PhD Thesis

Optimization under Uncertainty

Robust Parameter Estimation with Erroneous Measurements and Uncertain Model Coefficients

Inaugural-Dissertation zur Erlangung der Doktorwürde der
Fakultät für Mathematik und Informatik
der
Philipps-Universität Marburg

vorgelegt von
Tanja Binder
aus
Heidelberg

Oktober 2012

Gutachter:
Prof. Dr. Ekaterina Kostina, Philipps-Universität Marburg
Prof. Dr. Dr. h.c. Hans-Georg Bock, Ruprecht-Karls-Universität Heidelberg

Dissertation

Optimization under Uncertainty

Robust Parameter Estimation with Erroneous Measurements and Uncertain Model Coefficients

vorgelegt von

Tanja Binder

aus

Heidelberg

Philipps-Universität Marburg
Fachbereich Mathematik und Informatik
Hans-Meerwein-Straße, 35032 Marburg

Angefertigt mit Genehmigung des Fachbereichs Mathematik und Informatik der Philipps-Universität Marburg.

Gutachter:
Prof. Dr. Ekaterina Kostina, Philipps-Universität Marburg
Prof. Dr. Dr. h.c. Hans-Georg Bock, Ruprecht-Karls-Universität Heidelberg

Prüfungskommission:
Prof. Dr. Ekaterina Kostina, Philipps-Universität Marburg
Prof. Dr. Dr. h.c. Hans-Georg Bock, Ruprecht-Karls-Universität Heidelberg
Prof. Dr. Stephan Dahlke, Philipps-Universität Marburg
Prof. Dr. Manfred Sommer, Philipps-Universität Marburg

Datum der Abgabe: 29.10.2012
Datum der Disputation: 21.12.2012

Abstract

In particular in the last decade, optimization under uncertainty has engaged attention in the mathematical community and beyond. When trying to model the behavior of real-world processes mathematically, this behavior is often not fully understood. Uncertainty may concern the involved species of biological or chemical reactions, the kind of reactions taking place, and the numerical values of model coefficients. But also experimental measurements, which form the basis for the determination of the model parameters, contain unavoidable errors. In order to judge the validity and reliability of the numerical results of simulations and model predictions, the inherent uncertainties have at least to be quantified. But it would be even better to incorporate these uncertainties in the mathematical computations from the beginning to derive solutions that are insensitive with respect to such perturbations.

The present dissertation starts with an introductory chapter that examines the background of the thesis, in particular the distinct sources of uncertainties and errors in mathematical models. The structure of the remaining part of the thesis is guided by the different types of uncertainties that may be relevant in numerical optimization. Part I of the dissertation deals with measurement errors in the experimental data and comprises the following content:

- Chapter 2 yields an introduction to the theory of parameter estimation problems. Starting from the statistical background, the general problem is formulated on which the work of this thesis is based. A short recap of the classical least squares method motivates the necessity of alternative solution approaches that are more robust with respect to deviations from the model assumptions. The l_1 norm estimator and Huber's M-estimator are presented as two important instances of such methods. New in this chapter is the neat contrasting juxtaposition of the properties of the three methods.

- Chapter 3 introduces the reader to the theory of the generalized Gauss-Newton method for solving constrained nonlinear optimization problems. The transfer of this theory to the special cases of parameter estimation with the l_1 norm estimator and the Huber estimator constitutes the main part of this chapter. Here we go into detail about the optimality conditions of the two robust problems and about the local convergence of the resulting method. To our best knowledge, this has not been elaborated or published before for Huber's M-estimator.

- Chapter 4 contains an adaption of the generalized Gauss-Newton method to the case that the necessary derivatives can or must be computed only inexactly. The resulting error is compensated by a correction term in the cost function. The

transfer of this approach to parameter estimation with the l_1 norm estimator and the Huber estimator are innovations of this thesis. The proofs concerning the stability of the active sets and the local convergence of the inexact methods for the two robust estimators are published here for the first time.

Part II of the dissertation treats the case of uncertainties in the model coefficients and contains the following chapters:

- Chapter 5 introduces the worst-case approach of robust optimization. In particular, we give an approximative computation method of the worst-case solution of a particular class of problems. To our knowledge, the transfer of this computational method to ODE constrained parameter estimation problems is a new approach.
- Chapter 6 describes an alternative approach for solving uncertain ODE constrained parameter estimation problems that is based on the method of stochastic collocation. The combination of this method with multiple shooting is one of the main contributions of this thesis.

Part III of the dissertation, finally, contains numerical examples that illustrate the theoretical considerations of the previous chapters. Chapter 7 contains example problems referring to part I of the thesis concerning parameter estimation problems with due regard to measurement errors. In chapter 8 the case of part II of the thesis with uncertainties in the given model parameters is considered more closely by means of some examples.

The thesis ends with some conclusions and an outlook to possible future investigations in chapter 9.

Zusammenfassung

Besonders im letzten Jahrzehnt hat die numerische Optimierung unter Unsicherheiten immer mehr das Interesse auf sich gezogen. Beim Versuch, das Verhalten realer Prozesse mathematisch zu beschreiben, ist dieses Verhalten in vielen Fällen gar nicht mit absoluter Genauigkeit bekannt. Unsicherheiten können die reagierenden Substanzen, die Art der Reaktionen und die numerischen Werte der Modellkoeffizienten betreffen. Aber auch die experimentellen Messungen, aufgrund derer die Modellparameter im Allgemeinen bestimmt werden, enthalten unvermeidliche Fehler. Um die Aussagekraft und Zuverlässigkeit der numerischen Ergebnisse von Simulationen und Modellvorhersagen beurteilen zu können, müssen die inherenten Modellunsicherheiten zumindest quantifiziert werden. Noch besser ist es, diese Unsicherheiten in die mathematischen Berechnungen mit einzubeziehen, und Lösungen zu berechnen, die unempfindlich sind gegenüber solchen Störungen.

Die vorliegende Dissertation beginnt mit einem einleitenden Kapitel, das die Hintergründe der Arbeit, speziell die verschiedenen Ursachen von Unsicherheiten und Fehlern in mathematischen Modellen beleuchtet. Der Aufbau der restlichen Arbeit orientiert sich an den verschiedenen Arten von Unsicherheiten, die in der numerischen Optimierung relevant sein können. Teil I der Dissertation befasst sich mit Messfehlern in den experimentellen Daten und umfasst den folgenden Inhalt:

- Kapitel 2 bietet eine Einführung in die Theorie der Parameterschätzprobleme. Ausgehend vom statistischen Hintergrund wird das allgemeine grundlegende Problem dieser Arbeit formuliert. Eine kurze Zusammenfassung des klassischen Kleinste-Quadrate-Verfahrens motiviert die Notwendigkeit alternativer Lösungsansätze, die robuster gegenüber Störungen der Modellannahmen sind. Als zwei wichtige Beispiele derartiger Verfahren werden der l_1-Norm-Schätzer und Hubers M-Schätzer vorgestellt. Das Neue in diesem Kapitel ist die übersichtlichen Gegenüberstellung der Eigenschaften dieser drei Verfahren.

- Kapitel 3 führt in die Theorie des verallgemeinerten Gauß-Newton-Verfahrens zur Lösung restringierter nichtlinearer Optimierungsprobleme ein. Die Übertragung dieser Theorie auf die beiden Spezialfälle der Parameterschätzung mit dem l_1-Norm-Schätzer und dem Huber-Schätzer bilden den Hauptteil des Kapitels, wobei besonders auf die Optimalitätsbedingungen der beiden Probleme und die lokale Konvergenz der resultierenden Verfahren eingegangen wird. Unseres Wissen nach wurde das für den Huber-Schätzer so vorher noch nicht ausgearbeitet und veröffentlicht.

- Kapitel 4 enthält eine Adaption des verallgemeinerten Gauß-Newton-Verfahrens

für den Fall, dass die benötigten Ableitungen nur inexakt berechnet werden können oder sollen. Der dadurch entstehende Fehler wird duch einen Korrekturterm in der Zielfunktion ausgeglichen. Die Übertragung dieses Ansatzes auf die Parameterschätzung mit dem l_1-Norm-Schätzer und Hubers M-Schätzer sind Neuerungen dieser Arbeit. Die Beweise zur Stabilität der aktiven Mengen und der lokalen Konvergenz des inexakten Verfahrens für die beiden robusten Schätzer werden hier zum ersten Mal veröffentlicht.

Teil II der Dissertation behandelt den Fall von Unsicherheiten in den Modellkoeffizienten und enthält die folgenden Kapitel:

- Kapitel 5 stellt den Worst-Case-Ansatz der robusten Optimierung vor. Speziell wird eine näherungsweise Berechnung der Worst-Case-Lösung für eine spezielle Problemklasse dargestellt. Die Übertragung dieses approximativen Berechnungsverfahrens auf ODE-beschränkte Parameterschätzprobleme ist unseres Wissens nach ein neuer Ansatz.

- Kapitel 6 beschreibt einen alternativen Ansatz zur Lösung unsicherer Parameterschätzprobleme mit ODE-Beschränkungen, der auf der Methode der stochastischen Kollokation basiert. Die Kombination dieses Verfahrens mit der Mehrzielmethode ist einer der Hauptbeiträge dieser Arbeit.

Teil III der Dissertation enthält schließlich numerische Beispiele, die die theoretischen Behandlungen der vorangehenden Kapitel verdeutlichen. Kapitel 7 enthält Beispielprobleme zu Teil I der Arbeit über Parameterschätzprobleme unter Berücksichtigung der Messfehler. In Kapitel 8 wird der Fall aus Teil II der Arbeit mit Unsicherheiten in den gegebenen Modellparametern anhand von Beipielen näher untersucht.

Die Arbeit schließt mit einer Zusammenfassung und einem Ausblick auf weitere mögliche Untersuchungen in Kapitel 9.

Acknowledgements

I want to thank my advisor Prof. Dr. Ekaterina Kostina for the opportunity to pursue my Ph.D. under her guidance. Her many ideas and support have been very much appreciated. I have learned a lot during the last five years, not only in the mathematical field but also about myself.

I am further obliged to Max Nattermann, Dr. Alexandra Herzog, and Gregor Kriwet who joined me in Prof. Kostina's group after some time and also to numerous people from other groups at the University of Marburg. Max and Alex, I will always be indebted to you for many fruitful discussions, good cooperation, insight, and being real friends. Furthermore, I want to express my particular gratitude to the computer scientists Dr. Nicolas Menzel, Nico Grund, and Thomas Fober.

No less I have to acknowledge assistance of the people of my former group at the Interdisciplinary Center of Scientific Computing at the University of Heidelberg where I pursued my diploma under the guidance of Prof. Dr. Dr. h.c. Hans-Georg Bock. Especially I want to name Christian Hoffmann, Dr. Christian Kirches, Simon Lenz, Dr. Mario Mommer, Dr. Andreas Potschka, Andreas Sommer, and Leonard Wirsching. Though having left them after my diploma to continue my studies in Marburg, they have always welcomed me heartily in their group back home.

I am grateful having had Prof. Dr. Mira Mezini of the University of Darmstadt as a mentor who shared her experience and had an open ear for me and my co-mentees Eva Hörmann (University of Gießen) and Sarah Voß (University of Frankfurt/Main). I also want to express my thanks to the latter two for being a great peer group in our common interest in doing a good Ph.D. Furthermore, I am indebted to the team of SciMento hessenweit for initializing the contact and making this possible, for a great organization, and a very helpful workshop program.

Very special thanks go to my parents Monika and Gerhard Binder as well as my sister Claudia for their endless love and support. They have never stopped believing in me and have always been there when I needed them. I also thank my boyfriend Jens Peschek for his love and motivation and his good humor, for being happy with me when everything turned out as planned and consolidating me in times of failure.

Funding from the Federal Ministry of Education and Research during some time of my Ph.D. study in Marburg is thankfully acknowledged.

List of symbols

Stochastics

Z	sample of random variable $z_1, \ldots, z_n$
T	estimator for the sample Z
p	parameter to be estimated
$\mathbb{E}$	expected value
$\mathbb{B}$	bias
$\mathbb{P}$	probability
σ^2	variance
$\mathcal{N}$	normal distribution with expected value $\mathbb{E}$and variance σ^2
Φ	standard normal distribution
H	normal or other symmetric distribution
f	probability density
effect	effect of the replacement of m values of a sample Z on an estimator T
BDP	breakdown point of an estimator
p	assumed probability distribution
IF	influence function of an estimator
MSE	mean squared error of the estimate
efficiency	relative efficiency of an estimator compared to another estimator

Differential algebraic equations

n_y	number of differential variables
y	differential states $[\in \mathbb{R}^{n_y}]$
n_z	number of algebraic variables
z	algebraic states $[\in \mathbb{R}^{n_z}]$
n_x	number of state variables $n_x = n_y + n_z$
x	state variables $x(t)^T = (y(t)^T, z(t)^T)$ $[\in \mathbb{R}^{n_x}]$
n_p	number of parameters
p	parameters $[\in \mathbb{R}^{n_p}]$

f	right hand side of differential equations $[\mathbb{R} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_z} \times \mathbb{R}^{n_p} \to \mathbb{R}^{n_y}]$
g	right hand side of algebraic equations $[\mathbb{R} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_z} \times \mathbb{R}^{n_p} \to \mathbb{R}^{n_z}$ (index 1 DAE)]
t_0	initial time of integration interval
t_f	final time of integration interval
x_0	initial values of the state variables at t_0

Parameter estimation for DAEs

M	number of measurement times
t	measurement times $t_j,\ j = 1, \ldots, M$
M_j	number of measurements at time t_j
$\sum_j M_j$	total number of measurements
η	vector of measurements η_{ij} $[i = 1, \ldots, M_j,\ j = 1, \ldots, M]$
ε	vector of measurement errors ε_{ij} $[i = 1, \ldots, M_j,\ j = 1, \ldots, M]$
σ	vector of standard deviations σ_{ij} $[i = 1, \ldots, M_j,\ j = 1, \ldots, M]$
r_1	objective function of DAE parameter estimation problem
r_2	equality constraints of DAE parameter estimation problem
r_3	inequality cosntraints of DAE parameter estimation problem
K	number of times with interior or boundary points constraints
θ	times $\theta_i,\ i = 1, \ldots, K$, with interior or boundary points constraints
ρ	optimization criterion to be minimized $[\mathbb{R}^{n_x+n_p} \to \mathbb{R}]$
γ	tuning parameter for Huber's function $(\in [0, \infty))$
ρ_γ	Huber's function $[\mathbb{R} \to \mathbb{R}]$
$\mathbb{I}$	unit matrix
$\mathtt{1}$	vector of all ones of appropriate dimension

Multiple shooting

N	number of multiple shooting nodes
τ	multiple shooting nodes $\tau_j,\ j = 0, \ldots, N$
s^y	new multiple shooting variable representing the differential states at τ $(s^y_j = y(\tau_j))$
s^z	new multiple shooting variable representing the algebraic states at τ $(s^z_j = z(\tau_j))$

s	multiple shooting variable representing state variables at τ ($s_j = (s_j^y, s_j^z)$)
α	decreasing function with $\alpha(\tau_j) = 1,\ \alpha(\tau) \geq 0$ for consistency improvement of the algebraic equations during the integration
$x(t; s_j, p)$	subtrajectory for $t \in [\tau_j, \tau_{j+1}]$
$h_j(s_j, s_{j+1}, p)$	continuity conditions for multiple shooting ($h_j(s_j, s_{j+1}, p) = x(\tau_{j+1}; \tau_j, p) - s_{j+1} = 0,\ j = 0, \ldots, N-1$)

Nonlinear parameter estimation

n	number of variables to be optimized
X	variables to be optimized ($\in \mathbb{R}^n$)
m_1	number of components of the objective function
F_1	objective function of nonlinear parameter estimation problem ($\mathbb{R}^n \to \mathbb{R}^{m_1}$)
m_2	number of equality constraints of nonlinear parameter estimation problem
F_2	equality constraints of nonlinear parameter estimation problem ($\mathbb{R}^n \to \mathbb{R}^{numEqCon}$)
m_3	number of inequality constraints of nonlinear parameter estimation problem
F_3	inequality constraints of nonlinear parameter estimation problem ($\mathbb{R}^n \to \mathbb{R}^{numIneqCon}$)
ρ	optimization criterion to be minimized [$\mathbb{R}^n \to \mathbb{R}$]
$\|\cdot\|_1$	l_1 norm ($\|x\|_1 = \sum_i \|x_i\|$)
$\|\cdot\|_2$	l_2 norm ($\|x\|_2 = \sqrt{x^T x} = \sqrt{\sum_i x_i^2}$)
u, v	auxiliary vectors
e	vector of all ones of appropriate length ($e = (1, \ldots, 1)^T$)
$\mathcal{M}$	design matrix in linear standard regression model
ψ	derivative of optimization criterion ($\psi = \frac{\partial \rho}{\partial X}$)
w	weighting function ($w(r) = \frac{\psi(r)}{r}$)
$\bar{\sigma}$	scaling factor (estimate of the standard deviation)

Nonlinear programming

F	nonlinear cost function to be minimized ($\mathbb{R}^n \to \mathbb{R}$)
m_g	number of nonlinear equality constraints
G	nonlinear equality constraints ($\mathbb{R}^n \to \mathbb{R}^{m_g}$)
m_h	number of inequality constraints

H	nonlinear inequality constraints ($\mathbb{R}^n \to \mathbb{R}^{m_h}$)
$\hat{X}$	feasible point
X^*	minimizer
N	neighborhood of a specified point
I	set of indices of the active inequality constraints at a specified point
ϑ	path parameter ($\in \mathbb{R}$)
c	(continuously differentiable) arc/curve in $\mathbb{R}^n$ parameterized by ϑ
$\tilde{H}$	active inequalities ($\tilde{H}(X) = (H_i(X))_{i \in I(X)}$)
$\tilde{G}$	active constraints ($\tilde{G}(X) = \begin{pmatrix} G(X) \\ \tilde{H}(X) \end{pmatrix}$)
λ	Lagrange multiplier for equality constraints in Fritz John optimality conditions
μ	Lagrange multiplier for inequality constraints in Fritz John optimality conditions
μ_0	Lagrange multiplier for cost function in Fritz John optimality conditions
$\mathcal{L}$	Lagrange function
d	critical direction ($\in \mathbb{R}^n$)
D	set of critical directions d
k	iteration counter
t	step length
ΔX	increment / search direction
K	KKT matrix for equality constrained nonlinear program ($K(X, \lambda) = \begin{pmatrix} \nabla_X \mathcal{L}(X, \lambda) \\ G(X) \end{pmatrix}$)
J	Jacobian of $K (J = \nabla K)$
λ^{QP}	Lagrange multiplier for linearized equality constraints in quadratic problem of SQP iteration
μ^{QP}	Lagrange multiplier for linearized inequality constraints in quadratic problem of SQP iteration
L	approximation of Hessian of Langrangian ($L = \nabla_X \mathcal{L}(X, \lambda, \mu)$)
$\bar{H}$	linearized inequality constraints ($\bar{H}(\Delta X) = H(X^k) + \nabla_X H(X^k) \Delta X \geq 0$)
J_1	Jacobian of $F (J_1 = \nabla_X F)$
J_2	Jacobian of $F_2 (J_2 = \nabla_X F_2)$
J_3	Jacobian of $F_3 (J_3 = \nabla_X F_3)$

A_1	approximate Jacobian of $F (A_1 \approx \nabla_X F)$
A_2	approximate Jacobian of $F_2 (A_2 \approx \nabla_X F_2)$
A_3	approximate Jacobian of $F_3 (A_3 \approx \nabla_X F_3)$
R	rest term in Gauss-Newton approximation of Hessian of Langrangian
$\tilde{F}_2$	active constraints, $\tilde{F}_2 = \begin{pmatrix} F_2 \\ (F_{3i})_{i \in I_{F_3}} \end{pmatrix}$
$\tilde{J}_2$	Jacobian of the active constraints, $\tilde{J}_2 = \begin{pmatrix} J_2 \\ (J_{3i})_{i \in I_{F_3}} \end{pmatrix}$
$\mathcal{F}$	full vector of problem functions, $\mathcal{F} = \begin{pmatrix} F_1 \\ \tilde{F}_2 \end{pmatrix}$
$\mathcal{J}$	full Jacobian of the cost function and the active constraints, $\mathcal{J} = \begin{pmatrix} J_1 \\ \tilde{J}_2 \end{pmatrix}$
$\mathcal{J}^+$	generalized inverse of $\mathcal{J}$
G_L	linearized confidence region for given parameter estimate
ν	multiplier forrobust objective function in optimality conditions
T	merit function for stepsize selection in Gauss-Newton iterations
$\mathcal{N}$	neighborhood of a specified point

Robust optimization

n_a	number of uncertain parameters
α	uncertain parameter
$\mathcal{A}$	uncertainty set
δ	radius of the uncertainty set in one dimension
α_0	nominal value of the uncertainty
$\mathcal{P}(\alpha)$	probability measure on the uncertainty set $\mathcal{A}$
$\rho(\alpha)$	probability densitiy function associated with $\mathcal{P}(\alpha)$
$\mathcal{I}$	interpolation function
p	interpolation polynomial with degree indicated as index where necessary
$\mathcal{L}$	Lagrange basis polynomial
I	integration operator
Q	quadrature operator
ω	quadrature weights

Π	space of polynomials with number of variables indicated in subscript and maximal degree in superscript where necessary
m	number of interpolation or quadrature nodes
Θ	set of interpolation/quadrature nodes with space dimension indicated as subscript where necessary
Λ	Lebesgue constant in polynomial interpolation
$\mathcal{U}$	arbitrary interpolation or quadrature formula
$\mathbf{i}$	multi-index
a	basis function for interpolation or quadrature weight
Δ	difference formulae of two subsequent interpolation/quadrature formulae ($\Delta_i = \mathcal{U}_i - \mathcal{U}_{i-1}$)
A	Smolyak construction with level q and dimension n denoted in parantheses, $A(q, n)$
H	sparse grid with level q and dimension n denoted in parantheses, $H(q, n)$
ΔH	difference grid $\Delta H(q, n) = H(q, n) \setminus H(q-1, n)$
$\Delta\Theta$	difference set that contains only new support nodes, $\Delta\Theta_i = \Theta_i \setminus \Theta_{i-1}$

Contents

Abstract iii

Zusammenfassung v

Acknowledgements vii

Acknowledgements vii

List of symbols ix

Contents xiv

1. Introduction 1
- 1.1. Sources of uncertainties and errors in modeling and simulation 2
- 1.2. Stochastic vs. robust optimization 7
  - 1.2.1. Stochastic optimization 7
  - 1.2.2. Robust optimization 8
- 1.3. Outline of the thesis 10

I. Errors in Measurements 11

2. Parameter estimation: From classical to robust procedures 13
- 2.1. Statistical background 14
- 2.2. The problem formulation 16
- 2.3. Numerical treatment of the dynamics 20
  - 2.3.1. The initial value problem approach 20
  - 2.3.2. The boundary value problem approach 21
- 2.4. Classical least squares estimation 23
- 2.5. Need for robust parameter estimation 24
- 2.6. Least absolute deviations estimation 25
- 2.7. Huber estimation 26
- 2.8. Comparison of the three estimators 28
- 2.9. Other robust estimators 31

3. Solution of the robust parameter estimation problems **37**
3.1. The Gauss-Newton method in nonlinear programming 37
3.1.1. Optimality conditions for nonlinear programming problems . . 39
3.1.2. The SQP framework 42
3.1.3. The generalized Gauss-Newton method 45
3.1.4. Local convergence of the Gauss-Newton method 47
3.1.5. Numerical solution of the linearized problem 48
3.1.6. Statistical analysis of the solution and confidence regions 52
3.2. The Gauss-Newton method in robust optimization 54
3.2.1. Optimality conditions 54
3.2.2. Application of the GN method and the generalized inverse . . . 64
3.2.3. Local convergence properties 69
3.2.4. Solution of the linearized problems 70
3.3. Globalization strategies for the Gauss-Newton method 71
3.3.1. Line search methods 71
3.3.2. The restrictive monotonicity test 72

4. The Gauss-Newton method with inexact derivatives **73**
4.1. Inexact Gauss-Newton method for nonlinear optimization 73
4.2. Inexact Gauss-Newton for l_1 norm minimization 75
4.2.1. Derivation of the algorithm 76
4.2.2. Stability of the active set and local convergence 81
4.3. Inexact Gauss-Newton for Huber minimization 90
4.3.1. Derivation of the algorithm 91
4.3.2. Stability of the active set and local convergence 95

II. Uncertainties in model parameters and structure **105**

5. The worst-case approach of robust optimization **107**
5.1. Introduction to the problem 108
5.2. The robust counterpart problem and the worst-case approach 111
5.3. Approximation of the robust counterpart problem 113
5.4. Numerical solution of the approximated robust counterpart problem . 116
5.5. Special case without inequality constraints 118
5.6. Robustness and sparsity 119

6. Alternative approach: Stochastic collocation with multiple shooting **121**
6.1. Introduction to stochastic collocation 122
6.2. Review: Lagrange polynomial interpolation and quadrature 125
6.2.1. Polynomial interpolation methods 125
6.2.2. Univariate quadrature rules 128
6.2.3. Common structure 135
6.3. Tensor product formulae 135

6.4. Sparse grids . . . 137
6.5. Solution of the uncertain optimization problem with ODE constraints 143
6.6. The GN method for the doubly discretized uncertain optimization problem . . . 145

III. Applications and conclusions 149

7. Parameter estimation problems with erroneous measurements 151
7.1. The NF-κB pathway . . . 151
7.1.1. Outliers in the data . . . 153
7.1.2. White noise in the data . . . 155
7.2. Denitrogenization of pyridine . . . 157

8. Parameter estimation problems with uncertain model coefficients 165
8.1. The Lotka-Volterra predator-prey system . . . 165
8.2. Isomerization of α-pinene . . . 168
8.2.1. One uncertainty . . . 169
8.2.2. Two uncertainties . . . 173
8.2.3. Three uncertainties . . . 177
8.2.4. Four uncertainties . . . 181

9. Conclusion 185
9.1. Summary . . . 185
9.2. Future research directions . . . 187

List of Figures 189

List of Tables 190

References 192

1. Introduction

The mathematical representation of real-world processes from physics, chemistry, biology, engineering, or economics is typically based on both certain knowledge and less certain assumptions of the involved subprocesses and their behavior and interactions. These assumptions refer to the process structure in general, both the outer and inner conditions under which the process runs, the involved species in a biological or chemical reaction process and their correlations, as well as the defining parameters. Even small uncertainties may render the nominal optimal solution of an optimization problem completely useless for practical purposes.

A fundamental goal in mathematical modeling, simulation, and optimization is reliability in the predictions and estimations. This is of particular importance if expenditure or security restrictions are involved. But even predictions for deterministic physical systems are often uncertain as the systems are mostly heterogeneous and only insufficiently characterized by (erroneous) measurement data, see Lin et al. [92]. In order to give guarantees on the prediction accuracy of a mathematical model, effects caused by physical or numerical insecurities must be taken into account and quantified. A common approach for uncertainty quantification is the representation of uncertain parameters by random variables of which the statistical properties like distribution, mean, and variance are assessed from the data, cf. Lin et al. [92]. But this makes the deterministic equations become stochastic.

Costly function evaluations for the objective function or the constraints often prohibit the possibility to solve an optimization problem several hundreds or thousands of times, using different realizations of the inherent uncertainties, to analyze the behavioral spectrum of a process. Therefore, specially designed methods are needed, which account for these inaccuracies from the beginning in a cheap and fast manner.

A traditional approach to model validation, "My model is valid because it reproduces the test data with adequate accuracy ..." (quoted from Hemez and Döbling [71]), is insufficient as it limits the predictive value of the model to the small region of the design space from which the test data have been gathered. Any extrapolatory predictions are invalidated by this point of view, but often this extrapolation of information from only few observations is the purport of computer simulations. In effect, it is of minor importance whether the model reproduces measurements from a single experiment with sufficient accuracy. The significant task is having the model predict the system behavior with the same probability of occurence as is inferred from all available experimental data.

There are several conceptual questions involved in the task of accounting for inaccuracies, particularly verification, validation, and uncertainty quantification (see also the report of the Committee for Verification, Validation and Uncertainty Quantifi-

cation [38]). *Verification* corresponds to the problem of how accurate the solution of the mathematical equations or the computation of the quantities of interest can be done on a computer. This comprises both code verification, i.e. checking if the implementation of the algorithm has been done correctly, and solution verification, i.e. checking if the numerical algorithm really does what it is supposed to do. The aim is a justified confidence that the algorithms are correctly implemented and that the numerical solution approaches the true solution of the model (cf. Karniadakis and Glimm [82]). *Validation* is concerned with the accuracy as to which the model represents the physical reality of quantities it has been designed for. This is mostly done by a comparison of simulations and real experiments. *Uncertainty quantification*, finally, tries to measure how much influence the different sources of errors and uncertainties have on the model predictions with respect to the quantities of interest. It is the process of characterizing, estimating, propagating, and analyzing the different kinds of uncertainties in a model, cf. Booker et al. [28]. These three concepts are necessary tools to ensure performance, safety, and reliability of any mathematical or computer simulation. We will go into more detail with this in the next section.

It is important to note that these three assignments are not independent of one another. For instance, solution verification is completely useless unless code verification has shown the validity of the implementation. On the other hand, validation depends on an accurate estimation of the numerical error from the solution verification and of the propagation of input inaccuracies on the quantities of interest. Figure 1.1 gives a schematic representation of these interconnections, based on a graph from Oberkampf et al. [106] (also reproduced in the report of the Committee for Verification, Validation and Uncertainty Quantification [38]). In this report, also several principles and best practice strategies are collected for verification and validation or predition, respectively. The authors of this study also state open questions where further research is necessary in this field.

1.1. Sources of uncertainties and errors in modeling and simulation

As we have mentioned in the introductory paragraph above, predictions from computer simulations must contain concrete, quantifiable declarations of the contained amount of insecurity. In order to fully capture the spectrum of system behavior, all sources of uncertainties must be accounted for. The systematic treatment of model and data uncertainties and their propagation through the computer model for predicting quantities of interest with quantified uncertainty is part of *predictive simulation* procedures, cf. Oden et al. [107]. There are quite a lot possible sources of errors and uncertainties during this process. Measurements of physical processes contain observational errors while a corresponding theory and mathematical model will probably contain modeling errors. It is part of the validation task to trace these and estimate their influence on the outcome of the computation. Verification, on the other hand, is concerned with discretization errors that arise from the transfer of the

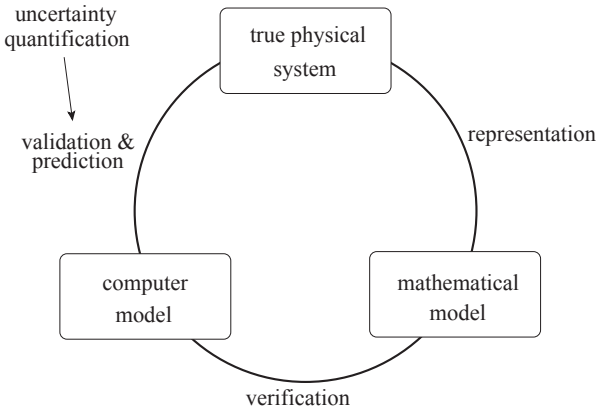


Figure 1.1.: Interconnections of verification, validation, and uncertainty quantification (based on a figure from Oberkampf et al. [106], reproduced in the report of the Committee for Verification, Validation and Uncertainty Quantification [38]).

(continuous, i.e. infinite dimensional) mathematical model into a (discrete, i.e. finite dimensional) computer model.

Schuëller and Jensen [123] further distinguish predictive simulation into reliability-based optimization, robust design optimization, and model updating and system identification. The term *reliability-based optimization* is used by them for the solution of optimization problems that quantitatively describes the effects of uncertainty by means of failure probabilities and expected values. In contrast, *robust design optimization* is the methodology of determining designs that are comparatively insensitive to changes of structural parameters and the like. *Model updating and system identification*, finally, aims at reducing the discrepancies that arise from comparing model responses with test data. We will not go that far in this thesis and restrict the further deliberations to the quite general task of identifying and assessing errors and uncertainties in all kinds of simulation and optimization problems.

For clarity of the used terminology, we define now the notions of uncertainties and errors which are not synonymous. An *uncertainty* is defined by Alvin et al. [2] as a "*potential* deficiency in any phase or activity of the modeling process that is due to *lack of knowledge*" (emphasis as used in the original). This means, an uncertainty may or may not occur in a specific instance of the process under consideration. But any extension of the information base for the process may help reduce the amount of uncertainty. It is often treated by giving a probability distribution associated with the frequency of its occurence and is thus prone to be dealt with by stochastic optimization in this case, see section 1.2. This concept can be further divided into *aleatory* and *epistemic* types of uncertainty. The former, aleatory (*stochastic*) uncertainty, refers to random or problem inherent variations that are inevitable and include statistical concepts of, e.g., variability and probability. So these uncertainties are unknowns that change their values each time an experiment

is run. So measurement errors belong to this class, for instance. The latter, epistemic (*systematic*) uncertainty, characterizes those uncertainties which arise from insufficient knowledge and can be reduced with increasing insight into the process under consideration. Disregard of certain effects in the model belongs to this class. An *error*, however, is a "*recognizable* deficiency in any phase or activity of modeling and simulation that is *not* due to lack of knowledge" (definition and emphasis again as in Alvin et al. [2]). This means, errors can be detected and probably corrected after closer investigation. We can further distinguish here between *acknowledged errors*, which are unavoidable—like rounding in finite precision arithmetics, necessary simplifications, finite iterations, discretization, prescribed error tolerances etc.—the order of magnitude of which can nevertheless be estimated or an upper bound given, and *unacknowledged errors*, which arise inadvertently—like human failure—without possibility of assessing their impact. It is often hard to distinguish between epistemic uncertainties and errors, but nevertheless they are not identical. Ignoring certain effects of the real process on purpose as a simplification of the resulting model is an error, while neglecting them out of unawareness of their existence can be counted as an epistemic uncertainty.

During the modeling process, all necessary choices are usually based on expert knowledge, mathematical and computational considerations, and requirements imposed by cost or security constraints. This means that errors and uncertainties do not only arise in form of the standardly considered measurement errors (measurement noise, image resolution and image processing—mostly aleatoric), but also in the model itself (choice and use of equations, point conditions and inputs, or weak formulations, discretizations, approximative algorithms, and rounding—often epistemic). So not only physical conditions and measuring devices and numerical methods are the sources of insecurities but also the human factor can introduce incertitude by his decisions in problem definition and modeling. This means, expert knowledge cannot only be used to reduce uncertainties in the modeling process, but is also a source of new uncertainties from, e.g., distorted perception or motivational bias (cf. also Booker et al. [28] for suggestions concerning bias minimizing elicitation methods).

But also the specific application that is to be dealt with has influence on the degree of accuracy with which the final computer model reflects reality. Analytical sources of uncertainty are the choice of the mathematical model, for instance, the type and number of equations used to describe the process, the adequacy of this model for the concrete application, and the degree up to which the finite-dimensional implementation on the computer is able to approximate the solution of the real model, cf. the report of the Committee for Verification, Validation and Uncertainty Quantification [38]. We can discriminate here numerical and physical sources of uncertainty, as done by Karniadakis and Glimm [82]. Numerical uncertainties include temporal and spatial discretization errors, erroneous solvers, e.g. due to incomplete iterations or loss of orthogonality, geometric discretizations, including the assumption of piecewise linearity, artificial boundary constraints and so on. Physical uncertainties come from inaccurate or unknown material properties, initial and boundary conditions, random

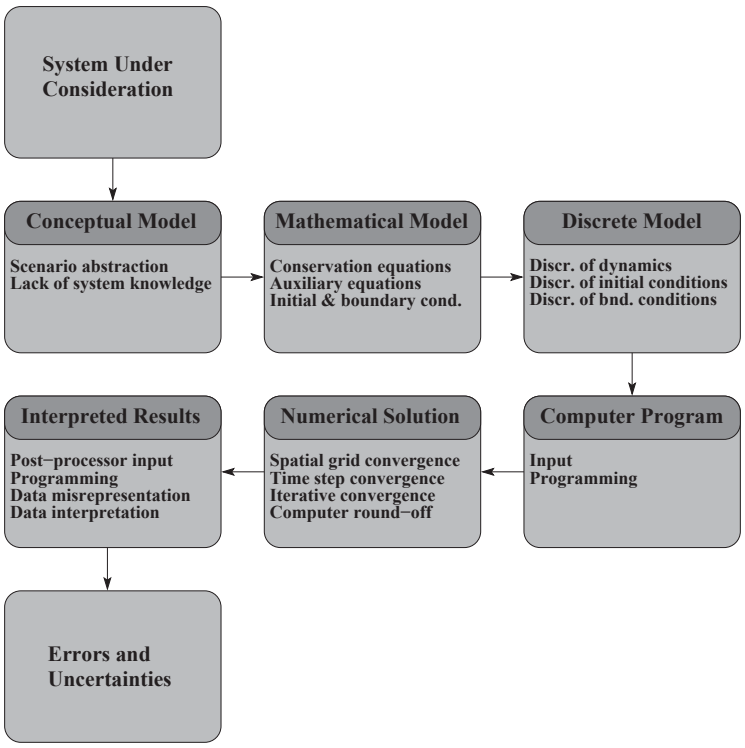


Figure 1.2.: Sources of uncertainties and errors in the different phases of the modeling and simulation process (illustration based on figures 1 and 2 by Alvin et al. [2]).

geometrical variations, e.g. in surface structure, state equations, material laws etc.

As this is rather a lot to consider, let us have a look at the different stages of the modeling and simulation process in turn and the errors and uncertainties associated with each of these. In the outline and description of the stages we follow the presentation for a PDE-based model by Alvin et al. [2]. For each phase we shortly summarize what it is done and where possibly errors might arise. Figure 1.2 shows these stages in a more schematic way.

Prestage The system under consideration is defined, as well as the goal of the simulation. Without a clear definition of the quantities of interest, it is not possible to validate the predictions afterwards with respect to their reliability, see the report of the Committee for Verification, Validation and Uncertainty Quantification [38].

Conceptual modeling The sequence of events in the process is determined with the respective influence factors and their correlations and connections. The analysis of the events being probably carried out without full knowledge of the system under consideration, this phase gives rise rather to uncertainties than to errors.

Two types of uncertainties be mentioned for this stage: scenario abstraction, i.e. incorrect description of the process structure, and lack of system knowledge, i.e. not recognizing even known structures to the full extend.

Mathematical modeling The (differential) equations that analytically describe the physical process are established, together with all initial and boundary or interior point conditions and the like. Usually, both uncertainties and errors originate from this phase. Errors are often caused by over-simplification; uncertainties have their origins in insufficient knowledge about the parameters and about the process itself.

Model discretization The continuous model is transferred into a discrete one, providing for consistency of the models and possible singularities, but without explicitly specifying spatio-temporal stepsizes. Note that this may also be part of the numerical solution process if not done in advance by hand. This phase gives rise rather to errors than uncertainties. These may come from discretization of the continuous dynamics and the initial and boundary conditions. The algorithms and procedures for the solution of the problem are determined in this stage as well.

Programming The developed algorithms and solution strategies are implemented on the computer. We have to be mindful of errors and uncertainties in this phase that come from inexpert programming and false or inaccurate inputs to the program.

Numerical solution The discretized numerical solutions are computed after all relevant quantities for model and procedure have been specified, i.e. grid sizes, material constants, coefficients, rates etc. Alvin et al. [2] divide the errors of this stage into four categories. Spatial grid convergence and time step convergence are both due to the finite discretization of the infinite-dimensional differential equations. Iterative convergence and computer round-off belong to the approximate solution of the discrete equations because of finite machine precision. These latter errors account for the difference that is observed between the exact solution of the discrete equations and the computer solution which has been obtained.

Interpretation Continuous solutions are reconstructed from the discrete data and the results are edited for presentation. Errors and uncertainties can come here from false post-processor inputs, inadequate programming, and misinterpretation or misrepresentation of the results.

We also need to note that errors and uncertainties may enter a model in two very different ways. They may concern either the data and parameters or the model structure. Uncertainties in the measurement data are mostly accounted for in parameter estimation procedures, though standard methods are based on assumptions

on the underlying distributions of the measurement errors that are almost never met in reality. We treat this problem in greater detail in chapter 2. Uncertainties in parameters, which obtain their meaning only in the specific context of the model and the application (cf. Draper [44]), are the field of interest in robust optimization which takes on ever greater significance in practical applications. We attend to this field in chapters 5 and 6. In contrast, uncertainties in the model structure are often neglected. One model is chosen to represent a physical process by just assuming it to be correct. One possibility to deal with structural uncertainty lies in the definition of several model formulations that are equally likely to describe the process correctly and then eliminating wrong models by means of applying a formal model discriminating procedure.

1.2. Stochastic vs. robust optimization

Both stochastic and robust optimization are concerned with optimization under uncertainties. As we attend only robust optimization in this thesis, and neglect stochastic optimization procedures, we define in this section the border that is used to distinguish one from the other. The main difference is the type of uncertainty that is treated by the two methodologies. Stochastic optimization assumes the uncertainty to follow (and be defined by) a known or unknown probability distribution. In contrast, robust optimization is concerned with models that do not contain stochastic but rather deterministic, set-based uncertainties (see also Bertsimas et al. [18]).

As a basic example, let us consider the linear program (LP)

$$\begin{array}{ll} \min\limits_{x} & c^T x \\ \text{s.t.} & Ax \leq b. \end{array} \tag{1.1}$$

The structure of this problem is mainly characterized by the dimensions of the matrix A (over-determined or under-determined system of equality constraints). The available data are the numerical values of (c, A, b). The corresponding uncertain problem is not a single LP, but rather the collection

$$\begin{array}{ll} \min\limits_{x} & c^T x \\ \text{s.t.} & Ax \leq b, \end{array} \qquad (c, A, b) \in \mathcal{A}, \tag{1.2}$$

where the uncertainty set $\mathcal{A}$ comprises all available knowledge of the "true" data. Problems (1.1) and (1.2) share both the structure and the data (c, A, b) from the uncertainty set $\mathcal{A}$. A solution of the uncertain LP (1.2),which is "immunized against uncertainty", is referred to as *robust feasible solution* and denoted by X^*, remaining feasible for all realization of the data in $\mathcal{A}$.

1.2.1. Stochastic optimization

Stochastic optimization is, just like sensitivity analysis, a more traditional approach to treating data uncertainty in optimization problems. The uncertain data is assumed

to be random, with its probability distribution being known in the simplest case, and there is a deterministic counterpart associated with the uncertain LP, for example, a chance constrained problem

$$\begin{array}{ll} \min\limits_{x,t} & t \\ \text{s.t.} & \mathbb{P}_{(c,A,b)\sim P}\big(c^T x \leq t \;\&\; Ax \leq b\big) \geq 1-\varepsilon, \end{array}$$

with a given tolerance $\varepsilon \ll 1$ and a distribution P of the data (c, A, b). In case that the probability distribution is only partly known, i.e. $P \in \mathcal{P}$ with a set $\mathcal{P}$ of possible distributions, an ambiguous chance constrained problem is considered,

$$\begin{array}{ll} \min\limits_{x,t} & t \\ \text{s.t.} & \mathbb{P}_{(c,A,b)\sim P}\big(c^T x \leq t \;\&\; Ax \leq b\big) \geq 1-\varepsilon \quad \forall P \in \mathcal{P}. \end{array}$$

In most cases, stochastic optimization has to rely on over-simplified estimations for the actual distributions, which gives rise to new uncertainties.The influence of these on the quality of the resulting optimal solution is hard to quantify.

In fact, the transition from the chance constrained to the ambiguous chance constrained problem is already a step toward robust optimization. The robust optimization problem as defined in the next section is nothing but an ambiguous chance constrained problem for the family $\mathcal{P}$ of all distributions for a given set $\mathcal{A}$.

A main difficulty with chance constrained problems is that they are often intractable, i.e. not solvable in polynomial time with respect to the size of the input data. Particularly for very small tolerances ε, feasibility of a given solution can be hardly verified. Furthermore, chance constraints mostly lead to non-convex feasible sets which also makes optimization difficult.

1.2.2. Robust optimization

Robust optimization is a complementary approach to stochastic optimization. A historical overview is given by Ben-Tal et al. [14] and Bertsimas et al. [18]. While stochastic optimization immunizes a solution of an optimization problem in a probabilistic sense against stochastic uncertainty, robust optimization constructs a solution that is feasible for any realization of the data uncertainty from a given set. Information about the stochastic nature of the uncertainty can be used as a guideline for the structural design of an appropriate uncertainty set $\mathcal{A}$, such that immunizing one candidate solution against all uncertainty realizations from the set $\mathcal{A}$ automatically leads to the immunization against "almost all" random perturbations (i.e. up to realizations with total probability mass not greater than ε). This makes the robust solution feasible for the chance constrained problem of stochastic optimization.

The value of the objective function of such a robust feasible solution can be interpreted as a quantification of the solution quality by means of giving a guaranteed objective value that will not be exceeded, i.e. the largest value $\sup c^T x$ for $(c, A, b) \in \mathcal{A}$.

Therefore, the most basic approach in robust optimization computes the best robust feasible solution as the solution of the optimization problem

$$\begin{array}{ll} \min\limits_{x} & \sup\limits_{(c,A,b)\in\mathcal{A}} c^T x \\ \text{s.t.} & Ax \leq b \quad \forall (c,A,b) \in \mathcal{A}. \end{array}$$

Note that instead of using the possibly uncertain objective function $\sup c^T x$, we can restrict our attention to problems with a certain linear objective function by rewriting the robust optimization problem as

$$\begin{array}{ll} \min\limits_{x,t} & t \\ \text{s.t.} & c^T x \leq t, \quad \forall (c,A,b) \in \mathcal{A}. \\ & Ax \leq b, \end{array}$$

This problem is also called the *robust counterpart* (RC) of the uncertain LP. Feasible or optimal solutions of the RC are referred to as *robust feasible* and *robust optimal solutions*, respectively. Robust counterparts of uncertain LPs are usually tractable if the underlying uncertainty set satisfies mild assumptions on its convexity and tractability.

This methodology can be applied to any generic optimization problem in which the numerical data, which are partially uncertain and only known to lie in a given uncertainty set, and the problem structure, which is known in advance and common for all instances of the uncertain problem, can be separated. See Ben-Tal et al. [14] for a list of problem classes for which robust optimization in this form is applicable. Tractability issues for the corresponding counterparts are also addressed there.

Note, however, that the given approach to robust optimization may be more conservative than stochastic optimization methods in some cases, where the uncertain data is of stochastic nature with the corresponding probability distribution being available, and where the probabilistic guarantees of the chance constraints are sufficient. These assumptions may be satisfied for some problems like, e.g., signal processing, but are much too restrictive for other applications. In engineering applications, decisive design parameters are often increased by 30-50% to saveguard against possible modeling inaccuracies or unexpected environmental influences (cp. Ben-Tal et al. [14]). Robust optimization meets this necessity for "staying on the safe side" by increasing the uncertainty set. With chance constrained problems, in contrast, the total amount of uncertainty is fixed as the total probability mass of all realizations of the uncertain data must equal 1. This means that in stochastic optimization increasing one uncertainty requires the decrease of another. Therefore, an ambiguous chance constrained problem is needed for such a task and, as we said above, this means a step toward robust optimization.

We will go into more detail with robust optimization and the computation of robust counterparts in chapter 5.

An alternative approach to this worst case robust optimization is described in chapter 6. It is especially useful when there are no security relevant constraints that

have to be met under all circumstances but rather an average solution is wanted as is most probable to occur for the majority of realizations of the uncertainty. For this approach it does not even matter if the uncertainty is set-based or described by its probability distribution as it will work in both cases though we restrict our considerations in chapter 6 to the case of deterministic uncertainties. Low-dimensional examples using this approach even for uncertainty quantification for steady-state single phase flow in randomly heterogeneous porous media, given as a Poisson equation, can be found in Lin et al. [92].

1.3. Outline of the thesis

Having thus introduced the problematic nature of numerical simulation and optimization, we will now give the outline of this thesis. It is divided into three parts: the first part in concerned with the treatment of measurement errors; the second part deals with uncertainties in model parameters; and the third part contains some illustrative applications.

Part I consists of three chapters. The first one, i.e. chapter 2 of this thesis, gives an introduction to parameter estimation problems. Both standard least squares estimation and robust estimation procedures are described and compared to one another. The drawbacks of least squares estimation are outlined and the ideas of the robust methods to overcome these problems are explained. Chapter 3 describes the adaption of the generalized Gauss-Newton method to the computation of the robust estimators. Local convergence properties are given and strategies for globalization of the method are sketched. In chapter 4 the methods of the previous chapter are generalized for the use of inexact derivatives. Again, local convergence is proved for inexact Jacobians that are in some sense "close enough" to the true one.

Part II is also divided into two chapters. The first one here, i.e. chapter 5 in the consecutive numbering, introduces the worst-case approach of robust optimization which already has its own tradition in the presence of model or parameter uncertainties. The following chapter 6 is dedicated to an alternative approach which employs a collocation method, as has been used for partial differential equations (PDEs) with stochastic coefficients before. This method leads to less conservative results than the standard worst-case approach while still strictly satisfying all constraints of the optimization problem.

Part III contains two more chapters which are related to the first two parts of this thesis. It treats problems, mainly from biology, that serve to illustrate the necessity for robust methods and their advantages over non-robust procedures. Chapter 7 describes applications of the robust parameter estimation procedures from part I and chapter 8 displays the solutions of some example problems with the methods from part II.

The thesis closes with a final chapter 9 that summarizes the findings of the prevailing chapters and offers an outlook on open questions for future consideration.

Part I.

Errors in Measurements

The first part of this thesis is concerned with robust optimization procedures that take into account the errors that are naturally inherent to measurements from experiments. We start with a chapter that reviews the development of parameter estimation procedures and how the focus changed from classical procedures like least squares to more robust methods that are less sensitive to measurement errors. It follows a chapter about the robust Gauss-Newton (GN) method, i.e. the computation of the robust estimators introduced in the previous chapter with the GN method. The last chapter of this part is dedicated to the problem how inexactly computed derivatives, which might distort the results of the GN method, can be compensated by adding an additional term to the objective function.

2. Parameter estimation: From classical to robust procedures

Parameter estimation is an important part in joining scientific theory and experimental observations. The mathematical models which are supposed to describe physical, chemical, biological, or some other processes contain unknown parameters which have to be determined in such a way that the model reproduces the data as good as possible.

All parameter estimation procedures are based on, possibly implicit, assumptions on the statistical distribution of the underlying data set. The well-known least squares method, for instance, was developed by Gauss [57] to match the standard normal distribution. As it has been known from the beginning of the twentieth century that even "good" samples are only approximately normally distributed but tend to be rather longer-tailed (see, e.g., Hampel [69]), there is need for alternative procedures which are better suited to cope with non-Gaussian distributions.

It is commonly agreed that an a priori correction of the data to eliminate outlying observations is not the way to proceed. Huber and Ronchetti [81] state two major reasons (amongst others) why this approach should be avoided:

- An incorrect determination of outliers will make the situation rather worse as there will ususally be both false rejections and false retentions.

- It is not always obvious what the outlying points are because several outliers may obscure each other and so escape detection.

Nevertheless, outliers have to be treated efficiently in order not to spoil the estimation from the beginning.

When building a mathematical model, we make the implicit assumption that small model errors (which can always be expected to arise from the necessary simplifications in the modeling process) will cause only minor errors in the conclusions. Beside that, there are also always basic assumptions on the underlying situation like randomness, independence, or the distribution of the errors. Neither of these assumptions has to be justified. But many common statistical procedures, least squares estimation being among them, are very sensitive to apparently marginal deviations from the underlying assumptions. A remedy for this problem are so-called "robust" methods. In this context, robustness means "insensitivity to small deviations from the assumptions" (cf. Huber and Ronchetti [81]) and that is also the definition on which this thesis is based.

We will be particularly concerned with distributional robustness, i.e. dealing with such problems where the form of the true underlying distribution slightly differs from the assumed models. This is the simplest case and also the best understood. About the deviations from other standard assumptions, like independence of the errors, much less is known, which leaves a huge space for further investigations. Deviations from the assumption of identically distributed observations seem to be much less critical, at least for methods that are symmetric in the observations, cf. Huber [77].

The goal of robust procedures is the spotting of outliers in a given data set and the reduction of their influence in the parameter estimation. As there are many different, more or less effective, means for this task and as different types of outliers have a varying degree of influence and different demands on the estimators, we will start this chapter with a section shortly introducing some of the most important statistical concepts in the estimation field. In the second section we consider the most general problem formulation we treat in this thesis and which constitutes the framework for the following theoretical observations. This problem may contain dynamical constraints of which the numerical treatment will be the topic of the third section. Then we define the common least squares parameter estimation problem in the fourth section of this chapter before we come to robust parameter estimators in the fifth section. In the two sections following this introductory description of robust estimation, we define the l_1 norm estimator and Huber's M-estimator and give reformulations of their (non-differentiable) objective functions as a linear and a quadratic function plus constraints, respectively. The eighth section is devoted to a comparison of the basic features of the three presented estimators, i.e. least squares, least absolute deviations, and Huber's estimator. Finally, we close this chapter with a section containing an overview of other robust estimators together with references to the relevant literature.

2.1. Statistical background

Assume that for any number n we have a set Z of random variables $z_1, \ldots, z_n$ and a corresponding estimator $T_n(z_1, \ldots, z_n)$ which yields an estimate of the parameter p. The estimator is called *unbiased* if its expected value equals the estimated parameter, i.e. if $\mathbb{E}(T_n) = p$. Otherwise, the difference $\mathbb{B}_{T_n}(p) := \mathbb{E}(T_n) - p$ is called the *bias* of the estimator T_n. Note that even maximum likelihood estimators are not necessarily unbiased, cp. Georgii [60]. Another important property for all estimators, robust or not, is their *consistency*. The sequence $\{T_n\}_n$ is then said to be consistent if for increasing sample size n it converges to the true parameter p. The formal definition, as given by Georgii [60], is that the probability that the absolute difference between the estimator T_n and the parameter p equals at most some $\varepsilon > 0$ goes to one as n goes to infinity,

$$\mathbb{P}\big(|T_n - p| \leq \varepsilon\big) \overset{n\to\infty}{\longrightarrow} 1 \qquad \forall\, \varepsilon > 0.$$

The *breakdown point* (BDP) of an estimator is a global measure of its resistance. It is formally defined as the smallest fraction or percentage of unusual observations

that an estimator can cope with without giving an arbitrary result (see, e.g., Hampel [70] or Andersen [3]). Assume that we have a finite sample Z of random variables and an estimator T. In this sample Z we randomly replace m data with arbitrary values, getting a new sample Z'. The effect of this replacement is defined as the supremum of the distance between the estimator for Z and that for Z' over all such replacements,

$$\text{effect}(m; T, Z) := \sup_{Z'} \|T(Z') - T(Z)\|_2. \tag{2.1}$$

Then the breakdown point of T is defined as the smallest fraction m/n for which the effect is unbounded,

$$\text{BDP}(T, Z) := \min \left\{ \frac{m}{n} \mid \text{effect}(m; T, Z) \text{ is infinite} \right\}. \tag{2.2}$$

Obviously, the highest possible breakdown point is 0.5 or 50%, respectively, and a breakdown point greater than zero seems to be a desirable property for a robust estimator. We will come back to this last statement later on.

In contrast to this global concept, the *influence function* (IF) or *influence curve* defined by Hampel [70] is a measure of an estimator's local resistance. It describes the impact that a single observation y contaminating the theoretically assumed distribution p has on the estimator T. Formally, it is defined as

$$\text{IF}(y, \text{p}, T) := \lim_{\lambda \to 0} \frac{T\left((1-\lambda)\text{p} + \lambda\delta_y\right) - T(\text{p})}{\lambda}, \tag{2.3}$$

where δ_y is the measure of contamination, i.e. 1 at the point y and 0 everywhere else, with probability mass λ. So the influence function indicates the change in an estimate caused by adding an arbitrary outlier at the point y, standardized by the proportion of contamination. Thus it should also stay bounded for a robust estimator. Also note the similarity of definition (2.3) to a partial derivative. In most cases the influence curve will simply be proportional to the first derivative of the objective function of an optimization problem (cf. Hampel [70]).

Another important aspect is the *efficiency* of an estimator. In the strictest sense, this is the ratio of the minimum possible variance of an estimator to its actual variance for a given problem. So this ratio should be one for a truely efficient estimator. *Asymptotic efficiency* is concerned with the efficiency of an estimator for large sample sizes. Different estimators may have varying efficiency under changing conditions. Therefore, the most important concept when the underlying error distribution is unknown, is *relative efficiency*. Suppose that under certain assumptions an estimator T_1 has maximum efficiency while another estimator T_2 is less efficient. Then the relative efficiency of T_2 compared to T_1 is defined as the ratio of the efficiency of T_2 to the efficiency of T_1 and can be computed as the ratio of the corresponding mean squared errors,

$$\text{efficiency}(T_1, T_2) := \frac{\text{MSE}(T_2)}{\text{MSE}(T_1)} = \frac{\mathbb{E}(\|T_2 - \hat{p}\|^2)}{\mathbb{E}(\|T_1 - \hat{p}\|^2)}, \tag{2.4}$$

where the mean squared error MSE is defined as

$$\mathrm{MSE}(T) := \mathbb{E}\left(\|T - \hat{p}\|_2^2\right)$$

and $\hat{p}$ denotes the true value of the parameter that is to be estimated.

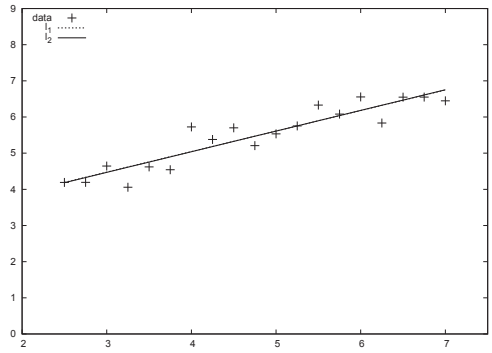
Note that this definition is meaningful only for large sample sizes. So it is a favorable feature of robust estimators to have a high relative efficiency with respect to the ordinary least squares estimator in order not to be too inefficient for a Gaussian error distribution.

Now we want to consider the outlying observations in question a little closer. Generally, there are distinct concepts for outliers that categorize their impact on an estimation, see e.g. Andersen [3]. A *univariate outlier* is an observation that stands apart from the rest of the distribution of a particular variable. Univariate outliers may occur both in the dependent and independent variables. They are often easy to detect and not necessarily problematic for the estimation. A *regression outlier* (or *vertical outlier*) is an observation that differs from the general pattern for the bulk of the data, meaning that it has a discrepant value in terms of the dependent variable. Typically, regression outliers are characterized by large residuals in least squares estimation though this is neither a necessary nor sufficient criterion for their identification. On the other hand, a *leverage point* is an observation with an unusual value in the independent variable. The further the observation stands away from the mean value of the independent variable, the greater is its leverage. Note, however, that also here there are "good" and "bad" leverage points. The most important quality of an observation is its *influence* which is mainly a combination of the discrepancy in the value of the dependent variable and its leverage. Its characteristic feature is that exclusion of influential points from the estimation always changes the estimates substantially. The following figures 2.1 demonstrate the different types of outliers for the example of a simple linear regression with the standard least squares (LS) estimator, which minimizes the squared l_2 norm of the deviations of the model response from the data, and the least absolute deviations (LAD) estimator, which minimizes the l_1 norm instead. We will go into further detail on these estimators in the following sections. The data were generated as points on the line $f(x) = \frac{2}{3}x + \frac{7}{3}$ with 5% white noise added to the function values. The numerical results of the estimation are given in table 2.1.

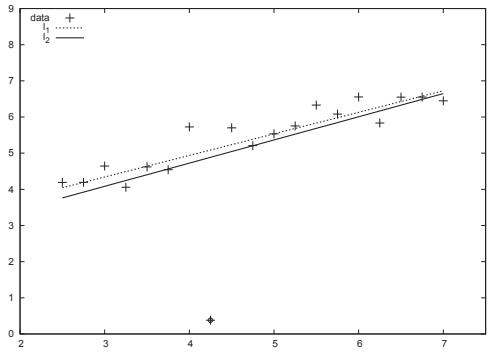
2.2. The problem formulation

Nonlinear differential equations play an important role in describing dynamic processes in the natural sciences as well as in medicine, engineering, or other application areas. Thus the most general problems we want to deal with in this thesis are processes that can be modeled by systems of nonlinear differential algebraic equations (DAEs) of the form

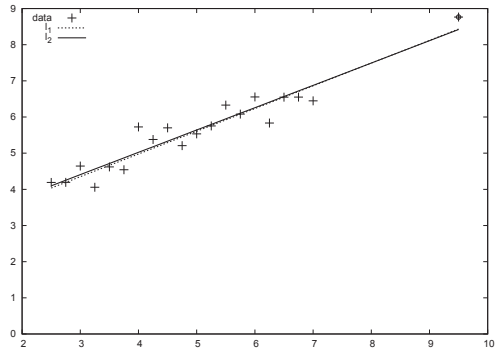
$$\begin{aligned} \dot{y}(t) &= f(t, y(t), z(t), p), \\ 0 &= g(t, y(t), z(t), p), \end{aligned} \tag{2.5}$$



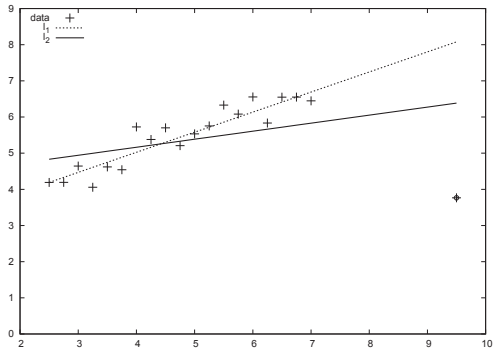
(a) If there are no outliers in the (normally distributed) data, least squares and least absolute deviations estimation give virtually the same result.



(b) A vertical outlier pulls the least squares estimate away from the data, but has only moderate influence, while it hardly affects the LAD estimate.



(c) A good leverage point rather improves the results for l_2 and l_1 estimation as it provides important additional information about the regression line.



(d) A bad leverage point alters the least squares estimate completely. In this example the least absolute deviations estimator is hardly affected though, if the leverage point lies sufficiently far away from the rest of the data, the LAD fit may even pass through it.

Figure 2.1.: Illustrative examples for the different types of outliers in the data for a simple linear regression with least squares and least absolute deviations estimator. Note that Huber's M-estimator will give results somewhere inbetween the l_2 and l_1 estimates, depending on the value of the tuning constant.

Table 2.1.: Numerical results for the examples concerning different types of outliers in a linear regression using least squares and least absolute deviations estimation.

	LS	LAD
m	0.571048	0.567881
c	2.756083	2.772490
obj. val.	4.741401	4.738015

(a) Estimation without outliers for $f(x) = mx + c$.

	LS	LAD
m	0.641223	0.596170
c	2.159592	2.553043
obj. val.	10.493292	9.251066

(b) Estimation with vertical outliers for $f(x) = mx + c$.

	LS	LAD
m	0.617288	0.628321
c	2.554693	2.464625
obj. val.	5.258736	5.139959

(c) Estimation with good leverage point for $f(x) = mx + c$.

	LS	LAD
m	0.221867	0.555235
c	4.276857	2.804103
obj. val.	12.623263	9.073248

(d) Estimation with bad leverage point for $f(x) = mx + c$.

with t from some given interval $[t_0, t_f]$. We consider t to represent *time*, although it could be a spatial direction as well, and we denote by $x(t) = (y(t), z(t)) \in \mathbb{R}^{n_x}$ the *state variables*, combining the *differential variables* $y \in \mathbb{R}^{n_y}$ and the *algebraic variables* $z \in \mathbb{R}^{n_z}$. The vector $p \in \mathbb{R}^{n_p}$ contains the unknown parameters, the values of which have to be determined by parameter estimation.

We restrict our considerations to DAEs of index 1 (i.e. $\partial g/\partial z$ is regular). More general dynamics, e.g. differential equations of higher order, DAEs of higher index, and the like, can often be transformed to the form (2.5).

It is further assumed that experiments have been performed and at given times $t_j,\ j = 1, \ldots, M$, measurements $\eta_{ij}, i = 1, \ldots, M_j,\ j = 1, \ldots, M$, have been recorded that correspond to observation functions m_{ij} which depend on the state variables $x(t) = (y(t), z(t))$ and the parameters p. These measurements are subject to measurement errors ε_{ij}, that is

$$\eta_{ij} = m_{ij}(t_j, x^{true}(t_j), p^{true}) + \varepsilon_{ij}, \tag{2.6}$$

where p^{true} denotes the "true" values of the parameters and x^{true} the corresponding state variables. Note that this problem formulation also allows for the case of several measurements being taken at the same time t_j.

The parameters are then estimated by minimizing the deviation between the model and the data,

$$\min_{p,x} \rho(r_1(x(t_1), \ldots, x(t_M), p)) := \rho \begin{pmatrix} \vdots \\ \frac{\eta_{ij} - m_{ij}(t_j, x(t_j), p)}{\sigma_{ij}} \\ \vdots \end{pmatrix}, \tag{2.7}$$

with appropriate weights σ_{ij}. Obviously, the observation functions m_{ij} may depend both on the differential and the algebraic variables. If it only depends on the algebraic z, then $u_{ij}(t) = m_{ij}(t, z(t), p)$ can simply be considered as additional algebraic variable. On the other hand, if m_{ij} only depends on the differential variables y, then $u_{ij}(t) = m_{ij}(t, y(t), p)$ can be treated as differential variable by adding the equation

$$\dot{u}_{ij} = (m_{ij})_y \dot{y} + \dot{m}_{ij} = (m_{ij})_y f + \dot{m}_{ij}$$

to the differential equations system, with $(m_{ij})_y = \left(\frac{\partial m_{ij}}{\partial y_1}, \ldots, \frac{\partial m_{ij}}{\partial y_{n_y}}\right)$ being a row vector, cf. Betts [19]. However, we do not follow this approach of enlarging the DAE system here.

The choice of the optimization criterion ρ is essential. It has to be chosen in such a way that the distance of the estimates $\bar{x}, \bar{p}$ from the true values is small, i.e. the values $\|\bar{x} - x^{true}\|_2$ and $\|\bar{p} - p^{true}\|_2$ are the important criteria for the quality of a fit rather than the actual value of the objective function (2.7). But what is the best optimization criterion for a given problem? The answer to this question depends on the probability distribution of the errors ε_{ij}. The efficiency of the various optimization criteria varies heavily for different distributions and there is none that is good (or even mediocre) for all situations, see e.g. Rice and White [112]. An experimental comparison of several l_p norms as optimization criterion ρ for different error distributions can be found in that paper, too.

The most popular approach is the use of the squared l_2 norm, the so-called least squares estimation, which yields a *maximum likelihood estimate* of the parameters if the measurement errors are independent, normally distributed with zero mean and known variances, $\varepsilon_{ij} \in \mathcal{N}(0, \sigma_{ij}^2)$. Another common estimator is defined by the use of the l_1 norm which yields a maximum likelihood estimate if the errors follow a Laplace distribution, see e.g. Birkes and Dodge [21]. We will present the most important features of both methods in the following sections. Additionally we will also consider the use of the Huber function,

$$\rho(r_1) = \sum_i \rho_\gamma(r_{1i}) \text{ with } \rho_\gamma(z) = \begin{cases} \frac{1}{2}z^2, & \text{if } |z| \le \gamma, \\ |z|\gamma - \frac{1}{2}\gamma^2, & \text{if } |z| > \gamma, \end{cases}$$

which is not a norm itself but can be seen as a combination of l_2 and l_1 norm optimization.

Additional information about the state variables and the parameters, like initial or boundary conditions, parameter bounds, positivity, and so on, can be included in the optimization by means of equality and inequality constraints at times θ_i, $i = 1, \ldots, K$,

$$\begin{aligned} r_2(x(\theta_1), \ldots, x(\theta_K), p) &= 0, \\ r_3(x(\theta_1), \ldots, x(\theta_K), p) &\ge 0. \end{aligned} \tag{2.8}$$

So all in all we get the following general parameter estimation problem:

$$\begin{array}{ll} \min\limits_{x,p} & \rho(r_1(x(t_1),\ldots,x(t_M),p)) \\ \text{s.t.} & (x,p) \text{ solves DAE (2.5)}, \\ & r_2(x(\theta_1),\ldots,x(\theta_K),p) = 0, \\ & r_3(x(\theta_1),\ldots,x(\theta_K),p) \geq 0. \end{array} \tag{2.9}$$

Note that beside measurement data, the user also has to provide standard deviations σ for the parameter estimation problem.

Before we describe the different parameter estimators mentioned above in more detail, we consider in the next section the numerical treatment of the dynamic constraints.

2.3. Numerical treatment of the dynamics

In problem (2.9) we face the problem how to solve the differential equations which define the state variables. We cannot solve the involved DAEs in advance or independent of the parameter estimation problem, both because they depend on the unknown parameters and because the implicitly defined states are to be optimized as well.

In this section we will discuss two different approaches, an initial value problem approach which treats the problem iteratively by alternate solution of the DAE as an initial value problem and subsequent improvement of the objective function and satisfaction of the constraints, and a boundary value problem approach which discretizes the dynamics and thus solves the differential equations and the optimization problem in one go.

2.3.1. The initial value problem approach

A common solution approach to parameter estimation problems with differential equations consists in the repeated solution of the DAE system as an initial value problem, i.e. the differential equation system is solved with estimated initial values and parameters, in combination with an iterative improvement of these estimates by some optimization procedure in order to minimize the objective function and to satisfy the remaining constraints (see e.g. Novak and Deuflhard [105] and Schittkowski [119]). Algorithmically this amounts to guessing initial parameter estimates and subsequently repeating the loop

(i) integrate the DAEs along the complete time horizon,

(ii) compute the values of the objective function and the constraints,

(iii) correct the parameter values to decrease the objective value,

until some termination criterion is satisfied.

The result and also the convergence properties of this approach heavily depend on the used integrator and optimization procedure, and the combination of these, respectively. Usually, the initial value problem approach is only used in practice when initial conditions are given explicitly for the differential equations. But even in the best possible circumstances the approach is rather inefficient as it is not capable of using all available data directly, e.g. interior point conditions, measurement values within the integration interval, and the like. Often the initial value problem is hard or may even be impossible to solve for certain initial guesses of the parameters and initial values so that severe stability problems may arise (see, e.g., Bock [24] and Betts [19]).

2.3.2. The boundary value problem approach

A numerically more stable approach is the boundary value problem or "all at once" approach (e.g. Bock [22, 23, 24], Schlöder [120]). Its basic idea is the parameterization of the dynamics as a boundary value problem, e.g., by collocation (Bär [7], Ascher [6]) or multiple shooting (Bock [22], Bulirsch [32]), and the simultaneous minimization of the objective function subject to the given constraints and additional nonlinear constraints induced by the discretized boundary value problem.

A great advantage of this approach is that it allows incorporation of information about the course of the DAE solution. Thus the influence of bad initial guesses for the parameters is reduced as they cannot propagate along the whole integration interval but are limited to the respective multiple shooting interval. This enables parameter estimations even for instable or chaotic DAE systems (Bock et al. [25]).

The solution methods used in this thesis are based on multiple shooting, so we give a short description of the resulting procedure in case of a general DAE model.

The whole integration interval $[t_0, t_f]$ is divided into N multiple shooting intervals $[\tau_i, \tau_{i+1}],\ i = 0, \ldots, N-1$, with suitably chosen multiple shooting nodes τ_i,

$$t_0 = \tau_0 < \tau_1 < \cdots < \tau_N = t_f.$$

At each node τ_j additional unknowns $s_j = (s_j^y, s_j^z)$ are introduced representing the values of the state variables in this point and N relaxed DAE initial value problems

$$\dot{y} = f(t, y, z, p), \tag{2.10a}$$
$$0 = g(t, y, z, p) - \alpha(t) g(\tau_i, s_i^y, s_i^z, p), \tag{2.10b}$$

with initial conditions $y(\tau_i) = s_i^y,\ z(\tau_i) = s_i^z$ are solved on each subinterval $[\tau_i, \tau_{i+1}]$. The function $\alpha(t)$ is a decreasing function with $\alpha(\tau_i) = 1,\ \alpha(t) \geq 0$ in order to improve consistency of the algebraic equations in the course of the integration. These problems yield solution trajectories $x(t; s_i, p)$ for $t \in [\tau_i, \tau_{i+1}]$. Usually the solutions are neither continuous at nodes τ_j nor feasible for the original algebraic equations.

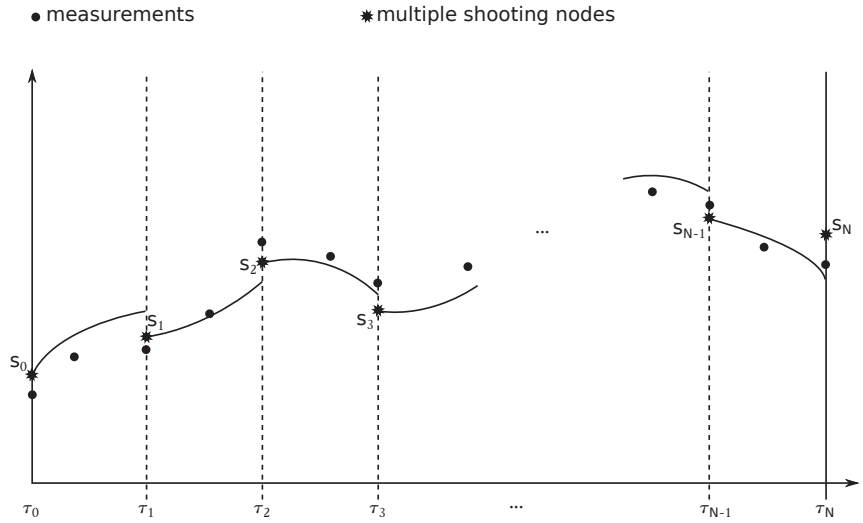


Figure 2.2.: Concept of the multiple shooting procedure for an ordinary differential equation.

We enforce continuity and feasibility by additional matching conditions,

$$\begin{aligned} h_j(s_j, s_{j+1}, p) := y(\tau_{j+1}; s_j, p) - s_{j+1}^y = 0, \quad j = 0, \ldots, N-1, \\ g(\tau_j, s_j, p) = 0, \quad j = 0, \ldots, N. \end{aligned}$$

The concept of the multiple shooting routine is depicted in figure 2.2.

