mapping from $\mathbb{R}^k$ to $\mathcal{X}_M$ and the operator $U : l^2(\mathbb{N}) \to \mathcal{X}_M$. To formulate the Rayleigh-Rith method in infinite dimensions we furthermore introduce the shift operator $J_k : l^2(\mathbb{N}) \to l^2(\mathbb{N})$ given by

$$J_k\xi = (0, \ldots, 0, \xi_1, \xi_2, \ldots), \qquad \xi = (\xi_j) \in l^2(\mathbb{N}),$$

where the first k entries of the vector are zero, and its adjoint operator

$$J_k^*\xi = (\xi_{k+1}, \xi_{k+2}, \ldots), \qquad \xi = (\xi_j) \in l^2(\mathbb{N}).$$

Moreover, we define $V_k := UJ_k^*$ and $\tilde{\xi}$ to be the first k components of $\xi \in l^2(\mathbb{N})$, that is $\tilde{\xi} := (\xi_1, \ldots, \xi_k)$. Now consider the operator $(U_k, V_k) : l^2(\mathbb{N}) \to \mathcal{X}_M$,

$$(U_k, V_k)\xi := U_k\tilde{\xi} + V_k\xi.$$

Obviously the equality $(U_k, V_k) = U$ is satisfied. To shorten the notation we define $\tilde{G} := M^{-1}G^*G$ and consider the partitioning

$$U^*\tilde{G}U = \begin{pmatrix} U_k^*\tilde{G}U_k & U_k^*\tilde{G}V_k \\ V_k^*\tilde{G}U_k & V_k^*\tilde{G}V_k \end{pmatrix}. \tag{3.35}$$

Notice that $S_k := U_k^*\tilde{G}U_k$ is represented by the matrix

$$S_k = [(u_i, \tilde{G}u_j)_{\mathcal{X}_M}]_{1\le i,j\le k} \in \mathbb{R}^{k\times k}. \tag{3.36}$$

It is clear that for $k = 1$,

$$S_1 = U_1^*\tilde{G}U_1$$

is just the Rayleigh quotient of $\tilde{G}$ with respect to u_1 and the inner product $(\,\cdot\,,\,\cdot\,)_{\mathcal{X}_M}$. So for $k > 1$, S_k is a natural generalization of the Rayleigh quotient. The Rayleigh-Ritz procedure consists in approximating the eigenvalues of $\tilde{G}$ by the eigenvalues of S_k.

Definition 3.11 *The eigenvalues $\theta_1 \ge \ldots \ge \theta_k$ of the symmetric matrix S_k are called Ritz values. Let $v_1, \ldots, v_k \in \mathbb{R}^k$ denote the corresponding orthonormal eigenvectors. The orthonormal vectors $U_kv_1, \ldots, U_kv_k \in \mathcal{X}_M$ are called Ritz vectors.*

The proceeding above is motivated by the assumption that the vectors $u_1, \ldots, u_k$ are known. Then S_k is a natural approximation of $\tilde{G}$ on the subspace $\text{span}\{u_1, \ldots, u_k\}$. Therefore the Ritz values and vectors are the natural approximations from the known part of the operator. The following theorem even states that they are optimal in a certain sense:

Theorem 3.12 *The following optimality result holds:*

$$\min_{S \in \mathbb{R}^{k \times k}} \|\tilde{G}U_k - U_k S\|_{\mathcal{X}_M \leftarrow \mathbb{R}^k} = \|\tilde{G}U_k - U_k S_k\|_{\mathcal{X}_M \leftarrow \mathbb{R}^k}. \tag{3.37}$$

Furthermore, we have the equality

$$\|\tilde{G}U_k - U_k S_k\|_{\mathcal{X}_M \leftarrow \mathbb{R}^k} = \|V_k^* \tilde{G} U_k\|_{l^2(\mathbb{N})}. \tag{3.38}$$

Proof: Let $S = S_k + B$ where $B \in \mathbb{R}^{k \times k}$. Then we can estimate

$$\begin{aligned}
\|(\tilde{G}U_k - U_k S)\xi\|_{\mathcal{X}_M}^2 &= (\tilde{G}U_k\xi - U_k S_k \xi - U_k B\xi, \tilde{G}U_k\xi - U_k S_k\xi - U_k B\xi)_{\mathcal{X}_M} \\
&= \|(\tilde{G}U_k - U_k S_k)\xi\|_{\mathcal{X}_M}^2 \\
&\quad -2(\tilde{G}U_k\xi - U_k S_k\xi, U_k B\xi)_{\mathcal{X}_M} + \|U_k B\xi\|_{\mathcal{X}_M}^2 \\
&= \|(\tilde{G}U_k - U_k S_k)\xi\|_{\mathcal{X}_M}^2 \\
&\quad -2((U_k^*\tilde{G}U_k - S_k)\xi, B\xi)_{\mathbb{R}^k} + \|U_k B\xi\|_{\mathcal{X}_M}^2 \\
&\geq \|(\tilde{G}U_k - U_k S_k)\xi\|_{\mathcal{X}_M}^2,
\end{aligned}$$

which holds for all $\xi \in \mathbb{R}^k$. This proves (3.37). Using the isometry of U, the equation $V_k^* U_k = J_k^* U^* U_k = 0$ and (3.38) we can conclude for all $\xi \in \mathbb{R}^k$

$$\begin{aligned}
\|(\tilde{G}U_k - U_k S_k)\xi\|_{\mathcal{X}_M} &= \|(U^*\tilde{G}U_k - U^* U_k S_k)\xi\|_{l^2(\mathbb{N})} \\
&= \left\| \left[\begin{pmatrix} U_k^*\tilde{G}U_k \\ V_k^*\tilde{G}U_k \end{pmatrix} - \begin{pmatrix} S_k \\ 0 \end{pmatrix} \right] \xi \right\|_{l^2(\mathbb{N})} \\
&= \left\| \left[\begin{pmatrix} S_k \\ V_k^*\tilde{G}U_k \end{pmatrix} - \begin{pmatrix} S_k \\ 0 \end{pmatrix} \right] \xi \right\|_{l^2(\mathbb{N})} \\
&= \|V_k^*\tilde{G}U_k\xi\|_{l^2(\mathbb{N})}.
\end{aligned}$$

□

Corollary 3.13 *Let $\theta_1 \geq \ldots \geq \theta_k$ be the eigenvalues of S_k and $v_1, \ldots, v_k$ the corresponding orthonormal eigenvectors. We define $\Lambda = \text{diag}(\theta_1, \ldots, \theta_k)$ and $V = (v_1, \ldots, v_k)$. Then we have the equality*

$$\min_{S \in \mathbb{R}^{k \times k}} \|\tilde{G}U_k - U_k S\|_{\mathcal{X}_M \leftarrow \mathbb{R}^k} = \|\tilde{G}U_k V - U_k V\Lambda\|_{\mathcal{X}_M \leftarrow \mathbb{R}^k} = \|V_k^*\tilde{G}U_k\|_{l^2(\mathbb{N})}.$$

Proof: To proof the corollary by formulas (3.37) and (3.38) we only have to show that $\|\tilde{G}U_k - U_kS_k\|_{\mathcal{X}_M \leftarrow \mathbb{R}^k} = \|\tilde{G}U_kV - U_kV\Lambda\|_{\mathcal{X}_M \leftarrow \mathbb{R}^k}$. By the assumption we have $V^TS_kV = \Lambda$, and since $V^TV = I_{\mathbb{R}^k}$ we have

$$\begin{aligned} \|\tilde{G}U_k - U_kS_k\|_{\mathcal{X}_M \leftarrow \mathbb{R}^k} &= \|\tilde{G}U_k - U_kV\Lambda V^T\|_{\mathcal{X}_M \leftarrow \mathbb{R}^k} \\ &= \|\tilde{G}U_kV - U_kV\Lambda\|_{\mathcal{X}_M \leftarrow \mathbb{R}^k}. \end{aligned}$$

This proves the assertion.

□

Theorem 3.12 and Corollary 3.13 justify the use of Ritz values and Ritz vectors as approximations to the eigenvalues and eigenvectors.

Lanczos' method is a particular case of the Rayleigh-Ritz method, where the operator U_k is given by Z_k. Naturally, the matrix $S_k \in \mathbb{R}^{k\times k}$ coincides in this case with $T_k \in \mathbb{R}^{k\times k}$ defined in (3.31) and it is easy to compute all the quantities of Theorem 3.12 and Corollary 3.13. This is because there are good algorithms for finding eigenvalues and eigenvectors of the symmetric tridiagonal matrix T_k and because the residual norm is simply $\|V_k^*\tilde{G}U_k\|_{l^2(\mathbb{N})} = \sqrt{\beta_k}/\alpha_k$, which is a consequence of formula (3.32) and the nonnegativity of $\sqrt{\beta_k}/\alpha_k$. Let us summarize these simplified error bounds on the approximate eigenvalues and eigenvectors in the following theorem.

Theorem 3.14 *Assume that S_k and U_k are determined by Lanczos' method, that is $S_k = T_k$ and $U_k = Z_k$. Let $T_k = V\Lambda V^T$ be the eigendecomposition of T_k, where $V = (v_1, \ldots, v_k)$ is orthogonal and $\Lambda = \operatorname{diag}(\theta_1, \ldots, \theta_k)$. Then the following statements hold true:*

a) There are k eigenvalues $\tilde{\mu}_1, \ldots, \tilde{\mu}_k$ of $\tilde{G}$ (not necessarily the largest k) such that

$$|\theta_i - \tilde{\mu}_i| \leq \frac{\sqrt{\beta_k}}{\alpha_k}, \qquad i = 1, \ldots, k.$$

b)

$$\|\tilde{G}(Z_kv_i) - (Z_kv_i)\theta_i\|_{\mathcal{X}_M} = \frac{\sqrt{\beta_k}}{\alpha_k}|v_i(k)|, \tag{3.39}$$

where $v_i(k)$ is the k-th (bottom) entry of v_i.

Proof: Recall the partitioning (3.35) with the operator U_k replaced by Z_k and the corresponding operator V_k. The eigenvalues of the linear operator

$$\hat{T} := \begin{pmatrix} Z_k^*\tilde{G}Z_k & 0 \\ 0 & V_k^*\tilde{G}V_k \end{pmatrix}$$

include $\theta_1 \geq \ldots \geq \theta_k$. Since

$$\|\hat{T} - U^*\tilde{G}U\|_{l^2(\mathbb{N})} = \left\| \begin{pmatrix} 0 & Z_k^*\tilde{G}V_k \\ V_k^*\tilde{G}U_k & 0 \end{pmatrix} \right\|_{l^2(\mathbb{N})} = \|V_k^*\tilde{G}Z_k\|_{l^2(\mathbb{N})},$$

by Weyl's Theorem (see [36, Theorem 32.6]), it follows that the eigenvalues of $\hat{T}$ and $U^*\tilde{G}U$ differ at most by $\|V_k^*\tilde{G}Z_k\|_{l^2(\mathbb{N})}$. But the eigenvalues of $U^*\tilde{G}U$ and $\tilde{G}$ are the same. This together with (3.32) shows a). b) is also a consequence of (3.32). □

Further estimates for the Rayleigh-Ritz method can be found for example in [51].

3.7 Kaniel-Paige Convergence Theory

In the previous section, we obtained computable a-posteriori error estimates for the Ritz values and Ritz vectors computed by Lanczos' method. But so far we know nothing about the rate of convergence. There is another error bound, due to Kaniel-Paige and Saad, that sheds light on this aspect. This error bound depends on the angle between the first normalized residual vector $\tilde{z}^0$ and the desired eigenvectors, the Ritz values, and the desired eigenvalues. In other words, it depends on quantities unknown during the computation, so it is of no practical use. But it shows that if $\tilde{z}^0$ is nearly orthogonal to the desired eigenvector, or if the desired eigenvalue is nearly multiple, then we can expect slow convergence.

To formulate this error bound we introduce Chebysheff polynomials of the first kind of degree j defined by

$$c_j(x) := \begin{cases} \cos(j \arccos x), & -1 \leq x \leq 1, \\ \frac{1}{2}\left[\left(x + \sqrt{x^2-1}\right)^j + \left(x - \sqrt{x^2-1}\right)^j\right], & |x| > 1. \end{cases} \tag{3.40}$$

Theorem 3.15 *Let us denote the eigenvalues of the bounded operator $\tilde{G} = M^{-1}G^*G$, which is self-adjoint and strictly coercive with respect to $(\,\cdot\,,\,\cdot\,)_M$, by $\tilde{\mu}_1 \geq \tilde{\mu}_2 \geq \ldots > \tilde{\mu}_r \geq \ldots$ and the corresponding orthonormal eigenvectors by $\tilde{\varphi}_1, \tilde{\varphi}_2, \ldots$. Let $\theta_1 > \ldots > \theta_k$ be the eigenvalues of T_k given by (3.31), then*

$$\tilde{\mu}_1 \geq \theta_1 \geq \tilde{\mu}_1 - (\tilde{\mu}_1 - \tilde{\mu})\frac{\tan(\chi_1)^2}{(c_{k-1}(1+2\rho_1))^2},$$

where

$$\cos(\chi_1) = |(\tilde{z}^0, \tilde{\varphi}_1)_M|, \qquad \tilde{\mu} = \inf_{s\in\mathbb{N}} \tilde{\mu}_s \qquad \textit{and} \qquad \rho_1 = \frac{\tilde{\mu}_1 - \tilde{\mu}_2}{\tilde{\mu}_2 - \tilde{\mu}}.$$

Proof: Recall formula (3.23). Hence, each vector $\tilde{z}^j$ can be written as $\tilde{z}^j = p_j(M^{-1}G^*G)\tilde{z}^0$, where p_j is a polynomial of degree j, $j = 0, \ldots, k-1$. Since θ_1 is the largest eigenvalue of T_k, this together with the definition $\tilde{G} = M^{-1}G^*G$ and (3.33) yields

$$\theta_1 = \max_{\xi \neq 0} \frac{(\xi, T_k\xi)_{l^2(k)}}{(\xi,\xi)_{l^2(k)}} = \max_{\xi \neq 0} \frac{(Z_k\xi, \tilde{G}Z_k\xi)_M}{(Z_k\xi, Z_k\xi)_M} = \max_{p \in \Pi_{k-1}} \frac{(p(\tilde{G})\tilde{z}^0, \tilde{G}p(\tilde{G})\tilde{z}^0)_M}{(p(\tilde{G})\tilde{z}^0, p(\tilde{G})\tilde{z}^0)_M}.$$

By Courant's Minimum-Maximum principle (see [36, Theorem 32.4]) it follows that $\theta_1 \leq \tilde{\mu}_1$. To establish the other bound on θ_1, we write $\tilde{z}^0 = \sum_{i=1}^{\infty}(\tilde{z}^0, \tilde{\varphi}_i)_M\tilde{\varphi}_i$ and obtain for $p \in \Pi_{k-1}$

$$\begin{aligned}
\frac{(p(\tilde{G})\tilde{z}^0, \tilde{G}p(\tilde{G})\tilde{z}^0)_M}{(p(\tilde{G})\tilde{z}^0, p(\tilde{G})\tilde{z}^0)_M} &= \frac{\sum_{i=1}^{\infty}(\tilde{z}^0, \tilde{\varphi}_i)_M^2 p(\tilde{\mu}_i)^2\tilde{\mu}_i}{\sum_{i=1}^{\infty}(\tilde{z}^0, \tilde{\varphi}_i)_M^2 p(\tilde{\mu}_i)^2} \\
&\geq \frac{\tilde{\mu}_1(\tilde{z}^0, \tilde{\varphi}_1)_M^2 p(\tilde{\mu}_1)^2 + \tilde{\mu}\sum_{i=2}^{\infty}(\tilde{z}^0, \tilde{\varphi}_i)_M^2 p(\tilde{\mu}_i)^2}{(\tilde{z}^0, \tilde{\varphi}_1)_M^2 p(\tilde{\mu}_1)^2 + \sum_{i=2}^{\infty}(\tilde{z}^0, \tilde{\varphi}_i)_M^2 p(\tilde{\mu}_i)^2} \\
&= \frac{\tilde{\mu}_1((\tilde{z}^0, \tilde{\varphi}_1)^2 p(\tilde{\mu}_1)^2 + \sum_{i=2}^{\infty}(\tilde{z}^0, \tilde{\varphi}_i)_M^2 p(\tilde{\mu}_i)^2)}{(\tilde{z}^0, \tilde{\varphi}_1)_M^2 p(\tilde{\mu}_1)^2 + \sum_{i=2}^{\infty}(\tilde{z}^0, \tilde{\varphi}_i)_M^2 p(\tilde{\mu}_i)^2} \\
&\quad - \frac{\tilde{\mu}_1\sum_{i=2}^{\infty}(\tilde{z}^0, \tilde{\varphi}_i)_M^2 p(\tilde{\mu}_i)^2 - \tilde{\mu}\sum_{i=2}^{\infty}(\tilde{z}^0, \tilde{\varphi}_i)_M^2 p(\tilde{\mu}_i)^2}{(\tilde{z}^0, \tilde{\varphi}_1)_M^2 p(\tilde{\mu}_1)^2 + \sum_{i=2}^{\infty}(\tilde{z}^0, \tilde{\varphi}_i)_M^2 p(\tilde{\mu}_i)^2} \\
&= \tilde{\mu}_1 - (\tilde{\mu}_1 - \tilde{\mu})\frac{\sum_{i=2}^{\infty}(\tilde{z}^0, \tilde{\varphi}_i)_M^2 p(\tilde{\mu}_i)^2}{(\tilde{z}^0, \tilde{\varphi}_1)_M^2 p(\tilde{\mu}_1)^2 + \sum_{i=2}^{\infty}(\tilde{z}^0, \tilde{\varphi}_i)_M p(\tilde{\mu}_i)^2}.
\end{aligned}$$

We can derive a sharp bound by choosing

$$p(x) := c_{k-1}\left(-1 + 2\frac{x - \tilde{\mu}}{\tilde{\mu}_2 - \tilde{\mu}}\right).$$

The argument to c_{k-1} is constructed to map the interval $[\tilde{\mu}, \tilde{\mu}_2]$ to $[-1, 1]$. Since the Chebysheff polynomial c_{k-1} satisfies $|c_{k-1}(x)| \leq 1$ for all $x \in [-1, 1]$ and grows rapidly outside of this interval, it follows that $p(x)$ is large at $\tilde{\mu}_1$ and small at all the other eigenvalues. More precisely, the estimate

$$0 \leq \frac{\tilde{\mu}_i - \tilde{\mu}}{\tilde{\mu}_2 - \tilde{\mu}} \leq 1, \qquad i = 2, 3, \ldots$$

leads to

$$|p(\tilde{\mu}_i)| = \left|c_{k-1}\left(-1 + 2\frac{\tilde{\mu}_i - \tilde{\mu}}{\tilde{\mu}_2 - \tilde{\mu}}\right)\right| \leq 1, \quad i = 2, 3, \ldots,$$

and the identity

$$1 + 2\frac{\tilde{\mu}_1 - \tilde{\mu}_2}{\tilde{\mu}_2 - \tilde{\mu}} = -1 + 2\frac{\tilde{\mu}_1 - \tilde{\mu}}{\tilde{\mu}_2 - \tilde{\mu}}$$

yields

$$p(\tilde{\mu}_1) = c_{k-1}(1 + 2\rho_1).$$

Now using $1 = (\tilde{z}^0, \tilde{z}^0)_M = \sum_{i=1}^{\infty} (\tilde{z}^0, \tilde{\varphi}_i)^2$ by Parseval's equality, we can estimate the quotient of the inequality above by

$$\begin{aligned} & \frac{\sum_{i=2}^{\infty} (\tilde{z}^0, \tilde{\varphi}_i)_M^2 p(\tilde{\mu}_i)^2}{(\tilde{z}^0, \tilde{\varphi}_1)_M^2 p(\tilde{\mu}_1)^2 + \sum_{i=2}^{\infty} (\tilde{z}^0, \tilde{\varphi}_i)_M^2 p(\tilde{\mu}_i)^2} \\ \leq & \frac{(\tilde{z}^0, \tilde{\varphi}_1)_M^2 + \sum_{i=2}^{\infty} (\tilde{z}^0, \tilde{\varphi}_i)_M^2 p(\tilde{\mu}_i)^2 - (\tilde{z}^0, \tilde{\varphi}_1)_M^2}{(\tilde{z}^0, \tilde{\varphi}_1)_M^2 p(\tilde{\mu}_1)^2} \\ \leq & \frac{1 - (\tilde{z}^0, \tilde{\varphi}_1)_M^2}{(\tilde{z}^0, \tilde{\varphi}_1)_M^2} \frac{1}{(c_{k-1}(1 + 2\rho_1))^2} = \frac{(\tan \chi_1)^2}{(c_{k-1}(1 + 2\rho_1))^2}. \end{aligned}$$

□

Note that in practice the vector $\tilde{z}^0$ is determined by Algorithm 3.6 and depends on the given right hand side. In the case where $\tilde{z}^0 = \tilde{\varphi}_1$, Theorem 3.15 shows that the corresponding Ritz Value θ_1 is the exact largest eigenvalue of $\tilde{G}$. Due to Theorem 3.10 the CG-method terminates in this case after only one step. This corresponds to our discussion in Section 3.5. On the other hand, if ρ_1 is "small" we expect slow convergence and therefore approximations of low quality.

Chapter 4

Complexity analysis of a preconditioned Newton method

For large-scale problems in three space dimensions usually the complexity of an iterative numerical reconstruction method depends for the most part on the number of evaluations of the operators $F(x_n^\delta)$ and its Fréchet derivative $F'[x_n^\delta]$ and the adjoint $F'[x_n^\delta]^*$ at some given vectors. For example, for the IRGNM with inner CG-iteration the construction of the right hand side in the n-th Newton step involves the evaluation of $F(x_n^\delta)$. Furthermore, usually several evaluations of $F'[x_n^\delta]$ and its adjoint $F'[x_n^\delta]^*$ at given vectors generated in the CG-algorithm are required to compute an approximation to the solution of the linearized and regularized equation.

For the IRGNM with inner CG-iteration a complexity analysis has not been done so far. It is the goal of this chapter to contribute to such a complexity analysis both for mildly and for exponentially ill-posed problems. In particular, our aim is to give an upper bound for the total complexity of this algorithm in terms of the noise level $\delta > 0$. As measure for the complexity we will count the total number of operator evaluations. This corresponds to results formulated in [16, Chapter 7], where linear ill-posed problems are solved by the CG-method applied to the normal equation using the regularizing properties of this method.

Moreover, we suggest a preconditioned version of the IRGNM, that is we construct a preconditioner for the inner CG-iteration, which accelerates the speed of convergence leading to a significant reduction of the number of operator evaluations. Finally, we compare the complexity of the standard IRGNM to its preconditioned version.

4.1 Standard error estimate for the conjugate gradient method

Recall that we use the CG-method as an iterative method for solving the linear systems (2.8). In the following we want to establish convergence rates of this CG-iteration and determine the number of CG-steps required to reach a desired error level. We will find out that both topics are closely related. For a more detailed convergence rate analysis of the CG-method for linear operators in Hilbert spaces we refer to [4].

Throughout this chapter we denote the true solution of (2.8) by $h_n^\dagger$. Recall that the iterates of Algorithm 3.5 and 3.6 respectively are given by h_n^k, $k = 0, 1, 2, \ldots$. The iteration error in the k-th step of the CG-method we denote throughout this chapter by $e^k := h^\dagger - h_n^k$. The rate of convergence of the iteration error $\|e^k\|_{G_n^*G_n}$ is measured by the average convergence factor

$$\left(\frac{\|e^k\|_{G_n^*G_n}}{\|e^0\|_{G_n^*G_n}}\right)^{1/k},$$

where

$$\|x\|_{G_n^*G_n} := \sqrt{\langle x, G_n^*G_n x\rangle_{\mathcal{X}}}, \qquad x \in \mathcal{X},$$

defines the $G_n^*G_n$–norm on $\mathcal{X}$, which is actually a norm due to Lemma 3.3 and since $G_n^*G_n$ is bounded and strictly coercive for all $n \in \mathbb{N}_0$. To prove convergence rates we make use of the close connection between the k-th iteration error of the CG-method and certain polynomials. Recall that due to (3.22) the iteration error e^k of the CG-method, is related to the initial error by

$$e^k = p_k(M_n^{-1}G_n^*G_n)e^0, \tag{4.1}$$

where p_k belongs to Π_k^1, the set of polynomials of degree k such that $p_k(0) = 1$. Furthermore, since by Theorem 3.4 for any $v \in \mathcal{K}_k(M_n^{-1}G_n^*G_n, z^0)$ there exist polynomials $\tilde{q}_{k-1} \in \Pi_{k-1}$ and $\tilde{p}_k = 1 - t\tilde{q}(t) \in \Pi_k^1$ such that

$$e^0 + v = e^0 + \tilde{q}_{k-1}(M_n^{-1}G_n^*G_n)M_n^{-1}G_n^*G_n e^0 = \tilde{p}_k(M_n^{-1}G_n^*G_n)e^0,$$

an application of (3.24) shows that the polynomial defined by (4.1) is optimal in the sense that it satisfies

$$\|e^k\|_{G_n^*G_n} = \min_{p\in\Pi_k^1} \|p(M_n^{-1}G_n^*G_n)e^0\|_{G_n^*G_n}. \tag{4.2}$$

Equation (4.2) implies monotone convergence of $\|e^k\|_{G_n^*G_n}$. Moreover, it can be used to derive various upper bounds on the rate of convergence and for the number of iterations required to reach a desired relative error level.

To this end, let $\{\tilde{\varphi}_j : j \in \mathbb{N}\}$ be a set of orthonormal eigenvectors of $M_n^{-1}G_n^*G_n$ with corresponding eigenvalues $\tilde{\mu}_j$, $j \in \mathbb{N}$, with respect to $(\,\cdot\,,\,\cdot\,)_{M_n}$. Then, the identity

$$e^0 = \sum_{i=1}^{\infty} \left(e^0, \tilde{\varphi}_i\right)_{M_n} \tilde{\varphi}_i,$$

and (4.1) yields

$$e^k = \sum_{i=1}^{\infty} \left(e^0, \tilde{\varphi}_i\right)_{M_n} p_k(\tilde{\mu}_i)\tilde{\varphi}_i.$$

Using the nonnegativity of the eigenvalues, we find

$$\begin{aligned}
\|e^k\|_{G_n^*G_n} &= \left\langle \sum_{i=1}^{\infty} \left(e^0, \tilde{\varphi}_i\right)_{M_n} p_k(\tilde{\mu}_i)\tilde{\varphi}_i, M_n M_n^{-1} G_n^* G_n \sum_{i=1}^{\infty} \left(e^0, \tilde{\varphi}_i\right)_{M_n} p_k(\tilde{\mu}_i)\tilde{\varphi}_i \right\rangle_{\mathcal{X}}^{1/2} \\
&= \left(\sum_{i=1}^{\infty} \left(e^0, \tilde{\varphi}_i\right)_{M_n} p_k(\tilde{\mu}_i)\tilde{\varphi}_i, \sum_{i=1}^{\infty} \left(e^0, \tilde{\varphi}_i\right)_{M_n} \tilde{\mu}_i p_k(\tilde{\mu}_i)\tilde{\varphi}_i \right)_{M_n}^{1/2} \\
&= \left(\sum_{i=1}^{\infty} \tilde{\mu}_i (p_k(\tilde{\mu}_i))^2 \left| \left(e^0, \tilde{\varphi}_i\right)_{M_n} \right|^2 \right)^{1/2} \\
&\leq \max_{i\in\mathbb{N}} |p_k(\tilde{\mu}_i)| \left(\sum_{i=1}^{\infty} \tilde{\mu}_i \left| \left(e^0, \tilde{\varphi}_i\right)_{M_n} \right|^2 \right)^{1/2} \\
&= \max_{i\in\mathbb{N}} |p_k(\tilde{\mu}_i)| \|e^0\|_{G_n^*G_n}. \qquad (4.3)
\end{aligned}$$

Due to the minimization property (4.2), it follows from (4.3) that

$$\|e^k\|_{G_n^*G_n} \leq \min_{p\in\Pi_k^1} \max_{i\in\mathbb{N}} |p(\tilde{\mu}_i)| \|e^0\|_{G_n^*G_n}. \qquad (4.4)$$

This inequality reduces the problem to estimate the error of the CG-iteration to the construction of polynomials which make this bound small. Actually, the bound is sharp in the sense that there exists an initial guess for which the bound will be attained at every step (see Greenbaum [26]). The standard convergence rate of the CG-method, which we will prove in the following, can be derived from (4.4) using Chebysheff polynomials of the first kind defined by (3.40). In the following two lemmas we prove fundamental results on the Chebysheff polynomials that are important for our convergence rate analysis of the CG-method.

Lemma 4.1 *Let $0 < a < b$ and $\tilde{c}_j$ be the normalized Chebysheff polynomial of the first kind of degree j, more precisely*

$$\tilde{c}_j(t) := \frac{c_j(x(t))}{c_j(x(0))} \qquad \textit{where} \qquad x(t) := 1 - \frac{2(t-a)}{b-a}, \qquad (4.5)$$

and c_j is defined in (3.40). Then we have the estimate

$$\|\tilde{c}_j\|_{\infty,[a,b]} \leq \frac{1}{c_j(x(0))} \leq \frac{2\kappa(a,b)^j}{1+\kappa(a,b)^{2j}}, \tag{4.6}$$

where $\kappa(a,b)$ is defined by

$$\kappa(a,b) := \frac{\sqrt{b}-\sqrt{a}}{\sqrt{b}+\sqrt{a}}. \tag{4.7}$$

Proof: Note, the mapping x defined in (4.5) maps the interval $[a,b]$ to $[-1,1]$ and satisfies $x(0) = \frac{b+a}{b-a} > 1$. Now, rewriting c_j for $|x| > 1$ as

$$c_j(x) = \frac{1}{2}\left[\left(x+\sqrt{x^2-1}\right)^j + \left(x-\sqrt{x^2-1}\right)^j\right] = \frac{1+(x+\sqrt{x^2-1})^{2j}}{2(x+\sqrt{x^2-1})^j}$$

and using the fact that $\|c_j\|_{\infty,[-1,1]} \leq 1$ we get

$$\begin{aligned}\|\tilde{c}_j\|_{\infty,[a,b]} &\leq \frac{1}{c_j(x(0))} = \left(\frac{2\left(\frac{\sqrt{b}+\sqrt{a}}{\sqrt{b}-\sqrt{a}}\right)^j}{1+\left(\frac{\sqrt{b}+\sqrt{a}}{\sqrt{b}-\sqrt{a}}\right)^{2j}}\right)\frac{\left(\frac{\sqrt{b}-\sqrt{a}}{\sqrt{b}+\sqrt{a}}\right)^{2j}}{\left(\frac{\sqrt{b}-\sqrt{a}}{\sqrt{b}+\sqrt{a}}\right)^{2j}} \\ &= \frac{2\kappa(a,b)^j}{1+\kappa(a,b)^{2j}}.\end{aligned}$$

□

Notice that the normalized Chebysheff polynomials satisfy $\tilde{c}_j \in \Pi_j^1$. Hence, they can be used to derive upper bounds in inequality (4.4). The next lemma implies that they can be used to determine the number of CG-steps until some relative error level is reached.

Lemma 4.2 *Let $0 < a < b$ and $0 < \varepsilon \leq 1$. Then*

$$\frac{2\kappa(a,b)^j}{1+\kappa(a,b)^{2j}} \leq \varepsilon,$$

if $j \geq j(\varepsilon,a,b)$, where

$$j(\varepsilon,a,b) := \left\lceil \ln\left(\frac{1}{\varepsilon}+\sqrt{\frac{1}{\varepsilon^2}-1}\right) / \ln\left(\kappa(a,b)^{-1}\right)\right\rceil. \tag{4.8}$$

Here, $\lceil x \rceil$ denotes the smallest integer not less than x. In particular the inequality $\|\tilde{c}_j\|_{\infty,[a,b]} \leq \varepsilon$ is satisfied.

Proof: Notice that the function $f_\varepsilon(t) := \varepsilon^t/(1+\varepsilon^{2t})$, $0 < \varepsilon \leq 1$, $t \geq 0$, is monotonically decreasing, since for all $t \geq 0$ we have

$$f'_\varepsilon(t) = \frac{(\ln \varepsilon)\varepsilon^t(1-\varepsilon^{2t})}{(1+\varepsilon^{2t})^2} \leq 0.$$

From this together with $0 < \kappa(a,b) < 1$ it follows that

$$\begin{aligned}
\frac{2\kappa(a,b)^j}{1+\kappa(a,b)^{2j}} &\leq \frac{2\kappa(a,b)^{j(\varepsilon,a,b)}}{1+\kappa(a,b)^{2j(\varepsilon,a,b)}} \\
&\leq \frac{2\left(\frac{1}{\varepsilon}+\sqrt{\frac{1}{\varepsilon^2}-1}\right)}{1+\left(\frac{1}{\varepsilon}+\sqrt{\frac{1}{\varepsilon^2}-1}\right)^2} \\
&= \frac{2}{\left(\frac{1}{\varepsilon}-\sqrt{\frac{1}{\varepsilon^2}-1}\right)+\left(\frac{1}{\varepsilon}+\sqrt{\frac{1}{\varepsilon^2}-1}\right)} \\
&= \varepsilon.
\end{aligned}$$

Then the assertion $\|\tilde{c}_j\|_{\infty,[a,b]} \leq \varepsilon$ follows from (4.6).

□

We formulate the most well known convergence estimate for the CG-method in the following theorem.

Theorem 4.3 *The iteration error e^j of the preconditioned CG-method satisfies for all $j \in \mathbb{N}_0$*

$$\|e^j\|_{G_n^*G_n} \leq 2\left(\frac{\sqrt{\mathrm{cond}(M_n^{-1}G_n^*G_n)}-1}{\sqrt{\mathrm{cond}(M_n^{-1}G_n^*G_n)}+1}\right)^j \|e^0\|_{G_n^*G_n}. \tag{4.9}$$

Proof: Let us assume that $\mu^* := \max_{i\in\mathbb{N}} \tilde{\mu}_i$ and let $\mu_* := \inf_{i\in\mathbb{N}} \tilde{\mu}_i$. By Lemma 3.3 the operator $M_n^{-1}G_n^*G_n$ is strictly coercive, in particular there exits an $\varepsilon > 0$ such that $\mu_* \geq \varepsilon > 0$. By a combination of inequality (4.4) and Lemma 4.1 we have

$$\begin{aligned}
\|e^j\|_{G_n^*G_n} &\leq \max_{i\in\mathbb{N}} |\tilde{c}_j(\tilde{\mu}_i)| \|e^0\|_{G_n^*G_n} \leq 2\left(\frac{\sqrt{\mu^*}-\sqrt{\mu_*}}{\sqrt{\mu^*}+\sqrt{\mu_*}}\right)^j \|e^0\|_{G_n^*G_n} \\
&= 2\left(\frac{\sqrt{\frac{\mu^*}{\mu_*}}-1}{\sqrt{\frac{\mu^*}{\mu_*}}+1}\right)^j \|e^0\|_{G_n^*G_n} \\
&= 2\left(\frac{\sqrt{\mathrm{cond}(M_n^{-1}G_n^*G_n)}-1}{\sqrt{\mathrm{cond}(M_n^{-1}G_n^*G_n)}+1}\right)^j \|e^0\|_{G_n^*G_n}.
\end{aligned}$$

□

Estimate (4.9) implies fast convergence for small condition numbers of $M_n^{-1}G_n^*G_n$. On the other hand, if the condition number is large we expect slow convergence since

$$\frac{\sqrt{\text{cond}(M_n^{-1}G_n^*G_n)}-1}{\sqrt{\text{cond}(M_n^{-1}G_n^*G_n)}+1} \approx 1.$$

We give an interpretation of this result for our problem for the case $M_n = I$. Here the condition number is given by

$$\text{cond}(G_n^*G_n) = \frac{\gamma_n + \lambda_1}{\gamma_n} = 1 + \frac{\lambda_1}{\gamma_n}.$$

Obviously, the condition number explodes as $\gamma_n \searrow 0$. This indicates that solving a linear system (2.8) after several Newton steps requires a lot more CG-steps than in the first steps of the IRGNM. This behavior is supported by according observations in numerical examples. Usually, in applications where some general linear system is given, estimate (4.9) is the best one can expect. If no further information of the operator of the linear system is known, one can for example compute a good approximation to the condition number by the power method or Lanczos' method to get an idea of the expected convergence rate.
However, estimate (4.9) is not sufficient to predict about the convergence behavior of the CG-method, since not only the bounds of the spectrum of the operator play a role in the convergence behavior, but also the entire eigenvalue distribution. In our situation we can take advantage of this dependency, since we even have a quantitative a-priori knowledge of the distribution of the eigenvalues, since for ill-posed problems the rate of decay of the eigenvalues is of particular interest. It is even used to classify the problem. We will discuss this topic in detail in Section 4.5.

For the rest of this chapter we pursue five main goals:

a) We formulate two different stopping criteria for the CG-method while performing the IRGNM. The first criterion shows that it is possible to satisfy the inequalities (2.24) and (2.32). The second one we use in practice for our computations.

b) We construct a preconditioner converting the eigenvalue distribution of $G_n^*G_n$ such that $\text{cond}(M_n^{-1}G_n^*G_n) \approx 1$, that is we can expect fast convergence of the inner CG-iteration in each Newton step.

c) We suggest a model algorithm based on the IRGNM that involves a preconditioner in several Newton steps.

d) We derive an improvement of estimate (4.9). This will be done by choosing a polynomial which is tailored to the eigenvalue distribution of linear operators occurring in ill-posed problems.

e) We analyze the complexity of the IRGNM and the model algorithm and compare them. As measurement for the complexity we count the number of operator evaluations of $F(x)$ and of $F'[x]$ and $F'[x]^*$ at some given vectors.

4.2 Stopping criteria

We now want to show that it is possible to formulate a stopping criterion for the CG-method that guarantees that the conditions (2.24) resp. (2.32) are satisfied in each Newton step. This result shows that at least from a theoretical point of view the assumptions of Proposition 2.4 and of Theorem 2.7 concerning the error e_n^{ls} defined in (2.13) and its image under $F'[x_n^\delta]$ can be satisfied.

Theorem 4.4 *For Algorithm 3.5 with $C(\gamma_n, k) := \gamma_n f(\gamma_n)$ applied to (3.2), where f is determined by the source condition (2.4), the resulting stopping criterion to iterate until*

$$\|r^k\| \leq \varepsilon \gamma_n f(\gamma_n) \tag{4.10}$$

is met after a finite number J_n of steps, and the update $h_n^{\mathrm{app}} := h_n^{J_n}$ (cf. (2.9)) satisfies the estimates

$$\begin{aligned} \|h_n^\dagger - h_n^{J_n}\|_{\mathcal{X}} &\leq \varepsilon f(\gamma_n), \\ \|G_n(h_n^\dagger - h_n^{J_n})\|_{\mathcal{X}\times\mathcal{Y}} &\leq \varepsilon \sqrt{\gamma_n} f(\gamma_n), \end{aligned}$$

where $h_n^\dagger$ denotes the true solution of (3.2). In particular (2.24) and (2.32) hold.

Proof: It follows from standard convergence theory of the CG-method as presented in Section 4.1 that $\lim_{k\to\infty} h_n^k = h_n^\dagger$ and $\lim_{k\to\infty} r^k = 0$. So the stopping criterion (4.10) is met after a finite number J_n of steps.

Note that due to (4.10) Algorithm 3.5 terminates if $\|r^k\|_{\mathcal{X}} \leq \varepsilon\gamma_n f(\gamma_n)$. Recall from (3.3) that the norm of $G_n^* G_n$ is bounded by $\|(G_n^* G_n)^{-1}\| \leq \gamma_n^{-1}$. Hence, we conclude

$$\begin{aligned} \|h_n^\dagger - h_n^{J_n}\|_{\mathcal{X}} &\leq \|(G_n^* G_n)^{-1}\| \|G_n^* G_n (h_n^\dagger - h_n^{J_n})\|_{\mathcal{X}} \\ &\leq \gamma_n^{-1} \|G_n^*(g_n^\delta - G_n h_n^{J_n})\|_{\mathcal{X}} \\ &= \varepsilon f(\gamma_n). \end{aligned}$$

This proves (2.24a) with the constant $C_{\mathrm{ls}} = \varepsilon$. Since $(G_n^*)^\dagger G_n^*$ is the orthogonal projection onto $\overline{R(G_n)}$ and

$$\|G_n^\dagger\| = \|(G_n^* G_n)^{-1} G_n^*\| = \|(G_n^* G_n)^{-1} (G_n^* G_n)^{1/2}\| = \|(G_n^* G_n)^{-1/2}\| \leq \gamma_n^{-1/2},$$

we can estimate

$$\begin{aligned} \|G_n(h_n^\dagger - h_n^{J_n})\|_{\mathcal{X}\times\mathcal{Y}} &= \|(G_n^*)^\dagger G_n^* G_n (h_n^\dagger - h_n^{J_n})\|_{\mathcal{X}\times\mathcal{Y}} \\ &\leq \|(G_n^*)^\dagger\| \|G_n^* G_n (h_n^\dagger - h_n^{J_n})\|_{\mathcal{X}} \\ &\leq \varepsilon \sqrt{\gamma_n} f(\gamma_n). \end{aligned}$$

The estimate (2.24b) with the constant C_{ls} given by ε now follows from the inequality

$$\|F'[x_n^\delta]x\|_{\mathcal{Y}}^2 \leq \|G_n x\|_{\mathcal{X}\times\mathcal{Y}}^2,$$

which holds for all $x \in \mathcal{X}$. The estimates (2.32a) and (2.32b) follow by choosing $f(\gamma_n) = \sqrt{\gamma_n}$ (see the discussion before Theorem 2.7).

□

In particular, note that for the choice $C(\gamma_n, k) := \gamma_n^{3/2}$ in Algorithm 3.5 the resulting stopping criterion does not require knowledge of the source condition.
In practice we will use a related stopping criterion yielding some relative error estimate, which is essential for the complexity analysis.

Theorem 4.5 *Let M_n be a preconditioner for $G_n^* G_n$. For Algorithms 3.5 and 3.6 with $C(\gamma_n, k) := \gamma_n \|h_n^k\|_{\mathcal{X}}$ the resulting stopping criterion to iterate until*

$$\|r^k\|_{\mathcal{X}} \leq \varepsilon\gamma_n \|h_n^k\|_{\mathcal{X}}, \tag{4.11}$$

is met after a finite number J_n of steps and we have the relative error estimate

$$\|h_n^{J_n} - h_n^\dagger\|_{\mathcal{X}} \leq \left(\frac{\varepsilon}{1-\varepsilon}\right) \|h_n^\dagger\|_{\mathcal{X}}. \tag{4.12}$$

Proof: By the same arguments as in the proof of Theorem 4.4 the stopping index J_n is finite. Due to the choice of the stopping criterion (4.11) we have the estimate

$$\|r^{J_n}\|_{\mathcal{X}} \leq \varepsilon\gamma_n \|h_n^{J_n}\|_{\mathcal{X}},$$

which together with (3.3) yields

$$\|h_n^\dagger - h_n^{J_n}\|_{\mathcal{X}} \leq \|(G_n^* G_n)^{-1}\| \|G_n^* G_n (h_n^\dagger - h_n^{J_n})\|_{\mathcal{X}} \leq \varepsilon \|h_n^{J_n}\|_{\mathcal{X}}. \tag{4.13}$$

By an application of the triangle inequality we have

$$\|h_n^{J_n}\|_{\mathcal{X}} \leq \|h_n^\dagger\|_{\mathcal{X}} + \|h_n^{J_n} - h_n^\dagger\|_{\mathcal{X}} \leq \|h_n^\dagger\|_{\mathcal{X}} + \varepsilon \|h_n^{J_n}\|_{\mathcal{X}},$$

that is $(1-\varepsilon)\|h_n^{J_n}\|_{\mathcal{X}} \leq \|h_n^\dagger\|_{\mathcal{X}}$. This in combination with (4.13) proves (4.12).

□

4.3 Definition of a preconditioner

Designing an efficient preconditioner for a linear system is not an easy task in general. Some textbooks claim that *"Constructing preconditioners is more an art than a science"* (see for example [64] and [77]). We will discuss this topic in more detail in Chapter 5.

