RICE UNIVERSITY

Trust-Region SQP Methods with Inexact Linear System Solves for Large-Scale Optimization

by

Denis Ridzal

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy

APPROVED, THESIS COMMITTEE:

Matthias Heinkenschloss, Chairman
Professor of Computational and Applied Mathematics

William W. Symes
Noah Harding Professor of Computational and Applied Mathematics

Timothy C. Warburton
Assistant Professor of Computational and Applied Mathematics

Matteo Pasquali
Associate Professor of Chemical and Biomolecular Engineering

HOUSTON, Texas
APRIL, 2006
The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.
Abstract

Trust-Region SQP Methods with Inexact Linear System Solves for Large-Scale Optimization

by

Denis Ridzal

This thesis extends the design and the global convergence analysis of a class of trust-region sequential quadratic programming (SQP) algorithms for smooth nonlinear optimization to allow for an efficient integration of inexact linear system solvers.

Each iteration within an SQP method requires the solution of several linear systems, whose system matrix/operator involves the linearized constraints. Most existing implementations of SQP algorithms use direct linear algebra methods to solve these systems. For many optimization problems in science and engineering this is infeasible, because the systems are too large or the matrices associated with the linearized constraints are not formed explicitly. Instead, iterative solvers, such as preconditioned Krylov subspace methods have to be applied for the approximate solution of the linear systems arising within the SQP algorithm. In this case, the optimization algorithm has to provide stopping tolerances for the iterative solver.

The existing literature on the treatment of inexact linear system solves in SQP algorithms is rather scarce. Most theoretical results either provide stopping tolerances for iterative solvers that cannot be easily implemented in practice, or are restricted to specific classes of optimization problems. This thesis provides concrete stopping criteria for the iterative solution of so-called augmented systems, which allows for
a wider applicability of the resulting SQP algorithm and a rigorous integration of available KKT preconditioners. A key contribution is the development of an inexact conjugate gradient algorithm for the solution of quadratic subproblems with linear constraints that are subject to arbitrary nonlinear perturbations that arise from the approximate computation of projections via Krylov subspace methods.

The resulting SQP algorithm dynamically adjusts stopping tolerances for iterative linear system solves based on its current progress toward a KKT point. The stopping tolerances can be easily implemented and efficiently computed, and are sufficient to guarantee first-order global convergence of the algorithm. The performance of the algorithm is examined on optimal control problems governed by Burgers and Navier–Stokes equations.
Acknowledgements

First, I would like to thank my advisor, Matthias Heinkenschloss, for his support and patience, guidance when I needed it, and infusions of enthusiasm during the rough stretches. I also wish to thank the rest of my committee for their interest in the topic and their appreciation for the difficulty of the issues that I attempted to tackle.

The beginning of this five-year journey was the doing of my undergraduate advisor, Ray Chin, who introduced me to the world of computational and applied mathematics, and recommended Rice University for my graduate studies. For this, I am forever indebted to him.

I would also like to thank Eddie, Fernando, Michael, John, and Dwayne for watching my back, and letting me return the favor. Many thanks to the fellow thirty-sixteeners, Hoang, Agata, Andreas, Patricia, and Kary, for providing an inspirational and productive working environment.

For friendship, love, and ultimate patience, I thank Heidi Thornquist.

Most importantly, I wish to thank my parents, Nada and Muharem Ridzal, for always believing in me. They taught me to be tough, work hard, and never give up, skills without which this work could have never been completed. I thank them for their constant support and their love.

This research was supported by NSF grant ACI-0121360.
Contents

Abstract

Acknowledgements

List of Figures

List of Tables

1 Introduction
  1.1 Existing Work ................................................. 2
    1.1.1 Available SQP Implementations .......................... 2
    1.1.2 The Issue of Inexactness in Optimization Algorithms ... 4
    1.1.3 Shortcomings of the Current Approaches to Inexactness Control
    1.2 Organization of the Thesis ................................... 7

2 The Composite–Step Trust–Region SQP Method ................ 8
  2.1 Analysis of the Trust-Region SQP Algorithm ................. 8
    2.1.1 Examples of Null-Space Representations $W_k$ ............. 11
    2.1.2 Lagrange Multipliers ........................................ 13
    2.1.3 Hessian Approximations ........................................ 14
    2.1.4 Solution of the Quadratic Subproblem ...................... 14
    2.1.5 Acceptance of the Trial Step and the Update of the Trust-Region Radius .............................. 24
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3  Inexactness in a Generalized Trust–Region SQP Algorithm</td>
<td>28</td>
</tr>
<tr>
<td>3.1 Inexactness in the Quasi–Normal Step</td>
<td>30</td>
</tr>
<tr>
<td>3.2 Inexactness in the Tangential Step</td>
<td>31</td>
</tr>
<tr>
<td>3.3 Balancing Progress in the Computation of the Tangential and the</td>
<td>33</td>
</tr>
<tr>
<td>Quasi–Normal Step</td>
<td></td>
</tr>
<tr>
<td>3.4 Statement of the Generalized Trust–Region SQP Algorithm with</td>
<td>36</td>
</tr>
<tr>
<td>Inexactness Control</td>
<td></td>
</tr>
<tr>
<td>3.5 Global Convergence of Alg. 3.4.2</td>
<td>37</td>
</tr>
<tr>
<td>3.6 A Generalized Conjugate Gradient Method for the Solution of the</td>
<td>41</td>
</tr>
<tr>
<td>Tangential Subproblem</td>
<td></td>
</tr>
<tr>
<td>4  Inexact Linear System Solves in an SQP Algorithm with Null–Space</td>
<td>56</td>
</tr>
<tr>
<td>Projections</td>
<td></td>
</tr>
<tr>
<td>4.1 Tangential Step</td>
<td>57</td>
</tr>
<tr>
<td>4.1.1 Satisfaction of General Convergence Requirements for Inexactly</td>
<td>59</td>
</tr>
<tr>
<td>Computed Tangential Steps</td>
<td></td>
</tr>
<tr>
<td>4.1.2 Practical Convergence Requirements for Inexactly Computed</td>
<td>68</td>
</tr>
<tr>
<td>Tangential Steps</td>
<td></td>
</tr>
<tr>
<td>4.2 Quasi–Normal Step</td>
<td>74</td>
</tr>
<tr>
<td>4.3 Lagrange Multipliers</td>
<td>77</td>
</tr>
<tr>
<td>4.4 Balancing the Tangential and the Quasi–Normal Step</td>
<td>78</td>
</tr>
<tr>
<td>4.5 Statement of the Trust–Region SQP Algorithm with Inexact Augmented</td>
<td>83</td>
</tr>
<tr>
<td>System Solves</td>
<td></td>
</tr>
<tr>
<td>5  Numerical Examples</td>
<td>87</td>
</tr>
<tr>
<td>5.1 Optimal Control of Burgers Equation in One Dimension</td>
<td>87</td>
</tr>
<tr>
<td>5.1.1 Problem Formulation</td>
<td>88</td>
</tr>
<tr>
<td>5.1.2 Problem Discretization</td>
<td>90</td>
</tr>
<tr>
<td>5.1.3 Numerical Results</td>
<td>94</td>
</tr>
</tbody>
</table>
5.2 Optimal Control of Navier–Stokes Equations in Two Dimensions  101
  5.2.1 Problem Formulation  101
  5.2.2 Numerical Results  105

6 Conclusion  113

Bibliography  115
List of Figures

2.1 The composite-step approach. ........................................... 15
3.1 Interaction between the quasi-normal and the tangential step. ........ 29
4.1 Balancing the quasi-normal and the effective tangential step. ....... 81
5.1 The matrix $N'(\bar{u})$. .................................................. 93
5.2 Absolute GMRES stopping tolerances – Burgers. ...................... 96
5.3 Relative GMRES stopping tolerances – Burgers. ...................... 97
5.4 Performance of preconditioned GMRES – Burgers. ................... 98
5.5 Geometry of the backward-facing step channel. ...................... 101
5.6 The finite element grid for the backstep channel. .................... 106
5.7 The velocity plot for the uncontrolled fluid flow. .................... 107
5.8 Uncontrolled flow: recirculation near the corner region. ............ 107
5.9 The velocity plot for the controlled fluid flow. ....................... 108
5.10 Absolute GMRES stopping tolerances – Navier–Stokes. ............. 108
5.11 Relative GMRES stopping tolerances – Navier–Stokes. .............. 109
5.12 Performance of preconditioned GMRES – Navier–Stokes. ........... 110
List of Tables

5.1 Inexactness control vs. fixed stopping tolerances – Burgers. . . . . . . . . . 97
5.2 SQP convergence history with inexactness control – Burgers. . . . . . . . . 99
5.3 SQP convergence history with exact solves – Burgers. . . . . . . . . . . . . . 100
5.4 Inexactness control vs. fixed stopping tolerances – Navier-Stokes. . . . . 109
5.5 SQP convergence history with inexactness control – Navier-Stokes. . . . 111
5.6 SQP convergence history with exact solves – Navier-Stokes. . . . . . . . 112
Chapter 1

Introduction

The sequential quadratic programming (SQP) methodology represents the state of the art for the solution of smooth nonconvex, nonlinear optimization problems. In combination with effective strategies for the treatment of inequality constraints, it is used in powerful optimization codes for the solution of a wide range of nonlinear optimization problems. For example, the popular optimization codes SNOP, KNITRO and LOQO utilize the SQP methodology.

Each iteration within an SQP algorithm requires the solution of several linear systems, whose system matrix/operator involves the linearized constraints. Most existing implementations of SQP algorithms, including those mentioned above, use direct linear algebra methods to solve these systems. Lately though, serious efforts have been made in applying the SQP methodology to large-scale optimization problems within parallel computing environments. In this scenario, iterative solvers, such as preconditioned Krylov methods have to be applied for the approximate solution of linear systems arising in the SQP algorithm.

Iterative linear system solvers are also required for problems whose structure does not allow for an explicit representation of linear operators, as is the case for many optimization problems in science and engineering. In addition, the use of specialized linear system solvers, such as multigrid or domain decomposition methods, is often required in optimization problems governed by partial differential equations (PDEs).
The use of iterative solvers results in several complications. First, global convergence theories for most concrete SQP schemes assume that linear systems can be solved exactly. Iterative solvers, however, are inherently inexact. Our task will be to examine how the global convergence properties of SQP algorithms are affected by inexactness in linear system solves and to determine generic global convergence conditions.

The second, more difficult task is to let the optimization algorithm dynamically determine stopping tolerances for iterative solvers. Here we insist on only using stopping tolerances that are easily implementable and that can be efficiently computed in practice. The computed tolerances must be fairly loose when the SQP iterate is far away from the solution, but sufficient to guarantee global convergence of the SQP algorithm. As SQP iterates approach the solution, the stopping tolerances for linear system solves must be lowered accordingly.

1.1 Existing Work

The purpose of this section to familiarize the reader with a representative subset of SQP algorithms and corresponding codes, the available work on inexact linear system solves in SQP algorithms (and the general issue of inexactness in optimization), and the shortcomings of the current approaches to inexactness control.

1.1.1 Available SQP Implementations

Strong convergence properties of SQP algorithms combined with their robustness make them successful methods for the solution of nonlinearly constrained optimization problems. Of course, efficient solution of the underlying linear systems plays an important role as well. Early SQP codes were based on dense linear algebra, while the codes of the 1990s rely on sparse direct solution techniques. On the other hand, some more recent implementations can utilize iterative linear solvers.
One of the earliest codes, based on dense linear algebra, is NPSOL, developed in the 1980s by Gill, et al. [40, 41]. To ensure convergence from remote starting points, line search is used, while inequality constraints are treated by an active-set approach. The applicability of NPSOL to large-scale applications is intrinsically limited by the use of dense linear algebra. Another SQP code utilizing dense linear algebra is NLPQL, developed in the early 1980's by Schittkowski [67].

In the mid-1990s Gill, et al. [42] developed SNOPT, one of the first commercially available SQP codes to utilize sparse direct linear algebra. Like NPSOL, SNOPT is a line search method with an active-set approach for the treatment of inequality constraints. An entirely different approach is implemented in KNITRO, developed by Byrd, et al. [19, 75, 20], where inequality constraints are handled via an interior-point method. Linear systems are solved using sparse LU factorizations. Yet another interior-point code is LOQO, developed by Vanderbei, et al. [68, 73]. At the lowest level, LOQO solves modified KKT systems using sparse Cholesky factorizations. Other SQP codes based on sparse direct linear algebra include SPRNLP, developed by Betts and Frank [9, 10] and used in the SOCS optimal control framework, and IPOPT, developed by Wächter, et al. [74].

This concludes our review of codes based on direct (dense or sparse) linear algebra. For an excellent overview of the history and practice of SQP methods from their first emergence in 1963 (Wilson [76]) to the developments of the early 1990s, we refer the reader to Boggs and Tolle [13]. An exhaustive theoretical study of modern algorithms is given by Gould and Toint [48] as well as Conn, Gould, and Toint [22, Ch.15], whereas a comparative performance evaluation of the latest implementations is provided by Benson, et al. [7]. Additional details about the above codes can be found on the NEOS server for optimization [23].

The reviewed methods have been effectively used in applications ranging from engineering to finance, and have greatly contributed to the popularization and success of SQP methods. On the other hand, their limitations are evident. They always tie
the underlying SQP algorithm to a specific linear algebra engine. This makes the use of problem–specific solvers impossible. All exclusively use direct linear algebra, which due to memory limitations, cannot be applied to large-scale problems involving, for example, PDE simulations of complex physical systems.

We now turn our attention to the existing work in the area of matrix–free SQP codes. MOOCHO, developed by Bartlett [6], is an object–oriented framework for the solution of large–scale nonlinear optimization problems using gradient–based methods. MOOCHO evolved from rSQP++, an implementation of a reduced–space SQP algorithm [3]. One of the most attractive features of the MOOCHO framework is that it supports fully abstract linear algebra, allowing complex problem simulations on parallel distributed–memory supercomputers. A disadvantage of the underlying optimization algorithm is that it does not take into account the issue of inexactness arising from iterative linear system solves. Linear systems are solved to a fixed stopping tolerance, which usually results in their oversolving.

One currently available SQP code that is built around a theoretical framework for the treatment of inexact linear system solves is TRICE, developed by Heinkenschloss, et al. [29, 55, 54]. TRICE is an interior–point, trust–region SQP code that uses a reduced Hessian approach for the formulation of quadratic subproblems. It is primarily geared toward the solution of optimal control problems, in which there is a clear separation between a state (basis) variable space and a control (design, non–basis) space. TRICE lacks a full integration of practical mechanisms for inexactness control in linear system solves. Consequently, no numerical results documenting their performance are available.

1.1.2 The Issue of Inexactness in Optimization Algorithms

There is a considerable amount of literature available on the subject of inexactness in the context of Newton methods for unconstrained optimization and nonlinear equations. A series of publications related to the treatment of local convergence of inexact
Newton methods for unconstrained optimization dates back to the early 1980s, see, for example, Dembo, et al. [24, 25] and Dennis and Walker [31]. The connection with inexact SQP methods is established shortly thereafter, by Dembo and Tulowitzki [26, 27] and Fontecilla [38], although only on the level of local convergence analysis. An early local convergence result for inexact Newton methods for nonlinear equations, due to Ypma, can be found in [77].

Global convergence behavior of inexact trust-region algorithms for unconstrained optimization is investigated by Carter in [21], while the first global convergence results in the context of nonlinear equations were obtained by Deuffhard [32], Eisenstat and Walker [33, 34], and Brown and Saad [17].

The first global convergence result for an SQP method is in the work of Jäger and Sachs [59], where they investigate a line-search reduced-space approach. They, however, do not give an implementable mechanism for the control of inexactness, due to the need to compute Lipschitz constants in order to establish certain bounds. In a similar fashion, Biros and Ghattas [11] establish convergence requirements for an inexact quasi-Newton reduced-space SQP method. Their approach requires tight bounds on derivative norms, which are hard to estimate in practice.

One known result that gives practical, easily enforceable accuracy requirements is due to Heinkenschloss and Vicente [55]. They focus on a composite-step trust-region SQP technique in the reduced space, and manage to effectively control inexactness in first-order derivative information as well as inexact linearized constraint equation solves. Their algorithm, albeit tied to the reduced-space approach, represents an excellent starting point for our discourse on inexactness in the full-space setting, in Chapter 3.

A recent study by Ulbrich [72] treats a generalized composite-step SQP method in the context of time-dependent PDE-constrained optimization. The SQP algorithm is based on a nonmonotone trust-region method without a penalty function, and offers a unique way of balancing feasibility and optimality. Some of the inexactness handling
mechanisms are similar to the inexactness framework of Heinkenschloss and Vicente [55]. Ulbrich's framework is suited for PDE-constrained optimization problems that rely on separate state and adjoint equation solves.

1.1.3 Shortcomings of the Current Approaches to Inexactness Control

As presented above, there are essentially only two theoretical studies of inexactness in SQP algorithms that are satisfactory from the point of view of implementability and computational efficiency of the devised stopping criteria for iterative linear system solves, [55, 72]. All others fall short of this goal, due to the need for estimates of Lipschitz constants and norms of derivative operators.

The approaches of Heinkenschloss and Vicente, and Ulbrich, however, have certain shortcomings. Both are tied to reduced-space SQP methods that rely closely on what is known as the basis–nonbasis decomposition of optimization variables. We consider a more general, full-space composite-step approach, in which this distinction is not necessary.

The second shortcoming of [55, 72] is in the lack of treatment of concrete algorithms for the computation of the trial step in composite-step trust-region methods. A trial step in these algorithms is composed as the sum of two sub-steps, whose computation requires the solution of several linear systems. One of these steps, known as the tangential step, is usually computed using a conjugate gradient method, and requires many linear system solves, often dominating the computational cost of the entire SQP algorithm. In [55, 72], the computation of the tangential step is not considered in sufficient detail. Our key contribution in the tangential step computation is the development of an inexact conjugate gradient algorithm that utilizes a series of approximate projections via Krylov subspace methods, whose accuracy is controlled by the SQP algorithm. The computation of approximate projections is related to the iterative solution of so called augmented systems, which allow for a wider appli-
cability of the resulting SQP algorithm and a rigorous integration of available KKT preconditioners.

1.2 Organization of the Thesis

Chapter 2 gives a brief overview of the composite-step trust-region SQP methodology and states the relevant global convergence conditions and results. The convergence requirements stated in this chapter will guide our analysis of inexactness, which is presented in Chapter 3. In this chapter, we identify the key components of the SQP algorithm that are affected by inexact linear system solves, and slightly generalize the framework for SQP methods with inexact linear system presented in [55] to handle general null-space representations. In Chapter 4 we present a concrete implementation of the SQP method using null-space projections and show how this implementation fits into the inexactness framework of Chapter 3. Chapter 5 contains numerical results, which demonstrate the efficiency of the devised mechanisms for inexactness control in linear system solves. Finally, Chapter 6 summarizes the highlights of our work and discusses potential extensions.
Chapter 2

The Composite–Step Trust–Region SQP Method

In this chapter we describe the basic structure of an SQP algorithm for equality–constrained optimization. We are particularly interested in algorithms that employ trust–region methods, due to the proven robustness of this globalization strategy. It should be noted that the theoretical foundations reviewed in this chapter serve merely as a stepping stone toward the inexactness–controlling schemes proposed in Chapter 3. Thus the primary purpose of this chapter is to familiarize the reader with our SQP approach, present elementary theoretical results, and introduce the core algorithmic modules.

The equality–constrained problem is discussed in Section 2.1. It is solved using a composite–step trust–region approach in the full variable space. Along with concrete algorithms, we will present the basic global convergence requirements and results.

2.1 Analysis of the Trust-Region SQP Algorithm

We consider the following generalized equality–constrained nonlinear programming problem (NLP):

\[ \begin{align*}
\text{min} & \quad f(x) \\
\text{s.t.} & \quad c(x) = 0,
\end{align*} \]  

(2.1a)  

(2.1b)
where \( f : \mathcal{X} \rightarrow \mathbb{R} \) and \( c : \mathcal{X} \rightarrow \mathcal{Y} \) for some Hilbert spaces \( \mathcal{X} \) and \( \mathcal{Y} \). To derive an SQP model for the problem (2.1), we proceed as follows.

First, we present the basic building blocks of the standard KKT theory, and introduce some notation. We assume that \( f \) and \( c \) are twice continuously differentiable with Lipschitz continuous second derivatives. It is necessary to define the Lagrangian functional \( \mathcal{L} : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R} \) for (2.1), given by the expression

\[
\mathcal{L}(x, \lambda) = f(x) + \langle \lambda, c(x) \rangle_{\mathcal{Y}}.
\]

If \( x_* \) satisfies the constraint equation (2.1b) and if \( c_x(x_*) \) is surjective, then \( x_* \) is called a regular point. If additionally \( x_* \) is a local solution of (2.1), then there exists a Lagrange multiplier \( \lambda_* \in \mathcal{Y} \) such that the first order necessary optimality conditions

\[
\nabla_x f(x_*) + c(x_*)^* \lambda_* = 0 \quad (2.2a)
\]
\[
c(x_*) = 0 \quad (2.2b)
\]

are satisfied. The system (2.2) is also known as the KKT system. Finally, if \( x_*, \lambda_* \) satisfy (2.2), and

\[
\nabla_{xx} \mathcal{L}(x_*, \lambda_*)[h, h] \geq \rho \|h\|^2 \quad \forall h \in \text{Null } (c_x(x_*)), \quad (2.3)
\]

for some \( \rho > 0 \), then \( x_* \) is a strict local minimum of (2.1). This statement represents the second order sufficient optimality conditions.

If Newton's method is used for the solution of the KKT system (2.2), then at the \( k \)-th iterate \((x_k, \lambda_k)\) the following system must be solved:

\[
\begin{pmatrix}
\nabla_{xx} \mathcal{L}(x_k, \lambda_k) & c_x(x_k)^* \\
(c_x(x_k) & 0
\end{pmatrix}
\begin{pmatrix}
s_k^x \\
s_k^\lambda
\end{pmatrix}
= - \begin{pmatrix}
\nabla_x f(x_k) + c_x(x_k)^* \lambda \\
c(x_k)
\end{pmatrix}. \quad (2.4)
\]

The foundation of the SQP methodology is the fact that the Newton system (2.4)
can be interpreted as the KKT system for the solution of the quadratic problem

\[
\begin{align*}
\min & \quad \frac{1}{2} \langle \nabla_x \mathcal{L}(x_k, \lambda_k) s_k^T, s_k^T \rangle_X + \langle \nabla_x \mathcal{L}(x_k, \lambda_k), s_k^T \rangle_X + \mathcal{L}(x_k, \lambda_k) \\
\text{s.t.} & \quad c_x(x_k) s_k^T + c(x_k) = 0,
\end{align*}
\]

(2.5a)

(2.5b)

assuming that there exists a \( \rho_k > 0 \) such that \( \nabla_{xx} \mathcal{L}(x_k, \lambda_k)[h, h] \geq \rho \|h\|^2 \), for all \( h \in \text{Null}(c_x(x_k)) \). In other words, the SQP approach amounts to repeatedly minimizing a quadratic approximation of the Lagrangian with respect to a linearization of the constraints. In order to achieve global convergence, we further constrain the step by a trust-region bound. At each iterate \( x_k \), this yields the problem

\[
\begin{align*}
\min & \quad \frac{1}{2} \langle H_k s_k^T, s_k^T \rangle_X + \langle \nabla_x \mathcal{L}(x_k, \lambda_k), s_k^T \rangle_X + \mathcal{L}(x_k, \lambda_k) \\
\text{s.t.} & \quad c_x(x_k) s_k^T + c(x_k) = 0 \quad (2.6b) \quad \|s_k^T\|_X \leq \Delta_k, \quad (2.6c)
\end{align*}
\]

where \( \Delta_k \) is a given trust-region radius. At this point, it is important to note two things. First, we have introduced a new operator \( H_k = H(x_k, \lambda_k) \) that in some sense "approximates" the exact Hessian of the Lagrangian \( \nabla_{xx} \mathcal{L}(x_k, \lambda_k) \). Section 2.1.3 discusses this modification briefly. Second, one should be aware of the fact that the constraints (2.6b)-(2.6c) in general might not be consistent, and that an algorithm for the solution of (2.6) must take this into account. Once the step \( s_k^T \) has been computed, if it gives sufficient decrease for a chosen reduction measure (this remains to be discussed in Section 2.1.5), it is accepted, otherwise, it is rejected, and the trust region radius is reduced.

While this covers the basics of trust-region SQP, we have not discussed the underlying computational algorithms yet. These can, of course, vary from one implementation to another. In general, we have to focus on four main components: computation of Lagrange multipliers, computation of Hessian approximations, solu-
tion of the trust-region QP problem (2.6), and acceptance of the trial step $s_k^x$. The theoretical requirements on each of these modules and the corresponding statement of convergence for our SQP scheme are based on the following general assumptions, as stated in [28] and modified to fit the Hilbert space setting:

(A1) For all iterations $k$, we assume that $x_k \in \Omega$ and $x_k + s_k^x \in \Omega$, where $\Omega$ is an open subset of $\mathcal{X}$.

(A2) The function(al)s $f$ and $c$ are twice continuously Fréchet differentiable with Lipschitz continuous second derivatives.

(A3) The operator $c_x(x)$ is surjective for all $x \in \Omega$.

(A4) The following function(al)s and operators are uniformly bounded over all $x \in \Omega$: $f(x), \nabla_x f(x), \nabla_{xx} f(x), c(x), c_x(x), (c_x(x)c_x(x)^*)^{-1},$ and $c_{xx}(x)$.

(A5) For every iterate $x_k$, there exists a bounded linear operator $W_k : \mathcal{Z} \to \mathcal{X}$, where $\mathcal{Z}$ is a Hilbert space, such that

$$\text{Range}(W_k) = \text{Null}(c_x(x_k)).$$

Moreover, the sequence of operators $W_k, k \in \mathbb{N}$, is bounded.

(A6) The sequence of operators $H_k, k \in \mathbb{N}$, is bounded.

(A7) The sequence of Lagrange multipliers $\lambda_k, k \in \mathbb{N}$, is bounded.

2.1.1 Examples of Null–Space Representations $W_k$

It is very important to note that the null–space operator $W_k$ introduced in the general assumption (A5) is not a mere theoretical construct. It represents a generic tool for handling the linear constraint $c_x(x_k)s_k^x + c(x_k) = 0$, given in (2.6b). Several concrete approaches can be found in the literature on SQP methods. We discuss them next.
The QR Approach

If $\mathcal{X} = \mathbb{R}^n$ and $\mathcal{Y} = \mathbb{R}^m$, $m \leq n$, then one can use the QR decomposition of the linearized constraints to compute $W_k$, see, e.g., [43]. Assume that $c_x(x_k) \in \mathbb{R}^{n \times n}$, $c_x(x_k)$ has full row rank $m$. We have

$$c_x(x_k)^* = Q_k R_k = [Q_k^1 \quad Q_k^2] \begin{bmatrix} R_k^1 \\ 0 \end{bmatrix},$$

(2.7)

where $Q_k \in \mathbb{R}^{n \times n}$ is orthogonal, $Q_k^1 \in \mathbb{R}^{n \times m}$ represents the first $m$ columns of $Q_k$, $Q_k^2 \in \mathbb{R}^{n \times (n-m)}$ represents the remaining $n - m$ columns of $Q_k$, and $R_k^1 \in \mathbb{R}^{m \times m}$ is upper triangular. In this case the the domain of the null-space representor $W_k$ is $\mathcal{Z} = \mathbb{R}^{n-m}$. A simple argument shows that the columns of the matrix

$$W_k = Q_k^2$$

form a basis for the null space of $c_x(x_k)$.

The Basis–Nonbasis Approach

A different null-space representation is often used in, e.g. PDE-constrained optimization problems, in which there is a clear separation between basis (state) and nonbasis (control, design) variables. The optimization problem can be formulated as

$$\min \quad f(u, g)$$

(2.8a)

subject to $c(u, g) = 0,$

(2.8b)

where $f : \mathcal{U} \times \mathcal{G} \to \mathbb{R}$ and $c : \mathcal{U} \times \mathcal{G} \to \mathcal{U}$ for some Hilbert spaces $\mathcal{U}$ and $\mathcal{G}$. In (2.8) $u$ represent the states and $g$ represent the control/design variables. We assume that $c_{u}(u, g)$ is invertible. In this case the variable space is $\mathcal{X} = \mathcal{U} \times \mathcal{G}$, the constraint space is $\mathcal{Y} = \mathcal{U}$, and the domain of $W_k$ is $\mathcal{Z} = \mathcal{G}$. The operator representing the null
space of the linearized constraints is given by $W_k : \mathcal{G} \rightarrow \mathcal{U} \times \mathcal{G}$, defined by

$$W_k = \begin{bmatrix}
-c_u(u_k, g_k)^{-1}c_g(u_k, g_k) \\
I
\end{bmatrix},$$

where $I : \mathcal{G} \rightarrow \mathcal{G}$ is the identity operator. For SQP algorithms based on this approach, see [29, 52, 55].

The Augmented System Approach

A third alternative, related to the approach used by Byrd, et al., in the KNITRO optimization package [19], is a full–space formulation based on the original problem formulation (2.1). Here $Z = \mathcal{X}$ and $W_k$ is the projection onto the null space of $c_x(x_k)$. The application $z = W_k r$ of $W_k : \mathcal{X} \rightarrow \mathcal{X}$ with $z, r \in \mathcal{X}$, is computed by solving the augmented system

$$\begin{pmatrix}
I & c_x(x_k)^* \\
c_x(x_k) & 0
\end{pmatrix}
\begin{pmatrix}
z \\
y
\end{pmatrix}
= \begin{pmatrix}
r \\
0
\end{pmatrix}, \quad (2.9)$$

where $I : \mathcal{X} \rightarrow \mathcal{X}$ is the identity operator. We should note that the above $W_k$ is an orthogonal projector, i.e. $W_k = W_k^* = W_k W_k$. This approach is attractive if efficient iterative solvers for the augmented system (2.9) are available [4, 8, 14, 15, 53].

