INAUGURAL-DISSERTATION
zur
Erlangung der Doktorwürde
der
Naturwissenschaftlich-Mathematischen Gesamtfakultät
der
Ruprecht-Karls-Universität Heidelberg

vorgelegt von
M.Sc. María Elena Suárez Garcés
aus Venezuela

Tag der mündlichen Prüfung:

Iterative Linear Algebra for Parameter Estimation

Betreuer: PD. Dr. Andreas Potschka
Prof. Dr. Dr. h.c. mult. Hans Georg Bock

Zusammenfassung

Das Hauptziel dieser Arbeit ist die Entwicklung und Analyse effizienter numerischer Methoden für große nichtlineare Parameterschätzprobleme. Solche Probleme haben eine hohe Relevanz in vielen Bereichen der angewandten Mathematik, die versuchen das zukünftige Verhalten von Prozessen basierend auf großen Datensätzen vorherzusagen indem zunächst ein mathematisches Modell erstellt wird und dieses dann extrapoliert wird. In dieser Arbeit interessieren wir uns für die Erstellung des mathematisches Modells. Dabei liegen die Schwierigkeiten in der Behandlung der Nichtlinearitäten und der schieren Größe der Datensätze und Unbekannten. Ein gängiger Ansatz zur numerischen Behandlung solcher Parameter-Schätzprobleme ist die Gauss-Newton-Methode, die darin besteht eine Reihe von linearisierten Unterproblemen zu lösen.

Ein Beitrag dieser Arbeit ist eine eingehende Analyse der Problemklasse auf Basis der kovarianten und kontravarianten κ-Theorie. Basierend auf dieser Analyse ist es möglich ein neues Abbruchkriterium für die iterativen Lösungen der inneren linearisierten Unterprobleme zu entwickeln. Die Auswertung zeigt, dass es ausreicht die inneren Unterprobleme nur mit geringer Genauigkeit zu lösen ohne dabei die Konvergenzgeschwindigkeit der äußeren Iterationen signifikant zu senken. Des Weiteren wird in dieser Arbeit gezeigt, dass das neue Abbruchkriterium ein quantitatives Maß dafür ist, wie genau die Lösung der Unterprobleme erfolgen muss um inexakte Gauss-Newton-Folgen zu erzeugen, die gegen eine statistisch stabile Abschätzung (deren Existenz wir voraussetzen) konvergieren. Daher liefert dieser Ansatz eine neuartige inexakte Gauss-Newton-Methode, die im Vergleich zu klassischen exakten Gauss-Newton-Methoden eine geringere Zahl innerer Iterationen zur Berechnung des inexakten Gauss-Newton Schritts benötigt. Auf diese Weise erhalten wir große Recheneinsparungen verglichen mit der klassischen exakten Gauss-Newton-Methode, die 100% innere Iterationen zur Berechnung des Gauss-Newton Schritts benötigt, was ungeheuer rechenintensiv ist, wenn die Zahl der Parameter zu groß ist. Des Weiteren verallgemeinern wir die lokalen Ideen dieses neuartige inexakten Gauss-Newton-Ansatzes und führen eine gedämpfte inexakte Gauss-Newton-Methode ein, indem wir die Backward Step Control for global Newton-type theory von Potschka benutzen.

Die Validierung unseres neuen Ansatzes erfolgt anhand zweier Beispiele. Zunächst betrachten wir ein Parameteridentifikationsproblem einer nichtlinearen, elliptischen, partiellen Differentialgleichung. Anschliessend untersuchen wir ein großes Parameterschätzungsproblem aus dem Bereich der Bildverarbeitung. Beide Beispiele sind schlecht konditioniert, weshalb eine günstige Regularisierung angewandt wird. Mithilfe unsere numerischen Experimente konnten wir bestätigen, dass, wie von unserer Theorie voausgesagt, der neueartige inexakte Gauss-Newton-Ansatz, die wenniger als 3% Zahl innerer Iterationen zur Berechnung des inexakten Gauss-Newton-Schritts benötigen um gegen eine statistisch stabile Abschätzung konvergieren.

Abstract

The principal goal of this thesis is the development and analysis of efficient numerical methods for large-scale nonlinear parameter estimation problems. These problems are of high relevance in all sciences that predict the future using big data sets of the past by fitting and then extrapolating a mathematical model. This thesis is concerned with the fitting part. The challenges lie in the treatment of the nonlinearities and the sheer size of the data and the unknowns. The state-of-the-art for the numerical solution of parameter estimation problems is the Gauss-Newton method, which solves a sequence of linearized subproblems.

One of the contributions of this thesis is a thorough analysis of the problem class on the basis of covariant and contravariant κ-theory. Based on this analysis, it is possible to devise a new stopping criterion for the iterative solution of the inner linearized subproblems. The analysis reveals that the inner subproblems can be solved with only low accuracy without impeding the speed of convergence of the outer iteration dramatically. In addition, I prove that this new stopping criterion is a quantitative measure of how accurate the solution of the subproblems needs to be in order to produce inexact Gauss-Newton sequences that converge to a statistically stable estimate provided that at least one exists. Thus, this new local approach results to be an inexact Gauss-Newton method that requires far less inner iterations for computing the inexact Gauss-Newton step than the classical exact Gauss-Newton method based on factorization algorithm for computing the Gauss-Newton step that requires to perform 100% of the inner iterations, which is computationally prohibitively expensive when the number of parameters to be estimated is large. Furthermore, we generalize the local ideas of this local inexact Gauss-Newton approach, and introduce a damped inexact Gauss-Newton method using the Backward Step Control for global Newton-type theory of Potschka.

We evaluate the efficiency of our new approach using two examples. The first one is a parameter identification of a nonlinear elliptical partial differential equation, and the second one is a real world parameter estimation on a large-scale bundle adjustment problem. Both of those examples are ill conditioned. Thus, a convenient regularization in each one is considered. Our experimental results show that this new inexact Gauss-Newton approach requires less than 3% of the inner iterations for computing the inexact Gauss-Newton step in order to converge to a statistically stable estimate.

Danksagungen

Mein besonderer Dank gilt dem Interdisziplinären Zentrum fur Wissenschaftliches Rechnen (IWR), der Fakultät für Mathematik und Informatik der Universität Heidelberg, und der Model-Based Optimizing Control (MOBOCON) Group für deren finanzielle Unterstützung. Ich danke Herrn Dr. Andreas Potschka, meinem Doktorvater, für die wissenschaftliche Betreuung, seine Hilfsbereitschaft und für die vielen anregenden Diskussionen. Jede Phase dieser Arbeit wurde von ihm intensiv, professionell, wie warmherzig begleitet. Weiterhin danke ich Herrn Prof. Dr. Georg Bock für sein Mitwirken, der freundlichen Hilfe und der mannigfachen Ideengebung, die mir einen kritischen Zugang zu dieser Thematik eröffnete.
Die wunderbare Atmosphäre am Interdisziplinären Zentrum für Wissenschaftliches Rechnen (IWR) spielte eine entscheidende Rolle während meines Forschens. Ich habe nützlichen Rat erfahren dürfen, wenn es zu Problemen bezüglich Arbeit und Leben gekommen ist.
An alle Kolleginnen und Kollegen, insbesondere den Mitglieder der MOBOCON Gruppe, geht mein herzlicher Dank für Ihre Freundschaft und Hilfe.
Schließlich bedanke ich mich auch bei meinen Eltern, meinen Brüdern, meinen Schwestern und meinem Ehemann, die mich fortwährend unterstützen.

Contents

Introduction

A large variety of natural, industrial, social and economical phenomena can be modeled by systems of partial differential equations (PDEs) where the solution describes the dynamic of such a phenomena. Most of the time the solution cannot be given explicitly and must be estimated from a finite number of indirect measurements. Thus, a discrete solution of such as PDEs that depends on a finite number of unknown parameters is proposed and an estimation process is implemented. If the discrepancy between the measurements and the discrete solution with real parameters is an aleatory variable, which is independent and normally distributed with expected value zero and variance-covariance matrix known, then we can obtain a plausible estimation of the real parameters through the solution of a large-scale nonlinear least squares problem [7], which is typically ill conditioned. Nevertheless, if a particular regularization for this optimization problems is available, we can reformulate it and obtain a well conditioned problem but large-scale nonlinear least squares problem. On the other hand, the estimation of discrete parameters can also yield a large-scale nonlinear least squares problems, as example we can consider the parameter estimation of large-scale bundle adjustment problems, whose may be ill conditioned. In this thesis, we focus on the treatment of the nonlinearities and the sheer size of measurement and unknown parameters. The state-of-the-art for numerically solving such as large-scale parameter estimation problems is the Gauss-Newton (GN) method, which is a variant of the Newton method for finding roots of a nonlinear equation in where second order derivative information is not taken into account. The GN method determines an estimation by solving a sequence of linearized subproblems whose solutions define the Gauss-Newton step. The principal drawback of such a GN approach is the computation at every iterate of the GN step, which may be computationally prohibitively expensive especially for large scale problems. Thus, inner iterative methods that determine an approximation of such GN step must be considered for ensuring numerical efficiency, which define different variations of the Gauss-Newton method known as the inexact Gauss-Newton (IGN) methods. In order to develop efficient IGN methods for large-scale nonlinear least squares problems, we require three ingredients.

(*i*) A cheap inner iterative method for approximately solving the linearized subproblems.

(*ii*) An early inner termination rule that only depends on cheaply available information, and that

(*iii*) the IGN sequence, which IGN step is generated using (i) and (ii), converges locally and linearly to a statistically stable solution.

An important question is: ($Q1$) What level of accuracy is required in approximately solving the linearized subproblems to preserve the local convergence of GN method?. The

answer is intimately related with the development of an inner termination rule that only depends on cheaply available information. Because an IGN sequence can also be considered as an inexact Newton (IN) sequence, we can reformulate the above question: $(Q2)$ What level of accuracy is required in approximately solving the linearized subproblems to preserve locally rapid convergence of Newton method?. The κ-theory is dedicated to give answer to $(Q2)$ question through the control of the discrepancies generated between the IN method and the Newton method. It classify the answer in two different approach: Covariant κ-Theorems that ensures locally rapid convergence of IN sequence, or contravariant κ-Theorems that ensure locally rapid convergence of the IN residual sequences. Thus, within the κ-theory the most popular measures of such discrepancies are given by: covariant error matrix, covariant inner residual relative error, contravariant error matrix, and contravariant inner residual relative error. Many authors presented Theorems that control one of the above errors and ensure local convergence of IN sequence for covariant approaches and IN residual sequence for contravariant approaches. Relevant examples are the following κ-Theorems:

- Ostrowski [64, Section 10.2.1] controlled how large must be the spectral radius of the contravariant error matrix or the spectral radius of the covariant error matrix, and concludes local convergence of the IN sequence with root factor of convergence.
- Dennis [24, Theorem 1] controlled how large must be the contravariant error matrix with $\|y\|$-norm, and concludes local and linear convergence of the IN sequence.
- Dembo, Stanley, Eisenstat, and Steihaug [22] controlled how large must be the contravariant inner residual relative error with $\|y\|$-norm, and conclude local and linear convergence of the IN sequence with a particular $\|y\|_*$-norm instead of $\|y\|$-norm.
- Bock [10] controlled how large must be the covariant inner residual relative error, and concludes local and linear convergence of the GN sequence to a statistically stable solution.

Furthermore, Cătinaş [16] studied what magnitudes can be allowed in perturbing the Newton matrix so that the convergence order of the resulting method does not decrease. Gratton, Lawless, and Nichols [40] introduced a deep analysis of truncated and perturbed GN methods. Deuflhard [27] studied how theoretical results from κ-Theorems can be exploited for the construction of adaptative algorithms. Hohmann [42] provided a computationally available stopping criterion that depends on a certain forcing sequence based on the calculation of sharpened contravariant quantities, and guarantees local and linear convergence. For a deeper study of κ-Theorems see e.g.,[64, 27, 22, 24, 70, 40, 16, 39, 31, 67, 18, 42].

The idea of inexact Gauss-Newton methods that satisfy (i), (ii), and (iii) is conceptually simple, but it is also surprisingly hard to propose a numerical method, which satisfies all those requirement. The principal problems are implementation, numerical efficiency, and statistically stable solutions.

Implementation. The issue in this part is the gap between the κ-theory results and practical implementations. Covariant κ-conditions deliver computationally unavailable termination rules, and the contravariant Theorem of Dembo, Stanley, Eisenstat, and

Steihaug [22] delivers a computationally available termination rule that control how large the contravariant inner residual relative error must be in order to conclude linearly and locally convergent with a particular $\|y\|_*$-norm of the inexact Newton sequence, which represents the principal drawback of this κ-condition since in such IGN methods second order derivative information is not available. Thus, if we want to provide an inner termination rule that satisfies (ii) based on a contravariant κ-condition, we must propose an affine contravariant κ-Theorem for our IGN method that controls how large the contravariant inner residual error with respect to the GN method must be in order to guarantee local convergence of IGN sequences.

Numerical efficiency of an IGN method depends on the numerical effort for the calculation of the IGN step at every iteration and the linear convergence factor of our IGN sequence. Thus, we must provide an inner termination criterion that is satisfied, as early as possible. In order words, we must ensure that the number of inner iterations necessary for computing IGN step using a certain numerical linear algebra for solving approximately the linearized problems is "small". An important question at this point is: How does the inaccuracy of an IGN method with respect to the GN method influence locally the convergence factor of IGN sequence.

Statistically stable solutions. We say that an estimation is statistically stable under statistical perturbations in the measurement data if it can be considered as a continuous deformation of the true parameter. Using a combination of the above κ-Theorems for determining an inner termination rule that satisfies (ii), we are interested in providing an IGN method, whose estimation is statistically stable. Nevertheless, within the κ-theory, we can conclude that IN sequence converges locally to an estimation or that IN residual sequences converges locally to zero, but we cannot conclude that such an estimation is statistically stable, in this scenario the κ-theory is not too wide. Bock [10] provides a local covariant κ-Theorem for GN method known as the local contraction Theorem, which ensures that the GN method converges linearly and locally to a statistically stable estimation, but as we said before covariant Theorems do not deliver computationally available termination rules for IGN methods.

The challenge in this thesis is that from covariant Theorems that guarantees locally rapid convergence of IGN sequence, we cannot obtain an available termination rule suitable for (ii), and (iii). On the other hand, there is a contravarinat Theorem that provides an available termination rule for inexact Newton methods, but not for IGN methods since here the second order derivative information is not available. Furthermore, not all κ-Theorem provides statistically stable solutions.

Contributions of the thesis

Assuming that the contravariant error matrix with $\|y\|$-norm introduced by the GN method is bounded by a κ_{GN}-constant less than one, we present a new IGN method that computes the IGN step using the LSQR [65] or the LSMR [33] Krylov subspace method as numerical linear algebra method for solving the inner linearized subproblems with an new early inner termination criterion that depends on cheaply available information, which implies that the contravariant inner residual relative error with respect to the

Newton method is bounded by a κ-constant less than one. Furthermore,

- The IGN step is computed using less inner iteration than the necessary for satisfying the standard termination rule of LSQR and LSMR based on the backward error provided by Stewart [78]. Indeed, we prove that the inner linearized subproblems can be solved with only low accuracy without impeding the speed of convergence of the outer iteration dramatically.

- We prove that there is a $\|y\|_*$-norm such that if the contravariant error matrix with $\|y\|$-norm is bounded by a κ-constant less than one, then the covariant error matrix with $\|y\|_*$-norm is bounded by a constant less than one. Reciprocally, there is a norm $\|y\|_\star$ such that if the covariant error matrix with $\|y\|$-norm is bounded by a constant less than one, then the contravariant error matrix with $\|y\|_\star$-norm is bounded by a constant less than one. Thus, we conclude: This new IGN method, which assumes that the contravariant error matrix with $\|y\|$-norm introduced by the GN method is bounded by κ_{GN} less than one, implies that the covariant error matrix with $\|y\|_*$-norm introduced by the GN method is bounded by a constant less than one. This result says that our first hypothesis is essentially a covariant hypothesis with $\|y\|_*$-norm. Moreover, this result can also be extended to our new stopping criterion since we prove that controlling the discrepancies between this IGN approach and the GN method, we can also conclude that our inner termination rule implies that the covariant inner realtive error introduced by our IGN method with norm $\|y\|_*$-norm is also bounded by a constant less than one. Both results allow to say that our IGN approach is essentially a covariant approach with $\|y\|_*$-norm.

- The results in the above item allow to conclude that the hypotheses with $\|y\|_*$-norm of the local contraction Theorem presented by Bock [10] for this IGN approach are valid. Therefore, it is possible to guarantee locally rapid convergences with $\|y\|_*$-norm of our IGN sequences. Moreover, we propose a κ-Theorem for our IGN approach that explains how the inaccuracy of this IGN approach with respect to the GN method influences locally and linearly the convergence factor with $\|y\|_*$-norm of the IGN sequence.

- We prove that this new IGN approach provides local statistically stable estimation provided that at least one exists.

- Because an efficient IGN method must deal with initial guesses that are not necessarily close to a local solution, we generalize the local ideas of this IGN approach, and introduce a damped IGN method using the Backward Step Control for Global Newton-type theory of Potschka [70], which ensures the existence of an inexact Gauss-Newton path $x(t)$ that connects a particular initial guess with some solution of our large scale nonlinear least squares problem and along it, the residual level function decreases exponentially. Furthermore, using a backward analysis argument based on following the above path $x(t)$, we provide a class of damped inexact Gauss-Newton sequences that converge to a particular local solution of our large-scale nonlinear least squares problem.

- We evaluate the efficiency of our new approach using two examples. The first one is a parameter identification of a nonlinear elliptic partial differential equation,

and the second one is a real world parameter estimation on a large-scale bundle adjustment problem. Both of those problems represent a challenge. The first one is a particular inverse problem where the parameter is in an infinite dimensional space, therefore a discrete form of the problem is considered using finite element methods, in this setting, we obtain a finite dimensional nonlinear least squares problem where only a finite number of unknown parameters can be estimated. In order to avoid ill posedness in our inverse problem, we focus on the regularization appoach proposed by Jun Zou [84], which provides at least a theoretical well posed problem such that its finite dimensional nonlinear reformulation is well conditioned. Our results show that our IGN approach requires just less than 3% of the inner iterations for computing the IGN step at every outer iteration, in spite of we work with a discretization that generated 1032 parameters to be estimated and the exact GN step at every outer iteration requires 1032 (100%) inner iterations to be computed. Furthermore, the estimation obtained is statistically stable. The challenge in the second example is that the Jacobian $J_f(x_k)$ is rank deficient at every outer iteration, and this problem is a large scale nonlinear least squares problem. We work with one experiment where 485013 parameters must be estimated, and obtain the IGN step at every outer iteration with just less than 1% of the inner iterations, which represent an enormous computational saves in comparison with the GN method based on factorization algorithm that requires to perform 485013 (100%) inner iterations in order to compute the GN step at every outer iteration.

Thesis overview

This thesis is organized as follows: Chapter 1 contains the parameter estimation formulation problem, as well as, the definition of Newton, GN, IGN, IN, and Newton-type method for numerically solving such a problem. We define the most popular errors that measure the discrepancy between GN and IGN method, and the discrepancy between IGN and Newton method, and set down the relation between the above different methods.

Chapter 2 presents the most popular numerical linear algebra for solving linear least squares problems: LSQR [65] or the LSMR [33], and introduces the new termination criterion that defines our IGN approach. We finalize this Chapter with the most relevant properties derived from this new IGN estrategy.

In Chapter 3, we prove that our IGN approach is essentially a covariant startegy with $\|y\|_*$-norm, and we discuss briefly when our IGN approach implies that the hypotheses with $\|y\|_*$-norm of the local contraction Theorem introduced by Bock [10] for our IGN approach are satisfied.

In Chapter 4, we prove that our IGN method guarantees statistically stable solutions provided that at least one exists.

In Chapter 5, we present an analysis based on the classical globalization strategies based on the popular Residual Monotonicity Test and on the Natural Monotonicity Test that reveals the principal drawbacks of globalization strategies based on a particular merit function. Rather, we focus on globalization strategies that follow the affine covariant Newton path $x(t)$. Thus, we survey two globalization strategies based on following

such a path $x(t)$, one of them was introduced by Bock, Kostina, and Schlöder [12] and is known as the Restrictive Monotonicity Test (RMT), and the other one was introduced by Potschka [70] and is known as the Backward Step Control (BSC) method, which provides, under reasonable assumptions, a global convergence Theorem from the basis of a backward step argument.

In Chapter 6, we introduce our damped IGN method based on BSC theory.

Finally, we evaluate the efficiency of our new IGN approach in Chapter 7 using two examples. The first example is a parameter identification of a nonlinear elliptical partial differential equation, and the second example is parameter estimation on a large-scale bundle adjustment problem.

Chapter 1

Preliminaries

1.1 Parameter Estimation Formulation Problem

The parameter estimation of a mathematical model is the process of finding a parameter which makes that our mathematical model reproduces, as close as possible, a collection of observed data. Let $D \subset \mathbb{R}^n$ be a nonempty open set, $h : D \subset \mathbb{R}^n \to \mathbb{R}^m$ be a twice continuously differentiable function representing such a mathematical model with $n \leq m$, and let us consider a series of observed data points

$$\eta_i \in \mathbb{R}, \quad i \in \{1, \cdots, m\},$$

which are obtained during the experimental phase. We define the measurement error $\epsilon \in \mathbb{R}^m$ introduced by the observations as the deviation of the model in the true but unknown parameter x_{true} and the observational data, i.e., the entry i of ϵ is defined by

$$\epsilon_i := \eta_i - h_i(x_{\text{true}}).$$

Let us assume that the Jacobian $J_h(x_{\text{true}})$ of $h(x)$ at x_{true} is full rank, that the model is structurally correct and the measurement error $\epsilon = (\epsilon_1, \cdots, \epsilon_m)$ is a random variable such as

- ϵ_i and ϵ_j are pairwise independent if $i \neq j$,
- ϵ_i is normally distributed for all j, with
- mean value zero, which means that the expectation value of the observational data is equal to the model responses.
- The variance-covariance matrix is known and equal to the diagonal matrix

$$\Sigma = \text{diag}(\sigma_1, \sigma_2, \cdots, \sigma_m).$$

Let $L(x)$ be the correspondent *Likelihood function* of the parameter x, i.e.,

$$L(x) := P(\epsilon|x) = \prod_{i=1}^{m} P_i(\epsilon_i),$$

where P_i is the probability density function of our normally distributed variable ϵ_i. It is well known (see Bard [7]) that the maximum likelihood function estimator can also be obtained solving the following nonlinear least squares problem:

$$\arg\min_{x\in D} \frac{1}{2}\|f(x)\|_2^2 \tag{1.1}$$

where $f(x) \in \mathbb{R}^m$, and its entry i is $f_i(x) = \frac{\eta_i - h_i(x)}{\sigma_i}$. Let us define the residual level function $T(x) = \frac{1}{2}\|f(x)\|_2^2$, and let us consider the following nonlinear equation

$$\nabla T(x) = 0.$$

We say that $x_* \in D$ is a stationary point of (1.1) if x_* is a root of the above equation.

The practical way to find a local solutions of (1.1) is given by the first-order necessary condition Theorem and second-order sufficient Theorem.

Theorem 1.1 (Fisrt-Order Necessary Condition). *If T is a continuously differentiable function, and $x_* \in D$ is a local minimizer of (1.1), then x_* is a stationary point of (1.1).*

Proof. Nocedal and Wright[63, Chapter 2].

■

Theorem 1.2 (Second-Order Necessary Condition). *If T is a twice continuously differentiable function, $x_* \in D$ is a local minimizer of (1.1), then x_* is a stationary point of (1.1) and $\nabla^2 T(x_*)$ is positive semidefinite.*

Proof. Nocedal and Wright[63, Chapter 2].

■

Let $J_f(x) \in \mathbb{R}^{m\times n}$ be the Jacobian matrix of f, i.e.,

$$J_f(x) = \begin{bmatrix} \nabla f_1(x)^T \\ \nabla f_2(x)^T \\ \vdots \\ \nabla f_m(x)^T \end{bmatrix}.$$

The first-order necessary condition Theorem establishes that in order to obtain a local solution of (1.1), we need to solve first of all the nonlinear equation,

$$F(x) := \nabla T(x) = \sum_{i=1}^{m} f_i(x)\nabla f_i(x) = J_f^T(x)f(x) = 0. \tag{1.2}$$

Nevertheless, not all the solutions of (1.2) are also local solution of (1.1). The following second-order sufficient condition Theorem gives us the sufficient conditions to know when a stationary point become a local minimizer.

Theorem 1.3 (Second-Order Sufficient Condition). *If T is a twice continuously differentiable function, $x_* \in D$ is a stationary point of (1.1), and the Hessian $\nabla^2 T(x_*)$ is positive definite, then we can guarantee that $x_* \in D$ is a strict local minimizer of (1.1).*

Proof. Nocedal and Wright[63, Chapter 2].

■

In our case, we can rewrite the Hessian as,

$$\begin{aligned}\nabla^2 T(x) &= \sum_{i=1}^m \nabla f_i(x)\nabla f_i(x)^T + \sum_{i=1}^m f_i(x)\nabla^2 f_i(x) \\ &= J_f(x)^T J_f(x) + \sum_{i=1}^m f_i(x)\nabla^2 f_i(x).\end{aligned}$$

Defining

$$Q_\epsilon(x) := \sum_{i=1}^m f_i(x)\nabla^2 f_i(x), \tag{1.3}$$

we obtain

$$J_F(x) := \nabla^2 T(x) = J_f(x)^T J_f(x) + Q_\epsilon(x). \tag{1.4}$$

Remark 1.4. *Let us consider a positive constant $\kappa < 1$. By definition, we know that*

$$f_i(x) = \frac{\eta_i - h_i(x)}{\sigma_i} \quad \textit{and} \quad \epsilon_i = \eta_i - h_i(x_{true}).$$

Thus, $J_f(x) = -[\Sigma]^{-1} J_h(x)$ and because $J_h(x_{true})$ is full rank, we obtain that for $\epsilon = 0$

$$Q_0(x_{true})[J_f^T(x_{true})J_f(x_{true})]^{-1} = 0,$$

and from the continuity properties, it follows that there is an $r_* > 0$ such that

$$V := \left\{ x \in D \;\middle|\; J(x) \textit{ is full rank and } \left\| Q_\epsilon(x)\left[J_f(x)^T J_f(x)\right]^{-1} \right\| \leq \kappa \right\}$$

is not empty for all $\epsilon \in B(0, r_*)$.

Let us assume that there is a stationary point $x_* \in V \subset D$ of (1.1). Using Theorem 1.3, it is possible to conclude that x_* is also a local minimizer of (1.1) if $\nabla^2 T(x_*)$ is positive definite. Note that the matrix $\nabla^2 T(x_*)$ is the sum of the positive definite matrix $J_f(x_*)^T J_f(x_*)$, and the symmetric matrix $Q_\epsilon(x_*)$. A natural question is: What property must $Q_\epsilon(x_*)$ satisfy in order to conclude that $\nabla^2 T(x_*)$ is positive definite?. The answer is given by the following Proposition, which requires a preparation Lemma given below.

Lemma 1.5. *Let M and N be two $n \times n$ symmetric matrices. If M and $MN + NM$ are positive definite, then N is also positive definite.*

Proof. Note that N is invertible. In fact, let us fix $x \in \text{Kern}\,(N)$, then $Nx = 0$. Because $MN + NM$ is positive definite, we obtain $x = 0$. Otherwise, we conclude that $x \neq 0$ and $x^T [MN + NM]\, x = 0$, which is not possible since $MN + NM$ is positive definite.

In the following lines, we prove that N is positive definite. Let $\lambda \in \mathbb{R} - \{0\}$ be a eigenvalue of N, then there is an unitary vector v such that $Nv = \lambda v$. From our hypothesis, it follows

$$0 < v^T \left[MN + NM\right] v = 2\lambda \left[v^T M v\right],$$

since M is positive definite, we obtain $0 < \lambda$, which said that all the eigenvalues of N are positive, i.e., N is positive definite.

■

Proposition 1.6. *Given a stationary point $x_* \in D$ of (1.1), let us assume that $J_f(x)$ is full rank in a neighborhood $V_* \subseteq D$ of x_*. If one of the following matrices*

$$\left[J_f(x)^T J_f(x)\right]^{-1} Q_\epsilon(x) = \left[J_f(x)^T J_f(x)\right]^{-1} \nabla^2 T(x) - I$$

or

$$Q_\epsilon(x) \left[J_f(x)^T J_f(x)\right]^{-1} = \nabla^2 T(x) \left[J_f(x)^T J_f(x)\right]^{-1} - I$$

has Euclidean norm or spectral radius less than one in V_*, then $\nabla^2 T(x)$ is positive definite for all $x \in V_*$.

Proof. From (1.4), we have

$$\nabla^2 T(x) = J_f(x)^T J_f(x) + Q_\epsilon(x),$$

therefore,

$$\nabla^2 T(x) \left[J_f(x)^T J_f(x)\right]^{-1} = I + Q_\epsilon(x) \left[J_f(x)^T J_f(x)\right]^{-1} \tag{1.5a}$$

$$\left[J_f(x)^T J_f(x)\right]^{-1} \nabla^2 T(x) = I + \left[J_f(x)^T J_f(x)\right]^{-1} Q_\epsilon(x). \tag{1.5b}$$

Let us define $M = \left[J_f(x)^T J_f(x)\right]^{-1}$ and $N = \nabla^2 T(x)$. Adding (1.5a) and (1.5b), we obtain,

$$MN + NM = 2I + \left[J_f(x)^T J_f(x)\right]^{-1} Q_\epsilon(x) + Q_\epsilon(x) \left[J_f(x)^T J_f(x)\right]^{-1}.$$

Statement: $MN + NM$ is a symmetric and positive definite matrix. $MN + NM$ is a symmetric matrix because M and N are symmetric matrices. Let us assume that

$$\rho\left(\left[J_f(x)^T J_f(x)\right]^{-1} Q_\epsilon(x)\right) = \rho\left(Q_\epsilon(x) \left[J_f(x)^T J_f(x)\right]^{-1}\right) < 1 \tag{1.6}$$

where $\rho(M)$ denotes the spectral radius of the matrix M. Let λ be an eigenvalue of $MN + NM$, then there is a nonzero vector u such that

$$2u + \left[J_f(x)^T J_f(x)\right]^{-1} Q_\epsilon(x) u + Q_\epsilon(x) \left[J_f(x)^T J_f(x)\right]^{-1} u = \lambda u,$$

or equivalently,

$$\left[\left[J_f(x)^T J_f(x)\right]^{-1} Q_\epsilon(x) + Q_\epsilon(x) \left[J_f(x)^T J_f(x)\right]^{-1}\right] u = (\lambda - 2) u.$$

From (1.6) it follows that $|(\lambda - 2)| < 2$, which implies that $0 < \lambda < 4$. We prove that all the eigenvalues of $MN + NM$ are positive, therefore $MN + NM$ is positive definite. Applying the Lemma 1.5, it follows that $N = \nabla^2 T(x)$ is a positive definite matrix for all $x \in V_*$.

■

Remark 1.7. *The Proposition 1.6 says that if x_* is a stationary point of* (1.1)*, $J_f(x_*)$ is full rank, and*

$$\left\| \left[J_f(x_*)^T J_f(x_*) \right]^{-1} Q_\epsilon(x_*) \right\| < 1, \text{ or } \left\| Q_\epsilon(x_*) \left[J_f(x_*)^T J_f(x_*) \right]^{-1} \right\| < 1,$$

then $\nabla^2 T(x_*) = J_F(x_*)$ is positive definite. Therefore, we conclude directly from Theorem 1.3 that x_* is also a strict local solution of (1.1)*.*

We finalize this section introducing a definition that classifies the convergence factor of a convergent sequence (y_k).

Definition 1.8 (Convergence factor)**.** *Let us consider a sequence (y_k) in $\mathbb{R}^n$ that converges to y_*. We said that,*

1. *(y_k) converges with superlinear convergence factor if*

$$\lim_{k\to\infty} \frac{\|y_{k+1} - y_*\|}{\|y_k - y_*\|} = 0,$$

2. *(y_k) converges with order $p \in [1, \infty)$ and with quotient convergence factor κ if*

$$\kappa := \limsup_{k\to+\infty} \frac{\|y_{k+1} - y_*\|}{\|y_k - y_*\|^p}$$

   exists. In particular, when $p = 1$ ($p = 2$) we say that the convergence is linear (quadratic).

3. *(y_k) converges with root convergence factor ρ if*

$$\rho = \limsup_{k\to+\infty} \|y_k - y_*\|^{\frac{1}{k}}$$

   exists.

4. *(y_k) converges with weak order $p \in [1, \infty)$ and root convergence factor ρ if*

$$\rho = \limsup_{k\to+\infty} \|y_k - y_*\|^{\frac{1}{p^k}}$$

   exists.

From this definition it follows that the convergence factor depends on the norm in *1.*, *2.*; and does not in *3.*.

1.2 Newton Method for Nonlinear Equations

In this section, we focus on numerically solving the nonlinear equation (1.2), i.e.,

$$F(x) = J_f^T(x) f(x) = 0,$$

with Jacobian

$$J_F(x) = \nabla^2 T(x) = J_f(x)^T J_f(x) \quad + \quad Q_\epsilon(x).$$

Since h is a twice continuously differentiable function then $J_F(x)$ is continuously differentiable function. Let us assume that $J_F(x)$ is invertible, the most relevant approach for numerically solving nonlinear equation problems is the Newton method. In our case, the Newton method solves (1.2) starting from an initial guess $x_0 \in D$, and consequently, form a linear model function $M_F(x)$ of $F(x)$ by taking the first two terms of its Taylor series approximation around the current iterate x_k. Using this linear model, we iteratively compute a sequence (x_k) according to

$$x_{k+1} = x_k + \Delta x_k, \text{ with } M_F(x_k + \Delta x_k) = 0.$$

In other words, this method approaches to the solution of (1.2) by solving a sequence of linear equation subproblems. We present its algorithm in Algorithm 1.1.

Algorithm 1.1 Newton's Algorithm for Nonlinear Equations

Step 0: Choose the initial guess x_0 close to a local solution x_* of (1.2).
Step 1: Repeat until convergence:
  Step 1.1: Solve $M_F(x_k + \Delta x_k) = 0$,
  i.e., $J_F(x_k)\Delta x_k = -J_f^T(x_k)f(x_k)$.
  Step 1.2: Set $x_{k+1} = x_k + \Delta x_k$.

The classical Theorems describing the convergence properties of the Newton sequence (x_k), as well as, the uniqueness of a local solution x_* of (1.2) are the *Newton-Kantorovich Theorem* [58] and the *Newton-Mysovskikh Theorem* [60], but the above Theorem is more attractive for the convergence analysis because it does not require the existence of such a local solution x_*.

Assuming that

(i) the Jacobian matrix $J_F(x)$ is a Lipschitz function in D, with Lipschitz constant γ_F,

(ii) there is a positive constant β_F, such that, $\| [J_F(x)]^{-1} \| \leq \beta_F$ for all $x \in D$, and

(iii) $x_0 \in B_{\gamma_F}$ where

$$B_{\gamma_F} := \{x \in D \mid \beta_F \gamma_F \|\Delta x_0\| < 2\},$$

Newton and Mysovskikh [60] proved that the Newton sequence (x_k) converges to a root $x_* \in D$ of (1.2) with quadratic convergence rate. Thus, B_{γ_F} define a neighborhood of $x_* \in D$ where quadratic convergence of the Newton method is guaranteed. Let us consider the following class of problems

$$AF(x) = 0 \text{ where } A \text{ is a invertible matrix.} \tag{1.7}$$

Therefore, giving an initial guess $x_0 \in D$, we obtain that the Newton sequences (x_k) of the class of problems (1.7) is calculated by,

$$x_{k+1} = x_k + \Delta x_k, \text{ with } AJ_F(x_k)\Delta x_k = -AF(x_k),$$

or $J_F(x_k)\Delta x_k = -F(x_k)$, which means that our Newton sequence (x_k) is invariant under transformations on the images space of $F(x)$, but $B_{\gamma_{AF}}$ is not invariant since the Lipschitz constant γ_{AF} of $AF(x)$ in (i) and β_{AF} in (ii) depend on A. We focus, at this point, on convergence Theorems of Newton method that provide a neighborhood N_* of $x_* \in D$ such that

- N_* is invariant under transformations on the images space of F, if our Newton sequence (x_k) is invariant under transformations on the images space of F, and
- if $x_0 \in N_*$, then the quadratic convergence of Newton sequence (x_k) with initial guess x_0 is guaranteed.

Thus, the Newton-Mysovskikh Theorem is not adecuate for our interest. Deuflhard [27] presented variations of Newton-Mysovskikh Theorem by restricting the convergence analysis of our Newton sequence (x_k) to *affine invariance* convergence Theorems.

1.3 Affine Invariance

In this section, we consider the problem

$$G(y) = AF(By) = 0, \quad x = By \tag{1.8}$$

with nonsingular matrices A and B, and we are interested in the study of affine invariance convergence properties of Newton's method. We observe that the Newton sequence (y_k) with initial guess $y_0 = B^{-1}x_0$ satisfies,

$$y_{k+1} := y_k + \Delta y_k, \quad \text{for all } k \in \mathbb{N},$$

where

$$J_G(y_k)\Delta y_k = -G(y_k) \text{ and } J_G(y_k) = AJ_F(By_k)B.$$

Note that the Newton sequence (x_k) of (1.2) and the above new Newton sequence (y_k) are related through

$$x_k = By_k, \quad \text{for all } k \in \mathbb{N}. \tag{1.9}$$

From here, it is the clear that the sequences (x_k) and (y_k) are invariant under transformations on the image spaces of F, an invariance property defined by Deuflhard [27] as *affine covariance*, and they are related through (1.9) under transformations on the domain D, an invariance property defined by Deuflhard [27] as *affine contravariance*.

In order to provide affine invariance convergence Theorems of Newton method, we need to guarantee that our convergence Theorems provide results that inherit such as affine invariance properties.

1.3.1 Affine Covariance

Here, we keep $B = I$ fixed in (1.8), i.e., we consider the class of problems,

$$G(y) = AF(x) = 0$$

generated by the class $\mathrm{GL}(n)$ of nonsingular matrices A. The above class of problems has the same roots and generate the same Newton sequences. The last ingredient necessary

for building an affine covariant Newton method theory is reduced to present a Theorem, which results are invariant under transformation on the images spaces. Deuflhard [27] presented a variant of the Newton-Mysovskikh Theorem, which is invariant under transformations on the images spaces.

Theorem 1.9 (Affine covariant Newton-Mysovskikh). *Let us assume that D is convex,*

(*i*) *There is some $\omega > 0$ such that $J_F(x)$ satisfies* ***the covariant Lipschitz condition***

$$\left\|J_F(z)^{-1}\left[J_F(x+t(y-x))-J_F(x)\right](y-x)\right\| \leq t\omega\|y-x\|^2 \text{ for all } x,y,z \in D,$$

and $t \in [0,1]$.

(*ii*) *Given some $\alpha < 2$, the initial guess x_0 satisfies*

$$x_0 \in V_N := \left\{ x \in D \,\middle|\, \omega \left\| [J_F(x)]^{-1} F(x) \right\| \leq \alpha < 2 \right\}, \text{ and}$$

$$\overline{B}(x_0,\rho) \subset D \text{ where } \rho := \frac{\|\Delta x_0\|}{1-[\omega\|\Delta x_0\|]/2}.$$

Then, the Newton sequence (x_k) stays in $B(x_0,\rho)$, and converges quadratically to a root $x_* \in D$ of (1.2) in the sense that

$$\|x_{k+1}-x_k\| \leq \frac{1}{2}\omega\|x_k-x_{k-1}\|^2.$$

Furthermore, the estimate x_k of x_* satisfies

$$\|x_k - x_*\| \leq \frac{\|x_k - x_{k-1}\|}{1-\frac{1}{2}\omega\|x_k-x_{k-1}\|}.$$

Proof. Deuflhard [27].

