Trust-Region Interior-Point SQP Algorithms for a Class of Nonlinear Programming Problems

J.E. Dennis
Matthias Heinkenschloss
Luís N. Vicente

December 1994
(revised November 1995 and December 1996)

TR94-45

TRUST–REGION INTERIOR–POINT SQP ALGORITHMS FOR A CLASS OF NONLINEAR PROGRAMMING PROBLEMS*

J. E. DENNIS †, MATTHIAS HEINKENSCHLOSS ‡, AND LUíS N. VICENTE §

Abstract. In this paper a family of trust–region interior–point SQP algorithms for the solution of a class of minimization problems with nonlinear equality constraints and simple bounds on some of the variables is described and analyzed. Such nonlinear programs arise e.g. from the discretization of optimal control problems. The algorithms treat states and controls as independent variables. They are designed to take advantage of the structure of the problem. In particular they do not rely on matrix factorizations of the linearized constraints, but use solutions of the linearized state equation and the adjoint equation. They are well suited for large scale problems arising from optimal control problems governed by partial differential equations.

The algorithms keep strict feasibility with respect to the bound constraints by using an affine scaling method proposed for a different class of problems by Coleman and Li and they exploit trust–region techniques for equality–constrained optimization. Thus, they allow the computation of the steps using a variety of methods, including many iterative techniques.

Global convergence of these algorithms to a first–order KKT limit point is proved under very mild conditions on the trial steps. Under reasonable, but more stringent conditions on the quadratic model and on the trial steps, the sequence of iterates generated by the algorithms is shown to have a limit point satisfying the second–order necessary KKT conditions. The local rate of convergence to a nondegenerate strict local minimizer is q–quadratic. The results given here include as special cases current results for only equality constraints and for only simple bounds.

Numerical results for the solution of an optimal control problem governed by a nonlinear heat equation are reported.

Keywords. Nonlinear programming, SQP methods, trust–region methods, interior–point algorithms, Dikin–Karmarkar ellipsoid, Coleman–Li affine scaling, simple bounds, optimal control problems.

AMS subject classifications. 49M37, 90C06, 90C30

1. Introduction. In this paper we introduce and analyze a family of algorithms for the solution of an important class of minimization problems which often arise from the discretization of optimal control problems. These problems are specially structured nonlinear programming problems of the following form:

$$\begin{array}{ll} \text{minimize} & f(y,u) \\ \text{subject to} & C(y,u)=0, \\ & u \in \mathcal{B} = \{u:\ a \le u \le b\}, \end{array} \tag{1.1}$$

* This document was generated January 23, 1997.

† Department of Computational and Applied Mathematics, Rice University, Houston, Texas 77005–1892, USA, E-Mail: `dennis@rice.edu`. Support of this author has been provided by DOE–FG03–93ER25178, CRPC CCR–9120008, and AFOSR–F49620–9310212.

‡ Department of Computational and Applied Mathematics, Rice University, Houston, Texas 77005–1892, USA, E-Mail: `heinken@rice.edu`. This author was supported by the NSF under Grant DMS–9403699, by the DoE under Grant DE-FG03-95ER25257 and in part by the AFOSR under Grant F49620–93–1–0280.

§ Departamento de Matemática, Universidade de Coimbra, 3000 Coimbra, Portugal. E-Mail: `lvicente@mat.uc.pt`. This work was developed while the author was a graduate student at the Department of Computational and Applied Mathematics of Rice University. Support has been provided by INVOTAN (NATO scholarship), CCLA (Fulbright scholarship), FLAD, (Portugal) and by DOE–FG03–93ER25178.

where $y \in \mathbb{R}^m$, $u \in \mathbb{R}^{n-m}$, $a \in (\mathbb{R} \cup \{-\infty\})^{n-m}$, and $b \in (\mathbb{R} \cup \{+\infty\})^{n-m}$. The functions $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ and $C : \mathbb{R}^n \longrightarrow \mathbb{R}^m$, $m < n$, are assumed to be at least continuously differentiable. As indicated above, minimization problems of the form (1.1) often arise from the discretization of optimal control problems. In this case y is the vector of state variables, u is the vector of control variables, and $C(y, u) = 0$ is the discretized state equation. Other applications, which might be viewed as special optimal control problems include optimal design and parameter identification problems. Minimization problems (1.1) originating from optimal control problems governed by large systems of ordinary differential equations, or partial differential equations are the targets of the algorithms in this paper.

Although there are algorithms available for the solution of nonlinear programming problems that are more general than (1.1), the family of algorithms presented in this paper is unique in the consequent use of structure inherent in many optimal control problems, the use of optimization techniques successfully applied in other contexts of nonlinear programming, and the rigorous theoretical justification.

Our algorithms are based on sequential quadratic programming (SQP) methods and use trust–region interior–point techniques to guarantee global convergence and to handle the bound constraints on the controls. SQP methods find a solution of the nonlinear programming problem (1.1) by solving a sequence of quadratic programming problems. It is known, see e.g. [31], [32], that the structure of optimal control problems can be used to implement and analyze SQP methods. In particular, to implement SQP methods, it is sufficient to compute quantities of the form $C_y(y,u)v_y$, $C_y(y,u)^T v_y$, $C_u(y,u)v_u$, $C_u(y,u)^T v_y$, and to compute solutions of the linearized state equation $C_y(y,u)v_y = r$, and of the "adjoint equation" $C_y(y,u)^T v_y = r$. Here C_y and C_u denotes the derivatives of C with respect to y and u. This is an important observation, because these are tasks that arise naturally in the context of optimal control problems. All of the early SQP algorithms, and many of the recent ones rely on matrix factorizations, like the QR, of the Jacobian $J(x)$ of $C(x)$. For the applications we have in mind this is not feasible. The involved matrices are too large to perform such computations and very often these matrices are not even available in explicit form. On the other hand, matrix–vector multiplications $C_y(x)v_y$, $C_y(x)^T v_y$, $C_u(x)v_u$, $C_u(x)^T v_y$ can be performed and efficient solvers for the linearized state equation $C_y(x)v_y = r$, and the adjoint equation $C_y(x)^T v_y = r$ often are available. For example, the partial Jacobian $C_y(x)$ in the application treated in Section 11 has a block bidiagonal structure with diagonal matrices being tridiagonal. Thus, while the Jacobian is large, the solution of the linearized state equation or the adjoint equation can be done by block forward substitution or block backward substitution, respectively. In each substitution step, only a relatively small system with tridiagonal system has to be solved. This is typical for many applications. A QR factorization of the entire Jacobian $J(x)$ of $C(x)$ is extremely costly and not feasible from a practical point of view, but the partial Jacobian $C_y(x)$ possesses a particular structure which allows efficient solution techniques for $C_y(y,u)v_y = r$ and $C_y(y,u)^T v_y = r$. These solution techniques may involve (LU–) factorizations of $C_y(y,u)^T$ or submatrices thereof.

A purely local analysis for the case with no bounds constraints has being given in [28], [30], [31], [33]. However, we consider here the much more difficult issue of incorporating all this structure into an algorithm that converges globally and handles bound constraints on the control variables u.

The global convergence of our algorithms is guaranteed by a trust–region strategy. In our framework the trust region serves a dual purpose. Besides ensuring global convergence, trust regions also introduce a regularization of the subproblems which is related to the Tikhonov regularization. For the solution of optimal control problems, the partitioning of the variables into states y and controls u motivates a partial decoupling of step components that leads to interesting alternatives for the choice of the trust region. In Sections 5.2.1 and 5.2.2 we will introduce a decoupled and a coupled trust–region approach. As indicated by the names, in the decoupled approach the trust region will act on step components separately. This allows a more efficient implementation of algorithms for the computation of these steps. However, for problems with ill–conditioned state equations, this decoupling does not give an accurate estimate of the size of the steps and might lead to poor performance. In this situation the coupled approach is better, and so we include both.

For the treatment of the bound constraints on u we use an affine scaling interior–point method introduced by Coleman and Li [9] for problems with simple bounds. Interior–point approaches are attractive for problems with a large number of bounds. In our context, the affine scaling interior–point method is also of interest, because it does not interfere with the structure of the problem. To apply this method, no additional information is required from the user. This or similar interior–point approaches have recently also been used e.g. in [3], [10], [35], [36], [41]. The advantage of the approach in [9] is that the scaling matrix is determined by the distance of the iterates to the bounds and by the direction of the gradient. This dependence on the direction of the gradient is important for global convergence and its good effect can be seen in numerical examples, see e.g. Figures 11.1 and 11.2.

Another important issue, that is addressed in the implementations of the algorithms presented in this paper is the problem scaling inherent in optimal control problems. As we have pointed out, the problems we are primarily interested in are discretizations of optimal control problems governed by partial differential equations. The infinite dimensional problem structure greatly influences the finite dimensional problem. In our implementation, we take this into account by choosing scalar products for the states y, the controls u, and the duality pairing needed to represent $\lambda^T C(y,u)$ that are discretizations of proper infinite dimensional ones. It is beyond the scope of this paper to give a comprehensive theoretical study of these issues, but it is important to notice that the formulation of the algorithms discussed here fully support the use of such scalar products without any changes. This is a great advantage. In some of our numerical experiments [7], [24] this improved the performance of our algorithms significantly, it avoided artificial ill–conditioning, and it enhanced the quality of the solution computed for a given stopping tolerance. Moreover, our numerical experiments also indicate the mesh independent behavior of our algorithms when this type of scaling is used.

We believe that the features and strong theoretical properties of these algorithms make them very attractive and powerful tools for the solution of optimal control problems. They have been successfully applied to a boundary control problem, see Section 11, a distributed nonlinear elliptic control problem [25], and optimal control problems arising in fluid flow [7], [24]. The software that produced these results currently is being beta–tested with the intent of electronic distribution [12].

Before we give an outline of this paper, it is worth discussing the relationship between the constrained minimization problem (1.1) and an equivalent reduced problem. Under the assumptions of the Implicit Function Theorem it is possible to solve $C(y,u) = 0$ for y. This defines a smooth function $y(u)$ and allows us to reduce the minimization problem (1.1). The reduced problem is given by

$$\begin{array}{ll} \text{minimize} & \hat{f}(u) \equiv f(y(u), u) \\ \text{subject to} & u \in \mathcal{B} = \{u : \ a \le u \le b\}. \end{array} \tag{1.2}$$

This leads to the so–called *black box* approach in which the nonlinear constraint $C(y, u) = 0$ is not visible to the optimizer. Its solution is part of the evaluation of the objective function $\hat{f}(u)$. The reduced problem can be solved by a gradient or a Newton–like method. For optimal control problems, many algorithms follow this approach. Often, projection techniques are used to handle the box constraints, see e.g. [22], [42].

Recently, so–called *all–at–once* approaches that treat both y and u as independent variables have been proposed to solve optimal control problems, see e.g. [23], [26], [27], [28], [29], [30], [31], [33]. Since they move towards optimality and feasibility at the same time, they offer significant advantages. SQP methods are of particular interest. They do not require the possibly very expensive solution of the nonlinear state equation in every step, but as indicated above allow use of the structure of optimal control problems. In addition, SQP methods have proven to be very successful for the solution of other nonlinear programming problems. See e.g. [2], [18], [34], [39], [41], [47].

As outlined before, we use SQP based methods for the solution of (1.1), i.e., the all–at–once approach. However, the reduced problem (1.2) is important to us for two reasons. Firstly, the relation between the full problem (1.1) and the reduced problem (1.2) gives important insight into the structure of (1.1) and allows us to extend techniques successfully applied to problems of the form (1.2). Secondly, black box approaches are used very often to solve the problems we have in mind. We want to use this expertise in designing more efficient codes. Specifically, our consequent use of the structure of the optimal control problems leads to our family of trust–region interior–point SQP algorithms. These algorithms only require information that the user has to provide anyway if a black–box approach is used with a Newton–like method for the solution of the nonlinear state equation and adjoint equations techniques for the computation of gradients. Thus we combine the possible implementational advantages of a black–box approach with the generally more efficient all–at–once approach. It will be seen that in our algorithms the step s is decomposed into two components: $s = s^{\mathsf{n}} + s^{\mathsf{t}}$, where s^{n} is called the quasi–normal component and s^{t} is called the tangential component. The role of quasi–normal component s^{n} is to move towards feasibility. It is of the form $s^{\mathsf{n}} = ((s_y^{\mathsf{n}})^T \ 0^T)^T$, where s_y^{n} is essentially a Newton step for the solution of the nonlinear state equation $C(y, u) = 0$ for given u. For most problems of interest here, the computation of a "true" normal component is not practical. The tangential component s^{t} moves towards optimality. This component is in the null–space of the linearized constraints and it is of the form $s^{\mathsf{t}} = ((-C_y(y, u)^{-1} C_u(y, u) s_u)^T \ s_u^T)^T$, where s_u is essentially a Newton–like step for the reduced problem (1.2).

This paper is organized as follows: In Section 2 we discuss the structure of the problem and motivate our SQP approach. We study the relationship between the all–at–once approach based on (1.1) and the black box approach for (1.2) and the relationship between SQP methods for (1.1) and Newton methods for (1.2). For problems without box–constraints, these connections are known, but for problems with box–constraints this will reveal useful new information. The first and second order Karush–Kuhn–Tucker (KKT) conditions for (1.1) are stated in Section 3. We will state them in a nonstandard form that will lead to the scaling matrix used in the affine scaling interior–point approach. In Section 4 we will discuss the application of Newton's method to the system of nonlinear equations arising from the first–order KKT conditions. This will be important for the derivation of our SQP method. In Section 5 we describe our trust–region interior–point SQP algorithms. Sections 5.1 and 5.2 contain a description of the quasi–normal component and of the tangential component. Using the derivations in Sections 2 and 4 the connections between the quasi–normal component s^{n} and the Newton step for the solution of the nonlinear state equation $C(y,u) = 0$ for given u and the relations between the tangential component s^{t} and Newton–like steps for the reduced problem (1.2) will be made precise. As noticed previously, the partial decoupling of the step components motivated by the partitioning of the variables into states y and controls u and the roles of the decoupled and coupled trust–region approaches will be exposed in Sections 5.2.1 and 5.2.2. A complete statement of the trust–region interior–point SQP algorithms is given in Section 5.4.

The convergence theory for these algorithms is given in Sections 6, 7, 8, and 9. Section 6 contains some technical results. In Section 7 we establish the existence of an accumulation point of the iterates which satisfies the first–order Karush–Kuhn–Tucker (KKT) conditions (Corollary 7.1). This result is established under very mild assumptions on the steps and on the Lagrange multipliers. It simultaneously extends the results presented recently by Coleman and Li [9] for simple bounds and those by Dennis, El–Alem, and Maciel [11] for equality constraints. Under additional conditions on the steps and on the quadratic model, we show that the accumulation point satisfying the first–order necessary KKT conditions also solves the second–order necessary KKT conditions (Theorem 8.2). This latter result simultaneously extends those by Coleman and Li [9] for simple bounds and those by Dennis and Vicente [14] for equality constraints. (See also [52].) Finally, we prove that if the sequence converges to a nondegenerate point satisfying the sufficient second–order KKT conditions, then the rate of convergence is q–quadratic (Corollary 9.1). Our analysis allows the application of a variety of methods for the computation of the step components s^{n} and s^{t}. In Section 10 we discuss practical algorithms for the computation of trial steps and the multiplier estimates that are currently used in our implementation. Numerical results obtained with our implementation of these algorithms, called TRICE (trust–region interior–point SQP algorithms for optimal control and engineering design problems) [12], are reported in Section 11. Section 12 contains conclusions and a discussion of future work.

We review the notation used in this paper. The vector x is given by

$$x = \begin{pmatrix} y \\ u \end{pmatrix}.$$

The Jacobian matrix of $C(x)$ is denoted by $J(x)$. We use subscripted indices to represent the evaluation of a function at a particular point of the sequences $\{x_k\}$ and $\{\lambda_k\}$. For instance, f_k represents $f(x_k)$, and ℓ_k is the same as $\ell(x_k, \lambda_k)$. The vector and matrix norms used are the ℓ_2 norms, and I_l represents the identity matrix of order l. Also $(z)_y$ and $(z)_u$ represent the subvectors of $z \in \mathbb{R}^n$ corresponding to the y and u components, respectively.

2. The structure of the minimization problem. The purpose of this section is to discuss some of the basic relationships between the problem (1.1) and its reduction (1.2). This will introduce fundamental quantities that are needed subsequently and it will support our claim that the basic quantities needed to implement our SQP approach are already available if one uses a gradient or Newton–like method for the solution of the reduced problem (1.2).

The Lagrange function $\ell : \mathbb{R}^{n+m} \longrightarrow \mathbb{R}^n$ associated with the objective function $f(x)$ and the equality constraint $C(x) = (c_1(x), \ldots, c_m(x))^T = 0$ is given by

$$\ell(x, \lambda) = f(x) + \lambda^T C(x),$$

where $\lambda \in \mathbb{R}^m$ are the Lagrange multipliers.

The linearized constraints are given by $J(x)s = -C(x)$ or equivalently by

$$\begin{pmatrix} C_y(x) & C_u(x) \end{pmatrix} \begin{pmatrix} s_y \\ s_u \end{pmatrix} = -C(x). \tag{2.1}$$

We say that

$$s = \begin{pmatrix} s_y \\ s_u \end{pmatrix}, \quad s_y \in \mathbb{R}^m, \; s_u \in \mathbb{R}^{n-m},$$

satisfies the linearized state equation if it is a solution to (2.1). If $C_y(x)$ is invertible, the solutions of the linearized state equation are of the form

$$s = s^{\mathsf{n}} + W(x)s_u, \tag{2.2}$$

where

$$s^{\mathsf{n}} = \begin{pmatrix} -C_y(x)^{-1}C(x) \\ 0 \end{pmatrix} \tag{2.3}$$

is a particular solution and

$$W(x) = \begin{pmatrix} -C_y(x)^{-1}C_u(x) \\ I_{n-m} \end{pmatrix}$$

is a matrix whose columns form a basis for the null space $\mathcal{N}(J(x))$ of $J(x)$. One can see that matrix–vector multiplications of the form $W(x)^T s$ and $W(x)s_u$ involve only the solution of linear systems with the matrices $C_y(x)$ and $C_y(x)^T$. Moreover, the y component of the particular solution s^{n} is just the step that one would compute if one would apply Newton's method for the solution of the nonlinear equation $C(y, u) = 0$ for given u.

The point we want to convey in this section has nothing to do with the presence or absence of the bound constraints $a \le u \le b$. Therefore, for the remainder of this section, we consider the simpler case where there are no bound constraints, i.e., where $\mathcal{B} = \mathbb{R}^{n-m}$. If we solve (1.1) with $\mathcal{B} = \mathbb{R}^{n-m}$ by an SQP method, then the quadratic programming subproblem we have to solve at every iteration is of the form

$$\begin{aligned} \text{minimize} \quad & \nabla f(x)^T s + \tfrac{1}{2} s^T \nabla^2_{xx}\ell(x, \lambda)\, s \\ \text{subject to} \quad & C_y(x)s_y + C_u(x)s_u + C(x) = 0. \end{aligned} \tag{2.4}$$

If the reduced Hessian $W(x)^T \nabla^2_{xx}\ell(x, \lambda)W(x)$ is nonsingular, the solution of (2.4) is given by (2.2) with

$$s_u = -\Big(W(x)^T \nabla^2_{xx}\ell(x, \lambda)\, W(x)\Big)^{-1} W(x)^T \Big(\nabla f(x) + \nabla^2_{xx}\ell(x, \lambda)s^{\mathsf{n}}\Big). \tag{2.5}$$

In practice the Hessian $\nabla^2_{xx}\ell(x,\lambda)$ or the reduced Hessian $W(x)^T\nabla^2_{xx}\ell(x,\lambda)\,W(x)$ are often approximated using quasi–Newton updates. In the latter case, when an approximation to $\nabla^2_{xx}\ell(x,\lambda)$ is not available, then the "cross–term" $W(x)^T\nabla^2_{xx}\ell(x,\lambda)s^{\mathsf{n}}$ has also to be approximated. This term can be approximated by zero, by finite differences, or by other quasi–Newton approximations, see e.g. [1]. In the case where this cross term is approximated by zero, the right hand side of the linear system (2.5) defining s_u can be written as

$$W(x)^T\nabla f(x) = -C_u(x)^T C_y(x)^{-T}\nabla_y f(x) + \nabla_u f(x).$$

Thus, if the Lagrange multiplier is computed by the adjoint formula

$$\lambda = -C_y(x)^{-T}\nabla_y f(x), \tag{2.6}$$

then

$$W(x)^T\nabla f(x) = C_u(x)^T\lambda + \nabla_u f(x) = \nabla_u \ell(x,\lambda).$$

Now we turn to the reduced problem with $\mathcal{B} = \mathbb{R}^{n-m}$. Suppose there exists an open set $\mathcal{U}$ such that for all $u \in \mathcal{U}$ there exists a solution y of $C(y,u) = 0$ and such that the matrix $C_y(x)$ is invertible for all $x = (y,u)$ with $u \in \mathcal{U}$ and $C(y,u) = 0$. Then the Implicit Function Theorem guarantees the existence of a differentiable function

$$y : \mathcal{U} \to \mathbb{R}^m$$

defined by

$$C(y(u),u) = 0$$

and the problem (1.1) can be reduced to (1.2). Since $y(\cdot)$ is differentiable, the function $\hat{f}$ is differentiable and its gradient is given by

$$\nabla \hat{f}(u) = W(y(u),u)^T\nabla f(y(u),u),$$

cf. [23]. Moreover, it can be shown that the Hessian of $\hat{f}$ is equal to the reduced Hessian

$$\nabla^2\hat{f}(u) = W(y(u),u)^T\nabla^2_{xx}\ell(y(u),u,\lambda)\,W(y(u),u),$$

provided that the Lagrange multiplier is computed from (2.6).

