Computational Optimization and Applications, 15, 103–123 (2000)

A Quasi-Newton Quadratic Penalty Method for Minimization Subject to Nonlinear Equality Constraints*

THOMAS F. COLEMAN
Computer Science Department and Cornell Theory Center, Cornell University, Ithaca, NY 14850, USA

JIANGUO LIU
Department of Mathematics, University of North Texas, Denton, TX 76203, USA

WEI YUAN
Center for Applied Mathematics, Cornell University, Ithaca, NY 14853, USA

Received January 28, 1997; Accepted November 24, 1998

Abstract. We present a modified quadratic penalty function method for equality constrained optimization problems. The pivotal feature of our algorithm is that at every iterate we invoke a special change of variables to improve the ability of the algorithm to follow the constraint level sets. This change of variables gives rise to a suitable block diagonal approximation to the Hessian which is then used to construct a quasi-Newton method. We show that the complete algorithm is globally convergent. Preliminary computational results are reported.

Keywords: nonlinearly constrained optimization, equality constraints, quasi-Newton methods, BFGS, quadratic penalty function, reduced Hessian approximation

1. Introduction

One of the great success stories in continuous optimization is the development of effective quasi-Newton methods for unconstrained minimization (at least for problems of moderate size). Three important reasons for this success are:

- Line search rules that ensure global convergence are consistent with positive definite quasi-Newton updates: the approximating matrix can be updated at every iteration (even at points far from the solution) and positive definiteness can be maintained.
- Ultimately the line search rules allow for unit step sizes which facilitates rapid local convergence.
- The true Hessian matrix is positive definite in a neighborhood of a strong local minimizer (thus "justifying" the preservation of positive definiteness of the approximating matrices).

*This research was partially supported by the Applied Mathematical Sciences Research Program (KC-04-02) of the Office of Energy Research the U.S. Department of Energy under grant DE-FG02-86ER25013.A000, and by the Computational Mathematics Program of the National Science Foundation under grant DMS-8706133.

Unfortunately, adaptation of the unconstrained quasi-Newton technology to the nonlinearly constrained problem,

$$\text{minimize } \{f(x) : c(x) = 0\}, \tag{1.1}$$

where $f : \Re^n \to \Re^1$ and $c(x) = [c_1(x)\ c_2(x) \cdots c_m(x)]^T : \Re^n \to \Re^m$ $(m < n)$, has not been an easy task. The problem is not in the asymptotics where there are now many effective choices, especially with respect to reduced Hessian approximations, e.g., [2, 4, 6, 7, 11, 12, 18]. The main problem lies in *smoothly connecting* global techniques, and line searches, with an effective asymptotic procedure. In particular, second-order optimality conditions strongly suggest that approximating the reduced Hessian of the Lagrangian function is the right thing to do. Indeed, the effective asymptotic methods do just this [7, 12, 18]. However, local minimizers of (1.1) are not necessarily local minimizers of the Lagrangian function; therefore, it is not possible to design consistent line search rules based on the Lagrangian function. Many global strategies for problem (1.1) are based on penalty or merit functions; however, direct approximation of the Hessian matrix of a penalty/merit function is usually a bad idea due to inherent ill-conditioning. Moreover, line search rules based on a penalty function to ensure global convergence are not usually consistent with positive definite quasi-Newton approximations to the Hessian, or even the reduced Hessian, of the Lagrangian function.

In this paper we propose a solution based on the quadratic penalty function:

$$p_\mu(x) = f(x) + \frac{1}{2\mu}\|c(x)\|_2^2.$$

Our solution rests on the observation that the reduced Hessian of the quadratic penalty function on a specified *curved surface* is closely approximated by the reduced Lagrangian Hessian (on a linear subspace). Therefore, it is possible to tie in *curvilinear* line search rules based on the quadratic penalty function with positive definite updates of a reduced Hessian approximation to the Lagrangian function.

In analogy to the unconstrained quasi-Newton methods, our proposed scheme has three important properties:

- (Curvilinear) line search rules that ensure global convergence are consistent with positive definite quasi-Newton updates: the approximating matrix can be updated at every iteration (even at points remote from the solution) and positive definiteness can be maintained. Positive definiteness yields descent directions.
- Ultimately the step size rules allow for unit step sizes: fast asymptotic convergence follows.
- The true (reduced) Hessian matrix of the Lagrangian function is positive definite in a neighborhood of a strong local minimizer (thus "justifying" the preservation of positive definiteness of the approximating matrices).

In addition to allowing a smooth tie-in with an effective asymptotic updating strategy, the curvilinear search idea has an additional benefit with regard to the quadratic penalty function. One of the usual drawbacks of this penalty function approach is that progress can

be very slow, even at points far from the solution. Slow progress can occur when the penalty parameter μ is small, the constraints are near to being satisfied, and yet the solution is still remote. In these circumstances straightline algorithms will exhibit zig-zagging behavior, the iterates crawl along (or near) the manifold defined by $c(x) = 0$, slowly headed toward the minimizer. Of course strategies for adjusting μ try to avoid this situation but it is often difficult to do so. A curvilinear search solves this zigzagging problem and allows for large steps.

Another traditional criticism of the quadratic penalty function concerns the asymptotic ill-conditioning of the Hessian matrix. However, several studies e.g., [8, 13] have shown how to circumvent possible negative effects of this ill-conditioning by either using "extended" systems, or orthogonal transformations to isolate the ill-conditioning. Our proposed approach is in line with the latter view. In particular, the cause of the asymptotic ill-conditioning is the decrease of the penalty parameter toward zero; however, the penalty parameter has no role in the reduced Hessian of the Lagrangian function (i.e., the Hessian of the quadratic penalty function along a specified curved surface) and this is the matrix that is approximated.

1.1. *Related and supporting work*

There is considerable literature on the quadratic penalty function. The fundamental reference is Fiacco and McCormick's [10] influential book. Numerically sound approaches for dealing with ill-conditioning of the Hessian are given in [8, 13]. An algorithm somewhat similar to ours for constrained optimization using penalty function and constrained step is proposed in [15]. Methods based on the union of quasi-Newton method and quadratic penalty function are also suggested in [1]. We can broadly classify this body of work into two categories: local projected quasi-Newton updating strategies for the nonlinearly constrained problem (1.1) and the use of the quadratic penalty function to force convergence from remote points. For example,

In the nonlinearly constrained minimization problem (1.1) we assume f and every c_i $(1 \leq i \leq m)$ are twice continuously differentiable in the domain of interest. For any given $x \in \Re^n$, we denote $A = A(x) \equiv \nabla c(x) = [\nabla c_1(x)\ \nabla c_2(x) \cdots \nabla c_m(x)]$.