■

1.3.2 Affine Contravariance

This setting is dual to the preceding one. Here, we keep $A = I$ fixed in (1.8), i.e., we consider the class of problems,

$$G(y) = F(By) = 0, \text{ where } B \in \mathrm{GL}(n) \text{ and } x = By.$$

In this case, we have that x_* is a root of (1.2), iff y_* is a root of $G(y) = F(By) = 0$ where $x_* = By_*$. The residual Newton sequence $(F(x_k))$ and the residual Newton sequence $(F(By_k))$ generated by the above class of problems coincide. In order to present a contravariant affine Newton method theory, we need to provide a residual convergence Theorem, which is invariant under transformations on the domain spaces. Deuflhard [27] presents a variant of the Newton-Mysovskikh Theorem, which is invariant under transformations on the domain spaces.

Theorem 1.10 (Affine contravariant Newton-Mysovskikh)**.** *Let us assume that D is convex. If there is a constant $\omega > 0$ such that $J_F(x)$ satisfies the* ***contravariant Lipschitz condition***

$$\| \left[J_F(y) - J_F(x) \right] (y - x) \| \leq \omega \| J_F(x)(y - x) \|^2,$$

for all $x, y \in D$; and the initial guess $x_0 \in L_\omega$ where

$$L_\omega := \left\{ x \in D \mid \|F(x)\| < \frac{2}{\omega} \right\} \text{ with } \overline{L}_\omega \subset D,$$

then the Newton sequence (x_k) stays in L_ω, and the residual sequence $(F(x_k))$ converges quadratically to zero.

Proof. Deuflhard [27].

■

Remark 1.11. *In the above Theorem, we can show that there is at least a subsequence (x_{k_n}) of (x_k) that converge to $x_* \in \overline{L}_\omega$ such that $F(x_*) = 0$. Unfortunately, neither can we ensure that (x_k) converges, nor can we determine the convergence rate of such an subsequence.*

1.4 Gauss-Newton Method

The Gauss-Newton Method is the most popular approach for numerically solving nonlinear least squares problems, in our case the problem (1.1). GN is a variant of the Newton method in which we are not taking into account the second order derivative information. The process of generating a Gauss-Newton sequence starts with an initial guess $x_0 \in D$, and consequently form a linear model function $M_f(x)$ of $f(x)$ by taking the first two terms of its Taylor approximation around of the current iterate x_k. Using this linear model, we construct iteratively a sequence (x_k) according to,

$$x_{k+1} = x_k + \Delta x_k,$$

where

$$\Delta x_k := \arg\min \frac{1}{2} \| M_f(x_k + \Delta x) \|_2^2 \quad \text{with } M_f(x_k + \Delta x_k) := J_f(x_k)\Delta x_k + f(x_k). \tag{1.10}$$

If $J_f(x_k)$ is full rank for all k, then the Gauss-Newton step Δx_k is given by the Moore-Penrose pseudoinverse of the function $f(x)$ at x_k, i.e.,

$$\Delta x_k = - \left[J_f^T(x_k) J_f(x_k) \right]^{-1} J_f^T(x_k) f(x_k).$$

Thus, the GN method approaches to a local solution x_* of (1.1) by solving a sequence of linear least squares problems. The following Theorem provides sufficient conditions that guaranty convergence of the GN method.

Theorem 1.12 (Spectral Radius)**.** *Let us assume that for our function f defined in* (1.1) *the following are valid,*

(O_1) *$x_* \in D$ is a stationary point of $T(x) = \frac{1}{2} \| f(x) \|_2^2$.*

(O_2) *The Jacobian $J_f(x)$ of f has full rank in a neighborhood of x_*.*

(O_3) *The spectral radius (SR) of the matrix $\left[J_f^T(x_*)J_f(x_*)\right]^{-1} Q_\epsilon(x_*)$ is less than one, i.e.,*

$$\rho_* = \rho\left(\left[J_f^T(x_*)J_f(x_*)\right]^{-1} Q_\epsilon(x_*)\right) < 1 \tag{1.11}$$

where $J_F(x) = \nabla^2 T(x) = J_f^T(x)J_f(x) + Q_\epsilon(x)$ with $Q_\epsilon(x)$ defined in (1.3).

Then, there is a particular neighborhood $V_{SR} \subset D$ of x_* such that for all $x_0 \in V_{SR}$ the correspondent Gauss-Newton sequence (x_k) converges to a local solution x_* of (1.2) with root coefficient factor ρ_*.

Proof. Orthega and Rheinboldt [64, Section 10.2].

■

Indeed, the SR Theorem tells more than the GN sequence converges to a stationary point x_* of (1.1). It conclude also that $\nabla^2 T(x_*)$ is positive define, which implies that x_* is also a local solution of (1.1).

Corollary 1.13. *Let (O_1), (O_2) and (O_3) in Theorem 1.12 hold. Then*

$$\nabla^2 T(x_*) = \left[J_f^T(x_*)J_f(x_*)\right] + Q_\epsilon(x_*)$$

is positive definite. Furthermore, there is a particular neighborhood $V_{SR} \subset D$ of x_* such that for all $x_0 \in V_{SR}$ the correspondent Gauss-Newton sequence (x_k) converges to a local solution x_* of (1.1) with root coefficient factor ρ_*..

Proof. Proposition 1.6 and Theorem 1.12.

■

Remark 1.14. *Giving a positive constant $\kappa < 1$, let us assume that there is a stationary point $x_* \in D$ of (1.1) such that $J_f(x_*)$ is full rank and*

$$\left\| Q_\epsilon(x_*)\left[J_f^T(x_*)J_f(x_*)\right]^{-1} \right\| < \kappa.$$

Let us define

$$V_\kappa := \left\{ x \in D \,\middle|\, J_f(x) \text{ is full rank and } \left\| Q_\epsilon(x)\left[J_f^T(x)J_f(x)\right]^{-1} \right\| \leq \kappa < 1 \right\}, \tag{1.12}$$

from remark (1.4), it follows that V_κ and its interior are nonempty set. Thus, starting from $x_0 \in V_\kappa$ such that $\|F(x_0)\|$ is sufficient close to zero, we obtain that the GN method generates sequences (x_k) that are well defined. In the following Theorem we explain with more detail how to choose such an initial guess x_0.

Theorem 1.15. *Let us assume that*

- *D is convex.*
- *(O_1) and (O_2) are valid.*

- *Given a positive constant $\kappa < 1$, $x_* \in D$ satisfies*

$$\left\| Q_\epsilon(x_*) \left[J_f^T(x_*) J_f(x_*) \right]^{-1} \right\| < \kappa < 1.$$

- *The closure of V_κ defined in* (1.12) *is compact and subset of D.*

If we define

$$\mathcal{L} := \left\{ x \in V_\kappa \mid \omega \| \left[J_f^T(x) J_f(x) \right]^{-1} F(x) \| \leq 2(1-\kappa) \right\} \tag{1.13}$$

where ω is the Lipschitz constant of $J_F(x)$ in $\overline{V_\kappa}$, then the Gauss-Newton sequence (x_k) starting from $x_0 \in \mathcal{L}$ converges linearly to x_* with convergence factor less than κ.

Proof. We omit the proof of this Theorem because it is a particular case of Theorem 3.2.

■

Remark 1.16. *In the above Theorem, there is a point that is not clear: Is Theorem 1.15 a covariant or a contravariant Theorem?. The set V_κ is a contravariant set but $\mathcal{L}$ is neither covariant nor contravariant. We answer this question in Chapter 3 in where we prove that there is a norm $\|y\|_*$ and a neighborhood V_δ of x_* such that*

$$V_\delta := \left\{ x \in D \,\middle|\, J_f(x) \text{ is full rank and } \left\| \left[J_f^T(x) J_f(x) \right]^{-1} Q_\epsilon(x) \right\|_* \leq \kappa + \delta < 1 \right\} \subset V_{\kappa_{GN}}$$

where V_δ is covariant with $\|y\|_*$-norm and the above Theorem is also valid for all

$$x_0 \in \mathcal{L}_\delta := \left\{ x \in V_\delta \,\middle|\, \omega \left\| \left[J_f^T(x) J_f(x) \right]^{-1} F(x) \right\|_* \leq 2(1-\kappa) \right\} \subset \mathcal{L}.$$

Thus, we can also say that the Theorem 1.15 is covariant with respect to $\|y\|_*$-norm.

Relation between the Gauss-Newton and the Newton Method

Unlike the Gauss-Newton method, the Newton method attacks the solution of (1.2) by solving a sequence of linear equation problems

$$M_F(x_k^N + \Delta x) = 0 \text{ where } M_F(x_k^N + \Delta x) := \left[J_f^T(x_k^N) J_f(x_k^N) + Q_\epsilon(x_k^N) \right] \Delta x + J_f^T(x_k^N) f(x_k^N),$$

i.e., $M_F(x_k^N + \Delta x)$ is an linear approximation of $F(x) = J_f^T(x) f(x)$ at the iterate x_k^N. Let us define a new linear approximation $M_f(x_k^{GN} + \delta x)$ of $F(x) = J_f^T(x) f(x)$ around x_k^{GN} by dropping the term $Q_\epsilon(x_k^{GN})$ in $M_F(x_k^{GN} + \delta x)$, i.e.,

$$M_f(x_k^{GN} + \delta x) := J_f^T(x_k^{GN}) J_f(x_k^{GN}) \delta x + J_f^T(x_k^{GN}) f(x_k^{GN}). \tag{1.14}$$

Therefore, we determine a new sequence (x_k^{GN}), which is defined by $x_{k+1}^{GN} = x_k^{GN} + \delta x_k^{GN}$ with $M_f(x_k^{GN} + \delta x_k^{GN}) = 0$ that satisfies the following properties,

(G1) The sequence (x_k^{GN}) is the Gauss-Newton sequence that approaches locally to a local solution x_* of (1.1) with root coefficient factor ρ_* if the hypotheses of Corollary 1.13 are valid.

(G2) The Newton step at the iterate x_k^{GN} satisfies

$$\left[J_f^T(x_k^{GN})J_f(x_k^{GN}) + Q_\epsilon(x_k^{GN})\right]\Delta x_k^N = -J_f^T(x_k^{GN})f(x_k^{GN})$$

while the Gauss-Newton step

$$J_f^T(x_k^{GN})J_f(x_k^{GN})\Delta x_k^{GN} = -J_f^T(x_k^{GN})f(x_k^{GN}).$$

(G3) The covariant discrepancy between Newton method and GN method is measured through the **covariant relative error** at the iterate x_k^{GN}

$$\frac{\|\Delta x_k^N - \Delta x_k^{GN}\|}{\|\Delta x_k^N\|}, \tag{1.15}$$

and the **covariant error matrix** with $\|y\|$-norm

$$\left\|I - \left[J_f^T(x)J_f(x)\right]^{-1}\nabla^2 T(x)\right\| = \left\|\left[J_f^T(x)J_f(x)\right]^{-1}Q_\epsilon(x)\right\| \text{ for all } x_k \in V_\kappa \tag{1.16}$$

where V_κ is defined in (1.12). Furthermore, the covariant relative error and the covariant matrix error are related through

$$\frac{\|\Delta x_k^N - \Delta x_k^{GN}\|}{\|\Delta x_k^N\|} \leq \left\|\left[J_f^T(x_k^{GN})J_f(x_k^{GN})\right]^{-1}Q_\epsilon(x_k^{GN})\right\| \text{ for all } x \in V_\kappa.$$

(G4) The contravariant discrepancy between Newton method and GN method is measured through the **contravariant error matrix** with $\|y\|$-norm

$$\left\|I - \nabla^2 T(x)\left[J_f^T(x)J_f(x)\right]^{-1}\right\| = \left\|Q_\epsilon(x)\left[J_f^T(x)J_f(x)\right]^{-1}\right\| \text{ for all } x \in V_\kappa \tag{1.17}$$

where V_κ is defined in (1.12).

(G5) The covariant Newton-Mysovskikh Theorem (1.9) ensures locally that the Newton sequence (x_k^N) converges to x_* with quadratic convergence factor, the spectral radius Corollary 1.13 ensures locally that the GN sequence (x_k^{GN}) converges to x_* with root convergence factor ρ_*, and Theorem 1.15 ensures locally that the sequence (x_k^{GN}) converges linearly to x_*. Thus, the spectral radius Corollary delivers locally faster convergence GN sequence than Theorem 1.15 since the spectral radius of the error matrix that measure the discrepancy between Newton method and GN method does not depend on the norm neither the case (covariant or contravariant), i.e.,

$$\rho_* = \rho\left(\left[J_f^T(x_*)J_f(x_*)\right]^{-1}Q_\epsilon(x_*)\right) = \rho\left(Q_\epsilon(x_*)\left[J_f^T(x_*)J_f(x_*)\right]^{-1}\right).$$

1.5 Inexact Gauss-Newton Method

Often, we work with large-scale nonlinear least squares problems that arise from problems like Bundle Adjustment [56], or inverse problems such as infinite dimensional parameter identification problems of Partial Differential Equations models [84] where the Gauss-Newton step cannot be computed directly, but has to be approximated. The inexact Gauss-Newton method approaches to a local solution of (1.1) by solving iteratively a

sequence of linear least square problems. We restrict our study to the use of iterative technique based on projection processes onto orthogonal Krylov subspaces.

The whole process starts with an initial guess $x_0 \in D$, and consequently, we define a sequence (x_k) by

$$x_{k+1} = x_k + \delta x_k$$

where δx_k solves approximately the linear least squares problem (1.10) using a certain orthogonal Krylov subspace method with a certain termination rule (or stopping criterion). In comparison with the exact Gauss-Newton step Δx_k^{GN} at the iterate x_k, the inexact Gauss-Newton (IGN) method introduce at every iterate x_k the following error

$$\delta x_k - \Delta x_k^{GN}.$$

On the other hand, defining

$$r(x_k) := J_f(x_k)\delta x_k + f(x_k),$$

the inner residual error introduced by our inexact Gauss-Newton step δx_k is given by

$$J_f^T(x_k)r(x_k) = J_f^T(x_k)J_f(x_k)\delta x_k + J_f^T(x_k)f(x_k).$$

Thus, we are able to measure the discrepancy between the inexact and the exact Gauss-Newton method through:

(1) The **covariant inner residual relative error**

$$\frac{\|J^+(x_k)r(x_k)\|}{\|J^+(x_k)f(x_k)\|} = \frac{\|\delta x_k - \Delta x_k^{GN}\|}{\|\Delta x_k^{GN}\|} \tag{1.18}$$

provided that $J_f(x_k)$ is full rank and $F(x_k) \neq 0$.

(2) The **contravarinat inner residual relative error**

$$\frac{\|J^T(x_k)r(x_k)\|}{\|J^T(x_k)f(x_k)\|} = \frac{\left\|\left[J_f^T(x_k)J_f(x_k)\right]\left[\delta x_k - \Delta x_k^{GN}\right]\right\|}{\left\|\left[J_f^T(x_k)J_f(x_k)\right]\Delta x_k^{GN}\right\|} \tag{1.19}$$

provided that $J_f(x_k)$ is full rank and $J^T(x_k)f(x_k) \neq 0$.

(3) If there is a function $M : V \subseteq D \to \mathrm{GL}(n)$ such that the IGN step can be written as $\delta x_k = -M(x_k)J_f^T(x_k)f(x_k)$ for all k and $J_f(x)$ is full rank in V, the **contravariant inverse error matrix** with $\|y\|$-norm is defined as

$$\left\|I - [M(x)]^{-1}\left[J_f^T(x)J_f(x)\right]^{-1}\right\| \text{ for all } x \in V. \tag{1.20}$$

(4) The **contravariant error matrix** with $\|y\|$-norm is defined as

$$\left\|I - \left[J_f^T(x)J_f(x)\right]M(x)\right\| \text{ for all } x \in V \tag{1.21}$$

where $M(x)$ is introduced in (3).

We introduce in Chapter 2 a new early inner termination criterion that only depends on cheaply available information for our IGN method, which implies that the errors (1.18), (1.19), (1.20), and (1.21) are bounded. Furthermore, our IGN approach delivers linearly and locally convergent IGN sequence that converges to a local statistically stable solution of (1.1) provided that at least one exists. We postpone the proof of all result to Chapter 2 and 3.

1.6 Newton-Type Method

We start the Newton-type method with our initial guess $x_0 \in D$ and propose a better approximation to a local solution x_* of (1.2) through

$$x_{k+1} := x_k + \delta x_k \text{ where } \delta x_k := -M(x_k)J_f^T(x_k)f(x_k) \tag{1.22}$$

where $M(x)$ could be interpreted as an approximation of the Jacobian inverse $[J_F(x)]^{-1}$ of $F(x)$, i.e.,

$$[J_F(x)]^{-1} = \left[\nabla^2 T(x)\right]^{-1} = \left[J_f^T(x)J_f(x) + Q_\epsilon(x)\right]^{-1} \text{ and } Q_\epsilon(x) \text{ defined in (1.3).}$$

If $M(x)=\left[\nabla^2 T(x)\right]^{-1}$, then the sequence (x_k) generated by (1.22) is the Newton sequence. If $M(x)=\left[J_f^T(x)J_f(x)\right]^{-1}$, then (x_k) is the Gauss-Newton sequence. If $M(x)$ is an approximation of $\left[J_f^T(x)J_f(x)\right]^{-1}$, then (x_k) is the inexact Gauss-Newton sequence. From this argument we conclude that the Newton method, the Gauss-Newton method and the inexact Gauss-Newton method are particular cases of Newton-type method.

Given a nonnegative sequence (η_k), *the inexact Newton method* is another particular case of Newton-type method, which defines an inexact Newton sequence starting from $x_0 \in D$ as follows:

$$x_{k+1} = x_k + \delta x_k \tag{1.23}$$

where the inexact Newton step δx_k solves approximately the following Newton equation

$$\nabla^2 T(x_k)\Delta x = -J_f^T(x_k)f(x_k),$$

and defining the inner residual error as $R_k := \nabla^2 T(x_k)\delta x_k + J_f^T(x_k)f(x_k)$, the inexact Newton satisfies δx_k satisfies

$$\|R_k\| \leq \eta_k \|J_f^T(x_k)f(x_k)\|.$$

Remark 1.17. *There is a matrix $M(x) \approx [J_F(x)]^{-1}$ (see for example [69, Lemma 5.1]) such that the inexact Newton step satisfies $\delta x_k = -M(x_k)J_f^T(x_k)f(x_k)$ for all k.*

An important question is: what level of accuracy is required in our inexact Newton step δx_k to preserve the rapid local convergence of Newton method?. Dembo, Stanley, Eisenstat, and Steihaug [22] answered the above question in the following Theorem

Theorem 1.18. *Let us assume*

$(D1)$ *There is an $x_* \in D$ such that $F(x_*) = 0$.*

$(D2)$ *$F : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ is differentiable in a neighborhood of x_*.*

$(D3)$ *$J_F(x_*)$ is nonsingular.*

if (η_k) is a positive sequence such that $\eta_k \leq \eta_{\max} < t < 1$, then there is an neighborhood $V_{\mathcal{D}}$ of x_* such that for all $x_0 \in V_{\mathcal{D}}$ the correspondent inexact Newton sequence (x_k) defined in (1.23) converges to x_*. Moreover, the convergence is linear in the sense

$$\|x_{k+1} - x_*\|_* \leq t\|x_k - x_*\|_*$$

where $\|y\|_* = \|J_F(x_*)y\|_*$.

Proof. Dembo, Stanley, Eisenstat, and Steihaug [22].

■

Remark 1.19. *Dembo, Stanley, Eisenstat, and Steihaug [22] not only analyzed the local behavior of the inexact Newton sequence (x_k) (1.23), but also characterized the order of convergence of (x_k) and indicated how to choose a forcing sequence (η_k) such that (x_k) preserves the rapid convergence of Newton method. In other words, in [22] there is a recipe of how to construct inexact Newton convergence sequences. In the following, we summarize the most important results of such a paper related to the order of convergence of (x_k). Let us assume that an inexact Newton sequence (x_k) satisfies Theorem 1.18 then (x_k) converges locally to x_*. Furthermore, if x_0 is sufficiently close to x_* then*

(i) (x_k) converges with superlinear factor of convergence if

$$\lim_{k\to\infty} \eta_k = 0.$$

(ii) (x_k) converges with strong order at least $1 + p$ if $J_F(x)$ is Hölder continuous with exponent p at x_*, and $\|R_k\| = O(\|F(x_k)\|^{1+p})$ when k goes to infinity.

(iii) (x_k) converges with weak order at least $1 + p$ if $J_F(x)$ is Hölder continuous with exponent p at x_*, and (R_k) converge to zero with weak order at least $1 + p$.

In this thesis, we do not take into account (i), (ii), (iii) because those conditions say that the inner residual sequence (R_k) converges to zero too quickly, but, we are interested in an early stopping criterion, which means that we need to keep the residual $\|R_k\|$ as large as possible at every iteration.

Remark 1.20. *Let us consider $\kappa < 1$ and an inexact Newton sequence $(x_k) \subset D$ such that*

$$\|R_k\| \leq \kappa\|F(x_k)\| \text{ and } J_F(x_k) \text{ nonsingular for all } k.$$

Then, we cannot ensure using Theorem 1.18 that (x_k) converges since this Theorem is local, i.e., x_0 must be sufficiently close to a root x_* of (1.2).

Given positive constants $\kappa_* < 1$, $\kappa < 1$, and $x_0 \in D$, let us consider the following IGN sequence (x_k^{IGN}) such that

$$x_{k+1}^{IGN} = x_k^{IGN} + \delta x_k^{IGN} \text{ where } \delta x_k^{IGN}$$

solves approximately

$$J_f^T(x_k^{IGN})J_f(x_k^{IGN})\Delta x = -J_f^T(x_k^{IGN})f(x_k^{IGN}), \tag{1.24}$$

using a certain iterative technique based on projection process onto Krylov subspaces with stopping criterion

$$\|R_k^{IGN}\| \leq \kappa_* \|J_f^T(x_k) f(x_k)\| \text{ where } R_k^{IGN} := \nabla^2 T(x_k^{IGN}) \delta x_k^{IGN} + J_f^T(x_k^{IGN}) f(x_k^{IGN}) \tag{1.25}$$

then such as stopping criterion is unavaible because the calculation of the residual R_k^{GN} requires knowledge of $Q_\epsilon(x)$, which is unavailabe for IGN methods. Instead of calculating R_k^{IGN} other authors as Lourakis, Manolis and Argyros [56] propose to calculate an approximation r_k^{IGN} of R_k^{IGN} that does take into account the term $Q_\epsilon(x)$ in $\nabla^2 T(x) = J_f^T(x) J_f(x) + Q_\epsilon(x)$, i.e.,

$$r_k^{IGN} = J_f^T(x_k^{IGN}) J_f(x_k^{IGN}) \delta x_k^{IGN} + J_f^T(x_k^{IGN}) f(x_k^{IGN}),$$

and introduce the following stopping criterion

$$\|r_k^{IGN}\| \leq \kappa \|J_f^T(x_k^{IGN}) f(x_k^{IGN})\|. \tag{1.26}$$

Natural questions are:

(Q1) Using the stopping criterion (1.26), does (x_k^{IGN}) converge?,

(Q2) What is the convergence factor of such a sequence?, in the case that it converges.

In this thesis, we propose a new IGN method where our IGN step δx_k^{IGN} is computed using LSQR [65] or LSMR [33] as iterative linear algebra method for appproximately solving the inner linearized least squares subproblem (1.24) with a new early inner termination criterion that only depends on cheaply available information, which implies that stopping criterion (1.25) and stopping criterion (1.26) are valid, and ensures linear and local convergence of (x_k^{IGN}) to a local solution x_* of (1.1). Furthermore, we propose a new damped IGN method based on our new local IGN approach and in the backward step control theory presented by Potschka [70]. We finalize this Chapter introducing definitions that allow to measure the discrepancy between the IGN method (with sequence (x_k)) and the Newton method.

(1) The **contravarinat inner residual relative error**

$$\frac{\|R(x_k)\|}{\|F(x_k)\|} = \frac{\left\|[J_F(x_k)]\left[\delta x_k - \Delta x_k^N\right]\right\|}{\left\|[J_F(x_k)]\,\Delta x_k^N\right\|} \tag{1.27}$$

where $J_F(x)$ is positive definite, $R(x_k) = J_F(x_k)\delta x_k + F(x_k)$ and $F(x_k) \neq 0$.

(2) If there is a function $M : V \subseteq D \to \mathrm{GL}(n)$ such that the IGN step can be written as $\delta x_k = -M(x_k) J_f^T(x_k) f(x_k)$ for all k; and $J_F(x)$ is nonsingular in V, the **covariant error matrix** with $\|y\|$-norm is defined as

$$\|I - M(x) J_F(x)\| \text{ for all } x \in V. \tag{1.28}$$

(3) The **contravariant error matrix** with $\|y\|$-norm is defined as

$$\|I - J_F(x) M(x)\| \text{ for all } x \in V \tag{1.29}$$

where $M(x)$ is defined in (2) and $J_F(x)$ is nonsingular in V.

(4) The **covarinat inner residual relative error**

$$\frac{\|M(x_{k+1})R(x_k)\|}{\|M(x_k)F(x_k)\|} \tag{1.30}$$

where $R(x_k)$ is defined in (1), $M(x)$ is defined in (2) and $F(x_k) \neq 0$.

Chapter 2

Iterative Linear Algebra for Parameter Estimation

In this Chapter, we are concerned with the development of efficient numerical methods for large-scale nonlinear parameter estimation problems. The state of the art for such parameter estimation problems is the Gauss-Newton method. We propouse a new inexact Gauss-Newton method for numericaly solving such large-scale nonlinear parameter estimation problems in where the inexact Gauss-Newton step is computed using LSQR [65] or LSMR [33] as iterative linear algebra method for appproximately solving the inner linearized least squares subproblems with a new early inner termination criterion that only depends on cheaply available information. The idea of such an inner termination criterion is based on the contravariant κ-theory result introduced by Dembo, Stanley, Eisenstat, and Steihaug [22] and local κ-theory introduced by Potschka [70]. Our new approach results to be an inexact Gauss-Newton method that guarantees statistically stable solutions provided that at least one exists.

This chapter is organized as follows, we introduce the most relevant Krylov subspace numerical solvers for linear systems of equations. Later, we present LSQR and LSMR Krylov space numerical methods for approximately solving linear least squares problems, which standard termination rule is based on backward error minimization properties [17]. It turns out that such a standard termination rule is too conservative for our inexact Gauss-Newton method since in this setting it is not necessary to have high precision when we compute our inexact Gauss-Newton step. Instead, it is fundamental control how large the inner residual generated in our inner linearized subproblem must be in order to ensure convergence of our IGN sequences (x_k). In the last section of this chapter, we introduce our new inner termination criterion and we study its principal properties.

2.1 Krylov Space Methods for Linear Systems

Let B be a nonsingular symmetric $n \times n$ matrix, $b \in \mathbb{R}^n$ a vector, and consider the following linear system of equations

$$B\Delta x = b. \tag{2.1}$$

The aims of this section are to survey efficient numerical methods for solving the above linear system of equation. When the matrix B is so large-scale where matrix-vector prod-

ucts can be evaluated efficiently, direct factorization methods such as Gaussian elimination or Cholesky decomposition are computationally expensive, and therefore numerical methods that depend on this kind of factorization are not a valid choice. Instead, we are interested in iterative methods for solving (2.1) based on projection process, both orthogonal and oblique, onto Krylov subspaces.

Definition 2.1. *We define the ith Krylov subspace of* (B, b)*, which is denoted by* $K_i(B, b)$*, as*

$$K_i(B, b) := span\left\{b, \quad Bb, \quad B^2b, \quad \cdots \quad , B^{i-1}b\right\} \tag{2.2}$$

where i *is a positive integer.*

Definition 2.2. *Let* L_i *be a subspace of* $\mathbb{R}^n$ *with the same dimension as* $K_i(B, b)$*. Starting with the initial guess* $x_0 = 0 \in \mathbb{R}^n$*, a Krylov method is an iterative method, which finds at every iterate* i *a better approximation* $\delta x_i \in K_i(B, b)$ *of* Δx *by imposing the Petrov-Galerkin condition,*

$$b - B\delta x_i \perp L_i$$

□

The different versions of Krylov subspace methods arise from different choices of the subspace L_i. We restrict our study to $L_i := K_i(B, b)$, and the popular minimum residual method given by $L_i := BK_i(B, b)$.

Lemma 2.3. *Let* B *be a positive definite matrix. If* $L_i := K_i(B, b)$ *then the following problems have the same solution,*

- *Find* $\delta x_i \in K_i(B, b)$*, such that* $b - B\delta x_i \perp L_i$*.*
- $\delta x_i = \underset{\delta x \in K_i(B,b)}{\arg\min} \|\Delta x - \delta x\|_B$ *where* $\|y\|_B := \left[y^T B y\right]^{1/2}$*.*

Proof: Saad [73, Proposition 5.2].

□

Lemma 2.4. *Let us assume that* B *is nonsingular. If* $L_i := BK_i(B, b)$ *then the following problems have the same solution,*

- *Find* $\delta x_i \in K_i(B, b)$*, such that* $b - B\delta x_i \perp L_i$*.*
- $\delta x_i = \underset{\delta x \in K_i(B,b)}{\arg\min} \|b - B\delta x\|_2$*.*

Proof: Saad [73, Proposition 5.3].

□

The previous Lemmas provide another way to classify our Krylov subspace methods based on the error and residual properties.

- **Minimum Error Krylov Method**

$$\delta x_i := \underset{\delta x \in K_i(B,b)}{\arg\min} \|\Delta x - \delta x\|_B$$

- **Minimum Residual Krylov Method**

$$\delta x_i := \underset{\delta x \in K_i(B,b)}{\arg\min} \|b - B\delta x\|_2$$

Remark 2.5. *Let $\{v_1, \cdots, v_i\}$ be a basis of $K_i(B,b)$ with $i \leq n$, and let us define its corresponding matrix as $V_i = [v_1 \mid v_2 \mid \cdots \mid v_i]$. If we want to solve approximately* (2.1) *using Minimum Error Krylov method then there is a $y_i \in \mathbb{R}^i$ such that $\delta x_i = V_i y_i$ and*

$$b - BV_i y_i \perp V_i$$

i.e.,

$$V_i^T [b - BV_i y_i] = 0,$$

which means that y_i is the solution of the following system

$$V_i^T B V_i y_i = V_i^T b. \tag{2.3}$$

When B is symmetric and positive definite, the spectral decomposition theorem allows to write the matrix B as follows,

$$B = Q_n \Lambda Q_n^T,$$

where the diagonal matrix $\Lambda = diag(\lambda_1, \cdots, \lambda_n)$ contains the eigenvalues of B, and $Q_n = [q_1 \mid \cdots \mid q_n]$ is a orthogonal matrix. If we assume without loss of generality that $\{q_1, \cdots, q_i\}$ is a basis of $K_i(B,b)$ then solve (2.3) *is equivalent to solve*

$$Q_i^T B Q_i y_i = Q_i^T b$$

where $Q_i^T B Q_i = \Lambda$ is diagonal, it means that the solution of the above system is calculated with just n steps. On the other hand, constructing a matrix Q_i is computationally expensive. In the following section we are interested in building a computationally cheaper basis V_i of our Krylov space such that the system (2.3) *can be solved efficiently.*

2.1.1 Lanczos Process

When B is a symmetric and positive definite matrix, the Lanczos process provides a recursive formula for building an orthogonal basis $\{v_1, \cdots, v_i\}$ of $K_i(B,b)$ such that $V_i^T B V_i$ is tridiagonal. From the above remark it is clear that this method approximates the solution of (2.1) through $\delta x_i = V_i y_i$ where y_i solves the tridiagonal system of $i \leq n$ equations (2.3). Formally, we present its recursive algorithm in Algorithm 2.1.

Properties:

(LP1) $Bv_i = \beta_i v_{i-1} + \alpha_i v_i + \beta_{i+1} v_{i+1}$, for all $i < n$.

Algorithm 2.1 Lanczos Process

1: $v_0 := 0$, $\beta_1 := b^T b$ and $\beta_1 v_1 := b$
2: **for** $i = 1, \cdots, n-1$ **do**
3: **if** $\beta_i \neq 0$ **then**
4: $p_i := Av_i$, and $\alpha_i := v_i^T p_i$
5: $\beta_{i+1} v_{i+1} := p_i - \alpha_i v_i - \beta_i v_{i-1}$, where β_{i+1} serves to normalize v_{i+1}
6: **end if**
7: **end for**

(LP2) $BV_i = [v_1 \mid v_2 \mid \cdots \mid v_{i-1} \mid v_i] \underbrace{\begin{bmatrix} \alpha_1 & \beta_2 & & & \\ \beta_2 & \alpha_2 & \beta_3 & & \\ & \beta_3 & \alpha_3 & \ddots & \\ & & \ddots & \ddots & \beta_i \\ & & & \beta_i & \alpha_i \end{bmatrix}_{i\times i}}_{T_i} + \beta_{i+1} v_{i+1} e_i^T$, which implies, $V_i^T B V_i = T_i$.

(LP3) $BV_i = V_{i+1} \underbrace{\begin{bmatrix} T_i \\ \hline \beta_{i+1} e_i^T \end{bmatrix}_{(i+1)\times i}}_{\underline{T}_i}$ i.e., $BV_i = V_{i+1} \underline{T}_i$.

2.1.2 Minimum Error Krylov Method

From (LP1) we obtain $V_i^T b = \beta_1 e_1$ where $e_1 \in \mathbb{R}^i$ is the canonical vector with first entry one and the rest zero. Using (LP2), the system given by (2.3), i.e.,

$$T_i y_i = \beta_1 e_1$$

is tridiagonal. If we apply the Cholesky factorization to the matrix T_i then our iterative method for solving (2.1) is knowing by **LanczosCG** method.

2.1.3 Minimum Residual Krylov Method

In this method, the orthogonal basis $\{v_1, \cdots, v_i\}$ of $K_i(B, b)$ is given by the Lanczos process and our approximation solution is given by $\delta x_i = V_i y_i$ where y_i solves the following linear least squares problem

$$y_i := \arg\min_{y \in \mathbb{R}^i} \|b - BV_i y\|_2.$$

Using the property (LP3) and that V_{i+1} is an orthogonal matrix, we have that

$$y_i := \arg\min_{y \in \mathbb{R}^i} \|\beta_1 e_1 - \underline{T}_i y\|_2$$

where $e_1 \in \mathbb{R}^{i+1}$ is the canonical vector with its first entry one and the rest zero. If we apply the a QR factorization to the matrix $\underline{T}_i$ then we obtain the popular Krylov space method **MINRES** to approximately solve (2.1).

	LanczosCG	MINRES
Subproblem	$T_i y_i = \beta_1 e_1$	$y_i = \arg\min_{y \in \mathbb{R}^i} \|\beta_1 e_1 - \underline{T}_i y\|$
Factorization	$T_i = L_i D_i L_i^T$	$Q_i \underline{T}_i = \begin{bmatrix} R_i \\ 0 \end{bmatrix}$
Estimate	$\delta x_i = V_i y_i \in K_i(B, b)$	$\delta x_i = V_i y_i \in K_i(B, b)$
New basis of the Krylov Spaces	$W_i = V_i L_i^{-T}$	$D_i = V_i R_i^{-1}$
New subproblem	$L_i D_i z_i = \beta_i e_1$	$R_i z_i = \beta_1 \begin{bmatrix} I_i & 0 \end{bmatrix} Q_i e_1$
New estimate δx_k	$\delta x_i = W_i z_i$	$\delta x_i = D_i z_i$
ith residual	$\min \|b - B y_i\|_{B^{-1}}$	$\min \|b - B y_i\|_2$
Orthogonality	$r_i \perp K_i(B, b)$	$B^T r_i \perp K_i(B, b)$
ith error	$\min \|\hat{x} - \delta x_i\|_B$	unknown

Table 2.1: LanczosCG vs MINRES. We resume the principal properties present in both methods and visualize its differences where V_i denote the basis generated by The Lanczos Process, $V_i^T B V_i = T_i$, and $B V_i = V_{i+1} \underline{T}_i$.

We finalize this section with a resumme of the principal properties present in the MINRES and LanczosCG methods, which is presented in Table 2.1. For a deeper study of the differents iterative methods for solving linear equation based on Krylov space, and its comparative tables, we sugesst Choi's dissertation thesis [20].

2.2 Krylov Space Methods for Solving Least-Squares Problems

Let $J \in \mathbb{R}^{m \times n}$ be a matrix with $n \leq m$, $f \in \mathbb{R}^n$ a non zero vector, and let us assume that J has full rank. We consider the following linear least squares (LS) problem,

$$\Delta x := \arg\min_{x \in \mathbb{R}^n} \|Jx + f\|_2 \tag{2.4}$$

In this section, we are interested in the study of computational efficient iterative methods based on orthogonal Krylov spaces for solving (2.4). Solving the above linear LS problem is equivalent to solving the following linear system of equations

$$J^T J \Delta x = -J^T f \tag{2.5}$$

The general approach is reduced to finding a basis V_i of $K_i(J^T J, J^T f)$, and an approximation $\delta x_i = V_i y_i$ of the solution Δx of (2.4) such that

$$J^T \left[J V_i y_i + f \right] \perp V_i.$$

In other words, y_i is the solution of the following linear system of equations,

$$V_i^T \left[J^T J \right] V_i y_i = -V_i^T J^T f. \tag{2.6}$$

A variant of the Lanczos process with matrix J^TJ and starting vector $-J^Tf$ provides a way to compute such as basis V_i.

Golub-Kahan bidiagonalization Process

The Golub-Kahan algorithm [37] builds a pair of unitary matrices $U_{i+1} = [u_1 \mid \cdots \mid u_{i+1}]$ and $V_i = [v_1 \mid \cdots \mid v_i]$ such that

$$U_{i+1}^T J V_i$$

is a bidiagonal matrix and $V_i^T \left[J^TJ\right] V_i$ is a tridiagonal matrix. In the following we introduce its algorithm and later its properties,

Algorithm 2.2 Golub-Kahan bidiagonalization Process

1: $\beta_1 u_1 = -f$ (shorthand $\beta_1 = \|f\|_2 \neq 0$ and u_1 unitary)
2: $\alpha_1 v_1 = J^T u_1$
3: **for** $i = 1, \cdots$ **do**
4: $\quad \beta_{i+1} u_{i+1} = J v_i - \alpha_i u_i$
5: $\quad \alpha_{i+1} v_{i+1} = J^T u_{i+1} - \beta_{i+1} v_i$
6: **end for**

Properties:

(GK1) The scalars α_i and β_{i+1} are positive.

(GK2) $JV_i = U_{i+1}B_i$, where

$$B_i = \begin{bmatrix} \alpha_1 & & & \\ \beta_2 & \alpha_2 & & \\ & \beta_3 & \ddots & \\ & & \ddots & \alpha_i \\ & & & \beta_{i+1} \end{bmatrix}_{(i+1)\times i}$$

(GK3) $J^TU_{i+1} = V_iB_i^T + \alpha_{i+1}v_{i+1}e_{i+1}^T$, which is equivalent to

$$J^TU_{i+1} = V_{i+1}\begin{bmatrix} B_i^T \\ \alpha_{i+1}e_{i+1}^T \end{bmatrix} = V_{i+1}\underbrace{\begin{bmatrix} B_i & \alpha_{i+1}e_{i+1} \end{bmatrix}^T}_{L_{i+1}^T}.$$

Using (GK2) and (GK3), we conclude

$$J^TJV_i = J^TU_{i+1}B_i = V_{i+1}\begin{bmatrix} B_i^T \\ \alpha_{i+1}e_{i+1}^T \end{bmatrix} B_i = V_{i+1}\begin{bmatrix} B_i^TB_i \\ \alpha_{i+1}\beta_{i+1}e_{i+1}^T \end{bmatrix}.$$

From the above equality, it is clear that the Golub-Kahan bidiagonalization process is equivalent to what would be generated by the Lanczos process applied to the matrix J^TJ and starting vector $-J^Tf$, and also that our matrix $V_i^T \left[J^TJ\right] V_i$ is tridiagonal.