One can see that the gradient and the Hessian information in the SQP method for (1.1) and in the Newton method for (1.2) are the same if (y,u) solves $C(y,u) = 0$. Thus, if Newton–like methods are applied for the solution of (1.2), then one has all the ingredients available necessary to implement an SQP method for the solution of (1.1). The important difference, of course, is that in the SQP method we do not have to solve the nonlinear constraints $C(y,u) = 0$ at every iteration.

In these considerations we neglected the bound constraints $a \le u \le b$. These will be analyzed in the following sections. We already point out that these relationships between (1.1) and (1.2) are basically the same with or without the bound constraints.

3. Optimality conditions. A point x_* satisfies the first–order Karush–Kuhn–Tucker (KKT) conditions if there exist $\lambda_* \in \mathbb{R}^m$ and $\mu_*^a, \mu_*^b \in \mathbb{R}^{n-m}$ such that

$$
\begin{aligned}
& C(x_*) = 0, \\
& a \le u_* \le b, \\
& \begin{pmatrix} \nabla_y f(x_*) \\ \nabla_u f(x_*) \end{pmatrix} + \begin{pmatrix} C_y(x_*)^T \lambda_* \\ C_u(x_*)^T \lambda_* \end{pmatrix} - \begin{pmatrix} 0 \\ \mu_*^a \end{pmatrix} + \begin{pmatrix} 0 \\ \mu_*^b \end{pmatrix} = 0, \\
& ((u_*)_i - a_i)\,(\mu_*^a)_i = (b_i - (u_*)_i)\,(\mu_*^b)_i = 0, \quad i = 1, \ldots, n-m, \quad \text{and} \\
& \mu_*^a \ge 0, \; \mu_*^b \ge 0.
\end{aligned}
$$

These KKT conditions are necessary conditions for x_* to be a local solution of (1.1). Note that the constraint qualifications are satisfied since the invertibility of $C_y(x_*)$ and the form of the bound constraints imply the linear independence of the active constraints. Under the assumption of the invertibility of $C_y(x_*)$, we can rewrite the first–order KKT conditions:

$$
\begin{array}{lll}
C(x_*) = 0, & & \\
a \le u_* \le b, & & \\
\lambda_* = -C_y(x_*)^{-T} \nabla_y f(x_*), & & \\
a_i < (u_*)_i < b_i & \Longrightarrow & \left(\nabla_u \ell(x_*, \lambda_*)\right)_i = 0, \\
(u_*)_i = a_i & \Longrightarrow & \left(\nabla_u \ell(x_*, \lambda_*)\right)_i \ge 0, \quad \text{and} \\
(u_*)_i = b_i & \Longrightarrow & \left(\nabla_u \ell(x_*, \lambda_*)\right)_i \le 0.
\end{array}
$$

One can obtain a useful form of the first–order KKT conditions by noting that

$$
\begin{aligned}
\nabla_u \ell(x_*, \lambda_*) &= \nabla_u f(x_*) + C_u(x_*)^T \lambda_* \\
&= \nabla_u f(x_*) - C_u(x_*)^T C_y(x_*)^{-T} \nabla_y f(x_*) \\
&= W(x_*)^T \nabla f(x_*).
\end{aligned}
$$

In other words, $\nabla_u \ell(x_*, \lambda_*)$ is just the reduced gradient corresponding to the u variables. Hence x_* is a first–order KKT point if

$$
\begin{array}{lll}
C(x_*) = 0, & & \\
a \le u_* \le b, & & \\
a_i < (u_*)_i < b_i & \Longrightarrow & \left(W(x_*)^T \nabla f(x_*)\right)_i = 0, \\
(u_*)_i = a_i & \Longrightarrow & \left(W(x_*)^T \nabla f(x_*)\right)_i \ge 0, \quad \text{and} \\
(u_*)_i = b_i & \Longrightarrow & \left(W(x_*)^T \nabla f(x_*)\right)_i \le 0.
\end{array}
$$

Furthermore, x_* satisfies the second–order necessary KKT conditions if it satisfies the first–order KKT conditions and if the principal submatrix of the reduced Hessian

$$
W(x_*)^T \nabla_{xx}^2 \ell(x_*, \lambda_*) W(x_*)
$$

corresponding to indices i such that $a_i < (u_*)_i < b_i$ is positive semi–definite, where the multipliers λ_* are given by $\lambda_* = -C_y(x_*)^{-T} \nabla_y f(x_*)$.

Now we adapt the idea of Coleman and Li [8] to this context and define $D(x) \in \mathbb{R}^{(n-m)\times(n-m)}$ to be the diagonal matrix with diagonal elements given by

$$\Big(D(x)\Big)_{ii} = \begin{cases} (b-u)_i^{\frac{1}{2}} & \text{if } \Big(W(x)^T\nabla f(x)\Big)_i < 0 \text{ and } b_i < +\infty, \\ 1 & \text{if } \Big(W(x)^T\nabla f(x)\Big)_i < 0 \text{ and } b_i = +\infty, \\ (u-a)_i^{\frac{1}{2}} & \text{if } \Big(W(x)^T\nabla f(x)\Big)_i \geq 0 \text{ and } a_i > -\infty, \\ 1 & \text{if } \Big(W(x)^T\nabla f(x)\Big)_i \geq 0 \text{ and } a_i = -\infty, \end{cases} \tag{3.1}$$

for $i = 1, \ldots, n-m$. In the following proposition we give the form of the first–order and second–order necessary KKT conditions that we use in this paper. To us, they indicate the suitability of (3.1) as a scaling for (1.1).

Proposition 3.1. *The point x_* satisfies the first–order KKT conditions if and only if*

$$\begin{aligned} &C(x_*) = 0, \quad a \leq u_* \leq b, \quad \textit{and} \\ &D(x_*)W(x_*)^T\nabla f(x_*) = 0. \end{aligned}$$

The point x_* satisfies the second–order necessary KKT conditions if and only if it satisfies the first–order KKT conditions and

$$D(x_*)W(x_*)^T\nabla^2_{xx}\ell(x_*,\lambda_*)W(x_*)D(x_*)$$

is positive semi–definite. The corresponding multiplier is given by $\lambda_* = -C_y(x_*)^{-T}\nabla_y f(x_*)$.

Remark 3.1. Proposition 3.1 remains valid for a larger class of diagonal matrices $D(x)$. The scalar 1 in the Definition (3.1) of D can be replaced by any other positive scalar and Proposition 3.1 also remains valid with $D(x)$ replaced by $D(x)^p$, $p > 0$. Most of our convergence results still hold true if $D(x)$ is replaced by $D(x)^p$, $p \geq 1$. See also Remark 8.1. However, the square roots in the definition of $D(x)$ will be necessary for the proof of local q–quadratic convergence of our algorithms.

The form of the sufficient optimality conditions used in this paper requires the definition of nondegeneracy or strict complementarity.

Definition 3.1. *A point x in $\mathcal{B}$ is said to be nondegenerate if $\Big(W(x)^T\nabla f(x)\Big)_i = 0$ implies $a_i < u_i < b_i$ for all $i \in \{1, \ldots, n-m\}$.*

We now define a diagonal $(n-m)\times(n-m)$ matrix $E(x)$ with diagonal elements given by

$$\Big(E(x)\Big)_{ii} = \left|\Big(W(x)^T\nabla f(x)\Big)_i\right| \quad \text{for} \quad i = 1, \ldots, n-m\,.$$

The significance of this matrix will become clear in the next section when we apply Newton's method to the system of nonlinear equations arising from the first–order KKT conditions. From the definitions of $D(x)$ and $E(x)$ we have the following property.

PROPOSITION 3.2. *A point x_* is nondegenerate and satisfies the second–order sufficient KKT conditions if and only if it is a first–order KKT point and*

$$D(x_*)W(x_*)^T\nabla^2_{xx}\ell(x_*,\lambda_*)W(x_*)D(x_*)+E(x_*)$$

is positive definite, where $\lambda_*=-C_y(x_*)^{-T}\nabla_y f(x_*)$.

4. Newton's method. One way to motivate the algorithms described in this paper is to apply Newton's method to the system of nonlinear equations

$$\begin{aligned} C(x)&=0,\\ D(x)^2W(x)^T\nabla f(x)&=0, \end{aligned} \tag{4.1}$$

where x is strictly feasible with respect to the bounds on the variables u, i.e., $a<u<b$. This is related to Goodman's approach [21] for an orthogonal null–space basis and equality constraints. Although $D(x)^2$ is usually discontinuous at points where $\left(W(x)^T\nabla f(x)\right)_i=0$, the function $D(x)^2W(x)^T\nabla f(x)$ is continuous (but not differentiable) at such points. The application of Newton's method to this type of nonlinear systems has first been suggested by Coleman and Li [8] in the context of nonlinear minimization problems with simple bounds. They have shown that this type of nondifferentiability still allows the Newton process to achieve local q–quadratic convergence. In order to apply Newton's method we first need to compute some derivatives.

To calculate the Jacobian of the reduced gradient $W(x)^T\nabla f(x)$, we write

$$W(x)^T\nabla f(x)=\nabla_u f(x)+C_u(x)^T\lambda,$$

where λ is given by $C_y(x)^T\lambda=-\nabla_y f(x)$ and has derivatives

$$\begin{aligned} \frac{\partial\lambda}{\partial y} &= -C_y(x)^{-T}\left(\textstyle\sum_{i=1}^m\nabla^2_{yy}c_i(x)\lambda_i+\nabla^2_{yy}f(x)\right)\\ &= -C_y(x)^{-T}\nabla^2_{yy}\ell(x,\lambda),\\ \frac{\partial\lambda}{\partial u} &= -C_y(x)^{-T}\left(\textstyle\sum_{i=1}^m\nabla^2_{yu}c_i(x)\lambda_i+\nabla^2_{yu}f(x)\right)\\ &= -C_y(x)^{-T}\nabla^2_{yu}\ell(x,\lambda). \end{aligned}$$

This implies the equalities

$$\begin{aligned} \frac{\partial}{\partial y}\left(W(x)^T\nabla f(x)\right) &= C_u(x)^T\frac{\partial\lambda}{\partial y}+\nabla^2_{uy}f(x)+\textstyle\sum_{i=1}^m\nabla^2_{uy}c_i(x)\lambda_i\\ &= W(x)^T\begin{pmatrix}\nabla^2_{yy}\ell(x,\lambda)\\ \nabla^2_{uy}\ell(x,\lambda)\end{pmatrix},\\ \frac{\partial}{\partial u}\left(W(x)^T\nabla f(x)\right) &= C_u(x)^T\frac{\partial\lambda}{\partial u}+\nabla^2_{uu}f(x)+\textstyle\sum_{i=1}^m\nabla^2_{uu}c_i(x)\lambda_i\\ &= W(x)^T\begin{pmatrix}\nabla^2_{yu}\ell(x,\lambda)\\ \nabla^2_{uu}\ell(x,\lambda)\end{pmatrix}, \end{aligned}$$

and we can conclude that

$$\frac{\partial}{\partial x}\Big(W(x)^T\nabla f(x)\Big) = W(x)^T\nabla^2_{xx}\ell(x,\lambda),$$

where $\lambda = -C_y(x)^{-T}\nabla_y f(x)$.

A linearization of (4.1) gives

$$C_y(x)s_y + C_u(x)s_u = -C(x), \tag{4.2}$$

$$\Big(D(x)^2W(x)^T\nabla^2_{xx}\ell(x,\lambda) + \big[0 \mid E(x)\big]\Big)\begin{pmatrix} s_y \\ s_u\end{pmatrix} = -D(x)^2W(x)^T\nabla f(x), \tag{4.3}$$

where 0 denotes the $(n-m)\times m$ matrix with zero entries. Equation (4.2) is the linearized state equation. The diagonal elements of $E(x)$ are the product of the derivative of the diagonal elements of $D(x)^2$ and the components of the reduced gradient $W(x)^T\nabla f(x)$. The derivative of $(D(x)^2)_{ii}$ does not exist if $\Big(W(x)^T\nabla f(x)\Big)_i = 0$. In this case we set the corresponding quantities in the Jacobian to zero (see references [8], [9]). This gives the equation (4.3).

By using (2.2) we can rewrite the linear system (4.2)–(4.3) as

$$s = s^{\mathsf{n}} + W(x)s_u,$$

$$\Big(D(x)^2W(x)^T\nabla^2_{xx}\ell(x,\lambda)W(x) + E(x)\Big)\, s_u = -D(x)^2W(x)^T\Big(\nabla^2_{xx}\ell(x,\lambda)s^{\mathsf{n}} + \nabla f(x)\Big). \tag{4.4}$$

We define our Newton–like step as the solution of

$$s = s^{\mathsf{n}} + W(x)s_u, \tag{4.5}$$

$$\Big(\bar{D}(x)^2W(x)^T\nabla^2_{xx}\ell(x,\lambda)W(x) + E(x)\Big)\, s_u = -\bar{D}(x)^2W(x)^T\Big(\nabla^2_{xx}\ell(x,\lambda)s^{\mathsf{n}} + \nabla f(x)\Big), \tag{4.6}$$

where $\bar{D}(x) \in \mathbb{R}^{(n-m)\times(n-m)}$ is the diagonal matrix defined by

$$\Big(\bar{D}(x)\Big)_{ii} = \begin{cases} (b-u)_i^{\frac{1}{2}} & \text{if} \quad \Big(W(x)^T\left(\nabla^2_{xx}\ell(x,\lambda)s^{\mathsf{n}} + \nabla f(x)\right)\Big)_i < 0 \text{ and } b_i < +\infty, \\ 1 & \text{if} \quad \Big(W(x)^T\left(\nabla^2_{xx}\ell(x,\lambda)s^{\mathsf{n}} + \nabla f(x)\right)\Big)_i < 0 \text{ and } b_i = +\infty, \\ (u-a)_i^{\frac{1}{2}} & \text{if} \quad \Big(W(x)^T\left(\nabla^2_{xx}\ell(x,\lambda)s^{\mathsf{n}} + \nabla f(x)\right)\Big)_i \geq 0 \text{ and } a_i > -\infty, \\ 1 & \text{if} \quad \Big(W(x)^T\left(\nabla^2_{xx}\ell(x,\lambda)s^{\mathsf{n}} + \nabla f(x)\right)\Big)_i \geq 0 \text{ and } a_i = -\infty, \end{cases}$$

for $i = 1, \ldots, n-m$. This change of the diagonal scaling matrix is based on the form of the right hand side of (4.4).

If x is close to a nondegenerate point x_* satisfying the second–order sufficient KKT conditions and if $W(x)^T\nabla^2_{xx}\ell(x,\lambda)s^{\mathsf{n}}$ is sufficiently small, a step s defined in this way is a Newton step on the following system of nonlinear equations

$$\begin{aligned} &C(x) = 0, \\ &D(x)^2_u W(x)^T\nabla f(x) = 0, \end{aligned} \tag{4.7}$$

where $D(x)_u$ depends on x_* as follows:

$$(D(x)_u)_{ii} = \begin{cases} 1 \text{ or } -1 \text{ or } (b-u)_i^{\frac{1}{2}} \text{ or } (u-a)_i^{\frac{1}{2}} & \text{if } \left(W(x_*)^T \nabla f(x_*)\right)_i = 0, \\ (b-u)_i^{\frac{1}{2}} & \text{if } \left(W(x_*)^T \nabla f(x_*)\right)_i < 0, \\ (u-a)_i^{\frac{1}{2}} & \text{if } \left(W(x_*)^T \nabla f(x_*)\right)_i > 0, \end{cases}$$

for $i = 1, \ldots, n-m$. If $\left(W(x_*)^T \nabla f(x_*)\right)_i = 0$, the i–th diagonal element of $D(x)_u$ has to be chosen so that $\bar{D}(x)$ and $D(x)_u$ are the same matrix. Of course, this depends on the sign of $\left(W(x)^T(\nabla^2_{xx}\ell(x,\lambda)s^{\mathsf{n}} + \nabla f(x))\right)_i$. As Coleman and Li [8] pointed out, $D(x)_u$ is just of theoretical use since x_* is unknown. One can see that $D(x)_u^2 W(x)^T \nabla f(x)$ is continuously differentiable with Lipschitz continuous derivatives in an open neighborhood of x_*, that $D(x_*)_u^2 W(x_*)^T \nabla f(x_*) = 0$, and that the Jacobian of $D(x)_u^2 W(x)^T \nabla f(x)$ at x_* is nonsingular, for all choices of $D(x)_u$. These conditions are those typically required to get q–quadratic convergence for the Newton iteration (see [13, Thm. 5.2.1]). Thus the sequence of iterates generated by the Newton step (4.5)–(4.6) will converge q–quadratically to a nondegenerate point that satisfies the sufficient KKT conditions. The interior–point process damps the Newton step so that it stays strictly feasible but this does not affect the rate of convergence. The details are provided in Corollary 9.1.

5. Trust–region interior–point SQP algorithms. The algorithms that we propose generate a sequence of iterates $\{x_k\}$ where

$$x_k = \begin{pmatrix} y_k \\ u_k \end{pmatrix},$$

and u_k is strictly feasible with respect to the bounds, i.e., $a < u_k < b$. At iteration k we are given x_k, and we need to compute a trial step s_k. If s_k is accepted, we set $x_{k+1} = x_k + s_k$. Otherwise we set x_{k+1} to x_k, reduce the trust–region radius, and compute a new trial step.

Following the application of Newton's method (4.5), each trial step s_k is decomposed as

$$s_k = s_k^{\mathsf{n}} + s_k^{\mathsf{t}} = s_k^{\mathsf{n}} + W_k (s_k)_u,$$

where s_k^{n} is called the quasi–normal component and s_k^{t} is the tangential component.

The role of s_k^{n} is to move towards feasibility. It will be seen that s_k^{n} is related to the Newton step for the solution of $C(y, u_k) = 0$ for fixed u_k. The role of s_k^{t} is to move towards optimality. The u component of s_k^{t} is related to the Newton step for the reduced problem (1.2). However, as made clear previously, we do not require feasibility with respect to the nonlinear equality constraints.

The global convergence is guaranteed by imposing an appropriate trust region on the step and monitoring the progress by a suitable merit function. The definition of the quasi–normal component, the tangential component, and the merit function as well as the complete formulation of our algorithms is the content of this section.

5.1. The quasi–normal component. Let δ_k be the trust radius at iteration k. The quasi–normal component s_k^{n} is related to the trust–region subproblem for the linearized constraints

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|J_k s^{\mathsf{n}} + C_k\|^2 \\ \text{subject to} & \|s^{\mathsf{n}}\| \leq \delta_k, \end{array}$$

and it is required to have the form

$$s_k^{\mathsf{n}} = \begin{pmatrix} (s_k^{\mathsf{n}})_y \\ 0 \end{pmatrix}. \tag{5.1}$$

Thus the displacement along s_k^{n} is made only in the y variables, and as a consequence, x_k and $x_k + s_k^{\mathsf{n}}$ have the same u components. Since $(s_k^{\mathsf{n}})_u = 0$, the trust–region subproblem introduced above can be rewritten as

$$\text{minimize} \quad \frac{1}{2}\|C_y(x_k)(s^{\mathsf{n}})_y + C_k\|^2 \tag{5.2}$$

$$\text{subject to} \quad \|(s^{\mathsf{n}})_y\| \leq \delta_k. \tag{5.3}$$

Thus, the quasi–normal component s_k^{n} is a trust–region globalization of the component s^{n} given in (2.3) of the Newton step (4.5). We do not have to solve (5.2)–(5.3) exactly, we only have to assume that the quasi–normal component satisfies the conditions

$$\|s_k^{\mathsf{n}}\| \leq \kappa_1 \|C_k\| \tag{5.4}$$

and

$$\|C_k\|^2 - \|C_y(x_k)(s_k^{\mathsf{n}})_y + C_k\|^2 \geq \kappa_2 \|C_k\| \min\{\kappa_3\|C_k\|, \delta_k\}, \tag{5.5}$$

where κ_1, κ_2, and κ_3 are positive constants independent of k. In Section 10.1, we describe several ways of computing a quasi–normal component that satisfies the requirements (5.1), (5.4), and (5.5). Condition (5.4) tell us that the quasi–normal component is small close to feasible points. Condition (5.5) is just a weaker form of Cauchy decrease or simple decrease for the trust–region subproblem (5.2), (5.3).