The idea behind a local projected quasi-Newton method for (1.1) is to approximate the reduction of the Hessian matrix of the Lagrangian on the null space of matrix $(A^{(k)})^T = (A(x^{(k)})^T$, where $x^{(k)}$ is the current iterate. Specifically, if the columns of matrix $Z^{(k)}$ form an orthonormal basis for the null space of $(A^{(k)})^T$, and $L(x^{(k)}, \lambda^{(k)}) = f(x^{(k)}) + (\lambda^{(k)})^T c(x^{(k)})$ is the Lagrangian function, with Lagrange multipliers $\lambda^{(k)}$, then the reduced Hessian of L with respect to $x^{(k)}$ can be written $H = (Z^{(k)})^T \nabla^2_{xx} L(x^{(k)}, \lambda^{(k)}) Z^{(k)}$. In a neighborhood of a strong local minimizer $H(x)$ is positive definite. Local projected quasi-Newton methods approximate $H(x^{(k)})$ with a positive definite matrix $B^{(k)}$. In most projected Hessian methods $B^{(k)}$ is updated using Broyden's class of formulas, e.g., BFGS. For example, Coleman and Conn [7] propose the following local quasi-Newton method:

$$\text{solve } B^{(k)} h^{(k)} = -\left(Z^{(k)}\right)^T \nabla f^{(k)} \tag{1.2}$$

$$v^{(k)} \leftarrow A^{(k)} \left(\left(A^{(k)}\right)^T A^{(k)}\right)^{-1} c\left(x^{(k)} + Z^{(k)} h^{(k)}\right) \tag{1.3}$$

$$x^{(k+1)} \leftarrow x^{(k)} + Z^{(k)} h^{(k)} + v^{(k)}. \tag{1.4}$$

Matrix $B^{(k)}$ is updated using the BFGS formula based on the pair $\{h^{(k)}, y^{(k)}\}$ where $y^{(k)} = (Z^{(k)})^T[\nabla(L(x^{(k)} + Z^{(k)}h^{(k)}) - \nabla L(x^{(k)})]$. The new Lagrange multipliers $\lambda^{(k+1)}$ can be calculated in any of a number of ways, e.g., a least-squares calculation.

An important point is that in a neighborhood of a strong minimizer to (1.1) the inner product $(y^{(k)})^T h^{(k)}$ is positive and therefore the reduced BFGS update is well-defined; positive definiteness is preserved. Coleman and Conn [7] establish a 2-step superlinear convergence result for this algorithm; subsequently, this result was strengthened [5, 6] to 1-step superlinear convergence of the intermediate sequence $\{x^{(k)} + Z^{(k)}h^{(k)}\}$. A number of variations of this basic scheme have now been proposed e.g., [18], using different definitions of $y^{(k)}$ and slightly different corrections for $x^{(k)}$. However, a common feature is the recurrence of a reduced matrix $B^{(k)}$ and the preservation of positive definiteness in a neighborhood of a strong minimizer due to $(y^{(k)})^T h^{(k)} > 0$. The relevance to our current paper is two-fold. First, asymptotically we would like the quasi-Newton method based on the quadratic penalty function to closely resemble the local procedure described above. Second, the crux of our challenge is to ensure a positivity condition $(y^{(k)})^T h^{(k)} > 0$ holds globally. We approach the quasi-Newton globalization problem in the context of the quadratic penalty function.

1.2. Organization

Our paper is organized as follows. In Section 2 we derive and discuss the proposed algorithm. Global convergence properties are established in Section 3; preliminary numerical results, to help establish viability, are provided in Section 4.

2. Algorithm

The most novel aspect of our approach is combining the curvilinear search idea with quasi-Newton updating. The reason these fit together is that the reduction of $\nabla^2 L$ onto the null space of the constraint gradients approximates, to high order, the Hessian of p_μ reduced to the local *curved space* defined by an approximation to the nonlinear constraints. Let us be more precise.

Given a point $x^{(k)} \in \Re^n$, define the QR-factorization of $A^{(k)}$,

$$A^{(k)} = Q^{(k)}\bar{R}^{(k)} = \begin{bmatrix} Y^{(k)} & Z^{(k)} \end{bmatrix} \begin{bmatrix} R^{(k)} \\ 0 \end{bmatrix} = Y^{(k)}R^{(k)}, \tag{2.5}$$

where $Q^{(k)} \in \Re^{n\times n}$ is orthogonal and $R^{(k)}$ is an $m \times m$ upper triangular matrix. Assume that $A^{(k)}$ has full column rank; hence, $R^{(k)}$ is nonsingular. For any vector $h \in \Re^{n-m}$ define a curved path $u(h) \equiv x^{(k)} + s^{(k)}(h)$ which approximately follows the level set $c(x) = c(x^{(k)})$, where

$$s^{(k)}(h) \equiv Z^{(k)}h + Y^{(k)}\big(R^{(k)}\big)^{-T}\big[c\big(x^{(k)}\big) - c\big(x^{(k)} + Z^{(k)}h\big)\big]. \tag{2.6}$$

Note that under our assumptions (2.6) defines a one-to-one mapping in a neighborhood of $x^{(k)}$. The following result shows some important properties of the constraint function $c(x)$ and the penalty function $p_\mu(x)$ along the path $u(h)$.