In order to introduce the algorithms that solve approximately our linear least squares problem (2.4), we need to explain how y_i must be calculated. We have two options that generate two completely different algorithms. First of all, y_i solves (2.6) if and only if

$$\textbf{LSQR: } y_i := \arg\min_{y\in\mathbb{R}^i} \|JV_i y + f\|_2.$$

and the other option is given by the following result: If

$$\textbf{LSMR: } y_i := \arg\min_{y\in\mathbb{R}^i} \|J^T JV_i y + J^T f\|_2.$$

then y_i solve (2.6).

2.2.1 LSQR: Sparse Linear Least Squares Iterative Algorithm Based on QR-factorization.

LSQR was introduced by C. Paige and M. Saunders, [65], and is a particular orthogonal Krylov Space method for numerically solving our linear least squares problem (2.4). The approximate solution is given by $\delta x_i = V_i y_i$, where

(i) V_i is a basis of our Krylov space $K_i := K_i(J^T J, J^T f)$ generated by the Golub-Kahan bidiagonalization process, and

(ii) $y_i := \arg\min_{y\in\mathbb{R}^i}\|JV_i y + f\| = \arg\min_{y\in\mathbb{R}^i}\|U_{i+1}\underbrace{(\beta_1 e_1 + B_i y)}_{t_{i+1}}\| = \arg\min_{y\in\mathbb{R}^i}\|t_{k+1}\|.$

LSQR is similar in style to the LanczosCG method applied to the normal equation (2.5), and the inner LSQR residuals $r_i := J\delta x_i + f$ are monotonically decreasing.

In the following, we build a new basis $\mathcal{D}_i = \{d_1, d_2, \cdots, d_i\}$ of our Krylov space K_i, which is easy to compute, and allows to write δx_i as a linear combinations of δx_{i-1} and d_i. Applying a QR factorization to the matrix B_i,

$$Q_i\,[B_i \;\; \beta_1 e_1] = \begin{bmatrix} R_i & f_i \\ 0 & \overline{\phi}_{i+1} \end{bmatrix} = \left[\begin{array}{ccccc|c} \rho_1 & \theta_2 & & & & \phi_1 \\ & \rho_2 & \theta_3 & & & \phi_2 \\ & & \ddots & \ddots & & \\ & & & \rho_{i-1} & \theta_i & \phi_{i-1} \\ & & & & \rho_i & \phi_i \\ \hline 0 & 0 & 0 & 0 & 0 & \overline{\phi}_{i+1} \end{array}\right]. \tag{2.7}$$

Then the vector y_i is given by,

$$R_i y_i = -f_i \text{ and } t_{i+1} = Q_i^T \begin{bmatrix} 0 \\ \overline{\phi}_{i+1} \end{bmatrix}.$$

Thereby,

$$\delta x_i = V_i y_i = V_i [R_i]^{-1} f_i = D_i f_i,$$

where the columns of $D_i := [d_1 \mid \; d_2 \mid \; \cdots d_i]$ can be iteratively calculated from the system $R_i^T D_i^T = V_i^T$ using forward substitution. In other words, starting from $d_0 = \delta x_0 = 0$, we can calculate the columns d_i of D_i using the following formula,

$$d_i = \frac{1}{\rho_i}(v_i - \theta_i d_{i-1})$$

therefore,

$$\delta x_i = V_i y_i = V_{i-1} y_{i-1} + \phi_i d_i = \delta x_{i-1} + \phi_i d_i.$$

2.2.2 LSMR: Sparse Linear Least Squares Iterative Algorithm Based on Double QR-factorization.

This method was introduced by Fong and Saunders [33], and is a particular orthogonal Krylov Space method for numerically solving our linear least squares problem (2.4). The approximate solution is given by $\delta x_i = V_i y_i$ where

(i) V_i is a basis of our Krylov space $K_i := K_i(J^T J, J^T f)$ generated by the Golub-Kahan bidiagonalization process, and

(ii) $y_i := \arg\min_{y\in\mathbb{R}^i} \|J^T J V_i y + J^T f\| = \arg\min_{y\in\mathbb{R}^i} \left\| V_{i+1} \left[\overline{\beta}_1 e_1 + \begin{bmatrix} B_i^T B_i \\ \overline{\beta}_{i+1} e_i^T \end{bmatrix} y \right] \right\|$, where $\overline{\beta}_i = \alpha_i \beta_i$.

LSMR method is equivalent to MINRES method applied to the normal equation (2.5), and the inner LSMR residuals $\|J^T r_i\|_2$ are monotonically decreasing where $r_i := J\delta x_i + f$. Furthermore, the stopping criterion using in LSQR and LSMR is obtained from a Backward error analysis, but such a stopping criterion is satisfied earlier in LSMR than in LSQR. We go into more detail about such a stopping criterion in Section 2.2.4.

In the following we build a new basis $\overline{\mathcal{W}}_i = \{\overline{w}_1, \overline{w}_2, \cdots, \overline{w}_i\}$ of our Krylov space K_i, which is easy to compute, and allows to write our estimate δx_i as a linear combinations of δx_{i-1} and $\overline{w}_i$. Applying a QR factorization to the matrix B_i,

$$Q_{i+1} B_i = \begin{bmatrix} R_i \\ 0 \end{bmatrix}, \text{ where } R_i := \begin{bmatrix} \rho_1 & \theta_2 & & & \\ & \rho_2 & \theta_3 & & \\ & & \ddots & \ddots & \\ & & & \rho_{i-1} & \theta_i \\ & & & & \rho_i \end{bmatrix} \tag{2.8}$$

therefore,

$$B_i^T B_i = \begin{bmatrix} R_i^T & 0 \end{bmatrix} Q_{i+1} Q_{i+1}^T \begin{bmatrix} R_i \\ 0 \end{bmatrix} = R_i^T R_i.$$

If we define $q_i \in \mathbb{R}^i$ such that $R_i^T q_i = \overline{\beta}_{i+1} e_i$, then $q_i = \frac{\overline{\beta}_{i+1}}{\rho_i} e_i = \varphi_i e_i$, and consequently, we obtain

$$\begin{aligned} y_i &= \arg\min_{y\in\mathbb{R}^i} \left\| V_{i+1} \left[\overline{\beta}_1 e_1 + \begin{bmatrix} B_i^T B_i \\ \overline{\beta}_{i+1} e_i^T \end{bmatrix} y \right] \right\| \\ &= \arg\min_{y\in\mathbb{R}^i} \left\| V_{i+1} \left[\overline{\beta}_1 e_1 + \begin{bmatrix} R_i^T R_i \\ q_i^T R_i \end{bmatrix} y \right] \right\| \\ &= \arg\min_{y\in\mathbb{R}^i} \left\| V_{i+1} \left[\overline{\beta}_1 e_1 + \begin{bmatrix} R_i^T \\ \varphi_i e_i^T \end{bmatrix} R_i y \right] \right\|. \end{aligned}$$

Introducing the variable change $t = R_i y$, we simplify the above problem to,

$$t_i := R_i y_i = \arg\min_{t\in\mathbb{R}^i} \left\| V_{i+1} \left[\overline{\beta}_1 e_1 + \begin{bmatrix} R_i^T \\ \varphi_i e_i^T \end{bmatrix} t \right] \right\|.$$

Performing a second QR factorization

$$\overline{Q}_{i+1}\begin{bmatrix} R_i^T & \overline{\beta}_1 e_1 \\ \varphi_i e_i^T & 0 \end{bmatrix} = \begin{bmatrix} \overline{R}_i & z_i \\ 0 & \overline{\xi}_{i+1} \end{bmatrix}, \text{ where } \overline{R}_i := \begin{bmatrix} \overline{\rho}_1 & \overline{\theta}_2 & & \\ & \overline{\rho}_2 & \ddots & \\ & & \ddots & \\ & & & \overline{\theta}_i \\ & & & \overline{\rho}_i \end{bmatrix}, \tag{2.9}$$

we obtain

$$y_i := R_i^{-1} \arg\min_{t\in\mathbb{R}^i} \left\| V_{i+1}\left[\overline{\beta}_1 e_1 + \begin{bmatrix} R_i^T \\ \varphi_i e_i^T \end{bmatrix} t\right]\right\| = R_i^{-1} \arg\min_{t\in\mathbb{R}^i} \left\| \begin{bmatrix} z_i \\ \overline{\xi}_{i+1} \end{bmatrix} + \begin{bmatrix} \overline{R}_i \\ 0 \end{bmatrix} t\right\|$$

and this subproblem is solved choosing t_i such that,

$$\overline{R}_i t_i = -z_i \text{ where } z_i := (\xi_1, \cdots, \xi_i)^T.$$

Let W_i and $\overline{W}_i$ be computed by forward substitution from

$$R_i^T W_i^T = V_i^T \text{ and } \overline{R}_i^T \overline{W}_i^T = W_i^T. \tag{2.10}$$

Then from $\delta x_i = V_i y_i$, $R_i y_i = t_i$, and $\overline{R}_i t_i = -z_i$, and $\delta x_0 = 0$, we have

$$\delta x_i = W_i R_i y_i = W_i t_i = \overline{W}_i \overline{R}_i t_i = -\overline{W}_i z_i = \delta x_{i-1} - \xi_i \overline{w}_i.$$

Note that we can compute the elements of our basis $\overline{\mathcal{W}}_i$ efficiently, because from (2.10) and (2.8), it follows that

$$w_i = \frac{1}{\rho_i}(v_i - \theta_i w_{i-1}) \text{ where } w_0 = \delta x_0 = 0,$$

and from (2.10) and (2.9), it follows

$$\overline{w}_i = \frac{1}{\overline{\rho}_i}(w_i - \overline{\theta}_i \overline{w}_{i-1}) \text{ where } \overline{w}_0 = 0.$$

We finalize this section with the Table 2.2 that compares the principal properties present in the LSQR and LSMR methods.

2.2.3 Krylov Solvers Based on Backward Error Minimization Properties

In this Section, we are interested in measuring the backward error introduced in the left-hand side of (2.5), i.e., $J^T J\Delta x = -J^T f$ when it is solved using LSQR and LSMR. Let δx_i be its approximate solution, $r_i := J\delta x_i + f$, then the contravariant inner residual $J^T r_i$ of δx_i satisfies

$$J^T J\delta x_i = -J^T f + J^T r_i.$$

In the following Lemma, we prove that there is a matrix E such that

$$\overline{J}^T \overline{J}\delta x_i = -J^T f \text{ where } \overline{J} := J + E.$$

and also that there is a matrix $\tilde{E}$, such that,

$$\left[J^T J - \tilde{E}\right]\delta x_i = -J^T f.$$

Those elemental results set up the basis of our consecutive Sections.

	LSQR	LSMR
Subproblem	$y_i := \arg\min_{y\in\mathbb{R}^i} \|JV_i y + f\|$	$y_i = \arg\min_{y\in\mathbb{R}^i} \|J^T J V_i y + J^T f\|$
Factorization	$Q_i \begin{bmatrix} B_i & \beta_1 e_1 \end{bmatrix} = \begin{bmatrix} R_i & f_i \\ 0 & \overline{\phi}_{i+1} \end{bmatrix}$	$Q_{i+1} B_i = \begin{bmatrix} R_i \\ 0 \end{bmatrix}$ $\overline{Q}_{i+1} \begin{bmatrix} R_i^T & \overline{\beta}_1 e_1 \\ \varphi_i e_i^T & 0 \end{bmatrix} = \begin{bmatrix} \overline{R}_i & z_i \\ 0 & \xi_{i+1} \end{bmatrix}$
Estimate	$\delta x_i = V_i y_i \in K_i(J^T J, J^T f)$	$\delta x_i = V_i y_i \in K_i(J^T J, J^T f)$
New subproblem	$\arg\min_{y\in\mathbb{R}^i} \|\beta_1 e_1 + B_i y\|$	$\arg\min_{y\in\mathbb{R}^i} \left\| \overline{\beta}_1 e_1 + \begin{bmatrix} B_i^T B_i \\ \overline{\beta}_{i+1} e_i^T \end{bmatrix} y \right\|$
New basis of the Krylov Spaces	$D_i = V_i R_i^{-1}$	$\overline{W}_i = V_i R_i^{-1} \overline{R}_i^{-1}$
Estimate in function of the new basis	$\delta x_i = D_i f_i$	$\delta x_i = -\overline{W}_i z_i$
Orthogonality	$J^T r_i \perp K_i(J^T J, J^T f)$	$J^T r_i \perp K_i(J^T J, J^T f)$

Table 2.2: LSQR vs LSMR. We resume the principal properties present in both methods and visualize its differences where V_i is a basis of our Krylov space $K_i := K_i(J^T J, J^T f)$ generated by the Golub-Kahan bidiagonalization process.

Lemma 2.6 (Backward error)**.** *Let $J \in \mathbb{R}^{m\times n}$ be a full rank matrix with $n \leq m$, and assume that $J^T f \neq 0$.*

(BE1) If $\delta x_i \neq 0$ is an approximate solution of (2.5) calculated via LSQR or LSMR with $r_i = J\delta x_i + f$, then

$$\overline{J}^T \overline{J} \delta x_i = -J^T f,$$

where

$$\overline{J} := J + E, \text{ and } E = -\frac{\left[J\delta x_i\right]\left[J^T r_i\right]^T}{\|J\delta x_i\|^2}.$$

(BE2) If $\tilde{E} = \dfrac{\left[J^T r_i\right]\left[J^T J\delta x_i\right]^T}{\|J\delta x_i\|^2}$, δx_i is the exact solution of the following problem,

$$\left[J^T J - \tilde{E}\right] \delta x_i = -J^T f.$$

(BE3) The matrix $A = [J^T J] - \tilde{E}$ is invertible, its inverse is

$$M := A^{-1} = \left[J^T J\right]^{-1} \left[I + \frac{\left[J^T r_i\right] \delta x_i^T}{\|J\delta x_i\|^2}\right],$$

and $\delta x_i = -M J^T f$.

Proof.
$(BE1)$ Since $\left[J^T r_i\right]^T \delta x_i = 0$, then

$$\begin{aligned}
\overline{J}^T \overline{J} \delta x_i = & \left[J - \frac{\left[J\delta x_i\right]\left[J^T r_i\right]^T}{\|J\delta x_i\|^2}\right]^T \left[J - \frac{\left[J\delta x_i\right]\left[J^T r_i\right]^T}{\|J\delta x_i\|^2}\right] \delta x_i \\
= & \left[J^T - \frac{\left[J^T r_i\right]\left[J\delta x_i\right]^T}{\|J\delta x_i\|^2}\right] J\delta x_i \\
= & J^T J\delta x_i - J^T r_i \\
= & -J^T f.
\end{aligned}$$

$(BE2)$ From the $(BE1)$, it follows that

$$\overline{J}^T \overline{J} \delta x_i = \left[J^T J - \frac{\left[J^T r_i\right]\left[J^T J\delta x_i\right]^T}{\|J\delta x_i\|^2}\right] \delta x_i = -J^T f.$$

$(BE3)$ Let us define the following matrices

$$A = \left[J^T J - \frac{\left[J^T r_i\right]\left[J^T J\delta x_i\right]^T}{\|J\delta x_i\|^2}\right] \text{ and } P := \left[I - \frac{\delta x_i \left[J^T r_i\right]^T}{\|J\delta x_i\|^2}\right],$$

then $A^T = \left[J^T J\right] P$ and let us prove that P is an invertible matrix. Given $u \in \mathrm{Ker}\,(P)$, we have by definition that

$$P(u) = u - \frac{\delta x_i \left[J^T r_i\right]^T u}{\|J\delta x_i\|^2} = 0,$$

i.e.,

$$u = \frac{\delta x_i \left[J^T r_i\right]^T u}{\|J\delta x_i\|^2}. \tag{2.11}$$

Multiplying (2.11) by $J^T r_i$, it follows that

$$\left[J^T r_i\right]^T u = \frac{\left[J^T r_i\right]^T \delta x_i \left[J^T r_i\right]^T u}{\|J\delta x_i\|^2} \overset{\delta x_i \perp J^T r_i}{=} 0,$$

i.e., $\left[J^T r_i\right]^T u = 0$, and substituting the above equality in (2.11), we conclude that $u = 0$. Therefore $\mathrm{Ker}\,(P) = \{0\}$, which implies that P is invertible. Consequently, A^T and A are invertible matrices. Because A is invertible, we can apply the Sherman-Morrison-Woodbury formula, which yields

$$M := A^{-1} = \left[J^T J\right]^{-1} + \frac{\left[J^T J\right]^{-1}\left[J^T r_i\right]\left[J^T J\delta x_i\right]^T\left[J^T J\right]^{-1}}{\|J\delta x_i\|^2 - \left[J^T J\delta x_i\right]^T\left[J^T J\right]^{-1}\left[J^T r_i\right]},$$

where

$$\|J\delta x_i\|^2 - \left[J^T J\delta x_i\right]^T \left[J^T J\right]^{-1} \left[J^T r_i\right] \overset{\delta x_i \perp J^T r_i}{=} \|J\delta x_i\|^2 = \delta x_i^T \left[J^T J\right] \delta x_i > 0,$$

it means

$$M = \left[J^T J\right]^{-1} \left[I + \frac{\left[J^T r_i\right] \delta x_i^T}{\|J\delta x_i\|^2}\right],$$

which complete the proof.

■

2.2.4 Error Estimate

Let δx_i be an approximate solution of (2.4) which is computed using LSQR or LSMR. A variant of the *acceptable least squares solution* definition introduced by Chang, Paige and Titley-Peloquin [17, Section 2] says that δx_i is a *v-acceptable least squares solution* if it is the exact solution of a linear least squares problem within an accepted range of relative errors in the data. In other words, δx_i is acceptable if $\xi(\delta x_i, r_i) < 1$ where

$$\xi(\delta x_i, r_i) := \min_E \left\{ \frac{\|E\|_2}{\|J^T J\|_2} : \left[J + E\right]^T \left[J + E\right] \delta x_i = -\left[J + E\right]^T f \text{ and } r_i = J\delta x_i + f \right\}.$$

It is well known from Golub and Van Loan [38, Section 5.3.7] that the relative error

$$\frac{\|\Delta x - \delta x_i\|}{\|\Delta x\|}$$

is bounded by a constant that is directly proportional to $\xi(\delta x_i, r_i)$. Because, finding an analytical expression of $\xi(\delta x_i, r_i)$ remains an open question, it was laid down stopping criteria from easily computable upper bound on $\xi(\delta x_i, r_i)$ see for example [33, 65, 17, 46]. The standard stopping criterion used in LSQR and LSQR is derived from the backward error introduced by Stewart [78], i.e.,

$$E_s = \frac{r_i [J^T r_i]^T}{\|r_i\|^2},$$

which satisfies

$$\left[J + E_s\right]^T \left[J + E_s\right] \delta x_i = -\left[J + E_s\right]^T f = -J^T f + J^T r_i \text{ and } \|E_s\|_2 = \frac{\|J^T r_i\|}{\|r_i\|}.$$

Therefore,

$$\|J^T J\|_2 \xi(\delta x_i, r_i) \leq \frac{\|J^T r_i\|}{\|r_i\|}.$$

Thus, from a given $ATOL > 0$, the standard stopping criterion used in LSQR and LSMR is provided the approximate solution δx_i of Δx if

$$\textbf{Standard Stopping Criterion}: \|J^T r_i\| \leq ATOL \|J^T J\|_2 \|r_i\| \tag{2.12}$$

where $\|J^T J\|_2$ can be estimated.

2.3 Inexact Gauss-Newton Method Based on LSQR and LSMR

If we use the inexact Gauss-Newton method for numerically solving the problem (1.1) where δx_k is calculated iteratively using LSQR or LSMR, we note that the stopping criterion (2.12) is too conservative since in this setting it is not necessary to have a high precision when we compute our inexact step δx_i, but it is rather fundamental to control the size of the contravarinat inner residual $J_f^T(x_k)r_k$ because in the Gauss-Newton method $J_f^T(x_k^{GN})r_k^{GN} = 0$ where $r_k^{GN} = J_f(x_k^{GN})\Delta x_k^{GN} + f(x_k^{GN})$.

Using the Lemma 2.6 part $(BE3)$, it follows that there is a matrix $M(x_k)$ such that $\delta x_k = -M(x_k)J_f^T(x_k)f(x_k)$, therefore

$$\begin{aligned} J_f^T(x_k)r_k &= J_f^T(x_k)\left[f(x_k) - J_f(x_k)M(x_k)J_f^T(x_k)\right] \\ &= \left[I - J_f^T(x_k)J_f(x_k)M(x_k)\right]J_f^T(x_k)f(x_k), \end{aligned}$$

thereby

$$\|J_f^T(x_k)r_k\| \leq \|I - [J_f^T(x_k)J_f(x_k)]M(x_k)\|\|J_f^T(x_k)f(x_k)\|.$$

If we assume that for all iterates x_k

$$\|I - [J_f^T(x_k)J_f(x_k)]M(x_k)\| \leq \alpha < 1,$$

then a new possible stopping criterion may be

$$\textbf{Stopping Criterion I: } \|J_f^T(x_k)r_k\| \leq \alpha\|J_f^T(x_k)f(x_k)\|. \tag{2.13}$$

A natural question is: Does the iterate $x_{k+1} = x_k + \delta x_k$ where δx_k is computed via LSQR or LSMR with stopping criterion (2.13) converge to a local solution of (1.1)?, If the answer is positive, what is the factor of convergence of such a sequence?. A partial answer is given by Gratton, Lawless and Nichols in [40], but in this case, it is assumed that

- There is a positive constant β such that

$$\|Q_\epsilon(x_k)[J_f^T(x_k)J_f(x_k)]^{-1}\|_2 \leq \beta < 1,$$

- the forcing sequence (β_k) satisfies

$$0 < \beta_k \leq \frac{\beta - \|Q_\epsilon(x_k)[J_f^T(x_k)J_f(x_k)]^{-1}\|}{1 + \|Q_\epsilon(x_k)[J_f^T(x_k)J_f(x_k)]^{-1}\|},$$

- and the stopping criterion is given by

$$\|J_f^T(x_k)r_k\| \leq \beta_k\|J_f^T(x_k)f(x_k)\|.$$

Therefore, it is proved there that the sequence (x_k) converges locally and linearly to x_*. Nevertheless, the forcing sequence (β_k) must be very small specially if we have that

$$\beta - \|Q_\epsilon(x_k)[J_f^T(x_k)J_f(x_k)]^{-1}\| \approx 0,$$

which means that the above stopping criterion is not satisfied early for iterates that are not close to solution, since in this cases β_k must be fast zero and thus it is required a large number of inner iteration in order to satisfy such a stooping criterion. But intuitively, we expect to implement less effort in obtained our IGN step when we are in an iterate far away from the solution. We conjecture that from the stopping criterion (2.13) with α very close to one it is not possible to prove convergence of our sequence (x_k). The problem is that in such a case the size of the inner residual $\|J_f^T(x_k)r_k\|$ may be too large that an extra condition that restricts the behavior of $\|J_f^T(x_k)r_k\|$ must be taken into account. In the Gauss-Newton method, we know that $J_f^T(x_k^{GN})r_k^{GN} = 0$ for all k, but it does mean that just bounding the contravariant inner relative residual using (2.13), we obtain convergence when α is close to one.

On the other hand, using the Theorem 1.18 given by Dembo, Stanley, Eisenstat, and Steihaug [22], we are able to characterize linear and local convergence of inexact Gauss-Newton sequence (x_k) through the following condition

$$\|\nabla^2 T(x_k)\delta x_k + \nabla T(x_k)\| \leq \kappa \|\nabla T(x_k)\| \text{ with } T(x) = \frac{1}{2}\|f(x)\|_2^2 \text{ and } \kappa \in (0,1).$$

But this stopping criterion is not available for IGN method since here

$$\nabla^2 T(x_k) = [J_f^T(x_k)J_f(x_k)] + Q_\epsilon(x_k)$$

where $Q_\epsilon(x_k)$ is unavailable for IGN method. In the following, we lay down sufficient conditions that make possible to conclude the above inequality provided that the Gauss-Newton method generates locally and linearly convergent sequence, which limit point x_* is a satistically stable solution of (1.1). We go into more detail about staistically stable properties in Chapter 4.

Hypotheses

($S1$) Let D be convex, open and nonempty set. Giving a positive constant $\kappa_{GN} < 1$, let us assume that there is a stationary point $x_* \in D$ of (1.1) such that $J_f(x_*)$ is full rank and

$$\left\| Q_\epsilon(x_*) \left[J_f^T(x_*)J_f(x_*)\right]^{-1} \right\| < \kappa_{GN}.$$

($S2$) The closure of $V_{\kappa_{GN}}$ is compact and contained in D where $V_{\kappa_{GN}}$ is defined in (1.12), i.e.,

$$V_{\kappa_{GN}} := \left\{ x \in D \,\middle|\, J_f(x) \text{ is full rank and } \left\| Q_\epsilon(x) \left[J_f^T(x)J_f(x)\right]^{-1} \right\| \leq \kappa_{GN} < 1 \right\}.$$

($S3$) Let us choose $\kappa \in (\kappa_{GN}, 1)$. We generate our Inexact Gauss-Newton sequence (x_k) according to $x_{k+1} = x_k + \delta x_k$ where δx_k solves via LSQR or LSMR the following linear problem

$$J_f^T(x_k)J_f(x_k)\Delta x = -J_f^T(x_k)f(x_k) \tag{2.14}$$

with **stopping criterion** given by

$$\|J_f^T(x_k)r_k\| \leq \kappa \|J_f^T(x_k)f(x_k)\| - \kappa_{GN}\|[J_f^T(x_k)J_f(x_k)]\delta x_k\| \tag{2.15}$$

where $r_k = J_f(x_k)\delta x_k + f(x_k)$.

Theorem 2.7. *Let* $(S1)$, $(S2)$, *and* $(S3)$ *hold. Then,* (x_k) *converges locally and linearly to* x_*. *Moreover, the convergence is linear in the sense that*

$$\|x_{k+1} - x_*\|_* \leq t\|x_k - x_*\|_*$$

where $\|y\|_* = \|J_F(x_*)y\|_*$ *and* $t \in (\kappa, 1)$.

Proof. Note that from stopping criterion (2.15) it follows that

$$\begin{aligned}\|J_F(x_k)\delta x_k + F(x_k)\| \leq & \|J_f^T(x_k)r_k\| + \|Q_\epsilon(x)[J_f^T(x_k)J_f(x_k)]^{-1}[J_f^T(x_k)J_f(x_k)]\delta x_k\| \\ \leq & \|J_f^T(x_k)r_k\| + \kappa_{GN}\|[J_f^T(x_k)J_f(x_k)]\delta x_k\| \\ \leq & \kappa\|J_f^T(x_k)f(x_k)\|\end{aligned}$$

The rest of the proof follows from Theorem 1.18.

■

Computational availability:

- The first question that arise is: Is the right side of stopping criterion (2.15) positive?. We prove in Lemma 2.9 that it is positive for all $\kappa \in (\kappa_{GN}, 1)$, and if $\kappa = \kappa_{GN}$, δx_k is the Gauss-Newton step.
- We assume in this thesis that $[J_f^T(x_k)J_f(x_k)]$ is large, sparse, and the matrix vector product $[J_f^T(x_k)J_f(x_k)]\delta x_k$ can be evaluated efficiently. Therefore, the stopping criterion (2.15) is an inner termination rule that only depends on available information if κ_{GN} is known. Thereby, our IGN sequences (x_k) converges linearly and locally to an estimate x_* of x_{true} if we choose an initial guess x_0 sufficiently close to x_*.
- A natural question is: How must we choose an initial guess in order to obtain linear convergence of our IGN sequence using such an IGN approach. Theorem 2.7 does not provide explicitly a set of initial guesses where the linear convergence is guaranteed. Nevertheless, in Chapter 3, we provide explicitly such a set using the hybrid Theorem 3.2 for Newton-type method introduced by Potschka [70]. Since all those results are locally valid and we cannot have at the beginning a good initial guess, we globalize this local IGN approach in Chapter 6.

We finalize this Section introducing some results that describe the principal properties derived from our new stopping criterion (2.15). The first one says that stopping criterion (2.15) implies stopping criterion (2.13), and the second property says that the right side of our stopping criterion (2.15) is not negative. The last properties are focus on measuring the discrepancy between the GN and IGN method trough the contravariant error matrix (1.21), and the contravariant inverse error matrix (1.20).

Lemma 2.8. *Let* $(S1)$, $(S2)$, *and* $(S3)$ *hold. Then*

$$\|J_f^T(x_k)r_k\| \leq (\kappa - \kappa_{GN})\|J_f^T(x_k)f(x_k)\|.$$

Proof. Because $J_f^T(x_k)r_k \perp K(J_f^T(x_k)J_f(x_k), J_f^T(x_k)f(x_k))$, we obtain that $J_f^T(x_k)r_k$ and $J_f^T(x_k)f(x_k)$ are orthogonal vectors. using

$$J_f^T(x_k)J_f(x_k)\delta x_k = -J_f^T(x_k)f(x_k) + J_f^T(x_k)r_k,$$

and Pythagorean theorem, it follows that

$$\|J_f^T(x_k)J_f(x_k)\delta x_k\|^2 = \|J_f^T(x_k)f(x_k)\|^2 + \|J_f^T(x_k)r_k\|^2,$$

i.e., $\|J_f^T(x_k)f(x_k)\| \leq \|J_f^T(x_k)J_f(x_k)\delta x_k\|$, which implies

$$\begin{aligned}\|J_f^T(x_k)r_k\| + \kappa_{GN}\|J_f^T(x_k)f(x_k)\| &\leq \|J_f^T(x_k)r_k\| + \kappa_{GN}\|J_f^T(x_k)J_f(x_k)\delta x_k\| \\ &\overset{(2.15)}{\leq} \kappa\|J_f^T(x_k)f(x_k)\|.\end{aligned}$$

■

Lemma 2.9. *Let* $(S1)$, $(S2)$, *and* $(S3)$ *hold. Then*

$$0 \leq RS = \kappa\|J_f^T(x_k)f(x_k)\| - \kappa_{GN}\|[J_f^T(x_k)J_f(x_k)]\delta x_k\|.$$

Furthermore, the equality of the above expression is only possible if $\kappa = \kappa_{GN}$.

Proof. Since $J_f^T(x_k)J_f(x_k)\delta x_k = -J_f^T(x_k)f(x_k) + J_f^T(x_k)r_k$ and using the triangle inequality, it follows

$$\|J_f^T(x_k)J_f(x_k)\delta x_k\| \leq \|J_f^T(x_k)f(x_k)\| + \|J_f^T(x_k)r_k\|,$$

Lemma 2.8 yields,

$$\begin{aligned}\|J_f^T(x_k)J_f(x_k)\delta x_k\| &\leq \|J_f^T(x_k)f(x_k)\| + \|J_f^T(x_k)r_k\| \\ &\leq [1 + (\kappa - \kappa_{GN})]\|J_f^T(x_k)f(x_k)\|.\end{aligned}$$

Therefore,

$$-\kappa_{GN}\|J_f^T(x_k)J_f(x_k)\delta x_k\| \geq -\kappa_{GN}[1 + (\kappa - \kappa_{GN})]\|J_f^T(x_k)f(x_k)\|,$$

which implies,

$$\begin{aligned}RS &\geq [\kappa - \kappa_{GN}[1 + (\kappa - \kappa_{GN})]]\,\|J_f^T(x_k)f(x_k)\| \\ &= (\kappa - \kappa_{GN})(1 - \kappa_{GN})\|J_f^T(x_k)f(x_k)\|,\end{aligned}$$

i.e., $RS > 0$ if $\kappa \in (\kappa_{GN}, 1)$. If $\kappa = \kappa_{GN}$, Lemma 2.8 guaranties that $J_f^T(x_k)r_k = 0$ and consequently, we have

$$J_f^T(x_k)J_f(x_k)\delta x_k = -J_f^T(x_k)f(x_k) + J_f^T(x_k)r_k = -J_f^T(x_k)f(x_k),$$

i.e., $RS = 0$.

■

Lemma 2.10. *Let* $(S1)$, $(S2)$, *and* $(S3)$ *hold. Then*

$$\|I - \left[J_f^T(x_k)J_f(x_k)\right] M(x_k)\| \leq (\kappa - \kappa_{GN})cond([J_f^T(x_k)J_f(x_k)])$$

for all $k \in \mathbb{N}$.

Proof. From the Lemma 2.6 part (BE3), we obtain that

$$M(x_k) = \left[J_f^T(x_k)J_f(x_k)\right]^{-1} \left[I + \frac{\left[J_f^T(x_k)r_k\right]\delta x_k^T}{\|J_f(x_k)\delta x_k\|^2}\right] \tag{2.16}$$

therefore,

$$\|I - \left[J_f^T(x_k)J_f(x_k)\right] M(x_k)\| \leq \frac{\|J_f^T(x_k)r_k\|}{\|\delta x_k\|} \frac{\delta x_k^T \delta x_k}{\delta x_k^T \left[J_f^T(x_k)J_f(x_k)\right]\delta x_k}. \tag{2.17}$$

Let λ_m and λ_M be the smallest and the biggest eigenvalue of $\left[J_f^T(x_k)J_f(x_k)\right]$ respectively, then

$$\lambda_m \leq \frac{\delta x_k^T \left[J_f^T(x_k)J_f(x_k)\right]\delta x_k}{\delta x_k^T \delta x_k} \leq \lambda_M$$

i.e.,

$$\frac{\delta x_k^T \delta x_k}{\delta x_k^T \left[J_f^T(x_k)J_f(x_k)\right]\delta x_k} \leq \frac{1}{\lambda_m} = \rho\left(\left[J_f^T(x_k)J_f(x_k)\right]^{-1}\right) \leq \| \left[J_f^T(x_k)J_f(x_k)\right]^{-1} \|. \tag{2.18}$$

From Lemma 2.8, it follows that $\|J_f^T(x_k)r_k\| \leq (\kappa - \kappa_{GN})\|J_f^T(x_k)f(x_k)\|$. Using the above information, we conclude

$$\begin{aligned}
\frac{\|J_f^T(x_k)r_k\|}{\|\delta x_k\|} &\leq (\kappa - \kappa_{GN})\frac{\|J_f^T(x_k)f(x_k)\|}{\|\delta x_k\|} \\
&\leq (\kappa - \kappa_{GN}) \left\|\left[J_f^T(x_k)J_f(x_k)\right]\right\| \frac{\|J_f^T(x_k)f(x_k)\|}{\left\|\left[J_f^T(x_k)J_f(x_k)\right]\right\| \|\delta x_k\|} \\
&\leq (\kappa - \kappa_{GN}) \left\|\left[J_f^T(x_k)J_f(x_k)\right]\right\| \frac{\|J_f^T(x_k)f(x_k)\|}{\left\|\left[J_f^T(x_k)J_f(x_k)\right]\delta x_k\right\|}.
\end{aligned} \tag{2.19}$$

By hypotheses $T_f^T(x_k)r_k \perp K(J_f^T(x_k)J_f(x_k), J_f^T(x_k)f(x_k))$, in particular $T_f^T(x_k)r_k \perp J_f^T(x_k)f(x_k)$ and because

$$J_f^T(x_k)J_f(x_k)\delta x_k = -J_f^T(x_k)f(x_k) + J_f^T(x_k)r_k,$$

we can apply the Pythagorean theorem, which yields

$$\|J_f^T(x_k)J_f(x_k)\delta x_k\|^2 = \|J_f^T(x_k)f(x_k)\|^2 + \|J_f^T(x_k)r_i\|^2.$$

Therefore $\|J_f^T(x_k)f(x_k)\| \leq \|J_f^T(x_k)J_f(x_k)\delta x_k\|$, which means

$$\frac{\|J_f^T(x_k)f(x_k)\|}{\left\|\left[J_f^T(x_k)J_f(x_k)\right]\delta x_k\right\|} \leq 1.$$

Substituting the above information in (2.19), we have

$$\frac{\|J_f^T(x_k)r_k\|}{\|\delta x_k\|} \leq (\kappa - \kappa_{GN}) \left\|\left[J_f^T(x_k)J_f(x_k)\right]\right\|. \tag{2.20}$$

Substituting (2.18) and (2.20) in (2.17), we obtain the result.

■

Lemma 2.11. *Let* $(S1)$, $(S2)$, *and* $(S3)$ *hold. Defining* $A(x_k) := [M(x_k)]^{-1}$, *we obtain*

$$A(x_k) = \left[I - \frac{\left[J_f^T(x_k)r_k\right]\delta x_k^T}{\|J_f(x_k)\delta x_k\|^2}\right]\left[J_f^T(x_k)J_f(x_k)\right],$$

and

$$\|I - A(x_k)\left[J_f^T(x_k)J_f(x_k)\right]^{-1}\| \leq (\kappa - \kappa_{GN})cond([J_f^T(x_k)J_f(x_k)]).$$

Proof. We prove in Lemma 2.6 part $(BE3)$ that

$$A(x_k) = \left[I - \frac{\left[J_f^T(x_k)r_k\right]\delta x_k^T}{\|J_f(x_k)\delta x_k\|^2}\right]\left[J_f^T(x_k)J_f(x_k)\right],$$

therefore,

$$I - A(x_k)\left[J_f^T(x_k)J_f(x_k)\right] = \frac{\left[J_f^T(x_k)r_k\right]\delta x_k^T}{\|J_f(x_k)\delta x_k\|^2}.$$

The rest of the proof follows analogy to Lemma 2.11.

■

Chapter 3

Different κ-Theories

In this Section, we are interested in measuring and bounding the discrepancy between our new local IGN approach ($S3$) and Newton method. Furthermore, we want to interpreted the meaning of such a results. Our IGN approach assumes two contravarinat hypotheses. The fisrt one is that the contravariant error matrix (1.17) introduced by the GN method is bounded by a positive constant $\kappa_{GN} < 1$ and the other one is related to our new stopping criterion (2.15), which implies that the contravariant inner residual relative error (1.27) generated by our IGN method is bounded by a constant less than one. An important question is: Why does we obtain local and linear convergence with our contravariant IGN method?. Typically, contravariant methods ensure faster convergence of the IGN residual sequence, but not of IGN sequence. The answer is that our IGN method is essentially a locally covariant approach. In order to justify this affirmation, we prove in this Chapter that in a vicinity of a stationary point of the nonlinear least squares problem (1.1) the following are valid: If the contravariant error matrix (1.17) introduced by GN method is bounded by κ_{GN}, then there is a certain $\|y\|_*$-norm such that the covarinat error matrix introduced by GN method (1.16) with $\|y\|_*$-norm is also bounded by a constant less than one; and if the covarinat error matrix introduced by GN method (1.16) with euclidean norm is bounded by a constant less than one, then there is a certain $\|\cdot\|_\star$-norm such that the contravarinat error matrix introduced by GN method (1.17) with $\|y\|_\star$-norm is also bounded by a constant less than one. This result say that our first hypothesis is essentially a covariant hypothesis with $\|y\|_*$-norm. Furthermore, controlling the discrepancy between the IGN method and the GN method, we prove that our new stopping criterion defined in (2.15) ensures that the contravariant inner residual relative error (1.27) and the covariant inner relative error with $\|y\|_*$-norm (1.30) are bounded by a constant less than one. Thus, we can also say that our stopping criterion is essentially a covariant stopping criterion with $\|y\|_*$-norm. In this sense, our IGN approach is a covariant approach with $\|y\|_*$-norm. Those results allow also to proof that the hypotheses with norm $\|y\|_*$-norm of the famous local covariant contraction Theorem presented by Bock [10] for our IGN method are valid. Thus, with our IGN approach, we are able to produce a class of locally and linearly convergent IGN residual sequences with euclidean norm, and also produce a class of locally and linearly convergent IGN sequence with $\|y\|_*$-norm.