By inserting the computed values $x(t; s_i, p),\ t \in [\tau_i, \tau_{i+1}]$, into the parameter estimation problem (2.9), we obtain a nonlinearly constrained optimization problem in the variables $(s, p) = (s_0, \ldots, s_N, p)$,

$$\begin{aligned} \min_{(s,p)} \quad & \rho(r_1(s,p)) \\ \text{s.t.} \quad & r_2(s,p) = 0, \\ & r_3(s,p) \geq 0, \\ & h_j(s_j, s_{j+1}, p) := y(\tau_{j+1}; s_j, p) - s_{j+1}^y = 0, \quad j = 0, \ldots, N-1, \\ & g(\tau_j, s_j, p) = 0, \quad j = 0, \ldots, N. \end{aligned} \tag{2.11}$$

So after parameterizing the boundary value problem, we end up with a finite dimensional, possibly large-scale, optimization problem (2.11) which belongs to the general class of nonlinear constrained approximation problems of the form

$$\begin{aligned} \min_X \quad & \rho(F_1(X)), \\ \text{s.t.} \quad & F_2(X) = 0, \\ & F_3(X) \geq 0, \end{aligned} \tag{2.12}$$

with $F_1 : \mathbb{R}^n \to \mathbb{R}^{m_1}$, $F_2 : \mathbb{R}^n \to \mathbb{R}^{m_2}$, $F_3 : \mathbb{R}^n \to \mathbb{R}^{m_3}$. Here the equality constraints $F_2(X) = 0$ also include the matching conditions induced by multiple shooting. The

number of variables n in problem (2.12) equals the number of differential and algebraic variables multiplied by the number of multiple shooting nodes plus the number of parameters. Note, however, that the method can be implemented in such a way that neither programming nor computational effort exceed that for the single shooting approach. Instead, stability and efficiency are greatly improved, cf. Bock [22]. The special structure of the equations suggests the approach for parallelization techniques (Bock et al. [25]). Also solution methods for sparse quadratic programs may be applied as described by Betts [19].

Throughout this thesis we assume the functions $F_i,\ i = 1, 2, 3$, to be at least twice continuously differentiable.

2.4. Classical least squares estimation

The most commonly used method for data fitting by means of a parameter estimation is the least squares estimator which uses the squared l_2 norm in the objective function,

$$\rho(F_1(X)) = \frac{1}{2}\|F_1(X)\|_2^2 = \frac{1}{2}F_1(X)^T F_1(X). \tag{2.13}$$

It was developed independently by Gauss [57] and Legendre [90] in the first decade of the 19th century.

For the case of statistically independent measurement errors ε_{ij} as introduced in (2.6) that follow a normal distribution with zero mean and constant variance σ_{ij}^2, i.e. their probability density is given by

$$f(\varepsilon_{ij}) = \frac{1}{\sigma_{ij}\sqrt{2\pi}} \exp\big(-\frac{\varepsilon_{ij}^2}{2\sigma_{ij}^2}\big),$$

the least squares estimator yields a maximum likelihood estimate of the parameters. Already the statistical derivation of this method by Gauss [57] relies heavily on these assumptions about the nature of the measurement errors. If a problem does not meet these assumptions, the least squares estimates will be biased as the squared residuals make this estimator very sensitive to outliers in the measurement data. These outliers may or may not be characterized by unproportionally large residuals but have an inappropriately high influence on the estimates. Other estimators do suffer as much from this problem, cf. Watson [138].

The minimization of squared deviations necessitates that very large deviations of the measurements from the resulting estimator are avoided. Otherwise, the consistency of the model with some good points is sacrificed for the accomodation of outliers. This makes real outliers even harder to identify and the harm is even worsened by the fact that the inclusion of outliers leads to significant errors in the parameter estimates.

2.5. Need for robust parameter estimation

Gross errors in the measurement data can arise from different sources. These can be measurement errors from inaccurate measurement devices as well as misplaced decimal points or copying errors, i.e. human inaccuracies. But they may also arise from unintentional measurements of a member of a different population or simply "something went wrong" (Hampel [69]). Gross errors can occur in form of clear outliers or as hidden contamination which usually cannot be detected. The frequency of occurence strongly depends on the quality of the data. A quantity of 5-10% wrong values are the rule rather than the exception (cf. Hampel [69]). Even high quality data are slightly but still conspicuously longer-tailed. As Geary [59] puts it: "*Normality is a myth; there never was, and never will be, a normal distribution.*"

But even small and innocent-looking deviations from the assumptions of an estimator (e.g. randomness, independence, distribution models, distribution of certain unknown parameters, etc.) can render the classical "optimal" methods inefficient and bad. Most statistical methods, in particular those that are optimized for an underlying normal distribution, react very sensitive to seemingly negligible deviations from these assumptions, cf. Huber and Ronchetti [81]. So in least squares estimation a normal distribution of the measurement errors with zero mean and known variance is assumed. In real-life applications, however, measurement errors tend to suffer from the reasons for gross errors mentioned in the last paragraph. Thus the normality assumption will usually not hold and least squares will not be a maximum likelihood estimator in this case. A remedy is provided by so-called "robust" optimization methods, where "robustness" means "insensitivity to small deviations from the assumptions" (Huber and Ronchetti [81]).

The effect of minor outliers can be simulated by means of a mixture model based on the standard normal distribution Φ with contaminating ϵ-fractions of normal (or just symmetric) distributions H,

$$\mathrm{p}(x) = (1-\epsilon)\Phi(x) + \epsilon H(x), \tag{2.14}$$

with $\epsilon \in (0,1)$ small.

Using non-robust estimators like least squares, small errors and inaccuracies are completely overlooked. It is possible to secure against large losses when the data are obviously non-normal by transformations and more or less formal rejection procedures neglecting obvious outliers or treating them separately, respectively. Still the question remains what the criterion should be that identifies gross errors from "good" data. Errors may even hide one another, making each other look like belonging to the bulk of data. Another danger lies in automatic data processing on computers without previous data analysis.

The main goal of robust estimation is the incorporation of safeguards against an unexpectedly large amount of gross errors, the bounding of the influence of hidden contamination and questionable measurements, as well as the isolation of clear outliers for separate treatment. Nevertheless, they shall still give approximate optimality

when, contrary to expectations, the standard assumptions, like normality of the errors, are strictly met. Robust estimation considers both small, barely detectable, and large, obvious contamination of the model. The main interest is concentrated on the bulk of data and procedures for its determination.

In practical applications, avoiding the worst is often more important than optimizing one or more aspects at all costs, cf. Hampel [69]. Also Anscombe [5] and Huber [77] consider robustness as a kind of insurance, i.e. a compromise sacrificing some efficiency (5-10% for the ideal model, say) in order to secure against bad effects from small deviations.

Two widely used robust estimators, namely the least absolut deviations estimator and Huber's M-estimator, are introduced in the following sections.

2.6. Least absolute deviations estimation

When estimating parameters by l_1 norm minimization, the optimization criterion is

$$\rho(F_1(X)) = \|F_1(X)\|_1 = \sum_{i=1}^{m_1} |F_{1i}(X)|. \tag{2.15}$$

We already mentioned above that l_1 norm minimization yields a maximum likelihood estimate of the parameters if the measurement errors ε_{ij} follow a Laplace distribution with zero mean and variance $2\sigma_{ij}^2$, i.e. their probability density is

$$f(\varepsilon_{ij}) = \frac{1}{2\sigma_{ij}} \exp\big(-\frac{|\varepsilon_{ij}|}{\sigma_{ij}}\big).$$

But even more important for our present purpose is its property to be insensitive to outliers in the measurement data (cp. Huber [78]) and therefore to be a robust estimation method in the sense of the previous section.

This robustness follows from the fact that a minimization of the l_1 norm interpolates the $\bar{n}$ "best" measurements, where $\bar{n}$ denotes the degrees of freedom of the optimization problem, and neglects all other data.

Historically, l_1 norm minimization dates back to 1755 when Boscovich investigated the figure of the earth from geodatic measurements in joint work with Maire (cf. Scales and Gersztenkorn [118]). Laplace [89] also used the methods developed by Boscovich in the following year and advanced the ideas on a real statistical basis. This means that the least absolute deviations (LAD) estimator dates back even longer than least squares estimation, cp. section 2.4. We want to mention, however, that Gauss [57] also described an algebraic method for computing the l_1 solution of an overdetermined linear equations system.

Until the middle of the 1970s, most solution methods for LAD minimization for overdetermined linear systems were based on relatively slow extensions of Gauss's method, e.g. Barrodale and Roberts [8, 9]. An iteratively reweighted least squares algorithm was published in 1973 by Schlossmacher [121].

The problem in the computation of the LAD estimator is the fact that (2.15) is a non-smooth function. This contradicts the standard assumption of the common optimization methods that all occuring functions are suffiently smooth in order to compute the derivatives needed in the course of the optimization. But by introducing additional optimization variables it is possible to reformulate the unconstrained problem

$$\min_X \|F_1(X)\|_1 = \sum_{i=1}^{m_1} |F_{1i}(X)|$$

as an optimization problem with equality and inequality constraints and linear objective function,

$$\begin{array}{ll} \min\limits_{X\in\mathbb{R}^n, u\in\mathbb{R}^{m_1}, v\in\mathbb{R}^{m_1}} & \sum\limits_{i=1}^{m_1}(u_i+v_i) \\ \text{s.t.} & F_{1i}(X) = u_i - v_i, \quad i=1,\ldots,m_1, \\ & u_i, v_i \geq 0, \quad i=1,\ldots,m_1. \end{array} \tag{2.16}$$

Note that in case that F_1 is a vector of linear functions, we even get a linear optimization problem this way.

Of course this reformulation is also possible in the presence of equality and inequality constraints in the original problem. But the introduction of the new variables u and v leads to an unnecessary enlargement of the problem dimension. Therefore, for practical purposes it may be preferable to use methods which are able to cope with the non-differentiability of the objective function (2.15). But the reformulation (2.16) will be of use for the following theoretical considerations.

2.7. Huber estimation

For Huber's M-estimator the objective function takes the form

$$\rho(F_1(X)) = \sum_{i=1}^{m_1} \rho_\gamma(F_{1i}(X)) \text{ with } \rho_\gamma(z) = \begin{cases} \frac{1}{2}z^2, & \text{if } |z| \leq \gamma, \\ |z|\gamma - \frac{1}{2}\gamma^2, & \text{if } |z| > \gamma. \end{cases} \tag{2.17}$$

Obviously, we evaluate a least squares term if the absolute value of the argument is less than a predefined bound γ and an l_1 term if it is greater. In the context of this thesis, where we want to fit a model function to given measurements by minimizing the deviation of the model prediction from the data which corresponds to minimizing the measurement error, this tuning constant γ can be computed from the assumed percentage of errors ϵ in the data by means of a nonlinear function (see Huber [78] and Eckblom and Madsen [45]) which reads

$$(1-\epsilon)^{-1} = \int_{-\gamma}^{\gamma} \phi(t)\,dt + 2\phi(\gamma)/\gamma, \tag{2.18}$$

where ϕ denotes the standard normal density. Note that larger contaminations ϵ point to a mixture model of the type (2.14) with an overly contribution of the distribution

H to the measurements. In this case some form of diagnosis is needed rather than a robust estimator, cf. Huber [80]. In practice, Huber's function (2.17) leads to a downweighing of the errors which are considered to be large and thus reduces their influence. This leads again to robust parameter estimates if the ratio of "bad" data points is correctly estimated and the tuning constant chosen accordingly. Compared to the l_1 norm minimization of the previous section, it has the advantage that it does not ignore any measurements which seem to suffer from gross errors but takes into account all available information with an appropriate scaling of its influence on the estimated parameters.

The idea behind this estimator is the following. Huber [76] investigated what would be the "worst" possible ϵ-contaminated distribution (2.14) among all normal distributions Φ with zero mean and unit variance with an ϵ-contamination of a symmetric distribution H. He then proposed to take the associated maximum likelihood estimator (*M-estimator*) as a robust estimator. The worst possible distribution in this sense means the distribution p from which the least possible information can be gained in each individual residual. This is known to be that p in the class (2.14) of which the probability density f minimizes the Fisher information,

$$\min_H \int_{-\infty}^{\infty} \frac{f'(z)^2}{f(z)}\, dz.$$

The minimizer f^* is given by

$$f^*(z) = (1-\epsilon)(2\pi)^{-\frac{1}{2}} e^{-\rho_\gamma(z)}.$$

This density f^* behaves like a normal density for small z and like an exponential density for large z. The corresponding distribution is

$$\mathrm{p}^*(x) = \begin{cases} \frac{1-\epsilon}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}), & |x| \le \gamma, \\ \frac{1-\epsilon}{\sqrt{2\pi}} \exp(\frac{\gamma^2}{2} - \gamma|x|), & |x| > \gamma, \end{cases}$$

where the constant γ depends on the contamination ϵ. The tuning constant varies from zero to infinity as the error ratio ϵ goes from 1 to 0. Maximizing the log-likelihood of f^* finally leads to the minimization

$$\min_x \rho(x) = \sum_i \rho_\gamma(x_i), \tag{2.19}$$

where $\rho_\gamma(\cdot)$ denotes the Huber function as given in (2.17).

Note that the development of this function occured under the assumption that the standard errors of the measurements are normally distributed from $\mathcal{N}(0, \mathbb{I})$. In order to adapt it to real data with non-unit variance, we must either divide the residuals $F_{1i}(X),\ i = 1, \ldots, m_1$, or multiply the threshold γ, respectively, by a *robust* estimate $\bar{\sigma}$ of the standard deviation of the residuals. For ease of notation we will assume in the following that we know this deviation and that we have scaled all residuals with the reciprocal of this value so that we can continue with the formulation (2.17).

When computing Huber's M-estimator, we are again confronted with the problem that the objective function in question is only once continuously differentiable, see also section 3.2.1 below. Via a derivation of the one-dimensional case, Mangasarian and Musicant [98] gave a reformulation of the unconstrained Huber minimization problem in case of a linear function $F_1(x) = Ax - b$ with x being a real d-vector, A being a given $l \times d$ real matrix and b a real l-vector. The following theorem shows that their formulation also carries over to the case of a general nonlinear function $F_1(X)$.

Theorem 2.7.1 (Reformulation of Huber's nonlinear estimator)**.** *A solution X of the unconstrained nonlinear Huber minimization problem (2.19) with $\rho_\gamma(\cdot)$ defined by (2.17) is given by any solution (X, v, t) of the following problem,*

$$\begin{array}{ll} \min\limits_{X\in\mathbb{R}^n, v\in\mathbb{R}^{m_1}, t\in\mathbb{R}^{m_1}} & \frac{1}{2}v^T v + \gamma e^T t \\ s.t. & -t \leq F_1(X) - v \leq t, \end{array} \tag{2.20}$$

where $e = (1, \ldots, 1)^T$ denotes the vector of all ones of appropriate dimension.

Note that in case that F_1 is a vector of linear functions, problem (2.20) is a convex quadratic program. We defer the proof of this theorem to section 3.2.1.

Note that problem (2.20) is equivalent to the problem

$$\min_{X\in\mathbb{R}^n, v\in\mathbb{R}^{m_1}} \frac{1}{2}\|v\|_2^2 + \gamma\|F_1(X) - v\|_1,$$

which clearly shows the hybrid nature of the Huber function as a compound of l_2 and l_1 components.

Like in case of the l_1 norm objective function, the introduction of the new variables leads to an unnecessary enlargement of the problem dimension. But we will see in the next chapter that we need the given reformulations (2.16) and (2.20) for theoretical purposes only while the actual computations will be done for the original problems.

2.8. Comparison of the three estimators

As we already said in the sections above, robustness particularly refers to the sensitivity of the estimator to outliers in the data, i.e. here large measurement errors. While the least squares estimator is drawn away from the bulk of data by any measurement lying sufficiently far away from the rest, this is not the case for the least absolute deviations estimator and Huber's M-estimator. However, this is only true for points largely deviating in the dependent variable. Both robust estimators are no less affected than the least squares estimator by points with outstanding independent variable values (e.g. Andersen [3], Rousseeuw and Leroy [115], Wilson [139]) and thus all three estimators have a breakdown point of BDP = 0.0. Note that the one-dimensional equivalent of the LAD estimator, the median, has a maximal breakdown point of 0.5, but this property does not transfer to the general higher dimensional

Table 2.2.: Comparison of least squares, least absolute deviations, and Huber estimator for linear problems.

	Least squares	Least absolute deviations	Huber
Robustness	not very robust	robust	robust (for correct choice of γ)
BDP	0.0	0.0	0.0
Stability	stable solution	unstable solution	depending on choice of γ
Uniqueness	always unique solution	possibly multiple solutions	depending on choice of γ
Location model	mean of data	median of data	in between
Assumed errors	Gaussian distribution	Laplace distribution	"worst possible ε-contaminated distribution"
References	Gauss [57], Legendre [90]	Maire and Boscovich [96], Laplace [89]	Huber [76, 78, 79]

case. For our applications, measurements can be supposed to be taken at specified time points where a single measurement long before or after all others is not likely to occur. So we are not bothered by outliers in the independent variable but hope to profit from the estimators' robustness with respect to outstanding values of the dependent variable.

For a nonlinear cost function F_1, all three estimators yield nonlinear optimization problems both in their original form and in form of the reformulations (2.16) and (2.20), respectively. In this case, we cannot tell in advance what differences the solutions of the estimators will exhibit. But if we have a linear cost function F_1 without additional constraints, there are significant differences between the three estimators.

We will now summarize the properties of the three estimators introduced in the previous sections for unconstrained problems with linear components of the cost function $F_1(X) = AX + b$, where the matrix $A \in \mathbb{R}^{m_1 \times n}$ is assumed to have full rank $m_1 < n$. They will be presented in a comparative table for ease of reference. A more detailed explanation is also given. In order to give a more illustrative view of the described features, we will also show the estimators' behavior for a small linear example. Note, however, that this example was only chosen to demonstrate some decisive differences and is not a representative problem for practical applications.

Table 2.2 shows the most important features of least squares, least absolute deviations, and Huber estimation for linear problem functions.

The stability of a solution means its sensitivity to initial parameter estimates. This point is also closely related to the uniqueness or non-uniqueness of the estimator. If x and y are two solutions of a linear l_p minimization problem, then so is any convex combination $tx + (1-t)y$ with $0 < t < 1$. This follows from Minkowski's inequality. Thus the problem is either uniquely solvable or it has infinitely many solutions, cf. Scales and Gersztenkorn [118]. The linear least squares solution is always unique and is thus reached from any set of initial values, if the estimator converges. The linear LAD estimator on the other hand possibly has infinitely many solutions. Mostly it cannot be predicted to which solution it is going to converge. It is very likely that in the presence of multiple solutions any small change in the initial conditions will lead to different parameter estimates. Compare also the example below for this case. For Huber's M-estimator the situation does not present itself so obviously. Stability and uniqueness of the estimator depend on the choice of the tuning constant γ. As for large values of γ the estimator converges to the solution of the least squares estimator, it will also have only one unique solution as soon as the tuning constant exceeds a problem dependent threshold. With decreasing γ, Huber's M-estimator approaches the LAD estimate and thus it will also have multiple solutions if this is the case for the least absolute deviations estimator (cf. also Li and Swetits [91]).

The tabular entries on the location model are mentioned to illustrate the estimators for the most simple example of the estimation of one location parameter by linear regression. It is obvious from statistics that in this case least squares estimation will return the mean of the measurements while the least absolute deviations estimator will give their median. Huber's M-estimator will result in a value in between the two, depending on the choice of the tuning constant γ.

The assumed error models for the estimators (for which these are maximum likelihood estimators) were already mentioned in the sections above and are only given for the sake of completeness of table 2.2. The same holds true for the given references.

In the rest of this section we will consider an illustrative example which shall especially demonstrate the non-uniqueness of the linear robust estimators. Suppose therefore that we have five (one-dimensional) measurements taken at three different times. Let the first time point have a single measurement and let the other two times have two measurements each. We assume that for both these times the measurements have the same distance from the measurement of the first time point. This situation is depicted in figure 2.3.

The filled area shows the range in which the least absolute deviations estimator, depicted by the solid line, may lie. To understand this, let $a_i,\ i = 2, 3$, denote the distance of the estimator from the upper points T_2, T_3 of the double measurements, and let b denote the distance between the two measurements T_2, T_4 at the middle time point and c the distance between the two measurements T_3, T_5 at the last time point. Then the objective function of the LAD estimator computes to

$$\text{LAD} = 0 + a_2 + (b - a_2) + a_3 + (c - a_3) = b + c.$$

Now shift the LAD estimator such that its difference from $T_i,\ i = 2, 3$, changes to $\tilde{a}_i$

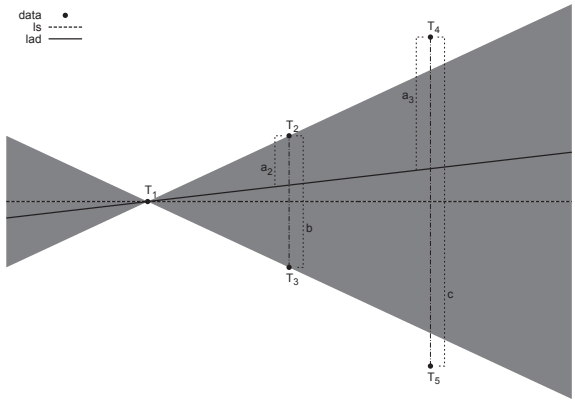


Figure 2.3.: Illustrative example for non-uniqueness of the linear robust estimators.

while it still stays in the colored range. The new estimator's objective function value then becomes

$$\widetilde{\text{LAD}} = 0 + \tilde{a}_2 + (b - \tilde{a}_2) + \tilde{a}_3 + (c - \tilde{a}_3) = b + c = \text{LAD}.$$

So the LAD estimator may be any line in the colored range, depending on the initial guesses of the line's defining constants. In contrast, the least squares estimator will always compute to the horizontal dashed line in figure 2.3. Huber's M-estimator is limited to a range symmetric around the least squares estimator that depends on the choice of the tuning constant and that converges to the whole filled area as γ approaches zero.

2.9. Other robust estimators

Robustness, i.e. insensitivity to deviations from the assumptions, particularly normality of the errors, has been in the mind of scientists almost from the beginning of the employment of well-defined procedures for parameter estimation (see also Stigler [130]). One of the oldest techniques in dealing with gross errors in measurement data is the rejection of outliers which was already applied by Legendre in the first published work on least squares [90]: "*Si parmi ces erreurs il s'en trouve que l'on juge trop grandes pour être admises, alors on rejettera les équations qui ont produit ces erreurs, comme venant d'expériences trop défectueuses, et on déterminera les inconnues par le moyen des équations restentes, qui alors donneront des erreurs beaucoup moindre.*"

In the translation by Ruger and Walker in Smith [127] (pp. 576–579), this quotation reads "*If among these errors are some which appear too large to be admissible, then those equations which produced these errors will be rejected, as coming from too faulty experiments, and the unkowns will be determined by means of the other equations, which will then give much smaller errors.*"

A first formal criterion for the identification of outliers was published in 1852 by Peirce [110]. This was followed only few years later by a paper by Airy [1] criticizing the use of outlier criteria in which he concludes, amongst other things, that "*No consideration of the magnitudes of residual errors* per se *will justify us in rejecting a result.*"

The following year saw an ongoing debate on the subject with some authors in favor of Peirce, some in favor with Airy (e.g. Glaisher [64, 65, 66], Stone [132]). Nevertheless, soon a variety of different criteria for outlier rejection were in use, one of the most widely used being that proposed by Chauvenet [35]. Ross [114] showed that a comparison of Peirce's and Chauvenet's criteria is in favor of Peirce, though. Yet, some of these rejection procedures perform so poorly that they are not even able to identify one single distant outlier (Hampel [68]).

Parallel to this development, other techniques for robustification were also employed. Most popular among these were different weighted means which advanced further until by the last half of the nineteenth century weighted least squares had become a standard procedure. So up to 1885 the cautious use of the sample mean, possibly with weights for the observations or rejected outlying data points, was considered the best approach to estimate parameters (cf. Stigler [130]). Stigler [130] also summarizes the history of robust estimation with special emphasis on the period 1885–1920 where he concentrates on statistics and historical development.

Two techniques that have been known already in the nineteenth century are trimming and Winsorizing. Both are closely related to least squares, to the two robust estimators described above and also have connections to outlier rejection. The simple idea of trimming is to a priori discard measurements at both ends of the observation range. The trimmed mean for instance corresponds to the estimator with the objective function

$$\rho(F_1(X)) = \sum_{i=1}^{m_1} \rho_k(F_{1i}(X)) \text{ with } \rho_k(z) = \begin{cases} \frac{1}{2}z^2, & \text{if } |z| \leq k, \\ \frac{1}{2}k^2, & \text{if } |z| > k. \end{cases}$$

This estimator is very sensitive with respect to the behavior of the underlying probability distribution at the points $\pm k$. If the corresponding probability density puts much weight at these points, the estimator produces arbitrary results, cf. Huber [76]. The Winsorized mean is similar in its computation but it does not discard observations with such rigorosity and is thus less sensitive to the underlying distribution. To understand its workings, assume that the observations have been sorted in non-decreasing order, $z_1 \leq z_2 \leq \ldots, \leq z_n$. The Winsorized mean is then defined as $T = \frac{1}{n}(gz_{g+1} + z_{g+1} + z_{g+2} + \cdots + z_{n-h} + hz_{n-h})$, i.e. the g leftmost and the h rightmost observations are Winsorized. Note that this estimator is asymptotically equivalent to Huber's M-estimator as the latter can be expressed by the same formula with the factors z_{g+1} and z_{n-h} in the first and last summand being replaced with u, v with $z_g \leq u \leq z_{g+1}$ and $z_{n-h} \leq v \leq z_{n-h+1}$, respectively. All constants g, h, u, v depend on the specific sample.

Today there is a variety of robust estimators in use. MATLAB® provides a routine for robust regression which lets the user choose an estimator for his specific problem from a class of robust multilinear regression estimators. The implemented estimators share the common framework of iteratively reweighted least squares and only differ in the used weight function. All these estimators incorporate a tuning constant, by which the downweighing of large residuals can be controlled. The implemented default values correspond to parameter estimates that are approximately 95% as efficient as ordinary least squares estimates in case of a normal error distribution without outliers.

Iteratively reweighted least squares (IRLS) being a widely used technique for robust regression, we will shortly summarize this approach here for the case of a standard regression model. So let us consider the model

$$\eta = \mathcal{M}X + \varepsilon, \tag{2.21}$$

where η is an $m \times 1$ vector of measurements, $\mathcal{M}$ is the $m \times n$ design matrix, X is the $n \times 1$ vector of unknowns, and ε is an $m \times 1$ vector of measurement errors. A robust estimate $\hat{X}$ for X is obtained by solving

$$\min_X \sum_{i=1}^{m} \rho\left(\frac{\eta_i - \mathcal{M}_i X}{\sigma}\right), \tag{2.22}$$

with a robust loss function ρ, an estimated or previously known scale parameter σ, and $\mathcal{M}_i$ denoting the ith row of $\mathcal{M}$. A necessary condition for $\hat{X}$ being a solution of (2.22) is that the derivative $\psi = \rho'$ is zero in $\hat{X}$,

$$\sum_{i=1}^{m} \mathcal{M}_{ij} \psi\left(\frac{\eta_i - \mathcal{M}_i X}{\sigma}\right) = 0 \quad \text{for all} \quad j = 1, \ldots, n. \tag{2.23}$$

So in general, this means that we have to solve a set of nonlinear equations which requires iterative methods. Starting with an initial guess X^0, Newton's method can be written in this context as

$$X^{i+1} = X^i + \sigma \left(\mathcal{M}^T \mathrm{diag}(\psi'(\frac{\eta - \mathcal{M}X^i}{\sigma}))\mathcal{M}\right)^{-1} \mathcal{M}^T \psi(\frac{\eta - \mathcal{M}X^i}{\sigma}).$$

Defining the weight function $w(r) = \psi(r)/r$, we can get another iteration scheme

$$X^{i+1} = X^i + \left(\mathcal{M}^T \mathrm{diag}(w(\frac{\eta - \mathcal{M}X^i}{\sigma}))\mathcal{M}\right)^{-1} \mathcal{M}^T \mathrm{diag}(w(\frac{\eta - \mathcal{M}X^i}{\sigma}))(\eta - \mathcal{M}X^i), \tag{2.24}$$

which is usually attributed to Beaton and Tukey [11]. As this method only requires the knowledge of the weight function $w(r)$, it can be used with any existing weighted least squares solver or even a standard least squares solver if routines for computing the square root of $w(r)$ and evaluating $\mathrm{diag}(w^{1/2})\mathcal{M}$ and $\mathrm{diag}(w^{1/2})\eta$ are provided.

Usually, the convergence of the IRLS methods will be slower than Newton's method. For nonconvex functions ρ, convergence cannot be guaranteed, and a good starting value becomes even more important, cf. Holland and Welsch [74]. The authors of that paper propose the use of the least absolute deviations estimator as initial guess X^0. The scale estimate $\bar{\sigma}$ used in the MATLAB® routine `robustfit` is an estimate of the standard deviation of the errors, computed as the median absolute deviation of the residuals from their median, divided by 0.6745 to make the estimate unbiased for the normal distribution,

$$\bar{\sigma} = \frac{1}{0.6745}\text{median}_i \left|(\eta_i - \mathcal{M}_i X^0) - \text{median}_i(\eta_i - \mathcal{M}_i X^0)\right|.$$

This scale is only computed once before starting the iterations.

The estimators which are implemented in this MATLAB® function use different weight functions w and can be categorized by the behavior of the corresponding function $\psi(r)$ in the following way:

hard redescenders: $\psi(r) = 0$ for $|r|$ sufficiently large

To this class belong the estimators with

$$w(r) = \begin{cases} \left(\frac{r}{c_A}\right)^{-1} \sin r c_A, & |r| \le \pi c_A, \\ 0, & |r| > \pi c_A, \end{cases} \text{ of Andrews et al. [4],}$$

$$w(r) = \begin{cases} \left(1 - (\frac{r}{c_B})^2\right)^2, & |r| \le c_B, \\ 0, & |r| > c_B, \end{cases} \text{ of Beaton and Tukey [11], and}$$

$$w(r) = \begin{cases} 1, & |r| \le c_T, \\ 0, & |r| > c_T \end{cases} \text{ of Hinrich and Talwar [72].}$$

soft redescenders: $\psi(r) \to 0$ for $|r| \to \infty$

This class contains the estimators with

$w(r) = \left(1 + (\frac{r}{c_C})^2\right)^{-1}$ giving a maximum likelihood estimate for t distributions (denoted C for Cauchy as the Cauchy distribution can be derived as a special case of the t distribution with one degree of freedom) and

$w(r) = \exp\left(-(\frac{r}{c_W})^2\right)$ of Dennis and Welsch [41].

others: monotone function ψ

In this class fall the estimators with

$$w(r) = \begin{cases} 1, & |r| \le c_H, \\ c_H |r|^{-1}, & |r| > c_H, \end{cases} \text{ of Huber [76],}$$

$w(r) = (1 + \frac{|r|}{c_F})^{-1}$ of Fair [48], and

$w(r) = \left(\frac{r}{c_L}\right)^{-1} \tanh(\frac{r}{c_L})$ which is called the logistic estimator.

We do not go further into detail here. The interested reader is referred to the literature, e.g. Beaton and Tukey [11], Holland and Welsch [74], and Street et al. [133].

3. Solution of the parameter estimation problem: The Gauss-Newton method in robust optimization

We have seen in the previous chapter that after discretization of the dynamics the parameter estimation problem with DAE constraints can be written in the form (2.11) of a general nonlinear programming problem. So this chapter is devoted to the solution of such general problems. In addition, we will treat the special cases of the two robust estimators, least absolute deviations and Huber's M-estimator, and show how these fit into this framework. Our method of choice for solving the nonlinear optimization problem (2.11) is a generalized Gauss-Newton method.

We take the view of considering the Gauss-Newton method for the solution of (3.1) as special kind of a *sequential quadratic programming* (SQP) algorithm. Therefore we first briefly describe the general SQP framework and its connection to the Gauss-Newton method in the first section. The second section contains the application of the Gauss-Newton method to robust parameter estimation with l_1 norm or Huber objective function. In the third section we will finally describe globalization methods that can be used to make the method converge from arbitrary initial values.

3.1. The Gauss-Newton method in nonlinear programming

For the beginning of this chapter, let us consider a general least squares estimation problem,

$$\begin{array}{ll} \min\limits_{X\in\mathbb{R}^n} & \frac{1}{2}\|F_1(X)\|_2^2 \\ \text{s.t.} & F_2(X) = 0, \\ & F_3(X) \geq 0, \end{array} \tag{3.1}$$

with $F_1 : \mathbb{R}^n \to \mathbb{R}^{m_1}$, $F_2 : \mathbb{R}^n \to \mathbb{R}^{m_2}$, $F_3 : \mathbb{R}^n \to \mathbb{R}^{m_3}$ all sufficiently smooth, i.e. at least once continuously differentiable and with continuous second partial derivatives. In the context of parameter estimation, the cost function F_1 typically describes the deviation of the response of some mathematical model from given measurements which we want to minimize in a suitable norm (or some other appropriate functional), cp. section 2.2. In this case, we need $\|F_1(X)\|$ to be "small" for the model to be valid

and we can expect the number of measurements to be much larger than the number of variables, i.e. $m_1 \gg n$ (cf. Gill et al. [63]).

We consider the most general nonlinear programming problem (NLP),

$$\begin{array}{ll} \min\limits_{X \in \mathbb{R}^n} & F(X) \\ \text{s.t.} & G(X) = 0, \\ & H(X) \geq 0, \end{array} \tag{3.2}$$

with $F : \mathbb{R}^n \to \mathbb{R}$, $G : \mathbb{R}^n \to \mathbb{R}^{m_g}$, $H : \mathbb{R}^n \to \mathbb{R}^{m_h}$ all suffienciently smooth as given for (3.1).

We introduce now some notations which will be used throughout this chapter.

A point $\hat{X} \in \mathbb{R}^n$ will be called a *feasible* point if it satisfies the equality and inequality constraints, i.e. $G(\hat{X}) = 0$ and $H(\hat{X}) \geq 0$. The point X^* is a *local minimizer* of problem (3.2) if X^* is feasible and if there exists a neighborhood $\mathsf{N}(X^*)$ such that $F(X^*) \leq F(X)$ for all feasible $X \in \mathsf{N}(X^*)$. If we have a strict inequality here, i.e. $F(X^*) < F(X)$ for all $X \neq X^*$ in this neighborhood, the point X^* is called a *strict local minimizer*.

Another important definition concerns the inequality constraints. The constraint $H_i(X) \geq 0$ is said to be *active* at a point $\hat{X}$ if $H_i(\hat{X}) = 0$. Otherwise it will be called *inactive* at $\hat{X}$. For any feasible point $\hat{X}$ we call the set $I_H(\hat{X}) = \{i \mid H_i(\hat{X}) = 0\}$ the *active set* at this point.

Due to the nonlinearity of the constraints, feasibility will in general only be retained along nonlinear paths in $\mathbb{R}^n$, so-called *feasible arcs*. A (continuously differentiable) arc is a directed (continuously differentiable) curve $c(\vartheta)$ in $\mathbb{R}^n$ parameterized by the variable $\vartheta \geq 0$ in an interval $[0, \epsilon]$ with $\epsilon > 0$. We always choose $c(\cdot)$ so that $c(0) = \hat{X}$. Feasibility with respect to the equality constraints is retained for $G(c(\vartheta)) \equiv 0$. This implies that

$$\left.\frac{d}{d\vartheta} G_i(c(\vartheta))\right|_{\vartheta=0} = \nabla G_i(c(0))c'(0) = \nabla G_i(\hat{X})y = 0, \quad i = 1, \ldots, m_g, \tag{3.3a}$$

where $y = c'(0)$, i.e. the tangent to the arc at $\hat{X}$, is a feasible direction. Inactive inequality constraints do not pose any restrictions on the set of feasible directions. For $H_i(\hat{X}) > 0$ there always exists a neighborhood of $\hat{X}$ with $H_i(X) > 0$ for all X in this neighborhood. For active inequalities at $\hat{X}$ we are restricted to the neighborhood of $\vartheta = 0$ in which the constraint is nondecreasing,

$$\left.\frac{d}{d\vartheta} H_i(c(\vartheta))\right|_{\vartheta=0} = \nabla H_i(c(0))c'(0) = \nabla H_i(\hat{X})y \geq 0, \quad i \in I_H(\hat{X}). \tag{3.3b}$$

Note that these conditions completely characterize the set of feasible directions only in the case of linear constraints. For nonlinear constraints they are still necessary for y to be a feasible direction but they do not guarantee the existence of a feasible arc with tangent y. This problem is overcome by *constraint qualifications*. One of

the most commonly used is the so-called *linear independence constraint qualification* (LICQ).

We say that the LICQ condition holds at a point $\hat{X}$ if the gradients of the equality constraints $\nabla G_i(\hat{X}),\ i = 1,\dots,m_g$, and of the active inequality constraints $\nabla H_i(\hat{X}),\ i \in I_H(\hat{X})$, are linearly independent. This means the matrix $\nabla\tilde{G}(X) = \begin{pmatrix}\nabla G(X)\\ \nabla\tilde{H}(X)\end{pmatrix}$ with $\tilde{H}(X) = (H_i(X))_{i\in I_H(X)}$ has full rank. A feasible point at which the LICQ condition is satisfied is called a *regular point*.

Lemma 3.1.1. *[101] If $y \in \mathbb{R}^n$ satisfies the conditions (3.3) with $\|y\| = 1$ and if the LICQ condition holds at $\hat{X}$, then y is a tangent to a feasible arc $c(\vartheta)$ at $c(0) = \hat{X}$.*

Another common constraint qualification is the so-called *Mangasarian Fromovitz constraint qualification* (MFCQ). Given a point $\hat{X}$ and the active set $I_H(\hat{X})$, we say that the MFCQ condition holds at $\hat{X}$ if there exists a vector $\omega \in \mathbb{R}^n$ such that $\nabla G_i(\hat{X})\omega = 0$ for all $i = 1,\dots,m_g$, and $\nabla H_i(\hat{X})\omega > 0$ for all $i \in I_H(\hat{X})$ and the set of equality constraint gradients $\{\nabla G_i(\hat{X}), i = 1,\dots,m_g\}$ is linearly independent.

MFCQ is a weaker condition than LICQ. If the LICQ condition is satisfied, then there exists a vector w such that $\nabla G_i(\hat{X})w = 0$ $(i = 1,\dots,m_g)$ and $\nabla H_j(\hat{X})w = 1$ $(j \in I_H(\hat{X}))$. Obviously the MFCQ condition is also satisfied by this w. The opposite direction is not true. Performance of the MFCQ condition does not imply performance of the LICQ condition.

3.1.1. Optimality conditions for nonlinear programming problems

Beside the widely used Karush-Kuhn-Tucker (KKT) optimality conditions, which require the optimization problem to satisfy a constraint qualification for being applicable, there are the less commonly used Fritz John (FJ) optimality conditions, which do not pose any restrictions on the problem under consideration (beside sufficient differentiability).

The original conditions were formulated only for inequality constrained problems and, in their original form, do not apply to optimization problems with equality constraints. Mangasarian and Fromovitz [97] derived a generalization of the Fritz John conditions that allows handling equalities and inequalities together. It is their formulation that we are going to use in the following.

For the general nonlinear programming problem (3.2) they take the following form:

Theorem 3.1.1 (Generalized Fritz John (FJ) necessary optimality conditions [97])**.** *If X^* is a solution of (3.2) then there exist vectors $\mu^* = (\mu_0^*, \mu_1^*, \dots, \mu_{m_h}^*)^T \in \mathbb{R}^{m_h+1}$,*

$\lambda^* = (\lambda_1^*, \ldots, \lambda_{m_g}^*)^T \in \mathbb{R}^{m_g}$ *such that*

$$\mu_0^* \nabla F(X^*) - \sum_{i=1}^{m_g} \lambda_i^* \nabla G_i(X^*) - \sum_{i=1}^{m_h} \mu_i^* \nabla H_i(X^*) = 0, \tag{3.4a}$$

$$\sum_{i=1}^{m_h} \mu_i^* H_i(X^*) = 0, \tag{3.4b}$$

$$\mu^* \geq 0, \tag{3.4c}$$

$$\begin{pmatrix} \mu^* \\ \lambda^* \end{pmatrix} \neq 0. \tag{3.4d}$$

We will refer to a point X^* satisfying (3.4) as a Fritz John (FJ) point.

Equation (3.4a) states the fact that the gradient of the modified Lagrange function

$$\mathcal{L}(X, \mu, \lambda) = \mu_0 F(X) - \lambda^T G(X) - \mu^T H(X) \tag{3.5}$$

has to vanish in any stationary point (X^*, μ^*, λ^*).

Although these conditions are necessary for local optimality, they are often not selective enough because these optimality criteria can be satisfied regardless of the objective function when $\mu_0^* = 0$. In order to ensure that $\mu_0^* \neq 0$, we need a constraint qualification, e.g. the LICQ or MFCQ conditions. Then we can choose $\mu_0^* = 1$ by a simple rescaling and are left with the standard Karush-Kuhn-Tucker (KKT) conditions. We already said above that the LICQ condition is the stronger requirement. While the LICQ condition guarantees the uniqueness of the multipliers λ, μ, the MFCQ condition is equivalent to the boundedness to the set of multipliers (cf. Nocedal and Wright [101]).

Note also that constraint qualifications are only sufficient conditions for the linear approximations (3.3) of the constraints being adequate representations of the local geometry of the feasible set. See Nocedal and Wright [101] for a simple example that they are not necessary.

The classification of stationary points given in theorem 3.1.1 is not sufficient to differentiate between their different natures, i.e. minimum, maximum, or saddle point. This is done by optimality conditions of second order which we will state next.

These conditions involving the Hessians of the problem functions were derived by Ben-Tal [12]. His second-order necessary conditions in principle contain the results of the previous theorem but only refer to the active constraints, i.e. equality constraints and inequality constraints that are satisfied as equality in the solution. The main difference to second-order results based on the Karush-Kuhn-Tucker conditions mentioned above is that we do not have any *fixed multipliers* now but rather multipliers that are *functions of the critical directions*, cf. Ben-Tal [12]. As critical directions we

denote all vectors $d \in D(X^*)$, where

$$\begin{aligned} D(X^*) = \{d \in \mathbb{R}^n \mid & \nabla F(X^*)d \leq 0, \\ & \nabla G_i(X^*)d = 0 \ (i = 1, \ldots, m_g), \\ & \nabla H_i(X^*)d \geq 0 \ (i \in I_H(X^*))\}. \end{aligned}$$

Ben-Tal's necessary conditions of second-order for the general optimization problem (3.2) are given in the following theorem.

Theorem 3.1.2 (Necessary conditions of second order [12]). *Suppose that X^* is a local minimizer of (3.2). Then, corresponding to every critical direction $d \in D(X^*)$, there exist multipliers*

$$\mu_i^* \geq 0, \quad i \in I_H(X^*) \cup \{0\}, \qquad \lambda_j^* \in \mathbb{R}, \quad j = 1, \ldots, m_g, \qquad \textit{not all zero},$$

satisfying

$$\mu_0^* \nabla F(X^*) - \sum_{i=1}^{m_g} \lambda_i^* \nabla G_i(X^*) - \sum_{i \in I_H(X^*)} \mu_i^* \nabla H_i(X^*) = 0, \tag{3.6a}$$

$$\mu_0^* \nabla F(X^*)d = 0, \tag{3.6b}$$

$$\forall i \in I_H(X^*): \quad \mu_i^* \nabla H_i(X^*)d = 0, \tag{3.6c}$$

$$d^T \left[\mu_0^* \nabla^2 F(X^*) - \sum_{i=1}^{m_g} \lambda_i^* \nabla^2 G_i(X^*) - \sum_{i \in I_H(X^*)} \mu_i^* \nabla^2 H_i(X^*) \right] d \geq 0. \tag{3.6d}$$

It is possible, e.g., if X^* is an isolated feasible point, that for every critical direction $d \in D(X^*)$ the corresponding multiplier of the objective function $\mu_0^* = \mu_0^*(d)$ is equal to zero. In this case the necessary conditions hold regardless of the objective function. In order to exclude this possibility, we need a constraint qualification as additional assumption. This constraint qualification may just be that for at least one critical direction d the multiplier $\mu_0^*(d)$ is strictly positive. Both the commonly used linear independence constraint qualification (LICQ) or the Mangasarian-Fromovitz constraint qualification (MFCQ) guarantee this condition to hold but they are more limited. Ben-Tal [12] proposes a natural extension of the so-called modified Arrow-Hurwitz-Uzawa condition. This constraint qualification reads as follows.

Lemma 3.1.2 (Constraint qualification [12]). *If the gradients of the equality constraints $\{\nabla G_i(X^*), i = 1, \ldots, m_g\}$ are linearly independent and if there exists a critical direction $d \in D(X^*)$ and a vector $z \in \mathbb{R}^n$ such that*

$$\begin{aligned} &\nabla H_i(X^*)z - d^T \nabla^2 H_i(X^*)d < 0, \quad i \in \Gamma(d), \\ &\nabla G_j(X^*)z - d^T \nabla^2 G_j(X^*)d = 0, \quad j = 1, \ldots, m_g, \end{aligned}$$

where $\Gamma(d) = \{i \in I_H(X^*) \mid \nabla H_i(X^*)d = 0\}$, then the second-order necessary conditions (3.6) can be satisfied with a multiplier $\mu_0^*(d) > 0$.

If there exists a vector $\nu \in \mathbb{R}^n$ with $\nabla H_i(X^*)\nu < 0$ for all $i \in I_H(X^*)$ and $\nabla G_j(X^*)\nu = 0$ for all $j = 1, \ldots, m_g$, then the requirements of this lemma are trivially satisfied with $d = 0$ and $z = \nu$. Note, however, that the constraint qualification of the lemma still does not guarantee the existence of fixed multipliers satisfying the necessary conditions of second order. Constraint qualifications which do guarantee the latter are termed *strong constraint qualifications* by Ben-Tal [12]. The LICQ condition, for instance, falls into this category.

The sufficient optimality conditions of Fritz John type are given in the next theorem.

Theorem 3.1.3 (Sufficient conditions of second order [12]). *Suppose that corresponding to every nonzero critical direction $d \in D(X^*)$, there exist multipliers*

$$\mu_i^* \geq 0, \quad i \in I_H(X^*) \cup \{0\}, \qquad \lambda_j^* \in \mathbb{R}, \quad j = 1, \ldots, m_g, \qquad \textit{not all zero},$$

satisfying

$$\mu_0^* \nabla F(X^*) - \sum_{i=1}^{m_g} \lambda_i^* \nabla G_i(X^*) - \sum_{i \in I_H(X^*)} \mu_i^* \nabla H_i(X^*) = 0, \tag{3.7a}$$

$$\mu_0^* \nabla F(X^*) d = 0, \tag{3.7b}$$

$$\forall i \in I_H(X^*): \quad \mu_i^* \nabla H_i(X^*) d = 0, \tag{3.7c}$$

$$d^T \left[\mu_0^* \nabla^2 F(X^*) - \sum_{i=1}^{m_g} \lambda_i^* \nabla^2 G_i(X^*) - \sum_{i \in I_H(X^*)} \mu_i^* \nabla^2 H_i(X^*) \right] d > 0. \tag{3.7d}$$

Then X^* is an isolated local minimizer of (3.2).

Being equipped with the necessary vocabulary and background knowledge on optimality, we can now consider how to actually solve the NLP (3.2).

3.1.2. The SQP framework

There is a wide range of algorithms for solving general nonlinear programming problems of the form (3.2). All nonlinear optimization algorithms share the common property of being an iterative procedure, that means, starting from a given initial guess for the solution, they generate a new iterate in each step which is "better" in some sense than the previous iterate. To do so, in each iteration k a step ΔX^k and possibly a steplength t^k are computed and the current iterate X^k is updated according to the rule

$$X^{k+1} = X^k + [t^k]\Delta X^k.$$

An efficient algorithm for nonlinear programming problems is the SQP approach which corresponds to a (local) quadratic approximation of the optimization problem from which then the step is computed.

For ease of notation and because we only want to give a short introduction to the concept of SQP algorithms, we first restrict our considerations to equality constrained problems,

$$\begin{aligned} \min_{X\in\mathbb{R}^n} \quad & F(X) \\ \text{s.t.} \quad & G(X) = 0, \end{aligned} \tag{3.8}$$

with F and G as given for (3.2).

Let us assume that a constraint qualification, e.g. the LICQ condition, is satisfied for (3.8) so that the Fritz John conditions can be satisfied with $\mu_0^* \neq 0$. Without loss of generality we can then choose $\mu_0^* = 1$. The FJ conditions (3.4) are reduced to

$$K(X,\lambda) = \begin{pmatrix} \nabla_X \mathcal{L}(X,\lambda) \\ G(X) \end{pmatrix} = \begin{pmatrix} \nabla F(X) - \nabla G(X)^T \lambda \\ G(X) \end{pmatrix} = 0, \tag{3.9}$$

i.e. the standard KKT conditions as shortly mentioned in the previous section. We can naturally solve this system with Newton's method. The Jacobian of (3.9) with respect to X and λ reads

$$J(X,\lambda) = \begin{pmatrix} \nabla_X^2 \mathcal{L}(X,\lambda) & -\nabla G(X)^T \\ \nabla G(X) & 0 \end{pmatrix},$$

which we call the *FJ or KKT matrix*, and the new iterate (X^{k+1}, λ^{k+1}) is computed as

$$\begin{pmatrix} X^{k+1} \\ \lambda^{k+1} \end{pmatrix} = \begin{pmatrix} X^k \\ \lambda^k \end{pmatrix} + \begin{pmatrix} \Delta X^k \\ \Delta\lambda^k \end{pmatrix}, \tag{3.10}$$

where the step $(\Delta X^k, \Delta\lambda^k)$ solves the system

$$J(X^k,\lambda^k) \begin{pmatrix} \Delta X \\ \Delta\lambda \end{pmatrix} = -K(X^k,\lambda^k). \tag{3.11}$$

This approach is also referred to as the *Newton-Lagrange method*. It is well-defined if the FJ matrix is non-singular. The following lemma gives conditions which guarantee this non-singularity.

Lemma 3.1.3 (see e.g. Bock [24] for this formulation)**.** *Let the LICQ condition hold at a given point X and let the Hessian matrix of the Lagrangian at this point be positive definite on the nullspace of the gradient matrix of the constraints, i.e.*

$$p^T \nabla_X^2 \mathcal{L}(X,\lambda) p > 0 \quad \text{for all } p \in Z = \{p \neq 0 \mid \nabla G(X) p = 0\}.$$

Then the FJ matrix $J(X,\lambda)$ is non-singular.

Rewriting the iteration (3.10), (3.11) in a different way, also allows another interpretation of the method. Let us rewrite (3.11) as

$$\nabla_X^2 \mathcal{L}(X^k,\lambda^k)\Delta X + \nabla F(X^k) - \nabla G(X^k)^T \underbrace{(\lambda^k + \Delta\lambda)}_{\lambda^{k+1}} = 0, \tag{3.12a}$$

$$\nabla G(X^k)\Delta X + G(X^k) = 0. \tag{3.12b}$$

Obviously, $(\Delta X, \lambda^{k+1})$ satisfies the FJ conditions, again with $\mu_0^* = 1$, of the quadratic program (QP)

$$\begin{array}{ll} \min\limits_{\Delta X} & \frac{1}{2}\Delta X^T \nabla_X^2 \mathcal{L}(X^k, \lambda^k)\, \Delta X + \nabla F(X^k)^T \Delta X \\ \text{s.t.} & \nabla G(X^k)\Delta X + G(X^k) = 0. \end{array} \tag{3.13}$$

If the solution $(\Delta X^*, \lambda^{QP})$ of this program is unique, we have $\Delta X^k = \Delta X^*$ and $\lambda^{k+1} = \lambda^{QP}$. So the iteration (3.10) can be written equivalently as

$$\begin{pmatrix} X^{k+1} \\ \lambda^{k+1} \end{pmatrix} = \begin{pmatrix} X^k \\ \lambda^k \end{pmatrix} + \begin{pmatrix} \Delta X^* \\ \lambda^{QP} - \lambda^k \end{pmatrix}. \tag{3.14}$$

See Nocedal and Wright [101] for a proof that the uniqueness of the QP solution can be guaranteed under exactly the same conditions as that of the Newton-Lagrange method above, i.e. lemma 3.1.3.

This relation of the solutions of the Newton-Lagrange method and of the quadratic program (3.13) can be interpreted as *equivalence of Newton's method and the SQP method*. Of course, the local convergence property of Newton's method hence transfers to the SQP method immediately (see also section 3.1.4).

This approach can be extended at once to the general nonlinear programming problem (3.2). Linearizing both the equality and inequality constraints leads to the quadratic problem

$$\begin{array}{ll} \min\limits_{\Delta X} & \frac{1}{2}\Delta X^T \nabla_X^2 \mathcal{L}(X^k, \lambda^k, \mu^k)\, \Delta X + \nabla F(X^k)^T \Delta X \\ \text{s.t.} & \nabla G(X^k)\Delta X + G(X^k) = 0, \\ & \nabla H(X^k)\Delta X + H(X^k) \geq 0. \end{array} \tag{3.15}$$

An explicit statement of the local SQP framework, though in a very simple form, is given in the following algorithm.

Algorithm 3.1.1 Local SQP algorithm

(S1) Choose an initial guess $(X^0, \lambda^0, \mu^0) \in \mathbb{R}^n \times \mathbb{R}^{m_g} \times \mathbb{R}^{m_h}$ and set $k := 0$.

(S2) If (X^k, λ^k, μ^k) is an FJ point of (3.2) for $\mu_0^* = 1$: STOP with $(X^*, \lambda^*, \mu^*) = (X^k, \lambda^k, \mu^k)$.

(S3) Compute an FJ point $(\Delta X^k, \lambda^{QP}, \mu^{QP})$ of (3.15) with $\mu_0^{QP} = 1$.

(S4) Set $X^{k+1} = X^k + \Delta X^k$, $\lambda^{k+1} = \lambda^{QP}$, $\mu^{k+1} = \mu^{QP} \setminus \{\mu_0^{QP}\}$, and $k = k+1$ and go to (S2).

It is possible to prove that this approach finally identifies the optimal active set of the general NLP (3.2).

Theorem 3.1.4 (Stability of the active set in the neighborhood of a solution). *Let X^* be a local minimizer of (3.2) and let the LICQ condition hold at X^*. Furthermore let the Hessian of the Lagrangian $\nabla^2_X \mathcal{L}(X^*, \lambda^*, \mu^*)$ be positive definite on the nullspace of the active gradients, i.e. let $p^T \nabla^2_X \mathcal{L}(X^*, \lambda^*, \mu^*) p > 0$ for all $p \in \tilde{Z} = \{p \neq 0 \mid \nabla \tilde{G}(X^*)p = 0\}$, and let strict complementarity hold, i.e.*

$$\mu_i^* = 0 \Leftrightarrow F_{3i}(X^*) > 0.$$

If (X^k, λ^k, μ^k) is sufficiently close to (X^*, λ^*, μ^*), then there exists a local solution of the QP (3.15) the active set $I_{\bar{H}}(\Delta X^k)$ of which is the same as the active set $I_H(X^*)$ of the NLP (3.2) at X^*, where $\bar{H}(\Delta X) = \nabla H(X^k)\Delta X + H(X^k) \geq 0$ denotes the linearized inequality constraints at X^k.

Therefore (3.15) will behave like an equality constrained QP as the iterates of the SQP method approach a minimizer of (3.2) that satisfies the conditions of this theorem. This means that we can eventually ignore the inequality constraints that are inactive at X^* and treat the inequality constraints that belong to the active set $I_H(X^*)$ as additional equality constraints.

In implementations of the SQP algorithm it is common practice to replace the Hessian of the Lagrangian $\nabla^2_X \mathcal{L}(X^k, \lambda^k, \mu^k)$ in the objective function of (3.15) by an approximation L^k. By the equivalence of the SQP method and the Newton-Lagrange method this corresponds to solving the problem by a Newton-type approach or, in case that the approximated Hessians are computed by means of an update procedure, by a quasi-Newton approach.

3.1.3. The generalized Gauss-Newton method

We return now to the special case of the general NLP (3.2) with which we started this section, i.e. the least squares estimation problem (3.1). The Lagrange function for this problem takes the form

$$\mathcal{L}(X, \lambda, \mu) = \frac{1}{2}\mu_0 \|F_1(X)\|_2^2 - \lambda^T F_2(X) - \mu^T F_3(X). \tag{3.16}$$

Deriving this function once and twice with respect to X yields its gradient,

$$\nabla_X \mathcal{L}(X, \lambda, \mu) = \mu_0 J_1(X)^T F_1(X) - J_2(X)^T \lambda - J_3(X)^T \mu, \tag{3.17a}$$

and its Hessian matrix,

$$\begin{aligned} \nabla^2_X \mathcal{L}(X, \lambda, \mu) &= \mu_0 J_1(X)^T J_1(X) \\ &\quad + \mu_0 \sum_{i=1}^{m_1} F_{1i}(X) \nabla^2 F_{1i}(X) - \sum_{i=1}^{m_2} \lambda_i \nabla^2 F_{2i}(X) - \sum_{i=1}^{m_3} \mu_i \nabla^2 F_{3i}(x) \\ &= \mu_0 J_1(X)^T J_1(X) + R(X, \lambda, \mu), \end{aligned} \tag{3.17b}$$

where $J_i(X) = \nabla F_i(X),\ i = 1,2,3$, denotes the Jacobian matrix of the function $F_i,\ i = 1,2,3$, respectively.

We see that the gradient and the Hessian of the Lagrangian function, especially the parts referring to the objective function, have a special structure. This is a good reason to use specialized procedures for this problem class instead of general methods for nonlinear optimization problems.

The idea of the constrained Gauss-Newton method now consists in neglecting the second order terms of the Lagrangian Hessian summarized in $R(X,\lambda,\mu)$ and choosing $L^k = \mu_0 J_1(X^k)^T J_1(X^k)$ as the approximation of the Hessian of the Lagrangian in the k-th iteration.

Assuming again that a constraint qualification like the LICQ condition holds for our problem, i.e. that we can choose $\mu_0 = 1$, then the QP to be solved at each iteration takes the form

$$\begin{array}{ll} \min\limits_{\Delta X} & \frac{1}{2}\Delta X^T J_1(X^k)^T J_1(X^k)\,\Delta X + F_1(X^k)^T J_1(X^k)\Delta X \\ \text{s.t.} & J_2(X^k)\Delta X + F_2(X^k) = 0, \\ & J_3(X^k)\Delta X + F_3(X^k) \geq 0. \end{array} \tag{3.18}$$

It is easy to see that this equivalent to the problem

$$\begin{array}{ll} \min\limits_{\Delta X} & \frac{1}{2}\|J_1(X^k)\Delta X + F_1(X^k)\|_2^2 \\ \text{s.t} & J_2(X^k)\Delta X + F_2(X^k) = 0, \\ & J_3(X^k)\Delta X + F_3(X^k) \geq 0. \end{array} \tag{3.19}$$

In practical implementations of the method it is rather inefficient to compute $L^k = J_1(X^k)^T J_1(X^k)$ explicitly at each iteration for a direct solution of the problem. A great disadvantage of this procedure is that the condition number of $J_1(X)^T J_1(X)$ is the square of the condition number of $J_1(X)$. This yields the risk of unnecessary errors in the computation of the search direction as a common feature of nonlinear least squares problems for parameter estimation is their inherent ill-conditioning which results from the underlying mathematical model being ill-defined (cf. Gill et al. [63]).

A better approach is solving the linear least squares problem (3.18) by means of orthogonal factorization or singular value decomposition. For the case that $J_1(X^k)$ does not have full column rank, special care has to be taken in the implementation. In this rank-deficient case, the search direction is no longer unique. It may be chosen as the solution of minimal norm. This requires a strategy for estimating the rank of $J_1(X^k)$ which is unknown in practice and thus has to be estimated in the course of the computation. But when using floating point arithmetic (as it is the case on any computer), the definition of the "rank" is problem dependent. An explicit estimation of the scaling is necessary, i.e. the determination of those quantities which are considered to be "neglectable" (cf. again Gill et al. [63]).

3.1.4. Local convergence of the Gauss-Newton method

It can be shown (e.g. Bock [24]) that in a neighborhood of an FJ point $(\hat{X}, \hat{\lambda}, \hat{\mu})$ of the nonlinear problem (3.2) the local behavior of the equality and inequality constrained linearized problem (3.19) is equivalent to that of an equality constrained problem where the equality and active inequality constraints have been framed together. We denote these equality constraints by $\tilde{F}_2 = \begin{pmatrix} F_2 \\ (F_{3i})_{i \in I_{F_3}} \end{pmatrix}$ and the number of these active constraints is denoted by $\tilde{m}_2 = m_2 + |I_{F_3}|$.

We will assume that for this equality constrained problem the standard regularity assumptions hold that

- the rank of the constraints Jacobian $\tilde{J}_2 = \nabla \tilde{F}_2$ equals the number of constraints,

$$\operatorname{rank}(\tilde{J}_2(X)) = \tilde{m}_2; \tag{3.20a}$$

- the rank of the full Jacobian of the cost function and the active constraints $\mathcal{J} = \begin{pmatrix} J_1 \\ \tilde{J}_2 \end{pmatrix}$ equals the number of unknowns,

$$\operatorname{rank}(\mathcal{J}(X)) = n. \tag{3.20b}$$

Then the solution of the linearized problem (3.19) can be given by means of a generalized inverse as described in the following theorem.

Theorem 3.1.5 (Existence of a generalized inverse [24])**.** *Consider the equality constrained linearized optimization problem*

$$\begin{array}{ll} \min\limits_{\Delta X} & \frac{1}{2}\|J_1 \Delta X + F_1\|_2^2 \\ s.t & \tilde{J}_2 \Delta X + \tilde{F}_2 = 0, \end{array} \tag{3.21}$$

where J_1 and $\tilde{J}_2$ satisfy the regularity assumptions given above. Then:

(i) For arbitrary $\mathcal{F} = \begin{pmatrix} F_1 \\ \tilde{F}_2 \end{pmatrix} \in \mathbb{R}^{m_1 + \tilde{m}_2}$ there exists one and only one FJ point $(\Delta X, \tilde{\lambda})$ of (3.21) and ΔX is a strict local minimum.

(ii) There exists a linear mapping $\mathcal{J}^+ : \mathbb{R}^{m_1 + \tilde{m}_2} \to \mathbb{R}^n$ such that $\Delta X = -\mathcal{J}^+ \mathcal{F}$ is a solution of (3.21) for arbitrary $\mathcal{F} \in \mathbb{R}^{m_1 + \tilde{m}_2}$.

(iii) The solution operator $\mathcal{J}^+$ is a generalized inverse that satisfies the condition

$$\mathcal{J}^+ \mathcal{J} \mathcal{J}^+ = \mathcal{J}^+.$$

The generalized inverse has the explicit representation

$$\mathcal{J}^+ = \begin{pmatrix} \mathbb{I} & 0 \end{pmatrix} \begin{pmatrix} J_1^T J_1 & \tilde{J}_2^T \\ \tilde{J}_2 & 0 \end{pmatrix}^{-1} \begin{pmatrix} J_1^T & 0 \\ 0 & \mathbb{I} \end{pmatrix}.$$

Under the regularity assumptions (3.20) and some further assumptions on $\mathcal{J}$ and $\mathcal{J}^+$, it can be shown that the Gauss-Newton method converges locally. The following theorem is a summary of the local contraction theorem given by Bock [24].

Theorem 3.1.6 (Local contraction theorem [24]). *Let the regularity assumptions (3.20) hold and assume further that, for all X, Y in the domain D of $\mathcal{F}$ with $Y - X = -\mathcal{J}^+(X)\mathcal{F}(X)$ and for all $t \in [0,1]$, the matrices $\mathcal{J}$ and $\mathcal{J}^+$ satisfy the following (weighted) Lipschitz conditions,*

$$\frac{\|\mathcal{J}^+(Y)\big(\mathcal{J}(X + t(Y - X)) - \mathcal{J}(X)\big)(Y - X)\|_2}{t\|Y - X\|_2^2} \leq \omega(X) \leq \omega, \tag{3.22}$$

$$\frac{\|(\mathcal{J}^+(Y) - \mathcal{J}^+(X))R(X)\|_2}{\|Y - X\|_2} \leq \kappa(X) \leq \kappa, \tag{3.23}$$

with $\omega < \infty$, $\kappa < 1$ and the residual $R(X) = \mathcal{F}(X) - \mathcal{J}(X)\mathcal{J}^+(X)\mathcal{F}(X)$. Then for any initial guess $X^0 \in D$ with

$$\delta_0 = \frac{\omega\|\Delta X^0\|_2}{2} + \kappa < 1, \qquad D_0 = \bar{B}\left(X^0, \frac{\|\Delta X^0\|_2}{1 - \delta_0}\right) \subset D,$$

there exists a stationary point $X^* \in D_0$, i.e. $\mathcal{J}^+(X^*)\mathcal{F}(X^*) = 0$, such that $X^k \to X^*$ as $k \to \infty$. Furthermore, the a priori estimate

$$\|\Delta X^{k+1}\|_2 \leq \left(\frac{\omega\|\Delta X^k\|_2}{2} + \kappa\right)\|\Delta X^k\|$$

holds, which means that the convergence is at least linear with rate κ.

3.1.5. Numerical solution of the linearized problem

The solution of the linearized problems (3.21) is decisive for the performance of the method. As the starting point of our optimization is a dynamically constrained minimization problem, the equality constraints contain the boundary value problem arising from the discretization of the DAE as described in section 2.3.2. Efficient solution methods thus have to take into account the special structure of the constraints. This block structure of the constraints Jacobian $\tilde{J}_2$ can be exploited by efficient condensing methods, as described by Bock [24] and Bock et al. [27] for multiple shooting, or by Schulz [125] or Cervantes and Biegler [34] for collocation methods. Such condensing procedures significantly reduce the dimensions of the variables and the constraints in the linearized problems that have to be solved at each iteration.

We now shortly describe a condensing approach for multiple shooting that is based on *LQ* decompositions of the block matrices in the constraints Jacobian, see also Bock et al. [27]. In chapter 6 we will come back to this approach and show its adaption to the systems arising in the treatment of uncertain optimization problems by means of stochastic collocation.

So, in effect, the optimization problem that we consider with a fixed active set has the form

$$\begin{array}{ll}\min\limits_{(s,p)} & \frac{1}{2}\|F_1(s,p)\|_2^2 \\ \text{s.t.} & \tilde{F}_2(s,p)=0, \\ & h_j(s_j,s_{j+1},p) := y(\tau_{j+1};s_j,p)-s_{j+1}^y=0, \quad j=0,\dots,N-1, \\ & g(\tau_j,s_j,p)=0, \quad j=0,\dots,N.\end{array}$$

where, like in problem (2.11), s denotes the multiple shooting variables, p are the unknown parameters, $h_j(s_j,s_{j+1},p)$, $j=0,\dots,N-1$, and g comprise the multiple shooting matching conditions, and $\tilde{F}_2$ contains all equality constraints and all active inequality constraints.

Then the linearized optimization problem that has to be solved at each iteration turns out to be equivalent to the linear equation system $\mathcal{J}(X^k)\Delta X = -\mathcal{F}(X^k)$, where $X^k=(s^k,p^k)$, $\Delta X=(\Delta s^T,\Delta p^T)^T$. With the following notations,

$$\begin{aligned} g_0 &= g(\tau_0,s_0,p), & \bar{h}_j &= \begin{pmatrix} h_j(s_j,s_{j+1},p) \\ g(\tau_{j+1},s_{j+1},p)\end{pmatrix}, \\ D_1^p &= \frac{\partial F_1}{\partial p}, & D_1^j &= \frac{\partial F_1}{\partial s_j},\ j=0,\dots,N, \\ D_2^p &= \frac{\partial \tilde{F}_2}{\partial p}, & D_2^j &= \frac{\partial \tilde{F}_2}{\partial s_j},\ j=0,\dots,N, \\ D_g^p &= \frac{\partial g(\tau_0,s_0,p)}{\partial p}, & D_g^0 &= \frac{\partial g(\tau_0,s_0,p)}{\partial s_0}, \\ H_j^p &= \begin{pmatrix} \partial y(\tau_{j+1};s_j,p)/\partial p \\ \partial g(\tau_{j+1};s_{j+1},p)/\partial p\end{pmatrix}, & H_j &= \begin{pmatrix} \partial y(\tau_{j+1};s_j,p)/\partial s_j \\ 0\end{pmatrix},\ j=0,\dots,N-1, \end{aligned}$$

$$G_j = \begin{pmatrix} -\mathbb{I} & 0 \\ \partial g(\tau_{j+1};s_{j+1},p)/\partial s_{j+1}^y & \partial g(\tau_{j+1};s_{j+1},p)/\partial s_{j+1}^z\end{pmatrix},\ j=0,\dots,N-1,$$

the matrix $\mathcal{J}$ and the right hand side $\mathcal{F}$ defining the linear system take the form

$$\mathcal{J} = \begin{pmatrix} D_1^0 & D_1^1 & D_1^2 & \cdots & D_1^{N-1} & D_1^N & D_1^p \\ D_2^0 & D_2^1 & D_2^2 & \cdots & D_2^{N-1} & D_2^N & D_2^p \\ D_g^0 & 0 & 0 & \cdots & 0 & 0 & D_g^p \\ H_0 & G_0 & & & & & H_0^p \\ & H_1 & G_1 & & & & H_1^p \\ & & \ddots & & & & \vdots \\ & & & \ddots & & & \vdots \\ & & & & H_{N-1} & G_{N-1} & H_{N-1}^p \end{pmatrix}, \quad \mathcal{F} = \begin{pmatrix} F_1 \\ \tilde{F}_2 \\ g_0 \\ \bar{h}_0 \\ \bar{h}_1 \\ \bar{h}_2 \\ \vdots \\ \vdots \\ \bar{h}_{N-1}\end{pmatrix}.$$

We now transfer this system by applying subsequent LQ-decompositions. The first step is an LQ decomposition of the blocks $[H_0,\ G_0]$. We thus get with corresponding

transformations of the first two block columns of $\mathcal{J}$ and also of the increments Δs_0 and Δs_1,

$$[H_0,\ G_0] = [L_0,\ 0]Q_0, \tag{3.24a}$$

with L_0 being a lower triangular matrix and an orthogonal matrix Q_0,

$$\begin{bmatrix}\Delta s_0\\ \Delta s_1\end{bmatrix} = Q_0^T \begin{bmatrix}\widetilde{\Delta s}_0\\ \widehat{\Delta s}_1\end{bmatrix} \tag{3.24b}$$

$$\begin{bmatrix}D_1^0 & D_1^1\\ D_2^0 & D_2^1\\ D_g^0 & 0\end{bmatrix} = \begin{bmatrix}\tilde{D}_1^0 & \hat{D}_1^1\\ \tilde{D}_2^0 & \hat{D}_2^1\\ \tilde{D}_g^0 & \hat{D}_g^1\end{bmatrix} Q_0, \tag{3.24c}$$

$$[0,\ H_1] = [T_0,\ \tilde{H}_1]Q_0. \tag{3.24d}$$

That way, the matrix $\mathcal{J}$ is transformed to

$$\mathcal{J}^{(1)} = \begin{pmatrix} \tilde{D}_1^0 & \hat{D}_1^1 & D_1^2 & \cdots & D_1^{N-1} & D_1^N & D_1^p\\ \tilde{D}_2^0 & \hat{D}_2^1 & D_2^2 & \cdots & D_2^{N-1} & D_2^N & D_2^p\\ \tilde{D}_g^0 & \hat{D}_g^1 & 0 & \cdots & 0 & 0 & D_g^p\\ L_0 & 0 & & & & & H_0^p\\ T_0 & \tilde{H}_1 & G_1 & & & & H_1^p\\ & & \ddots & & & & \vdots\\ & & & \ddots & & & \vdots\\ & & & & H_{N-1} & G_{N-1} & H_{N-1}^p \end{pmatrix}.$$

We continue by decomposing the blocks $[\tilde{H}_1,\ G_1]$ and corresponding block columns and increments to get

$$[\tilde{H}_1,\ G_1] = [L_1,\ 0]Q_1, \tag{3.25a}$$

again with L_1 being a lower triangular matrix and an orthogonal matrix Q_1,

$$\begin{bmatrix}\widehat{\Delta s}_1\\ \Delta s_2\end{bmatrix} = Q_1^T \begin{bmatrix}\widetilde{\Delta s}_1\\ \widehat{\Delta s}_2\end{bmatrix} \tag{3.25b}$$

$$\begin{bmatrix}\hat{D}_1^1 & D_1^2\\ \hat{D}_2^1 & D_2^2\\ \hat{D}_g^1 & 0\end{bmatrix} = \begin{bmatrix}\tilde{D}_1^1 & \hat{D}_1^2\\ \tilde{D}_2^1 & \hat{D}_2^2\\ \tilde{D}_g^1 & \hat{D}_g^2\end{bmatrix} Q_1, \tag{3.25c}$$

$$[0,\ H_2] = [T_1,\ \tilde{H}_2]Q_1. \tag{3.25d}$$

This leads to the matrix

$$\mathcal{J}^{(2)} = \begin{pmatrix} \tilde{D}_1^0 & \tilde{D}_1^1 & \hat{D}_1^2 & \cdots & D_1^{N-1} & D_1^N & D_1^p \\ \tilde{D}_2^0 & \tilde{D}_2^1 & \hat{D}_2^2 & \cdots & D_2^{N-1} & D_2^N & D_2^p \\ \tilde{D}_g^0 & \tilde{D}_g^1 & \hat{D}_g^2 & \cdots & 0 & 0 & D_g^p \\ L_0 & 0 & & & & & H_0^p \\ T_0 & L_1 & 0 & & & & H_1^p \\ & T_1 & \tilde{H}_2 & G_2 & & & \vdots \\ & & & \ddots & & & \vdots \\ & & & & H_{N-1} & G_{N-1} & H_{N-1}^p \end{pmatrix}.$$

We proceed in this manner and have in the last of these preprocessing steps the transformations

$$[\tilde{H}_{N-1},\ G_{N-1}] = [L_{N-1},\ 0]Q_{N-1}, \tag{3.26a}$$

with lower triangular matrix L_{N-1} and orthogonal matrix Q_{N-1},

$$\begin{bmatrix} \widehat{\Delta s}_{N-1} \\ \Delta s_N \end{bmatrix} = Q_{N-1}^T \begin{bmatrix} \widetilde{\Delta s}_{N-1} \\ \widehat{\Delta s}_N \end{bmatrix} \tag{3.26b}$$

$$\begin{bmatrix} \hat{D}_1^{N-1} & D_1^N \\ \hat{D}_2^{N-1} & D_2^N \\ \hat{D}_g^{N-1} & 0 \end{bmatrix} = \begin{bmatrix} \tilde{D}_1^{N-1} & \tilde{D}_1^N \\ \tilde{D}_2^{N-1} & \tilde{D}_2^N \\ \tilde{D}_g^{N-1} & \tilde{D}_g^N \end{bmatrix} Q_{N-1}. \tag{3.26c}$$

This leaves us with the final matrix

$$\mathcal{J}^{(N)} = \begin{pmatrix} \tilde{D}_1^0 & \tilde{D}_1^1 & \tilde{D}_1^2 & \cdots & \tilde{D}_1^{N-1} & \tilde{D}_1^N & D_1^p \\ \tilde{D}_2^0 & \tilde{D}_2^1 & \tilde{D}_2^2 & \cdots & \tilde{D}_2^{N-1} & \tilde{D}_2^N & D_2^p \\ \tilde{D}_g^0 & \tilde{D}_g^1 & \tilde{D}_g^2 & \cdots & \tilde{D}_g^{N-1} & \tilde{D}_g^N & D_g^p \\ L_0 & 0 & & & & & H_0^p \\ T_0 & L_1 & 0 & & & & H_1^p \\ & T_1 & L_2 & 0 & & & \vdots \\ & & \ddots & \ddots & & & \vdots \\ & & & T_{N-2} & L_{N-1} & 0 & H_{N-1}^p \end{pmatrix}.$$

Assuming that we know the step Δp in the parameters, we can compute the step $\widetilde{\Delta s}_0$ as

$$\widetilde{\Delta s}_0 = -L_0^{-1}(\bar{h}_0 + H_0^p \Delta p) = \tilde{h}_0 + \tilde{H}_0^p \Delta p,$$

where $\tilde{h}_0 = -L_0^{-1}\bar{h}_0$ and $\tilde{H}_0^p = -L_0^{-1}H_0^p$. From this, we get recursively the remaining components of the step vector $\widetilde{\Delta s}$ by computing for $i = 1, \ldots, N-1$,

$$\widetilde{\Delta s}_i = -L_i^{-1}(\bar{h}_i + H_i^p \Delta p + T_{i-1}\widetilde{\Delta s}_{i-1}) = \tilde{h}_i + \tilde{H}_i^p \Delta p, \tag{3.27}$$

where we have $\tilde{h}_i = -L_i^{-1}(\bar{h}_i + T_{i-1}\tilde{h}_{i-1})$ and $\tilde{H}_i^p = -L_i^{-1}(H_i^p + T_{i-1}\tilde{H}_{i-1}^p)$. The step component for the last multiple shooting node, $\widehat{\Delta s}_N$, cannot be computed in this way but has to be estimated together with the parameter step Δp from the objective function and the equality constraints of the condensed problem,

$$\begin{aligned} \min_{\widehat{\Delta s}_N, \Delta p} \quad & \tfrac{1}{2}\left\| \tilde{D}_1^N \widehat{\Delta s}_N + \tilde{D}_1^p \Delta p + \tilde{r}_1 \right\|_2^2 \\ \text{s.t.} \quad & \tilde{D}_2^N \widehat{\Delta s}_N + \tilde{D}_2^p \Delta p + \tilde{r}_2 = 0, \\ & \tilde{D}_g^N \widehat{\Delta s}_N + \tilde{D}_g^p \Delta p + \tilde{g}_0 = 0, \end{aligned}$$

where we used the abbreviations

$$\begin{aligned} \tilde{D}_1^p &:= D_1^p + \sum_{i=0}^{N-1} \tilde{D}_1^i \tilde{H}_i^p, & \tilde{D}_2^p &:= D_2^p + \sum_{i=0}^{N-1} \tilde{D}_2^i \tilde{H}_i^p, & \tilde{D}_g^p &:= D_g^p + \sum_{i=0}^{N-1} \tilde{D}_g^i \tilde{H}_i^p, \\ \tilde{r}_1 &:= F_1 + \sum_{i=0}^{N-1} \tilde{D}_1^i \tilde{h}_i, & \tilde{r}_2 &:= \tilde{F}_2 + \sum_{i=0}^{N-1} \tilde{D}_2^i \tilde{h}_i & \tilde{g}_0 &:= g_0 + \sum_{i=0}^{N-1} \tilde{D}_g^i \tilde{h}_i. \end{aligned}$$

Having solved this system and thus obtained $\widehat{\Delta s}_N$ and Δp, we first compute the remaining components $\widetilde{\Delta s}_i$ for $i = 1, \ldots, N-1$ from (3.27) and $\widetilde{\Delta s}_0$. Then it remains the task of undoing the orthogonal transformations of the matrices Q_i of the used LQ-decompositions. This is done by means of a backward recursion,

$$\begin{bmatrix} \widehat{\Delta s}_i \\ \Delta s_{i+1} \end{bmatrix} = Q_i^T \begin{bmatrix} \widetilde{\Delta s}_i \\ \widehat{\Delta s}_{i+1} \end{bmatrix}, \quad i = N-1, \ldots, 1, \qquad \begin{bmatrix} \Delta s_0 \\ \Delta s_1 \end{bmatrix} = Q_0^T \begin{bmatrix} \widetilde{\Delta s}_0 \\ \widehat{\Delta s}_1 \end{bmatrix}. \tag{3.28}$$

3.1.6. Statistical analysis of the solution and confidence regions

As the measurement errors ε are assumed to be random variables, the parameter estimates $X^*(\varepsilon)$ are also random variables. So to know their variances would be a valuable information.