5.2. The tangential component. The computation of the tangential component $(s_k)_u$ follows a trust–region globalization of the Newton step (4.6). Following Coleman and Li [9] we symmetrize (4.6) and get

$$\left(\bar{D}_k W_k^T H_k W_k \bar{D}_k + E_k\right) \bar{D}_k^{-1} s_u = -\bar{D}_k W_k^T \left(H_k s_k^{\mathsf{n}} + \nabla f_k\right),$$

where $\bar{D}_k = \bar{D}(x_k)$, $E_k = E(x_k)$, and H_k denotes a symmetric approximation to the Hessian matrix $\nabla^2_{xx}\ell_k$. This suggests the change of variables $\hat{s}_u = \bar{D}_k^{-1} s_u$ and the consideration in the scaled space $\hat{s}_u$ of the trust–region subproblem

$$\begin{array}{ll} \text{minimize} & \left(\bar{D}_k W_k^T \left(H_k s_k^{\mathsf{n}} + \nabla f_k\right)\right)^T \hat{s}_u + \frac{1}{2}\hat{s}_u^T \left(\bar{D}_k W_k^T H_k W_k \bar{D}_k + E_k\right) \hat{s}_u \\ \text{subject to} & \|\hat{s}_u\| \leq \delta_k. \end{array}$$

Now we can rewrite the previous subproblem in the unscaled space s_u as

$$
\begin{array}{ll}
\text{minimize} & \left(W_k^T\left(H_k s_k^{\mathsf{n}} + \nabla f_k\right)\right)^T s_u + \frac{1}{2} s_u^T \left(W_k^T H_k W_k + E_k \bar{D}_k^{-2}\right) s_u \\
\text{subject to} & \|\bar{D}_k^{-1} s_u\| \le \delta_k.
\end{array} \tag{5.6}
$$

Of course, we also have to require that the new iterate is in the interior of the box constraints. To ensure that $u_k + s_k$ is strictly feasible with respect to the box constraints we choose $\sigma_k \in [\sigma, 1)$, $\sigma \in (0,1)$, and compute s_u with $\sigma_k(a - u_k) \le s_u \le \sigma_k(b - u_k)$. However, one of the strength of this trust–region approach is that we can allow for approximate solutions of this subproblem. In particular, it is not necessary to solve the full trust–region subproblem including the box constraints. For example, one can compute the solution of the trust–region subproblem without the box constraints and then scale the computed solution back so that the resulting damped s_u obeys $\sigma_k(a - u_k) \le s_u \le \sigma_k(b - u_k)$, see e.g. Section 5.2.4. We will show that under suitable assumptions this strategy guarantees global convergence and local q–quadratic convergence. Another way to compute an approximate u component of the step is to use a modified conjugate–gradient algorithm applied to the trust–region subproblem without the box constraints that is truncated if one of the bounds $\sigma_k(a - u_k) \le s_u \le \sigma_k(b - u_k)$ is violated. See Section 10.2. More ways to compute the tangential component are possible. The conditions on the tangential component necessary to guarantee global convergence are stated in Section 5.2.3.

We now introduce a quadratic model

$$
q_k(s) = \ell_k + \nabla_x {\ell_k}^T s + \frac{1}{2} s^T H_k s
$$

of $\ell(x_k + s, \lambda_k)$ about (x_k, λ_k). A trivial manipulation shows that

$$
q_k(s_k^{\mathsf{n}} + W_k s_u) = q_k(s_k^{\mathsf{n}}) + \bar{g}_k^T s_u + \frac{1}{2} {s_u}^T W_k^T H_k W_k s_u, \tag{5.7}
$$

with

$$
\bar{g}_k = W_k^T \nabla q_k(s_k^{\mathsf{n}}) = W_k^T \left(H_k s_k^{\mathsf{n}} + \nabla f_k\right).
$$

For convenience we define

$$
\Psi_k(s_u) = q_k(s_k^{\mathsf{n}} + W_k s_u) + \frac{1}{2} s_u^T \left(E_k \bar{D}_k^{-2}\right) s_u. \tag{5.8}
$$

5.2.1. The decoupled trust–region approach. We can restate the trust–region subproblem (5.6) as

$$
\text{minimize} \quad \Psi_k(s_u) \tag{5.9}
$$

$$
\text{subject to} \quad \|\bar{D}_k^{-1} s_u\| \le \delta_k. \tag{5.10}
$$

We refer to the approach based on this subproblem as the decoupled approach. In this decoupled approach the trust–region constraint is of the form $\|\bar{D}_k^{-1} s_u\| \le \delta_k$ corresponding to the constraint $\|\hat{s}_u\| \le \delta_k$ in the scaled space. One can see from (5.3) and (5.10) that we are imposing the trust region separately on the y part of the quasi–normal component and on the u part of the tangential component. Moreover, if the cross–term $W_k^T H_k s_k^{\mathsf{n}}$ is set to zero, then the trust–region subproblems for the quasi–normal component and for the tangential component are completely separated.

5.2.2. The coupled trust–region approach. The approach we present now forces the y and u parts of the tangential component $s_k^{\mathsf{t}} = W_k(s_k)_u$ to lie inside the trust region of radius δ_k. The reference trust–region subproblem is given by

$$\text{minimize} \quad \Psi_k(s_u) \tag{5.11}$$

$$\text{subject to} \quad \left\| \begin{pmatrix} -C_y(x_k)^{-1}C_u(x_k)s_u \\ \bar{D}_k^{-1}s_u \end{pmatrix} \right\| \le \delta_k. \tag{5.12}$$

In the case where there are no bounds on u this trust–region constraint is of the form

$$\left\| \begin{pmatrix} -C_y(x_k)^{-1}C_u(x_k)s_u \\ s_u \end{pmatrix} \right\| \;=\; \|W_k s_u\| \;\le\; \delta_k.$$

As opposed to the decoupled case, one can see that the term $C_y(x_k)^{-1}C_u(x_k)s_u$ is present in the trust–region constraint (5.12). If W_k^+ denotes the Moore–Penrose pseudo inverse of W_k (see [19, [Sec. 5.5.4]), then

$$\frac{1}{\|W_k^+\|}\|s_u\| \le \|W_k s_u\| \le \|W_k\|\|s_u\|.$$

Thus, if the condition number $\kappa(W_k) = \|W_k^+\|\;\|W_k\|$ is small, then the decoupled and the coupled approach will generate similar iterates. In this case, the decoupled approach will be more efficient since it uses fewer linear system solves with the system matrix $C_y(x_k)$. See Section 10.2. However, if $\kappa(W_k)$ is large, e.g. if $C_y(x_k)$ is ill–conditioned, then the coupled approach will use the true size of the tangential component, whereas the decoupled approach may underestimate vastly the size of this step component. This can lead to poor performance of the decoupled approach when steps are rejected and the trust–region radius is reduced based on the incorrect estimate $\|s_u\|$ of the norm of $s^{\mathsf{t}} = W_k s_u$. This indicates that when $C_y(x)$ is ill–conditioned the coupled approach offers a better regularization of the step.

5.2.3. Cauchy decrease for the tangential component. To assure global convergence to a first–order KKT point, we consider analogs for the subproblems (5.9)–(5.10) and (5.11)–(5.12) of the fraction of Cauchy decrease or simple decrease conditions for the unconstrained minimization problem.

First we consider the decoupled trust–region subproblem (5.9)–(5.10). The Cauchy step c_k^{d} is defined for this case as the solution of

$$\begin{array}{rl} \text{minimize} & \Psi_k(s_u) \\ \text{subject to} & \|\bar{D}_k^{-1}s_u\| \le \delta_k, \quad s_u \in span\{-\bar{D}_k^2\bar{g}_k\}, \\ & \sigma_k(a-u_k) \le s_u \le \sigma_k(b-u_k), \end{array}$$

where $-\bar{D}_k^2\bar{g}_k$ is the steepest–descent direction for $\Psi_k(s_u)$ at $s_u = 0$ in the norm $\|\bar{D}_k^{-1}\cdot\|$. Here $\sigma_k \in [\sigma, 1)$ ensures that the Cauchy step c_k^{d} remains strictly feasible with respect to the box constraints. The parameter $\sigma \in (0,1)$ is fixed for all k. As in many trust–region algorithms, we require the tangential component $(s_k)_u$ with $\sigma_k(a-u_k) \le (s_k)_u \le \sigma_k(b-u_k)$ to give a decrease on $\Psi_k(s_u)$ smaller than a uniform fraction of the decrease given by c_k^{d} for the same function $\Psi_k(s_u)$. This condition is often called fraction of Cauchy decrease, and in this case is

$$\Psi_k(0) - \Psi_k((s_k)_u) \ge \beta_1^{\mathsf{d}}\left(\Psi_k(0) - \Psi_k(c_k^{\mathsf{d}})\right), \tag{5.13}$$

where β_1^{d} is positive and fixed across all iterations. It is not difficult to see that dogleg or conjugate–gradient algorithms can compute components $(s_k)_u$ conveniently that satisfy condition (5.13) with $\beta_1^{\mathsf{d}} = 1$. We leave these issues to Section 10.2.

In a similar way, the component $(s_k)_u$ with $\sigma_k(a - u_k) \leq (s_k)_u \leq \sigma_k(b - u_k)$ satisfies a fraction of Cauchy decrease for the coupled trust–region subproblem (5.11)–(5.12) if

$$\Psi_k(0) - \Psi_k((s_k)_u) \geq \beta_1^{\mathsf{c}} \Big(\Psi_k(0) - \Psi_k(c_k^{\mathsf{c}}) \Big), \tag{5.14}$$

for some β_1^{c} independent of k, where the Cauchy step c_k^{c} is the solution of

$$\begin{array}{ll} \text{minimize} & \Psi_k(s_u) \\ \text{subject to} & \left\| \begin{pmatrix} -C_y(x_k)^{-1} C_u(x_k) s_u \\ \bar{D}_k^{-1} s_u \end{pmatrix} \right\| \leq \delta_k, \quad s_u \in span\{-\bar{D}_k^2 \bar{g}_k\}, \\ & \sigma_k(a - u_k) \leq s_u \leq \sigma_k(b - u_k). \end{array}$$

In Section 10.2 we show how to use conjugate–gradients to compute components $(s_k)_u$ satisfying the condition (5.14).

One final comment is in order. In the coupled approach, the Cauchy step c_k^{c} was defined along the direction $-\bar{D}_k^2 \bar{g}_k$. To simplify this discussion, suppose that there are no bounds on u. In this case the trust–region constraint is of the form $\|W_k s_u\| \leq \delta_k$. The presence of W_k gives the trust region an ellipsoidal shape. The steepest–descent direction for the quadratic (5.8) in the norm $\|W_k \cdot\|$ at $s_u = 0$ is given by $-(W_k^T W_k)^{-1} \bar{g}_k$. Our analysis still holds for this case since $\{\|(W_k^T W_k)^{-1}\|\}$ is a bounded sequence. The reason why we avoid the term $(W_k^T W_k)^{-1}$ is that in many applications there is no reasonable way to solve systems with $W_k^T W_k$. We will show in Section 10.2 how this affects the use of conjugate gradients (see Remark 10.2). Finally, we point out that this problem does not arise if the decoupled approach is used.

5.2.4. Optimal decrease for the tangential component. The conditions in the previous subsection are sufficient to guarantee global convergence to a point satisfying first–order necessary KKT conditions, but they are too weak to guarantee global convergence to a point satisfying second–order necessary KKT conditions. To accomplish this, just as in the unconstrained case [38], [48], in the box–constrained case [9], and in the equality–constrained case [14], we need to make sure that s_u satisfies an appropriate fraction of optimal decrease condition.

First we consider the decoupled approach and let o_k^{d} be an optimal solution of the trust–region subproblem (5.9)–(5.10). It follows from the KKT conditions for this trust–region subproblem that there exists $\gamma_k \geq 0$ such that

$$W_k^T H_k W_k + E_k \bar{D}_k^{-2} + \gamma_k \bar{D}_k^{-2} \quad \text{is positive semi–definite,} \tag{5.15}$$

$$\left(W_k^T H_k W_k + E_k \bar{D}_k^{-2} + \gamma_k \bar{D}_k^{-2} \right) o_k^{\mathsf{d}} = -\bar{g}_k, \text{ and} \tag{5.16}$$

$$\gamma_k (\delta_k - \|\bar{D}_k^{-1} o_k^{\mathsf{d}}\|) = 0.$$

(For practical algorithms to compute o_k^{d} see references [44], [38], [46], [49]. These conditions are also sufficient for o_k^{d} to be an optimal solution [17], [48].) Since $u_k + o_k^{\mathsf{d}}$ might not be strictly feasible, we consider $\tau_k o_k^{\mathsf{d}}$, where τ_k is given by

$$\tau_k = \sigma_k \ \min \left\{ 1, \ \max \left\{ \frac{b_i - (u_k)_i}{(o_k^{\mathsf{d}})_i}, \ \frac{(u_k)_i - a_i}{(o_k^{\mathsf{d}})_i} \right\}, \ i = 1, \ldots, n - m \right\}. \tag{5.17}$$

The tangential component $(s_k)_u$ then is required to satisfy the following fraction of optimal decrease condition

$$\begin{aligned} &\Psi_k(0) - \Psi_k((s_k)_u) \geq \beta_2^{\mathsf{d}} \left(\Psi_k(0) - \Psi_k(\tau_k o_k^{\mathsf{d}}) \right) \quad \text{and} \\ &\|\bar{D}_k^{-1}(s_k)_u\| \leq \beta_3^{\mathsf{d}} \delta_k, \end{aligned} \tag{5.18}$$

where $\beta_2^{\mathsf{d}}, \beta_3^{\mathsf{d}}$ are positive parameters.

From conditions (5.15), (5.16), and (5.18), and $\tau_k < 1$, we can write

$$\begin{aligned} \Psi_k(0) - \Psi_k((s_k)_u) \;\; &\geq \;\; \beta_2^{\mathsf{d}} \left(-\tau_k \bar{g}_k^T o_k^{\mathsf{d}} - \frac{1}{2}\tau_k^2 (o_k^{\mathsf{d}})^T \left(W_k^T H_k W_k + E_k \bar{D}_k^{-2} \right) (o_k^{\mathsf{d}}) \right) \\ &\geq \;\; \beta_2^{\mathsf{d}} \tau_k \left(-\bar{g}_k^T o_k^{\mathsf{d}} - \frac{1}{2} (o_k^{\mathsf{d}})^T \left(W_k^T H_k W_k + E_k \bar{D}_k^{-2} + \gamma_k \bar{D}_k^{-2} \right) (o_k^{\mathsf{d}}) \right) \\ &\qquad\qquad + \frac{1}{2}\beta_2^{\mathsf{d}} \tau_k^2 \gamma_k (o_k^{\mathsf{d}})^T \bar{D}_k^{-2} (o_k^{\mathsf{d}}) \\ &\geq \;\; \frac{1}{2}\beta_2^{\mathsf{d}} \tau_k \|R_k o_k^{\mathsf{d}}\|^2 + \frac{1}{2}\beta_2^{\mathsf{d}} \tau_k^2 \gamma_k \delta_k^2 \\ &\geq \;\; \frac{1}{2}\beta_2^{\mathsf{d}} \tau_k^2 \gamma_k \delta_k^2, \end{aligned} \tag{5.19}$$

where $W_k^T H_k W_k + E_k \bar{D}_k^{-2} + \gamma_k \bar{D}_k^{-2} = R_k^T R_k$.

Now let us focus on the coupled approach and let o_k^{c} be the optimal solution of the trust–region subproblem (5.11)–(5.12). It follows from the KKT conditions for this trust–region subproblem and the equality

$$\left(C_y(x_k)^{-1} C_u(x_k) \right)^T C_y(x_k)^{-1} C_u(x_k) \;=\; W_k^T W_k - I_{n-m},$$

that there exists $\gamma_k \geq 0$ such that

$$W_k^T H_k W_k + E_k \bar{D}_k^{-2} + \gamma_k \left(\bar{D}_k^{-2} + W_k^T W_k - I_{n-m} \right) \qquad \text{is positive semi–definite,} \tag{5.20}$$

$$\left(W_k^T H_k W_k + E_k \bar{D}_k^{-2} + \gamma_k \left(\bar{D}_k^{-2} + W_k^T W_k - I_{n-m} \right) \right) o_k^{\mathsf{c}} = -\bar{g}_k, \text{ and} \tag{5.21}$$

$$\gamma_k \left(\delta_k - \left\| \begin{pmatrix} -C_y(x_k)^{-1} C_u(x_k) o_k^{\mathsf{c}} \\ \bar{D}_k^{-1} o_k^{\mathsf{c}} \end{pmatrix} \right\| \right) = 0.$$

Now we damp o_k^{c} with τ_k given as in (5.17) but with o_k^{d} replaced by o_k^{c}. Thus, the resulting step $u_k + \tau_k o_k^{\mathsf{c}}$ is strictly feasible. We impose the following fraction of optimal decrease condition on the tangential component $(s_k)_u$:

$$\begin{aligned} &\Psi_k(0) - \Psi_k((s_k)_u) \geq \beta_2^{\mathsf{c}} \Big(\Psi_k(0) - \Psi_k(\tau_k o_k^{\mathsf{c}}) \Big) \quad \text{and} \\ &\left\| \begin{pmatrix} -C_y(x_k)^{-1} C_u(x_k) (s_k)_u \\ \bar{D}_k^{-1} (s_k)_u \end{pmatrix} \right\| \leq \beta_3^{\mathsf{c}} \delta_k. \end{aligned} \tag{5.22}$$

In this case it can be shown in a way similar to (5.19) that

$$\Psi_k(0) - \Psi((s_k)_u) \geq \frac{1}{2}\beta_2^{\mathsf{c}} \tau_k^2 \gamma_k \delta_k^2. \tag{5.23}$$

5.3. Reduced and full Hessians. In the previous section we considered an approximation H_k to the full Hessian. The algorithms and theory presented in this paper are also valid if we use an approximation $\widehat{H}_k$ to the reduced Hessian $W_k^T \nabla^2_{xx} \ell_k W_k$. In this case we set

$$H_k = \begin{pmatrix} 0 & 0 \\ 0 & \widehat{H}_k \end{pmatrix}. \tag{5.24}$$

Due to the form of W_k, we have

$$W_k^T H_k W_k = \widehat{H}_k.$$

This allows us to obtain the expansion (5.7) in the context of a reduced Hessian approximation.

For the algorithms with reduced Hessian approximation the following observations are useful:

$$\begin{aligned} H_k d &= \begin{pmatrix} 0 \\ \widehat{H}_k d_u \end{pmatrix}, \\ d^T H_k d &= d_u^T \widehat{H}_k d_u, \\ W_k^T H_k d &= \widehat{H}_k d_u. \end{aligned} \tag{5.25}$$

5.4. Outline of the algorithms. We need to introduce a merit function and the corresponding actual and predicted reductions. The merit function used is the augmented Lagrangian

$$L(x, \lambda; \rho) = f(x) + \lambda^T C(x) + \rho C(x)^T C(x).$$

We follow [11] and define the actual decrease at iteration k as

$$ared(s_k; \rho_k) = L(x_k, \lambda_k; \rho_k) - L(x_k + s_k, \lambda_{k+1}; \rho_k),$$

and the predicted decrease as

$$pred(s_k; \rho_k) = L(x_k, \lambda_k; \rho_k) - \left(q_k(s_k) + \Delta\lambda_k^T (J_k s_k + C_k) + \rho_k \| J_k s_k + C_k \|^2 \right),$$

with $\Delta\lambda_k = \lambda_{k+1} - \lambda_k$.

REMARK 5.1. A possible redefinition of the actual and predicted decreases is obtained by subtracting the term $\frac{1}{2}(s_k)_u^T \left(E_k \bar{D}_k^{-2} \right) (s_k)_u$ from both $ared(s_k; \rho_k)$ and $pred(s_k; \rho_k)$. This type of modification has been suggested in [9] for minimization with simple bounds, and it does not affect the global and local results given in this paper.

To decide whether to accept or reject a trial step s_k, we evaluate the ratio

$$\frac{ared(s_k; \rho_k)}{pred(s_k; \rho_k)}.$$

To update the penalty parameter ρ_k we use the scheme proposed by El–Alem [15]. Other schemes to update the penalty parameter have been suggested in [16] and [34].

We can now outline the main procedures of the trust–region interior–point SQP algorithms and leave the practical computation of $s_k^{\mathfrak{n}}$, $(s_k)_u$, and λ_k to Section 10.

ALGORITHM 5.1 (TRUST–REGION INTERIOR–POINT SQP ALGORITHMS).

1 Choose x_0 such that $a < u_0 < b$, pick $\delta_0 > 0$, and calculate λ_0. Choose α_1, η_1, σ, δ_{min}, δ_{max}, $\bar{\rho}$, and ρ_{-1} such that $0 < \alpha_1, \eta_1, \sigma < 1$, $0 < \delta_{min} \leq \delta_{max}$, $\bar{\rho} > 0$, and $\rho_{-1} \geq 1$.

2 For $k = 0, 1, 2, \ldots$ do

2.1 Compute s_k^{n} such that $\|s_k^{\mathsf{n}}\| \leq \delta_k$.
Compute $(s_k)_u$ based on the subproblem (5.9)–(5.10) (or (5.11)–(5.12) for the coupled approach) satisfying

$$\sigma_k(a - u_k) \leq (s_k)_u \leq \sigma_k(b - u_k),$$

with $\sigma_k \in [\sigma, 1)$. Set $s_k = s_k^{\mathsf{n}} + s_k^{\mathsf{t}} = s_k^{\mathsf{n}} + W_k (s_k)_u$.