We organize this Chapter as follows: In the first Section, we presented the covariant local contraction Theorem for Newton-type method of Bock [10] and the hybrid Theorem for Newton-type method of Potschka [70]. Therefore, we prove locally that there is a relation between the covariant error matrix with $\|y\|$-norm introduced by Newton-type

method and the contravariant error matrix with $\|y\|_*$-norm introduced by Newton-type method. In the second Section, we are focus on our new IGN method, we prove that the discrepancy between our IGN method and GN method are bounded. Furthermore, we present a result that explain how the inaccuracy of our IGN method with respect to GN method influences the convergence factor of our IGN sequence. Finally, we discuss briefly when our IGN method implies that the hypotheses of the contravariant contraction Theorem of Bock with $\|y\|_*$-norm are valid.

3.1 Affine Covariant and Hybrid Convergence Theory for Newton-type method

In this secction, we work just with Newton-type method in where we assume that $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$ is a continuously differentiable function with invertible Jacobian $J_F(x)$ and $M(x)$ is a function that may be interpreted as an approximation of $[J_F(x)]^{-1}$, and defines our Newton-type method (see (1.22)). The following Theorem is a variant of the affine covariant Newton-Mysovskikh Theorem 1.9, which controls how large the covarinat inner residual relative error of our Newton-type sequence must be in order to guarantee local and linear convergence. Let us define

$$N := \left\{(x, x') \in D \times D \mid x' = x - M(x)F(x)\right\}$$

and consider the following condition

(*i*) **Covariant Lipschitz condition.** There is a positive and finite constant $\tilde{\omega}$ such that

$$\|M(x')\left[J_F(x + t(x' - x)) - J_F(x)\right](x' - x)\| \leq \tilde{\omega} t \|(x' - x)\|^2$$

for all $t \in [0, 1]$ and $(x, x') \in N$.

(*ii*) **$\tilde{\kappa}$-covariant condition.** There is a positive constant $\tilde{\kappa} < 1$ such that

$$\|M(x')\left[\mathbb{I} - J_F(x)M(x)\right]F(x)\| \leq \tilde{\kappa}\|(x' - x)\|$$

for all $(x, x') \in N$.

(*iii*) The initial guess $x_0 \in D$ is sufficiently close to a solution in the sense that $\tilde{c}_0 := \tilde{\kappa} + \dfrac{\tilde{\omega}}{2}\|\delta x_0\| < 1$ and the closed ball $D_0 := \overline{B}\left(x_0;\ \dfrac{\|\delta x_0\|}{1 - \tilde{c}_0}\right) \subseteq D$. Furthermore, we define $\tilde{c}_k = \tilde{\kappa} + \dfrac{\tilde{\omega}}{2}\|\delta x_k\|$.

Theorem 3.1 (Covariant). *If (i), (ii) and (iii) hold then the Newton-type sequence (x_k), which is defined in (1.22), satisfies the following*

- *$x_k \in D_0$, for all $k \in \mathbb{N}$ and the sequence (x_k) converges to some $x_* \in D_0$ with a convergence rate*

$$\|\delta x_{k+1}\| \leq \tilde{c}_k \|\delta x_k\|$$

- *Furthermore, the a-priori estimate*

$$\|x_{j+k} - x_*\| \leq \frac{(\tilde{c}_k)^j}{1 - \tilde{c}_k}\|\delta x_k\| \leq \frac{(\tilde{c}_0)^{j+k}}{1 - \tilde{c}_0}\|\delta x_0\|$$

holds and the limit x_* satisfies

$$M(x_*)F(x_*) = 0.$$

Proof. Bock [10].

■

In the following, we present the hybrid Theorem of Potschka [70] for Newton-type residual method.

Theorem 3.2 (Hybrid). *Let us assume that the following are valid*

(1) ***Hybrid Lipschitz condition.*** *There is a positive and finite constant ω such that*

$$\| \left[J_F(x + t(x' - x)) - J_F(x) \right] (x' - x) \| \leq \omega t \| (x' - x) \| \| F(x_k) \|$$

for all $t \in [0, 1]$ and $(x, x') \in N$.

(2) ***κ-contravariant condition.*** *There is a positive constant $\kappa < 1$ such that*

$$\| \left[\mathbb{I} - J_F(x) M(x) \right] F(x) \| \leq \kappa \| F(x) \|.$$

for all $(x, x') \in N$.

(3) *The initial guess $x_0 \in D$ is sufficiently close to a solution in the sense that*

$$x_0 \in \mathcal{L} := \{ x \in D \quad | \quad \omega \| M(x) F(x) \| \leq 2(1 - \kappa) \}.$$

Then, the residual sequence $(F(x_k))$ where (x_k) is the Newton-type sequence defined in (1.22) converges to zero with convergence rate

$$\| F(x_{k+1}) \| \leq \left[\kappa + \frac{\omega}{2} \| \delta x_k \| \right] \| F(x_k) \|,$$

and $(F(x_k)) \subset \mathcal{L}$.

■

Of course, the above Theorems are not the only Theorems that determine how large the discrepancy between the Newton method and the Newton-type method must be in order to guarantee local convergence. Indeed, the κ-theory is dedicated to provide κ-condition that controls the error produced by the Newton-type method. Particular examples are:

- Ostrowski κ-Theorem [64, Section 10.2.1] that controls how large must be the espectral radius of the contravariant error matrix or the espectral radius of the covariant error matrix, and concludes local convergence of the Newton-type sequence with root factor of convergence.
- Dennis κ-Theorem [24, Theorem 1] that controls how large must be the contravariant error matrix with $\|y\|$-norm, and conclude local and linear convergence of the Newton-type sequence.
- Dembo, Stanley, Eisenstat, and Steihaug [22] κ-Theorem that controls how large must be the contravariant inner residual relative error with $\|y\|$-norm, and concludes local and linear convergence of the Newton-type sequence with a particular $\|y\|_*$-norm instead of $\|y\|$-norm.

A natural question for our IGN approach is: how does the inexactness of our IGN method with respect to GN method influence the convergence factor of our IGN sequence?. We answer this question for our IGN approach described in $(S3)$ in the following Section, and prove that with certain $\|y\|_*$-norm such a convergence factor depends on the constant κ_{GN} and κ using in our new stopping criterion (2.15). We finalize this Section setting out a connection between the κ-covariant condition based on the covariant matrix error and κ-contravariant condition based on the contravariant matrix error.

Lemma 3.3. *Let $x_* \in D$ be a root of $F(x)$. Let us consider that the both functions $J_F(x)$ and $M(x)$ are continuous at x_*, and that $J_F(x)$ is nonsingular at x_*.*

(i) ***Covariant condition:*** *Let us define the following norm $\|y\|_\star := \| [J_F(x_*)]^{-1} y\|$. If $\|I - M(x_*)J_F(x_*)\| < \kappa < 1$, then for all $\delta \in (0, 1-\kappa)$ there is a vicinity V_δ of x_* that satisfies*

$$\|I - J_F(x)M(x)\|_\star \leq \kappa + \delta < 1 \text{ and } \|I - M(x)J_F(x)\| \leq \kappa + \delta < 1$$

for all $x \in V_\delta$ where $\|A\|_\star = \sup \dfrac{\|Ax\|_\star}{\|x\|_\star}$. Furthermore, $M(x)$ and $J_F(x)$ are nonsingular in V_δ.

(ii) *Reciprocally,* ***contravariant condition:*** *Let us define the following norm $\|y\|_* = \| [J_F(x_*)]\, y\|$. If $\|I - J_F(x_*)M(x_*)\| < \kappa < 1$, then for all $\delta \in (0, 1-\kappa)$ there is a vicinity V_δ of x_* which satisfies*

$$\|I - M(x)J_F(x)\|_* < \kappa + \delta < 1 \text{ and } \|I - J_F(x)M(x)\| \leq \kappa + \delta < 1$$

for all $x \in V_\delta$, where $\|A\|_* = \dfrac{\|Ax\|_*}{\|x\|_*}$. Furthermore, $M(x)$ and $J_F(x)$ are nonsingular in V_δ.

Proof. Because $J_F(x)$ is nonsingular at x_*, we obtain

$$\text{Part } (i) \quad \| [J_F(x_*)]^{-1} [I - J_F(x_*)M(x_*)] [J_F(x_*)] \| = \|I - M(x_*)J_F(x_*)\| < \kappa < 1,$$

$$\text{Part } (ii) \quad \| [J_F(x_*)] [I - M(x_*)J_F(x_*)] [J_F(x_*)]^{-1} \| = \|I - J_F(x_*)M(x_*)\| < \kappa < 1.$$

Given $\delta \in (0, 1-\kappa)$, it follows by continuity of $J_F(x)$ and $M(x)$ at x_* that there is a vicinity V_δ of x_* such that $J_F(x)$ is nonsingular in V_δ,

$$\text{Part } (i) \|I - M(x)J_F(x)\| \leq \kappa + \delta < 1$$

$$\text{Part } (ii) \quad \|I - J_F(x)M(x)\| \leq \kappa + \delta < 1,$$

in V_δ, furthermore,

$$\text{Part } (i) \quad \| [J_F(x_*)]^{-1} [I - J_F(x)M(x)] [J_F(x_*)] \| \leq \kappa + \delta < 1,$$

$$\text{Part } (ii) \quad \| [J_F(x_*)] [I - M(x)J_F(x)] [J_F(x_*)]^{-1} \| \leq \kappa + \delta < 1.$$

From the above inequalities we obtain the result because the following are valid

$$\text{Part } (i) \quad \|A\|_\star = \sup \frac{\|Ax\|_\star}{\|x\|_\star} = \sup \frac{\| [J_F(x_*)]^{-1} A [J_F(x_*)] [J_F(x_*)]^{-1} x\|}{\| [J_F(x_*)]^{-1} x\|}$$

$$= \| \left[J_F(x_*)\right]^{-1} A \left[J_F(x_*)\right] \|,$$

$$\text{Part } (ii) \quad \|A\|_* = \sup \frac{\|Ax\|_*}{\|x\|_*} = \sup \frac{\| \left[J_F(x_*)\right] A \left[J_F(x_*)\right]^{-1} \left[J_F(x_*)\right] x\|}{\| \left[J_F(x_*)\right] x\|}$$
$$= \| \left[J_F(x_*)\right] A \left[J_F(x_*)\right]^{-1} \|.$$

There is just a point that we need to clarify: Is $M(x)$ nonsingular in V_δ?. The answer is positive, we prove just Part (i). Proof by contradiction: Let us assume that $M(x)$ is singular in Part (i), then there is a $v \in V_\delta$ such that $\text{Kern}(M(v)) \neq 0$. Let us choose $w \in \text{Kern}(M(v))$, then

$$\|w\| = \|[I - J_F(v)M(v)]w\| \leq (\kappa + \delta)\|w\| < \|w\|,$$

which is a contradiction.

■

3.2 Relation between Covariant and Contravariant Gauss-Newton Type method

Let us assume that $(S1)$ is valid and consider the set $V_{\kappa_{GN}}$ defined in (1.12). Given a positive constant $\kappa \in (\kappa_{GN}, 1)$, we say that a Newton-type method for solving (1.2) is a Gauss-Newton type method if there is a function $M : V_{\kappa_{GN}} \to \text{GL}(n)$ continuous at x_* such that

$$\|I - J_F(x)M(x)\| \leq \kappa \text{ for all } x \in V_{\kappa_{GN}}, \tag{3.1}$$

and the GN-type sequence is defining by

$$x_{k+1} = x_k + \delta x_k \text{ where } \delta x_k = -M(x_k)F(x_k) \tag{3.2}$$

where $x_0 \in V_{\kappa_{GN}}$.

Relation between contravariant and covariant GN-type approach

Note that the above GN-type method controls how large must be the contravariant matrix error defined in (3.1), and also

$$\|Q_\epsilon(x)[J_f^T(x)J_f(x)]^{-1}\| = \|I - J_F(x)[J_f^T(x)J_f(x)]^{-1}\| \leq \kappa_{GN} \text{ for all } x \in V_{\kappa_{GN}},$$

which means using Proposition 1.6 that $J_F(x)$ is invertible for all $x \in V_{\kappa_{GN}}$, in particular at x_*. Lemma 3.3 ensures that there is a neighborhood $V_\delta \subset V_{\kappa_{GN}}$ of x_* such that GN-type method controls also how large the following covariant matrix errors must be

$$\|I - M(x)J_F(x)\|_* \leq \kappa + \delta < 1 \text{ and } \|[J_f^T(x)J_f(x)]^{-1}Q_\epsilon(x)\|_* \leq \kappa_{GN} + \delta < 1 \text{ for all } x \in V_\delta \tag{3.3}$$

where $\|A\|_*$ is defined in Lemma 3.3.

Reciprocally, let us define

$$\mathcal{V}_{\kappa_{GN}} := \left\{ x \in D \,\middle|\, J_f(x) \text{ is full rank and } \left\| \left[J_f^T(x)J_f(x)\right]^{-1} Q_\epsilon(x) \right\| \leq \kappa_{GN} < 1 \right\}.$$

and assume that $x_* \in \mathcal{V}_{\kappa_{GN}}$. If we define our GN-type method based on controlling the following covariant matrix error

$$\|I - M(x)J_F(x)\| \leq \kappa < 1 \text{ for all } x \in \mathcal{V}_{\kappa_{GN}}, \tag{3.4}$$

where $M(x)$ is continuous at x_* and $x_0 \in \mathcal{V}_{\kappa_{GN}}$, we can also conclude using Lemma 3.3 that the following contravariant matrix errors with $\|y\|_\star$-norm are bounded in a neighborhood $\mathcal{V}_\delta \subset \mathcal{V}_{\kappa_{GN}}$ of x_*

$$\|I - J_F(x)M(x)\|_\star \leq \kappa + \delta < 1 \text{ and } \|Q_\epsilon(x)[J_f^T(x)J_f(x)]^{-1}\|_\star \leq \kappa_{GN} + \delta < 1 \text{ for all } x \in \mathcal{V}_\delta \tag{3.5}$$

where $\|A\|_\star$ is defined in Lemma 3.3.

Theorem 3.4. *The following are valid:*

- *If $x_* \in V_{\kappa_{GN}}$ then there is a neighborhood $V_\delta \subset V_{\kappa_{GN}}$ of x_* such that GN-type methods based on controlling contravariant matrix error with $\|y\|$-norm (3.1) are also GN-type methods based on controlling covariant matrix errors with $\|y\|_*$-norm (3.3).*
- *If $x_* \in \mathcal{V}_{\kappa_{GN}}$ then there is a neighborhood $\mathcal{V}_\delta \subset \mathcal{V}_{\kappa_{GN}}$ of x_* such that GN-type methods based on controlling covariant matrix error with $\|y\|$-norm (3.4) are also GN-type methods based on controlling contravariant matrix errors with $\|y\|_\star$-norm (3.5).*

■

Relation between GN-type method and our IGN approach

Let $(S1)$, $(S2)$, and $(S3)$ hold. Then

(i) The following contravariant matrix error with $\|y\|$-norm is bounded by κ_{GN}

$$\|Q_\epsilon(x)[J_f^T(x)J_f(x)]^{-1}\| = \|I - J_F(x)[J_f^T(x)J_f(x)]^{-1}\| \leq \kappa_{GN} \text{ for all } x \in V_{\kappa_{GN}}.$$

(ii) Our IGN sequence $(x_k) \subset V_{\kappa_{GN}}$ satisfies

$$\|J_f^T(x_k)\left[J_f(x_k)\delta x_k + F(x_k)\right]\| \leq \kappa\|F(x_k)\| - \kappa_{GN}\|[J_f^T(x_k)J_f(x_k)]\delta x_k\|,$$

which implies that the following contravariant inner residual error is bounded by κ

$$\|J_F(x_k)\delta x_k + F(x_k)\| \leq \kappa\|F(x_k)\|. \tag{3.6}$$

Therefore, from Theorem 3.4, it follows that hypothesis (i) implies that the covariant error matrix with norm $\|y\|_*$-norm (1.17) is bounded by a constant less than one. Hypothesis (ii) says that our new stopping criterion implies (3.6), we would like to conclude locally that the following

$$\|M(x_{k+1})\left[J_F(x_k)\delta x_k + F(x_k)\right]\|_* \leq \tilde{\kappa}\|M(x_k)F(x_k)\|_* \text{ with } \tilde{\kappa} < 1 \tag{3.7}$$

is also valid for all k. We prove in the following Section that it is possible to conclude the above relation (3.7) from hypotheses (i) and (ii).

3.3 Inexact Gauss-Newton Contravariant Convergence Theory

Let $(S1)$, $(S2)$, and $(S3)$ hold. In this Section, we present a new Theorem for our IGN method that guarantees linear and local convergence of our IGN sequence (x_k). We measure and bound the discrepancy between our new IGN approach and the GN method through the covariant inner residual relative error (1.18), the contravarinat inner residual relative error (1.19), the contravariant inverse error matrix (1.20), and the contravariant error matrix (1.21). We present a result that explain how the inaccuracy of our IGN method with respect to GN method influences the convergence factor of our IGN sequence. Defining

$$\|y\|_* := \left\| \left[J_f^T(x_*) J_f(x_*) \right] y \right\|,$$

we explain when our new stopping criterion (2.15) implies that the covariant inner relative error with $\|y\|_*$-norm (1.30) is bounded by a constant less than one. Finally, we discuss briefly when our IGN method implies that the hypotheses with $\|y\|_*$-norm of the covariant local contraction Theorem of Bock are valid for our IGN sequence (x_k). We start the analysis proving the existence of a function $M_\kappa(x)$ such that $\delta x_k = -M_\kappa(x)F(x)$.

Lemma 3.5 (**Defining** $M_\kappa(x)$). *Let* $(S1)$, $(S2)$ *and* $(S3)$ *hold. We define the function* $g : V_{\kappa_{GN}} \subset D \to \mathbb{R}^n$ *as follows: for all* $x \in V_{\kappa_{GN}}$, $g(x)$ *solves via LSQR or LSMR the following linear problem*

$$J_f^T(x) J_f(x) \Delta x = -J_f^T(x) f(x) \tag{3.8}$$

with ***stopping criterion***

$$\|J_f^T(x) r(x)\| \leq \kappa \|J_f^T(x) f(x)\| - \kappa_{GN} \|[J_f^T(x) J_f(x)] g(x)\| \tag{3.9}$$

where $r(x) = J_f(x)g(x) + f(x)$. *Let us consider the function* $M_\kappa : V_{\kappa_{GN}} \subset D \to GL(\mathbb{R}^n)$ *such that*

$$M_\kappa(x) = \begin{cases} \left[J_f^T(x) J_f(x) \right]^{-1} \left[I + \dfrac{\left[J_f^T(x) r(x) \right] g(x)^T}{\|J_f(x) g(x)\|^2} \right] & \text{if } \|J_f^T(x) f(x)\| \neq 0 \\ [J_f^T(x) J_f(x)]^{-1} & \text{if } \|J_f^T(x) f(x)\| = 0 \end{cases}$$

Then, $M_\kappa(x)$ *is well defined, invertible with* $A_\kappa(x) = [M_\kappa(x)]^{-1}$, $\delta x_k = -M_\kappa(x_k) F(x_k)$ *for all* $k \in \mathbb{N}$, *and the following are valid,*

$$\begin{aligned} &\textbf{\textit{The contravariant error matrix with }} \|y\|\textbf{\textit{-norm}} \\ &\|I - \left[J_f^T(x) J_f(x) \right] M_\kappa(x)\| \leq (\kappa - \kappa_{GN}) cond([J_f^T(x) J_f(x)]), \text{ and} \end{aligned} \tag{3.10}$$

$$\begin{aligned} &\textbf{\textit{The contravariant inverse error matrix with }} \|y\|\textbf{\textit{-norm}} \\ &\|I - A_\kappa(x) \left[J_f^T(x) J_f(x) \right]^{-1} \| \leq (\kappa - \kappa_{GN}) cond([J_f^T(x) J_f(x)]). \end{aligned} \tag{3.11}$$

Proof. Lemma 2.10, and Lemma 2.11.

■

Remark 3.6. *$M_\kappa(x)$ **is bounded in** $V_{\kappa_{GN}}$.*
Proof. *Let us define*

$$\overline{K} = \max_{x\in\overline{V}_{\kappa_{GN}}} \|[J_f^T(x)J_f(x)]^{-1}\| < +\infty \text{ and } \underline{K} = \max_{x\in\overline{V}_{\kappa_{GN}}} \|[J_f^T(x)J_f(x)]\| < +\infty.$$

Using the Lemma 3.5 part (3.10)*, we obtain*

$$\begin{aligned}
\overline{M} := & \sup_{x\in\overline{V}_{\kappa_{GN}}} \|M(x)\| \\
\leq & \sup_{x\in\overline{V}_{\kappa_{GN}}} \left[(\kappa-\kappa_{GN})cond([J_f^T(x)J_f(x)])+1\right] \|[J_f^T(x)J_f(x)]^{-1}\| \\
\leq & \left[(\kappa-\kappa_{GN})\overline{K}\underline{K}+1\right]\overline{K}.
\end{aligned}$$

■

Lemma 3.7. *Let* $(S1)$*,* $(S2)$*, and* $(S3)$ *hold. The function* $g(x) = -M_\kappa(x)F(x)$ *defined in the above Lemma is continuous at* x_*.

Proof. Here, we need to show that for all sequence $(x_k) \subset V_{\kappa_{GN}}$ that converges to x_*, we have that

$$\lim_{k\to\infty} g(x_k) = g(x_*) = \left[J_f^T(x_*)J_f(x_*)\right]^{-1} F(x_*) = 0.$$

From Lemma 2.8 and construction of $g(x)$ it follows

$$\|J_f^T(x)r(x)\| = \| \left[J_f^T(x)J_f(x)\right] g(x) + F(x)\| \leq (\kappa-\kappa_{GN})\|F(x)\|.$$

Given $(x_k) \subset V_{\kappa_{GN}}$ such that (x_k) converges to x_*, we have by the above inequality that $(J_f^T(x_k)r(x_k))$ converges to zero. By construction of $g(x)$ we obtain that

$$\left[J_f^T(x_k)J_f(x_k)\right] g(x_k) + F(x_k) = J_f^T(x_k)r(x_k).$$

Thus,

$$g(x_k) = \left[J_f^T(x_k)J_f(x_k)\right]^{-1} J_f^T(x_k)r(x_k) - \left[J_f^T(x_k)J_f(x_k)\right]^{-1} F(x_k),$$

which implies that $g(x_k)$ converges to zero, i.e., $g(x)$ is continuous at x_*.

■

Remark 3.8. *Since $g(x)$ defined in Lemma 3.5 is continuous at x_* and $g(x_*) = 0$ then there is a δ_* such that*

$$g(x) = -M_\kappa(x)F(x) \in V_{\kappa_{GN}} \text{ for all } x \in B(x_*,\delta_*) \subset V_{\kappa_{GN}}, \tag{3.12}$$

and x_* is the only root of $F(x)$ in $B(x_*,\delta_*)$.

■

Lemma 3.9. *Let* $(S1)$ *holds. For all* $\delta \in (0, 1-\kappa_{GN})$ *there is a neighborhood* $V_\delta \subseteq V_{\kappa_{GN}}$ *of* x_* *such that for all* $x \in V_\delta$ *the following are valid*

$$\begin{aligned} &\textbf{\textit{The contravariant error matrix}}\ (1.17)\ \textbf{\textit{introduced by GN method:}} \\ &\left\|Q_\epsilon[J_f^T(x)J_f(x)]^{-1}\right\| = \left\|I - J_F(x)[J_f^T(x)J_f(x)]^{-1}\right\| \leq \kappa_{GN} + \delta < 1, \ \textit{and} \\ &\textbf{\textit{The covariant error matrix}}\ (1.16)\ \textbf{\textit{introduced by GN method:}} \\ &\left\|[J_f^T(x)J_f(x)]^{-1}Q_\epsilon(x)\right\|_* = \left\|I - [J_f^T(x)J_f(x)]^{-1}J_F(x)\right\|_* \leq \kappa_{GN} + \delta < 1, \end{aligned} \tag{3.13}$$

where $\|y\|_* = \left\|[J_f^T(x_*)J_f(x_*)]y\right\|$ *and* $\|A\|_* = \sup \dfrac{\|Av\|_*}{\|y\|_*}$.

Proof. Analogous to Lemma 3.3.

■

Remark 3.10. *The above Lemma says that our hypothesis* $(S1)$ *is essentially a covariant hypothesis with* $\|y\|_*$*-norm.*

■

Lemma 3.11. *Let* $(S1)$, $(S2)$, *and* $(S3)$ *hold. Then, for all* $x \in V_{\kappa_{GN}}$ *we have*

$$\|I - J_F(x)M_\kappa(x)\| \leq \kappa_{GN} + (1+\kappa_{GN})(k-\kappa_{GN})cond\left([J_f^T(x)J_f(x)]\right).$$

Proof.

$$\begin{aligned} I - J_F(x)M_\kappa(x) &= I - [J_f^T(x)J_f(x)]M_\kappa(x) + Q_\epsilon(x)M_\kappa(x) \\ &= I - [J_f^T(x)J_f(x)]M_\kappa(x) + Q_\epsilon(x)[J_f^T(x)J_f(x)]^{-1}[J_f^T(x)J_f(x)]M_\kappa(x). \end{aligned}$$

The result follows from Lemma 3.5.

■

Remark 3.12. *Let us define*

$$\tilde{\kappa} := \kappa_{GN} + (1+\kappa_{GN})(k-\kappa_{GN})cond\left([J_f^T(x_*)J_f(x_*)]\right).$$

If $\tilde{\kappa} < 1$, *from the above Lemma there is a positive constant* δ *and a neighborhood* $V_\delta \subset V_{\kappa_{GN}}$ *of* x_* *such that*

$$\|I - J_F(x)M_\kappa(x)\| \leq \tilde{\kappa} + \delta < 1$$

for all $x \in V_\delta$. *Because* $M_\kappa(x)$ *is not continuous at* x_*, *we cannot conclude using an argument analogous to Theorem 3.3 that*

$$\|I - M_\kappa(x)J_F(x)\|_* \leq \tilde{\kappa} + \delta < 1 \ \textit{in}\ V_\delta.$$

■

Lemma 3.13. *If* $(S1)$, $(S2)$, *and* $(S3)$ *hold, then there is a positive constant* ω *such that*

$$\|\left[J_F(x+tg(x)) - J_F(x)\right]M_\kappa(x)F(x)\| \leq \omega t\|g(x)\|\|F(x)\|$$

for all $t \in [0,1]$, $x \in V_{\kappa_{GN}}$, *and* $g(x) \in V_{\kappa_{GN}}$.

Proof. From hypothesis $J_F(x)$ is continuous and the closure of $V_{\kappa_{GN}}$ is compact then $J_F(x)$ is Lipschitz with Lipschitz constant $\tilde{\omega}$. Thus,

$$\|J_F(x - tM_\kappa(x)F(x)) - J_F(x)\| \leq t\omega\|M_\kappa(x)F(x)\|$$

if x and $M_\kappa(x)F(x)$ are in $x \in V_{\kappa_{GN}}$. Therefore,

$$\| \left[J_F(x - tM_\kappa(x)F(x)) - J_F(x)\right] M_\kappa(x)F(x)\| \leq t\omega\|M_\kappa(x)F(x)\|\overline{M}\|F(x)\|$$

defining $\omega := \tilde{\omega}\overline{M}$, we obtain the result.

■

Theorem 3.14. *Let (S1), (S2), and (S3) hold, and let us define*

$$\mathcal{L} := \left\{ x \in \cap \in B(x_*, \delta_*) \quad \middle| \quad \omega\overline{M}\|F(x)\| < 2(1-\kappa) \right\} \subset V_{\kappa_{GN}} \tag{3.14}$$

where $\overline{M}$ is defined in Remark 3.6, $B(x_*, \delta_*)$ in Remark 3.8, and ω in Lemma 3.13. We obtain that the following are valid.

(*i*) ***Contravariant result:*** *If $x_0 \in \mathcal{L}$ then the IGN sequence (x_k) stays in $\mathcal{L}$ and the IGN residual sequence $(F(x_k))$ converges to zero. Furthermore,*

$$\|F(x_{k+1})\| \leq \left[\kappa + \frac{\omega}{2}\|\delta x_k\|\right] \|F(x_k)\|. \tag{3.15}$$

(*ii*) ***Linear-result.*** *Let us choose $x_0 \in \mathcal{L}$ such that $c_0 := \kappa + \frac{\omega}{2}\overline{M}\|F(x_0)\| < 1$, and $\mathcal{B}_0 := B\left(x_0, \frac{\overline{M}\|F(x_0)\|}{1-c_0}\right) \subset \mathcal{L}$. Then, the IGN sequence (x_k) stays in $\mathcal{B}_0$. Furthermore, (x_k) converges to x_* with linear convergence factor and*

$$\|x_k - x_*\| \leq \overline{M}\,(c_0)^k \,\frac{\|F(x_0)\|}{1-c_0}. \tag{3.16}$$

(*iii*) ***Covariant result with $\|\cdot\|_*$-norm.*** *There is an $\overline{\kappa} \in (\kappa_{GN}, 1)$ such that for all $\kappa \in (\kappa_{GN}, \overline{\kappa})$ there is a vicinity $V_\kappa \subset \mathcal{L}$ of x_* that satisfies the following*

(R_1) *For all $x_0 \in V_\kappa$ the correspondent IGN sequence (x_k) stays in $\mathcal{B}_0$. Furthermore, (x_k) converges to x_*.*

(R_2) ***Descent argument.*** *We obtain*

$$\|\delta x_{k+1}\|_* \leq \tilde{c}(x_k)\|\delta x_k\|_* \text{ for all } k \in \mathbb{N}, \tag{3.17}$$

where

$$\|y\|_* := \left\| \left[J_f^T(x_*)J_f(x_*)\right] y \right\|,$$

$$\rho_* := \left\| \left[J_f^T(x_*)J_f(x_*)\right]^{-1} \right\|,$$

$$\mu_* := \left\| \left[J_f^T(x_*)J_f(x_*)\right]^{-1} \right\| \left\| \left[J_f^T(x_*)J_f(x_*)\right] \right\|.$$

and

$$\tilde{c}(x_k) := (1 + 2(\kappa - \kappa_{GN})\mu_*)^4 \left[\kappa + \frac{\omega\rho_*}{2}\|\delta x_k\|_*\right] < 1.$$

(R_3) ***a-priori estimate:***

$$\|x_{k+j} - x_*\|_* \leq \frac{(\tilde{c}(x_k))^j}{1-\tilde{c}(x_k)}\|\delta x_k\|_* \leq \frac{(c(x_0))^{k+j}}{1-c(x_0)}\|\delta x_0\|_*$$

Proof. (i) Note that

$$F(x_k + t_k\delta x_k) = F(x_k) + \int_0^{t_k} J(x_k + t\delta x_k)\delta x_k dt.$$

therefore,

$$\begin{aligned}
\|F(x_k + t_k\delta x_k)\| \leq & \|F(x_k) + t_k J_F(x_k)\delta x_k\| + \int_0^{t_k} \| \left[J(x_k + t\delta x_k) - J_F(x_k)\right]\delta x_k\| dt \\
\leq & (1-t_k)\|F(x_k)\| + \|t_k F(x_k) + t_k J_F(x_k)\delta x_k\| + \int_0^{t_k} \omega t\|\delta x_k\|\|F(x_k)\| dt \\
\leq & (1-t_k)\|F(x_k)\| + t_k\kappa\|F(x_k)\| + \omega\frac{t_k^2}{2}\|\delta x_k\|\|F(x_k)\| \\
\leq & \left[1 - (1-\kappa)t_k + \omega\frac{t_k^2}{2}\|\delta x_k\|\right]\|F(x_k)\|,
\end{aligned}$$

i.e.,

$$\|F(x_k + t_k\delta x_k)\| \leq \left[1 - (1-\kappa)t_k + \omega\frac{t_k^2}{2}\overline{M}\|F(x_k)\|\right]\|F(x_k)\|. \tag{3.18}$$

Using induction, we prove in what follows that $(x_k) \subset \mathcal{L}$. By hypothesis $x_0 \in \mathcal{L}$. Let us assume that $x_1, \cdots, x_k \in \mathcal{L}$, then

$$\begin{aligned}
\|F(x_k + t_k\delta x_k)\| \overset{(3.18)}{\leq} & \left[1 - (1-\kappa)t_k + \omega\frac{t_k^2}{2}\overline{M}\|F(x_k)\|\right]\|F(x_k)\| \\
\overset{x_k\in\mathcal{L}}{\leq} & \left[1 - (1-\kappa)t_k + (1-\kappa)t_k^2\right]\|F(x_k)\|.
\end{aligned} \tag{3.19}$$

Note that $\varphi(t) = 1 - (1-\kappa)t + (1-\kappa)t^2$ defines a convex parabola which has its minimum in $(1/2, \varphi(1/2))$, therefore $0 < 1 - \dfrac{1-\kappa}{4} = \varphi(1/2) \leq \varphi(t) \leq 1$ for all $t \in [0,1]$. Consequently, we obtain

$$\|F(x_k + \delta x_k)\| \leq \varphi(1)\|F(x_k)\| \leq \|F(x_k)\|,$$

which shows that $x_{k+1} \in \mathcal{L}$. Thus, the above inequality is valid for $i = 0, 1, \cdots, k-1, k, k+1$, which implies

$$\|F(x_{k+1})\| \leq \left[\kappa + \omega\frac{1}{2}\overline{M}\|F(x_k)\|\right]\|F(x_k)\| \leq \cdots \leq \left[\kappa + \frac{\omega}{2}\overline{M}\|F(x_0)\|\right]^k\|F(x_0)\|$$

and because $\kappa + \dfrac{\omega}{2}\overline{M}\|F(x_0)\| < 1$, we obtain the result.

(ii) Using induction, we prove in what follows that $(x_k) \subset \mathcal{B}_0$. Let us assume that $x_k \in \mathcal{B}_0$ then

$$\|x_{k+1} - x_0\| \leq \sum_{i=0}^{k} \|M_\kappa(x_i)F(x_i)\| \leq \overline{M} \sum_{i=0}^{k} \|F(x_i)\| \leq \overline{M} \sum_{i=0}^{k} (c_0)^i \|F(x_0)\|.$$

Thereby, it holds that

$$\|x_{k+1} - x_0\| \leq \overline{M}\|F(x_0)\| \frac{1-(c_0)^{k+1}}{1-c_0} < \frac{\overline{M}\|F(x_0)\|}{1-c_0}.$$

Consequently, $x_{k+1} \in \mathcal{B}_0$. Telescopic application of the triangle inequality yields

$$\|x_{k+m} - x_k\| \leq \sum_{i=k}^{k+m-1} \|M_\kappa(x_i)F(x_i)\| \leq \overline{M}\|F(x_0)\| \sum_{i=k}^{m-1} (c_0)^k (c_0)^i \leq \overline{M} (c_0)^k \frac{\|F(x_0)\|}{1-c_0},$$

which proves that (x_k) is a Cauchy sequence, therefore it converges to some root $\tilde{x}_*$ of $F(x)$. Note that $x_* = \tilde{x}_*$ since x_* is the only root of $F(x)$ in $\mathcal{L}$ (see Remark 3.8). The convergence rate (3.16) follows from the above inequality since

$$\|x_{k+m} - x_k\| \leq \overline{M} (c_0)^k \frac{\|F(x_0)\|}{1-c_0}$$

and taking the limit $m \to +\infty$ we obtain

$$\|x_* - x_k\| \leq \overline{M} (c_0)^k \frac{\|F(x_0)\|}{1-c_0}.$$

(iii) From hypothesis $(S1)$ we have that $F(x_*) = 0$. Let us define

$$1 \leq \mu_* := \left\| \left[J_f^T(x_*)J_f(x_*)\right] \right\| \left\| \left[J_f^T(x_*)J_f(x_*)\right]^{-1} \right\|$$

and let us choose an $\overline{\kappa} \in (\kappa_{GN}, 1)$ sufficient close to κ_{GN}, such that

$$(1 + 2(\overline{\kappa} - \kappa_{GN})\mu_*)^4 \overline{\kappa} < 1.$$

Let us fix κ in $(\kappa_{GN}, \overline{\kappa})$. Because $\left[J_f^T(x)J_f(x)\right]$ and $\left[J_f^T(x)J_f(x)\right]^{-1}$ are continuous at x_*, there is an $\epsilon_* \in (0,1)$ such that for all $x \in \mathcal{B}_* := B(x_*, \epsilon_*)$ the following are valid,

$$\left\| \left[J_f^T(x_*)J_f(x_*)\right] \left(\left[J_f^T(x)J_f(x)\right]^{-1} - \left[J_f^T(x_*)J_f(x_*)\right]^{-1} \right) \right\| \leq 2(\kappa - \kappa_{GN})\mu_*,$$

$$\left\| \left(\left[J_f^T(x)J_f(x)\right] - \left[J_f^T(x_*)J_f(x_*)\right] \right) \left[J_f^T(x_*)J_f(x_*)\right]^{-1} \right\| \leq 2(\kappa - \kappa_{GN})\mu_*,$$

$$\left| \left\| \left[J_f^T(x)J_f(x)\right] \right\| \left\| \left[J_f^T(x)J_f(x)\right]^{-1} \right\| - \mu_* \right| \leq \mu_*,$$

$$\|F(x)\| < 2\frac{1-\kappa}{\overline{M}\omega}. \tag{3.20}$$

Applying the triangle inequality, we obtain

$$\left\| \left[J_f^T(x_*)J_f(x_*)\right] \left[J_f^T(x)J_f(x)\right]^{-1} \right\| \leq 2(\kappa - \kappa_{GN})\mu_* + 1, \tag{3.21a}$$

$$\left\| \left[J_f^T(x)J_f(x)\right] \left[J_f^T(x_*)J_f(x_*)\right]^{-1} \right\| \leq 2(\kappa - \kappa_{GN})\mu_* + 1, \tag{3.21b}$$

$$\left\| \left[J_f^T(x)J_f(x)\right] \right\| \left\| \left[J_f^T(x)J_f(x)\right]^{-1} \right\| \leq 2\mu_*, \tag{3.21c}$$

for all $x \in \mathcal{B}_* := B(x_*, \epsilon_*)$. Let us define

$$\rho_* := \left\| \left[J_f^T(x_*)J_f(x_*)\right]^{-1} \right\|,$$

$$\tilde{c}(x) := (1 + 2(\kappa - \kappa_{GN})\mu_*)^4 \left[\kappa + \frac{\rho_*\omega}{2}\|M_\kappa(x)F(x)\|_*\right],$$

$$c(x) := \kappa + \frac{\omega}{2}\overline{M}\|F(x)\| \text{ and } c_0 := c(x_0).$$

Let us consider the following set

$$V_\kappa := \left\{ x \in B\left(x_*, \frac{\epsilon_*}{2}\right) \middle| \frac{\overline{M}\|F(x)\|}{1 - (1 + 2(\kappa - \kappa_{GN})\mu_*)^4 c(x)} < \frac{\epsilon_*}{2} \right\},$$

which by construction satisfies $V_\kappa \subseteq \mathcal{B}_* = B(x_*, \epsilon_*) \subseteq \mathcal{L}$ (see (3.20)). In the following lines, we prove that if $x_0 \in V_\kappa$, the sequence (x_k) stays in $\mathcal{B}_0 := B\left(x_0, \frac{\overline{M}\|F(x_0)\|}{1 - c_0}\right)$ and converges to x_*. Let us choose x_0 in V_κ, therefore $x_0 \in \mathcal{L}$ and it follows form the part (i) that

$$\|F(x_{k+1})\| \leq \left[\kappa + \frac{\omega}{2}\|\delta x_k\|\right] \|F(x_k)\| \text{ for all } k \in \mathbb{N} \tag{3.22}$$

and from part (ii) we conclude that (x_k) stays in $\mathcal{B}_0$ and converge to x_*. We finalize this part verifying that $\mathcal{B}_0 \subseteq \mathcal{B}_* = B(x_*, \epsilon)$, which is an important part in the rest of our argument. Let us fix $x \in \mathcal{B}_0$, because $x_0 \in V_\kappa$, we obtain

$$\begin{aligned} \|x - x_*\| &\leq \|x - x_0\| + \|x_0 - x_*\| \\ &\leq \frac{\overline{M}\|F(x_0)\|}{1 - c_0} + \frac{\epsilon_*}{2} \\ &\leq \frac{\overline{M}\|F(x_0)\|}{1 - (1 + 2(\kappa - \kappa_{GN})\mu_*)^4 c_0} + \frac{\epsilon_*}{2} \\ &\frac{\epsilon_*}{2} + \frac{\epsilon_*}{2} = \epsilon_*, \end{aligned}$$

which implies that $x \in \mathcal{B}_* = B(x_*, \epsilon)$.