Lemma 2.1. *Let* $p_\mu(x) = f(x) + \frac{1}{2\mu}\|c(x)\|_2^2$ *be the quadratic penalty function and let* $u(h) = x^{(k)} + s^{(k)}(h)$ *where* $s^{(k)}(h)$ *is defined in* (2.6). *Then*

$$c(u(h)) = c\big(x^{(k)}\big) + \mathcal{O}(\|h\|^3) \quad \text{as } h \to 0, \tag{2.7}$$

and

$$\begin{cases} \nabla_h p_\mu(u(0)) = \big(Z^{(k)}\big)^T \nabla f\big(x^{(k)}\big) \\ \nabla_h^2 p_\mu(u(0)) = \big(Z^{(k)}\big)^T \nabla^2 L\big(x^{(k)}\big) Z^{(k)} \equiv H^{(k)}, \end{cases} \tag{2.8}$$

where $\nabla^2 L(x^{(k)}) = \nabla^2 f(x^{(k)}) + \sum_{i=1}^m \lambda_i^{(k)} \nabla^2 c_i(x^{(k)})$ *and* $\lambda_i^{(k)}$ *is the ith component of the (least-squares) Lagrangian multiplier* $\lambda^{(k)} = -(R^{(k)})^{-1}(Y^{(k)})^T \nabla f(x^{(k)})$.

Proof: For simplicity we will omit the superscript k in the proof. It follows from (2.5) and (2.6) that $\nabla c(x)^T Z = 0$, $s(0) = 0$, and $u(0) = x$. Therefore, denoting

$$r = \frac{1}{2}[(Zh)^T \nabla^2 c_1(x)(Zh), (Zh)^T \nabla^2 c_2(x)(Zh), \ldots, (Zh)^T \nabla^2 c_m(x)(Zh)]^T,$$

we have $s(h) = Zh - Y R^{-T} r + \mathcal{O}(\|h\|^3)$ and

$$\nabla c(x)^T s(h) = R^T Y^T s(h) = -r + \mathcal{O}(\|h\|^3).$$

Hence

$$c(u(h)) = c(x + s(h)) = c(x) + \nabla c(x)^T s(h) + r + \mathcal{O}(\|h\|^3) = c(x) + \mathcal{O}(\|h\|^3),$$

Hence

$$c(u(h)) = c(x + s(h)) = c(x) + \nabla c(x)^T s(h) + r + \mathcal{O}(\|h\|^3) = c(x) + \mathcal{O}(\|h\|^3).$$

That proves (2.7).

By definition, $p_\mu(u(h)) = f(u(h)) + \frac{1}{2\mu}\sum_{i=1}^m c_i(u(h))^2$. Hence

$$\nabla_h p_\mu(u(h)) = \nabla u(h) \nabla f(u(h)) + \frac{1}{2\mu} \sum_{i=1}^m 2c_i(u(h)) \nabla u(h) \nabla c_i(u(h)).$$

It is easy to see that $\nabla_u(0) = Z^T$ and $Z^T \nabla c_i(x) = 0\ (1 \le i \le m)$. Therefore

$$\nabla_h p_\mu(u(0)) = Z^T \nabla f(x).$$

Now

$$\nabla_h^2 p_\mu(u(h)) = \nabla u(h) \nabla^2 f(u(h)) \nabla u(h)^T + \sum_{j=1}^n \nabla^2 u_j(h) \partial_j f(u(h)) + \frac{1}{\mu} \nabla t(h),$$

where $\partial_j f$ is the partial derivative with respect to the jth variable, and $t(h) = \sum_{i=1}^m c_i(u(h)) \nabla u(h) \nabla c_i(u(h))$. It follows from direct calculations that

$$\sum_{j=1}^n \nabla^2 u_j(0) \partial_j f(u(0)) = Z^T \left[\sum_{i=1}^m (-R^{-1} Y^T \nabla f(x))_i \nabla^2 c_i(x) \right] Z$$

and

$$\nabla t(0) = \sum_{i=1}^m c_i(x) Z^T \left[-\sum_{k=1}^m \beta_k \nabla^2 c_k(x) + \nabla^2 c_i(x) \right] Z,$$

where $\beta_k = (R^{-1} Y^T \nabla c_i(x))_k = 1$ if $k = i$ and 0 otherwise. Therefore $\nabla t(0) = 0$ and

$$\nabla_h^2 p_\mu(u(0)) = Z^T \nabla^2 f(x) Z + Z^T \sum_{i=1}^m \lambda_i \nabla^2 c_i(x) Z.$$

That completes the proof. □

Equation (2.7) says that along any curve $u(h)$ the value of $\|c(u(h))\|$ changes only very slightly. Equation (2.8) tells the story: at $x^{(k)}$ the Hessian of the quadratic penalty function p_μ, reduced to the curved surface defined by $u(h)$, is equal to $H^{(k)}$, the Hessian of the Lagrangian function reduced to the linear manifold defined by the columns of $Z^{(k)}$. This is important because $H^{(k)}$ does not involve the penalty parameter μ; moreover, following the discussion in Section 1, use of $H^{(k)}$ is consistent with superlinear convergence.

If we define any potential new iterate to be of the form $w(h, v) = u(h) + Y^{(k)} v$ for some $v \in \Re^m$, then, locally defining $\hat{p}_\mu(h, v) = p_\mu(w(h, v))$, it follows that

$$\nabla_{(h,v)} \hat{p}_\mu(0, 0) = \begin{bmatrix} \nabla_h \hat{p}(0, 0) \\ \nabla_v \hat{p}(0, 0) \end{bmatrix} = \begin{bmatrix} (Z^{(k)})^T \nabla p_\mu(x^{(k)}) \\ (Y^{(k)})^T \nabla p_\mu(x^{(k)}) \end{bmatrix} = \begin{bmatrix} (Z^{(k)})^T \nabla f(x^{(k)}) \\ (Y^{(k)})^T \nabla p_\mu(x^{(k)}) \end{bmatrix},$$

and

$$\nabla^2_{(h,v)} \hat{p}_\mu(0, 0) = \begin{bmatrix} (Z^{(k)})^T \nabla^2 L(x^{(k)}) Z^{(k)} & (Z^{(k)})^T \nabla^2 \hat{L}(x^{(k)}) Y^{(k)} \\ (Y^{(k)})^T \nabla^2 \hat{L}(x^{(k)}) Z^{(k)} & (Y^{(k)})^T \nabla^2 \hat{L}(x^{(k)}) Y^{(k)} + \frac{1}{\mu} R^{(k)} (R^{(k)})^T \end{bmatrix},$$

where $\nabla^2 \hat{L}(x^{(k)}) = \nabla^2 f(x^{(k)}) + \sum_{i=1}^m \frac{c_i(x^{(k)})}{\mu} \nabla^2 c_i(x^{(k)})$.