2.2 Compute λ_{k+1} and set $\Delta\lambda_k = \lambda_{k+1} - \lambda_k$.

2.3 Compute $pred(s_k; \rho_{k-1})$:

$$pred(s_k; \rho_{k-1}) = q_k(0) - q_k(s_k) - \Delta\lambda_k^T (J_k s_k + C_k) + \rho_{k-1} \left(\|C_k\|^2 - \|J_k s_k + C_k\|^2 \right).$$

If $pred(s_k; \rho_{k-1}) \geq \frac{\rho_{k-1}}{2} \left(\|C_k\|^2 - \|J_k s_k + C_k\|^2 \right)$ then set $\rho_k = \rho_{k-1}$. Otherwise set

$$\rho_k = \frac{2 \left(q_k(s_k) - q_k(0) + \Delta\lambda_k^T (J_k s_k + C_k) \right)}{\|C_k\|^2 - \|J_k s_k + C_k\|^2} + \bar{\rho}.$$

2.4 If $\frac{ared(s_k; \rho_k)}{pred(s_k; \rho_k)} < \eta_1$, set

$$\delta_{k+1} = \alpha_1 \max \left\{ \|s_k^{\mathsf{n}}\|, \|\bar{D}_k^{-1} (s_k)_u\| \right\} \text{ in the decoupled case or}$$
$$\delta_{k+1} = \alpha_1 \max \left\{ \|s_k^{\mathsf{n}}\|, \left\| \begin{pmatrix} -C_y(x_k)^{-1} C_u(x_k)(s_k)_u \\ \bar{D}_k^{-1} (s_k)_u \end{pmatrix} \right\| \right\} \text{ in the}$$

coupled case, and reject s_k.
Otherwise accept s_k and choose δ_{k+1} such that

$$\max\{\delta_{min}, \delta_k\} \leq \delta_{k+1} \leq \delta_{max}.$$

2.5 If s_k was rejected set $x_{k+1} = x_k$ and $\lambda_{k+1} = \lambda_k$. Otherwise set $x_{k+1} = x_k + s_k$ and $\lambda_{k+1} = \lambda_k + \Delta\lambda_k$.

Of course the rules to update the trust radius in the previous algorithm can be much more involved but the above suffices to prove convergence results and to understand the trust–region mechanism.

5.5. Assumptions. In order to establish local and global convergence results we need some general assumptions. We list these assumptions below. Let Ω be an open subset of $\mathbb{R}^n$ such that for all iterations k, x_k and $x_k + s_k$ are in Ω.

A.1 The functions $f(x)$, $c_i(x)$, $i = 1, \ldots, m$, are twice continuously differentiable in Ω.
A.2 The partial Jacobian $C_y(x)$ is nonsingular for all $x \in \Omega$.
A.3 The functions $f(x)$, $\nabla f(x)$, $\nabla^2 f(x)$, $C(x)$, $J(x)$, $\nabla^2 c_i(x)$, $i = 1, \ldots, m$ are bounded in Ω.
A.4 The sequences $\{W_k\}$, $\{H_k\}$, and $\{\lambda_k\}$ are bounded.
A.5 The matrix $C_y^{-1}(x)$ is uniformly bounded in Ω.

A.6 The sequence $\{u_k\}$ is bounded.

It is equivalent to Assumptions A.3–A.6, that there exist positive constants $\nu_0, \ldots, \nu_9$ independent of k such that

$$
\begin{array}{lllll}
|f(x)| \le \nu_0, & \|\nabla f(x)\| \le \nu_1, & \|\nabla^2 f(x)\| \le \nu_2, & \|C(x)\| \le \nu_3, & \|J(x)\| \le \nu_4, \\
\|\nabla^2 c_i(x)\| \le \nu_5, & i = 1, \ldots, m, & \text{and} & \|C_y(x)^{-1}\| \le \nu_6 &
\end{array}
$$

for all $x \in \Omega$, and

$$\|W_k\| \le \nu_6, \quad \|H_k\| \le \nu_7, \quad \|\lambda_k\| \le \nu_8, \quad \text{and} \quad \|\bar{D}_k\| \le \nu_9,$$

for all k.

For the rest of this paper we suppose that Assumptions A.1–A.6 are always satisfied.

As we have pointed out earlier, our approach is related to the Newton method presented in Section 4. The u component $(s_k^{\mathsf{N}})_u$ of the Newton step $s_k^{\mathsf{N}} = s_k^{\mathsf{n}} + W_k (s_k^{\mathsf{N}})_u$, whenever it is defined, is given by

$$
\begin{aligned}
(s_k^{\mathsf{N}})_u &= -\left(\bar{D}_k^2 W_k^T H_k W_k + E_k\right)^{-1} \bar{D}_k^2 \bar{g}_k \\
&= -\bar{D}_k \left(\bar{D}_k W_k^T H_k W_k \bar{D}_k + E_k\right)^{-1} \bar{D}_k \bar{g}_k,
\end{aligned}
\tag{5.26}
$$

where

$$s_k^{\mathsf{n}} = \begin{pmatrix} -C_y(x_k)^{-1} C_k \\ 0 \end{pmatrix}, \tag{5.27}$$

and $\bar{g}_k = W_k^T \left(H_k s_k^{\mathsf{n}} + \nabla f_k \right)$. From (5.26) we see that the Newton step is well defined in a neighborhood of a nondegenerate point that satisfies the second–order sufficient KKT conditions and for which $W_k^T H_k s_k^{\mathsf{n}}$ is sufficiently small. To guarantee strict feasibility of this step we consider a damped Newton step given by

$$s_k^{\mathsf{n}} + W_k \tau_k^{\mathsf{N}} (s_k^{\mathsf{N}})_u, \tag{5.28}$$

where $(s_k^{\mathsf{N}})_u$ and s_k^{n} are given by (5.26) and (5.27), and

$$\tau_k^{\mathsf{N}} = \sigma_k \min \left\{ 1, \; \max \left\{ \frac{b_i - (u_k)_i}{((s_k^{\mathsf{N}})_u)_i}, \frac{(u_k)_i - a_i}{((s_k^{\mathsf{N}})_u)_i} \right\}, \; i = 1, \ldots, n-m \right\}. \tag{5.29}$$

If Algorithms 5.1 are particularized to satisfy the following conditions on the steps, on the quadratic model, and on the Lagrange multipliers, then we can prove global and local convergence.

C.1 The quasi–normal component s_k^{n} satisfies conditions (5.1), (5.4), and (5.5).
The tangential component $(s_k)_u$ satisfies the fraction of Cauchy decrease condition (5.13) ((5.14) for the coupled approach).
The parameter σ_k is chosen in $[\sigma, 1)$, where $\sigma \in (0, 1)$ is fixed for all k.

C.2 The tangential component $(s_k)_u$ satisfies the fraction of optimal decrease condition (5.18) ((5.22) for the coupled approach).

C.3 The second derivatives of f and c_i, $i = 1, \ldots, m$ are Lipschitz continuous in Ω. The approximation to the Hessian matrix is exact, i.e., $H_k = \nabla^2_{xx}\ell(x_k, \lambda_k)$ with Lagrange multiplier $\lambda_k = -C_y(x_k)^{-T}\nabla_y f(x_k)$.

C.4 The step s_k is given by (5.28) provided $(s_k^{\mathsf{N}})_u$ exists, $(s_k^{\mathsf{n}})_y$ lies inside the trust region (5.3), and $\tau_k^{\mathsf{N}}(s_k^{\mathsf{N}})_u$ lies inside the trust region (5.10) ((5.12) for the coupled approach). The parameter σ_k is chosen such that $\sigma_k \geq \sigma$ and $|\sigma_k - 1|$ is $\mathcal{O}\left(\|\bar{D}_k\bar{g}_k\|\right)$.

Condition C.1 assures global convergence to a first–order KKT point. Global convergence to a point that satisfies the second–order necessary KKT conditions requires Conditions C.1–C.3. To prove local q–quadratic convergence, we need Conditions C.1, C.3, and C.4. It should be pointed out that the satisfaction of C.2 or C.4 does not necessarily imply the satisfaction of C.1.

6. Intermediate results. We start by pointing out that (5.5) with the fact that the tangential component lies in the null space of J_k, together imply

$$\|C_k\|^2 - \|J_k s_k + C_k\|^2 \geq \kappa_2\|C_k\| \min\{\kappa_3\|C_k\|, \delta_k\}. \tag{6.1}$$

We calculated the first derivatives of $\lambda(x) = -C_y(x)^{-T}\nabla_y f(x)$ in Section 4. It is clear that under Assumptions A.3 and A.5 these derivatives are bounded in Ω. Thus, if λ_k is computed as stated in Condition C.3, then there exists a positive constant ν_{10} independent of k such that

$$\|\Delta\lambda_k\| \leq \nu_{10}\|s_k\|. \tag{6.2}$$

From $\|s_k^{\mathsf{q}}\| \leq \delta_{max}$ and Assumptions A.3–A.4 we also have

$$\|\bar{g}_k\| = \left\|W_k^T\left(H_k s_k^{\mathsf{q}} + \nabla f_k\right)\right\| \leq \nu_{11}, \tag{6.3}$$

where $\nu_{11} = \nu_6(\nu_7\delta_{max} + \nu_1)$.

The following lemma is required for the convergence theory.

LEMMA 6.1. *Every trial step satisfies*

$$\|s_k\| \leq \kappa_4\delta_k \tag{6.4}$$

and, if s_k is rejected in Step 2.4 of Algorithms 5.1, then

$$\delta_{k+1} \geq \kappa_5\|s_k\|, \tag{6.5}$$

where κ_4 and κ_5 are positive constants independent of k.

Proof. In the coupled trust–region approach we bound s_k^{t} as follows:

$$\begin{aligned}\left\|\begin{pmatrix} -C_y(x_k)^{-1}C_u(x_k)s_u \\ s_u\end{pmatrix}\right\| &\leq \left\|\begin{pmatrix} I_m & 0 \\ 0 & \bar{D}_k\end{pmatrix}\right\| \left\|\begin{pmatrix} -C_y(x_k)^{-1}C_u(x_k)s_u \\ \bar{D}_k^{-1}s_u\end{pmatrix}\right\| \\ &\leq (1+\nu_9)\,\delta_k\,,\end{aligned}$$

where ν_9 is a uniform bound for $\|\bar{D}_k\|$, see Assumption A.6. Since $\|s_k^{\mathsf{n}}\| \leq \delta_k$, we obtain $\|s_k\| \leq (2+\nu_9)\,\delta_k$. It is not difficult to see now that in Step 2.4 we have $\delta_{k+1} \geq \frac{\alpha_1}{2}\min\left\{1, \frac{1}{1+\nu_9}\right\}\|s_k\|$.

In the decoupled approach, $\|s_k\| = \|s_k^{\mathsf{n}} + W_k(s_k)_u\| \le (1+\nu_6\nu_9)\delta_k$ and similarly $\delta_{k+1} \ge \frac{\alpha_1}{2}\min\left\{1, \frac{1}{\nu_6\nu_9}\right\}\|s_k\|$, where ν_6 is a uniform bound for $\|W_k\|$, see Assumption A.4.

We can combine these bounds to obtain

$$\begin{aligned} \|s_k\| &\le \max\{2+\nu_9, 1+\nu_6\nu_9\}\,\delta_k, \\ \delta_{k+1} &\ge \tfrac{\alpha_1}{2}\min\left\{1, \tfrac{1}{1+\nu_9}, \tfrac{1}{\nu_6\nu_9}\right\}\,\|s_k\|. \end{aligned}$$

In the case where fraction of optimal decrease (5.18) or (5.22) is imposed on $(s_k)_u$, the constants κ_4 and κ_5 depend also on β_3^{d} and β_3^{c}. ☐

In the following lemma we rewrite the fraction of Cauchy decrease conditions (5.13) and (5.14) in a more useful form for the analysis.

LEMMA 6.2. *If $(s_k)_u$ satisfies Condition C.1 then*

$$q_k(s_k^{\mathsf{n}}) - q_k(s_k^{\mathsf{n}} + W_k(s_k)_u) \ge \kappa_6\|\bar D_k\bar g_k\|\min\left\{\kappa_7\|\bar D_k\bar g_k\|, \kappa_8\delta_k\right\}, \tag{6.6}$$

where κ_6, κ_7, and κ_8 are positive constants independent of the iteration k.

Proof. From the definition (5.8) of Ψ_k we find

$$\begin{aligned} q_k(s_k^{\mathsf{n}}) - q_k(s_k^{\mathsf{n}} + W_k(s_k)_u) &\ge q_k(s_k^{\mathsf{n}}) - q_k(s_k^{\mathsf{n}} + W_k(s_k)_u) - \frac{1}{2}(s_k)_u^T\left(E_k\bar D_k^{-2}\right)(s_k)_u \\ &= \Psi_k(0) - \Psi_k((s_k)_u). \end{aligned} \tag{6.7}$$

Let $\tilde\delta_k$ be the maximum $\|\bar D_k^{-1}\cdot\|$ norm of a step, say $(\tilde s_k)_u$, along $-\bar D_k\frac{\tilde g_k}{\|\tilde g_k\|}$ allowed inside the trust region. Here $\tilde g_k = \bar D_k\bar g_k$.

If the trust region is given by (5.10), then

$$\delta_k = \tilde\delta_k. \tag{6.8}$$

If the trust region is given by (5.12), then we can use Assumptions A.4–A.6 to deduce the inequality

$$\begin{aligned} \delta_k^2 = \left\|\begin{pmatrix} -C_y(x_k)^{-1}C_u(x_k)(\tilde s_k)_u \\ \bar D_k^{-1}(\tilde s_k)_u \end{pmatrix}\right\|^2 &= \|-C_y(x_k)^{-1}C_u(x_k)\bar D_k\bar D_k^{-1}(\tilde s_k)_u\|^2 + \|\bar D_k^{-1}(\tilde s_k)_u\|^2 \\ &\le (\nu_6^2\nu_9^2+1)\|\bar D_k^{-1}(\tilde s_k)_u\|^2 \\ &= (\nu_6^2\nu_9^2+1)\,\tilde\delta_k^2 \end{aligned}$$

or, equivalently,

$$\tilde\delta_k \ge \frac{1}{\sqrt{\nu_6^2\nu_9^2+1}}\,\delta_k. \tag{6.9}$$

Define $\psi : \mathbb{R}^+ \longrightarrow \mathbb{R}$ as $\psi(t) = \Psi_k\left(-t\bar D_k\frac{\tilde g_k}{\|\tilde g_k\|}\right) - \Psi_k(0)$. Then $\psi(t) = -\|\tilde g_k\|t + \frac{r_k}{2}t^2$, where $r_k = \frac{\tilde g_k^T\tilde H_k\tilde g_k}{\|\tilde g_k\|^2}$ and $\tilde H_k = \bar D_k\left(W_k^T H_k W_k + E_k\bar D_k^{-2}\right)\bar D_k$. Now we need to minimize ψ in $[0, T_k]$ where T_k is given by

$$T_k = \min\left\{\tilde\delta_k,\ \sigma_k\min\left\{\frac{\|\bar D_k\bar g_k\|}{(\bar g_k)_i} : (\bar g_k)_i > 0\right\},\ \sigma_k\min\left\{-\frac{\|\bar D_k\bar g_k\|}{(\bar g_k)_i} : (\bar g_k)_i < 0\right\}\right\}.$$

Let t_k^* be the minimizer of ψ in $[0, T_k]$. If $t_k^* \in (0, T_k)$ then

$$\psi(t_k^*) = -\frac{1}{2}\frac{\|\tilde{g}_k\|^2}{r_k} \leq -\frac{1}{2}\frac{\|\tilde{g}_k\|^2}{\|\tilde{H}_k\|}. \tag{6.10}$$

If $t_k^* = T_k$ then either $r_k > 0$ in which case $\frac{\|\tilde{g}_k\|}{r_k} \geq T_k$ or $r_k \leq 0$ in which case $r_k T_k \leq \|\tilde{g}_k\|$. In either event,

$$\psi(t_k^*) = \psi(T_k) = -T_k\|\tilde{g}_k\| + \frac{r_k}{2}T_k^2 \leq -\frac{T_k}{2}\|\tilde{g}_k\|. \tag{6.11}$$

We can combine (6.7), (6.10), and (6.11) with

$$\Psi_k(0) - \Psi_k((s_k)_u) \geq \beta_1^{\mathsf{d}}\left(\Psi_k(0) - \Psi_k(c_k^{\mathsf{d}})\right) = -\beta_1^{\mathsf{d}}\psi(t_k^*)$$

to get

$$q_k(s_k^{\mathsf{n}}) - q_k(s_k^{\mathsf{n}} + W_k(s_k)_u) \geq \frac{1}{2}\beta_1^{\mathsf{d}}\|\tilde{g}_k\| \min\left\{\frac{\|\tilde{g}_k\|}{\|\tilde{H}_k\|}, T_k\right\}.$$

The facts that $\sigma_k \geq \sigma$ and $\|\bar{g}_k\| \leq \nu_{11}$ (see (6.3)) imply that

$$\begin{aligned}
&\Psi_k(0) - \Psi_k((s_k)_u) \\
&\quad\geq \frac{1}{2}\beta_1^{\mathsf{d}}\|\bar{D}_k\bar{g}_k\| \min\left\{\frac{\|\bar{D}_k\bar{g}_k\|}{\|\bar{D}_k^T\left(W_k^T H_k W_k + E_k\bar{D}_k^{-2}\right)\bar{D}_k\|}, \min\left\{\tilde{\delta}_k, \frac{\sigma}{\nu_{11}}\|\bar{D}_k\bar{g}_k\|\right\}\right\}.
\end{aligned}$$

To complete the proof, we use (6.8), (6.9), the Assumptions A.1–A.6, and the fact that $\delta_k \leq \delta_{max}$ to establish (6.6) with $\kappa_6 = \frac{1}{2}\min\left\{\beta_1^{\mathsf{d}}, \beta_1^{\mathsf{c}}\right\}$, $\kappa_7 = \min\left\{\frac{1}{\nu_7\nu_6^2\nu_9^2 + \nu_1\nu_6}, \frac{\sigma}{\nu_{11}}\right\}$, and $\kappa_8 = \min\left\{1, \frac{1}{\sqrt{\nu_6^2\nu_9^2+1}}\right\}$. ∎

Now we state the convenient form of the fraction of optimal decrease conditions (5.18) and (5.22).

LEMMA 6.3. *If $(s_k)_u$ satisfies Condition C.2 then*

$$q_k(s_k^{\mathsf{n}}) - q_k(s_k^{\mathsf{n}} + W_k(s_k)_u) \geq \kappa_9\tau_k^2\gamma_k\delta_k^2, \tag{6.12}$$

where κ_9 is a positive constant independent of the iteration k.

Proof. The proof follows immediately from observation (6.7) and conditions (5.19) and (5.23). ∎

We also need the following two inequalities.

LEMMA 6.4. *Under Condition C.1 there exists a positive constant κ_{10} such that*

$$q_k(0) - q_k(s_k^{\mathsf{n}}) - \Delta\lambda_k^T(J_k s_k + C_k) \geq -\kappa_{10}\|C_k\|. \tag{6.13}$$

Moreover if we assume Condition C.3, then

$$q_k(0) - q_k(s_k^{\mathsf{n}}) - \Delta\lambda_k^T(J_k s_k + C_k) \geq -\kappa_{11}\|C_k\|\left(\|s_k^{\mathsf{n}}\| + \|s_k\|\right). \tag{6.14}$$

Proof. The term $q_k(0) - q_k(s_k^{\mathsf{n}})$ can be bounded using (5.4) and $\|s_k^{\mathsf{n}}\| \leq \delta_k$ in the following way:

$$\begin{aligned} q_k(0) - q_k(s_k^{\mathsf{n}}) &= -\nabla_x \ell_k^T s_k^{\mathsf{n}} - \tfrac{1}{2}(s_k^{\mathsf{n}})^T H_k (s_k^{\mathsf{n}}) \\ &\geq -\kappa_1 \left(\|\nabla_x \ell_k\| + \tfrac{1}{2}\delta_k \|H_k\| \right) \|C_k\|. \end{aligned}$$

On the other hand, it follows from $\|J_k s_k + C_k\| \leq \|C_k\|$ that

$$-\Delta\lambda_k^T (J_k s_k + C_k) \geq -\|\Delta\lambda_k\| \, \|C_k\|. \tag{6.15}$$

Combining these two bounds with Assumptions A.3 and A.4 we get (6.13).

To prove (6.14) we first observe that, due to the definition of λ_k in Condition C.3 and to the form (5.1) of the quasi–normal component s_k^{n},

$$\nabla_x \ell_k^T s_k^{\mathsf{n}} = \begin{pmatrix} 0 \\ \nabla_u f_k + C_u(x_k)^T \lambda_k \end{pmatrix}^T \begin{pmatrix} (s_k^{\mathsf{n}})_y \\ 0 \end{pmatrix} = 0. \tag{6.16}$$

Thus

$$q_k(0) - q_k(s_k^{\mathsf{n}}) \geq -\frac{1}{2}\kappa_1 \|H_k\| \, \|C_k\| \, \|s_k^{\mathsf{n}}\| \geq -\frac{1}{2}\kappa_1 \nu_7 \, \|C_k\| \, \|s_k^{\mathsf{n}}\|. \tag{6.17}$$

Also, by appealing to (6.2) and (6.15),

$$-\Delta\lambda_k^T (J_k s_k + C_k) \geq -\nu_{10} \|s_k\| \, \|C_k\|. \tag{6.18}$$

The proof of (6.14) is complete by combining (6.17) and (6.18). ☐

The convergence theory for trust regions traditionally requires consistency of actual and predicted decreases. This is given in the following lemma.