Assume that the covariance of the measurement errors is known to be $\Sigma = \mathbb{E}(\varepsilon\varepsilon^T)$. Then consider the linearized system (3.21) at the solution,

$$\begin{aligned} \min_{\Delta X} \quad & \tfrac{1}{2}\|J_1(X^*)\Delta X + F_1(X^*)\|_2^2 \\ \text{s.t.} \quad & \tilde{J}_2(X^*)\Delta X + 0 = 0. \end{aligned} \tag{3.29}$$

Due to the definition of the objective function as $F_1(X) = \Sigma^{-\frac{1}{2}}(\eta - m(X)) = \Sigma^{-\frac{1}{2}}\varepsilon$, the vector $m(X)$ denoting the model responses corresponding to the measurements, the components of this function vector are also random variables with the expected value $\mathbb{E}(F_1) = \mathbb{E}(\Sigma^{-\frac{1}{2}}\varepsilon) = 0$ as $\mathbb{E}(\varepsilon) = 0$ and the variance $\mathbb{E}(F_1 F_1^T) = \mathbb{E}(\Sigma^{-\frac{1}{2}}\varepsilon\varepsilon^T\Sigma^{-\frac{1}{2}}) = \Sigma^{-\frac{1}{2}}\Sigma\Sigma^{-\frac{1}{2}} = \mathbb{I}$. So with the standard assumption that the measurement errors are distributed according to $\varepsilon \sim \mathcal{N}(0, \Sigma)$, the optimal values of the objective function follow a standard normal distribution $F_1(X^*) \sim \mathcal{N}(0, \mathbb{I})$.

We know from section 3.1.4 that the solution of (3.29) is given by means of a generalized inverse as $\Delta X = -\mathcal{J}^+(X^*)\begin{pmatrix} F_1(X^*) \\ 0 \end{pmatrix}$. So for its expected value and its variance follow

$$\begin{aligned} \mathbb{E}(\Delta X) &= 0, \\ \mathbb{E}(\Delta X \Delta X^T) &= \mathcal{J}^+(X^*)\mathbb{E}\left(\begin{pmatrix} F_1(X^*) \\ 0 \end{pmatrix}\begin{pmatrix} F_1(X^*) \\ 0 \end{pmatrix}^T\right)\mathcal{J}^+(X^*)^T \\ &= \mathcal{J}^+(X^*)\begin{pmatrix} \mathbb{I} & 0 \\ 0 & 0 \end{pmatrix}\mathcal{J}^+(X^*)^T. \end{aligned}$$

Defining $\mathcal{C} := \mathbb{E}(\Delta X \Delta X^T)$ as the covariance matrix of the estimate, we get a measure for the variance of the estimate subject to random (normally distributed) measurement errors. Note that we have to compute the generalized inverse anyway during application of the generalized Gauss-Newton method. Therefore we can approximate $\mathcal{C}$ cheaply using the generalized inverse $\mathcal{J}^+(X^{k^*})$ from the last iterate.

But often in practice the covariance Σ of the measurement errors is only known up to a common factor β^2, i.e. $\Sigma = \beta^2\hat{\Sigma}$ with $\Sigma = \mathbb{E}(\varepsilon\varepsilon^T)$, $\tilde{\varepsilon} = \Sigma^{-\frac{1}{2}}\varepsilon$. If, for instance, the measurement errors are independent, normally distributed with zero mean and relative, i.e. β percent with β unknown, the objective function is usually weighted with the measurements, $\hat{F}_{1i}(X) = (\eta_i - m_i(X))/\eta_i$, such that the variance matrix is given by $\hat{\Sigma} = \operatorname{diag}(\eta_1^2, \ldots, \eta_{m_1}^2)$. Using $\hat{F}_1(\varepsilon) = \hat{\Sigma}^{-\frac{1}{2}}(\eta - m(X)) = \hat{\Sigma}^{-\frac{1}{2}}\varepsilon$ as objective function does not change the estimate $X^*(\varepsilon)$, but the variance of the optimal objective values becomes $\mathbb{E}(\hat{F}_1\hat{F}_1^T) = \beta^2\mathbb{I}$. Therefore, also the covariance matrix of the estimates is changed to $\mathcal{C} = \mathcal{J}^+(X^*)\begin{pmatrix} \beta^2\mathbb{I} & 0 \\ 0 & 0 \end{pmatrix}\mathcal{J}^+(X^*)^T$. In this case we can use $b^2 = \frac{\|F_1(X^*)\|_2^2}{m_1-(n-m_2)}$ as an estimate for β^2. So the covariance matrix of the estimates is set to $\mathcal{C} := b^2\mathcal{J}^+(X^*)\begin{pmatrix} \mathbb{I} & 0 \\ 0 & 0 \end{pmatrix}\mathcal{J}^+(X^*)^T$.

Thus, for normally distributed measurement errors $\varepsilon \sim \mathcal{N}(0, \Sigma)$, the estimates $X^*(\varepsilon)$ are also normally distributed with $X^*(\varepsilon) \sim \mathcal{N}(X^{true}, \mathcal{C}^{true})$, where X^{true} is the unknown true value and $\mathcal{C}^{true}$ the covariance matrix at X^{true}.

We want to give a confidence region for the actual estimate X^* in which the true parameters X^{true} lie with an error probability α.

For unconstrained optimization problems, a linearized confidence region is given by the ellipsoid

$$G_L(\alpha) := \{X \mid (X - X^*)^T\mathcal{C}^{-1}(X - X^*) \leq \gamma(\alpha)\}.$$

By choosing $\gamma(\alpha) := \chi_n^2(1-\alpha)$ as the $(1-\alpha)$ quantile of the χ^2 distribution with n degrees of freedom, we ensure that the probability that the true parameters X^{true} lie in this linearized confidence region is

$$\mathbb{P}\big(X^{true} \in G_L(\alpha)\big) = \mathbb{P}\big((X^{true} - X^*)^T\mathcal{C}^{-1}(X^{true} - X^*) \leq \gamma(\alpha)\big) = 1 - \alpha.$$

This is due to the fact that under the assumption that $\mathcal{C} = \mathcal{C}^{true}$, i.e. $X^* \sim \mathcal{N}(X^{true}, \mathcal{C})$, the random variable $y := \mathcal{C}^{-\frac{1}{2}}(X - X^*)$ is normally distributed with zero mean and unit variance and thus $y^T y$ follows a χ^2 distribution with n degrees of freedom.

If we have a constrained optimization problem with $\tilde{m}_2 > 0$, the optimal estimate X^* always lies on the manifold described by the equalities $\tilde{F}_2(X^*) = 0$. In this case we define the linearized confidence region as

$$G_L(\alpha) := \{X^* + \delta X \mid \delta X^T J_1^T J_1 \delta X \leq \gamma(\alpha),\ \tilde{J}_2 \delta X = 0\}.$$

As δX now lies in a $(n - \tilde{m}_2)$-dimensional subspace of $\mathbb{R}^n$, the bound $\gamma(\alpha)$ is now chosen as the $(1-\alpha)$ quantile of the χ^2 distribution with $n - \tilde{m}_2$ degrees of freedom, $\gamma(\alpha) := \chi^2_{n-\tilde{m}_2}(1-\alpha)$. It hold the equalities

$$\begin{aligned} G_L(\alpha) &= \{X^* - \mathcal{J}^+(X^*) \begin{pmatrix} \varepsilon \\ 0 \end{pmatrix} \mid \|\varepsilon\|_2^2 \leq \gamma(\alpha)\} \\ &= \{X^* + \delta X \mid \tilde{J}_2 \delta X = 0,\ \|F_1(X^*) + J_1(X^*)\delta X\|_2^2 \leq \|F_1(X^*)\|_2^2 + \gamma(\alpha)\}. \end{aligned}$$

If again we only know the variance of the measurement errors up to some common factor, $\Sigma = \beta^2 \hat{\Sigma}$ with $\hat{\Sigma}$ as above and β^2 unknown, we can again use the estimate $b^2 = \frac{\|F_1(X^*)\|_2^2}{m_1 + \tilde{m}_2 - n}$ for β^2 and set

$$G_L(\alpha) = \{X^* + \delta X \mid \tilde{J}_2 \delta X = 0,\ \|F_1(X^*) + J_1(X^*)\delta X\|_2^2 \leq \|F_1(X^*)\|_2^2 + b^2\gamma(\alpha)\}.$$

But now $\gamma(\alpha)$ is a multiple of the the $(1-\alpha)$ quantile of the Fischer distribution F_{l_1,l_2} with $l_1 = n - \tilde{m}_2$ degrees of freedom in X and $l_2 = m_1 - (n - \tilde{m}_2)$ degrees of freedom in η, i.e. $\gamma(\alpha) := l_1 F_{l_1,l_2}(1-\alpha)$.

For further reading and more general nonlinear confidence regions see Bock [24] and Nattermann [99].

3.2. The Gauss-Newton method in robust optimization

We will now consider the solution of robust optimization problems with the l_1 norm or the Huber function as objective function. Each of these problems is treated on its own in the following sections in order to formally derive the corresponding optimality conditions. We will see that the problems share a common structure so that we can treat the two problems as special cases of a superordinate problem.

3.2.1. Optimality conditions

As we already said in the previous chapter, both robust estimators suffer from objective functions that are not differentiable to second order. The l_1 norm in the objective function of the LAD estimator is not continuously differentiable at all and Huber's function ρ_γ is only once continuously differentiable. Thus we have to rewrite these

estimation problems in order to get optimization problems that can be treated with the Gauss-Newton method. For the least absolute deviations estimator, we replace the non-smooth objective function (2.15) by the optimization problem

$$\begin{array}{ll} \min\limits_{X\in\mathbb{R}^n, u\in\mathbb{R}^{m_1}, v\in\mathbb{R}^{m_1}} & \sum\limits_{i=1}^{m_1}(u_i+v_i) \\ \text{s.t.} & F_{1i}(X) = u_i - v_i, \quad i=1,\ldots,m_1, \\ & \qquad u_i, v_i \geq 0, \quad i=1,\ldots,m_1. \end{array} \tag{3.30}$$

The Huber estimator does not involve nonsmooth optimization as it has a continuously differentiable objective function. Its derivative is given by

$$\rho'(X) = \sum_{i=1}^{m_1} \rho'_\gamma(F_{1i}(X))\nabla F_{1i}(X),$$

with

$$\rho'_\gamma(z) = \begin{cases} z, & |z| \leq \gamma, \\ \text{sign}\,(z)\cdot\gamma, & |z| > \gamma. \end{cases} \tag{3.31}$$

However, this problem still does not have a continuous second derivative.

Defining the m_1-vector ω by

$$\omega_i = \rho'_\gamma(F_{1i}(X)), \tag{3.32}$$

the necessary condition of the unconstrained problem is given by

$$\nabla F_1(X)\omega = 0. \tag{3.33}$$

We use theorem 2.7.1 to rewrite the unconstrained Huber optimization problem (2.19) in a more suitable form. Therefore, we start this section with a reminder of this theorem in a slightly different form and its proof.

Theorem 3.2.1 (Reformulations of Huber's nonlinear estimator)**.** *A solution X of the unconstrained nonlinear Huber minimization problem*

$$\min_X \rho(F_1(X)) = \sum_i \rho_\gamma(F_{1i}(X))$$

with $\rho_\gamma(\cdot)$ defined by

$$\rho_\gamma(z) = \begin{cases} \frac{1}{2}z^2, & \text{if } |z| \leq \gamma, \\ |z|\gamma - \frac{1}{2}\gamma^2, & \text{if } |z| > \gamma. \end{cases}$$

is given by any solution (X, w, u, v) of the following problem,

$$\begin{aligned} &\min_{X\in\mathbb{R}^n, w\in\mathbb{R}^{m_1}} \frac{1}{2}\|w\|_2^2 + \gamma\|F_1(X) - w\|_1 \\ \Longleftrightarrow &\begin{cases} \min\limits_{X\in\mathbb{R}^n, w, u, v\in\mathbb{R}^{m_1}} & \frac{1}{2}w^Tw + \gamma\sum\limits_{i=1}^{m_1}(u_i+v_i) \\ s.t. & F_1(X) - w = u - v, \\ & u, v \geq 0. \end{cases} \end{aligned} \tag{3.34}$$

Proof. With $y := (X^T, w^T, u^T, v^T)^T$, problem (3.34) has the corresponding modified Lagrange function

$$\begin{aligned}\mathcal{L}(y, \mu_0, \lambda, \mu^{(1)}, \mu^{(2)}) =& \mu_0\big(\frac{1}{2}w^T w + \gamma \sum_{i=1}^{m_1}(u_i + v_i)\big) \\ & - \lambda^T(F_1(X) - w - u + v) - \mu^{(1)^T} u - \mu^{(2)^T} v, \end{aligned} \tag{3.35}$$

and the FJ conditions

$$\nabla_y \mathcal{L}(y^*, \mu_0^*, \lambda^*, \mu^{(1)*}, \mu^{(2)*}) = \begin{pmatrix} -\nabla F_1(X^*)^T \lambda^* \\ \mu_0^* w^* + \lambda^* \\ \mu_0^* \gamma e + \lambda^* - \mu^{(1)*} \\ \mu_0^* \gamma e - \lambda^* - \mu^{(2)*} \end{pmatrix} = 0, \tag{3.36a}$$

$$F_1(X^*) - w^* - u^* + v^* = 0, \tag{3.36b}$$

$$u^*, v^* \geq 0, \tag{3.36c}$$

$$\mu_0^* \geq 0, \tag{3.36d}$$

$$\mu^{(1)*} \geq 0, \tag{3.36e}$$

$$\mu^{(2)*} \geq 0, \tag{3.36f}$$

$$\sum_{i=1}^{m_1} \mu_i^{(1)*} u_i^* + \mu_i^{(2)*} v_i^* = 0, \tag{3.36g}$$

$$\begin{pmatrix} \lambda \\ \mu \end{pmatrix} \neq 0. \tag{3.36h}$$

If we had $\mu_0^* = 0$, the second equation of (3.36a) would at once give $\lambda^* = 0$ and with this the third and forth equations would give $\mu^{(1)*} = 0 = \mu^{(2)*}$. As this is a clear contradiction to (3.36h), we must have $\mu_0^* \neq 0$. From the first two equations of (3.36a) obviously follows for the vector w^* that $\mu_0^* w^* = -\lambda^*$ and $\mu_0^* \nabla F_1(X^*)^T w^* = 0$, i.e. the vector w^* solves system (3.33). The third and fourth equation of (3.36a) give two equations for λ^* and thus also for w^*, namely $w^* = \mu_0^* \gamma e - \mu^{(1)*} = \mu^{(2)*} - \mu_0^* \gamma e$. This means at the same time that the inequality multipliers have to satisfy the condition $\mu^{(1)*} + \mu^{(2)*} = 2\mu_0^* \gamma e$. By a case distinction with respect to the active and inactive inequality constraints, we can also determine a precise formula for w^*:

1. Let $u_i^* > 0$ and $v_i^* > 0$. Then $\mu_i^{(1)*} = 0 = \mu_i^{(2)*}$ and thus due to $\mu_i^{(1)*} + \mu_i^{(2)*} = 2\mu_0^* \gamma$ also $\mu_0^* = 0$ as we always assume $\gamma > 0$. Otherwise we would have a standard least squares optimization problem. Therefore we also get $\mu_j^{(1)*} = 0 = \mu_j^{(2)*}$ for all $j \neq i$ and with this $w^* = 0 = -\lambda^*$ which is a clear contradiction to the last FJ condition (3.36h).

2. Let $u_i^* > 0$ and $v_i^* = 0$, i.e. $F_{1i}(X^*) > w_i^*$ and $\mu_i^{(1)*} = 0$, $\mu_i^{(2)*} \geq 0$. Therefore we get $\mu_i^{(2)*} = 2\mu_0^* \gamma$ and with this $\mu_0^* \geq 0$ due to $\gamma > 0$ and $w_i^* = \mu_0^* \gamma$. But this also means that $F_{1i}(X^*) > \mu_0^* \gamma > 0$. So we can write $w_i^* = \operatorname{sign}(F_{1i}(X^*))\mu_0^* \gamma$.

3. Let $u_i^* = 0$ and $v_i^* > 0$, i.e. $F_{1i}(X^*) < w_i^*$ and $\mu_i^{(1)*} \geq 0$, $\mu_i^{(2)*} = 0$. Therefore we get $\mu_i^{(1)*} = 2\mu_0^*\gamma$ and with this $\mu_0^* \geq 0$ due to $\gamma > 0$ and $w_i^* = -\mu_0^*\gamma$. This now means that $F_{1i}(X^*) < -\mu_0^*\gamma < 0$ and we can analogically write $w_i^* = \operatorname{sign}(F_{1i}(X^*))\mu_0^*\gamma$.

4. Let $u_i^* = 0$ and $v_i^* = 0$, i.e. $F_{1i}(X^*) = w_i^*$ and $\mu_i^{(1)*} \geq 0$, $\mu_i^{(2)*} \geq 0$. A second case destinction with respect to the signs of the multipliers $\mu_i^{(1)*}, \mu_i^{(2)*}$ shows that the cases where at least one of the multipliers is zero are equivalent to the three cases above. When both multipliers are non-zero, we get that $-\mu_0^*\gamma < w_i^* < \mu_0^*\gamma$ and thus also $|F_{1i}(X^*)| < \mu_0^*\gamma$.

Altogether, this leads to the equation

$$w_i^* = \begin{cases} F_{1i}(X^*), & \text{if } |F_{1i}(X^*)| \leq \gamma, \\ \mu_0^*\gamma \cdot \operatorname{sign}(F_{1i}(X^*)), & \text{if } |F_{1i}(X^*)| > \gamma. \end{cases} \tag{3.37}$$

As the Fritz John conditions for this problem impose no further conditions on μ_0^* beside $\mu_0^* \geq 0$, we can without loss of generality set $\mu_0^* = 1$. So equation (3.37) equals the conditions given by (3.31) and (3.32). So any solution (X^*, w^*) of (3.34) satisfies the necessary optimality conditions of the Huber optimization problem and from any solution $\tilde{X}$ of the latter problem we can readily compute the corresponding pair $(\tilde{X}, \tilde{w})$ of the reformulated problem. In order to finally show the equivalence of the two problems, we prove that values of the objective functions are equal in the respective solution points. The minimum of the Huber optimization problem is given by

$$\sum_{i=1}^{m_1} \rho_\gamma(F_{1i}(\tilde{X})) = \frac{1}{2}F_1(\tilde{X})^T D F_1(\tilde{X}) + \gamma\bar{\alpha}\bar{D}F_1(\tilde{X}) - \frac{1}{2}\gamma^2 e^T \bar{D} e, \tag{3.38}$$

where $D = \operatorname{diag}(d_{ii})$ with $d_{ii} = 1$ if $|F_{1i}(\tilde{X})| \leq \gamma$ and $d_{ii} = 0$ otherwise, $\bar{D} = \mathbb{I} - D$, and $\bar{\alpha}_i = \operatorname{sign}(F_{1i}(\tilde{X}))$. The minimum of the reformulated problem (3.34) on the other hand computes to be

$$\begin{aligned} \frac{1}{2}\|w^*\|_2^2 + \gamma\|F_1(X^*) - w^*\|_1 &= \frac{1}{2}F_1(X^*)^T D F_1(X^*) + \frac{1}{2}\gamma^2\alpha^{*T}\bar{D}\alpha^* \\ &\quad + \gamma\alpha^{*T}\bar{D}(F_1(X^*) - w^*) \\ &= \frac{1}{2}F_1(X^*)^T D F_1(X^*) + \frac{1}{2}\gamma^2\alpha^{*T}\bar{D}\alpha^* \\ &\quad + \gamma\alpha^{*T}\bar{D}F_1(X^*) - \gamma^2\alpha^{*T}\bar{D}\alpha^* \\ &= \frac{1}{2}F_1(X^*)^T D F_1(X^*) + \gamma\alpha^{*T}\bar{D}F_1(X^*) \\ &\quad - \frac{1}{2}\gamma^2 e^T\bar{D}e, \end{aligned} \tag{3.39}$$

where we inserted (3.37) for w^* and used the fact that $\alpha^{*T}\alpha^* = e^T e$. So, obviously, the minima (3.38) and (3.39) are equal in $X^* = \tilde{X}$ which proves the equivalence of the two problems. □

Note that we have also proven here that for the unconstrained Huber optimization problem we can always choose $\mu_0^* = 1$ in the FJ conditions without needing any constraint qualifications for the constraints in the problem reformulation (3.34).

For robust optimization problems with equality and inequality constraints, problems (3.30) and (3.34) can be simply extended by these additional constraints. Thus we have rewritten both robust optimization problems as special cases of the general nonlinear optimization problem

$$\begin{aligned} \min_X \quad & \rho(F_1(X)) \\ \text{s.t.} \quad & F_2(X) = 0, \\ & F_3(X) \geq 0, \end{aligned} \tag{3.40}$$

treated in the previous section and we can compute their specific optimality conditions by applying theorems 3.1.1, 3.1.2, and 3.1.3.

The l_1 norm estimator

First, we will show that also for the reformulation (3.30) of the l_1 norm minimization problem the Fritz John optimality conditions can be satisfied with $\mu_0^* = 1$. With $y^T := (X^T, u^T, v^T)$, the modified Lagrangian for this problem is

$$\mathcal{L}(y, \mu_0, \lambda, \mu^{(1)}, \mu^{(2)}) = \mu_0 \sum_{i=1}^{m_1} (u_i + v_i) - \lambda(F_1(X) - u + v) - {\mu^{(1)}}^T u - {\mu^{(2)}}^T v,$$

and the FJ optimality conditions are

$$\nabla_y \mathcal{L}(y^*, \mu_0^*, \lambda^*, \mu^{(1)*}, \mu^{(2)*}) = \begin{pmatrix} -\nabla F_1(X^*)^T \lambda^* \\ \mu_0^* e + \lambda^* - \mu^{(1)*} \\ \mu_0^* e - \lambda^* - \mu^{(2)*} \end{pmatrix} = 0, \tag{3.41a}$$

$$F_1(X^*) - u^* + v^* = 0, \tag{3.41b}$$

$$u^*, v^* \geq 0, \tag{3.41c}$$

$$\mu_0^* \geq 0, \tag{3.41d}$$

$$\mu^{(1)*} \geq 0, \tag{3.41e}$$

$$\mu^{(2)*} \geq 0, \tag{3.41f}$$

$$\sum_{i=1}^{m_1} \mu_i^{(1)*} u_i + \mu_i^{(2)*} v_i = 0, \tag{3.41g}$$

$$\begin{pmatrix} \lambda^* \\ \mu^* \end{pmatrix} \neq 0. \tag{3.41h}$$

If we had $\mu_0^* = 0$, the second and third equation of (3.41a) together with (3.41e) and (3.41f), respectively, would give $\lambda^* = \mu^{(1)*} \geq 0$ and $-\lambda^* = \mu^{(2)*} \geq 0$. This obviously means $\lambda^* = 0$ and therefore also $\mu^{(1)*} = 0 = \mu^{(2)*}$. This contradicts (3.41h) and thus

we must have $\mu_0^* \neq 0$. Then the first equation of (3.41a) defines λ^* as the solution of $\nabla F_1(X^*)^T\lambda^* = 0$. If this system had only the trivial solution $\lambda^* = 0$, then we would get $\mu_0^* = \mu_i^{(1)*} = \mu_i^{(2)*} \geq 0$ for all $i = 1, \ldots, m_1$. As they may not all be zero according to (3.41h), they must be all positive and thus $u = 0 = v$ due to (3.41c) and (3.41g). Then (3.41b) yields $F_1(X^*) = 0$. So in this case we can simply set $\mu_0^* = 1 = \mu_i^{(1)*} = \mu_i^{(2)*}$ for all $i = 1, \ldots, m_1$. But if there exists a $j \in \{1, \ldots m_1\}$ such that $F_{1j}(X^*) \neq 0$, we have to ensure that the equation system $\nabla F_1(X^*)^T\lambda^* = 0$ has a non-trivial solution in order that the FJ conditions can be satisfied. This means we have to impose some kind of constraint qualification on the rank of $\nabla F_1(X^*)$. Define

$$I_{F_1}(X) = \{i = 1, \ldots, m_1 \mid F_{1i}(X) = 0\} \text{ with } |I_{F_1}(X)| = m_1^0.$$

Then the condition $\text{rank}\nabla F_1(X^*) = |I_{F_1}(X^*)|$ guarantees that $\lambda^* \neq 0$ if $F_{1j}(X^*) \neq 0$ for at least one j. So we now get from (3.41a), (3.41e), and (3.41f) the conditions $\mu_i^{(1)*} = \mu_0^* + \lambda_i^* \geq 0$ and $\mu_i^{(2)*} = \mu_0^* - \lambda_i^* \geq 0$. In order to ensure positivity for both $\mu^{(1)*}$ and $\mu^{(2)*}$, we need $\mu_0^* > 0$, which we have already shown to hold. But if the FJ conditions are satisfied for some $\mu_0^* > 0$, they are also satisfied with $(\bar{\mu_0}^*, \bar{\lambda}^*, \bar{\mu}^{(1)*}, \bar{\mu}^{(2)*}) := (1, \lambda^*/\mu_0^*, \mu^{(1)*}/\mu_0^*, \mu^{(2)*}/\mu_0^*)$.

So, if the system of optimality conditions for the reformulation of the unconstrained LAD problem under the assumption $\text{rank}\nabla F_1(X^*) = |I_{F_1}(X^*)|$ has a solution with $\mu_0^* > 0$, it is without loss of generality solvable with $\mu_0^* = 1$ by a simple rescaling of the multipliers.

Now, for the constrained case, we modify the linear independence constraint qualification for the least absolute deviations minimization problem to comprise not only the gradients of the equality and active inequality constraints, but also the gradients of the *active measurements*, i.e. the gradients of those components of the objective function $F_{1i}(X)$ with $i \in I_{F_1}(X)$.

Thus a regular point of the LAD minimization problem is characterized by the rank condition

$$\text{rank}\nabla F_A(X) = m_1^0 + m_2 + m_3^0 \text{ where } \nabla F_A = \begin{pmatrix} \nabla \tilde{F}_1(X) \\ \nabla F_2(X) \\ \nabla \tilde{F}_3(X) \end{pmatrix}, \tag{3.42}$$

$$\tilde{F}_1(X) = (F_{1i}(X))_{i \in I_{F_1}(X)}, \quad \tilde{F}_3(X) = (F_{3i}(X))_{i \in I_{F_3}(X)}, \; m_3^0 = |I_{F_3}(X)|.$$

The following theorems give the explicit optimality conditions for the least absolute deviations estimator.

Theorem 3.2.2 (Necessary optimality conditions for the LAD estimator)**.** *Let X^* be a regular solution of the* LAD *estimation problem. Then there exist vectors $\nu \in \mathbb{R}^{m_1}$, $\lambda \in \mathbb{R}^{m_2}$, and $\mu \in \mathbb{R}^{m_3}$ such that the following first-order necessary optimality*

conditions are satisfied,

$$\nabla F_1(X^*)^T\nu^* - \nabla F_2(X^*)^T\lambda^* - \nabla F_3(X^*)^T\mu^* = 0, \tag{3.43a}$$
$$F_2(X^*) = 0, \tag{3.43b}$$
$$F_3(X^*) \geq 0, \tag{3.43c}$$
$$|\nu_i^*| \leq 1 \quad \forall i = 1, \ldots, m_1 \text{ and} \tag{3.43d}$$
$$\nu_i^* = sign(F_{1i}(X^*)) \quad \text{if } F_{1i}(X^*) \neq 0,$$
$$\mu^* \geq 0, \tag{3.43e}$$
$$\sum_{i=1}^{m_3} \mu_i^* F_{3i}(X^*) = 0. \tag{3.43f}$$

Furthermore, the point X^* and the Lagrange multipliers ν^*, λ^*, μ^* satisfy the second-order necessary conditions

$$d^T \left(\sum_{i=1}^{m_1} \nu_i^* \nabla^2 F_{1i}(X^*) - \sum_{i=1}^{m_2} \lambda_i^* \nabla^2 F_{2i}(X^*) - \sum_{i=1}^{m_3} \mu_i^* \nabla^2 F_{3i}(X^*) \right) d \geq 0, \tag{3.44a}$$
$$\forall i \in I_{F_3}(X^*): \quad \mu_i^* \nabla F_{3i}(X^*) d = 0, \tag{3.44b}$$

for all critical directions d defined by

$$\begin{aligned} D_1(X^*, \nu^*) = \{ d \in \mathbb{R}^n \mid & \nabla F_{1i}(X^*) d = 0 \text{ for all } i \text{ with } |\nu_i^*| < 1, \\ & \nu_i^* \nabla F_{1i}(X^*) d \geq 0 \text{ for all } i \in I_{F_1}(X^*) \text{ with } |\nu_i^*| = 1, \\ & \nabla F_2(X^*) d = 0, \\ & \nabla F_{3i}(X^*) d \geq 0 \text{ for all } i \in I_{F_3}(X^*) \}. \end{aligned} \tag{3.45}$$

Proof. The proof of this theorem is a direct extension of the proof given by Binder and Kostina [20] for the LAD estimation problem with equality constraints only. It applies the Fritz John optimality conditions of theorems 3.1.1, 3.1.2 to the reformulation (3.30) of the objective function (2.15) in (3.40) of the l_1 norm minimization problem. By case distinctions, the additional variables u, v can be eliminated such that the given result only depends on the original optimization variables X. □

Theorem 3.2.3 (Sufficient optimality conditions for the LAD estimator)**.** *Let the point $(X^*, \nu^*, \lambda^*, \mu^*)$ satisfy the first-order optimality conditions (3.43) and let the matrix*

$$H^* = \sum_{i=1}^{m_1} \nu_i^* \nabla^2 F_{1i}(X^*) - \sum_{i=1}^{m_2} \lambda_i^* \nabla^2 F_{2i}(X^*) - \sum_{i=1}^{m_3} \mu_i^* \nabla^2 F_{3i}(X^*)$$

be positive definite for all non-zero critical directions $d \in D_1(X^*, \nu^*)$, i.e.

$$d^T H d > 0 \quad \text{for all } d \in D_1(X^*, \nu^*),\ d \neq 0.$$

Then X^* is a strict local minimizer of the least absolute deviations optimization problem.

Proof. Like the corresponding theorem in Binder and Kostina [20], this theorem is a consequence of Theorem 4 by Fiacco and McCormick [50]. □

Corollary 3.2.1. *Let $(X^*, \nu^*, \lambda^*, \mu^*)$ satisfy the first-order optimality conditions (3.43) and let strict complementarity hold at $(X, \nu) = (X^*, \nu^*)$, i.e.*

$$|\nu_i| < 1 \text{ for all } i \in I_{F_1}(X). \tag{3.46}$$

Assume further that the Hessian H is positive definite for all d in the set

$$\begin{aligned} Z(X) = \{d \neq 0 \,|\, &\nabla F_{1i}(X)d = 0 \text{ for all } i \in I_{F_1}(X), \\ &\nabla F_2(X^*)d = 0, \\ &\nabla F_{3i}(X^*)d \geq 0 \text{ for all } i \in I_{F_3}(X^*)\}. \end{aligned}$$

Then X^* is a strict local minimizer of the least absolute deviations optimization problem.

Huber's M-estimator

Contrary to the l_1 norm estimator, we can use the standard LICQ condition to characterize a regular point for the Huber optimization problem as we have shown in the proof of theorem 3.2.1 that we need no constraint qualifications for the unconstrained Huber minimization problem and still set $\mu_0^* = 1$.

The following theorems give the explicit optimality conditions for the Huber estimator.

Theorem 3.2.4 (Necessary optimality conditions for the Huber estimator)**.** *Let X^* be a regular solution of the Huber estimation problem. Then there exist vectors $\nu^* \in \mathbb{R}^{m_1}$, $\lambda^* \in \mathbb{R}^{m_2}$, and $\mu^* \in \mathbb{R}^{m_3}$ such that the following first-order necessary optimality conditions are satisfied,*

$$\nabla F_1(X^*)^T\nu^* - \nabla F_2(X^*)^T\lambda^* - \nabla F_3(X^*)^T\mu^* = 0, \tag{3.47a}$$

$$F_2(X^*) = 0, \tag{3.47b}$$

$$F_3(X^*) \geq 0, \tag{3.47c}$$

$$\nu_i^* = \begin{cases} F_{1i}(X^*), & |F_{1i}(X^*)| \leq \gamma, \\ sign(F_{1i}(X^*))\gamma, & |F_{1i}(X^*)| > \gamma, \end{cases} \tag{3.47d}$$

$$\mu^* \geq 0, \tag{3.47e}$$

$$\sum_{i=1}^{m_3} \mu_i^* F_{3i}(X^*) = 0. \tag{3.47f}$$

Furthermore, the point X^* and the Lagrange multipliers ν^*, λ^*, μ^* satisfy the second-order necessary conditions

$$d^T H^* d \geq 0, \tag{3.48a}$$

$$\forall i \in I_{F_3}(X^*): \quad \mu_i^* \nabla F_{3i}(X^*)d = 0, \tag{3.48b}$$

with

$$H^* = \sum_{i \in I_{F_3}^{\leq \gamma}(X^*)} \nabla F_{1i}(X^*)^T \nabla F_{1i}(X^*) + \sum_{i=1}^{m_1} \nu_i^* \nabla^2 F_{1i}(X^*) - \sum_{i=1}^{m_2} \lambda_i^* \nabla^2 F_{2i}(X^*) - \sum_{i=1}^{m_3} \mu_i^* \nabla^2 F_{3i}(X^*)$$

for all critical directions d defined by

$$D_H(X^*, \nu^*) = \{d \in \mathbb{R}^n \mid \nabla F_2(X^*)d = 0,\ \nabla F_{3i}(X^*)d \geq 0 \ \text{ for all } i \in I_{F_3}(X^*)\}. \tag{3.49}$$

Proof. The proof of this theorem is again a direct extension of the proof given by Binder and Kostina [20] for the Huber estimation problem with equality constraints only. It applies the Fritz John optimality conditions of theorems 3.1.1, 3.1.2 to the reformulation (3.34) of the objective function (2.17) in (3.40) of the Huber minimization problem. By case distinctions, the additional variables w, u, v can be eliminated such that the given result only depends on the original optimization variables X. □

Theorem 3.2.5 (Sufficient optimality conditions for the Huber estimator)**.** *Let the point $(X^*, \nu^*, \lambda^*, \mu^*)$ satisfy the first-order optimality conditions (3.47) and let the matrix H^* be positive definite for all non-zero critical directions $d \in D_H(X^*, \nu^*)$, i.e.*

$$d^T H d > 0 \quad \text{for all } d \in D_H(X^*, \nu^*),\ d \neq 0.$$

Then X^* is a strict local minimizer of the Huber optimization problem.

Proof. Like for the LAD estimator, this theorem is a consequence of Theorem 4 by Fiacco and McCormick [50]. □

Common structure

Comparing the results for the two robust estimators, we see that the optimality conditions of the l_1 optimization problem and the Huber optimization problem also share a common structure. So apparently, we can state the optimality conditions for the robust optimization problem (3.40) in the following form.

Theorem 3.2.6 (General robust necessary optimality conditions of first order)**.** *Assume that X^* is a regular point of (3.40) and the constraint qualification (3.42) corresponding to the* LAD *estimator or the LICQ for Huber's M-estimator holds, respectively. Then there exist vectors $\nu_\rho^* \in \mathbb{R}^{m_1}$ $(\rho = 1, H)$, $\lambda^* \in \mathbb{R}^{m_2}$, $\mu^* \in \mathbb{R}^{m_3}$ such*

that

$$\nabla F_1(X^*)^T \nu_\rho^* - \nabla F_2(X^*)^T \lambda^* - \nabla F_3(X^*)^T \mu^* = 0, \tag{3.50a}$$
$$F_2(X^*) = 0, \tag{3.50b}$$
$$F_3(X^*) \geq 0, \tag{3.50c}$$
$$\mu^* \geq 0, \tag{3.50d}$$
$$\sum_{i=1}^{m_3} \mu_i^* F_{3i}(X^*) = 0, \tag{3.50e}$$

where the multiplier ν_ρ^* has to satisfy the further conditions, depending on the form of the objective function,

$$|\nu_{1_i}^*| \leq 1 \quad \textit{with } \nu_{1_i}^* = \textit{sign}(F_{1i}(X^*)) \textit{ if } F_{1i}(X^*) \neq 0, \tag{3.51a}$$
$$\nu_{H_i}^* = \begin{cases} F_{1i}(X^*), & \textit{if } |F_{1i}(X^*)| \leq \gamma, \\ \textit{sign}(F_{1i}(X^*)) \cdot \gamma, & \textit{if } |F_{1i}(X^*)| > \gamma. \end{cases} \tag{3.51b}$$

The critical directions that we need for the second-order optimality conditions are all directions d from the following set,

$$\begin{aligned} D_\rho(X) = \{ d \in \mathbb{R}^n \; & | \bar{\nu}_{\rho_i} \nabla F_{1i}(X) d = 0 \; (i \in I_{F_1}(X) \text{ with } |\nu_{\rho_i}| < 1), \\ & \bar{\nu}_{\rho_i} \nabla F_{1i}(X) d \geq 0 \; (i \in I_{F_1}(X) \text{ with } |\nu_{\rho_i}| = 1), \\ & \nabla F_2(X) d = 0, \\ & \nabla F_{3i}(X) d \geq 0 \; (i \in I_{F_3}(X))\}, \quad (\rho = 1, H), \end{aligned} \tag{3.52}$$

where for $i \in I_{F_1}(X)$ we defined

$$\bar{\nu}_{1_i} := \begin{cases} \nu_{1_i}, & \text{if } |\nu_{1_i}| = 1, \\ 1, & \text{otherwise}, \end{cases} \tag{3.53a}$$
$$\bar{\nu}_{H_i} := 0. \tag{3.53b}$$

Theorem 3.2.7 (General robust necessary optimality conditions of second order)**.** *Suppose that X^* is a local minimizer of (3.40) and that the LICQ holds for Huber and (3.42) for* LAD*. Then, corresponding to every critical direction $d \in D_\rho(X^*)$, $\rho = 1, H$, there exist multipliers $\nu^* \in \mathbb{R}^{m_1}$, $\lambda^* \in \mathbb{R}^{m_2}$, $\mu^* \in \mathbb{R}^{m_3}$, satisfying*

$$\nabla F_1(X^*)^T \nu^* - \nabla F_2(X^*)^T \lambda^* - \nabla F_3(X^*)^T \mu^* = 0, \tag{3.54a}$$
$$\forall i \in I_{F_3}(X^*): \quad \mu_i^* \nabla F_{3i}(X^*) d = 0, \tag{3.54b}$$
$$d^T H_\rho^* d \geq 0, \tag{3.54c}$$

with

$$H_1^* = \sum_{i=1}^{m_1} \nu_i^* \nabla^2 F_{1i}(X^*) - \sum_{i=1}^{m_2} \lambda_i^* \nabla^2 F_{2i}(X^*) - \sum_{i=1}^{m_3} \mu_i^* \nabla^2 F_{3i}(X^*),$$

$$H_H^* = \sum_{i \in I_{F_1}^{\leq \gamma}(X^*)} \nabla F_{1i}(X^*)^T \nabla F_{1i}(X^*) +$$

$$\sum_{i=1}^{m_1} \nu_i^* \nabla^2 F_{1i}(X^*) - \sum_{i=1}^{m_2} \lambda_i^* \nabla^2 F_{2i}(X^*) - \sum_{i=1}^{m_3} \mu_i^* \nabla^2 F_{3i}(X^*)$$

Theorem 3.2.8 (General robust sufficient optimality conditions of second order)**.** *Suppose that corresponding to every nonzero critical direction $d \in D_\rho(x^*)$, $\rho = 1, H$, there exist multipliers $\nu^* \in \mathbb{R}^{m_1}$, $\lambda^* \in \mathbb{R}^{m_2}$, $\mu^* \in \mathbb{R}^{m_3}$, satisfying*

$$\nabla F_1(X^*)^T \nu^* - \nabla F_2(X^*)^T \lambda^* - \nabla F_3(X^*)^T \mu^* = 0, \tag{3.55a}$$

$$\forall i \in I_{F_3}(X^*): \quad \mu_i^* \nabla F_{3i}(X^*) d = 0, \tag{3.55b}$$

$$d^T H_\rho^* d > 0. \tag{3.55c}$$

Then X^* is an isolated local minimizer of (3.40).

3.2.2. Application of the GN method and the generalized inverse

We apply a constrained Gauss-Newton method to these robust minimization problems in order to find the optimal values of the unknowns. Starting from a given initial guess $X^0 \in \mathbb{R}^n$, we proceed iteratively by computing $X^{k+1} = X^k + [t^k]\Delta X^k$ where ΔX^k solves the linearized problem

$$\begin{array}{ll} \min\limits_{\Delta X} & \rho(F_1(X^k) + J_1(X^k)\Delta X) \\ \text{s.t.} & F_2(X^k) + J_2(X^k)\Delta X = 0, \\ & F_3(X^k) + J_3(X^k)\Delta X \geq 0. \end{array} \tag{3.56}$$

The matrices $J_i(X^k)$, $i = 1, 2, 3$, denote the Jacobians of the problem functions, $J_i(X^k) = \nabla F_i(X^k)$, $i = 1, 2, 3$, and the scalar t^k is an optional stepsize for globalization of the method. We will treat possible stepsize selection strategies in section 3.3.

We already stated in section 3.1.4 above that in the neighborhood of a solution the equality and inequality constrained optimization problem is equivalent to a nonlinear optimzation problem with equality constraints only,

$$\begin{array}{ll} \min\limits_{\Delta X} & \rho(F_1(X^k) + J_1(X^k)\Delta X) \\ \text{s.t.} & \tilde{F}_2(X^k) + \tilde{J}_2(X^k)\Delta X = 0, \end{array} \tag{3.57}$$

where $\tilde{F}_2 = \begin{pmatrix} F_2 \\ (F_{3i})_{i \in I_{F_3}(\Delta X)} \end{pmatrix}$ and $\tilde{J}_2 = \begin{pmatrix} J_2 \\ (J_{3i})_{i \in I_{F_3}(\Delta X)} \end{pmatrix}$. Therefore, we restrict the following elaborations to this simpler case. If the initial guess X^0 is close enough to

the optimal solution X^*, the active set of the linearized problem equals the optimal active set and we can use $I_{F_3}(\Delta X) = I_{F_3}(X^*)$. If X^0 lies outside of this neighborhood, where the active set does not change, an active set strategy has to be applied (see e.g. Gill et al. [63] and Luenberger [94]).

Solution of the l_1 norm optimization problem with the GN method

Application of the necessary and sufficient optimality conditions according to theorems 3.2.6–3.2.8 to the linearized problem (3.57) with the l_1 norm in the objective function shows that ΔX^* is a solution of this problem if there are vectors ν^*, $\tilde{\lambda}^*$ such that

$$\begin{aligned} J_1(X)^T\nu^* - \tilde{J}_2(X)^T\tilde{\lambda}^* &= 0, && \text{(3.58a)}\\ \tilde{F}_2(X) + \tilde{J}_2(X)\Delta X^* &= 0, && \text{(3.58b)}\\ |\nu_i^*| &\le 1 \text{ with} && \text{(3.58c)}\\ \nu_i^* &= \operatorname{sign}\big((F_1(X) + J_1(X)\Delta X^*)_i\big)\\ &\quad \text{if } (F_1(X) + J_1(X)\Delta X^*)_i \neq 0. \end{aligned}$$

So obviously $(X^*, \nu^*, \tilde{\lambda}^*)$ is an FJ point of the l_1 norm optimization problem if and only if $(0, \nu^*, \tilde{\lambda}^*)$ is an FJ point of the corresponding linearized problem (cf. Kostina [86] and Binder and Kostina [20]), i.e. if X^* is a fixed point of the Gauss-Newton iteration.

For ease of notation and in addition to

$$I_{F_1}(X) = \{i = 1, \dots, m_1 \mid F_{1i}(X) = 0\} \text{ with } |I_{F_1}(X)| = m_1^0,$$

we introduce the index sets

$$\begin{aligned} I_{F_1}^+(X) &= \{i = 1, \dots, m_1 \mid F_{1i}(X) > 0\} \text{ with } |I_{F_1}^+(X)| = m_1^+ \text{ and} && \text{(3.59a)}\\ I_{F_1}^-(X) &= \{i = 1, \dots, m_1 \mid F_{1i}(X) < 0\} \text{ with } |I_{F_1}^-(X)| = m_1^-, && \text{(3.59b)} \end{aligned}$$

and

$$\begin{aligned} I_{F_1}(\Delta X) &= \{i = 1, \dots, m_1 \mid (F_1(X) + J_1(X)\Delta X)_i = 0\}, && \text{(3.59c)}\\ I_{F_1}^+(\Delta X) &= \{i = 1, \dots, m_1 \mid (F_1(X) + J_1(X)\Delta X)_i > 0\}, \text{ and} && \text{(3.59d)}\\ I_{F_1}^-(\Delta X) &= \{i = 1, \dots, m_1 \mid (F_1(X) + J_1(X)\Delta X)_i < 0\}. && \text{(3.59e)} \end{aligned}$$

The FJ conditions for the linearized problem at the point X^k are

$$\begin{pmatrix} 0 & \tilde{J}_1(X^k)^T & -\tilde{J}_2(X^k)^T \\ \tilde{J}_1(X^k) & 0 & 0 \\ \tilde{J}_2(X^k) & 0 & 0 \end{pmatrix} \begin{pmatrix} \Delta X^k \\ \tilde{\nu}^k \\ \tilde{\lambda}^k \end{pmatrix} = - \begin{pmatrix} \sum\limits_{i \notin I_{F_1}(\Delta X^k)} \nu_i^k (J_1(X^k))_i^T \\ \tilde{F}_1(X^k) \\ \tilde{F}_2(X^k) \end{pmatrix}, \tag{3.60}$$

where $\tilde{F}_1$ and $\tilde{J}_1$ comprise those components i of the respective functions for which $(F_1(X^k) + J_1(X^k)\Delta X^k)_i = 0$.

So in order to guarantee existence and uniqueness of the increment ΔX^k we require the matrix $\mathcal{J}_{\Delta X}(X^k) = \begin{pmatrix} \tilde{J}_1(X^k) \\ \tilde{J}_2(X^k) \end{pmatrix}$ to have full column rank n. Analogically, a sufficient condition for existence and uniqueness of the multipliers $\tilde{\nu}$ and $\tilde{\lambda}$ is that the matrix $\mathcal{J}_{\tilde{\nu},\tilde{\lambda}}(X^k)^T = (\tilde{J}_1(X^k)^T, \tilde{J}_2(X^k)^T)$ has full column rank together with strict complementarity (3.46). These conditions imply that $\mathcal{J}_{\Delta X}(X^k) = \mathcal{J}_{\tilde{\nu},\tilde{\lambda}}(X^k)$ must be quadratic and invertible for the increment ΔX^k and the multipliers $\tilde{\nu}^k$ and $\tilde{\lambda}^k$ to be unique.

Denote the system of problem functions by $\mathcal{F}(X) = \begin{pmatrix} F_1(X) \\ \tilde{F}_2(X) \end{pmatrix}$ and the corresponding full Jacobian by $\mathcal{J}(X) = \begin{pmatrix} J_1(X) \\ \tilde{J}_2(X) \end{pmatrix}$. Then the increment ΔX can be computed by means of a generalized inverse, $\Delta X = -\mathcal{J}^+(X)\mathcal{F}(X)$, and the generalized inverse $\mathcal{J}^+(X)$, satisfying $\mathcal{J}^+\mathcal{J}\mathcal{J}^+ = \mathcal{J}^+$, is a permutation of the matrix $(0, \mathcal{J}_{\Delta X}(X)^{-1})$ as proven by Kostina [86] and Bock et al. [27].

Like for the general case stated in theorem 3.1.4, a closer analysis of the optimality conditions reveals that a solution X^* of the nonlinear LAD optimization problem corresponds to an optimal active set that here not only identifies the active inequality constraints F_{3i}, $i \in I_{F_3}(X^*)$, but also the zero components of the objective function F_{1i}, $i \in I_{F_1}(X^*)$. In the neighborhood of any solution X^* that is also a regular point according to (3.42), satisfies $\text{rank}\nabla F_A(X^*) = n$ and the strict complementarity condition (3.46), the generalized Gauss-Newton method eventually identifies this optimal active set of the nonlinear l_1 norm problem. So the iterates X^k generated by this method are attracted by a local minimum that satisfies the regularity assumptions and neither the active sets of the linearized problems will change nor the non-zero components of the objective function will change signs when approaching such a minimizer X^*. These sets $I_{F_3}(\Delta X)$, $I_{F_1}(\Delta X)$ are even the same as the corresponding optimal sets $I_{F_3}(X^*)$, $I_{F_1}(X^*)$ at the solution X^*. This means that the full-step method with $t^k \equiv 1$ is essentially the standard Newton method for the nonlinear equation system

$$F_A^*(X) = 0, \qquad F_A^* = \begin{pmatrix} (F_{1i}(X))_{i \in I_{F_1}(X^*)} \\ F_2(X) \\ (F_{3i}(X))_{i \in I_{F_3}(X^*)} \end{pmatrix}.$$

Consequently, it inherits the quadratic rate of convergence of Newton's method, cf. Bock et al. [27] and Kostina [86] and theorem 3.2.10 in section 3.2.3.

Solution of the Huber optimization problem with the GN method

Similarly as for the LAD estimator, application of the necessary and sufficient optimality conditions according to theorems 3.2.6–3.2.8 to the linearized problem (3.57)

with the Huber function in the objective function shows that ΔX^* is a solution of this problem if there are vectors ν^*, $\tilde{\lambda}^*$ such that

$$J_1(X)^T\nu^* - \tilde{J}_2(X)^T\tilde{\lambda}^* = 0, \tag{3.61a}$$

$$\tilde{F}_2(X) + \tilde{J}_2(X)\Delta X^* = 0, \tag{3.61b}$$

$$\nu_i^* = \begin{cases} F_{1i}(X^*), & |F_{1i}(X^*)| \le \gamma, \\ \operatorname{sign}(F_{1i}(X^*)) \cdot \gamma, & |F_{1i}(X^*)| > \gamma, \end{cases}.$$

So obviously $(X^*, \nu^*, \tilde{\lambda}^*)$ is an FJ point of the Huber optimization problem if and only if $(0, \nu^*, \tilde{\lambda}^*)$ is an FJ point of the corresponding linearized problem (cf. Kostina [86] and Binder and Kostina [20]), i.e. if X^* is a fixed point of the Gauss-Newton iteration.

Like before, we introduce some index sets for ease of notation,

$$I_{F_1}^{\le\gamma}(X) = \{i = 1, \dots, m_1 \mid |F_{1i}(X)| \le \gamma\} \text{ with } |I_{F_1}^{\le\gamma}(X)| = m_1^{\le\gamma}, \tag{3.62a}$$

$$I_{F_1}^{\not\le\gamma}(X) = \{i = 1, \dots, m_1 \mid |F_{1i}(X)| > \gamma\} \text{ with } |I_{F_1}^{\not\le\gamma}(X)| = m_1^{\not\le\gamma}, \tag{3.62b}$$

$$I_{F_1}^{\le\gamma}(\Delta X) = \{i = 1, \dots, m_1 \mid |(F_1(X) + J_1(X)\Delta X)_i| \le \gamma\}, \text{ and} \tag{3.62c}$$

$$I_{F_1}^{\not\le\gamma}(\Delta X) = \{i = 1, \dots, m_1 \mid |(F_1(X) + J_1(X)\Delta X)_i| > \gamma\}. \tag{3.62d}$$

The FJ system of the linearized subproblem at X^k corresponding to Huber optimization results to be

$$\begin{pmatrix} J_1(X^k)^T D J_1(X^k) & -\tilde{J}_2(X^k)^T \\ \tilde{J}_2(X^k) & 0 \end{pmatrix} \begin{pmatrix} \Delta X^k \\ \tilde{\lambda}^k \end{pmatrix} = - \begin{pmatrix} J_1(X^k)^T \tilde{F}_1(X^k) \\ \tilde{F}_2(X^k) \end{pmatrix}, \tag{3.63}$$

with the diagonal matrix D,

$$D = \operatorname{diag}(d_{ii}), \quad d_{ii} = \begin{cases} 1, & \text{if } |(F_1(X^k) + J_1(X^k)\Delta X^k)_i| \le \gamma, \\ 0, & \text{if } |(F_1(X^k) + J_1(X^k)\Delta X^k)_i| > \gamma, \end{cases}$$

and the modified objective function

$$\tilde{F}_1(X) = \begin{pmatrix} F_1^{\le\gamma}(X) \\ \tilde{F}_1^{\not\le\gamma}(X) \end{pmatrix}, \quad \text{where} \quad \begin{aligned} F_1^{\le\gamma}(X) &= (F_{1i}(X))_{i \in I_{F_1}^{\le\gamma}(\Delta X)}, \\ \tilde{F}_1^{\not\le\gamma}(X) &= (\nu_i)_{i \in I_{F_1}^{\not\le\gamma}(\Delta X)}. \end{aligned}$$

Note that all components of the multiplier ν^k for this problem can be computed for a given value of X^k and therefore do not appear in the FJ system given here.

We make again the usual regularity assumptions that at any iterate X^k the rank of the matrix of constraints gradients equals the number of constraints,

$$\operatorname{rank}\left(\tilde{J}_2(X^k)\right) = m_2,$$

and that the rank of the composed Jacobian $\mathcal{J}(X^k) = \begin{pmatrix} \tilde{J}_1(X^k) \\ \tilde{J}_2(X^k) \end{pmatrix}$ equals the number of unknowns,

$$\operatorname{rank}\left(\mathcal{J}(X^k)\right) = n,$$

where $\tilde{J}_1(X^k) = DJ_1(X^k)$. In combination with the strict complementarity condition

$$|(F_1(X) + J_1(X)\Delta X)_i| \neq \gamma \ \forall i = 1, \ldots, m_1, \tag{3.64}$$

for the Huber estimation problem, these regularity assumptions guarantee existence and uniqueness of the increment ΔX^k and the multiplier $\tilde{\lambda}^k$. Then the solution of the linearized system can again be computed by means of a generalized inverse, $\Delta X^k = -\mathcal{J}^+(X^k)\mathcal{F}(X^k)$, where we use the notation $\mathcal{F}(X) = \begin{pmatrix} \tilde{F}_1(X) \\ \tilde{F}_2(X) \end{pmatrix}$ with

$$\tilde{F}_{1i}(X) = \begin{cases} F_{1i}(X), & \text{if } |(F_1(X) + J_1(X)\Delta X)_i| \leq \gamma, \\ \gamma \cdot \operatorname{sign}(F_1(X) + J_1(X)\Delta X)_i, & \text{if } |(F_1(X) + J_1(X)\Delta X)_i| > \gamma. \end{cases}$$

Similarly to the identification of the zero components of the objective function for the LAD estimator, it can also be shown that the generalized Gauss-Newton method eventually identifies not only the active inequality constraints but also the optimal partitioning of Huber's function with respect to the sets $I_{F_1}^{\leq\gamma}(X)$ and $I_{F_1}^{\not\leq\gamma}(X)$.

Theorem 3.2.9. *Assume that X^* with corresponding multipliers $(\nu^*, \lambda^*, \mu^*)$ is a solution of the nonlinear Huber minimization problem that satisfies the LICQ, the conditions*

$$\textit{rank}\left(\nabla F_A(X^*)\right) = n, \qquad \nabla F_A(X) = \begin{pmatrix} D(X)\nabla F_1(X) \\ F_2(X) \\ F_3(X) \end{pmatrix},$$

with the diagonal matrix $D(X) = \textit{diag}(d_{ii})$, $d_{ii} = \begin{cases} 1, & \textit{if } i \in I_{F_3}^{\leq\gamma}(X^*), \\ 0, & \textit{if } i \in I_{F_3}^{\not\leq\gamma}(X^*) \end{cases}$, and strict complementarity $|F_{1i}(X^*)| \neq \gamma$, $i = 1, \ldots, m_1$. Then there exists a neighborhood $\mathcal{N}$ of $(X^*, \nu^*, \lambda^*, \mu^*)$ such that for all $(X^k, \nu^k, \lambda^k, \mu^k) \in \mathcal{N}$ the linearized Huber problem has a unique solution ΔX^k with corresponding multipliers and the partitioning of the objective function $I_{F_1}^{\leq\gamma}(\Delta X^k)$, $I_{F_1}^{\not\leq\gamma}(\Delta X^k)$ for this solution is the same as the partitioning $I_{F_1}^{\leq\gamma}(X^k)$, $I_{F_1}^{\not\leq\gamma}(X^k)$ for the nonlinear problem at $X = X^k$, which is in turn the same as the partitioning $I_{F_1}^{\leq\gamma}(X^*)$, $I_{F_1}^{\not\leq\gamma}(X^*)$ of the nonlinear problem at the minimizer $X = X^*$.

So the iterates X^k generated by the Gauss-Newton method for the Huber minimization problem are attracted by a local minimum that satisfies the regularity assumptions and neither the active sets of the linearized problems nor the partitioning of the objective function will change when approaching such a minimizer X^*.

These sets $I_{F_3}(\Delta X)$, $I_{F_1}^{\leqq\gamma}(\Delta X)$, $I_{F_1}^{\not\leqq\gamma}(\Delta X)$ are even the same as the corresponding optimal sets $I_{F_3}(X^*)$, $I_{F_1}^{\leqq\gamma}(X^*)$, $I_{F_1}^{\not\leqq\gamma}(X^*)$ at the solution X^*. This means that the full-step method with $t^k \equiv 1$ is essentially the Gauss-Newton method for a modified least squares problem,

$$\begin{aligned} \min_X \quad & \tfrac{1}{2} \sum_{i \in I_{F_1}^{\leqq\gamma}(X^*)} F_{1i}^2(X) + \sum_{i \in I_{F_1}^{\not\leqq\gamma}(X^*)} \gamma \cdot \operatorname{sign}(F_{1i}(X)) F_{1i}(X), \\ \text{s.t.} \quad & F_2(X) = 0, \\ & F_{3i}(X) = 0, \quad i \in I_{F_3}(X^*), \\ & F_{3i}(X) > 0, \quad i \notin I_{F_3}(X^*). \end{aligned}$$

Consequently, it has a linear rate of convergence as shown by theorem 3.2.10 in the next section.

3.2.3. Local convergence properties

Both the Gauss-Newton method for least absolute deviations optimization and for Huber optimization converge locally with at least linear rate of convergence under the regularity assumptions given in the respective sections above. We summarize this property for both optimization problems in the following theorem.

Theorem 3.2.10 (Local contraction theorem, e.g. Bock [24]). *Let the regularity assumptions for* LAD *optimization or Huber optimization, respectively, given in the corresponding sections above, hold. Assume further that the following two Lipschitz conditions for $\mathcal{J}$ and $\mathcal{J}^+$ hold for all X, Y and all $t \in [0,1]$,*

$$(\mathcal{J}) \quad \frac{\|\mathcal{J}^+(Y)(\mathcal{J}(X+t(Y-X))-\mathcal{J}(X))(Y-X)\|}{t\|Y-X\|^2} \leq \omega(X) \leq \omega < \infty,$$

$$(\mathcal{J}^+) \quad \frac{\|\big(\mathcal{J}^+(Y)-\mathcal{J}^+(X)\big)R(X)\|}{\|Y-X\|} \leq \kappa(X) \leq \kappa < 1,$$

where $R(X)$ denotes the residual values of the problem functions, i.e. in case of l_1 norm optimization, we have $R(X) = 0$ (because we only consider equality constraints and active components of the objective function which naturally have to evaluate to zero), and in case of Huber optimization, we get $R(X) = \begin{pmatrix} \tilde{R}(X) \\ \tilde{F}_2(X) + \tilde{J}_2(X)\Delta X \end{pmatrix}$ with

$$\tilde{R}_i(X) = \begin{cases} (F_1(X) + J_1(X)\Delta X)_i, & \text{if } |(F_1(X) + J_1(X)\Delta X)_i| \leq \gamma, \\ \gamma \cdot sign(F_1(X) + J_1(X)\Delta X)_i, & \text{if } |(F_1(X) + J_1(X)\Delta X)_i| > \gamma. \end{cases}$$

Then it follows that the sequence of iterates (X^k) converges to a stationary point X^* with $\mathcal{J}^+(X^*)\mathcal{F}(X^*) = 0$ and the increments satisfy the following inequality,

$$\|\Delta X^{k+1}\| \leq \left(\frac{\omega}{2}\|\Delta X^k\| + \kappa\right) \|\Delta X^k\|.$$

So the convergence of the Gauss-Newton method is at least linear with rate κ for Huber's M-estimator and quadratic for the l_1 norm estimator, where from $R(X) = 0$ follows $\kappa = 0$.

For a proof of this theorem see Binder and Kostina [20].

The quantity ω is a measure for the nonlinearity of the problem, excluding those functions corresponding to "large" measurement errors, as the Lipschitz condition for $\mathcal{J}$ given in the theorem is nothing but a weighted second derivative. Its inverse ω^{-1} characterizes the region of validity of the linearized model.

The quantity κ, on the other hand, is an incompatibility constant that describes the compatibility of the model functions and the measurement data for parameter estimation, where in case of Huber estimation "large" measurement errors are replaced by $\pm\gamma$. The condition $\kappa < 1$ is necessary for the unknowns to be identifiable from the data. A stationary point with $\kappa < 1$ is statistically stable. An advantage of the constrained Gauss-Newton method we use here is that it does not converge to large residual solutions that are statistically unstable. For Huber's M-estimator, it is interesting to note that the incompatibility constant κ can be reduced by reducing the tuning constant γ as

$$\|\big(\mathcal{J}^+(X^{k+1}) - \mathcal{J}^+(X^k)\big)R(X^k)\| \leq \|\mathcal{J}^+(X^{k+1}) - \mathcal{J}^+(X^k)\|\gamma m_1.$$

3.2.4. Solution of the linearized problems

We already mentioned in section 3.1.5 that the efficient solution of the linearized problems at each iteration is decisive for the performance of the generalized Gauss-Newton method. So the special structure of the linearized problems (3.57) that arises due to the multiple shooting discretization of the dynamic system should be exploited. This can be done by applying a condensing strategy as described in section 3.1.5 to reduce the dimension of the linearized problem.

In case of equality constrained robust optimization problems with the l_1 norm or Huber's function as objective function, the resulting condensed problems have the general form

$$\begin{aligned} \min_{z} \quad & \rho(A_1 z + c_1) \\ \text{s.t.} \quad & A_2 z + c_2 = 0, \end{aligned} \tag{3.65}$$

with $z \in \mathbb{R}^{r_z}$, matrices $A_i \in \mathbb{R}^{r_i \times r_z}$, and vectors $c_i \in \mathbb{R}^{r_i}$, $i = 1, 2$. The optimization criterion $\rho(\cdot)$ is in each case the same as in the original robust problem, i.e. either the l_1 norm or Huber's function. For LAD optimization, this problem is equivalent to a linear programming problem, while for Huber optimization, it is equivalent to a quadratic programming problem.

Consider the dual problem of (3.65). It is given by

$$\begin{aligned} \min_{y_1 \in \mathbb{R}^{r_1}, y_2 \in \mathbb{R}^{r_2}} \quad & \tfrac{\gamma}{2} y_1^T y_1 - c_1^T y_1 - c_2^T y_2 \\ \text{s.t.} \quad & A_1^T y_1 + A_2^T y_2 = 0, \\ & |y_{1i}| \leq 1, \; i = 1, \ldots, r_1. \end{aligned}$$

The dual of the l_1 norm minimization problem simply corresponds to the case $\gamma = 0$. This dual problem can now be solved efficiently using a primal-dual active-set method

with so-called "long steps" as in Kostina [85] and Kostina and Kostyukova [87] or "flipping bounds" as in Koberstein [84].

We want to note that, for the Huber estimator, problem (3.65) can be reformulated as

$$\begin{aligned} \min_{w\in\mathbb{R}^{r_1}, z\in\mathbb{R}^{r_z}} \quad & \tfrac{1}{2}\|w\|_2^2 + \gamma\|A_1 z + c_1 - w\|_1 \\ \text{s.t.} \quad & A_2 z + c_2 = 0. \end{aligned}$$

This problem is closely related to problems arising in sparse signal or image reconstruction or LASSO problems, see Osborne et al. [109]. Thus it can also be solved by methods that have been developed for this problem class, see the overview article by Bruckstein et al. [31] and the references given there.

3.3. Globalization strategies for the Gauss-Newton method

In this section we finally give a very brief summary of different methods for computing the damping parameter t^k in the Gauss-Newton iterations $X^{k+1} = X^k + t^k \Delta X^k$. The considered methods are line search methods and monotonicity tests.

3.3.1. Line search methods

One possibility to compute the relaxation parameter t^k is a line search procedure. This approach chooses the stepsize t^k such that the new iterate X^{k+1} is in some sense "better" than the current iterate X^k by means of a merit function, i.e. $T(X^{k+1}) \leq T(X^k)$. A common choice is the usage of the exact penalty function

$$T_1(X) = \rho(F_1(X)) + \sum_{i=1}^{m_2} \alpha_i |F_{2i}(X)| + \sum_{i=1}^{m_3} \beta_i \max\{0, -F_{3i}(X)\} \tag{3.66}$$

as merit function, where the weights α_i and β_i have to be chosen sufficiently large.

Under the regularity assumptions given in the respective sections above on local convergence and strict complementarity for both the original and the linearized problem, it can be shown that the increment ΔX solving the linearized problem leads to a descent direction of the original l_1 norm or Huber optimization problem, i.e. the compatibility condition

$$\lim_{\varepsilon \to 0} \frac{T_1(X^k + \varepsilon \Delta X^k) - T_1(X^k)}{\varepsilon} < 0, \tag{3.67}$$

is satisfied for the exact penalty function T_1. Thus the method is globally convergent with this stepsize strategy.

A drawback of this approach is that already for only mildly ill-conditioned problems it can lead to very small stepsizes even in the region of full-step convergence and may thus be very inefficient. The next section gives a possibility to overcome this problem.

3.3.2. The restrictive monotonicity test

A remedy in case of least squares estimation is provided by stepsize selection based on natural level functions as first introduced by Deuflhard [42]. Bock [24] and Bock et al. [26] introduced the so-called "restrictive monotonicity test". The Gauss-Newton method is interpreted as Newton's method applied to the nonlinear system of equality constraints and active or γ-active components of the objective function, respectively. Then a line search based on the natural level function

$$T_k(X^k + t\Delta X^k) = \|\mathcal{J}^+(X^k)\mathcal{F}(X^k + t\Delta X^k)\|_2^2 \tag{3.68}$$

is conducted by means of the restrictive monotonicity test: we choose the maximal stepsize $t \leq 1$ such that

$$t\|\Delta X^k\| \leq \frac{\eta}{\omega(t)}$$

for a given $\eta < 2$, where $\omega(t)$ is an asymptotically correct estimate of the "curvature" of the problem, i.e. a measure for its nonlinearity,

$$\omega(t) = \frac{2\|\mathcal{J}^+(X^k)\left(\mathcal{F}(X^k + t\Delta X^k) - (1-t)\mathcal{F}(X^k)\right)\|}{t^2\|\Delta X^k\|^2}. \tag{3.69}$$

4. The Gauss-Newton method with inexact derivatives

For efficient solution of the linearized systems (3.21) in the Gauss-Newton method, another major effort is the computation of the Jacobians, both of the objective function and of the constraints, that are needed at each iteration. The computation of derivative information is numerically challenging and rather expensive in terms of computational time and memory. Thus it is worthwhile developing strategies that overcome this problem. One possibility is the classic use of matrix updates in Quasi-Newton methods (see, e.g., Nocedal and Wright [101]). But also other methods for approximating the Jacobians are feasible. We will show in this chapter that the theory on Gauss-Newton methods with inexact derivatives that exists for general nonlinear optimization problems has rather straightforward specializations for robust optimization problems.

In the first section we thus shortly review the Gauss-Newton method with inexact derivatives for the general case. The following two sections are then devoted to the extension of this approach to the LAD and Huber optimization problems. The corresponding algorithms will be derived and stability of the respective active sets and local convergence will be proved.

4.1. Inexact Gauss-Newton method for nonlinear optimization

We start out from the general nonlinear optimization problem

$$\begin{array}{ll} \min\limits_{X} & F(X) \\ \text{s.t.} & G(X) = 0, \\ & H(X) \geq 0, \end{array} \tag{4.1}$$

as we did in the previous chapter and utilize an SQP method for its solution. This means the problem is iteratively solved by computing the successive iterates $X^{k+1} = X^k + \Delta X^k$ where ΔX^k solves the quadratic problem

$$\begin{array}{lll} \min\limits_{\Delta X} & \frac{1}{2}\Delta X^T \mathcal{M}^k \Delta X + \nabla F(X^k)\Delta X & \\ \text{s.t.} & G_i(X^k) + \nabla G_i(X^k)\Delta X = 0 & (i = 1, \ldots, m_2), \\ & H_j(X^k) + \nabla H_j(X^k)\Delta X \geq 0 & (j = 1, \ldots, m_3), \end{array} \tag{4.2}$$

where the matrix $\mathcal{M}^k$ in the quadratic term is the Hessian of the Lagrange function, $\mathcal{M}^k = \nabla_X^2 \mathcal{L}(X^k, \lambda^k, \mu^k)$, or an approximation of it. The Lagrange function is

$$\mathcal{L}(X, \lambda, \mu) = F(X) - \lambda^T G(X) - \mu^T H(X). \tag{4.3}$$

We assume that the problems satisfy some constraint qualification that guarantees existence and uniqueness of the multipliers λ, μ and that allows the choice of $\mu_0 = 1$ in the Fritz John optimality conditions, cp. section 3.1.1.