2.1.2 Lagrange Multipliers

The computation of a Lagrange multiplier estimate $\lambda_k$ at each iterate $x_k$ is related to the least-squares problem

$$\min_{\lambda} \| \nabla_x f(x_k) + c_x(x_k)^* \lambda \|_{\mathcal{X}}. \quad (2.10)$$

The computation of $\lambda_k$ is typically related to the choice of the null–space representation $W_k$. 
If the null-space representation $W_k$ is computed using the QR approach, see Section 2.1.1, then the solution of (2.10) can be computed as

$$
\lambda_k = R_k^{-1}Q_k^T \nabla_x f(x_k).
$$

(2.11)

If instead we use the augmented system approach, then the solution of (2.10) can be computed by solving the linear system

$$
\begin{pmatrix}
I & c_x(x_k) \\
c_x(x_k) & 0
\end{pmatrix}
\begin{pmatrix}
z \\
\lambda_k
\end{pmatrix}
= 
\begin{pmatrix}
-\nabla_x f(x_k) \\
0
\end{pmatrix}.
$$

(2.12)

Alternatively, if the basis–nonbasis approach is used, see [52], then the Lagrange multipliers $\lambda_k$ can be estimated as

$$
\lambda_k = -c_u(u_k, g_k)^{-1} \nabla_u f(u_k, g_k).
$$

(2.13)

In this case $\lambda_k$ does not solve (2.10), but is still an excellent estimate.

### 2.1.3 Hessian Approximations

In addition to the exact Hessian, quasi-Newton approximations can be used to generate $\{H_k\}$. For a detailed discussion of the choices in the SQP context, see [22, p.627–630]. Also see [12].

### 2.1.4 Solution of the Quadratic Subproblem

In this section we describe a composite-step approach for the solution of the quadratic subproblem (2.6). The composite step $s_k$ is the combination of a quasi-normal step $n_k$ that aims to (approximately) satisfy the linear constraints, and a tangential step $t_k$ that lies in the tangent space of the constraints and that tries to achieve optimality. A simple illustration of the quasi-normal and tangential steps is given in Figure
2.1. We adopt a method similar to that of Byrd and Omojokun [62]. The core ideas of this method have been successfully used and refined by Byrd, Gilbert, and Nocedal [18], Byrd, Hribar, and Nocedal [19], Dennis, El-Alem, and Maciel [28], Dennis, Heinkenschloss, and Vicente [29, 30], and many others.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{figure2_1.png}
\caption{The composite-step approach. For illustrative purposes, it is assumed that \( n_k \) is actually normal to the null space of the linearized constraints.}
\end{figure}

2.1.4.1 Cauchy Decrease Condition

Before we discuss the details of the computation of the quasi-normal and tangential steps, we must turn our attention to a concept crucial to the convergence theory of trust-region methods in general. For this purpose, we consider the unconstrained problem

\[
\min \quad f(x)
\]  \hspace{1cm} (2.14)
where \( x \in \mathcal{X} \), and \( f : \mathcal{X} \rightarrow \mathbb{R} \) is a continuously differentiable functional. A trust-region algorithm computes for every iterate \( x_k \) and a given trust-region radius \( \Delta \) a trial step \( s_k \in \mathcal{X} \) as an approximate solution of the problem

\[
\begin{align*}
\min \quad m_f(x_k, s) & \equiv f(x_k) + \langle \nabla_x f(x_k), s \rangle_{\mathcal{X}} + \frac{1}{2} \langle B(x_k)s, s \rangle_{\mathcal{X}} \\
\text{s.t.} \quad & \|s\| \leq \Delta,
\end{align*}
\]

where \( B(x_k) \) is (or approximates) the Hessian operator \( \nabla_x f(x_k) \).

The step \( s_k \) is required to satisfy what is referred to as a fraction of Cauchy decrease condition to ensure global convergence. What this means is that, for some constant \( \tau > 0 \) independent of \( k \), we require that

\[
m_f(x_k, 0) - m_f(x_k, s_k) \geq \tau (m_f(x_k, 0) - m_f(x_k, s^\triangledown)), \tag{2.15}
\]

where the Cauchy step \( s^\triangledown \) is defined as \( s^\triangledown = -\gamma^\triangledown \nabla f(x_k) \) with

\[
\gamma^\triangledown = \begin{cases} 
\frac{\|\nabla f(x_k)\|_{\mathcal{X}}^2}{(B(x_k)\nabla f(x_k), \nabla f(x_k))_{\mathcal{X}}} & \text{if } \frac{\|\nabla f(x_k)\|_{\mathcal{X}}^2}{(B(x_k)\nabla f(x_k), \nabla f(x_k))_{\mathcal{X}}} \leq \Delta \text{ and } (B(x_k)\nabla f(x_k), \nabla f(x_k))_{\mathcal{X}} > 0, \\
\frac{\Delta}{\|\nabla f(x_k)\|_{\mathcal{X}}} & \text{otherwise.}
\end{cases}
\]

In the latter sections, we will make frequent use of the following lemma, due to Powell [64]. It gives a weaker, yet more compact statement that is used in proofs of convergence of trust-region algorithms instead of the condition (2.15).

**Lemma 2.1.1.** If a step \( s_k \) satisfies a fraction of Cauchy decrease condition, then

\[
m_f(x_k, 0) - m_f(x_k, s_k) \geq \frac{\tau}{2} \|\nabla f(x_k)\|_{\mathcal{X}} \min \left\{ \frac{\|\nabla f(x_k)\|_{\mathcal{X}}}{\|B(x_k)\|_{\mathcal{X}}}, \Delta \right\}, \tag{2.16}
\]
2.1.4.2 Quasi-Normal Step

The two requirements on the quasi-normal step are, first, that it lie well within the trust region, and second, satisfy the constraint (2.6b) in some sense. For example, we choose a fixed parameter $\zeta \in (0, 1)$ and ask that the quasi-normal step $n_k$ approximately solve a problem of the following type:

\[
\begin{align*}
\min & \quad \|c_x(x_k)n + c(x_k)\|_Y^2, \\
\text{s.t.} & \quad \|n\|_X \leq \zeta \Delta_k. 
\end{align*}
\]  

(2.17a)  

(2.17b)

Global convergence theory for composite-step SQP algorithms does not require that the quasi-normal step exactly satisfy (2.17). In fact, $n$ need not even be normal to the tangent space (thus the designation “quasi-normal”). Instead, we usually ask that it satisfy a fraction of Cauchy decrease condition on the quadratic model $m(x_k, n)$ of the linearized constraints,

\[ m(x_k, n) = \|c_x(x_k)n + c(x_k)\|_Y^2. \]  

(2.18)

In particular, (2.15) applied to the model (2.18) suggests that we require from $n_k$ to give as much decrease as $\alpha^{cp} = \alpha^{cp}c_x(x_k)^*c(x_k)$, where the step length $\alpha^{cp}$ is given by

\[
\alpha^{cp} = \begin{cases} 
\frac{\|c_x(x_k)^*c(x_k)\|_Y^2}{\|c_x(x_k)c_x(x_k)^*c(x_k)\|_Y^2} & \text{if } \frac{\|c_x(x_k)c_x(x_k)^*c(x_k)\|_Y^2}{\|c_x(x_k)c_x(x_k)^*c(x_k)\|_X} \leq \zeta \Delta_k \\
\frac{\zeta \Delta_k}{\|c_x(x_k)^*c(x_k)\|_Y} & \text{otherwise,}
\end{cases}
\]

(2.19)

and

\[ m(x_k, 0) - m(x_k, n_k) \geq \sigma_1 \left( \|c(x_k)\|_Y^2 - \|c_x(x_k)n^{cp}_k + c(x_k)\|_Y^2 \right), \]

(2.20)

for some $0 < \sigma_1 \leq 1$ independent of $k$. Additionally, due to possible nonnormality of $n_k$ with respect to the tangent space of the constraints, the following condition must
hold for $\kappa_1 > 0$ independent of $k$

$$\|n_k\|_X \leq \kappa_1 \|c(x_k)\|_Y.$$  \hfill (2.21)

Requirement (2.20) implies a more compact condition used by Dennis, El-Alem, and Maciel in [28] and Dennis, El-Alem, and Maciel in [28], stated in terms of the following lemma.

**Lemma 2.1.2.** If the quasi-normal step $n_k$ satisfies a fraction of Cauchy decrease condition associated with the trust-region problem (2.17), then

$$m(x_k, 0) - m(x_k, n_k) \geq \kappa_2 \|c(x_k)\|_Y \min \left\{\kappa_3 \|c(x_k)\|_Y, \Delta_k\right\},$$  \hfill (2.22)

for some positive constants $\kappa_2$ and $\kappa_3$ independent of $k$.

**Proof.** The proof is a direct consequence of the application of Lemma 2.1.1 to the model (2.18), followed by the use of the boundedness property of $c_x(x)$ in assumption (A4), the norm inequality

$$\|c(x_k)\|_Y = \|(c_x(x_k)c_x(x_k)^*)^{-1}(c_x(x_k)c_x(x_k)^*)c(x_k)\|_Y$$

$$\leq \|(c_x(x_k)c_x(x_k)^*)^{-1}\|_Y \|c_x(x_k)\|_X \|c_x(x_k)^*c(x_k)\|_Y,$$

and the use of the boundedness property of $(c_x(x_k)c_x(x_k)^*)^{-1}$ in assumption (A4). \hfill \square

Requirement (2.22) replaces the fraction of Cauchy decrease condition in convergence proofs for trust-region SQP methods.

A concrete procedure for the computation of the quasi-normal step is given in Algorithm 2.1.3. We use the popular *dogleg* method, proposed by Powell, [63]. It is based on the computation of the Cauchy point $n_k^CP$ and the so called *Newton step* $n_k^N$ (more generally referred to as the *feasibility step*). The computation of $n_k^N$ is typically tied to the choice of the null-space representation $W_k$. 
If the null-space representor $W_k$ from Section 2.1.1 is computed using the QR approach or the augmented system approach, then the feasibility step can be computed as the minimum norm minimizer of (2.18):

$$n_k^N = -c_x(x_k)^*(c_x(x_k)c_x(x_k)^*)^{-1}c(x_k).$$  \hspace{1cm} (2.23)

If the null-space representor $W_k$ from Section 2.1.1 is computed using the QR approach, then the QR factorization can be reused to compute the feasibility step. A simple argument shows that expression (2.23) simplifies to

$$n_k^N = -Q_k^1 R_k^{-T}c(x_k).$$  \hspace{1cm} (2.24)

If the augmented system approach is used instead, the feasibility step (2.23) can be computed by finding the solution $y \in \mathcal{Y}$, $z \in \mathcal{X}$ of the linear system

$$
\begin{pmatrix}
I & c_x(x_k)^* \\
c_x(x_k) & 0 \\
\end{pmatrix}
\begin{pmatrix}
z \\
y \\
\end{pmatrix}
=
\begin{pmatrix}
0 \\
c(x_k) \\
\end{pmatrix}
$$

and setting

$$n_k^N = c_x(x_k)^*y.$$  \hspace{1cm} (2.25)

If the basis–nonbasis approach is used, then

$$n_k^N = \begin{pmatrix}
-c_u(u_k, g_k)^{-1}c(u_k, g_k) \\
0 \\
\end{pmatrix},$$  \hspace{1cm} (2.26)

is the adequate representation of the feasibility step. In this case $n_k^N$ is not perpendicular to the null-space of $c_x(x_k)$.

The precise statement of the dogleg algorithm follows.

**Algorithm 2.1.3. (Dogleg Method for the Quasi-Normal Subproblem.)**

1. Compute $n_{k,full}^p = \frac{\|c_x(x_k)^*c(x_k)\|_2^2}{\|c_x(x_k)\|_2^2} c_x(x_k)^*c(x_k)$. 

2. If \( \|n_{k,full}^{p}\|_{\mathcal{X}} > \zeta \Delta_k \), then

Compute \( \theta_k^{p} \in (0, 1) \) such that \( \|\theta_k^{p} n_{k,full}^{p}\|_{\mathcal{X}} = \zeta \Delta_k \).

Set \( n_k = \theta_k^{p} n_{k,full}^{p} \).

Else

(a) Compute \( n_k^{N} \) via (2.24), (2.25) and (2.26), or (2.27).

(b) If \( \|n_k^{N}\|_{\mathcal{X}} > \zeta \Delta_k \), then

Compute \( \theta_k^{N} \in (0, 1) \) such that \( \|n_{k,full}^{p} + \theta_k^{N} (n_k^{N} - n_{k,full}^{p})\|_{\mathcal{X}} = \zeta \Delta_k \).

Set \( n_k = n_{k,full}^{p} + \theta_k^{N} (n_k^{N} - n_{k,full}^{p}) \).

Else

Set \( n_k = n_k^{N} \).

\( \square \)

2.1.4.3 Tangential Step

The standard requirements on the tangential step are that it lie in the tangent space of the constraints and that it move the quadratic model of the Lagrangian toward optimality. Thus we usually ask that the tangential step \( t_k \) approximately solve a problem of the type

\[
\begin{align*}
\min & \quad \frac{1}{2} \langle H_k(t + n_k), t + n_k \rangle_{\mathcal{X}} + \langle \nabla_x \mathcal{L}(x_k, \lambda_k), t + n_k \rangle_{\mathcal{X}} + \mathcal{L}(x_k, \lambda_k) \\
\text{s.t.} & \quad c_x(x_k)t = 0 \tag{2.28b} \\
& \quad \|t + n_k\|_{\mathcal{X}} \leq \Delta_k. \tag{2.28c}
\end{align*}
\]

Again, theoretical convergence results do not require an accurate solution of (2.28).

To state an analog of the condition (2.1.2) for the tangential step, we first define the quadratic model of the Lagrangian, given by

\[
q_k^\mathcal{X}(s) \equiv \mathcal{L}(x_k, \lambda_k) + \langle \nabla_x \mathcal{L}(x_k, \lambda_k), s \rangle_{\mathcal{X}} + \frac{1}{2} \langle H_k s, s \rangle_{\mathcal{X}}. \tag{2.29}
\]

Due to assumption (A5), for every \( t \in \mathcal{X}, c_x(x_k)t = 0 \) if and only if there exists a \( w \in \mathcal{Z} \) with the property \( W_k w = t \). Thus, problem (2.28) can be reformulated as the
one of finding the solution \( w_k \) of

\[
\begin{align*}
\min & \quad g_k^X(w) \equiv g_k^X(n_k) + \langle W_k^* (H_k n_k + \nabla_x \mathcal{L}(x_k, \lambda_k), w) \rangle_Z \\
& \quad + \frac{1}{2} \langle W_k^* H_k W_k w, w \rangle_Z \\
\text{s.t.} & \quad \|n_k + W_k w\|_X \leq \Delta_k,
\end{align*}
\]

(2.30a)  

(2.30b)

and setting \( t_k = W_k w_k \).

Note that since \( \|n_k\|_X \leq \zeta \Delta_k \), any \( w \) with \( \|W_k w\|_X \leq \Delta_k - \|n_k\|_X \) satisfies the trust region constraint in (2.30). The trust region constraints in (2.28), (2.30) can be replaced by \( \|t\|_X \leq \Delta_k - \|n_k\|_X \) and \( \|W_k w\|_X \leq \Delta_k - \|n_k\|_X \), respectively.

As usual, it is not necessary that \( w_k \) solves (2.30) exactly, but we only require that it satisfy a fraction of the Cauchy decrease condition on the model (2.30a) with the corresponding trust-region constraint (2.30b). We will skip the details of the derivation of the resulting compact condition, as it closely follows the procedure used in Section 2.1.4.2, and simply state the result.

**Lemma 2.1.4.** If the tangential step \( t_k \) satisfies a fraction of Cauchy decrease condition associated with the trust-region problem (2.30), then

\[
q_k^X(n_k) - q_k^X(n_k + t_k) \\
\geq \kappa_4 \|W_k^* \nabla q_k^X(n_k)\|_Z \min \left\{ \kappa_5 \|W_k^* \nabla q_k^X(n_k)\|_Z, \kappa_6 \Delta_k \right\},
\]

(2.31)

for some positive constants \( \kappa_4, \kappa_5, \) and \( \kappa_6 \) independent of \( k \).

Requirement (2.31) replaces the fraction of Cauchy decrease condition in convergence proofs for trust-region SQP methods.

We now present concrete algorithms for the solution of the tangential step subproblem (2.28). The first is based on the reduced-space reformulation (2.30). It is a conjugate gradient (CG) algorithm that utilizes Steihaug–Toint [71] stopping criteria to account for the possibility of negative curvature in the reduced Hessian \( W_k^* H_k W_k \)
and violations of the trust-region constraint. Unlike the original Steihaug–Toint scheme, for general null-space representation $W_k$, this particular version does not guarantee monotone increase of the norms of the iterates. Therefore, even though the trust region constraint $\|W_k w_{i+1}\|_\chi \leq \Delta_k - \|n_k\|_\chi$ might be violated at some iterate $w_{i+1}$, if the algorithm is not terminated, a subsequent iterate could again satisfy this condition. This means that we could be terminating Alg. 2.1.5 earlier than necessary. 

From the point of view of the global convergence theory, this is acceptable.

**Algorithm 2.1.5.** (*Steihaug–Toint CG method for the solution of (2.30)*)

0. Let $w_0 = 0 \in Z$. Let $g_k = \nabla \mathcal{L}(x_k, \lambda_k) + H_k n_k$, $r_0 = W_k^* g_k$, $p_0 = r_0$.

1. For $i = 0, 1, 2, ...$
   
   (a) If $\langle p_i, W_k^* H_k W_k p_i \rangle_Z \leq 0$, compute $\theta > 0$ such that $\|W_k (w_i + \theta p_i)\| = \Delta_k - \|n_k\|_\chi$ and return $w_{i+1} = w_i + \theta p_i$.

   (b) $\alpha_i = \langle r_i, r_i \rangle_Z / \langle p_i, W_k^* H_k W_k p_i \rangle_Z$

   (c) $w_{i+1} = w_i + \alpha_i p_i$

   (d) If $\|W_k w_{i+1}\|_\chi \geq \Delta_k - \|n_k\|_\chi$, compute $\theta > 0$ such that $\|W_k (w_i + \theta p_i)\| = \Delta_k - \|n_k\|_\chi$ and return $w_{i+1} = w_i + \theta p_i$.

   (e) $r_{i+1} = r_i - \alpha_i W_k^* H_k W_k p_i$

   (f) $\beta_i = \langle r_{i+1}, r_{i+1} \rangle_Z / \langle r_i, r_i \rangle_Z$

   (g) $p_{i+1} = r_{i+1} + \beta_i p_i$

If $W_k$ is the orthogonal null-space projector, such as those given by augmented system solves, see Section 2.1.1, then previous algorithm can be rewritten following [47]. First, we note that if $n_k$ is normal to the null space of the linearized constraints, which is for example true for quasi-normal steps produced by Algorithm 2.1.3 with $n_k^N$ computed via (2.25) and (2.26), then $\|t + n_k\|_\chi^2 = \|t\|_\chi^2 + \|n_k\|_\chi^2$ for all $t$ in the null-space of $c(x_k)$. Consequently, in this case, the trust-region constraint $\|t + n_k\|_\chi \leq \Delta_k$ can be written as

$$\|t\|_\chi \leq \sqrt{\Delta_k^2 - \|n_k\|_\chi^2}.$$
Thus we can formulate the tangential subproblem as

\[
\begin{align*}
\min & \quad \frac{1}{2} \langle H_k(t + n_k), t + n_k \rangle \chi + \langle \nabla_x \mathcal{L}(x_k, \lambda_k), t + n_k \rangle \chi + \mathcal{L}(x_k, \lambda_k) \\
\text{s.t.} & \quad c_x(x_k) t = 0 \\
& \quad \|t\|_\chi \leq \sqrt{\Delta_k^2 - \|n_k\|_\chi^2}.
\end{align*}
\] (2.32a, 2.32b, 2.32c)

It is solved using a Steihaug–Toint variant of a projected CG method in the full space [47]. Because of the properties of \(W_k\), this algorithm retains the monotonicity property of the norms of the iterates. Further details are given below.

**Algorithm 2.1.6. (Full-Space Projected Steihaug CG Method.)**

0. Set stopping tolerance \(tol\). Set \(t_0^j = 0\) and \(r_0 = \nabla_x \mathcal{L}(x_k, \lambda_k) + H_k n_k\). Let \(\Delta_k = \sqrt{\Delta_k^2 - \|n_k\|_\chi^2}\).

1. Compute \(g_0 = W_k r_0\), where \(W_k\) is an orthogonal null space projector.
   Set \(p_0 = -g_0\).

2. For \(j = 0, 1, 2, \ldots\)
   
   (a) If \(\langle H_k p_j, p_j \rangle \chi \leq 0\), then set \(t_k = t_k^j + \theta p_j\) where \(\theta > 0\) is such that \(\|t_k\|_\chi = \Delta_k\), and stop.
   
   (b) Set \(\alpha = \langle r_j, g_j \rangle \chi / \langle H_k p_j, p_j \rangle \chi\).
   
   (c) Set \(t_k^{j+1} = t_k^j + \alpha p_j\).
   
   (d) If \(\|t_k^{j+1}\|_\chi > \Delta_k\), then set \(t_k = t_k^j + \theta p_j\) where \(\theta > 0\) is such that \(\|t_k\|_\chi = \Delta_k\), and stop.
   
   (e) Set \(r_{j+1} = r_j + \alpha H_k p_j\).
   
   (f) Compute \(g_{j+1} = W_k r_{j+1}\).
   
   (g) If \(\langle g_{j+1}, r_{j+1} \rangle \chi < tol\), then set \(t_k = t_k^j\) and stop.
   
   (h) Set \(\beta = \langle r_{j+1}, g_{j+1} \rangle \chi / \langle r_j, g_j \rangle \chi\).
   
   (i) Set \(p_{j+1} = -g_{j+1} + \beta p_j\).

\[\square\]

We recall that \(W_k = W_k^* = W_k W_k\). The consequence of this property is that Alg. 2.1.6 applied to the full-space problem (2.32) is equivalent to Alg. 2.1.5 applied to the corresponding problem in the reduced space.
2.1.5 Acceptance of the Trial Step and the Update of the Trust–Region Radius

Global convergence of trust–region SQP methods is often ensured through the use of a merit function, which helps us determine whether a step is acceptable and whether the trust–region radius needs to be modified. A crucial property of a merit function is that its critical points should also be the critical points of the underlying problem. Additionally, it is highly desirable that the critical points of the two problems completely coincide.

A variety of merit functions have been used in SQP algorithms (for detailed reviews see [22, Ch.15] and [60, Ch.18]) as their performance varies depending on the details of the particular SQP scheme, features of target applications, etc. We use the augmented Lagrangian merit function

\[ \phi(x, \lambda; \rho) = f(x) + \langle \lambda, c(x) \rangle_y + \rho\|c(x)\|_y^2 = \mathcal{L}(x, \lambda) + \rho\|c(x)\|_y^2. \]  

(2.33)

We use this merit function because the few theoretical results on the treatment of inexact linear system solves in SQP algorithms are based on this particular choice. Additionally, the convergence theory presented here is mostly based on [28], where the authors use this merit function.

Let \( s_k^x = n_k + t_k \) be a trial step with \( n_k \) satisfying (2.22) and (2.21), and \( t_k \) satisfying (2.31), and let \( \lambda_{k+1} = \lambda_k + \Delta \lambda_k \) be an updated Lagrange multiplier. To measure the improvement in the merit function \( \phi \), we compare the actual reduction and the predicted reduction in moving from the current iterate \( x_k \) to the trial iterate \( x_k + s_k^x \). The actual reduction is defined by

\[ \text{ared}(s_k^x; \rho_k) = \phi(x_k, \lambda_k; \rho_k) - \phi(x_k + s_k^x, \lambda_{k+1}; \rho_k), \]  

(2.34)
and the predicted reduction is given by

\[ \text{pred}(s_k^e; \rho_k) = \phi(x_k, \lambda_k; \rho_k) \]
\[ - \left( q_k^X (s_k^e) + \langle \Delta \lambda_k, c_z(x_k) s_k^e + c(x_k) \rangle_Y + \rho_k \| c_z(x_k) s_k^e + c(x_k) \|_Y^2 \right). \]

(2.35)

Two issues remain to be discussed. The first is the examination of the computed step together with an update strategy for the trust-region radius. It is directly addressed in Algorithm 2.1.8. The second is the update of the penalty parameter \( \rho_k \). This is a crucial issue for SQP methods that employ the augmented Lagrangian merit function. We use El-Alem's scheme given in [35], and adopted by Dennis, El-Alem, and Maciel in [28], and Heinkenschloss and Vicente in [55]. The scheme guarantees a decrease in the merit function proportional to a fraction of Cauchy decrease in the quadratic model (2.18) of the constraints. It is stated in Algorithm 2.1.7, which is essentially the same as the penalty update method in [28]. An important observation is that a penalty parameter \( \rho_k \) generated by this algorithm always satisfies

\[ \text{pred}(s_k^e; \rho_k) \geq \frac{\rho_k}{2} (\| c(x_k) \|_Y^2 - \| c_z(x_k) s_k^e + c(x_k) \|_Y^2). \]

(2.36)

**Algorithm 2.1.7. (Penalty Parameter Update.)**

1. Assumption: Let \( \rho_{-1} = 1 \) and let \( \beta \) be a small positive constant.

2. If

\[ \text{pred}(s_k^e; \rho_{k-1}) \geq \frac{\rho_{k-1}}{2} (\| c(x_k) \|_Y^2 - \| c_z(x_k) s_k^e + c(x_k) \|_Y^2) \]

then set \( \rho_k = \rho_{k-1} \). Otherwise set

\[ \rho_k = \frac{2 (q_k^X (s_k^e) - q_k^X (0) + \langle \Delta \lambda_k, c_z(x_k) s_k^e + c(x_k) \rangle_Y)}{\| c(x_k) \|_Y^2 - \| c_z(x_k) s_k^e + c(x_k) \|_Y^2} + \beta. \]


\( \Box \)

We now have the necessary pieces to formulate the version of our trust-region SQP algorithm based on exact linear system solves, see Algorithm 2.1.8.

**Algorithm 2.1.8. (Trust-region SQP algorithm.)**
1. Initialization. Choose initial point $x_0$, initial trust-region radius $\Delta_0$, and constants $0 < \eta_1 < \eta_2 < 1$, $0 < \gamma_1 < \gamma_2 < 1$, $\rho_0 \geq 1$, and $\epsilon_{tol} > 0$. Compute an initial Lagrange multiplier estimate $\lambda_0$.

2. For $k = 0, 1, 2, ...$

   (a) Convergence Check. If
   \[ \|W_k^* \nabla_x \mathcal{L}(x_k, \lambda_k)\|_Z + \|C(x_k)\|_Y < \epsilon_{tol}, \]
   then terminate.

   (b) Trial Step Computation.
   
   i. Find a quasi-normal step $n_k$ satisfying (2.22) and (2.21).
   
   ii. Find a tangential step $t_k$ satisfying (2.31).
   
   iii. Set $s_k = n_k + t_k$.
   
   iv. Compute a new Lagrange multiplier estimate $\lambda_{k+1}$.

   (c) Acceptance Test.
   
   i. Update penalty parameter $\rho_k$ via Algorithm 2.1.7.
   
   ii. Compute actual reduction $\text{ared}(s_k; \rho_k)$ and predicted reduction $\text{pred}(s_k^p; \rho_k)$, and their ratio $\theta_k = \frac{\text{ared}(s_k; \rho_k)}{\text{pred}(s_k^p; \rho_k)}$.
   
   iii. If $\theta_k \geq \eta_1$, set $x_{k+1} = x_k + s_k$, otherwise set $x_{k+1} = x_k$ and reset $\lambda_{k+1} = \lambda_k$.

   (d) Trust-Region Radius Update. Set
   \[ \Delta_{k+1} \in \begin{cases} [\Delta_k, \infty] \quad &\text{if } \theta_k \geq \eta_2, \\ [\gamma_2 \Delta_k, \Delta_k] \quad &\text{if } \theta_k \in [\eta_1, \eta_2), \\ [\gamma_1 \Delta_k, \gamma_2 \Delta_k] \quad &\text{if } \theta_k < \eta_1. \end{cases} \]

\[ \square \]

We also state a basic convergence result. For a detailed proof, see [28].

**Theorem 2.1.9.** Under the general problem assumptions (A1)-(A7), the sequence of iterates generated by Algorithm 2.1.8 satisfies

\[ \liminf_{k \to \infty} (\|W_k^* \nabla_x \mathcal{L}(x_k, \lambda_k)\|_Z + \|C(x_k)\|_Y) = 0. \]
Moreover, given any $\epsilon_{tol} > 0$, the algorithm terminates in finitely many iterations, due to

$$
\| W_k^* \nabla z \mathcal{L}(x_k, \lambda_k) \|_Z + \| C(x_k) \|_Y < \epsilon_{tol}.
$$

This thesis is primarily concerned with the global convergence to a point satisfying the first order necessary optimality conditions. Global convergence of algorithms of the type 2.1.8 to a point that satisfies the second order necessary optimality conditions is shown in [30].

Finally, we remind the reader that the trust-region approach is only one of the possible globalization schemes for SQP algorithms. Line search techniques have been studied and effectively applied in this context, see, for example, [68, 73]. Additionally, in the late 1990s the concept of filter methods emerged in the study of SQP techniques, see Fletcher, et al. [36, 37].
Chapter 3

Inexactness in a Generalized Trust–Region SQP Algorithm

As presented in Chapter 2, the computation of Lagrange multipliers, the quasi–normal step, and the tangential step each involve (potentially multiple) linear system solves. For large–scale problems utilizing augmented system or basis–nonbasis null space representations in the tangential step algorithm, with corresponding Lagrange multiplier and quasi–normal step computations (see Section 2.1.1), all linear systems are solved iteratively. In this case, one can usually only guarantee that their approximate solutions give linear system residuals that are below a specified tolerance. Therefore, in this scenario, the quasi-normal step $n_k$, the Lagrange multipliers $\lambda_k$, and the tangential step $t_k$ are never computed exactly. The main question here is how to ensure global convergence of the optimization algorithm by controlling the size of the linear system residuals in the computation of the Lagrange multipliers, the quasi–normal step, and the tangential step. Additionally, the interaction between the inexactness in the quasi–normal and tangential step must be closely examined. One important issue is that the inexactley computed tangential step can leave the null space of the linearized constraints, see Figure 3.1, which can affect the amount of linear feasibility gained in the quasi-normal step computation, as well as the criteria for the acceptance of the composite step (penalty parameter update, definition of predicted reduction).
Figure 3.1: Interaction between the quasi-normal (thin arrows) and the tangential step (bold arrows), with exact and inexact linear system solves.