LEMMA 6.5. *Under Condition C.1 there exists a positive constant κ_{12} such that*

$$|ared(s_k;\rho_k) - pred(s_k;\rho_k)| \leq \kappa_{12} \left(\|s_k\|^2 + \rho_k \left(\|s_k\|^3 + \|C_k\| \, \|s_k\|^2 \right) \right). \tag{6.19}$$

Moreover, if Condition C.3 is also valid, then

$$|ared(s_k;\rho_k) - pred(s_k;\rho_k)| \leq \kappa_{13}\rho_k \left(\|s_k\|^3 + \|C_k\| \, \|s_k\|^2 \right). \tag{6.20}$$

Proof. Adding and subtracting $\ell(x_{k+1}, \lambda_k)$ to $ared(s_k;\rho_k) - pred(s_k;\rho_k)$ and using Taylor expansion we obtain

$$\begin{aligned} ared(s_k;\rho_k) - pred(s_k;\rho_k) &= \tfrac{1}{2} s_k^T \left(H_k - \nabla_{xx}^2 \ell(x_k + t_k^1 s_k, \lambda_k) \right) s_k \\ &\quad -\tfrac{1}{2}\textstyle\sum_{i=1}^m (\Delta\lambda_k)_i s_k^T \nabla^2 c_i(x_k + t_k^2 s_k) s_k \\ &\quad -\rho_k \Big(\textstyle\sum_{i=1}^m c_i(x_k + t_k^3 s_k)(s_k)^T \nabla^2 c_i(x_k + t_k^3 s_k)(s_k) \\ &\qquad +(s_k)^T J(x_k + t_k^3 s_k)^T J(x_k + t_k^3 s_k)(s_k) \\ &\qquad -(s_k)^T J(x_k)^T J(x_k)(s_k) \Big), \end{aligned}$$

where t_k^1, t_k^2, and t_k^3 are in $(0,1)$. By expanding $c_i(x_k + t_k^3 s_k)$ around $c_i(x_k)$ and using Assumptions A.3 and A.4 we get (6.19).

The estimate (6.20) follows from (6.2), $\rho_k \geq 1$, and the Lipschitz continuity of the second derivatives. □

The last result in this section is a direct consequence of the scheme that updates ρ_k in Step 2.3 of Algorithms 5.1.

LEMMA 6.6. *The sequence $\{\rho_k\}$ satisfies*

$$\rho_k \geq \rho_{k-1} \geq 1 \quad \textit{and}$$

$$pred(s_k;\rho_k) \geq \frac{\rho_k}{2}\Big(\|C_k\|^2 - \|J_k s_k + C_k\|^2\Big). \tag{6.21}$$

7. Global convergence to a first–order KKT point. The proof of the global convergence to a first–order KKT point (Theorem 7.1) established in this section follows the structure of the convergence theory presented in [11] for the equality–constrained optimization problem. This proof is by contradiction and is based on Condition C.1. We show that the supposition

$$\|\bar{D}_k \bar{g}_k\| + \|C_k\| > \epsilon_{tol},$$

for all k, leads to a contradiction.

The following three lemmas are necessary to bound the predicted decrease.

LEMMA 7.1. *Under Condition C.1 the predicted decrease in the merit function satisfies*

$$\begin{aligned} pred(s_k;\rho) \quad &\geq \quad \kappa_6 \|\bar{D}_k \bar{g}_k\| \min\Big\{\kappa_7 \|\bar{D}_k \bar{g}_k\|, \kappa_8 \delta_k\Big\} \\ &\quad -\kappa_{10}\|C_k\| + \rho\Big(\|C_k\|^2 - \|J_k s_k + C_k\|^2\Big), \end{aligned} \tag{7.1}$$

for every $\rho \geq 0$.

Proof. The inequality (7.1) follows from a direct application of (6.13) and from the lower bound (6.6). □

LEMMA 7.2. *Assume Condition C.1 and $\|\bar{D}_k \bar{g}_k\| + \|C_k\| > \epsilon_{tol}$ are satisfied. If $\|C_k\| \leq \alpha\delta_k$, where α is a positive constant satisfying*

$$\alpha \leq \min\left\{\frac{\epsilon_{tol}}{3\delta_{max}}, \frac{\kappa_6 \epsilon_{tol}}{3\kappa_{10}} \min\left\{\frac{2\kappa_7\epsilon_{tol}}{3\delta_{max}}, \kappa_8\right\}\right\}, \tag{7.2}$$

then

$$pred(s_k;\rho) \geq \frac{\kappa_6}{2}\|\bar{D}_k \bar{g}_k\| \min\Big\{\kappa_7\|\bar{D}_k \bar{g}_k\|, \kappa_8 \delta_k\Big\} + \rho\left(\|C_k\|^2 - \|J_k s_k + C_k\|^2\right), \tag{7.3}$$

for every $\rho \geq 0$.

Proof. From $\|\bar{D}_k \bar{g}_k\| + \|C_k\| > \epsilon_{tol}$ and the first bound on α given by (7.2), we get

$$\|\bar{D}_k \bar{g}_k\| > \frac{2}{3}\epsilon_{tol}. \tag{7.4}$$

If we use this, (7.1), and the second bound on α given by (7.2), we obtain

$$\begin{aligned} pred(s_k;\rho) \;&\geq\; \tfrac{\kappa_6}{2}\|\bar{D}_k\bar{g}_k\| \min\Big\{\kappa_7\|\bar{D}_k\bar{g}_k\|, \kappa_8\delta_k\Big\} + \tfrac{\kappa_6\epsilon_{tol}}{3}\min\Big\{\tfrac{2\kappa_7\epsilon_{tol}}{3}, \kappa_8\delta_k\Big\} \\ &\quad -\kappa_{10}\|C_k\| + \rho\Big(\|C_k\|^2 - \|J_k s_k + C_k\|^2\Big) \\ &\geq\; \tfrac{\kappa_6}{2}\|\bar{D}_k\bar{g}_k\| \min\Big\{\kappa_7\|\bar{D}_k\bar{g}_k\|, \kappa_8\delta_k\Big\} + \rho\Big(\|C_k\|^2 - \|J_k s_k + C_k\|^2\Big). \end{aligned}$$

□

We can use Lemma 7.2 with $\rho = \rho_{k-1}$ and conclude that if $\|\bar{D}_k\bar{g}_k\| + \|C_k\| > \epsilon_{tol}$ and $\|C_k\| \leq \alpha\delta_k$, then the penalty parameter at the current iteration does not need to be increased. See Step 2.3 of Algorithms 5.1. This is equivalent to Lemma 7.7 in [11]. The next lemma states the same result as Lemma 7.8 in [11] but with a different choice of α.

LEMMA 7.3. *Assume Condition C.1 and $\|\bar{D}_k\bar{g}_k\| + \|C_k\| > \epsilon_{tol}$. If $\|C_k\| \leq \alpha\delta_k$, where α satisfies (7.2), then there exists a positive constant $\kappa_{14} > 0$ such that*

$$pred(s_k;\rho_k) \geq \kappa_{14}\delta_k. \tag{7.5}$$

Proof. From (7.3) with $\rho = \rho_k$ and $\|\bar{D}_k\bar{g}_k\| \geq \frac{2}{3}\epsilon_{tol}$, cf. (7.4), we obtain

$$\begin{aligned} pred(s_k;\rho_k) \;&\geq\; \tfrac{\kappa_6\epsilon_{tol}}{3}\min\{\tfrac{2\kappa_7\epsilon_{tol}}{3}, \kappa_8\delta_k\} \\ &\geq\; \tfrac{\kappa_6\epsilon_{tol}}{3}\min\{\tfrac{2\kappa_7\epsilon_{tol}}{3\delta_{max}}, \kappa_8\}\delta_k. \end{aligned}$$

Hence (7.5) holds with

$$\kappa_{14} = \frac{\kappa_6\epsilon_{tol}}{3}\min\left\{\frac{2\kappa_7\epsilon_{tol}}{3\delta_{max}}, \kappa_8\right\}.$$

□

The following lemma is also required.

LEMMA 7.4. *Under Condition C.1, if $\|\bar{D}_k\bar{g}_k\| + \|C_k\| > \epsilon_{tol}$ for all k then the sequences $\{\rho_k\}$ and $\{L_k\}$ are bounded and δ_k is uniformly bounded away from zero.*

Proof. See Lemmas 7.9–7.13, 8.2 in [11]. □

Our first global convergence result follows.

THEOREM 7.1. *Under Condition C.1 the sequences of iterates generated by the trust–region interior–point SQP Algorithms 5.1 satisfy*

$$\liminf_k \Big(\|D_k W_k^T \nabla f_k\| + \|C_k\|\Big) = 0. \tag{7.6}$$

Proof. The proof is by contradiction. Suppose that for all k

$$\|\bar{D}_k\bar{g}_k\| + \|C_k\| > \epsilon_{tol}. \tag{7.7}$$

At each iteration k either $\|C_k\| \leq \alpha\delta_k$ or $\|C_k\| > \alpha\delta_k$, where α satisfies (7.2). In the first case we appeal to Lemmas 7.3 and 7.4 and obtain

$$pred(s_k;\rho_k) \geq \kappa_{14}\delta_*,$$

where δ_* is the lower bound on δ_k given by Lemma 7.4. If $\|C_k\| > \alpha\delta_k$, we have from $\rho_k \geq 1$, (6.1), (6.21), and Lemma 7.4, that

$$pred(s_k;\rho_k) \geq \frac{\kappa_2}{2}\alpha\min\{\kappa_3\alpha, 1\}\delta_*.$$

Hence $pred(s_k;\rho_k) \geq \kappa_{15}$ for all k, where the positive constant κ_{15} does not depend on k. From this and (6.19) we establish

$$\left|\frac{ared(s_k;\rho_k) - pred(s_k;\rho_k)}{pred(s_k;\rho_k)}\right| \;\leq\; \frac{\kappa_{12}}{\kappa_{15}}\left(\|s_k\|^2 + \rho_*\left(\|s_k\|^3 + \|C_k\|\,\|s_k\|^2\right)\right) \;\leq\; \kappa_{16}\delta_k^2,$$

where ρ_* is the upper bound on ρ_k guaranteed by Lemma 7.4. From the rules that update δ_k in Step 2.4 of Algorithms 5.1 this inequality tells us that an acceptable step always is found after a finite number of unsuccessful iterations. Using this fact, we can ignore the rejected steps and work only with successful iterates. So, without loss of generality, we have

$$L_k - L_{k+1} = ared(s_k;\rho_k) \geq \eta_1 pred(s_k;\rho_k) \geq \eta_1\kappa_{15}.$$

Now, if we let k go to infinity, this contradicts the boundedness of $\{L_k\}$ guaranteed by Lemma 7.4. Hence the supposition (7.7) is false, and we must have that

$$\liminf_k \left(\|\bar{D}_k\bar{g}_k\| + \|C_k\|\right) = 0. \tag{7.8}$$

To establish the desired result, we note that $D(x)W(x)^T\nabla f(x)$ is a continuous function of x. For a given bounded $H(x,\lambda)$, let us consider $\bar{D}(x)W(x)^T\left(H(x,\lambda)s^{\mathsf{n}} + \nabla f(x)\right)$, where $\bar{D}(x)$ is defined with the reduced gradient $W(x)^T\left(H(x,\lambda)s^{\mathsf{n}}+\nabla f(x)\right)$. It is then clear that $\bar{D}(x)W(x)^T\left(H(x,\lambda)s^{\mathsf{n}}+\nabla f(x)\right)$ is a continuous function of the pair (x, s^{n}) at $(x, 0)$. From this observation, and since $\|s_k^{\mathsf{n}}\| \leq \kappa_1\|C_k\|$ and $\liminf_{k\to+\infty}\|C_k\| = 0$, we see that the limit (7.8) implies the limit (7.6). □

Using the continuity of $C(x)$, $D(x)W(x)^T\nabla f(x)$, and Theorem 7.1, we can deduce the following result.

COROLLARY 7.1. *Let the conditions of Theorem 7.1 be valid. If $\{x_k\}$ is a bounded sequence, then $\{x_k\}$ has a limit point satisfying the first–order KKT conditions.*

8. Global convergence to a second–order KKT point. In this section we establish global convergence to a point that satisfies the second–order necessary KKT conditions.

THEOREM 8.1. *Under Conditions C.1–C.3, the sequences of iterates generated by the trust–region interior–point SQP Algorithms 5.1 satisfy*

$$\liminf_k \left(\|\bar{D}_k\bar{g}_k\| + \|C_k\| + \tau_k^2\gamma_k\right) = 0, \tag{8.1}$$

where γ_k is the Lagrange multiplier corresponding to the trust–region constraint, see (5.15), (5.20), and τ_k is the damping parameter defined in (5.17).

Proof. The proof is again by contradiction. Suppose that for all k,

$$\|\bar{D}_k \bar{g}_k\| + \|C_k\| + \tau_k^2 \gamma_k > \frac{5}{3}\epsilon_{tol}. \tag{8.2}$$

(i) Suppose that $\|C_k\| \le \alpha' \delta_k$, where

$$\alpha' = \min\left\{\alpha, \frac{\kappa_9 \epsilon_{tol}}{3\kappa_{11}(1+\kappa_4)}\right\} \tag{8.3}$$

and α satisfies (7.2). From the first bound on α in (7.2) we get

$$\|\bar{D}_k \bar{g}_k\| + \tau_k^2 \gamma_k > \frac{4}{3}\epsilon_{tol}.$$

Thus, either $\|\bar{D}_k \bar{g}_k\| > \frac{2}{3}\epsilon_{tol}$ or $\tau_k^2 \gamma_k > \frac{2}{3}\epsilon_{tol}$. In the first case we proceed exactly as in Lemmas 7.2, 7.3 and obtain

$$\begin{aligned} pred(s_k;\rho) &\ge \frac{\kappa_6}{2}\|\bar{D}_k \bar{g}_k\| \min\left\{\kappa_7 \|\bar{D}_k \bar{g}_k\|, \kappa_8 \delta_k\right\} + \rho\left(\|C_k\|^2 - \|J_k s_k + C_k\|^2\right) \\ &\ge \frac{\kappa_{14}}{\delta_{max}}\delta_k^2 \end{aligned} \tag{8.4}$$

for every $\rho > 0$. If $\tau_k^2 \gamma_k > \frac{2}{3}\epsilon_{tol}$ then from (6.4), (6.12), (6.14), $\|s_k^{\mathrm{n}}\| \le \delta_k$, and the second bound on α' given in (8.3), we can write

$$\begin{aligned} pred(s_k;\rho) &= q_k(s_k^{\mathrm{n}}) - q_k(s_k^{\mathrm{n}} + W_k(s_k)_u) + q_k(0) - q_k(s_k^{\mathrm{n}}) - \Delta\lambda_k^T(J_k s_k + C_k) \\ &\quad + \rho\left(\|C_k\|^2 - \|J_k s_k + C_k\|^2\right) \\ &\ge \frac{1}{2}\kappa_9 \tau_k^2 \gamma_k \delta_k^2 + \left(\frac{1}{3}\kappa_9 \epsilon_{tol}\delta_k - \kappa_{11}\|C_k\|(1+\kappa_4)\right)\delta_k + \rho\left(\|C_k\|^2 - \|J_k s_k + C_k\|^2\right) \\ &\ge \frac{1}{2}\kappa_9 \tau_k^2 \gamma_k \delta_k^2 + \rho\left(\|C_k\|^2 - \|J_k s_k + C_k\|^2\right) \\ &\ge \frac{\kappa_9 \epsilon_{tol}}{3}\delta_k^2 \end{aligned} \tag{8.5}$$

for every $\rho > 0$. From the two bounds (8.4), (8.5), we conclude that if $\|C_k\| \le \alpha' \delta_k$ then the penalty parameter does not increase. See Step 2.3 of Algorithms 5.1. Moreover, these two bounds on $pred(s_k;\rho_k)$ show the existence of a positive constant κ_{17} independent of k such that

$$pred(s_k;\rho_k) \ge \kappa_{17}\delta_k^2, \tag{8.6}$$

provided $\|C_k\| \le \alpha' \delta_k$.

(ii) Now we prove that $\{\rho_k\}$ is bounded. If ρ_k is increased at iteration k, then it is updated according to the rule

$$\rho_k = 2\left(\frac{q_k(s_k) - q_k(0) + \Delta\lambda_k^T(J_k s_k + C_k)}{\|C_k\|^2 - \|J_k s_k + C_k\|^2}\right) + \bar{\rho}.$$

We can write

$$
\begin{aligned}
\tfrac{\rho_k}{2}\left(\|C_k\|^2 - \|J_k s_k + C_k\|^2\right) &= q_k(s_k) - q_k(s_k^{\mathsf{n}}) \\
&\quad -\left(q_k(0) - q_k(s_k^{\mathsf{n}})\right) + \Delta\lambda_k^T (J_k s_k + C_k) \\
&\quad +\tfrac{\bar\rho}{2}\left(\|C_k\|^2 - \|J_k s_k + C_k\|^2\right).
\end{aligned}
$$

By applying (6.1) to the left hand side and (6.4), (6.12), (6.14), and $\|s_k^{\mathsf{n}}\| \le \delta_k$ to the right hand side, we obtain

$$
\begin{aligned}
\frac{\rho_k}{2}\kappa_2\|C_k\|\min\{\kappa_3\|C_k\|, \delta_k\} &\le \kappa_{11}(1+\kappa_4)\delta_k\|C_k\| + \frac{\bar\rho}{2}\left(-2(J_k^T C_k)^T s_k - \|J_k s_k\|^2\right) \\
&\le (\kappa_{11}(1+\kappa_4) + \bar\rho\nu_4\kappa_4)\delta_k\|C_k\|.
\end{aligned}
\tag{8.7}
$$

If ρ_k is increased at iteration k, then, because of part (i), $\|C_k\| > \alpha'\delta_k$. Now we use this fact to establish that

$$\left(\frac{\kappa_2}{2}\min\{\kappa_3\alpha', 1\}\right)\rho_k \le \kappa_{11}(1+\kappa_4) + \bar\rho\nu_4\kappa_4.$$

This proves that $\{\rho_k\}$ and $\{L_k\}$ are bounded sequences.

(iii) The next step is to prove that δ_k is bounded away from zero.

If s_{k-1} was an acceptable step, then $\delta_k \ge \delta_{min}$, see Step 2.4 in Algorithms 5.1.

If s_{k-1} was a rejected, then $\delta_k \ge \kappa_5\|s_{k-1}\|$, see (6.5). We consider two cases. In both cases we will use the fact that

$$1 - \eta_1 \le \left|\frac{ared(s_{k-1};\rho_{k-1})}{pred(s_{k-1};\rho_{k-1})} - 1\right|.$$

In the first case we will assume that $\|C_{k-1}\| \le \alpha'\delta_{k-1}$. From (8.6) we have $pred(s_{k-1};\rho_{k-1}) \ge \kappa_{17}\delta_{k-1}^2$. Thus we can use $\|s_{k-1}\| \le \kappa_4\delta_{k-1}$, see (6.4), and (6.20) with k replaced by $k-1$ to obtain

$$\left|\frac{ared(s_{k-1};\rho_{k-1})}{pred(s_{k-1};\rho_{k-1})} - 1\right| \le \frac{\kappa_{13}\rho_*\left(\kappa_4^2\delta_{k-1}^2 + \kappa_4\alpha'\delta_{k-1}^2\right)}{\kappa_{17}\delta_{k-1}^2}\|s_{k-1}\|.$$

This gives $\delta_k \ge \kappa_5\|s_{k-1}\| \ge \frac{\kappa_5(1-\eta_1)\kappa_{17}}{\kappa_{13}\rho_*(\kappa_4^2+\alpha'\kappa_4)} \equiv \kappa_{18}$.

The other case is $\|C_{k-1}\| > \alpha'\delta_{k-1}$. In this case we get from (6.1) and (6.21) with k replaced by $k-1$ that

$$
\begin{aligned}
pred(s_{k-1};\rho_{k-1}) &\ge \tfrac{\rho_{k-1}}{2}\kappa_2\|C_{k-1}\|\min\{\kappa_3\|C_{k-1}\|, \delta_{k-1}\} \\
&\ge \rho_{k-1}\kappa_{19}\delta_{k-1}\|C_{k-1}\| \\
&\ge \rho_{k-1}\alpha'\kappa_{19}\delta_{k-1}^2,
\end{aligned}
$$

where $\kappa_{19} = \frac{\kappa_2}{2}\min\{\kappa_3\alpha', 1\}$. Again we use $\rho_{k-1} \ge 1$ and (6.20) with k replaced by $k-1$, this time with the last two lower bounds on $pred(s_{k-1};\rho_{k-1})$, and we write

$$
\begin{aligned}
\left|\frac{ared(s_{k-1};\rho_{k-1})}{pred(s_{k-1};\rho_{k-1})} - 1\right| &\le \frac{\kappa_{13}\rho_{k-1}\|s_{k-1}\|^3}{|pred(s_{k-1};\rho_{k-1})|} + \frac{\kappa_{13}\rho_{k-1}\|C_{k-1}\|\,\|s_{k-1}\|^2}{|pred(s_{k-1};\rho_{k-1})|} \\
&\le \left(\frac{\kappa_{13}\rho_{k-1}\kappa_4^2\delta_{k-1}^2}{\rho_{k-1}\alpha'\kappa_{19}\delta_{k-1}^2} + \frac{\kappa_{13}\rho_{k-1}\kappa_4\delta_{k-1}\|C_{k-1}\|}{\rho_{k-1}\kappa_{19}\delta_{k-1}\|C_{k-1}\|}\right)\|s_{k-1}\|.
\end{aligned}
$$

Hence $\delta_k \geq \kappa_5 \|s_{k-1}\| \geq \frac{\kappa_5(1-\eta_1)\alpha'\kappa_{19}}{\kappa_{13}(\kappa_4^2+\alpha'\kappa_4)} \equiv \kappa_{20}$.

Combining the two cases yields

$$\delta_k \geq \delta_* = \min\{\delta_{min}, \kappa_{18}, \kappa_{20}\}$$

for all k.