The following consideration serves as a motivation for the preconditioner defined below. Using the CG-method as solver for the linear systems (2.8) Lanczos' method provides good approximations to some of the eigenvalues and eigenvectors of the linear operator. Moreover, this method tends to approximate the outliers in the spectrum of the operator very well, while eigenvalues in the bulk of the spectrum are typically harder to approximate (see Kuijlaars [53]). For simplicity we assume in the following that we have exact knowledge of the $k_n \in \mathbb{N}$ largest eigenvalues $\mu_1 \geq \ldots \geq \mu_{k_n}$ with corresponding eigenvectors $\varphi_1, \ldots, \varphi_{k_n}$ of $G_n^* G_n$. We use this information to define a preconditioner $M_n \in L(\mathcal{X}, \mathcal{X})$ given by

$$M_n x := \gamma_n x + \sum_{j=1}^{k_n} \left(\frac{\mu_j}{\zeta} - \gamma_n \right) \langle x, \varphi_j \rangle_{\mathcal{X}} \, \varphi_j, \tag{4.14}$$

which is designed such that the known outliers $\mu_1 \geq \ldots \geq \mu_{k_n}$ are mapped to some value $\zeta > 0$. Note that we have assumed that $k_n \in \mathbb{N}$, that is at least one eigenpair is known. Naturally, in the case where we do not have any spectral information we define $M_n = I$. Hence, if nothing else is said we always assume in the following $k_n > 0$ when we use M_n. Recall that the eigenvalues μ_j of $G_n^* G_n$ satisfy $\mu_j = \gamma_n + \lambda_j$, $j \in \mathbb{N}$, and that the corresponding set $\{\varphi_j : j \in \mathbb{N}\}$ of orthonormal eigenvectors is a complete orthonormal system in $\mathcal{X}$ (see Section 3.1).

The preconditioner M_n belongs to the class of so-called spectral preconditioners (see Section 5.2). In general a lot of spectral data of $G_n^* G_n$ will be necessary for such a preconditioner to be efficient in the sense that it reduces the number of CG-steps until the stopping criterion is reached significantly. Moreover, in practice it is more realistic to assume that we only have approximations to the spectral data available. Still, the theoretical discussion in this chapter serves as motivation for the realization in Chapter 6. Let us summarize the main properties of M_n:

Theorem 4.6 *The operator M_n defined by (4.14) has the following properties:*

a) $M_n \in L(\mathcal{X}, \mathcal{X})$ is strictly coercive and self-adjoint with respect to $\langle \cdot \, , \cdot \rangle_{\mathcal{X}}$ and its inverse is given by

$$M_n^{-1} x = \frac{1}{\gamma_n} x + \sum_{j=1}^{k_n} \left(\frac{\zeta}{\mu_j} - \frac{1}{\gamma_n} \right) \langle x, \varphi_j \rangle_{\mathcal{X}} \, \varphi_j. \tag{4.15}$$

b) $M_n^{-1} G_n^* G_n = G_n^* G_n M_n^{-1}$.

c) The preconditioned operator is given by

$$M_n^{-1} G_n^* G_n x = \zeta x + \sum_{j=k_n+1}^{\infty} \left(\frac{\mu_j}{\gamma_n} - \zeta \right) \langle x, \varphi_j \rangle_{\mathcal{X}} \, \varphi_j. \tag{4.16}$$

d) For the spectrum of the preconditioned operator $M_n^{-1}G_n^*G_n$ holds

$$\sigma(M_n^{-1}G_n^*G_n) = \left\{1 + \frac{\lambda_s}{\gamma_n} : s > k_n\right\} \cup \{1, \zeta\}. \tag{4.17}$$

e) The preconditioned operator $M_n^{-1}G_n^*G_n$ is self-adjoint and strictly coercive with respect to $(\,\cdot\,,\,\cdot\,)_{M_n}$.

Proof: By symmetry of the inner product $\langle\,\cdot\,,\,\cdot\,\rangle_{\mathcal{X}}$ it follows that M_n is self-adjoint, and the calculation

$$\begin{aligned}
\langle M_n x, x\rangle_{\mathcal{X}} &= \langle \gamma_n x, x\rangle_{\mathcal{X}} + \sum_{j=1}^{k_n}\left(\frac{\mu_j}{\zeta} - \gamma_n\right) |\,\langle x, \varphi_j\rangle_{\mathcal{X}}\,|^2 \\
&= \sum_{j=1}^{k_n}\left(\frac{\mu_j}{\zeta}\right) |\,\langle x, \varphi_j\rangle\,|^2 + \gamma_n \sum_{j=k_n+1}^{\infty} |\,\langle x, \varphi_j\rangle\,|^2 \\
&\geq \min\left\{\frac{\mu_{k_n}}{\zeta}, \gamma_n\right\} \|x\|_{\mathcal{X}}^2,
\end{aligned}$$

shows that M_n is strictly coercive. Finally, the computations

$$M_n\varphi_s = \begin{cases} \frac{\mu_j}{\zeta}\varphi_s, & 1 \leq s \leq k_n, \\ \gamma_n\varphi_s, & s > k_n, \end{cases} \qquad M_n^{-1}\varphi_s = \begin{cases} \frac{\zeta}{\mu_j}\varphi_s, & 1 \leq s \leq k_n, \\ \gamma_n^{-1}\varphi_s, & s > k_n \end{cases} \tag{4.18}$$

prove a). b) follows by a straightforward computation and c) by

$$M_n^{-1}G_n^*G_n\varphi_s = \mu_s M_n^{-1}\varphi_s = \begin{cases} \zeta\varphi_s, & 1 \leq s \leq k_n, \\ \gamma_n^{-1}\mu_s\varphi_s, & s > k_n \end{cases}$$

and comparison with the right hand side in (4.16). d) and e) are consequences of c) and the definition of M_n.

□

Note that the preconditioner defined by (4.14) satisfies all requirements we postulated in Section 3.3.

- $M_n \in L(\mathcal{X}, \mathcal{X})$ and it is self-adjoint.
- The approximation property of M_n depends on the number of known eigenvalues with corresponding eigenvectors. Naturally, M_n coincides with the operator $G_n^*G_n$ on the subspace $\mathrm{span}\{\varphi_1, \ldots, \varphi_{k_n}\} \subset \mathcal{X}$. Since the eigenvalues of $G_n^*G_n$ decay rapidly, we expect that on the orthogonal complement $\mathrm{span}\{\varphi_1, \ldots, \varphi_{k_n}\}^{\perp}$ M_n is a reasonable approximation to $G_n^*G_n$.
- To store M_n it is sufficient to keep record of the eigenvalues and eigenvectors, which is acceptable if k_n is small.
- Since we know an explicit inverse, the system $M_n z = c$ is efficiently solvable.

4.4 A model algorithm

In this section we derive a *model algorithm* to solve the operator equation (2.1) by some kind of frozen IRGNM. When solving the operator equation (3.2) we can use Lanczos' method to compute approximations to some of the extremal eigenvalues of $G_n^* G_n$ with corresponding eigenvectors. Then we keep the operator fixed for a few steps of the IRGNM by defining $x_* := x_n^\delta$ and

$$G_{n,*} := \begin{pmatrix} F'[x_*] \\ \sqrt{\gamma_n} I \end{pmatrix} \in L(\mathcal{X}, \mathcal{Y} \times \mathcal{X}). \tag{4.19}$$

Using the approximations to the extremal eigenvalues and corresponding eigenvectors determined by Lanczos' method we set up the preconditioner (4.14). Instead of (2.8) we now solve in several following Newton steps the linear systems

$$M_n^{-1} G_{n,*}^* G_{n,*} h_n = M_n^{-1} G_{n,*} g_n^\delta \tag{4.20}$$

by Algorithm 3.6. A Newton method of this kind was suggested by Hohage [40], and numerical examples illustrated a significant reduction of the total complexity, whereas the final iterates were comparable with those of a standard IRGNM. Unfortunately, no theoretical analysis of this algorithm has been performed. In the following we will carry out such an analysis under the following simplifying assumption.

Assumption 4.7 *We assume that for all $n \in \mathbb{N}_0$ there exists a method L to compute the exact $k_n \in \mathbb{N}_0$ largest eigenvalues $\mu_1 \geq \ldots \geq \mu_{k_n}$ of the operator $G_n^* G_n$ and corresponding eigenvectors $\varphi_1, \ldots, \varphi_{k_n}$ using k_n applications of $F'[x_n^\delta]$ and $F'[x_n^\delta]^*$.*

Remark 4.8 *Under the additional assumption that the largest eigenvalues of $G_n^* G_n$ satisfy $\mu_1 > \ldots > \mu_{k_n}$, that is they are simple, and the right hand side of (3.2) has a representation $\sum_{j=1}^{k_n} \alpha_j \varphi_j$ such that $\alpha_j \neq 0$, $j = 1, \ldots, k_n$, indeed the CG-method stops with the exact solution of (3.2) after exactly k_n steps and Lanczos' method determines the exact k_n largest eigenvalues of $G_n^* G_n$ with corresponding eigenvectors.*

Definition 4.9 *By f_{up} we denote an update criterion determining if the operator $G_{n,*}$ needs to be updated, which is equal to one if and only if the operator $G_{n,*}$ needs to be updated. The criterion f_{up} depends on the Newton step, on the right hand side g_n^δ and the operator F .*

We now formulate a *model algorithm* for a preconditioned IRGNM.

Algorithm 4.10

Compute $F(x_0)$;

$n = 0$;

while $\left(\left\|F(x_n^\delta) - y^\delta\right\| \geq \tau\delta\right)$

if $\left(f_{\text{up}}(n, g_n^\delta, F) = 1\right)$

$x_* := x_n^\delta$;

Solve $G_n h_n = g_n^\delta$ by Algorithm 3.5;

Compute eigenvalues $\mu_1 \geq \ldots \geq \mu_{k_n}$ with corresponding eigenvectors $\varphi_1, \ldots, \varphi_{k_n}$ of $G_n^* G_n$ by method L;

Set up preconditioner (4.14);

else

Solve $M_n^{-1} G_{n,*}^* G_{n,*} h_n = M_n^{-1} G_{n,*} g_n^\delta$ by Algorithm 3.6;

$x_{n+1}^\delta = x_n^\delta + h_n$;

$n = n + 1$;

Compute $F(x_n^\delta)$;

Note that the update criterion f_{up} has to balance the convergence speed in the outer Newton iteration with the complexity needed in the inner CG-iterations. Hence, the criterion f_{up} has to satisfy some optimization problem, which maybe depends on more variables than the variables we considered here. In [40], where the algorithm above was originally presented, it was suggested to update the operator whenever $\sqrt{n+1} \in \mathbb{N}$. For the examples given in this article this criterion worked efficiently. We will discuss this topic in more detail in Section 6.3.
Finally note that if no method L is available we set $f_{\text{up}} \equiv 0$ and $M_n = I$. Then Algorithm 4.10 simplifies to the standard IRGNM.

4.5 Uniform estimates on the number of inner iterations

As a first step to determine the total complexity of the IRGNM and Algorithm 4.10 we now establish an improvement of the standard estimate (4.9) for the CG-method. This estimate is based on the special decay behavior of the eigenvalues for ill-posed problems.
To simplify the notation we assume for the following that the eigenvalues λ_j, $j \in \mathbb{N}$, of the compact operator $F'[x_n^\delta]^* F'[x_n^\delta]$ are simple. We will discuss in Remark 4.18 the case of multiple eigenvalues. In the following we consider two different types of ill-posed problems (see [16, Chapter 2]).

i) For the first kind of ill-posed problems we assume that the eigenvalues of $F'[x_n^\delta]^* F'[x_n^\delta]$ decay at a polynomial rate, that is there exist constants

$c_p, C_p > 0$ such that

$$c_p j^{-\alpha} \leq \lambda_j \leq C_p j^{-\alpha}, \qquad \alpha > 0. \tag{4.21a}$$

An ill-posed problem satisfying (4.21a) belongs to the class of so called mildly ill-posed problems.

ii) For the second kind of ill-posed problems we assume that the eigenvalues of $F'[x_n^\delta]^* F'[x_n^\delta]$, decay at an exponential rate, that is there exist constants $c_e, C_e > 0$ such that

$$c_e \exp(-c_s j^\beta)) \leq \lambda_j \leq C_e \exp(-c_s j^\beta)), \qquad c_s, \beta > 0. \tag{4.21b}$$

An ill-posed problem satisfying (4.21b) is also called exponentially ill-posed and belongs to the class of so-called severely ill-posed problems.

To take advantage of this special eigenvalue distribution we start by giving a precise mathematical definition of what we understand by outliers in the spectrum of $G_n^* G_n$ and $M_n^{-1} G_n^* G_n$ and the cluster of eigenvalues.

Definition 4.11 *Throughout the rest of this chapter we denote by $m > 1$ some threshold parameter.*

Definition 4.12 *By $q_n \in \mathbb{N}_0$ we denote the number of "large" eigenvalues of the operator $G_n^* G_n$ defined by*

$$\mu_{q_n} > m\gamma_n \geq \mu_{q_n+1}. \tag{4.22}$$

We call the eigenvalues $\mu_1 \geq \ldots \geq \mu_{q_n}$ the outliers *in the spectrum of $G_n^* G_n$. The set of eigenvalues*

$$\sigma(G_n^* G_n) \setminus \{\mu_1, \ldots, \mu_{q_n}\},$$

which are in a neighborhood of the regularization parameter γ_n is called the cluster *of $G_n^* G_n$. In the case where M_n is given by (4.14) we substitute q_n by $q_n + k_n$. The number of "large" eigenvalues of the operator $M_n^{-1} G_n^* G_n$ is defined by (cf. (4.17))*

$$1 + \frac{\lambda_{q_n}}{\gamma_n} > m \geq 1 + \frac{\lambda_{q_n+1}}{\gamma_n}. \tag{4.23}$$

Analogous, the eigenvalues of $M_n^{-1} G_n^* G_n$ larger than the threshold parameter m are called the outliers *in the spectrum of $M_n^{-1} G_n^* G_n$ and set of eigenvalues*

$$\{\tilde{\mu} \in \sigma(M_n^{-1} G_n^* G_n) : m \geq \tilde{\mu}\}$$

the cluster.

Note that it depends on the size of the eigenvalue $\zeta > 0$ if it belongs to the outliers or to the cluster of $M_n^{-1}G_n^*G_n$. As we will see, this choice has some effect on the convergence property of the CG-method.
We now prove an improvement of estimate (4.9) taking advantage of the a-priori knowledge about the eigenvalue distribution (4.21). By virtue of inequality (4.4) we consider polynomials that treat the outliers and the cluster of the operators $G_n^*G_n$ and $M_n^{-1}G_n^*G_n$ in a special way.

Theorem 4.13 *We use the notation of Definition 4.12. Let $M_n = I$ or M_n be given by (4.14) and assume that ζ is in the cluster of $M_n^{-1}G_n^*G_n$, that is $\zeta \in [1, m]$. For both cases the estimate*

$$\|e^{q_n+j}\|_{G_n^*G_n} \leq \frac{2\kappa(1,m)^j}{1+\kappa(1,m)^{2j}}\|e^0\|_{G_n^*G_n}, \tag{4.24}$$

holds true, where κ is given by (4.7). Hence, $\|e^{q_n+j}\|_{G_n^*G_n}/\|e^0\|_{G_n^*G_n} \leq \varepsilon$ for some $0 < \varepsilon \leq 1$ holds, if

$$j = \left\lceil \ln\left(\frac{1}{\varepsilon} + \sqrt{\frac{1}{\varepsilon^2} - 1}\right) / \ln(\kappa(1,m)^{-1}) \right\rceil .$$

Proof: Consider first the case $M_n = I$. We define the polynomial $\Psi_{q_n+j} \in \Pi_{q_n+j}$ by

$$\Psi_{q_n+j}(t) := \frac{c_j(x(t))}{c_j(x(0))}\Pi_{s=1}^{q_n}\left(1 - \frac{t}{\mu_s}\right) \qquad \text{where} \qquad x(t) := 1 - \frac{2(t-\gamma_n)}{(m-1)\gamma_n},$$

(cf. (4.5)) where the eigenvalues $\mu_1, \ldots, \mu_{q_n}$ are determined by (4.22). Then the polynomial Ψ_{q_n+j} satisfies

$$\begin{aligned} \Psi_{q_n+j}(0) &= 1, \\ \Psi_{q_n+j}(\mu_k) &= 0, \qquad k = 1, \ldots, q_n, \\ |\Psi_{q_n+j}(\mu_k)| &\leq \left|\frac{c_j(x(\mu_k))}{c_j(x(0))}\right|, \qquad k > q_n. \end{aligned}$$

Hence, $\Psi_{q_n+j} \in \Pi_{q_n+j}^1$ can be used to derive an upper bound in inequality (4.4). Now Lemma 4.1 together with the definition of κ yields

$$\begin{aligned} \|e^{q_n+j}\|_{G_n^*G_n} &\leq \max_{k>q_n}\left[\left|\frac{c_j(x(\mu_k))}{c_j(x(0))}\right| \Pi_{s=1}^{q_n}\left(1 - \frac{\mu_k}{\mu_s}\right)\right] \|e^0\|_{G_n^*G_n} \\ &\leq \max_{t\in[\gamma_n, m\gamma_n]}\left|\frac{c_j(x(t))}{c_j(x(0))}\right| \|e^0\|_{G_n^*G_n} \\ &\leq \left(\frac{2\kappa(\gamma_n, m\gamma_n)^j}{1+\kappa(\gamma_n, m\gamma_n)^{2j}}\right)\|e^0\|_{G_n^*G_n} \\ &= \left(\frac{2\kappa(1,m)^j}{1+\kappa(1,m)^{2j}}\right)\|e^0\|_{G_n^*G_n}. \end{aligned}$$

In the case where M_n is given by (4.14) we replace the polynomial Ψ_{q_n+j} using the substitution $q_n \mapsto q_n + k_n$ by

$$\tilde{\Psi}_{q_n+j}(t) := \left(\frac{c_j(x(t))}{c_j(x(0))}\right) \Pi_{k=k_n+1}^{q_n} \left(1 - \frac{t}{1+\frac{\lambda_k}{\gamma_n}}\right), \tag{4.25}$$

where $x(t) := 1 - \frac{2(t-1)}{m-1}$. Then $\tilde{\Psi}_{q_n+j} \in \Pi^1_{q_n+j}$ and using (4.23) and $\zeta \in [1,m]$ we have

$$\begin{aligned}
\tilde{\Psi}_{q_n+j}(0) &= 1, \\
|\tilde{\Psi}_{q_n+j}(\zeta)| &\leq 1, \\
\tilde{\Psi}_{q_n+j}\left(1+\frac{\lambda_k}{\gamma_n}\right) &= 0, \qquad k = k_n+1, \ldots, q_n, \\
\left|\tilde{\Psi}_{q_n+j}\left(1+\frac{\lambda_k}{\gamma_n}\right)\right| &\leq \left|\frac{c_j\left(x\left(1+\frac{\lambda_k}{\gamma_n}\right)\right)}{c_j(x(0))}\right|, \qquad k > q_n.
\end{aligned}$$

Estimating as above the assertion follows. The proof on the number j follows along the lines of Lemma 4.2.

□

Another approach to improve the standard estimate (4.9) is given by the so-called "effective" condition number. Here the interplay between the approximation properties of the Ritz values computed by Lanczos' method and the polynomials from Theorem 3.10 c) is exploited to show that the convergence speed of the CG-method increases if some Ritz value starts to converge to some extremal eigenvalue (see van der Sluis & van der Vorst [83]).

Remark 4.14 *In the case where $\zeta \notin [1,m]$ the slightly worse estimate*

$$\|e^{q_n+j+1}\|_{G_n^* G_n} \leq \frac{2\kappa(1,m)^j}{1+\kappa(1,m)^{2j}} \|e^0\|_{G_n^* G_n},$$

can be proven by considering the polynomials $\tilde{\Psi}_{q_n+j+1} \in \Pi^1_{q_n+j+1}$ defined through

$$\tilde{\Psi}_{q_n+j+1}(t) := \left(\frac{c_j(x(t))}{c_j(x(0))}\right) \left(\frac{\zeta - t}{\zeta}\right) \Pi_{k=k_n+1}^{q_n} \left(1 - \frac{t}{1+\frac{\lambda_k}{\gamma_n}}\right).$$

where $x(t) := 1 - \frac{2(t-1)}{m-1}$. Since $\zeta \notin [1,m]$ the difficulty arises to ensure that

$$\left|\tilde{\Psi}_{q_n+j+1}\left(1+\frac{\lambda_k}{\gamma_n}\right)\right| \leq 1 \qquad \textit{for all} \qquad k > q_n,$$

yielding to the condition $|\zeta - (1 + \lambda_k/\gamma_n)| \leq \zeta$ *for all* $k > q_n$. *To this end one has to impose the artificial condition*

$$\zeta \geq \frac{1}{2}\left(1 + \frac{\lambda_{q_n+1}}{\gamma_n}\right). \tag{4.26}$$

Usually the eigenvalue λ_{q_n+1} *is unknown. Hence, it is unrealistic to determine* ζ *such that (4.26) is satisfied. Moreover, if only approximations to the spectral data are available the operator* $M_n^{-1}G_n^*G_n$ *does not have the eigenvalue* ζ *but many eigenvalues in a neighborhood of* ζ. *This usually leads to a larger number of CG-steps required to reach a desired error level compared with the preconditioner where* $\zeta \in [1, m]$.

So, for the following we restrict ourselves to the case $\zeta \in [1, m]$. This assumption also theoretically reduces the number of CG-steps by one. Recall that the residuals r^j, $j = 0, 1, 2, \ldots$ of the CG-method satisfy $r^j = G_n^*G_ne^j$ (see Section 3.2). The next corollary shows the correspondence between the error in the residual and e^0.

Corollary 4.15 *We use the notation of Definition 4.12. The estimate*

$$\|r^{q_n+j}\|_{\mathcal{X}} \leq \frac{2\kappa(1,m)^j}{1+\kappa(1,m)^{2j}}\sqrt{\mu_1}\|e^0\|_{G_n^*G_n}. \tag{4.27}$$

holds true.

Proof: The equality $r^j = G_n^*G_ne^j$, $j = 0, 1, 2, \ldots$ together with the estimate

$$\langle G_n^*G_nx, G_n^*G_nx\rangle_{\mathcal{X}} = \sum_{j=1}^{\infty}\mu_j\left\langle\langle x, \varphi_j\rangle_{\mathcal{X}}\varphi_j, G_n^*G_nx\right\rangle_{\mathcal{X}} \leq \mu_1\langle x, G_n^*G_nx\rangle_{\mathcal{X}},$$

which is true for all $x \in \mathcal{X}$, and (4.24) yields

$$\begin{aligned}
\|r^{q_n+j}\|_{\mathcal{X}}^2 &= \left\langle G_n^*G_ne^{q_n+j}, G_n^*G_ne^{q_n+j}\right\rangle_{\mathcal{X}} \\
&\leq \mu_1\left\langle e^{q_n+j}, G_n^*G_ne^{q_n+j}\right\rangle_{\mathcal{X}} \\
&= \mu_1\|e^{q_n+j}\|_{G_n^*G_n}^2 \\
&\leq \mu_1\left(\frac{2\kappa(1,m)^j}{1+\kappa(1,m)^{2j}}\right)^2\|e^0\|_{G_n^*G_n}^2.
\end{aligned}$$

This holds true in both cases $M_n = I$ and M_n given by (4.14) by the definition of q_n.

□

After having established convergence rates, we now want to focus on the complexity when solving (3.2) by the CG-method. Measuring the complexity by evaluations of the operators $F'[x_n^\delta]$ and $F'[x_n^\delta]^*$ at some given vectors, we need to know the number of steps until Algorithm 3.5 terminates. The following lemma gives an upper bound for the CG-steps corresponding to the outliers in the spectrum.

Lemma 4.16 *We use the notation of Definition 4.12. Assume that the regularization parameter is chosen by* $\gamma_n = \gamma_0\gamma^{-n}$ *and that* $M_n = I$ *or* M_n *is given by (4.14).*

a) If the decay rate of the eigenvalues of $G_n^*G_n$ *is given by (4.21a), then for all* $n \in \mathbb{N}_0$ *the following estimate holds:*

$$q_n \leq \left\lceil \left(\frac{C_p\gamma^n}{(m-1)\gamma_0} \right)^{1/\alpha} \right\rceil. \tag{4.28}$$

b) If the decay rate of the eigenvalues of $G_n^*G_n$ *is given by (4.21b), then for all* $n \in \mathbb{N}$ *the estimate*

$$q_n \leq \left\lceil \left(\frac{2}{c_s} \max\left\{ \ln\gamma, \left| \ln\left(\frac{(m-1)\gamma_0}{C_e} \right) \right| \right\} \right)^{1/\beta} n^{1/\beta} \right\rceil \tag{4.29}$$

holds true. For the case $n = 0$ *we assume that* m *and* γ_0 *are sufficiently large such that* $(m-1)\gamma_0 \geq C_e$*. Then the estimate*

$$q_0 \leq \left\lceil \left(\frac{1}{c_s} \ln\left(\frac{(m-1)\gamma_0}{C_e} \right) \right)^{1/\beta} \right\rceil \tag{4.30}$$

is satisfied.

Proof: To prove the assertions first note that (4.23) is equivalent to

$$\mu_{q_n} > m\gamma_n \geq \mu_{q_n+1}. \tag{4.31}$$

Hence, we can prove both cases $M_n = I$ and M_n given by (4.14) at once, since (4.31) is equivalent to (4.22) for the special case $k_n = 0$.

Using the relation $\mu_j = \gamma_n + \lambda_j$, $j \in \mathbb{N}$, and (4.21a) together with (4.31) we have the estimate

$$(m-1)\gamma_n < \lambda_{q_n} \leq C_p q_n^{-\alpha}, \tag{4.32}$$

which yields for the choice $\gamma_n = \gamma_0\gamma^{-n}$

$$q_n^\alpha < \frac{C_p\gamma^n}{(m-1)\gamma_0}.$$

Now (4.28) follows since $x \leq \left(\lceil x \rceil^{1/\alpha}\right)^\alpha$ for all $x \geq 0$.

To show (4.29) we use the following inequality: Let $a > 0$, $b \in \mathbb{R}$ and $n \in \mathbb{N}$. Then

$$-2n\max\{a, |b|\} \leq -na + b, \tag{4.33}$$

which is a consequence of the estimate

$$-2n\max\{a,|b|\} \leq -n(a+|b|) = -na - n|b| \leq -na + b.$$

Using the inequalities (4.33) and (4.32) together with (4.21b) and the choice $\gamma_n = \gamma_0\gamma^{-n}$ we can estimate for $n \in \mathbb{N}$

$$\begin{aligned}
& C_e \exp\left(-c_s\left(\left\lceil\left(\frac{2}{c_s}\max\left\{\ln\gamma, \left|\ln\left(\frac{(m-1)\gamma_0}{C_e}\right)\right|\right\}\right)^{1/\beta} n^{1/\beta}\right\rceil\right)^{\beta}\right) \\
\leq\ & C_e \exp\left(-2n\max\left\{\ln\gamma, \left|\ln\left(\frac{(m-1)\gamma_0}{C_e}\right)\right|\right\}\right) \\
\leq\ & C_e \exp\left(-n\ln\gamma + \ln\left(\frac{(m-1)\gamma_0}{C_e}\right)\right) \\
=\ & C_e \exp\left(\ln\left(\frac{(m-1)\gamma_n}{C_e}\right)\right) \\
=\ & (m-1)\gamma_n \\
<\ & \lambda_{q_n},
\end{aligned}$$

which proves (4.29). With the additional assumptions on m and γ_0 (4.30) follows by a similar computation, since

$$\frac{(m-1)\gamma_0}{C_e} \geq 1.$$

□

After having established upper bounds for the number of CG-steps for the outliers in the spectrum, we are now concerned with the number of steps concerning the cluster.

Theorem 4.17 *We use the notation of Definition 4.12. Assume that* $0 < \varepsilon < 1/2$. *If the Algorithms 3.5 and 3.6 are stopped by criterion (4.11), then the number of CG-steps is bounded by*

$$J_n \leq q_n + j\left(\tilde{\varepsilon}, 1, m\right), \tag{4.34}$$

where the function j *is given by (4.8) and*

$$\tilde{\varepsilon} := \frac{\gamma_n}{\mu_1}\left(\frac{1-2\varepsilon}{1-\varepsilon}\right)\varepsilon. \tag{4.35}$$

Proof: By Theorem 4.5 the number J_n is finite. To prove the assertion we decompose J_n into $J_n = q_n + j_n$.

To show (4.34) we start with estimating the residual. Using (4.27), $h_n^0 = 0$ and Lemma 4.1 we have for all $j \in \mathbb{N}_0$

$$\begin{aligned} \|r^{q_n+j}\|_{\mathcal{X}} &\leq \sqrt{\mu_1}\left(\frac{2\kappa(1,m)^j}{1+\kappa(1,m)^{2j}}\right)\|h_n^\dagger\|_{G_n^*G_n} \\ &\leq \mu_1\left(\frac{2\kappa(1,m)^j}{1+\kappa(1,m)^{2j}}\right)\|h_n^\dagger\|_{\mathcal{X}}. \end{aligned} \tag{4.36}$$

Due to (4.12) and an application of the second triangle inequality we have for the final iterate

$$\left|1 - \frac{\|h_n^{J_n}\|_{\mathcal{X}}}{\|h_n^\dagger\|_{\mathcal{X}}}\right| \leq \frac{\varepsilon}{1-\varepsilon}.$$

This estimate in particular implies

$$\frac{1-2\varepsilon}{1-\varepsilon} \leq \frac{\|h_n^{J_n}\|_{\mathcal{X}}}{\|h_n^\dagger\|_{\mathcal{X}}}. \tag{4.37}$$

Note that by the choice of ε we have $0 < \tilde{\varepsilon} < 1$. Now, for the choice $j = j(\tilde{\varepsilon}, 1, m)$ we can estimate using (4.36), (4.37) and Lemma 4.2

$$\|r^{q_n+j(\tilde{\varepsilon},1,m)}\|_{\mathcal{X}} \leq \mu_1\left(\frac{1-\varepsilon}{1-2\varepsilon}\right)\left(\frac{2\kappa(1,m)^{j(\tilde{\varepsilon},1,m)}}{1+\kappa(1,m)^{2j(\tilde{\varepsilon},1,m)}}\right)\|h_n^{J_n}\|_{\mathcal{X}} \leq \varepsilon\gamma_n\|h_n^{J_n}\|_{\mathcal{X}}.$$

Hence for this choice of j the stopping criterion (4.11) is satisfied. This proves the estimate (4.34).

□

Remark 4.18 *Actually, (4.34) also holds true if the linear operators $G_n^*G_n$ and $M_n^{-1}G_n^*G_n$ have multiple eigenvalues. To show this, first note that the proof of (4.24) is based on estimate (4.4), which is true for multiple eigenvalues. Hence, estimate (4.24) is also true in this case.*
Lemma 4.16 in the presence of multiple eigenvalues is only true in the case $M_n = I$. In this case we must define the number q_n in Definition 4.12 by the number of different eigenvalues of $G_n^*G_n$. Since the CG-method performs at most one step for a multiple eigenvalue (see Theorem 3.10 and [12]), it is clear that even in the presence of multiple eigenvalues the maximal number of CG-steps for the outliers in the spectrum is given by q_n. But the number q_n of different outliers is bounded by the bounds proven in Lemma 4.16.
In the case of the preconditioned operator $M_n^{-1}G_n^*G_n$ Lemma 4.16 is only true if the eigenvalues $\mu_1, \ldots, \mu_{k_n}$ are simple.

We now state the main result of this section establishing an upper bound on the total number of inner CG-steps of the IRGNM and the Levenberg-Marquardt algorithm, respectively.

Theorem 4.19 *We use the notation of Definition 4.12. Let the regularization parameter be chosen by* $\gamma_n = \gamma_0 \gamma^{-n}$ *and* $0 < \varepsilon < 1/2$ *and assume that in the case where* M_n *is given by (4.14) the eigenvalues* $\mu_1, \ldots, \mu_{k_n}$ *are simple.*

a) Let (4.21a) hold. In the case where $M_n = I$ *there exists a constant* $C \geq 0$ *independent of* n *such that the total number* J_n *of steps to reach the stopping criterion (4.11) is bounded by*

$$J_n \leq \left\lceil \gamma^{n/\alpha} \right\rceil + \lceil Cn \rceil, \qquad n \in \mathbb{N}_0. \tag{4.38}$$

In the case where M_n *is given by (4.14) the total number* J_n *of steps to reach the stopping criterion (4.11) is bounded by*

$$J_n \leq \left(\left\lceil \gamma^{n/\alpha} \right\rceil - k_n \right)_+ + \lceil Cn \rceil, \qquad n \in \mathbb{N}_0. \tag{4.39}$$

b) Let (4.21b) hold. In the case where $M_n = I$ *there exist constants* $C_1, C \geq 0$ *independent of* n *such that the total number* J_n *of steps to reach the stopping criterion (4.11) is bounded by* $J_0 \leq \lceil C \rceil$ *and*

$$J_n \leq \left\lceil C_1 n^{1/\beta} \right\rceil + \lceil Cn \rceil, \qquad n \in \mathbb{N}. \tag{4.40}$$

In the case where M_n *is given by (4.14) the total number* J_n *of steps to reach the stopping criterion (4.11) is bounded by*

$$J_n \leq \left(\left\lceil C_1 n^{1/\beta} \right\rceil - k_n \right)_+ + \lceil Cn \rceil, \qquad n \in \mathbb{N}. \tag{4.41}$$

Here $(x)_+ = x$ *if* $x \geq 0$ *and* $(x)_+ = 0$ *if* $x < 0$.

Proof: From (4.34) it is obvious that we need an estimate for q_n and $j(\tilde{\varepsilon}, 1, m)$, where $\tilde{\varepsilon}$ is given by (4.35). Consider first the case $M_n = I$, hence $k_n = 0$. In the case of (4.21a) we have due to (4.28) and the choice $m = 1 + \frac{C_p}{\gamma_0}$

$$q_n \leq \left\lceil \left(\frac{C_p}{(m-1)\gamma_n} \right)^{1/\alpha} \right\rceil = \left\lceil \gamma^{n/\alpha} \right\rceil,$$

which shows the first term on the right hand side of (4.38). By the choice $m = 1 + \frac{C_e}{\gamma_0}$ in the case of (4.21b) the estimate (4.30) simplifies to $q_0 = 0$ and (4.29) simplifies for all $n \geq 1$ to

$$q_n \leq \left\lceil \left(\frac{2}{c_s} \max \left\{ \ln \gamma, \left| \ln \left(\frac{(m-1)\gamma_0}{C_e} \right) \right| \right\} \right)^{1/\beta} n^{1/\beta} \right\rceil = \left\lceil \left(2 \frac{\ln \gamma}{c_s} \right)^{1/\beta} n^{1/\beta} \right\rceil.$$

This proves the first term on the right hand side of (4.40) with the constant $C_1 = (2 \ln \gamma / c_s)^{1/\beta}$. If M_n is given by (4.14) the first terms on the right hand side of (4.39) and (4.41) follow by the substitution $q_n \mapsto q_n + k_n$ (see Definition 4.12).

Note, to show the second term on the right hand sides of (4.38) – (4.41) we need to estimate $j(\tilde{\varepsilon}, 1, m)$. First recall that $0 < \tilde{\varepsilon} < 1$ (see 4.35) for all $n \in \mathbb{N} \cup \{0\}$ and for our choice of $\varepsilon > 0$. Then, by (4.8) we have for all $n \in \mathbb{N}$

$$\begin{aligned} j(\tilde{\varepsilon}, 1, m) &= \left\lceil \ln\left(\frac{1}{\tilde{\varepsilon}} + \sqrt{\frac{1}{\tilde{\varepsilon}^2} - 1}\right) / \ln\left(\kappa(1,m)^{-1}\right) \right\rceil \\ &\leq \left\lceil \ln\left(\frac{2}{\tilde{\varepsilon}}\right) / \ln\left(\kappa(1,m)^{-1}\right) \right\rceil \\ &= \left\lceil \ln\left(\frac{2}{\varepsilon}\left(\frac{1-\varepsilon}{1-2\varepsilon}\right)\frac{(\gamma_n + \lambda_1)}{\gamma_n}\right) / \ln\left(\kappa(1,m)^{-1}\right) \right\rceil \\ &\leq \left\lceil \left(\ln\left(\frac{2}{\varepsilon}\left(\frac{1-\varepsilon}{1-2\varepsilon}\right)\frac{(\gamma_0 + \lambda_1)}{\gamma_0}\right) + n\ln\gamma\right) / \ln\left(\kappa(1,m)^{-1}\right) \right\rceil \\ &\leq \lceil Cn \rceil, \end{aligned}$$

with the constant

$$C = \left(\ln\left(\frac{2}{\varepsilon}\left(\frac{1-\varepsilon}{1-2\varepsilon}\right)\frac{(\gamma_0 + \lambda_1)}{\gamma_0}\right) + \ln\gamma\right) / \ln\left(\kappa(1,m)^{-1}\right),$$

where

$$\kappa(1,m)^{-1} = \frac{\sqrt{1 + \frac{C_e}{\gamma_0}} + 1}{\sqrt{1 + \frac{C_e}{\gamma_0}} - 1} \qquad \text{or} \qquad \kappa(1,m)^{-1} = \frac{\sqrt{1 + \frac{C_p}{\gamma_0}} + 1}{\sqrt{1 + \frac{C_p}{\gamma_0}} - 1},$$

resp. if (4.21a) or (4.21b) hold. In both cases $\kappa(1,m)$ is independent of n. In the case where $n = 0$ we simply have the estimate $j(\tilde{\varepsilon}, 1, m) \leq C$.

□

4.6 The total complexity

The efficiency of the IRGNM for large-scale problems depends on its total number of inner CG-steps until the outer iteration is stopped by some stopping criterion. In Chapter 2 we investigated an a-priori stopping criterion given by (2.5a) and the discrepancy principle (2.5b). Moreover, in Corollary 2.5 it was shown that for a known noise level $\delta > 0$ for both stopping criteria the stopping index $N = N(\delta, y^\delta)$ for the IRGNM is finite and satisfies

$$N = O(-\ln(u^{-1}(\delta))) \tag{4.42}$$

where the function u is given by (1.11). Since the essential cost to perform one step of the IRGNM consists in iteratively solving the linear systems (2.8), we need to

measure the complexity for solving such a system. We will restrict ourselves here to Krylov subspace iterations. For large-scale problems the main cost to perform one step of a Krylov subspace method method usually consists in the evaluation of $F'[x_n^\delta]$ and $F'[x_n^\delta]^*$ at some given vector. Moreover, to set up the right hand side of the system (2.8) we need to evaluate $F(x_n^\delta)$. Therefore, we can compare different types of regularized Newton methods by counting the total number of operator evaluations until the stopping criterion (2.5a) or (2.5b) is satisfied.

Theorem 4.20 *Let the assumptions of Corollary 2.5 hold except the inequalities (2.24) and consider we measure the complexity of the IRGNM in terms of operator evaluations. If the linear systems (2.8) are solved by Algorithm 3.5 coupled with the stopping criterion (4.11), and if there exists a sufficiently small constant $\bar{C} \geq 0$ such that the estimates*

$$\|h_n^{J_n}\|_{\mathcal{X}} \leq \bar{C}\|h_n^\dagger\|_{\mathcal{X}}, \qquad n = 0, 1, \ldots, N-1, \tag{4.43}$$

are satisfied, then the complexity of the IRGNM is bounded by

a)

$$O\left(\left(u^{-1}(\delta)\right)^{(-\ln\gamma)/\alpha}\right), \qquad \delta \to 0, \tag{4.44}$$

in the case of (4.21a),

b)

$$O\left(\left(-\ln(u^{-1}(\delta))\right)^{\max\{1+1/\beta,2\}}\right), \qquad \delta \to 0. \tag{4.45}$$

in the case of (4.21b).

Proof: Note that for the true solutions $h_n^\dagger$, $n = 0, 1, \ldots, N-1$, of the linear systems (2.8) obviously the inequalities (2.24) are satisfied. Hence, for the IRGNM where the updates are given by the second term in (2.12) the estimate (2.26) is true. Therefore, for $n = 0, 1, \ldots, N-1$ we have the estimate

$$\|h_n^\dagger\|_{\mathcal{X}} = \|x_{n+1}^\delta - x_n^\delta\|_{\mathcal{X}} \leq \|x_{n+1}^\delta - x^\dagger\|_{\mathcal{X}} + \|x_n^\delta - x^\dagger\| \leq \tilde{C} f(\gamma_n),$$

which leads in combination with (4.43) to $\|h_n^{J_n}\|_{\mathcal{X}} \leq \bar{C}\tilde{C} f(\gamma_n)$. Since Algorithm 3.5 is stopped by criterion (4.11), the final residual satisfies

$$\|r^{J_n}\|_{\mathcal{X}} \leq \gamma_n \|h_n^{J_n}\|_{\mathcal{X}} \leq \bar{C}\tilde{C}\gamma_n f(\gamma_n).$$

Hence, if the constant $\bar{C}$ is sufficiently small the stopping criterion (4.10) is satisfied and therefore the inequalities (2.24) are satisfied. In particular Corollary 2.5 and therefore also (4.42) hold true.

To compute an upper bound for the total number of CG-steps we use the estimates (4.38) and (4.40). Consider first the case (4.21a). Then the total number of CG-steps is bounded by

$$\sum_{n=0}^{N-1} J_n \leq \sum_{n=0}^{N-1} \left(\lceil \gamma^{n/\alpha} \rceil + \lceil Cn \rceil \right) \leq \left\lceil \frac{\gamma^{N/\alpha} - 1}{\gamma^{\alpha} - 1} \right\rceil + N + \lceil C \rceil \frac{N(N-1)}{2}.$$

Thus, due to the evaluation of $F'[x_n^\delta]$ and $F'[x_n^\delta]^*$ at some given vector in each CG-step as well as the evaluation of $F(x_n^\delta)$ to set up the right hand side an upper bound for the complexity of the IRGNM and is given by

$$2 \left(\left\lceil \frac{\gamma^{N/\alpha} - 1}{\gamma^{\alpha} - 1} \right\rceil + \lceil C \rceil \frac{N(N-1)}{2} \right) + 2N = O(\gamma^{N/\alpha}).$$

So, (4.44) is a consequence of

$$\gamma^{-\ln(u^{-1}(\delta))/\alpha} = \left(\gamma^{\ln(u^{-1}(\delta))} \right)^{-1/\alpha} = \left(u^{-1}(\delta) \right)^{(-\ln \gamma)/\alpha} \tag{4.46}$$

together with (4.42). In the case of (4.21b) the total number of CG-steps is bounded by

$$\begin{aligned} \sum_{n=0}^{N-1} J_n &\leq \left(\sum_{n=1}^{N-1} \left\lceil C_1 n^{1/\beta} \right\rceil + \lceil Cn \rceil \right) + \lceil C \rceil \\ &\leq \lceil \tilde{C}_1 \rceil \lceil N^{1+1/\beta} \rceil + \lceil C \rceil \frac{N(N-1)}{2} + \lceil C \rceil + N \end{aligned}$$

with some constant $\tilde{C}_1$. Analogously as in the case (4.21a), this together with (4.42) proves (4.45).

□

Let us briefly discuss the assumption (4.43) implying (2.24) and therefore the validity of Corollary 2.5 which together with Theorem 4.19 are the key points to prove the upper bounds (4.44) and (4.45). Naturally, for all $n = 0, 1, \ldots, N-1$ we have the estimates $\|h_n^{J_n}\|_{\mathcal{X}} \leq C(n) \|h_n^\dagger\|_{\mathcal{X}}$ with some constants $C(n)$ depending on the Newton step n. Unfortunately, we could not give a strict proof that these constants can be uniformly bounded by some other constant. On the other hand, since γ_n tends to zero we expect that the approximation quality of the the final iterates of Algorithm 3.5 when coupled with stopping criterion (4.11) increases during Newton's method. Therefore, it seems reasonable to assume that the constants $C(n)$ can be uniformly bounded.
However, we could also assume to terminate the inner CG-iterations by (4.10) to avoid this heuristic argumentation. But since we used in practice (4.11) we decided to prove Theorem 4.20 for this stopping criterion.

To compare the IRGNM with Algorithm 4.10 it is necessary to determine a reasonable update criterion f_{up}. In [40] it was suggested to choose $f_{\text{up}}(n, g_n^\delta, F) = 1$ if and only if $\sqrt{n+1} \in \mathbb{N}$ and $f_{\text{up}}(n, g_n^\delta, F) = 0$ else. In the following theorem we will prove for this update criterion that Algorithm 4.10 for exponentially ill-posed problems is under certain conditions superior compared to the standard IRGNM when the total complexity is measured by operator evaluations.
Note that for other update criteria similar results could be easily obtained.