According to Wirsching [140] we can replace the quadratic problem (4.2) by

$$\begin{array}{ll} \min\limits_X & \frac{1}{2}(X - X^k)^T L^k (X - X^k) + \nabla F(X^k)(X - X^k) \\ & +\lambda_{QP}^T (B^k - \nabla G(X^k))(X - X^k) \\ & +\mu_{QP}^T (C^k - \nabla H(X^k))(X - X^k), \\ \text{s.t.} & B^k (X - X^k) + G(X^k) = 0, \\ & C^k (X - X^k) + H(X^k) \geq 0, \end{array} \tag{4.4}$$

which uses the inexact derivatives $L^k \approx \nabla_X^2 \mathcal{L}(X^k, \lambda^k, \mu^k)$, $B^k \approx \nabla G(X^k)$ and $C^k \approx \nabla H(X^k)$ but is still equivalent in some sense to the QP (4.2) due to the correction term in the objective function.

Inserting the specific functions which define the l_1 norm and Huber optimization problem, respectively, in the problem formulation (4.1), we derive in the following the corresponding SQP problems for using inexact derivatives for these problems.

So we now turn our attention to the robust optimization problem

$$\begin{array}{ll} \min\limits_X & \rho(F_1(X)) \\ \text{s.t.} & F_2(X) = 0, \\ & F_3(X) \geq 0, \end{array} \tag{4.5}$$

where the optimization criterion may again be the l_1 norm (2.15) or Huber's function (2.19).

We will need the following lemma which is a variant as used by Wirsching [140] of standard results that are similarly given, e.g., by Bock [24].

Lemma 4.1.1 (Local convergence for inexact Newton methods)**.** *Assume that $\phi : D \to \mathbb{R}^n$, $D \subset \mathbb{R}^n$ open, is a continuously differentiable function and consider the sequence $X^{k+1} = X^k + \Delta X^k$ with $\Delta X^k = -(\Psi^k)^{-1}\phi(X^k)$ with an invertible matrix Ψ^k starting with some $X^0 \in D$. Let the following assumptions be true:*

i) The sequence of invertible matrices (Ψ^k) is uniformly bounded and has uniformly bounded inverses.

ii) There exists a $\kappa < 1$ such that for all $k \in \mathbb{N}$ and all $t \in [0, 1]$ it can be guaranteed that

$$\left\| (\Psi^{k+1})^{-1} \big(\Psi^k - \nabla\phi(X^k + t\Delta X^k) \big) \Delta X^k \right\| \leq \kappa \|\Delta X^k\|.$$

iii) The set

$$\mathcal{B}\left(X^0, \frac{\|\Delta X^0\|}{1-\kappa}\right) = \left\{X \in \mathbb{R}^n \mid \|X - X^0\| \leq \frac{\|\Delta X^0\|}{1-\kappa}\right\}$$

is completely contained in D.

Then the sequence $(X^k)_{k\in\mathbb{N}}$ remains in D and converges towards a limit point $X^* \in \mathcal{B}(X^0, \frac{\|\Delta X^0\|}{1-\kappa})$ in which $\Delta X^* = -(\Psi^*)^{-1}\phi(X^*) = 0$. If

$$\lim_{k\to\infty} \frac{\left\|(\Psi^{k+1})^{-1}\left(\Psi^k - \nabla\phi(X^*)\right)\Delta X^k\right\|}{\|\Delta X^k\|} = 0,$$

then the convergence rate is q-superlinear.

4.2. Inexact Gauss-Newton for l_1 norm minimization

This section contains the derivation of the inexact Gauss-Newton algorithm for the nonlinear l_1 norm minimization problem

$$\begin{array}{ll} \min\limits_{X} & \|F_1(X)\|_1 = \sum\limits_{i=1}^{m_1} |F_{1i}(X)| \\ \text{s.t.} & F_2(X) = 0, \\ & F_3(X) \geq 0, \end{array} \tag{4.6}$$

where the problem functions $F_1 : \mathbb{R}^n \to \mathbb{R}^{m_1}$, $F_2 : \mathbb{R}^n \to \mathbb{R}^{m_2}$, $F_3 : \mathbb{R}^n \to \mathbb{R}^{m_3}$ are assumed to be at least twice continuously differentiable.

We start our elaborations from the smooth reformulation of problem (4.6) that is given by

$$\begin{array}{lll} \min\limits_{X\in\mathbb{R}^n, u,v\in\mathbb{R}^{m_1}} & \sum\limits_{i=1}^{m_1}(u_i + v_i) & \\ \text{s.t.} & F_{1i}(X) = u_i - v_i, & i = 1,\ldots,m_1, \\ & F_2(X) = 0, & \\ & F_3(X) \geq 0, & \\ & u_i, v_i \geq 0, & i = 1,\ldots,m_1. \end{array} \tag{4.7}$$

Obviously, this is a special case of the general nonlinear optimization problem (4.1) with the defining functions being specified by

$$F(X,u,v) = \sum_{i=1}^{m_1}(u_i + v_i), \tag{4.8a}$$

$$G(X,u,v) = \begin{pmatrix} F_1(X) - u + v \\ F_2(X) \end{pmatrix}, \tag{4.8b}$$

$$H(X,u,v) = \begin{pmatrix} F_3(X) \\ u \\ v \end{pmatrix}, \tag{4.8c}$$

and we can apply the corresponding theory here. The derivatives that are needed in the inexact SQP problem (4.4) can be computed as

$$\nabla F(X,u,v) = \begin{pmatrix} 0^T & 1^T & 1^T \end{pmatrix} \in \mathbb{R}^{n+2m_1}, \tag{4.9a}$$

$$\nabla G(X,u,v) = \begin{pmatrix} \nabla F_1(X) & -\mathbb{I} & \mathbb{I} \\ \nabla F_2(X) & 0 & 0 \end{pmatrix} \in \mathbb{R}^{(m_1+m_2)\times(n+2m_1)}, \tag{4.9b}$$

$$\nabla H(X,u,v) = \begin{pmatrix} \nabla F_3(X) & 0 & 0 \\ 0 & \mathbb{I} & 0 \\ 0 & 0 & \mathbb{I} \end{pmatrix} \in \mathbb{R}^{(m_3+2m_1)\times(n+2m_1)}, \tag{4.9c}$$

where 1 denotes the vector of all ones with suitable length and $\mathbb{I}$ is the unity matrix of suitable dimensions.

For ease of notation, we will again use $J_i(X^k) = \nabla F_i(X^k),\ i = 1,2,3$, to denote the Jacobians of the problem functions at the iterates and we introduce the new matrices $A_i^k \approx \nabla F_i(X^k), i = 1,2,3$, for approximations of these.

4.2.1. Derivation of the algorithm

We can now insert these problem functions into the inexact QP (4.4) and get

$$\begin{aligned} \min_{(\Delta X,\Delta u,\Delta v)} \quad & \frac{1}{2}\begin{pmatrix} \Delta X \\ \Delta u \\ \Delta v \end{pmatrix}^T L^k \begin{pmatrix} \Delta X \\ \Delta u \\ \Delta v \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}^T \begin{pmatrix} \Delta X \\ \Delta u \\ \Delta v \end{pmatrix} \\ & + \lambda_{QP}^T \begin{pmatrix} A_1^k - J_1(X^k) & 0 & 0 \\ A_2^k - J_2(X^k) & 0 & 0 \end{pmatrix} \begin{pmatrix} \Delta X \\ \Delta u \\ \Delta v \end{pmatrix} \\ & + \mu_{QP}^T \begin{pmatrix} A_3^k - J_3(X^k) & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \Delta X \\ \Delta u \\ \Delta v \end{pmatrix}, \\ \text{s.t.} \quad & \begin{pmatrix} F_1(X^k) - u^k + v^k \\ F_2(X^k) \end{pmatrix} + \begin{pmatrix} A_1^k & -\mathbb{I} & \mathbb{I} \\ A_2^k & 0 & 0 \end{pmatrix} \begin{pmatrix} \Delta X \\ \Delta u \\ \Delta v \end{pmatrix} = 0, \\ & \begin{pmatrix} F_3(X^k) \\ u^k \\ v^k \end{pmatrix} + \begin{pmatrix} A_3^k & 0 & 0 \\ 0 & \mathbb{I} & 0 \\ 0 & 0 & \mathbb{I} \end{pmatrix} \begin{pmatrix} \Delta X \\ \Delta u \\ \Delta v \end{pmatrix} \geq 0. \end{aligned} \tag{4.10}$$

As we will apply the Gauss-Newton method which uses linearizations of the problem functions and thus effectively ignores terms of second and higher order, we can immediately neglect the quadratic term in the problem's objective function. The multipliers are split into $\lambda_{QP}^T = (\lambda_{QP}^{(1)^T}, \lambda_{QP}^{(2)^T})$ and $\mu_{QP}^T = (\mu_{QP}^{(1)^T}, \mu_{QP}^{(2)^T}, \mu_{QP}^{(3)^T})$ corresponding to the distinct block rows of the constraints. The problem then figures out

to be

$$\begin{array}{ll} \min\limits_{(\Delta X,\Delta u,\Delta v)} & \sum\limits_{i=1}^{m_1}(\Delta u_i+\Delta v_i)+\lambda_{QP}^{(1)^T}(A_1^k-J_1(X^k))\Delta X \\ & +\lambda_{QP}^{(2)^T}(A_2^k-J_2(X^k))\Delta X+\mu_{QP}^{(1)^T}(A_3^k-J_3(X^k))\Delta X \\ \text{s.t} & F_1(X^k)+A_1^k\Delta X-(u^k+\Delta u)+(v^k+\Delta v)=0, \\ & F_2(X^k)+A_2^k\Delta X=0, \\ & F_3(X^k)+A_3^k\Delta X\geq 0, \\ & u^k+\Delta u\geq 0, \\ & v^k+\Delta v\geq 0. \end{array} \tag{4.11}$$

Implicitly adding the constant term $\sum_{i=1}^{m_1}(u_i^k+v_i^k)$ to the objective function, we recognize in the subproblem

$$\begin{array}{ll} \min\limits_{\Delta u,\Delta v} & \sum\limits_{i=1}^{m_1}\big((u_i^k+\Delta u_i)+(v_i^k+\Delta v_i)\big) \\ \text{s.t.} & F_1(X^k)+A_1^k\Delta X-(u^k+\Delta u)+(v^k+\Delta v)=0, \\ & u^k+\Delta u\geq 0, \\ & v^k+\Delta v\geq 0. \end{array}$$

a variant of the reformulation (4.7) of the l_1 norm minimization problem and can thus replace it by

$$\min_{\Delta X}\|F_1(X^k)+A_1^k\Delta X\|_1.$$

Note that this problem contains all terms depending on $\Delta u,\ \Delta v$ such that these auxiliary variables are effectively removed again by this transformation.

This leads to an equivalent formulation of (4.11) that reads

$$\begin{array}{ll} \min\limits_{\Delta X\in\mathbb{R}^n} & \|F_1(X^k)+A_1^k\Delta X\|_1 \\ & -\Big(\nu_{QP}^T(A_1^k-J_1(X^k))-\lambda_{QP}^T(A_2^k-J_2(X^k))-\mu_{QP}^T(A_3^k-J_3(X^k))\Big)\Delta X, \\ \text{s.t.} & F_2(X^k)+A_2^k\Delta X=0, \\ & F_3(X^k)+A_3^k\Delta X\geq 0, \end{array}$$

where we made the redefinitions $\nu_{QP}=-\lambda_{QP}^{(1)}$, $\lambda_{QP}=\lambda_{QP}^{(2)}$, and $\mu_{QP}=\mu_{QP}^{(1)}$. It can be easily shown that, like for the exact LAD optimization problem, the multiplier ν_{QP} for the objective function has to satisfy the optimality condition

$$\begin{aligned} &|(\nu_{QP})_i|\leq 1 \text{ with} \\ &(\nu_{QP})_i=\operatorname{sign}\big((F_1(X^k)+A_1^k\Delta X)_i\big) \text{ if } (F_1(X^k)+A_1^k\Delta X)_i\neq 0. \end{aligned}$$

In practical application, we use the current multiplier values $\nu^k,\ \lambda^k,\ \mu^k$ as approx-

imations for the unknown multipliers ν_{QP}, λ_{QP}, μ_{QP} and solve the problem

$$\begin{aligned} \min_{\Delta X \in \mathbb{R}^n} \quad & \|F_1(X^k) + A_1^k \Delta X\|_1 \\ & - \left(\nu^{kT}(A_1^k - J_1(X^k)) - \lambda^{kT}(A_2^k - J_2(X^k)) - \mu^{kT}(A_3^k - J_3(X^k))\right) \Delta X, \\ \text{s.t.} \quad & F_2(X^k) + A_2^k \Delta X = 0, \\ & F_3(X^k) + A_3^k \Delta X \geq 0. \end{aligned} \tag{4.12}$$

As a reminder of the notation, we repeat the definition of the used index sets in the context of l_1 norm minimization,

$$I_{F_1}(X) = \{i = 1, \ldots, m_1 \mid F_{1i}(X) = 0\}, \tag{4.13a}$$
$$I_{F_1}^+(X) = \{i = 1, \ldots, m_1 \mid F_{1i}(X) > 0\}, \tag{4.13b}$$
$$I_{F_1}^-(X) = \{i = 1, \ldots, m_1 \mid F_{1i}(X) < 0\}, \tag{4.13c}$$
$$I_{F_3}(X) = \{i = 1, \ldots, m_3 \mid F_{3i}(X) = 0\}, \tag{4.13d}$$

and

$$I_{F_1}(\Delta X) = \{i = 1, \ldots, m_1 \mid (F_1(X) + A_1 \Delta X)_i = 0\} \text{ with } |I_{F_1}(\Delta X)| = m_1^0, \tag{4.13e}$$
$$I_{F_1}^+(\Delta X) = \{i = 1, \ldots, m_1 \mid (F_1(X) + A_1 \Delta X)_i > 0\} \text{ with } |I_{F_1}^+(\Delta X)| = m_1^+, \tag{4.13f}$$
$$I_{F_1}^-(\Delta X) = \{i = 1, \ldots, m_1 \mid (F_1(X) + A_1 \Delta X)_i < 0\} \text{ with } |I_{F_1}^-(\Delta X)| = m_1^-, \tag{4.13g}$$
$$I_{F_3}(\Delta X) = \{i = 1, \ldots, m_3 \mid (F_3(X) + A_3 \Delta X)_i = 0\} \text{ with } |I_{F_3}(\Delta X)| = m_3^0. \tag{4.13h}$$

We will from here use the notations $\tilde{A}_1 = (A_{1i})_{i \in I_{F_1}(\Delta X)}$ and $\tilde{A}_3 = (A_{3i})_{i \in I_{F_3}(\Delta X)}$, where A_{ji}, $j = 1, 3$, is the ith row of the matrix A_j, $j = 1, 3$, respectively. Further we introduce the matrices

$$A = \begin{pmatrix} A_1 \\ A_2 \\ A_3 \end{pmatrix} \qquad \text{and} \qquad \tilde{A} = \begin{pmatrix} \tilde{A}_1 \\ A_2 \\ \tilde{A}_3 \end{pmatrix}.$$

The constraint qualification

$$\operatorname{rank}(\tilde{A}) = m_1^0 + m_2 + m_3^0 \tag{4.14}$$

ensures uniqueness of the Lagrange multipliers, cf. Kostina [86]. The vectors $\tilde{F}_1(X)$ and $\tilde{F}_3(X)$ will correspondingly contain the components of the objective function and of the inequality constraints corresponding to these sets, respectively,

$$\tilde{F}_1(X) = (F_{1i}(X))_{i \in I_{F_1}(\Delta X)} \qquad \text{and} \qquad \tilde{F}_3(X) = (F_{3i}(X))_{i \in I_{F_3}(\Delta X)}.$$

For ease of notation we further summarize all active functions in the vector

$$\tilde{F}(X) = \begin{pmatrix} \tilde{F}_1(X) \\ F_2(X) \\ \tilde{F}_3(X) \end{pmatrix}.$$

The objective function is complemented with its positive and negative components

$$F_1^+(X) = (F_{1i}(X))_{i \in I_{F_1}^+(\Delta X)} \qquad \text{and} \qquad F_1^-(X) = (F_{1i}(X))_{i \in I_{F_1}^-(\Delta X)}.$$

For the Lagrangian multipliers ν and μ we will use the corresponding notation

$$\tilde{\nu} = (\nu_i)_{i \in I_{F_1}(\Delta X)} \qquad \text{and} \qquad \tilde{\mu} = (\mu_i)_{i \in I_{F_3}(\Delta X)}.$$

As the application of the Gauss-Newton method corresponds to choosing $L^k = 0$ in problem (4.10), lemma 3.1.3 is not applicable here. The following lemma thus ensures existence and uniqueness of the solution ΔX^{AP} and the multipliers $\nu^{AP}, \lambda^{AP}, \mu^{AP}$ of problem (4.12).

Lemma 4.2.1. *Let the matrices of approximated Jacobians A_i, $i = 1, 2, 3$, be chosen such that the composed matrix*

$$\tilde{A}^k = \begin{pmatrix} \tilde{A}_1^k \\ A_2^k \\ \tilde{A}_3^k \end{pmatrix} \in \mathbb{R}^{(m_1^0 + m_2 + m_3^0) \times n}, \tag{4.15}$$

defined according to the active sets of the linearized problem at X^k, has full rank to meet the constraint qualification (4.14). If strict complementarity,

$$|\nu_i| < 1 \text{ for } i \in I_{F_1}(\Delta X^{AP}) \quad \text{and} \quad \mu_i > 0 \text{ for } i \in I_{F_3}(\Delta X^{AP}), \tag{4.16}$$

holds at X^k and the assumption $\operatorname{rank}(\tilde{A}^k) = n$ is satisfied, then the Fritz John point $(\Delta X^{AP}, \nu^{AP}, \lambda^{AP}, \mu^{AP})$ of (4.12) is unique.

Proof. As we have shown in section 3.2.1, under the constraint qualification (4.14) we can always set $\mu_0^{AP} = 1$ for the linearized problem such that the FJ optimality conditions become the standard KKT conditions. Consider thus the necessary optimality conditions of first order for problem (4.12) with approximated derivatives:

$$\begin{aligned}
\nabla_{\Delta X}\mathcal{L}^{AP}(\Delta X^{AP},\nu^{AP},\lambda^{AP},\mu^{AP}) &= A_1^{kT}\nu^{AP} - \Big((A_1^k - J_1(X^k))^T\nu^k \\
&\quad - (A_2^k - J_2(X^k))^T\lambda^k \\
&\quad -(A_3^k - J_3(X^k))^T\mu^k\Big) \\
&\quad - A_2^{kT}\lambda^{AP} - A_3^{kT}\mu^{AP} \\
&= 0, && \text{(4.17a)}\\
F_2(X^k) + A_2^k\Delta X^{AP} &= 0, && \text{(4.17b)}\\
F_3(X^k) + A_3^k\Delta X^{AP} &\geq 0, && \text{(4.17c)}\\
|\nu^{AP}| &\leq 1 \text{ with} && \text{(4.17d)}\\
\forall i \notin I_{F_1}(\Delta X^{AP}): \quad \nu_i^{AP} &= \operatorname{sign}\left(F_1(X^k) + A_1^k\Delta X^{AP}\right)_i, && \text{(4.17e)}\\
\mu^{AP} &\geq 0, && \text{(4.17f)}\\
\forall i = 1,\ldots,m_3: \quad \mu_i^{AP}(F_3(X^k) + A_3^k\Delta X^{AP})_i &= 0. && \text{(4.17g)}
\end{aligned}$$

The linearized LAD minimization problem is essentially a linear program. The increment ΔX^{AP} is the primal optimal solution and non-degenerate due to the constraint qualification. The dual solution $(\nu^{AP},\lambda^{AP},\mu^{AP})$ is also non-degenerate due to the strict complementarity condition (4.16). Thus according to the linear programming theory the solution of the linearized l_1 norm minimization problem is unique and satisfies the KKT conditions (4.17). □

The inexact Gauss-Newton method for the LAD minimization problem is summarized in algorithm 4.2.1.

Note that the correction term in the objective function of the inexact problem (4.12),

$$m^{kT}\Delta X = -\left(\nu^{kT}(A_1^k - J_1(X^k)) - \lambda^{kT}(A_2^k - J_2(X^k)) - \mu^{kT}(A_3^k - J_3(X^k))\right)\Delta X$$

can be written equivalently with

$$m^k = \nabla_X\mathcal{L}(X^k,\nu^k,\lambda^k,\mu^k) - A_1^{kT}\nu^k + A_2^{kT}\lambda^k + A_3^{kT}\mu^k. \tag{4.18}$$

We will show in the following section that there exists a neighborhood of the local solution X^* of the nonlinear problem in which the active set of the inequality constraints and the partitioning of the objective function are the same for the nonlinear and the inexact least absolute deviations estimation problem if only the approximations of the Jacobians are chosen close enough to the exact Jacobians. Then it is quite straightforward to prove local convergence of algorithm 4.2.1 within this neighborhood.

Algorithm 4.2.1 Local inexact GN method for l_1 norm optimization problems

(S1) Choose $(X^0, \nu^0, \lambda^0, \mu^0) \in \mathbb{R}^n \times \mathbb{R}^{m_1} \times \mathbb{R}^{m_2} \times \mathbb{R}^{m_3}$, $A_i^0, i = 1, 2, 3$, to have full rank and such that $\tilde{A}^0$ is nonsingular, and set $k := 0$.

(S2) If $(X^k, \nu^k, \lambda^k, \mu^k)$ is an FJ point of (4.6): STOP.

(S3) Compute an FJ point $(\Delta X^{AP}, \nu^{AP}, \lambda^{AP}, \mu^{AP}) \in \mathbb{R}^n \times \mathbb{R}^{m_1} \times \mathbb{R}^{m_2} \times \mathbb{R}^{m_3}$ of the subproblem

$$\begin{array}{ll} \min\limits_{\Delta X} & \|F_1(X^k) + A_1^k \Delta X\|_1 \\ & - \left({\nu^k}^T (A_1^k - J_1(X^k)) - {\lambda^k}^T (A_2^k - J_2(X^k)) - {\mu^k}^T (A_3^k - J_3(X^k)) \right) \Delta X \\ \text{s.t.} & F_2(X^k) + A_2^k \Delta X = 0, \\ & F_3(X^k) + A_3^k \Delta X \geq 0, \end{array}$$

and set $X^{k+1} := X^k + \Delta X^{AP}$, $\nu^{k+1} := \nu^k + (\nu^{AP} - \nu^k)$, $\lambda^{k+1} := \lambda^k + (\lambda^{AP} - \lambda^k)$, and $\mu^{k+1} := \mu^k + (\mu^{AP} - \mu^k)$.

(S4) If a convergence test fails, choose $A_i^{k+1}, i = 1, 2, 3$, to have full rank and such that $\tilde{A}^{k+1}$ is nonsingular, set $k \leftarrow k + 1$ and go to (S2).

4.2.2. Stability of the active set and local convergence

We define the matrices

$$\mathcal{A} = \begin{pmatrix} 0 & A_1^T & -A_2^T & -A_3^T \\ A_1 & 0 & 0 & 0 \\ A_2 & 0 & 0 & 0 \\ A_3 & 0 & 0 & 0 \end{pmatrix}, \qquad \tilde{\mathcal{A}} = \begin{pmatrix} 0 & {\tilde{A}_1}^T & -A_2^T & -{\tilde{A}_3}^T \\ \tilde{A}_1 & 0 & 0 & 0 \\ A_2 & 0 & 0 & 0 \\ \tilde{A}_3 & 0 & 0 & 0 \end{pmatrix}.$$

Now we show the stability of the active sets in a neighborhood of a local solution of the inexact LAD estimation problem.

Theorem 4.2.1 (Stability of the active sets of the problem with approximated derivatives near a solution of the l_1 norm minimization problem)**.** *Let $(X^*, \nu^*, \lambda^*, \mu^*)$ be an FJ point of the l_1-problem (4.6), $\|\cdot\|_2$ be the Euklidian vector norm and $\|\cdot\|_F$ be the Frobenius matrix norm. Further let X^* be a regular point and let the strict complementarity condition,*

$$|\nu_i| < 1 \text{ for all } i \in I_{F_1}(X), \qquad \mu_i \neq 0 \text{ for all } i \in I_{F_3}(X),$$

hold at $(X^*, \nu^*, \lambda^*, \mu^*)$. Further, let the following conditions hold for given constants $\delta_1, \delta_2 > 0$:

- $\|A_1\|_F, \|A_2\|_F, \|A_3\|_F \leq \delta_1$,

- *the matrix*

$$\tilde{\mathcal{A}} = \begin{pmatrix} 0 & \tilde{A_1}^T & -A_2^T & -\tilde{A_3}^T \\ \tilde{A_1} & 0 & 0 & 0 \\ A_2 & 0 & 0 & 0 \\ \tilde{A_3} & 0 & 0 & 0 \end{pmatrix}, \tag{4.19}$$

  is invertible, where $\tilde{A_1} = (A_{1i})_{i \in I_{F_1}(X^*)}$ *with* A_{1i} *being the* i*th row of* A_1 *and* $\tilde{A_3} = (A_{3i})_{i \in I_{F_3}(X^*)}$ *with* A_{3i} *being the* i*th row of* A_3*, and*

- $\|\mathcal{A}\|_F \leq \delta_2$.

Then there exists a neighborhood $\mathcal{N}$ *of* $(X^*, \nu^*, \lambda^*, \mu^*)$ *and a constant* $\kappa > 0$ *such that for all* $(X, \nu, \lambda, \mu) \in \mathcal{N}$ *the problem with approximated derivatives*

$$\begin{aligned} \min_{\Delta X} \quad & \|F_1(X) + A_1 \Delta X\|_1 \\ & - \left(\nu^T (A_1 - J_1(X)) - \lambda^T (A_2 - J_2(X)) - \mu^T (A_3 - J_3(X))\right) \Delta X \\ s.t. \quad & F_2(X) + A_2 \Delta X = 0, \\ & F_3(X) + A_3 \Delta X \geq 0, \end{aligned} \tag{4.20}$$

has a unique solution $(\Delta X^{AP}, \nu^{AP}, \lambda^{AP}, \mu^{AP})$ *which satisfies the inequality*

$$\begin{aligned} \|(\Delta X^{AP}, \nu^{AP}, \lambda^{AP}, \mu^{AP}) - (0, \nu^*, \lambda^*, \mu^*)\| \\ \leq \kappa \|(X^*, \nu^*, \lambda^*, \mu^*) - (X, \nu, \lambda, \mu)\|. \end{aligned} \tag{4.21}$$

Furthermore, the problem with approximated derivatives (4.20) has the same active set and the same positive and negative components of the objective function as the l_1*-solution* X^*.

Proof. Any solution ΔX^{AP} with the corresponding multipliers $\nu^{AP}, \lambda^{AP}, \mu^{AP}$ of (4.20) have to satisfy the FJ conditions with $\mu_0 = 1$, i.e. the KKT conditions

$$\begin{aligned} J_1(X)^T \nu - J_2(X)^T \lambda - J_3(X)^T \mu + \\ A_1^T(\nu^{AP} - \nu) - A_2^T(\lambda^{AP} - \lambda) - A_3^T(\mu^{AP} - \mu) = 0, & \quad (4.22a) \\ F_2(X) + A_2 \Delta X^{AP} = 0, & \quad (4.22b) \\ F_3(X) + A_3 \Delta X^{AP} \geq 0, & \quad (4.22c) \\ |\nu^{AP}| \leq 1 \text{ with} & \quad (4.22d) \\ i \notin I_{F_1}(\Delta X^{AP}): \quad |\nu_i^{AP}| = \operatorname{sign}(F_1(X) + A_1 \Delta X^{AP})_i, & \quad (4.22e) \\ \mu^{AP} \geq 0, & \quad (4.22f) \\ i = 1, \ldots, m_3: \quad \mu_i^{AP}(F_3(X) + A_3 \Delta X^{AP})_i = 0. & \quad (4.22g) \end{aligned}$$

Note here that the requirement of the matrix $\tilde{\mathcal{A}}$ being regular can be considered as a constraint qualification that allows choosing $\mu_0 = 1$ in the FJ conditions. For (4.22a) also remember that $J_1(X)^T \nu - J_2(X)^T \lambda - J_3(X)^T \mu = \nabla_X \mathcal{L}(X, \nu, \lambda, \mu)$.

In $(X, \nu, \lambda, \mu) = (X^*, \nu^*, \lambda^*, \mu^*)$ we obviously get an FJ point of (4.20) by setting $(\Delta X^{AP}, \nu^{AP}, \lambda^{AP}, \mu^{AP}) = (0, \nu^*, \lambda^*, \mu^*)$. The active set as well as the positive and negative components of the objective function and the validity of the strict complementarity conditions are clearly inherited from the l_1-solution. The constraint qualification is satisfied if

$$\tilde{A} = \begin{pmatrix} \tilde{A}_1 \\ A_2 \\ \tilde{A}_3 \end{pmatrix}$$

has full rank and the uniqueness of ΔX^{AP} and the Lagrange multipliers follows from the invertibility of $\tilde{\mathcal{A}}$.
Due to this uniqueness of ΔX^{AP} and the Lagrange multipliers, the FJ point is a strict local minimum of (4.20).
Obviously, due to $\Delta X^* = 0$ we always have $I_{F_1}(\Delta X^*) = I_{F_1}(X^*)$, $I^+_{F_1}(\Delta X^*) = I^+_{F_1}(X^*)$, $I^-_{F_1}(\Delta X^*) = I^-_{F_1}(X^*)$, and $I_{F_3}(\Delta X^*) = I_{F_3}(X^*)$. We will now show that there exists a neighborhood $\mathcal{N}^1_X$ of X^* in which the positive and negative components of the objective function do not change sign or become zero and in which these sets are the same for the nonlinear problem at X^* and the inexactly linearized minimization problem. In this neighborhood we then have $I_{F_1}(\Delta X^{AP}) = I_{F_1}(X^*)$, $I^+_{F_1}(\Delta X^{AP}) = I^+_{F_1}(X^*)$, $I^-_{F_1}(\Delta X^{AP}) = I^-_{F_1}(X^*)$, and thus $\nu^{AP}_i = \nu^{AP*}_i$ for $i \notin I_{F_1}(X^*)$.
Define $F^+_1(X) = (F_{1i}(X))_{i \in I^+_{F_1}(X^*)}$ and $A^+_1 = (A_{1i})_{i \in I^+_{F_1}(X^*)}$, i.e. the submatrix of A_1 that consists of those rows only that belong to the positive components of the objective function. We can make the estimation

$$\begin{aligned} |(F_1(X) + A_1\Delta X - F_1(X^*))_i| &\leq |F_{1i}(X) - F_{1i}(X^*)| + |(A_1\Delta X)_i| \\ &\leq L_1|X^*_i - X_i| + \delta_1|\Delta X_i| \\ &\leq (L_1 + \delta_1)|X^*_i - X_i|, \end{aligned} \tag{4.23}$$

which holds for all $i \in I^+_{F_1}(X^*)$. Denote $f_+ := \min_{i \in I^+_{F_1}(X^*)}\{F_{1i}(X^*)\} > 0$ and choose $0 < f_{1+} < f_+$. By choosing an appropriate neighborhood $\mathcal{N}^1_+$ of X^*, we can guarantee that $(L_1 + \delta_1)|X^*_i - X_i| \leq f_{1+}$ and thus $|(F_1(X) + A_1\Delta X)_i - F_{1i}(X^*)| \leq f_{1+}$ for all $X \in \mathcal{N}^1_+$. It follows that $(F_1(X) + A_1\Delta X)_i = F_{1i}(X^*) + [(F_1(X) + A_1\Delta X)_i - F_{1i}(X^*)] \geq f_+ - f_{1+} > 0$ for all $i \in I^+_{F_1}(X^*)$, i.e. $(F^+_1(X) + A^+_1\Delta X)_i > 0$, i.e. $I^+_{F_1}(\Delta X^{AP}) \supseteq I^+_{F_1}(X^*)$.
Further define $F^-_1(X) = (F_{1i}(X))_{i \in I^-_{F_1}(X^*)}$ and $A^-_1 = (a_{1i})_{i \in I^-_{F_1}(x^*)}$, i.e. the submatrix of A_1 that consists of those rows only that belong to the negative components of the objective function. Estimate (4.23) also holds for all $i \in I^-_{F_1}(X^*)$. Denote $f_- := -\max_{i \in I^-_{F_1}(X^*)}\{F_{1i}(X^*)\} > 0$ and choose $0 < f_{1-} < f_-$. By choosing an appropriate neighborhood $\mathcal{N}^1_-$ of X^*, we can guarantee that $(L_1 + \delta_1)|X^*_i - X_i| \leq f_{1-}$ and thus $|(F_1(X) + A_1\Delta X)_i - F_{1i}(X^*)| \leq f_{1-}$ for all $X \in \mathcal{N}^1_-$. It follows that $(F_1(X) + A_1\Delta X)_i = F_{1i}(X^*) + [(F_1(X) + A_1\Delta X)_i - F_{1i}(X^*)] \leq -f_- + f_{1-} < 0$ for all $i \in I^-_{F_1}(X^*)$, i.e. $(F^-_1(X) + A^-_1\Delta X)_i < 0$, i.e. $I^-_{F_1}(\Delta X^{AP}) \supseteq I^-_{F_1}(X^*)$.

If we now require $I_{F_1}(\Delta X^{AP}) = I_{F_1}(X^*)$, i.e. we try to solve the KKT conditions (4.22) with the additional conditions

$$(F_1(X) + A_1\Delta X)_i = 0 \ \forall i \in I_{F_1}(X^*), \tag{4.24}$$

we at once get for $X \in \hat{\mathcal{N}}_X^1 \subset \mathcal{N}_+^1 \cap \mathcal{N}_-^1$ that the positive and negative components of the objective function are the same as in X^*.
For the inequality constraints, define $\check{F}_3(X) = (F_{3i}(X))_{i \notin I_{F_3}(X^*)}$ and the matrix $\check{A}_3 = (a_{3i})_{i \notin I_{F_3}(X^*)}$, i.e. the submatrix of A_3 that consists of those rows only that belong to the inactive components of the inequality constraints. We can make the estimation

$$\begin{aligned} |(F_3(X) + A_3\Delta X - F_3(X^*))_i| &\leq |(F_{3i}(X) - F_{3i}(X^*)| + \|(A_3\Delta X)_i| \\ &\leq L_3|X_i^* - X_i| + \delta_1|\Delta X_i| \\ &\leq (L_3 + \delta_1)|X_i^* - X_i|, \end{aligned}$$

which holds for all $i \notin I_{F_3}(X^*)$. Denote $f := \min_{j \notin I_{F_3}(X^*)}\{F_{3j}(X^*)\} > 0$ and choose $0 < f_3 < f$. By choosing an appropriate neighborhood $\mathcal{N}_X^3$ of X^*, we can guarantee that $(L_3+\delta_1)|X_i^* - X_i| \leq f_3$ and thus $|(F_3(X)+A_3\Delta X)_i - F_{3i}(X^*)| \leq f_3$ for all $X \in \mathcal{N}_X^3$. It follows that $(F_3(X) + A_3\Delta X)_i = F_{3i}(X^*) + [(F_3(X) + A_3\Delta X)_i - F_{3i}(X^*)] \geq f - f_3 > 0$ for all $i \notin I_{F_3}(X^*)$, i.e. $\check{F}_3(X) + \check{A}_3\Delta X > 0$, i.e. $I_{F_3}(\Delta X^{AP}) \subset I_{F_3}(X^*)$.
If we now require $I_{F_3}(\Delta X^{AP}) = I_{F_3}(X^*)$, i.e. we try to solve the KKT conditions (4.22) with the additional conditions

$$(F_3(X) + A_3\Delta X)_i = 0 \ \forall i \in I_{F_3}(X^*), \tag{4.25}$$

we at once get for $X \in \hat{\mathcal{N}}_X^3 \subset \mathcal{N}_X^3$ that the active components of the inequality constraints are the same as in the solution of problem (4.20).
So we have shown that there exists a neighborhood $\mathcal{N}_X := \hat{\mathcal{N}}_X^1 \cap \hat{\mathcal{N}}_X^3$ in which the partitioning of the objective function and the inequality constraints for the inexactly linearized LAD optimization problem is the same as for the nonlinear problem at the optimum X^*.
Now let (4.20) be formulated in the point (X, ν, λ, μ) with $X \in \mathcal{N}_X$ and define

$$m(X, \nu, \lambda, \mu) := \nabla_X \mathcal{L}(X, \nu, \lambda, \mu) - A_1^T\nu + A_2^T\lambda + A_3^T\mu. \tag{4.26}$$

From the invertibility of $\tilde{\mathcal{A}}$ follows that the solution of the linear system

$$\tilde{\mathcal{A}} \begin{pmatrix} s \\ a \\ b \\ c \end{pmatrix} = \begin{pmatrix} -m(X, \nu, \lambda, \mu) - \sum_{i \notin I_{F_1}(X^*)} \nu_i^{AP*} {A_1}_i^T \\ -\tilde{F}_1(X) \\ -F_2(X) \\ -\tilde{F}_3(X) \end{pmatrix}, \tag{4.27}$$

is unique. Let us for short write

$$\bar{m}(X, \nu, \lambda, \mu) := m(X, \nu, \lambda, \mu) + \sum_{i \notin I_{F_1}(X^*)} \nu_i^{AP*} {A_1}_i^T. \tag{4.28}$$

Then we further have

$$\begin{pmatrix} s-0 \\ a-\tilde{\nu}^* \\ b-\lambda^* \\ c-\tilde{\mu}^* \end{pmatrix} = \tilde{\mathcal{A}}^{-1} \begin{pmatrix} \bar{m}(X^*,\nu^*,\lambda^*\mu^*) - \bar{m}(X,\nu,\lambda,\mu) \\ \tilde{F}_1(X^*) - \tilde{F}_1(X) \\ F_2(X^*) - F_2(X) \\ \tilde{F}_3(X^*) - \tilde{F}_3(X) \end{pmatrix}. \tag{4.29}$$

As we have assumed F_1, F_2, F_3 to be at least twice continuously differentiable in X, the gradient of the Lagrangian, $\nabla_X \mathcal{L}(X,\nu,\lambda,\mu)$, is continuously differentiable in (X,ν,λ,μ) and thus locally Lipschitz continuous in (X,ν,λ,μ) with a Lipschitz constant L. Further let $L_j,\ j=1,2,3$, be the Lipschitz constants of $F_j(x),\ j=1,2,3$, respectively. Then

$$\begin{aligned} &\|\bar{m}(X^*,\nu^*,\lambda^*,\mu^*) - \bar{m}(X,\nu,\lambda,\mu)\| \\ &\qquad \leq \|m(X^*,\nu^*,\lambda^*,\mu^*) - m(X,\nu,\lambda,\mu)\|. \end{aligned} \tag{4.30}$$

We can continue the estimation in the following way,

$$\begin{aligned} &\|m(X^*,\nu^*,\lambda^*,\mu^*) - m(X,\nu,\lambda,\mu)\| \\ &\qquad \leq \|\nabla_X \mathcal{L}(X^*,\nu^*,\lambda^*,\mu^*) - \nabla_X \mathcal{L}(X,\nu,\lambda,\mu)\| \\ &\qquad\quad + \|A_1^T(\nu^*-\nu)\| + \|A_2^T(\lambda^*-\lambda)\| + \|A_3^T(\mu^*-\mu)\| \\ &\qquad \leq L\|(X^*,\nu^*,\lambda^*,\mu^*) - (X,\nu,\lambda,\mu)\| \\ &\qquad\quad + \delta_1\|\nu^*-\nu\| + \delta_1\|\lambda^*-\lambda\| + \delta_1\|\mu^*-\mu\| \\ &\qquad \leq (L+3\delta_1)\|(X^*,\nu^*,\lambda^*,\mu^*) - (X,\nu,\lambda,\mu)\|, \end{aligned} \tag{4.31}$$

where we have used $\|A\|_F = \|A^T\|_F$. From this follows the estimate

$$\left\| \begin{pmatrix} s-0 \\ a-\tilde{\nu}^* \\ b-\lambda^* \\ c-\tilde{\mu}^* \end{pmatrix} \right\| \leq \delta_2(L+L_1+L_2+L_3+3\delta_1) \left\| \begin{pmatrix} X^*-X \\ \nu^*-\nu \\ \lambda^*-\lambda \\ \mu^*-\mu \end{pmatrix} \right\|. \tag{4.32}$$

Let us recall the FJ conditions (4.22) of the problem with approximated derivatives plus conditions (4.24) and (4.25) with the solution $(s^T,a^T,b^T,c^T)^T$ of (4.27):

- Set $\Delta X = s$. Then (4.22b) follows from (4.27) directly.
- Using esimate (4.32) we can make the estimation

  $$\begin{aligned} \|F_1(X) + A_1\Delta X - F_1(X^*)\| &\leq \|F_1(X) - F_1(X^*)\| + \|A_1\Delta X\| \\ &\leq L_1\|X^*-X\| + \delta_1\|\Delta X\| \\ &\leq (L_1+\delta_1\theta)\|(X^*,\nu^*,\lambda^*,\mu^*) - (X,\nu,\lambda,\mu)\| \end{aligned}$$

  with $\theta = \delta_2(L+L_1+L_2+L_3+3\delta_1)$. Analogically to the explanations above, where we derived the neighborhood $\mathcal{N}_X^1$ of X^* in which the positive and negative components of the objective function stay constant, we can now find a neighborhood $\mathcal{N}^1 = \mathcal{N}_X^1 \times \mathcal{N}_\nu^1 \times \mathcal{N}_\lambda^1 \times \mathcal{N}_\mu^1$ of $(X^*,\nu^*,\lambda^*,\mu^*)$ in which the same conclusions hold.

- Analogically, we get a neighborhood $\mathcal{N}^3 = \mathcal{N}_X^3 \times \mathcal{N}_\nu^3 \times \mathcal{N}_\lambda^3 \times \mathcal{N}_\mu^3$ of the point $(X^*, \nu^*, \lambda^*, \mu^*)$ in which condition (4.22c) holds with the same active components of the inequality constraints as in the solution of problem (4.20) defined in the optimum $(X^*, \nu^*, \lambda^*, \mu^*)$.

- Without loss of generality let $I_{F_3}(X^*) = \{1, \ldots, m_3^0\}$ and set $\mu^{AP} = (c^T, 0^T)^T$. As $\tilde{\mu}^* > 0$, from inequality (4.32) we can again find a neighborhood $\bar{\mathcal{N}}^3 = \bar{\mathcal{N}}_X^3 \times \bar{\mathcal{N}}_\nu^3 \times \bar{\mathcal{N}}_\lambda^3 \times \bar{\mathcal{N}}_\mu^3$ of $(X^*, \nu^*, \lambda^*, \mu^*)$ such that $c > 0$ for all $(X, \nu, \lambda, \mu) \in \bar{\mathcal{N}}^3$. Thus we get for all $(X, \nu, \lambda, \mu) \in \mathcal{N}^{F_3} := \mathcal{N}^3 \cap \bar{\mathcal{N}}^3$ conditions (4.22f) and (4.22g) with strict complementarity.

- Without loss of generality let $I_{F_1}(X^*) = \{1, \ldots, m_1^0\}$, as well as $I_{F_1}^+(X^*) = \{m_1^0 + 1, \ldots, m_1^0 + m_1^+\}$, and $I_{F_1}^-(X^*) = \{m_1^0 + m_1^+ + 1, \ldots, m_1\}$. Then set $\nu^{AP} = (a^T, 1^T, (-1)^T)^T$. As $|\tilde{\nu}^*| < 1$, from inequality (4.32) we can again find a neighborhood $\bar{\mathcal{N}}^1 = \bar{\mathcal{N}}_X^1 \times \bar{\mathcal{N}}_\nu^1 \times \bar{\mathcal{N}}_\lambda^1 \times \bar{\mathcal{N}}_\mu^1$ of $(X^*, \nu^*, \lambda^*, \mu^*)$ such that $|a| < 1$ for all $(X, \nu, \lambda, \mu) \in \bar{\mathcal{N}}^1$. Thus we get for all $(X, \nu, \lambda, \mu) \in \mathcal{N}^{F_1} := \mathcal{N}^1 \cap \bar{\mathcal{N}}^1$ conditions (4.22d) and (4.22e) with strict complementarity.

- With $\lambda^{AP} = b$ the Lagrangian gradient (4.22a) of the problem with approximated derivatives vanishes, as can also be seen from (4.27).

Thus for all $(X, \nu, \lambda, \mu) \in \mathcal{N} := \mathcal{N}^{F_3} \cap \mathcal{N}^{F_1}$ we get an FJ point of the inexact LAD minimization problem (4.20) in $(\Delta X^{AP}, \nu^{AP}, \lambda^{AP}, \mu^{AP})$ with

$$\begin{aligned} \Delta X^{AP} &= s, \\ \nu^{AP} &= (a^T, 1^T, -1^T)^T, \\ \lambda^{AP} &= b, \\ \mu^{AP} &= (c^T, 0^T)^T, \end{aligned}$$

which satisfies the necessary optimality conditions with the active set $I_{F_3}(X^*)$ and the positive and negative components of the objective function $I_{F_1}^+(X^*)$ and $I_{F_1}^-(X^*)$, respectively. Estimate (4.21) follows from (4.32), appropriately filled up with positive and negative ones and zeros. □

Due to the equivalence of all vector norms in a finite-dimensional normed vector space, we could have used any vector norm and corresponding matrix norm. With the choice of the Frobenius norm we followed the trick used by Wirsching [140] to be able to use the same bound for both a matrix and its transposed matrix. It was also along the lines of the proof in this work that we did our proof here which just needed a little expansion to our problem.

Theorem 4.2.1 allows a rather arbitrary choice for the Jacobian approximations A_1, A_2, and A_3 as long as they satisfy the conditions required by the theorem. In particular, no relation between the exact Jacobians and their approximations had to be assumed. This will change in the following where we will give requirements for

the iterative process to converge locally to the solution of the l_1 norm estimation problem.

We can now prove local convergence of algorithm 4.2.1 for the inexact l_1 norm minimization problem.

Theorem 4.2.2 (Local convergence of the inexact Gauss-Newton method for the LAD minimization problem with fixed active sets)**.** *Let $(X^*, \nu^*, \lambda^*, \mu^*)$ be an FJ point of the l_1-problem (4.6) and $\|\cdot\|$ be an arbitrary but fixed vector norm. Further assume that the following requirements are satisfied:*

(i) *X^* is a regular point and strict complementarity holds in $(X^*, \nu^*, \lambda^*, \mu^*)$.*

(ii) *(A_1^k), (A_2^k), and (A_3^k) are uniformly bounded sequences of matrices such that the sequence $\left((\tilde{\mathcal{A}}^k)^{-1}\right)$, with $\tilde{\mathcal{A}}^k$ defined as in theorem 4.2.1, i.e. according to the active sets of the nonlinear problem at the optimum X^*, is also uniformly bounded.*

(iii) *There exists a sequence $y^k = (X^k, \nu^k, \lambda^k, \mu^k)^T$ generated according to the rule $y^{k+1} = y^k + \Delta y^k$ with $\Delta y^k = (\Delta X^k, \nu^{AP,k} - \nu^k, \lambda^{AP,k} - \lambda^k, \mu^{AP,k} - \mu^k)^T$ where ΔX^k is a solution of the problem with approximated derivatives*

$$\begin{array}{ll} \min\limits_{\Delta X} & \|F_1(X^k) + A_1^k \Delta X\|_1 \\ & - \left(\nu^{kT}(A_1^k - J_1(X^k)) - \lambda^{kT}(A_2^k - J_2(X^k)) - \mu^{kT}(A_3^k - J_3(X^k))\right) \Delta X \\ s.t. & F_2(X^k) + A_2^k \Delta X = 0, \\ & F_3(X^k) + A_3^k \Delta X \geq 0, \end{array} \tag{4.33}$$

and $\nu^{AP,k}, \lambda^{AP,k}, \mu^{AP,k}$ are the Lagrange multipliers in the solution of this problem.

(iv) *There exists $\kappa < 1$ such that for all $k \in \mathbb{N}$ and all $t \in [0,1]$ it can be guaranteed that*

$$\left\| (\tilde{\mathcal{A}}^{k+1})^{-1} \left(\tilde{\mathcal{A}}^k - \nabla f(y^k + t\Delta y^k) \right) \Delta y^k \right\| \leq \kappa \|\Delta y^k\|,$$

where $y^k = (X^k, \nu^k, \lambda^k, \mu^k)^T$ and the function f is defined as

$$f(y^k) = \begin{pmatrix} -m(y^k) - \sum_{i \notin I_{F_1}(X^*)} \nu_i^{AP*} A_{1i}^{kT} \\ -\tilde{F}_1(X^k) \\ -F_2(X^k) \\ -\tilde{F}_3(X^k) \end{pmatrix},$$

where $m(y^k)$ is defined as in (4.26), $\tilde{F}_1(X) = (F_{1i}(X))_{i \in I_{F_1}(X^*)}$ and $\tilde{F}_3(X) = (F_{3i}(X))_{i \in I_{F_3}(X^*)}$

Then there exists a neighborhood $\mathcal{N}$ of $(X^*, \nu^*, \lambda^*, \mu^*)$ such that for all initial guesses $(X^0, \nu^0, \lambda^0, \mu^0) \in \mathcal{N}$ the sequence $(X^k, \nu^k, \lambda^k, \mu^k)$ converges q-linearly to the FJ point

$(X^*, \nu^*, \lambda^*, \mu^*)$ with rate κ and the solution of the problem with approximated derivatives (4.33) has the same partitioning of the objective function as the nonlinear problem at X^*.

Proof. As the sequences (A_1^k), (A_2^k), (A_3^k), and $\left((\tilde{\mathcal{A}}^k)^{-1}\right)$ are uniformly bounded, we can find constants $\delta_1, \delta_2 > 0$ such that $\|A_1^k\|, \|A_2^k\|, \|A_3^k\| \leq \delta_1$ and $\|(\mathcal{A}^k)^{-1}\| \leq \delta_2$ for all $k \in \mathbb{N}$. Therefore we can apply theorem 4.2.1 on the stability of the active set. From this theorem follows the existence of a neighborhood $\bar{\mathcal{N}}$ of $(X^*, \nu^*, \lambda^*, \mu^*)$ such that for all $(X^k, \nu^k, \lambda^k, \mu^k) \in \bar{\mathcal{N}}$ and all $k \in \mathbb{N}$ the problem with approximated derivatives (4.33) has a unique solution and the same active set as well as the same positive and negative components of the objective function as the l_1-solution. Without loss of generality we can again assume that $I_{F_3}(X^*) = \{1, \ldots, m_3^0\}$, $I_{F_1}(X^*) = \{1, \ldots, m_1^0\}$, $I_{F_1}^+(X^*) = \{m_1^0 + 1, \ldots, m_1^0 + m_1^+\}$, $I_{F_1}^-(X^*) = \{m_1^0 + m_1^+ + 1, \ldots, m_1\}$. It is easy to see that the solution of (4.33) cannot only be computed from system (4.27) but also by solving the equivalent system

$$\tilde{\mathcal{A}}^k \begin{pmatrix} s^k \\ a^k \\ b^k \\ c^k \end{pmatrix} = f(y^k), \quad y^k = (X^k, \nu^k, \lambda^k, \mu^k). \tag{4.34}$$

The equivalence of the two equation systems is a consequence of the equivalence of Newton's method and the SQP method.
The sequence being generated by choosing an initial guess

$$(y^0)^T := ((X^0)^T, (\tilde{\nu}^0)^T, 1^T, -1^T, (\lambda^0)^T, (\tilde{\mu}^0)^T, 0^T) \in \bar{\mathcal{N}}$$

and applying the iteration $y^{k+1} = y^k + \Delta y^k$ with the increment

$$(\Delta y^k)^T = ((s^k)^T, (a^k)^T, 0^T, 0^T, (b^k)^T, (c^k)^T, 0)^T$$

is well defined as long as it can be guaranteed that the iterates remain in $\bar{\mathcal{N}}$. To show this as well as convergence of the sequence toward $(X^*, \nu^*, \lambda^*, \mu^*)$, we apply lemma 4.1.1. To do so, we have to show that the requirements of the lemma are satisfied:
As the sequences (A_1^k), (A_2^k), (A_3^k) are uniformly bounded, it can easily be seen that also the sequence $(\tilde{\mathcal{A}}^k)$ is uniformly bounded. What remains to be done is the determination of the neighborhood $\mathcal{N}$ such that $\mathcal{B}(y^0, \frac{\|\Delta y^0\|}{1-\kappa}) \subset \bar{\mathcal{N}}$ for all $y^0 \in \mathcal{N}$.
As $\bar{\mathcal{N}}$ is a neighborhood of $(X^*, \nu^*, \lambda^*, \mu^*)$, there exists an $\varepsilon > 0$ such that $\mathcal{B}(X^*, \varepsilon) \subset \bar{\mathcal{N}}$. As $f(y^*) = 0$ we can choose the first step $\Delta y^0 = -(\tilde{\mathcal{A}}^0)^{-1} f(y^0)$ arbitrarily small if the distance $\|y^* - y^0\|$ is small enough, due to the differentiability of f and the boundedness of $(\tilde{\mathcal{A}}^0)^{-1}$. For a given $\varepsilon > 0$ we can thus always find a $\delta > 0$ such that $\|y^* - y^0\| + \frac{\|\Delta y^0\|}{1-\kappa} \leq \varepsilon$ whenever $\|y^* - y^0\| \leq \delta$. Therefore it can be guaranteed for all $y^0 \in \mathcal{N} := \mathcal{B}(y^*, \delta)$ that $\mathcal{B}(y^0, \frac{\|\Delta y^0\|}{1-\kappa}) \subset \mathcal{B}(y^*, \varepsilon) \subset \bar{\mathcal{N}}$. The convergence of the sequence $(X^k, \nu^k, \lambda^k, \mu^k)$ to $(X^*, \nu^*, \lambda^*, \mu^*)$ follows from lemma 4.1.1. □

Like for the exact method, the solution of the problem with approximated derivatives (4.33) can be computed by means of a generalized inverse. Therefore define

$$F(X) = \begin{pmatrix} F_1(X) \\ F_2(X) \\ F_3(X) \end{pmatrix}, \quad J(X) = \begin{pmatrix} J_1(X) \\ J_2(X) \\ J_3(X) \end{pmatrix}, \quad A = \begin{pmatrix} A_1 \\ A_2 \\ A_3 \end{pmatrix}. \tag{4.35}$$

Theorem 4.2.3 (Existence of a generalized inverse for the LAD minimization problem with fixed active sets)**.** *Let ΔX^k be a minimizer of (4.33) and let the following conditions hold:*

(i) the matrix $\tilde{A}^k$ has full row rank,

$$rank\,(\tilde{A}^k) = rank \begin{pmatrix} \tilde{A_1}^k \\ A_2^k \\ \tilde{A_3}^k \end{pmatrix} = n_0 := m_1^0 + m_2 + m_3^0,$$

where $\tilde{A_1}^k = (A_{1i})_{i \in I_{F_1}(X^*)}$ and $\tilde{A_3}^k = (A_{3i}))_{i \in I_{F_1}(X^*)}$,

(ii) there exist Lagrange multipliers $\nu^{AP,k}, \lambda^{AP,k}, \mu^{AP,k}$, such that the FJ conditions are satisfied with $\mu_0 = 1$ and such that

$$rank\,(\tilde{A}^k) = rank \begin{pmatrix} (A_{1i})_{i \in I_{F_1}(X^k) \text{ with } |\nu_i^k| < 1} \\ A_2 \\ (A_{3i})_{i \in I_{F_3}(X^k) \text{ with } \mu_i^k > 0} \end{pmatrix} = n.$$

Then

(a) $(\Delta X^k, \nu^{AP,k}, \lambda^{AP,k}, \mu^{AP,k})$ is a unique FJ point of (4.33) and ΔX^k is a strict minimum.

(b) There exists a matrix $J_k^+ \in \mathbb{R}^{n \times (m_1 + m_2 + m_3)}$ such that $\Delta X^k = -J_k^+ F(X^k)$.

(c) The solution operator is a generalized inverse:

$$J_k^+ J(X^k) J_k^+ = J_k^+.$$

Proof. The FJ system for (4.33) with $\mu_0 = 1$ is

$$\begin{pmatrix} 0 & \tilde{A_1}^{kT} & -A_2^{kT} & -\tilde{A_3}^{kT} \\ \tilde{A_1}^k & 0 & 0 & 0 \\ A_2^k & 0 & 0 & 0 \\ \tilde{A_3}^k & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \Delta X \\ \tilde{\nu}^{AP} - \tilde{\nu}^k \\ \lambda^{AP} - \lambda^k \\ \tilde{\mu}^{AP} - \tilde{\mu}^k \end{pmatrix}$$

$$= \begin{pmatrix} -m(X^k, \nu^k, \lambda^k, \mu^k) - \sum_{i \notin I_{F_1}(X^*)} \nu_i^{AP*} A_{1i}^{kT} \\ -\tilde{F}_1(X^k) \\ -F_2(X^k) \\ -\tilde{F}_3(X^k) \end{pmatrix},$$

with $m(X, \nu, \lambda, \mu)$ defined as in (4.26), $\tilde{F}_1(X) = (F_{1i}(X))_{i \in I_{F_1}(X^*)}$ and $\tilde{F}_3(X) = (F_{3i}(X))_{i \in I_{F_3}(X^*)}$. Under the assumptions of the theorem, the matrix

$$\tilde{A}^k = \begin{pmatrix} \tilde{A}_1^k \\ A_2^k \\ \tilde{A}_3^k \end{pmatrix}$$

as defined in the requirements is quadratic and nonsingular. Hence,

$$\Delta X^k = -\left(\tilde{A}^k\right)^{-1} \tilde{F}(X^k), \tag{4.36}$$

i.e. $J_k^+ = P \left[0, \left(\tilde{A}^k\right)^{-1}\right]$ with a permuation matrix P. □

4.3. Inexact Gauss-Newton for Huber minimization

This section contains the derivation of the inexact Gauss-Newton algorithm for the nonlinear Huber minimization problem

$$\begin{array}{ll} \min\limits_{X} & \sum\limits_{i=1}^{m_1} \rho_\gamma(F_{1i}(X)) \\ \text{s.t.} & F_2(X) = 0, \\ & F_3(X) \geq 0, \end{array} \tag{4.37}$$

using Huber's function

$$\rho_\gamma(z) = \begin{cases} \frac{1}{2} z^2, & \text{if } |z| \leq \gamma, \\ |z|\gamma - \frac{1}{2}\gamma^2, & \text{if } |z| > \gamma, \end{cases}$$

as optimization criterion, where the problem functions $F_1 : \mathbb{R}^n \to \mathbb{R}^{m_1}$, $F_2 : \mathbb{R}^n \to \mathbb{R}^{m_2}$, $F_3 : \mathbb{R}^n \to \mathbb{R}^{m_3}$ are again assumed to be at least twice continuously differentiable.

Like in the previous section on l_1 norm minimization with inexact derivatives, we use a smooth reformulation of the optimization problem as our starting point,

$$\begin{array}{ll} \min\limits_{X \in \mathbb{R}^n, v, t \in \mathbb{R}^{m_1}} & \frac{1}{2} v^T v + \gamma \mathbb{1}^T t \\ \text{s.t.} & F_2(X) = 0, \\ & F_3(X) \geq 0, \\ & t + F_1(X) - v \geq 0, \\ & t - F_1(X) + v \geq 0. \end{array} \tag{4.38}$$

This corresponds to a general NLP (4.1) with the defining functions

$$F(X, v, t) = \frac{1}{2} v^T v + \gamma \mathbb{1}^T t, \tag{4.39a}$$

$$G(X, v, t) = F_2(X), \tag{4.39b}$$

$$H(X, v, t) = \begin{pmatrix} F_3(X) \\ t + F_1(X) - v \\ t - F_1(X) + v \end{pmatrix}. \tag{4.39c}$$

Again, we apply the theory introduced in section 4.1 on the Gauss-Newton method with inexact derivatives. The derivatives of the problem functions (4.39) are

$$\begin{aligned}
\nabla F(X,v,t) &= \begin{pmatrix} 0 & v^T & \gamma 1^T \end{pmatrix} \in \mathbb{R}^{n+2m_1}, && \text{(4.40a)}\\
\nabla G(X,v,t) &= \begin{pmatrix} \nabla F_2(X) & 0 & 0 \end{pmatrix} \in \mathbb{R}^{m_2\times(n+2m_1)}, && \text{(4.40b)}\\
\nabla H(X,v,t) &= \begin{pmatrix} \nabla F_3(X) & 0 & 0 \\ \nabla F_1(X) & -\mathbb{I} & \mathbb{I} \\ -\nabla F_1(X) & \mathbb{I} & \mathbb{I} \end{pmatrix} \in \mathbb{R}^{(m_3+2m_1)\times(n+2m_1)}. && \text{(4.40c)}
\end{aligned}$$

We continue to use the notations $J_i(X^k) = \nabla F_i(X^k),\ i = 1,2,3$, for the Jacobians of the problem functions at the iterates and $A_i^k \approx \nabla F_i(X^k), i = 1,2,3$, for approximations of these.

4.3.1. Derivation of the algorithm

Inserting these problem functions into the inexact QP (4.4), we get

$$\begin{aligned}
\min_{(\Delta X,\Delta v,\Delta t)} \quad & \frac{1}{2}\begin{pmatrix}\Delta X\\ \Delta v\\ \Delta t\end{pmatrix}^T \begin{pmatrix}\tilde{\mathcal{L}}^k & 0 & 0\\ 0 & \mathbb{I} & 0\\ 0 & 0 & 0\end{pmatrix}\begin{pmatrix}\Delta X\\ \Delta v\\ \Delta t\end{pmatrix} + \begin{pmatrix}0\\ v^k\\ \gamma 1\end{pmatrix}^T \begin{pmatrix}\Delta X\\ \Delta v\\ \Delta t\end{pmatrix}\\
& +\lambda_{QP}^T \begin{pmatrix} A_2^k - J_2(X^k), & 0, & 0\end{pmatrix}\begin{pmatrix}\Delta X\\ \Delta v\\ \Delta t\end{pmatrix}\\
& +\mu_{QP}^T \begin{pmatrix} A_3^k - J_3(X^k) & 0 & 0\\ A_1^k - J_1(X^k) & 0 & 0\\ -A_1^k + J_1(X^k) & 0 & 0\end{pmatrix}\begin{pmatrix}\Delta X\\ \Delta v\\ \Delta t\end{pmatrix} \qquad \text{(4.41)}\\
\text{s.t.} \quad & \begin{pmatrix} A_2^k & 0, & 0\end{pmatrix}\begin{pmatrix}\Delta X\\ \Delta v\\ \Delta t\end{pmatrix} + F_2(X^k) = 0,\\
& \begin{pmatrix} A_3^k & 0 & 0\\ A_1^k & -\mathbb{I} & \mathbb{I}\\ -A_1^k & \mathbb{I} & \mathbb{I}\end{pmatrix}\begin{pmatrix}\Delta X\\ \Delta v\\ \Delta t\end{pmatrix} + \begin{pmatrix} F_3(X^k)\\ t^k + F_1(X^k) - v^k\\ t^k - F_1(X^k) + v^k\end{pmatrix} \geq 0.
\end{aligned}$$

Again, the use of the Gauss-Newton method corresponds to neglecting all derivatives of second order or higher, such that the quadratic term in ΔX in the objective function is ignored. Splitting the multipliers for the inequality constraints into $\mu_{QP}^T = (\mu_{QP}^{(1)^T}, \mu_{QP}^{(2)^T}, \mu_{QP}^{(3)^T})$, according to the distinct block rows, the rest then figures

out to be

$$\begin{array}{ll}\min\limits_{(\Delta X,\Delta v,\Delta t)} & \frac{1}{2}\Delta v^T\Delta v + (v^k)^T\Delta v + \gamma 1^T\Delta t \\ & +\lambda_{QP}^T(A_2^k - J_2(X^k))\Delta X + \mu_{QP}^{(1)^T}(A_3^k - J_3(X^k))\Delta X \\ & -(\mu_{QP}^{(3)^T} - \mu_{QP}^{(2)^T})^T(A_1^k - J_1(X^k))\Delta X \\ \text{s.t.} & A_2^k\Delta X + F_2(X^k) = 0, \\ & A_3^k\Delta X + F_3(X^k) \geq 0, \\ & (t^k + \Delta t) + (A_1^k\Delta X + F_1(X^k)) - (v^k + \Delta v) \geq 0, \\ & (t^k + \Delta t) - (A_1^k\Delta X + F_1(X^k)) + (v^k + \Delta v) \geq 0.\end{array} \tag{4.42}$$

Implicitly adding the constant term $\frac{1}{2}(v^k)^T v^k + \gamma 1^T t^k$ to the objective function, the subproblem

$$\begin{array}{ll}\min\limits_{\Delta X,\Delta v,\Delta t} & \frac{1}{2}(v^k + \Delta v)^T(v^k + \Delta v) + \gamma 1^T(t^k + \Delta t) \\ \text{s.t.} & (t^k + \Delta t) + (A_1^k\Delta X + F_1(X^k)) - (v^k + \Delta v) \geq 0, \\ & (t^k + \Delta t) - (A_1^k\Delta X + F_1(X^k)) + (v^k + \Delta v) \geq 0,\end{array}$$

corresponds to the Huber minimization problem

$$\min_{\Delta X} \rho(A_1^k\Delta X + F_1(X^k)),$$

where ρ is Huber's function.

So (4.42) can be equivalently reformulated as

$$\begin{array}{ll}\min\limits_{\Delta X} & \sum\limits_{i=1}^{m_1} \rho_\gamma\big((F_1(X^k) + A_1^k\Delta X)_i\big) \\ & -\Big(\nu_{QP}^T(A_1^k - J_1(X^k)) - \lambda_{QP}^T(A_2^k - J_2(X^k)) - \mu_{QP}^T(A_3^k - J_3(X^k))\Big)\Delta X \\ \text{s.t.} & A_2^k\Delta X + F_2(X^k) = 0, \\ & A_3^k\Delta X + F_3(X^k) \geq 0,\end{array} \tag{4.43}$$

where we used $\mu_{QP} = \mu_{QP}^{(1)}$ and

$$\begin{aligned}\nu_{QP} &= \mu_{QP}^{(3)} - \mu_{QP}^{(2)} \\ &= \begin{cases}(F_1(X^k) + A_1^k\Delta X)_i, & \text{if } |(F_1(X^k) + A_1^k\Delta X)_i| \leq \gamma, \\ \operatorname{sign}\big((F_1(X^k) + A_1^k\Delta X)_i\big)\cdot\gamma, & \text{if } |(F_1(X^k) + A_1^k\Delta X)_i| > \gamma.\end{cases}\end{aligned}$$

We use again the following index sets for brevity of notation,

$$I_{F_1}^{\leq\gamma}(X) = \{i \in \{1,\ldots,m_1\} \mid |F_{1i}(X)| \leq \gamma\}, \tag{4.44a}$$

$$I_{F_1}^{\not\leq\gamma}(X) = \{i \in \{1,\ldots,m_1\} \mid |F_{1i}(X)| > \gamma\}. \tag{4.44b}$$

We will refer to $I_{F_1}^{\leq\gamma}(X)$ as the γ-*active set* at X. For ease of notation, in the following we split the set $I_{F_1}^{\not\leq\gamma}(X)$ further into the two sets

$$I_{F_1}^{<-\gamma}(X) = \{i \in \{1,\dots,m_1\} \mid F_{1i}(X) < -\gamma\}, \tag{4.45a}$$
$$I_{F_1}^{>\gamma}(X) = \{i \in \{1,\dots,m_1\} \mid F_{1i}(X) > \gamma\}, \tag{4.45b}$$

which are both contained in $I_{F_1}^{\not\leq\gamma}(X)$ such that $I_{F_3}^{<-\gamma}(X) \cup I_{F_3}^{>\gamma}(X) = I_{F_3}^{\not\leq\gamma}(X)$.

Analogous notations are used for the partitioning of the objective function of the inexact Huber optimization problem (4.43),

$$I_{F_1}^{\leq\gamma}(\Delta X) = \{i \in \{1,\dots,m_1\} \mid |(F_1(X) + A_1\Delta X)_i| \leq \gamma\}, \tag{4.46a}$$
$$I_{F_1}^{\not\leq\gamma}(\Delta X) = \{i \in \{1,\dots,m_1\} \mid |(F_1(X) + A_1\Delta X)_i| > \gamma\}, \tag{4.46b}$$
$$I_{F_1}^{<-\gamma}(\Delta X) = \{i \in \{1,\dots,m_1\} \mid (F_1(X) + A_1\Delta X)_i < -\gamma\}, \tag{4.46c}$$
$$I_{F_1}^{>\gamma}(\Delta X) = \{i \in \{1,\dots,m_1\} \mid (F_1(X) + A_1\Delta X)_i > \gamma\}. \tag{4.46d}$$

For the inequality constraints the same notations (4.13d) and (4.13h) are used as introduced in section 4.2.1 for the LAD minimization problem.

Furthermore, we define the following matrices,

$$D = \operatorname{diag}(d_{ii}), \tag{4.47a}$$
$$d_{ii} = \begin{cases} 1, & \text{if } |F_{1i}(X)| \leq \gamma, \\ 0, & \text{if } |F_{1i}(X)| > \gamma, \end{cases}$$
$$\bar{D} = \operatorname{diag}(\bar{d}_{ii}), \tag{4.47b}$$
$$\bar{d}_{ii} = \begin{cases} \operatorname{sign}(F_{1i}(X)), & \text{if } |F_{1i}(X)| > \gamma, \\ 0, & \text{if } |F_{1i}(X)| \leq \gamma, \end{cases}$$
$$D^{AP} = \operatorname{diag}(d_{ii}^{AP}), \tag{4.47c}$$
$$d_{ii}^{AP} = \begin{cases} 1, & \text{if } |(F_1(X) + A_1\Delta X^{AP})_i| \leq \gamma, \\ 0, & \text{if } |(F_1(X) + A_1\Delta X^{AP})_i| > \gamma, \end{cases}$$
$$\bar{D}^{AP} = \operatorname{diag}(\bar{d}_{ii}^{AP}), \tag{4.47d}$$
$$\bar{d}_{ii}^{AP} = \begin{cases} \operatorname{sign}\big((F_1(X) + A_1\Delta X^{AP})_i\big), & \text{if } |(F_1(X) + A_1\Delta X^{AP})_i| > \gamma, \\ 0, & \text{if } |(F_1(X) + A_1\Delta X^{AP})_i| \leq \gamma. \end{cases}$$

And similarly to the notation used for inexact LAD minimization, we use the notations $\tilde{A}_1 = A_1^T D^{AP} A_1$ and $\tilde{A}_3 = (A_{3i})_{i \in I_{F_3}(\Delta X)}$, where A_{3i} is the ith row of the matrix A_3, respectively.

Note here that for parameter estimation with Huber's M-estimator the multiplier ν for the objective function is completely defined by the value of the objective function. Therefore it is also not an optimization variable in the ineaxct Huber optimization problem and can be computed after solving the Gauss-Newton problem to define the problem for the next iteration.

The inexact Gauss-Newton method with approximated derivatives is described in algorithm 4.3.1.

Algorithm 4.3.1 Local inexact GN method for Huber problems

(S1) Choose $(X^0, \lambda^0, \mu^0) \in \mathbb{R}^n \times \mathbb{R}^{m_2} \times \mathbb{R}^{m_3}$, $A_i^0, i = 1,2,3$, such that $\tilde{A}_1^{\,0} = A_1^{0T} D^{AP,0} A_1^0$ is positive semidefinite and the matrix

$$\tilde{A}^0 = \begin{pmatrix} \tilde{A}_1^{\,0} & -A_2^{0T} & -\tilde{A}_3^{\,0^T} \\ A_2^0 & 0 & 0 \\ \tilde{A}_3^{\,0} & 0 & 0 \end{pmatrix}$$

is regular with $\tilde{A}_3^{\,0} = (\tilde{A}_{3i}^{\,0})_{i \in I_{F_3}(\Delta X^0)}$, compute

$$\nu_i^0 = \begin{cases} F_{1i}(X^0), & \text{if } |F_{1i}(X^0)| \le \gamma, \\ \operatorname{sign}(F_{1i}(X^0)) \cdot \gamma, & \text{if } |F_{1i}(X^0)| > \gamma, \end{cases}$$

and set $k := 0$.

(S2) If $(X^k, \nu^k, \lambda^k, \mu^k)$ is an FJ point of (4.37): STOP.

(S3) Compute an FJ point $(\Delta X^{AP}, \nu^{AP}, \lambda^{AP}, \mu^{AP}) \in \mathbb{R}^n \times \mathbb{R}^{m_2} \times \mathbb{R}^{m_3}$ of the inexactly linearized problem

$$\begin{aligned} \min_{\Delta X} \quad & \sum_{i=1}^{m_1} \rho_\gamma\big((F_1(X^k) + A_1^k \Delta X)_i\big) \\ & - \Big(\nu^{kT}(A_1^k - J_1(X^k)) - \lambda^{kT}(A_2^k - J_2(X^k)) - \mu^{kT}(A_3^k - J_3(X^k))\Big) \Delta X \\ \text{s.t.} \quad & A_2^k \Delta X + F_2(X^k) = 0, \\ & A_3^k \Delta X + F_3(X^k) \ge 0, \end{aligned}$$

and set $X^{k+1} := X^k + \Delta X^{AP}$, $\nu^{k+1} := \nu^k + (\nu^{AP} - \nu^k)$, $\lambda^{k+1} := \lambda^k + (\lambda^{AP} - \lambda^k)$, and $\mu^{k+1} := \mu^k + (\mu^{AP} - \mu^k)$.

(S4) If a convergence test fails, choose $A_i^{k+1}, i = 1,2,3$, such that the matrix $\tilde{A}_1^{\,k+1} = A_1^{k+1^T} D^{AP,k+1} A_1^{k+1}$ is positive semidefinite and the matrix

$$\tilde{A}^{k+1} = \begin{pmatrix} \tilde{A}_1^{\,k+1} & -(A_2^{k+1})^T & -(\tilde{A}_3^{\,k+1})^T \\ A_2^{k+1} & 0 & 0 \\ \tilde{A}_3^{\,k+1} & 0 & 0 \end{pmatrix}$$

is regular with $\tilde{A}_3^{\,k+1} = (\tilde{A}_{3_i}^{\,k+1})_{i \in I_{F_3}(\Delta X^{k+1})}$, set $k \leftarrow k+1$ and go to (S2).

Note that the correction term in the objective function of the inexact Huber optimization problem (4.43),

$$m^{k^T}\Delta X = -\left(\nu^{k^T}(A_1^k - J_1(X^k)) - \lambda^{k^T}(A_2^k - J_2(X^k)) - \mu^{k^T}(A_3^k - J_3(X^k))\right)\Delta X, \tag{4.48}$$

can be written equivalently with

$$m^k = \nabla_X \mathcal{L}(X^k, \nu^k, \lambda^k, \mu^k) - A_1^{k^T}\nu^k + A_2^{k^T}\lambda^k + A_3^{k^T}\mu^k. \tag{4.49}$$

We will show in the following section that there exists a neighborhood of the local solution X^* of the nonlinear problem in which the active set of the inequality constraints and the partitioning of the objective function are the same for the nonlinear and the inexact Huber optimization problem if only the approximations of the Jacobians are chosen close enough to the exact Jacobians. Then it is quite straightforward to prove local convergence of algorithm 4.3.1 within this neighborhood.