(iv) The rest of the proof consists of proving that an acceptable trial step is always found after a finite number of iterations and then from this concluding that the supposition (8.2) is false. The proof of these facts is exactly the proof of Theorem 7.1 where α is now α' and $\kappa_{14}\delta_*$ is replaced by $\kappa_{17}\delta_*^2$. ∎

The following result finally establishes global convergence to a point satisfying the second–order necessary KKT conditions. The proof uses ideas applied in [9, Lem. 3.8]. However, we show that convergence to a limit point satisfies the second–order necessary conditions even in the degenerate case.

THEOREM 8.2. *Let $\{x_k\}$ be a bounded sequence of iterates generated by the trust–region interior–point SQP Algorithms 5.1 under Conditions C.1–C.3. Then $\{x_k\}$ has a limit point x_* satisfying the first–order KKT conditions. Furthermore, x_* satisfies the second–order necessary KKT conditions.*

Proof. Consider the subsequence of $\{x_k\}$ for which the limit in (8.1) is zero. Since this subsequence is bounded we can use the same arguments as in the proof of Theorem 7.1 to show that it has a convergent subsequence indexed by $\{k_j\}$ such that

$$\lim_j \left(\|\bar{D}_{k_j} \bar{g}_{k_j}\| + \|C_{k_j}\| \right) = \lim_j \left(\|D_{k_j} W_{k_j}^T \nabla f_{k_j}\| + \|C_{k_j}\| \right) = 0. \tag{8.8}$$

Moreover,

$$\lim_j \tau_{k_j}^2 \gamma_{k_j} = 0, \tag{8.9}$$

where τ_{k_j} is given by (5.17). Let x_* denote the limit of $\{x_{k_j}\}$. It follows from (8.8) and the continuity of $C(x)$ and $D(x)W(x)^T \nabla f(x)$ that x_* satisfies the first–order KKT conditions.

Next, we will prove that $\lim_j \gamma_{k_j} = 0$. First we consider the decoupled approach. Define the vector valued function h as follows:

$$h(x)_i = \begin{cases} 1 & \text{if } \left(W(x)^T \nabla f(x)\right)_i = 0 \text{ and } \left(D(x)_{ii}\right) = 0, \\ \left(W(x)^T \nabla f(x)\right)_i & \text{otherwise,} \end{cases}$$

for all $i = 1, \ldots, n-m$. The function h is used to identify the active indices. By definition of h and since x_* satisfies the first–order KKT conditions, the implications

$$D(x_*)_{ii} = 0 \iff h(x_*)_i \neq 0, \quad i = 1, \ldots, n-m \tag{8.10}$$

are valid. (If x_* is nondegenerate then $h(x_*) = W(x_*)^T \nabla f(x_*)$.) Moreover,

$$\lim_{x \to x_*} D(x)h(x) = 0. \tag{8.11}$$

Since $\lim_j x_{k_j} = x_*$, (8.10) implies the existence of $\epsilon_0 \in (0,1)$ such that

$$\min\left\{ (u_{k_j})_i - a_i,\, b_i - (u_{k_j})_i \right\} + \left| \left(h_{k_j} \right)_i \right| > 2\epsilon_0, \qquad i = 1, \ldots, n-m \tag{8.12}$$

for large enough j, and

$$2\epsilon_0 < \min\{b_i - a_i,\, i = 1, \ldots, n-m\}.$$

Without loss of generality, we will only consider the cases where $\tau_{k_j} \le \sigma_{k_j} < 1$. In the following the index i will be the index defining τ_{k_j} in (5.17). (The index i is really i_j but we drop the j from i_j to alleviate the notation.) We also assume that j is large enough such that

$$\left| \left(\bar{D}_{k_j}^2 h_{k_j} \right)_i \right| < \epsilon_0^2, \tag{8.13}$$

cf. (8.11).

Multiplying both sides of (5.16) by $\bar{D}_{k_j}^2$ gives

$$\left(E_{k_j} + \gamma_{k_j} I_{n-m} \right) o_{k_j}^{\mathsf{d}} = \bar{D}_{k_j}^2 \left(-\bar{g}_{k_j} - W_{k_j}^T H_{k_j} W_{k_j} o_{k_j}^{\mathsf{d}} \right),$$

which in turn implies

$$\gamma_{k_j} |(o_{k_j}^{\mathsf{d}})_i| \le (\bar{D}_{k_j}^2)_{ii} \left| \left(-\bar{g}_{k_j} - W_{k_j}^T H_{k_j} W_{k_j} o_{k_j}^{\mathsf{d}} \right)_i \right|. \tag{8.14}$$

Also, Assumption A.6 implies $\|o_{k_j}^{\mathsf{d}}\| \le \nu_9 \delta_{k_j} \le \nu_9 \delta_{max}$. From this, (6.3), and Assumptions A.3–A.4, we can write

$$\frac{1}{(o_{k_j}^{\mathsf{d}})_i} \ge \frac{\gamma_{k_j}}{\kappa_{21} (\bar{D}_{k_j})_{ii}^2} \tag{8.15}$$

for some κ_{21} independent of k. Now we distinguish between two cases.

In the first case we consider $\left| \left(h_{k_j} \right)_i \right| \le \epsilon_0$ and appeal to (8.12) to get $\min\{(u_{k_j})_i - a_i,\, b_i - (u_{k_j})_i\} > \epsilon_0$. Thus from (8.15) and the definition (5.17) of τ_{k_j} we obtain

$$\tau_{k_j} \ge \frac{\sigma_{k_j} \gamma_{k_j} \epsilon_0}{\kappa_{21} (\bar{D}_{k_j})_{ii}^2}. \tag{8.16}$$

Now we analyze the case $\left| \left(h_{k_j} \right)_i \right| > \epsilon_0$. Two possibilities can occur.

(i) The first possibility is that the value of the numerator defining τ_{k_j} is equal to $(\bar{D}_{k_j})_{ii}^2$. In this situation (8.15) immediately implies

$$\tau_{k_j} \ge \frac{\sigma_{k_j} \gamma_{k_j}}{\kappa_{21}}. \tag{8.17}$$

(ii) The other possibility is that the value of the numerator defining τ_{k_j} is not equal to $(\bar{D}_{k_j})_{ii}^2$. In this case we have from (8.13) that $(\bar{D}_{k_j})_{ii}^2 < \epsilon_0$ and since $b_i - a_i > 2\epsilon_0$, the numerator in the definition (5.17) of τ_{k_j} is bigger than ϵ_0. Thus

$$\tau_{k_j} \ge \frac{\sigma_{k_j} \gamma_{k_j} \epsilon_0}{\kappa_{21} (\bar{D}_{k_j})_{ii}^2}. \tag{8.18}$$

Using (8.9), (8.16), (8.17), (8.18), $\sigma_{k_j} \geq \sigma$, and the boundedness of $\bar{D}_{k_j}$ this proves that

$$\lim_j \gamma_{k_j} = 0.$$

By (5.15) we know that

$$\bar{D}_{k_j} W_{k_j}^T H_{k_j} W_{k_j} \bar{D}_{k_j} + E_{k_j} + \gamma_{k_j} I_{n-m}$$

is positive semi–definite. Hence condition (8.8), the continuity of $W(x)^T \nabla_{xx}^2 \ell(x, \lambda) W(x)$, the limits $\lim_j \|W_{k_j}^T H_{k_j} s_{k_j}^{\mathsf{n}}\| = 0$ and $\lim_j \gamma_{k_j} = 0$ imply that the limit of the principal submatrix of $W_{k_j}^T H_{k_j} W_{k_j}$ corresponding to indices l such that $a_l < (u_*)_l < b_l$ is positive semi–definite. Hence, the second–order necessary KKT conditions are satisfied at x_*. This completes the proof for the decoupled approach.

The proof for the coupled trust–region approach differs only from the proof for the decoupled approach in the use of equations (5.20) and (5.21) and in the use of $\|W_{k_j} o_{k_j}^{\mathsf{c}}\| \leq (1 + \nu_9)\delta_{max}$ to bound the right hand side of inequality (8.14). □

REMARK 8.1. The global convergence results of Sections 7 and 8 hold true if the quadratic $\Psi_k(s_u)$ is redefined as $\Psi_k(s_u) = q_k(s_k^{\mathsf{n}} + W_k s_u)$ (see (5.7) and (5.8)) without the Newton augmentation term $\frac{1}{2} s_u^T \left(E_k \bar{D}_k^{-2} \right) s_u$. They are valid also if the matrices D_k and $\bar{D}_k$ are redefined respectively as D_k^p and $\bar{D}_k^p$ with $p \geq 1$.

9. Local rate of convergence. We will now analyze the local behavior of Algorithms 5.1 under Conditions C.1, C.3, and C.4. We start by looking at the behavior of the trust radius close to a nondegenerate point that satisfies the second–order sufficient KKT conditions. For this purpose we require the following lemma.

LEMMA 9.1. *Under Condition C.1 the quasi–normal component satisfies*

$$\|s_k^{\mathsf{n}}\| \leq \kappa_{22} \|s_k\|, \tag{9.1}$$

where κ_{22} is positive and independent of the iteration counter k.

Proof. From $s_k = s_k^{\mathsf{n}} + W_k (s_k)_u$, we obtain

$$\|s_k^{\mathsf{n}}\| \leq \|s_k\| + \|W_k\| \, \|(s_k)_u\|.$$

But since $\|s_k\|^2 = \|(s_k)_y\|^2 + \|(s_k)_u\|^2$, we use Assumption A.4 to obtain

$$\|s_k^{\mathsf{n}}\| \leq (1 + \nu_6) \, \|s_k\|,$$

and (9.1) holds with $\kappa_{22} = 1 + \nu_6$. □

THEOREM 9.1. *Let $\{x_k\}$ be a sequence of iterates generated by the trust–region interior–point SQP Algorithms 5.1 under Conditions C.1 and C.3. If x_k converges to a nondegenerate point x_* satisfying the second–order sufficient KKT conditions, then δ_k is uniformly bounded away from zero and eventually all the iterations will be successful.*

Proof. It follows from $\lim_{k \to +\infty} x_k = x_*$ and $C(x_*) = 0$ that $\lim_{k \to +\infty} \|C_k\| = 0$. This fact, condition (5.4), and Assumptions A.3–A.4, together imply $\lim_{k \to +\infty} \|W_k^T H_k s_k^{\mathsf{n}}\| = 0$. Since x_k converges to a nondegenerate point that satisfies the second–order sufficient KKT conditions and

$\lim_{k\to+\infty} \|W_k^T H_k s_k^{\mathsf{n}}\| = 0$, there exists a $\bar{\gamma} > 0$ such that the smallest eigenvalue of $\bar{D}_k W_k^T H_k W_k \bar{D}_k + E_k$ is greater than $\bar{\gamma}$ for k sufficiently large.

First we will proof that $\{\rho_k\}$ is a bounded sequence. Since $\Psi_k(0) - \Psi_k((s_k)_u) \geq 0$, we obtain

$$\begin{aligned}\tfrac{1}{2}(\bar{D}_k^{-1}(s_k)_u)^T \left(\bar{D}_k W_k^T H_k W_k \bar{D}_k + E_k\right)(\bar{D}_k^{-1}(s_k)_u) &\leq -(\bar{D}_k^{-1}(s_k)_u)^T(\bar{D}_k \bar{g}_k) \\ &\leq \|\bar{D}_k^{-1}(s_k)_u\| \, \|\bar{D}_k \bar{g}_k\|,\end{aligned}$$

which, by using the upper bounds on W_k and $\bar{D}_k$ given by Assumptions A.4 and A.6, implies

$$\|s_k^{\mathsf{t}}\| = \|W_k(s_k)_u\| \leq \frac{2\nu_6\nu_9}{\bar{\gamma}} \|\bar{D}_k \bar{g}_k\|. \tag{9.2}$$

Using (6.6) and (9.2), we find that

$$\begin{aligned} q_k(s_k^{\mathsf{n}}) - q_k(s_k^{\mathsf{n}} + W_k(s_k)_u) &\geq \kappa_6 \|\bar{D}_k \bar{g}_k\| \min\{\kappa_7 \|\bar{D}_k \bar{g}_k\|, \kappa_8 \delta_k\} \\ &\geq \kappa_{23} \|s_k^{\mathsf{t}}\|^2, \end{aligned} \tag{9.3}$$

where $\kappa_{23} = \frac{\kappa_6 \bar{\gamma}}{2\nu_6\nu_9} \min\{\frac{\kappa_7\bar{\gamma}}{2\nu_6\nu_9}, \frac{\kappa_8}{\nu_6\nu_9}, \frac{\kappa_8}{1+\nu_9}\}$ accounts for the decoupled and coupled cases.

Next, we prove that if $\|C_k\| \leq \alpha''\|s_k\|$, where α'' will be defined later, then the penalty parameter does not need to be increased. From (5.4) and $\|C_k\| \leq \alpha''\|s_k\|$, we get

$$\begin{aligned}\|s_k\|^2 \leq (\|s_k^{\mathsf{n}}\| + \|s_k^{\mathsf{t}}\|)^2 &\leq 2\|s_k^{\mathsf{n}}\|^2 + 2\|s_k^{\mathsf{t}}\|^2 \\ &\leq 2\alpha''\kappa_1^2 \|C_k\| \, \|s_k\| + 2\|s_k^{\mathsf{t}}\|^2.\end{aligned}$$

This estimate, (5.4), (6.14), (9.3), and $\|C_k\| \leq \alpha''\|s_k\|$ yield

$$\begin{aligned} pred(s_k;\rho) &= q_k(s_k^{\mathsf{n}}) - q_k(s_k^{\mathsf{n}} + W_k(s_k)_u) + q_k(0) - q_k(s_k^{\mathsf{n}}) - \Delta\lambda_k^T(J_k s_k + C_k) \\ &\quad + \rho\left(\|C_k\|^2 - \|J_k s_k + C_k\|^2\right) \\ &\geq \frac{1}{4}\kappa_{23}\|s_k\|^2 + \left(\frac{1}{4}\kappa_{23}\|s_k\| - (\alpha''\kappa_1^2\kappa_{23} + \kappa_{11}(\alpha''\kappa_1 + 1))\|C_k\|\right)\|s_k\| \\ &\quad + \rho\left(\|C_k\|^2 - \|J_k s_k + C_k\|^2\right), \end{aligned} \tag{9.4}$$

for every $\rho > 0$. If $\|C_k\| \leq \alpha''\|s_k\|$, where α'' satisfies

$$(4\kappa_{11})\,\alpha'' + \left(4\kappa_1^2\kappa_{23} + 4\kappa_1\kappa_{11}\right)(\alpha'')^2 \leq \kappa_{23}\,, \tag{9.5}$$

then we set $\rho = \rho_{k-1}$ in (9.4) and deduce that the penalty parameter does not need to be increased. See Step 2.3 of Algorithms 5.1. Hence if ρ_k is increased then the inequality $\|C_k\| > \alpha''\|s_k\|$ must hold, and we can proceed as in Theorem 8.1, equation (8.7), and write

$$\frac{\rho_k}{2}\kappa_2\|C_k\| \min\left\{\kappa_3\|C_k\|, \frac{1}{\kappa_4}\|s_k\|\right\} \leq (\kappa_{11}(\kappa_{22}+1) + \bar{\rho}\nu_4)\|s_k\| \, \|C_k\|,$$

(here we used inequality (9.1)) which in turn implies

$$\left(\frac{\kappa_2}{2}\min\left\{\kappa_3\alpha'', \frac{1}{\kappa_4}\right\}\right)\rho_k \leq \kappa_{11}(\kappa_{22}+1) + \bar{\rho}\nu_4.$$

This gives the uniform boundedness of the penalty parameter:

$$\rho_k \le \rho_*$$

for all k.

Given the boundedness of $\{\rho_k\}$ we can complete the proof of the theorem. If $\|C_k\| > \alpha''\|s_k\|$, where α'' satisfies (9.5), then from (6.1) and (6.21) we find that

$$pred(s_k;\rho_k) \ge \rho_k \frac{\kappa_2}{2}\|C_k\| \min\{\kappa_3\|C_k\|, \delta_k\} \ge \rho_k\kappa_{24}\|s_k\|^2, \tag{9.6}$$

where $\kappa_{24} = \frac{\kappa_2\alpha''}{2}\min\{\kappa_3\alpha'', \frac{1}{\kappa_4}\}$. In this case it follows from (6.20) and (9.6) that

$$\left|\frac{ared(s_k;\rho_k)}{pred(s_k;\rho_k)} - 1\right| \le \frac{\kappa_{13}}{\kappa_{24}}\left(\|s_k\| + \|C_k\|\right). \tag{9.7}$$

Now, suppose that $\|C_k\| \le \alpha''\|s_k\|$. From (9.4) with $\rho = \rho_k$ we obtain $pred(s_k;\rho_k) \ge \frac{\kappa_{23}}{4}\|s_k\|^2$. Now we use (6.20) and $\rho_k \le \rho_*$, to get

$$\left|\frac{ared(s_k;\rho_k)}{pred(s_k;\rho_k)} - 1\right| \le \frac{4\kappa_{13}\rho_*}{\kappa_{23}}\left(\|s_k\| + \|C_k\|\right). \tag{9.8}$$

Finally from (9.7), (9.8), $\lim_{k\to+\infty} x_k = x_*$, and $\lim_{k\to+\infty}\|C_k\| = 0$, we get

$$\lim_{k\to+\infty}\frac{ared(s_k;\rho_k)}{pred(s_k;\rho_k)} = 1,$$

which by the rules for updating the trust radius given in Step 2.4 of Algorithms 5.1, shows that δ_k is uniformly bounded away from zero. ∎

We use the following straightforward globalization of the quasi–normal component s_k^{n} of the Newton step given in (5.27). The new quasi–normal component is given by:

$$s_k^{\mathsf{n}} = \begin{pmatrix} -\xi_k C_y(x_k)^{-1}C_k \\ 0 \end{pmatrix}, \tag{9.9}$$

where

$$\xi_k = \begin{cases} 1 & \text{if } \|C_y(x_k)^{-1}C_k\| \le \delta_k, \\ \frac{\delta_k}{\|C_y(x_k)^{-1}C_k\|} & \text{otherwise.} \end{cases} \tag{9.10}$$

Before we state the q–quadratic rate of convergence we prove the following important result.

LEMMA 9.2. *The quasi–normal component (9.9) satisfies conditions (5.1), (5.4), and (5.5) for some positive κ_1, κ_2, and κ_3 independent of k.*

Proof. It is obvious that (5.1) holds. Condition (5.4) is a direct consequence of the condition (5.5). In fact, using $\|C_y(x_k)(s_k^{\mathsf{n}})_y + C_k\| \le \|C_k\|$ and the boundedness of $\{C_y(x_k)^{-1}\}$ we find that

$$\begin{aligned} \|s_k^{\mathsf{n}}\| &= \|s_k^{\mathsf{n}} + C_y(x_k)^{-1}C_k - C_y(x_k)^{-1}C_k\| \\ &\le \|C_y(x_k)^{-1}\|\Big(\|C_y(x_k)(s_k^{\mathsf{n}})_y + C_k\| + \|C_k\|\Big) \;\le\; 2\nu_6\,\|C_k\|. \end{aligned} \tag{9.11}$$

So, let us prove (5.5). A simple manipulation shows that

$$
\begin{aligned}
\|C_k\|^2 - \|C_y(x_k)(s_k^{\mathsf{n}})_y + C_k\|^2 &= \|C_k\|^2 - \| - \xi_k C_y(x_k) C_y(x_k)^{-1} C_k + C_k\|^2 \\
&= \|C_k\|^2 - \Big((1-\xi_k)\|C_k\|\Big)^2 \\
&= \xi_k(2-\xi_k)\|C_k\|^2 \;\geq\; \xi_k\,\|C_k\|^2.
\end{aligned}
$$

We need to consider two cases. If $\xi_k = 1$, then

$$\|C_k\|^2 - \|C_y(x_k)(s_k^{\mathsf{n}})_y + C_k\|^2 \geq \|C_k\| \min\{\|C_k\|, \delta_k\}.$$

Otherwise, $\xi_k = \frac{\delta_k}{\|C_y(x_k)^{-1}C_k\|}$. In this case we get

$$\|C_k\|^2 - \|C_y(x_k)(s_k^{\mathsf{n}})_y + C_k\|^2 \geq \frac{1}{\nu_6}\|C_k\|\,\delta_k \geq \frac{1}{\nu_6}\|C_k\| \min\{\|C_k\|, \delta_k\}.$$

Thus the result holds with $\kappa_2 = \min\{1, \frac{1}{\nu_6}\}$ and $\kappa_3 = 1$. □

COROLLARY 9.1. *Let $\{x_k\}$ be a sequence of iterates generated by the trust–region interior–point SQP Algorithms 5.1 under Conditions C.1, C.3, and C.4. If x_k converges to a nondegenerate point x_* satisfying the second–order sufficient KKT conditions, then x_k converges q–quadratically.*

Proof. We start by showing that $|\tau_k^{\mathsf{N}} - 1|$ is $\mathcal{O}\left(\|x_k - x_*\|\right)$, where τ_k^{N} is given by (5.29). Since $\lim_{k\to+\infty} \|W_k^T H_k s_k^{\mathsf{n}}\| = 0$, we have that $\left|\frac{\tau_k^{\mathsf{N}}}{\sigma_k} - 1\right|$ is $\mathcal{O}(\|(s_k^{\mathsf{N}})_u\|)$ (see [8, Eq. (6.4) and Lem. 12]). Also since by Condition C.4 $|\sigma_k - 1|$ is $\mathcal{O}\left(\|\bar{D}_k \bar{g}_k\|\right)$, and $\bar{D}_k \bar{g}_k$ is $\mathcal{O}\left(\|(s_k^{\mathsf{N}})_u\|\right)$ (see (5.26)), we can see that $|\sigma_k - 1|$ is also $\mathcal{O}\left(\|(s_k^{\mathsf{N}})_u\|\right)$. Furthermore,

$$|\tau_k^{\mathsf{N}} - 1| \leq \sigma_k \left|\frac{\tau_k^{\mathsf{N}}}{\sigma_k} - 1\right| + |\sigma_k - 1|\,.$$

Hence $|\tau_k^{\mathsf{N}} - 1|$ is $\mathcal{O}\left(\|(s_k^{\mathsf{N}})_u\|\right)$. But $(s_k^{\mathsf{N}})_u$ is $\mathcal{O}\left(\|x_k + s_k^{\mathsf{n}} - x_*\|\right)$ and s_k^{n} is $\mathcal{O}\left(\|x_k - x_*\|\right)$ and this shows that $|\tau_k^{\mathsf{N}} - 1|$ is $\mathcal{O}\left(\|x_k - x_*\|\right)$.