It is instructive to consider the Newton system for $\nabla_{(h,v)} \hat{p}_\mu(h, v)$:

$$\begin{bmatrix} (Z^{(k)})^T \nabla^2 L(x^{(k)}) Z^{(k)} & (Z^{(k)})^T \nabla^2 \hat{L}(x^{(k)}) Y^{(k)} \\ (Y^{(k)})^T \nabla^2 \hat{L}(x^{(k)}) Z^{(k)} & (Y^{(k)})^T \nabla^2 \hat{L}(x^{(k)}) Y^{(k)} + \frac{1}{\mu} R^{(k)} (R^{(k)})^T \end{bmatrix} \begin{bmatrix} h \\ v \end{bmatrix} = - \begin{bmatrix} (Z^{(k)})^T \nabla p_\mu(x^{(k)}) \\ (Y^{(k)})^T \nabla p_\mu(x^{(k)}) \end{bmatrix}. \tag{2.9}$$

The following argument suggests an approximation scheme for system (2.9). First, it follows from Taylor's theorem that

$$\begin{aligned}\left(Z^{(k)}\right)^T \nabla p_\mu\left(x^{(k)}\right) + \left(\left(Z^{(k)}\right)^T \nabla^2 \hat{L}\left(x^{(k)}\right) Y^{(k)}\right) v &\approx \left(Z^{(k)}\right)^T \nabla p_\mu\left(x^{(k)} + Y^{(k)} v\right) \\ &\approx \left(Z^{(k)}\right)^T \nabla f\left(x^{(k)} + Y^{(k)} v\right). \quad (2.10)\end{aligned}$$

Thus the top equation in (2.9) can be approximately written as

$$\left(\left(Z^{(k)}\right)^T \nabla^2 L\left(x^{(k)}\right) Z^{(k)}\right) h = -\left(Z^{(k)}\right)^T \nabla f\left(x^{(k)} + Y^{(k)} v\right).$$

Because $A^{(k)} = Y^{(k)} R^{(k)}$ and $\lambda^{(k)} = -(R^{(k)})^{-1}(Y^{(k)})^T \nabla f(x^{(k)})$, we obtain

$$\begin{aligned}\left(Y^{(k)}\right)^T \nabla p_\mu\left(x^{(k)}\right) &= R^{(k)}\left(R^{(k)}\right)^{-1}\left(Y^{(k)}\right)^T \left[\nabla f\left(x^{(k)}\right) + A^{(k)} \frac{c\left(x\left(x^{(k)}\right)\right)}{\mu}\right] \\ &= -\frac{1}{\mu} R^{(k)}\left[\mu\lambda^{(k)} - c\left(x^{(k)}\right)\right]. \quad (2.11)\end{aligned}$$

It is clear that $\frac{1}{\mu} R^{(k)}(R^{(k)})^T$ plays a dominant role in the lower of (2.9) as μ tends to zero. Thus, the second equation in system (2.9) is approximated by

$$\begin{aligned}\frac{1}{\mu} R^{(k)}\left(R^{(k)}\right)^T v &= -\left(Y^{(k)}\right)^T \nabla p_\mu\left(x^{(k)}\right) \\ &= \frac{1}{\mu} R^{(k)}\left[\mu\lambda^{(k)} - c\left(x^{(k)}\right)\right] \\ &\approx -\frac{1}{\mu} R^{(k)} c\left(x^{(k)}\right). \quad (2.12)\end{aligned}$$

Therefore, the system (2.9) can be approximated:

$$\begin{cases} a) & \left(R^{(k)}\right)^T v = -c\left(x^{(k)}\right) \\ b) & B^{(k)} h = -\left(Z^{(k)}\right)^T \nabla f\left(x^{(k)} + Y^{(k)} v\right) \end{cases} \quad (2.13)$$

where $B^{(k)}$ is an approximation to the reduced Hessian matrix $H^{(k)}$. The equations in (2.13) are closely related to the local Newton computation illustrated in Section 1. Indeed the difference is merely whether the "tangential" step is performed first, and the "half-step" is then defined by $(x^{(k)} + Z^{(k)} h^{(k)})$, or the "normal" step is performed.

2.1. *The descent curve*

Suppose $B^{(k)}$ is a positive definite matrix of order $n - m$, $(Z^{(k)})^T \nabla f(x^{(k)}) \neq 0$, and $B^{(k)} d_h^{(k)} = -(Z^{(k)})^T \nabla f(x^{(k)})$. Clearly the direction $Z^{(k)} d_h^{(k)}$ is a descent direction for p_μ

at $x^{(k)}$ and $s^{(k)}(\alpha) \equiv \alpha Z^{(k)} d_h^{(k)} + Y^{(k)}(R^{(k)})^{-T}[c(x^{(k)}) - c(x^{(k)} + \alpha Z^{(k)} d_h^{(k)})]$ is a descent curve. Our step size procedure determines a positive scalar $\alpha^{(k)}$ such that

$$p_\mu\left(u\left(\alpha^{(k)} d_h^{(k)}\right)\right) - p_\mu(u(0)) \le \sigma \alpha^{(k)} \nabla_h p_\mu(u(0))^T d_h^{(k)} \tag{2.14}$$

$$\nabla_h p_\mu u\left(\alpha^{(k)} d_h^{(k)}\right)^T d_h^{(k)} \ge \omega \nabla_h p_\mu(u(0))^T d_h^{(k)} \tag{2.15}$$

where $0 < \alpha < \omega < 1$.

Assuming that p_μ is bounded below along the path $h^{(k)}(\alpha) = x^{(k)} + s^{(k)}(\alpha)$, it is easy to establish the existence of a contiguous set of positive values of α satisfying conditions (2.14, 2.15). We state this result formally: the proof is a straightforward adaptation of Thm 6.3.2. in Dennis and Schnabel [9].