Global convergence theory of trust-region methods for unconstrained optimization [21, 48] and trust-region SQP methods [28, 48] provides important guidelines for the treatment of the issue of inexactness in composite-step trust-region SQP algorithms. The work of Heinkenschloss and Vicente [55] builds on the generic trust-region SQP convergence theory and establishes a theoretical framework for the treatment of inexactness arising in iterative linear system solves. The mechanisms of inexactness control developed in [55] do not depend on Lipschitz constants and derivative norm bounds, as is the case with other attempts in this area (see [59, 11]), and are therefore computationally attractive. However, the authors do not fully analyze the practical implications of suggested theoretical mechanisms. Also, their work is limited to the reduced-space SQP approach, which assumes the basis–nonbasis partitioning of optimization variables.

The two goals of our work are to extend the theoretical inexactness framework of Heinkenschloss and Vicente to trust-region SQP methods for general equality–
constrained optimization problems, and to devise a practical inexact SQP framework, i.e. a concrete optimization algorithm that efficiently manages the stopping criteria for iterative linear system solves. This chapter establishes the theoretical basis for a generalized inexactness framework. Chapter 4 specializes this generic framework to the case in which the null-space representer $W_k$ is an orthogonal null-space projection.

Section 3.1 states the convergence conditions that must be satisfied by the quasi-normal step. Section 3.2 addresses the global convergence conditions for the tangential step. In Section 3.3, we address the issue of balancing progress in feasibility and optimality in presence of inexactness. A generalized SQP algorithm based on our theoretical framework for inexactness control is stated in Section 3.4. In Section 3.5, we prove that this algorithm is globally convergent. Section 3.6 focuses on the issue of inexact linear system solves in the tangential step computation. The results of this section are specialized to the case in which $W_k$ is a projector in Chapter 4.

3.1 Inexactness in the Quasi-Normal Step

For completeness, in this section we merely restate the global convergence requirements that must be imposed on the quasi-normal step, in the inexact setting. We require the boundedness condition

$$
\|n_k\|_X \leq \kappa_1 \|c(x_k)\|_Y, \tag{3.1}
$$

where $\kappa_1 > 0$ is independent of $k$, and the fraction of Cauchy decrease condition

$$
\|c(x_k)\|_Y^2 - \|c_x(x_k)n_k + c(x_k)\|_Y^2 \geq \kappa_2 \|c(x_k)\|_Y \min \{\kappa_3 \|c(x_k)\|_Y, \Delta_k\}, \tag{3.2}
$$

where $\kappa_2, \kappa_3 > 0$ are independent of $k$. 
3.2 Inexactness in the Tangential Step

The tangential subproblem can be formulated as

\[
\begin{align*}
\min & \quad \frac{1}{2} \langle H_k t, t \rangle_{\mathcal{X}} + \langle g_k, t \rangle_{\mathcal{X}} + q^X_k(n_k) \\
\text{s.t.} & \quad c_x(x_k)t = 0 \\
& \quad \|t\|_{\mathcal{X}} \leq \Delta_k - \|n_k\|_{\mathcal{X}},
\end{align*}
\]  

(3.3a) (3.3b) (3.3c)

where

\[ g_k = H_k n_k + \nabla_x \mathcal{L}(x_k, \lambda_k). \]

We now recall that we have previously introduced the (exact) null-space operator \( W_k : \mathcal{Z} \rightarrow \mathcal{X} \) that satisfies \( \text{Range}(W_k) = \text{Null}(c_x(x_k)) \). This operator is not a mere theoretical construct. Its purpose is to cover all existing practical implementations for handling the linear constraint \( c_x(x_k)t = 0 \), and at the same time generalize the analysis. We let \( t_k = W_k w_k \), and problem (3.3) becomes

\[
\begin{align*}
\min & \quad \frac{1}{2} \langle W_k^* H_k W_k w, w \rangle_{\mathcal{Z}} + \langle W_k^* g_k, w \rangle_{\mathcal{Z}} + q^X_k(n_k) \\
\text{s.t.} & \quad \|W_k w\|_{\mathcal{X}} \leq \Delta_k - \|n_k\|_{\mathcal{X}}.
\end{align*}
\]  

(3.4a) (3.4b)

We call (3.4) the reduced tangential subproblem.

With inexact linear system solves in the application of \( W_k \) and \( W_k^* \), the reduced gradient \( W_k^* g_k \) and the reduced Hessian \( W_k^* H_k W_k \) are no longer available. We hope to replace the exact null-space operator \( W_k \) with an inexact representation \( \widetilde{W}_k \), and introduce the inexact reduced gradient vector \( \widetilde{W}_k^* g_k \) and the inexact reduced Hessian operator \( \widetilde{W}_k^* H_k W_k \). This allows us to impose convergence criteria arising in the treatment of inexact trust-region methods for unconstrained optimization.

Under the assumption that we can introduce a fixed inexact reduced gradient and a fixed inexact reduced Hessian in the tangential subproblem (3.4), we obtain its
inexact counterpart

\[
\begin{align*}
\min \quad & \tilde{q}_k(w) \equiv \frac{1}{2} \left\langle \tilde{W}_k^* H_k \tilde{W}_k w, w \right\rangle_z + \left\langle \tilde{W}_k^* g_k, w \right\rangle_z + q_k^x(n_k) \\
\text{s.t.} \quad & \|\tilde{W}_k w\|_x \leq \Delta_k - \|n_k\|_x.
\end{align*}
\] (3.5a) (3.5b)

The theoretical work on the global convergence of trust-region algorithms using inexact gradient and Hessian information, in the context of unconstrained optimization [21, 22], allows us to pose convergence requirements on \(\tilde{W}_k g_k\) and \(\tilde{W}_k^* H_k \tilde{W}_k\). The inexact reduced gradient needs to satisfy

\[
\|\tilde{W}_k^* g_k - W_k^* g_k\|_Z \leq \xi_1 \min \left(\|\tilde{W}_k^* g_k\|_Z, \Delta_k \right),
\] (3.6)

for some \(\xi_1 > 0\) independent of \(k\). The inexact reduced Hessian must satisfy

\[
\left\langle \tilde{W}_k^* H_k \tilde{W}_k w_k, w_k \right\rangle_z \leq \xi_2 \|w_k\|_Z^2,
\] (3.7)

for some \(\xi_2 > 0\) independent of \(k\). Finally, we also need to impose the fraction of Cauchy decrease condition on \(w_k\) with respect to the inexact quadratic model \(\tilde{q}_k\),

\[
\tilde{q}_k(0) - \tilde{q}_k(w_k) \geq \kappa_4 \|\tilde{W}_k^* g_k\|_Z \min \left\{ \kappa_5 \|\tilde{W}_k^* g_k\|_Z, \kappa_6 \Delta_k \right\},
\] (3.8)

for \(\kappa_4, \kappa_5, \kappa_6 > 0\), independent of \(k\). Depending on the algorithm for the solution of the reduced quadratic subproblem, establishing the existence of \(\tilde{W}_k\), \(\tilde{W}_k^* g_k\), and \(\tilde{W}_k^* H_k \tilde{W}_k\) is by no means trivial, and is a subject of study of Section 3.6.
3.3 Balancing Progress in the Computation of the Tangential and the Quasi-Normal Step

Assuming that we have computed the reduced-space step \( w_k \in Z \) satisfying the requirements discussed in Section 3.2, we still have to perform the computation of the tangential step \( \tilde{t}_k \in X \). With exact linear system solves, \( \tilde{t}_k = W_k w_k \). Since the application of \( W_k \) requires the solution of linear systems, it is not possible to compute \( W_k w_k \) exactly. Hence, we need to determine how much the effective tangential step \( \tilde{t}_k \in X \) can differ from \( W_k w_k \).

In order to better understand how to balance feasibility and optimality in presence of inexactness, we revisit the step acceptance criteria from Section 2.1.5, in particular the measure of predicted reduction. In exact arithmetic, the following holds

\[
c_x(x_k)(n_k + W_k w_k) = c_x(x_k)n_k,
\]

thus we obtain for the predicted reduction defined in (2.35)

\[
\text{pred}(s^x_k; \rho_k) = \phi(x_k, \lambda_k; \rho_k) - \left[ \mathcal{L}(x_k, \lambda_k) + \langle \nabla_x \mathcal{L}(x_k, \lambda_k), s^x_k \rangle_X + \frac{1}{2} \langle H_k s^x_k, s^x_k \rangle_X \right] \\
+ \langle \lambda_{k+1} - \lambda_k, c_x(x_k)s^x_k + c(x_k) \rangle_Y + \rho_k \| c_x(x_k)s^x_k + c(x_k) \|_Y^2 \right]
\]

\[
= \mathcal{L}(x_k, \lambda_k) - \left( \mathcal{L}(x_k, \lambda_k) + \langle W^*_k(\nabla_x \mathcal{L}(x_k, \lambda_k) + H_k n_k), w_k \rangle_Z \right) \\
+ \frac{1}{2} \langle W^*_k H_k W_k w_k, w_k \rangle_Z - \langle \nabla_x \mathcal{L}(x_k, \lambda_k), n_k \rangle_X - \frac{1}{2} \langle H_k n_k, n_k \rangle_X \\
- \langle \lambda_{k+1} - \lambda_k, c_x(x_k)n_k + c(x_k) \rangle_Y \\
+ \rho_k \left( \| c(x_k) \|_Y^2 - \| c_x(x_k)n_k + c(x_k) \|_Y^2 \right).
\]
Therefore,

\[
pred(s_k^r; \rho_k) = - \langle W_k^* (\nabla_x \mathcal{L}(x_k, \lambda_k) + H_k n_k), w_k \rangle_Z - \frac{1}{2} \langle W_k^* H_k W_k w_k, w_k \rangle_Z
\]

\[
- \langle \nabla_x \mathcal{L}(x_k, \lambda_k), n_k \rangle_{\mathcal{H}} - \frac{1}{2} \langle H_k n_k, n_k \rangle_{\mathcal{H}}
\]

\[
- \langle \lambda_{k+1} - \lambda_k, c_x(x_k) n_k + c(x_k) \rangle_{\mathcal{Y}}
\]

\[
+ \rho_k \left( \|c(x_k)\|_{\mathcal{Y}}^2 - \|c_x(x_k) n_k + c(x_k)\|_{\mathcal{Y}}^2 \right)
\].

With inexact linear system solves the trial step is \( s_k^r = n_k + \tilde{t}_k \), with

\[
c_x(x_k)(n_k + \tilde{t}_k) = c_x(x_k) n_k + r_k^t \neq c_x(x_k) n_k,
\]

where we have defined

\[
r_k^t \equiv c_x(x_k) \tilde{t}_k.
\]

This will effect the term

\[
- \langle \lambda_{k+1} - \lambda_k, c_x(x_k) n_k + c(x_k) \rangle_{\mathcal{Y}} + \rho_k \left( \|c(x_k)\|_{\mathcal{Y}}^2 - \|c_x(x_k) n_k + c(x_k)\|_{\mathcal{Y}}^2 \right)
\]

in the representation of \( pred(s_k^r; \rho_k) \) given above.

For the predicted reduction with inexactness we replace

\[
- \langle W_k^* (\nabla_x \mathcal{L}(x_k, \lambda_k) + H_k n_k), w_k \rangle_Z - \frac{1}{2} \langle W_k^* H_k W_k w_k, w_k \rangle_Z
\]

by

\[
- \langle \widetilde{W_k^* g_k}, w_k \rangle_Z - \frac{1}{2} \langle \widetilde{W_k^* H_k W_k} w_k, w_k \rangle_Z = \widetilde{q_k}(0) - \widetilde{q_k}(w_k),
\]

and replace \( n_k \) by \( n_k + \tilde{t}_k \) in

\[
- \langle \lambda_{k+1} - \lambda_k, c_x(x_k) n_k + c(x_k) \rangle_{\mathcal{Y}} + \rho_k \left( \|c(x_k)\|_{\mathcal{Y}}^2 - \|c_x(x_k) n_k + c(x_k)\|_{\mathcal{Y}}^2 \right)
\].
We consequently obtain for the predicted reduction with inexactness

\[
\widehat{\text{pred}}(s_k^2; \rho_k) \equiv \text{pred}(n_k, w_k; \rho_k) + r\text{pred}(r_k^1; \rho_k),
\]

where

\[
\text{pred}(n_k, w_k; \rho_k) \equiv \tilde{g}_k(0) - \tilde{g}_k(w_k) - \langle \nabla_x \mathcal{L}(x_k, \lambda_k), n_k \rangle_X - \frac{1}{2} \langle H_k n_k, n_k \rangle_X \\
- \langle \lambda_{k+1} - \lambda_k, c_x(x_k)n_k + c(x_k) \rangle_Y \\
+ \rho_k \left( \|c(x_k)\|_Y^2 - \|c_x(x_k)n_k + c(x_k)\|_Y^2 \right), \tag{3.9}
\]

and

\[
r\text{pred}(r_k^1; \rho_k) \equiv -\langle \lambda_{k+1} - \lambda_k, r_k^1 \rangle_Y - \rho_k \|r_k^1\|_Y^2 - 2\rho_k \langle r_k^1, c_x(x_k)n_k + c(x_k) \rangle_Y. \tag{3.10}
\]

In practice, we first compute a penalty parameter \(\rho_k\) satisfying

\[
\text{pred}(n_k, w_k; \rho_k) \geq \frac{\rho_k}{2} \left( \|c(x_k)\|_Y^2 - \|c_x(x_k)n_k + c(x_k)\|_Y^2 \right),
\]

and then make sure that the effective tangential step \(\tilde{t}_k\) satisfies the requirement

\[
|r\text{pred}(r_k^1; \rho_k)| \leq \eta_0 \text{pred}(n_k, w_k; \rho_k), \tag{3.11}
\]

where \(\eta_0 \in (0, 1 - \eta_1)\), and \(\eta_1 \in (0, 1)\) is the smallest acceptable ratio of the actual and predicted reduction.

There are three additional, rather technical conditions that need to be satisfied by the effective tangential step. They are required for the proof of global convergence of Alg. 3.4.2. The first is related to how much the effective tangential step can deviate from \(W_k w_k\), and reads

\[
\|\tilde{t}_k - W_k w_k\|_X \leq \xi_3 \Delta_k \|s_k^2\|_X, \tag{3.12}
\]
for some $\xi_3 > 0$ independent of $k$. The second is a boundedness condition,

$$\|\hat{r}_k\|_\mathcal{X} \leq \xi_4 \Delta_k,$$

(3.13)

where $\xi_4 > 0$ is independent of $k$. The third condition is only indirectly related to the computation of the effective tangential step. The reduced-space tangential step $w_k$ must satisfy

$$\|w_k\|_\mathcal{Z} \leq \xi_5 s_k^e \|\chi\|,$$

(3.14)

for $\xi_3 > 0$ independent of $k$.

### 3.4 Statement of the Generalized Trust–Region SQP Algorithm with Inexactness Control

In this section we state a trust–region SQP algorithm that provides a generic theoretical framework for inexactness control.

For the update of the penalty parameter $\rho_k$, we use Alg. 3.4.1.

**Algorithm 3.4.1.** (*Penalty Parameter Update with Inexactness.*)

1. Assumption: Let $\rho_{-1} = 1$ and let $\bar{\rho}$ be a small positive constant.

2. If

$$\text{pred}(n_k, w_k; \rho_{k-1}) \geq \frac{\rho_{k-1}}{2} (\|c(x_k)\|^2_\mathcal{Y} - \|c_z(x_k)n_k + c(x_k)\|^2_\mathcal{Y})$$

then set $\rho_k = \rho_{k-1}$. Otherwise set

$$\rho_k = \frac{2\text{pred}(n_k, w_k; \rho_{k-1})}{\|c(x_k)\|^2_\mathcal{Y} - \|c_z(x_k)n_k + c(x_k)\|^2_\mathcal{Y}} + \bar{\rho}.$$

□

We now have all ingredients necessary to formulate the SQP algorithm in the inexact setting.

**Algorithm 3.4.2.** (*Trust-region SQP algorithm with inexact linear system solves.*)
1. Initialization. Choose initial point $x_0$, initial trust-region radius $\Delta_0$, constants $0 < \alpha_1, \eta_1 < 1$, $0 < \eta_0 < 1 - \eta_1$, $\rho_1 \geq 1$, and $tol^{SQP} > 0$. Set $\Delta_{min}, \Delta_{max}$ so that $0 < \Delta_{min} < \Delta_{max}$. Compute an initial Lagrange multiplier estimate $\lambda_0$.

2. For $k = 0, 1, 2, \ldots$
   
   (a) Convergence Check. If
   
   $$\|\nabla_x \mathcal{L}(x_k, \lambda_k)\|_x < tol^{SQP} \quad \text{and} \quad \|c(x_k)\|_y < tol^{SQP},$$
   
   then terminate.

   (b) Step Computation.
   
   i. Compute a quasi-normal step $n_k$ satisfying (3.1) and (3.2).
   
   ii. Compute a reduced-space tangential step $w_k$ satisfying (3.6)–(3.8).

   (c) Acceptance Test.
   
   i. Compute a new Lagrange multiplier estimate $\lambda_{k+1}$ based on $n_k + t_k$.
   
   ii. Update penalty parameter $\rho_k$ via Alg. 3.4.1.
   
   iii. Compute effective tangential step $\tilde{t}_k$ satisfying the conditions (3.11), (3.12), (3.13), and (3.14).
   
   iv. Compute trial step $s_k^T = n_k + \tilde{t}_k$.
   
   v. Compute actual reduction $\text{ared}(s_k^T, \rho_k)$ and predicted reduction $\text{pred}(n_k, w_k; \rho_k)$, and their ratio $\theta_k = \frac{\text{ared}(s_k^T, \rho_k)}{\text{pred}(n_k, w_k; \rho_k)}$.
   
   vi. If $\theta_k \geq \eta_1$, set $x_{k+1} = x_k + s_k^T$, and choose $\Delta_{k+1}$ such that
   
   $$\max\{\Delta_{min}, \Delta_k\} \leq \Delta_{k+1} \leq \Delta_{max}.$$
   
   Otherwise set $x_{k+1} = x_k$, reset $\lambda_{k+1} = \lambda_k$, and set
   
   $$\Delta_{k+1} = \alpha_1 \max\{\|n_k\|_x, \|\tilde{t}_k\|_x\}.$$

\[\Box\]

### 3.5 Global Convergence of Alg. 3.4.2

The global convergence property of Alg. 3.4.2 is stated in the following theorem.

**Theorem 3.5.1.** Let assumptions $(A1)$–$(A7)$ be satisfied. The sequences of iterates generated by Alg. 3.4.2 satisfy

$$\liminf_{k \to \infty} (\|W_k^* g_k\|_z + \|c(x_k)\|_y) = 0. \quad (3.15)$$
Additionally, we have

\[
\liminf_{k \to \infty} (\|W_k^* \nabla f(x_k)\|_Z + \|c(x_k)\|_Y) = 0.
\]

(3.16)

**Proof.** The proof of (3.15) is based on the global convergence analysis given in [28] and closely follows the modification given in [55]. The component ‘pred\((s_k^n, (s_u)_k; \rho_k)\)’ of the predicted decrease in [55] will be replaced by \(\text{pred}(n_k, w_k; \rho_k)\) in our notation.

As in [55], the first modification relates the size of the trial step \(s_k^x\) to the trust-region radius \(\Delta_k\). We need that

\[
\|s_k^x\|_X \leq \kappa_7 \Delta_k,
\]

and, if \(s_k^x\) is rejected,

\[
\|s_k^x\|_X \geq \kappa_8 \Delta_k.
\]

The first inequality follows from \(s_k^x = n_k + \tilde{t}_k\), the condition (2.17b) (which is always satisfied directly by the corresponding quasi-normal step algorithm), and the requirement (3.13). The second inequality follows trivially from the trust region update in Step 2(c)vi of Alg. 3.4.2.

The second modification concerns the estimates of the actual and the predicted reduction. By analogy to [55] we will be able to show

\[
|\text{ared}(s_k^x; \rho_k) - \text{pred}(n_k, w_k; \rho_k) - \tau\text{pred}(r_k^t; \rho_k)|
\leq \kappa_9 \Delta_k \|s_k^x\|_X + \kappa_{10} \rho_k \|s_k^x\|_X^2 + \kappa_{11} \rho_k \|s_k^x\|_X \|c(x_k)\|
\]

(3.17)

instead of [28, Lemma 7.4] and

\[
|\text{ared}(s_k^x; \rho_k) - \text{pred}(n_k, w_k; \rho_k) - \tau\text{pred}(r_k^t; \rho_k)| \leq \kappa_{12} \rho_k \Delta_k \|s_k\|_X
\]

(3.18)
instead of [28, Lemma 7.5]. In the analysis below, we will use the following notation

\[ q_k^x(s) = \mathcal{L}(x_k, \lambda_k) + \langle \nabla_x \mathcal{L}(x_k, \lambda_k), s \rangle_{\mathcal{X}} + \frac{1}{2} \langle H_k s, s \rangle_{\mathcal{X}}, \]

\[ q_k^z(w) = q_k^x(n_k) + \langle W_k^* (H_k n_k + \nabla_x \mathcal{L}(x_k, \lambda_k)), w \rangle_{\mathcal{Z}} + \frac{1}{2} \langle W_k^* H_k W_k w, w \rangle_{\mathcal{Z}}, \]

\[ x_{k+1} = x_k + s_k^x, \]

and recall the definition of the actual reduction

\[ \text{ared}(s_k^x; \rho_k) = \mathcal{L}(x_k, \lambda_k) + \rho_k \| c(x_k) \|_Y^2 - \mathcal{L}(x_{k+1}, \lambda_{k+1}) - \rho_k \| c(x_{k+1}) \|_Y^2. \]

With definitions (3.9) and (3.10), we obtain

\[
\text{pred}(n_k, w_k; \rho_k) + \text{rpred}(r^t_k; \rho_k)
\]

\[ = - \langle \widetilde{W}_k^* g_k, w_k \rangle_{\mathcal{Z}} - \frac{1}{2} \langle \widetilde{W}_k^* H_k w_k, w_k \rangle_{\mathcal{Z}} - \langle \nabla_x \mathcal{L}(x_k, \lambda_k), n_k \rangle_{\mathcal{X}} - \frac{1}{2} \langle H_k n_k, n_k \rangle_{\mathcal{X}}
\]

\[ - \langle \Delta \lambda_k, c_x(x_k) s_k^x + c(x_k) \rangle_Y + \rho_k (\| c(x_k) \|_Y^2 - \| c_x(x_k) s_k^x + c(x_k) \|_Y^2). \]

This expression, the fact \( W_k^* (H_k n_k + \nabla_x \mathcal{L}(x_k, \lambda_k)) = W_k^* g_k, \) and additional manipulations (see [55]) yield

\[
\text{ared}(s_k^x; \rho_k) - \text{pred}(n_k, w_k; \rho_k) - \text{rpred}(r^t_k; \rho_k)
\]

\[ = - \mathcal{L}(x_{k+1}, \lambda_k) + q_k^x(s_k^x) - q_k^x(s_k^x) + q_k^z(w_k)
\]

\[ + \langle \widetilde{W}_k^* g_k - W_k^* g_k, w_k \rangle_{\mathcal{Z}} + \frac{1}{2} \langle \widetilde{W}_k^* H_k W_k w_k, w_k \rangle_{\mathcal{Z}} - \frac{1}{2} \langle W_k^* H_k W_k w_k, w_k \rangle_{\mathcal{Z}}
\]

\[ + \langle \Delta \lambda_k, -c(x_{k+1}) + c_x(x_k) s_k^x + c(x_k) \rangle_Y - \rho_k (\| c(x_{k+1}) \|_Y^2 - \| c_x(x_k) s_k^x + c(x_k) \|_Y^2). \]

(3.19)

A Taylor expansion of \( \mathcal{L}(x_{k+1}, \lambda_k) \) gives

\[
| - \mathcal{L}(x_{k+1}, \lambda_k) + q_k^x(s_k^x) | \leq \frac{1}{2} \| H_k - \nabla_x \mathcal{L}(x_k + \tau_k s_k^x, \lambda_k) \| s_k^x \|_2^2, \]

(3.20)
where $0 < \tau_k^1 < 1$. Definitions of $q_k^X$ and $q_k^Z$ and norm inequalities give

$$| - q_k^X(s_k^r) + q_k^Z(w_k) | \leq \| H_k n_k + \nabla_x \mathcal{L}(x_k, \lambda_k) \|_X \| \tilde{t}_k - W_k w_k \|_X$$

$$+ \frac{1}{2} \| H_k \| \| \tilde{t}_k \|_X^2 + \frac{1}{2} \| W_k^* H_k W_k \| \| w_k \|_Z^2,$$

and so from (3.12), (3.14), and

$$\| \tilde{t}_k \|_X \leq \| \tilde{t}_k - W_k w_k \|_X + \| W_k w_k \|_X,$$

we obtain

$$| - q_k^X(s_k^r) + q_k^Z(w_k) |$$

$$\leq C_5 \| H_k n_k + \nabla_x \mathcal{L}(x_k, \lambda_k) \|_X \| s_k^r \|_X$$

$$+ \frac{1}{2} \| H_k \| (\xi_3^2 \Delta_k^2 + 2 \xi_3 \xi_0 \| W_k \| \Delta_k + \xi_0^2 \| W_k \|^2) \| s_k^r \|_X^2$$

$$+ \frac{1}{2} \xi_0^2 \| W_k^* H_k W_k \| \| s_k^r \|_X^2. \tag{3.21}$$

Inequalities (3.6) and (3.7) yield

$$\left\langle \widetilde{W_k^* g_k} - W_k^* g_k, w_k \right\rangle_Z + \frac{1}{2} \left\langle \widetilde{W_k^* H_k W_k w_k}, w_k \right\rangle_Z - \frac{1}{2} \left\langle W_k^* H_k W_k w_k, w_k \right\rangle_Z$$

$$\leq \xi_1 \Delta_k \| w_k \|_Z + \frac{1}{2} (\xi_2 + \| W_k^* H_k W_k \|) \| w_k \|_Z^2$$

$$\leq \xi_1 \xi_5 \Delta_k \| s_k^r \|_X + \frac{1}{2} \xi_0^2 (\xi_2 + \| W_k^* H_k W_k \|) \| s_k^r \|_X^2. \tag{3.22}$$

The final step follows [55] directly. After a series of Taylor expansions, we obtain

$$\left\langle \Delta \lambda_k, -c(x_{k+1}) + c_x(x_k) s_k^r + c(x_k) \right\rangle_Y - \rho_k (\| c(x_{k+1}) \|_Y^2 - \| c_x(x_k) s_k^r + c(x_k) \|_Y^2)$$

$$\leq \kappa_{10} \rho_k \| s_k^r \|_X^2 + \kappa_{11} \rho_k \| s_k^r \|_X \| c(x_k) \|_Y. \tag{3.23}$$

Substitution of (3.20)–(3.23) into (3.19) along with the assumptions (A4)–(A7) gives the estimate (3.17). As $\rho_k \geq 1$, (3.18) follows. The estimates of the difference between
the actual and the predicted reduction can be found in [55], which concludes the part of the proof related to (3.15). The limit (3.16) is obtained by combining (3.15), (3.6), and (3.1).

\[ \blacksquare \]

### 3.6 A Generalized Conjugate Gradient Method for the Solution of the Tangential Subproblem

The reduced tangential subproblem (3.4) is commonly solved using a conjugate gradient (CG) algorithm with Steihaug–Toint stopping conditions. This is Alg. 2.1.5, which we state here again with slightly different notation. This will be the starting point in the analysis of inexactness.

**Algorithm 3.6.1. (Steihaug–Toint CG method for the solution of (3.4))**

1. For \( i = 0, 1, 2, \ldots \)
   
   \begin{enumerate}[(a)]
   
   \item If \( \langle \tilde{p}_i, W_{k}^* H_k W_k \tilde{p}_i \rangle \leq 0 \), compute \( \theta > 0 \) such that \( \| W_k (w_i + \theta \tilde{p}_i) \| = \Delta_k - \| n_k \| \) and return \( w_{i+1} = w_i + \theta \tilde{p}_i \).
   
   \item \( \tilde{\alpha}_i = \langle \tilde{r}_i, \tilde{r}_i \rangle \) / \( \langle \tilde{p}_i, W_{k}^* H_k W_k \tilde{p}_i \rangle \)
   
   \item \( w_{i+1} = w_i + \tilde{\alpha}_i \tilde{p}_i \)
   
   \item If \( \| W_k w_{i+1} \| \geq \Delta_k - \| n_k \| \), compute \( \theta > 0 \) such that \( \| W_k (w_i + \theta \tilde{p}_i) \| = \Delta_k - \| n_k \| \) and return \( w_{i+1} = w_i + \theta \tilde{p}_i \).
   