We need to prove that Condition C.4 does not conflict with Condition C.1 so that Theorem 9.1 can be applied. In other words, we need to show that the decrease conditions given in Condition C.1 hold for the Newton damped step (5.28) whenever it is taken. In Lemma 9.2 we showed that the quasi–normal component s_k^{n} given in (9.9) satisfies (5.1), (5.4), and (5.5). From Condition C.4, s_k^{n} given by (5.27) is used when it coincides with the s_k^{n} given by (9.9). Thus s_k^{n} given by (5.27) satisfies also (5.1), (5.4), and (5.5). It remains to prove that $\tau_k^{\mathsf{N}}(s_k^{\mathsf{N}})_u$ satisfies the Cauchy decrease condition (5.13) ((5.14) for the coupled approach). This is indeed the case since

$$
\begin{aligned}
\Psi_k(0) - \Psi_k(\tau_k^{\mathsf{N}}(s_k^{\mathsf{N}})_u) &\geq -\tau_k^{\mathsf{N}} \bar{g}_k^T (s_k^{\mathsf{N}})_u - \frac{1}{2}(\tau_k^{\mathsf{N}})^2 ((s_k^{\mathsf{N}})_u)^T \left(W_k^T H_k W_k + E_k \bar{D}_k^{-2}\right)((s_k^{\mathsf{N}})_u) \\
&\geq \tau_k^{\mathsf{N}} \left(-\bar{g}_k^T (s_k^{\mathsf{N}})_u - \frac{1}{2}((s_k^{\mathsf{N}})_u)^T \left(W_k^T H_k W_k + E_k \bar{D}_k^{-2}\right)((s_k^{\mathsf{N}})_u)\right) \\
&\geq \tau_k^{\mathsf{N}} \left(\Psi_k(0) - \Psi_k(c_k^{\mathsf{d}})\right),
\end{aligned}
$$

and $|\tau_k^{\mathsf{N}} - 1|$ is $\mathcal{O}(\|x_k - x_*\|)$.

Now we need to show that eventually s_k is given by (5.28). Since $\{x_k\}$ converges to a nondegenerate point satisfying the second–order sufficient KKT conditions, $(s_k^{\mathsf{N}})_u$ exists for k sufficiently large. Furthermore $(s_k^{\mathsf{n}})_y = -C_y(x_k)^{-1}C_k$ for k large enough because $\lim_{k\to+\infty}\|C_y(x_k)^{-1}C_k\| = 0$, and from Theorem 9.1, δ_k is eventually bounded away from zero. Using a similar argument we see that $\tau_k^{\mathsf{N}}(s_k^{\mathsf{N}})_u$ is inside the trust region (5.10) for the decoupled approach or (5.12) for the coupled approach. So, from Condition C.4 we conclude that there exists a positive integer $\bar{k}$ such that s_k is given by (5.28) for $k \geq \bar{k}$.

Using the fact that $(s_k^{\mathsf{N}})_u$ is $\mathcal{O}(\|x_k - x_*\|)$, we conclude that $\tau_k^{\mathsf{N}}(s_k^{\mathsf{N}})_u - (s_k^{\mathsf{N}})_u$ is $\mathcal{O}(\|x_k - x_*\|^2)$. Thus

$$s_k - s_k^{\mathsf{N}} = \left(\begin{array}{c} s_k^{\mathsf{n}} - C_y(x_k)^{-1}C_u(x_k)\tau_k^{\mathsf{N}}(s_k^{\mathsf{N}})_u \\ \tau_k^{\mathsf{N}}(s_k^{\mathsf{N}})_u \end{array}\right) - \left(\begin{array}{c} s_k^{\mathsf{n}} - C_y(x_k)^{-1}C_u(x_k)(s_k^{\mathsf{N}})_u \\ (s_k^{\mathsf{N}})_u \end{array}\right)$$

is $\mathcal{O}(\|x_k - x_*\|^2)$. This completes the proof since s_k^{N} can be seen as a Newton step on a given vector function of the type (4.7). This function vanishes at x_* and is continuously differentiable with Lipschitz continuous derivatives and a nonsingular Jacobian matrix in an open neighborhood of x_*. See the discussion at the end of Section 4. Thus the q–quadratic rate of convergence follows from [13][Thm. 5.2.1] and from the fact that $s_k - s_k^{\mathsf{N}}$ is $\mathcal{O}(\|x_k - x_*\|^2)$. $\square$

10. Trial steps and multiplier estimates. When we described the trust–region interior–point SQP algorithms, we deferred the practical computation of the quasi–normal and tangential components and of the multiplier estimates. In the following sections we address these issues.

10.1. Computation of the quasi–normal component. The quasi–normal component s_k^{n} is an approximate solution of the trust–region subproblem

$$\begin{array}{lll} (10.1) \qquad & \text{minimize} & \frac{1}{2}\|C_y(x_k)(s^{\mathsf{n}})_y + C_k\|^2 \\ & \text{subject to} & \|(s^{\mathsf{n}})_y\| \leq \delta_k, \end{array}$$

and it is required for global convergence to a point that satisfies the necessary KKT conditions to satisfy conditions (5.1), (5.4), and (5.5). As we saw in equation (9.11) of the proof of Lemma 9.2, property (5.4) is a consequence of (5.5). Whether Property (5.5) holds depends on the way in which the quasi–normal component is computed. We will show below that (5.5) is satisfied for many reasonable ways to compute s_k^{n}.

There are various ways to compute the quasi–normal component s_k^{n} for large scale problems. For example, one can use the conjugate–gradient method as suggested in [50] and [51], or one can use the Lanczos bidiagonalization as described in [20]. Both methods compute an approximate minimizer to the least squares functional in (10.1) from a subspace which contains its negative gradient $-C_y(x_k)^TC_k$. Thus, the components s_k^{n} generated by these methods satisfy $||s_k^{\mathsf{n}}|| \leq \delta_k$ and

$$\frac{1}{2}\|C_y(x_k)(s_k^{\mathsf{n}})_y + C_k\|^2 \ \leq \ \min\left\{\frac{1}{2}\|C_y(x_k)s + C_k\|^2 : \ s \in span\{-C_y(x_k)^TC_k\}\,, \ \|s\| \leq \delta_k\right\}.$$

We can appeal to a classical result due to Powell, see [43, Thm. 4], [37, Lem. 4.8], to show that

$$\|C_k\|^2 - \|C_y(x_k)(s_k^{\mathsf{n}})_y + C_k\|^2 \geq \frac{1}{2}\|C_y(x_k)^TC_k\|\min\left\{\frac{\|C_y(x_k)^TC_k\|}{\|C_y(x_k)^TC_y(x_k)\|}, \delta_k\right\}.$$

Now one can use the fact that $\{C_y(x_k)\}$ and $\{C_y(x_k)^{-T}\}$ are bounded and write

$$\|C_k\|^2 - \|C_y(x_k)(s_k^{\mathsf{n}})_y + C_k\|^2 \geq \kappa_2 \|C_k\| \min\{\kappa_3 \|C_k\|, \delta_k\},$$

where κ_2 and κ_3 are positive and do not depend on k.

An alternative to the previous procedures is to compute the solution of $C_y(x_k)s = -C(x_k)$ and to scale this solution back into the trust region (see (9.9)). In Lemma 9.2, we proved that (9.9) satisfies conditions (5.1), (5.4), and (5.5).

10.2. Computation of the tangential component. In this section we show how to derive conjugate–gradient algorithms to compute $(s_k)_u$. Other practical algorithms to compute trial steps for box–constrained minimization trust–region subproblems are introduced in [4] using three dimensional subspace approximations and conjugate gradients.

Let us consider first the decoupled trust–region approach given in Section 5.2.1. If we ignore the bound constraints for the moment, we can apply the conjugate–gradient algorithm proposed by Steihaug [50] and Toint [51] to solve the problem

$$\begin{array}{ll} \text{minimize} & \Psi_k(s_u) \\ \text{subject to} & \|\bar{D}_k^{-1} s_u\| \leq \delta_k. \end{array}$$

However we also need to incorporate the constraints

$$\sigma_k(a - u_k) \leq s_u \leq \sigma_k(b - u_k).$$

This leads to the following algorithm:

ALGORITHM 10.1 (COMPUTATION OF $s_k = s_k^{\mathsf{n}} + W_k(s_k)_u$ (DECOUPLED APPROACH)).

1 Set $s_u^0 = 0$, $r_0 = -\bar{g}_k = -W_k^T \nabla q_k(s_k^{\mathsf{n}})$, $q_0 = \bar{D}_k^2 r_0$, $d_0 = q_0$, and $\epsilon > 0$.

2 For $i = 0, 1, 2, \ldots$ do

  2.1 Compute $\gamma_i = \frac{r_i^T q_i}{d_i^T (W_k^T H_k W_k + E_k \bar{D}_k^{-2}) d_i}$.

  2.2 Compute
  $$\begin{array}{rl} \tau_i = \max\{\tau > 0 \ : & \|\bar{D}_k^{-1}(s_u^i + \tau d_i)\| \leq \delta_k, \\ & \sigma_k(a - u_k) \leq s_u^i + \tau d_i \leq \sigma_k(b - u_k)\}. \end{array}$$

  2.3 If $\gamma_i \leq 0$, or if $\gamma_i > \tau_i$, then set $(s_k)_u = s_u^i + \tau_i d_i$, where τ_i is given as in 2.2 and go to 3; otherwise set $s_u^{i+1} = s_u^i + \gamma_i d_i$.

  2.4 Update the residuals: $r_{i+1} = r_i - \gamma_i (W_k^T H_k W_k + E_k \bar{D}_k^{-2}) d_i$ and $q_{i+1} = \bar{D}_k^2 r_{i+1}$.

  2.5 Check truncation criteria: if $\sqrt{\frac{r_{i+1}^T q_{i+1}}{r_0^T q_0}} \leq \epsilon$, set $(s_k)_u = s_u^{i+1}$ and go to 3.

  2.6 Compute $\alpha_i = \frac{r_{i+1}^T q_{i+1}}{r_i^T q_i}$ and set $d_{i+1} = q_{i+1} + \alpha_i d_i$.

3 Compute $s_k = s_k^{\mathsf{n}} + W_k(s_k)_u$ and stop.

Step 2 iterates entirely in the vector space of the u variables. After the u component of the step s_k has been computed, Step 3 finds its y component. The decoupled approach allows an efficient use of an approximation $\widehat{H}_k$ to the reduced Hessian $W_k^T \nabla_{xx}^2 \ell_k W_k$. In this case, only two linear systems are required, one with $C_y(x_k)^T$ in Step 1 to compute $\bar{g}_k$ and the other with $C_y(x_k)$ in Step 3 to compute $W_k(s_k)_u$. If it is the Hessian H_k that is being approximated, then the total number of linear systems is $2I(k) + 2$, where $I(k)$ is the number of conjugate–gradient iterations.

One can transform this algorithm to work in the whole space rather then in the reduced space by considering the coupled trust–region approach given in Section 5.2.2. This alternative is presented below.

ALGORITHM 10.2 (COMPUTATION OF $s_k = s_k^{\mathsf{n}} + W_k(s_k)_u$ (COUPLED APPROACH)).

1 Set $s^0 = 0$, $r_0 = -\bar{g}_k = -W_k^T \nabla q_k(s_k^{\mathsf{n}})$, $q_0 = \bar{D}_k^2 r_0$, $d_0 = W_k q_0$, and $\epsilon > 0$.

2 For $i = 0, 1, 2, \ldots$ do

2.1 Compute $\gamma_i = \frac{r_i^T q_i}{d_i^T H_k d_i + (d_i)_u^T E_k \bar{D}_k^{-2} (d_i)_u}$.

2.2 Compute

$$\tau_i = \max\{\tau > 0 \quad : \quad \left\| \begin{pmatrix} -C_y(x_k)^{-1} C_u(x_k) \tau (d_i)_u \\ \bar{D}_k^{-1} \tau (d_i)_u \end{pmatrix} \right\| \le \delta_k,$$

$$\sigma_k(a - u_k) \le s_u^i + \tau (d_i)_u \le \sigma_k(b - u_k)\}.$$

2.3 If $\gamma_i \le 0$, or if $\gamma_i > \tau_i$, then $s_k^{\mathsf{t}} = s^i + \tau_i d_i$, where τ_i is given as in 2.2 and go to 3; otherwise set $s^{i+1} = s^i + \gamma_i d_i$.

2.4 Update the residuals: $r_{i+1} = r_i - \gamma_i \left(W_k^T H_k d_i + E_k \bar{D}_k^{-2} (d_i)_u \right)$ and $q_{i+1} = \bar{D}_k^2 r_{i+1}$.

2.5 Check truncation criteria: if $\sqrt{\frac{r_{i+1}^T q_{i+1}}{r_0^T q_0}} \le \epsilon$, set $s_k^{\mathsf{t}} = s^{i+1}$ and go to 3.

2.6 Compute $\alpha_i = \frac{r_{i+1}^T q_{i+1}}{r_i^T q_i}$ and set $d_{i+1} = W_k(q_{i+1} + \alpha_i d_i)$.

3 Compute $s_k = s_k^{\mathsf{n}} + s_k^{\mathsf{t}}$ and stop.

Note that in Step 2 both the y and the u components of the tangential component are being computed. The coupled approach is suitable particularly when an approximation to the full Hessian H_k is used. The coupled approach can be used also with an approximation $\widehat{H}_k$ to the reduced Hessian $W_k^T \nabla_{xx}^2 \ell_k W_k$. In this case, we consider H_k that is given by (5.24) and use the equalities (5.25) to compute the terms involving H_k in Algorithm 10.2. If the Hessian H_k is approximated, the total number of linear systems is $2I(k) + 2$, where $I(k)$ is the number of conjugate–gradient iterations. If the reduced Hessian $W_k^T \nabla_{xx}^2 \ell_k W_k$ is approximated, this number is $I(k) + 2$.

Two final important remarks are in order.

REMARK 10.1. If $W_k^T W_k$ was included as a preconditioner in Algorithm 10.2, then the conjugate–gradient iterates would monotonically increase in the norm $\|W_k \cdot \|$. Dropping this preconditioner means that the conjugate–gradient iterates do not necessarily increase in this norm (see [50]). As a result, if the quasi–Newton step is inside the trust region, Algorithm 10.2 can terminate prematurely by stopping at the boundary of the trust region.

REMARK 10.2. Since the conjugate–gradient Algorithms 10.1, 10.2 start by minimizing the quadratic function $\Psi_k(s_u)$ along the direction $-\bar{D}_k^2 \bar{g}_k$, it is quite clear that they produce reduced tangential components $(s_k)_u$ that satisfy (5.13) and (5.14), respectively, with $\beta_1^{\mathsf{d}} = \beta_1^{\mathsf{c}} = 1$.

10.3. Multiplier estimates. A convenient estimate for the Lagrange multipliers is the adjoint update

$$\lambda_k = -C_y(x_k)^{-T} \nabla_y f_k, \tag{10.2}$$

which we use after each successful step. However we also consider the following update:

$$\lambda_{k+1} = -C_y(x_k)^{-T} \nabla_y q_k(s_k^{\mathsf{n}}) = -C_y(x_k)^{-T} \left((H_k s_k^{\mathsf{n}})_y + \nabla_y f_k \right). \tag{10.3}$$

Here the use of (10.3) instead of

$$\lambda_{k+1} = -C_y(x_k+s_k)^{-T}\nabla_y f(x_k+s_k), \tag{10.4}$$

might be justified since we obtain (10.3) without any further cost from the first iteration of any of the conjugate–gradient algorithms described above. The updates (10.2), (10.3), and (10.4) satisfy the requirement given by A.4 needed to prove global convergence to a first–order KKT point.

11. Numerical example. A typical application that has the structure described in this paper is the control of a heating process. In this section we introduce a simplified model for the heating of a probe in a kiln discussed in [5]. The temperature $y(x,t)$ inside the probe is governed by a nonlinear partial differential equation. The spatial domain is given by $(0,1)$. The boundary $x=1$ is the inside of the probe and $x=0$ is the boundary of the probe.

The goal is to control the heating process in such a way that the temperature inside the probe follows a certain desired temperature profile $y_d(t)$. The control $u(t)$ acts on the boundary $x=0$. The problem can be formulated as follows.

$$\text{minimize } \frac{1}{2}\int_0^T [(y(1,t)-y_d(t))^2+\gamma u^2(t)]dt \tag{11.1}$$

subject to

$$\begin{array}{rcl}
\tau(y(x,t))\frac{\partial y}{\partial t}(x,t) - \partial_x(\kappa(y(x,t))\partial_x y(x,t)) & = & q(x,t), \quad (x,t)\in(0,1)\times(0,T), \\
\kappa(y(0,t))\partial_x y(0,t) & = & g[y(0,t)-u(t)], \quad t\in(0,T), \\
\kappa(y(1,t))\partial_x y(1,t) & = & 0, \quad t\in(0,T), \\
y(x,0) & = & y_0(x), \quad x\in(0,1), \\
u_{low} \leq & u & \leq u_{upp},
\end{array}$$

where $y \in L^2(0,T;H^1(0,1))$, and $u\in L^2(0,T)$. The functions $\tau:\mathbb{R}\to\mathbb{R}$ and $\kappa:\mathbb{R}\to\mathbb{R}$ denote the specific heat capacity and the heat conduction, respectively, y_0 is the initial temperature distribution, q is the source term, g is a given scalar, and γ is a regularization parameter. Here $u_{low}, u_{upp}\in L^\infty(0,T)$ are given functions.

If the partial differential equation and the integral are discretized, we obtain an optimization problem of the form (1.1). The discretization uses finite elements and was introduced in [5] (see also [23] and [33]). The spatial domain $(0,1)$ is divided into N_x subintervals of equidistant length, and the spatial discretization is done using piecewise linear finite elements. The time discretization is performed by partitioning the interval $[0,T]$ into N_t equidistant subintervals. Then the backward Euler method is used to approximate the state space in time, and piecewise constant functions are used to approximate the control space. This leads to a discretized problem with dimension $n=N_t(N_x+1)+N_t$ and $m=N_t(N_x+1)$. Under the assumptions on the coefficient functions κ and τ stated in [5], [33] which guarantee the well–posedness of the infinite dimensional problem, it is shown in [33] that the constraints $C(y,u)$ of the discretized problem satisfy the assumptions A.3 and A.5 provided the discretization parameters N_x and N_t are chosen appropriately. For more details we refer to the comprehensive treatments in [5] and [33].

The algorithms studied in this paper have been implemented in `FORTRAN 77`. The resulting software package TRICE, trust–region interior–point SQP algorithms for optimal control and engineering design problems is available via the internet [12].

We use the formula (9.9) to compute the quasi–normal component, and Algorithms 10.1 and 10.2 to calculate the tangential component. The numerical test computations were done on a Sun Sparcstation 10 in double precision. These results demonstrate the effectiveness of the algorithms.

With this discretization scheme, $C_y(x)$ is a block bidiagonal matrix with tridiagonal blocks. Hence linear systems with $C_y(x)$ and $C_y(x)^T$ can be solved efficiently by block forward substitution or block backward substitution, respectively. In each substitution step, only a small system with tridiagonal system has to be solved. In the implementation we use the LINPACK subroutine DGTSL to solve the tridiagonal systems. Notice that direct factorizations are only applied to the small $(N_x + 1) \times (N_x + 1)$ tridiagonal subblocks of $C_y(x)$, but not to the entire $N_t N_x \times (N_t(N_x + 1))$ Jacobian matrix $(C_y(x) \; C_u(x))$. See also [33].

As we pointed out in Section 1, the inner products and norms used in the trust–region interior–point SQP algorithms are not necessarily the Euclidean ones. In our implementation [12], we call subroutines to calculate the inner products $\langle y^1, y^2 \rangle$ and $\langle u^1, u^2 \rangle$ with $y^1, y^2 \in \mathbb{R}^m$ and $u^1, u^2 \in \mathbb{R}^{n-m}$. The user may supply these subroutines to incorporate a specific scaling. If the inner product $\langle x^1, x^2 \rangle$ is required, then it is calculated as $\langle y^1, y^2 \rangle + \langle u^1, u^2 \rangle$. In this example, we used discretizations of the $L^2(0,T)$ and $L^2(0,T;H^1(0,1))$ norms for the control and the state spaces respectively. This is important for the correct computation of the adjoint and the appropriate scaling of the problem.