Lemma 2.2. *Suppose the functions $f, c_i : \Re^n \to \Re, i = 1, \ldots, m$ are continuously differentiable on $\Re^n$. Assume that $d_h^{(k)}$ is a direction satisfying $\nabla_h p_\mu(u(0))^T d_h^{(k)} < 0$, and $\{p_\mu(u(\alpha d_h^{(k)})) : \alpha > 0\}$ is bounded below. Then if $0 < \sigma < \omega < 1$ there exists α_h^l, α_h^u with $\alpha_h^u > \alpha_h^l > 0$ such that $\alpha d_h^{(k)}$ satisfies (2.14, 2.15) if $\alpha \in (\alpha_h^l, \alpha_h^u)$.*

The importance of this result is the implication that sufficient decrease along the *curved path* $u(d_h^{(k)}(\alpha))$ is compatible with the projected BFGS update e.g., [7]. This follows because (2.15) and $\nabla_h p_\mu(u(0))^T d_h^{(k)} < 0$ imply $(y^{(k)})^T s^{(k)} > 0$ where $y^{(k)} = \nabla p_\mu(h^{(k)}) - \nabla p_\mu(x^{(k)})$.

2.2. *The normal step*

Whereas the descent path $u(\alpha d_h^{(k)})$ decreases p_μ while approximately following $c(x) = c(x^{(k)})$, the normal direction decreases p_μ while simultaneously *decreasing* $\|c(x)\|$. In particular, a normal direction

$$d_v^{(k)} = -\left(R^{(k)}\right)^{-T} c\left(x^{(k)}\right)$$

is computed when $\|c(x^{(k)})\| > \Lambda^{(k)}\mu$, where $\Lambda^{(k)} = \max\{\|\lambda^{(k)}\|/\sigma, 1\}$ and the Lagrange multipliers λ are computed: $\lambda^{(k)} = -(R^{(k)})^{-1}(Y^{(k)})^T \nabla f(x^{(k)})$.

First we establish that $Y^{(k)} d_v^{(k)}$ is a descent direction.

Lemma 2.3. *Suppose the functions $f, c_i : \Re^n \to \Re, i = 1, \ldots, m$ are continuously differentiable on $\Re^n$ and $A^{(k)} = A(x^{(k)})$ is of full column rank. Assume that at $x^{(k)}$,*

$$\left\|c^{(k)}\right\| > \mu \left\|\lambda^{(k)}\right\|. \tag{2.16}$$

Then, the vector $Y^{(k)} d_v^{(k)}$ is a descent direction for p_μ at $x^{(k)}$.

Proof: Inequality (2.16) yields

$$\left(\lambda^{(k)}\right)^T c^{(k)} < \frac{\left\|c^{(k)}\right\|^2}{\mu}. \tag{2.17}$$

Therefore,

$$\begin{aligned}\nabla p_\mu\left(x^{(k)}\right)^T d_v^{(k)} &= Y^{(k)} d_v^{(k)} \\ &= \left(\lambda^{(k)}\right)^T c^{(k)} - \frac{\left\|c^{(k)}\right\|^2}{\mu} \\ &< 0. \end{aligned} \tag{2.18}$$

□

Assuming p_μ is bounded below in the direction $Y^{(k)} d_v^{(k)}$, a sufficient decrease step size condition follows (see Thm 6.3.2 in Dennis and Schnabel [9]):

Lemma 2.4. *Suppose the functions $f, c_i : \Re^n \to \Re$, $i = 1, \ldots, m$ are continuously differentiable on $\Re^n$. Assume $\{p_\mu(x^{(k)} + \beta Y^{(k)} d_v^{(k)}) : \beta > 0\}$ is bounded below. Then, if $0 < \sigma < \omega < 1$, there exist constants β_l^v, β_u^v with $\beta_u^v > \beta_l^v > 0$ such that $\beta \in (\beta_l^v, \beta_u^v)$ implies*

$$p_\mu\left(x^{(k)} + \beta Y^{(k)} d_v^{(k)}\right) \leq \sigma\beta \nabla p_\mu\left(x^{(k)}\right)^T Y^{(k)} d_v^{(k)}. \tag{2.19}$$

A simple backtracking procedure can be used to find a satisfying step length β. For example,

Algorithm βacktrack

- Let $0 < \tau < \tau' < 1$ be given. We perform the line search along the direction $Y_i^{(k)} v_i^{(k)}$ as follows.
  - Set $\beta := 1$;
  - Until the line search condition
    $$p_\mu\left(x^{(k)} + \beta Y^{(k)} v^{(k)}\right) - p_\mu\left(x^{(k)}\right) \leq \sigma\beta \nabla p_{\mu_i}\left(x_i^{(k)}\right)^T Y_i^{(k)} v_i^{(k)} \tag{2.20}$$
    is satisfied, choose a new $\beta \in [\tau\beta, \tau'\beta]$.

The penalty parameter μ must be driven to zero in the limit. For a given (fixed) value of μ the penalty function p_μ is approximately minimized. That is, first order necessary conditions for problem (1.1) are satisfied inexactly:

$$\begin{cases} a) & \left\|Z\left(x^{(k)}\right)^T \nabla f\left(x^{(k)}\right)\right\| \leq \mu^{1/2}, \\ b) & \left\|c\left(x^{(k)}\right)\right\| \leq \Lambda^{(k)} \mu. \end{cases} \tag{2.21}$$

Upon satisfaction of (2.21), μ is reduced, yielding μ_+ satisfying:

$$\mu^{6/5} \leq \mu_+ \leq \rho\mu, \tag{2.22}$$

where $\rho < 1$. Gould's analysis [14] underpins conditions (2.22).