   \item \( \tilde{r}_{i+1} = \tilde{r}_i - \tilde{\alpha}_i W_{k}^* H_k W_k \tilde{p}_i \)
   
   \item \( \tilde{\beta}_i = \langle \tilde{r}_{i+1}, \tilde{r}_{i+1} \rangle \) / \( \langle \tilde{r}_i, \tilde{r}_i \rangle \)
   
   \item \( \tilde{p}_{i+1} = \tilde{r}_{i+1} + \tilde{\beta}_i \tilde{p}_i \)
   
   \[ \blacksquare \]

We now make several important observations. First, if linear systems arising in the application of \( W_k \) and \( W_k^* \) are solved iteratively, the linear operators \( W_k \) and \( W_k^* \) are being effectively applied with a certain level of inexactness. Second, in general, every iteration of the CG algorithm 3.6.1 uses different inexact counterparts of \( W_k \) and \( W_k^* \). For example, if a Krylov subspace method is used for the solution of the underlying
linear systems, one cannot guarantee that the same Krylov polynomials are used in every application of $W_k$ and $W_k^*$, due to their dependence on the right-hand side of the linear system, which changes with every CG iteration. Third, if we could write down an inexact CG operator that replaces $W_k^* H_k W_k$, it would be nonsymmetric. In fact, the CG iteration would view this operator as being effectively nonlinear. Considering all these difficulties, it is nearly impossible to determine which quadratic functional (if any) is being minimized by Alg. 3.6.1 in presence of inexactness.

We should note that several studies of inexact Krylov methods [16, 69] and inexact preconditioned conjugate gradient methods [1, 46, 61] address similar issues in the context of the solution of linear systems. Here, either the inexact operator $W_k^* H_k W_k$ is viewed as a variable Krylov operator, or $W_k$ is applied inexactly as a variable nonsingular preconditioning operator. Both change with every Krylov iteration. Nonetheless, the results of these studies are not directly applicable to the trust-region minimization context. In the first approach, the loss of symmetry in the Krylov operator does not allow us to solve the minimization problem by replacing it with a linear system. In the latter approach, the origin of $W_k$, which is a rectangular null-space operator, and not a preconditioner, is problematic. In both cases, the possibility of negative curvature and the trust-region constraint represent additional obstacles.

We propose a new inexact conjugate gradient algorithm for the solution of the tangential subproblem. This algorithm uses a full-space approach, in which the CG operator is $H_k$, and can be applied exactly. The inexactness is moved into a "preconditioning" operator, for which we can prove the existence of a fixed linear representation. This fact along with the symmetry of the CG operator $H_k$ allow us to relate the new inexact CG algorithm to the minimization of a quadratic functional.

In the inexact regime, the most significant obstacle in applying Alg. 3.6.1 to the reduced tangential subproblem is the loss of symmetry (and potentially linearity) of the inexact reduced Hessian. This will be the case for any formulation that operates in
the reduced space \( \mathcal{Z} \). We thus try to move the nonlinearities away from the Hessian operator by considering problem (3.4) in the full space \( \mathcal{X} \). We let \( t = W_k w \), and obtain the full-space reformulation of (3.4),

\[
\begin{align*}
\min & \quad q_k(t) \equiv \frac{1}{2} \langle H_k t, t \rangle_{\mathcal{X}} + \langle g_k, t \rangle_{\mathcal{X}} + q_k^*(n_k) \\
\text{s.t.} & \quad t \in \text{Range}(W_k) \\
& \quad \|t\|_{\mathcal{X}} \leq \Delta_k - \|n_k\|_{\mathcal{X}}.
\end{align*}
\]

(3.24a) (3.24b) (3.24c)

Analytically, this formulation has no advantages over the original full-space problem (3.3), however, there is a major algorithmic advantage. In particular, if we apply a conjugate gradient scheme in which we can explicitly ensure that the constraints (3.24b) and (3.24c) are satisfied at every iteration, we will have solved (3.24). This can be done by treating the null-space operator \( W_k \) as an inexact preconditioner, for which we will find a fixed linear representation.

We start by introducing the variable null-space operators \( \mathcal{W}_k : \mathcal{Z} \to \mathcal{X} \) and \( \mathcal{W}_k^* : \mathcal{X} \to \mathcal{Z} \) that represent \( W_k \) and \( W_k^* \), respectively, but can change at every CG iteration. The composite operator \( \mathcal{V}_k : \mathcal{X} \to \mathcal{X} \), defined by

\[ \mathcal{V}_k = \mathcal{W}_k \circ \mathcal{W}_k^* \]

will be used in our inexact CG algorithm. We now formulate the inexact full-space CG algorithm with full orthogonalization, for the solution of (3.24) in presence of inexactness.

**Algorithm 3.6.2. (Full-space CG for the solution of (3.24))**

1. Let \( t_0 = 0 \in \mathcal{X} \). Let \( r_0 = g_k \). Set \( i_{\max} \), set \( i = 0 \).
2. While \((r_i \neq 0 \text{ and } \mathcal{V}_k(r_i) \neq 0 \text{ and } i < i_{\max}\)

   \[
   \begin{align*}
   (a) & \quad z_i = \mathcal{V}_k(r_i) \\
   (b) & \quad p_i = -z_i + \sum_{j=0}^{i-1} \frac{\langle z_i, H_k p_j \rangle_{\mathcal{X}}}{\langle p_j, H_k p_j \rangle_{\mathcal{X}}} p_j
   \end{align*}
   \]
(c) If \( \langle p_i, H_k p_i \rangle_X \leq 0 \) or \( \langle p_i, r_i \rangle_X \leq 0 \), extend \( t_i \) to boundary of trust–region and stop.

(d) \( \alpha_i = -\frac{\langle r_i, p_i \rangle_X}{\langle p_i, H_k p_i \rangle_X} \)

(e) \( t_{i+1} = t_i + \alpha_i p_i \)

(f) If \( \|t_{i+1}\|_X \geq \Delta_k - \|n_k\|_X \), extend \( t_i \) to boundary of trust–region and stop.

(g) \( r_{i+1} = H_k t_{i+1} + g_k = r_i + \alpha_i H_k p_i \)

(h) \( i \rightarrow i + 1 \)

\( \square \)

**Remark 3.6.3.** The termination condition \( \langle p_i, r_i \rangle_X < 0 \) in Step 1c of Alg 3.6.2 is necessary to ensure that the computed coefficients \( \alpha_i \) are nonpositive. This is a necessary (but in the above case not sufficient) condition for the monotone increase of the norms of the iterates \( t_i \).

The second part of the termination condition, \( \langle p_i, r_i \rangle_X = 0 \), is used in Lemma 3.6.10.

The control of inexactness in the inexact null–space operator is the subject of the next chapter. Nonetheless, several important properties of Alg. 3.6.2 can already be established. We first show the equivalence of Alg. 3.6.2 and Alg. 3.6.1 in the exact case. This property is important, as we would like to recover the convergence properties of the exact CG method if linear system solves can be performed with high accuracy. We then show that the search directions \( p_i \) are \( H_k \)-orthogonal, which further enables us to use a standard technique to prove that every iterate of the algorithm minimizes \( q \) over the set of the previously computed search directions. Finally, we show that, from the point of view of Alg. 3.6.2, the nonlinear operators \( \mathcal{W}_k \) and \( \mathcal{W}_k^* \) can be replaced with linear representations \( W_{2,k} \) and \( W_{1,k}^* \), respectively. The existence of \( W_{2,k} \) enables us to formulate an inexact quadratic functional which is minimized by Alg. 3.6.2.

We start with the proof of equivalence of Alg. 3.6.2 and Alg. 3.6.1 in the exact case. In order to simplify the presentation, we introduce an equivalent version of Alg. 3.6.1.
Algorithm 3.6.4. (Steinhaug-Toint CG for the solution of (3.4), alternate.)

0. Let \( w_0 = 0 \in Z \). Let \( \tilde{r}_0 = W_k^* g_k \).

1. For \( i = 0, 1, 2, \ldots \)

   (a) \( \tilde{p}_i = -\tilde{r}_i + \sum_{j=0}^{i-1} \frac{\langle \tilde{r}_j, W_k^* H_k W_k \tilde{p}_j \rangle_Z}{\langle \tilde{p}_j, W_k^* H_k W_k \tilde{p}_j \rangle_Z} \tilde{p}_j \)

   (b) If \( \langle \tilde{p}_i, W_k^* H_k W_k \tilde{p}_i \rangle_Z \leq 0 \) or \( \langle \tilde{p}_i, \tilde{r}_i \rangle_Z \leq 0 \) (see Remark 3.6.3), extend \( W_k w_i \) to boundary of trust-region and stop.

   (c) \( \tilde{\alpha}_i = \frac{\langle \tilde{r}_i, \tilde{p}_i \rangle_Z}{\langle \tilde{p}_i, W_k^* H_k W_k \tilde{p}_i \rangle_Z} \)

   (d) \( w_{i+1} = w_i + \tilde{\alpha}_i \tilde{p}_i \)

   (e) If \( \| W_k w_{i+1} \| \geq \Delta_k - \| n_k \|_X \), extend \( W_k w_i \) to boundary of trust-region and stop.

   (f) \( \tilde{r}_{i+1} = W_k^* H_k W_k w_{i+1} + W_k^* g_k \)

\[ \square \]

We should note that the sum in Step (1a) automatically truncates to its last term for \( i \geq 1 \), i.e.

\[
\sum_{j=0}^{i-1} \frac{\langle \tilde{r}_j, W_k^* H_k W_k \tilde{p}_j \rangle_Z}{\langle \tilde{p}_j, W_k^* H_k W_k \tilde{p}_j \rangle_Z} \tilde{p}_j = \frac{\langle \tilde{r}_i, W_k^* H_k W_k \tilde{p}_{i-1} \rangle_Z}{\langle \tilde{p}_{i-1}, W_k^* H_k W_k \tilde{p}_{i-1} \rangle_Z} \tilde{p}_{i-1},
\]

due to an argument similar to Remark 3.6.9. All other relationships necessary to establish the equivalence of Alg. 3.6.1 and Alg. 3.6.4 can be found in standard textbooks, see e.g. [65, p.187ff] or [60, p.108ff].

Lemma 3.6.5. If \( W_k = W_k \) and \( W_k^* = W_k^* \) are linear operators, then Alg. 3.6.2 applied to the full-space problem (3.24) generates at each iteration \( i \) the quantities \( p_i, \alpha_i, t_{i+1} \) and \( r_{i+1} \) such that

\[
p_i = W_k \tilde{p}_i, \quad \alpha_i = \tilde{\alpha}_i, \quad t_{i+1} = W_k w_{i+1}, \quad \text{and} \quad r_{i+1} = W_k^* r_{i+1},
\]

for the corresponding quantities \( \tilde{p}_i, \tilde{\alpha}_i, w_{i+1} \) and \( \tilde{r}_{i+1} \) generated by Alg. 3.6.4, and is therefore equivalent to Alg. 3.6.1 applied to the reduced-space problem (3.4).
Proof. We show by induction that the desired relationships hold for every iteration \( i \) of Alg. 3.6.2 and Alg. 3.6.4. We start at iteration \( i = 0 \). Before entering the iteration, we have already set \( \tilde{r}_0 = W_k^* g_k = W_k^* r_0 \). We have

\[
p_0 = -W_k W_k^* r_0 = -W_k (W_k^* r_0) = -W_k \tilde{r}_0 = W_k \tilde{p}_0, \]

\[
\alpha_0 = \frac{-\langle r_0, p_0 \rangle_X}{\langle p_0, H_k p_0 \rangle_X} = \frac{-\langle r_0, W_k \tilde{p}_0 \rangle_X}{\langle W_k \tilde{p}_0, H_k W_k \tilde{p}_0 \rangle_X} = \frac{-\langle W_k^* r_0, \tilde{p}_0 \rangle_Z}{\langle \tilde{p}_0, W_k^* H_k W_k \tilde{p}_0 \rangle_Z} = \frac{-\langle \tilde{r}_0, \tilde{p}_0 \rangle_Z}{\langle \tilde{p}_0, W_k^* H_k W_k \tilde{p}_0 \rangle_Z} = \bar{\alpha}_0, \]

\[
t_1 = \alpha_0 p_0 = \alpha_0 W_k \tilde{p}_0 = W_k (\alpha_0 \tilde{p}_0) = W_k w_1, \]

and

\[
\tilde{r}_1 = W_k^* H_k W_k w_1 + W_k^* g_k = W_k^* (H_k (W_k w_1) + g_k) = W_k^* (H_k t_1 + g_k) = W_k^* r_1. \]

We now assume that the claim is true for iteration \( i - 1 \) and prove it for iteration \( i \). By the induction hypothesis, before we enter iteration \( i \) we have \( \tilde{r}_i = W_k^* r_i \) and \( t_i = W_k w_i \). Then, following the same procedure as for \( i = 0 \), we get

\[
p_i = -W_k W_k^* r_i + \sum_{j=0}^{i-1} \frac{\langle W_k W_k^* r_j, H_k p_j \rangle_X}{\langle p_j, H_k p_j \rangle_X} p_j = -W_k \tilde{r}_i + \sum_{j=0}^{i-1} \frac{\langle W_k \tilde{r}_j, H_k W_k \tilde{p}_j \rangle_X}{\langle W_k \tilde{p}_j, H_k W_k \tilde{p}_j \rangle_X} W_k \tilde{p}_j \]

\[
= W_k \left( -\tilde{r}_i + \sum_{j=0}^{i-1} \frac{\langle \tilde{r}_j, W_k^* H_k W_k \tilde{p}_j \rangle_Z}{\langle \tilde{p}_j, W_k^* H_k W_k \tilde{p}_j \rangle_Z} \tilde{p}_j \right) = W_k \tilde{p}_i, \]

\[
\alpha_i = \frac{\langle r_i, p_i \rangle}{\langle p_i, H_k p_i \rangle_X} = \frac{\langle r_i, W_k \tilde{p}_i \rangle_X}{\langle W_k \tilde{p}_i, H_k W_k \tilde{p}_i \rangle_X} = \frac{\langle W_k^* r_i, \tilde{p}_i \rangle_Z}{\langle \tilde{p}_i, W_k^* H_k W_k \tilde{p}_i \rangle_Z} = \frac{\langle \tilde{r}_i, \tilde{p}_i \rangle_Z}{\langle \tilde{p}_i, W_k^* H_k W_k \tilde{p}_i \rangle_Z} = \bar{\alpha}_i, \]

\[
t_{i+1} = t_i + \alpha_i p_i = W_k w_i + \alpha_i W_k \tilde{p}_i = W_k (w_i + \bar{\alpha}_i \tilde{p}_i) = W_k w_{i+1}, \]

and

\[
\tilde{r}_{i+1} = W_k^* H_k W_k w_{i+1} + W_k^* g_k = W_k^* (H_k (W_k w_{i+1}) + g_k) = W_k^* (H_k t_{i+1} + g_k) = W_k^* r_{i+1}. \]

\( \square \)
The following lemma shows the $H_k$-conjugacy of the search directions, which will help us in proving the minimization property in Lemma 3.6.7.

**Lemma 3.6.6 (Conjugacy Property).** If $H_k$ is a self-adjoint linear operator then

$$\langle p_s, H_k p_l \rangle_X = 0,$$

for all $s \neq l$, where $0 \leq l, s \leq i$.

**Proof.** We start with $i = 1$. Since $H_k$ is self adjoint, we have

$$\langle p_0, H_k p_1 \rangle_X = \langle p_1, H_k p_0 \rangle_X = \left\langle -V_k(r_1) + \frac{\langle V_k(r_1), H_k p_0 \rangle_X p_0, H_k p_0 \rangle}{\langle p_0, H_k p_0 \rangle} \right\rangle_X = -\langle V_k(r_1), H_k p_0 \rangle_X + \langle V_k(r_1), H_k p_0 \rangle_X = 0.$$

We now assume that the claim is true for iteration $i - 1$, i.e. where $0 \leq l, s \leq i - 1$, and show its validity for iteration $i$, where $0 \leq l, s \leq i$. Without loss of generality, assume $l < s$. Due to the induction hypothesis, the claim need only be shown for the case $0 \leq l < s = i$. We obtain

$$\langle p_i, H_k p_i \rangle_X = \langle p_i, H_k p_i \rangle_X = \left\langle -V_k(r_i) + \sum_{j=0}^{i-1} \frac{\langle V_k(r_i), H_k p_j \rangle_X p_j, H_k p_i \rangle}{\langle p_j, H_k p_j \rangle_X} \right\rangle_X,$$

and notice that the last index in the sum above is $i - 1$, i.e. we can apply the induction hypothesis to obtain

$$\langle p_i, H_k p_i \rangle_X = \langle p_i, H_k p_i \rangle_X = \left\langle -V_k(r_i) + \frac{\langle V_k(r_i), H_k p_i \rangle_X p_i, H_k p_i \rangle}{\langle p_i, H_k p_i \rangle_X} \right\rangle_X = 0.$$

\[\Box\]

**Lemma 3.6.7 (Minimization Property).** The iterate $t_{i+1}$ minimizes $q_k$, defined in (3.24a), over $t = t_0 + \text{span} \{p_0, p_1, ..., p_i\}$. 
Proof. Any vector $t$ in the set $t_0 + \text{span} \{p_0, p_1, \ldots, p_i\}$ can be written as

$$t = t_0 + \sum_{j=0}^{i} \alpha_j p_j,$$

for some $\alpha_0, \ldots, \alpha_i$. Thus, minimizing $q_k$ over the given set is equivalent to solving

$$\min_{\alpha_{0}, \alpha_{1}, \ldots, \alpha_{i}} q_k \left( t_0 + \sum_{j=0}^{i} \alpha_j p_j \right).$$

We have

$$q_k \left( t_0 + \sum_{j=0}^{i} \alpha_j p_j \right) = \frac{1}{2} \left\langle H_k t_0 + H_k \sum_{j=0}^{i} \alpha_j p_j, t_0 + \sum_{j=0}^{i} \alpha_j p_j \right\rangle_{\chi} + \left\langle g_k, t_0 + \sum_{j=0}^{i} \alpha_j p_j \right\rangle_{\chi} + q_k^X (\eta_k)$$

$$= \frac{1}{2} \sum_{j=0}^{i} \alpha_j^2 \langle p_j, H_k p_j \rangle_{\chi} + \sum_{j=0}^{i} \alpha_j \langle H_k t_0 + g_k, p_j \rangle_{\chi} + \gamma,$$

where $\gamma$ does not depend on $\alpha_j$. All cross-terms vanish due to the conjugacy property. As Alg. 3.6.2 has not terminated, we have $\langle p_j, H_k p_j \rangle > 0$, for $j = 0, 1, \ldots, i$, thus a unique minimizer $(\alpha_0, \alpha_1, \ldots, \alpha_i)$ of the above expression exists and can be obtained by setting

$$\frac{\partial}{\partial \alpha_j} q_k (\alpha_0, \alpha_1, \ldots, \alpha_i) = 0.$$

For each $j = 0, 1, \ldots, i$ we get

$$\alpha_j^* \langle p_j, H_k p_j \rangle_{\chi} + \langle r_0, p_j \rangle_{\chi} = 0,$$

i.e.

$$\alpha_j^* = -\frac{\langle r_0, p_j \rangle_{\chi}}{\langle p_j, H_k p_j \rangle_{\chi}},$$
and therefore obtain
\[
  t^* = t_0 - \sum_{j=0}^{i} \frac{\langle r_0, p_j \rangle}{\langle p_j, H_k p_j \rangle} p_j.
\]

Step (1g) of Alg. 3.6.2 implies
\[
  r_j = r_0 - \sum_{s=0}^{j-1} \frac{\langle r_s, p_s \rangle}{\langle p_s, H_k p_s \rangle} H_k p_s,
\]
thus
\[
  \langle r_j, p_j \rangle = \left( r_0 - \sum_{s=0}^{j-1} \frac{\langle r_s, p_s \rangle}{\langle p_s, H_k p_s \rangle} H_k p_s, p_j \right) = \langle r_0, p_j \rangle,
\]
where we have used the conjugacy property. We finally obtain
\[
  t^* = t_0 - \sum_{j=0}^{i} \frac{\langle r_j, p_j \rangle}{\langle p_j, H_k p_j \rangle} p_j,
\]
which is equivalent to Step (1e). \(\square\)

In the remainder of this section, we turn to proving the existence of fixed linear operators that can replace their nonlinear counterparts \(W_k\) and \(W^*_k\) in Alg. 3.6.2. The existence of these operators is based on Lemma 3.6.15, which states that the action of any nonlinear operator on a linearly independent set of vectors can always be interpreted as the action of some linear operator on that set of vectors. As far as Alg. 3.6.2 is concerned, there are two sets of vectors to which nonlinear operators are applied. One, associated with the operator \(W^*_k\), is the set of residuals \(\{r_i\}\), the other is the set \(\{W_k^*(r_i)\}\), acted on by the operator \(W_k\). We will prove that the set \(\{r_i\}\) is linearly independent, by using Lemma 3.6.8. We will assume that the set \(\{W_k^*(r_i)\}\) is linearly independent. In practice, this turns out to be a rather weak assumption. One possible explanation for this is that if \(W_k = W_k\), the set \(\{W_k^*(r_i)\}\) is not only linearly independent, but in fact orthogonal, see Lemma 3.6.12.
Lemma 3.6.8. For all \( l \) and \( s \) such that \( 0 \leq l < s \leq i + 1 \), we have

\[
\langle r_s, p_l \rangle_{X} = 0.
\]

Proof. We start the inductive argument with \( i = 0 \), in which case we need to consider

\[
\langle r_1, p_0 \rangle_{X} = \left( r_0 - \frac{\langle r_0, p_0 \rangle_{X}}{\langle p_0, H_k p_0 \rangle_{X}} H_k p_0, p_0 \right)_{X} = \langle r_0, p_0 \rangle_{X} - \langle r_0, p_0 \rangle_{X} = 0.
\]

We suppose that the claim is true for iteration \( i - 1 \), i.e. for \( 0 \leq l < s \leq i \), and show its validity for iteration \( i \), i.e. the case \( 0 \leq l < s \leq i + 1 \). The induction hypothesis reduces this case to \( 0 \leq l < s = i + 1 \). We obtain

\[
\langle r_{i+1}, p_{i} \rangle_{X} = \left( r_{i} - \frac{\langle r_{i}, p_{i} \rangle_{X}}{\langle p_{i}, H_k p_{i} \rangle_{X}} H_k p_{i}, p_{i} \right)_{X} = \langle r_{i}, p_{i} \rangle_{X} - \frac{\langle r_{i}, p_{i} \rangle_{X}}{\langle p_{i}, H_k p_{i} \rangle_{X}} \langle p_{i}, H_k p_{i} \rangle_{X}
\]

\[
= \begin{cases} 
\langle r_{i}, p_{i} \rangle_{X} - \langle r_{i}, p_{i} \rangle_{X} = 0, & \text{for } l = i, \\
0 - \frac{\langle r_{i}, p_{i} \rangle_{X}}{\langle p_{i}, H_k p_{i} \rangle_{X}}, & \text{for } l < i,
\end{cases}
\]

using the induction hypothesis and the conjugacy property. \( \square \)

Remark 3.6.9. Let \( V_k = V_k \) be a self-adjoint linear operator. We can now easily show the reduction of \( \sum_{j=0}^{i-1} \frac{(V_k r_j, H_k p_j)_{X}}{(p_j, H_k p_j)_{X}} p_j \) to \( \frac{(V_k r_{i-1}, H_k p_{i-1})_{X}}{(p_{i-1}, H_k p_{i-1})_{X}} p_{i-1} \). For any \( j < i \), from Lemma 3.6.8 we get

\[
0 = \langle r_{i}, p_{j} \rangle_{X} = \left( r_{i}, -V_k r_{j} + \sum_{l=0}^{j-1} \frac{(V_k r_{j}, H_k p_{l})_{X}}{(p_{l}, H_k p_{l})_{X}} p_{l} \right)_{X},
\]

which by another application of Lemma 3.6.8 implies

\[
\langle r_{i}, V_k r_{j} \rangle_{X} = 0.
\]

Since \( V_k \) is self-adjoint,

\[
\langle r_{j}, V_k r_{i} \rangle_{X} = 0
\]
is also true. From Step (1g) of Alg. 3.6.2 we obtain

\[ \langle V_k r_i, H_k p_j \rangle_X = \frac{\langle p_i, H_k p_i \rangle_X}{\langle r_i, p_i \rangle_X} \langle V_k r_i, r_j - r_{j+1} \rangle_X = 0 \]

i.e.

\[ \langle V_k (r_i), H_k p_j \rangle_X = \langle V_k r_i, H_k p_j \rangle_X = 0, \]

for every iteration \( i \geq 2 \) and any \( 0 \leq j \leq i - 2 \), and the claim follows.

**Lemma 3.6.10.** For every iteration \( i \) of Alg. 3.6.2 the residual vectors \( r_0, \ldots, r_i \) are linearly independent.

**Proof.** It is obvious that \( r_i \neq 0 \), and \( i < i_{\text{max}} \); otherwise Alg. 3.6.2 would have terminated after iteration \( i - 1 \). Suppose that the vectors \( r_0, \ldots, r_i \) are linearly dependent. Then there exists a sequence of coefficients \( \{\xi_j\}_{j=0}^i \) with \( \sum_{j=0}^i |\xi_j| \neq 0 \), i.e. not all \( \xi_j \) can be zero, such that

\[ \sum_{j=0}^i \xi_j r_j = 0. \]

Without loss of generality let \( \xi_s \) be the first nonzero coefficient in \( \{\xi_j\}_{j=0}^i \), i.e. the sum above starts at index \( s \). Thus

\[ 0 = \left\langle \sum_{j=s}^i \xi_j r_j, p_s \right\rangle_X = \xi_s \langle r_s, p_s \rangle_X, \]

due to Lemma 3.6.8, i.e. \( \langle r_s, p_s \rangle_X = 0 \). This means that Alg. 3.6.2 would have terminated at iteration \( s \) in Step (1c) with \( \langle r_s, p_s \rangle_X = 0 \), and the linear combination from above reduces to \( \xi_s r_s = 0 \), which is a contradiction to the assumptions \( \xi_s \neq 0 \) and \( r_s \neq 0 \).

**Remark 3.6.11.** We observe that the iterate \( t_{i+1} \) is the unique global minimizer of \( q_k \) over \( t_0 + \text{span} \{p_0, p_1, \ldots, p_i\} \). Assuming that \( \langle r_i, p_i \rangle_X \neq 0 \), we have that \( t_{i+1} \neq t_i \). Therefore, \( q_k(t_{i+1}) < q_k(t_i) \).
Lemma 3.6.12. Assuming that $W_k = W_k$ is a linear operator, the vectors $W_k^r_0, ..., W_k^r_i$ are orthogonal for every iteration $i$ of Alg. 3.6.2.

Proof. From Lemma 3.6.8, $\langle r_{j_1}, p_{j_2} \rangle = 0$ for $j_2 < j_1$, thus

$$\langle r_{j_1}, p_{j_2} \rangle = \left( r_{j_1}, -V_k(r_{j_2}) + \sum_{l=0}^{j_2-1} \frac{\langle V_k(r_{j_2}), H_k p_l \rangle}{\langle p_l, H_k p_l \rangle} p_l \right) = \langle r_{j_1}, -V_k(r_{j_2}) \rangle = 0.$$

In the exact case, we have $V = W_k W_k^*$, and so we obtain

$$\langle r_{j_1}, W_k W_k^* r_{j_2} \rangle = \langle W_k^* r_{j_1}, W_k^* r_{j_2} \rangle = 0,$$

for all $j_1 \neq j_2$. \qed

Assumption 3.6.13. The vectors $W_k^*(r_0), ..., W_k^*(r_i)$ are linearly independent for every iteration $i$ of Alg. 3.6.2.

Remark 3.6.14. In many cases, with help of Lemma 3.6.12 and under certain mild requirements, Assumption 3.6.13 can be eliminated. For example, in Section 4.1, we use Lemma 3.6.12 to show the invertibility of a matrix related to the quantities $W_k^*(r_j)$. Additionally, in large-scale applications of interest, it is feasible to explicitly verify Assumption 3.6.13, and use it as a termination criterion for Alg. 3.6.2.

We now state a lemma that will help us prove a main result.

Lemma 3.6.15. Let $\mathcal{X}$ and $\mathcal{Y}$ be Hilbert spaces, let $\mathcal{X}$ be separable, and let $R = \{r_i\}$ be a countable linearly independent set of vectors in $\mathcal{X}$. Let $N : \mathcal{X} \to \mathcal{Y}$ be an arbitrary (generally nonlinear) mapping. Then there exists a fixed linear operator $N : \mathcal{X} \to \mathcal{Y}$ such that $N r_i = N(r_i)$ for all $r_i \in R$.

Proof. Since $\mathcal{X}$ is separable and $R$ is linearly independent, there exists a countable basis $B$ for $\mathcal{X}$ such that $R \subseteq B$. Moreover, any vector $x \in \mathcal{X}$ can be uniquely written as

$$x = \sum_{b_j \in B} x_j b_j,$$
where \( x_j \in \mathbb{R} \). Considering this linear expansion, the desired linear operator \( \mathcal{N} \) can be defined as follows:

\[
\mathcal{N}x = \sum_{b_j \in B} x_j \mathcal{N}(b_j),
\]

for any \( x \in \mathcal{X} \). First, we note that \( \mathcal{N} \) is well-defined, since \( B \) is countable. Second, it has the correct property with respect to the set \( R \), as for any \( r_i \in R \)

\[
\mathcal{N}r_i = 1 \cdot \mathcal{N}(r_i) = \mathcal{N}(r_i),
\]

due to the uniqueness of the basis expansion and the fact that \( r_i \) is a basis vector. Finally, we must verify that \( \mathcal{N} \) is indeed linear. We have for arbitrary real constants \( \alpha \) and \( \beta \)

\[
\mathcal{N}(\alpha x + \beta y) = \sum_{b_j \in B} (\alpha x_j + \beta y_j) \mathcal{N}(b_j) = \alpha \sum_{b_j \in B} x_j \mathcal{N}(b_j) + \beta \sum_{b_j \in B} y_j \mathcal{N}(b_j) = \alpha \mathcal{N}x + \beta \mathcal{N}y.
\]

\[\Box\]

**Theorem 3.6.16.** Let \( \mathcal{X} \) and \( \mathcal{Z} \) be Hilbert spaces, let \( \mathcal{X} \) be separable, and let \( \mathcal{W}_k^* : \mathcal{X} \to \mathcal{Z} \) be an arbitrary nonlinear operator. Then there exists a fixed linear operator \( \mathcal{W}_{1,k} \) such that

\[
\mathcal{W}_k^*(r_i) = \mathcal{W}_{1,k}^* r_i,
\]

for every iteration \( i \) of Alg. 3.6.2.