4.3.2. Stability of the active set and local convergence

All necessary preliminary studies and definitions were already made in the previous sections. So we can start directly with the theorem on the stability of the active sets.

Theorem 4.3.1 (Stability of the active sets of the problem with approximated derivatives near a solution of the Huber optimization problem)**.** *Let $(X^*, \nu^*, \lambda^*, \mu^*)$ be an FJ point of the Huber minimization problem (4.37). The norm $\|\cdot\|$ be the Euclidean vector norm, and $\|\cdot\|_F$ be the Frobenius matrix norm. Further let X^* be a regular point and let the strict complementarity conditions (3.64) hold at $(X^*, \nu^*, \lambda^*, \mu^*)$. Further, let the following conditions hold for given constants $\delta_1, \delta_2 > 0$:*

- $\|A_1\|_F, \|A_2\|_F, \|A_3\|_F \leq \delta_1$,
- *the matrix*

$$\tilde{\mathcal{A}} = \begin{pmatrix} A_1^T D A_1 & -A_2^T & -\tilde{A_3}^T \\ A_2 & 0 & 0 \\ \tilde{A_3} & 0 & 0 \end{pmatrix}, \tag{4.50}$$

  is invertible, where D is the matrix defined in (4.47a) for the index set $I_{F_1}(X^*)$, A_1 such that $A_1^T D A_1$ is a positive semidefinite matrix, and $\tilde{A_3} = (A_{3i})_{i \in I_{F_3}(X^*)}$ with A_{3i} being the i-th row of A_3, and

- $\|\tilde{\mathcal{A}}\|_F \leq \delta_2$,

Then there exists a neighborhood $\mathcal{N}$ of $(X^*, \nu^*, \lambda^*, \mu^*)$ and a constant $\kappa > 0$ such that for all $(X, \nu, \lambda, \mu) \in \mathcal{N}$ the problem with approximated derivatives

$$\begin{aligned} \min_{\Delta X} \quad & \sum_{i=1}^{m_1} \rho_\gamma\big((F_1(X) + A_1\Delta X)_i\big) \\ & - \big(\nu^T(A_1 - J_1(X)) - \lambda^T(A_2 - J_2(X)) - \mu^T(A_3 - J_3(X))\big)\Delta X \\ s.t. \quad & A_2\Delta X + F_2(X) = 0, \\ & A_3\Delta X + F_3(X) \geq 0, \end{aligned} \tag{4.51}$$

has a unique solution $(\Delta X^{AP}, \nu^{AP}, \lambda^{AP}, \mu^{AP})$ *which satisfies the inequality*

$$\|(\Delta X^{AP}, \nu^{AP}, \lambda^{AP}, \mu^{AP}) - (0, \nu^*, \lambda^*, \mu^*)\| \leq \kappa \|(X^*, \nu^*, \lambda^*, \mu^*) - (X, \nu, \lambda, \mu)\|. \tag{4.52}$$

Furthermore, the problem with approximated derivatives (4.51) has the same active set and the same partitioning of the objective function as the Huber solution X^*.

Proof. We define $\Gamma := \gamma \mathbb{1} \in \mathbb{R}^{m_1}$ to be the $\mathbb{R}^{m_1}$-vector with value γ in each component. Any solution ΔX^{AP} with the corresponding multipliers $\nu^{AP}, \lambda^{AP}, \mu^{AP}$ of (4.51) have to satisfy the FJ conditions with $\mu_0 = 1$, i.e. the KKT conditions

$$\begin{aligned}
&\nabla_X \mathcal{L}(X, \nu, \lambda, \mu) - A_2^T(\lambda^{AP} - \lambda) - A_3^T(\mu^{AP} - \mu) + \\
&\underbrace{A_1^T D^{AP} A_1 \Delta X^{AP} + A_1^T D^{AP} F_1(X) + A_1^T \bar{D}^{AP} \Gamma}_{= A_1^T \nu^{AP}} \underbrace{- A_1^T D F_1(X) - A_1^T \bar{D} \Gamma}_{= -A_1^T \nu} = 0, && \text{(4.53a)} \\
&F_2(X) + A_2 \Delta X^{AP} = 0, && \text{(4.53b)} \\
&F_3(X) + A_3 \Delta X^{AP} \geq 0, && \text{(4.53c)} \\
&\mu^{AP} \geq 0, && \text{(4.53d)} \\
&i = 1, \ldots, m_3: \quad \mu_i^{AP} \big(F_3(X) + A_3 \Delta X^{AP}\big)_i = 0, && \text{(4.53e)}
\end{aligned}$$

with the multiplier ν^{AP} being given by

$$\nu_i^{AP} = \begin{cases} \big(F_1(X) + A_1 \Delta X^{AP}\big)_i, & \big|\big(F_1(X) + A_1 \Delta X^{AP}\big)_i\big| \leq \gamma, \\ \operatorname{sign}\big[\big(F_1(X) + A_1 \Delta X^{AP}\big)_i\big] \cdot \gamma, & \big|\big(F_1(X) + A_1 \Delta X^{AP}\big)_i\big| > \gamma, \end{cases} \tag{4.53f}$$

for $i = 1, \ldots, m_1$. Remember that $\nabla_X \mathcal{L}(X, \nu, \lambda, \mu) = J_1(X)^T \nu - J_2(X)^T \lambda - J_3(X)^T \mu$. For the optimal point $(X, \nu, \lambda, \mu) = (X^*, \nu^*, \lambda^*, \mu^*)$ we obviously get an FJ point of the inexact Huber optimization problem (4.51) by setting $(\Delta X^{AP}, \nu^{AP}, \lambda^{AP}, \mu^{AP}) = (0, \nu^*, \lambda^*, \mu^*)$. The active set of inequality constraints and the γ-active set as well as the validity of the strict complementarity conditions are clearly inherited from the Huber solution. The constraint qualification is satisfied if $\tilde{A}_2 = \begin{pmatrix} A_2 \\ \tilde{A}_3 \end{pmatrix}$ has full rank and the uniqueness of ΔX^{AP} and the Lagrange multipliers follows from the invertibility of $\tilde{\mathcal{A}}$.

We can also derive from this fact the conclusion that the matrix $A_1^T D A_1$ is positive definite on the null space of the matrix $\tilde{A}_2$. For a proof by contradiction take a vector $z \in \mathbb{R}^n$, $z \neq 0$ with $\tilde{A}_2 z = 0$ and assume that $z^T A_1^T D A_1 z = 0$. Having chosen A_1 such that $\tilde{A}_1 := A_1^T D A_1$ is positive semidefinite, we can write $\tilde{A}_1 = (\tilde{A}_1^{1/2})^T \tilde{A}_1^{1/2}$. Obviously, z must then lie in the null space of $\tilde{A}_1^{1/2}$ and therefore in the null space of $\tilde{A}_1$. For the vector $\tilde{z}^T = (z^T, 0^T) \in \mathbb{R}^{n+m_2+m_3^0}$, we then get $\tilde{\mathcal{A}}\tilde{z} = 0$ contradicting the assumed invertibility of $\tilde{\mathcal{A}}$. Thus it must hold that $z^T \tilde{A}_1 z \neq 0$. As we have $\tilde{A}_1$ positive semidefinite, we get $z^T \tilde{A}_1 z > 0$.

Applying lemma 3.1.3, we can conclude that the FJ point is a strict local minimizer

of (4.51).
Obviously, due to $\Delta X^* = 0$ we always have $I^{\leq\gamma}_{\bar{F}_1}(\Delta X^*) = I^{\leq\gamma}_{\bar{F}_1}(X^*)$, $I^{<-\gamma}_{F_1}(\Delta X^*) = I^{<-\gamma}_{F_1}(X^*)$, $I^{>\gamma}_{F_1}(\Delta X^*) = I^{>\gamma}_{F_1}(X^*)$, and $I_{F_3}(\Delta X^*) = I_{F_3}(X^*)$. We will now show that there exists a neighborhood $\mathcal{N}^1_X$ of X^* in which the γ-active set and its complement stay constant. In this neighborhood we then have $I^{\leq\gamma}_{\bar{F}_1}(\Delta X^{AP}) = I^{\leq\gamma}_{\bar{F}_1}(X^*)$, $I^{<-\gamma}_{F_1}(\Delta X^{AP}) = I^{<-\gamma}_{F_1}(X^*)$, $I^{>\gamma}_{F_1}(\Delta X^{AP}) = I^{>\gamma}_{F_1}(X^*)$ and thus $D = D^* = D^{AP*} = D^{AP}$ and $\bar{D} = \bar{D}^* = \bar{D}^{AP*} = \bar{D}^{AP}$.
Define $\bar{F}_1(X) = (F_{1i}(X))_{i\in I^{\leq\gamma}_{\bar{F}_1}(X^*)}$, $\bar{A}_1 = (A_{1i})_{i\in I^{\leq\gamma}_{\bar{F}_1}(X^*)}$. We can make the estimation

$$\begin{aligned}|(F_1(X) + A_1\Delta X - F_1(X^*))_i| &\leq |F_{1i}(X) - F_{1i}(X^*)| + |(A_1\Delta X)_i| \\ &\leq L_1|X^*_i - X_i| + \delta_1|\Delta X_i| \\ &\leq (L_1+\delta_1)|X^*_i - X_i|,\end{aligned} \tag{4.54}$$

which holds for all $i \in I^{\leq\gamma}_{\bar{F}_1}(X^*)$. Denote $\bar{f} := \min_{i\in I^{\leq\gamma}_{\bar{F}_1}(X^*)}\{\gamma - |F_{1i}(X^*)|\} > 0$ and choose $0 < \bar{f}_1 < \bar{f}$. By choosing an appropriate neighborhood $\mathcal{N}^1_{\leq\gamma}$ of X^*, we can guarantee that $(L_1+\delta_1)|X^*_i - X_i| \leq \bar{f}_1$ and thus $|(\bar{F}_1(X) + \bar{A}_1\Delta X - \bar{F}_1(X^*))_i| \leq \bar{f}_1$ for all $X \in \mathcal{N}^1_{\leq\gamma}$. It follows that

$$\begin{aligned}\gamma - |(F_1(X) + A_1\Delta X)_i| &= \gamma - |F_{1i}(X^*) + [(F_1(X) + A_1\Delta X - F_1(X^*))_i]| \\ &\geq \gamma - |F_{1i}(X^*)| - |(F_1(X) + A_1\Delta X - F_1(X^*))_i| \\ &\geq \bar{f} - \bar{f}_1 > 0 \quad \text{for all } i \in I^{\leq\gamma}_{\bar{F}_1}(X^*),\end{aligned}$$

i.e. $|(F_1(X)+A_1\Delta X)_i| < \gamma$ for all $X \in \mathcal{N}^1_{\leq\gamma}$ and all $i \in I^{\leq\gamma}_{\bar{F}_1}(X^*)$, i.e. $I^{\leq\gamma}_{\bar{F}_1}(\Delta X^{AP}) \supseteq I^{\leq\gamma}_{\bar{F}_1}(X^*)$.
Define $\check{F}_1(X) = (F_{1i}(X))_{i\in I^{<-\gamma}_{F_1}(X^*)}$, $\check{A}_1 = (A_{1i})_{i\in I^{<-\gamma}_{F_1}(X^*)}$. Estimate (4.54) also holds for all $i \in I^{<-\gamma}_{F_1}(X^*)$. Denote $\check{f} := \min_{i\in I^{<-\gamma}_{F_1}(X^*)}\{-F_{1i}(X^*) - \gamma\} > 0$ and choose $0 < \check{f}_1 < \check{f}$. By choosing an appropriate neighborhood $\mathcal{N}^1_{<-\gamma}$ of X^*, we can guarantee that $(L_1+\delta_1)|X^*_i - X_i| \leq \check{f}_1$ and thus $|(F_1(X) + A_1\Delta X - F_1(X^*))_i| \leq \check{f}_1$ for all $X \in \mathcal{N}^1_{<-\gamma}$. It follows that

$$\begin{aligned}-(F_1(X) + A_1\Delta X)_i - \gamma &= -\big(F_{1i}(X^*) + [(F_1(X) + A_1\Delta X - F_1(X^*))_i]\big) - \gamma \\ &\geq \check{f} - \check{f}_1 > 0 \quad \text{for all } i \in I^{<-\gamma}_{F_1}(X^*),\end{aligned}$$

i.e. $(F_1(X) + A_1\Delta X)_i < -\gamma$ for all $X \in \mathcal{N}^1_{<-\gamma}$ and all $i \in I^{<-\gamma}_{F_1}(X^*)$, and thus $I^{<-\gamma}_{F_1}(\Delta X^{AP}) \supseteq I^{<-\gamma}_{F_1}(X^*)$.
Define $\hat{F}_1(X) = (F_{1i}(X))_{i\in I^{>\gamma}_{F_1}(X^*)}$, $\hat{A}_1 = (A_{1i})_{i\in I^{>\gamma}_{F_1}(X^*)}$. Estimate (4.54) also holds for all $i \in I^{>\gamma}_{F_1}(X^*)$. Denote $\hat{f} := \min_{i\in I^{>\gamma}_{F_1}(X^*)}\{F_{1i}(X^*) - \gamma\} > 0$ and choose $0 < \hat{f}_1 < \hat{f}$. By choosing an appropriate neighborhood $\mathcal{N}^1_{>\gamma}$ of X^*, we can guarantee that

$(L_1+\delta_1)|X_i^*-X_i| \leq \hat{f}_1$ and thus $|(F_1(X)+A_1\Delta X-F_1(X^*))_i| \leq \hat{f}_1$ for all $X \in \mathcal{N}^1_{>\gamma}$. It follows that

$$\begin{aligned}(F_1(X)+A_1\Delta X)_i-\gamma &= \big(F_{1i}(X^*)+[(F_1(X)+A_1\Delta X-F_1(X^*))_i]\big)-\gamma \\ &\geq \hat{f}-\hat{f}_1>0 \quad \text{for all } i \in I^{>\gamma}_{F_1}(X^*),\end{aligned}$$

i.e. $(F_1(X)+A_1\Delta X)_i>\gamma$ for all $X \in \mathcal{N}^1_{>\gamma}$ and all $i \in I^{>\gamma}_{F_1}(X^*)$, i.e. $I^{>\gamma}_{F_1}(\Delta X^{AP}) \supseteq I^{>\gamma}_{F_1}(X^*)$.
Together we obtain that for all $X \in \mathcal{N}^1_X := \mathcal{N}^1_{\leq\gamma} \cap \mathcal{N}^1_{<-\gamma} \cap \mathcal{N}^1_{>\gamma}$ the components of the objective function do not cross the $\pm\gamma$ boundaries and so the sets $I^{\leq\gamma}_{F_1}(X^*)$, $I^{<-\gamma}_{F_1}(X^*)$, and $I^{>\gamma}_{F_1}(X^*)$ also define the partitioning of the objective function in the inexactly linearized optimization problem within $\mathcal{N}^1_X$.
For the inequality constraints, define $\check{F}_3(X) = (F_{3i}(X))_{i\notin I_{F_3}(X^*)}$ and the matrix $\check{A}_3 = (A_{3i})_{i\notin I_{F_3}(X^*)}$. We can make the estimation

$$\begin{aligned}|(F_3(X)+A_3\Delta X-F_3(X^*))_i| &\leq |F_{3i}(X)-F_{3i}(X^*)|+|(A_3\Delta X)_i| \\ &\leq L_3|X_i^*-X_i|+\delta_1|\Delta X_i| \\ &\leq (L_3+\delta_1)|X_i^*-X_i|,\end{aligned}$$

which holds for all $i \notin I_{F_3}(X^*)$. Denote $f := \min_{i\notin I_{F_3}(X^*)}\{F_{3i}(X^*)\}>0$ and choose $0<f_3<f$. By choosing an appropriate neighborhood $\mathcal{N}^3_X$ of X^*, we can guarantee that $(L_3+\delta_1)|X_i^*-X_i| \leq f_3$ and thus $|(F_3(X)+A_3\Delta X-F_3(X^*))_i| \leq f_3$ for all $X \in \mathcal{N}^3_X$. It follows that

$$\begin{aligned}(F_3(X)+A_3\Delta X)_i &= F_{3i}(X^*)+[(F_3(X)+A_3\Delta X-F_3(X^*))_i] \\ &\geq f-f_3>0 \quad \text{for all } i \notin I_{F_3}(X^*),\end{aligned}$$

i.e. $\check{F}_3(X)+\check{A}_3\Delta X>0$, i.e. $I_{F_3}(\Delta X^{AP}) \subset I_{F_3}(X^*)$.
If we now require $I_{F_3}(\Delta X^{AP}) = I_{F_3}(X^*)$, i.e. we try to solve the KKT conditions (4.53) with the additional conditions

$$(F_3(X)+A_3\Delta X)_i = 0 \;\forall i \in I_{F_3}(X^*), \tag{4.55}$$

we at once get for $X \in \hat{\mathcal{N}}^3_X \subset \mathcal{N}^3_X$ that the active components of the inequality constraints are the same as at the solution of problem (4.51).
So we have shown that there exists a neighborhood $\mathcal{N}_X := \mathcal{N}^1_X \cap \hat{\mathcal{N}}^3_X$ in which the partitioning of the objective function and the inequality constraints for the inexactly linearized Huber optimization problem is the same as for the nonlinear problem at the optimum X^*.
Now let (4.51) be defined in the point (X,ν,λ,μ) with $X \in \mathcal{N}_X$ and define

$$m(X,\nu,\lambda,\mu) := \nabla_X\mathcal{L}(X,\nu,\lambda,\mu)-A_1^T\nu+A_2^T\lambda+A_3^T\mu.$$

Define $\tilde{F}_3(X) := (F_{3i}(X))_{i \in I_{F_3}(X^*)}$ and the function $\tilde{F}_1(X)$ with

$$\tilde{F}_1(X) = \begin{cases} F_1^{\leq\gamma}(X), \\ \tilde{F}_1^{\not\leq\gamma}(X), \end{cases} \quad \text{with} \quad \begin{aligned} F_1^{\leq\gamma}(X) &= \big(F_{1i}(X)\big)_{i \in I_{F_1}^{\leq\gamma}(X^*)}, \\ \tilde{F}_1^{\not\leq\gamma}(X) &= \big(\nu_i\big)_{i \in I_{F_1}^{\not\leq\gamma}(X^*)} = (\pm\gamma). \end{aligned} \tag{4.56}$$

From the invertibility of $\tilde{\mathcal{A}}$ follows that the solution of the linear system

$$\tilde{\mathcal{A}} \begin{pmatrix} s \\ b \\ c \end{pmatrix} = - \begin{pmatrix} m(X, \nu, \lambda, \mu) + A_1^T \tilde{F}_1(X) \\ F_2(X) \\ \tilde{F}_3(X) \end{pmatrix}, \tag{4.57}$$

is unique. We further have

$$\begin{pmatrix} s - 0 \\ b - \lambda^* \\ c - \tilde{\mu}^* \end{pmatrix} = \tilde{\mathcal{A}}^{-1} \begin{pmatrix} m(X^*, \nu^*, \lambda^*, \mu^*) + A_1^T \tilde{F}_1(X^*) - m(X, \nu, \lambda, \mu) - A_1^T \tilde{F}_1(X) \\ F_2(X^*) - F_2(X) \\ \tilde{F}_3(X^*) - \tilde{F}_3(X) \end{pmatrix}.$$

As we have assumed F_1, F_2, F_3 to be at least twice continuously differentiable in X, the gradient of the Lagrangian, $\nabla_X \mathcal{L}(X, \nu, \lambda, \mu)$, is continuously differentiable with respect to X and therefore locally Lipschitz continuous in (X, ν, λ, μ) with a Lipschitz constant L. Further let L_j, $j = 1, 2, 3$, be the Lipschitz constants of $F_j(X)$, $j = 1, 2, 3$, respectively. Consider

$$\begin{aligned} &\|m(X^*, \nu^*, \lambda^*, \mu^*) + A_1^T \tilde{F}_1(X^*) - m(X, \nu, \lambda, \mu) - A_1^T \tilde{F}_1(X)\| \\ &\qquad \leq \|m(X^*, \nu^*, \lambda^*, \mu^*) - m(X, \nu, \lambda, \mu)\| + \|A_1^T \tilde{F}_1(X^*) - A_1^T \tilde{F}_1(X)\|. \end{aligned} \tag{4.58}$$

We will consider the two parts of the second line of (4.58) separately and we start with the second summand. As we have shown that in the neighborhood $\mathcal{N}_X^1$ the partitioning of the objective function is maintained, this term reduces to

$$\begin{aligned} &\|A_1^T \tilde{F}_1(X^*) - A_1^T \tilde{F}_1(X)\| \\ &= \|A_1^T D F_1(X^*) + A_1^T \bar{D} \Gamma - A_1^T D^{AP} F_1(X) - A_1^T \bar{D}^{AP} \Gamma\| \\ &= \|A_1^T D (F_1(X^*) - F_1(X))\| \\ &\leq \delta_1 \|D\|_F L_1 \|X^* - X\| \\ &\leq \delta_1 \mathcal{D} L_1 \|(X^*, \nu^*, \lambda^*, \mu^*) - (X, \nu, \lambda, \mu)\|, \end{aligned} \tag{4.59}$$

where $\mathcal{D} = \|D\|_F$ equals the number of indices in the γ-active set $I_{F_1}^{\leq\gamma}(X^*)$ by definition of the Frobenius matrix norm. Let us now look at the first summand in (4.58). We have

$$\begin{aligned} &\|m(X^*, \nu^*, \lambda^*, \mu^*) - m(X, \nu, \lambda, \mu)\| \\ &= \|\nabla_X \mathcal{L}(X^*, \nu^*, \lambda^*, \mu^*) - \nabla_X \mathcal{L}(X, \nu, \lambda, \mu) \\ &\qquad - A_1^T (\nu^* - \nu) + A_2^T (\lambda^* - \lambda) + A_3^T (\mu^* - \mu)\|, \end{aligned}$$

where $A_1^T(\nu^* - \nu) = A_1^T DF_1(X^*) + A_1^T \bar{D}\Gamma - A_1^T DF_1(X) - A_1^T \bar{D}\Gamma = A_1^T D(F_1(X^*) - F_1(X))$ in $\mathcal{N}_X^1$. For the whole term we can therefore make the estimate

$$\begin{aligned}
&\|m(X^*, \nu^*, \lambda^*, \mu^*) - m(X, \nu, \lambda, \mu)\| \\
&\quad \leq \|\nabla_X \mathcal{L}(X^*, \nu^*, \lambda^*, \mu^*) - \nabla_X \mathcal{L}(X, \nu, \lambda, \mu)\| \\
&\qquad + \|A_2^T(\lambda^* - \lambda)\| + \|A_3^T(\mu^* - \mu)\| + \|A_1^T D(F_1(X^*) - F_1(X))\| \\
&\quad \leq L\|(X^*, \nu^*, \lambda^*, \mu^*) - (X, \nu, \lambda, \mu)\| + \delta_1\|\lambda^* - \lambda\| + \delta_1\|\mu^* - \mu\| \\
&\qquad + \delta_1 \mathcal{D} L_1 \|X^* - X\| \\
&\quad \leq \big(L + \delta_1(2 + \mathcal{D}L_1)\big)\|(X^*, \nu^*, \lambda^*, \mu^*) - (X, \nu, \lambda, \mu)\|.
\end{aligned} \tag{4.60}$$

In both (4.59) and (4.60) we have used $\|A\|_F = \|A^T\|_F$. And from these two estimates follows the norm estimate

$$\left\| \begin{pmatrix} s - 0 \\ b - \lambda^* \\ c - \tilde{\mu}^* \end{pmatrix} \right\| \leq \delta_2(L + L_2 + L_3 + 2\delta_1(1 + \mathcal{D}L_1)) \left\| \begin{pmatrix} X^* - X \\ \nu^* - \nu \\ \lambda^* - \lambda \\ \mu^* - \mu \end{pmatrix} \right\|. \tag{4.61}$$

For ease of notation we define $\theta := \delta_2(L + L_2 + L_3 + 2\delta_1(1 + \mathcal{D}L_1))$. Now let us recall the KKT conditions (4.53) of the problem with approximated derivatives plus condition (4.55) with the solution $(s^T, b^T, c^T)^T$ of (4.57):

- Set $\Delta X^{AP} = s$. Then (4.53b) follows directly from (4.57).
- Using estimate (4.61) we can make the estimation

$$\begin{aligned}
\|F_1(X) + A_1 \Delta X - F_1(X^*)\| &\leq \|F_1(X) - F_1(X^*)\| - \|A_1 \Delta X\| \\
&\leq L_1\|X^* - X\| + \delta_1\|\Delta X\| \\
&\leq (L_1 + \delta_1\theta)\|(X^*, \nu^*, \lambda^*, \mu^*) - (X, \nu, \lambda, \mu)\|.
\end{aligned} \tag{4.62}$$

  Analogically to the explanations above, where we derived the neighborhood $\mathcal{N}_X^1$ of X^* in which the γ-active set stays constant, we can now find a neighborhood $\mathcal{N}^1 = \mathcal{N}_X^1 \times \mathcal{N}_\nu^1 \times \mathcal{N}_\lambda^1 \times \mathcal{N}_\mu^1$ of $(X^*, \nu^*, \lambda^*, \mu^*)$ in which the same conclusions hold.

- Analogically, we get a neighborhood $\check{\mathcal{N}}^3 = \check{\mathcal{N}}_X^3 \times \check{\mathcal{N}}_\nu^3 \times \check{\mathcal{N}}_\lambda^3 \times \check{\mathcal{N}}_\mu^3$ of $(X^*, \nu^*, \lambda^*, \mu^*)$ in which condition (4.53c) holds with the same active components of the inequality constraints as in the solution of problem (4.37) defined in the FJ point $(X^*, \nu^*, \lambda^*, \mu^*)$.
- Without loss of generality let $I_{F_3}(X^*) = \{1, \ldots, m_3^0\}$ and set $\mu^{AP} = (c^T, 0^T)^T$. As $\tilde{\mu}^* > 0$, it follows from inequality (4.61) that we can again find a neighborhood $\bar{\mathcal{N}}^3 = \bar{\mathcal{N}}_X^3 \times \bar{\mathcal{N}}_\nu^3 \times \bar{\mathcal{N}}_\lambda^3 \times \bar{\mathcal{N}}_\mu^3$ of $(X^*, \nu^*, \lambda^*, \mu^*)$ such that $c > 0$ for all $(X, \nu, \lambda, \mu) \in \bar{\mathcal{N}}^3$. Thus we get for all $(X, \nu, \lambda, \mu) \in \mathcal{N}^3 = \check{\mathcal{N}}^3 \cap \bar{\mathcal{N}}^3$ the validity of conditions (4.53d) and (4.53e) with strict complementarity.

- With $\lambda^{AP} = b$ the Lagrangian gradient (4.53a) of the problem with approximated derivatives vanishes, as can also be seen from (4.57).

Thus for all $(X, \nu, \lambda, \mu) \in \mathcal{N} := \mathcal{N}^3 \cap \mathcal{N}^1$ we get an FJ point of problem (4.51) in $(\Delta X^{AP}, \nu^{AP}, \lambda^{AP}, \mu^{AP})$ with

$$\begin{aligned} \Delta X^{AP} &= s, \\ \nu^{AP} &\text{ given by } (4.53f), \\ \lambda^{AP} &= b, \\ \mu^{AP} &= (c^T, 0^T)^T, \end{aligned}$$

which satisfies the necessary conditions with the active set $I_{F_3}(X^*)$ and the γ-active set $I_{F_1}^{\leqslant\gamma}(X^*)$ and its complement $I_{F_1}^{\not\leqslant\gamma}(X^*)$. Estimate (4.52) follows from (4.61), appropriately supplemented with zeros, and (4.62) because

$$\begin{aligned} \left\| \begin{pmatrix} \Delta X^{AP} - 0 \\ \nu^{AP} - \nu^* \\ \lambda^{AP} - \lambda^* \\ \mu^{AP} - \mu^* \end{pmatrix} \right\| &\leq \left\| \begin{pmatrix} \Delta X^{Ap} - 0 \\ \lambda^{AP} - \lambda^* \\ \mu^{AP} - \mu^* \end{pmatrix} \right\| + \|\nu^{AP} - \nu^*\| \\ &\leq (\theta + L_1 + \delta_1\theta) \left\| \begin{pmatrix} X^* - X \\ \nu^* - \nu \\ \lambda^* - \lambda \\ \mu^* - \mu \end{pmatrix} \right\|. \end{aligned}$$

□

Due to the equivalence of all vector norms in a finite-dimensional normed vector space, we could have used any vector norm and corresponding matrix norm. With the choice of the Frobenius norm we followed the trick used by Wirsching [140] to be able to use the same bound for both a matrix and its transposed matrix. It was also along the lines of the proof in this work that we did our proof here which just needed a little expansion to our problem.

Like in the case of l_1 norm minimization, theorem 4.3.1 allows a rather arbitrary choice for the approximations A_1, A_2, A_3 of the Jacobians as long as they satisfy the conditions required by the theorem. In particular, no relation between the exact Jacobians and their approximations has to be assumed. This will change in the following where we will give requirements for the iterative process to converge locally to the solution of the Huber optimization problem.

Theorem 4.3.2 (Local convergence of the inexact Gauss-Newton method for the Huber problem with fixed active sets)**.** *Let $(X^*, \nu^*, \lambda^*, \mu^*)$ be an FJ point of the Huber optimization problem (4.37) and $\|\cdot\|$ be an arbitrary but fixed vector norm. Further assume that the following requirements are fulfilled:*

i) X^* is a regular point and strict complementarity holds at $(X^*, \nu^*, \lambda^*, \mu^*)$.

ii) (A_1^k), (A_2^k), *and* (A_3^k) *are uniformly bounded sequences of matrices such that the sequence* $\big((\tilde{\mathcal{A}}^k)^{-1}\big)$, *with* $\tilde{\mathcal{A}}^k$ *defined as in theorem 4.3.1, i.e. according to the active sets of the nonlinear problem at the optimum* X^*, *is also uniformly bounded and such that the matrices* ${A_1^k}^T D^{AP*} A_1^k$ *are positive semidefinite for diagonal matrices* $D^{AP*} = diag\,(d_{ii}^k)$ *defined according to (4.47a).*

iii) There exists a sequence $z^k = (X^k, \lambda^k, \mu^k)^T$ *generated according to the rule* $z^{k+1} = z^k + \Delta z^k$ *with* $\Delta z^k = (\Delta X^k, \lambda^{AP,k} - \lambda^k, \mu^{AP,k} - \mu^k)^T$ *where* ΔX^k *is a solution of the problem with approximated derivatives*

$$\begin{aligned} \min_{\Delta X} \quad & \sum_{i=1}^{m_1} \rho_\gamma\big((F_1(X^k) + A_1^k \Delta X)_i\big) \\ & - \Big({\nu^k}^T (A_1^k - J_1(X^k)) - {\lambda^k}^T (A_2^k - J_2(X^k)) - {\mu^k}^T (A_3^k - J_3(X^k))\Big)\,\Delta X \\ s.t. \quad & A_2 \Delta X + F_2(X^k) = 0, \\ & A_3 \Delta X + F_3(X^k) \geq 0, \end{aligned} \tag{4.63}$$

where $\nu^{AP,k}$, $\lambda^{AP,k}$, $\mu^{AP,k}$ *are the Lagrange multipliers in the solution of this problem with* $\nu^{AP,k}$ *being given by*

$$\nu_i^{AP,k} = \begin{cases} \big(F_1(X^k) + A_1^k \Delta X^k\big)_i, & |\big(F_1(X^k) + A_1^k \Delta X^k\big)_i| \leq \gamma, \\ sign\big((F_1(X^k) + A_1^k \Delta X^k)_i\big) \cdot \gamma, & |\big(F_1(X^k) + A_1^k \Delta X^k\big)_i| > \gamma. \end{cases}$$

iv) There exists a $\kappa < 1$ *such that for all* $k \in \mathbb{N}$ *and all* $t \in [0,1]$ *it can be guaranteed that*

$$\Big\|(\tilde{\mathcal{A}}^{k+1})^{-1}\big(\tilde{\mathcal{A}}^k - \nabla f(z^k + t\Delta z^k)\big)\Delta z^k\Big\| \leq \kappa \|\Delta z^k\|,$$

where $z^k = (X^k, \lambda^k, \mu^k)^T$ *and the function* f *is defined as*

$$f(z^k) = \begin{pmatrix} -m(z^k, \nu^k) - {A_1^k}^T \tilde{F}_1(X^k) \\ -F_2(X^k) \\ -\tilde{F}_3(X^k) \end{pmatrix},$$

where $m(z^k, \nu^k) := \nabla_X \mathcal{L}(X^k, \nu^k, \lambda^k, \mu^k) - {A_1^k}^T \nu^k + {A_2^k}^T \lambda^k + {A_3^k}^T \mu^k$, *the function* $\tilde{F}_1$ *is defined as in (4.56) in the proof of theorem 4.3.1 and* $\tilde{F}_3(X) = (F_{3i}(X))_{i \in F_3(X^*)}$.

Then there exists a neighborhood $\mathcal{N}$ *of* $(X^*, \nu^*, \lambda^*, \mu^*)$ *such that for all initial guesses* $(X^0, \nu^0, \lambda^0, \mu^0) \in \mathcal{N}$ *with*

$$\nu_i^0 = \begin{cases} F_{1i}(X^0), & |F_{1i}(X^0)| \leq \gamma, \\ sign\big(F_{1i}(X^0)\big) \cdot \gamma, & |F_{1i}(X^0)| > \gamma, \end{cases}$$

the sequence $(X^k, \nu^k, \lambda^k, \mu^k)$ *converges q-linearly to* $(X^*, \nu^*, \lambda^*, \mu^*)$ *with rate* κ *and the solution of the problem with approximated derivatives (4.63) has the same active set and the same partitioning of the objective function as the nonlinear problem at* X^*.

Proof. As the sequences (A_1^k), (A_2^k), (A_3^k), and $\left((\tilde{\mathcal{A}}^k)^{-1}\right)$ are uniformly bounded, we can find constants $\delta_1, \delta_2 > 0$, such that $\|A_1^k\|_F$, $\|A_2^k\|_F$, $\|A_3^k\|_F \leq \delta_1$ and $\|(\tilde{\mathcal{A}}^k)^{-1}\|_F \leq \delta_2$ for all $k \in \mathbb{N}$. Therefore we can apply theorem 4.3.1 on the stability of the active set. From this theorem follows the existence of a neighborhood $\bar{\mathcal{N}}$ of $(X^*, \nu^*, \lambda^*, \mu^*)$ such that for all $(X^k, \nu^k, \lambda^k, \mu^k) \in \bar{\mathcal{N}}$ and all $k \in \mathbb{N}$ the problem with approximated derivatives (4.63) has a unique solution and the same active set and the same γ-active set as the Huber solution. (Without loss of generality we can again assume that $I_{F_3}(X^*) = \{1, \ldots, m_3^0\}$.) It can easily be shown that the solution of (4.63) can not only be computed by means of (4.57) but also by solving the equivalent system

$$\tilde{\mathcal{A}}^k \begin{pmatrix} s^k \\ b^k \\ c^k \end{pmatrix} = f(z^k), \quad z^k = (X^k, \lambda^k, \mu^k)^T. \tag{4.64}$$

The equivalence of these two equations is a consequence of the equivalence of Newton's method and the SQP method.
The sequence being generated by choosing an initial guess

$$(z^0)^T := \left((X^0)^T, (\lambda^0)^T, (\tilde{\mu}^0)^T, 0^T\right) \in \bar{\mathcal{N}}$$

with corresponding ν^0 and applying the iteration $z^{k+1} = z^k + \Delta z^k$, $\nu^{k+1} = \nu^k + \Delta\nu^k$, with $\Delta z^k = (s^k, b^k, c^k, 0)^T$ and $\Delta\nu^k = \nu^{AP,k} - \nu^k$ is well defined as long as it can be guaranteed that the iterates remain in $\bar{\mathcal{N}}$. To show this as well as convergence of the sequence toward $(X^*, \nu^*, \lambda^*, \mu^*)^T$, we apply lemma 4.1.1. To do so, we have to show that the requirements of the lemma are satisfied:
First, define $\bar{z}^k$ as a permutation of z^k and of ν^k, i.e. $\bar{z}^k := (X^k, \nu^k, \lambda^k, \mu^k)$, and $\bar{\Delta z}^k$ accordingly as $\bar{\Delta z}^k := (s^k, \Delta\nu^k, b^k, c^k, 0)$. As the sequence (A_1^k) is uniformly bounded, it follows by submultiplicativity of the Frobenius norm that the same is true for the sequence of matrices $(A_1^{kT} D^* A_1^k)$ for diagonal matrices $D^* = \operatorname{diag} d_{ii}^*$ defined according to (4.47a) with the partitioning of the optimum X^*. And since also the sequences (A_2^k), (A_3^k) are uniformly bounded, it can easily be seen that also the sequence $(\tilde{\mathcal{A}}^k)$ is uniformly bounded. What remains to be done is the determination of the neighborhood $\mathcal{N}$ such that $\mathcal{B}\left(\bar{z}^0, \frac{\|\bar{\Delta z}^0\|}{1-\kappa}\right) \subset \bar{\mathcal{N}}$ for all $\bar{z}^0 \in \mathcal{N}$.
As $\bar{\mathcal{N}}$ is a neighborhood of $(X^*, \nu^*, \lambda^*, \mu^*)$, there exists an $\varepsilon > 0$ such that $\mathcal{B}(X^*, \varepsilon) \subset \bar{\mathcal{N}}$. As $f(z^*) = 0$, we can choose the first step $\Delta z^0 = -(\tilde{\mathcal{A}}^0)^{-1} f(z^0)$ arbitrarily small if the distance $\|z^* - z^0\|$ is small enough, due to the differentiability of f and the boundedness of $(\tilde{\mathcal{A}}^0)^{-1}$. Further, we can choose X^0 with $\|F_1(X^*) - F_1(X^0)\|$ equally small. Then $\|\bar{z}^* - \bar{z}^0\| \leq \|z^* - z^0\| + \|\nu^* - \nu^0\|$ will also be small. For a given $\varepsilon > 0$ we can thus always find a $\delta > 0$ such that $\|\bar{z}^* - \bar{z}^0\| + \frac{\|\bar{\Delta z}^0\|}{1-\kappa} \leq \varepsilon$ whenever $\|\bar{z}^* - \bar{z}^0\| \leq \delta$. Therefore it can be guaranteed for all $\bar{z}^0 \in \mathcal{N} := \mathcal{B}(\bar{z}^*, \delta)$ that $\mathcal{B}\left(\bar{z}^0, \frac{\|\bar{\Delta z}^0\|}{1-\kappa}\right) \subset \mathcal{B}(\bar{z}^*, \varepsilon) \subset \bar{\mathcal{N}}$. The convergence of the sequence $(X^k, \nu^k, \lambda^k, \mu^k)$ to $(X^*, \nu^*, \lambda^*, \mu^*)$ follows from lemma 4.1.1. □

Unlike for the exact method, the solution of the problem with approximated derivatives (4.63) is not defined by means of a simple generalized inverse. Instead the following theorem holds.

Theorem 4.3.3 (Solution of the inexactly linearized problems with fixed active sets for the Huber estimator)**.** *Let ΔX^k be a minimizer of (4.63) and let the conditions of lemma 3.1.3 hold that guarantee the invertibility of the FJ matrix*

$$\tilde{\mathcal{A}}^k = \begin{pmatrix} A_1^{kT} D^* A_1^k & -A_2^{kT} & -\tilde{A}_3^{kT} \\ A_2^k & 0 & 0 \\ \tilde{A}_3^k & 0 & 0 \end{pmatrix},$$

with D^* as defined by (4.47a) with the partitioning at the optimum X^* and $\tilde{A}_3^k = (A_{3i}^k)_{i \in I_{F_3}(X^*)}$. Then there exists a unique FJ point $(\Delta X^k, \nu^{AP,k}, \lambda^{AP,k}, \mu^{AP,k})$ with $\mu_0^{AP,k} = 1$ such that ΔX^k is a strict minimum. This minimizer can be computed by means of the formula

$$\begin{aligned} \Delta X^k = &- \begin{pmatrix} \mathbb{I} & 0 & 0 \end{pmatrix} \begin{pmatrix} A_1^{kT} D^* A_1^k & -A_2^{kT} & -\tilde{A}_3^{kT} \\ A_2^k & 0 & 0 \\ \tilde{A}_3^k & 0 & 0 \end{pmatrix}^{-1} \begin{pmatrix} A_1^{kT} & 0 & 0 \\ 0 & \mathbb{I} & 0 \\ 0 & 0 & \mathbb{I} \end{pmatrix} \begin{pmatrix} \tilde{F}_1(X^k) \\ F_2(X^k) \\ \tilde{F}_3(X^k) \end{pmatrix} \\ &- \begin{pmatrix} \mathbb{I} & 0 & 0 \end{pmatrix} \begin{pmatrix} A_1^{kT} D^* A_1^k & -A_2^{kT} & -\tilde{A}_3^{kT} \\ A_2^k & 0 & 0 \\ \tilde{A}_3^k & 0 & 0 \end{pmatrix}^{-1} \begin{pmatrix} m(X^k, \nu^k, \lambda^k, \mu^k) \\ 0 \\ 0 \end{pmatrix}, \end{aligned}$$

where $m(X^k, \nu^k, \lambda^k, \mu^k) := \nabla_X \mathcal{L}(X^k, \nu^k, \lambda^k, \mu^k) - (A_1^k)^T \nu^k - (A_2^k)^T \lambda^k - (A_3^k)^T \mu^k$ like in the proof of theorem 4.3.1.

Proof. The formula for ΔX^k is a direct consequence of equation (4.57). □

Part II.

Uncertainties in model parameters and structure

The second part of the thesis extends the range of errors and uncertainties from measurement errors alone to erroneous measurements in combination with uncertain model parameters. The first chapter of this part summarizes the well-known worst-case approach of robust optimization that ensures that the solution of the optimization problem will still be valid under the worst possible circumstances. Then we present in the following chapter an alternative approach that combines the multiple shooting technique described in section 2.3.2 of this thesis with the method of stochastic collocation. In contrast to worst-case optimization, this approach gives a solution that is closer to the average case of all problem instances over the space of uncertain parameters.

5. The worst-case approach of robust optimization

As we have seen in the previous part, experimental data in optimization problems are usually uncertain and prone to measurement errors. These errors can be coped with by means of robust parameter estimation methods. Standard estimators that are based on a normality assumption for the measurement errors are very sensitive to deviations from this assumption and often a single outlier can make the resulting estimate completely worthless. Thus we presented the least absolute deviations estimator and Huber's M-estimator as two important examples of robust alternatives that are less sensitive to wild data points and analyzed their properties. But still the uncertainty in the data is to some extent transferred to the estimates which is the reason that beside the nominal estimate usually some confidence region is given around it. For least squares parameter estimates computed with the Gauss-Newton method, a linear approximation of the covariance matrix in the solution can be used to compute a linear approximation of the nonlinear confidence region that is associated with the $(1-\alpha)$-quantile of the χ^2 distribution with a corresponding number of degrees of freedom. This confidence region contains the true parameters with a probability of $1-\alpha$ per cent. So not only the nominal parameter estimate but any point in the surrounding confidence region could be the true parameter value. When using these estimates in later computations, like e.g. in optimum experimental design, the nominal parameter values might not be the "best choice". In particular in security critical applications, it is important to know when the influence of the perturbations is particularly large in order to compute a robust solution, i.e. one that remains feasible under all circumstances. Ben-Tal and Nemirovski [16] show for a linear example from the NETLIB library what desastrous consequences even small deviations in the uncertain problem data can have on the constraints and the cost function value. In the same paper they also quantify the effects that small uncertainties cause in the feasibility of 90 NETLIB problem and the impact of a robustification of these problems. Their results underscore the importance of the robust optimization methodology that is the subject matter of this chapter.

We start with a general introduction to this field of optimization and with the definition of the problem that we are going to consider in the following. We define our specific uncertainty set and relate it to other frequently used uncertainty sets. In the second section we define the robust counterpart problem that corresponds to our uncertain optimization problem and explain the idea of the worst-case approach for its solution. The third section is devoted to an approximation technique for the resulting worst-case problem that reduces it to a nonlinear, non-smooth minimization

problem in standard form. We show how this problem is addressed numerically in the fourth section and the fifth section treats its simpler case without inequality constraints. The sixth section, finally, contains a short outlook on the relationship of robustness and sparsity in the solution of robust regression problems.

5.1. Introduction to the problem

There are several possibilities how to deal with uncertainty in the nominal data of optimization problems. The easiest and often used approach is replacing the uncertain parameter by its nominal value which corresponds to simply ignoring the uncertainty. In stochastic optimization, the uncertainty is described by a probability distribution. But mostly this approach does not secure against violations of the constraints but accepts minor violations of the feasibility. Robust mathematical programming considers several scenarios with penalty terms in the cost function for violations of the constraints. Though this approach may also find infeasible solutions, it can be considered as a special case of robust optimization under weak structural assumptions, see Ben-Tal and Nemirovski [15].

Robust optimization is concerned with deterministic, set-based uncertainty. Much of the motivation for the development of this methodology comes from optimal control which is a related but independent field, cf. Bersimas et al. [18]. An overview of the developments in robust optimization and the field of robustness relevant to operations research and management science of the last five years is given by Gabrel et al. [52]. Have another look at the introductory section 1.2 for a concise summary of the connection between stochastic and robust optimization.

The general idea of robust optimization is the following. For a problem with uncertain parameters define an uncertainty set in which these parameters are allowed to vary and demand the constraints to hold for all realizations of the parameters within the given set. The corresponding optimization problem is called *robust counterpart problem*. In general, this robust counterpart has infinitely many constraints, but for many choices of the uncertainty set it is at least *tractable*, i.e. solvable in polynomial time.

This chapter is concerned with problems of the form (2.9) with least squares objective function that are governed by ODE constraints with uncertain coefficients,

$$\begin{array}{ll} \min\limits_{x,p} & \frac{1}{2}\|r_1(x,p)\|_2^2 \\ \text{s.t.} & \dot{x}(t) = f(t,x,p,\alpha), \\ & r_2(x(\theta_1),\ldots,x(\theta_K),p,\alpha) = 0, \\ & -r_3(x(\theta_1),\ldots,x(\theta_K),p) \leq 0, \end{array} \qquad \alpha \in \mathcal{A}, \tag{5.1}$$

with the problem functions $r_1 : \mathbb{R}^{n_x} \times \mathbb{R}^{n_p} \to \mathbb{R}^{m_{r1}}$, $r_2 : \left(\mathbb{R}^{n_x}\right)^K \times \mathbb{R}^{n_p} \times \mathbb{R}^{n_a} \to \mathbb{R}^{m_{r2}}$, $r_3 : \left(\mathbb{R}^{n_x}\right)^K \times \mathbb{R}^{n_p} \to \mathbb{R}^{m_{r3}}$. We assume that there are as many equality constraints as there are state variables, i.e. $m_{r2} = n_x$. This means that if the dynamic system is characterized more specifically by an initial value condition, this is included in the

equality constraints r_2 here and there may not be any further equality constraints. This is quite a reasonable assumption as with dynamic constraints from biological or chemical processes often the initial state is known and thus fixed while there are no further restrictions on the system. On the other hand, the initial value may be uncertain also while there are other defining conditions in form of equality constraints. This is also allowed in this problem formulation as long as the condition $m_{r2} = n_x$ remains satisfied.

As an example for such a situation we may consider an optimum experimental design problem. An experiment shall be planned ensuing from already available data in order to gain more information about the process. This task can be formulated as a mathematical optimization problem in which the experimental conditions are the parameters that are to be estimated while the mathematical model of the considered process is fixed. These experimental conditions include control parameters like temperature and pressure under which the process takes place, influx of substances, and so on, but also decision variables like measurement times and substances to be recorded at these times. In biology or chemistry the initial concentrations of the involved species can then be chosen in dependence on the experimental conditions as these starting conditions can also influence the process behavior significantly.

Alternatively, it is also conceivable to have a problem where the initial value of the concentrations is completely free and subject to optimization only via the state variables while there are other certain or uncertain equality constraints on the system behavior. The only requirement that may not be relaxed for our approach is the number of equality constraints, $m_{r2} = n_x$.

The uncertainty set $\mathcal{A}$ has the form

$$\mathcal{A} = \bigotimes_{j=1}^{n_a} \mathcal{A}_j \quad \text{with} \quad \mathcal{A}_j = \{a : |a - \alpha_0^j| \leq \delta_j\} = [\alpha_0^j - \delta_j, \alpha_0^j + \delta_j] \subset \mathbb{R}, \tag{5.2}$$

i.e. each uncertain parameter α^j lies in an interval of length $2\delta_j$ around a nominal value α_0^j. This corresponds to a uniform distribution of the uncertainty α around some nominal value $\alpha_0 = (\alpha_0^j)_{j=1,\dots,n_a}$. Thus the expected (mean) value of each uncertainty is $\mathbb{E}(\alpha^j) = \alpha_0^j$ and correspondingly $\mathbb{E}(\alpha) = \alpha_0$. Note, however, that $\mathcal{A}$ gives a deterministic description of the uncertainty α so that (5.1) is not a stochastic optimization problem although it could be treated as one. This stochastic interpretation of the constraint set is exploited in the next chapter.

Again, this structure of the uncertainty set is typical for optimal experimental design problems where a parameter estimation for the process under consideration has been performed before and beside the nominal parameter values also the corresponding confidence regions are given. Particularly in the case when these confidence regions are large, i.e. the parameters could not be estimated with sufficient accuracy, it is probably advantageous with regard to the possible information gain to include these regions in the computations for the experimental conditions instead of just continuing with the nominal parameter values.

In general, the choice of the uncertainty set is of vital importance in robust optimization and should satisfy three basic requirements (cf. Ben-Tal et al. [13] and see also Bertsimas et al. [18]), namely

- consistency with all available data and information concerning the uncertain parameters,
- statistical meaning,
- tractability of the resulting robust counterpart problem.

If the uncertainty set is based on linear confidence regions of former parameter estimations, cp. section 3.1.6, the first requirement is satisfied. Its statistical significance can be controlled by means of the choice of the size of the confidence level. A 95% confidence level will lead to uncertainty sets that are statistically meaningful. The form of the confidence set will in this case not hinder tractability of the robust counterpart, but this point also depends on the problem functions and cannot be answered so easily. Beside our formulation (5.2), typical uncertainty sets are (Bertsimas et al. [18])

- the quadratic or ellipsoidal set around the mean α_0 with radius δ that is tilted according to the covariance Σ of the uncertainties,

$$\mathcal{A}^Q(\delta) := \{a : (a - \alpha_0)^T \Sigma (a - \alpha_0) \leq \delta^2\},$$

- the set with all a^i lying in the interval $[\underline{\alpha}^i, \bar{\alpha}^i]$, under the requirement that the total weight of the deviations from α_0^i over all i is not greater than δ,

$$\mathcal{A}^D(\delta) := \{a : \exists u \in \mathbb{R}^{n_a} a^i = \alpha_0^i - (\underline{\alpha}^i - \bar{\alpha}^i)u^i,\ u^i \leq 1,\ \sum_{i=1}^{n_a} u^i \leq \delta\}.$$

The third possibility given by Bertsimas et al. [18] does not apply to our optimization problem. Note that for $\delta = 0$ both sets reduce to the one-point set that only contains the nominal parameter value α_0. Statistically, the set $\mathcal{A}^Q$ puts most weight on the first two moments of the uncertainty, the set $\mathcal{A}^D$ corresponds to a worst-case outcome over a subset of the uncertainties. In contrast to this stochastic approach to defining uncertainty sets, they can also be classified according to their geometric properties (Ben-Tal et al. [13]), e.g.

- boxes $\|\alpha\|_\infty \leq 1$,
- balls $\|\alpha\|_2 \leq 1$,
- polytopes $C\alpha + d \geq 0$,
- (closed, convex, acute) cones $C\alpha + d \in \mathcal{K}$,

- separable functions $\sum_i g_{li}(\alpha^i) \leq 0 \; \forall l \in \{1, \ldots, L\}$ with g_{li} convex.

Obviously, our uncertainty set $\mathcal{A} = \bigotimes_{j=1}^{n_a} [\alpha_0^j - \delta_j, \alpha_0^j + \delta_j]$ is a polytope and can be written like in the list above with

$$C = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ 0 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & -1 & 0 \\ 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 0 & -1 \end{pmatrix}, \quad d = \begin{pmatrix} \delta_1 - \alpha_0^1 \\ \delta_1 + \alpha_0^1 \\ \delta_2 - \alpha_0^2 \\ \delta_2 + \alpha_0^2 \\ \vdots \\ \delta_{n_a-1} - \alpha_0^{n_a-1} \\ \delta_{n_a-1} + \alpha_0^{n_a-1} \\ \delta_{n_a} - \alpha_0^{n_a} \\ \delta_{n_a} + \alpha_0^{n_a} \end{pmatrix}.$$

Another reformulation of our uncertainty set uses the interval boundaries as upper and lower bounds for the uncertainties, $\alpha_0 - \delta \leq \alpha \leq \alpha_0 + \delta$, with $\delta = (\delta_j)_{j=1}^{n_a}$, such that the uncertainty set can also be written as (Ma and Braatz [95])

$$\begin{aligned} \mathcal{A} &= \{a : \alpha_0 - \delta \leq a \leq \alpha_0 + \delta, \; \delta = (\delta_j)_{j=1}^{n_a}\} \\ &= \{a : a = \alpha_0 + d\alpha, \; \|W_\alpha^{-1} d\alpha\|_\infty \leq 1\} \\ &= \{a : \|W_\alpha^{-1}(a - \alpha_0)\|_\infty \leq 1\} \end{aligned} \tag{5.3}$$

with the diagonal matrix $W_\alpha = \operatorname{diag}(\delta_j)$. We will use this last representation (5.3) of $\mathcal{A}$ when formulating the robust counterpart of problem (5.1) and computing an approximate solution for it in the following sections.

Obviously, the solution of the uncertain ODE constraints in (5.1) depends on the parameters p and the realization of the uncertainty α and any solution is a function of the uncertainty α,

$$p = p(\alpha), \quad x(t) = x(t; p(\alpha), \alpha).$$

For ease of notation, we will omit in the following the explicit mentioning of this dependency for the parameters and will write $x^\alpha(t)$ for the dynamic states where the dependency on the uncertainty needs to be emphasized.

5.2. The robust counterpart problem and the worst-case approach

We discretize the dynamic system in the uncertain problem (5.1) by means of the multiple shooting approach described in section 2.3 to get a standard least squares problem with equality and inequality constraints. For the resulting problem, we can define a corresponding robust counterpart problem that computes some kind of worst-case solution that still satisfies the constraints of the problem. To this end, the objective function and the inequality constraints are replaced by their respective

maxima over the set the of uncertainties and still required to satisfy the equality constraints. It will become clear in the course of this section what is meant by this.

We start with discretizing the dynamic system, using $N+1$ multiple shooting nodes $\tau_j,\ j=0,\dots,N$ to subdivide the integration interval. Thus the optimization problem is transformed to the parametric optimization problem

$$\begin{array}{ll}\min\limits_{s,p} & \frac{1}{2}\|r_1(s,p)\|_2^2 \\ \text{s.t.} & r_2(x(\theta_1;s_{j_1},p,\alpha),\dots,x(\theta_K;s_{j_K},p,\alpha),p,\alpha)=0, \\ & h_j(s_j,s_{j+1},p,\alpha):=x(\tau_{j+1};s_j,p,\alpha)-s_{j+1}=0, \quad j=0,\dots,N-1, \\ & -r_3(x(\theta_1;s_{j_1},p,\alpha),\dots,x(\theta_K;s_{j_K},p,\alpha),p)\le 0,\end{array} \tag{5.4}$$

with uncertain parameters $\alpha\in\mathcal{A}$, where the solution $x(t;s_j,p,\alpha)$ is the solution of the initial value problem

$$\dot{x}(t)=f(t,x,p,\alpha),\quad x(\tau_j)=s_j(\alpha),$$

on the $(j+1)$-st multiple shooting interval. The robust counterpart of this problem requires the constraints to hold for all uncertainties in $\mathcal{A}$ and the optimum to be minimal over this set,

$$\left.\begin{array}{ll}\min\limits_{s,p} & \frac{1}{2}\|r_1(s,p)\|_2^2 \\ & r_2(x(\theta_1;s_{j_1},p,\alpha),\dots,x(\theta_K;s_{j_K},p,\alpha),p,\alpha)=0, \\ \text{s.t.} & h_j(s_j,s_{j+1},p,\alpha):=x(\tau_{j+1};s_j,p,\alpha)-s_{j+1}=0, \quad j=0,\dots,N-1, \\ & -r_3(x(\theta_1;s_{j_1},p,\alpha),\dots,x(\theta_K;s_{j_K},p,\alpha),p)\le 0,\end{array}\right\}\ \forall\alpha\in\mathcal{A}. \tag{5.5}$$

So minimization of $\frac{1}{2}\|r_1(s,p)\|_2^2$ with respect to s and p refers to an optimization with respect to all realizations of $\alpha\in\mathcal{A}$. Thus, having a continuous set of uncertainties, we get an infinite continuous set of discretized solutions $s=s(\alpha)$ and $p=p(\alpha)$ and thus also $x=x(\alpha)$, i.e. a whole solution space the size of which depends on the size of the uncertainty set $\mathcal{A}$. Our aim is therefore the computation of minimizing solutions (s^*,p^*) from the set of all solutions that satisfy the constraints for all realizations of the uncertainties $\alpha\in\mathcal{A}$.

The point (s,p) is called a *robust feasible* or *robust optimal* solution, respectively. The robust feasible set is given by

$$\mathcal{S}(\mathcal{A})=\{(s,p)\mid r_2=0\ \&\ h_j=0\ (j=0,\dots,N-1)\ \&\ -r_3\le 0\ \forall\alpha\in\mathcal{A}\subset\mathbb{R}^{n_a}\}. \tag{5.6}$$

Such a robustification increases the computational difficulty enormously. Even for convex optimization problems the robust counterpart is generally not computable in polynomial time.

For practical purposes, the robust counterpart has to be tractable. See Ben-Tal and Nemirovski [15] for a proof that appropriate (re)formulations exist for several problem classes with suitable (ellipsoidal) uncertainty sets, and Bertsimas et al. [18] for a list of exemplary classes of optimization problems with tractable robust counterparts.

However, the equality conditions in our problem further complicate the robustification as they are much harder to satisfy for all realizations of the uncertainty than inequalities. Note that the reformulation and the choice of suitable uncertainty sets may be a challenging task already for optimization problems without equality constraints, cf. Ben-Tal and Nemirovski [17]. To a great extent, i.e. under a Slater condition, polynomial computability for problems with inequality constraints only corresponds to convexity of the robust feasible set (5.6) at (s,p) with an efficiently computable separation test, cf. Bertsimas et al. [18]. Due to the discretized dynamic system we cannot discard the equality constraints from our considerations here.

The idea of the worst-case approach that we describe in this chapter is that the scenario which corresponds to the worst possible instance of uncertainty in $\mathcal{A}$ shall be optimized. A solution (s,p) is feasible for this problem if it satisfies all possible realizations of the constraints $r_2(x(\cdot;s,p,\alpha),p,\alpha)=0$ and $h_j(s_j,s_{j+1},p)=0$ for all $\alpha\in\mathcal{A}$. This is apparently the same notion of feasibility as known from robust control. The vector (s,p) is an optimal solution of this problem if is yields a best guaranteed value $\sup_{\alpha\in\mathcal{A}}\frac{1}{2}\|r_1(s,p)\|_2^2$ for the cost function under all realizations of the uncertain constraints. Therefore, we define the maximum of the objective function over the set of uncertainties subject to the equality constraints by

$$\begin{aligned}\phi_0(p) := \max_{s,\alpha}\quad & \tfrac{1}{2}\|r_1(s,p)\|_2^2\\ \text{s.t.}\quad & r_2(x(\theta_1;s_{j_1},p,\alpha),\dots,x(\theta_K;s_{j_K},p,\alpha),p,\alpha)=0,\\ & h_j(s_j,s_{j+1},p,\alpha)=0,\quad j=0,\dots,N-1,\\ & \|W_\alpha^{-1}(\alpha-\alpha_0)\|_\infty\le 1,\end{aligned}\tag{5.7}$$

and the corresponding maximum of the inequality constraints,

$$\begin{aligned}\phi_i(p) := \max_{s,\alpha}\quad & -r_{3i}(x(\theta_1;s_{j_1},p,\alpha),\dots,x(\theta_K;s_{j_K},p,\alpha),p)\\ \text{s.t.}\quad & r_2(x(\theta_1;s_{j_1},p,\alpha),\dots,x(\theta_K;s_{j_K},p,\alpha),p,\alpha)=0,\\ & h_j(s_j,s_{j+1},p,\alpha)=0,\quad j=0,\dots,N-1,\\ & \|W_\alpha^{-1}(\alpha-\alpha_0)\|_\infty\le 1,\end{aligned}\tag{5.8}$$

for $i=1,\dots,m_{r3}$. These are the worst-possible values that the objective function and the inequalities may take over the uncertainty set $\mathcal{A}$ subject to the given equality constraints. The worst-case solution of the uncertain optimization problem is then given by the solution of the problem

$$\begin{aligned}\min_{p}\quad & \phi_0(p)\\ \text{s.t.}\quad & \phi_i(p)\le 0,\quad i=1,\dots,m_{r3}.\end{aligned}\tag{5.9}$$

5.3. Approximation of the robust counterpart problem

The double optimization structure of problem (5.9) makes this problem mathematically and numerically very challenging. Therefore we choose to solve it approximately.

For better readability we summarize the equality constraints of the discretized problem (5.5) in one function,

$$g(s,p,\alpha) := \begin{pmatrix} r_2(x(\theta_1; s_{j_1}, p, \alpha), \ldots, x(\theta_K; s_{j_K}, p, \alpha), p, \alpha) \\ h_0(s_0, s_1, p, \alpha) \\ \vdots \\ h_{N-1}(s_{N-1}, s_N, p, \alpha) \end{pmatrix}.$$

As we assumed that there are as many equality constraints as there are state variables, $m_{r2} = n_x$, we have $(N+1)n_x = Nn_x + m_{r2}$ and thus the Jacobian $\frac{\partial g}{\partial s}$ is quadratic. We further assume that this matrix is nonsingular everywhere. Then the discretized states s can be treated as implicit functions of the parameters p and the uncertainties α.

Under these assumptions we can apply the technique described by Diehl et al. [43]. This amounts to linearizing the functions $\phi_i,\ i = 0, \ldots, m_{r3}$, around some point $(\bar{s}, p, \alpha_0)$ that satisfies $g(\bar{s}, p, \alpha_0) = 0$ and solving the linearized problems $\tilde{\phi}_i,\ i = 0, \ldots, m_{r3}$, explicitly. This leads to an approximated robust counterpart problem with standard optimization structure.

So let, for given p and the nominal uncertainty α_0, the point $(\bar{s}, p, \alpha_0)$ be such that the equality constraints are satisfied, $g(\bar{s}, p, \alpha_0) = 0$. We now linearize not only the objective function and the inequality constraints but also the equations $g(\bar{s}, p, \alpha_0)$ that contain the discretized dynamic system. For brevity, we introduce the notation

$$\bar{x}(\theta_k) := x(\theta_k; \bar{s}_{j_k}, p, \alpha_0).$$

The linearized maximization problems then read

$$\tilde{\phi}_0(p) := \begin{cases} \max\limits_{\Delta s, \Delta\alpha} & \frac{1}{2}\|r_1(\bar{s}, p)\|_2^2 + r_1(\bar{s}, p)^T \frac{\partial r_1}{\partial s}(\bar{s}, p)\Delta s \\ \text{s.t.} & \frac{\partial r_2}{\partial s}(\bar{x}(\theta_1), \ldots, \bar{x}(\theta_K), p, \alpha_0)\Delta s + \frac{\partial r_2}{\partial \alpha}(\bar{x}(\theta_1), \ldots, \bar{x}(\theta_K), p, \alpha_0)\Delta\alpha = 0, \\ & \frac{\partial h_j}{\partial s}(\bar{s}_j, \bar{s}_{j+1}, p, \alpha_0)\Delta s + \frac{\partial h_j}{\partial \alpha}(\bar{s}_j, \bar{s}_{j+1}, p, \alpha_0)\Delta\alpha = 0, \quad j = 0, \ldots, N-1, \\ & \|W_\alpha^{-1}\Delta\alpha\|_\infty \leq 1, \end{cases}$$

and for $i = 1, \ldots, m_{r3}$,

$$\tilde{\phi}_i(p) := \begin{cases} \max\limits_{\Delta s, \Delta\alpha} & -r_{3i}(\bar{x}(\theta_1), \ldots, \bar{x}(\theta_K), p) - \frac{\partial r_{3i}}{\partial s}(\bar{x}(\theta_1), \ldots, \bar{x}(\theta_K), p)\Delta s \\ \text{s.t.} & \frac{\partial r_2}{\partial s}(\bar{x}(\theta_1), \ldots, \bar{x}(\theta_K), p, \alpha_0)\Delta s + \frac{\partial r_2}{\partial \alpha}(\bar{x}(\theta_1), \ldots, \bar{x}(\theta_K), p, \alpha_0)\Delta\alpha = 0, \\ & \frac{\partial h_j}{\partial s}(\bar{s}_j, \bar{s}_{j+1}, p, \alpha_0)\Delta s + \frac{\partial h_j}{\partial \alpha}(\bar{s}_j, \bar{s}_{j+1}, p, \alpha_0)\Delta\alpha = 0, \quad j = 0, \ldots, N-1, \\ & \|W_\alpha^{-1}\Delta\alpha\|_\infty \leq 1. \end{cases}$$

We can solve these convex problems by eliminating the state increments Δs via the equality constraints by computing

$$\Delta s = -\left(\frac{\partial g}{\partial s}(\bar{x}(\theta_1), \ldots, \bar{x}(\theta_K), p, \alpha_0)\right)^{-1} \frac{\partial g}{\partial \alpha}(\bar{x}(\theta_1), \ldots, \bar{x}(\theta_K), p, \alpha_0)\Delta\alpha.$$

Inserting this into the objective functions of the $\tilde{\phi}_i$, $i = 0, \ldots, m_{r3}$, we get

$$\begin{aligned} \tilde{\phi}_0(p) = \max_{\Delta\alpha} \quad & \tfrac{1}{2}\|r_1(s,p)\|_2^2 - r_1(\bar{s},p)^T \tfrac{\partial r_1}{\partial s}(\bar{s},p) \left(\tfrac{\partial g}{\partial s}\right)^{-1} \tfrac{\partial g}{\partial \alpha}\Delta\alpha \\ \text{s.t.} \quad & \|W_\alpha^{-1}\Delta\alpha\|_\infty \leq 1, \end{aligned} \tag{5.10}$$

and for $i = 1, \ldots, m_{r3}$,

$$\begin{aligned} \tilde{\phi}_i(p) = \max_{\Delta\alpha} \quad & -r_{3i}(x(\theta_1), \ldots, x(\theta_K), p) + \tfrac{\partial r_{3i}}{\partial s}(x(\theta_1), \ldots, x(\theta_K), p) \left(\tfrac{\partial g}{\partial s}\right)^{-1} \tfrac{\partial g}{\partial \alpha}\Delta\alpha \\ \text{s.t.} \quad & \|W_\alpha^{-1}\Delta\alpha\|_\infty \leq 1, \end{aligned} \tag{5.11}$$

where both the derivatives $\frac{\partial g}{\partial s}$ and $\frac{\partial g}{\partial \alpha}$ are evaluated at $(\bar{x}(\theta_1), \ldots, \bar{x}(\theta_K), p, \alpha_0)$. Note that due to the origin of the equality constraints from the discretization of the dynamic system the matrix $\frac{\partial g}{\partial s}$ has a special structure,

$$\frac{\partial g}{\partial s} = \begin{pmatrix} \frac{\partial r_2}{\partial s_0} & \frac{\partial r_2}{\partial s_1} & & \frac{\partial r_2}{\partial s_{N-1}} & \frac{\partial r_2}{\partial s_N} \\ \frac{\partial x}{\partial s_0} & -\mathbb{I} & & & \\ & \frac{\partial x}{\partial s_1} & -\mathbb{I} & & 0 \\ 0 & & \ddots & \ddots & \\ & & & \frac{\partial x}{\partial s_{N-1}} & -\mathbb{I} \end{pmatrix}. \tag{5.12}$$

The derivative of g with respect to the uncertainties α reads

$$\frac{\partial g}{\partial \alpha} = \begin{pmatrix} \frac{\partial r_2}{\partial \alpha} \\ \frac{\partial x}{\partial \alpha}(\tau_1) \\ \cdots \\ \frac{\partial x}{\partial \alpha}(\tau_N) \end{pmatrix}.$$

Lemma 5.3.1. *The problems $\tilde{\phi}_i(p)$, $i = 0, \ldots, m_{r3}$, have the explicit solutions*

$$\begin{aligned} \tilde{\phi}_0(p) &= \frac{1}{2}\|r_1(\bar{s},p)\|_2^2 + \left\| -W_\alpha \left(\frac{\partial g}{\partial \alpha}\right)^T \left(\frac{\partial g}{\partial s}\right)^{-T} \left(\frac{\partial r_1}{\partial s}\right)^T r_1 \right\|_1, \\ \tilde{\phi}_i(p) &= -r_{3i}(\bar{x}(\theta_1), \ldots, \bar{x}(\theta_K), p) - \left\| -W_\alpha \left(\frac{\partial g}{\partial \alpha}\right)^T \left(\frac{\partial g}{\partial s}\right)^{-T} \left(\frac{\partial r_{3i}}{\partial s}\right)^T \right\|_1, \\ & \quad i = 1, \ldots, m_{r3}. \end{aligned}$$

Proof. Problems (5.10) and (5.11) can be rewritten as

$$\begin{aligned} \tilde{\phi}_0(p) = \tfrac{1}{2}\|r_1(s,p)\|_2^2 + \max_{\Delta\alpha} \quad & -r_1(\bar{s},p)^T \tfrac{\partial r_1}{\partial s}(\bar{s},p) \left(\tfrac{\partial g}{\partial s}\right)^{-1} \tfrac{\partial g}{\partial \alpha}\Delta\alpha \\ \text{s.t.} \quad & \|W_\alpha^{-1}\Delta\alpha\|_\infty \leq 1, \end{aligned}$$

and for $i = 1, \ldots, m_{r3}$,

$$\begin{aligned} \tilde{\phi}_i(p) = -r_{3i}(x(\theta_1), \ldots, x(\theta_K), p) + \max_{\Delta\alpha} \quad & \tfrac{\partial r_{3i}}{\partial s}(x(\theta_1), \ldots, x(\theta_K), p) \left(\tfrac{\partial g}{\partial s}\right)^{-1} \tfrac{\partial g}{\partial \alpha}\Delta\alpha \\ \text{s.t.} \quad & \|W_\alpha^{-1}\Delta\alpha\|_\infty \leq 1. \end{aligned}$$

Now we can exploit the fact that $\|\cdot\|_1$ is the dual norm of $\|\cdot\|_\infty$, i.e. the mapping

$$\begin{aligned}\|\cdot\|_* : \quad & \mathbb{R}^{n_a} \to \mathbb{R}, \\ & u \mapsto \|u\|_* := \max_{\alpha \in \mathbb{R}^{n_a}} u^T \alpha, \text{ s.t.} \|\alpha\| \leq 1.\end{aligned}$$

□

So the approximated robust counterpart problem of (5.4) is given by

$$\begin{aligned}\min_{p,s} \quad & \tfrac{1}{2}\|r_1(s,p)\|_2^2 + \left\| -W_\alpha \left(\tfrac{\partial g}{\partial \alpha}\right)^T \left(\tfrac{\partial g}{\partial s}\right)^{-T} \left(\tfrac{\partial r_1}{\partial s}\right)^T r_1 \right\|_1 \\ \text{s.t.} \quad & -r_{3i}(x(\theta_1), \ldots, x(\theta_K), p) - \left\| -W_\alpha \left(\tfrac{\partial g}{\partial \alpha}\right)^T \left(\tfrac{\partial g}{\partial s}\right)^{-T} \left(\tfrac{\partial r_{3i}}{\partial s}\right)^T \right\|_1 \leq 0, \\ & \quad i = 1, \ldots, m_{r3}, \\ & r_2(x(\theta_1; s_{j_1}, p, \alpha_0), \ldots, x(\theta_K; s_{j_K}, p, \alpha_0), p, \alpha_0) = 0, \\ & h_j(s_j, s_{j+1}, p, \alpha_0) = 0, \quad j = 0, \ldots, N-1,\end{aligned} \tag{5.13}$$

where all derivatives are evaluated at the point (s, p, α_0). Note, however, that these derivatives depend on the unknown multiple shooting states s and the parameters p and are therefore not fixed but subject to optimization.

5.4. Numerical solution of the approximated robust counterpart problem

Numerically, it is not sensible to compute the inverse of the Jacobian $\frac{\partial g}{\partial s}$ explicitly as this would almost certainly destroy the sparse structure of the matrix (5.12). A computationally more reasonable way to access the problem consists in solving problem (5.13) using sensitivity equations. We summarize the two possible approaches via direct or adjoint sensitivity matrices. Note their similarity to the forward and reverse mode of automatic differentiation (cf. Diehl et al. [43]).

For the direct approach, we define the derivative matrix

$$D(s, p, \alpha_0) := \left[-\left(\frac{\partial g}{\partial s}(s, p, \alpha_0)\right)^{-1} \left(\frac{\partial g}{\partial \alpha}(s, p, \alpha_0)\right) \right] \in \mathbb{R}^{(N+1)n_x \times n_a},$$

and introduce it as a new optimization variable. The direct approach to the approximated robust counterpart problem via sensitivity equations thus amounts to solving the problem

$$\begin{aligned}\min_{p,s,D} \quad & \tfrac{1}{2}\|r_1(s,p)\|_2^2 + \left\| W_\alpha D^T \left(\tfrac{\partial r_1}{\partial s}\right)^T r_1 \right\|_1 \\ \text{s.t.} \quad & -r_{3i}(x(\theta_1), \ldots, x(\theta_K), p) - \left\| W_\alpha D^T \left(\tfrac{\partial r_{3i}}{\partial s}\right)^T \right\|_1 \leq 0, \quad i = 1, \ldots, m_{r3}, \\ & r_2(x(\theta_1; s_{j_1}, p, \alpha_0), \ldots, x(\theta_K; s_{j_K}, p, \alpha_0), p, \alpha_0) = 0, \\ & h_j(s_j, s_{j+1}, p, \alpha_0) = 0, \quad j = 0, \ldots, N-1, \\ & \tfrac{\partial g}{\partial s}(s, p, \alpha_0) D + \tfrac{\partial g}{\partial \alpha}(s, p, \alpha_0) = 0.\end{aligned} \tag{5.14}$$

Note that the last equation of this problem is a matrix equation in $\mathbb{R}^{(N+1)n_x \times n_a}$.