(R_2) **Descent argument proof.** Multiplying both sides of the inequality (3.22) by

$$\left\| \left[J_f^T(x_*)J_f(x_*)\right] M_\kappa(x_{k+1}) \right\|$$

we obtain

$$\|M_\kappa(x_{k+1})F(x_{k+1})\|_* \leq \| \left[J_f^T(x_*)J_f(x_*)\right] M_\kappa(x_{k+1})\| \| \left[\kappa + \frac{\omega}{2}\|\delta x_k\|\right] \|F(x_k)\| \quad (3.23)$$

Lemma 3.5 and the triangle inequality yield

$$\| \left[J_f^T(x_{k+1})J_f(x_{k+1})\right] M_\kappa(x_{k+1})\| \leq (\kappa - \kappa_{GN})\text{cond}\left(\left[J_f^T(x_{k+1})J_f(x_{k+1})\right]\right) + 1. \quad (3.24)$$

Using (3.21c), and that $(x_k) \subset \mathcal{B}_0 \subseteq \mathcal{B}_*$, it follows

$$\text{cond}\left(\left[J_f^T(x_{k+1})J_f(x_{k+1})\right]\right) \leq 2\mu_*. \quad (3.25)$$

Substituting (3.25) in (3.24), we obtain

$$\| \left[J_f^T(x_{k+1})J_f(x_{k+1})\right] M_\kappa(x_{k+1})\| \leq 2(\kappa - \kappa_{GN})\mu_* + 1. \quad (3.26)$$

From (3.21a) and using that $(x_k) \subset \mathcal{B}_0 \subseteq \mathcal{B}_*$, we obtain

$$\|\left[J_f^T(x_*)J_f(x_*)\right] \left[J_f^T(x_{k+1})J_f(x_{k+1})\right]^{-1}\| \leq 2(\kappa - \kappa_{GN})\mu_* + 1. \quad (3.27)$$

Because,

$$\left[J_f^T(x_*)J_f(x_*)\right] M_\kappa(x_{k+1}) = \left[J_f^T(x_*)J_f(x_*)\right] \left[J_f^T(x_{k+1})J_f(x_{k+1})\right]^{-1} \left[J_f^T(x_{k+1})J_f(x_{k+1})\right] M_\kappa(x_{k+1}), \quad (3.28)$$

and using (3.26) and (3.27), we conclude

$$\| \left[J_f^T(x_*)J_f(x_*)\right] M_\kappa(x_{k+1})\| \leq (2(\kappa - \kappa_{GN})\mu_* + 1)^2 \quad (3.29)$$

the above inequality and (3.23) yields,

$$\|M_\kappa(x_{k+1})F(x_{k+1})\|_* \leq (1 + 2(\kappa - \kappa_{GN})\mu_*)^2 \left[\kappa + \frac{\omega}{2}\|\delta x_k\|\right] \|F(x_k)\| \quad (3.30)$$

Note that,

$$\begin{aligned}
\|F(x_k)\| = & \|A_\kappa(x_k) \left[J_f^T(x_*)J_f(x_*)\right]^{-1} \left[J_f^T(x_*)J_f(x_*)\right] M_\kappa(x_k)F(x_k)\| \\
\leq & \|A_\kappa(x_k) \left[J_f^T(x_*)J_f(x_*)\right]^{-1} \|\|\delta x_k\|_* \\
\leq & \|A_\kappa(x_k) \left[J_f^T(x_k)J_f(x_k)\right]^{-1} \|\| \left[J_f^T(x_k)J_f(x_k)\right] \left[J_f^T(x_*)J_f(x_*)\right]^{-1} \|\|\delta x_k\|_*
\end{aligned}$$

From Lemma 3.5, triangle inequality, (3.21c),(3.21b) and using that $(x_k) \subseteq \mathcal{B}_0 \subseteq \mathcal{B}_*$ it follows

$$\|F(x_k)\| \leq (2(\kappa - \kappa_{GN})\mu_* + 1)^2\|\delta x_k\|_* \quad (3.31)$$

Substituting the above information in (3.30) and using that

$$\|\delta x_k\| = \| \left[J_f^T(x_*)J_f(x_*)\right]^{-1} \left[J_f^T(x_*)J_f(x_*)\right] \delta x_k\| \leq \rho_*\|\delta x_k\|_*,$$

we obtain

$$\begin{aligned}\|\delta x_{k+1}\|_* \leq & \quad (1+2(\kappa-\kappa_{GN})\mu_*)^4\left[\kappa+\frac{\omega\rho_*}{2}\|\delta x_k\|_*\right]\|\delta x_k\|_* \\ \leq & \quad \tilde{c}(x_k)\|\delta x_k\|_*\end{aligned}$$

i.e.,

$$\|\delta x_{k+1}\|_* \leq \tilde{c}(x_k)\|\delta x_k\|_* \tag{3.32}$$

we have proved that from (3.22) it follows (3.32), since (3.22) is valid for all $k \in \mathbb{N}$, we conclude that (3.32) is also valid for all $k \in \mathbb{N}$. Note that $\tilde{c}(x_k) \leq c(x_0) < 1$ for all $k \in \mathbb{N}$, and rest of the proof follows similar ideas as we did in part (ii).

■

Discussion: In our IGN approach it is assumed

(i) The following contravariant matrix error with $\|y\|$-norm is bounded by κ_{GN}

$$\|Q_\epsilon(x)[J_f^T(x)J_f(x)]^{-1}\| = \|I - J_F(x)[J_f^T(x)J_f(x)]^{-1}\| \leq \kappa_{GN} \text{ for all } x \in V_{\kappa_{GN}}.$$

Using Lemma 3.9, we conclude that there is a positive constant δ and a neighborhood $V_\delta \subset V_{\kappa_{GN}}$ such that

$$\|I-J_F(x)[J_f^T(x)J_f(x)]^{-1}\| \leq \kappa_{GN} \text{ and } \|I-[J_f^T(x)J_f(x)]^{-1}J_F(x)\|_* \leq \kappa_{GN}+\delta < 1$$

for all $x \in V_{\kappa_{GN}}$, which means that our hypothesis (i) implies that the above convariant error matrix with $\|y\|_*$-norm is valid. Thus, hypothesis ($S1$) is essentially a covarinat hypothesis with $\|y\|_*$-norm.

(ii) Our IGN sequence $(x_k) \subset V_{\kappa_{GN}}$ satisfies

$$\|J_f^T(x_k)\,[J_f(x_k)\delta x_k + F(x_k)]\,\| \leq \kappa\|F(x_k)\| - \kappa_{GN}\|[J_f^T(x_k)J_f(x_k)]\delta x_k\|,$$

which implies that the following contravariant inner residual error is bounded by κ

$$\|J_F(x_k)\delta x_k + F(x_k)\| \leq \kappa\|F(x_k)\|. \tag{3.33}$$

From the proof of Theorem 3.14, we have that for a x_0 sufficiently close to x_* the IGN sequence (x_k) satisfies (3.33) and also

$$\|[J_f^T(x_*)J_f(x_*)]M_\kappa(x_{k+1})\| \leq (2(\kappa-\kappa_{GN})\mu_*+1)^2 \text{ see (3.29)},$$

which implies using (3.33) that

$$\|M_\kappa(x_{k+1})[J_F(x_k)\delta x_k + F(x_k)]\|_* \leq (2(\kappa-\kappa_{GN})\mu_*+1)^2\kappa\|F(x_k)\|.$$

Using (3.31), we obtain

$$\|M_\kappa(x_{k+1})[J_F(x_k)\delta x_k + F(x_k)]\|_* \leq (2(\kappa-\kappa_{GN})\mu_*+1)^4\kappa\|M_\kappa(x_k)F(x_k)\|_*. \tag{3.34}$$

Thus, locally, our new stopping criterion (2.15) implies that the contravariant inner residual error with $\|y\|$-norm (3.33) is bounded by $\kappa < 1$, and the covariant inner residual error with $\|y\|_*$-norm (3.34) is bounded by $\tilde{\kappa} := (2(\kappa-\kappa_{GN})\mu_*+1)^4\kappa$ provided that $\tilde{\kappa} < 1$.

Therefore, our new stopping criterion is essentially a covariant stopping criterion with $\|y\|_*$-norm if $\tilde{\kappa} < 1$. Furthermore, from the proof of Theorem 3.14 it follows also that

$$\|M_\kappa(x_{k+1})\left[J_F(x_k - tM_\kappa(x_k)F(x_k)) - J_F(x_k)\right]M_\kappa(x_k)F(x_k)\|_* \leq \tilde{w}t\|M_\kappa(x_k)F(x_k)\|_*^2 \quad (3.35)$$

for all k where $\tilde{w} := (2(\kappa - \kappa_{GN})\mu_* + 1)^4\rho_*\overline{M}\omega$. A natural question is: Is there a relation between the local contraction Theorem for our local IGN approach (3.1) and Theorem (3.14)?. The answer is that locally, our IGN approach implies that the hypotheses of such a Theorem are valid with $\|y\|_*$-norm if $\tilde{k} = (2(\kappa - \kappa_{GN})\mu_* + 1)^4\kappa < 1$.

Chapter 4

Sensitivity Analysis of the Solution

In this Section, we are focus on proving the following result: If we apply the local IGN approach ($S3$) for numerically solving the nonlinear least squares problem (1.1), the solution obtained is statistically stable provided that at least one exists. Using the same notation as in Chapter 1, we have that $h(x) \in \mathbb{R}^m$ represents a mathematical model with a true but unknown parameter $x_{\text{true}} \in \mathbb{R}^n$, $\eta \in \mathbb{R}^m$ represents the observational data, and the measurement error $\epsilon \in \mathbb{R}^m$ is defined as $\epsilon_i = \eta_i - h_i(x_{\text{true}})$, which is assumed to be independent and normally distributed with expected value zero and known variance-covariance matrix $\Sigma = \text{diag}(\sigma_1, \cdots, \sigma_m)$, i.e.,

$$\epsilon \sim \mathcal{N}(0, \Sigma).$$

The discrepancy between the observational data and the model are measured trough the residual function

$$f_\epsilon(x) := \Sigma^{-1}[\eta - h(x)],$$

and a plausible estimation x_*^ϵ of x_{true} is obtained if we solve the following nonlinear least squares problem

$$x_*^\epsilon := \arg\min_{x \in D} \frac{1}{2}\|f_\epsilon(x)\|_2^2 =: T_\epsilon(x). \tag{4.1}$$

Defining $\hat{y} := h(x_{\text{true}})$, we have

$$f_\epsilon(x) = \Sigma^{-1}[\eta - h(x)] = \Sigma^{-1}[\epsilon + h(x_{\text{true}}) - h(x)] = \Sigma^{-1}[\epsilon + \hat{y} - h(x)], \tag{4.2}$$

which means that $y_*^\epsilon = f_\epsilon(x_*^\epsilon)$ is a perturbation of $0 = f_0(x_{\text{true}})$. A natural question is: how close are x_*^ϵ and x_{true}?, or in another words: how good is the estimation x_*^ϵ of our true parameter x_{true}?. Intuitively, if we are making small perturbations on the observational data of our problem (4.1), i.e, $\|\epsilon\|$ is small, we must conclude that the distance between x_*^ϵ and x_{true} is small. Otherwise, x_*^ϵ is not a reliable estimation of x_{true}. In order to answer the above question, we focus in this Chapter on estimation x_*^ϵ of x_{true} that satisfies the following definition.

Definition 4.1. *We say that a local solution x_*^ϵ of (4.1) is statistically stable under perturbation makes in the measurement error ϵ if x_*^ϵ is a continuously deformation of x_{true}, i.e., there is a continuously function $\phi : B(0, r_*) \to D$ such that $x_*^\epsilon = \phi(\epsilon)$, $x_{true} = \phi(0)$, and x_*^ϵ is locally the unique solution of (4.1) for all $\epsilon \in B(0, r_*)$.*

Note that from the hypotheses assumed in Chapter 1, we have that the following are valid.

(*i*) The function $f_\epsilon(x)$ is twice continuously differentiable with Jacobian $J_{f_\epsilon}(x)$, and there is a neighborhood V of x_{true} such that $J_{f_\epsilon}(x)$ is full rank for all $x \in V$.

(*ii*) Defining

$$H_\epsilon(x) := \nabla^2 T_\epsilon(x) \overset{(1.3)}{=} \left[J_{f_\epsilon}^T(x) J_{f_\epsilon}(x)\right] + Q_\epsilon(x),$$

we obtain that $H_0(x_{\text{true}}) = J_{f_0}^T(x_{\text{true}}) J_{f_0}(x_{\text{true}})$ is positive definite.

Lemma 4.2. *There is a positive constant r_*, and a continuously differentiable path $\phi : B(0, r_*) \to D$ such that $x_*^\epsilon = \phi(\epsilon)$ and $x_{true} = \phi(0)$, where $x_*^\epsilon = \phi(\epsilon)$ satisfies the following*

(*L*1) *$\phi(\epsilon)$ is locally the unique stationary point of* (4.1)*, i.e.,*

$$\nabla T_\epsilon(\phi(\epsilon)) = \left[J_{f_\epsilon}(\phi(\epsilon))\right]^T f_\epsilon(\phi(\epsilon)) = 0. \tag{4.3}$$

(*L*2) *$H_\epsilon(\phi(\epsilon))$ is positive definite.*

Proof. Let us consider the following functions

$$\mathcal{F}(x, \epsilon) = f_\epsilon(x) \text{ and } \mathcal{T}(x, \epsilon) = T_\epsilon(x) \text{ with } (x, \epsilon) \in V \times \mathbb{R},$$

and note that from (*i*) it follows that $\mathcal{F}$ is twice continuously differentiable. Let us consider the following nonlinear equation

$$\nabla_x \mathcal{T}(x, \epsilon) = \nabla T_\epsilon(x) = \left[J_{f_\epsilon}(x)\right]^T f_\epsilon(x) = 0,$$

then from (*i*) and (*ii*) we obtain that

$$\nabla_x \mathcal{T}(x_{\text{true}}, 0) = 0 \text{ and } \nabla_x^2 \mathcal{T}(x_{\text{true}}, 0) = H_0(x_{\text{true}}) \text{ is positive definite,}$$

which implies by virtue of the implicit function Theorem the existence of a ball $B(0, r_*) \subset V \subset D$ and a continuously differentiable path $\phi : B(0, r_*) \to V$ such that $\phi(\epsilon)$ satisfies (*L*1). Because $H_0(x_{\text{true}})$ is a positive definite matrix and also a continuous function with respect to ϵ, we can reduce the value of r_* if it is necessary and conclude that $H_\epsilon(\phi(\epsilon))$ is a positive definite matrix for all $\epsilon \in B(0, r_*)$.

■

Remark 4.3. *The second-order sufficient condition Theorem 1.3 guarantees that $x_*^\epsilon = \phi(\epsilon)$ defined in the above Lemma is locally the unique solution of the nonlinear least squares problem* (4.1) *for all $\epsilon \in B(0, r_*)$. Furthermore, $x_*^\epsilon = \phi(\epsilon)$ is a statistically stable solution of* (4.1) *for all $\epsilon \in B(0, r_*)$.*

We organize the rest of this chapter as follows: firts, we present the covariant $\tilde{\kappa}$-Theorem of Bock [10] known as the local contraction Theorem for GN method, which determines locally when the GN method converges to a statistically stable x_*^ϵ. Later, we prove that our IGN approach (*S*3) locally provides statistically stable solutions provided that at least one exists.

4.1 Statistically Stable κ-Theorems

Let us consider the Moore-Penrose pseudoinverse $J_\epsilon^+(x)$ of $J_{f_\epsilon}(x)$ in V defined in (i), i.e.,

$$J_\epsilon^+(x) = \left[J_{f_\epsilon}^T(x) J_{f_\epsilon}(x)\right]^{-1} J_{f_\epsilon}^T(x) \quad \text{for all } x \in V,$$

the Gauss-Newton step $\Delta x = -J_\epsilon^+(x) f_\epsilon(x)$, and the Gauss-Newton inner residual $R_\epsilon(x) = \left[I - J_{f_\epsilon}(x) J_{f_\epsilon}^+(x)\right] f_\epsilon(x)$.

Theorem 4.4 (Local Contraction). *Let us assume*

$(B1)$ ***Lipschitz covariant condition.*** *There is a constant $\omega < \infty$ such that*

$$\|J_\epsilon^+(x') \left[J_{f_\epsilon}(x') - J_{f_\epsilon}(x)\right](x' - x)\| \leq \omega t \|x' - x\|^2 \text{ for all } x', x \in V.$$

$(B2)$ ***$\tilde{\kappa}$-covariant condition.*** *There is a constant $\tilde{\kappa} < 1$ such that*

$$\left\|J_\epsilon^+(x') R_\epsilon(x)\right\| \leq \tilde{\kappa} \|x' - x\| \text{ for all } x', x \in V. \tag{4.4}$$

$(B3)$ *The initial guess $x_0 \in V$ is sufficiently close to a solution that*

$$c_0 := \tilde{\kappa} + \frac{\omega}{2}\|\Delta x_0\| < 1 \text{ and } B_0 := B\left(x_0, \frac{\|\Delta x_0\|}{1 - c_0}\right) \subset V.$$

Defining $c_k = \tilde{\kappa} + \frac{\omega}{2}\|\Delta x_k\|$, we obtain

$(R1)$ *$x_k \in B_0$, for all $k \in \mathbb{N}$ and the Gauss Newton sequence (x_k) converges linearly to $x_*^\epsilon \in B_0$ with descent argument*

$$\|\Delta x_{k+1}\| \leq c_k \|\Delta x_k\|.$$

$(R2)$ *Furthermore, the a-priori estimate*

$$\|x_{j+k} - x_*^\epsilon\| \leq \frac{(c_k)^j}{1 - c_k}\|\Delta x_k\| \leq \frac{(c_0)^{j+k}}{1 - c_0}\|\Delta x_0\|$$

holds.

$(R3)$ *The limit x_*^ϵ satisfies*

$$F_\epsilon(x_*^\epsilon) = \nabla T_\epsilon(x_*^\epsilon) = \left[J_{f_\epsilon}(x_*^\epsilon)\right]^T f_\epsilon(x_*^\epsilon) = 0.$$

Proof. Bock [10].

■

Theorem 4.5 (GN-Statistically Stable Solution). *If the following condition*

$$\left\|J_\epsilon^+(x) R_\epsilon(x_*^\epsilon)\right\| \leq \tilde{\kappa} \|x - x_*^\epsilon\|$$

is valid for all x in a neighborhood of x_{true} then the matrix $H_\epsilon(x_*^\epsilon)$ is positive definite.

Proof. Bock [10].

■

Remark 4.6. *Note that if Theorem 4.4 is valid, then the hypothesis of Theorem 4.5 is also valid, and consequently, we conclude that the Gauss-Newton sequence (x_k) that satisfies the conditions of Theorem 4.4 converges to statistically stable solutions.*

Remark 4.7. *Bock [10] proposed that the statistically stable solutions x_*^ϵ of (4.1) are defined by the $100\alpha\%$ confidence region*

$$G := \left\{\epsilon \mid \|\epsilon\|^2 \leq \gamma^2(\alpha)\right\}$$

where $\alpha \in [0,1]$ and $\gamma^2(\alpha) = \mathcal{X}_n^2(1-\alpha)$ is the quantile of the $\mathcal{X}^2$-distribution with n degree of freedom. For example, we can choose in most of the case $\alpha = 0.95$.

Remark 4.8. *If is clear that if $\epsilon \in G$ then $\epsilon \approx 0$, which implies from (4.2) that $\epsilon \approx f_\epsilon(x_{true}) - f_\epsilon(x_*^\epsilon)$, therefore the confidence region is defined by other authors [54, 8, 51] as*

$$G(\alpha, x_{true}) := \left\{x \in V \mid \|f_\epsilon(x)\|^2 - \|f_\epsilon(x_{true})\|^2 \leq \gamma^2(\alpha)\right\},$$

which is a more adequate presentation since we handle directly with the statistically stable solutions. Because the parameter x_{true} is unknown, it is used $G_L(\alpha, x_*^\epsilon)$ instead of $G(\alpha, x_{true})$ where $G_L(\alpha, x_*^\epsilon)$ describe the confidence ellipsoid which is a first order approximation of the nonlinear confident region $G(\alpha, x_{true})$, i.e.,

$$G_L(\alpha, x_*^\epsilon) = \left\{x \in V \mid \|f_\epsilon(x_*^\epsilon) + J_{f_\epsilon}(x_*^\epsilon)\,[x - x_*^\epsilon]\,\|^2 - \|f_\epsilon(x_*^\epsilon)\|^2 \leq \gamma^2(\alpha)\right\}.$$

Körkel proved in his dissertation thesis [54] that

$$G_L(\alpha, x_*^\epsilon) = \left\{x \in V \mid x - x_*^\epsilon = J_{f_\epsilon}^+(x_*^\epsilon)\delta w \text{ and } \|\delta w\|^2 \leq \gamma^2(\alpha)\right\}.$$

For a deeper study of statistically stable solution and its applications see e.g., [8, 54, 52, 51, 11].

Theorem 4.9. *Let $(S1)$, $(S2)$, and $(S3)$ hold. Then our IGN method produces statistically stable solutions.*

Proof. Let us define

$$\kappa(x_*^\epsilon) := \left\| Q_\epsilon(x_*^\epsilon)\left[J_{f_\epsilon}^T(x_*^\epsilon) J_{f_\epsilon}(x_*^\epsilon)\right]^{-1} \right\|, \tag{4.5}$$

and note that $\kappa(x_{\text{true}}) = 0$ since in this case $Q_0(x) = 0$, thus there is a $\tilde{r}_* > 0$ such that

$$\kappa(x_*^\epsilon) \leq \kappa_{GN} \text{ for all } \|\epsilon\| \leq \tilde{r}_*.$$

Using Theorem 3.14, we conclude that the IGN sequence (x_k) converges linearly and locally to x_*^ϵ for all $\|\epsilon\| \leq \tilde{r}_*$, and from the proposition 1.6 it follows that $H_\epsilon(x_*^\epsilon)$ is positive definite. Thus, the implicit function Theorem applies to the nonlinear equation (4.3) delivers the ball $B(0, \tilde{r}_*)$, and a continuously differentiable path $\phi : B(0, \tilde{r}_*) \to D$ such that $x_*^\epsilon = \phi(\epsilon)$ and $x_{\text{true}} = \phi(0)$ where $x_*^\epsilon = \phi(\epsilon)$ is the unique local solution of (4.3). Choosing an α sufficient close to one, we conclude that the solutions x_*^ϵ of (4.1) with $\epsilon \in G$ are statistically stable solutions.

■

Chapter 5

Global Newton Methods

Let $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$ be a continuously differentiable function with Jacobian $J_F(x)$, and let us consider the nonlinear equation

$$F(x) = 0. \tag{5.1}$$

In this Chapter, we restrict our study to globalization strategies based on the Newton method with damping strategy, i.e.,

$$x_{k+1} = x_k + t_k \Delta x_k, \quad \text{with } t_k \in [0, 1], \text{ and } J_F(x_k)\Delta x_k = -F(x_k), \tag{5.2}$$

where $x_0 \in D$ is pre-chosen and not necessarily close to a solution x_* of (5.1), and t_k is the step size or damping factor. An analysis made in the classical globalization strategies based on the popular residual monotonicity test, and on the natural monotonicity test reveals the principal drawbacks of the globalization strategies based on a particular merit function. Rather, we focus on globalization strategies that follow the affine covariant Newton path $\mathcal{P}(t)$, which connects the initial guess x_0 with a solution x_* of our problem (5.1) and produces along of it an exponential descent in every general level function [27]. Two globalization strategies based on following such a path $\mathcal{P}(t)$ are presented, one of them was introduced by Bock, Kostina, and Schlöder [12] and is known as Restrictive Monotonicity Test (RMT). The other one was introduced by Potschka [70] and it is known as Backward Step Control (BSC), which provides, under reasonable assumptions, a global convergence Theorem on the basis of a backward step argument.

5.1 Residual Based Descent

In this section, we focus on a line search strategy based on the **residual level function**,

$$T_F(x) = \frac{1}{2}\|F(x)\|_2^2 \tag{5.3}$$

with associated level set,

$$G(x_0) := \{y \in D \mid T_F(y) \leq T_F(x_0)\} \tag{5.4}$$

where $x_0 \in D$ is a pre-chosen initial guess. Note that in this case the Newton step Δx_k is a descent direction of our residual level function $T_F(x)$ if $\Delta x_k \neq 0$, since

$$[\nabla T_F(x_k)]^T \Delta x_k = -2T_F(x_k) < 0.$$

In order to produce the maximal descent in this direction, we must choose t_k such that

$$t_k^* := \min_{t\in[0,1]} T_F(x_k + t\Delta x_k).$$

An exact minimization is expensive and unnecessary. Instead, we apply an approximate line search which determines the largest damping factor t_k such that the residual reduction is in some sense optimal. The following Lemma tells us how to choose it.

Lemma 5.1. *Let $F : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ be a continuously differentiable function, with D open, convex, and $J_F(x)$ nonsingular for all $x \in D$. Assuming that the* ***affine contravariant Lipschitz condition***

$$\| [J_F(x) - J_F(y)] (y - x)\| \leq \omega \|J_F(x)(x - y)\|^2 \quad \textit{for all} \quad x, y \in D$$

is valid and defining the convenient notation $h_k := \omega\|F(x_k)\|$, we conclude that

$$\|F(x_k + t\Delta x_k)\| \leq \rho_k(t)\|F(x_k)\| \quad \textit{for all} \quad t \in [0, 2/h_k]\,,$$

where $\rho_k(t) := 1 - t + \frac{1}{2}t^2 h_k$, and the ***optimal choice of the damping factor*** *is given by*

$$t_k^o := \min(1, 1/h_k).$$

Proof. Deuflhard [27, Theorem 3.7].

■

The following Theorem says that the damped Newton sequence (x_k) with damping factor $t_k \approx t_k^o$ converges to a solution x_* of (5.1) for all pre-chosen initial guess $x_0 \in G_0$, where G_0 denote the path-connected component of $G(x_0)$ containing x_0.

Theorem 5.2. *If the hypotheses of Lemma 5.1 are valid and $G_0 \subseteq D$ is compact, then the damped Newton iteration* (5.2) *with damping factor in the range*

$$t_k \in [\epsilon, 2t_k^o - \epsilon]\,,$$

where $\epsilon > 0$ is sufficiently small and depends on G_0, converges to some solution point x_* of (5.1).

Proof. Deuflhard [27, Theorem 3.8].

■

We discourage the use of this line search strategy first of all because it can not be implemented directly since the quantity h_k is computationally unavailable due to the affine contravariant Lipschitz constant ω. On the other hand, if we know the value of ω then this line search strategy applied to mildly ill-conditioned problems may produce small stepsizes even in a vicinity of x_* where the quadratic convergence of the Newton sequence with full-step is guaranteed. The reason for this behavior is that the damping factor depends directly on the merit function. We explain in more detail such a behavior in the following section.

5.2 Error Oriented Descent

Here, we focus on a line search strategy based on the **general level function**

$$T_F(x \mid A) = \frac{1}{2}\|AF(x)\|_2^2 \tag{5.5}$$

with $A \in \mathrm{GL}(n)$, and associated level set,

$$G(x_0 \mid A) := \{y \in D \mid T_F(y \mid A) \leq T_F(x_0 \mid A)\}$$

where $x_0 \in D$ is a pre-chosen initial guess.

Remark 5.3. *In this case, the Newton step* $\Delta x_k = -\left[J_F(x_k)\right]^{-1} F(x_k)$ *is a descent direction with respect to all such level functions* (5.5)*, since*

$$\Delta x_k^T \nabla T_F(x_k \mid A) = -2T_F(x_k \mid A) < 0.$$

The following Lemma reveals the reason why the line search strategy based on the residual level function may be inefficient already in mildly ill-conditioned problems.

Lemma 5.4. *Let* $F : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ *be a continuously differentiable function, with* D *open, convex, and* $J_F(x)$ *nonsingular for all* $x \in D$*. If* $x_k \in G(x_k \mid A)$ *for some* $A \in GL(n)$*, and the* ***affine covariant Lipschitz condition***

$$\| \left[J_F(x)\right]^{-1} \left[J_F(y) - J_F(x)\right](y - x)\| \leq \omega \|(x - y)\|^2$$

is valid for all $x, y \in D$*, then with the convenient notation*

$$h_k := \omega\|\Delta x_k\|, \quad \overline{h}_k := h_k cond\left(AJ_F(x_k)\right),$$

we obtain that

$$T_F(x_k + t\Delta x_k \mid A) \leq \rho_k(t \mid A) T_F(x_k \mid A) \;\; \textit{for all} \;\; t \in \left[0, \min\left(1, 2/\overline{h}_k\right)\right] \tag{5.6}$$

where $\rho_k(t \mid A) := 1 - t - \frac{1}{2}t^2\overline{h}_k$*, and* ***the optimal choice of the damping factor*** *is given by*

$$t_k^o(A) := \min\left(1, 1/\overline{h}_k\right).$$

Proof. Deuflhard [27, Theorem 3.12].

■

The optimal choice of the damping factor $t_k^o(I) = t_k^o$ when we are working with a line search strategy based on the residual level function ($A = I$) may be inefficient, since even for mildly ill-conditioned problems $cond(J_F(x_k))$ may be too large, and consequently the interval where (5.6) is valid may be very small. Thereby, the damping factor t_k may be tiny even when the quadratic convergence of the Newton sequences with full Newton step is guaranteed. Therefore such an approach is inefficient. An example (Rosenbrock-type) that describes this behavior was given by Bock in [10].

On the other hand, if we choose as merit function the **Natural Level Function**,

$$T_k(x) := T_F(x \mid [J_F(x_k)]^{-1}) = \frac{1}{2}\| \left[J_F(x_k)\right]^{-1} F(x)\|^2$$

then from the Lemma 5.4 the line search strategy based on Natural level function at the iterate x_k satisfies the following properties:

1. **Extremal properties** (Deuflhard [27]), For all $A \in \mathrm{GL}(n)$ the reduction factor $\rho_k(t \mid A)$ and the theoretical optimal damping factor $t_k^o(A)$ satisfy,

$$\rho_k(t \mid [J_F(x_k)]^{-1}) \leq \rho_k(t \mid A),$$

$$t_k^o([J_F(x_k)]^{-1}) = \min\left(1, 1/\overline{h}_k\right) \geq t_k^o(A).$$

2. **Steepest Descent** (Deuflhard [27]): The Newton correction Δx_k given by $J_F(x_k)\Delta x_k = -F(x_k)$ is a descent direction with respect to all such level functions (5.5).

3. **Full step in a vicinity of** x_* (Deuflhard [27]). From the affine covariant Newton-Mysovskikh Theorem 1.9, we know that local quadratic convergence of the Newton sequence is guaranteed in the vicinity V_* of some solution x_* where

$$V_* := \left\{ x \in D \mid \omega \| \left[J_F(x)\right]^{-1} F(x) \| \leq \alpha < 2 \right\}.$$

   If $h_k = \omega \|\Delta x_k\| < 1$, then $t_k^o([J_F(x_k)]^{-1}) = \min\left(1, 1/\overline{h}_k\right) = 1$, i.e., using the Natural level function at the iterate x_k, our damping factor is equal to one in the vicinity V_*.

4. **Asymptotic distance function** (Bock [12]). For F twice continuously differentiable, and assuming that (x_k) converge to x_*, then

$$\left\| \left[J_F(x_k)\right]^{-1} F(x) \right\|_2 = \|x - x_*\|_2 \left[1 + O(\|x - x_*\|_2) + O(\|x_k - x_*\|_2)\right].$$

Both line search strategies have their disadvantages. The line search strategy based on the natural level function at the iterate x_k satisfies the above outstanding property but all of them are just valid at the iterate x_k, therefore we cannot guarantee with this strategy descent of our merit function $T_k(x)$ for all iterate of our Newton damping sequence, which is the major drawback of this approach, since the classical arguments of global convergence cannot be applied. Indeed, Ascher and Osborne [4] constructed a theoretical example that show the existence of two-cycles for successive quadratic programming solver method based on the natural level function at the iterate x_k. On the other hand, a global convergence proof exists for Newton damping iterate based on the residual function Deuflhard [27, Theorem 3.13]. Nevertheless, we have seen that this approach produces tiny damping factors even in the vicinity of the solution where full-step Newton sequence converges quadratically.

5.3 The Newton Path

In this section, we present the Newton-path, which is an affine covariant path that produces descent in the whole class of merit functions (5.5). Furthermore, we introduce the intuitive ideas that motivate the line search strategies based on RMT and BSC.

In order to avoid the arbitrary choice of a merit function introduced by the whole class (5.5), we focus on the intersection of all correspondent level sets

$$\overline{G}(x) := \bigcap_{A \in \mathrm{GL}(n)} G(x \mid A).$$

It is clear that this set is affine covariant since by construction it do not depend on $A \in \mathrm{GL}(n)$. The following Theorem reveals its properties.

Theorem 5.5. *Let $F : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ be a continuously differentiable function with $J_F(x)$ nonsingular for all $x \in D$. For some $\widehat{A} \in GL(n)$, let the path-connected component of $G(x_0 \mid \widehat{A})$ in x_0 be compact and contained in D. Then the path-connected component of $\overline{G}(x_0)$ is a topological path $\overline{x}_0 : [0, 2] \to \mathbb{R}^n$, which satisfies*

$$F(\overline{x}_0(\lambda)) = (1 - \lambda)F(x_0), \tag{5.7}$$

$$T_F(\overline{x}_0(\lambda) \mid A) = (1 - \lambda)^2 T_F(x_0 \mid A), \tag{5.8}$$

$$\begin{aligned} \frac{d\overline{x}_0}{d\lambda} &= -\left[J_F(\overline{x}_0)\right]^{-1} F(x_0), \\ \overline{x}_0(0) &= x_0, \\ \overline{x}_0(1) &= x_* \text{ where } F(x_*) = 0, \end{aligned} \tag{5.9}$$

$$\frac{d\overline{x}_0}{d\lambda}(0) = \Delta x_0 \tag{5.10}$$

where Δx_0 is the ordinary Newton correction.

Proof. Deuflhard [27, Theorem 3.6].

■

Remark 5.6. *In the proof of the above Theorem the differential equation (5.9) is derived from the homotopy*

$$H(x, \lambda) = F(x) - (1 - \lambda)F(x_0) = 0, \tag{5.11}$$

which defines the function $\overline{x}_0(\lambda)$ upon the invocation of the implicit function Theorem. A natural strategy to construct our damped Newton factor λ_k would be: choose it such that

$$F(x_k + \lambda_k \Delta x_k) - (1 - \lambda_k)F(x_k) \approx 0.$$

From (5.7), it follows that $\overline{x}(\lambda_k) \approx x_{k+1}$. This approach will be explained in more detail in the following section, and with additional requirements, it is known as the Restrictive Monotonicity Test (RMT).

Remark 5.7. *Introducing the reparametrization $\lambda(t) = 1 - e^{-t}$, we obtain from the above homotopy (5.11) the so called continuous Newton method or Davidenko differential equation [21],*

$$\dot{\hat{x}}_0(t) = -\left[J_F(\hat{x}_0(t))\right]^{-1} F(\hat{x}_0(t)), \quad t \in [0, +\infty), \quad \hat{x}_0(0) = x_0. \tag{5.12}$$

Furthermore, $\lim_{t \to +\infty} \hat{x}_0(t) = x_*$.

Remark 5.8. *From the remark* (5.7) *we conclude that the Newton path* $\hat{x}_0(t)$ *is connecting continuously a starting guess* x_0 *with a local solution* x_* *of our problem* (5.1) *in some optimal sense:*

(i) For any $t \in [0, +\infty)$, $\hat{x}_0(t)$ *is the unique local solution of the following equation,*

$$H(x, \lambda(t)) = F(x) - e^{-t}F(x_0) = 0,$$

which is not our original nonlinear equation but its solution $\hat{x}_0(t)$ *is very close to* x_* *when* t *is large.*

(ii) From property (5.8)*, and the reparametrization* $\lambda(t)$*, it is clear that the Newton path* $\hat{x}_0(t)$ *produces exponential descent in the whole class of merit functions* (5.5) *when* $t \in [0, +\infty)$.

Remark 5.9. *Another strategy for choosing our damping factor, which is not so natural as the introduced in remark* (5.6)*, would be to construct our iterate* $x_{k+1} = x_k + t_k\Delta_k$*, with* $t_k \in [0, 1]$*, such that*

$$\lim_{k\to+\infty} \hat{x}_0\left(\sum_{i=1}^{k-1} t_i\right) - x_k = 0, \tag{5.13}$$

and

$$\lim_{k\to\infty} \sum_{i=1}^{k-1} t_i = +\infty. \tag{5.14}$$

From (5.13) *and remark* (5.8) *we conclude that our sequence* (x_k) *converges to* x_*.

A way to construct such an damped sequence (x_k) is possible if we assume some hypotheses introduced by Potschka [70].

5.4 The Restrictive Monotonicity Test

The Monotonicity Restrictive Test (RMT) introduced by Bock, Kostina, and Schlöder [12] is a Global Newton Strategy, which chooses as damping factor t_k a positive number such that $x_{k+1} = x_k + t_k\Delta x_k$ is close to the Newton path $\overline{x}_k(t)$ that emanates from x_k. In fact, it was proved in [12, Lemma 7] that if

$$\epsilon_k(t) := \overline{x}_k(t) - x_k - t\Delta x_k$$

and $F(x)$ is twice continuously differentiable then

$$\epsilon_k(t) = -\left[J(x_k)\right]^{-1}\left[F(x_k + t\Delta x_k) - (1-t)F(x_k)\right] + O(t^3).$$

From this result, the damping factor can be chosen such that

$$\|t\Delta x_k\|\eta_* \leq \| \left[J(x_k)\right]^{-1}\left[F(x_k + t\Delta x_k) - (1-t)F(x_k)\right] \| \leq \|t\Delta x_k\|\eta^*.$$

The above relation is controlled by the choice of the positive numbers η_*, $\eta^* < 2$. This RMT strategy, which can be interpreted as a step size control for integration of the Davidenko differential equation (5.12) with the explicit Euler method, shows very good practical applications, in particular, it does not lead to two cycles. The explicit Euler

method can be extended by a number of so-called back projection steps which diminish the distance of the iterate x_{k+1} to the Newton path $\overline{x}_k(t)$, therefore this RMT method can be extended through repeated back projections step that provide us the benefit of a global convergence Theorem. Nevertheless, numerical experience shows that more than one back projection step does not improve convergence considerable and thereby it should be avoided [12].

5.5 Backward Step Control for Damped Newton Methods

Let us assume that $J_F(x)$ is invertible for all $x \in G(x_0)$ and let us define

$$g(x) = [J_F(x)]^{-1} F(x) \text{ and } h(x,t) = g(x - tg(x)) - g(x).$$

Let us consider the Davidenko equation (5.12) with starting point x_k and $k \geq 1$, i.e.,

$$\begin{array}{rcl} \dot{\hat{x}}_k(t) & = & -g(\hat{x}_k(t)) \\ \hat{x}_k(0) & = & x_k. \end{array} \tag{5.15}$$

Motivation: Given a positive constant H, we are interested in choosing a step size $t_k \in [0,1]$ such that the following properties are valid:

1. Starting from x_k, the *forward Euler* method applied to (5.15) provides the iterate $x_{k+1} = x_k - t_k g(x_k)$.

2. Starting from x_{k+1}, the *backward Euler* method applied to

$$\begin{array}{rcl} \dot{\hat{x}}_{k+1}(t) & = & -g(\hat{x}_{k+1}(t)) \\ \hat{x}_{k+1}(t_k) & = & x_{k+1}. \end{array} \tag{5.16}$$

   provides the backward iterate $\tilde{x}_k(0) = x_{k+1} + t_k g(x_{k+1})$.