**Proof.** The proof follows from Lemma 3.6.15 and Lemma 3.6.10. \[\Box\]

**Theorem 3.6.17.** Let \( \mathcal{X} \) and \( \mathcal{Z} \) be Hilbert spaces and let \( \mathcal{Z} \) be separable. Furthermore, let \( \mathcal{W}_k^* : \mathcal{X} \to \mathcal{Z} \) be a nonlinear operator that satisfies Assumption 3.6.13 and let \( \mathcal{W}_k : \mathcal{Z} \to \mathcal{X} \) be an arbitrary nonlinear operator. Then there exists a fixed linear operator \( \mathcal{W}_{2,k} \) such that

\[
\mathcal{W}_k(\mathcal{W}_k^*(r_i)) = \mathcal{W}_{2,k} \mathcal{W}_k^*(r_i),
\]

(3.26)
for every iteration $i$ of Alg. 3.6.2.

Proof. Follows directly from Lemma 3.6.15. \hfill \square

**Remark 3.6.18.** Under the assumptions of Theorem 3.6.17 the existence of a linear operator $V_k = W_{2,k}W_{1,k}^*$ that replaces the nonlinear operator $V_k = W_k \circ W_k^*$ with respect to Alg. 3.6.2 is obvious.

Having shown the existence of a fixed linear representation of the null-space operator, we can now revisit the global convergence criteria for the tangential step computation, as given in Section 3.2.

**Lemma 3.6.19.** Under Assumption 3.6.13, every iterate $t_i$ of Alg. 3.6.2 satisfies

$$ t_i \in \text{Range}(W_{2,k}). $$

Proof. Step (1b) of Alg. 3.6.2 yields $p_i \in \text{Range}(W_{2,k})$. Furthermore, $t_0 = 0 \in \text{Range}(W_{2,k})$. From Step (1e) we directly obtain $t_i \in \text{Range}(W_{2,k})$. \hfill \square

Lemma 3.6.19 implies that Alg. 3.6.2 effectively sees problem (3.24) as

\begin{align}
\min \quad & \frac{1}{2} \langle H_k t, t \rangle_{\mathcal{X}} + \langle g_k, t \rangle_{\mathcal{X}} + q_k^X(n_k) \quad (3.27a) \\
\text{s.t.} \quad & t \in \text{Range}(W_{2,k}) \quad (3.27b) \\
& \|t\|_{\mathcal{X}} \leq \Delta_k - \|n_k\|_{\mathcal{X}} \quad (3.27c)
\end{align}

Additionally, for every iterate $t_i$ of Alg. 3.6.2, there exists a $w_i \in Z$, such that $t_i = W_{2,k}w_i$. Therefore, due to Lemma 3.6.7, we have that $\tilde{q}_k(w_i) < \tilde{q}_k(w_{i+1})$ with

$$ \tilde{q}_k(w) = \frac{1}{2} \langle W_{2,k}^*H_kW_{2,k}w, w \rangle_Z + \langle W_{2,k}^*g_k, w \rangle_Z + q_k^X(n_k), $$

which enables us to define the inexact reduced gradient and Hessian, respectively, as follows,

$$ \tilde{W}_k^*g_k = W_{2,k}^*g_k, \quad \tilde{W}_k^*H_k \tilde{W}_k = W_{2,k}^*H_kW_{2,k}. $$
The global convergence requirements for the tangential step are therefore

\[ \| W_{2,k}^* g_k - W_k^* g_k \|_Z \leq \xi_1 \min \{ \| W_{2,k}^* g_k \|_Z, \Delta_k \}, \]  
(3.28)

for some \( \xi_1 > 0 \) independent of \( k \),

\[ \langle W_{2,k}^* H_k W_{2,k} w_k, w_k \rangle_z \leq \xi_2 \| w_k \|_Z^2, \]  
(3.29)

for some \( \xi_2 > 0 \) independent of \( k \), and

\[ - \frac{1}{2} \langle W_{2,k}^* H_k W_{2,k} w_k, w_k \rangle_Z - \langle W_{2,k}^* g_k, w_k \rangle_Z \geq \kappa_4 \| W_{2,k}^* g_k \|_Z \min \{ \kappa_5 \| W_{2,k}^* g_k \|_Z, \kappa_6 \Delta_k \}, \]  
(3.30)

for constants \( \kappa_4, \kappa_5, \kappa_6 > 0 \), independent of \( k \). These conditions can be satisfied by carefully controlling the size of \( \| W_{2,k} - W_k \| \) and \( \| W_{2,k} - W_{1,k} \| \). The next chapter presents the analysis for the case in which \( W_k \) is an orthogonal null–space projector.
Chapter 4

Inexact Linear System Solves in an SQP Algorithm with Null–Space Projections

This chapter focuses on the derivation of a concrete SQP algorithm with inexactness control for the case in which the null–space representation $W_k$ is an orthogonal projector. The projections in the tangential step computation are applied via an iterative solution of augmented systems discussed in Section 2.1.1. The algorithm efficiently manages the stopping criteria for iterative linear system solves. The developed stopping criteria are dynamically adjusted based on the current progress toward a KKT point, trade gains in feasibility for gains in optimality and vice versa, can be easily implemented, and are sufficient to guarantee first–order global convergence of the SQP algorithm. As a side benefit, they allow for a rigorous integration of preconditioners for KKT systems.

Section 4.1 specializes the tangential step convergence requirements (3.6)–(3.8) to the projector case. A major contribution in our work is in the derivation of practical mechanisms that are sufficient for (3.6)–(3.8). Section 4.2 specializes the quasi–normal step convergence requirements (3.1)–(3.2) to the projector case. Section 4.3 addresses the boundedness of the sequence of Lagrange multipliers. Section 4.4 interprets the convergence requirements (3.11)–(3.14). The trust–region SQP algorithm with inexactness control in augmented system solves is stated in Section 4.5.
4.1 Tangential Step

This section discusses the special case in which $Z = X$ and the exact null-space operator $W_k$ is an orthogonal projector, satisfying

$$W_k^* = W_k = W_k W_k.$$

(4.1)

This case is motivated by the full-space CG approach in [19, 47], which allows the use of KKT preconditioners.

As in the previous section, we formulate an inexact algorithm that in exact arithmetic reduces to Alg. 3.6.1. In this particular case, due to property (4.1), the resulting Alg. 4.1.1 is a trivial modification of Alg. 3.6.2.

**Algorithm 4.1.1.** *(Projected inexact CG with full orthogonalization.)*

0. Let $t_0 = 0 \in X$. Let $r_0 = g_k$. Set $i_{max}$, set $i = 0$.

1. While ($r_i \neq 0$ and $W_k(r_i) \neq 0$ and $i < i_{max}$)
   (a) $z_i = W_k(r_i)$
   (b) $p_i = -z_i + \sum_{j=0}^{i-1} \frac{(z_i, H_k p_j)_X}{(p_j, H_k p_j)_X} p_j$
   (c) If $(p_i, H_k p_i)_X \leq 0$ or $(p_i, r_i)_X \leq 0$, extend $t_i$ to boundary of trust–region and stop.
   (d) $\alpha_i = -\frac{(r_i, p_i)_X}{(p_i, H_k p_i)_X}$
   (e) $t_{i+1} = t_i + \alpha_i p_i$
   (f) If $\|t_{i+1}\|_X \geq \Delta_k$, extend $t_i$ to boundary of trust–region and stop.
   (g) $r_{i+1} = H_k t_{i+1} + g_k = r_i + \alpha_i H_k p_i$
   (h) $i \rightarrow i + 1$

\[ \Box \]

If $W_k$ is an orthogonal null-space projector whose application involves iterative linear system solves, then following the augmented system approach, every projection $z_i$ in Step (1a) of Alg. 4.1.1 satisfies

$$
\begin{pmatrix}
I & c_x(x_k)^* \\
c_x(x_k) & 0
\end{pmatrix}
\begin{pmatrix}
z_i \\
y
\end{pmatrix}
= 
\begin{pmatrix}
r_i \\
0
\end{pmatrix}
+ 
\begin{pmatrix}
e^1_i \\
e^2_i
\end{pmatrix}.
$$

(4.2)
Simply put, the main goal of this section is to develop a practical mechanism that controls global convergence requirements for the tangential step computation by controlling the size of \( \|e_i\|_{\mathcal{X} \times \mathcal{Y}} \).

First, we state a few necessary remarks. We can immediately conclude that Alg. 4.1.1 will satisfy Lemma 3.6.6 (conjugacy property), Lemma 3.6.7 (minimization property), and Lemma 3.6.10 (linear independence of residual vectors). The latter gives us the following existence theorem.

**Theorem 4.1.2.** Let \( \mathcal{X} \) be a separable Hilbert space, and let \( \mathcal{W}_k : \mathcal{X} \rightarrow \mathcal{X} \) be an arbitrary nonlinear operator. Then there exists a fixed linear operator \( \widetilde{W}_k \) such that

\[
\mathcal{W}_k(r_i) = \widetilde{W}_k r_i,
\]

for every iteration \( i \) of Alg. 4.1.1.

Analogous to Lemma 3.6.19, every iterate \( t_i \) of Alg. 4.1.1 satisfies

\[
t_i \in \text{Range}(\widetilde{W}_k),
\]

thus Alg. 4.1.1 effectively sees problem (3.24) as

\[
\begin{align}
\min & \quad \frac{1}{2} \langle H_k t, t \rangle_{\mathcal{X}} + \langle g_k, t \rangle_{\mathcal{X}} + q_k^X(n_k) \\
\text{s.t.} & \quad t \in \text{Range}(\widetilde{W}_k) \\
& \quad \|t\|_{\mathcal{X}} \leq \Delta_k - \|n_k\|_{\mathcal{X}}.
\end{align}
\]

Therefore, we can replace the exact reduced-space quadratic model with

\[
\tilde{q}_k(w) = \frac{1}{2} \left\langle \widetilde{W}_k^* H_k \widetilde{W}_k w, w \right\rangle_{\mathcal{X}} + \left\langle \widetilde{W}_k^* g_k, w \right\rangle_{\mathcal{X}} + q_k^X(n_k).
\]
Consequently, the inexact reduced gradient and Hessian are given by

\[ \widetilde{W}_k^* g_k = \widetilde{W}_k^* g_k, \quad \text{and} \quad \widetilde{W}_k^* H_k \widetilde{W}_k = \widetilde{W}_k^* H_k \widetilde{W}_k. \]

It should be noted that \( \widetilde{W}_k \) is, in general, neither self-adjoint nor a projector. Finally, for completeness, we restate the SQP convergence requirements,

\[ \| \widetilde{W}_k^* g_k - W_k g_k \| \leq \xi_1 \min \left\{ \| \widetilde{W}_k^* g_k \|, \Delta_k \right\}, \quad (4.6) \]

for some \( \xi_1 > 0 \) independent of \( k \),

\[ \left\langle \widetilde{W}_k^* H_k \widetilde{W}_k w_k, w_k \right\rangle_{\mathcal{H}} \leq \xi_2 \| w_k \|^2, \quad (4.7) \]

for some \( \xi_2 > 0 \) independent of \( k \), and

\[ -\frac{1}{2} \left\langle \widetilde{W}_k^* H_k \widetilde{W}_k w_k, w_k \right\rangle_{\mathcal{H}} - \left\langle \widetilde{W}_k^* g_k, w_k \right\rangle_{\mathcal{H}} \geq \kappa_4 \| \widetilde{W}_k^* g_k \| \min \left\{ \kappa_5 \| \widetilde{W}_k^* g_k \|, \kappa_6 \Delta_k \right\}, \quad (4.8) \]

for constants \( \kappa_4, \kappa_5, \kappa_6 > 0 \), independent of \( k \).

### 4.1.1 Satisfaction of General Convergence Requirements for Inexactly Computed Tangential Steps

In order to analyze the inexactness in the application of the null-space operator, we seek an explicit representation of \( \widetilde{W}_k \). Suppose that \( \mathcal{W}_k \) is applied \( m \) times in Alg. 4.1.1. As vectors \( r_i, i = 0, 1, \ldots, m \), are linearly independent, the matrix

\[ R_m = [r_0, r_1, \ldots, r_m] \quad (4.9) \]
has full column rank. Furthermore, we introduce matrices

\[ Y_m = [W_k r_0, W_k r_1, ..., W_k r_m], \quad \widetilde{Y}_m = [\mathcal{W}_k(r_0), \mathcal{W}_k(r_1), ..., \mathcal{W}_k(r_m)]. \]  

(4.10)

We have shown that there exists a linear operator \( \widetilde{W}_k \) such that

\[ \widetilde{W}_k R_m = \widetilde{Y}_m. \]

Any linear operator that satisfies this condition could be chosen as the explicit representation of \( \widetilde{W}_k \). For example, one possible choice, involving the pseudo inverse of \( R_m \), is as follows,

\[ \widetilde{W}_k = W_k + (\widetilde{Y}_m - Y_m)(R_m^* R_m)^{-1} R_m^*. \]

In [16], Brown uses this idea to analyze the local convergence of inexact–Newton/finite-difference methods. In our case, the expression

\[ \widetilde{W}_k = W_k + (\widetilde{Y}_m - Y_m)(\widetilde{Y}_m^* R_m)^{-1} \widetilde{Y}_m^* \]  

(4.11)

is analytically more convenient, provided that the inverse of \( \widetilde{Y}_m^* R_m \) exists. After establishing sufficient conditions for the invertibility of \( \widetilde{Y}_m^* R_m \), the explicit representation (4.11) will be used for the remainder of this section.

We briefly address the use (4.11) as the chosen explicit representation of \( \widetilde{W}_k \). To simplify the notation, we introduce the diagonal matrix \( D_m \in \mathbb{R}^{m+1 \times m+1} \), whose \( i \)-th diagonal entry is given by

\[ d_i = \frac{1}{\|\mathcal{W}_k(r_i)\|}, \]

for \( i = 0, 1, ..., m \). The advantage of representation (4.11) is that if exact linear system solves are used in Alg. 4.1.1, then the matrix \( \widetilde{Y}_m^* R_m \) simplifies to

\[ \widetilde{Y}_m^* R_m = D_m^2, \]
i.e. it is an invertible diagonal matrix. This is due to \( W_k^* = W_k = W_k W_k \) and Lemma 3.6.12, which shows the orthogonality of projected residual vectors \( W_k r_0, ..., W_k r_i \). With inexact linear system solves, we hope to control, in theory and in practice, the loss of orthogonality between the projected residual vectors by monitoring how much \( \widetilde{Y}_m^* R_m \) deviates from a diagonal matrix.

The following three lemmas are necessary to prove the main results.

**Lemma 4.1.3.** Let \( g_k \) and \( H_k \) be fixed. Suppose that Alg. 4.1.1 terminates after \( m \) applications of the inexact operator \( W_k \), on the residual vectors given by the first \( m \) columns of \( R_m \). Then there exists a matrix \( T_m \in \mathbb{R}^{m+1 \times m+1} \) such that

\[
\widetilde{Y}_m^* R_m = D_m (I_m + T_m) D_m.
\]

If \( \| T_m \| < 1 \), then \( \widetilde{Y}_m^* R_m \) is invertible. Moreover, if \( \| T_m \| < \frac{1}{2} \), there exists a matrix \( S_m \in \mathbb{R}^{m+1 \times m+1} \) such that

\[
(\widetilde{Y}_m^* R_m)^{-1} = D_m^{-1} (I_m + S_m) D_m^{-1},
\]

where \( \| S_m \| \leq 1 \).

**Proof.** We can write

\[
\widetilde{Y}_m^* R_m = D_m^2 + L_m,
\]

where \( L_m = \widetilde{Y}_m^* R_m - D_m^2 \) is likely to be nonzero if inexact linear system solves are used in Alg. 4.1.1. Let \( T_m \) be a matrix such that for every entry \( T_m^{ij} \) of \( T_m \) and every entry \( L_m^{ij} \) of \( L_m \)

\[
T_m^{ij} = \frac{L_m^{ij}}{d_i d_j},
\]

where \( d_i, d_j \) represent the \( i \)-th and \( j \)-th diagonal entries of \( D_m \), respectively, as defined above. Then, \( L_m = D_m T_m D_m \) and we have shown \( \widetilde{Y}_m^* R_m = D_m (I_m + T_m) D_m \).
If \( \| T_m \| < 1 \), it is known that the inverse of \( \widetilde{Y}_m^* R_m \) exists and is given by

\[
(\widetilde{Y}_m^* R_m)^{-1} = D_m^{-1}(I_m + T_m)^{-1} D_m^{-1}.
\]

We now assume that \( \| T_m \| < \frac{1}{2} \). Introducing

\[
S_m = -T_m(I_m + T_m)^{-1},
\]

it is easily verified that \( (I_m + T_m)^{-1} = I_m + S_m \). Moreover,

\[
\| S_m \| \leq \| T_m \| (I_m + T_m)^{-1} \| \leq \frac{1}{2} \cdot \frac{1}{1 - \| T_m \|} \leq \frac{1}{2} \cdot \frac{1}{1 - \frac{1}{2}} = 1,
\]

which completes the proof. \( \square \)

**Lemma 4.1.4.** Let \( U_m \) be a matrix, such that

\[
U_m^* = \widetilde{Y}_m^* D_m^{-1} = \begin{bmatrix} \frac{\mathcal{W}_k(r_0)}{\| \mathcal{W}_k(r_0) \|_{\mathcal{X}}}, \frac{\mathcal{W}_k(r_1)}{\| \mathcal{W}_k(r_1) \|_{\mathcal{X}}}, \ldots, \frac{\mathcal{W}_k(r_m)}{\| \mathcal{W}_k(r_m) \|_{\mathcal{X}}} \end{bmatrix}.
\]

Then, under the assumptions of Lemma 4.1.3,

\[
\| \widetilde{W}_k - W_k \| \leq 2 \| (\widetilde{Y}_m - Y_m) D_m^{-1} \| \| U_m \|.
\]

**Proof.** From (4.11), we have

\[
\widetilde{W}_k - W_k = (\widetilde{Y}_m - Y_m)(\widetilde{Y}_m^* R_m)^{-1} \widetilde{Y}_m^*.
\]
Due to Lemma 4.1.3,

\[\|\tilde{W}_k - W_k\| = \|(\tilde{Y}_m - Y_m)D_m^{-1}(I_m + S_m)D_m^{-1}\tilde{Y}_m\|\]
\[= \|(\tilde{Y}_m - Y_m)D_m^{-1}(I_m + S_m)U_m\|\]
\[\leq \|(\tilde{Y}_m - Y_m)D_m^{-1}\|\|U_m\| + \|(\tilde{Y}_m - Y_m)D_m^{-1}\|\|S_m\|\|U_m\|\]
\[\leq 2\|(\tilde{Y}_m - Y_m)D_m^{-1}\|\|U_m\|\].

\[\square\]

**Lemma 4.1.5.** Let \(A_m : \mathbb{R}^{m+1} \to \mathcal{X}\) be an arbitrary linear operator whose action on any \(x \in \mathbb{R}^{m+1}\) is defined by

\[A_m x = \sum_{i=0}^{m} x_i (A_m)_i,\]

where \((A_m)_i \in \mathcal{X}\), for \(i = 0, \ldots, m\). Let \(\| \cdot \|_{\mathcal{X},*}\) denote a norm induced by the inner product norm in \(\mathcal{X}\) and some arbitrary vector norm \(\| \cdot \|_*\) defined on \(\mathbb{R}^{m+1}\), i.e.

\[\|A_m\|_{\mathcal{X},*} = \sup_{x \neq 0} \frac{\|A_m x\|_{\mathcal{X}}}{\|x\|_*}.\]

If \(\|(A_m)_i\|_{\mathcal{X}} \leq \varsigma_1\), for all \(i = 0, \ldots, m\), then there exists a constant \(\varsigma_2 > 0\) dependent on \(m\) only, such that

\[\|A_m\|_{\mathcal{X},*} \leq \varsigma_1 \varsigma_2.\]

Moreover, if \(\| \cdot \|_* = \| \cdot \|_1\), i.e. \(\| \cdot \|_*\) is the standard 1-norm in \(\mathbb{R}^{m+1}\), then

\[\|A_m\|_{\mathcal{X},*} = \|A_m\|_{\mathcal{X},1} \leq \varsigma_1.\]

**Proof.** The proof is a simple consequence of the definition of induced matrix norms.
We have

\[ \|A_m\|_{\lambda,*} = \sup_{x \neq 0} \frac{\|A_m x\|_\lambda}{\|x\|_\lambda} = \sup_{x \neq 0} \frac{\sum_{i=0}^{m} x_i (A_m)_i \|x\|_\lambda}{\|x\|_\lambda} \leq \sup_{x \neq 0} \frac{\sum_{i=0}^{m} x_i \|x\|_\lambda}{\|x\|_\lambda} \leq \sup_{x \neq 0} \frac{\sum_{i=0}^{m} x_i \xi_1}{\|x\|_\lambda} \leq \xi_1 \xi_2, \]

where \( \xi_2 \) depends only on \( m \), due to equivalence of norms in \( \mathbb{R}^{m+1} \). Additionally, if \( \| \cdot \|_\star = \| \cdot \|_1 \), then \( \|A_m\|_{\lambda,*} \leq \xi_1 \). 

We can now state the theorems that introduce conditions on inexactness in the tangential step computation, such that the global convergence of the SQP algorithm is guaranteed. Under the assumptions of Lemma 4.1.3, we have confirmed the existence of the explicit operator \( \widetilde{W}_k \) given by (4.11). Therefore, we can use the notation \( \widetilde{W}_k r \) in place of \( W_k (r) \). Without loss of generality, wherever \( \|A\| \) appears for a linear operator \( A \) described by Lemma 4.1.5, the induced norm \( \| \cdot \|_{\lambda,1} \) is implied (note the 1-norm).

**Theorem 4.1.6.** Suppose that every iteration \( i \) of Alg. 4.1.1 satisfies

\[ \|\widetilde{W}_k r_i - W_k r_i\|_{\lambda} \leq C_1 \min \left\{ \frac{\|\widetilde{W}_k g_k\|_{\lambda}}{\|g_k\|_{\lambda}}, \frac{\Delta_k}{\|g_k\|_{\lambda}} \right\} \|\widetilde{W}_k r_i\|_{\lambda}, \quad (4.12) \]

for a fixed constant \( C_1 > 0 \) independent of \( i \) and \( k \). Further assume that

\[ \|\widetilde{W}_k g_k\|_{\lambda} \leq C_2 \|\widetilde{W}_k^* g_k\|_{\lambda}. \quad (4.13) \]

for \( C_2 > 0 \) independent of \( k \). Then, under the assumptions of Lemma 4.1.3, the convergence requirement (4.6) is satisfied, i.e. there exists \( \xi_1 > 0 \) independent of \( k \) such that

\[ \|\widetilde{W}_k^* g_k - W_k g_k\|_{\lambda} \leq \xi_1 \min \left\{ \|\widetilde{W}_k^* g_k\|_{\lambda}, \Delta_k \right\}. \]
Proof. We have

\[
\| \widetilde{W}_k^* g_k - W_k g_k \|_X \leq \| \widetilde{W}_k^* - W_k \|_X \| g_k \|_X = \| (\widetilde{W}_k - W_k)^* \|_X \| g_k \|_X = \| \widetilde{W}_k - W_k \|_X \| g_k \|_X.
\]

Thus, from Lemma 4.1.4, we obtain

\[
\| \widetilde{W}_k^* g_k - W_k g_k \|_X \leq 2\| U_m \|_X \| (\widetilde{Y}_m - Y_m) D_m^{-1} \|_X \| g_k \|_X.
\]

Since \( \| \widetilde{W}_k r_i \|_X = 1 \), by definition of \( U_m \) and using Lemma 4.1.5 (and the fact \( \| U \| = \| U^* \| )\), we have \( \| U_m \| \leq 1 \). Furthermore,

\[
(\widetilde{Y}_m - Y_m) D_m^{-1} = \left[ \frac{\widetilde{W}_k r_0 - W_k r_0}{\| \widetilde{W}_k r_0 \|_X}, \frac{\widetilde{W}_k r_1 - W_k r_1}{\| \widetilde{W}_k r_1 \|_X}, \ldots, \frac{\widetilde{W}_k r_m - W_k r_m}{\| \widetilde{W}_k r_m \|_X} \right].
\]

The stated assumption (4.12) yields

\[
\frac{\| \widetilde{W}_k r_i - W_k r_i \|_X}{\| \widetilde{W}_k r_i \|_X} \leq C_1 \min \left\{ \frac{\| \widetilde{W}_k g_k \|_X}{\| g_k \|_X}, \frac{\Delta_k}{\| g_k \|_X} \right\},
\]

for all \( i = 0, \ldots, m \). Using the assumption \( \| \widetilde{W}_k g_k \|_X \leq C_2 \| \widetilde{W}_k^* g_k \|_X \), we have

\[
\frac{\| \widetilde{W}_k r_i - W_k r_i \|_X}{\| \widetilde{W}_k r_i \|_X} \leq C_1 \max \{ 1, C_2 \} \min \left\{ \frac{\| \widetilde{W}_k^* g_k \|_X}{\| g_k \|_X}, \frac{\Delta_k}{\| g_k \|_X} \right\}.
\]

Therefore, due to Lemma 4.1.5,

\[
\| (\widetilde{Y}_m - Y_m) D_m^{-1} \| \leq C_1 \max \{ 1, C_2 \} \min \left\{ \frac{\| \widetilde{W}_k^* g_k \|_X}{\| g_k \|_X}, \frac{\Delta_k}{\| g_k \|_X} \right\}.
\]
We finally obtain
\[
\|\tilde{W}_k^* g_k - W_k g_k\|_X \leq 2C_1 \max\{1, C_2\} \min\left\{ \frac{\|\tilde{W}_k^* g_k\|_X}{\|g_k\|_X}, \frac{\Delta_k}{\|g_k\|_X} \right\} \|g_k\|_X
\]
\[
= \xi_1 \min\left\{ \|\tilde{W}_k^* g_k\|_X, \Delta_k \right\},
\]
with \(\xi_1 = 2C_1 \max\{1, C_2\} > 0\), independent of \(k\). This completes the proof. \(\square\)

**Theorem 4.1.7.** Suppose that every iteration \(i\) of Alg. 4.1.1 satisfies
\[
\frac{\|\tilde{W}_k r_i - W_k r_i\|_X}{\|\tilde{W}_k r_i\|_X} \leq C_3. \tag{4.14}
\]

Then, under the assumptions of Lemma 4.1.3, the convergence requirement (4.7) is satisfied, i.e. there exists \(\xi_2 > 0\) independent of \(k\) such that
\[
\left\langle \tilde{W}_k^* H_k \tilde{W}_k w_k, w_k \right\rangle_X \leq \xi_2 \|w_k\|_X^2.
\]

**Proof.** From Lemma 4.1.4 and \(\|U\| \leq 1\) (see proof of Theorem 4.1.6), we have
\[
\|\tilde{W}_k - W_k\| \leq 2\|\tilde{W}_m - Y_m\| D_m^{-1}.
\]

Moreover, we have defined
\[
(\tilde{Y}_m - Y_m) D_m^{-1} = \left[ \begin{array}{c} \tilde{W}_k r_0 - W_k r_0 \\ \tilde{W}_k r_1 - W_k r_1 \\ \vdots \\ \tilde{W}_k r_m - W_k r_m \\ \end{array} \right] \left[ \begin{array}{c} \|\tilde{W}_k r_0\|_X \\ \|\tilde{W}_k r_1\|_X \\ \vdots \\ \|\tilde{W}_k r_m\|_X \\ \end{array} \right],
\]
thus by (4.14) and Lemma 4.1.5, we obtain
\[
\|\tilde{W}_m - Y_m D_m^{-1}\| \leq C_3.
\]
which implies
\[
\|\tilde{W}_k - W_k\| \leq 2C_3.
\]
From $\|\tilde{W}_k - W_k\| \geq \|\tilde{W}_k\| - \|W_k\|$, we get

$$\|\tilde{W}_k\| \leq 2C_3 + \|W_k\|.$$  

As $W_k$ are bounded, boundedness of inexact operators $\tilde{W}_k$ follows. The latter along with the boundedness of $H_k$ yields the claim of the theorem. \qed

**Theorem 4.1.8.** Suppose that the first step of Alg. 4.1.1 satisfies

$$q_k(0) - q_k(t_1) = -\frac{1}{2} \langle Hkt_1, t_1 \rangle_{X} - \langle g_k, t_1 \rangle_{X} \geq \frac{1}{2} \|\tilde{W}_kg_k\|_X \min \left\{ \|\tilde{W}_kg_k\|_X, \Delta_k \right\}. \tag{4.15}$$

Further assume that

$$\|\tilde{W}_kg_k\|_X \geq C_4 \|\tilde{W}_kg_k^*\|_X,$$  

for $C_4 > 0$ independent of $k$. Then the convergence requirement (4.8) is satisfied, i.e.

$$-\frac{1}{2} \left\langle \tilde{W}_kg_k^* H_k \tilde{W}_kw_k, w_k \right\rangle_X - \left\langle \tilde{W}_kg_k^* g_k, w_k \right\rangle_X \geq \kappa_4 \|\tilde{W}_kg_k^*\|_X \min \left\{ \kappa_5 \|\tilde{W}_kg_k^*\|_X, \kappa_6 \Delta_k \right\},$$

for constants $\kappa_4, \kappa_5, \kappa_6 > 0$, independent of $k$.