Theorem 4.21 *Assume that the stopping index of Algorithm 4.10 is determined by (4.42) and that for $n = 0, 1, \ldots, N-1$ the eigenvalues of the operators $G_n^* G_n$ are simple and that there exists a constant C_m and a sufficiently large constant C_{exp} such that the number of outliers of $G_n^* G_n$ can be estimated by*

$$q_n \geq C_m \gamma^{n/\alpha} \qquad \text{in the case of (4.21a)}, \tag{4.47a}$$
$$q_n \geq C_{exp} n^{1/\beta} \qquad \text{in the case of (4.21b)}. \tag{4.47b}$$

Furthermore, let Assumption 4.7 hold in the sense that the method L determines exactly the q_n outliers and let the update criterion be given by

$$\begin{cases} f_{\text{up}}(n, g_n^\delta, F) = 1, & \sqrt{n+1} \in \mathbb{N}, \\ f_{\text{up}}(n, g_n^\delta, F) = 0, & \text{else.} \end{cases}$$

Then, if the linear systems (2.8) are solved by Algorithm 3.5 in the case where $M_n = I$ and if the linear systems (4.20) are solved by Algorithm 3.6 in the case where M_n is given by (4.14) coupled with the stopping criterion (4.11), measuring the complexity of Algorithm 4.10 by operator evaluations, the total complexity of this algorithm is bounded by

a)

$$O\left(\left(u^{-1}(\delta)\right)^{(-\ln\gamma)/\alpha}\right), \qquad \delta \to 0, \tag{4.48}$$

in the case of (4.21a),

b)

$$O\left(\left(-\ln(u^{-1}(\delta))\right)^{\max\{1/2+1/\beta, 2\}}\right), \qquad \delta \to 0. \tag{4.49}$$

in the case of (4.21b).

Proof: Let us consider first the case of (4.21a). Following the lines of the proof of Theorem 4.20 using (4.38), (4.39) and the update criterion the total number of

CG-steps of Algorithm 4.10 is bounded by

$$\begin{aligned}\sum_{n=0}^{N-1} J_n &\leq \left(\sum_{n=0,\sqrt{n+1}\in\mathbb{N}}^{N-1} \lceil \gamma^{n/\alpha}\rceil + \lceil Cn\rceil\right)\\ &\quad + \left(\sum_{k=1,\sqrt{k}\in\mathbb{N}}^{N-1}\sum_{n=k}^{(\sqrt{k}+1)^2-2} \left(\lceil \gamma^{n/\alpha}\rceil - q_{k-1}\right)_+ + \lceil Cn\rceil\right)\\ &\leq \sum_{n=0}^{\sqrt{N}} \left\lceil \gamma^{n^2/\alpha}\right\rceil + \lceil \tilde{C}\rceil \frac{N(N-1)}{2}\\ &\quad + \left(\sum_{k=1,\sqrt{k}\in\mathbb{N}}^{N-1}\sum_{n=k}^{(\sqrt{k}+1)^2-2} \left(\lceil \gamma^{n/\alpha}\rceil - C_m\gamma^{(k-1)/\alpha}\right)_+\right). \qquad (4.50)\end{aligned}$$

It turns out that already the first term is of order $O(\gamma^{N/\alpha})$, so the order of complexity of the unpreconditioned IRGNM is not improved, since on the other hand the complexity obviously does not increase under our assumptions. Hence, we obtain (4.48) from (4.46). In the case of (4.21b) we estimate using (4.40) and (4.41)

$$\begin{aligned}\sum_{n=0}^{N-1} J_n &\leq \left(\sum_{n=0,\sqrt{n+1}\in\mathbb{N}}^{N-1} \lceil C_1 n^{1/\beta}\rceil + \lceil C_2 n\rceil\right)\\ &\quad + \left(\sum_{k=1,\sqrt{k}\in\mathbb{N}}^{N-1}\sum_{n=k}^{(\sqrt{k}+1)^2-2} \left(\lceil C_1 n^{1/\beta}\rceil - q_{k-1}\right)_+ + \lceil Cn\rceil\right)\\ &\leq \sum_{n=0}^{\sqrt{N}} \lceil C_1 n^{2/\beta}\rceil + \lceil \tilde{C}\rceil \frac{N(N-1)}{2}\\ &\quad + \left(\sum_{k=1,\sqrt{k}\in\mathbb{N}}^{N-1}\sum_{n=k}^{(\sqrt{k}+1)^2-2} \left(\lceil C_1 n^{1/\beta}\rceil - C_{exp}(k-1)^{1/\beta}\right)_+\right). \qquad (4.51)\end{aligned}$$

Using $(\sqrt{k}+1)^2 - 2 = k + 2\sqrt{k} - 1$ and the assumtion that the constant C_{exp} is so large such that $C_{exp} \geq C_1$ we obtain

$$\begin{aligned} & C_1(k+2\sqrt{k}-1)^{1/\beta}) - C_{exp}(k-1)^{1/\beta}\\ \leq\ & C_{exp}[(k+2\sqrt{k}-1)^{1/\beta}) - (k-1)^{1/\beta}]\\ =\ & O(k^{1/\beta-1/2}).\end{aligned}$$

Hence, we can estimate the second sum on the right hand side of (4.51) by

$$\bar{C} \sum_{k=1,\sqrt{k}\in\mathbb{N}}^{N-1} k^{1/\beta-1/2} \leq \bar{C} N N^{1/\beta-1/2} = O(N^{1/\beta+1/2})$$

with some constant $\bar{C} > 0$. Note that a better estimate could be obtained by estimating with the help of an integral. But for our purposes this estimate is sufficient which is due to the fact that the first term on the right hand side of (4.51) can also be estimated by

$$\sum_{n=0}^{\sqrt{N}} \lceil C_1 n^{2/\beta} \rceil = O(N^{1/\beta+1/2}),$$

that is both sums are of order $O(N^{1/\beta+1/2})$. Following the lines of the proof of Theorem 4.20 we finally obtain (4.49).

□

Comparing the results of Theorem 4.20 and Theorem 4.21 we obtain that in the case of exponentially ill-posed problems Algorithm 4.10 is superior compared to the standard IRGNM, that is the complexity is significantly reduced, if the assumptions formulated above hold. This at least is true in the case where $\beta \leq 2/3$. In the numerical examples considered in Chapter 7 this condition is satisfied.

We do not want to hide that the assumptions of Theorem 4.21 are in practice not realistic. Furthermore, we did not prove that (4.42) holds true for a frozen Newton method and that the inequalities (4.47) are satisfied. On the other hand, since our convergence analysis includes the case that the linearized equations do not need to be solved exactly, we think that a convergence proof for a frozen Newton method leading to similar results as the standard Newton method could be obtained by the results formulated in this thesis.

Unfortunately, in Theorem 4.16 we only proved an upper bound for the number of outliers. Naturally, to ensure efficiency of the preconditioner (4.14) sufficiently many eigenvalues are required. This is expressed by the inequalities (4.47). We think that by similar ideas as they were used in the proof of Theorem 4.16 such lower bounds could also be obtained. Moreover, in practice we can choose in the steps where $f_{\text{up}} = 1$ the parameter $\varepsilon > 0$ in the stopping criterion (4.11) so small that sufficiently many Ritz pairs are determined. This heuristic justifies the inequalities (4.47).

Unfortunately, for the mildly ill-posed problems the results we obtained in Theorems 4.20 and 4.21 are the same. The reason for this is the eigenvalue distribution of the linearized operators. Still, although the asymptotic behavior of the complexity of a standard IRGNM and Algorithm 4.10 are the same, estimate (4.50) shows that also in this case the total complexity can possibly be reduced.

Note that this discussion serves as a model and motivation for a preconditioned IRGNM. Actually, in numerical examples we can observe that the final iterates of the standard IRGNM and the preconditioned IRGNM are comparable (see Chapter 7). Moreover, although in practice often only a few approximations to the eigenvalues and eigenvectors are known the number of CG-steps can be significantly reduced. Furthermore, the developed methods presented in Chapter 6 are

able to compute additional approximations to eigenvalues of $G_{n,*}^* G_{n,*}$ and eigenvectors when solving the preconditioned linear system (4.20). Hence, we can update the preconditioner during Newton's method.
A thorough analysis of the efficiency of the preconditioner (4.14) given only approximations to the eigenvalues and eigenvectors is the main topic of the next chapter.

Chapter 5

Sensitivity analysis for spectral preconditioners

While for well-posed problems acceleration of iterative solution methods for linear systems is well-studied, the design and analysis of preconditioners for ill-posed problems is not so well understood. Naturally, the linear systems (2.8) are formally well-posed. On the other hand, if the regularization parameter γ_n is small the systems (2.8) will be ill-conditioned due to the ill-posed nature of the original linear systems (2.6).
To this end, an analysis of the behavior of the preconditioned linear systems (4.20) in the presence of inexact eigenpairs to construct M_n given by (4.14) seems to be essential for the construction of efficient preconditioners. In particular we are interested in the behavior of the eigenvalues of the preconditioned operator $M_n^{-1}G_{n,*}^*G_{n,*}$ given only approximations to the eigenpairs of $G_{n,*}^*G_{n,*}$. This particular interest is motivated by the complexity analysis of Chapter 4, since we can only assume that computational cost is saved if the eigenvalues targeted by the preconditioner M_n are shifted into a small neighborhood of ζ, which is not clear for inexact eigenpairs. Therefore, following the articles by Giraud & Gratton [22, 23] in this chapter we carry out a first order analysis to investigate the dependency of these targeted eigenvalues assuming that M_n is constructed by inexact eigenpairs.

In this chapter and the following we are interested in an implementation of the IRGNM. Hence, we investigate the corresponding discretized version of this algorithm. To this end, we start this chapter by discussing briefly how to convert the continuous problem into a discrete one. Moreover, interested in an efficient realization at the outset we are open for any kind of preconditioning technique. Because in the literature many preconditioning techniques are discussed, we shortly recall some of these different techniques to point out the convenience of preconditioners of the form (4.14) in our situation. Subsequently we examine the behavior of the spectrum of the preconditioned operator in the case where the preconditioner is constructed only by approximations. This investigation illuminates important facts

which need to be considered to ensure fast convergence of the CG-method.

5.1 Discretization

Throughout this and the following chapter we focus on an implementable version of Algorithm 4.10. Therefore, we start by shortly giving the idea how to reach a discretized version of the IRGNM. For a detailed description of this discretization process we refer to [16, Chapter 9].

For a numerical realization of the IRGNM a discretization of the operator equations (3.2) and (3.26) respectively is required in each Newton step. This, for instance, can be done by choosing finite dimensional subspaces $\mathcal{X}_N \subset \mathcal{X}$ and $\mathcal{Y}_S \subset \mathcal{Y}$ with bases

$$\mathcal{X}_N = \operatorname{span}\{\varphi_1, \dots, \varphi_N\} \qquad \text{and} \qquad \mathcal{Y}_S = \operatorname{span}\{\psi_1, \dots, \psi_S\}.$$

Introducing projections $Q_N : \mathcal{X} \to \mathcal{X}_N$ and $P_S : \mathcal{Y} \to \mathcal{Y}_S$, we approximate vectors $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ by their projections $Q_N x \in \mathcal{X}_N$ and $P_S y \in \mathcal{Y}_S$, and represent these vectors by their coordinate vectors $\mathbf{x} = (\xi_1, \dots, \xi_N)^T \in \mathbb{R}^N$ and $\mathbf{y} = (\eta_1, \dots, \eta_S)^T \in \mathbb{R}^S$ with respect to the bases $\{\varphi_1, \dots, \varphi_N\}$ and $\{\psi_1, \dots, \psi_S\}$,

$$Q_N x = \sum_{j=1}^{N} \xi_j \varphi_j \qquad \text{and} \qquad P_S y = \sum_{j=1}^{S} \eta_j \psi_j.$$

The linear operator $F'[x_n^\delta]$ can be approximated by the discrete operator $P_S F'[x_n^\delta] Q_N$. Choosing suitable norms in the finite dimensional spaces measuring the smoothness of $\mathbf{x}$ so as to be consistent with $\|x\|_{\mathcal{X}}$ and representing the discrete approximations of the operators $F'[x_n^\delta]$ and $F'[x_n^\delta]^*$ by the matrices $\mathbf{A}_n \in \mathbb{R}^{S \times N}$ and $\mathbf{A}_n^T \in \mathbb{R}^{N \times S}$, we end up with a finite dimensional approximation of the operator equation (3.2) in standard form,

$$\mathbf{G}_n^T \mathbf{G}_n \mathbf{h}_n = \mathbf{G}_n^T \mathbf{g}_n^\delta. \tag{5.1}$$

Here, corresponding to (3.1), we have defined

$$\mathbf{G}_n := \begin{pmatrix} \mathbf{A}_n \\ \sqrt{\gamma_n}\, \mathbf{I} \end{pmatrix} \in \mathbb{R}^{(N+S) \times N}.$$

The symmetric matrix $\mathbf{G}_n^T \mathbf{G}_n$ has to be interpreted as a mapping from $(\mathbb{R}^N, \|\cdot\|_2)$ to $(\mathbb{R}^N, \|\cdot\|_2)$, where $\|\cdot\|_2$ denotes the Euclidean norm. Since the operator $G_n^* G_n$ is strictly coercive, for sufficiently large $S, N \in \mathbb{N}$ the matrix $\mathbf{G}_n^T \mathbf{G}_n$ is positive definite and the system (5.1) has a unique solution denoted by $\mathbf{h}_n^\dagger$. The IRGNM in its discretized version can be formulated by

Algorithm 5.1 (Discretized IRGNM)

Input: Initial guess $\mathbf{x}_0$;

$n = 0, \quad \mathbf{x}_0^\delta := \mathbf{x}_0;$

while $\left(\left\|\mathbf{F}(\mathbf{x}_n^\delta) - \mathbf{y}^\delta\right\| \geq \tau\delta\right)$

Solve $\mathbf{G}_n^T\mathbf{G}_n\mathbf{h}_n = \mathbf{G}_n^T\mathbf{g}_n^\delta$;

$\mathbf{x}_{n+1}^\delta := \mathbf{x}_n^\delta + \mathbf{h}_n^\dagger$;

$n := n + 1$;

Naturally, under our constraints the matrices $\mathbf{A}_n$ and $\mathbf{A}_n^T$ are not available, we can only evaluate $\mathbf{A}_n\mathbf{x}$ and $\mathbf{A}_n^T\mathbf{y}$ for given vectors $\mathbf{x} \in \mathbb{R}^N$ and $\mathbf{y} \in \mathbb{R}^S$. Hence, in general we can only compute an approximation $\mathbf{h}_n^{\text{app}}$ to the true solution $\mathbf{h}_n^\dagger$ of (5.1) by an iterative method. To determine $\mathbf{h}_n^{\text{app}}$ we apply Algorithm 3.5 coupled with a suitable stopping criterion to the linear system (5.1) (see Section 4.2). A realization of this algorithm just needs a "black box" evaluating $\mathbf{A}_n\mathbf{x}$ and $\mathbf{A}_n^T\mathbf{y}$. Following the idea of Section 4.4, to speed up the IRGNM different kind of preconditioning techniques may be successful. An efficient preconditioning technique usually depends on the coefficient matrix as well as the iterative method used to solve the linear system. In the next section we sum up some popular preconditioning techniques, which can be found in the literature. Beforehand we introduce some notation. To formulate a discrete frozen IRGNM we define corresponding to (4.19) the matrix

$$\mathbf{G}_{n,i} := \begin{pmatrix} \mathbf{A}_n \\ \sqrt{\gamma_{n+i}}\,\mathbf{I} \end{pmatrix} \in \mathbb{R}^{(N+S)\times N}, \qquad i \in \mathbb{N}_0. \tag{5.2}$$

Note that in particular $\mathbf{G}_{n,0} = \mathbf{G}_n$ holds. Throughout this and the following chapter we denote the nonnegative eigenvalues of the matrix $\mathbf{A}_n^T\mathbf{A}_n$ by λ_j, $j \in \mathbb{N}$, enumerated in nonincreasing order with multiplicity and the corresponding orthonormal eigenvectors by $\mathbf{u}_j$. The eigenvalues of $\mathbf{G}_{n,i}^T\mathbf{G}_{n,i}$ are given by $\mu_j := \gamma_{n+i} + \lambda_j$, $j \in \mathbb{N}$.

5.2 Preconditioning techniques

Finding an efficient preconditioner for solving a given linear system is often viewed as a combination of art and science. In general theoretical results are rare and some methods work surprisingly well, often despite expectations. Note, at the outset there are virtually no limits to available options for obtaining good preconditioners. Often they are built from the original coefficient matrix. If this matrix is available, its analyzation can yield information which can be used to set up an efficient preconditioner. If it is not available the only choice to exploit information are matrix-free iterative methods such as Lanczos' or Arnoldi's method.

Since in the literature many preconditioning techniques have been discussed, we give in this section a short overview on some of these methods. This discussion shall illuminate why preconditioners of the form (4.14) under the constraints of large-scale ill-posed problems are more promising compared to other preconditioning techniques. For a more detailed introduction to preconditioning techniques we refer to the textbooks of Saad [77, Chapter 10] and Fischer [17] and the articles by Axelsson [1, 3].

Let us consider for the following the linear system

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{5.3}$$

with a symmetric and positive definite matrix $\mathbf{A} : \mathbb{R}^d \to \mathbb{R}^d$ and a given right hand side $\mathbf{b} \in \mathbb{R}^d$. Roughly speaking, a (left) preconditioner $\mathbf{P}$ is any form of implicit or explicit modification of an original linear system which makes it "easier" to solve by a given iterative method, that is the resulting system

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{x} = \mathbf{P}^{-1}\mathbf{b}$$

should require less iteration steps until the approximate solution satisfies some (relative) error level than the original system (5.3). One general idea to define a preconditioner is to perform some Incomplete LU (ILU) factorization of the original matrix $\mathbf{A}$, that is we compute a sparse lower triangular matrix $\mathbf{L}$ and a sparse upper triangular matrix $\mathbf{U}$ such that the residual $\mathbf{R} := \mathbf{L}\mathbf{U} - \mathbf{A}$ satisfies certain constraints. A general algorithm for building incomplete LU factorizations can be derived by performing Gaussian elimination and dropping some elements in predetermined nondiagonal positions. Different implementations of Gaussian elimination lead to different ILU factorizations. Obviously, to perform Gaussian elimination the coefficient matrix $\mathbf{A}$ must be available. Hence, under our constraints this preconditioning technique is no option.
Another simple idea for finding an approximate inverse of the matrix $\mathbf{A}$ is to attempt to find a matrix $\mathbf{P}$, which minimizes the residual matrix $\mathbf{I} - \mathbf{A}\mathbf{P}$ with respect to some norm on $\mathbb{R}^{d\times d}$. For example, one can define the functional

$$J(\mathbf{P}) := \|\mathbf{I} - \mathbf{A}\mathbf{P}\|_F^2,$$

where the Frobenius norm $\|\cdot\|_F$ is defined by

$$\|\mathbf{A}\|_F := \left(\sum_{j,k=1}^{d} |\mathbf{a}_{jk}|^2 \right)^{1/2}.$$

A matrix $\mathbf{P}$ whose value $J(\mathbf{P})$ is small would be a right-approximate inverse of $\mathbf{A}$. Analogously we can define a left-approximate inverse. With the notation $\mathbf{I} = (\mathbf{e}_1, \ldots, \mathbf{e}_d)$ and $\mathbf{P} = (\mathbf{p}_1, \ldots, \mathbf{p}_d)$ the functional J takes the form

$$J(\mathbf{P}) = \sum_{j=1}^{d} \|\mathbf{e}_j - \mathbf{A}\mathbf{p}_j\|_2^2. \tag{5.4}$$

Hence, to minimize the functional J there are two different ways to proceed, either it can be minimized globally as a function of the matrix $\mathbf{P}$, for instance by a gradient-type method or the individual functions $\|\mathbf{e}_j - \mathbf{A}\mathbf{p}_j\|_2^2$, $j = 1, 2, \ldots, d$, can be minimized. Both minimization processes involve in general many evaluations of the matrix-vector product $\mathbf{A}\mathbf{x}$. Therefore, under our assumptions the construction of a preconditioner in this way would be far too complex.

Another class of preconditioning techniques is given by so-called polynomial preconditioning. The basic idea is as follows: instead of solving the system (5.3) by the CG-method, the CG-iteration is applied to

$$\psi(\mathbf{A})\mathbf{A}\mathbf{x} = \psi(\mathbf{A})\mathbf{b}. \tag{5.5}$$

Here ψ is a suitably chosen polynomial of small degree. Moreover, it is required that none of the zeros of ψ coincides with an eigenvalue of $\mathbf{A}$. This guarantees that the preconditioned system (5.5) is equivalent to (5.3). Polynomial preconditioning goes back to the 1950s and it has been suggested in many different ways (see for example [18, 25, 56] and the textbook of Fischer [17]). The standard approach for the design of preconditioners is to choose the polynomial ψ such that $\psi(\mathbf{A})\mathbf{A}$ is, in some sense, as close as possible to the identity matrix $\mathbf{I}$. For instance, one could attempt to minimize the Euclidean norm $\|\mathbf{I} - \psi(\mathbf{A})\mathbf{A}\|_2$. The solution of this problem would require the knowledge of all eigenvalues of $\mathbf{A}$. Therefore, one usually substitutes for the spectrum of $\mathbf{A}$ an interval $[a, b]$, $a, b > 0$, which is known to contain all eigenvalues of $\mathbf{A}$. This approach leads to the Chebysheff approximation problem on $[a, b]$, which can be rewritten in terms of the unit interval $[-1, 1]$ (cf. Lemma 4.1),

$$\min_{\psi \in \Pi_k^1} \max_{t \in [-1,1]} |\psi(t)|.$$

In the case of ill-posed problems following the lines of the proof of Lemma 4.13 it is possible to construct suitable polynomials. Unfortunately, solving the linear system (5.5) by the CG-algorithm involves a lot more evaluations of $\mathbf{A}$ to some given vector compared with solving the original system (5.3). Hence, in our situation polynomial preconditioning would increase the complexity drastically.

The last class of preconditioners we want to discuss are so-called spectral preconditioners. To introduce spectral preconditioners let us denote the eigenvalues of $\mathbf{A}$ by $\beta_1 \geq \ldots \geq \beta_d > 0$ and the corresponding orthonormal eigenvectors by $\mathbf{w}_1, \ldots, \mathbf{w}_d$. Spectral preconditioners can be split into two main families, depending on their effect on the spectrum of $\mathbf{A}$. They are referred to as coarse grid preconditioners if they attempt only to shift a subset of the eigenvalues of $\mathbf{A}$ close to some $\zeta > 0$ (see [10, 21, 70]). The name of these preconditioners comes from domain decomposition, where they were originally introduced. The second families are called deflation preconditioners (see [21]). Here a subset of eigenvalues is attempted to be moved to some $\zeta > 0$. These preconditioners have been proven to

be successful when there are a few isolated extremal eigenvalues (see for example Mansfield [60] and Nicolaides [69]).
There exist many different types of spectral preconditioner, which often reduce to the same expression if exact spectral information is used (see [23] for a number of examples of such preconditioners). We restrict ourselves to discuss spectral preconditioners corresponding to (4.14),

$$\mathbf{P} = \mathbf{I} + \sum_{j=1}^{k_{ev}} \left(\frac{\beta_j}{\zeta} - 1 \right) \mathbf{w}_j \mathbf{w}_j^T, \tag{5.6}$$

which belong to the class of deflation based preconditioners. Note that instead of exact spectral information in applications often only approximate spectral information is available to set up (5.6). It is the goal of the next section to study the efficiency of the preconditioner (5.6) in the presence of inexact spectral information. An amazing example for the efficiency of a spectral preconditioner is provided by the atmosphere data assimilation area (see [19]). In this application, nonlinear least-squares problems with more than 10^7 unknowns are solved day-to-day using a Gauss-Newton approach with inner CG-iteration. Similar to our idea Lanczos' method is exploited to extract approximate spectral information, which is used to construct a deflation spectral preconditioner for the subsequent linear least-squares problems.

5.3 Sensitivity analysis

To study the performance of the preconditioner (5.6) in the presence of inexact spectral information we mainly follow the articles by Giraud & Gratton [22, 23]. To this end, we assume that the eigenvalues $\beta_1 \geq \ldots \geq \beta_{k_{ev}}$ and corresponding orthonormal eigenvectors $\mathbf{w}_1, \ldots, \mathbf{w}_{k_{ev}}$ used to construct $\mathbf{P}$ given by (5.6) are not related to $\mathbf{A}$, but to a nearby matrix $\mathbf{A} + t\mathbf{E}$, where t is a real parameter and the symmetric matrix $\mathbf{E} \in \mathbb{R}^{d \times d}$ is normalized such that $\|\mathbf{E}\| = 1$. It is the aim of this section to carry out a first-order perturbation analysis, which shows the asymptotic sensitivity of the eigenvalues of the preconditioned matrix for small enough values of the parameter t.

For this purpose we denote by $\beta_i(t)$ the eigenvalues and by $\mathbf{w}_i(t)$, $i = 1, \ldots, d$, the corresponding eigenvectors of $\mathbf{A} + t\mathbf{E}$. To guarantee that the eigenvalues of $\mathbf{A} + t\mathbf{E}$ are differentiable functions of t in a small neighborhood $U_\varepsilon(0)$ for some $\varepsilon > 0$, we assume that $\mathbf{A}$ has only simple eigenvalues (see [24] and [79]). If the eigenvectors are normalized using

$$\mathbf{w}_i^T(t)\mathbf{w}_i = 1, \tag{5.7}$$

the eigenvectors are also differentiable functions of t in a neighborhood of $t = 0$.

The matrices $\tilde{\mathbf{W}} \in \mathbb{R}^{d\times d}$ and $\tilde{\Lambda} \in \mathbb{R}^{d\times d}$ are defined by

$$\tilde{\mathbf{W}}(t) := (\mathbf{w}_1(t), \ldots, \mathbf{w}_d(t)) \qquad \text{and} \qquad \tilde{\Lambda}(t) := \operatorname{diag}(\beta_i(t)).$$

Then, for sufficiently small $t \in U_\varepsilon(0)$ we have

$$(\mathbf{A} + t\mathbf{E})\tilde{\mathbf{W}}(t) = \tilde{\mathbf{W}}(t)\tilde{\Lambda}(t)$$

and it holds $\tilde{\mathbf{W}}(0) = \mathbf{W}$, where $\mathbf{W} := (\mathbf{w}_1, \ldots, \mathbf{w}_d)$. A first order expansion of the eigenvalues and eigenvectors in the direction $\mathbf{E}$ (see [24] and [79]) is given by

$$\begin{aligned} \tilde{\mathbf{W}}(t) &= \mathbf{W} + t\Delta\mathbf{W} + o(t), && (5.8)\\ \beta_i(t) &= \beta_i + t\Delta\beta_i + o(t), && (5.9) \end{aligned}$$

where the i-th column of $\Delta\mathbf{W}$ denoted by $\Delta\mathbf{w}_i$, and $\Delta\beta_i$ are given by

$$\begin{aligned} \Delta\mathbf{w}_i &= \mathbf{W}(\neg i)(\beta_i\mathbf{I} - \mathbf{B}_i)^{-1}\mathbf{W}(\neg i)^T\mathbf{E}\mathbf{w}_i, && (5.10)\\ \Delta\beta_i &= \mathbf{w}_i^T\mathbf{E}\mathbf{w}_i. && (5.11) \end{aligned}$$

Here the diagonal matrix $\mathbf{B}_i \in \mathbb{R}^{d-1\times d-1}$ is given by

$$\mathbf{B}_i = \operatorname{diag}(\beta_1, \ldots, \beta_{i-1}, \beta_{i+1}, \ldots, \beta_d).$$

and for any square matrix $\mathbf{C} \in \mathbb{R}^{d\times d}$ the matrix $\mathbf{C}(\neg i) \in \mathbb{R}^{d\times(d-1)}$ denotes the matrix whose columns are those of $\mathbf{C}$ except for the i-th.

As preparation to prove the main theorem of this section we formulate the following lemma establishing some useful equalities.

Lemma 5.2 *The following equalities hold:*

$$\begin{aligned} \mathbf{w}_j^T\Delta\mathbf{w}_i &= (\beta_i - \beta_j)^{-1}\mathbf{w}_j^T\mathbf{E}\mathbf{w}_i, \qquad j \neq i, && (5.12)\\ \Delta\mathbf{w}_j^T\mathbf{w}_i &= (\beta_j - \beta_i)^{-1}\mathbf{w}_j^T\mathbf{E}^T\mathbf{w}_i, \qquad j \neq i, && (5.13)\\ \mathbf{w}_j^T\Delta\mathbf{w}_j &= \Delta\mathbf{w}_j^T\mathbf{w}_j = 0. && (5.14) \end{aligned}$$

Proof: For $j \neq i$ consider first the case $j < i$. Then we have

$$\mathbf{w}_j^T\mathbf{W}(\neg i) = (0, \ldots, 0, 1, 0, \ldots, 0),$$

where the 1 is at position j. Therefore,

$$\begin{aligned} \mathbf{w}_j^T\Delta\mathbf{w}_i &= \mathbf{w}_j^T\mathbf{W}(\neg i)(\beta_i\mathbf{I} - \mathbf{B}_i)^{-1}\mathbf{W}(\neg i)^T\mathbf{E}\mathbf{w}_i\\ &= (\beta_i - \beta_j)^{-1}(0, \ldots, 0, 1, 0, \ldots, 0)\mathbf{W}(\neg i)^T\mathbf{E}\mathbf{w}_i\\ &= (\beta_i - \beta_j)^{-1}\mathbf{w}_j^T\mathbf{E}\mathbf{w}_i. \end{aligned}$$

In the case where $i < j$ we have $\mathbf{w}_j^T \mathbf{W}(\neg i) = (0, \ldots, 0, 1, 0, \ldots, 0)$, where the 1 is at position $j-1$. Notice, in this case the $(j-1)$-th entry on the diagonal of the matrix $\mathbf{B}_i$ is β_j and the $(j-1)$-th column of $\mathbf{W}(\neg i)$ is $\mathbf{w}_j$. So, with the same computation as above (5.12) follows. Analogously we can prove (5.13). The equation $\mathbf{w}_j^T \mathbf{W}(\neg j) = 0$ yields

$$\mathbf{w}_j^T \Delta \mathbf{w}_j = \mathbf{w}_j^T \mathbf{W}(\neg j)(\beta_j \mathbf{I} - \mathbf{B}_j)^{-1} \mathbf{W}(\neg j)^T \mathbf{E} \mathbf{w}_j = 0,$$

which proves (5.14).

□

Definition 5.3 *The* Hadamard product *of two matrices* $\mathbf{A} = (\mathbf{a}_{ij}) \in \mathbb{C}^{m \times n}$ *and* $\mathbf{B} = (\mathbf{b}_{ij}) \in \mathbb{C}^{m \times n}$ *is denoted by* $\circ$*,*

$$\mathbf{A} \circ \mathbf{B} := (\mathbf{a}_{ij} \mathbf{b}_{ij}) \in \mathbb{C}^{m \times n}.$$

Theorem 5.4 *The preconditioner*

$$\mathbf{P}(t) := \mathbf{I} + \sum_{j=1}^{k_{ev}} \left(\frac{\beta_j(t)}{\zeta} - 1 \right) \frac{\mathbf{w}_j(t) \mathbf{w}_j(t)^T}{\mathbf{w}_j(t)^T \mathbf{w}_j(t)}$$

is such that the eigenvalues $\beta_i^{\mathrm{pre}}(t)$*of the preconditioned matrix* $\mathbf{P}(t)^{-1}\mathbf{A} \in \mathbb{R}^{d \times d}$ *satisfy*

$$\begin{cases} \beta_i^{\mathrm{pre}}(t) = \zeta + \beta_i(\mathbf{T})\, t + o(t), & i \le k_{ev}, \\ \beta_i^{\mathrm{pre}}(t) = \beta_i + o(t), & i > k_{ev}. \end{cases} \tag{5.15}$$

Here the matrix

$$\mathbf{T} := (\mathbf{Y} + \mathbf{J}) \circ \mathbf{R} + (\mathbf{Y}^T \circ \mathbf{R}^T) \in \mathbb{R}^{k_{ev} \times k_{ev}}$$

is defined by the matrices $\mathbf{Y} = (\mathbf{y}_{\ell s}) \in \mathbb{R}^{k_{ev} \times k_{ev}}$*,* $\mathbf{J} \in \mathbb{R}^{k_{ev} \times k_{ev}}$ *and* $\mathbf{R} \in \mathbb{R}^{k_{ev} \times k_{ev}}$ *given through*

$$\mathbf{y}_{\ell\ell} = 0, \qquad \mathbf{y}_{\ell s} = \frac{\zeta - \beta_s}{\beta_s - \beta_\ell} \sqrt{\frac{\beta_\ell}{\beta_s}} \quad \textit{for} \quad \ell \neq s,$$

$\mathbf{J} = -\mathrm{diag}(\zeta/\beta_1, \ldots, \zeta/\beta_{k_{ev}})$ *and* $\mathbf{R} = \mathbf{W}_{k_{ev}}^T \mathbf{E} \mathbf{W}_{k_{ev}}$ *with* $\mathbf{W}_{k_{ev}} := (\mathbf{w}_1, \ldots, \mathbf{w}_{k_{ev}})$ *and* $\beta_i(\mathbf{T})$ *denotes the i-th eigenvalue of* $\mathbf{T}$ *such that* $|\beta_1(\mathbf{T})| \ge \ldots \ge |\beta_{k_{ev}}(\mathbf{T})|$.

Proof: It follows by a straightforward computation (cf. the proof of Theorem 4.6) that the inverse of $\mathbf{P}(t)$ is given by

$$\mathbf{P}(t)^{-1} = \mathbf{I} + \sum_{j=1}^{k_{ev}} \left(\frac{\zeta}{\beta_j(t)} - 1 \right) \frac{\mathbf{w}_j(t) \mathbf{w}_j(t)^T}{\mathbf{w}_j(t)^T \mathbf{w}_j(t)}.$$

Since the eigenvalues of an arbitrary $d \times d$ matrix coincide with the eigenvalues of its transposed, the equality

$$\left(\mathbf{P}(t)^{-1}\mathbf{A})\right)^T = \mathbf{A}^T(\mathbf{P}(t)^{-1})^T = \mathbf{A}\mathbf{P}(t)^{-1}$$

shows that the eigenvalues of $\mathbf{P}(t)^{-1}\mathbf{A}$ are those of $\mathbf{A}\mathbf{P}(t)^{-1}$. Using the first order expansion of the eigenvectors (5.8) together with (5.14) we obtain

$$\begin{aligned} \mathbf{w}_j(t)\mathbf{w}_j(t)^T &= \mathbf{w}_j\mathbf{w}_j^T + t\left(\Delta\mathbf{w}_j\mathbf{w}_j^T + \mathbf{w}_j^T\Delta\mathbf{w}_j\right) + o(t), && (5.16)\\ \mathbf{w}_j(t)^T\mathbf{w}_j(t) &= 1 + o(t). && (5.17) \end{aligned}$$

Therefore, using a Taylor expansion of the function $g(t) := \zeta/t$, that is

$$g(t+h) - g(t) = -\frac{\zeta}{t^2}h + O(h^2),$$

we can expand the inverse of the preconditioner $\mathbf{P}(t)^{-1}$ in a small neighborhood of $t = 0$ using (5.9), (5.16), (5.17) and (5.11) through

$$\begin{aligned} \mathbf{P}(t)^{-1} &= \mathbf{I} + \sum_{j=1}^{k_{ev}} \left(\frac{\zeta}{\beta_j} - t\frac{\zeta}{\beta_j^2}\Delta\beta_j - 1\right)\left(\mathbf{w}_j\mathbf{w}_j^T + t\mathbf{w}_j\Delta\mathbf{w}_j^T + t\Delta\mathbf{w}_j\mathbf{w}_j^T\right) + o(t)\\ &= \mathbf{P}(0)^{-1} + t\Delta(\mathbf{P}^{-1}) + o(t), \end{aligned}$$

where the matrix $\Delta(\mathbf{P}^{-1})$ is given by

$$\Delta(\mathbf{P}^{-1}) := \sum_{j=1}^{k_{ev}} \left[\left(\frac{\zeta}{\beta_j} - 1\right)\left(\mathbf{w}_j\Delta\mathbf{w}_j^T + \Delta\mathbf{w}_j\mathbf{w}_j^T\right) - \zeta\frac{\mathbf{w}_j^T\mathbf{E}\mathbf{w}_j}{\beta_j^2}\mathbf{w}_j\mathbf{w}_j^T\right].$$

Multiplying from the left by $\mathbf{A}$ yields

$$\mathbf{A}\mathbf{P}(t)^{-1} = \mathbf{A}\mathbf{P}(0)^{-1} + t\mathbf{A}\Delta(\mathbf{P}^{-1}) + o(t),$$

with

$$\mathbf{A}\Delta(\mathbf{P})^{-1} = \sum_{j=1}^{k_{ev}} \left[\left(\frac{\zeta}{\beta_j} - 1\right)\left(\mathbf{A}\mathbf{w}_j\Delta\mathbf{w}_j^T + \mathbf{A}\Delta\mathbf{w}_j\mathbf{w}_j^T\right) - \zeta\frac{\mathbf{w}_j^T\mathbf{E}\mathbf{w}_j}{\beta_j^2}\mathbf{A}\mathbf{w}_j\mathbf{w}_j^T\right].$$

For $i > k_{ev}$ the first-order approximation of the simple eigenvalues reads

$$\beta_i^{\text{pre}}(t) = \beta_i + t\mathbf{w}_i^T\mathbf{A}\Delta(\mathbf{P}^{-1})\mathbf{w}_i + o(t) = \beta_i + o(t), \tag{5.18}$$

due to (5.14) and the orthogonality of $\mathbf{W}$. For $i \leq k_{ev}$, using general perturbation results for matrices concerning multiple eigenvalues (see [54, Chapter 11]) the first order approximation of the multiple eigenvalue ζ reads

$$\beta_i^{\text{pre}}(t) = \zeta + t\beta_i(\mathbf{W}_{k_{ev}}^T\mathbf{A}\Delta(\mathbf{P}^{-1})\mathbf{W}_{k_{ev}}) + o(t),$$

where $\beta_i(\mathbf{W}_{k_{ev}}^T\mathbf{A}\Delta(\mathbf{P}^{-1})\mathbf{W}_{k_{ev}})$ denotes the i-th eigenvalue of $\mathbf{W}_{k_{ev}}^T\mathbf{A}\Delta(\mathbf{P}^{-1})\mathbf{W}_{k_{ev}}$ such that

$$|\beta_1(\mathbf{W}_{k_{ev}}^T\mathbf{A}\Delta(\mathbf{P}^{-1})\mathbf{W}_{k_{ev}})| \geq \ldots \geq |\beta_{k_{ev}}(\mathbf{W}_{k_{ev}}^T\mathbf{A}\Delta(\mathbf{P}^{-1})\mathbf{W}_{k_{ev}})|.$$

Using the orthogonality of $\mathbf{W}$ and (5.14) the diagonal elements of

$$\mathbf{W}_{k_{ev}}^T\mathbf{A}\Delta(\mathbf{P}^{-1})\mathbf{W}_{k_{ev}} \in \mathbb{R}^{k_{ev}\times k_{ev}}$$

are determined by

$$(\mathbf{W}_{k_{ev}}^T\mathbf{A}\Delta(\mathbf{P}^{-1})\mathbf{W}_{k_{ev}})_{\ell\ell} = -\zeta\frac{\mathbf{w}_\ell^T\mathbf{E}\mathbf{w}_\ell}{\beta_\ell}, \qquad \ell = 1,\ldots,k_{ev}. \tag{5.19}$$

For the (ℓ, s) off-diagonal element $\mathbf{w}_\ell^T\mathbf{A}\Delta(\mathbf{P}^{-1})\mathbf{w}_s$ using (5.12) and (5.13) we compute

$$\begin{aligned}
&\mathbf{w}_\ell^T\left\{\sum_{j=1}^{k_{ev}}\left[\left(\frac{\zeta}{\beta_j}-1\right)\left(\mathbf{A}\mathbf{w}_j\Delta\mathbf{w}_j^T+\mathbf{A}\Delta\mathbf{w}_j\mathbf{w}_j^T\right)-\zeta\frac{\mathbf{w}_j^T\mathbf{E}\mathbf{w}_j}{\beta_j^2}\mathbf{A}\mathbf{w}_j\mathbf{w}_j^T\right]\right\}\mathbf{w}_s\\
=&\left\{\sum_{j=1}^{k_{ev}}\left(\frac{\zeta}{\beta_j}-1\right)\mathbf{w}_\ell^T\mathbf{A}\mathbf{w}_j\Delta\mathbf{w}_j^T\right\}\mathbf{w}_s+\left\{\mathbf{w}_\ell^T\sum_{j=1}^{k_{ev}}\left(\frac{\zeta}{\beta_j}-1\right)\mathbf{A}\mathbf{w}_j\Delta\mathbf{w}_j^T\mathbf{w}_s\right\}\\
&-\sum_{j=1}^{k_{ev}}\zeta\frac{\mathbf{w}_j^T\mathbf{E}\mathbf{w}_j}{\beta_j^2}\mathbf{w}_\ell^T\mathbf{A}\mathbf{w}_j\mathbf{w}_j^T\mathbf{w}_s\\
=&\left(\frac{\zeta}{\beta_\ell}-1\right)\mathbf{w}_\ell^T\mathbf{A}\mathbf{w}_\ell\Delta\mathbf{w}_\ell^T\mathbf{w}_s+\left(\frac{\zeta}{\beta_s}-1\right)\mathbf{w}_\ell^T\mathbf{A}\Delta\mathbf{w}_s\\
=&\frac{t}{\beta_\ell-\beta_s}\left(\frac{\zeta}{\beta_\ell}-1\right)\left(\mathbf{w}_\ell^T\mathbf{A}\mathbf{w}_\ell\mathbf{w}_\ell^T\mathbf{E}^T\mathbf{w}_s\right)^T+\frac{t}{\beta_s-\beta_\ell}\left(\frac{\zeta\beta_\ell}{\beta_s}-\beta_\ell\right)\mathbf{w}_\ell^T\mathbf{E}\mathbf{w}_s\\
=&\frac{1}{\beta_s-\beta_\ell}\left[\frac{\beta_\ell}{\beta_s}(\zeta-\beta_s)\,\mathbf{w}_\ell^T\mathbf{E}\mathbf{w}_s-(\zeta-\beta_\ell)\,\mathbf{w}_\ell^T\mathbf{E}^T\mathbf{w}_s\right]t\\
=&\frac{\zeta-\beta_s}{\beta_s-\beta_\ell}\sqrt{\frac{\beta_\ell}{\beta_s}}\mathbf{w}_\ell^T\mathbf{E}\mathbf{w}_s\sqrt{\frac{\beta_\ell}{\beta_s}}+\frac{\zeta-\beta_\ell}{\beta_\ell-\beta_s}\sqrt{\frac{\beta_s}{\beta_\ell}}\left(\mathbf{w}_s^T\mathbf{E}\mathbf{w}_\ell\right)^T\sqrt{\frac{\beta_\ell}{\beta_s}}\\
=&\mathbf{y}_{\ell s}(\mathbf{R})_{\ell s}\sqrt{\frac{\beta_\ell}{\beta_s}}+\mathbf{y}_{s\ell}(\mathbf{R})_{s\ell}\sqrt{\frac{\beta_\ell}{\beta_s}}.
\end{aligned}\tag{5.20}$$

Note that the matrices

$$\mathbf{D}_{k_{ev}}^{-1/2}\mathbf{W}_{k_{ev}}^T\mathbf{A}\Delta(\mathbf{P}^{-1})\mathbf{W}_{k_{ev}}\mathbf{D}_{k_{ev}}^{1/2} \qquad \text{and} \qquad \mathbf{W}_{k_{ev}}^T\mathbf{A}\Delta(\mathbf{P}^{-1})\mathbf{W}_{k_{ev}},$$

where $\mathbf{D}_{k_{ev}} := \mathrm{diag}(\beta_1,\ldots,\beta_{k_{ev}})$ have the same eigenvalues by similarity. Now, (5.19) together with (5.20) yields

$$\mathbf{D}_{k_{ev}}^{-1/2}\mathbf{W}_{k_{ev}}^T\mathbf{A}\Delta(\mathbf{P}^{-1})\mathbf{W}_{k_{ev}}\mathbf{D}_{k_{ev}}^{1/2} = (\mathbf{Y}+\mathbf{J})\circ\mathbf{R}+(\mathbf{Y}^T\circ\mathbf{R}^T),$$

which proves the assertion.