3. $\|\hat{x}_k(0) - \tilde{x}_k(0)\| = \|x_k - (x_{k+1} + t_k g(x_{k+1}))\| = t_k \|h(x_k, t_k)\| = H.$

Intuitively, the center piece of BSC strategy focuses on keeping control the distance

$$\|\hat{x}_k(t_k + t) - \hat{x}_{k+1}(t)\|,$$

provided that the backward distance

$$\|\hat{x}_k(0) - \tilde{x}_k(0)\| = t_k \|h(x_k, t_k)\| = H.$$

This argument combining with Telescopic application of the triangle inequality allow us to bounded the distance

$$\left\| x_0 \left(\sum_{i=0}^{k-1} t_i \right) - x_k \right\|,$$

for a constant that contains the factor H, and therefore, a global convergence Theorem proof bases on follow the Gauss-Newton path $\hat{x}_0(t)$ is possible.

Backward Step Control: The BSC strategy keeps control the distance between x_k and $\tilde{x}_k(0)$ using a damping factor t_k defined as follows

$$t_k = \min \mathcal{B}_H(x_k) \text{ where } B_H(x_k) := \{t \in [0,1] \mid H = t\|h(x_k,t)\|\} \cup \{1\} \tag{5.17}$$

■

The first result derived from BSC strategy is that we can guarantee the existence of a lower bound for all damping factors t_k, which depends on H. In the rest of this Chapter we are working with the Newton step $\Delta x_k = -\left[J_F(x_k)\right]^{-1} F(x_k)$.

Lemma 5.10 (Potschka [70]). *If $g(x)$ is a Lipschitz function with Lipschitz constant $L > 0$ and $x_k \in G(x_0)$ then*

$$t_k \geq \sqrt{\frac{H}{L\overline{M}\|F(x_0)\|}}$$

where

$$\overline{M} = \max_{x \in G(x_0)} \| \left[J_F(x)\right]^{-1} \|.$$

■

In the following, we introduce reasonable hypotheses that allow a global convergence proof based on following the Newton path $\hat{x}_0(t)$, and later we present a Theorem that contains the principal properties of this strategy. We omit the proof of any result in this Section because we study all of them in more detail in the next Chapter.

Hypotheses:

(H1) $G(x_0)$ is compact and connected.

(H2) There is a Lipschitz constant $L < +\infty$ such that

$$\|g(x) - g(y)\| \leq L\|x - y\| \text{ for all } x, y \in G(x_0).$$

(H3) There is a constant $\omega < +\infty$ such that

$$\| \left[J_F(x) - J_F(x - tg(x))\right] \left[J_F(x)\right]^{-1} \| \leq \omega t \|g(x)\|$$

for all $x \in G(x_0)$ and $t \in [0, 1]$.

(H4) For all $\Delta > 0$ there are constants $\gamma, t_\gamma > 0$ such that

$$\|g(x - tg(x)) - g(x)\| \geq \gamma t$$

for all $x \in G(x_0)$ (see equation (5.4)) such that $\|g(x)\| > \Delta$ and $t \in [0, t_\gamma]$.

The following Theorem resumes the principal properties derived from the above hypotheses and the BSC strategy.

Theorem 5.11 (Potschka [70]). *Let $(H1)$, $(H2)$, $(H3)$ and $(H4)$ hold. If F is continuously differentiable with $J_F(x)$ invertible for all $x \in G(x_0)$ then the principal properties of the BSC strategy are*

*(i) Given an $H > 0$ **there is a positive lower step size bound** for the sequence defined by $t_k = \min \mathcal{B}_H(x_k)$.*

(ii) *For given* $\theta, \bar{t} \in (0,1)$ *there is an* $\overline{H} > 0$ *such that for all* $H \in (0, \overline{H}]$ *the following is valid*

$$\min \mathcal{B}_H(x) \leq \bar{t}$$

for all $x \in G(x_0)$ *with* $\omega \|F(x)\| \geq 2\theta(1-\kappa)$.

(iii) *For a given* $\theta \in (0,1)$ *there is an* $\overline{H} > 0$ *such that for all* $H \in (0, \overline{H}]$, *we obtain*

$$\omega \min \mathcal{B}_H(x_k) \|g(x_k)\| \leq 2\theta$$

if $x_k \in G(x_0)$.

(iv) ***Linear Contravarinat Theorem*** *There is an* $\overline{H} > 0$ *such that for all* $H \in (0, \overline{H}]$ *the residual sequence* $(F(x_k))$ *converge linear to zero. Furthermore, if there is* $\theta \in (0,1)$ *such that*

$$\omega \min \mathcal{B}_H(x_k) \|g(x_k)\| \leq 2\theta \text{ for all } k$$

then

$$\|F(x_{k+1})\| \leq \theta \|F(x_k)\|.$$

(v) ***Existence of a-priori estimate*** *using the same* H *of the above item, there is a constant* $c > 0$ H*-independent such that*

$$\sqrt{\|F(x_k)\|} \leq \sqrt{\|F(x_0)\|} - kc\sqrt{H} \text{ for all } k \leq \sqrt{\frac{\|F(x_0)\|}{c^2 H}}.$$

(vi) ***Linear covariant Theorem*** *There is an* $\overline{H} > 0$ *such that for all* $H \in (0, \overline{H}]$ *the sequence* (x_k) *with damping factor* $t_k = \min \mathcal{B}_H(x_k)$ *converges to* $x_* := \lim_{t \to \infty} \hat{x}_0(t)$.

■

Full step in the vicinity of the solution: A drawback of the damped Newton methods based on residual monotonocity is that we derive a damping factor that is tiny even in the vicinity of our solution x_* where full step Newton sequence converges. Nevertheless, the above strategy ensures not only linear convergence of our residual sequence $(F(x_k))$, but also ensures convergence of our sequence (x_k) to the solution x_*. Furthermore, full Newton step of our sequence is guaranteed in a vicinity of the solution x_*. Indeed, if $\|F(x_k)\| < H/(L\overline{M})$ then it follows from the hypotheses $(H1)$ and $(H2)$ that

$$t\|h(x_k,t)\| \leq Lt^2 \|g(x_k)\| \leq L\overline{M}\|F(x_k)\| < H.$$

Therefore, $t_k = B_H(x_k) = 1$.

Chapter 6

Global Inexact Gauss-Newton Methods

In this Chapter, we are focus on introducing a damped IGN strategy that globalizes our local IGN approach introduced in Chapter 2 for numerically solving the nonlinear least squares problem (1.1). The functions $f(x)$ and $F(x)$ are defined in (1.1) and (1.2) respectively, and $T(x) = \frac{1}{2}\|f(x)\|_2^2$. Furthermore, $J_f(x)$ is the Jacobian of $f(x)$, and

$$F(x) = \nabla T(x), \quad J_F(x) = \left[J_f^T(x)J_f(x)\right] + Q_\epsilon(x)$$

where $Q_\epsilon(x)$ is defined in (1.3). We assume that $(S1)$ is valid. Using the Backward Step Control theory of Potschka [70], we introduce an ineaxct Gauss-Newton path $x(t)$ that connect our initial guess $x_0 \in D$ with a local solution x_* of the nonlinear least squares problem (1.1) that satisfies $(S1)$ and along it, the residual level function $T_F(x) = \frac{1}{2}\|F(x)\|_2^2$ decreases exponentially. Furthermore, using a backward analysis argument based on following the above path $x(t)$, we provide a class of damped IGN-type sequences that converge to x_*. The classical Gauss-Newton path $x^*(t)$ Deuflhard [27, Theorem 4.11], or the classical IGN paths $\tilde{x}(t)$ Deuflhard [27, Theorem 4.12] are not convenient for our purpose because $x^*(t)$ depends on the unknown local solution x_* of $F(x) = 0$ and $\tilde{x}(t)$ does not connect x_0 with x_*.

The damped IGN-type strategy with damping factor t_k for numerically finding a root of $F(x)$ assumes that there is a function $M : V_{\kappa_{GN}} \subset \mathbb{R}^n \to \mathrm{GL}(n)$, and it is defined as follows

$$x_{k+1} = x_k + t_k \delta x_k, \text{ with } \delta x_k = -M(x)F(x) \text{ and } t_k \in (0, 1]. \tag{6.1}$$

where $x_0 \in V_{\kappa_{GN}}$ is pre-chosen, $V_{\kappa_{GN}}$ is defined in (1.12), and $M(x)$ can be interpreted as an approximation of $\left[J_f^T(x)J_f(x)\right]^{-1}$. We organize this Chapter as follows: We introduce reasonable hypotheses that allow to conclude a global inexact Gauss-Newton Theorem for the damped IGN-type sequence (6.1) with Backward Step Control (BSC) damping factor. We prove the existence of an inexact Gauss-Newton path $x(t)$, which connects x_0 with x_*. Later, we introduce the definition of our inexact Gauss Newton Backward Step Control damping factor t_k that defines our IGN-BSC globalization strategy. Consequently, we prove that the above IGN-BSC approach converges to x_*. We finalize this section explaining some details of the algorithm realization, and why this strategy is adequate for a globalization of our local IGN approach introduced in Chapter 2.

Hypotheses: Let us assume that

$$G(x_0) := \{x \in V_{\kappa_{GN}} \mid T_F(x) \leq T_F(x_0)\}$$

is compact and connected, and the following are valid,

$(A1)$ There is a constant $\kappa \in [\kappa_{GN}, 1)$ such that

$$\| [I - J_F(x)M(x)] F(x)\| \leq \kappa \|F(x)\| \text{ for all } x \in G(x_0).$$

$(A2)$ The function $g : G(x_0) \subset \mathbb{R}^n \to \mathbb{R}^n$ defines as $g(x) = M(x)F(x)$ is Lipschitz with Lipschitz constant L.

$(A3)$ There is a constant $\omega < +\infty$ such that

$$\| [J_F(x) - J_F(x - tg(x))] g(x)\| \leq \omega t \|g(x)\| \|F(x)\|$$

for all $x \in G(x_0)$ and $t \in [0, 1]$.

$(A4)$ Defining $h(x, t) := g(x - tg(x)) - g(x)$, we assume that for all $\Delta > 0$ there are constants $\gamma, t_\gamma > 0$ such that

$$\|h(x, t)\| \geq \gamma t \tag{6.2}$$

for all $x \in G(x_0)$ such that $\|g(x)\| > \Delta$ and $t \in [0, t_\gamma]$.

Remark 6.1. *Note that the GN method satisfies hypotheses:* $(A1)$ *with* $\kappa = \kappa_{GN}$, $(A2)$, *and* $(A3)$ *since in this case*

$$M(x) = \left[J_f^T(x)J_f(x)\right]^{-1},$$

which implies

$$I - J_F(x)M(x) = I - \left[\left[J_f^T(x)J_f(x)\right] + Q(x)\right] \left[J_f^T(x)J_f(x)\right]^{-1} = -Q(x) \left[J_f^T(x)J_f(x)\right]^{-1},$$

and from hypothesis $(S1)$ *it follows that*

$$\| [I - J_F(x)M(x)] F(x)\| \leq \kappa_{GN} \|F(x)\|$$

for all $x \in G(x_0)$. *Furthermore,* $M(x) = \left[J_f^T(x)J_f(x)\right]^{-1}$ *satisfies also* $(A2)$ *and* $(A3)$ *since* $J_F(x)$ *and* $g(x)$ *are continuous and* $G(x_0)$ *is compact.*

The condition $(A4)$ *says that if* $\dfrac{\partial g}{\partial x}(x)g(x)$ *exists, it must be bounded away from zero, i.e.,*

$$\left\| \frac{\partial g}{\partial x}(x)g(x) \right\| \geq \gamma \text{ for all } x \in G(x_0) \text{ with } \|g(x)\| > \Delta,$$

which is a relevant condition that excludes pathological examples from our analysis. We conclude from this remark that $M(x) = \left[J_f^T(x)J_f(x)\right]^{-1}$ *satisfies* $(A1)$, $(A2)$, *and* $(A3)$, *but* $(A4)$ *must be assumed.*

The following Theorem proves that it is possible to define an IGN path $x(t)$ that connect our initial guess x_0 with x_*.

Theorem 6.2 (The Inexact Gauss-Newton path)**.** *Let* $(A1)$*, and* $(A2)$ *hold. Then the following differential equation*

$$\begin{aligned} \dot{x} &= -g(x), \\ x(0) &= x_0 \end{aligned} \tag{6.3}$$

defines a path $x : [0, \infty) \to \mathbb{R}^n$ *such that*

$$x_* := \lim_{t\to\infty} x(t) \text{ and } F(x_*) = 0.$$

Furthermore,

$$T_F(x(t)) \leq T_F(x_0)e^{-(1-k)t} \text{ for all } t \in [0, \infty).$$

Proof. Potschka [70, see Lemma 5.5 and Theorem 5.6] .

■

6.1 Inexact Gauss-Newton Backward Step Control (IGN-BSC)

Let us define our damped IGN sequence as follows

$$x_{k+1} = x_k - t_k g(x_k) \tag{6.4}$$

with inexact Gauss-Newton BSC damping factor

$$t_k := \min \mathcal{B}_H(x_k) \text{ where } \mathcal{B}_H(x_k) := \{t \in [0, 1] \mid H = t\|h(x_k, t)\|\} \cup \{1\}\,. \tag{6.5}$$

The geometrical interpretation of our IGN damping factor follows in analogous to the BSC for Newton method introduced in the Chapter 5, but here we take into account the IGN path $x(t)$ defined in Theorem 6.2 instead of the Newton path.

■

The first result, which is given by the following Lemma, derived from this IGN-BSC strategy is the existence of a lower bound for all damping factors t_k.

Lemma 6.3. *Let* $(A1)$*, and* $(A4)$ *hold and let* H *be a positive constant. If* $t_k := \min \mathcal{B}_H(x_k)$ *for all* x_k*, then*

$$t_k \geq \sqrt{\frac{H}{L\overline{M}\|F(x_k)\|}} \geq \sqrt{\frac{H}{L\overline{M}\|F(x_0)\|}} =: c_H$$

where,

$$\overline{M} := \max_{x\in G(x_0)} \|M(x)\|.$$

Proof. Potschka [70].

■

Lemma 6.4. *Let (A1), and (A3) hold. If $x_k \in G(x_0)$ then*

$$\|F(x_{k+1})\| \leq \left[1 - (1-\kappa)t_k + \frac{\omega}{2}t_k^2\|g(x_k)\|\right] \|F(x_k)\|.$$

Furthermore, if there is an $\theta \in (0,1)$ such that the damping factor t_k satisfies

$$\omega t_k \|g(x_k)\| \leq 2\theta(1-\kappa)$$

then $x_{k+1} \in G(x_0)$ and

$$\|F(x_{k+1})\| \leq [1 - (1-\theta)(1-\kappa)t_k] \|F(x_k)\|. \tag{6.6}$$

Proof. Potschka [70, see Lemma 6.2].

■

The following result says that we can control how large our IGN damping factor $\min \mathcal{B}_H(x)$ is when $x \in G(x_0)$ is far away from a local solution of (1.2).

Lemma 6.5. *Let (A1), (A2), (A4). If $\theta, \bar{t} \in (0,1)$, there is an $\overline{H} > 0$ such that for all $H \in (0, \overline{H}]$ the following is valid*

$$\min \mathcal{B}_H(x) \leq \bar{t}$$

where $x \in G(x_0)$ and $\omega\|g(x)\| \geq 2\theta(1-\kappa)$.

Proof. Potschka [70, see Lemma 8.2].

■

Theorem 6.6 (BSC Potschka [70])**.** *Let (A1), (A2), and (A4) hold. For a given $\theta \in (0,1)$, there is an $\overline{H} > 0$ such that for every $H \in (0, \overline{H}]$ our inexact Gauss-Newton BSC sequence (x_k) defined in (6.5) has the following properties,*

(i) Our inexact Gauss-Newton BSC damping factor $t_k = \min \mathcal{B}_H(x_k)$ satisfies

$$\omega\|g(x_k)\|t_k \leq 2\theta(1-\kappa) \text{ for all } k \in \mathbb{N}. \tag{6.7}$$

(ii) Our residual IGN damped sequence $(F(x_k))$ converges to zero, (x_k) converges to a local solution of (1.1), and $(x_k) \subset G(x_0)$.

(iii) Let us define

$$\hat{c} = \frac{1}{2}\frac{(1-\theta)(1-\kappa)}{\sqrt{\overline{M}L}},$$

then the first iterates of our sequence (x_k) satisfy that

$$\sqrt{\|F(x_k)\|} \leq \sqrt{\|F(x_0)\|} - k\hat{c}\sqrt{H} \text{ for all } k \leq \sqrt{\frac{\|F(x_0)\|}{\hat{c}^2 H}}.$$

Proof. Let us define

$$\bar{t} := \frac{2\theta(1-\kappa)}{\omega\overline{M}\|F(x_0)\|}$$

where $\overline{M}$ is defined in Lemma 6.3. In the first step of this proof we focus on showing that (6.7) is valid and therefore we prove that the sequences $(F(x_k))$ and (x_k) are convergent. We have two case:

- If $\omega\|g(x_0)\| \leq 2\theta(1-\kappa)$ then we can choose an arbitrary positive constant $\overline{H}$.
- Otherwise, Lemma 6.5 delivers a constant $\overline{H}$ such that

$$\omega\|g(x_0)\| \min \mathcal{B}_H(x_0) \leq 2\theta(1-\kappa) \quad \text{for all} \quad H \in (0, \overline{H}].$$

Let us fix $H \in (0, \overline{H}]$ and assume by induction that (6.7) is valid for $x_k \in G(x_0)$ then from Lemma 6.4 it follows,

$$\|F(x_{k+1})\| \leq [1 - t_k(1-\theta)(1-\kappa)] \, \|F(x_k)\| < \|F(x_0)\| \text{ with } t_k = \min \mathcal{B}_H(x_k),$$

i.e., $x_{k+1} \in G(x_0)$.

If $t_{k+1} := \min \mathcal{B}_H(x_{k+1}) \leq \bar{t}$, then

$$\omega\|g(x_{k+1})\|t_{k+1} \leq \omega\overline{M}\|F(x_{k+1})\|t_{k+1} < \omega\overline{M}\|F(x_0)\|t_{k+1} \leq 2\theta(1-\kappa).$$

If $\bar{t} < t_{k+1}$, then $\bar{t} \leq 1$, which implies $\omega\overline{M}\|F(x_0)\| \leq 2\theta(1-\kappa)$ and from here, we conclude

$$\omega\|g(x_{k+1})\|t_{k+1} \leq \omega\overline{M}\|F(x_{k+1})\| < \omega\overline{M}\|F(x_0)\| \leq 2\theta(1-\kappa).$$

Thus, we prove that (6.7) is valid for all x_k. In the following line we prove that $(F(x_k))$ converges to zero. In fact, from Lemma 6.4, we have

$$\|F(x_{k+1})\| \leq [1 - t_k(1-\theta)(1-\kappa)] \, \|F(x_k)\| \text{ for all } k,$$

and using that our damping factors are lower bounded by $c_H < 1$ defined in Lemma 6.3, we conclude

$$\|F(x_{k+1})\| \leq [1 - c_H(1-\theta)(1-\kappa)]^k \, \|F(x_0)\|,$$

which implies that $(F(x_k))$ converges to zero. Let us define

$$c := [1 - c_H(1-\theta)(1-\kappa)] < 1$$

and let us show that (x_k) is a Cauchy sequence.

$$\|x_{k+m} - x_k\| \leq \sum_{i=k}^{k+m-1} \|M(x_i)F(x_i)\| \leq \overline{M}\|F(x_0)\| \sum_{i=k}^{m-1} c^k c^i \leq \frac{\overline{M}\|F(x_0)\|c^k}{1-c},$$

which prove that (x_k) is a Cauchy sequence, therefore it converges to $x_* \in G(x_0)$.

The proof of part (iii) follows in analogous to Potschka [70, Theorem 8.4].

■

Up until now, we have proved that our damped IGN-BSC sequence (x_k) converges to a local solution x_* of (1.2). Nevertheless, we cannot predict how close x_* is to x_0. The following Theorem says that x_* and x_0 are connected by the IGN path defined in (6.2).

Lemma 6.7. *Let (A1), (A2), (A3), and (A4) hold. Then $t_k = \min \mathcal{B}_H(x_k)$ satisfies*

$$\|x_k(t_k) - x_{k+1}\| \leq \frac{t_k^2}{2} L\|g(x_k)\|e^{Lt_k}, \tag{6.8}$$

and

$$\|x_k(t_k + t) - x_{k+1}(t)\| \leq \frac{t_k^2}{2} L\|g(x_k)\|e^{L(t_k+t)} \text{ for all } t \geq 0. \tag{6.9}$$

Proof. Potschka [70, Lemma 7.1 and Lemma 8.5]

■

Theorem 6.8 (IGN-BSC covariant result, Potschka [70]). *Let (A1), (A2), (A3), and (A4) hold. Then there is an $\overline{H} > 0$ such that for every $H \in (0, \overline{H}]$ our IGN-BSC sequence (x_k) converges to $x_* := \lim_{t\to\infty} x(t)$ where $x(t)$ is the IGN path emanating from x_0.*

Proof. Let us assume $\|F(x_0)\| > 0$. Theorem 6.2 ensure the existence of a solution x_* of our equation $F(x) = 0$ such that x_0 and x_* are connected by the IGN path $x(t)$. Because $J_F(x)$ is invertible for all $x \in G(x_0)$, we have that x_* is an isolate point, therefore, there is a $\epsilon > 0$ such that x_* is the unique root of F or unique equilibrium point of the Davidenko equation (6.3) in $B(x_*, r)$ with

$$r = \epsilon \left[1 + \frac{\overline{M}L}{1-\kappa}\right]$$

where $\overline{M}$ is defined in Lemma 6.3. Let us fix $\theta \in (0,1)$ and assume without loss of generality that ϵ is sufficient small to satisfy

$$\omega L \overline{M} \epsilon \leq 2\theta(1-\kappa). \tag{6.10}$$

Because $x_* = \lim_{t\to\infty} x(t)$, we can choose T_* such that

$$\|x(t) - x_*\| \leq \frac{\epsilon}{2} \text{ for all } t \geq T_* - 1. \tag{6.11}$$

Note that $\|g(x(t))\| > 0$ for all $t \in [0, T_*]$. Otherwise, it follows that $g(x_0) = 0$, which implies $F(x_0) = 0$ contrary to our assumption. Thus, we can choose $\tilde{\epsilon} \in (0, \epsilon)$ that implies the existence of a constant $\Delta_{\tilde{\epsilon}} > 0$ such that

$$\Delta_{\tilde{\epsilon}} \leq \|g(x)\| \text{ for all } t \in [0, T_*], x \in G(x_0) \text{ with } \|x(t) - x\| \leq \tilde{\epsilon}. \tag{6.12}$$

Hypothesis (A4) yields the existence of constants γ, t_γ that satisfies (6.2). It follows from Lemma 6.5 that there is a constant $\overline{H}$ such that

$$t_k = \min \mathcal{B}_H(x_k) \leq \bar{t} = \min\left\{t_\gamma, \frac{2\theta(1-\kappa)}{\omega \overline{M}}\right\}$$

for all $H \in (0, \overline{H}]$. Let us assume without loss of generality (we can decrease $\overline{H}$, it is necessary) that

$$0 < \overline{H} \leq \min\left\{L\overline{M}\|F(x_0)\|, \frac{[\gamma \tilde{e}^2]}{[T_* e^{LT_*}]^2 [L\overline{M}\|F(x_0)\|]^3}\right\}. \tag{6.13}$$

Basically, the proof consists in two step:

(*i*) we prove that there is a $\overline{k}$ such that $x_{\overline{k}} \in B(x_*, \epsilon)$.

(*ii*) we prove that there is a $\tilde{k} \geq \overline{k}$ such that $(x_k)_{k\geq\tilde{k}} \subset B(x_*, r)$ and (x_k) converges to x_*.

(i) Let us fix $H \in (0, \overline{H}]$. By virtue of Lemma 6.3, we can choose $\overline{k}$ depend on H such that

$$T_* - 1 \leq \sum_{i=0}^{\overline{k}-1} t_i \leq T_*. \tag{6.14}$$

Using once more Lemma 6.3, and the above equation, it follows

$$\overline{k}\sqrt{\frac{H}{L\overline{M}\|F(x_0)\|}} \leq \sum_{i=0}^{\overline{k}-1} t_i \leq T_*, \tag{6.15}$$

therefore

$$\overline{k}\sqrt{\frac{H}{L\overline{M}\|F(x_0)\|}} \leq T_*,$$

which implies

$$\overline{k} \leq T_*\sqrt{\frac{L\overline{M}\|F(x_0)\|}{H}}. \tag{6.16}$$

In the following argument, it is proved that

$$\left\| x\left(\sum_{i=0}^{\overline{k}-1} t_i\right) - x_{\overline{k}} \right\| \leq \frac{\tilde{\epsilon}}{2} \tag{6.17}$$

Telescopic application of the triangle inequality and (6.9) yields,

$$\begin{aligned} \left\| x\left(\sum_{i=0}^{\overline{k}-1} t_i\right) - x_{\overline{k}} \right\| &\leq \sum_{i=0}^{\overline{k}-1} \left\| x_i\left(\sum_{j=i}^{\overline{k}-1} t_j\right) - x_{i+1}\left(\sum_{j=i+1}^{\overline{k}-1} t_j\right) \right\| \\ &\leq \sum_{i=0}^{\overline{k}-1} \frac{t_i^2}{2} L\|g(x_i)\|e^{LT_*}, \end{aligned} \tag{6.18}$$

since $t_i := \min \mathcal{B}(x_i) \leq t_\gamma$ for all $i \leq \overline{k}$, we conclude from (6.2) and definition of t_i that

$$\gamma t_i^2 \leq t_i\|h(t_i, x_i)\| = H. \tag{6.19}$$

Thereby,

$$t_i^2 \leq \frac{H}{\gamma}. \tag{6.20}$$

Subtitling (6.20) in (6.18), we obtain

$$\begin{aligned} \left\| x\left(\sum_{i=0}^{\overline{k}-1} t_i\right) - x_{\overline{k}} \right\| \leq & \sum_{i=0}^{\overline{k}-1} \frac{H}{2\gamma} L\|g(x_i)\|e^{LT_*} \\ = & \ \overline{k}\frac{H}{2\gamma} L\overline{M}\|F(x_0)\|e^{LT_*} \\ \overset{(6.16)}{\leq} & \ \frac{T_* e^{LT_*}[L\overline{M}\|F(x_0)\|]^{3/2}}{2\gamma}\sqrt{H} \leq \frac{\tilde{\epsilon}}{2}. \end{aligned}$$

Thus,

$$\|x_{\overline{k}} - x_*\| \leq \left\| x\left(\sum_{i=0}^{\overline{k}-1} t_i\right) - x_k \right\| + \left\| x\left(\sum_{i=0}^{\overline{k}-1} t_i\right) - x_* \right\| \leq \frac{\tilde{\epsilon}}{2} + \frac{\epsilon}{2} = \epsilon.$$

(*ii*) From Theorem 6.6, it follows that (x_k) converges to $\overline{x}_*$ and $F(x_k)$ converges to zero. Using an argument analogous to Chapter 5 (page 65), we can guarantee that there is a vicinity $B\left(0, \frac{H}{L\overline{M}}\right)$ and $\tilde{k} \geq \overline{k}$ such that $F(x_k) \in B\left(0, \frac{H}{L\overline{M}}\right)$ and $t_k = 1$ for all $k \geq \tilde{k}$. Thus, if $k \geq \tilde{k}$ and defining $c = 1 - (1-\theta)(1-\kappa) < 1$, we obtain

$$\|x_k - x_*\| \leq \|x_{\overline{k}} - x_*\| + \sum_{i=\overline{k}}^{k-1} \|x_{i+1} - x_i\| \leq \epsilon + \frac{\overline{M}\|F(x_{\overline{k}})\|}{1-c} \leq \epsilon + \frac{\overline{M}L}{(1-\theta)(1-\kappa)} < \epsilon\left[1 + \frac{\overline{M}L}{1-\kappa}\right],$$

which implies $(x_k)_{k\geq\tilde{k}} \subset B(\epsilon, r)$, and because x_* is the only solution of $F(x) = 0$ in $B(x_*, r)$ it follows that $\overline{x}_* = x_*$.

■

Remark 6.9. *In contrast to Lemma 5.1 and Lemma 5.4 presented in Chapter 5, we can conclude that IGN-BSC is an essentially affine covariant strategy since*

$$t_k = \min\left(\left\{t \in [0,1] \mid \overline{H} = t\|g(x_k + t\delta x_k) - g(x_k)\| = t\|h(x_k, t)\|\right\} \cup \{1\}\right),$$

which means that $t\|h(x_k,t)\|$ is affine covariant because $(A2)$ and $(A4)$ are covariant properties, and $\overline{H}$ given in the above Theorem is a constant that depends on $(A1)$ (contravariant property) and $(A2)$. Thus, the damped factor t_k in IGN-BSC startegy does not depend on transformations on the images spaces of $F(x)$.

Algorithm Realization

In the following, we explain how to compute at every iteration x_k the damping factor $t_k = \min \mathcal{B}_H(x_k)$ where $t_k = 1$ or t_k is the smallest solution of

$$t_k\|h(x_k, t_k)\| = H.$$

As described in [70], a rigoruos approach is based on monotone iteration [35] for which we require an overestimate L' of the Lipschitz constant L defined in $(A2)$. From [70, Lemma 10.1], it follows that

$$[t_k]_{j+1} = \sqrt{[t_k]_j^2 + \frac{H - [t_k]_j\|h(x_k, [t_k]_j)\|}{L'\|g(x_k)\|}}.$$

Furthermore, from [35], we conclude $([t_k]_j)$ is monotonically increasing and either converge in $[0,1]$ or leaves the interval in a finite number of step. Because finding an adequate estimate L' of L is not of all an easy task, this rigorous approach is not implemented. Instead, Potschka [70, Secction 10.2] implements a simple root finding procedure for approximately solving $H = t\|h(x_k,t)\|$ using a bracketing procedure with exponentially smoothed step size prediction. Basically, we find a t_k that satisfies

$$\begin{aligned} &t_k\|h(x_k,t_k)\| \in [H^l, H^u] \text{ where } H^l < H \text{ and } H < H^u \text{ are close to } H, \text{ or} \\ &t_k = 1 \text{ if } t\|h(x_k,t)\| < H \text{ for all } t \in [0,1] \end{aligned} \tag{6.21}$$

with step size prediction, which is proposed by [70, Section 10.2], given by

$$[t_k]_0 = \min\left(1, t_{k-1}\left[\alpha + (1-\alpha)\frac{H}{t_{k-1}\|h(x_k, t_{k-1})\|}\right]\right)$$

where $\alpha \in [0,1]$ is the smoothing factor. The advantage of working with this strategy is that often the step size prediction satisfies (6.21) and thus $t_k = [t_k]_0$. It is not the case then with a few number of iteration and thus almost no extra computational effort in term of the residual evaluation $F(x_k + [t_k]_i \delta x_k)$ and the increment evaluation $g(x_k + [t_k]_i \delta x_k)$ we obtain a damping IGN-BSC factor $t_k = [t_k]_i$ that satisfies (6.21).

Discussion: In the following, we justify why this IGN-BSC strategy is adequate for a globalization of the local IGN approach introduced in Chapter 2. We have proved in Chapter 3 that there is a matrix $M_\kappa(x)$ such that $g_\kappa(x) = M_\kappa(x)F(x)$, and $\delta x_k^{IGN} = -g_\kappa(x_k^{IGN})$ (see Lemma 3.5). Nevertheless, we cannot ensure that $g_\kappa(x)$ satisfies $(A2)$ since δx_k^{IGN} depends on the number of inner iteration m necessary for ensuring that stopping criterion (2.15) is valid, which may chance from one k-iteration to another. However, a small modification $\delta\tilde{x}_k$ of the IGN step δx_k defined in $(S3)$ allows to conclude that $(A2)$ is valid. Let us define the following IGN step

$$\delta\tilde{x}_k := (1-\alpha_k)\,[\delta x_k]^{m-1} + \alpha_k\,[\delta x_k]^m\,,$$

where $[\delta x_k]^m$ solves (2.14) via LSQR or LSMR with stopping criterion (2.15) but $[\delta x_k]^{m-1}$ does not fulfill (2.15), and α_k is the smaller value in $[0,1]$ such that $\varphi(\alpha) = 0$ where

$$\varphi(\alpha) := \|J_f^T(x_k)\,[f(x_k) + J_f(x_k)\delta\tilde{x}_k]\,\| \\ - \kappa\|J_f^T(x_k)f(x_k)\| + \kappa_{GN}\left\|\left[J_f^T(x)J_f(x)\right]\delta\tilde{x}_k\right\|,$$

and the existence of such a value $\alpha_k \in [0,1]$ is guaranteed by virtud of the vale intermediate Theorem since $\varphi(1) < 0$, $\varphi(0) > 0$ and $\varphi(\alpha)$ is continuous. It turns out that defining $\tilde{g}(x_k) = -\delta\tilde{x}_k$, we obtain that the hypotheses $(A1)$, $(A2)$, and $(A3)$ are valid for $g(x) := \tilde{g}(x)$. Therefore, if $(A4)$ is valid for the above $\tilde{g}(x)$, all the results of this Chapter are valid, which means that

$$x_* := \lim_{t\to\infty} x(t) = \lim_{k\to\infty} \tilde{x}_k$$

where $x(t)$ is the IGN path emanating from x_0, and $\tilde{x}_{k+1} = \tilde{x}_k + \min \mathcal{B}_H(\tilde{x}_k)\delta\tilde{x}_k$.

Chapter 7

Applications and numerical results

7.1 Parameter Identification of nonlinear steady-state diffusion equation

Given a measuring data function z, we are interested in identifying the unknown coefficient $c(x,y) \in \mathbb{R}$ of the nonlinear steady-state diffusion equation

$$-\nabla \cdot (c(x,y)\nabla u(x,y)) = \mathrm{f}(x,y), \forall (x,y) \in \Omega \subseteq \mathbb{R}^2, \tag{7.1}$$

where $\Omega := B((1,1);1)$, $c(x,y) \in H^1(\Omega)$, $u(x,y) \in H_0^1(\Omega)$ and with Dirichlet condition: $u(x,y) = 0, \forall (x,y) \in \partial\Omega$.

The above partial differential equation (7.1) may describe the flow of a fluid (e.g., groundwater) through some medium with permeability $c(x,y)$.

Statement: If the above problem is formulated as a constrained optimization problem using the output least squares methods with a particular $H^1(\Omega)$ regularization and a penalty term, then this particular problem is well posed in the sense of Hadamard, i.e.;

1. The problem has at least one solution.
2. The solution is locally unique,
3. The solution depends continuously on the data.

In this section, we explain briefly the result given by Jun Zou [84], which ensures that the above statement is valid. Later, we use a finite element method to discretize the above problem, which yields a sequence of unconstrained minimization problems. Defining the following set

$$\mathcal{K} := \left\{ c(x,y) \in L^1(\Omega) \mid \|c(x,y)\|_{H^1(\Omega)} < \infty \text{ and } \alpha_1 \leq c(x,y) \leq \alpha_2, \text{ a.e.} \right\}, \tag{7.2}$$

we formulate our parameter identification problem as the following constrained minimization problem introduced by Jun Zou [84],

$$\underset{c\in\mathcal{K}, u\in H_0^1(\Omega)}{\text{minimize}} J(c) = \frac{1}{2}\int_\Omega c\|\nabla u - \nabla z\|_2^2 d(x,y) + \frac{\gamma}{2}\|u-z\|_{H_0^1(\Omega)}^2 + \frac{\epsilon}{2}\int_\Omega \mathcal{P}_c^- d(x,y) \tag{7.3a}$$

$$\text{subject to } \int_\Omega c\nabla u\cdot\nabla\varphi d(x,y)=\int_\Omega \mathrm{f}\varphi d(x,y) \text{ for all } \varphi\in H_0^1(\Omega). \tag{7.3b}$$

where the function $z(x,y)\in H_0^1(\Omega)$ is the measured data, $\gamma>0$ is a regulation weights, and $\epsilon>0$ is a penalty parameter with,

$$\mathcal{P}_c^-(x,y)=[c(x,y)-\alpha_1]_-^2+[\alpha_2-c(x,y)]_-^2,\quad \text{where } [y]_-=\max\{-y,0\}.$$

Discretization

In the following, the problem (7.3) is discretized using a piecewise linear finite element method and then the constrained finite element minimization problem is reduced to a sequence of unconstrained minimization subproblems. Let us consider a triangulation $\mathcal{T}_h=\{\mathcal{P},\mathcal{E},\mathcal{T},\mathcal{V}_h\}$ of Ω, where

$\mathcal{P}$ denotes the set of all nodal points of the triangulation, i.e.,
$\mathcal{P}=\{e_1,e_2,\cdots,e_{n_\mathcal{E}},v_1,v_2,\cdots,v_{n_\mathcal{P}}\}$.

$\mathcal{E}$ denotes the set of all nodes of the triangulation that are in $\partial\Omega$, i.e.,
$\mathcal{E}=\mathcal{P}\cap\partial\Omega=\{e_1,e_2,\cdots,e_{n_\mathcal{E}}\}$.

$\mathcal{T}$ denotes the set of all triangles of our triangulation, i.e.,
$\mathcal{T}=\{\triangle_1(v_1,u_1,w_1),\triangle_2(v_2,u_2,w_2),\cdots,\triangle_{n_\mathcal{T}}(v_{n_\mathcal{T}},u_{n_\mathcal{T}},w_{n_\mathcal{T}})\}$, where $\{v_i,u_i,w_i\}\subseteq\mathcal{P}$ denotes the vertices of the triangle $\triangle_i(v_i,u_i,w_i)\in\mathcal{T}$.

$\mathcal{X}_h$ denotes the classical test function generator of $\mathcal{C}(\Omega)$, which are defined as
$\mathcal{X}_h=\left\{\chi_{\triangle_1},\chi_{\triangle_2},\cdots,\chi_{\triangle_{n_\mathcal{T}}}\right\}$ where $\chi_{\triangle_i}(x)=1$ for all $x\in\triangle_i$ and
$\chi_{\triangle_i}(x)=0$ for all $x\in\Omega\setminus\triangle_i$.

$\mathcal{M}$ denotes the set of all triangle centers of $\mathcal{T}$, i.e.,
$\mathcal{M}=\{m_{\triangle_1},m_{\triangle_2},\cdots m_{\triangle_{n_\mathcal{T}}}\}$.

$\mathcal{V}_h$ denotes the classical test function generator of $H^1(\Omega)$, i.e., $\varphi\in\mathcal{V}_h$ if and only if there is $v\in\mathcal{P}$ such that $\varphi(w)=\delta_v(w)$ for all $w\in\mathcal{P}$ and $\varphi\big|_\triangle\in\mathbb{P}_1$ for all $\triangle\in\mathcal{T}$.
$\mathcal{V}_h=\left\{\varphi_{e_1},\varphi_{e_2},\cdots,\varphi_{e_{n_\mathcal{E}}},\varphi_{v_1},\varphi_{v_2},\cdots,\varphi_{v_{n_\mathcal{P}}}\right\}$.

$\mathring{\mathcal{V}}_h$ denotes the classical test functions generator of $H_0^1(\Omega)$, i.e.,
$\mathring{\mathcal{V}}_h=\left\{\varphi_{v_1},\varphi_{v_2},\cdots,\varphi_{v_{n_\mathcal{P}}}\right\}$.