Choose values $\mu_1 > 0, 0 < \sigma < 1 - \frac{1}{\sqrt{2}}, 0 < \rho < 1,$ and $\sigma < \omega < 1$.
Choose a point $x_0^* \in R^n$ and an $n \times n$ positive definite matrix $B_1^{(0)}$. Set $i \leftarrow 1$.
while (μ_i is not sufficiently small)
  Set $k \leftarrow 0, x_i^{(0)} \leftarrow x_{i-1}^*$;
  while either of the criteria in (2.21) does not hold
    if (2.21 b) does not hold
      Solve $(R_i^{(k)})^T d_{v_i}^{(k)} = -c(x_i^{(k)})$;
      βacktrack: Find a $\beta_i^{(k)} > 0$ satisfying (2.19)
      $x_i^{(k+)} \leftarrow x_i^{(k)} + \beta_i^{(k)} Y_i^{(k)} d_{v_i}^{(k)}$;
    else $x_i^{(k+)} \leftarrow x_i^{(k)}$;
    end;
    Solve $B_i^{(k)} d_{h_i}^{(k)} = -(Z_i^{(k+)})^T \nabla f(x_i^{(k+)})$;
    Path search: Find an $\alpha_i^{(k)} > 0$ satisfying (2.14) and (2.15);
    $x_i^{(k+1)} \leftarrow u(\alpha_i^{(k)} d_{h_i}^{(k)})$;
    $y_i^{(k)} \leftarrow \nabla_h p_{\mu_i}(x_i^{(k+1)}) - \nabla_h p_{\mu_i}(x_i^{(k+)})$,
    $B_i^{(k+1)} \leftarrow \mathrm{BFGS}(B_i^{(k)}, \alpha_i^{(k)}, d_{h_i}^{(k)}, y_i^{(k)})$;
    $k \leftarrow k + 1$;
  end;
  $x_i^* \leftarrow x_i^{(k)}$; $\mu_{i+1} \leftarrow \max\{\mu_i^{6/5}, \rho\|(Z_i^*)^T \nabla f_i^*\|^2\}$;
  $i \leftarrow i + 1$;
end;
Set $x^* \leftarrow x_i^*$ and STOP;

Figure 1. Algorithm 1.

2.3. *The algorithm*

Next we present the algorithm which mixes (tangential) path searches with normal steps. Note that the bracketed *superscripts* refer to inner loops—i.e., updating the iterate x—and the *subscripts* $i, i+1$ refer to the outer loop (where the penalty parameter μ_i is adjusted).

A monotone decrease result, for fixed μ_i, is easy to establish.

Lemma 2.5. *Assume a sequence* $\{x_i^{(k)}\}$ *is generated by Algorithm* 1 *with index i fixed. Then,*

$$p_{\mu_i}(x_i^{(k+1)}) - p_{\mu_i}(x_i^{(k)}) \le 0. \tag{2.23}$$

Furthermore, if

$$\|c(x_i^{(k)})\| > \Lambda_i^{(k)} \mu_i \tag{2.24}$$

where $\Lambda_i^{(k)} = \max\{\frac{\|\lambda(x_i^{(k)})\|}{\sigma}, 1\}$, *then*

$$\nabla p_{\mu_i}(x_i^{(k)})^T Y_i^{(k)} d_{v_i}^{(k)} = \lambda(x_i^{(k)})^T c(x_i^{(k)}) - \frac{1}{\mu_i}\|c(x_i^{(k)})\|^2. \tag{2.25}$$

Proof: The direction $d_{h_i}^{(k)}$ satisfies $B_i^{(k)} d_{h_i}^{(k)} = -\nabla_h p_{\mu_i}(u(0))$; this, along with (2.14), and the positive definiteness of $B_i^{(k)}$ implies that

$$p_{\mu_i}\left(x_i^{(k+1)}\right) - p_{\mu_i}\left(x_i^{(k+)}\right) \leq -\sigma\alpha\, \nabla_h p_{\mu_i}(u(0))^T \left(B_i^{(k)}\right)^{-1} \nabla_h p_{\mu_i}(u(0)) \leq 0. \tag{2.26}$$

If (2.24) does not hold, then $x_i^{(k+)} = x_i^{(k)}$ and (2.23) follows from (2.26).

If (2.24) holds, Eq. (2.25) follows from (2.11) and $d_{v_i}^{(k)} = -(R_i^{(k)})^{-T} c_i^{(k)}$. Since $\sigma < 1$, it follows from (2.20) that

$$\begin{aligned} p_{\mu_i}\left(x_i^{(k+)}\right) - p_{\mu_i}\left(x_i^{(k)}\right) &\leq \sigma\beta\left[\lambda\left(x_i^{(k)}\right)^T c\left(x_i^{(k)}\right) - \frac{1}{\mu_i}\left\|c\left(x_i^{(k)}\right)\right\|^2\right] \\ &\leq \sigma\beta\left[\left\|\lambda\left(x_i^{(k)}\right)\right\| \left\|c\left(x_i^{(k)}\right)\right\| - \frac{1}{\mu_i}\left\|c\left(x_i^{(k)}\right)\right\|^2\right] \\ &\leq \sigma\beta\left(\frac{\sigma - 1}{\mu_i}\right)\left\|c\left(x_i^{(k)}\right)\right\|^2 \leq 0. \end{aligned} \tag{2.27}$$

□

3. Global convergence

In this section we analyze the global convergence of Algorithm 1. We call $x^* \in \Re^n$ a *stationary point* of problem (1.1) if it satisfies

$$Z(x^*)^T \nabla f(x^*) = 0 \quad \text{and} \quad c(x^*) = 0. \tag{3.28}$$

The main result in this section, given in Theorem 3.1, states that under reasonable assumptions all limit points of the sequence $\{x_i^{(k)}\}$ are stationary points of problem (1.1). Moreover, in Theorem 3.2 we show that if there is a finite number of limit points then the whole sequence $\{x_i^{(k)}\}$ converges to a stationary point.

Assumption 3.1. The sequence $\{x_i^{(k)}\}$ generated by Algorithm 1 is contained in a bounded convex set D with the following properties:

1. The functions $f : \Re^n \to \Re$, and $c : \Re^n \to R^m$ and their first and second derivatives are uniformly bounded in norm over D.
2. The matrix $A(x)$ has full column rank for all $x \in D$, and there is a constant K_0 such that

$$\|A(x)[A(x)^T A(x)]^{-1}\| \leq K_0 \tag{3.29}$$

for all $x \in D$.

Note that (3.29) implies that $\|Y_i^{(k)}(R_i^{(k)})^{-1}\| \leq K_0$ for all i and k. We now prove that the step lengths $\beta_i^{(k)}$ are bounded away from zero.

Lemma 3.1. *Suppose Assumption* 3.1 *is satisfied. Then there is a constant* $\beta > 0$, *such that*

$$\beta_i^{(k)} \geq \beta > 0$$

whenever $\beta_i^{(k)}$ *is computed in Algorithm* 1.