In case that there are much more uncertainties than equality constraints in the original problem, $n_a \gg m_{r2} = n_x$ in our case, it is computationally more efficient to use adjoint sensitivity equations. We therefore define the adjoint sensitivity matrix

$$\Lambda(s,p,\alpha_0) := \left[-\left(\frac{\partial r}{\partial s}(s,p,\alpha_0)\right) \left(\frac{\partial g}{\partial s}(s,p,\alpha_0)\right)^{-1} \right]^T \in \mathbb{R}^{(N+1)n_x \times (1+m_{r3})},$$

$$\text{with } r := \begin{pmatrix} \frac{1}{2}\|r_1(s,p)\|_2^2 \\ r_{31} \\ \vdots \\ r_{3m_{r3}} \end{pmatrix},$$

as a new optimization variable. Let $\lambda_0, \lambda_1, \ldots, \lambda_{m_{r3}}$ denote the columns of Λ. Then the adjoint apporach to the approximated robust counterpart problem via sensitivity equations amounts to solving the problem

$$\begin{aligned} \min_{p,s,\Lambda} \quad & \tfrac{1}{2}\|r_1(s,p)\|_2^2 + \left\| W_\alpha \left(\tfrac{\partial g}{\partial \alpha}\right)^T \lambda_0 \right\|_1 \\ \text{s.t.} \quad & -r_{3i}(x(\theta_1),\ldots,x(\theta_K),p) - \left\| W_\alpha \left(\tfrac{\partial g}{\partial \alpha}\right)^T \lambda_i \right\|_1 \le 0, \quad i = 1,\ldots,m_{r3}, \\ & r_2(x(\theta_1;s_{j_1},p,\alpha_0),\ldots,x(\theta_K;s_{j_K},p,\alpha_0),p,\alpha_0) = 0, \\ & h_j(s_j,s_{j+1},p,\alpha_0) = 0, \quad j = 0,\ldots,N-1, \\ & \left(\tfrac{\partial g}{\partial s}(s,p,\alpha_0)\right)^T \Lambda + \left(\tfrac{\partial r}{\partial s}(s,p,\alpha_0)\right)^T = 0. \end{aligned} \tag{5.15}$$

Note that the last equation of this problem is a matrix equation in $\mathbb{R}^{(N+1)n_x \times (1+m_{r3})}$.

In both cases, we are faced with the problem of nonsmooth problem functions due to the l_1 norm in the objective function and the inequality constraints. But we can apply the reformulation introduced in (2.16) in section 2.6 on least absolute deviations estimation. To do this for the direct approach, we introduce new optimization variables $u_i, v_i \in \mathbb{R}^{n_a}$, $i = 0,1,\ldots,m_{r3}$, and reformulate problem (5.14) as

$$\begin{aligned} \min_{\substack{p,s,u_i,v_i,\\ i=0,1,\ldots,m_{r3}}} \quad & \tfrac{1}{2}\|r_1(s,p)\|_2^2 + 1^T(u_0+v_0) \\ \text{s.t.} \quad & -r_{3i}(x(\theta_1),\ldots,x(\theta_K),p) - 1^T(u_i+v_i) \le 0, \quad i = 1,\ldots,m_{r3}, \\ & W_\alpha \left(\tfrac{\partial g}{\partial \alpha}(s,p,\alpha_0)\right)^T \left(\tfrac{\partial g}{\partial s}(s,p,\alpha_0)\right)^{-T} \left(\tfrac{\partial r_1}{\partial s} r_1\right) = u_0 - v_0, \\ & W_\alpha \left(\tfrac{\partial g}{\partial \alpha}(s,p,\alpha_0)\right)^T \left(\tfrac{\partial g}{\partial s}(s,p,\alpha_0)\right)^{-T} \left(\tfrac{\partial r_{3i}}{\partial s}\right)^T = u_i - v_i, \quad i = 1,\ldots,m_{r3}, \\ & r_2(x(\theta_1;s_{j_1},p,\alpha_0),\ldots,x(\theta_K;s_{j_K},p,\alpha_0),p,\alpha_0) = 0, \\ & h_j(s_j,s_{j+1},p,\alpha_0) = 0, \quad j = 0,\ldots,N-1, \\ & u_i, v_i \ge 0, \quad i = 0,1,\ldots,m_{r3}. \end{aligned} \tag{5.16}$$

The analogous reformulation of the adjoint problem (5.15) is straightforward.

5.5. Special case without inequality constraints

The robust counterpart problem (5.5) simplifies significantly if there are no inequality constraints present. In this case we only have

$$\begin{array}{ll} \min\limits_{s,p} & \frac{1}{2}\|r_1(s,p)\|_2^2 \\ \text{s.t.} & r_2(x(\theta_1; s_{j_1}, p, \alpha), \ldots, x(\theta_K; s_{j_K}, p, \alpha), p, \alpha) = 0, \qquad \forall \alpha \in \mathcal{A}, \\ & h_j(s_j, s_{j+1}, \alpha) := x(\tau_{j+1}; s_j, p, \alpha) - s_{j+1} = 0, \quad j = 0, \ldots, N-1, \end{array}$$

and thus in the worst-case approach we do not have to consider problems $\phi_i(p)$ defined in (5.8). We only need to solve the minimax problem

$$\begin{array}{lll} \min\limits_{s,p} \max\limits_{\alpha} & \frac{1}{2}\|r_1(s,p)\|_2^2 & \\ \text{s.t.} & r_2(x(\theta_1; s_{j_1}, p, \alpha), \ldots, x(\theta_K; s_{j_K}, p, \alpha), p, \alpha) = 0, & \\ & h_j(s_j, s_{j+1}, \alpha) = 0, \quad j = 0, \ldots, N-1, & \\ & \|W_\alpha^{-1}(\alpha - \alpha_0)\|_\infty \leq 1. & \end{array} \tag{5.17}$$

Applying the method of the previous section, we reduce this problem to a standard optimization problem of the form (5.16) without the constraints referring to the original inequality constraints,

$$\begin{array}{ll} \min\limits_{\substack{p,s,u_i,v_i,\\ i=0,1,\ldots,m_{r3}}} & \frac{1}{2}\|r_1(s,p)\|_2^2 + \mathbb{1}^T(u_0 + v_0) \\ \text{s.t.} & W_\alpha \left(\frac{\partial g}{\partial \alpha}(s,p,\alpha_0)\right)^T \left(\frac{\partial g}{\partial s}(s,p,\alpha_0)\right)^{-T} \left(\frac{\partial r_1}{\partial s} r_1\right) = u_0 - v_0, \\ & r_2(x(\theta_1; s_{j_1}, p, \alpha_0), \ldots, x(\theta_K; s_{j_K}, p, \alpha_0), p, \alpha_0) = 0, \\ & h_j(s_j, s_{j+1}, p, \alpha_0) = 0, \quad j = 0, \ldots, N-1, \\ & u_i, v_i \geq 0, \quad i = 0, 1, \ldots, m_{r3}. \end{array} \tag{5.18}$$

This problem can then be solved by means of a Gauss-Newton method as described in section 3.1. But due to the constraint

$$W_\alpha \left(\frac{\partial g}{\partial \alpha}(s,p,\alpha_0)\right)^T \left(\frac{\partial g}{\partial s}(s,p,\alpha_0)\right)^{-T} \left(\frac{\partial r_1}{\partial s} r_1\right) = u_0 - v_0,$$

this means we have to compute second derivatives for the formulation of the linearized problems in the Gauss-Newton iterations. So either these have to be provided in explicit form or we should have a routine for automatic differentiation at hand.

It is not a topic of this thesis how automatic differentiation works in detail, so we just refer the reader to the standard literature like the textbook by Griewank and Walther [67].

As the worst-case solution of problem (5.1) shall only serve as a reference solution for comparison with the method described in the following chapter, we draw on existing solvers for problem (5.18) and do not go into further detail here.

5.6. Robustness and sparsity

We want to end this chapter with a short digression on the connection between robustness and sparsity. Consider therefore the l_1-penalized least squares regression problem,

$$\min_{x\in\mathbb{R}^n} \|AX - b\|_2^2 + \lambda\|x\|_1. \tag{5.19}$$

This problem is also known under the name of LASSO (Least Absolute Shrinkage and Selection Operator) which is equivalent to the standard least squares regression problem with the additional constraints that the l_1 norm of the parameter vector x may not be greater than a given constant. Xu et al. [142] showed that the solution of (5.19) is also a solution of a robust counterpart problem

$$\min_x \max_{\|\Delta A\|_{\infty,2}\leq\rho} \|(A + \Delta A)x - b\|_2^2, \tag{5.20}$$

where $\|\Delta A\|_{\infty,2}$ denotes the maximum norm of the l_2 norm of the columns of the matrix ΔA. So sparsity can also be considered as a consequence of robustness, see also Bertsimas et al. [18].

Similarly, the solution of the l_2-penalized least squares regression problem,

$$\min_x \|Ax - b\|_2^2 + \rho\sqrt{\|x\|_2^2 + 1},$$

is also the solution of a robust counterpart problem,

$$\min_x \max_{\|\Delta A\Delta b\|_F\leq\rho} \|(A + \Delta A)x - (b + \Delta b)\|_2^2,$$

where $\|\cdot\|_F$ is the Frobenius matrix norm, cf. El Ghaoui and Lebret [46] and Bersimas et al. [18]. So robustification of a solution brings along regularity properties as well.

We can even generalize problem (5.20) to

$$\min_x \max_{\Delta A\in\mathcal{A}} \|b - (A + \Delta A)x\|_2^2, \tag{5.21}$$

with an uncertainty set $\mathcal{A}$ of the form

$$\mathcal{A} = \{A = (a_1, \ldots, a_n) : \|a_i\|_2 \leq c_i,\ i = 1, \ldots, n\}.$$

For given constants $c_i,\ i = 1, \ldots, n$, this problem is equivalent to the l_1-penalized least squares regression problem

$$\min_x \|b - Ax\|_2^2 + \sum_{i=1}^n c_i|x_i|.$$

Even more general, we can use an arbitrary norm $\|\cdot\|_a$ on the columns of the matrix in the uncertainty set $\mathcal{A}$. Then the robust counterpart problem

$$\min_x \max_{\Delta A\in\mathcal{A}_a} \|b - (A + \Delta A)x\|_a$$

with $\mathcal{A}_a = \{A = (a_1, \ldots, a_n) : \|a_i\|_a \leq c_i,\ i = 1, \ldots, n\}$ is equivalent to the penalized regression problem

$$\min_x \|b - Ax\|_a + \sum_{i=1}^{n} c_i |x_i|,$$

see again Bertsimas et al. [18] where still further going generalizations are given. In problem (5.21), a solution component x_i^* becomes zero if there is a perturbation matrix $\Delta A \in \mathcal{A},\ \Delta A = (a_1, \ldots, a_n)$ with $a_i \neq 0$ that makes the i-the column of the matrix A irrelevant. This is the case, e.g., when it is "almost" orthogonal (i.e. within such a perturbation) to the vector b and all relevant columns $A_j,\ j \in J \subset \{1, \ldots, n\}$.

6. Alternative approach: Stochastic collocation with multiple shooting

In this chapter we are concerned with a superclass of the problems considered in the previous one, where we restrict our considerations to problems without inequality constraints. The latter would require some kind of active set strategy. However, due to the uncertain model parameters a standard active set strategy would not be sufficient as there is not one well-defined active set that is valid for all realizations of the uncertainty. Instead each choice of the uncertain model coefficients leads to another active set and we would have to trace all of them in one way or another in order to cope with inequality constraints in the approach that will be described in the following.

Nevertheless we can say that we treat a superclass of the equality constrained version of problem (5.1) because we now do not only treat uncertain ODE constraints and uncertain equality constraints in least squares estimation problems here, but also allow the objective function to depend on uncertain coefficients α,

$$\begin{aligned} \min_{x,p} \quad & \Phi(x,p,\alpha) := \tfrac{1}{2}\|r_1(x,p,\alpha)\|_2^2 \\ \text{s.t.} \quad & \dot{x}(t) = f(t,x,p,\alpha), \quad x(t_0) = x_0(\alpha), \quad \alpha \in \mathcal{A}, \\ & r_2(x(\theta_1),\ldots,x(\theta_K),p,\alpha) = 0, \end{aligned} \tag{6.1}$$

with the problem functions $r_1 : \mathbb{R}^{n_x} \times \mathbb{R}^{n_p} \times \mathbb{R}^{n_a} \to \mathbb{R}^{m_{r1}}$, $r_2 : \left(\mathbb{R}^{n_x}\right)^K \times \mathbb{R}^{n_p} \times \mathbb{R}^{n_a} \to \mathbb{R}^{m_{r2}}$ and the uncertainty set

$$\mathcal{A} = \bigotimes_{j=1}^{n_a} \mathcal{A}_j \quad \text{with} \quad \mathcal{A}_j = \{a : |a - \alpha_0^j| \leq \delta_j\} = [\alpha_0^j - \delta_j, \alpha_0^j + \delta_j] \subset \mathbb{R},$$

as in the previous chapter. We also lift the restriction on the number of equality constraints from the previous chapter. This is possible because the following approach takes another point of view which properties the solution of the uncertain optimization problem should have. Instead of computing the worst possible solution that still satisfies the constraints, we compute now an average solution that probably better describes the behavior of the dynamic system in the majority of instances. This idea is particularly suitable in cases where getting close to constraint bounds is not a security risk and where extreme problem instances are not expected to occur frequently.

The main idea of our approach is the following: As before, the dynamic system is discretized by means of the multiple shooting approach and thus transferred into a sequence of initial value problems along the temporal dimension. The multidimensional

uncertainty space described by the uncertainty set $\mathcal{A}$ is approximated by means of mutually orthogonal interpolation functions. In the collocation points, i.e. the nodes of the used interpolation formulae, the initial value problems have a deterministic form that can be solved with standard integrators. From these distinct deterministic solutions, which correspond to solutions in some grid points that are determined by specific realizations of the uncertainty and thus discretize the solution space, an interpolation function is computed that best approximates the average solution of the dynamic system (cf. also Ganapathysubramanian and Zabaras [53]). These two discretization strategies, multiple shooting plus stochastic collocation, reduce the dynamically constrained optimization problem to a large-scale linear equation system with a very special structure for which we will show an efficient structure exploiting solution strategy. Before we come to this, we have to address two other difficulties: The best choice of the collocation points with respect to both computational efficiency and accuracy and the construction of the multivariate interpolation function.

The chapter thus starts with a general introduction to the idea of the stochastic collocation method. As the whole approach is based on the approximation of the objective function by its expected value over the uncertainty set and as this value is approximated by means of multivariate quadrature, the second section contains a review on interpolation and quadrature in one and more dimensions. It is shown that multivariate interpolation and cubature share a common structure which can be used as a starting point for defining tensor products of these rules. The forth section then leads over to sparse grid methods that allow a significant reduction of support nodes compared to the full tensor product grids. In the fifth section we show how these sparse grid methods can be combined with the multiple shooting approach to solve problem (6.1). And in the sixth section, finally, we describe how the resulting large-scale problem is solved numerically by an adaption of the condensing algorithm from section 3.1.5.

6.1. Introduction to stochastic collocation

The method of stochastic collocation is most notably known in the context of solving partial differential equations with stochastic coefficients where it is used as an alternative to Monte Carlo methods or stochastic Galerkin methods, see e.g. Xiu and Hesthaven [141]. Monte Carlo methods, which are the classical solution methods for differential equations with random inputs, generate solutions for varying realizations of the uncertainty und compute the needed statistics from these. Their big drawback is the very slow convergence of the methods as very many deterministic problems have to be solved in this context. Galerkin methods, on the other hand are generalizations of the polynomial chaos expansion and use various polynomial or wavelet bases expansions. These methods are very fast but they result in a coupled system of differential equations which is very hard to solve. Stochastic collocation combines the fast convergence of Galerkin methods with the simple implementation of Monte Carlo methods by using multivariate polynomial interpolation in such a way that the

dynamic system only has to be solved in the collocation points. As the convergence depends only weakly on the dimension, this approach is much more efficient than Monte Carlo, and the resulting system of differential equation is much easier to solve than with Galerkin methods as it is not coupled though it may be very large-scale.

Note that for a functional F that depends on a random variable y with mean $\mathbb{E}(y)$ and standard deviation σ_y the Taylor expansion of second order of its expected value is

$$\mathbb{E}(F(y)) = F(\mathbb{E}(y)) + \frac{1}{2}\frac{\partial^2}{\partial y^2}F(\mathbb{E}(y))\sigma_y^2 + O(\sigma_y^4),$$

see Putko et al. [111]. So the approach of only inserting a nominal value for the uncertainty in the states or the cost function,

$$\begin{aligned}\mathbb{E}(x(t)) &\doteq x^{\alpha_0}(t),\\ \mathbb{E}(\Phi(x,p,\alpha)) &\doteq \Phi(x^{\alpha_0},p,\alpha_0),\end{aligned}$$

is a first-order approximation of the states or the cost function, respectively. This effectively corresponds to a negligence of the uncertain nature of the optimization problem. In order to take the uncertainty into account when solving the optimization problem, we need at least an approximation of second order.

In principle, stochastic collocation replaces the function of interest that depends on the uncertainties, i.e. in our case either the states $x(t,p,\alpha)$ or the cost function $\Phi(x,p,\alpha)$, by a linear combination, which means interpolation, of deterministic individual solutions for different realizations of the uncertainty in the collocation points (see also Nobile et al. [100]). We will restrict the further elaborations to the second possibility though everything can be transferred immediately to direct interpolation of the state variables.

Assuming that the ODE has already been reduced to a sequence of initial value problems by multiple shooting, we only have to solve the uncertain ODE or rather the uncertain initial value problem for certain realizations $\alpha_i,\ i = 1,\dots,m$, of the uncertainty,

$$\dot{x}^{\alpha_i}(t) = f(t, x^{\alpha_i}, p, \alpha_i), \quad x^{\alpha_i}(\tau_j) = s_j^{\alpha_i}, \qquad i = 1,\dots,m,\ j = 0,\dots,N.$$

Finally, a first-order statistic of the cost function $\mathbb{E}(\Phi(x,p,\alpha))$ is evaluated.

A second-order approximation of the cost function is given by

$$\begin{aligned}\tilde{\Phi}(x,p,\alpha) \approx \Phi(x^{\alpha_0},p,\alpha_0) &+ \frac{\partial}{\partial\alpha}\Phi(x^{\alpha_0},p,\alpha_0)(\alpha-\alpha_0)\\ &+ \frac{1}{2}(\alpha-\alpha_0)^T\frac{\partial^2}{\partial\alpha^2}\Phi(x^{\alpha_0},p,\alpha_0)(\alpha-\alpha_0),\end{aligned}$$

and with this follows (cp. Schulz and Schillings [124])

$$\mathbb{E}(\tilde{\Phi}(x,p,\alpha)) = \int_{\mathcal{A}}\tilde{\Phi}(x,p,\alpha)\,d\mathcal{P}(\alpha) = \Phi(x^{\alpha_0},p,\alpha_0) + \frac{1}{2}\sum_{j=1}^{n_a}\frac{\partial^2}{\partial(\alpha^j)^2}\mathrm{Var}(\alpha^j),$$

where $\mathcal{P}(\alpha)$ denotes the probability measure of the uncertainty set $\mathcal{A}$ and $\mathrm{Var}(\alpha^j)$ denotes the variance of the j-th component of the uncertainty α.

Another possibility of computing a higher-order approximation of the expected value of the cost function, which we will use in the following, is the application of an integration procedure,

$$\hat{\Phi}(x,p,\alpha) = \mathcal{I}_\Phi(x,p,\alpha) = \sum_{i=1}^{m} \Phi(x^{\alpha_i},p,\alpha_i)\mathcal{L}_i(\alpha), \tag{6.2}$$

to compute

$$\begin{aligned}\mathbb{E}(\Phi)(x,p,\alpha) &\approx \mathbb{E}(\hat{\Phi})(x,p,\alpha)\\ &= \sum_{i=1}^{m} \Phi(x^{\alpha_i},p,\alpha_i)\mathbb{E}(\mathcal{L}_i(\alpha)) = \sum_{i=1}^{m} \Phi(x^{\alpha_i},p,\alpha_i)\int_{\mathcal{A}} \mathcal{L}_i(\alpha)\,d\mathcal{P}(\alpha)\\ &= \sum_{i=1}^{m} \Phi(x^{\alpha_i},p,\alpha_i)\int_{\mathcal{A}} \mathcal{L}_i(\alpha)\rho(\alpha)\,d\alpha,\end{aligned} \tag{6.3}$$

where $\rho(\alpha)$ is the associated probability density function. This expected value is then used in the objective function of the optimization problem, which means we want to solve the optimization problem

$$\begin{aligned}\min_{x,p}\quad & \mathbb{E}(\hat{\Phi})(x,p,\alpha)\\ \text{s.t.}\quad & r_2(x(\theta_1;s_{j_1},p,\alpha),\dots,x(\theta_K;s_{j_K},p,\alpha),p,\alpha) = 0, \qquad \alpha\in\mathcal{A}.\\ & h_j(s_j,s_{j+1},\alpha) := x(\tau_{j+1};s_j,p,\alpha) - s_{j+1} = 0, \quad j=0,\dots,N-1,\end{aligned} \tag{6.4}$$

Note that in case of a linear cost function Φ it does not make a difference whether we interpolate the cost function as in (6.2) or the dynamic states while for nonlinear cost functions the results will usually differ.

The problem we are faced with now is the computation of the interpolation basis functions together with their corresponding integrals over the uncertainty set. This task is computationally rather expensive. But we can overcome this difficulty and reduce the computational complexity by choosing the same grid for the interpolation in (6.2) and the quadrature in (6.3). We then get

$$\mathbb{E}(\hat{\Phi})(x,p,\alpha) = \sum_{i=1}^{m} \Phi(x^{\alpha_i},p,\alpha_i)\omega_i,$$

with the quadrature weights ω_i that depend on the nodes $\{\alpha_i\}_{i=1}^m$.

So, the bottom line of stochastic collocation is the decoupling of the computations for the physical states from those for the uncertain coefficients, cf. Ganapathysubramanian and Zabaras [54].

Before we give a more detailed description of the stochastic collocation method and its application to our uncertain optimization problem, we shortly review interpolation and quadrature methods as these form the basis of this approach and thus determine its computational performance.

6.2. Review: Lagrange polynomial interpolation and quadrature

As the first step for a more detailed description of the stochastic collocation approach, we shortly review polynomial interpolation as well as quadrature rules as these form the basis of this method. Thus let Π_n^d be the space of all n-variate polynomials of degree at most d.

6.2.1. Polynomial interpolation methods

Univariate interpolation

Lagrange polynomial interpolation in one dimension can be summarized as follows. Given a function $f : [a, b] \to \mathbb{R}$ and a set of support nodes $a \leq y_1 < y_2 < \cdots < y_m \leq b$, there exists a unique polynomial $p_{m-1}(f) \in \Pi_1^{m-1}$ such that $p_{m-1}(f)(y_i) = f(y_i)$ for all $i = 1, \ldots, m$, and this polynomial can be written as $p_{m-1}(f)(y) = \sum_{i=1}^m f(y_i)\mathcal{L}_i(y)$ with the Lagrange basis polynomials

$$\mathcal{L}_i(y) = \prod_{\substack{j=1 \\ j \neq i}}^{m} \frac{y - y_j}{y_i - y_j}, \qquad i = 1, \ldots, m.$$

So the interpolation polynomial depends linearly on the function values $f(y_i)$, $i = 1, \ldots, m$. A similar representation with other basis polynomials is also possible, see e.g. Stoer [131]. The interpolation error for $f \in C^m([a, b])$ in a point $y \in [a, b]$ is given by $f(y) - p_{m-1}(f)(y) = (f^{(m)}(\xi))/(m!) \prod_{j=1}^m (y - y_j)$ for a $\xi = \xi(y) \in [a, b]$.

We want to mention that an efficient and numerically stable computation method for the Lagrange polynomials $\mathcal{L}_i(y)$ and thus the interpolation polynomial is given by barycentric formulae as explained by Klimke [83].

Let $\Delta = \{a = y_1 < y_2 < \cdots < y_m = b\}$ be an interval decomposition of $[a, b]$ with a corresponding interpolation polynomial $p_\Delta \in \Pi_1^{m-1}$ such that $p_\Delta(y_j) = f(y_j)$, $j = 1, \ldots, m$. Consider now a sequence of such interval decompositions $\Delta_k = \{a = y_1^{(k)} < y_2^{(k)} < \ldots, < y_{m_k}^{(k)} = b\}$ with increasing fineness $\|\Delta_k\| := \max_j |y_{j+1}^{(k)} - y_j^{(k)}|$ going to 0 for $k \to \infty$ and corresponding interpolation polynomials p_{Δ_k}. One could assume that in this case the p_{Δ_k} converge to the function f but this is not the case. On the contrary, Faber's theorem states the following (see also Stoer [131]:

Theorem 6.2.1 (Faber [47]). *For any sequence of interval decompositions Δ_k of $[a, b]$ there exists a function f that is continuous on $[a, b]$ and for which the polynomials p_{Δ_k} on $[a, b]$ do not converge uniformly to f with $k \to \infty$.*

Only for entire functions f the interpolation polynomials p_{Δ_k} for all decompositions Δ_k of $[a, b]$ converge uniformly to f for $k \to \infty$. On the other hand, for any continuous function f there exist specific sequences Δ_k such that the p_{Δ_k} converge to f, cf. Stoer [131].

In case of equidistant interval decompositions with $y_j = a + (j-1)(b-a)/(m-1)$, $j = 1, \ldots, m$, not even pointwise convergence can be guaranteed for non-entire functions f, consider e.g. $f(y) = \sqrt{y}$ on $[a, b] = [0, 1]$.

But not even uniform convergence, i.e. $\|f - p_{m-1}(f)\|_\infty \to 0$ for $m \to \infty$, can be guaranteed for arbitrary distributions of the support nodes. Nevertheless, there are node distributions for which uniform convergence can be proved for $f \in C^1([a, b])$, cf. Klimke [83].

Multivariate interpolation

It is important to note that the theory of polynomial interpolation in multivariate space is not as simple as in the one-dimensional case. First of all it is not clear which polynomial space one should consider as a suitable interpolation space. While for the univariate case there exists a unique interpolation polynomial of degree at most $m-1$ for any m distinct points in $\mathbb{R}$, there is no corresponding universal multivariate polynomial space, cf. de Boor and Ron [39]. Another difficulty is that not any set of m distinct points in n-dimensional space allows an interpolation of degree $m-1$ or possibly not even of a smaller degree.

Definition 6.2.1. *A set of points $\Theta \subset \mathbb{R}^n$, $\Theta = \{y_1, \ldots, y_m\}$ allows an interpolation of degree at most k if for any function $f : \Theta \to \mathbb{R}$ there exists a polynomial $p_{\leq k}(f)$ of degree at most k such that $p_{\leq k}(f)(y_i) = f(y_i)$, $i = 1, \ldots, m$. The point set Θ allows a unique interpolation of degree at most k if there exists exactly one such polynomial $p_{\leq k}$.*

Let $m = \binom{n+k}{k}$ and assume that the point set $\Theta = \{y_1, \ldots, y_m\} \subset \mathbb{R}^n$ is not a subset of an algebraic surface of degree at most k. Chung and Yao [36] show that this is a necessary and sufficient condition for Θ allowing a unique interpolation of degree at most k. They also show that if $m = \binom{n+k}{k}$ and if Θ allows an interpolation of degree at most k, then Θ allows a unique interpolation of degree at most k and the interpolation polynomial can be written as $p_{\leq k}(f)(y) = \sum_{i=1}^m p_i(y) f(y_i)$ with polynomials $p_i(y)$ for which the following conditions hold:

- the polynomials p_i, $i = 1, \ldots, m$, only depend on the point set Θ and not on the function f,
- $p_i(y_j) = \delta_{ij}$, where δ_{ij} denotes the Kronecker symbol,
- the polynomials p_i, $i = 1, \ldots, m$, are real polynomials of degree k,
- the polynomials p_i, $i = 1, \ldots, m$, do not have a factor of the form u^2 with a real polynomial u of degree ≥ 1.

There is also a geometric characterization of point sets allowing unique polynomial interpolation:

Condition 6.2.1 (Geometric characterization (GC), Chung and Yao [36]). *For any point $y_i \in \mathbb{R}^n$ in $\Theta = \{y_1, \ldots, y_m\} \subset \mathbb{R}^n$ with $m = \binom{n+k}{k}$ there exist k distinct hyperplanes $G_{i1}, \ldots, G_{ik}$ such that y_i does not lie on any of these hyperplanes and all other points from $\Theta \setminus \{y_i\}$ lie on at least on of these hyperplanes, i.e.*

$$y_j \in \bigcup_{l=1}^{k} G_{il} \quad \Leftrightarrow \quad i \neq j \qquad (i, j = 1, \ldots, m).$$

Theorem 6.2.2 (Chung and Yao [36]). *If the point set $\Theta \subset \mathbb{R}^n$ satisfies condition 6.2.1, then Θ allows a unique interpolation of degree at most k. Furthermore, each polynomial p_i in the representation $p_{\leq k}(f)(y) = \sum_{i=1}^{m} p_i(y) f(y_i)$ is a product of real polynomials of degree one, $p_i = u_{i1} \cdots u_{ik}$, where $u_{ij}(t) = 0$, $j = 1, \ldots, k$, is the equation of the hyperplane G_{ij} in condition 6.2.1. The reverse argument is also true: If Θ allows a unique interpolation of degree at most k and if the polynomials p_i in the representation $p_{\leq k}(f)(y) = \sum_{i=1}^{m} p_i(y) f(y_i)$ are each a product of real polynomials of degree one, then the point set Θ satisfies condition 6.2.1.*

For some examples of point sets satisfying the geometric characterization condition, see also Chung and Yao [36]. Note that at least in $\mathbb{R}^2$, condition 6.2.1 is sufficient but not necessary for a unique solvability of the Lagrange interpolation problem, cf. Gasca and Maeztu [56].

If we choose the algebra $\mathbb{R}[y^1, \ldots, y^n]$ of all n-variate polynomials with coefficients in $\mathbb{R}$ as polynomial space Π for the interpolation, then the multivariate Lagrange interpolation problem has infinitely many solutions, cf. Sauer [117]. For each specific solution $\bar{p}(f)$ the entirety of all solutions can be written as $\bar{p}(f) + I_\Theta$, where I_Θ denotes the ideal of all polynomials that vanish on the point set Θ, i.e.

$$I_\Theta := \{p \in \Pi = \mathbb{R}[y^1, \ldots, y^n] : p(y_i) = 0 \ \forall y_i \in \Theta\}.$$

For having a unique solution, a finite-dimensional subspace of $\mathbb{R}[y^1, \ldots, y^n]$ has to be determined which constitutes an interpolation space with respect to Θ. In an interpolation space, there exists a unique polynomial $p(f)$ for all sets of function values $\{f(y_i)\}_{y_i \in \Theta}$ which solves the Lagrange interpolation problem. A pair (Θ, Π) of a point set $\Theta \subset \mathbb{R}^n$ and a (polynomial) space Π of functions over $\mathbb{R}^n$ is called *correct* if for any function f that is at least defined on Θ there exists exactly one $p \in \Pi$ such that $p(y) = f(y)$ for all $y \in \Theta$. So the non-uniqueness of the multivariate interpolation corresponds to the notion that in $\mathbb{R}^n$ a correct polynomial space Π cannot be determined from the cardinality of the point set Θ alone but the distribution of the points plays a decisive role, cp. e.g. de Boor and Ron [40, 39] or Sauer [117]. The standard choice $\Pi = \Pi_n^k$ necessitates the cardinality of Θ being equal to the dimension of the polynomial space, $|\Theta| = \dim \Pi_n^k = \binom{n+k}{n}$, but this condition alone does not guarantee correctness of (Θ, Π_n^k). A polynomial space Π is called *minimal correct* for the point set Θ (or is *solution of minimal degree*) if Π is contained in the set of solutions of the Lagrange interpolation problem, $IP(\Theta)$, based on Θ and if $\dim(P \cap \Pi_n^k) \leq \dim(\Pi \cap \Pi_n^k)$ for all $P \in IP(\Theta)$ and all $k \in \mathbb{N}$.

The approximation quality of an interpolation polynomial of degree $m-1$ is assessed by comparing it to the interpolation polynomial p^*_{m-1} for which the support nodes have been chosen in an optimal way, i.e. that polynomial for which the maximal interpolation error is minimal,

$$\|f - p^*_{m-1}(f)\|_\infty \leq \|f - p_{m-1}(f)\|_\infty \qquad \forall p_{m-1}(f).$$

The distance between any interpolation polynomial and this best approximation polynomial is measured by means of the Lebesgue constant Λ,

$$\|f - p_{m-1}(f)\|_\infty \leq (1+\Lambda)\|f - p^*_{m-1}(f)\|_\infty,$$

where $\Lambda = \sup_{f\neq 0} \|p_{m-1}(f)\|_\infty / \|f\|_\infty = \|\sum_{i=1}^m \mathcal{L}_i(y)\|_\infty$ with the Lagrange basis polynomials $\mathcal{L}_i(y)$, cf. Klimke [83]. So, obviously, the Lebesgue constant and thus the measure for the approximation quality of the interpolation polynomial does not depend on the function f but only on the distribution of the support nodes $y_1, \ldots, y_m$.

Xiu and Hesthaven [141] mention that the optimal distribution of the support nodes is not explicitly known and the determination of the best approximation polynomial p^*_{m-1} for a given function f is still an unsolved problem. Nevertheless, some sets of support nodes with a small Lebesgue constant are known, which are well suited for practical applications, like e.g. the extrema of the Chebyshev polynomials, see e.g. Barthelmann et al. [10], Xiu and Hesthaven [141], Borzì [29], and Borzi et al. [30],

$$y^j = -\cos\frac{(j-1)\pi}{m-1}, \quad j = 1, \ldots, m, \tag{6.5}$$

for $m > 1$. In case of $m = 1$ one usually chooses $y^1 = 0$, i.e. the center point of the interpolation interval. In the literature, these points are also called Chebyshev Gauss-Lobatto points by Klimke [83] or Clenshaw-Curtis points by Nobile et al. [100] and Schulz and Schillings [124] for specific choice of the number m, see below.

Polynomials are only appropriate for local approximations and thus the interpolation interval should not be too large. As we apply a multiple shooting technique, which divides the possibly large range of the dynamic system in our optimization problem into smaller subintervals, this should not be a problem in our applications where we essentially treat each subinterval independently. Poorly performed polynomial approximations, e.g. using equidistant support nodes, are practically never recommendable, while good polynomial approximation methods, e.g. using the extrema of Chebyshev polynomials as support nodes, belong to the most efficient local approximation procedures, cf. de Boor and Ron [39].

6.2.2. Univariate quadrature rules

The most elementary univariate integration methods, i.e. Newton-Cotes formulae, replace the integrand f in $I(f) = \int_a^b f(y)\,dy$ by its interpolation polynomial $p_m(y)$ with equidistant support nodes,

$$Q(f) = \int_a^b p_m(y)\,dy = \sum_{j=1}^m f(y_j) \int_a^b \mathcal{L}_j(y)\,dy = \sum_{i=1}^m f(y_j)\omega_j.$$

It is commonly known that Newton-Cotes formulae of degree m exactly integrate all polynomials $p \in \Pi_1^m$ of degree at most m for m odd. For m even, all polynomials $p \in \Pi_1^{m+1}$ of degree at most $m+1$ are integrated exactly. The integration error for f can be written as

$$R_n(f) = \begin{cases} \frac{R_n(y^{m+1})}{(m+1)!} f^{(m+1)}(\xi), & \text{for } m \text{ odd}, \\ \frac{R_n(y^{m+2})}{(m+2)!} f^{(m+2)}(\xi), & \text{for } m \text{ even}, \end{cases}$$

with some $\xi \in [a,b]$.

Considering more general integrals $I(f) = \int_a^b \omega(y) f(y)\, dy$ with a general weight function ω on $[a,b]$ allows the development of higher-order quadrature formulae. Therefore, the weight function is assumed to satisfy the following prerequisites, see Stoer [131]:

- $\omega(y) \geq 0$ on $[a,b]$ and measureable.
- All moments $\mu_k = \int_a^b y^k \omega(y)\, dy,\ k = 0, 1, \ldots$, exist and are finite.
- For any polynomial $s(y)$ with $\int_a^b \omega(y) s(y)\, dy = 0$ and $s(y) \geq 0$ for $y \in [a,b]$ follows $s(y) \equiv 0$.

These conditions are satisfied, e.g., for all weight functions $\omega(y)$ that are positive and continuous on $[a,b]$.

Let now $\bar{\Pi}_1^m$ denote the set of normed real polynomials of degree m, i.e.

$$p \in \bar{\Pi}_1^m \quad \Leftrightarrow \quad p(y) = y^m + b_{m-1} y^{m-1} + \cdots + b_1 y + b_0.$$

Let further $p_i \in \bar{\Pi}_1^i$ denote polynomials that are orthogonal with respect to the scalar product induced by $\omega(y)$,

$$\langle f, g \rangle := \int_a^b \omega(y) f(y) g(y)\, dy,$$

i.e. $\langle p_i, p_j \rangle = 0$ for $i \neq j$, and such that they are defined by the following recursion,

$$\begin{aligned} p_0(y) &= 1, \\ p_{i+1}(y) &= (y - \delta_{i+1}) p_i(y) - \gamma_{i+1}^2 p_{i-1}(y), \quad i \geq 1, \quad \text{where } p_{-1}(y) := 0, \\ \delta_{i+1} &:= \frac{\langle y p_i, p_i \rangle}{\langle p_i, p_i \rangle}, \\ \gamma_{i+1}^2 &:= \begin{cases} 1, & \text{for } i = 0, \\ \frac{\langle p_i, p_i \rangle}{\langle p_{i-1}, p_{i-1} \rangle}, & \text{for } i \geq 1. \end{cases} \end{aligned}$$

With this definition, each polynomial $p \in \Pi_1^k$ can be written as a linear combination of such $p_i,\ i \leq k$, and thus we have $\langle p, p_m \rangle = 0$ for all $p \in \Pi_1^{m-1}$.

All zeros $y_i,\ i = 1, \ldots, m$, of p_m are real, simple, and lie in the open interval (a, b). The interpolation task of finding a polynomial p of the form $p(y) = \sum_{j=0}^{m-1} c_j p_j(y)$ such that $p(y_i) = f(y_i),\ i = 1, \ldots, m$, is always uniquely solvable. For the integration with these basis polynomials the following theorem holds.

Theorem 6.2.3 (Stoer [131]). *Let $y_1, \ldots, y_m$ be the zeros of p_m and let $\omega_1, \ldots, \omega_m$ solve the linear equation system*

$$\sum_{i=1}^{m} p_k(y_i)\omega_i = \begin{cases} \langle p_0, p_0 \rangle, & \text{for } k = 0, \\ 0, & \text{for } k = 1, \ldots, m-1. \end{cases} \tag{6.6}$$

Then $\omega_i > 0,\ i = 1, \ldots, m$, and

$$\int_a^b \omega(y) p(y)\, dy = \sum_{i=1}^{m} \omega_i p(y_i) \tag{6.7}$$

for all polynomials $p \in \Pi_1^{2m-1}$.

Reversely, it holds that if for certain $\omega_i, y_i,\ i = 1, \ldots, m$, equation (6.7) is satisfied for all $p \in \Pi_1^{2m-1}$, then the y_i are the zeros of p_m and the ω_i solve system (6.6).

There are no real numbers $y_i, \omega_i,\ i = 1, \ldots, m$, such that (6.7) holds for all $p \in \Pi_1^{2m}$.

The zeros $y_i,\ i = 1, \ldots, m$, of p_m are the eigenvalues of the tridiagonal matrix

$$J_m = \begin{pmatrix} \delta_1 & \gamma_2 & & \\ \gamma_2 & \delta_2 & \ddots & \\ & \ddots & \ddots & \gamma_m \\ & & \gamma_m & \delta_m \end{pmatrix}.$$

The weights are $\omega_i = (v_1^{(i)})^2,\ i = 1, \ldots, m$, where $v^{(i)} = (v_1^{(i)}, \ldots, v_m^{(i)})^T$ is the eigenvector to the eigenvalue y_i of J_m, normed such that $(v^{(i)})^T v^{(i)} = \langle p_0, p_0 \rangle = \int_a^b \omega(y)\, dy$, cf. Stoer [131].

For $f \in C^{2m}([a, b])$ this Gaussian integration method results in an error term of the form

$$\int_a^b \omega(y) f(y)\, dy - \sum_{i=1}^{m} \omega_i f(y_i) = \frac{f^{(2m)}(\xi)}{(2m)!} \langle p_m, p_m \rangle$$

with some $\xi \in [a, b]$.

Some common weight functions $\omega(y)$ and the corresponding orthogonal polynomials $p_i,\ i = 0, 1, \ldots$, are given in table 6.1.

We will see below that in the stochastic collocation approach with sparse grids it is advantageous to have nested sets of support nodes for the univariate quadrature rules, i.e. $\Theta_i \subset \Theta_{i+1}$. This condition is not satisfied for Gaussian quadrature formulae and with each increasing of the quadrature order of the univariate formula the optimal

Table 6.1.: Common quadrature weight functions and their corresponding orthogonal polynomials (adapted from Stoer [131]).

$[a, b]$	$\omega(y)$	p_k
$[-1, 1]$	1.0	Legendre polynomials $P_k(y)$
$[-1, 1]$	$(1 - y^2)^{-1/2}$	Chebyshev polynomials $T_k(y)$
$[0, \infty]$	e^{-y}	Laguerre polynomials $L_k(y)$
$[-\infty, \infty]$	e^{-y^2}	Hermite polynomials $H_k(y)$

support nodes have to be computed completely from scratch. This is why in the following our main focus will lie on the quadrature rule described by Clenshaw and Curtis [37]. Clenshaw-Curtis quadrature is an interpolatory quadrature rule of the usual form $\int_{-1}^{1} f(y)\,dy = \sum_{i=1}^{m} \omega_i f(y_i) + R_m$, i.e. with $\omega(y) \equiv 1.0$ and with R_m being the approximation error, which uses the extrema of the Chebyshev polynomials $T_m(y)$ as support nodes, see equation (6.5), instead of the optimal points as in the case of Gauss quadrature. There are three main variants of this kind of formula, cf. Trefethen [135]:

- the "classical" formula by Fejér [49] that uses the roots of the Chebyshev polynomials in the open interval $(-1, 1)$,
- the formula 2 by Fejér [49] that uses the Chebyshev extrema in $(-1, 1)$, and
- the "practical" formula by Clenshaw and Curtis [37] that employs the extrema of the Chebyshev polynomials in the closed interval $[-1, 1]$.

Considering the fact that the two Fejér formulae were developed almost thirty years before the Clenshaw-Curtis rule, Trefethen [135] correctly remarks that this class of quadrature formulae should have been named Fejér quadrature or at least Chebyshev quadrature according to the used orthogonal polynomials. Nevertheless, Clenshaw-Curtis quadrature is the name that is commonly used in the literature and most authors simply neglect any reference to the older paper by Fejér [49]. A great advantage of this kind of quadrature rules over Gauss quadrature is not only that they yield nested sets of support nodes for growing integration order, but their points and weights are also computable very efficiently using fast Fourier transforms. The computation of Gaussian support nodes and quadrature weights takes a multiple of the time. The fact that these Gauss points are not reusable for increasing order of quadrature accuracy even worsens the situation when applying such rules to multivariate quadrature based on sparse grids. A comparison of one-dimensional Clenshaw-Curtis and Gaussian quadrature rules is also given by Trefethen [135]. Note, however, that there are variants of Gauss quadrature, like Gauss-Patterson formulae, which use nested sets of support nodes and are thus well-suited for multivariate quadrature on sparse grids, cf. Gerstner and Griebel [61].

Theoretically, Clenshaw-Curtis quadrature rules have the same order of integration accuracy as Newton-Cotes formulae, i.e. the formula that is based on $m + 1$ support

nodes integrates exactly all polynomials of degree m. In practice, on the other hand, it usually works much more accurately and comes close to the accuracy of Gaussian quadrature rules for integrands f that are not analytic in a larger neighborhood of the integration interval $[-1, 1]$. This is not only true for the order of integration accuracy, but also holds for the rate of convergence and the number of function evaluations needed as observed by Trefethen [135]. A first try to explain this behavior was given by O'Hara and Smith [108] and is based on the magnitude of the coefficients in the Chebyshev polynomials. Two further lines of thought concerning this can be found in Trefethen [135].

The weights ω_i of the Clenshaw-Curtis rule can be determined analytically by integrating the interpolation polynomial of degree $m-1$ through the m nodes $(y_i, f(y_i))$. Explicit weights for these quadrature rules are given by Waldvogel [137] who cites Gautschi [58] as the original source that also gives positive bounds for the following weights. For the first Fejér formula, the weights are

$$\omega_i^{f1} = \frac{2}{m-1}\left(1 - 2\sum_{j=1}^{\lfloor\frac{m-1}{2}\rfloor}\frac{1}{4j^2-1}\cos\frac{j(2i-1)\pi}{m-1}\right), \quad i = 1, \ldots, m-1,$$

while for the second Fejér formula they are

$$\omega_i^{f2} = \frac{4}{m-1}\sin\frac{(i-1)\pi}{m-1}\sum_{j=1}^{\lfloor\frac{m-1}{2}\rfloor}\frac{\sin(2j-1)\frac{(i-1)\pi}{(m-1)}}{2j-1}, \quad i = 1, \ldots, m,$$

which automatically gives $\omega_1^{f2} = \omega_m^{f2} = 0$ excluding the boundary points. The weights for the Clenshaw-Curtis formula are

$$\omega_i^{cc} = \frac{c_i}{m-1}\left(1 - \sum_{j=1}^{\lfloor\frac{m-1}{2}\rfloor}\frac{b_j}{4j^2-1}\cos\frac{2j(i-1)\pi}{m-1}\right), \quad i = 1, \ldots, m, \tag{6.8a}$$

where

$$b_j = \begin{cases} 1, & j = \frac{m-1}{2}, \\ 2, & j < \frac{m-1}{2}, \end{cases}, \qquad c_i = \begin{cases} 1, & i = 1 \mod (m-1), \\ 2, & \text{otherwise.} \end{cases} \tag{6.8b}$$

Note that this leads to the expression $\omega_1^{cc} = \omega_m^{cc} = 1/\big((m-1)^2 - 1 + \mod(m-1, 2)\big)$. A proof of convergence for functions with singularities at the interval boundaries as well as a comparison of the first Fejér formula with the Gauss-Chebyshev and the midpoint rule are also given by Gautschi [58].

As we already mentioned above, these quadrature weights can be computed more efficiently using discrete Fourier transformation (DFT), cf. Waldvogel [137]. Let $\omega := \exp(\frac{2i\pi}{m-1})$ and

$$s_k := \begin{cases} (-1)^k(1-\omega^k), & \text{for } m-1 \text{ even}, \\ (1-\omega^k), & \text{for } m-1 \text{ odd}, \end{cases} \quad k = 1, \ldots, m-1,$$

yielding the matrix $S_m = \mathrm{diag}(s_k)$. Define further

$$u := \begin{cases} (-\frac{1}{m-1}, \ldots, -\frac{1}{3}, -1, 1\frac{1}{3}, \ldots, \frac{1}{m-1}), & \text{for } m-1 \text{ even}, \\ (1, \frac{1}{3}, \ldots, \frac{1}{m-2}, 0, -\frac{1}{m-1}, \ldots, -\frac{1}{3}, -1), & \text{for } m-1 \text{ odd}, \end{cases}$$

and the DFT matrices $F_m = (\omega^{-kl})_{k,l=1}^{m-1}$ and $F_m^{-1} = \frac{1}{m}(\omega^{kl})_{k,l=1}^{m-1}$. Then the vector ω^{f2} of weights for the second formula by Fejér can be computed as

$$\omega^{f2} = S_m F_m^{-1} u.$$

A further simplification can be achieved by defining the matrix $T_m = (t_{kl})_{k,l=1}^{m-1}$ with

$$t_{kl} := \begin{cases} 1, & \text{for } l-k = h_m \mod (m-1), \\ -1, & \text{for } l-k = (h_m - 1) \mod (m-1), \\ 0, & \text{otherwise}, \end{cases}$$

where

$$h_m := \begin{cases} \frac{m-1}{2}, & \text{for } m-1 \text{ even}, \\ 0, & \text{for } m-1 \text{ odd}. \end{cases}$$

Using now the definition $v^{f2} = T_m u$, we get

$$\begin{aligned} v_k^{f2} &= \frac{2}{1-4k^2}, \quad k = 1, \ldots, \lfloor \frac{m-1}{2} \rfloor - 1, \\ v_{\lfloor \frac{m-1}{2} \rfloor}^{f2} &= \frac{m-4}{2\lfloor \frac{m-1}{2} \rfloor - 1} - 1, \\ v_{n-1-k}^{f2} &= \bar{v}_k, \quad k = 1, \ldots, \lfloor \frac{m-2}{2} \rfloor, \end{aligned}$$

and $\omega^{f2} = F_m^{-1} v^{f2}$, where $\bar{v}$ denotes the complex conjugate of v. The quadrature weights for the first Fejér rule can be computed similarly but with the vector v defined by

$$\begin{aligned} v_k^{f1} &= \frac{2}{1-4k^2} e^{ik\pi/(m-1)}, \quad k = 1, \ldots, \lfloor \frac{m-2}{2} \rfloor, \\ v_{\frac{m-1}{2}}^{f1} &= 0, \\ v_{n-1-k}^{f1} &= \bar{v}_k, \quad k = 1, \ldots, \lfloor \frac{m-2}{2} \rfloor. \end{aligned}$$

The weight vector ω^{cc} for the Clenshaw-Curtis rule is given by $\omega^{cc} = F_m^{-1}(v^{f2} + g)$ where the vector g is defined by

$$\begin{aligned} g_k &= -\omega_0^{cc}, \quad k = 1, \ldots, \lfloor \frac{m-1}{2} \rfloor - 1, \\ g_{\lfloor \frac{m-1}{2} \rfloor} &= \omega_0^{cc}[(2 - \mod (m-1, 2))(m-1) - 1], \\ g_{m-1-k} &= \bar{g}_k, \quad k = 1, \ldots, \lfloor \frac{m-2}{2} \rfloor. \end{aligned}$$

Waldvogel [137] even gives the MATLAB code for computing these weights by means of the MATLAB routine for inverse fast Fourier transformations.

As in the case of numerical interpolation, also for numerical quadrature the problem gets more complicated as the number of variables is increased above one. Necessary and sufficient conditions concerning the distribution of support nodes for a quadrature formula of degree 2 or 3, respectively, in m-dimensional Euklidean space are given in the following theorems.

Theorem 6.2.4 (Stroud [134]). *A necessary and sufficient condition for m equally weighted points $y_1, \ldots, y_m$ giving a numerical integration formula of degree 2 on a symmetric region or a regular $(m-1)$-dimensional simplex R is that the points must form the vertices of a regular $(m-1)$-dimensional simplex of which the center of mass coincides with the center of mass of the considered region and which lie on the surface of a ball with radius $r = \sqrt{(m-1)I_2/I_0}$ with $I_0 = \int_R 1\,dv$, $I_2 = \int_R y_2^2\,dv = \cdots = \int_R y_m^2\,dv$.*

Theorem 6.2.5 (Stroud [134]). *A necessary and sufficient condition for an even number of m points $y_1, \ldots, y_m$ with $y_{m-j} = -y_{1+j}$, $j = 0, \ldots, \frac{m}{2} - 1$, giving an equally weighted integration formula of degree 3 on a symmetric region R is that these points must form the vertices of a region $Q_{\frac{m}{2}}$ that is defined by the $2^{\frac{m}{2}}$ inequalities $\pm y_1 \pm y_2 \pm \cdots \pm y_{\frac{m}{2}} \leq a$ (or a rotation thereof) and of which the center of mass coincides with the center of mass of the integration region and which lie on a ball with radius $r = \sqrt{\frac{m}{2} I_2/I_0}$ with I_0 and I_2 being defined as in theorem 6.2.4.*

Stroud [134] constructed from these theorems two cubature formulae of degree two and three, respectively. The degree-2 formula is based on the support nodes $\Gamma_k = (\gamma_k^1, \ldots, \gamma_k^n)$, $k = 0, \ldots, n$, with

$$\gamma_k^{2r-1} = \sqrt{\frac{2}{3}} \cos \frac{2rk\pi}{n+1}, \quad \gamma_k^{2r} = \sqrt{\frac{2}{3}} \sin \frac{2kr\pi}{n+1}, \quad r = 1, \ldots, \lfloor \frac{n}{2} \rfloor,$$

and for n odd $\gamma_k^n = (-1)^k/\sqrt{3}$. The degree-3 formula has the support nodes $\Sigma_k = (\sigma_k^1, \ldots, \sigma_k^n)$, $k = 1, \ldots, 2n$, with

$$\sigma_k^{2r-1} = \sqrt{\frac{2}{3}} \cos \frac{(2r-1)k\pi}{n}, \quad \sigma_k^{2r} = \sqrt{\frac{2}{3}} \sin \frac{(2r-1)k\pi}{n}, \quad r = 1, \ldots, \lfloor \frac{n}{2} \rfloor,$$

and for n odd $\sigma_k^n = (-1)^k/\sqrt{3}$. Both formulae are not only exact for all n-variate polynomials of degree 2 or 3, respectively, but also use the minimal number of points for the respective integration accuracy, cf. Xiu and Hesthaven [141].

Applying the tensor product approach described below, any univariate quadrature formula can be used in multivariate intergration and with it also any corresponding distribution of nodes that ensures a high degree of exactness for this quadrature rule.

With Π_n^d denoting the space of all n-variate polynomials of total degree $\leq d$ as before we have $\dim \Pi_n^d = \binom{d+n}{n}$. Let now N_n^d denote the minumum number of nodes

that is needed by any quadrature formula Q_n of degree of exactness d. Then the following bounds hold for this number (see Novak and Ritter [104]),

$$\dim \Pi_n^{\lfloor \frac{d}{2} \rfloor} \leq N_n^d \dim \Pi_n^d. \tag{6.9}$$

6.2.3. Common structure

Both univariate interpolation and quadrature rules can be written in the general form

$$\mathcal{U}(f)(y) = \sum_{j=1}^{m} f(y_j) a_j,$$

where a_j denotes the interpolation basis functions or the quadrature weights, respectively. For polynomial interpolation, the basis functions a_j are supposed to be continuous on the whole interpolation interval and must satisfy $a_j(y_j) = 1$ and $a_j(y_k) = 0$ for $k \neq j$. For quadrature rules, the weights are assumed to be chosen in such a way that the formula integrates all polynomials of degree m exactly. In the following section we will describe how univariate interpolation and quadrature formulae can be combined to multivariate rules using tensor products.

6.3. Tensor product formulae

In the following we will restrict our considerations to the case where the interpolation or quadrature interval is the unit interval $[-1, 1]$ in one dimension or its n-dimensional analogue $[-1, 1]^n$. As each general interval $[a, b]$ can be transferred to the unit interval by a linear transformation $g : [a, b] \to [-1, 1], g(t) = -\frac{2}{a-b}t + \frac{a+b}{a-b}$, this is no loss of generality. The re-transformation from $[-1, 1]$ to $[a, b]$ corresponds to the inverse function $g^{-1} : [-1, 1] \to [a, b], g^{-1}(t) = -\frac{a-b}{2}t + \frac{a+b}{2}$. For the multivariate case, analogous arguments hold.

Let $\Theta_n = \Theta_1^1 \times \cdots \times \Theta_1^n$ denote the tensor product grid of the n one-dimensional point sets $\Theta_1^i \subset [-1, 1]$, $i = 1, \ldots, n$, that means, Θ_n contains all points $y = (y_1^{i_1}, \ldots, y_n^{i_n})$ with $y_j^{i_j} \in \Theta_1^j$. Then the tensor product space

$$P_n = \mathrm{span}_{\mathbb{R}}\{y^{\mathbf{i}} : \mathbf{i} = (i_1, \ldots, i_n),\ i_j = 1, \ldots, m_j,\ m_j = |\Theta_1^j|\},$$

that is spanned by all these points in $\mathbb{R}^n$ is an interpolation space for the support nodes Θ_n, see Sauer [117]. Of course, there are also other interpolation spaces for Θ_n but *least interpolation* as described by de Boor and Ron [39] for getting minimal correct interpolation spaces always leads to tensor product spaces. The tensor product space P_n contains the multivariate Lagrange basis polynomials,

$$\mathcal{L}_{\mathbf{i}}(y) = \frac{\prod_{j=1}^{n} \prod_{k_j=1}^{m_j} (y_j - y_j^{k_j})}{\prod_{j=1}^{n} (y_j - y_j^{i_j})},$$

which vanish exactly in $\Theta_n \setminus \{y^{\mathbf{i}}\}$ but not in $y^{\mathbf{i}}$.

Equivalently, this interpolation space can be characterized as the space $P_l([-1,1]^n)$ of degree $\leq l = (l_1, \ldots, l_n)$ that is spanned by all tensor product polynomials over $[-1,1]^n$,

$$P_l([-1,1]^n) = \bigotimes_{j=1}^{n} P_{l_j}([-1,1]),$$

where $P_{l_j}([-1,1]) = \text{span}\{y \in \Theta_1^j,\ j = 1, \ldots, n\}$. From this representation we can directly see that its dimension is $\dim P_l([-1,1]^n) = \prod_{j=1}^n m_j$ with $m_j = |\Theta_1^j|$. Furthermore, Nobile et al. [100] show that $P_l([-1,1]^n) \subset L_{\mathcal{P}}(\alpha)^2([-1,1]^n)$ where $L_{\mathcal{P}}(\alpha)^2([-1,1]^n)$ denotes the (stochastic) Banach space of Banach-valued functions with finite second moments for the probability measure $\mathcal{P}(\alpha)$ of the uncertainty space $\mathcal{A}$.

Accordingly, we can construct multivariate formulae for interpolation or quadrature using tensor products. So let $\Theta_1^i = \{y_1^i, \ldots, y_{m_i}^i\} \subset \mathbb{R}$ denote the set of support nodes for the one-dimensional formula

$$\mathcal{U}^i(f)(y) = \sum_{j=1}^{m_i} a_j^i f(y_j^i).$$

Assume further that we have given a sequence of such formulae such that $\mathcal{U}^{i+1}$ has a smaller error than $\mathcal{U}^i$ and $m_{i+1} > m_i$, cf. Novak and Ritter [103]. Then the multivariate tensor product formula is

$$(\mathcal{U}^{i_1} \otimes \cdots \otimes \mathcal{U}^{i_n})(f)(y) = \sum_{j_1=1}^{m_1} \cdots \sum_{j_n=1}^{m_n} (a_{j_1}^{i_1}(y) \otimes \cdots \otimes a_{j_n}^{i_n}(y)) f(y_{j_1}^{i_1}, \ldots, y_{j_n}^{i_n}).$$

Obviously, this formula needs $M = \prod_{j=1}^n m_j$ support nodes on a regular grid. This number rapidly grows with increasing dimension and with the number of points used for each univariate formula. Therefore, this formula only works efficiently for low dimensional problems and one-dimensional basis functions that use only few points. On the other hand, the formula has the advantage that it is easy to compute as the collocation points and space are tensor products of the univariate ones. Note also that univariate Lagrange interpolation polynomials can be easily extended to the higher dimensional case, see Xiu and Hesthaven [141].

Defining $\mathcal{U}^0 := 0$ and the difference formulae $\Delta^i := \mathcal{U}^i - \mathcal{U}^{i-1}$, offers a way for reformulating the tensor product formula. Using the telescopic property of the difference formulae, we can replace the univariate formula $\mathcal{U}^i$ by a sum of difference formulae according to

$$\begin{aligned} \mathcal{U}^i(f) &= \mathcal{U}^i(f) - \mathcal{U}^{i-1}(f) + \mathcal{U}^{i-1}(f) - \cdots - \mathcal{U}^i(f) + \mathcal{U}^1(f) - \mathcal{U}^0(f) \\ &= \sum_{j=1}^{i} \Delta^j(f). \end{aligned} \tag{6.10}$$

The tensor product formula then transforms to

$$\mathcal{U}_n(f) = (\mathcal{U}^{i_1} \otimes \cdots \otimes \mathcal{U}^{i_n})(f) = \left(\sum_{j_1=1}^{i_1} \Delta^{j_1} \otimes \cdots \otimes \sum_{j_n=1}^{i_n} \Delta^{j_n} \right) (f)$$
$$= \sum_{j_1=1}^{i_1} \cdots \sum_{j_n=1}^{i_n} (\Delta^{j_1} \otimes \cdots \otimes \Delta^{j_n})(f). \tag{6.11}$$

Smolyak's method for sparse grid construction now employs this tensor product formula in a special way to reduce the required number of support nodes by several orders of magnitude at best.

However, non-tensorial cubature formulae of Clenshaw-Curtis type are even more exact than tensor product Clenshaw-Curtis or Gauss-Legendre formulae, see Sommariva et al. [129]. This construction is based on the concept of hyperinterpolation as described by Sloan [126]. We do not follow this approach any further here but mention it just for completeness.

The full n-dimensional grid of order I with $i_1 = \cdots = i_n = I$ is formed by means of the tensor product of the underlying one-dimensional grids and has the total order $\sum_{j=1}^{n} i_j = nI$. Sparse grids of order I, as described in the next section, are formed by strict subsets of this full grid, where I denotes the order of the largest admissible grid and the orders of all underlying one-dimensional grids sum up to a total order of $I + n - 1$, cp. e.g. Borzì [29].

6.4. Sparse grids

The general idea for the construction of sparse grids is to take the telescopic sum (6.10) or (6.11), respectively, and leave out some terms such that effort and accuracy are in adequate proportion. Different rules, which terms are to be left out and wich are to be kept in the approximation, lead to different sparse grids. We will concentrate on Smolyak's construction here and refer to Gerstner and Griebel [62] for generalizations and to Holtz [75] for other examples.

Sparse grid methods can also be seen as a specific discretization technique that compensates the curse of dimensionality of grid based approaches up to a certain degree. They have been applied successfully in the context of interpolation and approximation of functions, integral equations, partial differential equations and stochastic differential equations, the Maxwell equation or parabolic problems. A modification of the method, using a combination technique instead of the usual hierarchical bases, has been used for eigenvalue problems, numerical integration, options pricing, machine learning, data analysis, or solving the stochastic master equation in gene regulation networks, see Garcke [55] and the references therein.

Related concepts that have been developed independently for different application areas but are all based on the Smolyak construction, which will be explained in this section, are Biermann interpolation, hyperbolic crosses, Boolean methods, discrete

blending methods, or splitting extrapolation methods, see Novak and Ritter [102, 103] and the references given by them.

Smolyak's construction, which we will denote as $A(q,n)$, is a linear combination of tensor product formulae that satisfies two requirements:

(i) they only use formulae with a comparatively small number of points and

(ii) the linear combinations are chosen such that the interpolation or integration property for one dimension is preserved for multiple dimensions.

For $q \geq n$ the Smolyak construction is

$$\begin{aligned} A(q,n) &= \sum_{|\mathbf{i}|\leq q} (\Delta^{i_1} \otimes \cdots \otimes \Delta^{i_n}) \\ &= \sum_{q-n+1\leq|\mathbf{i}|\leq q} (-1)^{q-|\mathbf{i}|} \binom{n-1}{q-|\mathbf{i}|} (\mathcal{U}^{i_1} \otimes \cdots \otimes \mathcal{U}^{i_n}), \end{aligned} \tag{6.12}$$

where $\mathbf{i} = (i_1, \ldots, i_n) \in \mathbb{N}^n$ and $|\mathbf{i}| = \sum_{j=1}^n i_j$. The index i_j gives the order of the univariate formula used for direction j. So the algorithm combines only the tensor products of those one-dimensional formulae the indices of which satisfy the constraint that the total sum over all directions is greater than $q-n$ but less than $q+1$. Note that in case of a cubature formula being computed by means of Smolyak's construction the quadrature weights are replaced by the difference weights $\eta^i = a^i - a^{i-1}$ and are thus no longer all positive as is the case in standard univariate or tensor product quadrature formulae.

When simply computing a multivariate integral with this approach, one has to bear in mind that these negative weights can lead to numerical cancellation. Gerstner and Griebel [61] suggest a coordinate-wise summation instead of summing up with increasing q because their numerical experiments led to much smaller rounding errors in that case. When combining Smolyak's sparse grid construction with multiple shooting, as will be done in the next section, we have to resort the sum in (6.12) anyway. This relates to the fact that in the formulation given above some points or function values $f(y_{j_1}^{i_1}, \ldots, y_{j_n}^{i_n})$ can occur several times with differents weights and we will have to group these together to have just one summand for each point in the sparse grid.

Obviously, the computation of $A(q,n)$ only needs function evaluations on the sparse grid

$$H(q,n) = \bigcup_{q-n+1\leq|\mathbf{i}|\leq q} (\Theta_1^{i_1} \times \cdots \times \Theta_1^{i_n}), \tag{6.13}$$

where $\Theta_1^{i_j} = \{y_i^1, \ldots, y_i^{m_i}\} \subset [-1,1]$ denotes the support nodes used by formula $\mathcal{U}_i$, and the Smolyak construction $A(q,n)$ depends linearly on the function values on this union. Formula $A(q,n)$ is called the *level q interpolant/integral in n dimensions* and $H(q,n)$ is called the *sparse grid of level $q-n$*.

Note that this construction by Smolyak cannot only be applied to tensor products of integration or quadrature formulae. It was originally developed by Smolyak [128] and is not limited to the task considered here. A more general analysis for mappings on general Cartesian products of arbitrary sets is given by Schreiber [122]. The tensor product operator considered in this thesis is only a special case, although it is the one treated most frequently in the literature. For general vector spaces $\mathcal{X}_1, \ldots, \mathcal{X}_{q-n+1}$, the *Smolyak space* is the vector space $S\mathcal{V}_{q,n}^{\mathcal{X}_1,\ldots,\mathcal{X}_{q-n+1}} := \sum_{|\mathbf{i}|\leq q} \big((\mathcal{X}_{i_1} - \mathcal{X}_{i_1-1}) \otimes \cdots \otimes (\mathcal{X}_{i_n} - \mathcal{X}_{i_n-1})\big)$.

A list of 36 different univariate quadrature rules that can be used for constructing cubature formulae on sparse grids are listed by Burkhardt [33]. For each formula not only its standard interval and the corresponding weight function are given, but also the number of points used for several orders and accuracies of the integration together with further properties of the formulae (open or closed, nested or not, stability and the like).

Ganapathysubramanian and Zabaras [53] show that the Smolyak construction can also be written in the form

$$A(q,n) = \sum_{|\mathbf{i}|\leq q} (\Delta^{i_1} \otimes \cdots \otimes \Delta^{i_n}) = A(q-1,n) + \sum_{|\mathbf{i}|=q} (\Delta^{i_1} \otimes \cdots \otimes \Delta^{i_n}),$$

where $A(n-1,n) := 0$. Thus it allows the reuse of all previously computed results when going from level $q-1$ to q. In case of nested formulae, i.e. with $\Theta_1^i \subset \Theta_1^{i+1}$, e.g. when using the extrema of the Chebyshev polynomials in Clenshaw-Curtis quadrature, the sparse grid (6.13) even simplifies to

$$H(q,n) = \bigcup_{|\mathbf{i}|=q} (\Theta_1^{i_1} \times \cdots \times \Theta_1^{i_n}),$$

as in this case also $H(q,n) \subset H(q+1,n)$. The points $y \in H(q,n)$ are called *hyperbolic crosses*, see also Novak and Ritter [102]. Having already computed formula $\mathcal{U}^i$, the step to $\mathcal{U}^{i+1}$ only requires function evaluations on the difference grid $\Delta H(q+1,n)$ that contains the points which are in $H(q+1,n)$ but were not in $H(q,n)$. Denote by $\Delta\Theta$ the newly needed points, i.e. $\Delta\Theta_1^{i+1} = \Theta_1^{i+1}$ for non-nested formulae and $\Delta\Theta_1^{i+1} = \Theta_1^{i+1} \backslash \Theta_1^i$ for nested formulae. Then the step from the Smolyak construction of order q to order $q+1$ requires only function evaluations on the diffenence grid

$$\Delta H(q+1,n) = \bigcup_{|\mathbf{i}|=q} (\Delta\Theta_1^{i_1} \times \cdots \times \Delta\Theta_1^{i_n}),$$

see also Ganapathysubramanian and Zabaras [53].

We already mentioned in the respective sections above that both for interpolation and quadrature rules the extrema of Chebyshev polynomials (6.5) are good choices for the support nodes. Equidistant nodes should be avoided, like for one-dimensional formulae, especially when approximating nonlinear functions. In order to get nested sequences of support nodes and thus nested sparse grids, the numbers of Chebyshev

points can be chosen as $m_1 = 1$, $m_i = 2^{i-1} + 1$ for $i > 1$, as used by Clenshaw and Curtis [37], Novak and Ritter [103], Barthelmann et al. [10], Nobile et al. [100], and others. In case of $m_1 = 1$, the single support node is chosen as $y_1^1 = 0$, i.e. the center point of the approximation interval. Particularly in higher dimensions, $n \gg 1$, it is important to start with an initial one-point formula, $m_1 = 1$. Otherwise the number of points used by $A(q,n)$ grows too fast with the dimension. As $m_i = 2^{i-1}+1$ is odd for $i > 1$, it follows from (6.8) that the quadrature weights for this choice are given by

$$a_1^i = a_{m_i}^i = \frac{1}{m_i(m_i-2)},$$

$$a_j^i = a_{m_1+1-j}^i = \frac{2}{m_i-1}\left(1 - \frac{\cos((j-1)\pi)}{m_i(m_i-2)} - 2\sum_{k=1}^{(m_i-3)/2} \frac{1}{4k^2-1}\cos\frac{2k(j-1)\pi}{m_i-1}\right),$$

$$j = 2, \ldots, \frac{m_i+1}{2},$$

see also Novak and Ritter [102]. The number of support nodes $m(q,n)$ used by the Smolyak construction $A(q,n)$ with $q = k+n$ for fixed dimension n and growing integration order $k \to \infty$ is shown by Novak and Ritter [103] to behave like

$$m(k+n,n) \equiv \frac{1}{(n+1)!2^{n-1}} 2^k k^{n-1},$$

where $\equiv$ denotes the strong equivalence of sequences, $v_j \equiv w_j \Leftrightarrow \lim_{j\to\infty} \frac{v_j}{w_j} = 1$.

The total number of points used by $A(q,n)$ with $q = n+k$ is proportional to $\frac{2^k}{k!}n^k$ for fixed k, see Novak and Ritter [102] and Xiu and Hesthaven [141]. On the other hand, the dimension of the polynomial space Π_n^k is $\binom{n+k}{n}$, which is proportional to $\frac{n^k}{k!}$. So the Smolyak construction $A(q,n)$ needs about 2^k times the number of support nodes as there are degrees of freedom in Π_n^k. As this factor is independent of the dimension n, the Smolyak construction can be considered as optimal, cf. Xiu and Hesthaven [141].

The exact number of points in the sparse grid is computed by Schreiber [122] as

$$m(q,n) = \sum_{(h_1,\ldots,h_{q-n+1})\in J_{q-n+1,n}} \frac{n!}{h_1!\cdots h_{q-n+1}!} \prod_{k=1}^{q-n+1} (m_k - m_{k-1})^{h_k},$$

with the index set

$$J_{l,n} = \{h = (h_1,\ldots,h_l) \in \{0,\ldots,n\}^l : \sum_{k=1}^{l} h_k = n \;\&\; \sum_{k=1}^{l}(k-1)h_k \le k-1\}.$$

For the special case of the Clenshaw-Curtis grid with $m_1 = 1$, $m_i = 2^{i-1}+1$ for $i > 1$ this formula becomes

$$m^{cc}(q,n) = \sum_{j=0}^{q-n} 2^{j-n} \sum_{(h_1,\ldots,h_{q-n+1})\in J_{q-n+1,n}^j} \frac{n!}{h_1!\cdots h_{q-n+1}!} 2^{h_1+h_2},$$

with the index set

$$J_{l,n}^{j} = \{h = (h_1, \ldots, h_l) \in \{0, \ldots, n\}^l : \sum_{k=1}^{l} h_k = n \;\&\; \sum_{k=1}^{l} (k-1)h_k = j\}.$$

Schreiber [122] also gives other formulations for the number of points as well as upper and lower bounds that are easier to compute. These lead to the following bounds for the order q of the interpolation or integration,

$$\begin{aligned} &\log((n-1)!) + (n-1) - (n-1)\log(2n + \log(m^{cc}(q,n))) + \log(m^{cc}(q,n)) \\ &\leq q \\ &\leq 2n + \log(m^{cc}(q,n)). \end{aligned}$$

Figure 6.1 gives an overview how the number $m(q, 2)$ of support nodes in a two-dimensional grid changes with the order of the interpolation or quadrature and thus with the number of points used for the one-dimensional interpolation or quadrature rules.

Concerning the exactness of the Smolyak construction, Novak and Ritter [102] and Barthelmann et al. [10] prove the following result.

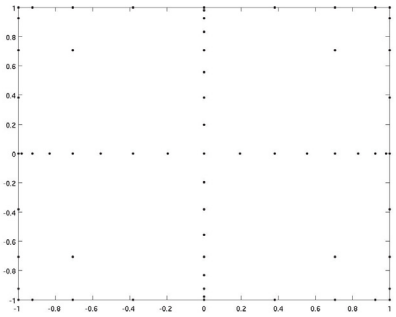
Theorem 6.4.1 (Novak and Ritter [102]). *Let again Π_n^k denote the space of all n-variate polynomials of degree at most k. If $\mathcal{U}_i$ is an exact interpolation formula on $\Pi_1^{m_i}$, as it is the case for Clenshaw-Curtis quadrature rules, then $A(q,n)$ is an exact interpolation formula on $E(q,n) := \sum_{|\boldsymbol{i}|=q} (\Pi_1^{i_1} \otimes \cdots \otimes \Pi_1^{i_n})$.*

This theorem has the consequence that the Smolyak interpolation $A(n+k, n)$ is exact for all polynomials of degree k, see Novak and Ritter [104] and Barthelmann et al. [10]. Xiu and Hesthaven [141] show the same exactness result for $\mathcal{U}_i$ being a quadrature rule. In this context, the number k is also called the *level of the Smolyak construction*.