Given $v_i\in\mathcal{P}$, let us consider the set $D_{\varphi_{v_i}}$ of all triangle in $\mathcal{T}$ such that one of its vertices is v_i, i.e.,

$$D_{\varphi_{v_i}}=\left\{\triangle_{i_1}(v_i,v_{i_1},v_{i_2}),\triangle_{i_2}(v_i,v_{i_2},v_{i_3}),\cdots,\triangle_{i_{\hat{n}-1}}(v_i,v_{i_{\hat{n}-1}},v_{i_{\hat{n}_i}}),\triangle_{i_{\hat{n}_i}}(v_i,v_{i_{\hat{n}_i}},v_{i_1})\right\}.$$

With the above preparation we define the standard nodal value interpolation function u_h associated with the finite element space $\mathrm{span}(\mathring{\mathcal{V}}_h)$ as,

$$u_h=\sum_{i=1}^{n_\mathcal{P}}\overline{u}_i\varphi_{v_i}\text{ where }\overline{u}=(u_h(v_1),u_h(v_2),\cdots,u_h(v_{n_\mathcal{P}})) \tag{7.4}$$

where the derivative ∇u_h is given by

$$\nabla u_h = \sum_{i=1}^{n_{\mathcal{P}}} \overline{u}_i \nabla \varphi_{v_i} = \sum_{i=1}^{n_{\mathcal{P}}} \overline{u}_i \sum_{\triangle \in \mathring{D}_{\varphi_{v_i}}} \nabla \varphi_{v_i}\big|_{\triangle}. \tag{7.5}$$

Let $\mathcal{R}_h : H_0^1(\Omega) \to \mathring{\mathcal{V}}_h$ be the projection operator on $\mathring{\mathcal{V}}_h$ defined by

$$\mathcal{R}_h(g) = \sum_{i=1}^{n_{\mathcal{P}}} g(v_i)\varphi_{v_i},$$

and let us consider the interpolation function c_h associated with the finite element space $span(\mathcal{X}_h)$ such as,

$$c_h = \sum_{i=1}^{n_{\mathcal{T}}} \overline{c}_i \chi_{\triangle_i} \text{ where } \overline{c} = (c_h(m_{\triangle_1}), c_h(m_{\triangle_2}), \cdots, c_h(m_{\triangle_{n_{\mathcal{T}}}})). \tag{7.6}$$

Consequently, the discrete constrained subset $\mathcal{K}_h$ of $\mathcal{K}$ (see (7.2)) is defined by,

$$\mathcal{K}_h = \left\{ \sum_{i=1}^{n_{\mathcal{T}}} \overline{c}_i \chi_{\triangle_i} \mid \alpha_1 \le \overline{c}_i \le \alpha_2 \quad \text{for all } i \in \{1, 2, \cdots, n_{\mathcal{T}}\} \right\},$$

$$\mathcal{K}_h^{\overline{c}} = \{ \overline{c} \in \mathbb{R}^{n_{\mathcal{T}}} \mid \alpha_1 \le \overline{c}_i \le \alpha_2 \quad \text{for all } i \in \{1, 2, \cdots, n_{\mathcal{T}}\} \}.$$

The discretization of the minimization problem (7.3)

Defining $\mathcal{P}_{\overline{c}}^{-}(x,y) = \sum_{i=1}^{n_{\mathcal{T}}} [\overline{c}_i - \alpha_1]_{-}^2 \chi_{\triangle_i}(x,y) + [\alpha_2 - \overline{c}_i]_{-}^2 \chi_{\triangle_i}(x,y)$, and

$$J_h(\overline{c}, \overline{u}) = \frac{1}{2} \int_{\Omega} c_h \|\nabla u_h - \nabla \mathcal{R}_h(z)\|_2^2 d(x,y) + \frac{\gamma}{2} \|u_h - \mathcal{R}_h(z)\|_{H_0^1(\Omega)}^2 + \frac{\epsilon}{2} \int_{\Omega} \mathcal{P}_{\overline{c}}^{-} d(x,y),$$

the discretization of the minimization problem (7.3) is given by,

$$\underset{\overline{c} \in \mathcal{K}_h^{\overline{c}}, \overline{u} \in \mathbb{R}^{n_{\mathcal{P}}}}{\text{minimize}} \; J_h(\overline{c}, \overline{u}) \tag{7.7a}$$

$$\text{subject to } \int_{\Omega} c_h \nabla u_h \cdot \nabla \varphi d(x,y) = \int_{\Omega} \mathrm{f} \varphi d(x,y) \text{ for all } \varphi \in \mathring{\mathcal{V}}_h. \tag{7.7b}$$

Theorem 7.1 (Jun Zou [84]). *The following are valid,*

(i) There is at least one local minimizer to the optimization problem (7.3).

(ii) There is at least one minimizer to the optimization problem (7.7).

(iii) If the sequence (c_h) *in* $\mathcal{K}_h$ *converges to some* $c \in \mathcal{K}$ *as* h *tends to zero, then the sequence* (u_h) *defining by* (7.7b) *converges to* u *weakly in* $H_0^1(\Omega)$ *and*

$$\lim_{h \to 0} \int_{\Omega} c_h \|\nabla u_h - \nabla \mathcal{R}_h(z)\|_2^2 d(x,y) = \int_{\Omega} c \|\nabla u - \nabla z\|_2^2 d(x,y).$$

(iv) If the sequence (c_n) in $\mathcal{K}$ converges to some $c \in \mathcal{K}$ in $L^1(\Omega)$ as n tends to infinity, then the sequence (u_n) defining by (7.3b) converge to u weakly in $H_0^1(\Omega)$ and

$$\lim_{n\to\infty} \int_\Omega c_n \|\nabla u_n - \nabla z\|_2^2 d(x,y) = \int_\Omega c \|\nabla u - \nabla z\|_2^2 d(x,y).$$

(v) Let (c_n^*) be a sequence of local minimizers of the discrete minimization problem (7.7), then each subsequence of (c_n^*) has a subsequence converging to one local minimizer of the continuous problem (7.3).

Unconstrained Minimization Problem

Let us consider the following unconstrained optimization problem

$$\underset{\overline{c}\in\mathcal{K}_h^{\overline{c}}}{\text{minimize}}\ J_h^\epsilon(\overline{c}) := \tilde{J}_h(\overline{c}) + \frac{\epsilon}{2}\int_\Omega \mathcal{P}_{\overline{c}}^- d(x,y) \tag{7.8}$$

where the functional $\tilde{J}_h$ is defined as following

$$\tilde{J}_h(\overline{c}) = \frac{1}{2}\int_\Omega |c_h| \|\nabla u_h - \nabla \mathcal{R}_h(z)\|_2^2 d(x,y) + \frac{\gamma}{2}\|u_h - \mathcal{R}_h(z)\|_{H_0^1(\Omega)}^2$$

and u_h satisfies the following linear system of equations,

$$\int_\Omega |c_h| \nabla u_h \cdot \varphi_{v_i} d(x,y) = \int_\Omega \mathrm{f}\varphi_{v_i} d(x,y) \text{ for all } i \in \{1, 2, \cdots, n_{\mathcal{P}}\}. \tag{7.9}$$

Remark 7.2. *In (7.8), we use absolute value of c_h in the functional $\tilde{J}_h$, but in (7.7a) was not taking into account, the principal reason for this change is to ensure that u_h that satisfies (7.7b) is well defined for each c_h. If we keep the original c_h instead of $|c_h|$, u_h may be undefined, say when c_h is very close to zero or negative in some subregion.*

Theorem 7.3 (Jun Zou [84])**.** *Let (ϵ_n) be a strictly monotone increasing sequence converging to infinity, and $(\overline{c}_{\epsilon_n}^*)$ be a local minimizer of (7.8), then each subsequence of $(\overline{c}_{\epsilon_n}^*)$ has a subsequence converging to a local minimizer of (7.7).*

Nonlinear least-squares subproblem

In this part, we explain how from (7.8), we obtain a nonlinear least squares problem. First, we obtain from (7.9) a system of linear equation, whose solution defines u_h. Later, we define a vector field function $f(\overline{c})$ such that $T(\overline{c}) = \frac{1}{2}\|f(\overline{c})\|_2^2 = J_h^\epsilon(\overline{c})$.

Reducing (7.9) to a system of linear equation
from (7.4) and (7.6), it follows

$$|c_h| \nabla u_h \cdot \nabla \varphi_{v_i} = \sum_{j=1}^{n_{\mathcal{P}}} \overline{u}_j \sum_{\triangle_s \in D_{\varphi_{v_j}}} |\overline{c}_s| \nabla \varphi_{v_j}\big|_{\triangle_s} \cdot \nabla \varphi_{v_i}\big|_{\triangle_s}$$

substituting the above information in (7.9), we obtain

$$\int_\Omega |c_h| \nabla u_h \cdot \nabla \varphi_{v_i} = \sum_{j=1}^{n_{\mathcal{P}}} \overline{u}_j \sum_{\triangle_s \in D_{\varphi_{v_j}}} |\overline{c}_s| \int_{\triangle_s} \nabla \varphi_{v_j}\big|_{\triangle_s} \cdot \nabla \varphi_{v_i}\big|_{\triangle_s}.$$

Defining the matrix $K_1(\overline{c}) \in \mathbb{R}^{n_{\mathcal{P}} \times n_{\mathcal{P}}}$ and $\mathcal{F} \in \mathbb{R}^{n_{\mathcal{P}}}$ such that

$$(K_1(\overline{c}))_{ij} = \sum_{\triangle_s \in D_{\varphi_{v_j}}} |\overline{c}_s| \int_{\triangle_s} \nabla\varphi_{v_j}\big|_{\triangle_s} \cdot \nabla\varphi_{v_i}\big|_{\triangle_s},$$

and

$$\mathcal{F}_i = \int_\Omega f\varphi_{v_i} d(x,y),$$

we obtain from the definitions of $M_1(\overline{c})$, $\mathcal{F}$, and (7.9) that $\overline{u}$ solve the following system of linear equations,

$$K_1(\overline{c})\overline{u} = \mathcal{F}.$$

Writing (7.8) as a level function $T(\overline{c}) = \frac{1}{2}\|f(\overline{c})\|_2^2$. For such purpose, we work with every term of our functional (7.8) where

$$\begin{aligned}
J_1(\overline{c}) &= \frac{1}{2}\int_\Omega |c_h| \|\nabla u_h - \nabla \mathcal{R}_h(z)\|_2^2 d(x,y), \\
J_2(\overline{c}) &= \frac{\gamma}{2}\int_\Omega \|u_h - \mathcal{R}_h(z)\|_2^2 d(x,y) + \frac{\gamma}{2}\int_\Omega \|\nabla u_h - \nabla\mathcal{R}_h(z)\|_2^2 d(x,y), \\
J_3(\overline{c}) &= \frac{\epsilon}{2}\int_\Omega \sum_{i=1}^{n_{\mathcal{T}}} [\overline{c}_i - \alpha_1]_-^2 \chi_{\triangle_i} d(x,y), \text{ and} \\
J_4(\overline{c}) &= \frac{\epsilon}{2}\int_\Omega \sum_{i=1}^{n_{\mathcal{T}}} [\alpha_2 - \overline{c}_i]_-^2 \chi_{\triangle_i} d(x,y).
\end{aligned}$$

Functional J_1:

$$\nabla u_h - \nabla \mathcal{R}_h(z) = \sum_{i=0}^{n_{\mathcal{P}}} (\overline{u}_i - \overline{z}_i) \left[\sum_{\triangle_s \in D_{\varphi_{v_i}}} \nabla\varphi_{v_i}\big|_{\triangle_s} \right]$$

therefore,

$$|c_h| \left[\nabla u_h - \nabla \mathcal{R}_h(z)\right] = \sum_{i=0}^{n_{\mathcal{P}}} (\overline{u}_i - \overline{z}_i) \left[\sum_{\triangle_s \in D_{\varphi_{v_i}}} |\overline{c}_s| \nabla\varphi_{v_i}\big|_{\triangle_s} \right].$$

Defining the sparse matrix $M_1(\overline{c}) \in \mathbb{R}^{n_{\mathcal{P}} \times n_{\mathcal{P}}}$ such that

$$(M_1(\overline{c}))_{ij} = \left[\sum_{\triangle_s \in D_{\varphi_{v_i}}} |\overline{c}_s| \nabla\varphi_{v_i}\big|_{\triangle_s} \right] \cdot \left[\sum_{\triangle_s \in D_{\varphi_{v_j}}} \nabla\varphi_{v_j}\big|_{\triangle_s} \right],$$

we obtain

$$J_1(c_h) = \frac{1}{2}(\overline{u} - \overline{z})^T M_1(\overline{c})(\overline{u} - \overline{z}).$$

Let us assume that $|c_s| > 0$ for all $s \in \{1, 2, \cdots, n_{\mathcal{P}}\}$ then by construction $M_1(\overline{c})$ is symmetric and positive define, which means that we can apply the Cholesky factorization,

i.e., there is a lower triangular matrix $L(\overline{c}) \in \mathbb{R}^{n_{\mathcal{P}} \times n_{\mathcal{P}}}$ such that $L(\overline{c})^T L(\overline{c}) = M_1(\overline{c})$. Defining $f_1(\overline{c}) = L(\overline{c})[\overline{u} - \overline{z}]$, then

$$\frac{1}{2}\|f_1(\overline{c})\|_2^2 = J_1(c_h).$$

Functional J_2: Let us define the mass matrix $M \in \mathbb{R}^{n_{\mathcal{P}} \times n_{\mathcal{P}}}$ and the Stiffness $S \in \mathbb{R}^{n_{\mathcal{P}} \times n_{\mathcal{P}}}$ of triangulation $\mathcal{T}$ as follow

$$M_{ij} = \int_\Omega \varphi_{v_i}\varphi_{v_j} \text{ and } S_{ij} = \int_\Omega \nabla\varphi_{v_i} \cdot \nabla\varphi_{v_j}.$$

Because the above matrices are positive defined, we apply the Cholesky factorization to both of them, therefore there are lower triangular matrices $L_M \in \mathbb{R}^{n_{\mathcal{P}} \times n_{\mathcal{P}}}$ and $L_S \in \mathbb{R}^{n_{\mathcal{P}} \times n_{\mathcal{P}}}$ such that $L_M^T L_M = M$ and $L_S^T L_S = S$. Since

$$u_h - \mathcal{R}_h(z) = \sum_{i=0}^{n_{\mathcal{P}}} (\overline{u}_i - \overline{z}_i)\varphi_{v_i}$$

and

$$\nabla u_h - \nabla\mathcal{R}_h(z) = \sum_{i=0}^{n_{\mathcal{P}}} (\overline{u}_i - \overline{z}_i) \left[\sum_{\triangle_s \in D_{\varphi_{v_i}}} \nabla\varphi_{v_i}\big|_{\triangle_s} \right],$$

then we conclude

$$\|\sqrt{\gamma} L_M[\overline{u} - \overline{z}]\|_2^2 + \|\sqrt{\gamma} L_S[\overline{u} - \overline{z}]\|_2^2 = J_2.$$

Defining $f_2(\overline{c}) = \sqrt{\gamma}\left[L_M^T, L_S^T\right]^T (\overline{u} - \overline{z})$, we obtain

$$\frac{1}{2}\|f_2(\overline{c})\|_2^2 = J_2.$$

Functional J_3 and J_4: From (7.6), we know that

$$R_h(c) = \sum_{i=1}^{n_{\mathcal{T}}} \overline{c}_i \chi_{\triangle_i}.$$

Defining the diagonal matrix $M_3 \in \mathbb{R}^{n_{\mathcal{T}} \times n_{\mathcal{T}}}$, such that

$$(M_3)_{ii} = \int_\Omega \chi_{\triangle_i} d(x, y),$$

we obtain

$$J_3(\overline{c}) = \frac{\epsilon}{2}\left[\overline{c} - \overline{\alpha}_1\right]_-^T M_3 \left[\overline{c} - \overline{\alpha}_1\right]_- \text{ and } J_4(\overline{c}) = \frac{\epsilon}{2}\left[\overline{\alpha}_2 - \overline{c}\right]_-^T M_3 \left[\overline{\alpha}_2 - \overline{c}\right]_-$$

where $\overline{\alpha}_1, \overline{\alpha}_2 \in \mathbb{R}^{n_{\mathcal{T}}}$ such that $\overline{\alpha}_1 = (\alpha_1, \alpha_1, \cdots, \alpha_1)$ and $\overline{\alpha}_2 = (\alpha_2, \alpha_2, \cdots, \alpha_2)$. Defining $f_3(\overline{c}) = \sqrt{\epsilon}[\sqrt{M_3}(\overline{c} - \overline{\alpha}_2), \sqrt{M_3}(\overline{\alpha}_1 - \overline{c})]^T$, it follows that

$$\frac{1}{2}\|f_3(\overline{c})\|_2^2 = J_3(\overline{c}) + J_4(\overline{c}).$$

Defining the vector field: Let us consider the following vector field

$$f(\overline{c}) = [f_1(\overline{c}), f_2(\overline{c}), f_3(\overline{c})]^T$$

then by construction our problem (7.8), it is reduced to minimize the following,

$$\min_{\overline{c} \in \mathcal{K}_h^{\overline{c}}} \frac{1}{2}\|f(\overline{c})\|_2^2 =: T(\overline{c}) \text{ with } F(\overline{c}) := \nabla T(\overline{c}). \tag{7.10}$$

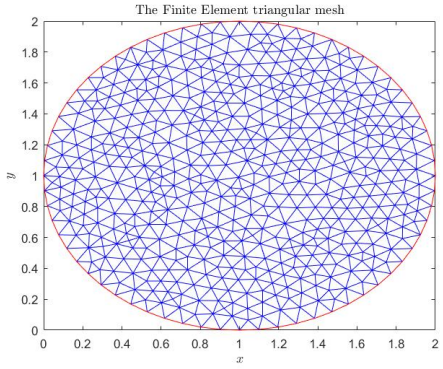


Figure 7.1: An automatic generated triangulation of $B((1,1);1)$ with maximum edged size 0.1, 485 nodes ($n_{\mathcal{P}}$) in the interior of Ω, 64 nodes ($n_{\mathcal{E}}$) in the rand of Ω, and 1032 triangles ($n_{\mathcal{T}}$).

Experiment I.

Let us define the projection operator $R_h : L^1(\Omega) \to span\,(\mathcal{X}_h)$ such that

$$R_h(c) = \sum_{i=1}^{n_{\mathcal{T}}} c(m_{\triangle_i})\chi_{\triangle_i},$$

and let us choose for our elliptic differential equation (7.1) the function

$$\mathrm{f}(x,y) = 1 - x^2 - y^2,$$

which has as exact solution

$$c_{true}(x,y) = 0.1 + 0.9\left[0.5 + 0.5\sin\left(10\pi\sqrt{x^2+y^2}\right)\right].$$

We choose as observed data $z = R_h(c_{true}) + 10^{-2}R_h(\varphi)$, where φ is the standard normal distribution. The penalty parameter ϵ and the step size control parameter H are taken to be $\epsilon = 10^4$ and $H = 0.5\max(1, \|\delta\bar{c}_0\|)$. The finite element triangular mesh is generated using an automatic Delanuary mesh generation approach [36], which provides a triangulation mesh with maximum edged size 0.1, 485 nodes ($n_{\mathcal{P}}$) in the interior of Ω, 64 nodes ($n_{\mathcal{E}}$) in the rand of Ω, 1032 triangles ($n_{\mathcal{T}}$), therefore our variable $\bar{c} \in \mathbb{R}^{1032}$ and $f(\bar{c}) \in \mathbb{R}^{n_{\mathcal{P}}+2n_{\mathcal{T}}}$, we show such a mesh in Figure 7.1. The lower and upper bound α_1 and α_2 in the constrained set $\mathcal{K}_h$ are taken to be 10^{-4} and 1 to ensure the identifying parameter lies in the above range. The initial guess $\bar{c}_0$ is the constant one everywhere and the termination criterion for the outer iterates is $\|F(\bar{c}_k)\| \leq 10^{-4}$. The damped inexact Gauss-Newton sequence $(\bar{c}_k)$ is defined by

$$\bar{c}_{k+1} = \bar{c}_k + t_k\bar{c}_k \text{ where } t_k = \min\mathcal{B}_H(\bar{c}_k), \text{ and } \delta\bar{c}_k := [\delta\bar{c}_k]_{i^k_{\max}}$$

satisfies the following conditions:

(1) $[\delta\bar{c}_k]_i$ solves approximately the linear equation

$$\left[J_f^T(\bar{c}_k)J_f(\bar{c}_k)\right]\Delta\bar{c}_k = -F(\bar{c}_k) \text{ for all } i = 0,1,\cdots,i^k_{\max}.$$

(2) Defining the inner residual error $J_f^T(\overline{c}_k)[r_k]_i := J_f^T(\overline{c}_k)[\delta\overline{c}_k]_i + F(\overline{c}_k)$, we obtain that $[\delta\overline{c}_k]_{i^k_{\max}}$ fulfill the stopping criterion (2.15), but $[\delta\overline{c}_k]^i$ do not for all $i = 0, \cdots, i^k_{\max} - 1$.

We compare the performance of three different IGN methods:

- IGN-LSQR, which uses LSQR as numerical linear algebra for computing the IGN step $\delta\overline{c}_k$ with stopping criterion (2.15), $\kappa = 0.55$ and $\kappa_{GN} = 0.5$.
- IGN-LSMR, which uses LSMR as numerical linear algebra for computing the IGN step $\delta\overline{c}_k$ with stopping criterion (2.15), $\kappa = 0.55$ and $\kappa_{GN} = 0.5$.
- IGN-CGS, which uses conjugate gradient squared (CGS) method [63] as numerical linear algebra for computing the IGN step $\delta\overline{c}_k$ with stopping criterion (2.15), $\kappa = 0.1$ and $\kappa_{GN} = 0$. Note that this approach does not fix with the theory presented in this thesis, but this way to compute the IGN step was proposed in [2]. Thus, we want to compare its performance with the new IGN methods (IGN-LSQR and IGN-LSMR) proposed in this thesis.

Inner Iteration: We start the analysis saying that IGN-LSMR requires for most of the outer iterations less inner iterations $i^k_{\max}$ for computing the IGN step $[\delta\overline{c}_k]_{i^k_{\max}}$ than IGN-LSQR and IGN-CGS. The reason is that in LSMR the inner residual $\|J_f^T(\overline{c}_k)[r_k]_i\|$ is monotonically decreasing at every inner iteration i. Thus, we can ensure that our stopping criterion (2.15), i.e.,

$$\|J_f^T(\overline{c}_k)[r_k]_i\| \leq \kappa\|F(\overline{c}_k)\| - \kappa_{GN}\| \left[J_f^T(\overline{c}_k)J_f(\overline{c}_k)\right] [\delta\overline{c}_k]_i\|,$$

which implies by virtue of Lemma 2.8 that

$$\|J_f^T(\overline{c}_k)[r_k]_i\| \leq (\kappa - \kappa_{GN})\|F(\overline{c}_k)\|,$$

must be earlier satisfied in IGN-LSMR. On the other hand, IGN-CGS is for most of the outer iterations the most computational expensive IGN method of the above. The reason is that in CGS the minimum error

$$\|\Delta\overline{c}_k - [\delta\overline{c}_k]_i\|_{\left[J_f^T(\overline{c}_k)J_f(\overline{c}_k)\right]} \text{ where } \|y\|_{\left[J_f^T(\overline{c}_k)J_f(\overline{c}_k)\right]} = \left(y^T \left[J_f^T(\overline{c}_k)J_f(\overline{c}_k)\right] y\right)^{1/2}$$

is monotonically decreasing at every inner iteration i, but we cannot guarantee here that the inner residual is monotonically decreasing. Thus, it is needed more inner iterations in order to fulfill (2.15) with $\kappa = 0.05$ and $\kappa_{GN} = 0$, i.e.,

$$\|J_f^T(\overline{c}_k)[r_k]_i\| \leq 0.05\|F(\overline{c}_k)\|.$$

Despite IGN-LSMR and IGN-LSQR use the same stopping criterion, we obtain that IGN-LSQR requires more inner iterations for computing the IGN step since in LSQR $\|[r_k]_i\|$ is monotonically decreasing but $\|J_f^T(\overline{c}_k)[r_k]_i\|$ is not. Clearly, Fig. 7.2c shows that the number of inner iterations necessary for computing the IGN-LSMR step $\delta\overline{c}_k$ is modestly small in comparison to IGN-LSQR and IGN-CGS. Indeed, IGN-LSMR requires less than 30 (3%) LSMR-inner iterations for computing $\delta\overline{c}_k$, which represents a huge savings in comparison with the $n = 1032$ (100%) inner iterations that requires the exact GN method for computing the GN step $\Delta\overline{c}_k$.

Speed of Convergence: The six largest singular values of $J_f(\bar{c}_k)$ are given by

IGN-LSMR $J_f(\bar{c}_{43})$	1.0333, 0.7774, 0.7585, 0.6616, 0.6133, 0.5951.
IGN-CGS $J_f(\bar{c}_{45})$	1.0333, 0.7766, 0.7584, 0.6615, 0.6134, 0.5947.
IGN-LSQR $J_f(\bar{c}_{46})$	1.0336, 0.7774, 0.7586, 0.6617, 0.6131, 0.5944.

and the six smallest singular values of $J_f(\bar{c}_k)$ are given by

IGN-LSMR $J_f(\bar{c}_{43})$	$6.68(10)^{-11}$, $1.6(10)^{-11}$, $1.16(10)^{-11}$, $2.7(10)^{-12}$, 0, 0.
IGN-CGS $J_f(\bar{c}_{45})$	$4.6(10)^{-11}$, $1.4(10)^{-11}$, $5.2(10)^{-12}$, $8.8(10)^{-15}$, 0, 0.
IGN-LSQR $J_f(\bar{c}_{46})$	$9.3(10)^{-10}$, $2.8(10)^{-10}$, $7.8(10)^{-12}$, 0, 0, 0.

Thus, we can expect that the above IGN methods be slow as we can appreciate in Fig. 7.2b. Nevertheless, IGN-LSMR reaches with just 43 outer iteration and less than 30 inner iterations for computing the IGN step a statistically stable solution (see Fig. 7.3). On the other hand, note that from Fig. 7.2c and the above smallest singular values, we can conclude that IGN-LSMR requires less inner iterations for computing the IGN step in most of the outer iterations, and in some outer iterations IGN-LSMR requires more inner iterations for computing the IGN step. The reason is that there is some outer iterations in IGN-LSQR and IGN-CGS where $J_f(\bar{c}_k)$ is rank deficient, which means that the Krylov subspaces $\mathcal{K}_i^k\left(\left[J_f(\bar{c}_k)^T J_f(\bar{c}_k)\right], F(\bar{c}_k)\right)$ in IGN-LSQR and IGN-CGS has lower dimension than the Krylov subspaces $\mathcal{K}_i^k\left(\left[J_f(\bar{c}_k)^T J_f(\bar{c}_k)\right], F(\bar{c}_k)\right)$ in IGN-LSMR, and therefore it is necessary less inner iterations for computing the IGN step in IGN-LSQR and IGN-CGS.

Difference between IGN and GN step: A natural question when we work with IGN methods is: How inaccurate must the IGN step be in order to ensure convergence of the IGN method?. Let us define

$$\kappa_k := \frac{\|J_f(\bar{c}_k)^T J_f(\bar{c}_k)\delta\bar{c}_k + J_f(\bar{c}_k)^T f(\bar{c}_k)\|}{\|J_f(\bar{c}_k)^T f(\bar{c}_k)\|},$$

and note that κ_k is measured how far away the IGN step $\delta\bar{c}_k$ is from the GN step $\Delta\bar{c}_k$ since

$$J_f(\bar{c}_k)^T J_f(\bar{c}_k)\Delta\bar{c}_k = -J_f(\bar{c}_k)^T f(\bar{c}_k),$$

which implies

$$\kappa_k := \frac{\|\delta\bar{c}_k - \Delta\bar{c}_k\|_{\left[J_f(\bar{c}_k)^T J_f(\bar{c}_k)\right]}}{\|\Delta x_k\|_{\left[J_f(\bar{c}_k)^T J_f(\bar{c}_k)\right]}} \text{ where } \|y\|_{\left[J_f(\bar{c}_k)^T J_f(\bar{c}_k)\right]} = \sqrt{\bar{c}_k^T\left[J_f(\bar{c}_k)^T J_f(\bar{c}_k)\right]\bar{c}_k}.$$

In Lemma 2.8, we prove that the stopping criterion (2.15) implies that

$$\kappa_k \leq \kappa - \kappa_{GN},$$

which means that such a stopping criterion provides implicitly an upper bound that predict how inaccurate our IGN step $\delta\bar{c}_k$ must be in order to produce convergence of the IGN sequence $(\bar{c}_k)$. Fig. 7.2e shows that the relative error between IGN step and the GN step for the above IGN methods is not bigger than 0.05 using $\|y\|_{[J_f(\bar{c}_k)^T J_f(\bar{c}_k)]}$ as norm. We finalize this section with another experiment that uses the IGN-LSMR approach described in the above experiment I with $\kappa = 0.9$. We want to show that the numerical results that we obtain in IGN-LSMR with $\kappa = 0.55$ do not depend on $\kappa \in (\kappa_{GN}, 1)$. Indeed, if κ is close to one, then the number of inner LSMR-iteration necessary for computing the IGN step $\delta\bar{c}_k$ decrease, but the number of outer iteration k necessary for satisfies our outer stopping criterion $\|F(\bar{c}_k)\| \leq 10^{-4}$ increase, the results obtained in this experiment are presented in Fig 7.4. We did not perform IGN-CGS with $\kappa = 0.4$ because this method does not converge.

7.2 Large-Scale Bundle Adjustment Problems

We start with a quote form [2] "Recent work in Structure from Motion (SfM) has demonstrated the possibility of reconstructing geometry from large-scale community photo collections. Bundle adjustment, the joint non-linear refinement of camera and point parameters, is a key component of most SfM systems, and one which can consume a significant amount of time for large problems. As the number of photos in such collections continues to grow into the hundreds of thousands or even millions, the scalability of bundle adjustment algorithms has become a critical issue."

Given a set of measured image feature locations and correspondences, the goal of Bundle adjustment is to find 3D point positions and camera parameters that minimize the projection error. This optimization problem is formulated as a nonlinear least-squares problem, where the error is the squared L_2-norm of the difference between the observed feature location and the projection of the corresponding 3D point on the image plane of the camera.

Camera Model

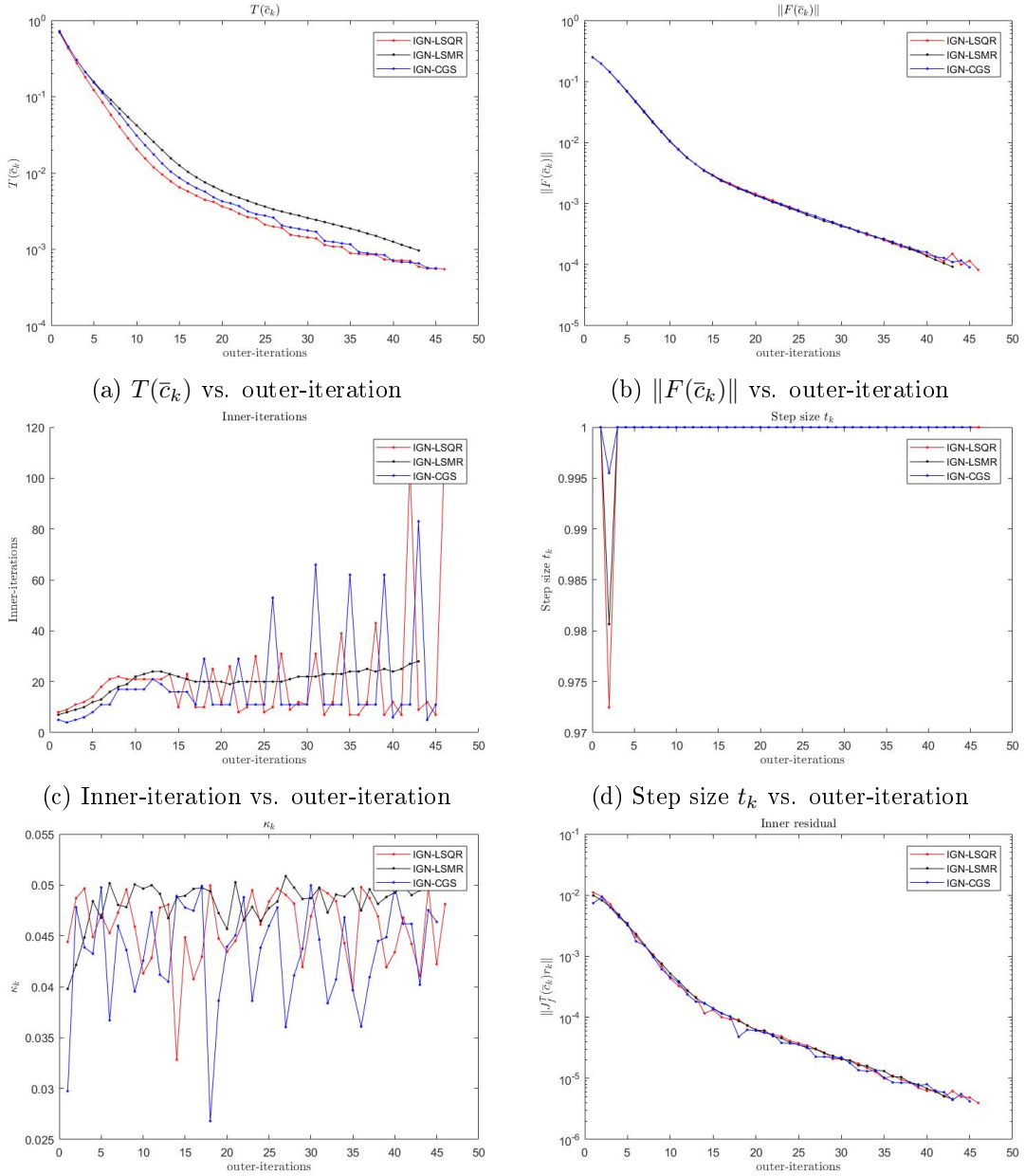
In [2] is used a pinhole camera model where the parameters to be estimated correspond to the camera c are (x_c, y_c, z_c) for the rotation matrix $R(x_c, y_c, z_c) \in \mathbb{R}^{3\times3}$, a translation vector $t_c \in \mathbb{R}^3$, a focal length $\mathrm{f}_c \in \mathbb{R}$ and two radial distortion parameters $k_1^c,\ \ k_2^c \in \mathbb{R}$. The formula for projecting a 3D point v into a pixel coordinates is:

$$P_c(v) = R(x_c, y_c, z_c)v + t_c \qquad \text{(conversion from world to camera coordinates)}$$

$$p_c(v) = -\frac{1}{(P_c(v))_3}[(P_c(v))_1, (P_c(v))_2] \qquad \text{(perspective division) where } (P_c(v))_i \text{ is the entry } i \text{ of } P_c(v) \in \mathbb{R}^3.$$

$$\mathcal{P}_c(v) = \mathrm{f}_c r_c(p_c(v)) p_c(v) \qquad \text{(conversion to pixel coordinates)}$$

$r_c(p)$ is a function that computes a scaling factor to undo the radial distortion and is defined by



(a) $T(\bar{c}_k)$ vs. outer-iteration

(b) $\|F(\bar{c}_k)\|$ vs. outer-iteration

(c) Inner-iteration vs. outer-iteration

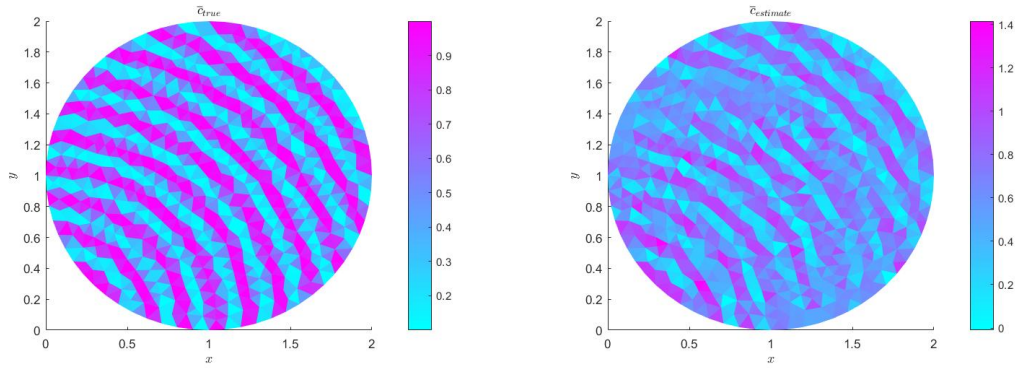
(d) Step size t_k vs. outer-iteration

(e) κ_k vs. outer-iteration

(f) Inner Residual Error vs. outer-iteration

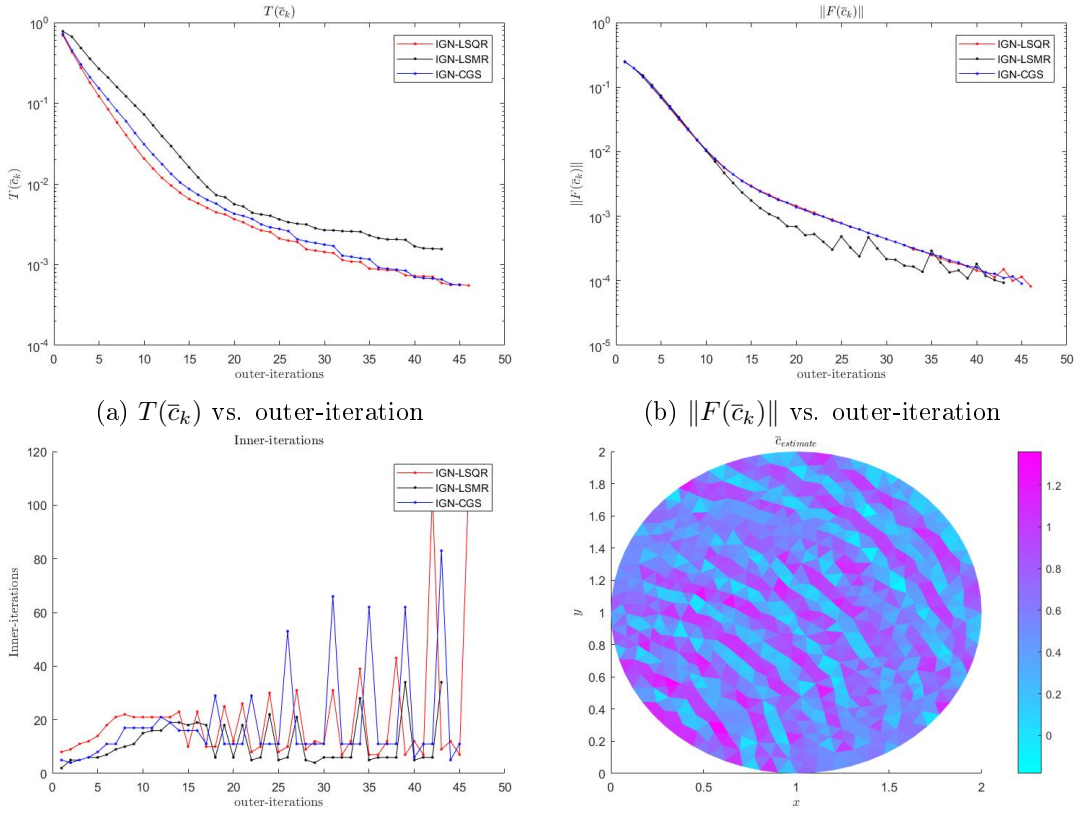
Figure 7.2: **Experiment I.** Parameter identification of the nonlinear steady-state differential equation (7.1) with f $= 1 - x^2 - y^2$. We solve the discretization unconstrained minimization sub-problem (7.10) that estimate our parameter on the finite element subspace generated by piecewise constant function with a regulation weight $\gamma = 10^{-4}$, penalty parameter $\epsilon = 10^4$, triangulation mesh given by Fig. 7.1, and outer stopping criterion $\|F(\bar{c}_k)\| \leq 10^{-4}$. The IGN step is calculated using three different numerical linear algebra with stopping criterion (2.15) that defines three different IGN methods: • ING-LSQR uses LSQR with $\kappa = 0.55$, and $\kappa_{GN} = 0.5$, • IGN-LSMR uses LSMR with $\kappa = 0.55$, and $\kappa_{GN} = 0.5$, • IGN-CGS uses CGS with $\kappa = 0.05$, and $\kappa_{GN} = 0$.

$$r_c(p) = 1 + k_1^c \|p_c\|^2 + k_2^c \|p_c\|^4.$$



(a) $\overline{c}_{true}$ (b) IGN-LSMR $\overline{c}_{estimate}$

Figure 7.3: **Experiment I.** $\overline{c}_{true}$ vs. $\overline{c}_{estimate}$. In each figure is drawn the triangulation mesh Fig. 7.1 in the plane xy and the color bar represents the value of the parameter on the triangle centers.