□

To investigate the sensitivity of the multiple eigenvalue $\beta_1^{\text{pre}}(0) = \ldots = \beta_{k_{ev}}^{\text{pre}}(0)$ of $P(0)^{-1}A$, it is possible to define a condition number $\text{cond}(\beta_i^{\text{pre}})$ for these eigenvalues in the direction of $\mathbf{E}$. Note that due to the multiplicity of $\beta_i^{\text{pre}} := \beta_i^{\text{pre}}(0)$, $i = 1, \ldots, k_{ev}$, at the outset it was not clear if the mappings $\beta_i^{\text{pre}}(t)$, $i = 1, \ldots, k_{ev}$, are differentiable in a small neighborhood of zero, which is now an obvious consequence of (5.15). Considering this remark we give the following definition of a condition number (see [73]), which in our case simplifies to the derivatives $(\beta_i^{\text{pre}})'(0)$.

Definition 5.5 *For $i = 1, \ldots, k_{ev}$ we define the condition number of the eigenvalues β_i^{pre} of the preconditioned matrix $\mathbf{P}(0)^{-1}\mathbf{A}$ in the direction $\mathbf{E}$ by*

$$\text{cond}(\beta_i^{\text{pre}}) := \lim_{u \to 0} \sup_{0<|t|<u} \frac{|\beta_i^{\text{pre}}(t) - \beta_i^{\text{pre}}(0)|}{|t|}. \tag{5.21}$$

Remark 5.6 $\text{cond}(\beta_i^{\text{pre}})$ *in our context is not the usual condition number of the eigenvalue of a matrix as it is used in the literature (see [24, Chapter 7]), but the condition number of the mapping $\beta_i^{\text{pre}}(t)$.*

With the help of Theorem 5.4 it is possible to compute upper bounds for $\text{cond}(\beta_i^{\text{pre}})$.

Corollary 5.7 *The condition number of the eigenvalue β_i^{pre}, $i = 1, \ldots, k_{ev}$, of the preconditioned matrix $\mathbf{P}(0)^{-1}\mathbf{A}$ satisfies the estimate*

$$\text{cond}(\beta_i^{\text{pre}}) \leq 2\|\mathbf{Y}\| + \|\mathbf{J}\|. \tag{5.22}$$

In the case where the approximations $\beta_1(t) \geq \ldots \geq \beta_{k_{ev}}(t)$ with corresponding orthogonal vectors $\mathbf{w}_1(t), \ldots, \mathbf{w}_{k_{ev}}(t)$ to the eigenpairs $(\beta_j, \mathbf{w}_j)$ satisfying (5.7) are known, for sufficiently small t we have for $i = 1, \ldots, k_{ev}$ the estimate

$$|\beta_i^{\text{pre}}(t) - \beta_i^{\text{pre}}| \leq (2\|\mathbf{Y}\| + \|\mathbf{J}\|)\, \|\mathbf{A}\tilde{\mathbf{W}}_{k_{ev}} - \tilde{\mathbf{W}}_{k_{ev}}\tilde{\Lambda}_{k_{ev}}\|. \tag{5.23}$$

Proof: By (5.15) and (5.21) it follows that

$$\text{cond}(\beta_i^{\text{pre}}) = \left|\beta_i \left((\mathbf{Y} + \mathbf{J}) \circ \mathbf{R} + (\mathbf{Y}^T \circ \mathbf{R}^T)\right)\right|, \qquad i = 1, \ldots, k_{ev}.$$

Taking norms the submultiplicativity of the Euclidean norm with respect to the Hadamard product (see [42]), the orthogonality of $\mathbf{W}_{k_{ev}}$ and $\|\mathbf{E}\| = 1$ yield for $i = 1, \ldots, k_{ev}$

$$\begin{aligned}
\text{cond}(\beta_i^{\text{pre}}) &\leq \left\|(\mathbf{Y} + \mathbf{J}) \circ \mathbf{R} + \mathbf{Y}^T \circ \mathbf{R}^T\right\| \\
&\leq \left(\|\mathbf{Y} + \mathbf{J}\| + \|\mathbf{Y}^T\|\right) \|\mathbf{R}\| \\
&\leq (2\|\mathbf{Y}\| + \|\mathbf{J}\|)\, \|\mathbf{W}_{k_{ev}}^T \mathbf{E} \mathbf{W}_{k_{ev}}\| \\
&\leq 2\|\mathbf{Y}\| + \|\mathbf{J}\|,
\end{aligned}$$

which proves (5.22). If $\beta_1(t), \ldots, \beta_{k_{ev}}(t)$ and $\mathbf{w}_1(t), \ldots, \mathbf{w}_{k_{ev}}(t)$ are known, these approximations are exact eigenvalues with corresponding eigenvectors of the matrix

$$\mathbf{A} - (\mathbf{A}\tilde{\mathbf{W}}_{k_{ev}} - \tilde{\mathbf{W}}_{k_{ev}}\tilde{\Lambda}_{k_{ev}})\tilde{\mathbf{W}}_{k_{ev}}^T. \tag{5.24}$$

Hence, in this case the matrix $t\mathbf{E}$ is explicitly known and given by

$$t\mathbf{E} = -(\mathbf{A}\tilde{\mathbf{W}}_{k_{ev}} - \tilde{\mathbf{W}}_{k_{ev}}\tilde{\Lambda}_{k_{ev}})\tilde{\mathbf{W}}_{k_{ev}}^T.$$

Using (5.15), the representation of $t\mathbf{E}$ and (5.17) we can estimate for sufficiently small t

$$\begin{aligned} |\beta_i^{\text{pre}}(t) - \beta_i^{\text{pre}}| &\leq (2\|\mathbf{Y}\| + \|\mathbf{J}\|)\, \|\mathbf{W}_{k_{ev}}^T t\mathbf{E}\mathbf{W}_{k_{ev}}\| \\ &\leq (2\|\mathbf{Y}\| + \|\mathbf{J}\|)\, \|\mathbf{A}\tilde{\mathbf{W}}_{k_{ev}} - \tilde{\mathbf{W}}_{k_{ev}}\tilde{\Lambda}_{k_{ev}}\|, \end{aligned}$$

which proves (5.23).

□

Remark 5.8 *Note that the estimate*

$$\left\|(\mathbf{Y}+\mathbf{J}) \circ \mathbf{R} + \mathbf{Y}^T \circ \mathbf{R}^T\right\| \leq \left(\|\mathbf{Y}+\mathbf{J}\| + \|\mathbf{Y}^T\|\right) \|\mathbf{R}\|$$

is a worst case bound and may be pessimistic in the case where $\mathbf{R}$ *has zero entries corresponding to large entries in* $\mathbf{Y}$ *and* $\mathbf{J}$*.*

Remark 5.9 *The theoretical study has been made assuming that all the eigenvalues of* $\mathbf{A}$ *are simple. Actually, the results are still true if some of the* β_i *for* $i > k_{ev}$ *are multiple, that is the eigenvalues which are not targeted by the preconditioner.*

Due to inequality (5.22) small entries in $\mathbf{Y}$ and $\mathbf{J}$ imply a small condition number of the eigenvalues β_i, $i = 1, \ldots, k_{ev}$. Hence, we expect that the preconditioned matrix $\mathbf{P}(t)^{-1}\mathbf{A}$ has a cluster of eigenvalues in a neighborhood of ζ leading to a reduction of the CG-steps. On the other hand, the preconditioner $\mathbf{P}(t)$ may be unstable if

a) for a pair (s, ℓ) the ratio

$$\frac{\zeta - \beta_s}{\beta_s - \beta_\ell}\sqrt{\frac{\beta_\ell}{\beta_s}}$$

b) and/or for an s the ratio ζ/β_s

is large. This instability may happen if the preconditioner $\mathbf{P}(t)$ targets small and/or clustered eigenvalues of $\mathbf{A}$ and if the parameter ζ is chosen too far outside of the spectrum of $\mathbf{A}$. The consequences of these results for Algorithm 4.10 will be discussed in the next section. Before we want to illustrate the instabilities by an example discussed in [78].

Example 5.10 Consider for sufficiently small $\varepsilon > 0$ the symmetric and positive definite matrices $\mathbf{C} := \mathrm{diag}(1, 1, 2)$,

$$\begin{aligned}
\mathbf{C}' &:= \mathbf{c}'(\varepsilon) \begin{pmatrix} 1+3\varepsilon^2+\varepsilon^3 & \varepsilon-\varepsilon^2+\varepsilon^3 & -\varepsilon-\varepsilon^2 \\ \varepsilon-\varepsilon^2+\varepsilon^3 & 1+3\varepsilon^2+\varepsilon^3 & \varepsilon+\varepsilon^2 \\ -\varepsilon-\varepsilon^2 & \varepsilon+\varepsilon^2 & 2+2\varepsilon^2-2\varepsilon^3 \end{pmatrix}, \\
\mathbf{C}'' &:= \mathbf{c}''(\varepsilon) \begin{pmatrix} 5-3\varepsilon+30\varepsilon^2-20\varepsilon^3 & 4\varepsilon+10\varepsilon^2+10\varepsilon^3 & 5\varepsilon(1-\varepsilon) \\ 4\varepsilon+10\varepsilon^2+10\varepsilon^3 & 5+3\varepsilon+45\varepsilon^2-5\varepsilon^3 & 10\varepsilon(1-\varepsilon) \\ 5\varepsilon(1-\varepsilon) & 5+3\varepsilon+45\varepsilon^2-5\varepsilon^3 & 10+25\varepsilon^2+25\varepsilon^3 \end{pmatrix},
\end{aligned}$$

where

$$\mathbf{c}'(\varepsilon) = \frac{1}{1+2\varepsilon^2} \qquad \text{and} \qquad \mathbf{c}''(\varepsilon) = \frac{1}{5(1+5\varepsilon^2)}.$$

Both matrices $\mathbf{C}'$ and $\mathbf{C}''$ differ from $\mathbf{C}$ by terms of order ε, and both have eigenvalues $1+\varepsilon$, $1-\varepsilon$ and 2. But, the matrices of eigenvectors of $\mathbf{C}'$ and $\mathbf{C}''$ normalized such that the largest element in each column is 1 are

$$\begin{pmatrix} 1 & 1 & -\varepsilon \\ 1 & -1 & \varepsilon \\ 0 & 2\varepsilon & 1 \end{pmatrix} \qquad \text{and} \qquad \begin{pmatrix} 1/2 & 1 & \varepsilon \\ 1 & -1/2 & 2\varepsilon \\ -5\varepsilon/2 & 0 & 1 \end{pmatrix}.$$

No matter how small is ε, the eigenvectors of $\mathbf{C}'$ and $\mathbf{C}''$ corresponding to the eigenvalues $1+\varepsilon$ and $1-\varepsilon$ differ by quantities of order 1. Hence, we cannot expect the eigenvectors of nearby matrices to lie near one another when their corresponding eigenvalues belong to clusters of poorly separated eigenvalues. The reason for this is that both $\mathbf{C}'$ and $\mathbf{C}''$ are near the matrix $\mathbf{C}$, which has the double eigenvalue 1. Since any vector in $U := \mathrm{span}\{(1,0,0)^T, (0,1,0)^T\}$ is an eigenvector of $\mathbf{C}$, two different perturbations will cause the plane U to coalesce into two different sets of two distinct eigenvectors.

This discussion has impact on the number of CG-steps when solving the linear system $\mathbf{Cx} = \mathbf{b}$ when preconditioned with $\mathbf{C}'$ or $\mathbf{C}''$. Note that for any right hand side by Theorem 3.10 the original system is solved by the CG-method in at most two steps with the exact solution. The matrices $\mathbf{C}'$ and $\mathbf{C}''$ could have been constructed by knowledge of inexact spectral data about $\mathbf{C}$. In general for sufficiently small $\varepsilon > 0$ the matrices $(\mathbf{C}')^{-1}\mathbf{C}$ and $(\mathbf{C}'')^{-1}\mathbf{C}$ have three distinct eigenvalues. Then we expect the CG-method to perform three steps to solve $(\mathbf{C}')^{-1}\mathbf{C}x = (\mathbf{C}')^{-1}\mathbf{b}$ for any arbitrary right hand side $\mathbf{b} \in \mathbb{R}^3$.

This example illustrates that spectral preconditioners constructed by approximations to poorly separated or multiple eigenvalues of the original matrix possibly destroy convergence properties of the CG-method.

5.4 The regularized and preconditioned Newton equation

Let us consider the consequences of the results formulated in Section 5.3 to the preconditioned and regularized linear systems (4.20) arising in a frozen IRGNM for example given by Algorithm 4.10. To this end corresponding to (4.14) we define a preconditioner by exact spectral data

$$\mathbf{M}_n^{\text{exc}} = \gamma_n \mathbf{I} + \sum_{j=1}^{k_n} \left(\frac{\mu_j}{\zeta} - \gamma_n \right) \mathbf{u}_j \mathbf{u}_j^T \tag{5.25}$$

and a preconditioner which is only constructed by Ritz pairs $(\theta_1, \mathbf{v}_1), \ldots, (\theta_{k_n}, \mathbf{v}_{k_n})$ approximating eigenpairs of $\mathbf{G}_n^T \mathbf{G}_n$,

$$\mathbf{M}_n^{\text{iexc}} = \gamma_n \mathbf{I} + \sum_{j=1}^{k_n} \left(\frac{\theta_j}{\zeta} - \gamma_n \right) \mathbf{v}_j \mathbf{v}_j^T. \tag{5.26}$$

Recall that $\mathbf{v}_i^T \mathbf{v}_j = \delta_{ij}$. Using results of Section 5.3 we now investigate the regularized and preconditioned linear system

$$(\mathbf{M}_n^{\text{iexc}})^{-1} \mathbf{G}_n^T \mathbf{G}_n \mathbf{h}_n = \mathbf{M}_n^{-1} \mathbf{G}_n^T \mathbf{g}_n^\delta \tag{5.27}$$

with respect to the condition number of the multiple eigenvalue ζ of $(\mathbf{M}_n^{\text{exc}})^{-1} \mathbf{G}_n^T \mathbf{G}_n$. To this end let us define the matrices

$$\begin{aligned} \mathbf{M}_n^{\text{out}} &:= \mathbf{I} + \sum_{j=1}^{k_n} \left(\frac{\mu_j}{\zeta} - 1 \right) \mathbf{u}_j \mathbf{u}_j^T, \\ \mathbf{M}_n^{\text{clu}} &:= \mathbf{I} + \sum_{j=k_n+1}^{N} (\gamma_n - 1) \, \mathbf{u}_j \mathbf{u}_j^T. \end{aligned}$$

The computation

$$\begin{aligned} \mathbf{M}_n^{\text{clu}} \mathbf{M}_n^{\text{out}} &= \mathbf{I} + \sum_{j=k_n+1}^{N} (\gamma_n - 1) \, \mathbf{u}_j \mathbf{u}_j^T + \sum_{j=1}^{k_n} \left(\frac{\mu_j}{\zeta} - 1 \right) \mathbf{u}_j \mathbf{u}_j^T \\ &= \sum_{j=1}^{k_n} \left(\frac{\mu_j}{\zeta} \right) \mathbf{u}_j \mathbf{u}_j^T + \sum_{j=1}^{N} \gamma_n \mathbf{u}_j \mathbf{u}_j^T - \sum_{j=1}^{k_n} \gamma_n \mathbf{u}_j \mathbf{u}_j^T \\ &= \gamma_n \mathbf{I} + \sum_{j=1}^{k_n} \left(\frac{\mu_j}{\zeta} - \gamma_n \right) \mathbf{u}_j \mathbf{u}_j^T \end{aligned}$$

shows the factorization $\mathbf{M}_n^{\text{exc}} = \mathbf{M}_n^{\text{clu}} \mathbf{M}_n^{\text{out}}$. Hence, $\mathbf{M}_n^{\text{exc}}$ can be written as a composition of a deflation preconditioner given by $\mathbf{M}_n^{\text{out}}$ and a coarse grid preconditioner

given by $\mathbf{M}_n^{\text{clu}}$. In particular, $\mathbf{M}_n^{\text{out}}$ is of the form (5.6). So, the results of Theorem 5.4 and Corollary 5.7 can be applied to $\mathbf{M}_n^{\text{out}}$ if the corresponding eigenvalues of $\mathbf{G}_n^T\mathbf{G}_n$ are simple, what we assume in the following.

Corollary 5.11 *Assume that $\zeta \notin \sigma(\mathbf{G}_n^T\mathbf{G}_n)$. Then the matrix $(\mathbf{M}_n^{\text{out}})^{-1}\mathbf{G}_n^T\mathbf{G}_n$ has the multiple eigenvalue ζ with multiplicity k_n. Moreover,*

$$\text{cond}(\zeta) \leq 2\|\mathbf{Y}\| + \|\mathbf{J}\|, \tag{5.28}$$

where $\mathbf{Y} = (\mathbf{y}_{\ell s}) \in \mathbb{R}^{k_n \times k_n}$ and $\mathbf{J} \in \mathbb{R}^{k_n \times k_n}$

$$\mathbf{y}_{\ell\ell} = 0, \qquad \mathbf{y}_{\ell s} = \frac{\zeta - (\gamma_n + \lambda_s)}{\lambda_s - \lambda_\ell}\sqrt{\frac{\gamma_n + \lambda_\ell}{\gamma_n + \lambda_s}} \qquad \textit{for} \qquad \ell \neq s,$$

$\mathbf{J} = -\text{diag}(\zeta/(\gamma_n + \lambda_1), \ldots, \zeta/(\gamma_n + \lambda_{k_n}))$.

Proof: The first assertion follows from the computation

$$\begin{aligned}(\mathbf{M}_n^{\text{out}})^{-1}\mathbf{G}_n^T\mathbf{G}_n &= \gamma_n\mathbf{I} + \sum_{j=1}^{N}\lambda_j\mathbf{u}_j\mathbf{u}_j^T + \sum_{j=1}^{k_n}(\gamma_n + \lambda_j)\left(\frac{\zeta}{\mu_j} - 1\right)\mathbf{u}_j\mathbf{u}_j^T \\ &= \sum_{j=1}^{k_n}\zeta\mathbf{u}_j\mathbf{u}_j^T + \sum_{j=k_n+1}^{N}(\gamma_n + \lambda_j)\mathbf{u}_j\mathbf{u}_j^T,\end{aligned}$$

and the assumption on ζ. The estimate (5.28) is a consequence of Corollary 5.7 and Theorem 5.4.

□

To discuss the consequences of Corollary 5.11 first note that $\mathbf{M}_n^{\text{clu}}$ does not influence the multiple eigenvalue ζ and therefore not its condition number. We now consider the two cases $\gamma_n \gg 0$ corresponding to the starting phase and $\gamma_n \approx 0$ corresponding to the final phase of the IRGNM. For sufficiently large $S, N \in \mathbb{N}$ we can assume that the eigenvalues of $\mathbf{A}_n^T\mathbf{A}_n$ satisfy (4.21) (see also Section 5.1). As a consequence, picking up the idea of Definition 4.12 we distinguish the eigenvalues of $\mathbf{G}_n^T\mathbf{G}_n$ through

a) the well separated outliers $\gamma_n + \lambda_1 \gg \ldots \gg \gamma_n + \lambda_{k_n}$ and

b) the cluster of eigenvalues $\{\gamma_n + \lambda_j : j > k_n\}$.

Let us assume in the following that $\zeta > 0$ is not too far outside the spectrum of $\mathbf{G}_n^T\mathbf{G}_n$ and note that due to (4.26) ζ cannot be chosen arbitrary small. Consider first the case $\gamma_n \gg 0$. Then by Corollary 5.11 $\|\mathbf{J}\|$ is small and therefore cond(ζ) is small if

$$\frac{\zeta - (\gamma_n + \lambda_s)}{\lambda_s - \lambda_\ell}\sqrt{\frac{\gamma_n + \lambda_\ell}{\gamma_n + \lambda_s}} \approx \frac{\zeta}{\lambda_s - \lambda_\ell}\sqrt{\frac{\gamma_n + \lambda_\ell}{\gamma_n + \lambda_s}}, \qquad \ell \neq s, \tag{5.29}$$

is small. Here we have assumed that $\gamma_n + \lambda_s$ if far away from ζ, which seems to be a reasonable assumption, as otherwise we would not have targeted this eigenvalue. Moreover, we can conclude that if we just use Ritz pairs $(\theta_j, \mathbf{v}_j)$ corresponding to the well separated outliers in the spectrum of $\mathbf{G}_n^T\mathbf{G}_n$ to construct the preconditioner the quantity (5.29) is small because the gaps $|\lambda_s - \lambda_\ell|$, $s \neq \ell$, do not become arbitrary small and the quotients $(\gamma_n+\lambda_\ell)/(\gamma_n+\lambda_s) \leq 1+\lambda_\ell/\gamma_n$, $s \neq \ell$, do not explode by the assumption on γ_n. Hence, the preconditioned operator $(\mathbf{M}_n^{\text{out}})^{-1}\mathbf{G}_n^T\mathbf{G}_n$ has several eigenvalues in a neighborhood of ζ. This usually leads to a significant reduction on the number of CG-steps required to satisfy some stopping criterion. For exact spectral information this was proven in Theorem 4.19.
Assume there exist multiple outliers in the spectrum of $\mathbf{G}_n^T\mathbf{G}_n$ or some of the outliers are not well separated. If we use Ritz pairs $(\theta_j, \mathbf{v}_j)$ corresponding to these outliers or Ritz pairs corresponding to eigenvalues in the cluster of $\mathbf{G}_n^T\mathbf{G}_n$ the quantity (5.29) explodes and we cannot exclude an instability. This possibly slows down the convergence rate of the preconditioned CG-method although the regularization parameter satisfies $\gamma_n \gg 0$. On the other hand, in practice where Ritz pairs $(\theta_j, \mathbf{v}_j)$ are computed by Lanczos' method it is hard to decide if determined Ritz values correspond to a multiple eigenvalue. To get an impression we refer to the Tables 7.1 and 7.3, where Ritz values from practical computations are plotted. As one can observe the size of the small Ritz values is about $1e-12$ and therefore the difference is about $1e-12$. Hence, in practice it is impossible to decide if the corresponding eigenvalues are multiple and if one of those Ritz values is just a bad approximation or if the corresponding eigenvalues are simple.

In the case where $\gamma_n \approx 0$ Theorem 4.19 implies that a lot of spectral information is required for the preconditioner (5.26) to be efficient. Therefore, the index k_n needs to be chosen larger. Since due to (4.21) the eigenvalues decay rapidly the intersection between the outliers and the cluster coalesces. Hence, $\|\mathbf{Y}\|$ and $\|\mathbf{J}\|$ possibly explodes if Ritz pairs corresponding to eigenvalues $\lambda_j \approx 0$ are used to set up the preconditioner. This can impair the convergence behavior of the preconditioned CG-method.

In practice normally the approximation property of the Ritz pairs $(\theta_j, \mathbf{v}_j)$ corresponding to the largest eigenvalues and eigenvectors of $\mathbf{G}_n^T\mathbf{G}_n$ are far better than the approximations to eigenvalues in a neighborhood of γ_n. Hence, to ensure stability and good convergence rates of the preconditioned CG-method the preconditioner $\mathbf{M}_n^{\text{iexc}}$ should be set up only by Ritz pairs where θ_j is not in a small neighborhood of γ_n. Unfortunately, in real computations it is hard to define a small neighborhood of γ_n, in particular if γ_n is small. Again, as impression for this serve the Tables 7.1 and 7.3. Choosing the neighborhood too large possibly means to lose a lot of valuable information which could be used to construct an efficient preconditioner, choosing it too small possibly leads to instabilities impairing the convergence behavior.

Anyways, we can combine estimate (5.23) and equality (3.39). For the construction of the preconditioner we will use only those Ritz pairs where the residual $\|(\mathbf{M}_n^{\text{iexc}})^{-1}\mathbf{G}_n^T\mathbf{G}_n\mathbf{v}_j - \theta_j\mathbf{v}_j\|$ is sufficiently small. As we can see from the Tables 7.1 and 7.3 for the outliers in the spectrum these residuals become actually rather small, whereas for the Ritz values in a neighborhood of γ_n the approximation quality impairs.

As a consequence of the discussion above we can conclude that in particular in the case where the regularization parameter γ_n is small the linear system (5.27) has a tendency to instabilities because of two main reasons:

a) Usually the approximations of Lanczos' method to small and clustered eigenvalues are of poor quality.

b) Clustered and poorly separated eigenvalues targeted by the preconditioner $\mathbf{M}_n^{\text{iexc}}$ are very sensitive to errors in the approximations.

Unfortunately, the first reason amplifies the second. As indicated in the introduction to this chapter although the systems (5.27) are formally well-posed for small γ_n the linear systems are ill-conditioned, which has an effect on the construction of the preconditioner. In this case it is recommended to use only approximations of high quality for constructing the preconditioner. In this sense this chapter also serves as a step into a better understanding for the design of preconditioners suited for ill-posed problems, in particular for linear systems arising from Tikhonov regularization.

We want to close this chapter with a final remark: This discussion shall serve as a warning to be careful with the selection of the Ritz values for the construction of the preconditioner. Even if the Ritz values and therefore the corresponding eigenvalues are rather small, in the numerical examples we considered preconditioning with the spectral preconditioner $\mathbf{M}_n^{\text{iexc}}$ worked quite well and usually led in practice to a significant reduction of the total complexity (see for example Figure 7.4), once again illustrating their adequateness for large-scale exponentially ill-posed problems. On the other hand, the analysis in this chapter also shows limitations of preconditioners of the form of $\mathbf{M}_n^{\text{iexc}}$.

Chapter 6

A preconditioned Newton method

In order to obtain a stable approximation to the solution $x^\dagger$ of the nonlinear ill-posed operator equation (2.1) we derive in this chapter a numerical realization of Algorithm 4.10. In particular our interest is dedicated to an efficient solver for large-scale nonlinear ill-posed problems reducing significantly the total complexity when compared to a standard IRGNM or Levenberg-Marquardt algorithm. Naturally, the final iterates should be comparable. As in the last chapter, dealing with an implementation we restrict ourselves to finite dimensional linear systems.

The realization of Algorithm 4.10 is based on the close connection of the CG-method and Lanczos' method as presented in Chapter 3. Its fundamental idea can be summarized shortly as followed: We compute in each step of the IRGNM by the CG-method an approximation to the solution of the occurring linear system (5.1) until the stopping criterion (4.11) is satisfied. In some Newton steps we additionally determine by Lanczos' method approximations to eigenpairs of $\mathbf{G}_n^T\mathbf{G}_n$, which we use to construct a preconditioner of the form (5.26) for the matrices $\mathbf{G}_{n,i}^T\mathbf{G}_{n,i}$ in the following Newton steps.

To analyze Algorithm 4.10 theoretically we had made simplifying assumptions such as the knowledge of the largest eigenvalues with corresponding eigenvectors. Naturally, for a realization we are not in a position to assume any simplifications. This causes several consequences, which need to be incorporated into a numerical realization to ensure fast convergence of the inner preconditioned CG-iterations. Hence, it is one of the goals of this chapter to indicate to the difficulties arising in a realization of Algorithm 4.10 and methods to deal with them.

6.1 Fundamentals

In this section we list the main aspects which need to be considered for an implementation of Algorithm 4.10 and their resulting consequences. Subsequently we will present solutions to deal with them taking into account the theory presented so far

in this thesis. Finally, the methods will be combined to a *preconditioned Newton method*.

Let us start with the general framework and main difficulties for an implementation of such an algorithm.

a) For all $n \in \mathbb{N}_0$ the matrices $\mathbf{A}_n$ and $\mathbf{A}_n^T$ are unknown. Only a "black box" evaluating $\mathbf{A}_n\mathbf{x}$ and $\mathbf{A}_n^T\mathbf{y}$ for some given vectors $\mathbf{x} \in \mathbb{R}^N$ and $\mathbf{y} \in \mathbb{R}^S$ is available.

b) Solving the linear system (5.1) is a complex task, in particular for small γ_n (see Theorem 4.19). To reduce the complexity it is desirable to solve the preconditioned system (5.27) instead. To this end we need to incorporate into the IRGNM a cheap method to compute Ritz pairs of $\mathbf{A}_n^T\mathbf{A}_n$, for instance Lanczos' method.

c) Usually round-off errors cause loss of orthogonality in the residual vectors $\mathbf{r}^j$ and $\mathbf{z}^j$ computed in Algorithm 3.5 or 3.6. This loss of orthogonality is closely related to the convergence of the Ritz vectors.

d) To ensure efficiency of the preconditioner by Corollaries 5.7 and 5.11 it is necessary that the approximations of the Ritz pairs used to set up the preconditioner are of high quality.

e) Recall from Section 3.5 that after having performed k CG-steps we are able to compute by Lanczos' method approximations to k eigenpairs of $\mathbf{G}_n^T\mathbf{G}_n$. Unfortunately, it remains open which eigenvalues of $\mathbf{G}_n^T\mathbf{G}_n$ are approximated (see Theorem 3.14). Moreover, because the approximations depend on $\mathbf{g}_n^\delta$ we possibly do not even approximate the largest eigenvalue of $\mathbf{G}_n^T\mathbf{G}_n$ and in the presence of multiple or not well separated eigenvalues Theorem 3.15 indicates that the convergence to this eigenvalue can be rather slow.

f) If the matrix $\mathbf{G}_n^T\mathbf{G}_n$ has multiple large eigenvalues by Theorem 3.10 Lanczos' method approximates only one Ritz pair corresponding to this multiple eigenvalue.

g) During Newton's method the regularization parameter γ_n tends to zero (see (1.25)). Hence, given a fixed number $k_{ev} \in \mathbb{N}_0$ of known eigenpairs of $\mathbf{A}_n^T\mathbf{A}_n$ we expect due to Theorem 4.19 that the number of CG-steps will increase rapidly after a few Newton steps.

The conclusions of points a) – d) can be discussed straightforward. To handle the consequences arising from e) – g) is more difficult and Section 6.2 is dedicated to this topic.

To a) As a consequence of remark a) an approximate solution of (5.1) and (5.27) can only be obtained by matrix-free iterative methods. Therefore, for an efficient implementation of such a method it is of basic importance that the evaluations $\mathbf{A}_n\mathbf{x}$ and $\mathbf{A}_n^T\mathbf{y}$ can be efficiently implemented. Throughout this thesis we consider the CG-method and its preconditioned version as iterative methods to solve (5.1) and (5.27), since it is a natural choice for self-adjoint and strictly coercive operators.

To b) We only want to justify that in our situation Lanczos' method is the most efficient way to determine spectral data of $\mathbf{A}_n^T\mathbf{A}_n$. To this end note that the additional complexity we have to invest to determine spectral data needs to be saved in the following Newton steps. At the outset several iterative methods, for example the power method and the inverse Rayleigh iteration are also possible choices to approximate eigenpairs. But usually these methods require many iterations until the approximates are of high quality yielding many evaluations of $\mathbf{A}_n\mathbf{x}$ and $\mathbf{A}_n^T\mathbf{y}$ causing too much complexity.

The close connection of the CG-method and Lanczos' method avoids this drawback. We have to solve the linear system (5.1) to compute an update for Newton's method anyway. Therefore, the additional complexity to determine Ritz pairs by Lanczos' method is negligible. Still, some additional complexity has to be invested. In the Newton steps where spectral data is computed we choose a sufficiently small $\varepsilon > 0$ in Algorithm 3.5 yielding approximations of higher quality. Such a proceeding is recommended and usually profitable due to the sensitivity analysis presented in Chapter 5.

To c) Recall from Section 3.5 that the Ritz values and Ritz vectors computed by Lanczos' method depend on the residuals $\mathbf{r}^j$ and $\mathbf{z}^j$ of the CG-method (see Section 3.5). The residual vectors are orthogonal by construction (see Theorem 3.2). Unfortunately, it is a well known fact that there is loss of orthogonality among the residuals $\mathbf{r}^j$ and $\mathbf{z}^j$ while performing the CG-method. Moreover, it is crucial that orthogonality is well maintained until one of the Ritz vectors starts to converge (see [71, Chapter 13]). Naturally, this loss of orthogonality does not only destroy the convergence of the CG-method, but also the approximation quality of the Ritz values and Ritz vectors. Several reorthogonalization algorithms have been proposed in the literature to regain orthogonality (see for example [12, Chapter 7]). In our implementation we use a complete reorthogonalization scheme as proposed in [24], which is based on Householder transformations.

Algorithm 6.1 (Complete reorthogonalization)

- Input: $\mathbf{z}^0, \ldots, \mathbf{z}^k$, the set of vectors, which need to be reorthogonalized;
- Determine Householder matrix $\mathbf{H}_0$ such that $\mathbf{H}_0\mathbf{z}^0 = \mathbf{e}_1$;

- for $j = 1, 2, \ldots$
  - $\mathbf{w} = \mathbf{H}_{j-1} \cdots \mathbf{H}_0 \mathbf{z}^j$;
  - Determine Householder matrix $\mathbf{H}_j$ such that

$$\mathbf{H}_j \mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_j, \|\mathbf{z}^j\|_2, 0 \ldots, 0)^T;$$

  - $\mathbf{z}^j := \mathbf{H}_0 \cdots \mathbf{H}_j \mathbf{e}_j$;

Note that in the CG-method the vectors $\mathbf{z}^j$, $j = 0, 1, \ldots, k$, are computed successively. Algorithm 6.1 avoids a rapid loss of convergence, but it does not prevent it altogether. Thus, in addition it turned out to be necessary to compute the angle between $\mathbf{z}^j$ and the subspace $\text{span}\{\mathbf{z}^0, \ldots, \mathbf{z}^{j-1}\}$. This angle acts as an indicator of the loss of orthogonality. If the angle is too large we stop the CG-algorithm.

To d) In order to ensure that the Ritz pairs used for setting up the preconditioner are good approximations we choose a sufficiently small $\varepsilon > 0$ in the stopping criterion of Algorithm 3.5. This leads to a large number of CG-steps implying on the one hand a refinement of the Ritz pairs, but on the other hand an increase of the complexity. Motivated by the discussion in Sections 5.3 and 5.4 we expect that additional accuracy in the Ritz pairs saves computational cost in the following Newton steps which is supported by numerical examples. Using Theorem 3.14 we can determine the quantities (3.39) indicating the approximation quality of the Ritz pairs. Theorem 5.4 and Corollary 5.7 imply to use only those Ritz pairs satisfying (3.39) for some given upper bound. We will include this important fact into the algorithms presented in the next section.

6.2 Iterated Lanczos' method

This section is devoted to discuss the consequences of remarks e) – g). To sustain efficiency of the spectral preconditioner for several successive Newton steps it is important to take care of these points. Furthermore, they illustrate once more that the assumptions formulated to prove the complexity result in Theorem 4.21 are in practice too restrictive.

As noted in e), there are no strict results which eigenvalues are approximated by Lanczos' method. Since our theory is motivated by the fact that we have knowledge of the largest eigenvalues, we need a justification that for our class of problems Lanczos' method tends to target these eigenvalues.
Indeed, usually Lanczos' method approximates outliers in the spectrum of a given matrix very well, while eigenvalues in the bulk of the spectrum are typically harder

to approximate and are normally of poor quality. This behavior could be theoretically confirmed (see [53]) and is supported by numerical experience (see [12, Chapter 7]). Moreover, usually the more isolated an eigenvalue is, the better the approximation is. A convergence rate for the largest eigenvalue has been proven in Theorem 3.15.

Dealing with ill-posed problems we can take advantage of this behavior of Lanczos' method, in particular for exponentially ill-posed problems. Due to the rapid decay behavior (4.21b) of the eigenvalues the spectrum of $\mathbf{G}_n^T\mathbf{G}_n$ consists of a small number of large eigenvalues and a cluster at γ_n. In the preconditioned case the matrix $(\mathbf{M}_n^{\text{exc}})^{-1}\mathbf{G}_n^T\mathbf{G}_n$ has a similar eigenvalue distribution where the cluster is at one. Recall that in the preconditioned case we always consider the inner product which is implied by the preconditioner (see Lemma 3.3). Hence, we expect some of the Ritz pairs to be high quality approximations to the largest eigenvalues and corresponding eigenvectors of $\mathbf{G}_n^T\mathbf{G}_n$ and $(\mathbf{M}_n^{\text{exc}})^{-1}\mathbf{G}_n^T\mathbf{G}_n$. Theorem 3.15 even implies that the approximation quality to the largest eigenvalue increases rapidly with each CG-step. Due to round-off errors we can assume that we will always approximate the largest eigenvalue. But we cannot expect to get an approximation of high quality to each of the outliers. If this is the case not every outlier in the spectrum is shifted into a neighborhood of ζ by the preconditioner, an undesirable effect usually causing more CG-steps to solve the corresponding linear system. Lemma 6.2 which follows below is the key to handle this situation.

We now discuss f). In the case where outliers in the spectrum of $\mathbf{A}_n^T\mathbf{A}_n$ are multiple eigenvalues or not well separated, by Theorem 3.10 Lanczos' method will either approximate only one of each of these multiple eigenvalues or the approximations are poor caused for example by round-off errors. By Theorem 5.4, Corollary 5.7 and Example 5.10 usage of these approximations for preconditioning is in general not recommended. Unfortunately, we cannot check while performing the algorithm if we approximate a multiple eigenvalue. Thus, our only choice is to decide by criterion (3.39) if we use these approximations. Therefore at most one of each multiple eigenvalue is shifted into a neighborhood of ζ in the preconditioned case and we expect that the preconditioned CG-method performs one step for each multiple eigenvalue. Hence, in the case where several of the outliers in the spectrum of $\mathbf{A}_n^T\mathbf{A}_n$ are multiple eigenvalues the preconditioner does not reduce the complexity significantly, an unpleasant effect. In this situation further spectral data is required.

To consider the consequences of remark g) assume we have exact knowledge of the $k_{ev} \in \{1, \ldots, N\}$ largest eigenvalues of $\mathbf{A}_n^T\mathbf{A}_n$, where k_{ev} is fixed. Since γ_n tends to zero by (4.39) and (4.41) we expect after a few Newton steps depending on the number k_{ev} a rapid increase of the number of CG-steps required to solve the preconditioned linear system (5.27).
To avoid an explosion of the total complexity to solve (2.1) a method to update the preconditioner is necessary. The following observation shows such a way.

Lemma 6.2 *Let $J \subset \{1, \dots N\}$, $J \neq \emptyset$ and*

$$\mathbf{M}_n^{\text{exc}} := \gamma_n \mathbf{I} + \sum_{j \in J} \left(\frac{\mu_j}{\zeta} - \gamma_n \right) \mathbf{u}_j \mathbf{u}_j^T.$$

For $i \in \mathbb{N}_0$ we denote by η_j, $j = 1, \dots, N$, the eigenvalues of the preconditioned matrix $(\mathbf{M}_{n+i}^{\text{exc}})^{-1} \mathbf{G}_{n,i}^T \mathbf{G}_{n,i}$ with multiplicity and corresponding orthonormal eigenvectors $\tilde{\mathbf{u}}_j$, where $\mathbf{G}_{n,i}$ is given by (5.2). Assume that $\zeta \notin \sigma(\mathbf{G}_{n,i}^T \mathbf{G}_{n,i})$.

a) $\zeta \in \sigma\left((\mathbf{M}_{n+i}^{\text{exc}})^{-1} \mathbf{G}_{n,i}^T \mathbf{G}_{n,i}\right)$ has multiplicity $\#J$.

b) Eigenvectors of $\mathbf{G}_{n,i}^T \mathbf{G}_{n,i}$ are also eigenvectors of $(\mathbf{M}_{n+i}^{\text{exc}})^{-1} \mathbf{G}_{n,i}^T \mathbf{G}_{n,i}$.

c) If a pair $(\mu_\ell, \mathbf{u}_\ell) \notin \{(\mu_j, \mathbf{u}_j) : j \in J\}$ then there exists an index $k \in \mathbb{N}$ such that $\mu_\ell = \gamma_{n+i} \eta_k$ and $\mathbf{u}_\ell = \tilde{\mathbf{u}}_k$.

Proof: Assertion a) follows by a similar computation as in the proof of Corollary 5.11, assertion b) is clear. Assume now $\mathbf{u}_\ell \notin \{\mathbf{u}_j : j \in J\}$. Note, by b) it follows that there exists an index $k \in \mathbb{N}$ such that $\mathbf{u}_\ell = \tilde{\mathbf{u}}_k$, then

$$\eta_k \tilde{\mathbf{u}}_k = (\mathbf{M}_{n+i}^{\text{exc}})^{-1} \mathbf{G}_{n,i}^T \mathbf{G}_{n,i} \tilde{\mathbf{u}}_k = (\mathbf{M}_{n+i}^{\text{exc}})^{-1} \mathbf{G}_{n,i}^T \mathbf{G}_{n,i} \mathbf{u}_\ell = \mu_\ell (\mathbf{M}_{n+i}^{\text{exc}})^{-1} \mathbf{u}_\ell = \frac{\mu_\ell}{\gamma_{n+i}} \mathbf{u}_\ell.$$

□

Remark 6.3 *To construct the preconditioner $\mathbf{M}_n^{\text{exc}}$ we do not assume any more to have knowledge of the largest eigenvalues but only of a subset.*

An application of Lemma 6.2 yields an iterated Lanczos' algorithm, which we exploit to update the preconditioner while performing the IRGNM. To describe the idea of this algorithm we assume that we have exact knowledge of some of the eigenvalues of $\mathbf{G}_n^T \mathbf{G}_n$, but at most one of each multiple eigenvalue. This corresponds to the situation where we have solved (5.1) by the CG-method and have computed by Lanczos' method approximations to the eigenpairs. Hence, using the notation of Lemma 6.2 the set J is such that $\mu_j \neq \mu_i$, $i, j \in J$, $i \neq j$. The corresponding set of eigenvalues we denote by

$$\mathcal{U}_{kn} := \{\mu_j : j \in J\}.$$

In particular we have $\mathcal{U}_{kn} \neq \emptyset$. Naturally, our algorithm shall include the case of multiple eigenvalues, although this may cause instabilities.
In the situation described above the preconditioned matrix $(\mathbf{M}_{n+i}^{\text{exc}})^{-1} \mathbf{G}_{n,i}^T \mathbf{G}_{n,i}$ in the $(n+i)$-th Newton step followed has the eigenvalues $\mathcal{U}_{ukn} \cup \{\zeta\}$ where

$$\mathcal{U}_{ukn} := \left\{ \frac{\mu_j}{\gamma_{n+i}} : j \in \{1, \dots, N\} \setminus J \right\}.$$

In particular, we observe that not only the remaining unknown largest eigenvalues which have not been found in the n-th Newton step, but also the small eigenvalues of $\mathbf{G}_n^T\mathbf{G}_n$ are amplified by the factor $1/\gamma_{n+i}$. So, when solving in the $(n+i)$-th Newton step the linear system

$$(\mathbf{M}_{n+i}^{\text{exc}})^{-1}\mathbf{G}_{n,i}^T\mathbf{G}_{n,i}\mathbf{h}_{n+i} = \mathbf{M}_{n+i}^{-1}\mathbf{G}_{n,i}^T\mathbf{g}_{n+i}^{\delta} \tag{6.1}$$

we expect Lanczos' method to target the largest eigenvalues of $(\mathbf{M}_{n+i}^{\text{exc}})^{-1}\mathbf{G}_{n,i}^T\mathbf{G}_{n,i}$ given by a subset of $\mathcal{U}_{ukn}$. Therefore, not only the remaining large eigenvalues of $\mathbf{G}_n^T\mathbf{G}_n$ are possibly found, but also originally small eigenvalues which are now large since they are amplified by the factor $1/\gamma_{n+i}$. By Lemma 6.2 the corresponding eigenvectors to these eigenvalues coincide with eigenvectors of $\mathbf{A}_n^T\mathbf{A}_n$ and if $\mu \in \mathcal{U}_{ukn}$ is an eigenvalue of $(\mathbf{M}_{n+i}^{\text{exc}})^{-1}\mathbf{G}_{n,i}^T\mathbf{G}_{n,i}$ we can compute the corresponding eigenvalue of $\mathbf{A}_n^T\mathbf{A}_n$ by the formula

$$\lambda_k = \gamma_{n+i}\,(\mu - 1) \tag{6.2}$$

for some $k \in \{1, \ldots, N\}$. Exploiting this proceeding we can detect further large and multiple eigenvalues of $\mathbf{A}_n^T\mathbf{A}_n$, which have not been approximated in the previous Newton steps. Before we can add these additional eigenvalues to the preconditioner we have to make sure that the corresponding normalized eigenvectors $\mathbf{u}$ do not satisfy

$$\mathbf{u} \in \operatorname{span}\{\mathbf{u}_j : j \in J\}.$$

Naturally, when solving the linear system (6.1) Lanczos' method possibly approximates the eigenvalue $\zeta \in \sigma\left((\mathbf{M}_{n+i}^{\text{exc}})^{-1}\mathbf{G}_{n,i}^T\mathbf{G}_{n,i}\right)$, which has multiplicity $\#J$ under the assumption that $\zeta \notin \sigma(\mathbf{G}_{n,i}^T\mathbf{G}_{n,i})$. A corresponding eigenvector is given by any $\mathbf{u} \in \operatorname{span}\{\mathbf{u}_j : j \in J\}$. When approximating ζ by Lanczos' method due to the multiplicity of ζ we expect that the approximation quality is poor. Moreover, if some computed Ritz value already lies in a neighborhood of ζ we would not target the corresponding eigenvalue in the next Newton steps and by the results formulated in Chapter 5 it is strongly recommended to omit this Ritz pair. Note that given only approximations for preconditioning we possibly determine some Ritz pairs corresponding to eigenvalues in a neighborhood of ζ. Concluding, before we can add $\mathbf{u}$ together with its corresponding eigenvalue computed by (6.2) to the preconditioner we must guarantee the orthogonality

$$\mathbf{u} \perp \operatorname{span}\{\mathbf{u}_j : j \in J\} \tag{6.3}$$

and that the corresponding eigenvalue does not lie in a neighborhood of ζ.