Proof: First, note that in Algorithm 1 the normal step, and thus the step length $\beta_i^{(k)}$, is computed only when $\|c(x_i^{(k)})\| > \Lambda_i^{(k)}\mu_i$. Suppose that $\beta_i^{(k)} < 1$. If step length $\beta \leq 1$ is the most recent failure of (2.20) in algorithm βacktrack, then

$$p_{\mu_i}\big(x_i^{(k)} + \tilde{\beta} Y_i^{(k)} d_{v_i}^{(k)}\big) - p_{\mu_i}\big(x_i^{(k)}\big) > \sigma\tilde{\beta}\big[\nabla p_{\mu_i}\big(x_i^{(k)}\big)\big]^T Y_i^{(k)} d_{v_i}^{(k)} \tag{3.30}$$

and $\tau\tilde{\beta} \leq \beta_i^{(k)}$. Since matrices $\nabla^2 c_i(x)$ and $\nabla^2 f(x)$ are bounded and

$$\nabla^2 p_{\mu_i}(x) = \nabla^2 f(x) + \sum_{i=1}^{m} \frac{c_i(x)}{\mu_i}\nabla^2 c_i(x) + \frac{1}{\mu_i} A(x)A(x)^T,$$

Taylor's theorem yields that for some $\tilde{x}_i^{(k)}$ near $x_i^{(k)}$

$$\begin{aligned} &p_{\mu_i}\big(x_i^{(k)} + \tilde{\beta} Y_i^{(k)} d_{v_i}^{(k)}\big) - p_{\mu_i}\big(x_i^{(k)}\big) \\ &\leq \tilde{\beta}\big[\nabla p_{\mu_i}\big(x_i^{(k)}\big)\big]^T Y_i^{(k)} d_{v_i}^{(k)} + \frac{\tilde{\beta}^2}{2}\big(Y_i^{(k)} d_{v_i}^{(k)}\big)^T\big[\nabla^2 p_{\mu_i}\big(\tilde{x}_i^{(k)}\big)\big]\big(Y_i^{(k)} d_{v_i}^{(k)}\big) \\ &\leq \tilde{\beta}\big[\nabla p_{\mu_i}\big(x_i^{(k)}\big)\big]^T Y_i^{(k)} d_{v_i}^{(k)} + \tilde{\beta}^2 \frac{K}{\mu_i}\big\|Y_i^{(k)} d_{v_i}^{(k)}\big\|^2 \end{aligned} \tag{3.31}$$

where K is a constant independent of μ_i. Noting that $Y_i^{(k)} d_{v_i}^{(k)} = -Y_i^{(k)}(R_i^{(k)})^{-T} c(x_i^{(k)})$, it then follows, using (3.30) and (3.29), that

$$-(1-\sigma)\big[\nabla p_{\mu_i}\big(x_i^{(k)}\big)\big]^T Y_i^{(k)} d_{v_i}^{(k)} < \tilde{\beta}\frac{K}{\mu_i}\big\|Y_i^{(k)} d_{v_i}^{(k)}\big\|^2 \leq \tilde{\beta}\frac{K K_0^2}{\mu_i}\big\|c\big(x_i^{(k)}\big)\big\|^2. \tag{3.32}$$

On the other hand, since $\|\lambda_i^{(k)}\| \leq \Lambda_i^{(k)} < \frac{\|c(x_i^{(k)})\|}{\mu_i}$ whenever $\beta_i^{(k)}$ is computed, Eq. (2.25) implies that

$$\begin{aligned} \nabla p_{\mu_i}\big(x_i^{(k)}\big)^T Y_i^{(k)} d_{v_i}^{(k)} &\leq \big\|c\big(x_i^{(k)}\big)\big\|\big\|\cdot\lambda_i^{(k)}\big\| - \frac{\big\|c\big(x_i^{(k)}\big)\big\|^2}{\mu_i} \\ &\leq -(1-\sigma)\frac{\big\|c\big(x_i^{(k)}\big)\big\|^2}{\mu_i}. \end{aligned} \tag{3.33}$$

Combining (3.32) and (3.33), we obtain that

$$\beta_i^{(k)} \geq \tau\ \tilde{\beta} > \tau \frac{(1-\sigma)^2}{K K_0^2}.$$

□

Next we show that for any fixed $\mu_i > 0$ the criteria in Algorithm 1, i.e., (2.21), can be satisfied after a finite number of iterations. Similar to most convergence analysis for quasi-Newton methods, we need to make some boundedness assumptions on the matrices $\{B_i^{(k)}\}$. It should be noted that such assumptions may not always hold, say, for the matrices generated by the BFGS update.

Lemma 3.2. *Let Assumption* 3.1 *hold, suppose that sequence* $\{x_i^{(k)}\}$ *is generated by Algorithm* 1, *and* μ_i *is held at a constant value (by Algorithm* 1). *Furthermore, assume that there exists a constant* $M > 0$ *such that*

$$\mathrm{eig}_{\max}\left(B_i^{(k)}\right) \leq M^{\frac{1}{2}}, \qquad \mathrm{eig}_{\min}\left(B_i^{(k)}\right) \geq M^{-\frac{1}{2}}, \tag{3.34}$$

where $\mathrm{eig}_{\max}$ *and* $\mathrm{eig}_{\min}$ *denote the greatest and the least eigenvalues, respectively. Then there exists an integer* $\bar{k}$, *such that for* $k \geq \bar{k}$

$$\left\|c\left(x_i^{(k)}\right)\right\| \leq \Lambda_i^{(k)} \mu_i \quad \text{and} \quad \left\|Z\left(x_i^{(k)}\right)^T \nabla f\left(x_i^{(k)}\right)\right\|^2 \leq \mu_i. \tag{3.35}$$

Proof: First we prove, by contradiction, that there exists an integer $\bar{k} > 0$ such that for all $k \geq \bar{k}$,

$$\left\|c\left(x_i^{(k)}\right)\right\| \leq \Lambda_i^{(k)} \mu_i. \tag{3.36}$$

If (3.36) does not hold for all k sufficiently large, there exists a subsequence $\{k_s\}$ such that

$$\left\|c\left(x_i^{(k_s)}\right)\right\| > \Lambda_i^{(k_s)} \mu_i \geq \mu_i.$$

Thus, it follows from (2.23), (2.26), and (2.27) that

$$\begin{aligned} p_{\mu_i}\left(x_i^{k_s+1}\right) - p_{\mu_i}\left(x_i^{(k_s)}\right) &\leq p_{\mu_i}\left(x_i^{k_s+1}\right) - p_{\mu_i}\left(x_i^{(k_s)}\right) \\ &\leq -\sigma(1-\sigma)\beta_i^{(k_s)} \frac{\left\|c\left(x_i^{(k_s)}\right)\right\|^2}{\mu_i} \\ &< -\sigma(1-\sigma)\beta_i^{(k_s)} \mu_i. \end{aligned}$$

Thus Lemma 3.1 implies that

$$p_{\mu_i}\left(x_i^{(k_s+1)}\right) - p_{\mu_i}\left(x_i^{(k_s)}\right) \leq -\sigma(1-\sigma)\beta\mu_i$$

which contradicts the fact that $p_{\mu_i}(x)$ is bounded below for any fixed $\mu_i > 0$.