(a) $T(\overline{c}_k)$ vs. outer-iteration (b) $\|F(\overline{c}_k)\|$ vs. outer-iteration

(c) Inner-iteration vs. outer-iteration (d) IGN-LSMR $\overline{c}_{estimate}$

Figure 7.4: Performance of three different IGN methods: • ING-LSQR with $\kappa = 0.55$, and $\kappa_{GN} = 0.5$, • IGN-LSMR with $\kappa = 0.9$, and $\kappa_{GN} = 0.5$, • IGN-CGS with $\kappa = 0.05$, and $\kappa_{GN} = 0$.

$\mathcal{P}_c$ gives a projection in pixels where the origin of the image is the center of the image, the positive x-axis points right, and the positive y-axis points up (in addition, in the camera coordinate system, the positive z-axis points backwards, so the camera is looking

down the negative z-axis).

$R(x_c, y_c, z_c)$ is given by a particular Rodrigues's rotation matrix where (x_c, y_c, z_c) describes the axis of rotation about which v rotates by the angle $\theta_c = \sqrt{x_c^2 + y_c^2 + z_c^2}$ according to the right hand rule. Defining the following matrix,

$$K = \begin{bmatrix} 0 & -z_c & y_c \\ z_c & 0 & -x_c \\ -y_c & x_c & 0 \end{bmatrix}$$

the Rodrigues's rotation matrix in this case is given by

$$R(x_c, y_c, z_c) = I + \frac{\sin(\theta_c)}{\theta_c} K + \frac{(1 - \cos(\theta_c))}{\theta_c^2} K^2 \text{ where } \theta_c = \sqrt{x_c^2 + y_c^2 + z_c^2}$$

Data

We experimented with two sources of data taken from
http://grail.cs.washington.edu/projects/bal/

- Images captured at a regular rate using a Ladybug camera mounted on a moving vehicle. Image matching was done by exploiting the temporal order of the images and the GPS information captured at the time of image capture.

More information about the data setting it is available in [2, 1, 76, 55].

The data format in this problem is provided as a matrix D of four column where the first row it is stored the following information,

$D(1,1)$ the total number of cameras to be used.

$D(1,2)$ the total number of points v to be estimated.

$D(1,3)$ the total number of pixel coordinates observed.

Therefore, the column one contains all the information referring to camera index, column two is related to all the information referring to the points index, the columns four and five are related to the pixel points that were already observed. Thus, column four for the coordinate x of our pixel point and column five for the coordinate y. Using the above information we work with the following group of cameras $\mathcal{C} = \{c_1, c_2, \cdots, c_{D(1,1)}\}$, the group of points in 3D $V = \{v_1, v_2, \cdots, v_{D(1,2)}\}$ to be estimated, and group of observations $\mathcal{O} = \{\mathcal{O}_1, \cdots, \mathcal{O}_{D(1,3)}\}$. The row $i \in \{2, \cdots, D(1,3)+1\}$ of D stores the following information,

$D(i,1)$ the camera index with which the pixel $(D(i,3), D(i,4))$ was generated.

$D(i,2)$ the index of a point $v_{D(i,2)} \in V$ to be estimated, which approximately generated through the camera $c_{D(i,1)}$ the pixel information $(D(i,3), D(i,4))$. In other words, $\mathcal{P}_{c_{D(i,1)}}(v_{D(i,2)}) \approx (D(i,3), D(i,4))$.

Residual Function $f(x)$

We define our residual function $f(x)$ using the data information D described in the above section where the variable vector was stored as in [2], i.e., x has a block structure $x = [y_1, y_2, \cdots, y_{D(1,1)}, v_1, v_2, \cdots v_{D(1,2)}]$ where y_i correspond to nine parameter related to the camera c_i, i.e., $y_i = (x_{c_i}, y_{c_i}, z_{c_i}, t_1^{c_i}, t_2^{c_i}, t_3^{c_i}, \mathrm{f}_{c_i}, k_1^{c_i}, k_2^{c_i})$ and $v_j = (v_1^j, v_2^j, v_3^j)$ correspond to the point parameters.

We define our residual function as $f(x) = (f_1(x), f_2(x), \cdots, f_{D_{(1,3)}}(x))$ where

$$f_i(x) = \mathcal{P}_{c_{D(i+1,1)}}(v_{D(i+1,1)})^T - (D(i+1,3), D(i+1,4))^T.$$

then our bundle adjustment problem is reduced to solve the following optimization problem,

$$\text{minimize } T(x), \text{ where } T(x) = \frac{1}{2}\|f(x)\|_2^2. \tag{7.11}$$

Let $J_f(x)$ be the Jacobian of $f(x)$, then our damped IGN-BSC sequence (x_k) is generated using $(S3)$ with $t_k = \min \mathcal{B}_H(x_k)$, where $H = H_{rel} \max(1, \|\delta x_0\|)$. Clearly, the matrix is sparse since in f_i there is not a single term that includes two or more cameras or point blocks. We use LSMR for generated our Inexact Gauss-Newton step δx_k with stopping criterion (2.15).

One of the challenges of this particular problem is that the Jacobian $J_f(x_k)$ is rank deficient at every iteration, which means that the dimension of our Krylov subspace $\mathcal{K}_n\left(J_f^T(x_k)J_f(x_k), J_f^T(x_k)f(x_k)\right)$ is smaller than n, and thus the inner iteration $[\delta x_k]_i$ may stagnate at zero, for an example of such a behavior see [16, Example 4.1]. In order to avoid this scenario we introduce a regularization term when we compute our inexact Gauss-Newton step, thus, instead of solving approximately via LSMR the following linear least squares problem

$$\min_{\Delta x} \frac{1}{2}\|J_f(x_k)\Delta x_k + f(x_k)\|_2^2, \tag{7.12}$$

we solve

$$\min_{\Delta x} \frac{1}{2}\|J_f(x_k)\Delta x_k + f(x_k)\|_2^2 + \gamma\|D(x_k)\Delta x_k\|_2^2 \tag{7.13}$$

where $D(x_k)$ is the square root of the diagonal of the matrix $J_f^T(x_k)J_f(x_k)$ and $\gamma > 0$ represents a parameter such that $H(x_k) = J_f^T(x_k)J_f(x_k) + \gamma D(x_k)^T D(x_k)$ is positive definite, which is a typical regularization term using in the Levenberg Marquardt Algorithm [63]. In our case, we have been keeping γ fixed at every iteration because it is sufficient to add a regularization term such that $J_f^T(x_k)J_f(x_k) + \gamma D(x_k)^T D(x_k)$ is positive definite.

Experiment I

The Ladybug Dataset were taken from http://grail.cs.washington.edu/projects/bal/ with file name problem-1723-156502-pre.txt.bz2, which is a problem with 1723 cameras, 156502 points, and 678718 pixel observations, therefore the total number of parameters to be estimated is $n := 1723 \times 9 + 156502 \times 3 = 485013$ and the total number of residual errors is $m := 2 \times 678718 = 1357436$. Thus, in this case the Jacobian matrix size is

1357436×485013. We fix the following parameters, $\kappa_{GN} = 0.2$, $\kappa = 0.3$, $\gamma = 0.01$, the initial guess is taken from problem-1723-156502-pre.txt.bz2 and $H = 0.3 * \max(1, \|\delta x_0\|)$. We use as outer stopping criterion

$$\|J_f^T(x_k)[J_f(x_k)\delta x_k + f(x_k)]\| < 0.5.$$

At every iterate x_k we use the scaling matrix $D_k = \text{diag}(d_1^k, \cdots, d_n^k)$ [62, Seccion 6, Equation (6.3)] where

$$\begin{aligned} d_i^0 &:= \|J_f(x_0)e_i\|, \\ d_i^1 &:= \max\{\|J_f(x_0)e_i\|, \|J_f(x_1)e_i\|\}, \\ &\vdots \\ d_i^k &:= \max\{\|J_f(x_{k-1})e_i\|, \|J_f(x_k)e_i\|\}. \end{aligned}$$

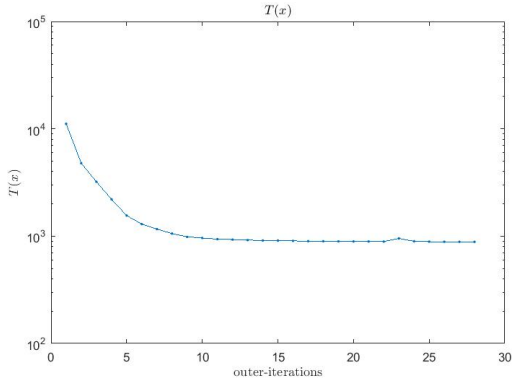
since some parameters (e.g. distortion) are up to 20 orders of magnitude more sensitive than others (e.g. rotations). The goal of this experiment is to show that the new damped IGN-BSC approach performs well despite this problem does not fix with the theory presented in Chapter 6. The results are presented in Fig. 7.5 and 7.6. As linear iterative solver for computing the IGN step we choose LSMR, which required less than 35 (less than 1%) inner iterations (see Fig. 7.5b) for computing our IGN step despite the GN method requires in this example 485013 (100%) inner iterations for computing the GN step. Fig. In 7.5e is measured how inexact the IGN step δx_k is in comparison with the GN step Δx_k since

$$H(x_k)\Delta x_k = -F(x_k), \text{ and } H(x_k)\delta x_k = -F(x_k) + J_f^T(x_k)r_k$$

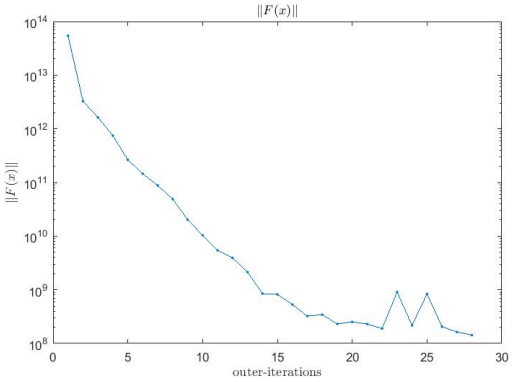
where $r_k = J_f(x_k)\delta x_k + f(x_k)$ which implies

$$J_f^T(x_k)r_k = H(x_k)[\Delta x_k - \delta x_k],$$

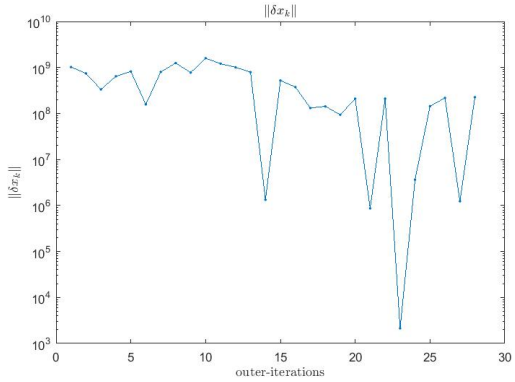
thus, if $\|J_f^T(x_k)r_k\| = 0$, then $\Delta x_k = \delta x_k$. As we observe, Fig. 7.5e shows than the inner residual error is decreasing from $\|J_f^T(x_0)r_0\| = 1.338533 * 10^3$ to $\|J_f^T(x_{28})r_{28}\| = 6.968297 * 10^{-1}$. A important question is: How the inacuracy of our IGN step δx_k does the convergence rate of our damped IGN-BSC approach influence?, the answer is given in Fig. 7.5b in where we observe that the speed of convergence of the outer iterates are not slow since with just 28 outer iteration we decrease from $5.4 * 10^{13}$ to $1.4 * 10^{08}$. This result allows to say that it is possible to solve the inner subproblems with low accuracy without reducing the seep of convergence of the outer iteration dramatically. Finally Fig. 7.5g and Fig. 7.5h show the entries of the initial residual error $f(x_0)$ and the entries of the Final residual error $f(x_{28})$ respectively and Fig. 7.6 shows a 3D representation of our estimation x_{28}.



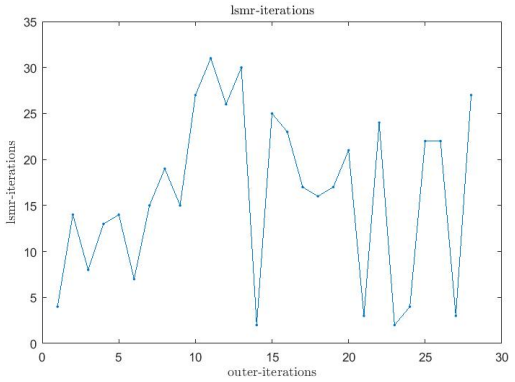
(a) $T(x_k)$ vs. outer-iteration



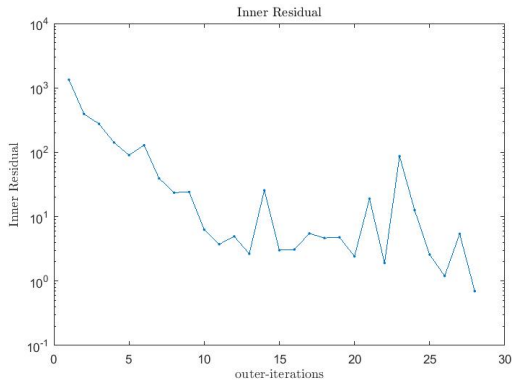
(b) $\|F(x_k)\|_2$ vs. outer-iteration



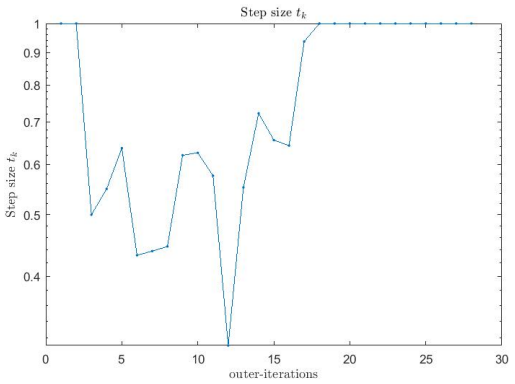
(c) $\|\delta x_k\|_2$ vs. outer-iteration



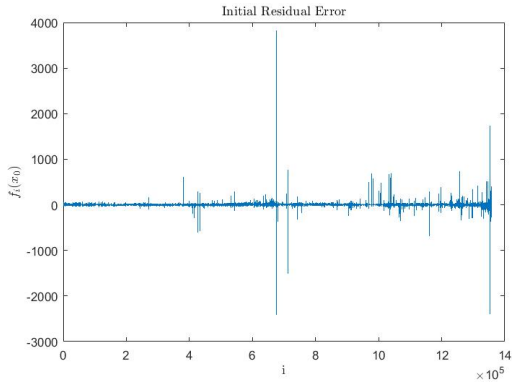
(d) inner iteration vs. outer-iteration



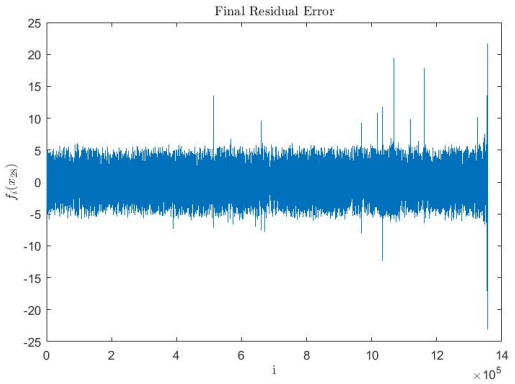
(e) Inner Residual vs. outer-iteration



(f) t_k vs. outer-iteration



(g) Initial Residual Error



(h) Final Residual Error

Figure 7.5: Parameter estimation of Ladybug problem with 1723 cameras, 156502 points, and 678718 pixel observations. $T(x) = \frac{1}{2}\|f(x)\|^2$ is the objective function, $F(x) = \nabla T(x)$, $H = 0.3 * \max(1, \|\delta x_0\|)$, $\kappa_{GN} = 0.2$, $\kappa = 0.3$, $\gamma = 0.01$ with a particular initial guess taken from problem-1723-156502-pre.txt.bz2. As outer stopping criterion we require that the inner residual Error be less than 1.

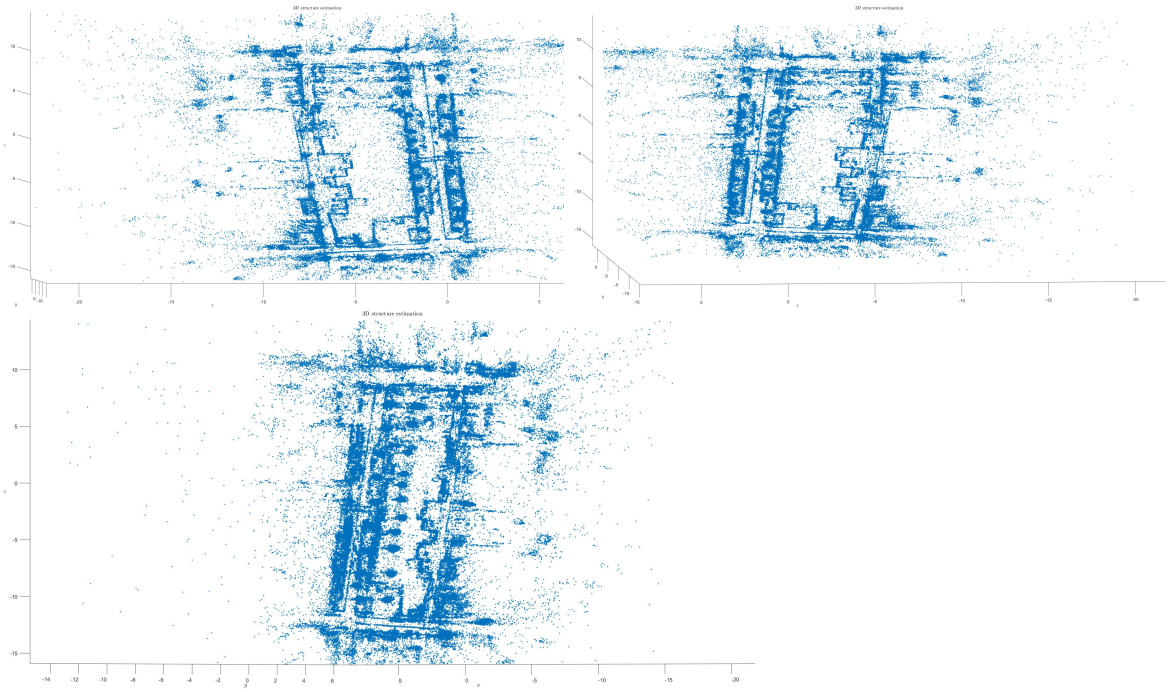


Figure 7.6: 3D solution of Experiment I.

Chapter 8

Conclusions and outlook

In this Thesis, we have introduced a local inexact Gauss-Newton approach for solving nonlinear least squares problems, which uses LSQR [65] or LSMR [33] as numerical linear algebra for solving approximately the linear least squares linearized subproblems with a new and early computationally available termination rule. Furthermore, we have proved that locally the above IGN approach is essentially a covariant approach with $\|y\|_*$-norm where the hypotheses with $\|y\|_*$-norm of the famous local contraction theorem introduced by Bock [10] are valid. Thus, we can ensure that this approach converges locally and linearly with $\|y\|_*$-norm to a statistically stable solution. Finally, We have generalized the local ideas of this new local IGN method and introduced a damped IGN method based on the Backward Step Control theory of Potschka [70]. This new approach results to be a damped inexact Gauss-Newton globalization strategy that requires far less inner-iterations for computing the IGN step than the classical exact Gauss-Newton method based on factorization algorithm for computing the GN step that requires 100% of the inner iterations. In our experiments, we have showed that this new damped IGN approach requires less than the 3% of inner iterations for computing the IGN step in order to converge to a statistically stable solution, which represents a huge computational savings in comparison with the classical exact Gauss-Newton.

Outlook

we observe in the experimental example given by the steady state equation (7.1) that for $\kappa \approx \kappa_{GN}$ the damped IGN sequence needs more inner iterations for calculated the IGN step and needs less outer iterations to satisfy the outer stopping criterion than the damped IGN sequence computed using an $\kappa \approx 1$. Thus, it would be interested to know what is the optimal κ than produce the minimum numbers of inner iteration and the minimum numbers of outer iterations.

We have proved that locally our new IGN approach is essentially a covariant approach with $\|y\|_*$-norm if $\tilde{k} = (2(\kappa - \kappa_{GN})\mu_* + 1)^4\kappa < 1$ where $\kappa \in (\kappa_{GN}, 1)$, and $\mu_* = \text{cond}\left(J_f^T(x_*)J_f(x_*)\right)$. We conjecture that this result is valid for all $\kappa \in (\kappa_{GN}, 1)$. Unfortunately, a proof of such a result is outside the scope of this thesis and is theme of future research.

The new IGN approach is based on a known linear algebra for solving linear least squares problems and a new early stopping criterion. Thus, the following question arises:

Is there a numerical linear algebra for solving linear least squares problems that requires less inner iterations for computing the IGN step than the IGN approach presented in this thesis and defines a new local IGN approach that provides statistically stable solutions provided that one exists?.

As continuation of this dissertation research, we would like to apply our IGN globalization strategy to large scale parameter estimation problems with nonlinear constraints and also be able to prove in this general setting that our IGN approach converges to a statistically stable solution provided that one exists. Let us write formally such a problem as follows: Let $f_1 : D \subseteq \mathbb{R}^n \to \mathbb{R}^{m_1}$, and $f_2 : D \subseteq \mathbb{R}^n \to \mathbb{R}^{m_2}$ be two twice differential functions with correspondent Jacobian $J_1(x)$ and $J_2(x)$, such that $J(x) = [J_1^T(x), J_2^T(x)]^T$ and $J_2(x)$ are full rank matrices for all $x \in D$ with $n < m := m_1 + m_2$ and $m_2 < n$. Defining $f(x) = [f_1^T(x), f_2^T(x)]$, we present our parameter estimation problems with nonlinear constraints as,

$$\begin{aligned} &\underset{x \in D}{\text{minimize}} \ \frac{1}{2}\|f(x)\|_2^2 \\ &f_2(x) = 0. \end{aligned} \tag{8.1}$$

In order to solve the above problem, we need to find the $k - k - t$ points defined by the following nonlinear system of equation where λ represent the Lagrange multiplier,

$$F(x) := \begin{bmatrix} J^T(x)f(x) - J_2^T(x)\lambda \\ \\ f_2(x) \end{bmatrix} = 0. \tag{8.2}$$

Given an initial guess $x_0 \in D$, the exact Gauss-Newton method finds a possible local solution of (8.1) according to $x_{k+1} = x_k + \Delta x_k$, where Δx_k is the only solution of

$$\begin{aligned} \min_{\Delta x \in \mathbb{R}^n} \quad & \frac{1}{2}\|J(x_k)\Delta x + f(x_k)\|_2^2 \\ \text{s.t} \quad & J_2(x_k)\Delta x_k + f_2(x_k) = 0. \end{aligned} \tag{8.3}$$

The principal issue here is that LSQR or LSMR cannot be applied to solved approximately (8.3). Nevertheless, a possible strategy to deal with this drawback may be: The GN step can be write as

$$\Delta x_k = \Delta x_k^1 + \Delta x_k^2 \text{ where } \Delta x_k^1 \in \text{Rank}(J_2(x_k)) \text{ and } \Delta x_k^2 \in \text{Null}(J_2(x_k)).$$

Thus, we can easily calculate the value of Δx_k^1 solving the following equation

$$J_2(x_k)\Delta x_k^1 = -f_2(x_k), \tag{8.4}$$

which solution is given by $\Delta x_k^1 = -J_2^T(x_k)\left[J_2(x_k)J_2^T(x_k)\right]^{-1} f_2(x_k)$, and

$$\Delta x_k^2 = \underset{\Delta x^2 \in \mathbb{R}^n}{\arg\min} \frac{1}{2}\left\|J(x_k)\Delta x^2 + J(x_k)\Delta x_k^1 + f(x_k)\right\|_2^2. \tag{8.5}$$

Thus, an Inexact Gauss-Newton approach for solving (8.1) would be according to

$$x_{k+1} = x_k + \delta x_k \text{ with } \delta x_k = \delta x_k^1 + \delta x_k^2$$

where δx_k^1 is calculated using LSQR or LSMR as numerical linear algebra with some particular early inner termination rule that only depends on cheaply available information, which is the topic of a future research. δx_k^2 is calculated using LSQR or LSMR as numerical linear algebra with a variant of the stopping criterion that we developed in this work. A globalization of the above approach is always possible based on the Backward Step Control theory introduced by Potschka [70].

Bibliography

[1] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. M. Seitz, and R. Szeliski. Building Rome in a Day. *Commun. ACM*, 54(10):105–112, October 2011.

[2] S. Agarwal, N. Snavely, S.M. Seitz, and R. Szeliski. Bundle Adjustment in the Large. In *Proceedings of the 11th European Conference on Computer Vision: Part II*, pages 29–42, Berlin, Heidelberg, 2010. Springer Verlag.

[3] M. Arioli, I. Duff, and D. Ruiz. Stopping Criteria for Iterative Solvers. *SIAM J. Matrix Anal. Appl.*, 13(1):138–144, January 1992.

[4] U. Ascher and M.R. Osborne. A note on solving nonlinear equations and the natural criterion function. *Journal of Optimization Theory and Applications*, 55:147–152, 1987.

[5] K. Astala and L. Päivärinta. Calderón's Inverse Conductivity Problem in the Plane. *Annals of Mathematics*, 163(1):265–299, 2006.

[6] K. Astala, L. Päivärinta, and M. Lassas. Calderón's Inverse Problem for Anisotropic Conductivity in the Plane. *Communications in Partial Differential Equations*, 30(1-2):207–224, 2005.

[7] Y. Bard. *Nonlinear parameter estimation*. Academic Press, 1974.

[8] I. Bauer, H.G. Bock, S. Körkel, and J.P. Schlöder. Numerical methods for optimum experimental design in DAE systems. *Journal of Computational and Applied Mathematics*, 120(1):1 – 25, 2000.

[9] Å. Björck. Numerical Methods for Least Squares Problems. Society for Industrial and Applied Mathematics, 1996.

[10] H.G. Bock. *Randwertproblemmethoden zur Parameteridentifizierung in Systemen nichtlinearer Differentialgleichungen*, volume 183 of *Bonner Mathematische Schriften*. Universität Bonn, Bonn, 1987.

[11] H.G. Bock, E. Kostina, and J.P. Schlöder. Numerical Methods for Parameter Estimation in Nonlinear Differential Algebraic Equations. *GAMM-Mitteilungen*, 30(2):376–408, 2007.

[12] H.G. Bock, E. Kostina, and J.P. Schlöder. *On the Role of Natural Level Functions to Achieve Global Convergence for Damped Newton Methods*, pages 51–74. Springer US, Boston, MA, 2000.

[13] H. Borouchaki, P.L. George, F. Hecht, P. Laug, and E. Saltel. Delaunary mesh generation governed by metric specifications. Part I. Algorithms. *Finite Element in Analysis and Desing*, pages 61–83, 1997.

[14] P. N. Brown. A Theoretical Comparison of the Arnoldi and GMRES Algorithms. *SIAM Journal on Scientific and Statistical Computing*, 12(1):58–78, 1991.

[15] P.N. Brown and Y. Saad. Convergence Theory of Nonlinear Newton–Krylov Algorithms. *SIAM Journal on Optimization*, 4(2):297–330, 1994.

[16] E. Cătinaş. Inexact perturbed Newton methods and applications to a class of Krylov solvers. *Journal of Optimization Theory and Applications*, 108(3):543–570, Mar 2001.

[17] X.W. Chang, C.C. Paige, and D. Titley-Péloquin. Stopping criteria for the iterative solution of Linear Least Squares problems. *SIAM J. Matrix Analysis Applications*, 31(2):831–852, 2009.

[18] J. Chen. The convergence analysis of inexact Gauss-Newton methods for nonlinear problems. *Computational Optimization and Applications*, 40(1):97–118, May 2008.

[19] J. Chen and W. Li. Convergence of Gauss-Newton's method and uniqueness of the solution. *Applied mathematics and computation*, 170(1):686–705, 2005.

[20] Sou-Cheng (Terrya) Choi. *Iterative Methods for Singular Linear Equations and Least-square Problems.* PhD dissertation, Stanford Univertity, 2006.

[21] D.F. Davidenko. On a new method of numerical solution of systems of nonlinear equations. *Doklady Akademii nauk SSSR*, 88:601–602, 1953.

[22] R.S. Dembo, S.C. Eisenstat, and T. Steihaug. Inexact Newton Methods. *SIAM Journal on Numerical Analysis*, 19(2):400–408, 1982.

[23] J. Dennis and R. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations.* Society for Industrial and Applied Mathematics, 1996.

[24] J.E. Dennis. On Newton-like methods. *Numerische Mathematik*, 11(4):324–330, May 1968.

[25] J.E. Dennis and J.J. Moré. A Characterization of Superlinear Convergence and Its Application to Quasi-Newton Methods. *Mathematics of Computation*, 28(126):549–560, 1974.

[26] P. Deuflhard. Global inexact newton methods for very large scale nonlinear problems. *IMPACT of Computing in Science and Engineering*, 3(4):366 – 393, 1991.

[27] P. Deuflhard. *Newton Methods for Nonlinear Problems. Affine Invariance and Adaptive Algorithms.* 2004.

[28] P. Deuflhard, R. Freund, and A. Walter. Fast secant methods for the iterative solution of large nonsymmetric linear systems. *IMPACT of Computing in Science and Engineering*, 2(3):244 – 276, 1990.

[29] S.C. Eisenstat and H.F. Walker. Globally Convergent Inexact Newton Methods. *SIAM Journal on Optimization*, 4(2):393–422, 1994.

[30] S.C. Eisenstat and H.F. Walker. Choosing the Forcing Terms in an Inexact Newton Method. *SIAM Journal on Scientific Computing*, 17(1):16–32, 1996.

[31] G. Fasano, F. Lampariello, and M. Sciandrone. A truncated nonmonotone Gauss-Newton method for Large-Scale Nonlinear Least-Squares Problems. *Computational Optimization and Applications*, 34(3):343–358, Jul 2006.

[32] O.P. Ferreira, M.L.N. Gonçalves, and P.R. Oliveira. Local convergence analysis of the Gauss–Newton method under a majorant condition. *Journal of Complexity*, 27(1):111 – 125, 2011.

[33] D. Fong and M. Saunders. LSMR: An Iterative Algorithm for Sparse Least-Squares Problems. *SIAM Journal on Scientific Computing*, 33(5):2950–2971, 2011.

[34] D. Fong and M. Saunders. CG versus MINRES: An empirical comparison. *Sultan Qaboos University Journal for Science*, 17(1):44–62, 2012.

[35] A. Galántai and J. Abaffy. Always convergent iteration methods for nonlinear equations of Lipschitz functions. *Numerical Algorithms*, 69(2):443–453, Jun 2015.

[36] P.L. George. Automatic mesh generation: Application to finite element methods. *International Journal for Numerical Methods in Engineering*, 1991.

[37] G. Golub and W. Kahan. Calculating the Singular Values and Pseudo-Inverse of a Matrix. *Journal of the Society for Industrial and Applied Mathematics Numerical Analysis*, 2(2):205–224, 1965.

[38] G.H. Golub and C.F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, 2nd edition, 1989.

[39] M.L.N. Gonçalves. Inexact Gauss-Newton like methods for injective-overdetermined systems of equations under a majorant condition. *Numerical Algorithms*, 72(2):377–392, Jun 2016.

[40] S. Gratton, A.S. Lawless, and N.K. Nichols. Approximate Gauss–Newton Methods for Nonlinear Least Squares Problems. *SIAM Journal on Optimization*, 18(1):106–132, 2007.

[41] Ming Gu. Backward Perturbation Bounds for Linear Least Squares Problems. *SIAM Journal on Matrix Analysis and Applications*, 20(2):363–372, 1998.

[42] A. Hohmann. *Inexact Gauss Newton methods for parameter dependent nonlinear problems*. Shaker Aachen, 1994.

[43] Ilse Ipsen and Carl D. Meyer. The Idea Behind Krylov Methods. 105:1–16, 11 1997.

[44] Kazufumi Ito. Functional Analysis and Optimization, November 2014.

[45] Jr. J.E. Dennis and J.J. Moré. Quasi-Newton Methods, Motivation and Theory. *SIAM Review*, 19(1):46–89, 1977.

[46] P. Jiránek and D. Titley-Peloquin. Estimating the Backward Error in LSQR. *SIAM J. Matrix Anal. Appl.*, 31(4):2055–2074, may 2010.

[47] B. Kaltenbacher. Parameter Identification in Partial Differential Equations, WS 2005/06.

[48] E.M. Kasenally. GMBACK: A generalised minimum backward error algorithm for nonsymmetric linear systems. *SIAM J. Sci. Comput.*, 16(3):698–719, 1995.

[49] E.M. Kasenally and V. Simoncini. Analysis of a Minimum Perturbation Algorithm for Nonsymmetric Linear Systems. *SIAM Journal on Numerical Analysis*, 34(1):48–66, 1997.

[50] C. Kelley. *Iterative Methods for Linear and Nonlinear Equations.* Society for Industrial and Applied Mathematics, 1995.

[51] S. Körkel and E. Kostina. Numerical Methods for Nonlinear Experimental Design. In H.G. Bock, H.X. Phu, E. Kostina, and R. Rannacher, editors, *Modeling, Simulation and Optimization of Complex Processes*, pages 255–272, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[52] E. Kostina. Robust Parameter Estimation in Dynamic Systems. *Optimization and Engineering*, 5(4):461–484, Dec 2004.

[53] E.A. Kostina, M.A. Saunders, and I. Schierle. Computation of covariance matrices for constrained parameters. Technical report, Universität Heidelberg, 2009.

[54] S. Körkel. *Numerische Methoden für optimale Versuchsplanungsprobleme bei nichtlinearen DAE-Modellen.* PhD thesis, Universitä Heidelberg, 2002.

[55] X. Li, C. Wu, C. Zach, S. Lazebnik, and J. Frahm. Modeling and Recognition of Landmark Image Collections Using Iconic Scene Graphs. In *Proceedings of the 10th European Conference on Computer Vision: Part I*, pages 427–440, Berlin, Heidelberg, 2008. Springer Verlag.

[56] M.I.A. Lourakis and A.A. Argyros. SBA: A Software Package for Generic Sparse Bundle Adjustment. *ACM Trans. Math. Softw.*, 36(1):2:1–2:30, March 2009.

[57] M.L.A. Lourakis and A.A. Argyros. Is Levenberg-Marquardt the most efficient optimization algorithm for implementing bundle adjustment? In *Tenth IEEE International Conference on Computer Vision*, volume 1. IEEE, 2005.

[58] G.P. Akilov L.V. Kantorovich. Functional Analysis in Normed Spaces. *Fizmatgiz, Moscow, 1959. German translation: Berlin, Academie-Verlag*, 1964.

[59] H. Martínez, Z. Parada, and R.A. Tapia. On the characterization of Q-superlinear convergence of Quasi-Newton interior-point methods for nonlinear programming. 1:16, 04 1995.

[60] I. Misovskikh. On convergence of Newton's method. *(Russia). Trudy Mat. Inst. Steklov*, 28:145–147, 1949.

[61] J.J. More. Levenberg–Marquardt algorithm: implementation and theory. 1 1977.

[62] J.J. Moré. The Levenberg-Marquardt algorithm: Implementation and theory. In G. Watson, editor, *Numerical Analysis*, pages 105–116, New York, 1978. Springer Verlag.

[63] J. Nocedal and S.J. Wright. *Numerical Optimization*. Springer-Verlag, New York, NY, USA, 2nd edition, 2006.

[64] J.M. Ortega and W.C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 1970.

[65] C.C. Paige and M.A. Saunders. LSQR: An Algorithm for Sparse Linear Equations and Sparse Least Squares. *ACM Trans. Math. Softw.*, 8(1):43–71, March 1982.

[66] V. Pereyra. Iterative Methods for Solving Nonlinear Least Squares Problems. volume 4, pages 27–36, 1967.

[67] M. Porcelli. On the convergence of an inexact Gauss–Newton trust-region method for nonlinear least-squares problems with simple bounds. *Optimization Letters*, 7(3):447–465, Mar 2013.

[68] F.A. Potra. On Q-order and R-order of convergence. *Journal of Optimization Theory and Applications*, 63(3):415–431, Dec 1989.

[69] A. Potschka. *A direct method for the numerical solution of optimization problems with time-periodic PDE constraints*. PhD thesis, Universität Heidelberg, 2011.

[70] A. Potschka. Backward Step Control for Global Newton-Type Methods. *SIAM Journal on Numerical Analysis*, 54(1):361–387, 2016.

[71] E. Ramirez. *Finite element methods for parameter identification problem of linear and nonlinear steady-state diffusion equations*. PhD thesis, Virginia Tech, 1997.

[72] W.C. Rheinboldt. *Methods for Solving Systems of Nonlinear Equations*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2nd edition, 1998.

[73] Y. Saad. *Iterative Methods for Sparse Linear Systems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1st edition, 1996.

[74] F.J. Sayas. A gentle introduction to the Finite Element Method. *Lecture notes, University of Delaware*, 2008.

[75] A.H. Sherman. On Newton-Iterative Methods for the Solution of Systems of Nonlinear Equations. *SIAM Journal on Numerical Analysis*, 15(4):755–771, 1978.

[76] N. Snavely, S.M. Seitz, and R. Szeliski. Skeletal graphs for efficient structure from motion. In *Proc. Computer Vision and Pattern Recognition*, 2008.

[77] G.W. Stewart. On the Perturbation of Pseudo-Inverses, Projections and Linear Least Squares Problems. *SIAM Review*, 19(4):634–662, 1977.

[78] G.W. Stewart. Research development and LINPACK. In J. R. Rice, editor, *Mathematical Software III*, pages 1–14. Academic Press, New York, 1977.

[79] G.W. Stewart and J.G. Sun. *Matrix Perturbation Theory*. Academic Press, 1990.

[80] P.-Å. Wedin. On the Gauss-Newton Method for the Nonlinear Least Squares Problems. Working Paper 24, Institute for Applied Mathematics, Stockholm, 1974. Cited in Åke Björck's bibliography on least squares, which is available by anonymous ftp from `math.liu.se` in `pub/references`.

[81] Per-Åke Wedin. Perturbation theory for pseudo-inverses. *BIT Numerical Mathematics*, 13(2):217–232, Jun 1973.

[82] S.J. Wright and J.N. Holt. An inexact Levenberg-Marquardt method for large sparse nonlinear least squres. *The Journal of the Australian Mathematical Society. Series B. Applied Mathematics*, 26(4):387–403, 1985.

[83] X. Xun, J. Cao, B. Mallick, A. Maity, and R.J. Carroll. Parameter Estimation of Partial Differential Equation Models. *Journal of the American Statistical Association*, 108(503):1009–1020, 2013.

[84] Jun Zou. Numerical methods for elliptic inverse problems. *International Journal of Computer Mathematics*, 70(2):211–232, 1998.