**Proof.** Assumptions (4.15) and (4.16) immediately yield

$$q_k(0) - q_k(t_1) \geq \frac{1}{2} C_4 \|\tilde{W}_kg_k^*\|_X \min \left\{ C_4 \|\tilde{W}_kg_k^*\|_X, \Delta_k \right\}.$$

Due to Remark 3.6.11 which shows monotone decrease of the inexact quadratic functional, we know $-q_k(t_i) \geq -q_k(t_1)$, for $i > 2$. On the other hand, assuming that Alg. 4.1.1 terminates after $m$ iterations (without loss of generality),

$$-\tilde{q}_k(w_k) = -q_k(t_m) \geq -q_k(t_1).$$
\[ \tilde{q}_k(0) - \tilde{q}_k(w_k) = -\frac{1}{2} \langle \tilde{W}_k^* H_k \tilde{W}_k w_k, w_k \rangle_X - \langle \tilde{W}_k^* g_k, w_k \rangle_X \geq \frac{1}{2} C_4 \| \tilde{W}_k^* g_k \| \min \left\{ C_4 \| \tilde{W}_k^* g_k \|_X, \Delta_k \right\}. \]

4.1.2 Practical Convergence Requirements for Inexactly Computed Tangential Steps

In the previous sections we established theoretical convergence requirements that must be satisfied in the computation of the tangential step with inexact linear system solves. In this section we reexamine these requirements from a practical standpoint.

Requirements of Lemma 4.1.3

In Lemma 4.1.3, we showed the existence of

\[ (\tilde{Y}_m^* R_m)^{-1} = D_m^{-1}(I_m + S_m)D_m^{-1}, \]

with \( \| S_m \| \leq 1 \), under the assumption that \( \| T_m \| < \frac{1}{2} \), where

\[ \tilde{Y}_m^* R_m = D_m(I_m + T_m)D_m. \]

Let \( T_i, S_i, I_i, \tilde{Y}_i, R_i, \) and \( D_i \) be the analogues of \( T_m, S_m, I_m, \tilde{Y}_m, R_m, \) and \( D_m \), respectively, at every iteration \( i \) of Alg. 4.1.1. In practice, one option is to monitor the norm of

\[ T_i = D_i^{-1}\tilde{Y}_i^* R_i D_i^{-1} - I_i, \]

and terminate Alg. 4.1.1 if \( \| T_i \| \geq \frac{1}{2} \). We should note that this computation is relatively inexpensive, as \( T_i \) is of size at most \( i_{\text{max}} \times i_{\text{max}} \), where \( i_{\text{max}} \) is the maximum
allowed number of iterations in Alg. 4.1.1, usually less than 100 (we remind the reader that in comparison, the size of linear systems that need to be solved in order to apply the null-space operator can easily be on the order of $10^6$). Any convenient matrix norm can be used here. This stopping criterion is intuitive, as $T_i$ provides a good measure of the loss of orthogonality between the projected residuals, however, it can be somewhat stringent. Another option is to monitor the norm of $S_i$, where

$$S_i = D_i \tilde{Y}_i^* R_i)^{-1} D_i - I_i,$$

with respect to a preset uniform bound, which can be on the order of 10, for example. This is sufficient, as the proof of Theorem 4.1.6 only requires boundedness of $S_m$ (also see proof of Lemma 4.1.4).

**Condition (4.12)**

Here we devise a practical approach for ensuring that

$$\|\tilde{W}_k r_i - W_k r_i\|_{\mathcal{X}} \leq C_1 \min \left\{ \frac{\|\tilde{W}_k g_k\|_{\mathcal{X}}}{\|g_k\|_{\mathcal{X}}}, \frac{\Delta_k}{\|g_k\|_{\mathcal{X}}} \right\} \|\tilde{W}_k r_i\|_{\mathcal{X}},$$

at every iteration $i$ of Alg. 4.1.1. We focus on the case when $\tilde{W}_k$ is applied through augmented system solves, as in (4.2). Let $G_k : \mathcal{X} \times \mathcal{Y} \to \mathcal{X} \times \mathcal{Y}$ be the composite linear operator

$$G_k = \begin{pmatrix} I & c_\varepsilon(x_k)^* \\ c_\varepsilon(x_k) & 0 \end{pmatrix},$$

and let $P : \mathcal{X} \to \mathcal{X} \times \mathcal{Y}$ be a fixed identity prolongation operator such that

$$P r_i = \begin{pmatrix} r_i \\ 0 \end{pmatrix}, \quad P^T \begin{pmatrix} z_i \\ y \end{pmatrix} = z_i.$$
Then, with exact linear system solves, every projection $z_i$ in Step (1a) of Alg. 4.1.1 is computed as follows,

$$z_i = W_k r_i = P^T G^{-1}_k P r_i.$$ 

A generic iterative linear system solver does not solve the system $G_k v_i = P r_i$ exactly. Instead, assuming that it uses an invertible left preconditioner $M_k : X \times Y \rightarrow X \times Y$ to accelerate its convergence, it returns a solution vector $v_i$ satisfying

$$M_k^{-1}G_k v_i = M_k^{-1} P r_i + e_i,$$

where $e_i \in X \times Y$ is known as the preconditioned residual vector. Solving for $v_i$ above gives

$$z_i = \tilde{W}_k r_i = P^T G_k^{-1}(Pr_i + M_k e_i),$$

which implies

$$\tilde{W}_k r_i - W_k r_i = P^T G_k^{-1} M_k e_i.$$ (4.17)

Therefore, the iterative augmented system solve can be stopped at iteration $\ell_i$ if

$$\|e_i^{(\ell_i)}\|_{X \times Y} \leq \min \left\{ \frac{\|\tilde{W}_k g_k\|_X}{\|g_k\|_X}, \frac{\Delta_k}{\|g_k\|_X} \right\} \|z_i^{(\ell_i)}\|_X,$$ (4.18)

as in this case,

$$\|\tilde{W}_k r_i - W_k r_i\|_X \leq \|G_k^{-1} M_k\| \min \left\{ \frac{\|\tilde{W}_k g_k\|_X}{\|g_k\|_X}, \frac{\Delta_k}{\|g_k\|_X} \right\} \|\tilde{W}_k r_i\|_X,$$

and we can assume that $G_k^{-1}$ and $M_k$ are bounded (boundedness of $G_k^{-1}$ follows from the boundedness properties of $c_x(x)$ and $(c_x(x)c_x(x)^*)^{-1}$; boundedness of $M_k$ is an explicit, although very reasonable assumption).

In the first iteration of Alg. 4.1.1, we have $z_0 = \tilde{W}_k g_k$, thus we monitor the preconditioned residual at every iteration of the iterative augmented system solve,
and stop the solve at iteration $\ell_0$ if
\[
\|e_0^{(\ell_0)}\|_{X \times Y} \leq \min \left\{ \frac{\|z_0^{(\ell_0)}\|_X}{\|g_k\|_X}, \frac{\Delta_k}{\|g_k\|_X} \right\} \|z_0^{(\ell_0)}\|_X.
\]

(4.19)

We should note that from the point of view of the iterative linear system solver, this is a fairly intrusive stopping condition. Nonetheless, many popular iterative solver libraries, in particular the Trilinos' [66] parallel Krylov solver package AztecOO [56], as well as PETSc's [2] parallel solvers allow for an efficient implementation of such stopping criteria. An alternative to (4.19) would be to perform the augmented solve with a fixed high accuracy in the first iteration of Alg. 4.1.1.

For subsequent iterations of Alg. 4.1.1, we use the value $\|z_0^{(\ell_0)}\|$ and terminate the augmented system solve at iteration $\ell_i$ if
\[
\|e_i^{(\ell_i)}\|_{X \times Y} \leq \min \left\{ \frac{\|z_0^{(\ell_0)}\|_X}{\|g_k\|_X}, \frac{\Delta_k}{\|g_k\|_X} \right\} \|z_i^{(\ell_i)}\|_X.
\]

(4.20)

In practice, the convergence behavior of Alg. 4.1.1 as well as that of the SQP algorithm seems to be unchanged if at every iteration $i$ of Alg. 4.1.1 the already available value $\|z_{i-1}^{(\ell_{i-1})}\|_X$ is used in place of $\|z_i^{(\ell_i)}\|_X$, when applying criterion (4.20) (see Chapter 5). This amounts to a simple stopping criterion that is nonintrusive with respect to the iterative augmented system solver.

**Conditions (4.13) and (4.16)**

Condition (4.16),
\[
\|\widetilde{W}_k g_k\|_X \geq C_4 \|\widetilde{W}_k g_k\|_X,
\]
follows from the assumption (4.12) of Thm. 4.1.6. We have

\[
\|W_k^* g_k\|_x \leq \|W_k^* g_k - W_k g_k\|_x + \|W_k g_k\|_x \\
\leq \|W_k^* - W_k\|_x \|g_k\|_x + \|W_k g_k\|_x \\
\leq \left( \|W_k^* - W_k\|_x + \|W_k - W_k\| \right) \|g_k\|_x + \|W_k g_k\|_x \\
= 2\|W_k - W_k\|_x \|g_k\|_x + \|W_k g_k\|_x \\
\leq 4C_1 \frac{\|W_k g_k\|_x}{\|g_k\|_x} \|g_k\|_x + \|W_k g_k\|_x,
\]

where the last inequality follows from Lemma 4.1.4 and the proof of Thm. 4.1.6 (ignoring the parts that relate to (4.13)). Therefore, we obtain

\[
\|W_k g_k\|_x \geq \frac{1}{1 + 4C_1} \|W_k^* g_k\|_x.
\]

Similarly, assumption (4.12) of Thm. 4.1.6 motivates a sufficient condition for (4.13),

\[
\|W_k g_k\|_x \leq C_2 \|W_k^* g_k\|_x,
\]

We obtain

\[
\|W_k g_k\|_x \leq \|W_k g_k - W_k^* g_k\|_x + \|W_k^* g_k\|_x \\
\leq 2\|W_k - W_k\|_x \|g_k\|_x + \|W_k^* g_k\|_x \\
\leq 4C_1 \frac{\|W_k g_k\|_x}{\|g_k\|_x} \|g_k\|_x + \|W_k^* g_k\|_x,
\]

thus

\[
\|W_k g_k\|_x \leq \frac{1}{1 - 4C_1} \|W_k^* g_k\|_x.
\]

Unfortunately, as $C_1 < \frac{1}{4}$ is implicit in the last inequality, and the previous section implies $C_1 = \|G_k^{-1} M_k\|$, we would need to change the main termination criterion
(4.20) for the augmented system solver to

\[ \|e^{(i)}\|_{X \times Y} \leq \min \left\{ \frac{\|W_kg_k\|_X}{g_k\|_X}, \frac{\Delta_k}{\|g_k\|_X} \right\} \min \left\{ \frac{\|z^{(i)}\|_X}{4\|G_k^{-1}M_k\|} \right\}. \]

This modification was not necessary for any of the numerical tests. We believe that condition (4.13) is implied in other convergence requirements and/or can be enforced in a more convenient fashion. Intuitively, as we approach the solution of the optimization problem, \(W_kg_k\) tends to zero, and the term \(\frac{1}{4\|G_k^{-1}M_k\|}\) above can be omitted.

**Condition (4.14)**

Explicitly ensuring that

\[ \frac{\|W_kr_i - W_kr_i\|_X}{\|W_kr_i\|_X} \leq C_3, \]

is fairly easy. From (4.17), we get

\[ \frac{\|W_kr_i - W_kr_i\|_X}{\|W_kr_i\|_X} = \frac{\|PTG_k^{-1}M_ke_i\|_X}{\|z_i\|_X} \leq \frac{\|G_k^{-1}\| \|M_k\| \|e_i\|_{X \times Y}}{\|z_i\|_X}, \]

therefore the only necessary requirement is that if the augmented system solve terminates at its \(i\)-th iteration, then the following must be satisfied,

\[ \|e^{(i)}\|_{X \times Y} \leq C^{hess}\|z^{(i)}\|_X, \quad (4.21) \]

where \(C^{hess} > 0\) is some constant independent of \(i\) and \(k\). In practice, choosing \(C^{hess}\) to be a small constant (e.g. \(10^{-3}\)) or tying it to the stopping tolerance of Alg. 4.1.1 proves very effective in improving the local convergence behavior of Alg. 4.1.1 (and consequently that of the SQP algorithm). As before, the already available value \(\|z^{(i-1)}\|_X\) can be used in place of \(\|z^{(i)}\|_X\) when applying criterion (4.21). Additionally, stopping condition (4.21) can be combined with stopping conditions (4.19) and (4.20).
Condition (4.15)

It is easy and computationally inexpensive to verify the condition (4.15),

\[-\frac{1}{2} \langle H_k t_1, t_1 \rangle_X - \langle g_k, t_1 \rangle_X \geq \frac{1}{2} \langle \tilde{W}_k g_k \rangle \min \left\{ \| \tilde{W}_k g_k \|_X, \Delta_k \right\},\]

as all required quantities are readily available (note that \(\tilde{W}_k g_k = z_0\) in Alg. 4.1.1). We should note however that the choice of the constant \(\frac{1}{2}\) in front of \(\| \tilde{W}_k g_k \|\) is not arbitrary. It can be verified that with exact linear system solves, the Cauchy point \(w^C_k = -\omega^C W^*_k g_k\) with

\[
\omega^C = \begin{cases} 
\frac{\|W^*_k g_k\|_X^2}{\langle (W^*_k H_k W_k) W^*_k g_k, W^*_k g_k \rangle_X} & \text{if } \frac{\|W^*_k g_k\|_X^2}{\langle (W^*_k H_k W_k) W^*_k g_k, W^*_k g_k \rangle_X} \leq \Delta_k \text{ and } \\
\frac{\Delta_k}{\|W^*_k g_k\|_X} & \langle (W^*_k H_k W_k) W^*_k g_k, W^*_k g_k \rangle_X > 0,
\end{cases}
\]

satisfies

\[-\frac{1}{2} \langle W^*_k H_k W_k w^C_k, w^C_k \rangle_X - \langle W^*_k g_k, w^C_k \rangle_X \geq \frac{1}{2} \langle W^*_k g_k \rangle \min \{ \| W^*_k g_k \|_X, \Delta_k \},\]

see [22, p.125-127]. In all example problems, condition (4.15) was never violated. If this happens, however, the iterate \(t_1\) should be discarded, and the first iteration of Alg. 4.1.1 should be repeated with increasingly more accurate applications of \(\tilde{W}_k\), until (4.15) holds. It is also possible to multiply the term on the right-hand side of (4.15) by a small constant independent of \(k\), and thereby obtain an even milder condition.

### 4.2 Quasi–Normal Step

The requirements (3.1)

\[\|n_k\|_X \leq \kappa_1 \|c(x_k)\|_Y,\]
and (3.2)

\[ \|c(x_k)\|_Y^2 - \|c(x_k)n_k + c(x_k)\|_Y^2 \geq \kappa_2 \|c(x_k)\|_Y \min \{\kappa_3 \|c(x_k)\|_Y, \Delta_k\}, \]

on the quasi-normal step are rather weak and can be satisfied using many algorithms for the minimization of a quadratic function subject to a spherical constraint. These algorithms can either be simple modifications of their "exact" counterparts (our dog-leg approach is a good candidate), or explicitly inexact schemes, usually based on Krylov subspace methods (for example, see Golub and von Matt [45] and Sorensen [70]).

Here we state a theoretical result related to an inexact variant of the dogleg algorithm 2.1.3.

**Lemma 4.2.1.** Let the inexact quasi-normal step \( n_k \) be computed using a modification of Algorithm 2.1.3, in which the inexact Newton step \( \tilde{n}_k \) is obtained by solving the system

\[
\begin{pmatrix}
I & c_x(x_k)^* \\
c_x(x_k) & 0
\end{pmatrix}
\begin{pmatrix}
z \\
y
\end{pmatrix} =
\begin{pmatrix}
r_k^1 \\
c(x_k) + r_k^2
\end{pmatrix},
\]

and setting \( \tilde{n}_k = c_x(x_k)^*y \), with \( r_k^1 \in \mathcal{X}, \|r_k^1\|_X \leq \epsilon \) and \( r_k^2 \in \mathcal{Y}, \|r_k^2\|_Y \leq \epsilon \) for some \( \epsilon > 0 \) independent of \( k \). Also assume that for every \( k \)

\[ \|c_x(x_k)r_k^1 - r_k^2\|_Y \leq \|c_x(x_k)n_k^{qp} + c(x_k)\|_Y. \]

Then \( n_k \) satisfies

\[ \|n_k\|_X \leq \kappa_1 \|c(x_k)\|_Y \]

and

\[ \|c(x_k)\|_Y^2 - \|c_x(x_k)n_k + c(x_k)\|_Y^2 \geq \|c(x_k)\|_Y^2 - \|c_x(x_k)n_k^{qp} + c(x_k)\|_Y^2, \]

thereby satisfying the requirements (3.1) and (3.2).
Proof. It is easily shown from (4.22) that the inexact Newton step $\tilde{n}_k^N$ is given by

$$\tilde{n}_k^N = n_k^N + c_x(x_k)^*(c_x(x_k)c_x(x_k)^*)^{-1}(c_x(x_k)r_k^1 - r_k^2)$$

$$= -c_x(x_k)^*(c_x(x_k)c_x(x_k)^*)^{-1}c(x_k) + c_x(x_k)^*(c_x(x_k)c_x(x_k)^*)^{-1}(c_x(x_k)r_k^1 - r_k^2).$$

The inexact quasi-normal step $n_k$ then satisfies

$$n_k = n_k^{op} + \tilde{n}_k^N \tilde{n}_k^{np}(n_k^N - n_k^{op}),$$

where $0 \leq \tilde{n}_k^N \leq 1$, $n_k^{op} = \alpha_k^{op}c_x(x_k)^*c(x_k)$, and $|\alpha_k^{op}| \leq \alpha$, where $\alpha$ is independent of $k$ and $\|c(x_k)\|_Y$. This is a trivial consequence of (2.19) and general boundedness assumptions (A4), which imply that there exist positive constants $\nu_1$, $\nu_2$, and $\nu_3$ such that $\|c_x(x_k)c_x(x_k)^*\|_Y \leq \nu_1$, $\|c(x_k)\|_X \leq \nu_2$, and $\|c_x(x_k)^*\|_Y \leq \nu_3$ for all $x \in \Omega$.

Consequently, for any $k$

$$\|n_k\| \leq 2\|n_k^{op}\|_X + \|\tilde{n}_k^N\|_X$$

$$\leq 2\alpha \nu_3\|c(x_k)\|_Y + \nu_3\nu_1\|c(x_k)\|_Y + \nu_3\nu_1\|c_x(x_k)n_k^{op} + c(x_k)\|_Y$$

$$\leq 2\alpha \nu_3\|c(x_k)\|_Y + \nu_3\nu_1\|c(x_k)\|_Y + \nu_3\nu_1(2\nu_2\nu_3\|c(x_k)\|_Y + \|c(x_k)\|_Y)$$

$$= \kappa \|c(x_k)\|_Y,$$

where $\kappa = \nu_3(2\alpha + \nu_1 + \nu_1(2\nu_2\nu_3 + 1))$.

The second part of the proof is only slightly more involved. Using the fact that $c_x(x_k)n_k^N + c(x_k) = 0$, after a simple calculation we obtain

$$\|c(x_k)\|_Y^2 - \|c_x(x_k)n_k + c(x_k)\|_Y^2$$

$$= \|c(x_k)\|_Y^2 - \|(1 - \tilde{n}_k^N)(c_x(x_k)n_k^{op} + c(x_k)) + \tilde{n}_k^N(c_x(x_k)r_k^1 - r_k^2)\|_Y^2.$$ (4.24)

In order to simplify the notation we introduce

$$\nu_k^{op} = c_x(x_k)n_k^{op} + c(x_k), \quad \nu_k^* = c_x(x_k)r_k^1 - r_k^2,$$
where from the assumptions of the lemma \( \|v_k^p\|_Y^2 \leq \|v_k^{op}\|_Y^2 \). After expanding the last term of the expression (4.24), we have

\[
\|c(x_k)\|_Y^2 - \|c_x(x_k)n_{k} + c(x_k)\|_Y^2 \\
= \|c(x_k)\|_Y^2 - \|c_x(x_k)n_{k}^{op} + c(x_k)\|_Y^2 + \\
\left\{ 2\theta_k^N \|v_k^{op}\|_Y^2 - \theta_k^N \|v_k^{op}\|_Y^2 - \theta_k^N \|v_k^r\|_Y^2 - 2\theta_k^N \langle v_k^{op}, v_k^r \rangle_Y + 2\theta_k^N \langle v_k^{op}, v_k^r \rangle_Y \right\} \\
\geq \|c(x_k)\|_Y^2 - \|c_x(x_k)n_{k}^{op} + c(x_k)\|_Y^2 + \left\{ \theta_k^N \|v_k^{op}\|_Y^2 - 2\theta_k^N \langle v_k^{op}, v_k^r \rangle_Y + \right.
\]

\[
\theta_k^N \|v_k^r\|_Y^2 - \theta_k^N \|v_k^{op}\|_Y^2 + 2\theta_k^N \langle v_k^{op}, v_k^r \rangle_Y - \theta_k^N \|v_k^r\|_Y^2 \right\} \\
\geq \|c(x_k)\|_Y^2 - \|c_x(x_k)n_{k}^{op} + c(x_k)\|_Y^2 + \left\{ \theta_k^N \|v_k^{op} - v_k^r\|_Y^2 - \theta_k^N \|v_k^{op} - v_k^r\|_Y^2 \right\} \\
\geq \|c(x_k)\|_Y^2 - \|c_x(x_k)n_{k}^{op} + c(x_k)\|_Y^2,
\]

where we have used \( \theta_k^N \geq \theta_k^N \).

It should be noted that the requirement (4.23) can be easily implemented as a stopping criterion for an iterative linear system solver.

### 4.3 Lagrange Multipliers

The computation of the Lagrange multipliers involves inexact linear system solves. Here we assume that the augmented system (2.12) in Section 2.1.2 is solved iteratively. On the other hand, global convergence proofs for SQP methods are merely based on the boundedness of Lagrange multipliers, as stated in assumption (A7). From the computational point of view, this requirement is quite easily imposed even when inexact linear system solvers are used. The key is usually in the remaining assumptions (A1)-(A6). For example, we can formulate the following lemma.

**Lemma 4.3.1.** Let the sequence \( \lambda_k \) of inexact Lagrange multipliers be generated by
solving the linear system

\[
\begin{pmatrix}
I & c_x(x_k)^* \\
c_x(x_k) & 0
\end{pmatrix}
\begin{pmatrix}
z \\
\lambda_k
\end{pmatrix} =
\begin{pmatrix}
-\nabla_x f(x_k) + r_k^1 \\
r_k^2
\end{pmatrix},
\] (4.25)

with \(r_k^1 \in \mathcal{X}, \|r_k^1\|_\mathcal{X} \leq \epsilon\) and \(r_k^2 \in \mathcal{Y}, \|r_k^2\|_\mathcal{Y} \leq \epsilon\) for some \(\epsilon > 0\) independent of \(k\).

Then the sequence \(\{\lambda_k\}\) of Lagrange multipliers is bounded.

Proof. From (4.25) we directly obtain

\[
\lambda_k = (c_x(x_k)c_x(x_k)^*)^{-1} (-c_x(x_k)\nabla_x f(x_k) + c_x(x_k)r_k^1 - r_k^2) \\
= - (c_x(x_k)c_x(x_k)^*)^{-1} c_x(x_k)\nabla_x f(x_k) + (c_x(x_k)c_x(x_k)^*)^{-1} (c_x(x_k)r_k^1 - r_k^2).
\]

General assumptions (A4) give uniform boundedness of \(\|(c_x(x)c_x(x)^*)^{-1}\|_\mathcal{Y}, \|c_x(x)\|_\mathcal{X}\),

and \(\|\nabla_x f(x)\|_\mathcal{X}\) for all \(x\). The claim follows from elementary norm inequalities and the boundedness of the residual norms \(\|r_k^1\|_\mathcal{X}\) and \(\|r_k^2\|_\mathcal{Y}\). \(\square\)

4.4 Balancing the Tangential and the Quasi-Normal Step

Having implicitly obtained a step \(w_k\) (through the computed step \(t_k\)), we focus on the computation of the effective tangential step \(\tilde{t}_k\).

We should immediately note that the computed step \(t_k = \tilde{W}_k w_k\) need not equal the effective step \(\tilde{t}_k\). In fact, as \(t_k\) (and clearly \(w_k\)) can have a significant component outside of the range of \(W_k\), it makes sense to perform an additional projection of the computed step \(t_k\) (and thus implicitly project \(w_k\)) onto the null space of the constraints, in order to reduce potential loss of linear feasibility.

With exact linear system solves, this would amount to computing the effective tangential step \(\tilde{t}_k = W_k t_k = W_k W_k w_k = W_k w_k = t_k\), i.e. we take the step \(t_k\) as the effective tangential step. No additional projection is necessary. With inexactness, we compute

\[
\tilde{t}_k = \tilde{W}_k t_k = \tilde{W}_k \tilde{W}_k w_k,
\] (4.26)
i.e. an additional projection must be performed, with sufficient accuracy so that the component of the effective tangential step outside of $\text{Null}(c_x(x_k))$ does not significantly violate the progress made by the quasi-normal step.

In particular, we have to make sure that the general condition (3.11),

$$|r_{\text{pred}}(r_k^t; \rho_k)| \leq \eta_{\text{pred}}(n_k, w_k; \rho_k),$$

is satisfied. This is done by performing the aforementioned projection $\tilde{t}_k = \overline{W}_k t_k$ with sufficient accuracy. The requirement (3.11) can always be satisfied, as $\overline{W}_k \rightarrow W_k$ implies $\|r_k^t\|_Y \rightarrow 0$. We also note that $w_k$ is not required explicitly to compute the quantity $\text{pred}(n_k, w_k; \rho_k)$, as $t_k$ is readily available, and $\overline{q}_k(w_k) = q_k(t_k)$.

The practical mechanisms addressing the remaining general conditions imposed on the effective tangential step are discussed below.

**Condition (3.14)**

In order to give a practical mechanism for satisfying

$$\|w_k\|_X \leq \xi_5 \|s_k^x\|_X,$$

we require that

$$\|w_k\|_X \leq C_5 \|\tilde{t}_k\|_X,$$

where $C_5 > 0$ is independent of $k$. We believe that this is a very reasonable assumption, as without inexactness in linear system solves $\tilde{t}_k = t_k = w_k$.

Consequently, we need to ensure that

$$\|\tilde{t}_k\|_X \leq C_6 \|s_k^x\|_X. \quad (4.27)$$

where we recall that the trial step $s_k^x$ is given by $s_k^x = n_k + \tilde{t}_k$. Our first considerations concern the quasi-normal step $n_k$. If the inexact quasi-normal step is computed
using, for example, the procedure given in Section 4.2, then

\[ n_k \in \text{Range}(c_x(x_k)^\top), \]

as \( n_k \) is a linear combination of \( \hat{n}_k^N \) and \( n_k^O \), and \( \hat{n}_k^N, n_k^O \in \text{Range}(c_x(x_k)^\top) \). This means that the inexact quasi-normal step is in fact normal to the null space of the linearized constraints, if computed as in Section 4.2. This implies that we can control the angle between \( n_k \) and \( \tilde{t}_k \) by changing the accuracy of the projection \( \tilde{t}_k = W_k t_k \).

We present two arguments. The first argument is merely stated in order to better illustrate the nature of the requirement (4.27); it does not tell us how to enforce (4.27). Let \( 0 < \varrho < 1 \) (for instance \( \varrho = 0.1 \)), and assume that

\[ \|\tilde{t}_k\|_X \leq (1 - \varrho)\|n_k\|_X \quad \text{or} \quad \|\tilde{t}_k\|_X \geq (1 + \varrho)\|n_k\|_X. \quad (4.28) \]

The first inequality implies \( \|s_k^F\|_X \geq \varrho\|n_k\|_X \), which by another application of the first inequality means that \( \|s_k^F\|_X \geq \frac{\varrho}{1 - \varrho}\|\tilde{t}_k\|_X \), finally yielding

\[ \|\tilde{t}_k\|_X \leq \frac{1 - \varrho}{\varrho}\|s_k^F\|_X. \]

From the second inequality, we obtain \( \|n_k\|_X \leq \frac{1}{1 + \varrho}\|\tilde{t}_k\|_X \), which implies \( \|s_k^F\|_X \geq (1 - \frac{1}{1 + \varrho})\|\tilde{t}_k\|_X \), and therefore

\[ \|\tilde{t}_k\|_X \leq \frac{1 + \varrho}{\varrho}\|s_k^F\|_X. \]

In summary, if (4.28) holds, the relation (4.27) is satisfied.

The second argument is of practical importance and covers all scenarios, in particular the one in which (4.28) is violated. If \( \tilde{t}_k \) is computed exactly, then \( n_k \perp \tilde{t}_k \), and thus \( \|s_k^F\|_X = \sqrt{\|n_k\|_X^2 + \|\tilde{t}_k\|_X^2} \). With inexactness, the angle \( \alpha \) between \( n_k \) and \( \tilde{t}_k \), see Figure 4.1, can be very small, causing \( \|s_k^F\|_X \) to shrink. In this case, we would like to ensure that the acute angle \( \alpha \) between \( n_k \) and \( \tilde{t}_k \) is greater than some fixed
value, in order to avoid excessively small trial steps $s_k^x$.

![Diagram](image)

**Figure 4.1:** Balancing the quasi-normal and the effective tangential step.

Let $0 < \alpha_0 < \frac{\pi}{2}$ be a fixed constant, independent of $k$ (for example $\alpha_0 = 0.1$). In Figure 4.1, let $r \perp \tilde{t}_k$ and let $n_k + z = r$. Without loss of generality, assume $\|z\|_\mathcal{X} \leq \|\tilde{t}_k\|_\mathcal{X}$. We have

$$\|z\|_\mathcal{X} = \frac{1}{\tan \alpha} \|r\|_\mathcal{X}, \quad \|\tilde{t}_k - z\|_\mathcal{X} \leq \|s_k^x\|_\mathcal{X}, \quad \text{and} \quad \|r\|_\mathcal{X} \leq \|s_k^x\|_\mathcal{X},$$

therefore

$$\|\tilde{t}_k\|_\mathcal{X} = \|z\|_\mathcal{X} + \|\tilde{t}_k - z\|_\mathcal{X} \leq \left(\frac{1}{\tan \alpha} + 1\right) \|s_k^x\|_\mathcal{X}.$$ 

In practice, we follow the following procedure:

If $\langle n_k, \tilde{t}_k \rangle < 0$ (acute angle $\alpha$), compute $\alpha$ and ensure that $\alpha > \alpha_0$.  \hspace{1cm} (4.29)
The condition \( \alpha > \alpha_0 \) can be used directly as a stopping criterion for the augmented system solve associated with the application of \( \overline{W}_k \).