So far we have illustrated the basic idea how to update the preconditioner assuming we have exact spectral data. We now turn to the realistic situation where the spectral data of $\mathbf{A}_n^T\mathbf{A}_n$ is approximated by Ritz pairs.

Notation 6.4 *Assume that in the n-th Newton step the Ritz pairs*

$$\{(\theta_j^{(1)}, \mathbf{v}_j^{(1)}) : j \in J_1\},$$

where $J_1 = \{j_1, \ldots, j_k\} \subset \{1, \ldots, N\}$, $J_1 \neq \emptyset$, *of* $\mathbf{A}_n^T \mathbf{A}_n$ *have been determined and in the following Newton steps we solve (6.1) with* $\mathbf{M}_{n+i}^{\text{exc}}$ *replaced by*

$$\mathbf{M}_{n+i}^{\text{iexc}} = \gamma_{n+i}\, \mathbf{I} + \sum_{j \in J} \left(\frac{\gamma_{n+i} + \theta_j^{(1)}}{\zeta} - \gamma_{n+i} \right) \mathbf{v}_j^{(1)} (\mathbf{v}_j^{(1)})^T, \qquad i = 1, 2, \ldots. \tag{6.4}$$

Moreover, we assume that for some $k \in \mathbb{N}$ *we determine in the* $(n+k)$*-th Newton step Ritz pairs*

$$\{(\theta_j^{(2)}, \mathbf{v}_j^{(2)}) : j \in J_2\}$$

of the matrix $(\mathbf{M}_{n+k}^{\text{iexc}})^{-1} \mathbf{G}_{n,k}^T \mathbf{G}_{n,k}$, *where* $J_2 = \{\tilde{j}_1, \ldots, \tilde{j}_{\tilde{k}}\} \subset \{1, \ldots, N\}$ $J_2 \neq \emptyset$. *By* $\mathbf{V}_1$ *and* $\mathbf{V}_2$ *we denote the corresponding subspaces*

$$\mathbf{V}_1 := \operatorname{span}\{\mathbf{v}_j^{(1)} : j \in J_1\} \qquad \textit{and} \qquad \mathbf{V}_2 := \operatorname{span}\{\mathbf{v}_j^{(2)} : j \in J_2\}.$$

Dealing just with approximations in general there does neither exist a $\tilde{j}_n \in J_2$ such that $\mathbf{v}_{\tilde{j}_n}^{(2)} \perp \mathbf{V}_1$ nor a $\tilde{j}_{\tilde{n}} \in J_2$ such that $\mathbf{v}_{\tilde{j}_{\tilde{n}}}^{(2)} \in \mathbf{V}_1$ is satisfied. If our approximations are of high quality there will at least exist indices such that both relations will be approximately satisfied. Indicated by (6.3) we need a criterion to test if a vector $\mathbf{v} \in \{\mathbf{v}_j^{(2)} : j \in J_2\}$ is approximately orthogonal on the subspace $\mathbf{V}_1$. We denote this by

$$\mathbf{v} \perp_{\approx} \mathbf{V}_1. \tag{6.5}$$

There are several possibilities to formulate reasonable criterions. The first one, which is easy to implement and works quite well in practice consists in checking if $\theta_j^{(2)} \gg 1$ for some $j \in J_2$. More precisely, we can distinguish three cases:

i) $\theta_j^{(2)} \gg 1$ and $\theta_j^{(2)} \not\approx \zeta$,

ii) $\theta_j^{(2)} \approx 1$,

iii) $\theta_j^{(2)} \approx \zeta$.

If i) is satisfied the Ritz value $\theta_j^{(2)}$ is not an approximation to an eigenvalue in the cluster at one and the Ritz vector satisfies (6.5). Thus, we can add the Ritz pair to our preconditioner if the approximation quality is acceptable which we can check by (3.39). In the case of ii) Lanczos' method has approximated an eigenvalue in the cluster of the preconditioned operator usually yielding an approximation of low quality. iii) indicates that the corresponding Ritz vector does not satisfy (6.5). Naturally, if ii) or iii) is satisfied we do not use the corresponding Ritz pair for updating the preconditioner.
A second possibility to check (6.5) is to compute for $j \in J_2$ the angle α_j between $\mathbf{v}_j^{(2)}$ and $\mathbf{V}_1$. Again we can distinguish three cases:

i) $\alpha_j \approx \pi/2$,

ii) $\alpha_j \approx 0$,

iii) $\alpha_j \in (a, b)$ where $0 \ll a < b \ll \pi/2$.

If i) is satisfied we expect (6.5) to be satisfied. Testing with (3.39) and using (6.2) we may add this Ritz pair to our preconditioner. In the case of ii) and iii) we expect that (6.5) is not satisfied and therefore we do not add the Ritz pair to our preconditioner. Naturally, we can combine both methods to check (6.5).

A third method to check (6.5) is to compute the numerical rank of the matrices $(\mathbf{v}_{j_1}^{(1)}, \ldots, \mathbf{v}_{j_k}^{(1)}, \mathbf{v}) \in \mathbb{R}^{N\times(k+1)}$, $\mathbf{v} \in \{\mathbf{v}_j^{(2)} : j \in J_2\}$. This can be realized by Householder transformations. For details we refer to [13, Chapter 3].

Now assume we have determined a subset $\mathbf{V} \subset \{\mathbf{v}_j^{(2)} : j \in J_2\}$ such that (6.5) is satisfied for all $\mathbf{v} \in \mathbf{V}$ and $\mathbf{V} \neq \emptyset$. Since only (6.5) holds, we can decompose each $\mathbf{v} \in \mathbf{V}$ into $\mathbf{v} = \mathbf{v}' + \mathbf{v}''$, $\mathbf{v}' \neq 0$, $\mathbf{v}'' \neq 0$ with $\mathbf{v}' \in \mathbf{V}_1$ and $\mathbf{v}'' \in \mathbf{V}_1^{\perp}$. Thus, a complete reorthogonalization of the total set of eigenvectors $\{\mathbf{v}_{j_1}^{(1)}, \ldots, \mathbf{v}_{j_k}^{(1)}, \mathbf{v}\}$ is recommended. Since the set of vectors $\{\mathbf{v}_{j_1}^{(1)}, \ldots, \mathbf{v}_{j_k}^{(1)}\}$ is an orthogonal basis of of $\mathbf{V}_1$ it is a natural proceeding to reorthogonalize $\mathbf{v}$ against these vectors, which can be done by Algorithm 6.1.

Naturally, the result of Algorithm 6.1 differs if we change the order of the vectors. Therefore, we could also choose the order of the vectors by their approximation quality. As an indicator for the quality we can again use (3.39). Such a proceeding is justified by the fact that if the reorthogonalization process is started with approximations of low quality these vectors deteriorate the approximation quality of the remaining vectors. But numerical experience has shown that the first way is usually sufficient. If there are several vectors $\mathbf{v}_1^{(2)}, \ldots, \mathbf{v}_k^{(2)} \in \mathbf{V}$ which shall be added to the preconditioner they should be ordered by the size of the Ritz values, that is $\mathbf{v}_1^{(2)}$ should correspond to the largest Ritz value, $\mathbf{v}_2^{(2)}$ to the next smaller Ritz value and so on. Due to Theorem 3.10 this can always be realized.

Moreover, in general Ritz vectors determined by $\mathbf{G}_n^T\mathbf{G}_n$, that is when no preconditioner is applied, are far better approximations than those computed with a preconditioner. In the preconditioned case we have observed that Ritz vectors computed in former Newton steps are usually better approximations than those computed several Newton steps later (see Chapter 7 for an example). Both experiences correspond to Theorem 3.15 and to the observation that in the preconditioned case the outliers in the spectrum are not so well separated and closer to the cluster than in the non-preconditioned case. This observation again supports to reorthogonalize the new Ritz vectors against the old ones.

We summarize these ideas in the following algorithm.

Algorithm 6.5 (Iterated Lanczos algorithm I)

Input: $\Omega := \{(\theta_j, \mathbf{v}_j) : j = 1, \ldots, k\}$, approximations to eigenpairs of $\mathbf{A}_n^T \mathbf{A}_n$;

Solve $(\mathbf{M}_{n+i}^{\text{iexc}})^{-1} \mathbf{G}_{n,i}^T \mathbf{G}_{n,i} \mathbf{h}_{n+i} = (\mathbf{M}_{n+i}^{\text{iexc}})^{-1} \mathbf{G}_{n,i}^T \mathbf{g}_{n+i}^{\delta}$ by Algorithm 3.6 $\rightsquigarrow \mathbf{h}_{n+i}^{\text{app}}$;

Compute Ritz pairs $(\theta_1^{(2)}, \mathbf{v}_1^{(2)}), \ldots, (\theta_{\tilde{k}}^{(2)}, \mathbf{v}_{\tilde{k}}^{(2)})$ of $(\mathbf{M}_{n+i}^{\text{iexc}})^{-1} \mathbf{G}_{n,i}^T \mathbf{G}_{n,i}$ by Lanczos' method;

for each Ritz pair $(\theta_j^{(2)}, \mathbf{v}_j^{(2)})$ satisfying some approximation condition

$k := k + 1$;

if (6.5) is satisfied

$\theta_k := \gamma_{n+i}(\theta_j^{(2)} - 1)$, $\mathbf{v}_k := \mathbf{v}_j^{(2)}$;
$\Omega := \Omega \cup \{(\theta_k, \mathbf{v}_k)\}$;
$k = k + 1$;

Output: $\mathbf{h}_{n+i}^{\text{app}}$ and Ω, approximations to eigenpairs of $\mathbf{A}_n^T \mathbf{A}_n$;

We consider another method to update the preconditioner, which is slightly different from the one discussed above and motivated by the following intention. Maybe it is possible to use the new computed approximations in the $(n+i)$-th Newton step to the eigenpairs of $\mathbf{A}_n^T \mathbf{A}_n$ to improve the approximation quality of the Ritz pairs determined in former Newton steps. This method is based on the approximation property of the preconditioner $\mathbf{M}_{n+i}^{\text{iexc}}$ to the matrix $\mathbf{G}_{n,i}^T \mathbf{G}_{n,i}$. We consider only the case $\zeta = 1$. By Notation 6.4 we have

$$\mathbf{G}_{n,i}^T \mathbf{G}_{n,i} = \gamma_{n+i}\, \mathbf{I} + \sum_{j=1}^{N} \lambda_j \mathbf{u}_j \mathbf{u}_j^T \approx \gamma_{n+i}\, \mathbf{I} + \sum_{j \in J_1} \theta_j^{(1)} \mathbf{v}_j^{(1)} (\mathbf{v}_j^{(1)})^T = \mathbf{M}_{n+i}^{\text{iexc}}.$$

Using (4.16) and assuming that (6.5) is satisfied for all $\mathbf{v} \in \{\mathbf{v}_j^{(2)} : j \in J_2\}$ the approximation

$$(\mathbf{M}_{n+i}^{\text{iexc}})^{-1} \mathbf{G}_{n,i}^T \mathbf{G}_{n,i} \;\approx\; \sum_{j \in J_1} \mathbf{v}_j^{(1)} (\mathbf{v}_j^{(1)})^T + \sum_{j \in J_2} \theta_j^{(2)} \mathbf{v}_j^{(2)} (\mathbf{v}_j^{(2)})^T + \mathbf{P}_{\mathbf{U}^\perp}, \qquad (6.6)$$

follows, where $\mathbf{U}$ denotes the subspace

$$\mathbf{U} := \operatorname{span}\left\{ \mathbf{v}_{j_1}^{(1)}, \ldots, \mathbf{v}_{j_k}^{(1)}, \mathbf{v}_{\tilde{j}_1}^{(2)}, \ldots, \mathbf{v}_{\tilde{j}_{\tilde{k}}}^{(2)} \right\}.$$

and $\mathbf{P}_{\mathbf{U}^\perp}$ the orthogonal projection on $\mathbf{U}^\perp$. Multiplying (6.6) from the left by $\mathbf{M}_{n+i}^{\text{iexc}}$ we have

$$\begin{aligned}
\mathbf{G}_{n,i}^T\mathbf{G}_{n,i} &\approx \mathbf{M}_{n+i}^{\text{iexc}}\left(\sum_{j\in J_1}\mathbf{v}_j^{(1)}(\mathbf{v}_j^{(1)})^T + \sum_{j\in J_2}\theta_j^{(2)}\mathbf{v}_j^{(2)}(\mathbf{v}_j^{(2)})^T + \mathbf{P}_{\mathbf{U}^\perp}\right)\\
&\approx \gamma_{n+i}\sum_{j\in J_1}\mathbf{v}_j^{(1)}(\mathbf{v}_j^{(1)})^T + \sum_{j\in J_1}\theta_j^{(1)}\mathbf{v}_j^{(1)}(\mathbf{v}_j^{(1)})^T\\
&\quad + \sum_{j\in J_2}\gamma_{n+i}\theta_j^{(2)}\mathbf{v}_j^{(2)}(\mathbf{v}_j^{(2)})^T + \gamma_{n+i}\mathbf{P}_{\mathbf{U}^\perp}\\
&\approx \gamma_{n+i}\,\mathbf{I} + \sum_{j\in J_1}\theta_j^{(1)}\mathbf{v}_j^{(1)}(\mathbf{v}_j^{(1)})^T + \sum_{j\in J_2}\gamma_{n+i}(\theta_j^{(2)}-1)\mathbf{v}_j^{(2)}(\mathbf{v}_j^{(2)})^T =: \mathbf{C}. \quad (6.7)
\end{aligned}$$

Now, the idea is the following. Since $\mathbf{C}$ approximates $\mathbf{G}_{n,i}^T\mathbf{G}_{n,i}$ we assume that the eigenpairs of $\mathbf{C}$ approximate eigenpairs of $\mathbf{G}_{n,i}^T\mathbf{G}_{n,i}$. Unfortunately, the matrix $\mathbf{C}\in\mathbb{R}^{N\times N}$ is high dimensional and an eigendecomposition would be rather complex. Furthermore, by construction we expect that $\mathbf{C}$ is only a good approximation to $\mathbf{G}_{n,i}^T\mathbf{G}_{n,i}$ on the subspace $\mathbf{U}$. On the other hand, our only attention attracts to the eigenpairs on the subspace $\mathbf{U}$.

To this end we compute via a QU-decomposition an orthonormal basis $\mathbf{Q} := \left(\mathbf{q}_1,\ldots,\mathbf{q}_{k+\tilde{k}}\right)$ of the subspace $\mathbf{U}$ and determine the low dimensional matrix

$$\mathbf{B} := \mathbf{Q}^T\mathbf{C}\mathbf{Q} \in \mathbb{R}^{(k+\tilde{k})\times(k+\tilde{k})}$$

approximating $\mathbf{Q}^T\mathbf{G}_{n,i}^T\mathbf{G}_{n,i}\mathbf{Q}$. Note that the computation of $\mathbf{B}$ is realizable without setting up the high dimensional matrix $\mathbf{C}$.
Assume that $\tilde{\lambda}_1,\ldots,\tilde{\lambda}_{k+\tilde{k}}$ are the eigenvalues and $\tilde{\mathbf{v}}_1,\ldots,\tilde{\mathbf{v}}_{k+\tilde{k}}$ the corresponding orthonormal eigenvectors of $\mathbf{B}$. Then we have

$$\mathbf{G}_{n,i}^T\mathbf{G}_{n,i}\mathbf{Q}\tilde{\mathbf{v}}_j \approx \mathbf{C}\mathbf{Q}\tilde{\mathbf{v}}_j = \tilde{\lambda}_j\mathbf{Q}\tilde{\mathbf{v}}_j, \qquad j=1,\ldots,k+\tilde{k}.$$

Thus, $\tilde{\lambda}_1,\ldots,\tilde{\lambda}_{k+\tilde{k}}$ approximate the eigenvalues and $\mathbf{Q}\tilde{\mathbf{v}}_1,\ldots,\mathbf{Q}\tilde{\mathbf{v}}_{k+\tilde{k}}$ the corresponding eigenvectors of $\mathbf{G}_{n,i}^T\mathbf{G}_{n,i}$. For $s\geq n+i$ the update of the preconditioner is given by

$$\mathbf{M}_s := \gamma_s\mathbf{I} + \sum_{j=1}^{k+\tilde{k}}(\tilde{\lambda}_j-\gamma_{n+i})\mathbf{w}_j\mathbf{w}_j^T, \qquad \mathbf{w}_j := \mathbf{Q}\tilde{\mathbf{v}}_j.$$

These considerations lead to the following algorithm.

Algorithm 6.6 (Iterated Lanczos algorithm II)

Input: $\Omega := \{(\theta_j,\mathbf{v}_j) : j=1,\ldots,k\}$, approximations to eigenpairs of $\mathbf{A}_n^T\mathbf{A}_n$;

$k_{old} := k$;

Solve $(\mathbf{M}_{n+i}^{\text{iexc}})^{-1}\mathbf{G}_{n,i}^T\mathbf{G}_{n,i}\mathbf{h}_{n+i} = (\mathbf{M}_{n+i}^{\text{iexc}})^{-1}\mathbf{G}_{n,i}^T\mathbf{g}_{n+i}^{\delta}$ by Algorithm 3.6 $\rightsquigarrow \mathbf{h}_{n+i}^{\text{app}}$;

Compute Ritz pairs $(\theta_1^{(2)}, \mathbf{v}_1^{(2)}), \ldots, (\theta_{\tilde{k}}^{(2)}, \mathbf{v}_{\tilde{k}}^{(2)})$ of $(\mathbf{M}_{n+i}^{\text{iexc}})^{-1}\mathbf{G}_{n,i}^T\mathbf{G}_{n,i}$ by Lanczos' method;

$\mathbf{U} := \{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$, $\Theta := \{\theta_1, \ldots, \theta_k\}$;

for each Ritz pair $(\theta_j^{(2)}, \mathbf{v}_j^{(2)})$ satisfying some approximation condition

$k := k + 1$;

if (6.5) is satisfied

$\mathbf{v}_k := \mathbf{v}_j^{(2)}$, $\theta_k := \theta_j^{(2)}$;
$\mathbf{U} := \mathbf{U} \cup \{\mathbf{v}_k\}$, $\Theta := \Theta \cup \{\theta_k\}$;
$k := k + 1$;

$k := k - 1$;

Compute orthonormal basis $\mathbf{Q} := (\mathbf{q}_1, \ldots, \mathbf{q}_k)$ of $\text{span}(\mathbf{U})$;

Determine $\mathbf{B} := \mathbf{Q}^T\mathbf{C}\mathbf{Q} \in \mathbb{R}^{k\times k}$ where

$$\mathbf{C} := \gamma_{n+i}\,\mathbf{I} + \sum_{j=1}^{k_{old}} \theta_j \mathbf{v}_j\mathbf{v}_j^T + \sum_{j=k_{old}+1}^{k} \gamma_{n+i}(\theta_j - 1)\mathbf{v}_j\mathbf{v}_j^T;$$

Compute eigenvalues $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_k$ with corresponding eigenvectors $\tilde{\mathbf{v}}_1, \ldots, \tilde{\mathbf{v}}_k$ of $\mathbf{B}$;

Output: $\mathbf{h}_{n+i}^{\text{app}}$ and $\Omega := \left\{\left(\tilde{\lambda}_1 - \gamma_{n+i}, \mathbf{Q}\tilde{\mathbf{v}}_1\right), \ldots, \left(\tilde{\lambda}_k - \gamma_{n+i}, \mathbf{Q}\tilde{\mathbf{v}}_k\right)\right\}$, approximations to eigenpairs of $\mathbf{A}_n^T\mathbf{A}_n$;

Remark 6.7 *The algorithm formulated above can be interpreted as the Rayleigh-Ritz method we described in Section 3.6 applied to the matrix* $\mathbf{C}$*. The matrix* S_k *given by (3.36) corresponds to* $\mathbf{B}$*. Hence, Theorem 3.12 and Corollary 3.13 hold true justifying to interpret for* $j = 1, \ldots, k$ *the pairs* $\left(\tilde{\lambda}_j, \mathbf{Q}\tilde{\mathbf{v}}_j\right)$ *as approximations to eigenpairs of* $\mathbf{C}$*. Since* $\mathbf{C} \approx \mathbf{G}_{n,i}^T\mathbf{G}_{n,i}$ *the elements of* Ω *approximate eigenpairs of* $\mathbf{A}_n^T\mathbf{A}_n$*.*

The Algorithms 6.5 and 6.6 give us the possibility to determine further approximations to eigenpairs of $\mathbf{A}_n^T\mathbf{A}_n$, which we use to make the preconditioner more efficient during Newton's method. It is only left to discuss the approximation condition which the Ritz pairs should satisfy. To this end recall *the notation of*

Theorem 3.14, that is using the pseudo-residuals $\mathbf{z}^j$, $j = 0, 1, \ldots, \tilde{k}-1$, occurring in Algorithm 3.6 we define $\tilde{\mathbf{z}}^j := \mathbf{z}^j / \|\mathbf{z}^j\|_{\mathbf{M}_n^{\text{iexc}}}$ and the matrix

$$\mathbf{Z}_{\tilde{k}} = (\tilde{\mathbf{z}}^0, \ldots, \tilde{\mathbf{z}}^{\tilde{k}-1}) \in \mathbb{R}^{N \times \tilde{k}}.$$

If $\mathbf{W}\mathbf{\Lambda}\mathbf{W}^T$ is an eigendecomposition of the matrix $\mathbf{T}_k = \mathbf{Z}_{\tilde{k}}^T (\mathbf{M}_{n+i}^{\text{iexc}})^{-1} \mathbf{G}_{n,i}^T \mathbf{G}_{n,i} \mathbf{Z}_{\tilde{k}}$, that is $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_{\tilde{k}})$ is orthogonal and $\mathbf{\Lambda} = \text{diag}(\theta_1^{(2)}, \ldots, \theta_{\tilde{k}}^{(2)})$ by (3.39) we have

$$\|\mathbf{G}_{n,i}^T \mathbf{G}_{n,i} (\mathbf{Z}_{\tilde{k}} \mathbf{w}_i) - (\mathbf{Z}_{\tilde{k}} \mathbf{w}_i) \theta_i^{(2)}\| = \frac{\sqrt{\beta_{\tilde{k}}}}{\alpha_{\tilde{k}}} |\mathbf{w}_i(\tilde{k})|, \qquad i = 1, \ldots, \tilde{k}, \tag{6.8}$$

where $\mathbf{w}_i(\tilde{k})$ denotes the bottom entry of $\mathbf{w}_i$. Hence, to test the approximation quality we choose some threshold parameter $\delta > 0$ and select only those Ritz pairs for which $\sqrt{\beta_{\tilde{k}}}/\alpha_{\tilde{k}} |\mathbf{w}_i(\tilde{k})| < \delta$. The choice of the parameter δ as well as the decision if (6.5) is satisfied are crucial for the success of the resulting preconditioner. It has been often observed in numerical examples that if δ was chosen to sloppily usually convergence of the CG-method was lost after already one or two updates of the preconditioner.
Note that in the non-preconditioned case we select Ritz pairs by the same idea replacing the pseudo-residuals by the residuals $\mathbf{r}^j$ occurring in Algorithm 3.5.

6.3 A preconditioned frozen IRGNM

We are now in a position to formulate a numerical realization of Algorithm 4.10. The algorithm links the interplay of a frozen Newton method as outer iteration with the CG-method as inner iteration. Using Lanczos' method and Algorithm 6.5 or 6.6 we have the possibility to build and update the preconditioner $\mathbf{M}_n^{\text{iexc}}$ whenever it seems to be necessary.
Before we describe the implementation details involving our experience, let us formulate a model of such an algorithm illustrating the main idea.

Algorithm 6.8 (Preconditioned frozen IRGNM)

$\mathbf{x}_0^\delta := \mathbf{x}_0$; ($\mathbf{x}_0$ is initial guess)

$m := 0$;

while ($\|\mathbf{F}(\mathbf{x}_m^\delta) - \mathbf{y}^\delta\| \geq \tau\delta$)

if $f_{\text{up}}(\mathbf{F}, m, \mathbf{g}_m^\delta) = 1$

$n := m$, $i := 0$; (just for notation)

Solve $\mathbf{G}_m^T \mathbf{G}_m \mathbf{h}_m = \mathbf{G}_m^T \mathbf{g}_m^\delta$ by CG-method;

Compute via Lanczos' method Ritz pairs of $\mathbf{A}_m^T \mathbf{A}_m$;

$\mathbf{x}_{m+1}^{\delta} := \mathbf{x}_m^{\delta} + \mathbf{h}_m$;

else

$i := i + 1$;

Construct preconditioner $\mathbf{M}_{n+i}^{\text{iexc}}$ by available Ritz pairs of $\mathbf{A}_m^T\mathbf{A}_m$;

Solve $(\mathbf{M}_{n+i}^{\text{iexc}})^{-1}\mathbf{G}_{n,i}^T\mathbf{G}_{n,i}\mathbf{h}_{n+i} = (\mathbf{M}_{n+i}^{\text{iexc}})^{-1}\mathbf{G}_{n,i}^T\mathbf{g}_{n+i}^{\delta}$ by CG-method;

Compute via Lanczos' method Ritz pairs of $(\mathbf{M}_{n+i}^{\text{iexc}})^{-1}\mathbf{G}_{n,i}^T\mathbf{G}_{n,i}$;

Determine out of the Ritz pairs of $(\mathbf{M}_{n+i}^{\text{iexc}})^{-1}\mathbf{G}_{n,i}^T\mathbf{G}_{n,i}$ Ritz pairs of $\mathbf{A}_m^T\mathbf{A}_m$ (see Lemma 6.2);

$\mathbf{x}_{m+1}^{\delta} := \mathbf{x}_m^{\delta} + \mathbf{h}_{n+i}$;

$m := m + 1$;

Remark 6.9 *We have formulated this algorithm corresponding to the IRGNM. Naturally, the same idea applies for the Levenberg-Marquardt algorithm or other kind of Newton methods.*

Algorithm 6.8 serves as a model. A straightforward implementation would not yield a significant reduction of the total complexity when compared with the algorithm presented in [40]. Before we will discuss implementation details below tuning Algorithm 6.8 let us take a closer look at the choice of the update criterion f_{up}.

The aim is to find a criterion f_{up} such that the total complexity of Algorithm 6.8 is minimized and the final iterate is comparable to the standard IRGNM. Hence, the criterion f_{up} needs to satisfy the following two contradicting conditions:

a) On the one hand it is advantageous to keep the matrix $\mathbf{A}_m^T\mathbf{A}_m$ for several Newton steps fixed. During these Newton steps we can collect a lot of spectral data of $\mathbf{A}_m^T\mathbf{A}_m$ making our preconditioner efficient yielding a significant reduction of the total complexity.

b) On the other hand it is advantageous to change the matrix often. This usually yields to better convergence rates in Newton's method. Hence, Algorithm 6.8 stops earlier which also reduces the total complexity.

In other words, the function f_{up} balances the convergence speed of the outer Newton iteration with the complexity we need for the inner iterations. Therefore, an optimal choice of f_{up} depends at least on the operator $\mathbf{F}$, the right hand side $\mathbf{g}_m^{\delta}$, the regularization parameter γ_n and the complexity required to solve the actual linear system. This consideration indicates that an optimal choice of f_{up} is unrealistic in practice. Unfortunately, we were not able to determine a better choice than

$$\begin{cases} f_{\text{up}} = 1, & \sqrt{m+1} \in \mathbb{N}, \\ f_{\text{up}} = 0, & \text{else}, \end{cases} \tag{6.9}$$

which was originally suggested in [40]. Out of this reason we have formulated Theorem 4.21 for this choice. Obviously, for other choices of f_{up} we could easily obtain similar results.

Implementation details:

For an efficient realization of Algorithm 6.8 several parameters need to be chosen in a proper way with respect to the following key point:

> **For an efficient preconditioner $\mathbf{M}_n^{\text{iexc}}$ Ritz pairs of high approximation quality are essential**.

Recall that Lanczos' method can only determine Ritz pairs of high approximation quality if

a) the eigenvalue distribution is suitable and

b) a sufficiently "large" number of CG-steps are performed.

That is, whenever candidates of Ritz pairs are determined which possibly be added to the preconditioner these conditions need to be satisfied. To ensure a sufficiently "large" number of CG-steps there are roughly three possibilities at hand:

i) Choose $\varepsilon > 0$ sufficiently small in Algorithms 3.5 and 3.6.

ii) Iterate the CG-iteration as long as $\sqrt{\beta_k}/\alpha_k > \tilde{\delta}$, where $\tilde{\delta}$ is some suitable threshold parameter. This criterion corresponds to (3.39) and ensures usually some good approximations.

iii) Without any change of the stopping criterion the CG-method performs on its own a sufficiently "large" number of steps.

In our implementation we used all three possibilities. When no Ritz pairs for the preconditioner need to be computed, we choose $\varepsilon = 1/3$. In the case where $f_{\text{up}} = 0$ we use $\varepsilon = 1e{-}9$ ensuring a sufficiently large number of CG-steps. In the case where an update of the preconditioner is planned we iterate as long as $\sqrt{\beta_k}/\alpha_k > 0.1$. Both criterions usually yield an acceptable extra complexity for a refinement of the Ritz pairs which turns out to be profitable in the following Newton steps.
Naturally, when the eigenvalue distribution is suitable even in the case where $\varepsilon = 1/3$ Ritz pairs with high approximation quality can be determined, although an update of the preconditioner is not planned. If this is the case without any extra complexity an update of the preconditioner can be performed.

To distinguish these different cases is necessary to make the algorithm efficient, since the additional effort spent on the computation on the refinement of Ritz pairs needs to be saved in the following Newton steps.
When additional CG-steps for a refinement are made we do not update in these additional CG-steps the approximate solution $\mathbf{h}_m^{\text{app}}$ of the linear system. This avoids a

loss of convergence in the outer Newton iteration, since for small ε or small $\sqrt{\beta_k}/\alpha_k$, that is a large number of CG-steps, the approximate solution $\mathbf{h}_m^{\text{app}}$ possibly deteriorates rapidly due to the ill-posedness and the loss of orthogonality of the residual vectors.

Even more important than the refinement of the Ritz pairs which can be realized by performing additional CG-steps is the eigenvalue distribution. Only if it is suitable we can expect that Lanczos' method will deliver good approximations. Hence, we need to incorporate into Algorithm 6.8 a criterion when possibly the next update of the preconditioner should be realized and the extra complexity for a refinement of the Ritz pairs seems profitable. Let us list two indicators:

i) A large number of CG-steps.

ii) If there are one or several Ritz values far away from the cluster after the CG-method with $\varepsilon = 1/3$ has been stopped by (4.11).

In both cases the eigenvalue distribution seems to be suitable and a refinement may be profitable. Note that in the non-preconditioned case both criterions are obviously always satisfied.
Numerical experience has shown that usually several steps after a new preconditioner has been constructed or an update of the preconditioner has been performed the *eigenvalue distribution is not suitable*. Hence, if the number of CG-steps is not too large we recommend to wait about four of five Newton steps until considering about an update of the preconditioner. This often ensures a suitable eigenvalue distribution.

Let us formulate a general warning. The parameters selecting Ritz pairs of high approximation quality should not be chosen too sloppily in practice. It has turned out that a too sloppy choice yields inefficiency of the preconditioner after already one or two updates, which is supported by the theory presented in Sections 5.3 and 5.4. As a consequence the total complexity possibly explodes and Algorithm 6.8 is inferior when compared with a standard IRGNM or, which is even worse, sometimes we can even observe loss of convergence.

Besides ensuring the approximation quality of the Ritz pairs we need to decide when an update of the preconditioner is necessary. We have realized this by an indicator function. Since the regularization parameter γ_n tends to zero and therefore the condition number of the corresponding original matrix differs significantly during Newton's method, we accept in the starting phase, middle phase and final phase a different number of inner CG-iterations of the preconditioned system. That is, given five integers $n_1, n_2, n_3, a_1, a_2 \in \mathbb{N}$ we define a function

$$K(n) := \begin{cases} n_1, & 0 \le n < a_1, \\ n_2, & a_1 \le n < a_2, \\ n_3, & n \ge a_2. \end{cases} \tag{6.10}$$

The numbers a_1 and a_2 determine the phases of Newton's method, n_1, n_2 and n_3 the number of accepted CG-steps. The number of accepted CG-steps should be so large, that an update is not done too often. Usually after two or three updates the approximation quality of the Ritz pairs is so weak that the preconditioner starts losing its efficiency. This observation corresponds to the results of Chapter 5. In this situation a further update of the preconditioner is useless and an update of the operator restarting the update process of the preconditioner is necessary. Naturally, we also do not perform an update of the preconditioner a few steps before we restart the process, since the additional complexity spent on this update cannot be saved any more.

Finally, let us discuss the choice of the parameter $\zeta > 0$. To be more general we have formulated the whole theory for some arbitrary ζ. Remark 4.14 already indicated that the choice of ζ is not arbitrary and that the convergence speed of the preconditioned CG-method depends on this choice. As expected if ζ is not in the cluster of eigenvalues at one the matrix $(\mathbf{M}_{n+i}^{\text{iexc}})^{-1}\mathbf{G}_{n,i}^T\mathbf{G}_{n,i}$ has a second cluster point of eigenvalues around ζ. Again, by Remark 4.14 this implies one additional CG-step to solve the corresponding linear system. Numerical experience has shown that usually not only one but several additional CG-steps are performed. On this account it has turned out that $\zeta = 1$ in general leads to the most significant reduction of complexity.

6.4 A preconditioned IRGNM

Basically with respect to one issue Algorithm 6.8 is not satisfactory. The proof of convergence and convergence rates for a frozen IRGNM is still open. In particular for the complexity analysis a result comparable with Corollary 2.5 is desirable. If such a result holds we have shown in Theorem 4.21 that for the special choice of f_{up} given by (6.9) the preconditioned frozen IRGNM theoretically leads to a significant reduction of the total complexity when compared with a standard IRGNM. Replacing f_{up} in Theorem 4.21 by any other reasonable update function not significantly influencing the convergence speed of the outer Newton iteration easily similar results can be obtained. Moreover, such a convergence analysis would also close the gap in the proof of Theorem 4.21.

Hence, we would like to replace the preconditioned frozen IRGNM by a preconditioned standard IRGNM. That is, picking up the idea of the preconditioned frozen IRGNM a reformulation of Algorithm 6.8 as a standard IRGNM where in each step of the Newton iteration an efficient spectral preconditioner is available describes our goal. Formulating such an Algorithm is no problem.

Algorithm 6.10 (Preconditioned IRGNM)

$\mathbf{x}_0^\delta := \mathbf{x}_0$; ($\mathbf{x}_0$ is initial guess)

$k_n := 0$; (number of Ritz pairs)

$p \in \mathbb{N}$; (each p steps the operator is updated)

$n = 0$;

while $\left(\left\|\mathbf{F}(\mathbf{x}_n^\delta) - \mathbf{y}^\delta\right\| \geq \tau\delta\right)$

if $k_n = 0$ and $n \bmod p = 0$

Solve $\mathbf{G}_n^T\mathbf{G}_n\mathbf{h}_n = \mathbf{G}_n^T\mathbf{g}_n^\delta$ by Algorithm 3.5;

Compute via Lanczos' method Ritz pairs of of $\mathbf{A}_n^T\mathbf{A}_n$;

$k_n :=$ number of Ritz pairs with good approximation quality;

else

Construct preconditioner $\mathbf{M}_n^{\text{iexc}}$ by available Ritz pairs;

Solve $(\mathbf{M}_n^{\text{iexc}})^{-1}\mathbf{G}_n^T\mathbf{G}_n\mathbf{h}_n = (\mathbf{M}_n^{\text{iexc}})^{-1}\mathbf{G}_n^T\mathbf{g}_n^\delta$ by Algorithm 3.6;

Compute via Lanczos' method Ritz pairs of $(\mathbf{M}_n^{\text{iexc}})^{-1}\mathbf{G}_n^T\mathbf{G}_n$;

Determine out of the Ritz pairs of $(\mathbf{M}_n^{\text{iexc}})^{-1}\mathbf{G}_n^T\mathbf{G}_n$ and the old k_n Ritz pairs approximations to eigenpairs of $\mathbf{A}_n^T\mathbf{A}_n$;

$k_n :=$ number of determined approximations to eigenpairs of $\mathbf{A}_n^T\mathbf{A}_n$;

$\mathbf{x}_{n+1}^\delta := \mathbf{x}_n^\delta + \mathbf{h}_n^{\text{app}}$;

$n := n + 1$;

Whereas the formulation of a preconditioned standard IRGNM is straightforward, an efficient realization seems to be a hard task and we were not able to implement a satisfactory version of Algorithm 6.10. Moreover, the implementation we will suggest below is usually far inferior when compared with Algorithm 6.8 with respect to the complexity. The major problem arises at the command

> "Determine out of the Ritz pairs of $(\mathbf{M}_n^{\text{iexc}})^{-1}\mathbf{G}_n^T\mathbf{G}_n$ and the old k_n Ritz pairs approximations to eigenpairs of $\mathbf{A}_n^T\mathbf{A}_n$."

Let us describe the main difficulties when replacing the "fixed" operators $\mathbf{G}_{n,i}$ used in the frozen IRGNM by the "varying" operators $\mathbf{G}_n$ used in the algorithm above (see Section 5.1 for the definition). To this end recall Notation 6.4 and assume furthermore that as above $\mathbf{M}_n^{\text{iexc}}$ is given by (6.4) and $\mathbf{M}_n^{\text{exc}}$ is given by

$$\mathbf{M}_n^{\text{exc}} := \gamma_n \mathbf{I} + \sum_{j \in J_1} \lambda_j \mathbf{u}_j \mathbf{u}_j^T,$$

where the pairs $(\lambda_j, \mathbf{u}_j)$ are exact eigenpairs of $\mathbf{A}_m^T\mathbf{A}_m$, where the index m is fixed. This corresponds to the situation where we have determined in the m-th Newton step Ritz pairs of $\mathbf{A}_m^T\mathbf{A}_m$ and use these Ritz pairs for preconditioning the operators $\mathbf{G}_n^T\mathbf{G}_n$, $n > m$, in the following Newton steps.

Note that Lemma 6.2 does not hold true in this situation causing three problems. Firstly, the effect of the preconditioner $\mathbf{M}_n^{\text{exc}}$ on the eigenvalues of $\mathbf{G}_n^T\mathbf{G}_n$ is not clear. Secondly, without having knowledge of the correspondence between the eigenpairs of $\mathbf{G}_n^T\mathbf{G}_n$ and $(\mathbf{M}_n^{\text{exc}})^{-1}\mathbf{G}_n^T\mathbf{G}_n$ a realization updating the preconditioner during Newton's method is not realizable under the condition that we do not want to spend too much complexity on it. Lanczos' method computes approximations to eigenpairs of $(\mathbf{M}_n^{\text{exc}})^{-1}\mathbf{G}_n^T\mathbf{G}_n$, that is we solve the generalized eigenvalue problem

$$\mathbf{G}_n^T\mathbf{G}_n\mathbf{v} = \lambda\mathbf{M}_n^{\text{exc}}\mathbf{v}. \tag{6.11}$$

But for the construction of the spectral preconditioner eigenpairs of $\mathbf{G}_n^T\mathbf{G}_n$ are required. Thirdly, interpreting the eigenpairs used for constructing $\mathbf{M}_n^{\text{exc}}$ as inexact spectral data of $\mathbf{G}_n^T\mathbf{G}_n$ by the results of Chapter 5 Ritz pairs corresponding to small and clustered eigenvalues should be only used rather carefully for constructing $\mathbf{M}_n^{\text{exc}}$. In practice where $\mathbf{M}_n^{\text{exc}}$ is replaced by $\mathbf{M}_n^{\text{iexc}}$ the effects described above are usually enforced.

However, having these difficulties in mind let us formulate a heuristic argument motivating a preconditioned standard IRGNM. From perturbation theory of symmetric matrices it is well known that eigenpairs corresponding to well separated and simple eigenvalues are stable against small perturbations. Dealing with ill-posed problems all the occurring matrices $\mathbf{G}_n^T\mathbf{G}_n$ possess the eigenvalue distribution (4.21). Hence, if only Ritz pairs corresponding to some of the largest well separated and simple eigenvalues are used for constructing $\mathbf{M}_n^{\text{iexc}}$, then we may assume that the approximation (6.7) also holds true for $\mathbf{G}_n^T\mathbf{G}_n$, that is

$$\mathbf{G}_n^T\mathbf{G}_n \approx \gamma_n\,\mathbf{I} + \sum_{j\in J_1}\theta_j^{(1)}\mathbf{v}_j^{(1)}(\mathbf{v}_j^{(1)})^T + \sum_{j\in J_2}\gamma_n(\theta_j^{(2)}-1)\mathbf{v}_j^{(2)}(\mathbf{v}_j^{(2)})^T. \tag{6.12}$$

In this situation the scalars occurring in the second sum of the right hand side of (6.12) do not need to satisfy

$$\gamma_n(\theta_j^{(2)}-1) > 1, \qquad j \in J_2,$$

any longer, which was the case in (6.7). For this reason we have to select those vectors $\mathbf{v} \in \{\mathbf{v}_{j_1}^{(1)},\dots,\mathbf{v}_{j_k}^{(1)},\mathbf{v}_{\tilde{j}_1}^{(2)},\dots,\mathbf{v}_{\tilde{j}_{\tilde{k}}}^{(2)}\}$, which correspond to "fixed" eigenpairs of the matrices $\mathbf{G}_n^T\mathbf{G}_n$. This for instance can be realized by constructing the matrix

$$\begin{pmatrix} \arccos\left(|(\mathbf{v}_1^{(1)})^T\mathbf{v}_1^{(2)}|\right) & \cdots & \arccos\left(|(\mathbf{v}_1^{(1)})^T\mathbf{v}_{\tilde{j}_{\tilde{k}}}^{(2)}|\right) \\ \vdots & & \vdots \\ \vdots & & \vdots \\ \arccos\left(|(\mathbf{v}_{j_k}^{(1)})^T\mathbf{v}_1^{(2)}|\right) & \cdots & \arccos\left(|(\mathbf{v}_{j_k}^{(1)})^T\mathbf{v}_{\tilde{j}_{\tilde{k}}}^{(2)}|\right) \end{pmatrix} \tag{6.13}$$

measuring the angles between the "old" and the "new" computed approximations to the eigenvectors. An approximation corresponding to a "fixed" eigenpair should satisfy that it is approximately orthogonal on all the other approximations. Using this constraint we can formulate an Algorithm corresponding to Algorithm 6.6, where in a first step the "non fixed" eigenpairs are sorted out and subsequently the "fixed" eigenpairs are used to determine approximations to spectral data of $\mathbf{A}_n^T\mathbf{A}_n$, $n > m$.