In our numerical example we use the functions

$$\tau(y) = q_1 + q_2 y, \quad y \in \mathbb{R}, \quad \kappa(y) = r_1 + r_2 y, \quad y \in \mathbb{R},$$

with parameters $r_1 = q_1 = 4$, $r_2 = -1$, and $q_2 = 1$. The desired and initial temperatures, and the right hand side are given by

$$\begin{aligned} y_d(t) &= 2 - e^{\eta t}, \\ y_0(x) &= 2 + \cos \pi x, \quad \text{and} \\ q(x,t) &= [\eta(q_1 + 2q_2) + \pi^2(r_1 + 2r_2)] e^{\eta t} \cos \pi x \\ &\quad - r_2 \pi^2 e^{2\eta t} + (2 r_2 \pi^2 + \eta q_2) e^{2\eta t} \cos^2 \pi x, \end{aligned}$$

with $\eta = -1$. The final temperature is chosen to be $T = 0.5$ and the scalar $g = 1$ is used in the boundary condition. The functions in this example are those used in [33, Ex. 4.1]. The size of the problem tested is $n = 2200$, $m = 2100$ corresponding to the values $N_t = 100$, $N_x = 20$.

The scheme used to update the trust radius is the following fairly standard one:

- If $\mathrm{ratio}(s_k; \rho_k) < 10^{-4}$, reject s_k and set $\delta_{k+1} = 0.5 \, \mathrm{norm}(s_k)$;
- If $10^{-4} \le \mathrm{ratio}(s_k; \rho_k) < 0.1$, reject s_k and set $\delta_{k+1} = 0.5 \, \mathrm{norm}(s_k)$;
- If $0.1 \le \mathrm{ratio}(s_k; \rho_k) < 0.75$, accept s_k and set $\delta_{k+1} = \delta_k$;
- If $\mathrm{ratio}(s_k; \rho_k) \ge 0.75$, accept s_k and set $\delta_{k+1} = \min\{2\delta_k, \, 10^{10}\}$;

where $\mathrm{ratio}(s_k; \rho_k) = \frac{ared(s_k;\rho_k)}{pred(s_k;\rho_k)}$,

$$\mathrm{norm}(s_k) = \max\left\{ \|s_k^{\mathrm{n}}\|, \|\bar{D}_k^{-1}(s_k)_u\| \right\}$$

in the decoupled approach, and

$$\mathrm{norm}(s_k) = \max\left\{ \|s_k^{\mathrm{n}}\|, \left\| \begin{pmatrix} -C_y(x_k)^{-1} C_u(x_k)(s_k)_u \\ \bar{D}_k^{-1}(s_k)_u \end{pmatrix} \right\| \right\}$$

TABLE 11.1
Numerical results for $\gamma = 10^{-2}$.

	Decoupled		Coupled	
	Reduced $\widehat{H}_k$	Full H_k	Reduced $\widehat{H}_k$	Full H_k
number of iterations k^*	14	20	17	18
$\Vert C_{k^*}\Vert$	$.5082E-11$	$.1370E-10$	$.7122E-12$	$.8804E-11$
$\Vert D_{k^*}W_{k^*}^T\nabla f_{k^*}\Vert$	$.4033E-08$	$.1389E-08$	$.6365E-10$	$.2641E-08$
$\Vert s_{k^*-1}\Vert$	$.1230E-04$	$.1461E-04$	$.3546E-05$	$.1445E-04$
δ_{k^*-1}	$.1638E+05$	$.1049E+07$	$.1311E+06$	$.2621E+06$
ρ_{k^*-1}	$.1000E+01$	$.1000E+01$	$.1000E+01$	$.1000E+01$

in the coupled approach. The algorithms are stopped if the trust radius gets below 10^{-8}.

We have used $\sigma_k = \sigma = 0.99995$ for all k; $\delta_0 = 1$ as initial trust radius; $\rho_{-1} = 1$ and $\bar{\rho} = 10^{-2}$ in the penalty scheme. The tolerance used in the conjugate–gradient iteration was $\epsilon = 10^{-4}$. The upper and lower bounds were $b_i = 10^{-2}$, $a_i = -1000$, $i = 1, \ldots, n-m$. The starting vector was $x_0 = 0$.

For both the decoupled and the coupled approaches, we did tests using approximations to reduced and to full Hessians. We approximate these matrices with the limited memory BFGS representations given in [6] with a memory size of 5 pairs of vectors. For the reduced Hessian we use a null–space secant update (see [40], [53]). The initial approximation chosen was γI_{n-m} for the reduced Hessian and γI_n for the full Hessian, where γ is the user specified regularization parameter in the objective function (11.1).

In our implementation we use the following form of the diagonal matrix $\bar{D}_k$

$$\left(\bar{D}_k\right)_{ii} = \begin{cases} \min\{1,\, (b-u_k)_i\} & \text{if} \quad (\bar{g}_k)_i < 0, \\ \min\{1,\, (u_k-a)_i\} & \text{if} \quad (\bar{g}_k)_i \geq 0, \end{cases} \tag{11.2}$$

for $i = 1, \ldots, n-m$. This form of $\bar{D}_k$ gives a better transition between the infinite and finite bound and is less sensitive to the introduction of meaningless bounds. See also Remark 3.1.

The algorithms were stopped when

$$\Vert D_k W_k^T \nabla f_k\Vert + \Vert C_k\Vert < 10^{-8}.$$

The results are shown in Tables 11.1 and 11.2 corresponding to the values $\gamma = 10^{-2}$ and $\gamma = 10^{-3}$, respectively. There were no rejected steps. The different alternatives tested performed quite similarly. The decoupled approach with reduced Hessian approximation seems to be the best for this example. Note that in this case the computation of each trial step costs only three linear system solvers with $C_y(x_k)$ and $C_y(x_k)^T$, one to compute the quasi–normal component and two for the computation of the tangential component.

We made an experiment to compare the use of the Coleman–Li affine scaling with the Dikin–Karmarkar affine scaling. When applied to our class of problems, the Coleman–Li affine scaling

TABLE 11.2
Numerical results for $\gamma = 10^{-3}$.

	Decoupled		Coupled	
	Reduced $\widehat{H}_k$	Full H_k	Reduced $\widehat{H}_k$	Full H_k
number of iterations k^*	16	18	17	19
$\|C_{k^*}\|$	$.6233E-11$	$.1115E-10$	$.6487E-11$	$.1246E-09$
$\|D_{k^*}W_{k^*}^T\nabla f_{k^*}\|$	$.5161E-08$	$.2539E-08$	$.7282E-09$	$.4696E-08$
$\|s_{k^*-1}\|$	$.1626E-04$	$.1703E-04$	$.1530E-04$	$.4659E-04$
δ_{k^*-1}	$.6554E+05$	$.2621E+06$	$.1311E+06$	$.5243E+06$
ρ_{k^*-1}	$.1000E+01$	$.1000E+01$	$.1000E+01$	$.1000E+01$

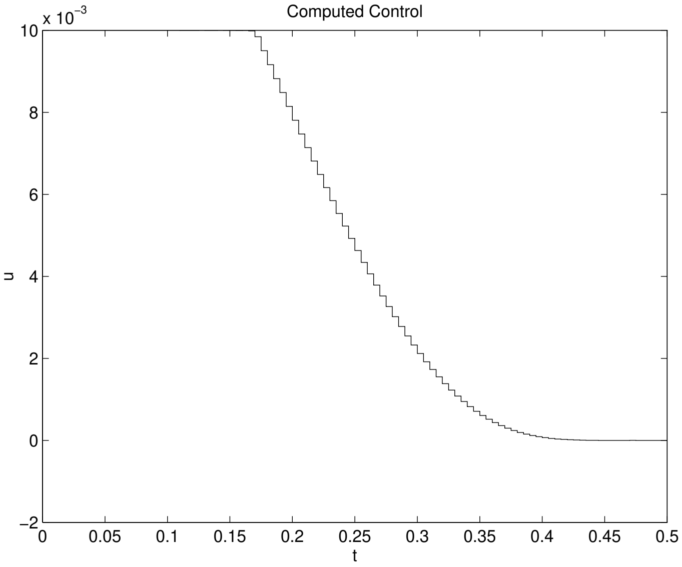

FIG. 11.1. *Coleman–Li affine scaling.*

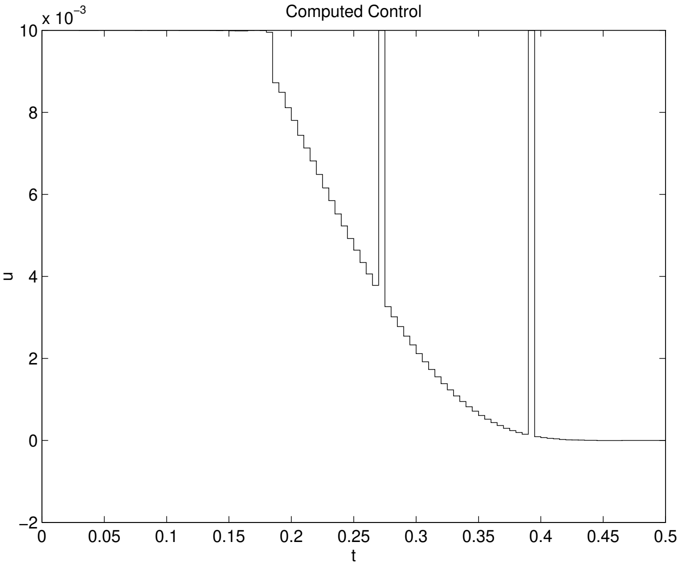

FIG. 11.2. *Dikin–Karmarkar affine scaling.*

is given by the matrices D_k and $\bar{D}_k$. A study of the Dikin–Karmarkar affine scaling for steepest descent is given in [45]. For our class of problems, this scaling is given by

$$\left(K_k\right)_{ii} = \min\{1,\ (u_k - a)_i,\ (b - u_k)_i\},\ i = 1, \ldots, n-m, \tag{11.3}$$

and has no dual information built in. We ran the trust–region interior–point SQP algorithm with the decoupled and reduced Hessian approximation and (11.2) replaced by (11.3). The algorithm took only 11 iterations to reduce $\|K_k W_k^T \nabla f_k\| + \|C_k\|$ to 10^{-8}. However, as we can see from the plots of the controls in Figures 11.1 and 11.2, the algorithm did not find the correct solution when it used the Dikin–Karmarkar affine scaling (11.3). Some of the variables are at the wrong bound corresponding to negative multipliers.

12. Conclusions. In this paper we have introduced and analyzed some trust–region interior–point SQP algorithms for an important class of nonlinear programming problems that appear in many engineering applications. These algorithms use the structure of the problem, and they combine trust–region techniques for equality–constrained optimization with an affine scaling interior–point approach for simple bounds. We have proved global and local convergence results for these algorithms that includes as special cases both the results established for equality constraints [11], [14] and those for simple bounds [9].

We have implemented the trust–region interior–point SQP algorithms covering several trial step computations and second–order approximations. In this paper we have reported numerical results for the solution of a specific optimal control problem governed by a nonlinear heat equation. In [7], [24], [25], these algorithms have been applied to other optimal control problems. The numerical results have been quite satisfactory.

We are investigating extensions of these algorithms to handle bounds on the state variables y. We also are developing an inexact analysis to deal with trial step computations that allow for inexact linear system solvers and inexact directional derivatives [25]. The formulation and analysis of these methods in an infinite dimensional framework is also part of our current studies.

Acknowledgements. This research collaboration on the tailoring of current optimization algorithm design to exploit the structure of problems in computational design and control was begun as the result of conversations held during the April 1994 Workshop on Optimal Design and Control held in Blacksburg at Virginia Polytechnic Institute and State University. This workshop was organized by the Interdisciplinary Center for Applied Mathematics at Virginia Tech and sponsored by the AFOSR.

We would like to thank two referees for their patient and careful reading and their helpful comments on an earlier draft of this paper.

REFERENCES

[1] L. T. Biegler, J. Nocedal, and C. Schmid, *A reduced Hessian method for large–scale constrained optimization*, SIAM J. Optim., 5 (1995), pp. 314–347.

[2] P. T. Boggs, *Sequential quadratic programming*, in Acta Numerica 1995, A. Iserles, ed., Cambridge University Press, Cambridge, London, New York, 1995, pp. 1–51.

[3] J. F. Bonnans and C. Pola, *A trust region interior point algorithm for linearly constrained optimization*, Tech. Rep. 1948, INRIA, 1993.

[4] M. A. Branch, T. F. Coleman, and Y. Li, *A subspace, interior, and conjugate gradient method for large–scale bound–constrained minimization problems*, Tech. Rep. CTC95TR217, Advancing Computing Research Institute, Cornell University, 1995.

[5] J. Burger and M. Pogu, *Functional and numerical solution of a control problem originating from heat transfer*, J. Optim. Theory Appl., 68 (1991), pp. 49–73.

[6] R. H. Byrd, J. Nocedal, and R. B. Schnabel, *Representations of quasi–Newton matrices and their use in limited memory methods*, Math. Programming, 63 (1994), pp. 129–156.

[7] E. M. Cliff, M. Heinkenschloss, and A. Shenoy, *An optimal control problem for flows with discontinuities*, Journal of Optimization Theory and Applications, 94 (1997), pp. (to appear, August).

[8] T. F. Coleman and Y. Li, *On the convergence of interior–reflective Newton methods for nonlinear minimization subject to bounds*, Math. Programming, 67 (1994), pp. 189–224.

[9] ———, *An interior trust region approach for nonlinear minimization subject to bounds*, SIAM J. Optim., 6 (1996), pp. 418–445.

[10] T. F. Coleman and J. Liu, *An interior Newton method for quadratic programming*, Tech. Rep. TR93–1388, Department of Computer Science, Cornell University, 1993.

[11] J. E. Dennis, M. El-Alem, and M. C. Maciel, *A global convergence theory for general trust–region–based algorithms for equality constrained optimization*, SIAM J. Optim., (To appear).

[12] J. E. Dennis, M. Heinkenschloss, and L. N. Vicente, *TRICE: Trust–region interior–point SQP algorithms for optimal control and engineering design problems.* `http://www.caam.rice.edu/trice`.

[13] J. E. Dennis and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice–Hall, Englewood Cliffs, New Jersey, 1983.

[14] J. E. Dennis and L. N. Vicente, *On the convergence theory of general trust–region–based algorithms for equality–constrained optimization*, SIAM J. Optim., (To appear).

[15] M. El-Alem, *A global convergence theory for the Celis–Dennis–Tapia trust–region algorithm for constrained optimization*, SIAM J. Numer. Anal., 28 (1991), pp. 266–290.

[16] ——, *A robust trust–region algorithm with a non–monotonic penalty parameter scheme for constrained optimization*, SIAM J. Optim., 5 (1995), pp. 348–378.

[17] D. M. Gay, *Computing optimal locally constrained steps*, SIAM J. Sci. Statist. Comput., 2 (1981), pp. 186–197.

[18] P. E. Gill, W. Murray, M. A. Saunders, and M. H. Wright, *User's guide for NPSOL (version 4.0): A FORTRAN package for nonlinear programming*, Technical Report SOL 86–2, Systems Optimization Laboratory, Department of Operations Research, Stanford University, Stanford, CA 94305–4022, 1986.

[19] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The John Hopkins University Press, Baltimore and London, second ed., 1989.

[20] G. H. Golub and U. von Matt, *Quadratically constrained least squares and quadratic problems*, Numer. Math., 59 (1991), pp. 561–580.

[21] J. Goodman, *Newton's method for constrained optimization*, Math. Programming, 33 (1985), pp. 162–171.

[22] W. A. Gruver and E. W. Sachs, *Algorithmic Methods In Optimal Control*, Pitman, London, 1980.

[23] M. Heinkenschloss, *Projected sequential quadratic programming methods*, SIAM J. Optim., 6 (1996), pp. 373–417.

[24] ——, *SQP methods for the solution of optimal control problems governed by the Navier Stokes equations*, Tech. Rep. in preparation, Interdisciplinary Center for Applied Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, 1996.

[25] M. Heinkenschloss and L. N. Vicente, *Analysis of inexact trust–region interior–point SQP algorithms*, Tech. Rep. TR95–18, Department of Computational and Applied Mathematics, Rice University, 1995. Revised April 1996. Appeared also as Tech. Rep. 95–06–01, Interdisciplinary Center for Applied Mathematics, Virginia Polytechnic Institute and State University.

[26] K. Ito and K. Kunisch, *The augmented Lagrangian method for parameter estimation in elliptic systems*, SIAM J. Control Optim., 28 (1990), pp. 113–136.

[27] C. T. Kelley and E. W. Sachs, *Solution of optimal control problems by a pointwise projected Newton method*, SIAM J. Control Optim., 33 (1995), pp. 1731–1757.

[28] C. T. Kelley and S. J. Wright, *Sequential quadratic programming for certain parameter identification problems*, Math. Programming, 51 (1991), pp. 281–305.

[29] K. Kunisch and G. Peichl, *Estimation of a temporally and spatially varying diffusion coefficient in a parabolic system by an augmented Lagrangian technique*, Numer. Math., 59 (1991), pp. 473–509.

[30] K. Kunisch and E. Sachs, *Reduced SQP methods for parameter identification problems*, SIAM J. Numer. Anal., 29 (1992), pp. 1793–1820.

[31] F.-S. Kupfer, *An infinite–dimensional convergence theory for reduced SQP methods in Hilbert space*, SIAM J. Optim., 6 (1996), pp. 126–163.

[32] F.-S. Kupfer and E. W. Sachs, *A prospective look at SQP methods for semilinear parabolic control problems*, in Optimal Control of Partial Differential Equations, Irsee 1990, K.-H. Hoffmann and W. Krabs, eds., vol. 149, Springer Lect. Notes in Control and Information Sciences, 1991, pp. 143–157.

[33] ——, *Numerical solution of a nonlinear parabolic control problem by a reduced SQP method*, Comput. Optim. and Appl., 1 (1992), pp. 113–135.

[34] M. Lalee, J. Nocedal, and T. Plantenga, *On the implementation of an algorithm for large–scale equality constrained optimization.* Submitted for publication, 1994.

[35] Y. Li, *On global convergence of a trust region and affine scaling method for nonlinearly constrained minimization*, Tech. Rep. CTC94TR197, Advanced Computing Research Institute, Cornell University, 1994.

[36] ——, *A trust region and affine scaling method for nonlinearly constrained minimization*, Tech. Rep. CTC94TR198, Advanced Computing Research Institute, Cornell University, 1994.

[37] J. J. Moré, *Recent developments in algorithms and software for trust regions methods*, in Mathematical pro-

gramming. The state of art, A. Bachem, M. Grotschel, and B. Korte, eds., Springer Verlag, New York, 1983, pp. 258–287.

[38] J. J. Moré and D. C. Sorensen, *Computing a trust region step*, SIAM J. Sci. Statist. Comput., 4 (1983), pp. 553–572.

[39] W. Murray and F. J. Prieto, *A sequential quadratic programming algorithm using an incomplete solution of the subproblem*, SIAM J. Optim., 5 (1995), pp. 590–640.

[40] J. Nocedal and M. L. Overton, *Projected Hessian updating algorithms for nonlinearly constrained optimization*, SIAM J. Numer. Anal., 22 (1985), pp. 821–850.

[41] T. Plantenga, *Large–Scale Nonlinear Constrained Optimization using Trust Regions*, PhD thesis, Northwestern University, Evanston, Illinois, 1994.

[42] E. Polak, *Computational Methods in Optimization. A Unified Approach*, Academic Press, New York, London, Paris, San Diego, San Francisco, 1971.

[43] M. J. D. Powell, *A new algorithm for unconstrained optimization*, in Nonlinear Programming, J. B. Rosen, O. L. Mangasarian, and K. Ritter, eds., Academic Press, New York, 1970.

[44] F. Rendl and H. Wolkowicz, *A semidefinite framework for trust region subproblems with applications to large scale minimization*, Tech. Rep. 94–32, CORR, 1994.

[45] C. M. Samuelson, *The Dikin–Karmarkar Principle for Steepest Descent*, PhD thesis, Department of Computational and Applied Mathematics, Rice University, Houston, Texas 77251, USA, 1992. Tech. Rep. TR92–29.

[46] S. A. Santos and D. C. Sorensen, *A new matrix–free algorithm for the large–scale trust–region subproblem*, Tech. Rep. TR95–20, Department of Computational and Applied Mathematics, Rice University, 1994.

[47] K. Schittkowski, *NLPQL: A* FORTRAN *subroutine solving constrained nonlinear programming problems*, Annals of Operations Research, 5 (1985), pp. 485–500.

[48] D. C. Sorensen, *Newton's method with a model trust region modification*, SIAM J. Numer. Anal., 19 (1982), pp. 409–426.

[49] ———, *Minimization of a large scale quadratic function subject to an spherical constraint*, SIAM J. Optim., (To appear).

[50] T. Steihaug, *The conjugate gradient method and trust regions in large scale optimization*, SIAM J. Numer. Anal., 20 (1983), pp. 626–637.

[51] P. L. Toint, *Towards an efficient sparsity exploiting Newton method for minimization*, in Sparse Matrices and Their Uses, I. S. Duff, ed., Academic Press, New York, 1981, pp. 57–87.

[52] L. N. Vicente, *Trust–Region Interior–Point Algorithms for a Class of Nonlinear Programming Problems*, PhD thesis, Department of Computational and Applied Mathematics, Rice University, Houston, Texas 77251, USA, 1996.

[53] Y. Xie, *Reduced Hessian Algorithms for Solving Large–Scale Equality Constrained Optimization Problems*, PhD thesis, Dept. of Computer Science, University of Colorado, 1991.