**Conditions (3.12) and (3.13)**

The requirement (3.13)

\[
\| \tilde{t}_k \|_{\mathcal{X}} \leq \xi_4 \Delta_k,
\]

follows from the definition \( \tilde{t}_k = \overline{W}_k t_k \), the standard algorithmic assumption \( \| t_k \|_{\mathcal{X}} \leq \Delta_k \) (see Step 1f of Alg. 3.6.2), and the boundedness of \( \overline{W}_k \), which can be ensured through a uniform bound on the norm of the residual in the augmented system solve and general assumptions (A4) and (A5).

The effective tangential step \( \tilde{t}_k \) also needs to satisfy the condition (3.12)

\[
\| \tilde{t}_k - W_k w_k \|_{\mathcal{X}} \leq \xi_3 \Delta_k \| s^e \|_{\mathcal{X}},
\]

which implies only minor modifications of the existing stopping criteria. Using the property \( W_k \overline{W}_k = W_k \), we note

\[
\| \tilde{t}_k - W_k w_k \|_{\mathcal{X}} = \| \overline{W}_k \tilde{t}_k - W_k W_k w_k \|_{\mathcal{X}} = \| \overline{W}_k \tilde{t}_k - W_k \tilde{t}_k - W_k W_k w_k \|_{\mathcal{X}} \leq \| (\overline{W}_k - W_k) \tilde{t}_k \|_{\mathcal{X}} + \| W_k (\overline{W}_k - W_k) w_k \|_{\mathcal{X}}.
\]

The second term in the last expression is easily handled. By imposing the condition

\[
\| e^{(e_i)} \|_{\mathcal{X} \times \mathcal{Y}} \leq \Delta_k \| z^{(e_i)} \|_{\mathcal{X}},
\]

on the \( \ell_i \)-th iteration of the augmented system solve, we can ensure that

\[
\| \overline{W}_k - W_k \| \leq C_7 \Delta_k,
\]
for $C_7 > 0$ independent of $k$. We subsequently use the fact that $W_k$ is bounded and recall relation (3.14).

In order to analyze the first term, we note $\tilde{t}_k = W_k^T W_k w_k = W_k t_k$. The projection is applied by solving a preconditioned augmented system, thus $\tilde{t}_k$ satisfies

$$\tilde{t}_k = P^T G_k^{-1}(P t_k + M_k e_k),$$

where $e_k \in \mathcal{X} \times \mathcal{Y}$ is the corresponding preconditioned residual vector. Analogously to (4.17),

$$W_k t_k - W_k t_k = P^T G_k^{-1} M_k e_k.$$

Therefore, if we ensure that at the last iteration of the augmented system solve

$$\| e_k \|_{\mathcal{X} \times \mathcal{Y}} \leq \Delta_k \| n_k + \tilde{t}_k \| \mathcal{X},$$

we obtain

$$\| (W_k - W_k) W_k w_k \| \mathcal{X} \leq \| G_k^{-1} M_k \| \Delta_k \| s_k^2 \| \mathcal{X}.$$ 

Condition (4.31) can be satisfied directly by monitoring the preconditioned residual vector and the current iterate.

### 4.5 Statement of the Trust–Region SQP Algorithm with Inexact Augmented System Solves

This section focuses on the statement of a complete trust–region SQP algorithm that controls the accuracy in inexact augmented system solves.

We start by reminding the reader that the tangential step computation is based
on projections \( z_i = W_k(r_i) \), such that
\[
\begin{pmatrix}
I & c_x(x_k)^* \\
0 & 0
\end{pmatrix}
\begin{pmatrix}
z_i \\
y
\end{pmatrix}
= \begin{pmatrix}
r_i \\
0
\end{pmatrix} + \begin{pmatrix}
e_i^1 \\
e_i^2
\end{pmatrix}.
\]

Here we collect conditions discussed in Section 4.1.2. Let \( i \) be the running index of the CG iteration, let \( \ell_i \) be the running index of the linear solver iteration. We require the orthogonality condition
\[
\|T_i\| = \|D_i^{-1} \tilde{Y}_i^* R_i D_i^{-1} - I_i\| < \frac{1}{2},
\]
(4.32)

or conditions
\[
\tilde{Y}_i^* R_i \text{ is invertible} \quad \text{and} \quad \|S_i\| = \|D_i(\tilde{Y}_i^* R_i)^{-1} D_i - I_i\| < C_{\text{ortho}}.
\]
(4.33)

Conditions (4.18), (4.21), and (4.30) which must hold at the final linear solver iteration \( \ell_i \) for every CG iteration \( i \), can be combined into
\[
\|e_i^{(\ell_i)}\|_{\mathcal{X} \times \mathcal{Y}} \leq \min \left\{ \frac{\|z_0^{(\ell_i)}\|_{\mathcal{X}}}{\|g_k\|_{\mathcal{X}}} \frac{\Delta_k}{\max\{1, \|g_k\|_{\mathcal{X}}\}}, C_{\text{hess}} \right\} \|z_i^{(\ell_i)}\|_{\mathcal{X}}.
\]
(4.34)

We note that the recycling of \( \|z_i^{(k_{i-1})}\| \) is possible and performs well in practice. Finally, in the first iteration of Alg. 4.1.1, the fraction of Cauchy decrease condition (4.15) has to be satisfied. It can be stated as follows,
\[
-\frac{1}{2} (H_k t_1, t_1)_{\mathcal{X}} - (g_k, t_1)_{\mathcal{X}} \geq \frac{1}{2} \|z_0\|_{\mathcal{X}} \min \{\|z_0\|_{\mathcal{X}}, \Delta_k\}.
\]
(4.35)

For the following algorithm, we choose parameters \( i_{\text{max}} = 100 \), \( tol^{CG} = 10^{-2} \), \( C_{\text{ortho}} = 10 \), and \( C_{\text{hess}} = 10^{-3} \).

**Algorithm 4.5.1.** *(A practical full-space CG for the solution of (3.24))*

0. Let \( t_0 = 0 \in \mathcal{X} \). Let \( r_0 = g_k \). Set \( i_{\text{max}}, tol^{CG}, C_{\text{ortho}}, \) and \( C_{\text{hess}} \).
1. For $i = 0, 1, \ldots, i_{\text{max}}$
   
   (a) Compute $z_i = \mathcal{W}_k(r_i)$ satisfying condition (4.34).
   
   (b) Convergence check: If $\frac{\|z_i\|_X}{\|z_0\|_X} < \text{tol}^{CG}$ or $r_i = 0$ then stop.
   
   (c) Full orthogonalization: $p_i = -z_i + \sum_{j=0}^{i-1} \frac{\langle z_i, H_k p_j \rangle_X}{\langle p_j, H_k p_j \rangle_X} p_j$.
   
   (d) If $\langle p_i, H_k p_j \rangle_X \leq 0$ or $\langle p_i, r_i \rangle_X \leq 0$ or condition (4.32) (alternatively (4.33))
   is violated, extend $t_i$ to the boundary of the trust-region and stop.
   
   (e) Compute $\alpha_i = -\frac{\langle r_i, p_i \rangle_X}{\langle p_i, H_k p_i \rangle_X}$.
   
   (f) Iterate update: $t_{i+1} = t_i + \alpha_i p_i$.
   
   (g) For $i = 0$ only, if condition (4.35) is violated, lower stopping tolerance for
   application of $\mathcal{W}_k$ and repeat Steps 1a–1f until (4.35) is satisfied.
   
   (h) If $\|t_{i+1}\|_X \geq \Delta_k$, extend $t_i$ to boundary of trust-region and stop.
   
   (i) Compute residual $r_{i+1} = H_k t_{i+1} + g_k = r_i + \alpha_i H_k p_i$.

$\square$

We must then ensure that the effective tangential step $\tilde{t}_k$ satisfies (3.11) and (4.31).
Additionally, we need to check if (4.28) is satisfied; if not, we enforce (4.29).

We now have the necessary pieces to formulate the full trust-region SQP algorithm
based on inexact linear system solves. The parameters used in Alg. 4.5.2, Alg. 3.4.1,
and condition (3.11), for all numerical experiments in Chapter 5, are as follows:
$\alpha_1 = 0.5$, $\Delta_0 = 10^2$, $\Delta_{\text{min}} = 10^{-8}$, $\Delta_{\text{max}} = 10^8$, $\eta_1 = 10^{-8}$, $\eta_0 = 0.5$, $\rho_{-1} = 1$,
$\bar{\rho} = 10^{-8}$, and $\text{tol}^{SQP} = 10^{-6}$.

Algorithm 4.5.2. (Trust-region SQP algorithm with inexact linear system solves.)

1. Initialization. Choose initial point $x_0$, initial trust-region radius $\Delta_0$, constants
   $0 < \alpha_1, \eta_1 < 1$, $0 < \eta_0 < 1 - \eta_1$, $\rho_{-1} \geq 1$, and $\text{tol}^{SQP} > 0$. Set $\Delta_{\text{min}}, \Delta_{\text{max}}$ so
   that $0 < \Delta_{\text{min}} < \Delta_{\text{max}}$. Compute an initial Lagrange multiplier estimate $\lambda_0$.

2. For $k = 0, 1, 2, \ldots$
   
   (a) Convergence Check. If
   $\|\nabla_x L(x_k, \lambda_k)\|_X < \text{tol}^{SQP}$ and $\|c(x_k)\|_Y < \text{tol}^{SQP}$,
   
   then terminate.
   
   (b) Step Computation.
i. Compute a quasi-normal step $n_k$ following Section 4.2.
ii. Compute a tentative tangential step $t_k$ via Alg. 4.5.1.

(c) Acceptance Test.

i. Compute a new Lagrange multiplier estimate $\lambda_{k+1}$ based on $n_k + t_k$, following Section 4.3.
ii. Update penalty parameter $\rho_k$ via Alg. 3.4.1.
iii. Compute effective tangential step $\hat{t}_k$ satisfying (3.11), (4.29), and (4.31).
iv. Compute trial step $s_k = n_k + \hat{t}_k$.
v. Compute actual reduction $ared(s_k; \rho_k)$ and predicted reduction $pred(n_k, w_k; \rho_k)$, and their ratio $\theta_k = \frac{ared(s_k; \rho_k)}{pred(n_k, w_k; \rho_k)}$.
vi. If $\theta_k \geq \eta_1$, set $x_{k+1} = x_k + s_k$, and choose $\Delta_{k+1}$ such that

$$\max\{\Delta_{min}, \Delta_k\} \leq \Delta_{k+1} \leq \Delta_{max}.$$  

Otherwise set $x_{k+1} = x_k$, reset $\lambda_{k+1} = \lambda_k$, and set

$$\Delta_{k+1} = \alpha_1 \max\{\|n_k\|, \|\hat{t}_k\|\}.$$  

□
Chapter 5

Numerical Examples

The purpose of this chapter is to test the performance of Alg 4.5.2 on two PDE-constrained optimization problems. The first involves optimal control of Burgers equation in one spatial dimension, and allows us to quickly assess the validity of the devised stopping criteria for iterative linear system solves. The second is a two-dimensional vorticity minimization problem, commonly used as a test case in the control of fluid flows in computational fluid dynamics, and involves Navier–Stokes equations.

5.1 Optimal Control of Burgers Equation in One Dimension

The first set of numerical experiments focuses on a simple optimal control problem involving the steady–state Burgers equation in one dimension. The main goal here is to test the performance of mechanisms for inexactness control in the tangential step computation, on a numerical example that can be modeled rather quickly.

In Section 5.1.1 we discuss a generic problem formulation. Section 5.1.2 describes a problem discretization with piecewise linear finite elements and considers the quantities that need to be computed in order to apply our SQP algorithm. The precise setup of the problem and corresponding numerical results are given in Section 5.1.3.
5.1.1 Problem Formulation

We consider an optimal control problem involving the steady-state Burgers equation in one dimension. In the case of distributed controls the problem is given by

\[
\min \quad \frac{1}{2} \int_0^1 (u(x) - u_d(x))^2 \, dx + \frac{\alpha}{2} \int_0^1 g^2(x) \, dx \tag{5.1a}
\]

subject to

\[
-\nu u_{xx}(x) + u(x)u_x(x) = f(x) + g(x) \quad x \in (0, 1) \tag{5.1b}
\]

\[
u(0) = 0, \quad u(1) = 0. \tag{5.1c}
\]

Here \( u \) and \( g \) play the role of states and controls, respectively, \( \alpha > 0 \) and \( \nu > 0 \) are given parameters, and \( u_d \) and \( f \) are given functions.

The state equations (5.1b)–(5.1c) are interpreted in the standard weak sense, which is introduced next. We define the state and control spaces

\[
U = \{ u \in H^1([0,1]) : u(0) = 0 \text{ and } u(1) = 0 \}, \quad G = L^2([0,1]), \tag{5.2}
\]

respectively, and the space of test functions \( V = U \). The weak form of the state equations is given by

\[
a(u, v) + b(u; u, v) - \langle g, v \rangle = \langle f, v \rangle \quad \forall v \in V, \tag{5.3}
\]
where
\[
a(u, v) = \int_0^1 \nu \frac{d}{dx} u(x) \frac{d}{dx} v(x) dx,
\]
\[
b(u; v, w) = \int_0^1 u(x) \frac{d}{dx} v(x) w(x) dx,
\]
\[
\langle g, v \rangle = \int_0^1 g(x) v(x) dx,
\]
\[
\langle f, v \rangle = \int_\Omega f(x) v(x) dx.
\]

The weak form of the optimal control problem can be stated as follows,
\[
\min \frac{1}{2} \int_0^1 (u(x) - u_d(x))^2 dx + \frac{\alpha}{2} \int_0^1 g^2(x) dx
\]

subject to
\[
a(u, v) + b(u; u, v) - \langle g, v \rangle = \langle f, v \rangle \quad \forall v \in V.
\]

The Lagrangian associated with the optimal control problem is given by
\[
L(u, g, \lambda) = \frac{1}{2} \int_0^1 (u(x) - u_d(x))^2 dx + \frac{\alpha}{2} \int_0^1 g^2(x) dx
+ a(u, \lambda) + b(u; u, \lambda) - \langle g, \lambda \rangle - \langle f, \lambda \rangle.
\]

If \( u \in U \) and \( g \in G \) solve (5.1), then there exists a Lagrange multiplier \( \lambda \in U \) such that the following equations are satisfied:
\[
a(v, \lambda) + b(v; u, \lambda) + b(u; v, \lambda) = -\int_0^1 (u(x) - u_d(x)) v(x) dx \quad \forall v \in V,
\]
\[
\alpha g - \lambda = 0 \quad \text{in } \Omega,
\]
\[
a(u, v) + b(u; u, v) - \langle g, v \rangle = \langle f, v \rangle \quad \forall v \in V.
\]
5.1.2 Problem Discretization

We approximate \( u \) and \( g \) by functions of the form

\[
  u_h(x) = \sum_{j=1}^{n_u} u_j \varphi_j(x) \tag{5.11}
\]

and

\[
  g_h(x) = \sum_{j=1}^{n_g} u_j \psi_j(x) \tag{5.12}
\]

where \( \varphi_j(0) = \varphi_j(1) = 1, j = 1, \ldots, n_u \), and \( \psi_j(0) = \psi_j(1) = 1, j = 1, \ldots, n_g \). We set

\[
  \bar{u} = (u_1, \ldots, u_{n_u})^T \quad \text{and} \quad \bar{g} = (g_1, \ldots, g_{n_g})^T.
\]

If we insert these approximations into (5.1b), we obtain

\[
  A \bar{u} + N(\bar{u}) + B \bar{g} = \bar{f}, \tag{5.13}
\]

where \( A \in \mathbb{R}^{n_u \times n_u} \), \( B \in \mathbb{R}^{n_u \times n_g} \), \( \bar{f} \in \mathbb{R}^{n_u} \), and \( N(\bar{u}) \in \mathbb{R}^{n_u} \) are matrices or vectors given by

\[
  A_{ij} = \nu \int_0^1 \frac{d}{dx} \varphi_j(x) \frac{d}{dx} \varphi_i(x) \, dx,
\]

\[
  B_{ij} = -\int_0^1 \psi_j(x) \varphi_i(x) \, dx,
\]

\[
  (N(\bar{u}))_i = \sum_{j=1}^{n_u} \sum_{k=1}^{n_g} \int_0^1 \frac{d}{dx} \varphi_j(x) \varphi_k(x) \varphi_i(x) \, dx,
\]

\[
  \bar{f}_i = \int_0^1 f(x) \varphi_i(x) \, dx.
\]

If we additionally define the matrices \( M \in \mathbb{R}^{n_u \times n_u} \), \( Q \in \mathbb{R}^{n_g \times n_g} \),

\[
  M_{ij} = \int_0^1 \varphi_j(x) \varphi_i(x) \, dx, \quad Q_{ij} = \int_0^1 \psi_j(x) \psi_i(x) \, dx,
\]
we obtain the discretized optimal control problem

$$
\min_{\bar{u}, \bar{g}} \quad F(\bar{u}, \bar{g}) \equiv \frac{1}{2}(\bar{u} - \bar{u}_d)^T \mathbf{M}(\bar{u} - \bar{u}_d) + \frac{\alpha}{2} \bar{g}^T \mathbf{Q} \bar{g} \\
\text{s.t.} \quad \mathbf{A} \bar{u} + \mathbf{N}(\bar{u}) + \mathbf{B} \bar{g} = \bar{f}.
$$

(5.14a)

(5.14b)

For our numerical implementation, we use piecewise linear functions on equidistant subintervals, i.e., we subdivide $[0, 1]$ into $n_x$ subintervals of length $h = 1/n_x$ and define

$$
\varphi_i(x) = \begin{cases} 
    h^{-1}(x - (i - 1)h) & x \in [(i - 1)h, ih] \cap [0, 1], \\
    h^{-1}(-x + (i + 1)h) & x \in [ih, (i + 1)h] \cap [0, 1], \\
    0 & \text{else},
\end{cases}
$$

where $i = 1, \ldots, n_u = n_x - 1$, and

$$
\psi_i(x) = \begin{cases} 
    h^{-1}(x - (i - 2)h) & x \in [(i - 2)h, (i - 1)h] \cap [0, 1], \\
    h^{-1}(-x + ih) & x \in [(i - 1)h, ih] \cap [0, 1], \\
    0 & \text{else},
\end{cases}
$$

where $i = 1, \ldots, n_g = n_x + 1$. This choice yields

$$
\mathbf{M} = \frac{h}{6}
\begin{pmatrix}
4 & 1 & 0 & 0 & \cdots & 0 \\
1 & 4 & 1 & 0 & \cdots & 0 \\
0 & 1 & 4 & 1 & \cdots & 0 \\
\vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\
0 & 0 & \cdots & 1 & 4 & 1 \\
0 & 0 & \cdots & 1 & 4 & 1 \\
\end{pmatrix} \in \mathbb{R}^{n_u \times n_u}, \quad \mathbf{A} = \frac{\nu}{h}
\begin{pmatrix}
2 & -1 & 0 & 0 & \cdots & 0 \\
-1 & 2 & -1 & 0 & \cdots & 0 \\
0 & -1 & 2 & -1 & \cdots & 0 \\
\vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\
0 & 0 & \cdots & -1 & 2 & -1 \\
0 & 0 & \cdots & -1 & 2 & -1 \\
\end{pmatrix} \in \mathbb{R}^{n_u \times n_u},
$$
\[ B = \frac{h}{6} \begin{pmatrix} 1 & 4 & 1 \\ 1 & 4 & 1 \\ \vdots & \cdots & \vdots \\ 1 & 4 & 1 \\ 1 & 4 & 1 \end{pmatrix} \in \mathbb{R}^{n_u \times n_g}, \quad Q = \frac{h}{6} \begin{pmatrix} 2 & 1 \\ 1 & 4 & 1 \\ \vdots & \cdots & \vdots \\ 1 & 4 & 1 \\ 1 & 2 \end{pmatrix} \in \mathbb{R}^{n_g \times n_g}, \]

and

\[ N(\bar{u}) = \frac{1}{6} \begin{pmatrix} u_1 u_2 + u_2^2 \\ -u_1^2 - u_1 u_2 + u_2 u_3 + u_3^2 \\ \vdots \\ -u_{i-1}^2 - u_{i-1} u_i + u_i u_{i+1} + u_{i+1}^2 \\ \vdots \\ -u_{n_u-2}^2 - u_{n_u-2} u_{n_u-1} + u_{n_u-1} u_{n_u} + u_{n_u}^2 \\ -u_{n_u-1}^2 - u_{n_u-1} u_{n_u} \end{pmatrix} \in \mathbb{R}^{n_u}. \]

The matrix \( N'(\bar{g}) \in \mathbb{R}^{n_u \times n_u} \) is shown in Figure 5.1. For the approximation of the integral arising in the definition of \( \bar{f} \), the composite trapezoidal rule yields

\[ \bar{f}_i = hf(ih), \]

for \( i = 1, \ldots, n_u \).

For a complete implementation of this example within our SQP framework, we need to compute several additional quantities. The discretized Lagrangian functional is given by

\[ \mathcal{L}(\bar{u}, \bar{g}, \bar{\lambda}) = \frac{1}{2} (\bar{u} - \bar{u}_d)^T M (\bar{u} - \bar{u}_d) + \frac{\alpha}{2} \tilde{g}^T Q \tilde{g} + \bar{\lambda}^T (A \bar{u} + N(\bar{u}) + B \bar{g} - \bar{f}), \]

where \( \bar{\lambda} \in \mathbb{R}^{n_u} \) is the discrete representation of the Lagrange multiplier vector. The
Figure 5.1: The matrix $N(u)$.

\[
N(u) = \begin{pmatrix}
-2u_1 - u_2 & u_1 + 2u_2 & -2u_1 - u_2 \\
u_3 - u_1 & u_3 - u_1 & -2u_3 \\
-2u_{n-2} - u_{n-1} & u_{n-1} + 2u_{n-1} & -2u_{n-2} - u_{n-1}
\end{pmatrix}
\]
gradient of the objective function is

$$\nabla_{\bar{u}, \bar{g}} F = \begin{pmatrix} M(\bar{u} - \bar{u}_d) \\ \alpha Q \bar{g} \end{pmatrix},$$

and the Jacobian of the constraints is

$$J(\bar{u}, \bar{g}) = \begin{pmatrix} A + N'(\bar{u}) & B \end{pmatrix}.$$

For the computation of the Hessian of the Lagrangian, by the trilinearity properties of \( b(u; v, w) \) we have for an arbitrary vector \( \bar{s}_u \in \mathbb{R}^n_u \)

$$\nabla^2_{\bar{u}, \bar{g}} (\bar{\lambda}^T N(\bar{u})) \bar{s}_u = N'(\bar{s}_u)^T \bar{\lambda},$$

and so for \( \bar{s} = (\bar{s}_u, \bar{s}_g) \in \mathbb{R}^{n_u + n_g} \)

$$\nabla^2_{\bar{u}, \bar{g}} (\mathcal{L}) \bar{s} = \begin{pmatrix} MS_u + N'(\bar{s}_u)^T \bar{\lambda} \\ \alpha Q \bar{s}_g \end{pmatrix}.$$

The solution of augmented systems needed to compute the Lagrange multipliers, the quasi-normal step, and the tangential step involves the matrix

$$\begin{pmatrix} I_{n_u \times n_u} & 0_{n_u \times n_g} & (A + N'(\bar{u}))^T \\ 0_{n_g \times n_u} & I_{n_g \times n_g} & B^T \\ A + N'(\bar{u}) & B & 0_{n_u \times n_u} \end{pmatrix}.$$

5.1.3 Numerical Results

For our numerical experiments, we choose the number of subintervals \( n_x = 100 \), the viscosity parameter \( \nu = 10^{-2} \), the control penalty parameter \( \alpha = 10^{-5} \), the right-hand side forcing term \( f = 0 \), and the desired state function \( u_d(x) = \sin(2\pi x) \). The code is implemented in Matlab.
As discussed previously, we focus on inexactness control in the computation of the tangential step only. The Lagrange multipliers and the quasi-normal step are computed using the algorithms of Chapter 2 with sparse direct linear system solves. The convergence conditions for the quasi-normal step are thus automatically satisfied.

The tangential step is computed using Alg. 4.5.1, with the option to turn inexactness control on or off. Our goal is to compare the two approaches. In particular, with inexactness control on, the full set of termination criteria and the step acceptance module properly modified for inexact linear system solves are used. The stopping tolerance for the first augmented system solve is computed based on the norm of the true projected residual, for all others, the norm of the projected residual from the previous iteration of Alg 4.5.1 is recycled. With inexactness control off, a fixed preset absolute stopping tolerance is used for all augmented system solves, along with a standard step acceptance scheme, i.e. without the additional null space projection and the rpred modification. In both cases, we have noticed that the full reorthogonalization procedure in Step 1c of Alg 4.5.1 was necessary. If omitted, the CG scheme tends to diverge after only a few iterations.

For augmented system solves, we use GMRES (even though our systems are symmetric), as it was readily available. Any iterative solver that can handle symmetric indefinite systems can be used. It is well known that KKT (and augmented) systems arising from optimal control problems are notoriously ill-conditioned, and usually require $O(n)$ linear solver iterations for convergence, where $n$ is the size of the system. To accelerate convergence of GMRES, we use an incomplete LU preconditioner with a drop tolerance of $5 \times 10^{-3}$.

The first experiment examines the convergence of the SQP algorithm 4.5.2 with inexactness control in the tangential step. Figure 5.2 shows absolute GMRES stopping tolerances for the entire SQP run, for every CG iteration. For every SQP iteration, the symbol □ denotes the controlled tolerance in the first CG iteration. The symbol * denotes the controlled tolerances in all other CG iterations. We make the distinc-
tion for two reasons: one is purely informative (so that one can distinguish between successive SQP iterations), the other is in the fundamental difference in how the first stopping tolerance is computed in Alg 4.5.1.

The desired SQP convergence tolerance ($10^{-6}$ for both the norm of the constraints and the norm of the gradient of the Lagrangian) is achieved in 16 SQP iterations, and 100 CG iterations. The total number of GMRES iterations in all tangential step computations (and the additional null space projections) is 1910. Figure 5.2 shows that when the SQP iterate is far from the optimum, fairly loose GMRES stopping tolerances are allowed, roughly on the order of $10^{-6}$. As the SQP algorithm progresses toward a KKT point, absolute GMRES stopping tolerances are decreased, to approximately $10^{-12}$. This behavior is clearly very encouraging.

![Figure 5.2: Absolute GMRES stopping tolerances – Burgers.](image)

In order to properly gauge the computational cost of the GMRES solves with controlled tolerances, we consider Figure 5.3, which depicts relative stopping tolerances (with respect to the size of the right-hand side vectors in augmented system solves). What is remarkable about Figure 5.3 is that the required relative tolerances are mostly on the order of $10^{-4}$ to $10^{-5}$, and never below $10^{-7}$, which is only slightly smaller than the required SQP stopping tolerances.

The second set of experiments aims to contrast the inexactness control approach
with the current state of the art, which amounts to choosing a fixed linear solver tolerance that “works well”, purely by trial and error. We start with very small stopping tolerances (we could go as low as $5 \times 10^{-16}$ while still converging within the chosen maximum number of GMRES iterations), and raise them by one or two orders of magnitude. The results are presented in Table 5.1.

<table>
<thead>
<tr>
<th></th>
<th>inx. ctrl</th>
<th>5e-16</th>
<th>1e-15</th>
<th>1e-14</th>
<th>1e-13</th>
<th>1e-12</th>
<th>1e-10</th>
<th>1e-8</th>
</tr>
</thead>
<tbody>
<tr>
<td>converges</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>GMRES iter’s</td>
<td>1910</td>
<td>2898</td>
<td>2629</td>
<td>2073</td>
<td>1978</td>
<td>3253</td>
<td>5563</td>
<td>&gt; 10000</td>
</tr>
<tr>
<td>CG iter’s</td>
<td>100</td>
<td>101</td>
<td>101</td>
<td>101</td>
<td>101</td>
<td>199</td>
<td>389</td>
<td>&gt; 500</td>
</tr>
<tr>
<td>SQP iter’s</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>20</td>
<td>37</td>
<td>&gt; 50</td>
</tr>
</tbody>
</table>

**Table 5.1**: Inexactness control vs. fixed stopping tolerances – Burgers.

The first observation is that the inexactness control approach seems to recover similar local convergence behavior as high-precision fixed tolerance cases, see Tables 5.2 and 5.3. This is somewhat unexpected, as the devised stopping tolerances for linear solvers are solely based on global SQP convergence criteria. Second, the inexactness control approach always outperforms fixed tolerance runs in terms of the number of required GMRES iterations. Third, who could have guessed that a fixed
stopping tolerance of $10^{-13}$ would be our best fixed-tolerance option, and that it would yield the same results as $10^{-12}$ in only half the number of GMRES iterations? Finally, how could we have known that a fixed GMRES stopping tolerance of $10^{-8}$ (which is still rather small) would yield no convergence to the desired SQP stopping tolerances?