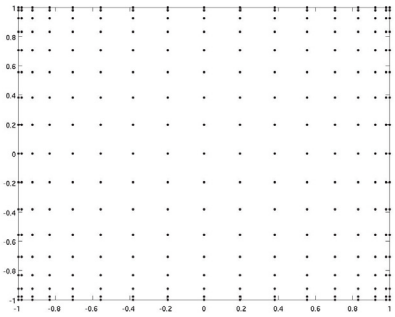
Barthelmann et al. [10] show that the interpolation property of the univariate case is preserved by the multivariate Smolyak construction in case of nested point sets. More precisely, they prove the following theorem.

Theorem 6.4.2 (Barthelmann et al. [10]). *Assume that $\Theta_i \subset \Theta_{i+1}$ and $\mathcal{U}_i(f)(y) = f(y)$ for all $f \in C([-1,1])$ and all $y \in \Theta_i$. Then $A(q,n)(f)(y) = f(y)$ for all $f \in C([-1,1]^n)$ and all $y \in H(q,n)$.*

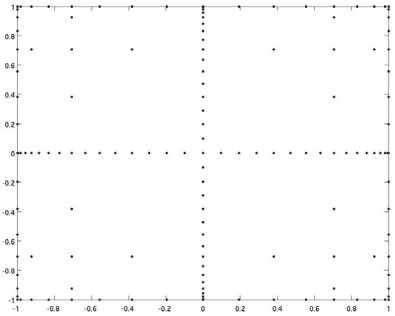
So if $A(q,n) \in E(q,n)$, it follows that for all $f \in C([-1,1]^n)$ there exists a polynomial $p \in E(q,n)$ such that $p(y) = f(y)$ for all $y \in H(q,n)$. The space $E(q,n)$ is the "least solution" in the sense of de Boor and Ron [40] for interpolation on $H(q,n)$ and the elements of $E(q,n)$ have smallest possible degree, i.e. for any polynomial space P that allows unique interpolation on $H(q,n)$ we have $\dim(P \cap \Pi_n^k) \leq \dim(E(q,n) \cap \Pi_n^k)$ for all k. It follows further that if p is a multivariate polynomial of degree at most k with respect to the j-th variable, the same holds true for the Smolyak construction $A(q,n)(p)$ for all $q > n$, and in particular $A(q,n)$ is degree reducing, i.e. for



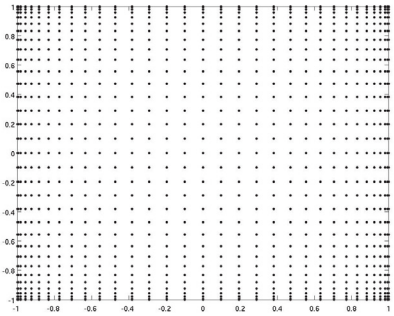
(a) Two-dimensional sparse grid of level 4 with 65 points.



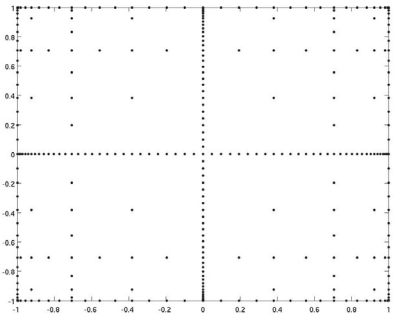
(b) Two-dimensional full grid with 289 points.



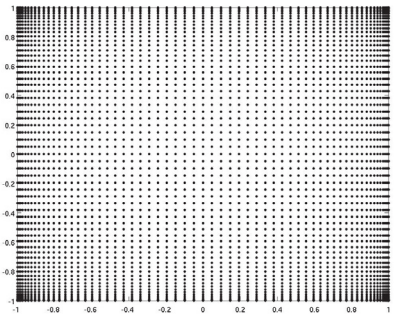
(c) Two-dimensional sparse grid of level 5 with 145 points.



(d) Two-dimensional full grid with 1089 points.



(e) Two-dimensional sparse grid of level 6 with 321 points.



(f) Two-dimensional full grid with 4225 points.

Figure 6.1.: Comparison of sparse grids with Clenshaw-Curtis points and corresponding full grids.

$p \in \Pi_n^k$ it follows that also $A(q,n)(p) \in \Pi_n^k$. Accordingly, it follows that $E(q,n)$ is the interpolation space of smallest degree in the sense of Sauer [116].

For multivariate interpolation for functions in the class

$$F_n^r := \{f : [-1,1]^n \to \mathbb{R} : D^\beta f \text{ continuous if } \beta_i \leq r \text{ for all } i\}$$

with the norm

$$\|f\| := \max\{\|D^\beta f\|_\infty : \beta \in \mathbb{N}_0^n,\ \beta_i \leq r\},$$

Barthelmann et al. [10] prove the error bound

$$\|I_n - A(q,n)\| \leq c_{n,r} m(q,n)^{-r} \big(\log m(q,n)\big)^{(r+2)(n-1)+1},$$

where $c_{n,r}$ is a constant that only depends on n and r and $m(q,n)$ denotes the number of points used by Smolyak's construction. For multivariate quadrature, Novak and Ritter [102] prove the error bound

$$\|I_n - A(q,n)\| = O\left(m(q,n)^{-r} \big(\log m(q,n)\big)^{(n-1)(r+1)}\right).$$

On the other hand, no sequence of interpolation or quadrature rules can result in errors on F_n^r that tend to zero faster than $m(q,n)^{-r}$. Thus Smolyak's construction leads to approximations that are optimal up to a logarithmic factor.

6.5. Solution of the uncertain optimization problem with ODE constraints

So we have seen in the last sections that when discretizing the uncertainty space with a tensor product approach, it is advantageous to apply a sparse grid technique in order to reduce the number of support nodes and thus the computational costs. With this background we can now apply such a stochastic collocation procedure with Smolyak's grid construction to our uncertain optimization problem.

The first step is the discretization of the dynamic constraints by means of multiple shooting as described in section 2.3.2. Like in the previous chapter on the worst-case approach, this gives us an uncertain optimization problem with equality constraints only,

$$\begin{aligned} \min_{s,p} \quad & \tfrac{1}{2}\|r_1(s,p,\alpha)\|_2^2 \\ \text{s.t.} \quad & r_2(x(\theta_1; s_{j_1}, p, \alpha), \ldots, x(\theta_K; s_{j_K}, p, \alpha), p, \alpha) = 0, \qquad \alpha \in \mathcal{A}, \\ & h_j(s_j, s_{j+1}, p, \alpha) := x(\tau_{j+1}; s_j, p, \alpha) - s_{j+1} = 0, \quad j = 0, \ldots, N-1, \end{aligned} \tag{6.14}$$

where the solution $x(t; s_j, p, \alpha)$ is the solution of the initial value problem

$$\dot{x}(t) = f(t, x, p, \alpha), \quad x(\tau_j) = s_j(\alpha),$$

on the $(j+1)$-st multiple shooting interval.

But now, in contrast to the previous chapter, we do not compute a worst-case solution that is feasible for all realizations of the uncertainties. Instead, we want to compute an average solution by computing the expected value of the obseravtion functions in the objective function over the uncertainty space. This means that in the objective function,

$$r_1(s,p,\alpha) = \begin{pmatrix} \vdots \\ \frac{\eta_{ij} - m_{ij}(t_j, x(t_j), p, \alpha)}{\sigma_{ij}} \\ \vdots \end{pmatrix},$$

we replace the uncertain model outputs $m_{ij}(t_j, x(t_j), p, \alpha)$ by their expected value over $\mathcal{A}$,

$$\mathbb{E}(m_{ij}(t_j, x(t_j), p, \alpha)) = \int_{\mathcal{A}} m_{ij}(t_j, x(t_j), p, \alpha)\, d\mathcal{P}(\alpha).$$

In order to compute this expected value, we employ the stochastic collocation method described above. Our choice for the n_a-variate cubature rule for the computation of the integral is a sparse grid Smolyak construction based on univariate Clenshaw-Curtis quadrature rules. A small integration level should usually be enough for this application as the integration takes place on each multiple shooting interval separately and these intervals should not span too long a time horizon each. The discretized dynamics and the equality constraints are only required to hold at the chosen collocation points in Θ_{n_a}. Suppose that the sparse grid contains m_a points altogether. Then the original problem is transformed into the following standard least-squares problem with equality constraints,

$$\begin{aligned} \min_{s^{\alpha_1},\dots,s^{\alpha_{m_a}},p} \quad & \tfrac{1}{2}\|r_1(s^{\alpha_1},\dots,s^{\alpha_{m_a}},p,\alpha_1,\dots,\alpha_{m_a})\|_2^2 \\ \text{s.t.} \quad & \left.\begin{aligned} & r_2(x(\theta_1; s_{j_1}^{\alpha_i}, p, \alpha_i),\dots,x(\theta_K; s_{j_K}^{\alpha_i}, p, \alpha_i), p, \alpha_i) = 0, \\ & h_j(s_j^{\alpha_i}, s_{j+1}^{\alpha_i}, p, \alpha_i) = 0, \quad j = 0,\dots,N-1, \end{aligned}\right\} \forall \alpha_i \in \Theta_{n_a}, \end{aligned} \tag{6.15}$$

where the objective function is given by

$$\begin{aligned} & r_1(s^{\alpha_1},\dots,s^{\alpha_{m_a}},p,\alpha_1,\dots,\alpha_{m_a}) \\ = & \begin{pmatrix} \vdots \\ \frac{\eta_{ij} - \frac{1}{\prod_l 2\delta_l} \sum_k m_{ij}(t_j, x(t_j; s^{\alpha_k}, p, \alpha_k), p, \alpha_k) a_k}{\sigma_{ij}} \\ \vdots \end{pmatrix}. \end{aligned}$$

The factor in front of the sum relates to the fact that we assumed the uncertainties to be uniformly distributed in the uncertainty set $\mathcal{A}$. For the expected value of the observation functions evaluated at the times t_j we thus have to divide by the volume of the polytope spanned by the uncertainties.

Note, however, that due to the double discretization of the dynamics and the uncertainty space this problem is probably very large-scale. Thus the discretization structure should be exploited in its numerical solution.

6.6. The Gauss-Newton method for the doubly discretized uncertain optimization problem

For applying a generalized Gauss-Newton method to problem (6.15), we need an initial guess for the parameters p_0 and all multiple shooting variables, i.e. for all realizations of the uncertainties defined by the sparse grid points in Θ_{n_a}. As there is probably no other information available beside the measurements η_{ij}, we can initialize them all with the same values, $s_0^{\alpha_1} = \cdots = s_0^{\alpha_{m_a}} = s_0$. In case that the behavior of the system for different values of the uncertain coefficients is known more specifically, the different multiple shooting variables should be initialized accordingly.

For the k-th iteration, we introduce the notation

$$x_k^{\alpha_i}(t) := x(t; s_{k,j}^{\alpha_i}, p_k, \alpha_i),$$

where $s_{k,j}^{\alpha_i}$ denotes the initial value of the multiple shooting interval $j+1$ that corresponds to the measurement time t.

Then the iteration for updating the multiple shooting states and the parameters,

$$(s_{k+1}^{\alpha_1}, \ldots, s_{k+1}^{\alpha_{m_a}}, p_{k+1}) = (s_k^{\alpha_1}, \ldots, s_k^{\alpha_{m_a}}, p_k) + t^k(\Delta s_k^{\alpha_1}, \ldots, \Delta s_k^{\alpha_{m_a}}, \Delta p_k),$$

depends on an optional damping parameter t^k and the increment at the k-th iteration, $(\Delta s_k^{\alpha_1}, \ldots, \Delta s_k^{\alpha_{m_a}}, \Delta p_k)$, that solves the linearized problem

$$\begin{aligned}
\min_{\Delta s^{\alpha_1}, \ldots, \Delta s^{\alpha_{m_a}}, \Delta p} \quad & \frac{1}{2} \Big\| r_1(s_k^{\alpha_1}, \ldots, s_k^{\alpha_{m_a}}, p_k, \alpha_1, \ldots, \alpha_{m_a}) \\
& \quad + \sum_{i=1}^{m_a} \frac{\partial r_1}{\partial s^{\alpha_i}}(s_k^{\alpha_1}, \ldots, s_k^{\alpha_{m_a}}, p_k, \alpha_1, \ldots, \alpha_{m_a}) \Delta s^{\alpha_i} \\
& \quad + \frac{\partial r_1}{\partial p}(s_k^{\alpha_1}, \ldots, s_k^{\alpha_{m_a}}, p_k, \alpha_1, \ldots, \alpha_{m_a}) \Delta p \Big\|_2^2 \\
\text{s.t.} \quad & r_2(x_k^{\alpha_i}(\theta_1), \ldots, x_k^{\alpha_i}(\theta_K), p_k, \alpha_i) \\
& \quad + \frac{\partial r_2}{\partial s^{\alpha_i}}(x_k^{\alpha_i}(\theta_1), \ldots, x_k^{\alpha_i}(\theta_K), p_k, \alpha_i) \Delta s^{\alpha_i} \\
& \quad + \frac{\partial r_2}{\partial p}(x_k^{\alpha_i}(\theta_1), \ldots, x_k^{\alpha_i}(\theta_K), p_k, \alpha_i) \Delta p = 0, \\
& h_j(s_{k,j}^{\alpha_i}, s_{k,j+1}^{\alpha_i}, p_k, \alpha_i) \\
& \quad + \frac{\partial h_j}{\partial s^{\alpha_i}}(s_{k,j}^{\alpha_i}, s_{k,j+1}^{\alpha_i}, p_k, \alpha_i) \Delta s^{\alpha_i} \\
& \quad + \frac{\partial h_j}{\partial p}(s_{k,j}^{\alpha_i}, s_{k,j+1}^{\alpha_i}, p_k, \alpha_i) \Delta p = 0, \ j = 0, \ldots, N-1, \\
& \forall \alpha_i \in \Theta_{n_a}.
\end{aligned} \tag{6.16}$$

For $i = 1, \ldots, m_a$ and $j = 0, \ldots, N-1$ we define the matrices

$$G_j^i := \frac{\partial x}{\partial s_j^{\alpha_i}}(\tau_{j+1}; s_j^{\alpha_i}, p, \alpha_i),$$

$$Gp_j^i := \frac{\partial x}{\partial p}(\tau_{j+1}; s_j^{\alpha_i}, p, \alpha_i),$$

and for $j = 0, \dots, N$ the matrices

$$\begin{aligned} A_j^i &:= \frac{\partial r_1}{\partial s_j^{\alpha_i}}(s^{\alpha_1}, \dots, s^{\alpha_{m_a}}, p, \alpha_1, \dots, \alpha_{m_a}), \\ A_p &:= \frac{\partial r_1}{\partial p}(s^{\alpha_1}, \dots, s^{\alpha_{m_a}}, p, \alpha_1, \dots, \alpha_{m_a}), \\ B_j^i &:= \frac{\partial r_2}{\partial s_j^{\alpha_i}}(x^{\alpha_i}(\theta_1), \dots, x^{\alpha_i}(\theta_K), p, \alpha_i), \\ B_p^i &:= \frac{\partial r_2}{\partial p}(x^{\alpha_i}(\theta_1), \dots, x^{\alpha_i}(\theta_K), p, \alpha_i). \end{aligned}$$

With these the Jacobian of the linearized constraints is given by

$$J_G := \left(\begin{array}{cccc|c|ccccc} B_0^1 & B_1^1 & \cdots & B_N^1 & & & & & B_p^1 \\ & & & & \ddots & & & & \vdots \\ & & & & & B_0^{m_a} & B_1^{m_a} & \cdots & B_N^{m_a} & B_p^{m_a} \\ \hline G_0^1 & -\mathbb{I} & & & & & & & & Gp_0^1 \\ & \ddots & & & & & & & & \vdots \\ & & \ddots & & & & & & & \vdots \\ & & G_{N-1}^1 & -\mathbb{I} & & & & & & Gp_{N-1}^1 \\ \hline & & & & \ddots & & & & & \vdots \\ \hline & & & & & G_0^{m_a} & -\mathbb{I} & & & Gp_0^{m_a} \\ & & & & & & \ddots & & & \vdots \\ & & & & & & & \ddots & & \vdots \\ & & & & & & & G_{N-1}^{m_a} & -\mathbb{I} & Gp_{N-1}^{m_a} \end{array}\right)$$

and the Jacobian of the linearized objective function is

$$J_{r_1} := \left(\begin{array}{cccc|c|cccc|c} A_0^1 & A_1^1 & \cdots & A_N^1 & \cdots & A_0^{m_a} & A_1^{m_a} & \cdots & A_N^{m_a} & A_p \end{array}\right)$$

Like in section 3.1 on the generalized Gauss-Newton method above, we make the standard regularity assumptions that the rank of the constraint Jacobian equals the number of constraints,

$$\operatorname{rank} J_G = m_{r2} + N n_x m_a,$$

and that the rank of the combined Jacobian from the objective function and the constraints equals the number of unknowns,

$$\operatorname{rank} \mathcal{J} = (N+1) n_x m_a, \qquad \mathcal{J} := \begin{pmatrix} J_{r_1} \\ J_G \end{pmatrix}.$$

Then according to theorem 3.1.5 the solution of the linearized problem (6.16) is defined by means of a linear solution operator $\mathcal{J}^+$, which is a generalized inverse of

the full Jacobian $\mathcal{J}$, i.e.

$$\Delta X := \begin{pmatrix} \Delta s^{\alpha_1} \\ \vdots \\ \Delta s^{\alpha_{m_a}} \\ \Delta p \end{pmatrix} = -\mathcal{J}^{+}\mathcal{F}, \qquad \mathcal{F} := \begin{pmatrix} r_1 \\ r_2^{\alpha_1} \\ \vdots \\ r_2^{\alpha_{m_a}} \\ h_0^{\alpha_1} \\ \vdots \\ h_{N-1}^{\alpha_1} \\ \vdots \\ h_0^{\alpha_{m_a}} \\ \vdots \\ h_{N-1}^{\alpha_{m_a}} \end{pmatrix}.$$

This means, we can determine the increments of the multiple shooting variables and the parameters by solving the linear system

$$\mathcal{J}\Delta X = -\mathcal{F}. \tag{6.17}$$

It is important to note that each constraint in (6.15) and (6.16) depends on only one realization of the uncertainty. The only coupling elements are the parameters p or Δp in the linearized system, respectively. We can thus apply the first step of the condensing technique introduced in section 3.1.5, i.e. the LQ-decompositions of equations (3.24)–(3.26), to each column block in $\mathcal{J}$ separately. This leads to a transformed matrix of the form

$$\tilde{\mathcal{J}} = \left(\begin{array}{ccccc|c|ccccc|c}
\hat{A}_0^1 & \hat{A}_1^1 & \cdots & \cdots & \hat{A}_N^1 & \cdots & \hat{A}_0^{m_a} & \hat{A}_1^{m_a} & \cdots & \cdots & \hat{A}_N^{m_a} & A_p \\ \hline
\hat{B}_0^1 & \hat{B}_1^1 & \cdots & \cdots & \hat{B}_N^1 & & & & \cdots & & & B_p^1 \\
 & & & & & \ddots & & & & & & \vdots \\
 & & & & & & \hat{B}_0^{m_a} & \hat{B}_1^{m_a} & & \cdots & \hat{B}_N^{m_a} & B_p^{m_a} \\ \hline
L_0^1 & 0 & & & & & & & & & & Gp_0^1 \\
T_0^1 & L_1^1 & & & & & & & & & & \vdots \\
 & \ddots & \ddots & & & & & & & & & \vdots \\
 & & T_{N-2}^1 & L_{N-1}^1 & 0 & & & & & & & Gp_{N-1}^1 \\ \hline
 & & & & & \ddots & & & & & & \vdots \\ \hline
 & & & & & & L_0^{m_a} & 0 & & & & Gp_0^{m_a} \\
 & & & & & & T_0^{m_a} & L_1^{m_a} & & & & \vdots \\
 & & & & & & & \ddots & \ddots & & & \vdots \\
 & & & & & & & & T_{N-2}^{m_a} & L_{N-1}^{m_a} & 0 & Gp_{N-1}^{m_a}
\end{array}\right).$$

Like before, we can now eliminate all multiple shooting variables except for those of the last multiple shooting node, $\widehat{\Delta s}_N^{\alpha_i}$, $i = 1, \ldots, m_a$, by assuming that the vector

Δp is known. In this case we get

$$\widetilde{\Delta s}_0^{\alpha_i} = -\left(L_0^i\right)^{-1}\left(h_0^{\alpha_i} + Gp_0^i\Delta p\right) = \tilde{h}_0^i + \tilde{Gp}_0^{\,i}\Delta p,$$

where $\tilde{h}_0^i = -\left(L_0^i\right)^{-1} h_0^{\alpha_i}$ and $\tilde{Gp}_0^{\,i} = -\left(L_0^i\right)^{-1} Gp_0^i$. From this, we get recursively the remaining components of the step vector, $\widetilde{\Delta s}_j^{\alpha_i}$, by computing for $j = 1, \ldots, N-1$,

$$\widetilde{\Delta s}_j^{\alpha_i} = -\left(L_j^i\right)^{-1}\left(h_j^{\alpha_i} + Gp_j^i\Delta p + T_{j-1}^i\widetilde{\Delta s}_{j-1}^{\alpha_i}\right) = \tilde{h}_j^i + \tilde{Gp}_j^{\,i}\Delta p, \tag{6.18}$$

where $\tilde{h}_j^i = -\left(L_j^i\right)^{-1}\left(h_j^{\alpha_i} + T_{j-1}^i\tilde{h}_{j-1}^i\right)$ and $\tilde{Gp}_j^{\,i} = -\left(L_j^i\right)^{-1}\left(Gp_j^i + T_{j-1}^i\tilde{Gp}_{j-1}^{\,i}\right)$. Inserting this in the block rows corresponding to the objective function and the original equality constraints r_2 transforms the matrices in the right-most block corresponding to Δp and also the block rows in the right hand side of the linear system (6.17). The matrices A_p, abd B_p^i, $i = 1, \ldots, m_a$, are transformed to

$$\begin{aligned}
\hat{A}_p &:= A_p + \sum_{i=1}^{m_a}\sum_{j=0}^{N-1}\hat{A}_j^i\tilde{Gp}_j^{\,i},\\
\hat{B}_p^i &:= B_p^i + \sum_{j=0}^{N-1}\hat{B}_j^i\tilde{Gp}_j^{\,i}, \quad i = 1, \ldots, m_a,
\end{aligned}$$

and the right hand side vectors r_1 and $r_2^{\alpha_i}$, $i = 1, \ldots, m_a$, become

$$\begin{aligned}
a &:= r_1 + \sum_{i=1}^{m_a}\sum_{j=0}^{N-1}\hat{A}_j^i\tilde{h}_j^i,\\
b^i &:= r_2^{\alpha_i} + \sum_{j=0}^{N-1}\hat{B}_j^i\tilde{h}_j^i, \quad i = 1, \ldots, m_a.
\end{aligned}$$

The remaining shooting increments $\widehat{\Delta s}_N^{\alpha_i}$, $i = 1, \ldots, m_a$, can then be estimated together with the parameter step Δp from the objective function and the equality constraints of the condensed problem,

$$\begin{aligned}
\min_{\widehat{\Delta s}_N^{\alpha_1}, \ldots, \widehat{\Delta s}_N^{\alpha_{m_a}}, \Delta p} \quad & \frac{1}{2}\Big\|\sum_{i=1}^{m_a}\hat{A}_N^{\alpha_i}\widehat{\Delta s}_N^{\alpha_i} + \hat{A}_p\Delta p + a\Big\|_2^2\\
\text{s.t.} \quad & \hat{B}_N^i\widehat{\Delta s}_N^{\alpha_i} + \hat{B}_p^i\Delta p + b^i = 0, \quad i = 1, \ldots, m_a.
\end{aligned}$$

Having solved this system and thus obtained $\widehat{\Delta s}_N^{\alpha_i}$, $i = 1, \ldots, m_a$, and Δp, we can compute the remaining components $\widetilde{\Delta s}_j^{\alpha_i}$ for $j = 1, \ldots, N-1$ and $i = 1, \ldots, m_a$ from (6.18). The shooting increments for the original system can then be computed by applying the backward recursion (3.28) for all realizations of the uncertainty to undo the transformations with the orthogonal matrices Q_j^i, $j = 0, \ldots, N-1$, $i = 1, \ldots, m_a$.

Part III.

Applications and conclusions

The third and last part of this thesis finally applies the methods of the first and second part to real-world problems. The first chapter in this part treats robust parameter estimation problems and the second chapter is concerned with problems that involve uncertain model coefficients.

7. Parameter estimation problems with erroneous measurements

This chapter contains example problems to part I of this thesis. It demonstrates the effects of robust parameter estimation compared to standard least squares estimation for noisy data and in the presence of outliers in the measurements.

The two examples are taken from systems biology and chemistry. The first example describes the NF-κB signaling system, which is an important part of several intracellular processes. The second example covers the denitrogenization of pyridine, a benchmark problem from chemistry.

7.1. The NF-κB pathway

For this example we consider the NF-κB pathway, which is a benchmark problem from systems biology. It has gained much interest in the last decade as it is involved in many cellular processes that range from immune reactions over inflammatory response to apoptosis, cf. Krishna et al. [88]. It also plays a decisive role in the understanding of cancer and autoimmune diseases.

Several researchers developed mathematical descriptions of this signaling system, including Hoffmann et al. [73], Lipniacki et al. [93], and Krishna et al. [88]. The model we consider here is a simplified model that Waldherr [136] derived from the model described by Krishna et al. [88].

There are seven chemical species involved in the core process that presumably causes the oscillations that are an important feature of the experimentally observed system behavior. The most important substances are NF-κB, which is present both in the cytoplasm and the nucleus of the cell, and its inhibitor IκB, also in cytoplasmic and nuclear form, as well as IκB kinase (IKK). The NF-κB transcription factor and its inhibitor form a complex. This complex can be transported from the nucleus to the cytoplasm but not vice versa. On the other hand, free NF-κB can only enter the nucleus but does not leave it. The inhibitor IκB, however, can enter and leave the nucleus in its free form. IKK phosphorylates the inhibitor and leads to its degradation and thus to a dissolving of NF-κB from its complex with IκB. Cytoplasmic IκB is produced from its corresponding mRNA. IκB mRNA in turn is activated by NF-κB dimers and degrades independently.

We use the denotations $[\cdot]_c$ and $[\cdot]_n$ to denote cytosolic and nuclear concentrations,

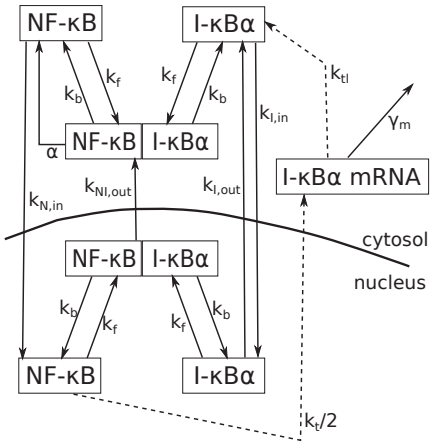


Figure 7.1.: NK-κB pathway. Dashed arrows signify indirect dependency.

respectively. The chemical reaction system reads

$$
\begin{aligned}
[\text{NF-}\kappa\text{B}]_\text{c} + [\text{I-}\kappa\text{B}\alpha]_\text{c} &\underset{k_b}{\overset{k_f}{\rightleftharpoons}} [\text{NF-}\kappa\text{B/I-}\kappa\text{B}\alpha]_\text{c} \\
[\text{NF-}\kappa\text{B}]_\text{n} + [\text{I-}\kappa\text{B}\alpha]_\text{n} &\underset{k_b}{\overset{k_f}{\rightleftharpoons}} [\text{NF-}\kappa\text{B/I-}\kappa\text{B}\alpha]_\text{n} \\
[\text{NK-}\kappa\text{B/I-}\kappa\text{B}\alpha]_\text{n} &\xrightarrow{k_{NI,out}} [\text{NK-}\kappa\text{B/I-}\kappa\text{B}\alpha]_\text{c} \\
[\text{NF-}\kappa\text{B}]_\text{c} &\xrightarrow{k_{N.in}} [\text{NF-}\kappa\text{B}]_\text{n} \\
[\text{I-}\kappa\text{B}\alpha]_\text{c} &\underset{k_{I,out}}{\overset{k_{I,in}}{\rightleftharpoons}} [\text{I-}\kappa\text{B}\alpha]_\text{n} \\
[\text{NF-}\kappa\text{B/I-}\kappa\text{B}\alpha]_\text{c} &\xrightarrow{\alpha} [\text{NF-}\kappa\text{B}]_\text{c} \\
[\text{I-}\kappa\text{B}\alpha \text{ mRNA}] &\xrightarrow{k_{tl}} [\text{I-}\kappa\text{B}\alpha]_\text{c} \\
[\text{NF-}\kappa\text{B}]_\text{n} + [\text{NF-}\kappa\text{B}]_\text{n} &\xrightarrow{\frac{1}{2}k_t} [\text{I-}\kappa\text{B}\alpha \text{ mRNA}] \\
[\text{I-}\kappa\text{B}\alpha \text{ mRNA}] &\xrightarrow{\gamma_m} \emptyset.
\end{aligned}
$$

A sketch of the system behavior is depicted in figure 7.1.

According to Waldherr [136], it is not necessary to model the full system but it suffices to consider the following four species:

- cytosolic I-κBα, denoted x_1,
- I-κBα mRNA, denoted x_2,
- total nuclear I-κBα, denoted x_3, and
- total nuclear NF-κB, denoted x_4.

Table 7.1.: Coefficients in model (7.1) for the NF-κB signaling system.

Coefficient	Meaning	Unit
N_{tot}	total concentration of NF-κB in the system	μM
k_{tl}	translation rate of I-κBα mRNA	min^{-1}
α	conversion rate of NF-κB/I-κBα to NF-κB in the cytosol	min^{-1}
$k_{I,in}$	uptake rate of I-κBα into the nucleus	min^{-1}
$k_{I,out}$	export rate of I-κBα from the nucleus	min^{-1}
$k_{NI,out}$	export rate of NF-κB/I-κBα from the nucleus	min^{-1}
k_t	NF-κB-dependent synthesis of I-κBα mRNA	$\mu\text{M}^{-1}\ \text{min}^{-1}$
γ_m	degradation rate of I-κBα mRNA	min^{-1}
$k_{N,in}$	uptake rate of NF-κB into the nucleus	min^{-1}

The dynamical system for these four species is then given by

$$\dot{x}_1 = k_{tl}x_2 - \frac{\alpha(N_{tot} - x_4)x_1}{K_I + x_1} - k_{I,in}x_1 + \frac{k_{I,out}K_N x_3}{K_N + x_4}, \tag{7.1a}$$

$$\dot{x}_2 = k_t x_4^2 - \gamma_m x_2, \tag{7.1b}$$

$$\dot{x}_3 = k_{I,in}x_1 - \frac{k_{I,out}K_N x_3}{K_N + x_4} - \frac{k_{NI,out}x_3 x_4}{k_N + x_4}, \tag{7.1c}$$

$$\dot{x}_4 = \frac{k_{N,in}K_I(N_{tot} - x_4)}{K_I + x_1} - \frac{K_{NI,out}x_3 x_4}{K_N + x_4}, \tag{7.1d}$$

where $K_I = \frac{k_b+\alpha}{k_f}$ and $K_N = \frac{k_b+k_{NI,out}}{k_f}$ with the dissociation rate $k_b = 0.03$ of the NF-κB/IκB complex and the association rate $k_f = 30.0$ of NF-κB and IκB. The meaning of the coefficients as well as their units are given in table 7.1. Waldherr [136] claims that the behavior of this model is essentially similar to that of the original full model by Krishna et al. [88].

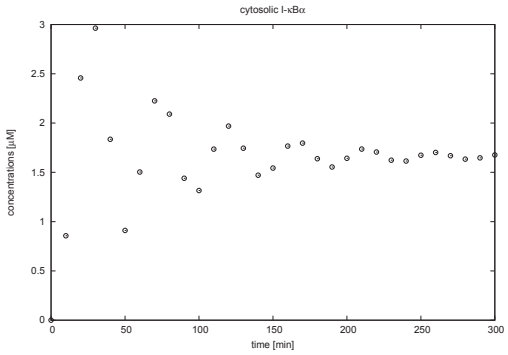
The initial state is $(0.0, 0.0, 0.0, 1.0)$. Measurements were simulated in the time interval $[0.0, 300.0]$ at points $t_i = 10.0i$, $i = 0, \ldots, 30$. This corresponds to an experiment running over five hours with measurements being taken every ten minutes during this time.

We consider two scenarios how these data might be perturbed. The first one uses the exact simulated values but contains three artificial outliers in nuclear NF-κB (x_4), which simulates the case of few gross errors. The second scenario adds 10% artificial measurement errors to the data set, which simulates the case of little noise in each data point.

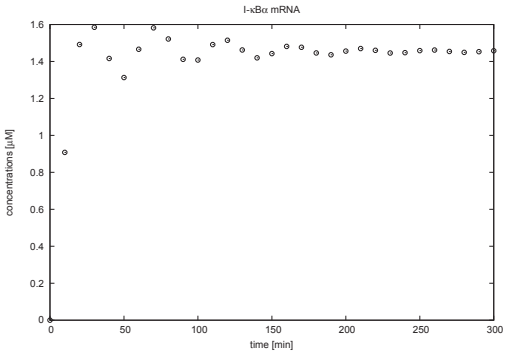
7.1.1. Outliers in the data

As measurement data, we chose here simulations using the "true" parameter values without artificial measurement errors but corrupted by three outliers introduced to the measurements of total nuclear NF-κB. Figure 7.2 shows the measurement points with the outliers indicated by filled dots.

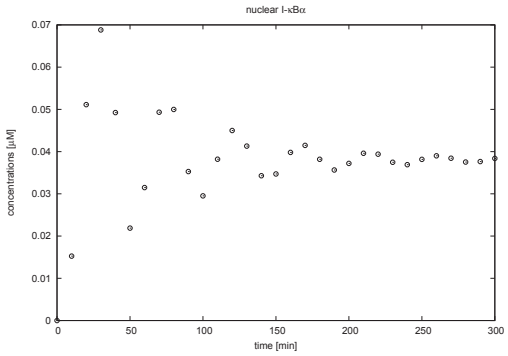
As for nuclear NF-κB (x_4) we have three outliers in a sample of 31 points, this corresponds to a fraction of about 0.097, i.e. roughly 10% bad data points in this



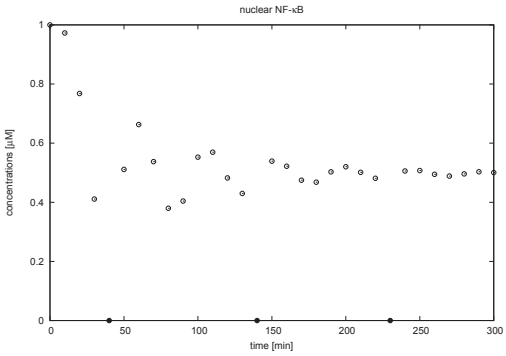
(a) Measurements of cytosolic I-κBα (x_1).



(b) Measurements of I-κBα mRNA (x_2).



(c) Measurements of nuclear I-κBα (x_3).



(d) Measurements of nuclear NF-κB (x_4).

Figure 7.2.: Measurement data for the NF-κB pathway with three gross outliers.

Table 7.2.: Estimates of the parameters and numbers of iterations for the NF-κB pathway with three outliers in nuclear NF-κB (x_4).

	$k_{N,in}$	$k_{I,in}$	$k_{I,out}$	$k_{NI,out}$	k_t
"true"	5.400	0.018	0.012	0.830	0.100
l_1	5.400	0.018	0.013	0.830	0.100
l_2	6.248	0.020	*-0.525*	0.981	0.090
Huber	5.412	0.018	0.001	0.832	0.100
	k_{tl}	α	γ_m	N_{tot}	iter
"true"	0.200	0.525	0.017	1.000	
l_1	0.200	0.525	0.017	1.000	6
l_2	0.201	0.446	0.013	1.039	11
Huber	0.200	0.524	0.017	1.000	9

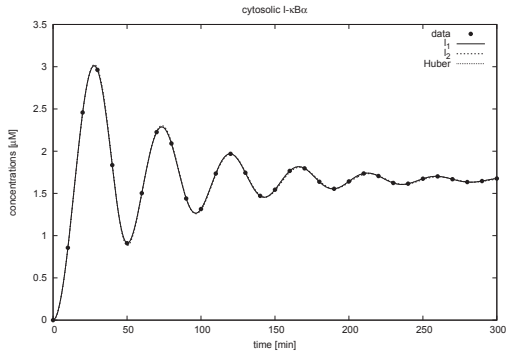
solution component. According to the tabular given by Ekblom and Madsen [45], the corresponding choice of the tuning constant for Huber's M-estimator is $\gamma = 1.14$. This value, however, leads to the same results as the least squares estimator and thus to a negative value for the constant $k_{I,out}$, cf. table 7.2. This can be helped by further reducing the tuning constant and table 7.2 gives the estimates corresponding to a value of $\gamma = 0.002$, the largest value of γ that gives positive values for all reaction rates. In figure 7.3, the trajectories from Huber's M-estimator are indistinguishable from that of the LAD estimator.

We see that very few outliers are sufficient to invalidate the least squares estimation. Although the trajectories do not differ too much from the true solution—only the concentration of nuclear NF-κB (x_4) is drawn down by the outliers—the negative reation rate $k_{I,out}$ for the export of I-κBα from the nucleus makes the result worthless in practical application.

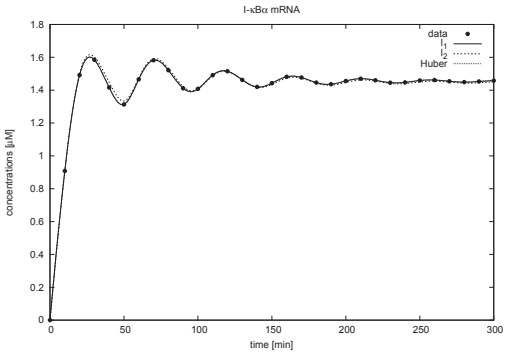
7.1.2. White noise in the data

In contrast to the previous section, we now do not consider a few gross outliers in the data but the case when all data points are perturbed by small measurement errors. We therefore added white noise with mean zero and standard deviation $\sigma = 0.1$ to all simulated measurements.

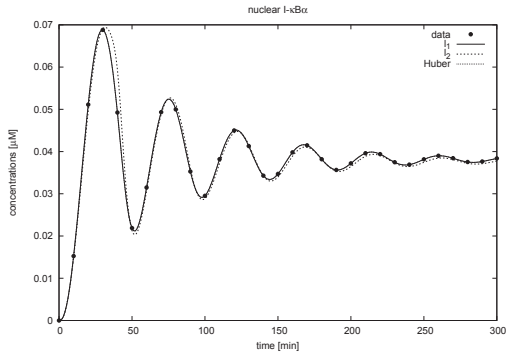
This scenario makes it hard to tell how many data points in our sample should be considered as outliers. Thus formula (2.18) for the connection of the fraction of wild points and the value of the Huber tuning constant and the corresponding tabular given by Ekblom and Madsen [45] are difficult to apply for choosing a good value of γ. A usual, though somewhat arbitrary, rule of thumb is to consider all values outside the double or triple fold of the standard deviation as outliers. This is due to the fact that for normally distributed random variables, 95.4% of these lie in an interval of length 2σ and 99.7% of the sample lie in an interval of length 3σ around the mean. For this reason we chose here a tuning constant of $\gamma = 0.25$ for the Huber



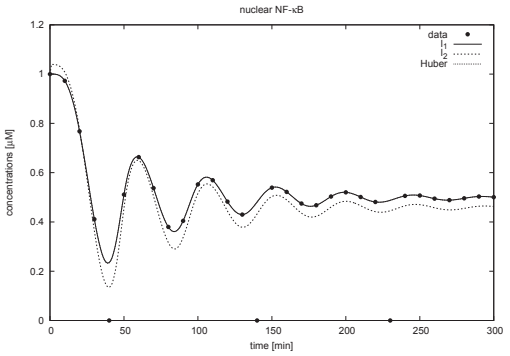
(a) Trajectory of cytosolic I-κBα (x_1).



(b) Trajectory of I-κBα mRNA (x_2).



(c) Trajectory of nuclear I-κBα (x_3).



(d) Trajectory of nuclear NF-κB (x_4).

Figure 7.3.: Trajectories corresponding to the estimated parameters for the NF-κB pathway with three gross outliers.

Table 7.3.: Estimates of the parameters and numbers of iterations for the NF-κB pathway with 10% noise.

	$k_{N,in}$	$k_{I,in}$	$k_{I,out}$	$k_{NI,out}$	k_t
"true"	5.400	0.018	0.012	0.830	0.100
l_1	5.279	0.018	0.243	0.792	0.105
l_2	4.230	0.017	*1.590*	0.621	0.111
Huber	4.243	0.017	*1.716*	0.646	0.112
	k_{tl}	α	γ_m	N_{tot}	iter
"true"	0.200	0.525	0.017	1.000	
l_1	0.205	0.551	0.017	0.965	5
l_2	0.214	0.641	0.019	0.934	15
Huber	0.212	0.642	0.019	0.929	11

estimator.

The results of the parameter estimation are presented in table 7.3 and figure 7.4.

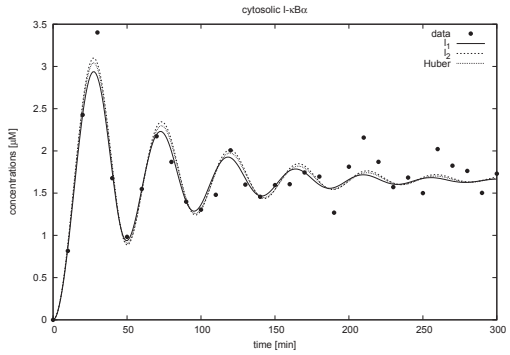
Here we see nicely, in particular for the trajectory of nuclear I-κBα, how Huber's M-estimator lies in between the solutions of the l_2 and the l_1 estimator. Note that although the numerical parameter values from Huber's estimator are closer to those of the least squares estimator than to the LAD estimate, this does not mean that its trajectory is also closer to that from l_2 estimation than to the l_1 solution. Figure 7.4 shows that this is not the case.

7.2. Denitrogenization of pyridine

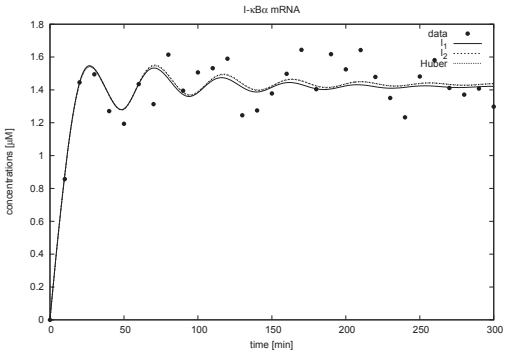
As our second numerical example we chose the chemical process of the denitrogenization of pyridine, taken from Bock [24]. A schematic sketch of the reaction is given in figure 7.5.

Pyridine often emerges in large quantities as a byproduct of other chemical reactions. Due to its high content of nitrogen, its combustion is rather detrimental to the environment and should be avoided. Other forms of recycling are to be preferred. The reaction considered here converts pyridine into ammonia and pentane using three catalysts.

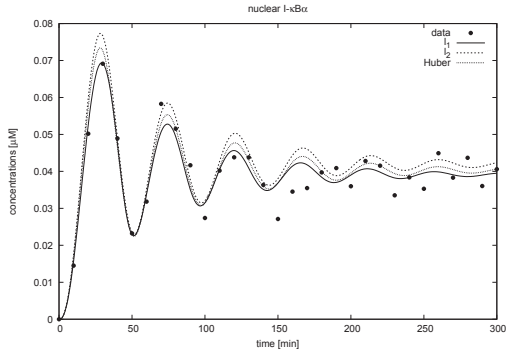
The process can be described mathematically by a system of seven ordinary differ-



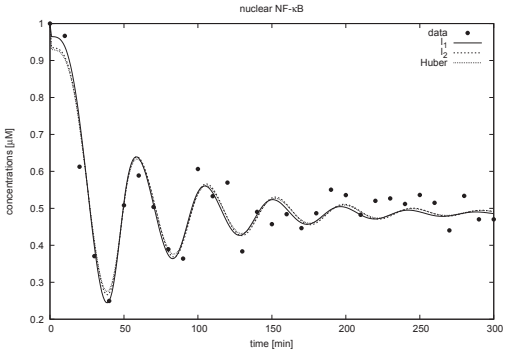
(a) Trajectory of cytosolic I-κBα (x_1).



(b) Trajectory of I-κBα mRNA (x_2).



(c) Trajectory of nuclear I-κBα (x_3).



(d) Trajectory of nuclear NF-κB (x_4).

Figure 7.4.: Trajectories corresponding to the estimated parameters for the NF-κB pathway with 10% overall noise.

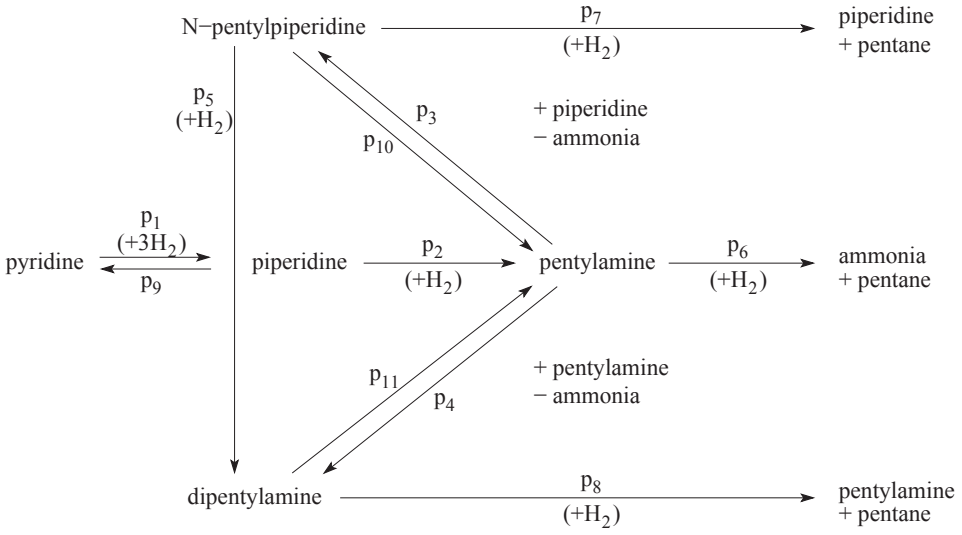


Figure 7.5.: Schematic sketch of the denitrogenization of pyridine.

ential equations,

$$\begin{aligned}
\text{pyridine:}\quad & \dot{A} = -p_1 A + p_9 B,\\
\text{piperidine:}\quad & \dot{B} = p_1 A - p_2 B - p_3 BC + p_7 D - p_9 B + p_{10} DF,\\
\text{pentylamine:}\quad & \dot{C} = p_2 B - p_3 BC - 2p_4 CC - p_6 C + p_8 E + p_{10} DF + 2p_{11} EF,\\
\text{N-pentylpiperidine:}\quad & \dot{D} = p_3 BC - p_5 D - p_7 D - p_{10} DF,\\
\text{dipentylamine:}\quad & \dot{E} = p_4 CC + p_5 D - p_8 E - p_{11} EF,\\
\text{ammonia:}\quad & \dot{F} = p_3 BC + p_4 CC + p_6 C - p_{10} DF - p_{11} EF,\\
\text{pentane:}\quad & \dot{G} = p_6 C + p_7 D + p_8 E\,.
\end{aligned}$$

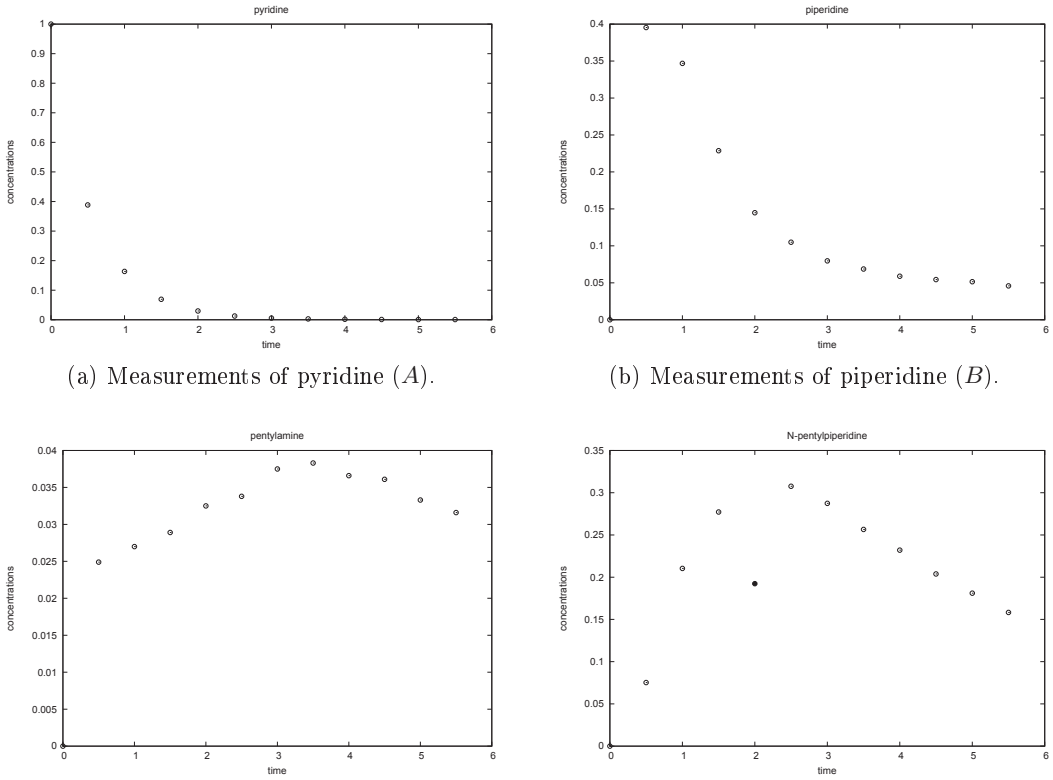
The reaction coefficients $p_1, \ldots, p_{11}$ are the unknowns of the reaction which are to be estimated from given measurements of concentrations of the species participating in the reaction.

The initial state is $(1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0)^T$. Measurements were taken in the time interval $[0.0, 5.5]$ at points $t_i = 0.5i,\ i = 0, \ldots, 11$.

As measurement data in numerical experiments we chose simulations using "true" parameter values which were corrupted by four outliers. Figure 7.6 shows the measurement points with the outliers indicated by filled dots.

The results of parameter estimation using l_2, l_1, and Huber estimators are given in table 7.4 and figure 7.7. The artificial 4 outliers in the total of 84 measurements correspond to a ratio of a little less than 5% bad data points. But tuning constants for the Huber estimator down to $\gamma = 1.14$, which would correspond to a ratio of 10% bad data points, still result in the same solution as the least squares estimator with negative parameter estimates, cf. table 7.4. So a further reduction is necessary. The results given below correspond to a Huber constant of $\gamma = 0.1$. This inadequacy of the mathematically derived tuning constant may be attributed to the fact that formula (2.18) does not take into account the nature of the outliers, cf. section 2.1. Single points with high bad leverage are widely deviating values in the dependent variable and have a worse impact on the results of a least squares estimation than good leverage points, no matter how far they are away from the data. We see that a simple mathematically motivated choice of Huber's tuning constant will not meet reality in many applications. Rather the data should be analyzed for outliers before performing a parameter estimation. Here statistical methods for data analysis may be a great help.

We see again how a few outliers can draw the least squares estimate away from the true solution. The trajectory of dipentylamine (E) demonstrates this especially dramatically. And also the numerical estimates show clearly that this l_2 estimation must be considered a failure, with two estimates of the reaction coefficients being negative and two others differing by one order of magnitude from the true values. The LAD estimator, on the other hand, is rather little perturbed by the outliers and follows the bulk of the data nicely, though not perfectly. Huber's M-estimator can again be observed to give results that lie somewhere in between the least squares and



(a) Measurements of pyridine (A).

(b) Measurements of piperidine (B).

(c) Measurements of pentylamine (C).

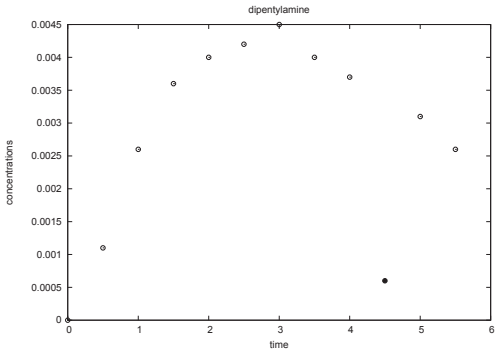
(d) Measurements of N-pentylpiperidine (D).

Figure 7.6.: Measurement data for the denitrogenization of pyridine.

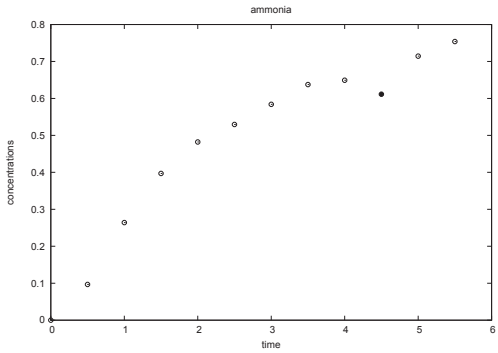
Table 7.4.: Estimates for the parameter estimation and numbers of iterations of the denitrogenization of pyridine with four outliers.

	p_1	p_2	p_3	p_4	p_5	p_6
"true"	1.810	0.894	29.400	9.210	0.058	2.430
l_1	1.805	0.899	29.508	7.750	0.057	2.510
l_2	1.809	0.847	32.677	*-1.613*	0.093	2.639
Huber	1.808	0.886	29.623	5.893	0.060	2.516

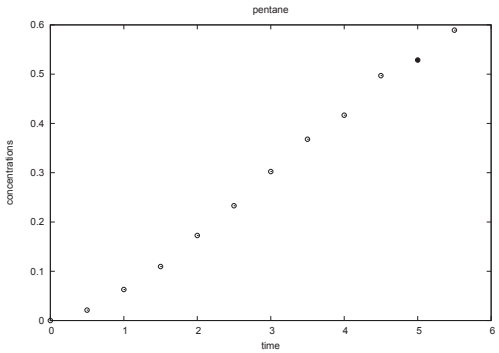
	p_7	p_8	p_9	p_{10}	p_{11}	iter
"true"	0.0644	5.550	0.0201	0.577	2.150	
l_1	0.0751	4.252	0.0201	0.559	4.379	15
l_2	*0.1558*	*-1.276*	0.0205	0.497	*22.929*	18
Huber	0.0948	2.994	0.0203	0.527	6.764	12



(e) Measurements of dipentylamine (E).

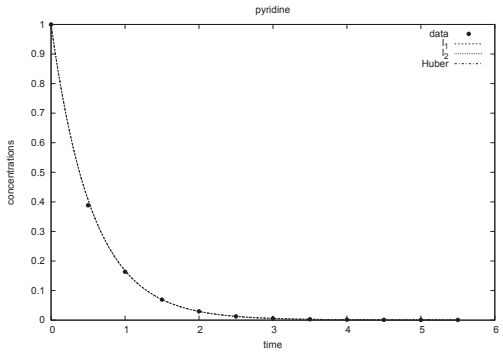


(f) Measurements of ammonia (F).

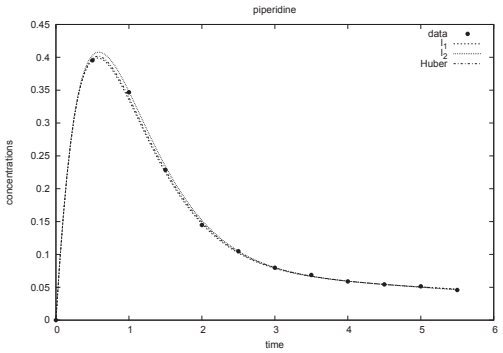


(g) Measurements of pentane (G).

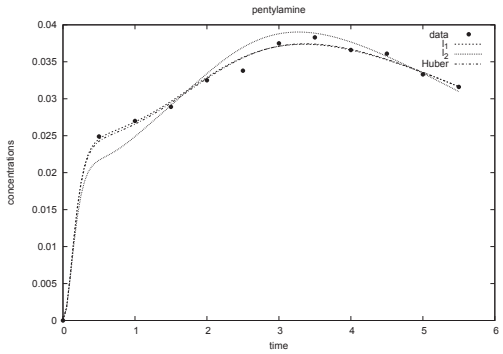
Figure 7.6.: Cont.: Measurement data for the denitrogenization of pyridine.



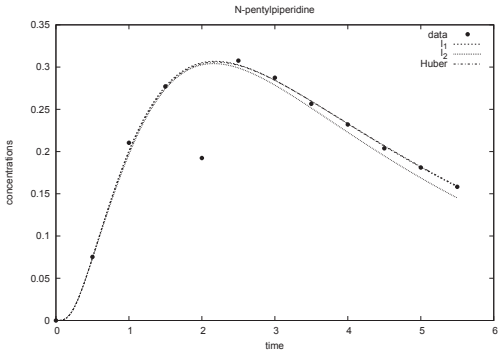
(a) Trajectory of pyridine (A).



(b) Trajectory of piperidine (B).

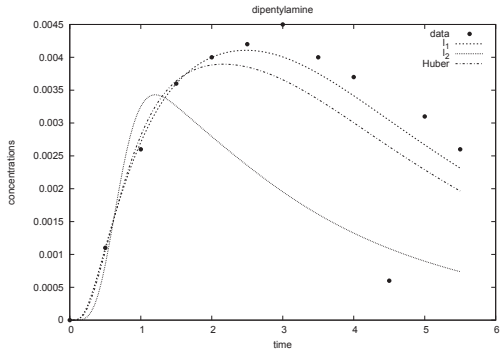


(c) Trajectory of pentylamine (C).

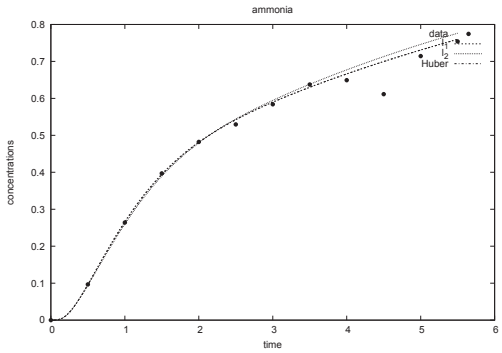


(d) Trajectory of N-pentylpiperidine (D).

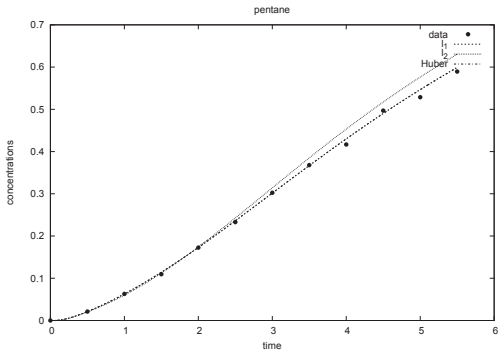
Figure 7.7.: Trajectories corresponding to the estimated parameters for the denitrogenization of pyridine with four outliers.



(e) Trajectory of dipentylamine (E).



(f) Trajectory of ammonia (F).



(g) Trajectory of pentane (G).

Figure 7.7.: Cont.: Trajectories corresponding to the estimated parameters for the denitrogenization of pyridine with four outliers.

the least absolute deviations solution. In this example it is closer to the l_1 solution due to the rather low tuning constant.

8. Parameter estimation problems with uncertain model coefficients

This chapter contains example problems to part II of this thesis. It demonstrates the effects of the stochastic collocation approach compared to the approximated solution of the worst-case approach for parameter estimation problems with uncertain coefficients in the model. We refrain from giving the numbers of iterations for the examples because we only have a very basic implementation of the methods, i.e. parameter estimation with the approximate worst-case approach and stochastic collocation combined with multiple shooting. In particular the stepsize control is overly conservative in many cases such that iteration numbers would be too high anyway. But as a proof of concept the implementations are sufficient at any rate.

The two example problems of this chapter are again taken from biology and chemistry. The first one described here is the Lotka-Volterra predator-prey system. The second example is a homogeneous chemical reaction system for the isomerization of α-pinene.

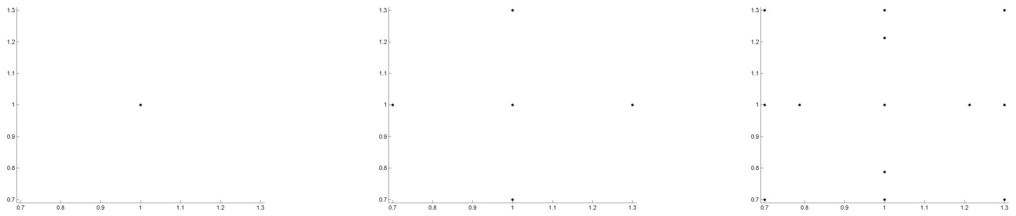
8.1. The Lotka-Volterra predator-prey system

As a first very basic example we consider the well-known Lotka-Volterra model describing the population development of a two-species environment in which one species preys on the other. The corresponding differential equations are

$$\begin{aligned} \text{prey:} \quad & \dot{x} = \alpha x - \beta xy, \\ \text{predator:} \quad & \dot{y} = -\gamma y + \delta xy. \end{aligned}$$

We simulated measurements for the system corresponding to the true parameter values $\alpha = \beta = \gamma = \delta = 1.0$ and added gentle noise with mean zero and standard deviation $\sigma = 0.02$. From these we tried to recover the coefficients α, β of the prey species such that the values of animals for each species best match the simulated observations. The coefficients γ, δ determining the dynamics of the predator species were considered uncertain in the interval $[0.7, 1.3]$ each. In all simulations with the stochastic collocation approach we only used every second measurement as a multiple shooting node to keep the equation systems that have to be solved during the Gauss-Newton iterations of moderate size.

Using stochastic collocation with integration level 0 corresponds to a standard parameter estimation with the uncertainty being ignored and a nominal value of



(a) Sparse grid of level 0 for the Lotka-Volterra example. This corresponds to a standard least squares estimation ignoring the uncertainty and using the nominal values $\gamma = 1.0 = \delta$ instead.

(b) Sparse grid of level 1 for the Lotka-Volterra example.

(c) Sparse grid of level 2 for the Lotka-Volterra example.

Figure 8.1.: Sparse grids of levels 0–2 for the Lotka-Volterra example.

$\gamma = \delta = 1.0$, i.e. the center point of the uncertainty set, being used as the only point in the sparse grid, cf. figure 8.1a.

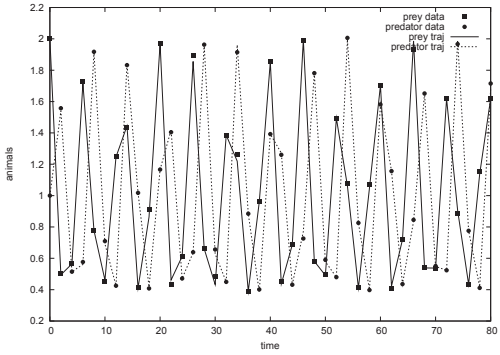
It is not surprising that in this case the parameters α and β are estimated very exactly, see the numerical results in table 8.1 and figure 8.2a, as the nominal values for γ and δ match the true values and there is only little noise in the simulated data. Figure 8.2a clearly shows how the trajectories match the slightly perturbed data almost perfectly well.

Table 8.1.: Estimates of the parameters in the prey equation of the Lotka-Volterra model with the stochastic collocation approach with integration levels 0–2 and the worst-case solution.

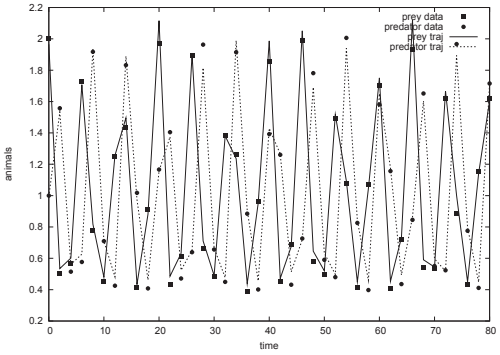
level	α	β	obj. value
0	1.000000	1.000000	0.011185
1	1.000445	0.989189	0.114201
2	0.968702	0.966258	0.048071
WC	1.000597	0.979820	0.052949

Using stochastic collocation with integration level 1, the expected values of the numbers of animals are computed over a sparse grid with 5 grid points, cf. figure 8.1b.

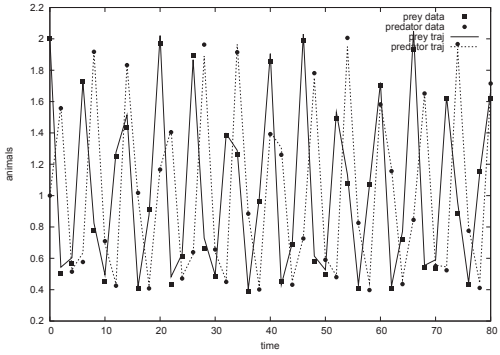
Here we can see that the estimates for the coefficients in the prey equation no longer exactly result to be the true values. The numerical data are given in table 8.1 and figure 8.2b. Still the value of the objective function in the optimum seems reasonably small, though one order of magnitude larger than for the true parameter values. In figure 8.2b we recognize at once several times where the estimated trajectories differ from measurement data. As this low integration level combines univariate integration



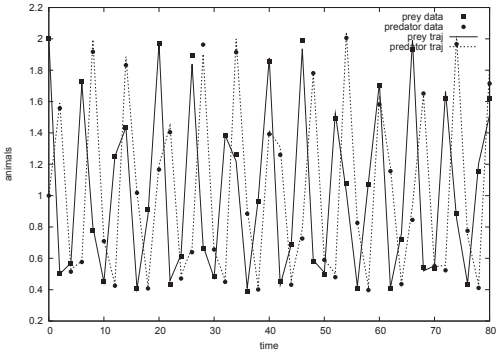
(a) Trajectories of the parameter estimation for the Lotka-Volterra model with the stochastic collocation approach with integration level 0.



(b) Trajectories of the parameter estimation for the Lotka-Volterra model with the stochastic collocation approach with integration level 1.



(c) Trajectories of the parameter estimation for the Lotka-Volterra model with the stochastic collocation approach with integration level 2.



(d) Trajectories of the parameter estimation for the Lotka-Volterra model with the approximated worst-case approach.

Figure 8.2.: Trajectories of the parameter estimation for the Lotka-Volterra model with the stochastic collocation approach with integration levels 0–2 and the worst-case solution.

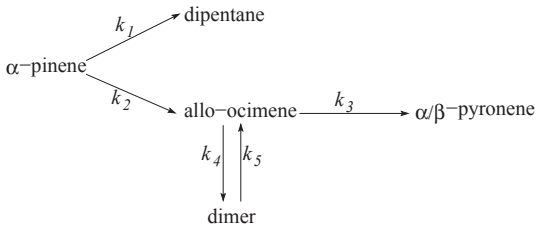


Figure 8.3.: Thermal isomerization of α-pinene.

methods that use at most linear interpolation polynomials in the integrand, it could not be expected that we would get a really good result. So it is rather surprising that the solution is that good at all.

Increasing the integration level of the stochastic collocation approach to 2 leads to a solution that is based on a sparse grid with 13 points, see figure 8.1c.

The optimal value of the objective function is improved compared to the lower integration level but does not reach the optimal value for the true parameters. The numerical values of the parameter estimates even further recede from their true values though, cf. table 8.1 and figure 8.2c. Comparing this figure to figure 8.2b for stochastic collocation with level-1 integration, though, shows that with level-2 integration the estimated trajectories get closer to the measured data in agreement with the optimal objective value.

For comparison, we also computed a worst-case solution with the approximation technique described in chapter 5. We used the same conditions as with the stochastic collocation approach to keep the solutions comparable. Note, however, that due to the approximate nature of the worst-case solution the meaningfulness of this comparison is limited.

The numerical results are given in table 8.1 and the corresponding trajectories are shown in figure 8.2d. The estimated values of the parameters are close to those for stochastic collocation with level-1 integration while the objective value in this optimum is rather close to that of the solution from stochastic collocation with level-2 integration though slightly worse.

8.2. Isomerization of α-pinene

Our second example is taken from Fuguitt and Hawkins [51] and was revisited by Rodriguez-Fernandez et al. [113]. It describes the chemical reactions when heating α-pinene in its liquid phase in a temperature range between 189.5° and 285°. By thermal isomerization α-pinene yields dipentane and allo-ocimene and the latter in turn yields α- and β-pyronene and a dimer. The reaction process is depicted in figure 8.3 and the corresponding mathematical equations are given in system (8.1).

Table 8.2.: Parameter estimates for the isomerization of α-pinene at temperature 189.5° found by Rodriguez-Fernandez et al. [113].

parameter	estimate
k_1	$5.9259e{-}5 \pm 1.4391e{-}6$
k_2	$2.9634e{-}5 \pm 1.3039e{-}6$
k_3	$2.0473e{-}5 \pm 6.6657e{-}6$
k_4	$2.7449e{-}4 \pm 5.5314e{-}5$
k_5	$3.9980e{-}5 \pm 1.9514e{-}5$

$$\alpha\text{-pinene:} \quad \dot{y}_1 = -(k_1 + k_2)y_1, \tag{8.1a}$$
$$\text{dipentane:} \quad \dot{y}_2 = k_1 y_1, \tag{8.1b}$$
$$\text{allo-ocimene:} \quad \dot{y}_3 = k_2 y_1 - (k_3 + k_4)y_3 + k_5 y_5, \tag{8.1c}$$
$$\alpha/\beta\text{-pyronene:} \quad \dot{y}_4 = k_3 y_3, \tag{8.1d}$$
$$\text{dimer:} \quad \dot{y}_5 = k_4 y_3 - k_5 y_5. \tag{8.1e}$$

The following parameter estimations are based on the measurement data originally given by Fuguitt and Hawkins [51] for temperature 189.5° and 204.5°, respectively. The uncertainty intervals are taken from the parameter estimation results given by Rodriguez-Fernandez et al. [113]. We repeat their estimated values and confidence intervals in table 8.2.

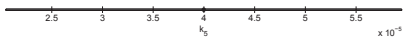
8.2.1. One uncertainty

For the first estimation only the fifth coefficient k_5 is assumed to be uncertain in the uncertainty interval $[2.0466e{-}5, 5.9494e{-}5]$. The corresponding sparse grids for stochastic collocation with integration levels 0 to 3 are shown in figure 8.4.

The numerical results from the parameter estimation with the stochastic collocation approach with integration levels 0–3 and the worst-case solution are given in tables 8.3 and 8.4 for the two different temperatures. Comparative plots of the corresponding trajectories are shown in figures 8.5 and 8.6, respectively.

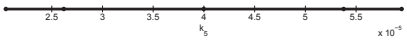
For both temperatures the trajectories for all solution approaches are practically identical. The plots in figure 8.5 for the reaction taking place at 189.5° show only slight differences in the trajectories for α- and β-pyronene (y_4) and even less for allo-ocimene (y_3) where increasing the integration level leads to small improvements. This is also mirrored by the decreasing optimal values of the objective function in table 8.3.

At the higher temperature of 204.5° the reaction apparently shows quite a different behavior with less dipentane but more allo-ocimene being produced. This case was not considered by Rodriguez-Fernandez et al. [113] such that their parameter estimates are insufficient to recover the system behavior. This can also be seen in figure

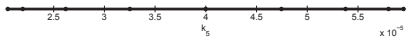


(a) Sparse grid of level 0 in one dimension with 1 point.

(b) Sparse grid of level 1 in one dimension with 3 points.



(c) Sparse grid of level 2 in one dimension with 5 points.



(d) Sparse grid of level 3 in one dimension with 9 points.

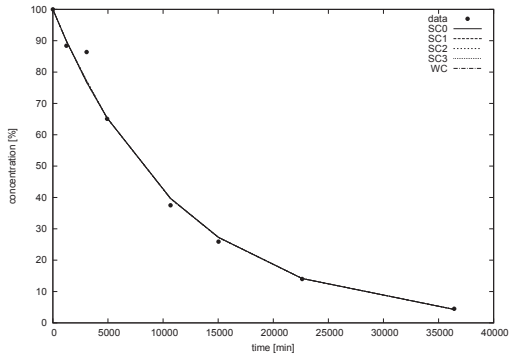
Figure 8.4.: Sparse grids in one dimension for stochastic collocation with integration levels 0–3.

Table 8.3.: Estimates of the reaction rates for the thermal isomerization of α-pinene at the temperature of 189.5° for uncertain k_5 with the stochastic collocation approach with integration levels 0–3 and the worst-case solution.

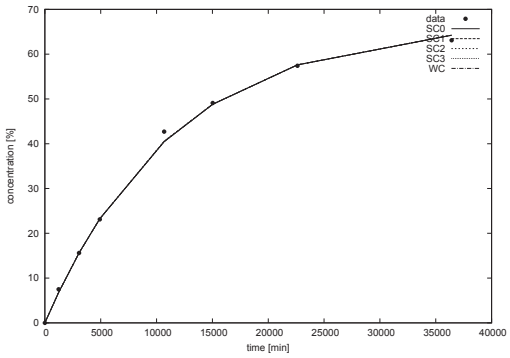
lev.	k_1	k_2	k_3	k_4	obj. val.
0	$5.800815e{-5}$	$2.842333e{-5}$	$1.870428e{-5}$	$2.793498e{-4}$	60.28230
1	$5.804764e{-5}$	$2.845724e{-5}$	$1.878630e{-5}$	$2.782224e{-4}$	56.79532
2	$5.807097e{-5}$	$2.847904e{-5}$	$1.881059e{-5}$	$2.781311e{-4}$	54.74702
3	$5.807610e{-5}$	$2.848396e{-5}$	$1.881717e{-5}$	$2.781140e{-4}$	54.29301
WC	$5.800617e{-5}$	$2.842444e{-5}$	$1.872404e{-5}$	$2.794354e{-4}$	60.28283

Table 8.4.: Estimates of the reaction rates for the thermal isomerization of α-pinene at the temperature of 204.5° for uncertain k_5 with the stochastic collocation approach with integration levels 0–3 and the worst-case solution.