Algorithm 6.11 (Sorting out Ritz pairs)

Input: $(\theta_{j_1}^1, \mathbf{v}_{j_1}^1), \ldots, (\theta_{j_k}^1, \mathbf{v}_{j_k}^1), (\theta_{\tilde{j}_1}^2, \mathbf{v}_{\tilde{j}_1}^2), \ldots, (\theta_{\tilde{j}_{\tilde{k}}}^2, \mathbf{v}_{\tilde{j}_{\tilde{k}}}^2)$; (Ritz pairs)

$\mathbf{U} := \emptyset$, $\mathbf{\Theta} := \emptyset$, $\ell := 1$;

for $k = j_1, j_2, \ldots, j_k$ (sorting out "old" Ritz pairs)

if $\mathbf{v}_k^{(1)} \perp_{\approx} \mathbf{v}_i^{(2)}$ for all $i = \tilde{j}_1, \tilde{j}_2, \ldots, \tilde{j}_{\tilde{k}}$

$\mathbf{v}_\ell := \mathbf{v}_k^{(1)}$, $\theta_\ell := \theta_k^{(1)}$

$\mathbf{U} := \mathbf{U} \cup \mathbf{v}_\ell$, $\mathbf{\Theta} := \Theta \cup \{\theta_\ell\}$;

$\ell := \ell + 1$;

$\ell_{\text{old}} := \ell - 1$;

for $i = \tilde{j}_1, \tilde{j}_2, \ldots, \tilde{j}_{\tilde{k}}$ (sorting out "new" Ritz pairs)

if $\mathbf{v}_k \perp_{\approx} \mathbf{v}_i^{(2)}$ for all $k = 1, 2, \ldots, \ell_{old}$

$\mathbf{v}_\ell := \mathbf{v}_k^{(2)}$, $\theta_\ell := \theta_k^{(2)}$

$\mathbf{U} := \mathbf{U} \cup \mathbf{v}_\ell$, $\mathbf{\Theta} := \Theta \cup \{\theta_\ell\}$;

$\ell := \ell + 1$;

$\ell := \ell - 1$;

Compute orthonormal basis $\mathbf{Q} := (\mathbf{q}_1, \ldots, \mathbf{q}_\ell)$ of span$\{\mathbf{U}\}$;

Determine $\mathbf{B} := \mathbf{Q}^T\mathbf{C}\mathbf{Q} \in \mathbb{R}^{\ell \times \ell}$ where

$$\mathbf{C} := \gamma_n \mathbf{I} + \sum_{j=1}^{\ell_{old}} \theta_j \mathbf{v}_j \mathbf{v}_j^T + \sum_{j=\ell_{old}+1}^{\ell} \gamma_n(\theta_j - 1)\mathbf{v}_j\mathbf{v}_j^T;$$

Compute eigenvalues $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_\ell$ with corresponding eigenvectors $\tilde{\mathbf{v}}_1, \ldots, \tilde{\mathbf{v}}_\ell$ of $\mathbf{B}$;

Output: $\left(\tilde{\lambda}_1 - \gamma_n, \mathbf{Q}\tilde{\mathbf{v}}_1\right), \ldots, \left(\tilde{\lambda}_\ell - \gamma_n, \mathbf{Q}\tilde{\mathbf{v}}_\ell\right)$, approximations to eigenpairs of $\mathbf{A}_n^T\mathbf{A}_n$;

When solving a nonlinear inverse problem with Algorithm 6.10 in combination with Algorithm 6.11 we need to decide numerically when two vectors are approximately orthogonal. Choosing the corresponding threshold parameter too small, many approximations are sorted out yielding an inefficient preconditioner. If the threshold parameter is chosen too sloppily, the eigenvalue distribution of the preconditioned operator is often worse than the distribution of the original operator. Both facts yield a significant increase of the complexity.

Motivated by the preconditioned frozen IRGNM we incorporated into Algorithm 6.10 the possibility to keep the operator fixed for several Newton steps. In our experiments for different choices of p the results were similar (see Figure 7.21).

Finally, we do not have much hope that a preconditioned IRGNM in a way we suggested it here can yield comparable results when compared with Algorithm 6.8 with respect to the complexity. Usually, when an update of the preconditioner is performed, many eigenpairs are sorted out, whereas only a few a are left for setting up the preconditioner, in particular if the regularization parameter γ_n is small. This effect is demonstrated in a numerical example in Section 7.3. As a consequence of this loss of information in the following Newton steps approximately as many CG-steps are required as if no preconditioner is available. Moreover, usually the eigenvalue distribution of the occurring preconditioned operators is not suitable for Lanczos' method to compute many approximations of high quality. So, to improve the efficiency of Algorithm 6.10 in combination with Algorithm 6.11 in some way the significant loss of Ritz pairs from step to step needs to be avoided which can only be realized if the successive operators $\mathbf{G}_n^T\mathbf{G}_n$, $n = 0, 1, 2, \ldots$, do not differ too much, which seems to be a too strict assumption.

Chapter 7

Numerical examples

An important class of inverse problems in practice are inverse scattering problems. Scattering theory has attracted scientists and mathematicians over a hundred years. It is concerned with the effect of an inhomogeneity on an incident particle or wave. As inverse problem we consider here the problem of identifying a spatially varying refractive index profile. We examine both, the acoustic scattering case and the electromagnetic scattering case.

We use these two problems to illustrate the efficiency of Algorithm 6.8. To this end we restrict ourselves to the derivation of an operator equation of the form (2.1) for the inverse problem and the characterization of the Fréchet derivative. For details on these topics we refer to [11], [40] and [41].

7.1 The inhomogeneous medium scattering problem

The mathematical modeling of the scattering of time-harmonic acoustic waves by a penetrable inhomogeneous medium of compact support leads to the following problem: Given an incident plane wave $u^i(x) := \exp(-ik\langle x, d\rangle)$ with propagation direction $d \in \Omega$ where $\Omega := \{x \in \mathbb{R}^3 : |x| = 1\}$, the direct scattering problem for an inhomogeneous medium is to find the total field u such that

$$\begin{aligned} \Delta u + k^2 n(x) u &= 0 \quad \text{in} \quad \mathbb{R}^3, && \text{(7.1a)} \\ u^i + u^s &= u \quad \text{in} \quad \mathbb{R}^3, && \text{(7.1b)} \\ \lim_{r\to\infty} \left(\frac{\partial u^s}{\partial r} - iku^s \right) &= 0. && \text{(7.1c)} \end{aligned}$$

Here $r = |x|$, n is the refractive index of the medium, u^s is the scattered field and $u = u^i + u^s$ is the total field. (7.1c) is called Sommerfeld radiation condition, which guarantees that the scattered wave is outgoing. Absorbing media are modeled

by complex-valued refractive indices. We assume that $\Re(n) \geq 0$ and $\Im(n) \geq 0$ and n is constant and equal to 1 outside of the ball $B_\rho := \{x \in \mathbb{R}^3 : |x| \leq \rho\}$ for some $\rho > 0$, that is

$$n = 1 - a, \qquad \operatorname{supp}(a) \subset B_\rho. \tag{7.2}$$

It can be shown that the scattering problem (7.1a) – (7.1c) has an equivalent formulation as an integral equation of the second kind called the Lippmann-Schwinger integral equation,

$$u(x) = u^i(x) - k^2 \int_{B_\rho} \Phi(x,y) a(y) u(y) \, \mathrm{d}y, \qquad x \in \mathbb{R}^3, \tag{7.3}$$

where

$$\Phi(x,y) := \frac{1}{4\pi} \frac{e^{ik|x-y|}}{|x-y|}, \qquad x \neq y, \tag{7.4}$$

is the fundamental solution to the Helmholtz equation in $\mathbb{R}^3$ (see [11, Theorem 8.3]). Hence, in order to establish existence and uniqueness of a solution to the scattering problem (7.1a) – (7.1c) for all positive values of the wave number k, it is sufficient to establish existence and uniqueness of a solution to the Lippmann-Schwinger integral equation (7.3). The proof is based on the unique continuation principle for solutions of the equation (7.1a) and Riesz-Fredholm theory. We just cite here the main theorems. For details we again refer to [11].

Theorem 7.1 *Let G be a domain in $\mathbb{R}^3$ and suppose $u \in C^2(G)$ is a solution of*

$$\Delta u + k^2 n(x) u = 0$$

in G such that $n \in C(\overline{G})$ and u vanishes in a neighborhood of some $x_0 \in G$. Then u is identically zero in G.

Proof: See [11, Theorem 8.6].

□

Theorem 7.2 *For each $k > 0$ there exists a unique solution to (7.1a) – (7.1c) and the solution depends continuously with respect to the maximum norm on the incident field u^i.*

Proof: See [11, Theorem 8.7].

□

We now turn to the inverse inhomogeneous scattering problem. The Sommerfeld radiation condition (7.1c) implies the asymptotic behavior

$$u^s(x) = \frac{e^{ik|x|}}{|x|} u_\infty(\hat{x}) + O\left(\frac{1}{|x^2|}\right), \qquad |x| \to \infty,$$

where $u_\infty : \Omega \to \mathbb{C}$ is called the far field pattern of the scattered wave u^s. u_∞ is given by $E(au)$ where E denotes the linear operator

$$(Ev)(\hat{x}) := -\frac{k^2}{4\pi} \int_{\mathbb{R}^3} e^{-ik\langle \hat{x}, y\rangle} v(y)\, \mathrm{d}y \tag{7.5}$$

for $\hat{x} = x/|x|$ on the unit sphere Ω. We indicate the dependency of the far field pattern on the direction d of the incident plane wave u^i by writing $u_\infty(\hat{x}) = u_\infty(\hat{x}; d)$ and similarly $u^s(x) = u^s(x; d)$ and $u(x) = u(x; d)$.

Now we are in a position to formulate the inverse medium problem:

> **The inverse medium problem for acoustic waves is to determine** n **from** $u_\infty(\hat{x}; d)$ **for all** $\hat{x}, d \in \Omega$**.**

This problem is ill-posed in the sense of Hadamard [28] and is also nonlinear. The reasons for these facts will be given below. The first and only issue that needs to be addressed is uniqueness.

Theorem 7.3 *The refractive index n is uniquely determined by the far field pattern $u_\infty(\hat{x}; d)$ for all $\hat{x}, d \in \Omega$ and a fixed wave number k.*

Proof: See [11, Theorem 10.5].

□

To solve the inverse scattering problem we consider the operator

$$\begin{aligned} F : D(F) \subset H_0^s(B_\rho) &\to L^2(\Omega \times \Omega), \qquad s > \frac{3}{2}, \\ a &\mapsto u_\infty, \end{aligned}$$

which maps a perturbation $a = 1 - n$ of the refractive index n to the corresponding far field pattern u_∞. Since we need a Hilbert space setting, we choose the domain of definition $D(F)$ of F to be the set of all functions a in the Sobolev space $H^s(B_\rho)$ with $a < 1$ and $s > 3/2$. To get an explicit representation of the operator F, we once again take a look at the Lippmann-Schwinger equation (7.3). Introducing the volume potential operator

$$(V\psi)(x) := \kappa^2 \int_{B_\rho} \Phi(x, y)\psi(y)\, \mathrm{d}y,$$

a multiplication by a carries (7.3) over to the linear operator equation of the second kind

$$(I + aV)au = au^i. \tag{7.6}$$

So, given a refractive index n the far field pattern of the corresponding solution u of the inhomogeneous medium problem is determined by

$$u_\infty(\,\cdot\,;d) = E(I+aV)^{-1}au^i(\,\cdot\,;d), \qquad d \in \Omega. \tag{7.7}$$

In other words, given only one incident wave $u^i(\,\cdot\,;d)$ the operator F to which we apply Newton's iteration to is defined through the right hand side of (7.7). Using this representation, the analytic kernel of the operator E serves as indicator that the inverse medium problem is exponentially ill-posed. Moreover, it can be also seen that the operator F is nonlinear. To reformulate F as an operator for not only given one incident plane wave but also for given all incident plane waves $u^i(\,\cdot\,;d)$, $d \in \Omega$, and for a detailed analysis of this operator, we refer to [40]. We cite here the main results.

Theorem 7.4 *The operator F has the following properties:*

a) The operator F is Fréchet differentiable on its domain $D(F)$.

b) The Fréchet derivative $F'[x]$ is injective for all $x \in D(F)$.

c) The Fréchet derivative $F'[x] : H_0^s(B_\rho) \to L^2(\Omega\times\Omega)$, $s > 3/2$, is compact and the singular values of $F'[x]$ satisfy

$$\sigma_j(F'[x]) = O(\exp(-cj^{1/4})), \qquad j \to \infty,$$

for some $c > 0$.

Proof: See [40].

□

To avoid setting up the derivative matrix for $F'[a_n^\delta]$ and $F'[a_n^\delta]^*$ in each step of the IRGNM, we compute the Fréchet derivative of F. Differentiating F in direction h for fixed u^i gives

$$(I+aV)v_h' = hu \tag{7.8}$$

since $u = u^i - Vv$. This yields $u_\infty' = E(I+aV)^{-1}hu$, that is

$$F'[a]h = E(I+aV)^{-1}hu. \tag{7.9}$$

Now the advantage of solving the linearized equation (2.8) in each Newton step by an iterative method – in our context by the CG-method – is evident. Instead of setting up the derivative matrix, formula (7.9) shows that computing $F'[a]h$ involves essentially the solution of (7.8) in a first step. In a second step the far field pattern is determined by an application of the operator E defined in (7.5). Note that we have to solve a similar problem for the evaluation of $F'[a]^*\tilde{h}$. Moreover, to set up the right hand side in each Newton step, equation (7.6) has to be solved. These processes involve solving the Lippmann-Schwinger equation (7.3) in three space

dimensions, which is rather time consuming and causes the main complexity in our algorithm. This discussion justifies the argumentation in the proof of Theorem 4.20 for the total complexity of our algorithm.

For a fast numerical solution method of the Lippmann-Schwinger equation and for the evaluations of $F'[a]h$ and $F'[a]^*\tilde{h}$ to some given vectors $h, \tilde{h}$ we refer to [40] and [82].
In practice many incident waves u^i from different incident directions d are necessary to get a good reconstruction of the refractive index. A heuristic argument for this fact is the following: The far field pattern depends on the incident direction d of the plane wave u^i and the observation point $\hat{x}$, that is a function of two variables, whereas the unknown refractive index n naturally depends on three variables.

7.2 Electromagnetic waves in an inhomogeneous medium

As a second example we consider the electromagnetic scattering problem of time-harmonic electromagnetic waves in an inhomogeneous, non-magnetic, isotropic medium without free charges in $\mathbb{R}^3$. By $\varepsilon = \varepsilon(x) > 0$ we denote the electric permittivity, by $\sigma = \sigma(x)$ the electric conductivity and by μ_0 the constant magnetic permeability. We assume that there exists a $\rho > 0$ such that $\varepsilon(x) = \varepsilon_0$ and $\sigma(x) = 0$ for all x outside the ball B_ρ, that is the inhomogeneity is supported inside B_ρ. The time dependence of the electric field $\mathcal{E}$ can be described by

$$\mathcal{E}(x,t) = \Re\left(E(x)e^{-i\omega t}\right),$$

where ω is the angular frequency and the vector field $E : \mathbb{R}^3 \to \mathbb{C}^3$ satisfies the differential equation

$$\operatorname{curl}\operatorname{curl} E - \kappa^2(1 - a(x))E = 0 \tag{7.10}$$

in $\mathbb{R}^3$ where the wave number κ is defined by $\kappa^2 = \varepsilon_0\mu_0\omega^2$ and the refractive index $n = n(x)$ is given by

$$n(x) = 1 - a(x) = \frac{1}{\varepsilon_0}\left(\varepsilon(x) + i\frac{\sigma(x)}{\omega}\right), \qquad x \in \mathbb{R}^3,$$

where again $\operatorname{supp}(a) \subset B_\rho$. We can now formulate the corresponding direct scattering problem: Given an incident field

$$E^i(x) := \exp(-i\kappa\langle x, d\rangle)p \tag{7.11}$$

with direction $d \in \Omega$ and polarization $p \in \mathbb{C}^3$ such that $p \cdot d = 0$ and the refractive index $n \in C^{1,\alpha}(\mathbb{R}^3)$, $0 < \alpha < 1$, with $\operatorname{supp}(a) \subset B_\rho$, the scattering problem for time harmonic electromagnetic waves for an inhomogeneous medium is to find the

scattered field $E^s \in C^2(\mathbb{R}^3, \mathbb{C}^3)$ such that the total field $E := E^i + E^s$ solves the Maxwell equations (7.10) and the scattered field satisfies the Silver-Müller radiation condition

$$\lim_{|x|\to\infty} (\text{curl } E^s(x) \times x - i\kappa|x|E^s(x)) = 0 \tag{7.12}$$

uniformly for all directions $\hat{x} = x/|x| \in \Omega$.

As in the acoustic case it can be shown that this scattering problem has an equivalent formulation as an integral equation. If $E \in C^2(\mathbb{R}^3, \mathbb{C}^3)$ is a solution to the scattering problem (7.10) and (7.12) where the incident field is given by (7.11), then it satisfies the (electromagnetic) Lippmann-Schwinger equation (see [11, Chapter 9])

$$\begin{aligned} E(x) &= E^i(x) - \kappa^2 \int_{\mathbb{R}^3} \Phi(x,y) a(y) E(y)\, \mathrm{d}y \\ &\quad + \text{grad} \int_{\mathbb{R}^3} \Phi(x,y) \left\langle \frac{\text{grad } a(y)}{1-a(y)}, E(y) \right\rangle \mathrm{d}y, \qquad x \in \mathbb{R}^3, \end{aligned} \tag{7.13}$$

where Φ is given by (7.4). Vice versa, if the total field E satisfies (7.13), then E solves the scattering problem. Moreover, the following theorem can be proven:

Theorem 7.5 *The scattering problem (7.10) and (7.12) where the incident field is given by (7.11) has a unique solution and the solution E depends continuously on the incident field with respect to the maximum norm.*

Proof: See [11, Theorem 9.5].

□

We now turn to the inverse electromagnetic inhomogeneous medium problem. As in the acoustic case we assume that the data is given by the far field pattern E_∞ of E_s in the representation

$$E^s(x;d,p) = \frac{e^{i\kappa|x|}}{|x|} E_\infty(x;d,p) + O\left(\frac{1}{|x|^2}\right), \qquad |x| \to \infty. \tag{7.14}$$

With the additional arguments d and p we indicate the dependency of E, E^s and E_∞ on the incident field $E^i = E^i(x;d,p)$.

The inverse medium problem for electromagnetic waves is to determine n from $E_\infty(\hat{x};d,p)$ for all $\hat{x}, d \in \Omega$ and $p \in \mathbb{C}^3$.

By similar reasons as in the acoustic case this problem is nonlinear and ill-posed. Corresponding to Theorem 7.3 it can be shown that the refractive index n is uniquely determined by the far field pattern $E^\infty(\hat{x};d,p)$ for all $\hat{x}, d \in \Omega$, $p \in \mathbb{C}^3$.

From the Lippmann-Schwinger equation (7.13) we obtain by a short computation the formula

$$\begin{aligned} E_\infty(\hat{x};d,p) &= -\kappa^2 \int_{B_\rho} \frac{e^{-ik\langle \hat{x},y\rangle}}{4\pi} a(y)E(y;d,p)\,\mathrm{d}y \\ &\quad -i\kappa\hat{x}\int_{B_\rho} \frac{e^{-i\kappa\langle \hat{x},y\rangle}}{4\pi}\left\langle \frac{\operatorname{grad} a(y)}{1-a(y)}, E(y;d,p)\right\rangle \mathrm{d}y. \end{aligned}$$

Since the far-field pattern is a tangential field, it can be rewritten as $E^\infty = Z(aE)$ where $Z : L^2(B_\rho)^3 \to L^2(\Omega)^3$ denotes the far-field operator defined by

$$(Zu)(\hat{x}) := -\kappa^2 \hat{x} \times \int_{B_\rho} \frac{e^{-i\kappa\langle \hat{x},y\rangle}}{4\pi} u(y)\,\mathrm{d}y \times \hat{x} \tag{7.15}$$

(see [41]). To solve the inverse scattering problem we consider the operator

$$\begin{aligned} F : D(F) \subset H_0^s(B_\rho) &\to L^2(\Omega\times\Omega)^3 \\ a &\mapsto E_\infty, \end{aligned}$$

where $D(F) := \{a \in H^s(B_\rho) : a < 1\}$. To ensure that $a \in C^{1,\alpha}(B_\rho)$ we let $s > 5/2$. As in the acoustic case our goal is to reformulate the operator F with the help of (7.13). Note that a multiplication of (7.13) with the function a and using aE as new unknown as we did in the acoustic case does not work in the electromagnetic case due to the third term on the right hand side. This term is also responsible for worse mapping properties of the integral operator.

Following [41] we define the 4ρ-periodic functions $f : \mathbb{R}^3 \to \mathbb{C}^3$ and $k : \mathbb{R}^3 \to \mathbb{C}$,

$$f(x) := \chi(x)E^i(x), \qquad k(x) := \begin{cases} \kappa^2\Phi(x,0), & |x| < 2\rho, \\ 0, & |x| \geq 2\rho \end{cases}$$

for $x \in G_{2\rho} := \{x \in \mathbb{R}^3 : |x_j| < 2\rho,\ j = 1,2,3\}$ where $\chi : \mathbb{R}^3 \to \mathbb{R}$ denotes a smooth cut-off function satisfying $\chi(x) = 1$ for $x \in B_\rho$ and $\operatorname{supp}(\chi) \subset G_{2\rho}$. For the construction of such a function χ we refer to [57, Theorem 2.15]. Introducing the function

$$b(x) := \frac{\operatorname{grad} a(x)}{\kappa^2(1-a(x))}, \qquad x \in \mathbb{R}^3,$$

which is well defined since $\Re(n(x)) > 0$ for all $x \in \mathbb{R}^3$, it can be shown (see [41]) that instead of solving (7.13) it is sufficient to solve the periodic Lippmann-Schwinger equation

$$U(x) + \int_{G_{2\rho}} k(x-y)a(y)U(y)\,\mathrm{d}y + \operatorname{grad}\int_{G_{2\rho}} k(x-y)\,\langle b(y),U(y)\rangle\,\mathrm{d}y = f(x) \tag{7.16}$$

for all $x \in G_{2\rho}$. The unique solution of this integral equation can be used to compute a solution of (7.13). With the convolution operator $K : L^2(G_{2\rho}) \to L^2(G_{2\rho})$,

$$(Kv)(x) := \int_{G_{2\rho}} k(x-y)v(y)\,\mathrm{d}y, \qquad x \in G_{2\rho},$$

and its component-wise application $\mathbf{K} : L^2(G_{2\rho})^3 \to L^2(G_{2\rho})^3$ equation (7.16) takes the form

$$U + \mathbf{K}(aU) + \text{grad } K(\langle b, U\rangle) = f \qquad \text{in} \quad G_{2\rho}. \tag{7.17}$$

In this notation the operator F can expressed by

$$F(a) = Za(I + \mathbf{K}(a\,\cdot) + \text{grad } K(\langle b,\,\cdot\,\rangle))^{-1}(\chi E^i) \qquad \text{in} \quad G_{2\rho}.$$

With this formula it can be shown that F is Fréchet differentiable (see [41]). Differentiating the periodic Lippmann-Schwinger equation (7.16) at a in the direction h of the periodic version $U = U_a$ of the electric field E gives

$$U'_{a,h} + \mathbf{K}(aU'_{a,h}) + \text{grad } K\left(\left\langle \frac{\text{grad } a}{\kappa^2(1-a)}, U'_{a,h}\right\rangle\right) = R_a h \qquad \text{in} \quad G_{2\rho}, \tag{7.18}$$

with the right hand side

$$R_a h = -\mathbf{K}(hU_a) - \text{grad } K\left(\frac{1}{\kappa_2}\left\langle \text{grad } h + h\frac{\text{grad } a}{1-a}, \frac{U_a}{1-a}\right\rangle\right).$$

Thus, the computation of $U'_{a,h}$ can be done by solving the the periodic Lippmann-Schwinger equation (7.16) with right hand side $R_a h$. Now, by an application of the product rule the Fréchet derivative is given by

$$F'[a]h = Z(aU'_{a,h} + hU_{a,h}).$$

Thus, to evaluate $F'[a]h$ we have to solve (7.17) in a first step and (7.18) in a second step. Solving these equations is rather time consuming and causes the main complexity of the IRGNM. Subsequently we have to apply the operator Z. Similarly a formula for the evaluation of the adjoint operator $F'[a]^* g$ to some given vector g can be derived. Again, these formulas are the fundamentals for a matrix-free Newton method and the applicability of our algorithm derived in Chapter 6.

As in the acoustic case usually many incident waves are necessary for a good reconstruction. This is due to the fact that the far-field pattern E_∞ is a function of two variables whereas the unknown refractive index is a function of three variables.

7.3 Numerical results

For numerical examples we need a tool to construct synthetic data, that is refractive indices satisfying condition (7.2). We just consider here real valued refractive indices. Complex valued refractive indices can be constructed by the same idea. Using a smooth partition of the one (see [20, Chapter 3] for details) defined through the function

$$H(t) := \frac{g(t)}{G(t)}, \qquad t \in \mathbb{R},$$

where $g \in C_0^\infty(\mathbb{R})$,

$$g(t) := \begin{cases} e^{-1/(1-t^2)}, & |t| < 1, \\ 0, & |t| \geq 1, \end{cases} \quad \text{and} \quad G(t) := \sum_{k=-\infty}^{\infty} g(t-k),$$

it can be shown that the functions

$$f_{q,\varepsilon}(x) := \Pi_{j=1}^3 H\left(\frac{x_j}{\varepsilon} - q_j\right)$$

satisfy $f_{q,\varepsilon} \in C_0(\mathbb{R}^3)$, $\operatorname{supp}(f_{q,\varepsilon}) = \{x \in \mathbb{R}^3 : |x_j - q_j\varepsilon| \leq \varepsilon,\ j = 1,2,3\}$ and we have

$$\sum_{q \in \mathbb{Z}^3} f_{q,\varepsilon}(x) \quad = \quad 1$$

for all $x \in \mathbb{R}^3$. Hence, functions of the form of $f_{q,\varepsilon}$ are suitable to construct smooth refractive indices. An example used in [40] and [41] is given by

$$\begin{aligned} a_1^\dagger(x) \quad := \quad & \frac{0.25(\sin(5(x_1-1)x_2 + x_3(x_3-2)))}{1.2 - \cos(x_1(x_1-2) + (x_2-0.5)x_2(x_2+1) + x_3)} \\ & \cdot \tilde{H}(-0.8(1.5x_1 + x_2 + x_3 - 0.5))\tilde{H}(2.5(|x| - 0.55)), \end{aligned} \tag{7.19}$$

where

$$\tilde{H}(t) := \frac{\sum_{j=0}^{\infty} g(t-j)}{G(t)}.$$

To compare our algorithms with the one presented in [40] we will use the refractive index given by $a_1^\dagger$ for the examples considered in the case of acoustic scattering. For the electromagnetic case we consider the refractive index $a_2^\dagger$ defined through

$$\begin{aligned} \tilde{a}(x) \quad := \quad & (2 + 0.2x_1^2 + 0.1x_2^2 + 0.8x_3^2 - x_1 - x_2)^{-1} \\ & \cdot \left[0.8\sin(x_1 + x_2 + 0.5 + x_1^2 + (x_3^2 - 0.1)) \right. \\ & \left. - \cos(x_1^2 - 2x_2^2 + x_3^2 + 0.3) - \cos(x_1x_2x_3)\right], \\ a_2^\dagger(x) \quad := \quad & \tilde{a}(x)\tilde{H}(0.6x_1 + 0.3x_2 + 0.9x_3) \\ & \tilde{H}(-0.8(1.5x_1 + x_2) + 0.1x_3)\tilde{H}(1.5(|x| - 0.3)). \end{aligned} \tag{7.20}$$

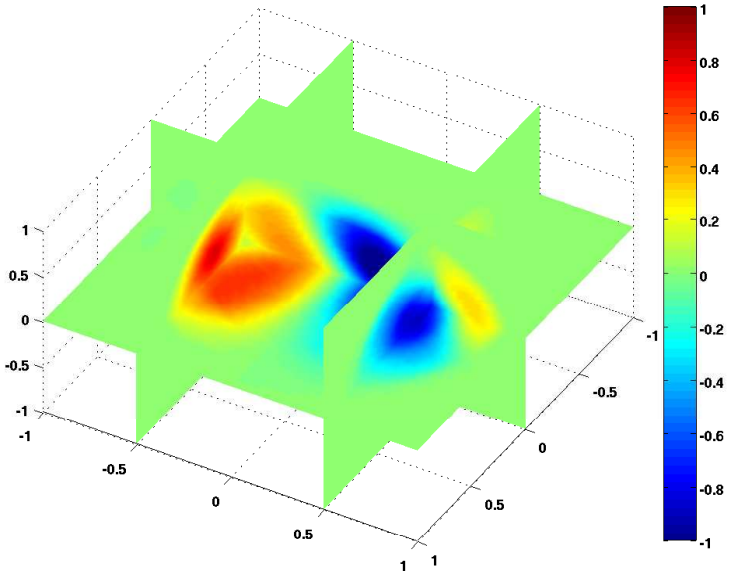

Figure 7.1: Plot of the refractive index $a_1^\dagger$

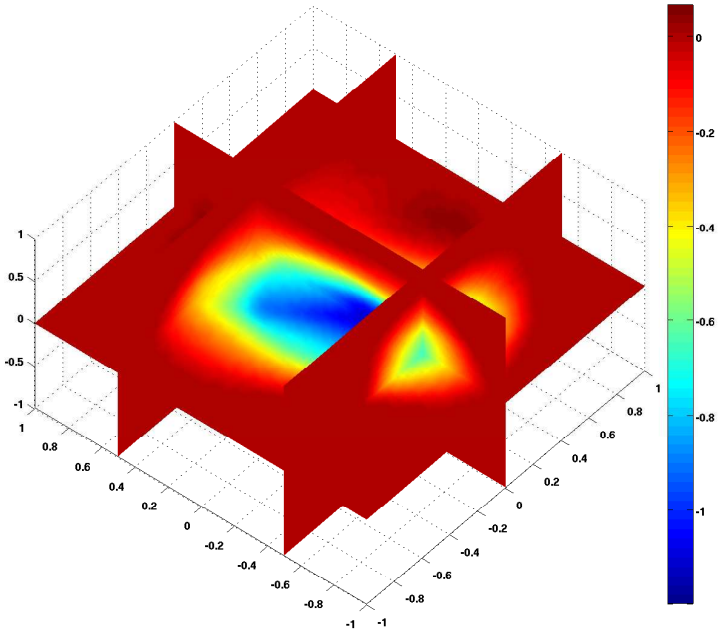

Figure 7.2: Plot of the refractive index $a_2^\dagger$

In the following we give a short description of the general framework used for the numerical examples, our intention what we want to illustrate and some notation used throughout this section.
First of all, in all the experiments with respect to the *acoustic* scattering problems the regularization parameter was chosen by

$$\gamma_n = 2^{-n}, \qquad n = 0, 1, 2, \ldots .$$

This corresponds to (1.25) with $\gamma_0 = 1$ and $\gamma = 2$. For the implementation of the matrices representing the operators occurring in the inverse acoustic scattering problems we refer to [40], for the electromagnetic case to [41]. Moreover, for the computations we used a C++-class library designed for iterative regularization methods, which already included the implementation of these operators. This library has been made available to us by Prof. Dr. Thorsten Hohage.
To test numerical algorithms for inverse problems synthetic data have to be produced. If these synthetic data are obtained by the same method that is used in the algorithm for the inverse problem, one often obtains unrealistically good results, especially if the exact solution is chosen from the approximating subspace. To this end precautions against *inverse crimes* have to be taken. This for instance can be done by using a different ansatz and a different number of grid points for producing the synthetic data. Furthermore, the exact solution should not be in the finite-dimensional approximating subspace.
To avoid *inverse crimes* such precautions have been implemented into the C++-class library at our disposal.

Recall that in our experiments we used (6.9) as update criterion and that for the inner CG-iterations in the case where $f_{\text{up}} = 1$ we chose (4.11) as stopping criterion with $\varepsilon = 1e-9$ and in the case where $f_{\text{up}} = 0$ we chose $\varepsilon = 1/3$ (see implementation details in Section 6.3). To ensure a sufficiently large number of inner CG-iterations when an update of the preconditioner seemed necessary we iterated as long as $\sqrt{\beta_k}/\alpha_k > 0.1$. Moreover, any iteration was stopped as soon as the residual vectors started losing their orthogonality. The indicator function (6.10) was given through

$$K(n) := \begin{cases} 5, & 0 \le n < 25, \\ 7, & 25 \le n < 43, \\ 9, & n \ge 43. \end{cases}$$

In our discussion of the numerical examples we want to focus on the two following points:

a) The effectiveness of Algorithm 6.8 when compared with a standard IRGNM and the algorithm presented in [40], which is basically given by Algorithm 4.10 and

b) a detailed description of the update process of the preconditioner.

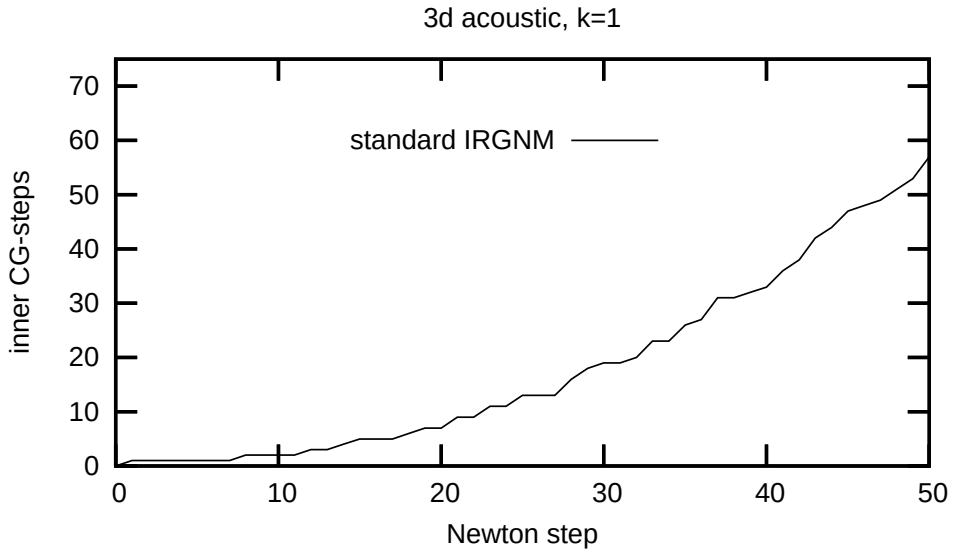

Figure 7.3: Inner CG-iterations for a standard IRGNM

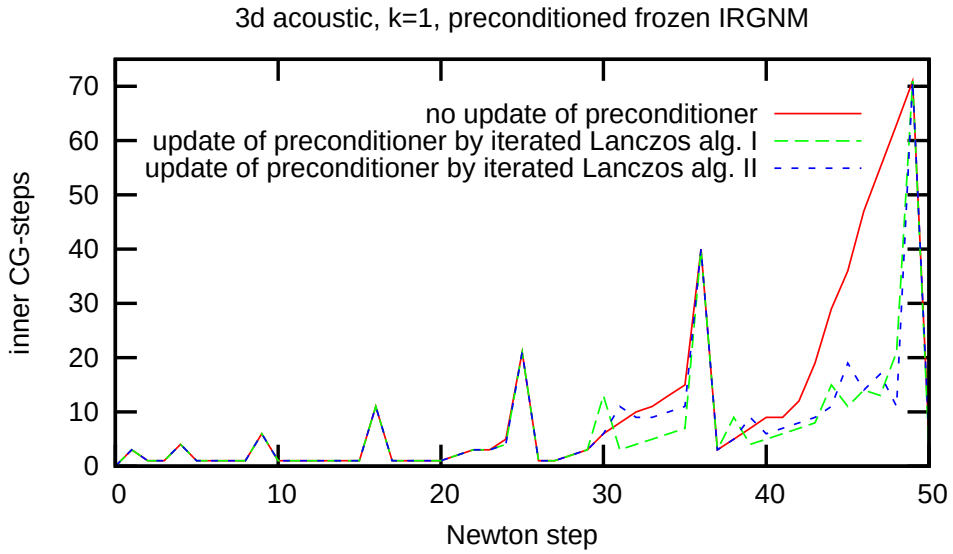

Figure 7.4: The effect of preconditioning and updating on the inner CG-iterations

As a first example we consider the inverse acoustic scattering problem presented in Section 7.1 for the wavenumber $k = 1$. The refractive index is defined through (7.19). To compare the different algorithms an early stopping by the discrepancy principle had to be avoided. To this end we only used exact data in the first experiments. To compute the reconstructions we used 50 incident waves.
Let us start by having a look at the total complexity of the different algorithms. In Figure 7.3 for a standard IRGNM the number of CG-steps are plotted over the Newton step. The progress of the line is as expected. Since the regularization parameter decreases during the IRGNM the number of inner CG-iterations to solve the linear systems increases. This behavior corresponds to the results formulated in Theorem 4.19. After 50 Newton steps a total number of 922 CG-steps have been performed.
Figure 7.4 shows the effect of preconditioning and updating on the inner CG-iterations. The red line represents the algorithm presented in [40], the green dashed line Algorithm 6.8 coupled with Algorithm 6.5 and the blue dashed line Algorithm 6.8 coupled with Algorithm 6.6. The peaks in the curves occur whenever an update of the operator has been performed, that is $\sqrt{n+1} \in \mathbb{N}$, since in these Newton steps we chose $\varepsilon = 10^{-9}$ to approximate eigenpairs for constructing a new preconditioner. Note that the peaks lie above the curve plotted in Figure 7.3. If we would have chosen $\varepsilon = 1/3$ the peaks would only differ a bit or would even coincide with the number of CG-steps plotted in Figure 7.3.
Let us compare the different curves. As we can observe following the red line between the Newton steps $25 - 36$ and $36 - 49$ the effectiveness of the original preconditioner reduces. This is expected because of the reasons we discussed in point g) in Section 6.1. Therefore, the number of inner CG-steps increases rapidly. The green dashed line starts to differ significantly from the red line at Newton step 30, that is when the original preconditioner starts losing its efficiency. Now the effect of updating the preconditioner improving its efficiency starts. The peak at Newton step 30 is explained by further inner CG-iterations in order to improve the approximations. Subsequently the green dashed line proceeds significantly below the red line. Hence, these additional CG-steps in the 30-th Newton step are profitable, since the saved number of CG-steps in the following Newton steps is definitely larger. A similar observation is true for the blue dashed line. In order to obtain better approximations additional CG-steps have been performed. This explains the peak. In an analogous way the preconditioner has been updated between the Newton steps $36 - 49$. As a consequence the green dashed line proceeds significantly under the red line. The total number of inner CG-steps for the different algorithms is given by

- 922 for the standard IRGNM,
- 554 for the preconditioned frozen IRGNM without updating the preconditioner,

- 348 for the preconditioned frozen IRGNM where the preconditioner is updated by the iterated Lanczos algorithm I,
- 377 for the preconditioned frozen IRGNM where the preconditioner is updated by the iterated Lanczos algorithm II.

As we can see by the total number of CG-steps, in this example after 50 Newton steps updating the preconditioner yields about a reduction of 35% of the CG-steps when compared with the preconditioned frozen IRGNM without updating the preconditioner, which was originally suggested in [40]. Furthermore, when compared with a standard Newton method the total complexity could have been reduced to about 1/3 of the original complexity.

To illustrate the dependency of the number of inner CG-steps on the update criterion we also considered the functions

$$f_{\mathrm{up},1}(n) = \begin{cases} 1, & n = 0, n = 33 \quad \text{and} \quad n = 43, \\ 0, & \text{else} \end{cases}$$

and

$$f_{\mathrm{up},2}(n) = \begin{cases} 1, & n = 0 \\ 0, & \text{else.} \end{cases}$$

Note that in the preconditioned frozen IRGNM with $f_{\mathrm{up},2}$ no operator update is performed, the function $f_{\mathrm{up},1}$ was chosen arbitrarily without any mathematical motivation. In Figure 7.5 we have plotted the inner CG-steps for the preconditioned frozen IRGNM coupled with the iterated Lanczos algorithm I and the update criterions $f_{\mathrm{up},1}$ and $f_{\mathrm{up},2}$. The total number of inner CG-steps for these choices are given by

- 326 for the update criterion $f_{\mathrm{up},1}$,
- 293 for the update criterion $f_{\mathrm{up},2}$.

Hence, for these choices additional complexity could have been saved. This shows that when we keep the operator fixed for a long period of Newton steps additional complexity can be saved. On the other hand we will illustrate in Figure 7.7 that the reconstructions are not as satisfactory when compared with the update criterion (6.9), that is additionally complexity needs to be spent to reach the approximation quality of this update criterion. These examples indicate that the update criterion f_{up} balances the convergence speed of the outer Newton iteration with the number of inner CG-steps.

To illustrate the real valued refractive index given by (7.19) and its reconstructions we plot them on slices through the sphere. This plotting technique is illustrated

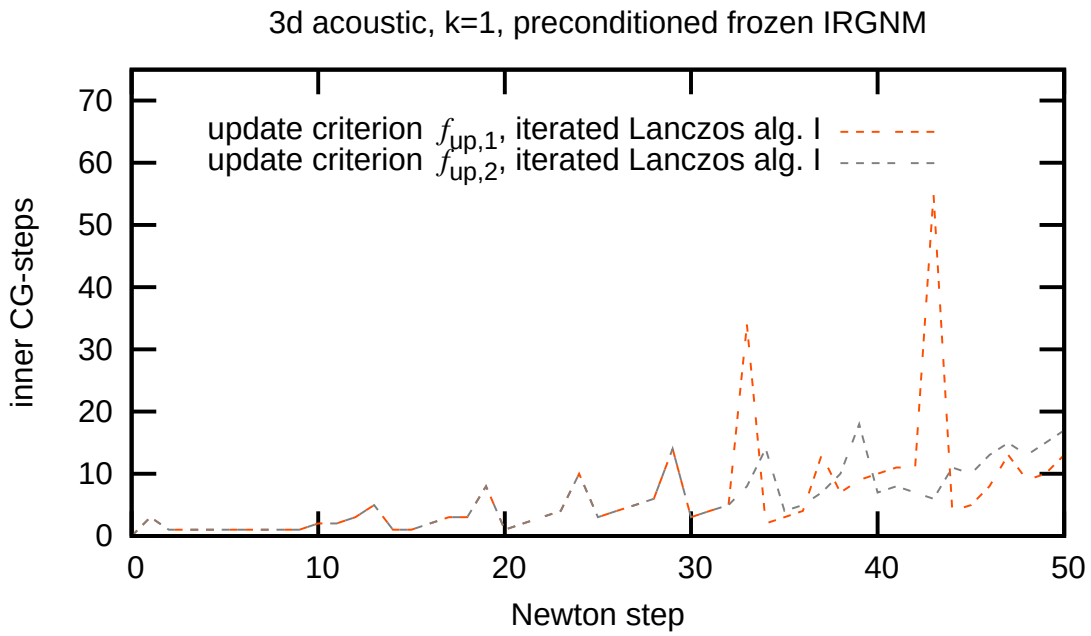

Figure 7.5: The effect of the choice of the update criterion f_{up} in Algorithm 6.8

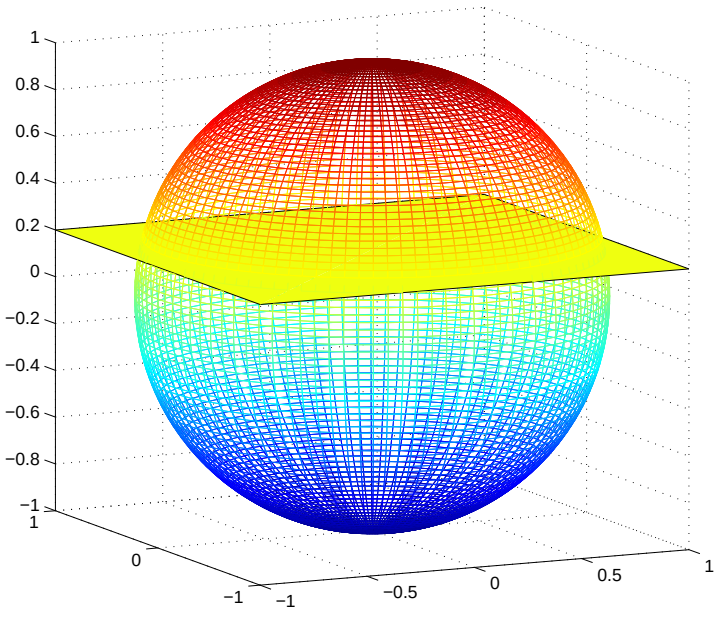

Figure 7.6: Sliced sphere

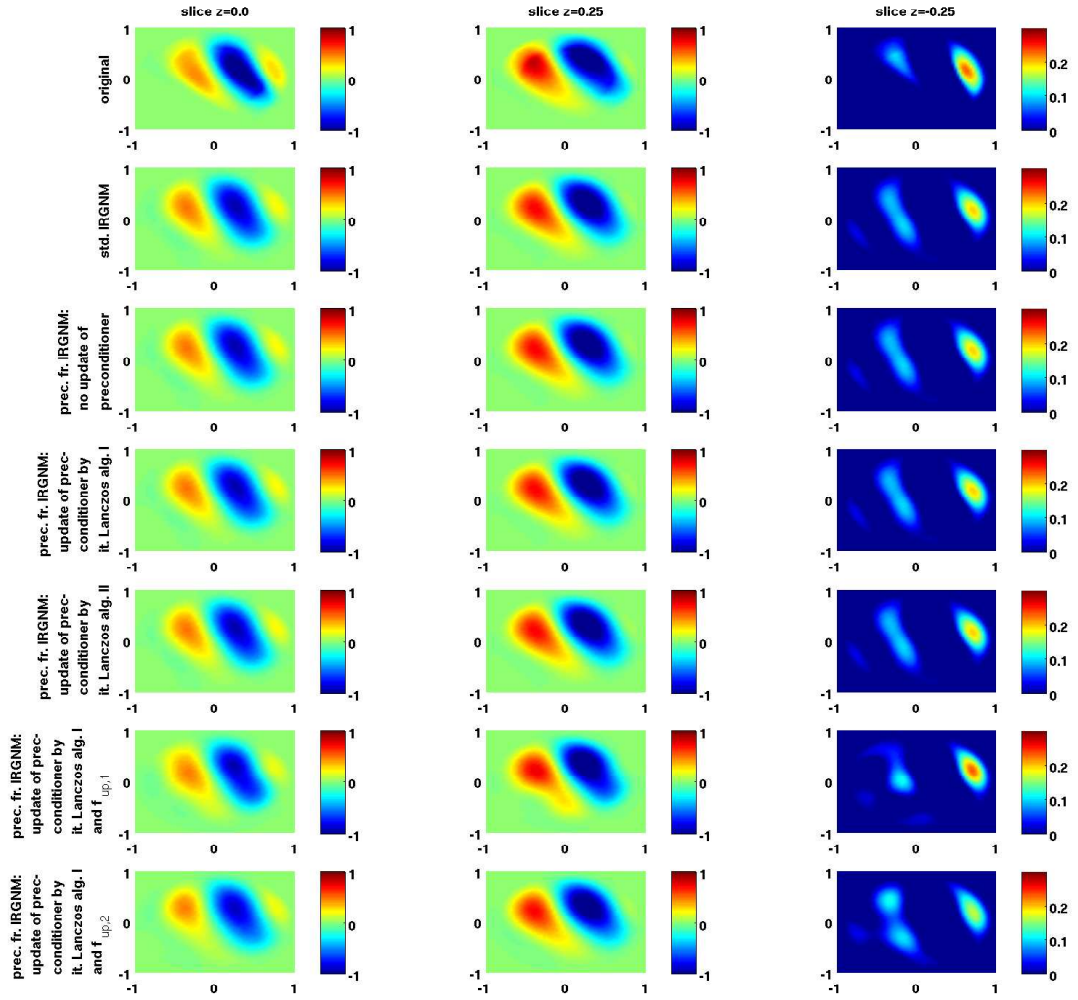

Figure 7.7: Reconstructions of the refractive index after 50 Newton steps, $k = 1$

in Figure 7.6. Figure 7.7 above shows the original refractive index and its reconstructions determined by the different methods. In the first column we plotted the refractive index on the slice $z = 0.0$, in the second column on $z = 0.25$ and in the third column on $z = -0.25$. It can be seen that the differences between the final reconstructions determined by the standard IRGNM and the preconditioned frozen IRGNM with the update criterion (6.9) are negligible.
For the update criterions $f_{\text{up},1}$ and $f_{\text{up},2}$ the reconstructions are slightly worse. This shows that for these update criterions the convergence speed of the outer Newton iteration is slowed down. Hence, additional Newton steps are necessary to get

comparable results with the other update criterions. Note that the reconstruction for the update criterion $f_{\text{up},1}$ on the slice $z = -0.25$ is surprisingly good when compared with the other algorithms.