In summary, the inexactness control strategy performs quite well on this example. Inexactness control tests indicate a globally convergent algorithm, whose local convergence behavior rivals that of an exact scheme. We would also like to note that the incomplete LU preconditioner performed almost too well on this example, with most GMRES solves ending in less than 20 iterations, and the gap between an absolute residual of $10^{-8}$ and that of $10^{-12}$ being closed by only three or four iterations. For a typical GMRES convergence plot, see Figure 5.4. For more complicated prob-

![Figure 5.4: Performance of preconditioned GMRES – Burgers.](image)

lems, and less efficient preconditioners, we expect our inexactness control scheme to outperform the best fixed-tolerance guess by a significantly wider margin.
<table>
<thead>
<tr>
<th>iter</th>
<th>$f(x_k)$</th>
<th>$|c(x_k)|$</th>
<th>$|\nabla_x \mathcal{L}_k|$</th>
<th>$\Delta_k$</th>
<th>$|n_k|$</th>
<th>$|l_k|$</th>
<th>CG iter</th>
<th>accept</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>7.43174e-01</td>
<td>1.480582e+00</td>
<td>6.064963e-02</td>
<td>1.000000e+02</td>
<td>1.000000e+02</td>
<td>1.000000e+02</td>
<td>1</td>
<td>N</td>
</tr>
<tr>
<td>1</td>
<td>7.43174e-01</td>
<td>1.480582e+00</td>
<td>6.064963e-02</td>
<td>5.000000e+01</td>
<td>3.998102e+00</td>
<td>5.000000e+01</td>
<td>1</td>
<td>N</td>
</tr>
<tr>
<td>2</td>
<td>7.43174e-01</td>
<td>1.480582e+00</td>
<td>6.064963e-02</td>
<td>2.500002e+01</td>
<td>3.998102e+00</td>
<td>2.500002e+01</td>
<td>1</td>
<td>Y</td>
</tr>
<tr>
<td>3</td>
<td>1.80268e-01</td>
<td>1.044880e+00</td>
<td>5.137368e-02</td>
<td>5.063544e+01</td>
<td>3.998102e+00</td>
<td>2.500002e+01</td>
<td>1</td>
<td>Y</td>
</tr>
<tr>
<td>4</td>
<td>1.80268e-01</td>
<td>1.044880e+00</td>
<td>5.137368e-02</td>
<td>2.531772e+01</td>
<td>9.999874e+00</td>
<td>2.531772e+01</td>
<td>2</td>
<td>N</td>
</tr>
<tr>
<td>5</td>
<td>1.80268e-01</td>
<td>1.044880e+00</td>
<td>5.137368e-02</td>
<td>1.265887e+01</td>
<td>9.999874e+00</td>
<td>1.265887e+01</td>
<td>2</td>
<td>N</td>
</tr>
<tr>
<td>6</td>
<td>1.80268e-01</td>
<td>1.044880e+00</td>
<td>5.137368e-02</td>
<td>6.329431e+00</td>
<td>9.999874e+00</td>
<td>6.329431e+00</td>
<td>2</td>
<td>N</td>
</tr>
<tr>
<td>7</td>
<td>5.95319e-02</td>
<td>5.060074e-01</td>
<td>1.848318e-02</td>
<td>1.621114e+01</td>
<td>5.063545e+00</td>
<td>5.063545e+00</td>
<td>1</td>
<td>Y</td>
</tr>
<tr>
<td>8</td>
<td>1.56000e-02</td>
<td>1.025080e-01</td>
<td>6.250593e-03</td>
<td>1.271603e+02</td>
<td>8.197468e+00</td>
<td>8.197468e+00</td>
<td>5</td>
<td>Y</td>
</tr>
<tr>
<td>9</td>
<td>1.56000e-02</td>
<td>1.025080e-01</td>
<td>6.250593e-03</td>
<td>6.358013e+01</td>
<td>6.953706e-01</td>
<td>6.953706e-01</td>
<td>5</td>
<td>N</td>
</tr>
<tr>
<td>10</td>
<td>1.56000e-02</td>
<td>1.025080e-01</td>
<td>6.250593e-03</td>
<td>3.179006e+01</td>
<td>6.953706e-01</td>
<td>6.358012e+01</td>
<td>5</td>
<td>N</td>
</tr>
<tr>
<td>11</td>
<td>1.56000e-02</td>
<td>1.025080e-01</td>
<td>6.250593e-03</td>
<td>1.589503e+01</td>
<td>6.953706e-01</td>
<td>3.179006e+01</td>
<td>5</td>
<td>N</td>
</tr>
<tr>
<td>12</td>
<td>5.64170e-03</td>
<td>1.117806e-01</td>
<td>1.513000e-03</td>
<td>1.589503e+01</td>
<td>6.953706e-01</td>
<td>1.589503e+01</td>
<td>5</td>
<td>Y</td>
</tr>
<tr>
<td>13</td>
<td>4.96091e-04</td>
<td>4.968187e-02</td>
<td>4.203988e-04</td>
<td>3.183471e+01</td>
<td>8.429511e+00</td>
<td>8.429511e+00</td>
<td>6</td>
<td>Y</td>
</tr>
<tr>
<td>14</td>
<td>3.70395e-05</td>
<td>1.755739e-03</td>
<td>4.359252e-05</td>
<td>4.325318e+01</td>
<td>5.983391e-01</td>
<td>6.149971e+00</td>
<td>10</td>
<td>Y</td>
</tr>
<tr>
<td>15</td>
<td>2.50740e-05</td>
<td>1.987861e-04</td>
<td>3.317000e-06</td>
<td>4.325318e+01</td>
<td>1.777556e-02</td>
<td>2.335972e+00</td>
<td>18</td>
<td>Y</td>
</tr>
</tbody>
</table>

**Table 5.2:** SQP convergence history with inexactness control – Burgers.
<table>
<thead>
<tr>
<th>iter</th>
<th>$f(x_k)$</th>
<th>$|c(x_k)|$</th>
<th>$|\nabla_x L_k|$</th>
<th>$\Delta_k$</th>
<th>$|n_k|$</th>
<th>$|t_k|$</th>
<th>CG iter</th>
<th>accept</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>7.43174e-01</td>
<td>1.480582e+00</td>
<td>6.064963e-02</td>
<td>1.000000e+02</td>
<td>3.998102e+00</td>
<td>1.000000e+02</td>
<td>1</td>
<td>N</td>
</tr>
<tr>
<td>1</td>
<td>7.43174e-01</td>
<td>1.480582e+00</td>
<td>6.064963e-02</td>
<td>5.000000e+01</td>
<td>3.998102e+00</td>
<td>5.000000e+01</td>
<td>1</td>
<td>N</td>
</tr>
<tr>
<td>2</td>
<td>7.43174e-01</td>
<td>1.480582e+00</td>
<td>6.064963e-02</td>
<td>2.500000e+01</td>
<td>3.998102e+00</td>
<td>2.500000e+01</td>
<td>1</td>
<td>Y</td>
</tr>
<tr>
<td>3</td>
<td>1.80267e-01</td>
<td>1.044879e+00</td>
<td>5.137359e-02</td>
<td>5.063536e+01</td>
<td>3.998102e+00</td>
<td>5.063536e+01</td>
<td>2</td>
<td>N</td>
</tr>
<tr>
<td>4</td>
<td>1.80267e-01</td>
<td>1.044879e+00</td>
<td>5.137359e-02</td>
<td>2.531768e+01</td>
<td>9.999860e+00</td>
<td>2.531768e+01</td>
<td>2</td>
<td>N</td>
</tr>
<tr>
<td>5</td>
<td>1.80267e-01</td>
<td>1.044879e+00</td>
<td>5.137359e-02</td>
<td>1.265884e+01</td>
<td>9.999860e+00</td>
<td>1.265884e+01</td>
<td>2</td>
<td>N</td>
</tr>
<tr>
<td>6</td>
<td>1.80267e-01</td>
<td>1.044879e+00</td>
<td>5.137359e-02</td>
<td>6.329419e+00</td>
<td>9.999860e+00</td>
<td>6.329419e+00</td>
<td>1</td>
<td>Y</td>
</tr>
<tr>
<td>7</td>
<td>5.95339e-02</td>
<td>5.060184e-01</td>
<td>1.848418e-02</td>
<td>1.621122e+01</td>
<td>5.063536e+00</td>
<td>1.621122e+01</td>
<td>5</td>
<td>Y</td>
</tr>
<tr>
<td>8</td>
<td>1.55872e-02</td>
<td>1.023382e-01</td>
<td>6.272712e-03</td>
<td>1.271618e+02</td>
<td>8.197469e+00</td>
<td>8.197469e+00</td>
<td>1</td>
<td>Y</td>
</tr>
<tr>
<td>9</td>
<td>1.55872e-02</td>
<td>1.023382e-01</td>
<td>6.272712e-03</td>
<td>6.358088e+01</td>
<td>6.962314e-01</td>
<td>6.962314e-01</td>
<td>5</td>
<td>N</td>
</tr>
<tr>
<td>10</td>
<td>1.55872e-02</td>
<td>1.023382e-01</td>
<td>6.272712e-03</td>
<td>3.179044e+01</td>
<td>6.962314e-01</td>
<td>3.179044e+01</td>
<td>5</td>
<td>N</td>
</tr>
<tr>
<td>11</td>
<td>1.55872e-02</td>
<td>1.023382e-01</td>
<td>6.272712e-03</td>
<td>1.589522e+01</td>
<td>6.962314e-01</td>
<td>1.589522e+01</td>
<td>5</td>
<td>Y</td>
</tr>
<tr>
<td>12</td>
<td>5.66448e-03</td>
<td>1.121051e-01</td>
<td>1.158361e-03</td>
<td>1.589522e+01</td>
<td>6.962314e-01</td>
<td>1.589522e+01</td>
<td>5</td>
<td>Y</td>
</tr>
<tr>
<td>14</td>
<td>3.96082e-05</td>
<td>1.899466e-03</td>
<td>3.582297e-05</td>
<td>4.091844e+01</td>
<td>5.987695e-01</td>
<td>5.987695e-01</td>
<td>9</td>
<td>Y</td>
</tr>
<tr>
<td>15</td>
<td>2.50772e-05</td>
<td>1.927458e-04</td>
<td>9.868444e-07</td>
<td>4.091844e+01</td>
<td>1.792903e-02</td>
<td>2.445594e+00</td>
<td>17</td>
<td>Y</td>
</tr>
<tr>
<td>16</td>
<td>2.50116e-05</td>
<td>4.725581e-07</td>
<td>1.265484e-08</td>
<td>4.091844e+01</td>
<td>6.361227e-04</td>
<td>1.143115e+00</td>
<td>31</td>
<td>Y</td>
</tr>
</tbody>
</table>

Table 5.3: SQP convergence history with exact solves – Burgers.
5.2 Optimal Control of Navier–Stokes Equations in Two Dimensions

In this section we consider an optimal control problem governed by the steady-state incompressible Navier–Stokes equations. Again, our goal is to test the performance of inexactness control in the tangential step computation. The structure of the problem is similar to that of optimal control of Burgers equation. It is, however, computationally much more demanding and from that point of view closer to a “real-world” application.

In Section 5.2.1 we discuss a generic problem formulation and set up a problem discretization while omitting low-level details. The precise setup of the test problem and corresponding numerical results are given in Section 5.2.2.

5.2.1 Problem Formulation

Let \( \Omega \in \mathbb{R}^2 \) be the channel with a backward facing step shown in Figure 5.5. We

![Figure 5.5: Geometry of the backward-facing step channel.](image-url)
consider an optimal control problem governed by the Navier–Stokes equations

\[-\nu \Delta u + (u \cdot \nabla)u + \nabla p = f \quad \text{in} \quad \Omega, \tag{5.15a}\]
\[\text{div } u = 0 \quad \text{in} \quad \Omega, \tag{5.15b}\]
\[(\nu \nabla u - pI)n = 0 \quad \text{on} \quad \Gamma_{out}, \tag{5.15c}\]
\[u = g \quad \text{on} \quad \Gamma_c, \tag{5.15d}\]
\[u = b \quad \text{on} \quad \Gamma_{in}, \tag{5.15e}\]
\[u = 0 \quad \text{on} \quad \partial \Omega \setminus (\Gamma_{in} \cup \Gamma_c \cup \Gamma_{out}). \tag{5.15f}\]

Later, \( g \) will be the control.

The Dirichlet boundary conditions (5.15d,5.15e,5.15f) can be implemented through interpolation [39], weakly through a Lagrange multiplier technique [50], or via a penalty approach [58, 57]. We use interpolation to implement the Dirichlet boundary conditions (5.15e,5.15f) with fixed data, and replace (5.15d) by the penalized Neumann boundary condition

\[(\nu \nabla u - pI)n + \frac{1}{\delta} u = \frac{1}{\delta} g \quad \text{on} \quad \Gamma_c, \tag{5.16}\]

The state space is related to the function spaces

\[H^1 = H^1(\Omega) = \left\{ v_j \in L^2(\Omega) \mid \frac{\partial v_j}{\partial x_k} \in L^2(\Omega) \text{ for } j, k = 1, \ldots, N \right\},\]
\[H^1_0 = H^1_0(\Omega) = \{ v \in H^1 \mid v = 0 \text{ on } \partial \Omega \setminus (\Gamma_c \cup \Gamma_{out}) \} ,\]

for the velocity vector, and

\[L^2 = L^2(\Omega)\]

for the pressure.

The weak form of the Navier-Stokes equations (5.15a,5.15c,5.15e, 5.15f,5.16) is given as follows. Find \( u \in H^1(\Omega) \) with \( u = b \) on \( \Gamma_{in} \) and \( u = 0 \) on \( \partial \Omega \setminus (\Gamma_c \cup \Gamma_{out}) \)
and \( p \in L^2(\Omega) \), such that

\[
\begin{align*}
\int_{\Omega} \nu \nabla u : \nabla v \, dx &+ \int_{\Omega} (u \cdot \nabla)u \cdot v \, dx - \int_{\Omega} p \nabla \cdot v \, dx \\
&- \int_{\Omega} q \nabla \cdot u \, dx + \frac{1}{\delta} \int_{\Gamma_e} v \cdot (u - g) \, dx \\
&= \int_{\Omega} f \cdot v \, dx
\end{align*}
\]  
(5.17)

for all \( v \in H^1_0(\Omega) \) and all \( q \in L^2(\Omega) \). With the definitions

\[
\begin{align*}
a(u, v) &:= \int_{\Omega} \nu \nabla u : \nabla v \, dx, \quad (5.18) \\
b(u; v, w) &:= \int_{\Omega} ((u \cdot \nabla)v) \cdot w \, dx, \quad (5.19) \\
c(v, q) &:= -\int_{\Omega} \nabla \cdot v \, q \, dx, \quad (5.20) \\
\langle f, v \rangle &:= \int_{\Omega} f \cdot v \, dx, \quad (5.21) \\
h, v)_{\Gamma_e} &:= \int_{\Gamma_e} h \cdot v \, dx \quad (5.22)
\end{align*}
\]

(5.17) is written as

\[
a(u, v) + b(u; u, v) + c(v, p) + c(u, q) + \frac{1}{\delta} \langle u, v \rangle_{\Gamma_e} - \frac{1}{\delta} \langle g, v \rangle_{\Gamma_e} = \langle f, v \rangle \quad (5.23)
\]

for all \( v \in H^1_0(\Omega) \) and all \( q \in L^2(\Omega) \).

We want to find a control \( g \) such that the vorticity in a region \( D = [1, 3] \times [0, 1/2] \) behind the step is reduced. We define

\[
J_1(u) = \frac{1}{2} \int_{D} (\partial_{x_1} u_2 - \partial_{x_2} u_1)^2 \, dx.
\]  
(5.24)
For a given $\alpha > 0$ and $\delta > 0$ we consider the optimal control problem

\begin{align}
\text{Minimize} \quad & J_1(u) + \frac{\alpha}{2} \int_{\Gamma_c} |g|^2 dx, \\
\text{subject to} \quad & a(u, v) + b(u; u, v) + c(v, p) + c(u, q) + \frac{1}{\delta} \langle u, v \rangle_{\Gamma_c} \\
& - \frac{1}{\delta} \langle g, v \rangle_{\Gamma_c} = \langle f, v \rangle \quad \forall v \in H^1_D(\Omega), q \in L^2(\Omega). \tag{5.25b}
\end{align}

The Lagrangian associated with the optimal control problem is given by

\begin{align}
L(u, p, g, \lambda, \theta) &= J_1(u) + \frac{\alpha}{2} \int_{\Gamma_c} |g|^2 dx \\
&+ a(u, \lambda) + b(u; u, \lambda) \\
&+ c(\lambda, p) + c(u, \theta) + \frac{1}{\delta} \langle u, \lambda \rangle_{\Gamma_c} \\
&- \frac{1}{\delta} \langle g, \lambda \rangle_{\Gamma_c} - \langle f, \lambda \rangle. \tag{5.26}
\end{align}

If $u \in H^1(\Omega)$ with $u = b$ on $\Gamma_{in}$ and $u = 0$ on $\partial\Omega \setminus (\Gamma_c \cup \Gamma_{out})$, $p \in L^2(\Omega)$ and $g \in L^2(\Gamma_c)$ solve (5.25), then there exist $\lambda \in H^1_D(\Omega)$ and $\theta \in L^2(\Omega)$ such that the following equations are satisfied:

\begin{align}
a(v, \lambda) + b(v; u, \lambda) + b(u; v, \lambda) \\
+ c(\lambda, q) + c(v, \theta) + \frac{1}{\delta} \langle v, \lambda \rangle_{\Gamma_c} = -D_u J_1(u) v \\
&\quad \forall v \in H^1_D(\Omega), q \in L^2(\Omega), \tag{5.27a}
\end{align}

\begin{align}
\alpha g - \frac{1}{\delta} \lambda &= 0 \quad \text{on } \Gamma_c, \tag{5.27b}
\end{align}

\begin{align}
a(u, v) + b(u; u, v) + c(v, p) + c(u, q) \\
+ \frac{1}{\delta} \langle u, v \rangle_{\Gamma_c} - \frac{1}{\delta} \langle g, v \rangle_{\Gamma_c} = \langle f, v \rangle \\
&\quad \forall v \in H^1_D(\Omega), q \in L^2(\Omega). \tag{5.27c}
\end{align}
For the objective function (5.24), $D_u J_1(u)v$ is given by

$$D_u J_1(u)v = \int_D (\partial_{x_1} u_2 - \partial_{x_2} u_1)(\partial_{x_1} v_2 - \partial_{x_2} v_1)dx.$$ 

We discretize the Navier–Stokes equations using Taylor–Hood finite elements [44, 49, 51, 58]. This discretization of the optimal control problem (5.25) leads to a nonlinear programming problem of the form

Minimize  \[
\frac{1}{2} \bar{u}^T Q \bar{u} + \frac{\alpha}{2} \bar{g}^T R \bar{g},
\]  

subject to  

\[
\begin{pmatrix}
A & B^T \\
B & 0
\end{pmatrix}
\begin{pmatrix}
\bar{u} \\
\bar{p}
\end{pmatrix}
+
\begin{pmatrix}
N(\bar{u}) \\
0
\end{pmatrix}
+
\begin{pmatrix}
M_u \\
0
\end{pmatrix}
\bar{u}
-
\begin{pmatrix}
M_g \\
0
\end{pmatrix}
\bar{g}
=
\begin{pmatrix}
f \\
0
\end{pmatrix}.
\]

(5.28b)

For brevity, we omit the details of the discretization. The quantities required for the application of our SQP algorithm can be computed as in Section 5.1.2, where the discrete representation (5.14) is replaced by (5.28).

### 5.2.2 Numerical Results

As mentioned previously, this test example considers vorticity minimization in fluid flows. The two-dimensional flow is set in a computational domain known as the backward-facing step channel, see Figure 5.5. This is a fairly common setup in the area of computational fluid dynamics.

At the inflow boundary $\Gamma_{in} = \{0\} \times [0.5, 1]$ we assume that the flow is a parabolic function of $x_2$ only with $u(x_2) = 8(x_2 - 0.5)(1 - x_2)$. The viscosity parameter is set to $\nu = 5 \times 10^{-3}$. It is known that for such viscosity levels, the flow separates and recirculation occurs close to the corner region. The uncontrolled flow is depicted in Figure 5.7 with a more detailed snapshot of the corner region in Figure 5.8.
The control is applied on the boundary $\Gamma_c = \{1\} \times [0, 0.5]$. We study the case in which suction/blowing is applied at $\Gamma_c$ with zero tangential velocity, i.e. the case of normal velocity control. This problem has been considered in more detail in [58]. The computational domain is divided into 352 triangles with a finer mesh across the area in which recirculation forms, see Figure 5.6. The parameter penalizing the magnitude of the velocity of the control fluid is set to $\alpha = 0.1$. The penalty parameter for the Dirichlet boundary conditions is taken to be $\delta = 10^{-5}$.

![Figure 5.6: The finite element grid for the backstep channel.](image)

For both uncontrolled and controlled flows, we have been able to reproduce the results of [58], which gives us confidence in our implementation of the Taylor–Hood finite element discretization for the solution of the Navier–Stokes equations, as well as the implementation of the quantities related to the objective function and the control terms. The controlled flow is given in Figure 5.9. The computed vorticity levels in the domain $D$ indicate that normal velocity control on the boundary $\Gamma_c$ can be very effective in eliminating recirculation.

We now focus on the true objective of this study, the inexactness control in the tangential step computation. The Lagrange multipliers and the quasi–normal step are computed using exact linear system solves. In a large–scale setting, exact solves can be replaced with high–accuracy iterative solves. This choice is particularly justified for the Navier–Stokes example, as the cost of the tangential step computation accounts for more than 85% of the total computational cost of the SQP algorithm. The experimental setup resembles that of Section 5.1. The drop tolerance for the incomplete LU preconditioner is lowered to $5 \times 10^{-5}$. 
Figure 5.7: The velocity plot for the uncontrolled fluid flow.

Figure 5.8: Uncontrolled flow: recirculation near the corner region.

The first experiment examines the convergence of the SQP algorithm with inexactness control. Figure 5.10 shows absolute GMRES stopping tolerances for the entire SQP run. The desired SQP convergence tolerance is achieved in 9 SQP iterations, and 162 CG iterations. The total number of GMRES iterations in all tangential step computations (and the additional null space projections) is 2672. As in the case of Burgers equation, when the SQP iterate is far from the optimum, fairly loose GMRES stopping tolerances are allowed, roughly on the order of $10^{-5}$. As the SQP algorithm moves toward a KKT point, absolute GMRES stopping tolerances are automatically
Figure 5.9: The velocity plot for the controlled fluid flow.

decided to approximately $10^{-12}$. At the same time, the required relative tolerances, see Figure 5.11, are mostly on the order of $10^{-4}$ to $10^{-5}$ (and as low as $10^{-3}$), and only twice below $10^{-6}$.

Figure 5.10: Absolute GMRES stopping tolerances – Navier–Stokes.

The second set of experiments compares the inexactness control approach with several choices of fixed linear solver tolerances. We start with the stopping tolerance of $10^{-12}$ and increase it by one order of magnitude to $10^{-8}$. The results are presented in Table 5.4.
Figure 5.11: Relative GMRES stopping tolerances – Navier–Stokes.

<table>
<thead>
<tr>
<th></th>
<th>inx. ctrl</th>
<th>1e-12</th>
<th>1e-11</th>
<th>1e-10</th>
<th>1e-9</th>
<th>1e-8</th>
</tr>
</thead>
<tbody>
<tr>
<td>converges</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
</tr>
<tr>
<td>GMRES iter's</td>
<td>2672</td>
<td>4020</td>
<td>3728</td>
<td>3404</td>
<td>&gt;10000</td>
<td>&gt;10000</td>
</tr>
<tr>
<td>CG iter's</td>
<td>162</td>
<td>142</td>
<td>142</td>
<td>142</td>
<td>&gt;500</td>
<td>&gt;500</td>
</tr>
<tr>
<td>SQP iter's</td>
<td>9</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>&gt;50</td>
<td>&gt;50</td>
</tr>
</tbody>
</table>

Table 5.4: Inexactness control vs. fixed stopping tolerances – Navier–Stokes.

As in Section 5.1, the inexactness control strategy outperforms the best fixed-tolerance guess, this time by about 20%, in terms of the number of GMRES iterations. (This is not unexpected, as the incomplete LU preconditioner is less effective when compared to Section 5.1, see Figure 5.12). Our algorithm requires one additional SQP iteration and 20 additional CG iterations, whose cost is offset by the low tolerance solves throughout the SQP run. Again, fast local convergence in the last few SQP iterations is evident, from Tables 5.5 and 5.6, although not yet supported by a theoretical convergence rate estimate.
Figure 5.12: Performance of preconditioned GMRES – Navier–Stokes.
<table>
<thead>
<tr>
<th>iter</th>
<th>( f(x_k) )</th>
<th>( |c(x_k)| )</th>
<th>( |\nabla_x \mathcal{L}_k| )</th>
<th>( \Delta_k )</th>
<th>( |n_k| )</th>
<th>( |t_k| )</th>
<th>CG iter</th>
<th>accept</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.00000e+00</td>
<td>7.531966e-02</td>
<td>0.000000e+00</td>
<td>1.000000e+02</td>
<td>2.561887e+00</td>
<td>17</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1.12883e-01</td>
<td>2.390632e-02</td>
<td>7.154109e-02</td>
<td>1.000000e+02</td>
<td>3.266380e+00</td>
<td>19</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3.49110e-01</td>
<td>7.949475e-03</td>
<td>1.105058e-01</td>
<td>1.000000e+02</td>
<td>1.468207e+00</td>
<td>19</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3.49110e-01</td>
<td>7.949475e-03</td>
<td>1.105058e-01</td>
<td>1.538637e+00</td>
<td>6.962600e-01</td>
<td>19</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>3.09049e-01</td>
<td>4.470019e-03</td>
<td>1.820603e-01</td>
<td>1.538637e+00</td>
<td>6.962600e-01</td>
<td>19</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1.82559e-01</td>
<td>7.790192e-04</td>
<td>1.389707e-02</td>
<td>5.495695e+00</td>
<td>5.904807e-01</td>
<td>17</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1.80549e-01</td>
<td>2.076801e-05</td>
<td>7.733582e-04</td>
<td>5.495695e+00</td>
<td>1.266288e-01</td>
<td>17</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1.80439e-01</td>
<td>1.489214e-08</td>
<td>4.471905e-06</td>
<td>5.495695e+00</td>
<td>1.871022e-03</td>
<td>19</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1.80439e-01</td>
<td>4.740854e-13</td>
<td>1.714431e-06</td>
<td>5.495695e+00</td>
<td>3.620395e-06</td>
<td>20</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>1.80439e-01</td>
<td>1.417097e-13</td>
<td>1.383309e-08</td>
<td>5.495695e+00</td>
<td>3.646422e-11</td>
<td>15</td>
<td>Y</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.5: SQP convergence history with inexactness control – Navier–Stokes.
<table>
<thead>
<tr>
<th>iter</th>
<th>$f(x_k)$</th>
<th>$|c(x_k)|$</th>
<th>$|\nabla_x C_k|$</th>
<th>$\Delta_k$</th>
<th>$|n_k|$</th>
<th>$|t_k|$</th>
<th>CG iter</th>
<th>accept</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.00000e+00</td>
<td>7.531966e-02</td>
<td>0.000000e+00</td>
<td>1.000000e+02</td>
<td>2.572190e+00</td>
<td>19</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1.12909e-01</td>
<td>2.389013e-02</td>
<td>7.184774e-02</td>
<td>1.000000e+02</td>
<td>3.266380e+00</td>
<td>17</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3.46935e-01</td>
<td>7.848287e-03</td>
<td>1.090567e-01</td>
<td>1.000000e+02</td>
<td>1.467597e+00</td>
<td>19</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3.46935e-01</td>
<td>7.848287e-03</td>
<td>1.090567e-01</td>
<td>1.528168e+00</td>
<td>6.871760e-01</td>
<td>19</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>3.08161e-01</td>
<td>4.413714e-03</td>
<td>1.762772e-01</td>
<td>1.528168e+00</td>
<td>6.871760e-01</td>
<td>6</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1.82320e-01</td>
<td>7.825016e-04</td>
<td>1.404086e-02</td>
<td>5.594572e+00</td>
<td>5.838284e-01</td>
<td>17</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1.80602e-01</td>
<td>2.018716e-05</td>
<td>4.951633e-04</td>
<td>5.594572e+00</td>
<td>1.322191e-01</td>
<td>18</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1.80439e-01</td>
<td>2.627475e-08</td>
<td>3.693730e-06</td>
<td>5.594572e+00</td>
<td>2.200917e-03</td>
<td>19</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1.80439e-01</td>
<td>2.868842e-10</td>
<td>5.904543e-12</td>
<td>5.594572e+00</td>
<td>6.585515e-06</td>
<td>20</td>
<td>Y</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.6: SQP convergence history with exact solves – Navier-Stokes.
Chapter 6

Conclusion

Our work extends the design and the global convergence analysis of a class of trust-region SQP algorithms for smooth nonlinear optimization to allow for an efficient integration of inexact linear system solvers. Our approach does not rely on the estimation of Lipschitz constants and derivative norm bounds, and it does not require a basis–nonbasis decomposition of optimization variables. It extends the range of applicability of SQP algorithms with inexact linear system solves and allows for a rigorous integration of preconditioners for KKT systems in the SQP framework.

The main theoretical contribution is in the careful treatment of the tangential step computation, which often represents the most expensive algorithmic step in SQP methods based on the composite-step trust-region approach. For this purpose, we develop an inexact conjugate gradient algorithm that utilizes a series of approximate projections via Krylov subspace methods, whose accuracy is controlled by the SQP algorithm. The resulting SQP algorithm dynamically adjusts stopping tolerances for iterative linear system solves based on its current progress toward a KKT point. The stopping tolerances can be easily implemented and efficiently computed, and are sufficient to guarantee first-order global convergence of the algorithm.

The devised stopping tolerances for iterative linear system solvers are fully implemented in an SQP algorithm, whose performance is examined on optimal control problems governed by Burgers and Navier–Stokes equations. Global convergence is
observed. Additionally, our inexactness control strategy always outperforms the best possible fixed-tolerance guess, in terms of the total number of Krylov solver iterations, for the same quality of the solution. The local convergence behavior of our approach is similar to that of exact schemes, in which all linear systems are solved using a sparse direct solver.

It is rather interesting that our SQP algorithm exhibits excellent local convergence properties, even though the inexactness framework aims to only satisfy the global convergence requirements. A detailed theoretical investigation of the local behavior is therefore necessary. In addition, it would be interesting to test the performance of the algorithm on important large-scale problems in a parallel computing environment. Preliminary large-scale results obtained using a C++ implementation of a simplified version of the algorithm coupled with parallel Krylov solvers and overlapping KKT domain decomposition preconditioners are very encouraging.
Bibliography