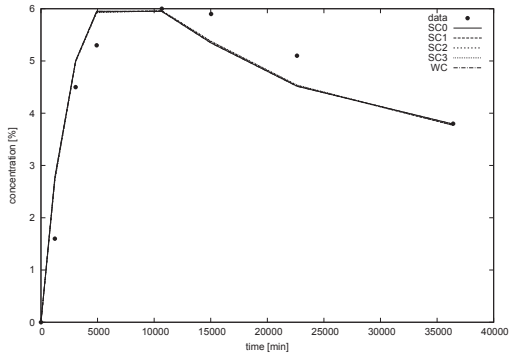
lev.	k_1	k_2	k_3	k_4	obj. val.
0	$2.240241e{-4}$	$1.309212e{-4}$	$9.194628e{-5}$	$5.873153e{-4}$	9.326250
1	$2.240914e{-4}$	$1.309526e{-4}$	$9.225405e{-5}$	$5.880603e{-4}$	9.338465
2	$2.240430e{-4}$	$1.309263e{-4}$	$9.198711e{-5}$	$5.872757e{-4}$	9.339362
3	$2.240753e{-4}$	$1.309439e{-4}$	$9.213459e{-5}$	$5.877118e{-4}$	9.338911
WC	$2.241008e{-4}$	$1.309627e{-4}$	$9.252261e{-5}$	$5.889562e{-4}$	9.328130



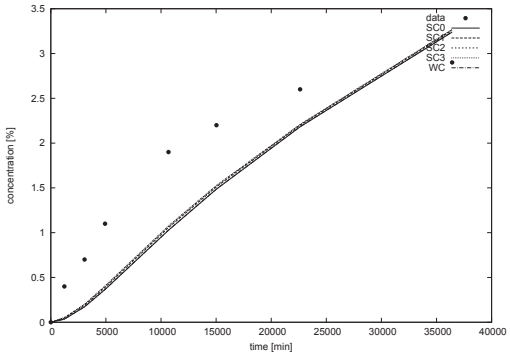
(a) Trajectories for α-pinene (y_1).



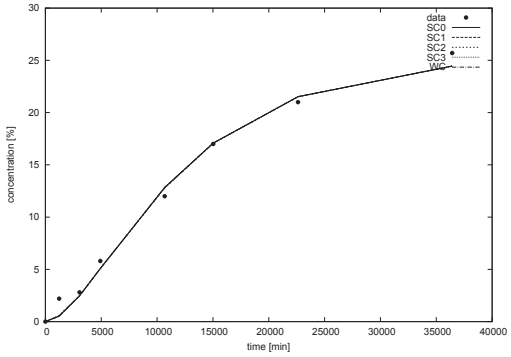
(b) Trajectories for dipentane (y_2).



(c) Trajectories for allo-ocimene (y_3).

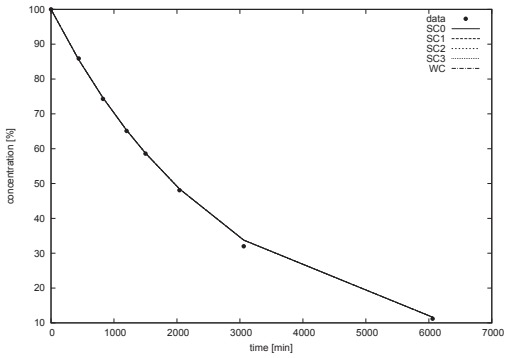


(d) Trajectories for α/β-pyronene (y_4).

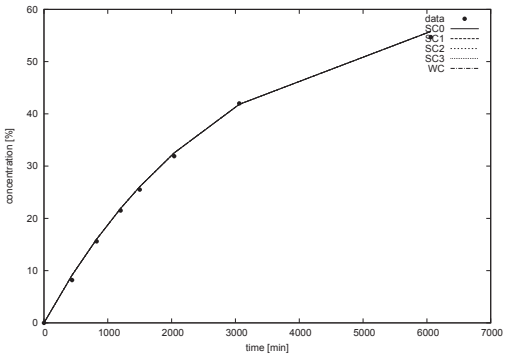


(e) Trajectories for the dimer (y_5).

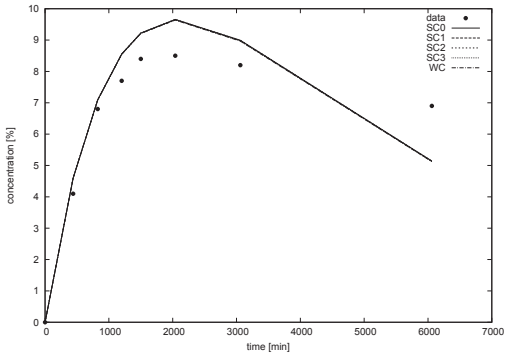
Figure 8.5.: Comparison of the trajectories for the thermal isomerization of α-pinene at the temperature of 189.5° for uncertain k_5 with the stochastic collocation approach with integration levels 0–3 and the worst-case solution.



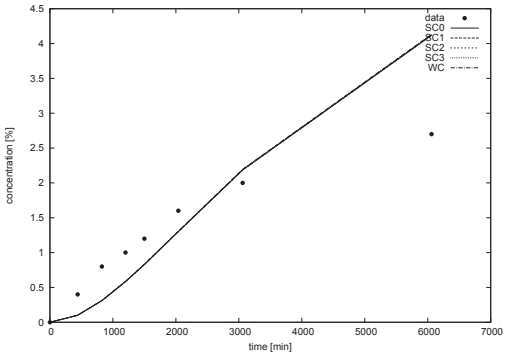
(a) Trajectories for α-pinene (y_1).



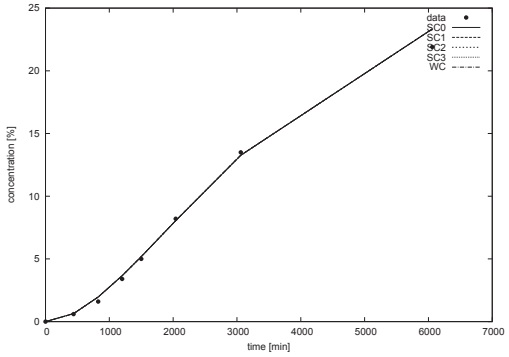
(b) Trajectories for dipentane (y_2).



(c) Trajectories for allo-ocimene (y_3).



(d) Trajectories for α/β-pyronene (y_4).



(e) Trajectories for the dimer (y_5).

Figure 8.6.: Comparison of the trajectories for the thermal isomerization of α-pinene at the temperature of 204.5° for uncertain k_5 with the stochastic collocation approach with integration levels 0–3 and the worst-case solution.

Table 8.5.: Estimates of the reaction rates for the thermal isomerization of α-pinene at the temperature of 189.5° for uncertain k_4 and k_5 with the stochastic collocation approach with integration levels 0–3 and the worst-case solution.

level	k_1	k_2	k_3	obj. val.
0	$5.800775e{-5}$	$2.845436e{-5}$	$1.338442e{-5}$	60.30268
1	$5.807904e{-5}$	$2.853087e{-5}$	$1.839092e{-5}$	54.01515
2	$5.811355e{-5}$	$2.856171e{-5}$	$1.844615e{-5}$	51.06767
3	$5.818919e{-5}$	$2.863034e{-5}$	$1.856403e{-5}$	44.88195
WC	$5.800564e{-5}$	$2.845638e{-5}$	$1.839596e{-5}$	60.30533

8.6 where again the trajectories for allo-ocimene (y_3) and α- and β-pyronene follow the measurement data rather roughly no matter what integration level is used. Also the optimal values of the objective function as given in table 8.4 show no improvement for increased integration level in the stochastic collocation approach.

Note, however, that the absolute values of the objective function at the optimum alone may be misleading. Although this value for temperature 189.5° is almost six times the value for temperature 204.5° in all five tested approaches, this does not mean that the parameter estimation was less successfull for the lower temperature. Rather there are single points in the measurements at temperature 189.5° where the measurement error seems to be larger than for the rest of the points, e.g. at time 3060 for α-pinene (y_1) or time 10680 for α-pinene (y_1) and dipentane (y_2). Note here the different scales of the vertical axes for the five species.

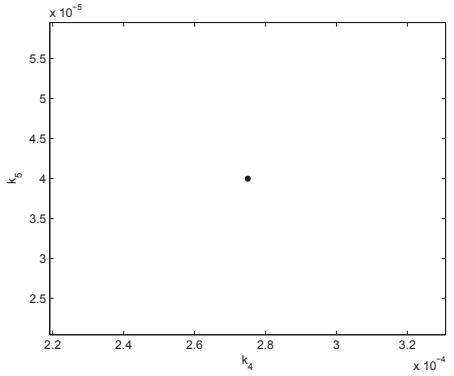
8.2.2. Two uncertainties

For the second estimation the fourth coefficient k_4 is also assumed to be uncertain, varying in the uncertainty interval $[2.19176e{-4}, 3.30804e{-4}]$, in addition to the uncertain coefficient k_5 from the first estimation. The corresponding sparse grids for stochastic collocation with integration levels 0 to 3 are shown in figure 8.7.

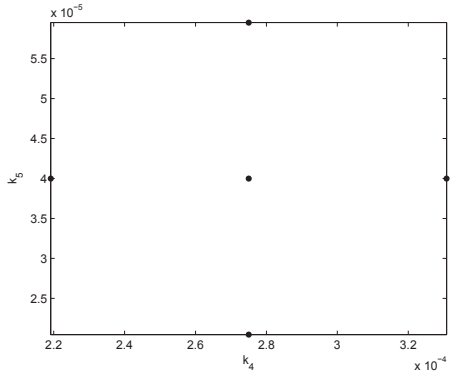
The numerical results from the parameter estimation with the stochastic collocation approach with integration levels 0–3 and the worst-case solution are given in tables 8.5 and 8.6 for the two different temperatures. Comparative plots of the corresponding trajectories are shown in figures 8.8 and 8.9, respectively.

This estimation confirms the findings of the first estimation with only one uncertainty and shows even greater effects.

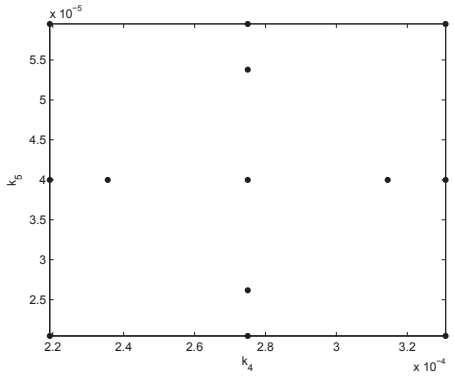
At the lower temperature, increasing the integration level in the stochastic collocation approach visibly improves the estimations. The trajectories for allo-ocimene (y_3) and α- and β-pyronene, which differ the most from the course of the measurement data, are drawn closer to the measurements, see figure 8.8, and the optimal value of the objective function is significantly improved at the higher levels of integration, see table 8.5. Yet the optimal parameter values for k_1, k_2, and k_3 do not differ much



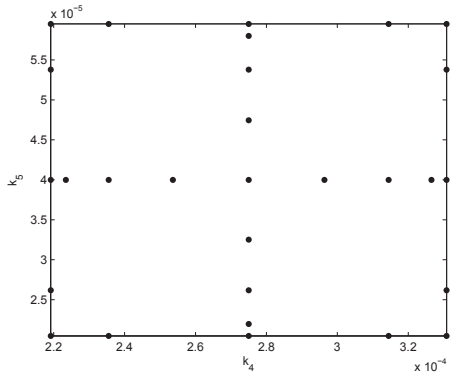
(a) Sparse grid of level 0 in two dimensions with 1 point.



(b) Sparse grid of level 1 in two dimensions with 5 points.



(c) Sparse grid of level 2 in two dimensions with 13 points.

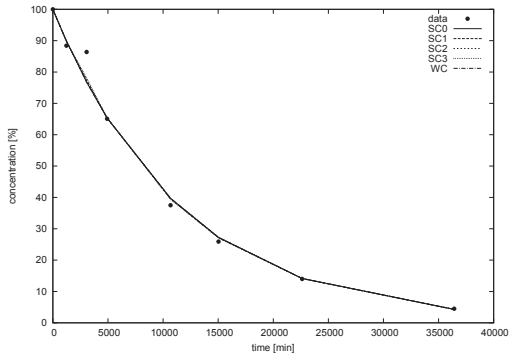


(d) Sparse grid of level 3 in two dimensions with 29 points.

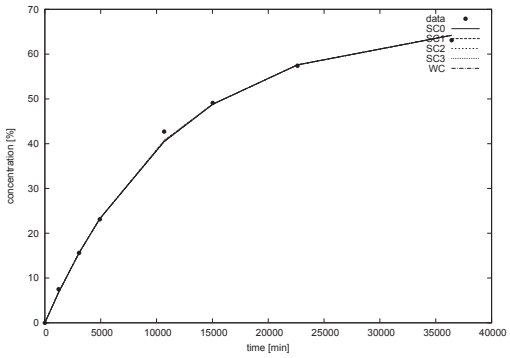
Figure 8.7.: Sparse grids in two dimensions for stochastic collocation with integration levels 0–3.

Table 8.6.: Estimates of the reaction rates for the thermal isomerization of α-pinene at the temperature of 204.5° for uncertain k_4 and k_5 with the stochastic collocation approach with integration levels 0–3 and the worst-case solution.

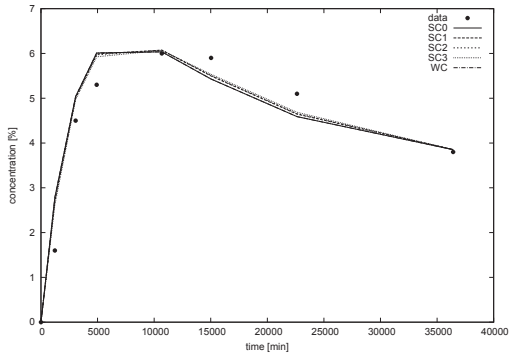
level	k_1	k_2	k_3	obj. val.
0	$2.250295e{-4}$	$1.281324e{-4}$	$7.305482e{-5}$	86.52683
1	$2.250547e{-4}$	$1.281206e{-4}$	$7.324688e{-5}$	87.35838
2	$2.229933e{-4}$	$1.269142e{-4}$	$7.296550e{-5}$	85.63706
3	$2.227828e{-4}$	$1.267906e{-4}$	$7.292478e{-5}$	85.45488
WC	$2.250549e{-4}$	$1.281464e{-4}$	$7.303541e{-5}$	86.57937



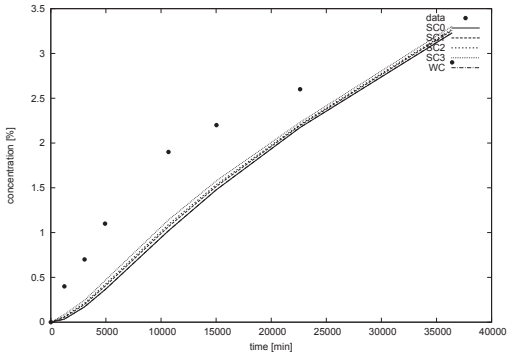
(a) Trajectories for α-pinene (y_1).



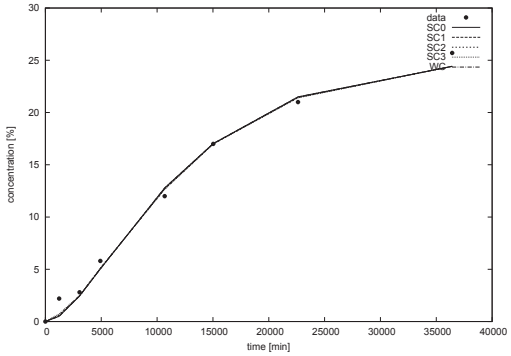
(b) Trajectories for dipentane (y_2).



(c) Trajectories for allo-ocimene (y_3).

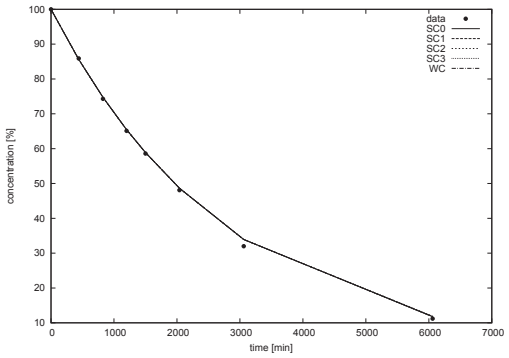


(d) Trajectories for α/β-pyronene (y_4).

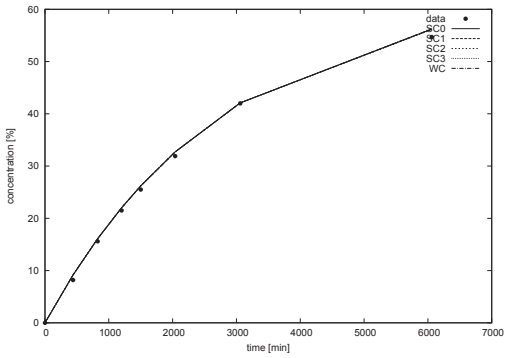


(e) Trajectories for the dimer (y_5).

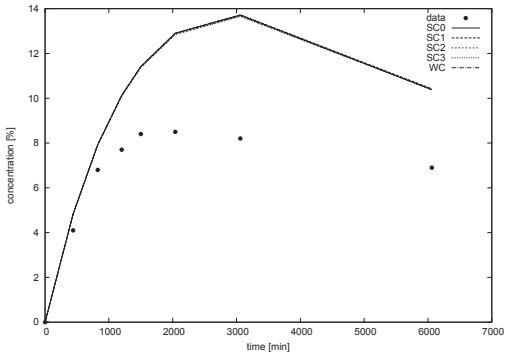
Figure 8.8.: Comparison of the trajectories for the thermal isomerization of α-pinene at the temperature of 189.5° for uncertain k_4 and k_5 with the stochastic collocation approach with integration levels 0–3 and the worst-case solution.



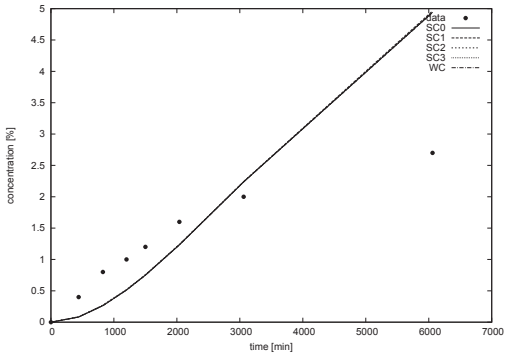
(a) Trajectories for α-pinene (y_1).



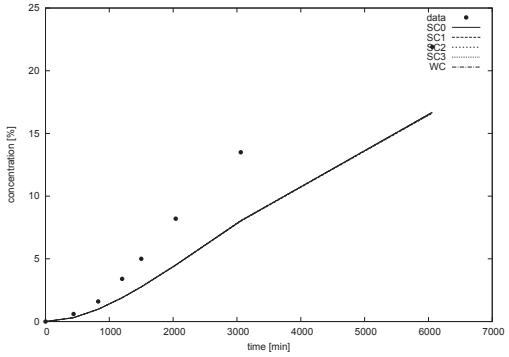
(b) Trajectories for dipentane (y_2).



(c) Trajectories for allo-ocimene (y_3).



(d) Trajectories for α/β-pyronene (y_4).



(e) Trajectories for the dimer (y_5).

Figure 8.9.: Comparison of the trajectories for the thermal isomerization of α-pinene at the temperature of 204.5° for uncertain k_4 and k_5 with the stochastic collocation approach with integration levels 0–3 and the worst-case solution.

from those of the previous estimation. This may also indicate that they are close to the true values.

At the higher temperature, quite the opposite is true. The trajectories are still indistinguishable for all five estimation approaches in figure 8.9 and also the values of the objective function show no improvement for higher integration levels, see table 8.6. Compared to the parameter estimation with only one uncertainty the parameter estimates have changed significantly and the optimal values of the objective function have also grown one order of magnitude. This results from the fact that the uncertainty intervals for k_4 and k_5 are chosen like for the lower temperature. So they do not contain the true values of the reaction rates at the higher temperature. The estimated value of k_4 from the previous estimation is about $5.88e{-}4$ and thus far outside the considered uncertainty interval $[2.19176e{-}4, 3.30804e{-}4]$. These wrong assumptions on the values of the uncertain parameters obviously cannot be compensated by an optimal choice of the remaining parameters.

This example shows the importance of an appropriate choice of the uncertainty intervals for the uncertain model parameters when using them for the stochastic collocation approach. If they do not contain the true parameter values or are at least very close to them, the results with our new approach will always be in the vicinity of the results from the approximated worst-case approach, independently of the chosen level of integration.

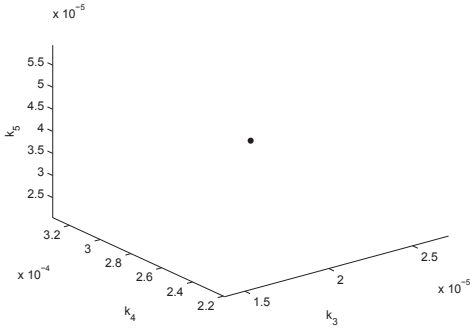
8.2.3. Three uncertainties

The third estimation considers the case of three uncertain coefficients k_3, k_4, and k_5, where the uncertainty intervals for the latter two are the same as in the previous estimations and k_3 is assumed to vary in $[1.38073e{-}5, 2.71387e{-}5]$. The corresponding sparse grids for stochastic collocation with integration levels 0 to 3 are shown in figure 8.10.

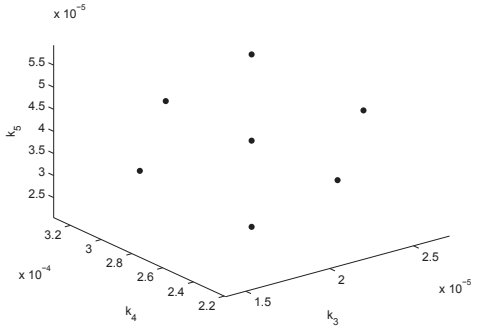
The numerical results from the parameter estimation with the stochastic collocation approach with integration levels 0–3 and the worst-case solution are given in tables 8.7 and 8.8 for the two different temperatures. Comparative plots of the corresponding trajectories are shown in figures 8.11 and 8.12, respectively.

The estimation with three uncertain coefficients further corroborates the findings of the preceding estimations. At 189.5°, the estimated parameter values for k_1 and k_2 given in table 8.7 are the same size as in tables 8.3 and 8.5. The optimal values of the objective function again decrease with growing order of integration in the stochastic collocation method, see table 8.7, and the trajectories depicted in figure 8.11 approach the measurement data even closer than before. The solution of the approximated worst-case approach is pretty close to the solution from the stochastic collocation method with level-0 integration that uses just the nominal values of the uncertain coefficients and ignores their uncertain nature. This is the case for all estimations in this section.

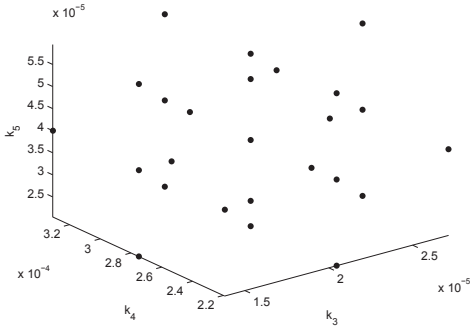
At 204.5°, in contrast, the estimation again cannot recover the system behavior. Though the optimal values of the objective function are only the same order of



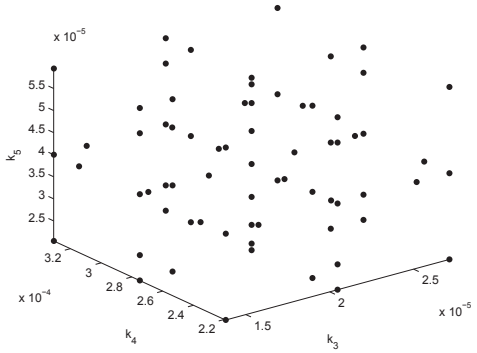
(a) Sparse grid of level 0 in three dimensions with 1 point.



(b) Sparse grid of level 1 in three dimensions with 7 points.



(c) Sparse grid of level 2 in three dimensions with 25 points.

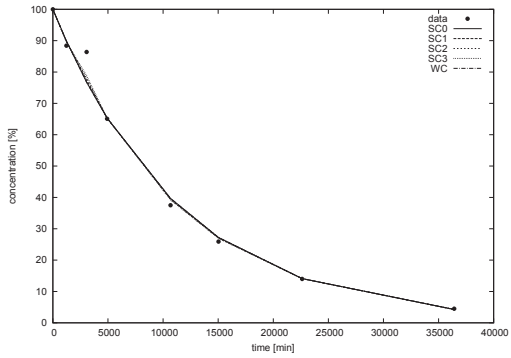


(d) Sparse grid of level 3 in three dimensions with 69 points.

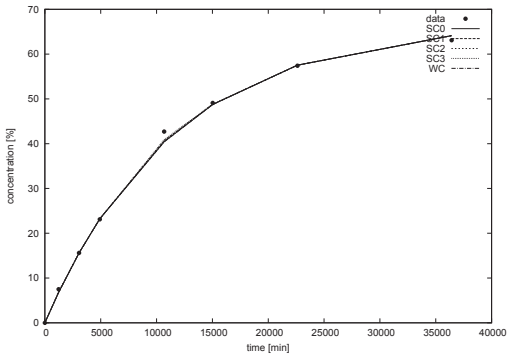
Figure 8.10.: Sparse grids in three dimensions for stochastic collocation with integration levels 0–3.

Table 8.7.: Estimates of the reaction rates for the thermal isomerization of α-pinene at the temperature of 189.5° for uncertain k_3, k_4, and k_5 with the stochastic collocation approach with integration levels 0–3 and the worst-case solution.

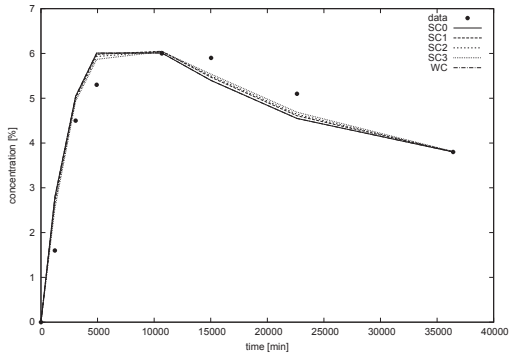
level	k_1	k_2	obj. val.
0	$5.798804e{-5}$	$2.857028e{-5}$	60.44601
1	$5.807218e{-5}$	$2.865334e{-5}$	53.13311
2	$5.815539e{-5}$	$2.872015e{-5}$	46.33259
3	$5.826777e{-5}$	$2.881172e{-5}$	37.85496
WC	$5.798643e{-5}$	$2.857173e{-5}$	60.44980



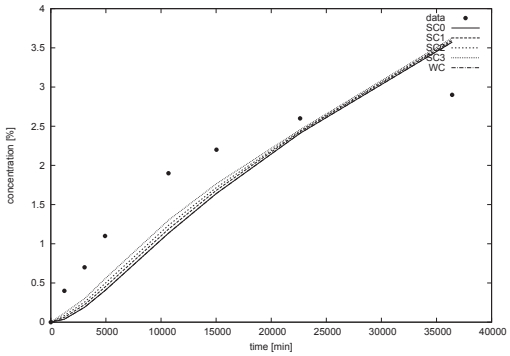
(a) Trajectories for α-pinene (y_1).



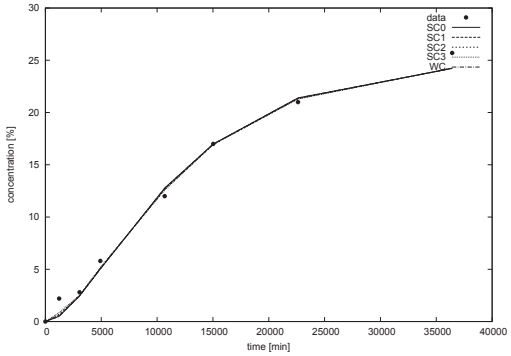
(b) Trajectories for dipentane (y_2).



(c) Trajectories for allo-ocimene (y_3).

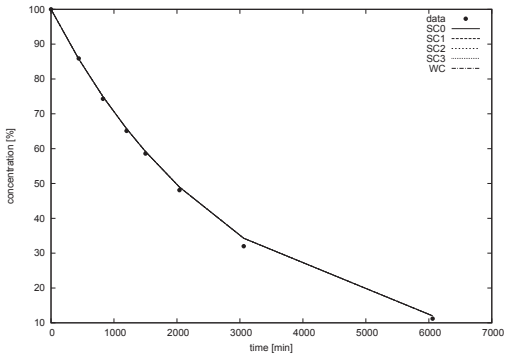


(d) Trajectories for α/β-pyronene (y_4).

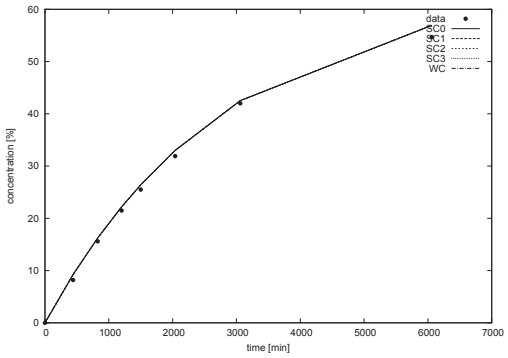


(e) Trajectories for the dimer (y_5).

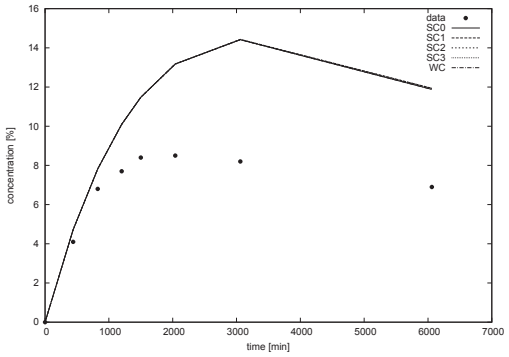
Figure 8.11.: Comparison of the trajectories for the thermal isomerization of α-pinene at the temperature of 189.5° for uncertain k_3, k_4, and k_5 with the stochastic collocation approach with integration levels 0–3 and the worst-case solution.



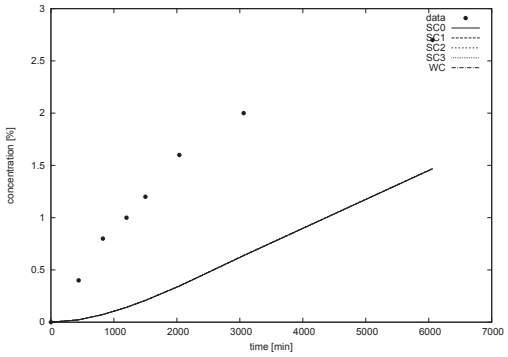
(a) Trajectories for α-pinene (y_1).



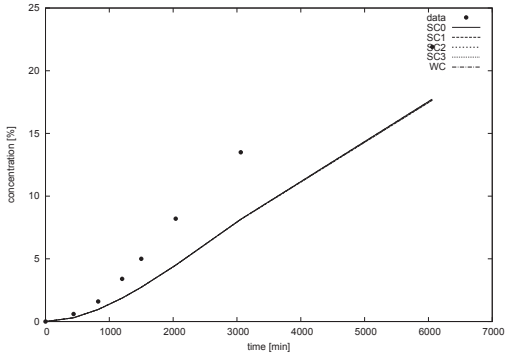
(b) Trajectories for dipentane (y_2).



(c) Trajectories for allo-ocimene (y_3).



(d) Trajectories for α/β-pyronene (y_4).



(e) Trajectories for the dimer (y_5).

Figure 8.12.: Comparison of the trajectories for the thermal isomerization of α-pinene at the temperature of 204.5° for uncertain k_3, k_4, and k_5 with the stochastic collocation approach with integration levels 0–3 and the worst-case solution.

Table 8.8.: Estimates of the reaction rates for the thermal isomerization of α-pinene at the temperature of 204.5° for uncertain k_3, k_4, and k_5 with the stochastic collocation approach with integration levels 0–3 and the worst-case solution.

level	k_1	k_2	obj. val.
0	$2.263740e{-4}$	$1.235756e{-4}$	96.89640
1	$2.263776e{-4}$	$1.235216e{-4}$	97.84858
2	$2.258535e{-4}$	$1.232300e{-4}$	97.32755
3	$2.261734e{-4}$	$1.234118e{-4}$	97.63303
WC	$2.263841e{-4}$	$1.235819e{-4}$	96.95529

magnitude as in the previous section with only two uncertainties, just a little larger, in particular the trajectories of α- and β-pyronene are quite different in figure 8.12 when compared to figure 8.9. The qualitative behavior is the same but quantitatively the trajectories differ. This again underlines the importance of choosing the correct interval for the uncertain parameter. In this example, the coefficient k_3 is the only one defining the solution for the differential equation corresponding to this species and thus the wrong uncertainty interval completely spoils the estimation.

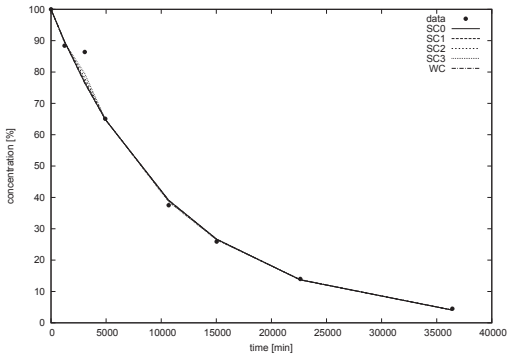
8.2.4. Four uncertainties

The final fourth estimation only tries to determine the optimal value of k_1 while the other coefficients k_2, k_3, k_4, and k_5 are assumed to be uncertain. The uncertainty interval for the new uncertainty k_2 is set to $[2.83301e{-5}, 3.09379e{-5}]$ and the intervals of the remaining uncertainties are unchanged with respect to the former estimations.

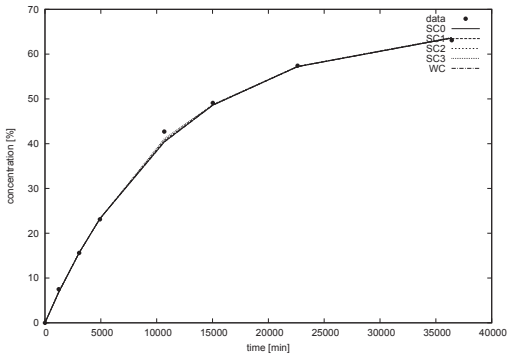
The estimations with three and two uncertain coefficients have already shown that at temperature 204.5°neither the stochastic collocation method nor the approximated worst-case method can estimate the unknown parameters because of the wrong uncertainty intervals for the uncertain coefficients. Already here computational time gets disproportionally long. Thus we do not try to estimate k_1 with the other four parameters being uncertain for the higher temperature.

The numerical results from the parameter estimation with the stochastic collocation approach with integration levels 0–3 and the worst-case solution for the lower temperature are given in table 8.9. As the resulting four-dimensional sparse grids cannot be presented in pictorial form, the numbers of points in the sparse grid are included in this table. Comparative plots of the corresponding trajectories are shown in figure 8.13.

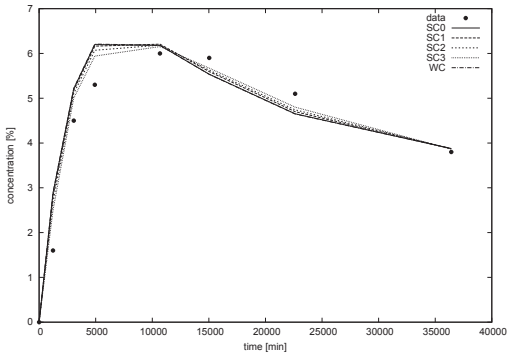
At 189.5°, the results of the parameter estimation with four uncertainties reveal no new results. The estimates of k_1 are pretty much the same as in the preceding estimations with fewer uncertainties. The optimal values of the objective function decrease a little further, see table 8.9, and the trajectories again get closer to the measurements with increasing integration level. The approximate worst-case solu-



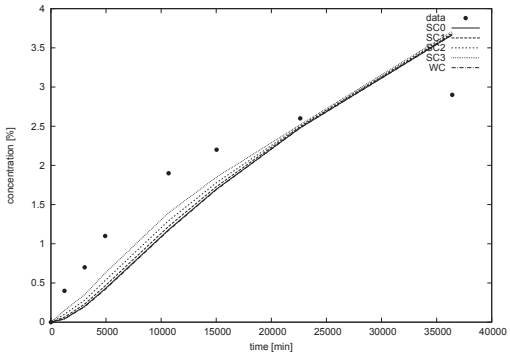
(a) Trajectories for α-pinene (y_1).



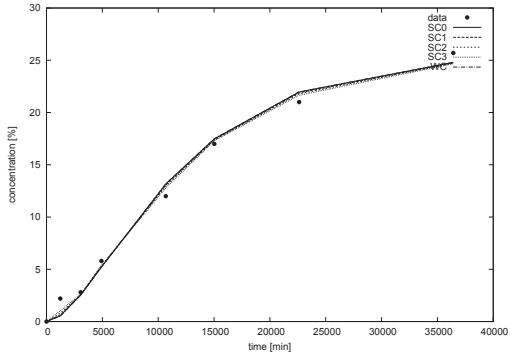
(b) Trajectories for dipentane (y_2).



(c) Trajectories for allo-ocimene (y_3).



(d) Trajectories for α/β-pyronene (y_4).



(e) Trajectories for the dimer (y_5).

Figure 8.13.: Comparison of the trajectories for the thermal isomerization of α-pinene at the temperature of 189.5° for uncertain k_2, k_3, k_4, and k_5 with the stochastic collocation approach with integration levels 0–3 and the worst-case solution.

Table 8.9.: Estimates of the reaction rates for the thermal isomerization of α-pinene at the temperature of 189.5° for uncertain k_2, k_3, k_4, and k_5 with the stochastic collocation approach with integration levels 0–3 and the worst-case solution.

level	points	k_1	obj. val.
0	1	$5.826678e{-5}$	60.52783
1	9	$5.831997e{-5}$	54.07495
2	41	$5.842482e{-5}$	43.41017
3	137	$5.856877e{-5}$	30.49980
WC	–	$5.825164e{-5}$	61.90875

tion again corresponds to the solution from the stochastic collocation approach with integration of level 0.

9. Conclusion

We conclude this thesis with a summary of the major points and contributions and a list of remaining questions that offer opportunities for future research.

9.1. Summary

In this thesis we addressed the problems that arise in numerical optimization due to errors in the measurement data and due to uncertainties concerning the mathematical model. In real world applications, parameter estimations for determining model coefficients are usually based on experimental observations that are naturally subject to measurement errors. But there is often also some unclarity about the correct mathematical model itself because of lacking knowledge about the process under consideration. These topics were treated independently in the preceding chapters.

The task of identifying parameters in process models from experimental measurements is one of the main application areas for numerical optimization methods. Classical estimation procedures like least squares estimation are very sensitive to deviations from the basic assumptions. These primarily include randomness, independence, and normal distribution of the measurement errors. None of these assumptions need to be satisfied in reality and often they are not. We showed in simple examples how even single outliers in the data can completely spoil a least squares estimation and render its solution useless. As a remedy, we revisited two other parameter estimators, namely least absolute deviations estimation and Huber's M-estimator that are more robust in the sense that they can cope with measurements that deviate from the bulk of the data. Their properties were compared to those of the least squares estimator in a concise way.

After a general introduction to the generalized Gauss-Newton method for solving nonlinear constrained optimization problems, we showed how this method can be applied to the two robust estimators. We did this by deriving the actual Gauss-Newton algorithms for least absolute deviations and Huber estimation from the optimality conditions of the respective optimization problems and demonstrating their common structure. For both estimators the linearized problems at the Gauss-Newton iterations can be solved by means of a generalized inverse under some regularity assumptions. Thus also local convergence of the respective methods follows from the general case.

For the case that the derivative matrices needed in the Gauss-Newton iterations cannot or shall not be computed exactly, there exists some theory for general nonlinear optimization problems how the errors that are introduced due to this inexactness

can be compensated by an additional term in the objective function of an SQP problem. We adapted this approach to the special cases of solving the nonlinear least absolute deviations problem and Huber's nonlinear M-estimator by means of the Gauss-Newton method. We derived the precise local algorithms and proved stability of the active sets of the optimization problems in the neighborhood of a solution and from this also local convergence of the algorithms.

Often not only the measurement data are perturbed by errors but there is also some uncertainty concerning the values of model coefficients that are not subject to optimization. A common treatment of this problem is the worst-case approach of the robust optimization methodology. This means, a robust counterpart is formulated for the problem under consideration that requires the constraints to hold for all realizations of the uncertain parameters. Then the worst possible solution for the robust counterpart is computed by solving a minimax optimization problem. We show how under some assumptions on the number of equality constraints the robust counterpart can be approximated by a first-order Taylor expansion such that the resulting inner optimization problem of the minimax problem can be solved analytically. That way, we are just left with a standard nonlinear minimization problem that approximates the worst-case solution. This approximate worst-case solution can be used as a reference solution for the alternative approach developed afterwards in this thesis.

As the worst-case approach is often overly conservative for many practical applications, in particular if there are no critical restrictions where we have to stay safely away from the bounds, we developed an alternative approach which is closer to an average solution over the set of uncertainties. We therefore combined the stochastic collocation approach that has been successfully used for optimizing PDE constrained estimation and control problems before with the multiple shooting approach for discretizing the ODE constraints of the problems considered in this thesis. The result is an optimization problem that defines an average solution of the original optimization problem over the uncertainty set that still satisfies all constraints. As this average solution is based on multivariate quadrature rules for computing the expected values of the model responses, we also show how sparse grid methods can be applied in order to reduce the number of support nodes needed by the cubature formula. We show how the resulting large-scale nonlinear optimization problem can be solved by means of a generalized Gauss-Newton method and how the special structure of the linearized problems can be exploited at each iteration for efficient solution.

Finally, the theory of parts I and II of this thesis is put into practice by means of some illustrative example problems. We first demonstrated the effects of the robust parameter estimators in comparison to least squares estimation by two examples from systems biology and chemistry with different kinds of measurement errors and outliers in the data. It was clearly visible in the example problems how these errors negatively affect the least squares solutions while the robust solutions are much less influenced by bad points.

Then we considered the problem of uncertain model coefficients on the basis of two examples that were again taken from biology and chemistry. Here we compared the

solutions of our stochastic collocation adaption using several levels of integration with that of the approximated worst-case approach for uncertainty sets of different sizes. In particular the example of the chemical reaction system showed nicely that stochastic collocation approaches the measurements closer that the worst-case solution and that this effect gets the more pronounced the higher the integration level for computing the expected values of the model responses is chosen. However, higher integration levels also need more points and thus increase computational time significantly.

9.2. Future research directions

While the investigations of this thesis were able to shed light on several problems that arise in numerical optimization due to measurement errors and uncertain model coefficients, they also rose some new questions. We see several points where future research is necessary and can help to gain new insights.

When computing parameter estimates with the least squares method, it is common practice also to give confidence intervals around the nominal estimates that contain the true parameter value with a certain high probability. These confidence regions are derived with a view to the statistics of the parameter estimates and are based on linear approximations of the covariance matrix. For the robust estimators treated in this thesis, these confidence intervals have no statistical meaning such that the reliability or inherent uncertainty of the results cannot be assessed for their solutions. Although there is some theory on robust confidence intervals, to our knowledge there is no theory that connects these robust confidence regions with the results of the robust estimators in such a way as the confidence regions that are usually computed for least squares estimates are connected to their covariance matrix. Further research in this direction might be worthwhile in order to strengthen the validity of parameter estimates computed by least absolute deviations estimation of Huber's M-estimator or some other robust estimation procedure.

In case that the derivative matrices for the iterations of the Gauss-Newton method are computed only inexactly, we were only able to show local convergence of the method described in this thesis. Here further work is necessary to globalize the convergence results. This concerns in particular proof of level functions being compatible with the inexact Gauss-Newton method with the additional terms in the objective function and also concrete rules for the choice of the approximated Jacobians.

Having uncertain coefficients in the mathematical model of the process under consideration is usually treated by computing a solution that ensures that the constraints are satisfied for all realizations of the uncertainties. But solving the corresponding robust counterpart problem is mathematically challenging and there are even problem classes for which no tractable formulation of the robust counterpart is known. Still, the approach presented in this thesis is only valid for a very limited class of optimization problems with its restriction on the number of equality constraints and, furthermore, the linear approximation used here may be insufficient for many problems even within this class. Here it would be nice to derive and implement a more

general procedure in order to get more realistic reference solutions to which the results of the adapted stochastic collocation method can be compared. But this is also an interesting and important problem in its own right.

For the combination of the stochastic collocation method with the multiple shooting approach a deeper mathematical analysis of the properties of the resulting procedure would be nice. Also our very basic implementation offers several starting points for improvement. More efficient grid computation for allowing higher integration levels is one of them. But as this also leads to much larger linear systems that have to be solved at the Gauss-Newton iterations, this is only useful when at the same time speeding up the numerical solution of these systems. Here, the special structure of the Jacobian resulting from multiple shooting lends itself to parallelization of the method.

List of Figures

1.1. Interconnections of verification, validation, and uncertainty quantification (based on a figure from Oberkampf et al. [106], reproduced in the report of the Committee for Verification, Validation and Uncertainty Quantification [38]). . . . 3
1.2. Sources of uncertainties and errors in the different phases of the modeling and simulation process (illustration based on figures 1 and 2 by Alvin et al. [2]). . . . 5

2.1. Illustrative examples for the different types of outliers in the data for a simple linear regression with least squares and least absolute deviations estimator. Note that Huber's M-estimator will give results somewhere inbetween the l_2 and l_1 estimates, depending on the value of the tuning constant. . . . 17
2.2. Concept of the multiple shooting procedure for an ordinary differential equation. . . . 22
2.3. Illustrative example for non-uniqueness of the linear robust estimators. 31

6.1. Comparison of sparse grids with Clenshaw-Curtis points and corresponding full grids. . . . 142

7.1. NK-κB pathway. Dashed arrows signify indirect dependency. . . . 152
7.2. Measurement data for the NF-κB pathway with three gross outliers. . 154
7.3. Trajectories corresponding to the estimated parameters for the NF-κB pathway with three gross outliers. . . . 156
7.4. Trajectories corresponding to the estimated parameters for the NF-κB pathway with 10% overall noise. . . . 158
7.5. Schematic sketch of the denitrogenization of pyridine. . . . 158
7.6. Measurement data for the denitrogenization of pyridine. . . . 160
7.7. Trajectories corresponding to the estimated parameters for the denitrogenization of pyridine with four outliers. . . . 162

8.1. Sparse grids of levels 0–2 for the Lotka-Volterra example. . . . 166
8.2. Trajectories of the parameter estimation for the Lotka-Volterra model with the stochastic collocation approach with integration levels 0–2 and the worst-case solution. . . . 167
8.3. Thermal isomerization of α-pinene. . . . 168

8.4. Sparse grids in one dimension for stochastic collocation with integration levels 0–3. . . . 170
8.5. Comparison of the trajectories for the thermal isomerization of α-pinene at the temperature of 189.5° for uncertain k_5 with the stochastic collocation approach with integration levels 0–3 and the worst-case solution. . . . 171
8.6. Comparison of the trajectories for the thermal isomerization of α-pinene at the temperature of 204.5° for uncertain k_5 with the stochastic collocation approach with integration levels 0–3 and the worst-case solution. . . . 172
8.7. Sparse grids in two dimensions for stochastic collocation with integration levels 0–3. . . . 174
8.8. Comparison of the trajectories for the thermal isomerization of α-pinene at the temperature of 189.5° for uncertain k_4 and k_5 with the stochastic collocation approach with integration levels 0–3 and the worst-case solution. . . . 175
8.9. Comparison of the trajectories for the thermal isomerization of α-pinene at the temperature of 204.5° for uncertain k_4 and k_5 with the stochastic collocation approach with integration levels 0–3 and the worst-case solution. . . . 176
8.10. Sparse grids in three dimensions for stochastic collocation with integration levels 0–3. . . . 178
8.11. Comparison of the trajectories for the thermal isomerization of α-pinene at the temperature of 189.5° for uncertain k_3, k_4, and k_5 with the stochastic collocation approach with integration levels 0–3 and the worst-case solution. . . . 179
8.12. Comparison of the trajectories for the thermal isomerization of α-pinene at the temperature of 204.5° for uncertain k_3, k_4, and k_5 with the stochastic collocation approach with integration levels 0–3 and the worst-case solution. . . . 180
8.13. Comparison of the trajectories for the thermal isomerization of α-pinene at the temperature of 189.5° for uncertain k_2, k_3, k_4, and k_5 with the stochastic collocation approach with integration levels 0–3 and the worst-case solution. . . . 182

List of Tables

2.1. Numerical results for the examples concerning different types of outliers in a linear regression using least squares and least absolute deviations estimation. . . . 18
2.2. Comparison of least squares, least absolute deviations, and Huber estimator for linear problems. . . . 29

6.1. Common quadrature weight functions and their corresponding orthogonal polynomials (adapted from Stoer [131]). . . . 131

7.1. Coefficients in model (7.1) for the NF-κB signaling system. . . . 153
7.2. Estimates of the parameters and numbers of iterations for the NF-κB pathway with three outliers in nuclear NF-κB (x_4). . . . 155
7.3. Estimates of the parameters and numbers of iterations for the NF-κB pathway with 10% noise. . . . 157
7.4. Estimates for the parameter estimation and numbers of iterations of the denitrogenization of pyridine with four outliers. . . . 160

8.1. Estimates of the parameters in the prey equation of the Lotka-Volterra model with the stochastic collocation approach with integration levels 0–2 and the worst-case solution. . . . 166
8.2. Parameter estimates for the isomerization of α-pinene at temperature 189.5° found by Rodriguez-Fernandez et al. [113]. . . . 169
8.3. Estimates of the reaction rates for the thermal isomerization of α-pinene at the temperature of 189.5° for uncertain k_5 with the stochastic collocation approach with integration levels 0–3 and the worst-case solution. . . . 170
8.4. Estimates of the reaction rates for the thermal isomerization of α-pinene at the temperature of 204.5° for uncertain k_5 with the stochastic collocation approach with integration levels 0–3 and the worst-case solution. . . . 170
8.5. Estimates of the reaction rates for the thermal isomerization of α-pinene at the temperature of 189.5° for uncertain k_4 and k_5 with the stochastic collocation approach with integration levels 0–3 and the worst-case solution. . . . 173

8.6. Estimates of the reaction rates for the thermal isomerization of α-pinene at the temperature of 204.5° for uncertain k_4 and k_5 with the stochastic collocation approach with integration levels 0–3 and the worst-case solution. 174
8.7. Estimates of the reaction rates for the thermal isomerization of α-pinene at the temperature of 189.5° for uncertain k_3, k_4, and k_5 with the stochastic collocation approach with integration levels 0–3 and the worst-case solution. 178
8.8. Estimates of the reaction rates for the thermal isomerization of α-pinene at the temperature of 204.5° for uncertain k_3, k_4, and k_5 with the stochastic collocation approach with integration levels 0–3 and the worst-case solution. 181
8.9. Estimates of the reaction rates for the thermal isomerization of α-pinene at the temperature of 189.5° for uncertain k_2, k_3, k_4, and k_5 with the stochastic collocation approach with integration levels 0–3 and the worst-case solution. 183

Bibliography

[1] G. Airy. Letter from professor Airy, Astronomer Royal, to the editor. *Astronomical Journal*, 4(18):137–138, 1856.

[2] K. F. Alvin, W. L. Oberkampf, K. V. Diegert, and B. M. Rutherford. Uncertainty quantification in computational structural dynamics: A new paradigm for model validation. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 3243, pages 1191–1198, 1998.

[3] R. Andersen. *Modern Methods for Robust Regression*, volume 152 of *Quantitative Application in the Social Sciences*. SAGE Publications, 2008.

[4] D. Andrews, F. Bickel, F. Hampel, P. Huber, W. Rogers, and J. Tukey. *Robust Estimates of Location: Survey and Advances*. Princeton University Press, 1972.

[5] F. J. Anscombe. Rejection of outliers. *Technometrics*, 2(1):123–147, 1960.

[6] U. Ascher. Collocation for two-point boundary value problems revisited. *SIAM Journal on Numerical Analysis*, 23(3):596–609, 1986.

[7] V. Bär. Ein Kollokationsverfahren zur numerischen Lösung allgemeiner Mehrpunktrandwertaufgaben mit Schalt- und Sprungbedingungen mit Anwendungen in der Optimalen Steuerung und der Parameteridentifizierung. Diploma thesis, University of Bonn, 1983.

[8] I. Barrodale and F. D. K. Roberts. An improved algorithm for discrete l_1 linear approximation. *SIAM Journal on Numerical Analysis*, 10(5):839–848, 1973.

[9] I. Barrodale and F. D. K. Roberts. Algorithm 478: Solution of an overdetermined system of linear equations in the l_1 norm. *Communications of the ACM*, 17:319–320, 1974.

[10] V. Barthelmann, E. Novak, and K. Ritter. High dimensional polynomial interpolation on sparse grids. *Advances in Computational Mathematics*, 12:273–288, 2000.

[11] A. E. Beaton and J. W. Tukey. The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, 16(2):147–185, 1974.

[12] A. Ben-Tal. Second-order and related extremality conditions in nonlinear programming. *Journal of Optimization Theory and Applications*, 31(2):143–165, 1980.

[13] A. Ben-Tal, D. den Hertog, A. de Waegenaere, B. Melenberg, and G. Rennen. Robust solutions of optimization problems affected by uncertain probabilities. CentER Working Paper Series 2011-061, Tilburg University, 2011. Available at SSRN: http://ssrn.com/abstract=1853428.

[14] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust Optimization*. Princeton Series in Applied Mathematics. Princeton University Press, 2009.

[15] A. Ben-Tal and A. Nemirovski. Robust convex optimization. *Mathematics of Operations Research*, 23(4):769–805, 1998.

[16] A. Ben-Tal and A. Nemirovski. Robust solutions of Linear Programming problems contaminated with uncertain data. *Mathematical Programming*, 88:411–421, 2000.

[17] A. Ben-Tal and A. Nemirovski. Robust optimization - methodology and applications. *Mathematical Programming, Series B*, 92:453–480, 2002.

[18] D. Bertsimas, D. B. Brown, and C. Caramanis. Theory and applications of robust optimization. *SIAM Review*, 53(3):464–501, 2011.

[19] J. T. Betts. *Practical Methods for Optimal Control and Estimation Using Nonlinear Programming*. Advances in Design and Control. SIAM, 2010.

[20] T. Binder and E. Kostina. Gauss-Newton methods for robust parameter estimation. In *Model Based Parameter Estimation: Theory and Applications*. Springer, to appear.

[21] D. Birkes and Y. Dodge. *Alternative Methods of Regression*. John Wiley & Sons, 1993.

[22] H. G. Bock. Numerical treatment of inverse problems in chemical reaction kinetics. In K.-H. Ebert, P. Deuflhard, and W. Jäger, editors, *Modelling of Chemical Reaction Systems*, volume 18 of *Springer Series in Chemical Physics*, pages 102–125. Springer, 1981.

[23] H. G. Bock. Recent advances in parameter identification techniques for ODE. In P. Deuflhard and E. Hairer, editors, *Numerical Treatment of Inverse Problems in Differential and Integral Equations*, pages 95–121. Birkhäuser, 1983.

[24] H. G. Bock. Randwertproblemmethoden zur Parameteridentifizierung in Systemen nichtlinearer Differentialgleichungen. Bonner Mathematische Schriften 187, 1987.

[25] H. G. Bock, S. Körkel, E. Kostina, and J. P. Schlöder. Robustness aspects in parameter estimation, optimal design of experiments and optimal control. In W. Jäger, R. Rannacher, and J. Warnatz, editors, *Reactive Flows, Diffusion and Transport*, pages 117–146. Springer, 2007.

[26] H. G. Bock, E. Kostina, and J. P. Schlöder. On the role of natural level functions to achieve global convergence for damped Newton methods. In M. Powell, editor, *System Modelling and Optimization. Methods, Theory and Applications*. Kluwer, 2000.

[27] 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.

[28] J. M. Booker, M. C. Anderson, and M. A. Meyer. The role of expert knowledge in uncertainty quantification (Are we adding more uncertainty or more understanding?). In B. Bodt and E. Wegman, editors, *Proceedings of the Seventh Annual U.S. Army Conference on Applied Statistics*, pages 155–161, 2001.

[29] A. Borzì. Multigrid and sparse-grid schemes for elliptic control problems with random coefficients. *Computation and Visualization in Science*, 13(4):153–160, 2010.

[30] A. Borzì, V. Schulz, C. Schillings, and G. von Winckel. On the treatment of distributed uncertainties in PDE-constrained optimization. *GAMM-Mitteilungen*, 33(2):230–246, 2010.

[31] A. M. Bruckstein, D. L. Donoho, and M. Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Review*, 51(1):34–81, 2009.

[32] R. Bulirsch. Die Mehrzielmethode zur numerischen Lösung von nichtlinearen Randwertproblemen und Aufgaben der optimalen Steuerung. Technical report, Carl-Cranz-Gesellschaft, 1971.

[33] J. Burkardt. 1D quadrature rules for sparse grids. http://people.sc.fsu.edu/~jburkardt/presentations/sgmga_1d_rules.pdf, July 2010.

[34] A. Cervantes and L. T. Biegler. Large-scale DAE optimization using a simultaneous NLP formulation. *AIChE Journal*, 44(5):1038–1050, 2004.

[35] W. Chauvenet. *A Manual of Spherical and Practical Astronomy*, volume II – Theory and Use of Astronomical Instruments, chapter Appendix – Criterion for the rejection of doubtful observations, pages 558–566. J. B. Lippincott & Co., Philadelphia, 4 edition, 1868. First edition: 1863.

[36] K. Chung and T. Yao. On lattices admitting unique Lagrange interpolations. *SIAM Journal on Numerical Analysis*, 14(4):735–743, 1977.

[37] C. W. Clenshaw and A. R. Curtis. A method for numerical integration on an automatic computer. *Numerische Mathematik*, 2:197–205, 1960.

[38] Committee on Mathematical Foundations of Verification, Validation, and Uncertainty Quantification and Board on Mathematical Sciences and Their Applications, Division of Engineering and Physical Sciences, National Research Council. Assessing the reliability of complex models: Mathematical and statistical foundations of verification, validation, and uncertainty quantification. Prepublication copy, The National Academics Press, 2012. http://www.nap.edu/catalog.php?record_id=13395.

[39] C. de Boor and A. Ron. Computational aspects of polynomial interpolation in several variables. *Mathematics of Computation*, 58(198):705–727, 1992.

[40] C. de Boor and A. Ron. The least solution for the polynomial interpolation problem. *Mathematische Zeitschrift*, 210:347–378, 1992.

[41] J. E. Dennis and R. E. Welsch. Techniques for nonlinear least squares and robust regression. *Communications in Statistics - Simulation and Computation*, 7(4):345–359, 1978.

[42] P. Deuflhard. A modified Newton method for the solution of ill-conditioned systems of nonlinear equations with application to multiple shooting. *Numerische Mathematik*, 22:289–315, 1974.

[43] M. Diehl, H. G. Bock, and E. Kostina. An approximation technique for robust nonlinear optimization. *Mathematical Programming*, 107:213–230, 2006.

[44] D. Draper. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):45–97, 1995.

[45] H. Ekblom and K. Madsen. Algorithms for non-linear Huber estimation. *BIT Numerical Mathematics*, 29(1):60–76, 1989.

[46] L. El Ghaoui and H. Lebret. Robust solutions to least squares problems with uncertain data. *SIAM Journal on Matrix Analysis and Applications*, 18:1035–1064, 1997.

[47] G. Faber. Über die interpolatorische Darstellung stetiger Funktionen. *Jahresbericht der Deutschen Mathematikervereinigung*, 23:192–210, 1914.

[48] R. C. Fair. On the robust estimation of econometric models. *Annals of Economic and Social Measurement*, 3(4):667–677, 1974.

[49] L. Fejér. Mechanische Quadraturen mit positiven Cotesschen Zahlen. *Mathematische Zeitschrift*, 37(1):287–309, 1933.

[50] A. V. Fiacco and G. P. McCormick. *Nonlinear programming: sequential unconstrained minimization techniques*. Classics in Applied Mathematics. SIAM, 1990.

[51] R. E. Fuguitt and J. E. Hawkins. Rate of the thermal isomerization of α-pinene in the liquid phase. *Journal of the American Chemical Society*, 69(2):319–322, 1947.

[52] V. Gabrel, C. Murat, and A. Thiele. Recent advances in robust optimization and robustness: An overview. Technical report, LAMSADE, Université Paris-Dauphine, Paris, France, 2012.

[53] B. Ganapathysubramanian and N. Zabaras. Modeling diffusion in random heterogeneous media: Data-driven models, stochastic collocation and the variational multiscale method. *Journal of Computational Physics*, 226:326–353, 2007.

[54] B. Ganapathysubramanian and N. Zabaras. Sparse grid collocation schemes for stochastic natural convection problems. *Journal of Computational Physics*, 225:652–685, 2007.

[55] J. Garcke. Sparse grid tutorial. Technical report, Technical University of Berlin, 2011.

[56] M. Gasca and J. I. Maeztu. On Lagrange and Hermite interpolation in R^k. *Numerische Mathematik*, 39:1–14, 1982.

[57] C. F. Gauss. *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium*. F. Perthes & J.H. Besser, Hamburg, 1809.

[58] W. Gautschi. Numerical quadrature in the presence of a singularity. *SIAM Journal on Numerical Analysis*, 4(3):357–362, 1967.

[59] R. C. Geary. Testing for normality. *Biometrika*, 34:209–242, 1947.

[60] H.-O. Georgii. *Stochastik – Einführung in die Wahrscheinlichkeitstheorie und Statistik*. de Gruyter, 4 edition, 2009.

[61] T. Gerstner and M. Griebel. Numerical integration using sparse grids. *Numerical Algorithms*, 18:209–232, 1998.

[62] T. Gerstner and M. Griebel. Dimension-adaptive tensor-product quadrature. *Computing*, 71:65–87, 2003.

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

[64] J. W. L. Glaisher. On the law of facility of errors of observations, and on the method of least squares. *Memoirs of the Royal Astronomical Society*, 39(Part II):75–124, 1872.

[65] J. W. L. Glaisher. On the rejection of discordant observations. *Monthly Notices of the Royal Astronomical Society*, 33:391–402, 1873.

[66] J. W. L. Glaisher. Note on a paper by Mr. Stone, 'On the rejection of discordant observations'. *Monthly Notices of the Royal Astronomical Society*, 34:251, 1874.

[67] A. Griewank and A. Walther. *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*. SIAM, 2008.

[68] F. Hampel. Is statistics too difficult? Research Report 81, ETH Zürich, 1997.

[69] F. R. Hampel. Robust estimation: A condensed partial survey. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 27:87–104, 1973.

[70] F. R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69:383–393, 1974.

[71] F. M. Hemez and S. W. Doebling. Model validation and uncertainty quantification. In *Proceedings of the 19th International Model Analysis Conference*, 2001.

[72] M. J. Hinich and P. P. Talwar. A simple method for robust regression. *Journal of the American Statistical Association*, 70:113–119, 1975.

[73] A. Hoffmann, A. Levchenko, M. L. Scott, and D. Baltimore. The IκB–NF-κB signaling module: Temporal control and selective gene activation. *Science*, 298:1241–1245, 2002.

[74] P. W. Holland and R. E. Welsch. Robust regression using iteratively reweighted least-squares. *Communications in Statistics - Theory and Methods*, 6(9):813–827, 1977.

[75] M. Holtz. *Sparse Grid Quadrature in High Dimensions with Applications in Finance and Insurance*, volume 77 of *Lecture Notes in Computational Science and Engineering*. Springer Verlag, 2011.

[76] P. J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

[77] P. J. Huber. Robust statistics: A review. *The Annals of Mathematical Statistics*, 43(4):1041–1067, 1972.

[78] P. J. Huber. *Robust Statistics*. John Wiley & Sons, 1981.

[79] P. J. Huber. Minimax aspects of bounded-influence regression. *Journal of the American Statistical Association*, 78:66–72, 1983.

[80] P. J. Huber. Robustness: Where are we now? In *L_1-Statistical Procedures and Related Topics*, volume 31 of *IMS Lecture Notes - Monograph Series*, pages 487–498. IMS, 1997.

[81] P. J. Huber and E. M. Ronchetti. *Robust Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, 2. edition, 2009.

[82] G. E. Karniadakis and J. Glimm. Uncertainty quantification in simulation science. Preface of Journal of Computational Physics 217, pp. 1–4, 2006.

[83] W. Klimke. *Uncertainty Modeling using Fuzzy Arithmetic and Sparse Grids*. PhD thesis, University of Stuttgart, 2005.

[84] A. Koberstein. *The Dual Simplex Method, Techniques for a Fast and Stable Implementation*. PhD thesis, University of Paderborn, 2005.

[85] E. Kostina. The long step rule in the bounded-variable dual simplex method: Numerical experiments. *Mathematical Methods of Operations Research*, 55:413–429, 2002.

[86] E. Kostina. Robust parameter estimation in dynamic systems. *Optimization and Engineering*, 5(4):461–484, 2004.

[87] E. Kostina and O. Kostyukova. A primal-dual active-set method for convex quadratic programming. Preprint, University of Heidelberg, 2003.

[88] S. Krishna, M. H. Jensen, and K. Sneppen. Minimal model of spiky oscillations in NF-κB signaling. *Proceedings of the National Academy of Sciences*, 103:10840–10845, 2006.

[89] P. S. Laplace. *Théorie Analytique des Probabilités*. V. Courcier, 2. edition, 1814.

[90] A. Legendre. *Nouvelles méthodes pour la détermination des orbites des comètes*. Firmin Didot, 1805.

[91] W. Li and J. J. Swetits. The linear l_1 estimator and the Huber M-estimator. *SIAM Journal on Optimization*, 8(2):457–475, 1998.

[92] G. Lin, A. M. Tartakovsky, and D. M.Tartakovsky. Uncertainty quantification via random domain decomposition and probabilistic collocation on sparse grids. *Journal of Computational Physics*, 229:6995–7012, 2010.

[93] T. Lipniacki, P. Paszek, A. R. Brasier, B. Luxon, and M. Kimmel. Mathematical model of NF-κB regulatory module. *Journal of Theoretical Biology*, 228:195–215, 2004.

[94] D. G. Luenberger. *Linear and Nonlinear Programming*. Addison-Wesley, 1984.

[95] D. L. Ma and R. D. Braatz. Worst-case analysis of finite-time control policies. *IEEE Transactions on Control Systems Technology*, 9(5):766–774, 2001.

[96] C. Maire and R. Boscovich. *De litteraria expeditione per pontificiam ditionem.* 1755.

[97] O. L. Mangasarian and S. Fromovitz. The Fritz John necessary optimality conditions in the presence of equality and inequality constraints. *Journal of Mathematical Analysis and Applications*, 17:37–47, 1967.

[98] O. L. Mangasarian and D. R. Musicant. Robust linear and support vector regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(9):950–955, 2000.

[99] M. Nattermann. Sensitivitätsanalyse und optimale Versuchsplanung für Parameterschätzprobleme, basierend auf der quadratischen Approximation der Konfidenzgebiete. Diploma thesis, University of Marburg, 2010.

[100] F. Nobile, R. Tempone, and C. G. Webster. A sparse grid stochastic collocation method for partial differential equations with random input data. *SIAM Journal on Numerical Analysis*, 46(5):2309–2345, 2008.

[101] J. Nocedal and S. J. Wright. *Numerical Optimization.* Springer, New York, 1999.

[102] E. Novak and K. Ritter. High dimensional integration of smooth functions over cubes. *Numerische Mathematik*, 75:79–97, 1995.

[103] E. Novak and K. Ritter. The curse of dimension and a universal method for numerical integration. In G. Nürnberger, J. Schmidt, and G. Walz, editors, *Multivariate Approximations and Splines*, pages 177–188. Birkhäuser, 1997.

[104] E. Novak and K. Ritter. Simple cubature formulas with high polynomial exactness. *Constructive Approximation*, 15:499–522, 1999.

[105] U. Novak and P. Deuflhard. Numerical identification of selected rate constants in large chemical reaction systems. *Applied Numerical Mathematics*, 1(1):59–75, 1985.

[106] W. L. Oberkampf, M. N. Sindir, and A. T. Conlisk. Guide for the verification and validation of computational fluid dynamics simulations. Tech Report G-077-98, American Institute for Aeronautics and Astronautics, 1998.

[107] T. Oden, R. Moser, and O. Ghattas. Computer predictions with quantified uncertainty, Part I. *SIAM News*, 43(9):1–4, 2010.

[108] H. O'Hara and F. J. Smith. Error estimation in the Clenshaw-Curtis quadrature formula. *The Computer Journal*, 11(2):213–219, 1968.

[109] M. R. Osborne, B. Presnell, and B. A. Turlach. On the LASSO and its dual. *Journal of Computational and Graphical Statistics*, 9(2):319–337, 2000.

[110] B. Peirce. Criterion for the rejection of doubtful observations. *The Astronomical Journal*, II(45):161–163, 1852.

[111] M. M. Putko, P. A. Newman, A. C. Taylor III, and L. L. Green. Approach for uncertainty propagation and robust design in CFD using sensitivity derivatives. AIAA Paper 2001-2528, American Institute of Aeronautics and Astronautics, 2001.

[112] J. R. Rice and J. S. White. Norms for smoothing and estimation. *SIAM Review*, 6(3):243–256, 1964.

[113] M. Rodriguez-Fernandez, J. A. Egea, and J. R. Banga. Novel metaheuristic for parameter estimation in nonlinear dynamic biological systems. *BMC Bioinformatics*, 7:483, 2006.

[114] S. M. Ross. Peirce's criterion for the elimination of suspect experimental data. *Journal of Engineering Technology*, 20:38–41, 2003.

[115] P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection*. Wiley, 1987.

[116] T. Sauer. Polynomial interpolation of minimal degree. *Numerische Mathematik*, 78:59–85, 1997.

[117] T. Sauer. Lagrange interpolation on subgrids of tensor product grids. *Mathematics of Computation*, 73:181–190, 2004.

[118] J. A. Scales and A. Gersztenkorn. Robust methods in inverse theory. *Inverse Problems*, 4:1071–1091, 1988.

[119] K. Schittkowski. Data fitting in partial differential algebraic equations: Some academic and industrial applications. *Journal of Computational and Applied Mathematics*, 163(1):29–57, 2004.

[120] J. P. Schlöder. Numerische Methoden zur Behandlung hochdimensionaler Aufgaben der Parameteridentifizierung. Bonner Mathematische Schriften 187, 1988.

[121] E. J. Schlossmacher. An iterative technique for absolute deviations curve fitting. *Journal of the American Statistical Association*, 68:857–859, 1973.

[122] A. Schreiber. *Die Methode von Smolyak bei der multivariaten Interpolation*. PhD thesis, University of Göttingen, 2000.

[123] G. I. Schuëller and H. A. Jensen. Computational methods in optimization considering uncertainties - an overview. *Computational Methods in Applied Mechanics and Engineering*, 198:2–13, 2008.

[124] V. Schulz and C. Schillings. Optimal aerodynamic design under shape uncertainties. Tech Report 09-5, University of Trier, 2009.

[125] V. H. Schulz. Ein effizientes Kollokationsverfahren zur numerischen Behandlung von Mehrpunktrandwertaufgaben in der Parameteridentifizierung und Optimalen Steuerung. Diploma thesis, University of Augsburg, 1990.

[126] I. H. Sloan. Polynomial interpolation and hyperinterpolation over general regions. *Journal of Approximation Theory*, 83:238–254, 1995.

[127] D. E. Smith, editor. *A Source Book in Mathematics*. Dover Publications, 1959.

[128] S. Smolyak. Quadrature and interpolation formulas for tensor products of certain classes of functions. *Dokl. Akad. Nauk SSSR*, 4:240–243, 1963.

[129] A. Sommariva, M. Vianello, and R. Zanovello. Nontensorial Clenshaw-Curtis cubature. *Numerical Algorithms*, 49:409–427, 2008.

[130] S. M. Stigler. Simon Newcomb, Percy Daniell, and the history of robust estimation 1885–1920. *Journal of the American Statistical Association*, 68:872–879, 1973.

[131] J. Stoer. *Numerische Mathematik 1*. Springer, 1999.

[132] E. J. Stone. On the rejection of discordant observations. *Monthly Notices of the Royal Astronomical Society*, 34:9–15, 1873.

[133] J. O. Street, R. J. Carroll, and D. Ruppert. A note on computing robust regression estimates via iteratively reweighted least squares. *The American Statistician*, 42(2):152–154, 1988.

[134] A. H. Stroud. Remarks on the disposition of points in numerical integration formulas. *Mathematical Tables and Other Aids to Computation*, 11(60):257–261, 1957.

[135] L. N. Trefethen. Is Gauss quadrature better than Clenshaw-Curtis? *SIAM Review*, 50(1):67–87, 2008.

[136] S. Waldherr. *Uncertainty and robustness analysis of biochemical reaction networks via convex optimisation and robust control theory*. PhD thesis, University of Stuttgart, 2009.

[137] J. Waldvogel. Fast construction of the Fejér and Clenshaw-Curtis quadrature rules. *BIT Numerical Mathematics*, 46:195–202, 2006.

[138] G. A. Watson. Choice of norms for data fitting and function approximation. *Acta Numerica*, 7:337–377, 1998.

[139] H. G. Wilson. Least-squares versus minimum absolute deviations estimation in linear models. *Decision Sciences*, 9:322–335, 1978.

[140] L. Wirsching. An SQP algorithm with inexact derivatives for a direct multiple shooting method for optimal control problems. Diploma thesis, University of Heidelberg, 2006.

[141] D. Xiu and J. S. Hesthaven. High-order collocation methods for differential equations with random inputs. *SIAM Journal of Scientific Computing*, 27(3):1118–1139, 2005.

[142] H. Xu, C. Caramanis, and S. Manner. Robust regression and LASSO. *IEEE Transactions on Information Theory*, 56(7):3561–3574, 2010.

Curriculum Vitae

Diese Seite enthält persönliche Daten und ist daher nicht Bestandteil der Online-Veröffentlichung.

This page contains personal data und is thus excluded from the online version of this thesis.

Statutory declaration / Eidesstattliche Erklärung

I hereby affirm that the thesis at hand is my own written work and that I have used no other sources and aids other than those indicated. All passages, which are quoted from publications or paraphrased from these sources, are indicated as such, i.e. cited, attributed. This declaration extends to drawings, sketches, illustrations and the like in this thesis.

This thesis was not submitted to another examination board and only partially published in academic releases.

Ich versichere an Eides statt durch meine eigene Unterschrift, dass ich die vorstehende Arbeit selbstständig und ohne fremde Hilfe angefertigt und alle Stellen, die wörtlich oder annähernd wörtlich aus Veröffentlichungen genommen sind, als solche kenntlich gemacht habe. Die Versicherung bezieht sich auch auf in der Arbeit gelieferte Zeichnungen, Skizzen, bildliche Darstellungen und dergleichen.

Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und nur teilweise in akademischen Veröffentlichungen publiziert.

Leiferde, October 29, 2012