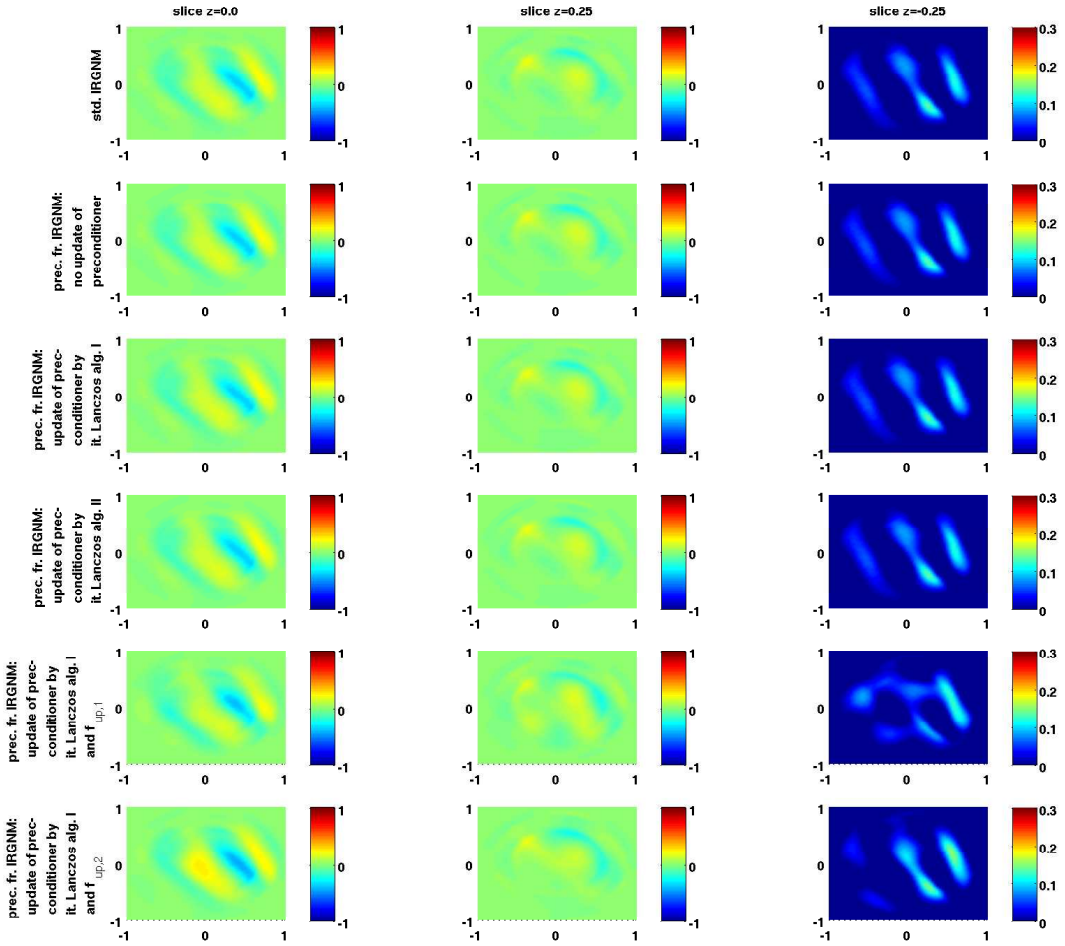

Figure 7.8: Error plot, $k = 1$

In Figure 7.8 we have plotted the error for the reconstructions determined by the different methods once again indicating the comparability of the final iterates of the different methods.

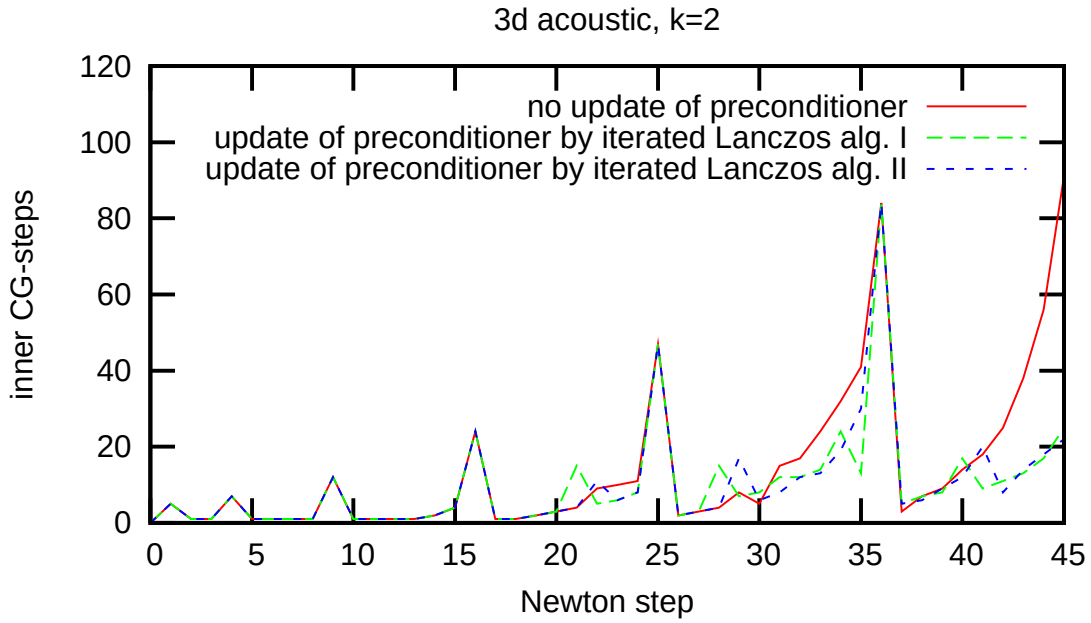

Figure 7.9: The effect of preconditioning and updating on the inner CG-iterations, $k = 2$

For the wavenumber $k = 2$ we obtained similar results, which are illustrated in the Figures 7.9, 7.10 and 7.11. The red line in Figure 7.9 again represents the algorithm presented in [40], the green dashed line Algorithm 6.8 coupled with Algorithm 6.5 and the blue dashed line Algorithm 6.8 coupled with Algorithm 6.6. The corresponding reconstructions and errors are shown in the Figures 7.10 and 7.11. Note that in this case we already stopped the iteration after 45 steps, since due to round-off errors and the ill-posedness the iterates started to deteriorate rapidly after that number of steps. The total number of inner CG-steps for the different algorithms is given by

- 648 for the preconditioned frozen IRGNM without updating the preconditioner,
- 458 for the preconditioned frozen IRGNM where the preconditioner is updated by the iterated Lanczos algorithm I,
- 459 for the preconditioned frozen IRGNM where the preconditioner is updated by the iterated Lanczos algorithm II.

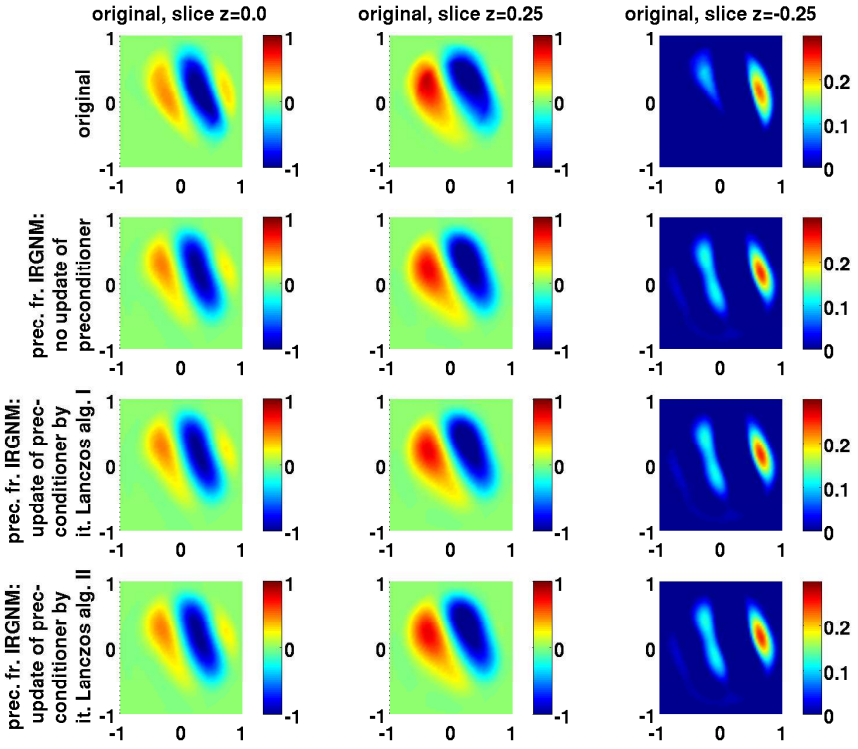

Figure 7.10: Reconstructions of the refractive index after 45 Newton steps, $k = 2$

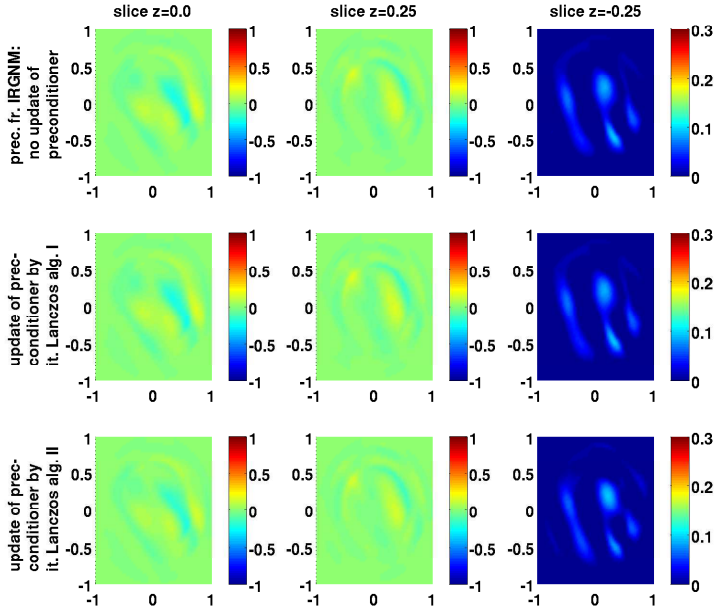

Figure 7.11: Error plot, $k = 2$

In further experiments, we tested the convergence of Algorithm 6.8 in combination with Algorithms 6.5 and 6.6 for different noise levels $\delta > 0$. The results are plotted in the Figures 7.12, 7.13 and 7.14. Let us first have a closer look at the number of inner CG-steps. It can be seen that just for the very small noise level $\delta = 0.0001$ an update of the preconditioner is necessary at Newton step 31. In all the other cases no update of the preconditioner is performed, since the discrepancy principle stops the outer Newton iteration before the number of inner CG-steps exceeds the threshold level given through the indicator function (6.10).

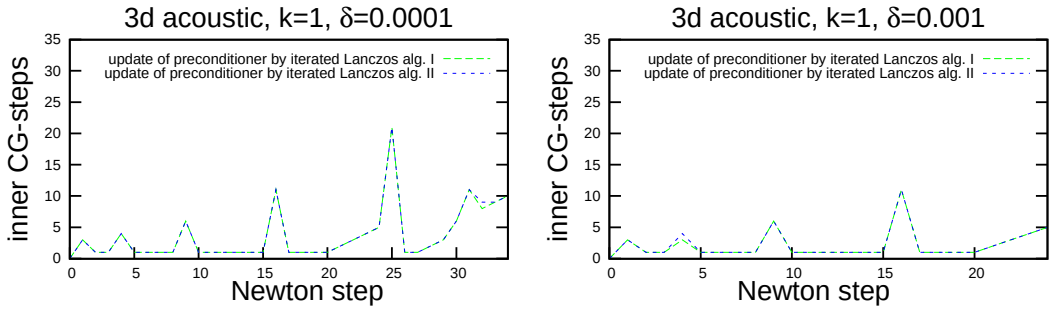

Figure 7.12: Number of inner CG-iterations for noisy data

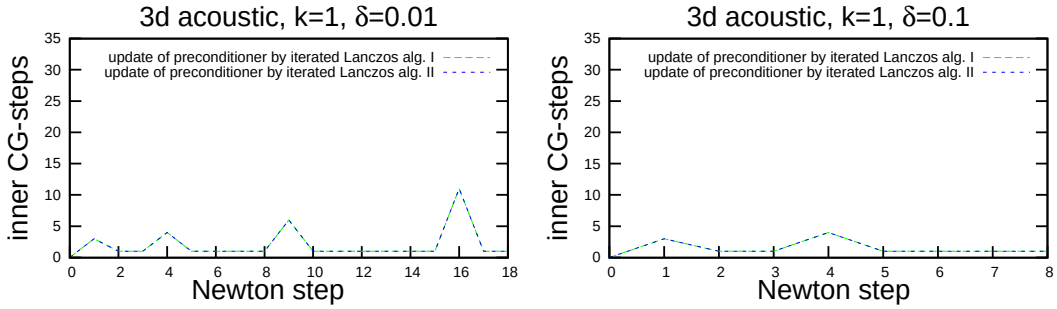

Figure 7.13: Number of inner CG-iterations for noisy data

Figure 7.14 shows that for the small noise level $\delta = 0.0001$ the reconstruction is a good approximation to the true solution. As expected with an increase of the noise level the main features of the scatterer are smoothed out and details get lost, for example for the rather high noise level $\delta = 0.1$ one cannot really identify the scatterer any more. On the other hand a high noise level together with the discrepancy principle enforces an early stopping of Algorithm 6.8. Naturally, this

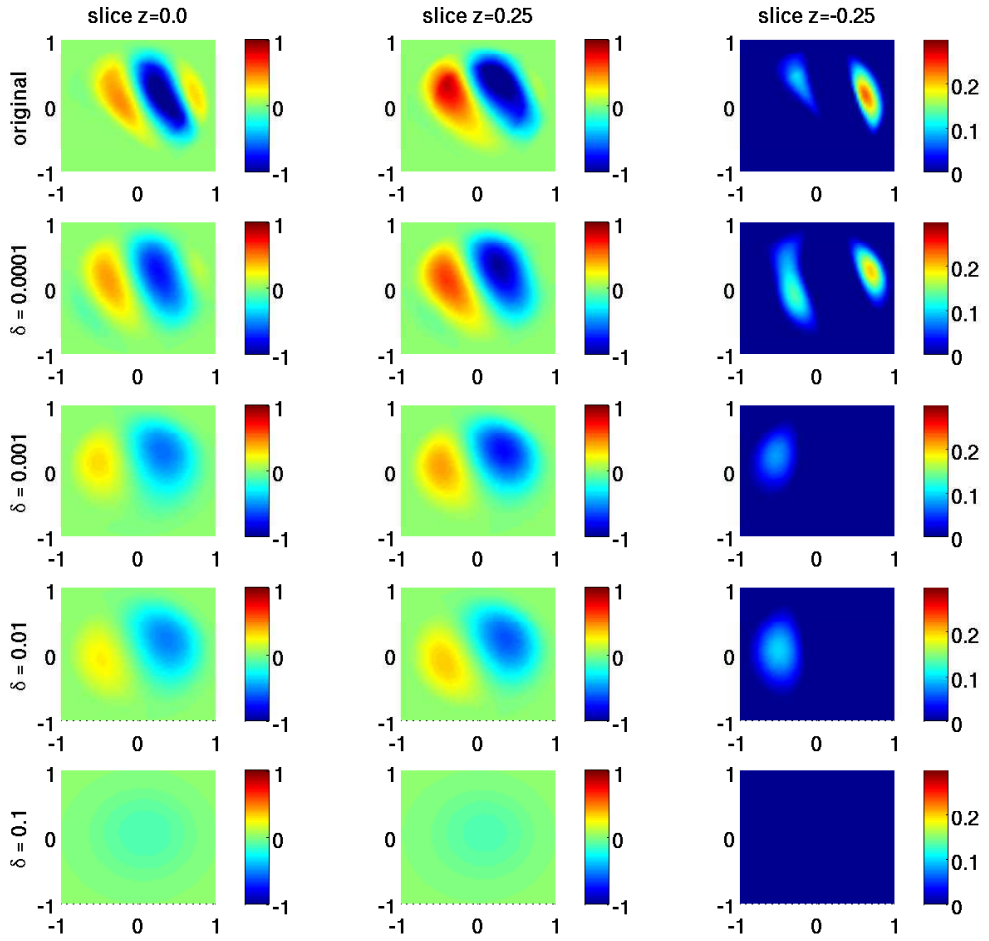

Figure 7.14: Reconstructions with different relative noise levels $\delta > 0$

reduces the total complexity. We only plotted the reconstructions determined by Algorithm 6.8 coupled with Algorithm 6.6, for Algorithm 6.5 the reconstructions are identical.

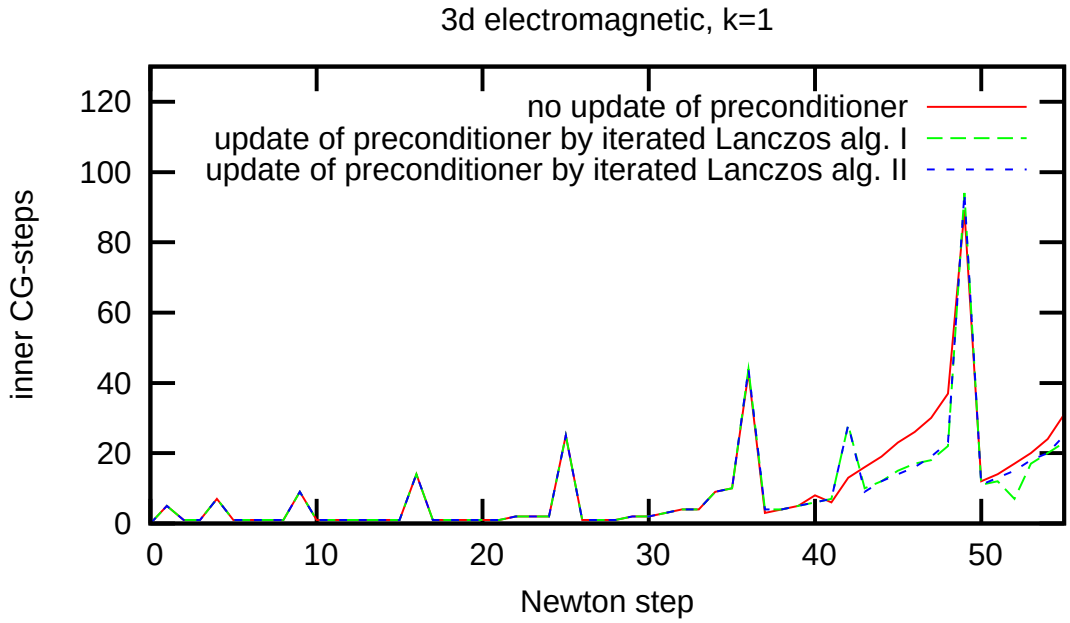

Figure 7.15: The effect of preconditioning in the electromagnetic scattering problem

In the case of the inverse *electromagnetic* scattering problem we only present experiments with exact data. Experiments with noisy data would be similar to those in the acoustic case. The regularization parameter in these experiments was chosen by

$$\gamma_n = 0.01 \cdot 2^{-n}, \qquad n = 0, 1, 2, \ldots.$$

To compute the reconstructions we used 100 incident waves. The indicator function (6.10) in these experiments was simply given by

$$K(n) := 10, \qquad n \geq 0.$$

The reconstructions and the error plot are illustrated in Figures 7.16 and 7.17. The inner CG-steps are plotted in Figure 7.15 above. It can be seen that in these experiments Algorithm 6.8 in combination with Algorithm 6.6 and 6.5 is again superior when compared with the algorithm where no update of the preconditioner is performed. The total number of inner CG-steps for the different algorithms is given by

- 560 for the preconditioned frozen IRGNM without updating the preconditioner,

- 496 for the preconditioned frozen IRGNM where the preconditioner is updated by the iterated Lanczos algorithm I,
- 507 for the preconditioned frozen IRGNM where the preconditioner is updated by the iterated Lanczos algorithm II.

That is, in this example after 55 Newton steps updating the preconditioner yields about a reduction of 10% of the inner CG-steps when compared with the preconditioned frozen IRGNM without updating the preconditioner. That the updating of the preconditioner is less efficient in this example when compared with the acoustic scattering problem can be explained by the following two reasons.

As we can observe in Figure 7.15 the original preconditioner works until the 40-th Newton step rather efficient. Hence, the complexity required for an update in the Newton steps before the 40-th step could not have been saved in the following Newton steps. To this end we chose the indicator function K in such a way that the updating process starts when we can ensure to save the additional complexity spent for the update. The other reason is that the update of the preconditioner in this example does not lead to such a significant difference in the inner CG-steps when compared with the acoustic scattering problem. This is possibly due to the fact the the linear operators arising in this electromagnetic example have a lot of degenerated eigenvalues. Such a property can influence the efficiency of the preconditioner and can reduce the effect when it is updated. This corresponds to the theory presented in Chapters 4, 5 and 6.

However, in all the examples we presented our preconditioning technique yielded a significant reduction of the total complexity when compared with a standard IRGNM and even when compared with the preconditioned frozen IRGNM presented in [40]. Naturally, this also reduced the computational time significantly. Since the reconstruction of the refraction index in the electromagnetic inverse scattering problem for the wavenumber $k = 1$ on a fine grid took us about 12 hours up to a whole day, even 10% reduction of the total complexity yielded a reduction of the computational time for more than one hour or two (or even more).
Moreover, for larger wavenumbers k usually the computational time increases rapidly for both the acoustic and the electromagnetic inverse scattering problem. This fact makes it necessary to have adequate preconditioning available reducing the total computational time such that these kind of problems are solvable by regularized Newton methods in appropriate time periods.

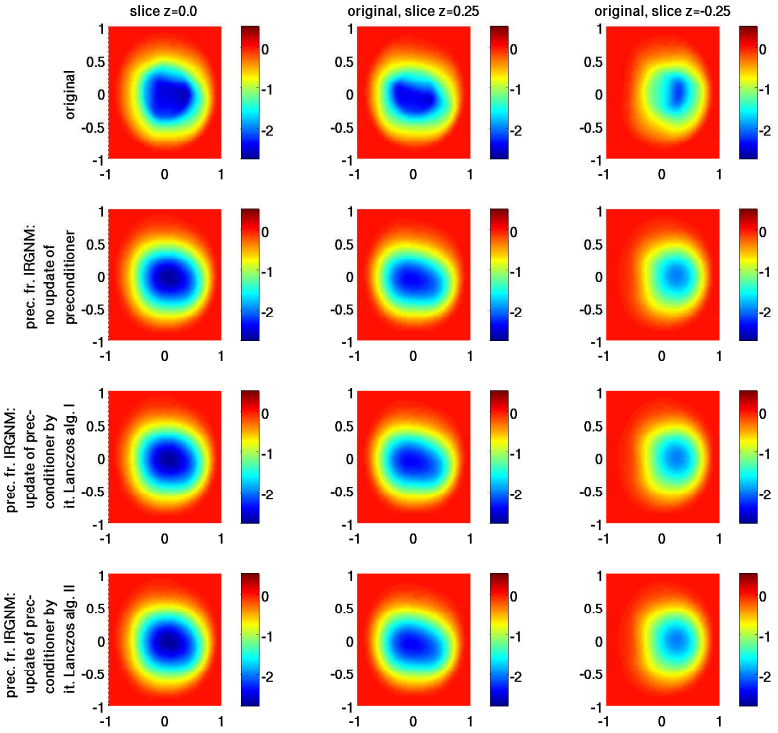

Figure 7.16: Reconstructions of the refractive index for the electromagnetic scattering problem, $k = 1$

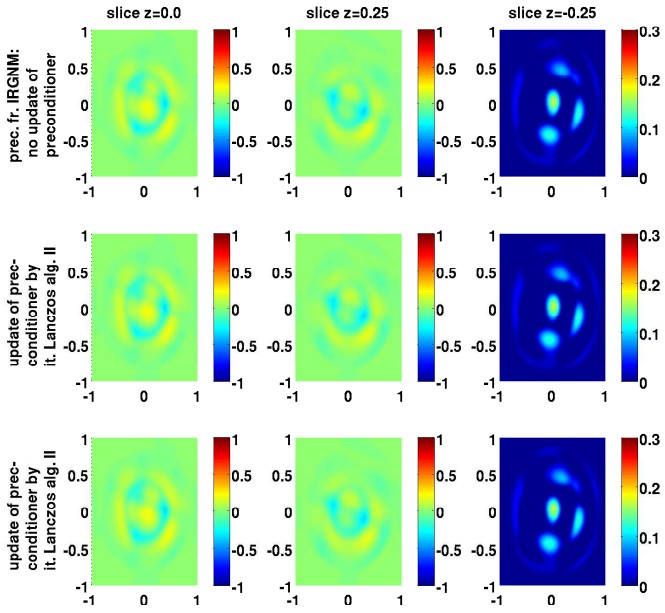

Figure 7.17: Error plot, $k = 1$

Let us now turn to the illustration of the exact behavior of Algorithm 6.8 when coupled with Algorithm 6.5. To this end we look at the approximations to the eigenvalues determined by Lanczos' method within the Newton steps 25 – 36 and 36 – 49. The values in the tables correspond to the inverse acoustic scattering problem for the wavenumber $k = 1$. In the tables we use the following notation:

- λ: Ritz value of non-preconditioned matrix
- μ: Ritz value of preconditioned matrix
- $\gamma_n(\mu - 1)$: see formula (6.2)
- appr. qual.: right hand side of (6.8)

Step	λ	appr. qual.	Step	μ	appr. qual.	$\gamma_n(\mu-1)$
25	2.27632e-02	0.0	30	3.67725	3.70816e-14	4.98677e-09
	1.28441e-03	0.0		2.57832	7.02890e-11	2.93983e-09
	4.88432e-04	0.0		1.99764	3.89232e-07	1.85825e-09
	3.01355e-04	0.0		1.96529	5.17614e-07	1.79799e-09
	2.69159e-05	2.77020e-41		1.85129	1.22699e-05	1.58565e-09
	2.61366e-05	6.78665e-41		1.73306	7.88985e-07	1.36543e-09
	1.69453e-05	3.58730e-30		1.65290	1.12727e-06	1.21612e-09
	3.99508e-06	2.71733e-30		*1.08626	5.11918e-04	1.60672e-10
	3.87455e-06	3.95705e-19		*1.06015	6.91324e-03	1.12038e-10
	5.16012e-07	5.97496e-19		*1.04872	7.92293e-04	9.07482e-11
	5.10298e-07	1.94985e-18		*1.01191	4.50290e-03	2.21841e-11
	3.34366e-07	2.80166e-16		*1.00000	3.41492e-05	—
	2.45728e-07	8.27110e-16		*0.91488	1.04662e-06	—
	2.00425e-07	3.46067e-15				
	1.96591e-07	6.64087e-13				
	2.93306e-08	7.17179e-13				
	2.52469e-08	4.22266e-10				
	5.34717e-09	6.18163e-10				
	4.23484e-09	4.71299e-10				
	*1.52702e-09	5.48059e-10				

Table 7.1: Ritz values computed in the 25-th and 30-th Newton step

Table 7.1 shows the Ritz values computed in the 25-th and 30-th step of Algorithm 6.8 together with their quantitative approximation quality given by the right hand side of (6.8). In the second column of Table 7.1 the computed approximations to the eigenvalues of $\mathbf{A}_{25}^T\mathbf{A}_{25}$ are listed, the third column shows the computed

values given by (6.8). The fifth column lists the approximations to the eigenvalues of $(\mathbf{M}_{30}^{\text{iexc}})^{-1}\mathbf{G}_{25,5}^T\mathbf{G}_{25,5}$, the sixth column again shows the approximation quality given by (6.8) and the seventh column the computed approximations to the eigenvalues of $\mathbf{A}_{25}^T\mathbf{A}_{25}$ (see Section 5.1 and Lemma 6.2 for the notation).
First note that the more separated and further away from the cluster the eigenvalues are, the better the approximations are. Moreover, since the eigenvalue distribution in the preconditioned case is more uniformly than in the non-preconditioned case and since the eigenvalues are not that well separated, in average the approximation quality of the Ritz values in the non-preconditioned case is far better than in the preconditioned case.
We mark the Ritz values, which were not used for constructing the preconditioner with a " * " in front. To determine these Ritz values we use two criterions. On the one hand the approximation quality should be acceptable. On the other hand the computed Ritz value should be well separated from the cluster of eigenvalues.
Note that both criterions imply that we need to choose two parameters. With the knowledge that the efficiency of the preconditioner depends sensitive on errors in the approximations in particular for small and clustered eigenvalues (see Corollary 5.11), we only choose eigenvalues which are at least 10% away from the cluster and the approximation quality satisfies at least (see 6.8)

$$\frac{\sqrt{\beta_k}}{\alpha_k}|\mathbf{w}_i(k)| \leq 0.0001.$$

From the computed approximations in the 25-th step only one was not used for setting up the preconditioner. It was sorted out not because of its approximation property, but since it was assumed to lie in the cluster of eigenvalues. Since the regularization parameter is given by

$$\gamma_{25} = 2^{-25} \approx 2.98023\text{e-}08,$$

we have that $\gamma_{25} + 1.52702\text{e-}09 \in [\gamma_{25}, 1.1\gamma_{25}]$, that is the computed approximation is less than 10% away from the cluster at γ_{25} of $\mathbf{G}_{25}^T\mathbf{G}_{25}$.

In Figure 7.18 we have focused on the number of inner CG-steps of Algorithm 6.8 coupled with Algorithm 6.5 within the Newton steps 24 – 36. Following the red line in Figure 7.18 it can be observed that the preconditioner constructed only by spectral data computed in the 25-th step starts losing its efficiency from the 30-th step on.
Hence, additional approximations to the spectral data are needed. The peak at step 30 indicates that an update of the preconditioner has been performed in this Newton step. As a consequence the green dashed line proceeds in the following Newton steps significantly below the red line.

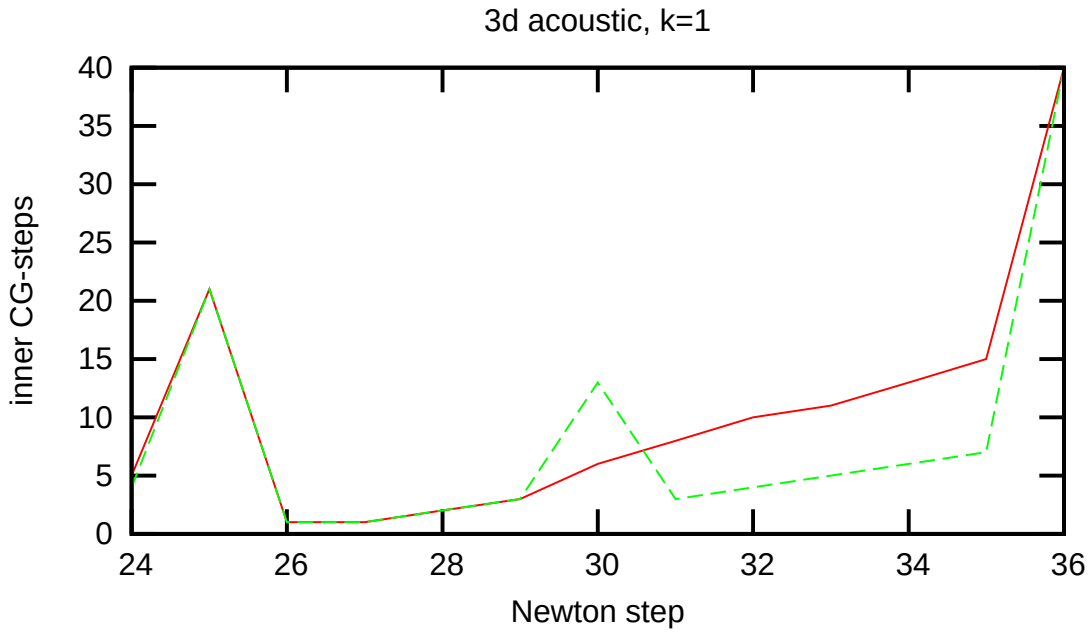

Figure 7.18: Inner CG-iterations between steps 24–36

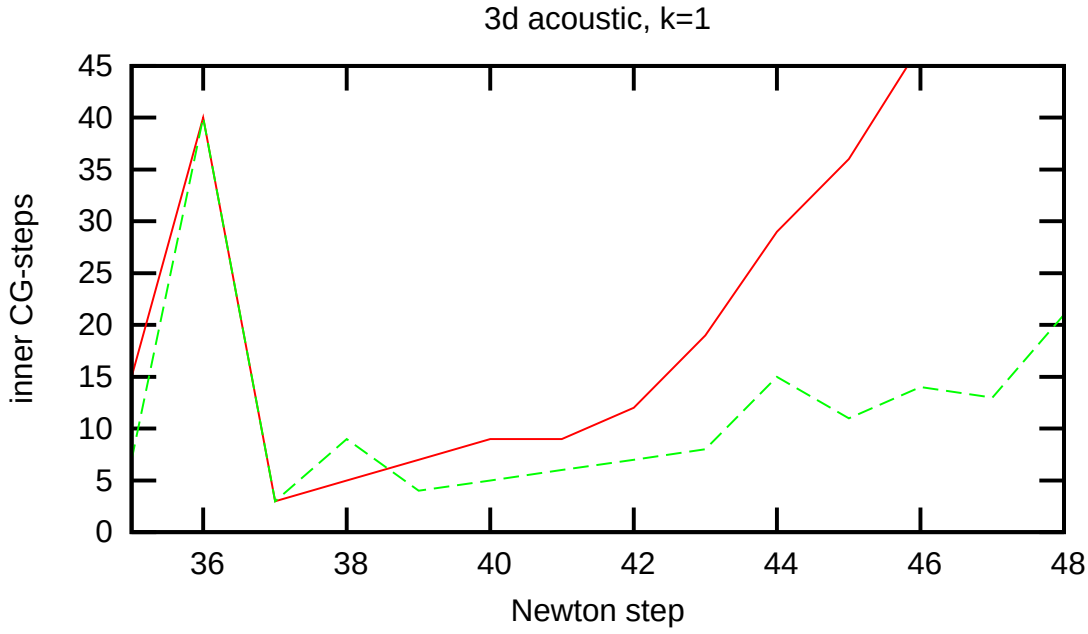

Figure 7.19: Inner CG-iterations between steps 35–48

Step	μ	appr. qual.	Step	μ	appr. qual.
32	*1.32156	0.501434	33	*1.68998	0.0394335
	*1.24305	0.863644		*1.35283	0.2066
	*1.00019	0.0274566		*1.10437	0.976432
	*0.715662	0.0439202		*1.00006	0.0381966
				*0.665045	0.0297071
34	*2.38008	0.0110668	35	*3.76015	0.000183982
	*1.96792	0.307096		*2.93652	0.00630761
	*1.46107	0.245884		*1.67745	0.126451
	*1.14774	0.918331		*1.45341	0.604292
	*1.00003	0.0270733		*1.17871	0.78636
	*0.6077	0.0323268		*1.00002	0.0133295
				*0.538816	0.0161668

Table 7.2: Ritz values computed after an update of the preconditioner

After an update of the preconditioner has been performed, one usually has to perform several Newton steps until the next update is necessary and profitable. As an indicator for this serves Table 7.2. It can be seen from this table that the approximations are of low quality. This behavior is supported by many numerical examples we computed and the theoretical knowledge of Lanczos' method. Naturally, to improve the approximations we could artificially impose further inner CG-iterations. But since the eigenvalue distribution for Lanczos' method is not adequate many additional CG-iterations would be necessary to determine approximations of high quality. Usually there would be no chance to save this additional complexity in the following Newton steps.

Hence, such a procedure would only increase the complexity of the actual Newton step without having any hope to save this additional complexity in the following Newton steps because of an improved preconditioner. Furthermore, we can also see from Table 7.2 that since the regularization parameter decreases in a natural manner more and more CG-steps are required to solve the linear systems (cf. Theorem 4.19). This automatically leads to an improvement in the approximations. Hence, the number of updates of the preconditioner needs to be balanced with the complexity required for an update.

The behavior we just explained can also be observed within the Newton steps 36 – 48. Table 7.3 corresponds to Table 7.1 and shows the approximations to the eigenvalues computed in the 36-th and 38-th Newton step. Indicated by Figure 7.19 the approximations to the eigenvalues computed in the 36-th Newton step were not enough for an efficient preconditioner. To this end additional information was

Step	λ	appr. qual.	Step	μ	appr. qual.	$\gamma_n(\mu-1)$
36	2.27837e-02	0.0	38	3.19664	2.96328e-08	1.59827e-11
	1.28463e-03	0.0		2.25952	4.62865e-06	9.16422e-12
	4.88529e-04	0.0		1.92191	3.19232e-05	6.70778e-12
	3.01202e-04	0.0		1.70738	1.67861e-04	5.14687e-12
	2.69246e-05	0.0		1.53378	1.47544e-04	3.88376e-12
	2.61326e-05	0.0		*1.10793	1.75094e-03	—
	1.69333e-05	0.0		*1.0227	5.19410e-03	—
	3.99668e-06	0.0		*1.0	3.42611e-05	—
	3.87180e-06	0.0		*0.78284	3.30233e-05	—
	5.16173e-07	0.0				
	5.10301e-07	0.0				
	3.34475e-07	0.0				
	2.45606e-07	0.0				
	2.00343e-07	0.0				
	1.96624e-07	0.0				
	2.93269e-08	0.0				
	2.52438e-08	0.0				
	5.61789e-09	2.11131e-45				
	5.01608e-09	1.67902e-44				
	4.05941e-09	1.39805e-42				
	2.89864e-09	5.94092e-39				
	1.89332e-09	1.22114e-34				
	1.80280e-09	1.79692e-33				
	1.63027e-09	8.84960e-34				
	1.36847e-09	9.37498e-32				
	1.30099e-09	1.19886e-32				
	1.21294e-09	2.10932e-32				
	1.60645e-10	4.45360e-25				
	1.12678e-10	1.97599e-22				
	3.4298e-11	3.98918e-15				
	2.9916e-11	1.38529e-13				
	2.9304e-11	2.39465e-13				
	2.5220e-11	4.93997e-14				
	1.6928e-11	2.29909e-13				
	1.2381e-11	2.18788e-12				
	9.132e-12	1.29050e-12				
	5.826e-12	2.82929e-12				
	4.217e-12	1.60296e-12				
	2.207e-12	4.22321e-12				
	8.5e-14	2.53055e-13				

Table 7.3: Ritz values computed in the 36-th and 38-th Newton step

computed in the 38-th step and added to the preconditioner. Hence, the green dashed line lies significantly below the red line.
Table 7.4 shows that the approximations in the following Newton steps are of low quality. Note, in the 42-nd step one approximation is used for updating the preconditioner, before in the 43-rd step a lot of additional approximations are determined. In Table 7.5 we introduced a column named "angle". This column illustrates the effect we described in Section 6.1 in point c). In this case the CG-iteration stopped because of loss of orthogonality in the residual vectors. This is a typical behavior, which can be observed and cannot be avoided.

Step	μ	appr. qual.	st.	μ	appr. qual.
39	*1.82089	0.187555	40	*2.08548	0.343045
	*1.42948	0.325559		*1.85745	0.0390014
	*1.08067	0.925097		*1.21598	0.654463
	*1	0.00695484		*1.12278	0.672576
	*0.354917	0.054602		*1	0.00200436
				*0.250638	0.0106955
41	*2.72602	0.00158443	42	4.44217	2.05333e-05
	*2.41614	0.0222835		*2.91016	0.0170637
	*1.45467	0.239468		*1.91554	0.129417
	*1.29053	0.41377		*1.65201	0.609985
	*1.15321	0.878018		*1.51993	0.49133
	*1	0.00111933		*1.19495	0.607775
	*0.175651	0.00580906		*1	0.000870562
				*0.124562	0.00904441

Table 7.4: Ritz values computed after an update of the preconditioner

Finally, in the following Newton steps 44 and 45 only a few Ritz values are used for updating the preconditioner making it more efficient.
Note that when we would have used all the approximations marked with a " * " our numerical results with respect to the complexity would have been much worse. This corresponds to the sensitivity results presented in Chapter 5, since the low approximation quality of these Ritz pairs destroys the convergence rates of the preconditioned CG-method. Moreover, one can also observe that sometimes even convergence of the CG-algorithm gets lost. These numerical observations justify the final remark of Chapter 5.

We have refrained from a detailed description of Algorithm 6.8 coupled with Algorithm 6.6, since the observations are similar and do not improve the comprehension for the algorithm.

Step	μ	appr. qual.	angle	$\gamma_n(\mu-1)$
43	3.6889	8.39201e-09	—	3.05692e-13
	2.8327	5.2245e-07	—	2.08354e-13
	2.44073	0.00000316	—	1.63792e-13
	2.29505	0.00001184	—	1.47230e-13
	2.09111	0.00007509	—	1.24045e-13
	*1.99421	0.0051594	—	—
	*1.86095	0.030205	—	—
	*1.72785	0.0217048	—	—
	*1.42969	0.0898282	—	—
	*1.37498	0.203838	—	—
	*1.24207	0.0532351	1.5706	—
	*1.11697	0.0316343	1.5693	—
	*1.01169	0.0244099	1.5646	—
	*1.0	0.000117223	1.5029	—
	*0.0884446	8.54277e-08	1.1968	—

Table 7.5: Eigenvalues computed in the 43-rd Newton step

Step	μ	appr. qual.	Step	μ	appr. qual.
44	4.98798	0.000156811	45	7.23743	3.68018e-05
	3.7315	0.00078983		6.39212	9.15147e-06
	*3.10275	0.0484111		*5.19581	0.00145043
	*2.95196	0.322901		*4.98565	0.00887455
	*2.67928	0.300166		*4.45983	0.0171487
	*2.30294	0.848033		*4.05674	0.0399935
	*1.83687	0.183679		*3.21413	0.754341
	*1.46772	0.194604		*2.68226	0.073008
	*1.16662	0.111337		*2.12736	0.549818
	*1.0	0.000300321		*1.95554	0.294801
	*0.06151	0.0112634		*1.44196	0.110447
				*1.23733	0.149149
				*1.0	0.000107254
				*0.0413965	0.00369196

Table 7.6: Eigenvalues computed after an update of the preconditioner

Finally, as a last example let us consider Algorithm 6.10 coupled with Algorithm 6.11. To understand the behavior of this algorithm we restrict ourselves to the *two dimensional* inverse acoustic scattering problem for the wavenumber $k = 1$. Our intention is to explain on this example where the difficulties of this algorithm arise and why we do not have much hope that this algorithm can yield comparable results to the preconditioned frozen IRGNM.
The refractive index in our experiments was defined through

$$a^{\dagger}(x) := \frac{\sin(5(x-1)y)}{(1.2-\cos(x^2+y^3))}\tilde{H}(-0.8(1.5x+y-0.5))\tilde{H}(2.5(|x|-0.55))$$

and the regularization parameter was chosen by

$$\gamma_n = 0.01 \cdot 2^{-n}, \qquad n = 0, 1, 2, \ldots.$$

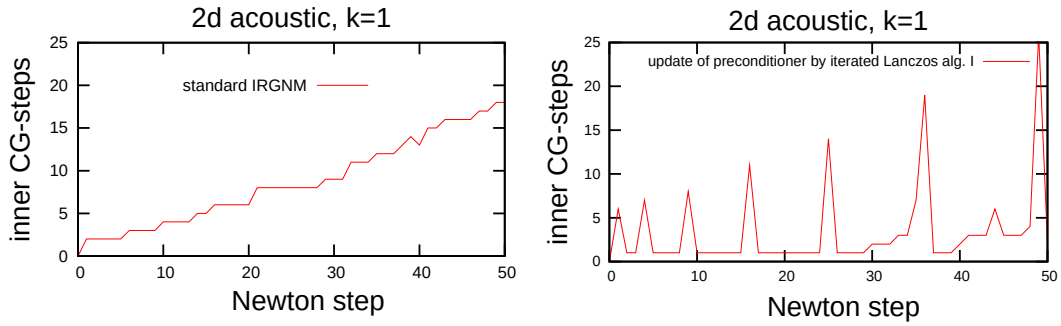

Figure 7.20: Inner CG-iterations for a standard IRGNM and Algorithm 6.8

As a reference for Algorithm 6.10 (preconditioned IRGNM) act the standard IRGNM and Algorithm 6.8 for which the number of inner CG-steps are plotted in Figure 7.20. The behavior of these values corresponds to the three dimensional case. To understand the inferiority of Algorithm 6.10 with respect to the complexity when compared with Algorithm 6.8 let us first take a closer look at Figure 7.21. Here we plotted for different update conditions, that is $p = 2, 3, 4, 5$, the behavior of the number of inner CG-steps together with the number of Ritz pairs used for constructing the preconditioner. The red line in the pictures on the right hand side shows the number of Ritz pairs used in Algorithm 6.10 and the green dashed line the number of Ritz pairs used in Algorithm 6.8. Two main reasons are responsible for the inferiority of Algorithm 6.10:

a) When we neglect the steps where in Algorithm 6.8 the linear systems are solved without a preconditioner, the number of Ritz pairs used for preconditioning is in average much larger when compared to Algorithm 6.10.

b) The approximation quality of the Ritz pairs used in Algorithm 6.10 is worse when compared with Algorithm 6.8.

Note that reason a) cannot be fixed by making more inner CG-iterations, since the loss of Ritz pairs is caused by changing the operator. For this reason it can be seen in Figure 7.21 that for the final Newton steps the number of Ritz pairs starts to oscillate yielding an increase in the number of inner CG-iterations.
To illustrate the operating mode of Algorithm 6.11 we plotted exemplarily the matrix (6.13) in the case of $n \mod 2$ at the Newton step 45. In Table 7.7 the corresponding Ritz values used in Newton step 45 and those who are left used in Newton step 46 for preconditioning are shown.

$$\begin{pmatrix} 1.570 & 1.570 & 1.567 & 1.327 & 1.510 & 1.413 & 1.178 & 1.495 & 1.383 \\ 1.570 & 1.570 & 1.569 & 1.539 & 0.849 & 1.472 & 1.516 & 1.416 & 1.560 \\ 1.570 & 1.570 & 1.570 & 1.014 & 1.545 & 1.278 & 1.240 & 1.518 & 1.520 \\ 1.570 & 1.570 & 1.570 & 1.427 & 1.498 & 0.913 & 1.248 & 1.555 & 1.570 \\ 1.570 & 1.570 & 1.569 & 1.569 & 1.570 & 1.566 & 1.560 & 1.560 & 1.558 \\ 1.570 & 1.570 & 1.570 & 1.018 & 1.442 & 1.300 & 1.254 & 1.493 & 1.521 \\ 1.570 & 1.570 & 1.570 & 1.547 & 1.540 & 1.555 & 1.554 & 1.526 & 1.551 \\ 1.570 & 1.570 & 1.569 & 1.520 & 0.868 & 1.412 & 1.474 & 1.425 & 1.550 \\ 1.570 & 1.570 & 1.570 & 1.541 & 1.542 & 1.529 & 1.545 & 1.569 & 1.521 \\ 1.570 & 1.570 & 1.570 & 1.546 & 1.564 & 1.539 & 1.497 & 1.539 & 1.138 \\ 1.570 & 1.570 & 1.570 & 1.501 & 1.520 & 1.549 & 1.557 & 1.407 & 1.479 \\ 1.570 & 1.570 & 1.569 & 1.460 & 1.474 & 0.935 & 1.212 & 1.531 & 1.545 \\ 1.570 & 1.570 & 1.570 & 1.437 & 1.515 & 1.514 & 1.403 & 1.490 & 1.288 \\ 1.568 & 1.569 & 1.569 & 1.482 & 1.556 & 1.519 & 1.452 & 1.519 & 1.398 \end{pmatrix} \tag{7.21}$$

As it can be seen in Figure 7.21 the loss of Ritz pairs in this situation is the most drastically one for the examples we considered. Therefore we chose this example.
Let us have a closer look at the matrix (7.21). As threshold parameter in Algorithm 6.11 we chose $\varepsilon = 0.17$. This corresponds to a deviance in the angle of about 10° which already seems to be rather high with respect to the sensitivity analysis of Chapter 5.
Note that the last six columns of (7.21) contain a value outside our threshold limit. To this end six of the old Ritz values are sorted out by Algorithm 6.11 and only three of the new computed Ritz values are selected for constructing a new preconditioner. By Table 7.7 it can be seen that unfortunately the largest eigenvalues are sorted out. Moreover, by the size of the new computed Ritz pairs in Newton step 46 it is obvious that in the Newton step before also some of the largest eigenvalues were thrown away. This loss of information according to the largest Ritz values is a very undesirable effect leading to a significant increase on the number of inner CG-iterations, which is illustrated in Figure 7.21. Hence, the change in the operator seems to be so significant, that only a few Ritz pairs are "fixed", whereas the major

Step	λ	Step	μ	appr. qual.	λ
45	0.092241586369000	46	7.81901e05	0.0	0.13968e-03
	0.003843801602560		2.20527e04	0.0	0.56733e-04
	0.003386553274810		2.86609e03	0.0	0.39255e-05
	0.000138375226890		2.17325	2.43705e-09	0.39393e-05
	0.000056733886196		2.08264	7.48568e-08	0.51187e-06
	0.000006531602490		1.96437	1.469e-07	0.49039e-06
	0.000003925510064		*1.8253	1.08846e-07	—
	0.000001077174137		*1.42781	2.05453e-04	—
	0.000000490396280		*1.30236	1.29391e-03	—
	0.000000084430925		*1.23326	4.29812e-03	—
	0.000000083177444		*1.09129	3.59033e-02	—
	0.000000051602120		*1.06282	5.54218e-01	—
	0.000000005669668		*1.0247	6.17442e-01	—
	0.000000002908520		*1.00182	3.00278e-01	—
	—		*0.976278	3.01239e-01	—
	—		*0.931081	4.20695e-02	—
	—		*0.832238	9.61324e-04	—
	—		*0.575748	2.63934e-07	—
	—		*0.052883	1.94661e-06	—
	—		*0.000705631	6.96143e-06	—
	—		*0.000113012	4.80095e-05	—

Table 7.7: Ritz values before and after an update of the preconditioner

part of the computed spectral information seems to be useless. Unfortunately, even some of the largest Ritz values are affected by the change of the operator.
Concluding we can say, that the heuristic argument that the largest well separated eigenvalues are "fixed" though the operator changes fails in practice. Therefore, Algorithm 6.10 cannot yield satisfactory results. Still, for this example this algorithm was superior when compared with a standard IRGNM. Maybe it is possible in future work to refine the selection of the Ritz pairs possibly yielding better results. Moreover, one could also think of an improvement by combining Algorithms 6.8 and 6.11, where for example in a starting phase one applies Algorithm 6.11 and in a final phase Algorithm 6.8. However, these ideas just try to combine the advantages of both algorithms, but they do not overcome the problems mentioned above.

Finally, Figure 7.22 shows that the final iterate of Algorithm 6.10 is comparable to the final iterates of Algorithm 6.8 and a standard IRGNM.

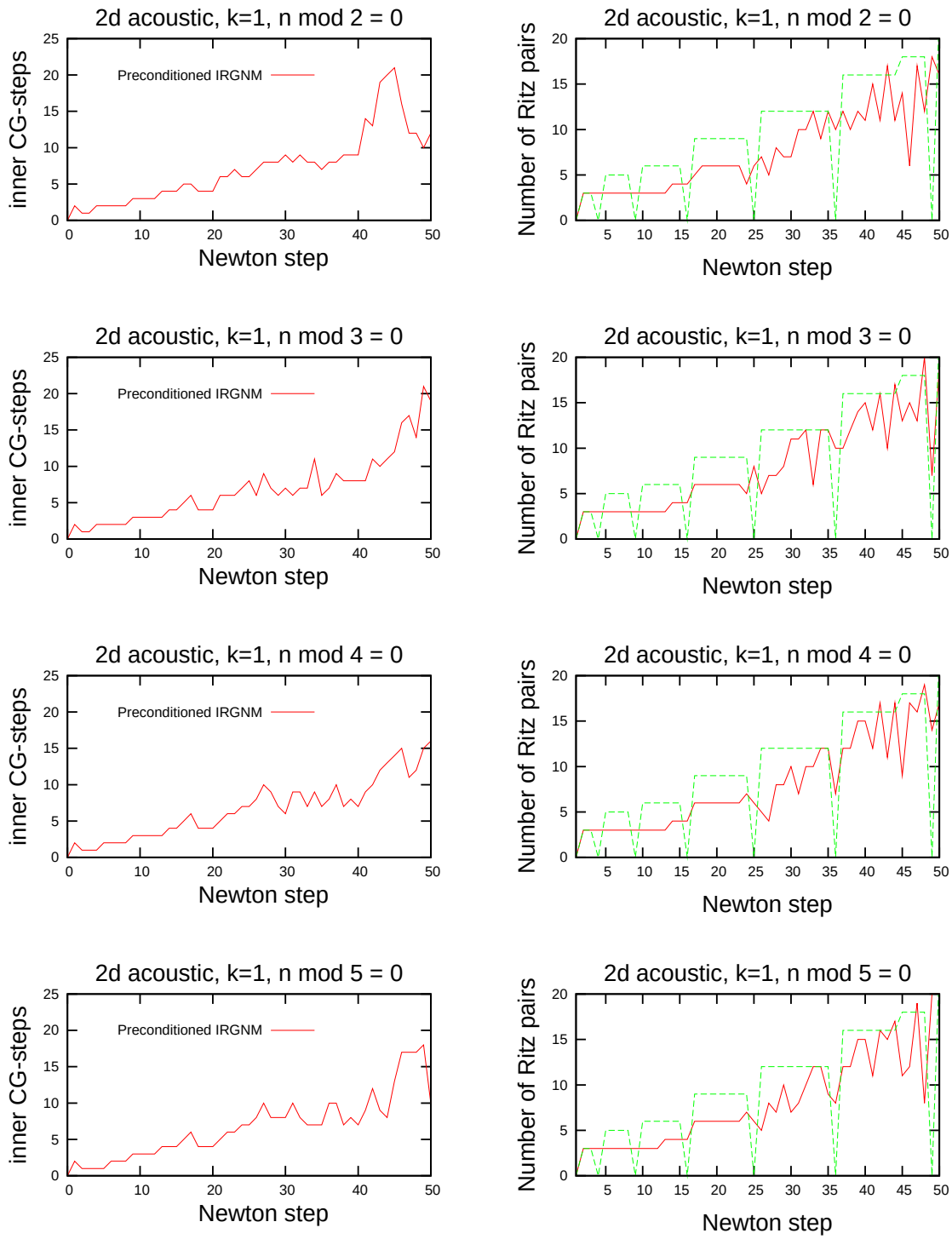

Figure 7.21: Number of inner CG-iterations compared with the number of Ritz pairs

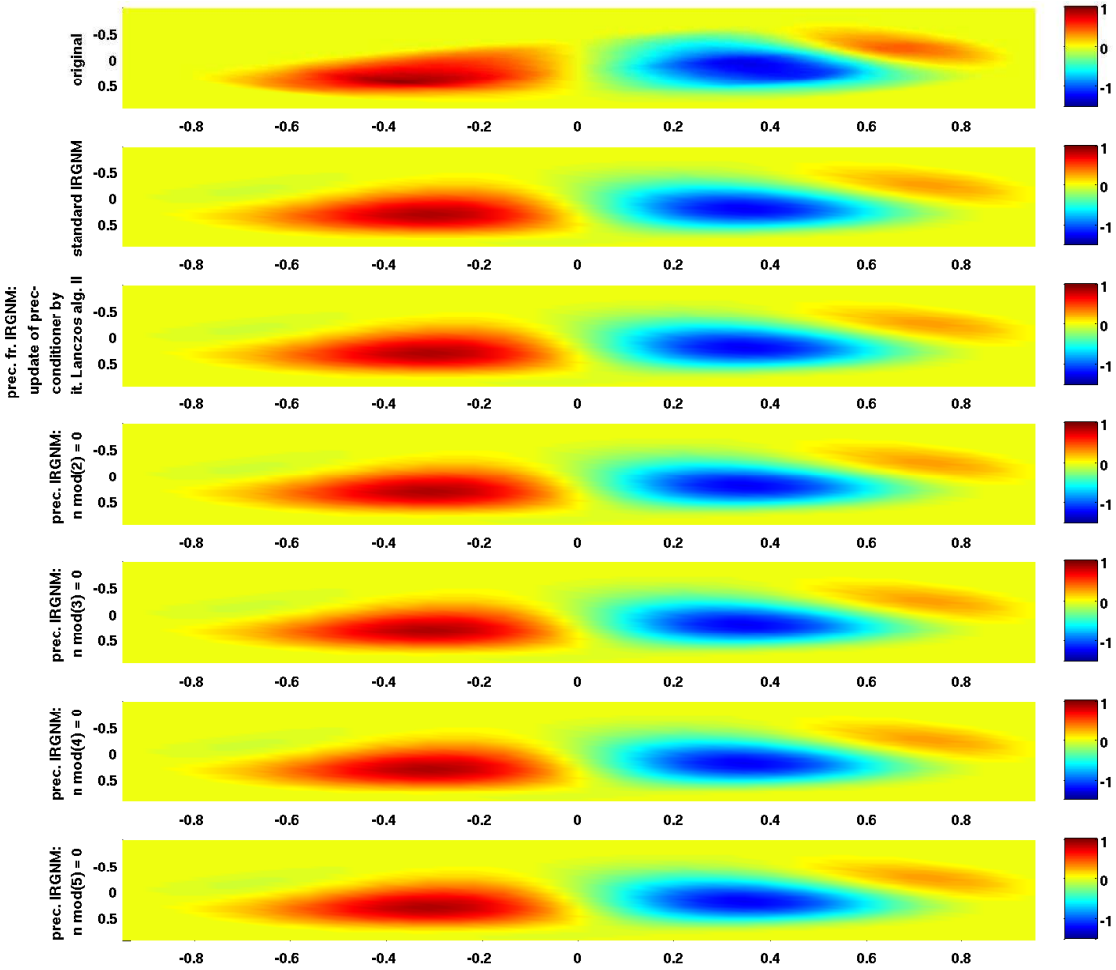

Figure 7.22: Reconstructions of the refractive index after 50 Newton steps for the standard IRGNM, the preconditioned frozen IRGNM and the preconditioned IRGNM for different choices of p

7.4 Conclusion

Concluding we can say that in average the preconditioned frozen IRGNM in combination with the iterated Lanczos algorithm I was the most efficient one of the considered algorithms. To update the preconditioner by the iterated Lanczos algorithm II often yielded slightly inferior results when we compared the algorithms by the total complexity. In the examples considered in this thesis and in many more experiments we performed, these algorithms significantly reduced the total complexity when compared with a standard IRGNM. Even when compared with the algorithm presented in [40] these algorithms were usually superior. Hence, these algorithms in fact are adequate for solving large-scale nonlinear ill-posed problems. The significant reduction of computational time we obtained by these algorithms is not negligible.

We do not want to hide that the success of the updating process of the preconditioner depends on many parameters such as the function (6.10), which usually need to be chosen a-priori. Unfortunately we were not able to implement a fully automatic choice of these parameters. On the other hand, we could often observe that a change of the parameters did not lead to totally different results.

To overcome the problems arising from the choice of a certain update criterion f_{up} and the choice of the different parameters we spent a lot of time trying to implement a satisfactory version of Algorithm 6.10. Unfortunately, it finally turned out that the approximation quality of the Ritz pairs determined by Lanczos' method in such an algorithm were not adequate for constructing efficient spectral preconditioners. Moreover, we do not have much hope that a satisfactory version of such an algorithm can be realized with techniques presented in this thesis. The considered spectral preconditioners react too sensitive to errors in the eigenelements. Hence, maybe other preconditioning techniques are more promising.

Chapter 8

Conclusion and outlook

This final chapter is devoted to a reconsideration of the IRGNM with respect to the three key points discussed in the introduction:

a) **Accuracy**,

b) **Complexity**,

c) **Acceleration**.

We claimed that the questions concerning these three aspects would be answered in this thesis. Therefore, it is the goal of this last chapter to discuss in what sense we have been able to answer these questions and what is still left open and should possibly be targeted in future work.

Let us start with the **accuracy**. The main result of this thesis concerning this topic has been to establish convergence and optimal rates of convergence for the IRGNM under general source conditions for both an a-priori and an a-posteriori stopping rule. Our results involve the practically important case that the linearized equations are not solved exactly in each Newton step. Moreover, explicit bounds on this additional error have been formulated and we could prove that when the arising regularized linear systems are solved by the CG-method, a reasonable stopping criterion could be posed such that the error bound is satisfied. In this sense we were able to generalize not only the well-known convergence and convergence rate results of the IRGNM with respect to the source condition, but also to carry over these results to an *inexact* IRGNM.
This achievement is particularly important for large-scale problems where usually only approximate solutions to the linear systems arising in Newton's method can be determined using iterative methods. From this point of view we answered the question of **accuracy**.
Still, unfortunately the convergence proofs for the IRGNM we presented are based on the nonlinearity conditions (2.11) which for many interesting problems are open.

For example for the inverse scattering problems discussed in Chapter 7 these conditions could not be proven so far and it is questionable if these conditions are at all satisfied. Hence, the local convergence proof of the IRGNM for many applications is not complete at this time and therefore not satisfactory so far. For examples for which the nonlinearity conditions (2.11) could be proven we refer to [44].
However, even if the nonlinearity conditions (2.11) are possibly unrealistic for many nonlinear ill-posed problems, the main idea of the convergence proof is the splitting of the total error into several components which can be analyzed separately. The nonlinearity conditions (2.11) are only one possibility to estimate the terms concerned with the nonlinearity. For certain problems one possibly finds other ways to handle these terms.

A further important assumption of the convergence proof which we think is worthwhile to be reconsidered is the smoothness assumption on the true solution which is expressed by the source condition (2.4). Note that in particular because of this assumption we have not proved that the IRGNM is an iterative regularization method in the sense of Definition 2.1. Furthermore, in practice it is usually unknown if such a condition is satisfied, since not only the true solution is not available, but also one usually does not have exact knowledge of a function determining an appropriate source set for the given problem. In this sense the source condition is an assumption which cannot be verified a-priori in general. Hence, if only given a set of noisy measurements it is a-priori not clear if the IRGNM reconstructs model parameters which generated the measured data. On the other hand, if the measured data are generated by some smooth parameters, then we have convergence and we know that the final iterates computed by the IRGNM are optimal in the sense that optimal rates of convergence are achieved.

The last drawback of the IRGNM we want to mention is the local character of its convergence which is inherited from Newton's method. On the one hand this remark possibly seems trivial, on the other hand for realistic applications it is important to keep it in mind, since for ill-posed problems it is usually a hard task to determine a good initial guess ensuring convergence to the reconstruction of the true solution. For example, if one considers the problem of reconstructing the shape of some scatterers usually a-priori the number of scatterers is unknown. Hence, before we can apply the IRGNM in a first step the number of scatterers needs to be determined, and moreover, a good initial guess of their position is essential for convergence. For inverse acoustic scattering problems *sampling and probe methods* have been shown to be successful methods for these tasks. Some of these methods do even work without knowledge of the boundary conditions. Naturally the reconstructions one obtains are usually of lower quality than the ones we can achieve with a Newton-type method.
Still, due to the local convergence behavior the IRGNM may have to be combined with methods which are able to compute sufficiently good initial guesses. Given some good initial guess, Newton-type methods usually yield stable and reasonably

good reconstructions. Furthermore, one obtains a parametrized final iterate which is often more suitably for further applications than a set of points determined by *sampling and probe methods*. Still, the IRGNM can only be an efficient method if either from the applications on its own or by other mathematical or heuristic methods sufficiently good initial guesses are at hand.

To discuss which results we have achieved with respect to the **complexity** of the IRGNM recall Theorems 4.20 and 4.21. The proofs of these theorems are based on the fundamental upper bounds on the number of CG-steps in the n-th Newton step shown in Theorem 4.19, which are a consequence of the stopping criterion (4.11) for the CG-method. Since we were able to establish these upper bounds both for mildly and exponentially ill-posed problems, we finally could express the total complexity of the IRGNM and its preconditioned version in terms of the noise level $\delta > 0$ by a combination of Theorem 4.19 and Corollary 2.5.
Moreover, the upper bounds presented in Theorem 4.19 are flexible in the sense that for any other stopping criterion than (2.5a) or (2.5b) for the outer Newton iteration the total complexity of the IRGNM can easily be determined. One just has to sum up the inner CG-steps in each Newton step until the stopping rule for Newton's iteration is reached. Moreover, the upper bounds in Theorem 4.19 have the advantage that they just rely on (4.11), but not on the nonlinearity conditions (2.11).
Note that our complexity result also includes the case of linear ill-posed problems when they are solved by the algorithm presented in Section 1.3, that is we have proven an upper bound for the total complexity to determine algorithmically the regularization parameter such that the discrepancy principle is satisfied. For linear ill-posed problems this complexity can be significantly reduced when we combine the algorithm presented in Section 1.3 with preconditioning techniques of Chapter 6.

We do not want to hide that there is a gap in our complexity result. To prove the assertions for the standard IRGNM formulated in Theorem 4.20 we had to impose the additional estimate (4.43). The gap is due to the fact that we could only show that the estimates (2.24) and (2.32) are satisfied for the stopping criterion (4.10), but not for the stopping criterion (4.11). Estimate (4.43) was the tool to conclude out of (4.11) the validity of (4.10). Hence, an application of Corollary 2.5 was possible. Although we gave some heuristic arguments that estimate (4.43) is reasonable, a general proof for such an estimate is still missing. On the other hand, we only wanted to prove Theorem 4.20 for the stopping criterion (4.11) since we used this in practice. For the theory we could exchange (4.11) against (4.10). Then no heuristic argumentation is necessary.
Recall that in order to obtain the the complexity results of Theorem 4.21 we formulated many assumptions which seem to be unrealistic in practice. On the other hand, Theorem 4.21 serves as a good motivation for the possible success of a preconditioned frozen Newton method, that is a significant reduction of the total complexity compared to a standard IRGNM.

Moreover, having the numerical examples of Chapter 7 in mind many of the formulated assumptions do not seem to be too digressive. In the examples we considered the number of determined Ritz pairs appeared to be sufficiently large to justify for example the estimate (4.47b). And since the final iterates throughout the algorithms were comparable the application of Corollary 2.5 is legitimated. Moreover, we think that the assertion of Corollary 2.5 could also be obtained theoretically for a frozen Newton method.
We always had the hope that the result of Theorem 4.19 could be improved such that the number of inner CG-iterations concerned with the cluster of eigenvalues in a neighborhood of γ_n grows slower than linearly with the Newton step n. As a consequence of such a result (4.45) and (4.49) could be significantly improved. Unfortunately this problem appeared to be harder than it seemed and the improvement is desirable.
Concluding, we could not only show in numerical examples that the implementation of the preconditioned frozen IRGNM was superior to a standard one, but we also gave theoretical arguments to this end. Our considerations with respect to this topic delivered the missing analysis of Algorithm 4.10 which was originally published in [40].

Finally, let us recapitulate the results on **acceleration** techniques for the IRGNM. Due to our discussion in Section 5.2 we once again want to emphasize that spectral preconditioners of the form (5.6) are in particular adequate for the linear systems arising in the IRGNM for large-scale problems in three space dimensions. Unfortunately the sensitivity analysis shows limitations of these kind of preconditioners we have to deal with. We have realized this by an application of the computable a-posteriori error bounds from Lanczos' method to obtain indicators for the approximation quality of the Ritz pairs. These indicators are exploited to select the Ritz pairs of high approximation quality to set up the spectral preconditioner.
After a careful consideration of the numerical examples in Chapter 7 the presented preconditioning techniques work quite convincingly, since indeed the total complexity of the standard IRGNM could be significantly reduced. Moreover, even when compared with the preconditioned frozen IRGNM presented in [40] it has turned out that the updating procedure we implemented once again yields a significant reduction of the total complexity. In particular for the inverse acoustic scattering problem in three space dimensions with the update technique of the preconditioner the original complexity of the standard IRGNM could be reduced to about 1/3.

In summary Algorithm 6.8 coupled with Algorithm 6.5 has turned out to be the most efficient among the algorithms we presented. We do not want to hide the fact that an implementation of these algorithms as well as Algorithm 6.6 involves the choice of many parameters, which can be seen as their major drawback, since we were not able to realize a fully automatic choice of all the threshold parameters arising in the different algorithms used for its realization. On the other hand, the choice of these tuning parameters is not as important as the choice of the stopping

index as they only influence the efficiency of the algorithm, but not the accuracy of the final result. Moreover, in further numerical experiments we performed it could be seen that the change of the parameters often yielded comparable results, as long as the parameter choice took into account the sensitivity of the preconditioner with respect to approximation quality of the Ritz pairs.
More crucial for an efficient implementation of Algorithm 6.8 is the choice of the update criterion f_{up}, that is a function balancing the convergence speed of the outer Newton iteration and the total complexity. As already mentioned in Chapter 4 and Chapter 6 we think that an optimal choice of this function depends on many variables usually unknown a-priori. On the other hand, given a certain large-scale problem, which is computationally complex, it is surely recommended to think about a reasonable choice of the update function. Otherwise the reduction of the complexity in each step of a frozen Newton method possibly has to be paid by an increase on the total number of Newton steps until some stopping criterion terminates the outer Newton iteration. To this end in future work it would be interesting to investigate the convergence of a frozen IRGNM, and furthermore if a similar result as presented in Corollary 2.5 can be obtained for such a method. Such a result could maybe give some hints on the choice of general update functions, which is naturally not optimal for a certain problem but optimal in average.
For the inverse scattering problems we considered in this thesis we also tried different update functions. Another natural choice for the update function with respect to the complexity would be to perform an update of the operator if some upper bound on the number of inner CG-iterations during the IRGNM is exceeded although the the operator is preconditioned. Such a procedure is motivated by the observation that after several updates usually the preconditioner starts losing its efficiency due to increasing errors in the approximations to the eigenpairs. Often such criteria yielded comparable or worse results than the criterion suggested in [40]. Finally, the determination of a general update criterion is an open problem.

Note that the question for an update function would not arise if Algorithm 6.10 coupled with Algorithm 6.11 yielded comparable results to Algorithm 6.8. Unfortunately, in the way we have realized this algorithm this is not the case. The difficulties arising in an implementation were already discussed in Chapter 6 and Chapter 7 and we do not have much hope that the procedure sorting out the Ritz pairs can be refined in such a way that the resulting preconditioners turn out to be more efficient in general.
Hence, in our opinion maybe other preconditioning techniques can be more successful. For example, instead of setting up a spectral preconditioner one could consider preconditioning by projection, that is to solve the linear systems arising in the IRGNM in a first step on the subspace defined by the known Ritz vectors. Under the condition that enough and good approximations corresponding to the largest eigenpairs are known and the weight of the components in the right hand side vector does not lie on the small eigenvalues we should obtain a good approximation

to the solution of the original linear system. In a second step this approximation could be refined by using it as initial guess for an iterative method. Naturally, this serves just as an idea. There exist many other preconditioning techniques which seem worthwhile to be attempted.

The last point we want to mention is that the convergence and complexity theory we presented also includes the case of mildly ill-posed problems, although we did not give an example for this case. This is simply due to the fact that we did not have an interesting nonlinear large-scale mildly ill-posed example at hand. In particular it would be interesting to investigate if Algorithm 6.8 works for mildly ill-posed problems just as well as for exponentially ill-posed problems. This is not clear since the eigenvalue distribution is not as well suited for Lanczos' method in this case. Hence, in future work Algorithm 6.8 should be applied to some mildly ill-posed problems to find out if the presented preconditioning techniques are also successful in this situation.

Bibliography

[1] O. Axelsson, *A survey of preconditioned iterative methods for linear systems of algebraic equations*, BIT Numerical Mathematics, 25 (1985), pp. 165–187.

[2] ——, *Iterative solution methods*, Cambridge University Press, 1994.

[3] ——, *Optimal preconditioners based on rate of convergence estimates for the conjugate gradient method*, Num. Funct. Anal. and Optimiz., 22 (2001), pp. 277–302.

[4] ——, *On the rate of convergence of the conjugate gradient method for linear operators in Hilbert spaces*, Num. Funct. Anal. and Optimiz., 3 (2002), pp. 285–302.

[5] A. Bakushinskii, *The problem of the convergence of the iterativley regularized Gauß-Newton method*, Comput. Maths. Math. Phys., 32 (1992), pp. 1353–1359.

[6] ——, *Iterative methods without saturation for solving degenerate nonlinear operator equations*, Dokl. Akad. Nauk, 1 (1995), pp. 7–8.

[7] A. Bakushinskii and M. Kokurin, *Iterative methods for approximate solution of inverse problems*, Springer, 2004.

[8] F. Bauer and T. Hohage, *A Lepskij-type stopping rule for regularized newton methods*, Inverse Problems, 21 (2005), pp. 1975–1991.

[9] B. Blaschke, A. Neubauer, and O. Scherzer, *On convergence rates for the iteratively regularized Gauß-Newton method*, IMA Journal of Numerical Analysis, 17 (1997), pp. 421–436.

[10] B. Carpentieri, I. S. Duff, and L. Giraud, *A class of two-level preconditioners*, SIAM J. Sci. Comput., 25 (2003), pp. 749–765.

[11] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, Springer Verlag, Berlin, Heidelberg, New York, second ed., 1997.

[12] J. Demmel, *Applied Numerical Linear Algebra*, Society for Industrial and Applied Mathematics, 1997.

[13] P. Deuflhard, *Numerische Mathematik I*, de Gruyter, 2002.

[14] P. Deuflhard, H. W. Engl, and O. Scherzer, *A convergence analysis of iterative methods for the solution of nonlinear ill-posed problems under affinely invariant conditions*, Inverse Problems, 14 (1998), pp. 1081–1106.

[15] H. Egger and A. Neubauer, *Preconditioning Landweber iteration in Hilbert scales*, Numerische Mathematik, 101 (2005), pp. 643–662.

[16] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Kluwer Academic Publisher, Dordrecht, Boston, London, 1996.

[17] B. Fischer, *Polynomial Based Iteration Methods for Symmetric Linear Systems*, Wiley Teubner, Sussex, Stuttgart, 1996.

[18] B. Fischer and R. W. Freund, *On adaptive weighted polynomial preconditioning for hermitian positive definite matrices*, SIAM J. Sci. Comput., 15 (1994), pp. 408–426.

[19] M. Fisher, *Minimization algorithms for variational data assimilation*, Proceedings of the ECMWF seminar "Recent developments in numerical methods for atmospheric modelling", UK, (1998), pp. 364–385.

[20] O. Forster, *Analysis 3*, Vieweg, Braunschweig, 1984.

[21] J. Frank and C. Vuik, *On the construction of deflation-based preconditioners*, SIAM J. Sci. Comput., 23 (2001), pp. 442–462.

[22] L. Giraud and S. Gratton, *On the sensitivity of some spectral preconditioners*, CERFACS Technical report TR/PA/01/108, (2004).

[23] ——, *On the sensitivity of some spectral preconditioners*, SIAM J. Matrix Anal. Appl., 27 (2006), pp. 1089–1105.

[24] G. H. Golub and C. F. van Loan, *Matrix Computations*, The John Hopkins University Press, Baltimore, second ed., 1983.

[25] G. H. Golub and R. S. Varga, *Chebyshef semi-iterative methods, successive overrelaxation iterative methods, and second order richardson iterative methods*, Numer. Math., 3 (1961), pp. 147–168.

[26] A. Greenbaum, *Comparison of splittings used with the conjugate gradient algorithm*, Numerische Mathematik, 33 (1979), pp. 181–194.

[27] N. Grinberg, *Factorization method in inverse obstacle scattering*, University of Karlsruhe, Habilitation thesis, 2004.

[28] J. Hadamard, *Lectures on Cauchy's problem in linear partial differential equations*, Dover publications, New York, 1952.

[29] P. Hähner and T. Hohage, *New stability estimates for the inverse acoustic inhomogeneous medium problem and applications*, SIAM J. Math. Anal., 33 (2001), pp. 670–685.

[30] M. Hanke, *A regularizing Levenberg-Marquardt scheme, with applications to inverse groundwater filtration problems*, Inverse Problems, 13 (1997), pp. 79–95.

[31] ——, *Regularizing properties of a truncated Newton-CG algorithm for nonlinear inverse problems*, Numer. Funct. Anal. Optim., 18 (1997), pp. 971–993.

[32] M. Hanke and J. Nagy, *Restoration of atmospherically blurred images by symmetric indefinite conjugate gradient techniques*, Inverse Problems, 12 (1996), pp. 157–173.

[33] M. Hanke, A. Neubauer, and O. Scherzer, *A convergence analysis of the Landweber iteration for nonlinear ill-posed problems*, Numer. Math., 72 (1995), pp. 21–37.

[34] R. Hayes, *Iterative methods of solving linear problems in Hilbert space*, Nat. Bur. Standards Appl. Math. Ser., 39 (1954), pp. 71–104.

[35] M. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Research Nat. Bur. Standards, 49 (1952), pp. 409–436.

[36] H. Heuser, *Funktionalanalysis*, Teubner Verlag, Stuttgart, 3rd ed., 1991.

[37] T. Hohage, *Logarithmic convergence rates of the iteratively regularized Gauß-Newton method for an inverse potential and an inverse scattering problem*, Inverse Problems, 13 (1997), pp. 1279–1299.

[38] ——, *Iterative Methods in Inverse Obstacle Scattering: Regularization Theory of Linear and Nonlinear Exponentially Ill-Posed Problems*, PhD thesis, University of Linz, 1999.

[39] ——, *Regularization of exponentially ill-posed problems*, Numer. Funct. Anal. Optim., 21 (2000), pp. 439–464.

[40] ——, *On the numerical solution of a three-dimensional inverse medium scattering problem*, Inverse Problems, 17 (2001), pp. 1743–1763.

[41] ——, *Fast numerical solution of the electromagnetic medium scattering problem and applications to the inverse problem*, Journal of Computational Physiscs, 214 (2006), pp. 224–238.

[42] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge Univerity Press, Cambridge, 2nd ed., 1990.

[43] V. Ivanov and T. Korolyuk, *Error estimates for solutions of incorrectly posed linear problems*, USSR Comput. Math. Math. Phys., 9 (1969), pp. 35–43.

[44] B. Kaltenbacher, *Some Newton-type methods for the regularization of nonlinear ill-posed problems*, Inverse Problems, 13 (1997), pp. 729–753.

[45] ——, *A posteriori parameter choice strategies for some Newton type methods for the regularization of nonlinear ill-posed problems*, Numer. Math, 79 (1998), pp. 501–528.

[46] A. Kirsch, *Characterization of the shape of the scattering obstacle by the spectral data of the far field operator*, Inverse Problems, 14 (1998), pp. 1489–1512.

[47] ——, *Factorizations of the far field operator for the inhomogeneous medium case and an application in inverse scattering theory*, Inverse Problems, 15 (1999), pp. 413–429.

[48] A. Kirsch and R. Kress, *On an integral equation of the first kind in inverse scattering*, Inverse problems (Cannon and Hornung, eds.), ISNM, 77 (1986), pp. 93–102.

[49] ——, *A numerical method for an inverse scattering problem*, (Engl and Groetsch, eds.) Academic Press, Orlando, (1987), pp. 279–290.

[50] ——, *An optimization method in inverse acoustic scattering*, Boundary Elements IX, Fluid Flow and Potential Applications (Brebbia et al., eds.), Springer Verlag, Berlin, Heidelberg, New York, 3 (1987), pp. 3–18.

[51] A. Knyazev, *New estimates for ritz vectors*, Math. Comp. 66, 219 (1997), pp. 985–995.

[52] R. Kress, *Linear Integral Equations*, Springer Verlag, Berlin, Heidelberg, New York, 2nd ed., 1999.

[53] A. Kuijlaars, *Which eigenvalues are found by the Lanczos method?*, SIAM J. Matrix Anal. Appl., 22 (2000), pp. 306–321.

[54] P. Lancaster and M. Tismenetsky, *The theory of matrices*, Academic Press , Inc., Orlando, Florida 32887, 2nd ed., 1985.

[55] C. Lanczos, *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*, J. Res. Nat. Bur. Stand., 45 (1950), pp. 255–282.

[56] ——, *Chebyshev polynomials in the solution of large-scale linear systems*, Proceedings of the Association for Computing Machinery, Toronto, Sauls Lithograph Co., Washington, DC, (1953), pp. 124–133.

[57] S. Langer, *Integralgleichungsmethode für das Neumann Problem in einem Gebiet mit Schlitz*, University of Göttingen, Diplomarbeit, 2003.

[58] A. K. Louis, *Inverse und schlecht gestellte Probleme*, Teubner Verlag, Stuttgart, 1989.

[59] B. A. Mair, *Tikhonov regularization for finitely and infinitely smoothing operators*, SIAM J. Math. Anal., 25 (1994), pp. 135–147.

[60] L. Mansfield, *On the conjugate gradient solution of the schur complement system obtained from domain decomposition*, SIAM J. Numer. Anal., 20 (1990), pp. 1612–1620.

[61] P. Mathé and S. Pereverzev, *Geometry of linear ill-posed problems in variable Hilbert scales*, Inverse Problems, 19 (2003), pp. 789–803.

[62] ——, *The discretized discrepancy principle under general source conditions*, Journal of Complexity, 22 (2006), pp. 371–381.

[63] A. A. Melkman and C. A. Micchelli, *Optimal estimation of linear operators in Hilbert spaces from inaccurate data*, SIAM J. Numer. Anal., 16 (1979), pp. 87–105.

[64] G. Meurant, *Computer solution of large linear systems*, North-Holland Publishing Co, Amsterdam, 1999.

[65] V. Morozov, *On the solution of functional equations by the method of regularization*, Soviet Math. Dokl., 7 (1966), pp. 414–417.

[66] J. Nagy and D. O'Leary, *Restoring images degraded by spatially variant blur*, SIAM J. Sci. Comp., 19 (1998), pp. 1063–1082.

[67] F. Natterer, *The Mathematics of Computerized Tomography*, Teubner Verlag, Stuttgart, 1986.

[68] A. Neubauer, *Tikhonov regularization for nonlinear ill-posed problems: optimal convergence and finite-dimensional approximation*, Inverse Problems, 5 (1989), pp. 541–557.

[69] R. A. Nicolaides, *Deflation of conjugate gradients with applications to boundary value problems*, SIAM J. Numer. Anal., 24 (1987), pp. 355–365.

[70] A. Padiy, O. Axelsson, and B. Polman, *Generalized augmented matrix preconditioning approach and its application to iterative solution of ill conditioned algebraic systems*, SIAM J. Matrix Anal. Appl., 22 (2000), pp. 793–818.

[71] B. Parlett, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Inc., Englewood Cliffs, N.J. 07632, 1980.

[72] R. Potthast, *A point source method for inverse acoustic and electromagnetic obstacle scattering problems*, IMA J. Appl. Math., 2 (1998), pp. 119–140.

[73] J. R. Rice, *A theory of condition*, SIAM J. Numer. Anal., 3 (1966), pp. 287–310.

[74] A. Rieder, *On the regularization of nonlinear ill-posed problems via inexact Newton iterations*, Inverse Problems, 15 (1999), pp. 309–327.

[75] ——, *Keine Probleme mit Inversen Problemen*, Vieweg Verlag, Wiesbaden, 2003.

[76] ——, *Inexact Newton regularization using conjugate gradients as inner iteration*, SIAM J. Numer. Anal., 43 (2005), pp. 604–622.

[77] Y. Saad, *Iterative Methods for Sparse Linear Systems*, International Thomson Publishing, Boston, 1996.

[78] G. Stewart, *Error and perturbation bounds for subspaces associated with certain eigenvalue problems*, SIAM Rev., 15 (1973), pp. 727–764.

[79] ——, *Matrix Algorithms, Volume II: Eigensystems*, SIAM, Philadelphia, 2001.

[80] G. Szegö, *Orthogonal polynomials*, Volume 23 of Colloqium Publications, AMS, Providence, RI, 1975.

[81] U. Tautenhahn, *Optimality for linear ill-posed problems under general source conditions*, Numer. Funct. Anal. Optim., 19 (1998), pp. 377–398.

[82] G. M. Vainikko, *Fast solvers of the Lippmann-Schwinger equation*, in Direct and Inverse Problems of Mathematical Physics, R. P. Gilbert, J. Kajiwara, and Y. S. Xu, eds., Kluwer Acad. Publ., Dordrecht, 1999.

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

[84] H. van der Vorst, *Iterative Krylov Methods for Large Linear Systems*, Cambridge University Press, Cambridge, 2003.

[85] R. Winther, *Some superlinear convergence results for the conjugate gradient method*, SIAM Journal on Numerical Analysis, 17 (1980), pp. 14–17.

Curriculum vitae — Lebenslauf

Name:	Stefan Langer
geboren:	27. Dezember 1977
Geburtsort:	Kassel
Familienstand:	ledig
Staatsangehörigkeit:	deutsch
Wohnsitz:	Göttingen

Akademische Ausbildung

06/97:	Abitur am Engelsburggymnasium in Kassel
10/98 – 07/03:	Studium der Mathematik an der Georg-August-Uninversität Göttingen
seit 10/03:	Promotionsstudium an der mathematischen Fakultät in Göttingen
10/03 – 03/04:	wissenschaftlicher Mitarbeiter am Institut für NAM der Georg-August-Uninversität Göttingen
04/04 – 09/05:	Stipendiat im DFG Graduiertenkolleg 1023 "Identification in Mathematical Models: Synergy of Stochastic and Numerical Methods"
seit 10/05:	wissenschaftlicher Mitarbeiter am Institut für NAM der Georg-August-Universität Göttingen

_____ _____

Datum, Ort Stefan Langer