Since μ_i is not further decreased by Algorithm 1, by assumption, and (3.36) holds for all $k \geq \bar{k}$, it must be that

$$\left\| \left(Z_i^{(k)}\right)^T \nabla f\left(x_i^{(k)}\right) \right\| > \mu_i^{1/2} \tag{3.37}$$

for all $k \geq \bar{k}$.

Inequality (2.14) and the positive definiteness of $B_i^{(k)}$ yield that

$$\begin{aligned} p_{\mu_i}\left(x_i^{(k+1)}\right) - p_{\mu_i}\left(x_i^{(k)}\right) &\leq \sigma\left[Z\left(x_i^{(k)}\right)^T \nabla p_{\mu_i}\left(x_i^{(k)}\right)\right]^T \left(\alpha_i^{(k)} d_{h_i}^{(k)}\right) \\ &\leq -\sigma \alpha_i^{(k)} \left(d_{h_i}^{(k)}\right)^T B_i^{(k)} d_{h_i}^{(k)} < 0. \end{aligned}$$

Since $p_{\mu_i}(x)$ is bounded below, it follows that

$$\sum_{k=0}^{\infty} \left| \left[\left(Z_i^{(k)}\right)^T \nabla f\left(x_i^{(k)}\right)\right]^T \left(\alpha_i^{(k)} d_{h_i}^{(k)}\right) \right| < +\infty.$$

Therefore, $\lim_{k\to\infty} |[(Z_i^{(k)})^T \nabla f(x_i^{(k)})]^T (\alpha_i^{(k)} d_{h_i}^{(k)})| = 0$. Combining with (3.37), we get $\lim_{k\to\infty} \|\alpha_i^{(k)} d_{h_i}^{(k)}\| = 0$.

Inequality (2.15) implies that

$$\begin{aligned} &-\left[Z\left(x_i^{(k)}\right)^T \nabla f\left(x_i^{(k)}\right)\right]^T \left(\alpha_i^{(k)} d_{h_i}^{(k)}\right) \\ &\quad \leq \frac{\left[\nabla_h p_{\mu_i}\left(u\left(\alpha_i^{(k)} d_{h_i}^{(k)}\right)\right) - \nabla_h p_{\mu_i}(u(0))\right]^T \left[\alpha_i^{(k)} d_{h_i}^{(k)}\right]}{1-\omega} \\ &\quad \leq \frac{\left\|\nabla_h p_{\mu_i}\left(u\left(\alpha_i^{(k)} d_{h_i}^{(k)}\right)\right) - \nabla_h p_{\mu_i}(u(0))\right\| \left\|\alpha_i^{(k)} d_{h_i}^{(k)}\right\|}{1-\omega}. \end{aligned} \tag{3.38}$$

It follows from (3.37), (3.38), and the uniform continuity assumption that, as $k \to \infty$,

$$\frac{\left|\left[Z\left(x_i^{(k)}\right)^T \nabla f\left(x_i^{(k)}\right)\right]^T \left(\alpha_i^{(k)} d_{h_i}^{(k)}\right)\right|}{\left\|Z\left(x_i^{(k)}\right)^T \nabla f\left(x_i^{(k)}\right)\right\| \left\|\alpha_i^{(k)} d_{h_i}^{(k)}\right\|} \leq \frac{\left\|\nabla_h p_{\mu_i}\left(u\left(\alpha_i^{(k)} d_{h_i}^{(k)}\right)\right) - \nabla_h p_{\mu_i}(u(0))\right\|}{\mu_i^{1/2} \cdot (1-\omega)} \to 0. \tag{3.39}$$

On the other hand, since $Z(x_i^{(k)})^T \nabla f(x_i^{(k)}) = -B_i^{(k)} d_{h_i}^{(k)}$, it follows from (3.34) that for any $\bar{k} \geq 0$ there exists an integer $k \geq \bar{k}$ such that

$$-\frac{\left[Z\left(x_i^{(k)}\right)^T \nabla f\left(x_i^{(k)}\right)\right]^T \left(\alpha_i^{(k)} d_{h_i}^{(k)}\right)}{\left\|\left(Z_i^{(k)}\right)^T \nabla f\left(x_i^{(k)}\right)\right\| \left\|\alpha_i^{(k)} d_{h_i}^{(k)}\right\|} \geq \frac{\left(d_{h_i}^{(k)}\right)^T B_i^{(k)} d_{h_i}^{(k)}}{\left\|B_i^{(k)} d_{h_i}^{(k)}\right\| \left\|d_{h_i}^{(k)}\right\|} \geq \frac{1}{M} > 0.$$

This inequality contradicts (3.39). □

It clearly follows from Lemma 3.2 that Algorithm 1 generates an infinite sequence of finite sequences:

$$\{x_i^{(k)}\} = \{x_1^{(0)}, x_1^{(1)}, \ldots, x_1^{(k_1-1)}, x_1^{(k_1)} = x_2^{(0)}, x_2^{(1)}, \ldots, x_2^{(k_2-1)}, \ldots, x_i^{(0)}, x_i^{(1)}, \ldots, x_i^{(k_i)} = x_{i+1}^{(0)}, \ldots\}$$

Lemma 3.3. *Let Assumption* 3.1 *hold and assume that there exists a positive constant M such that* (3.34) *is valid. Then*

$$\lim_{i\to\infty} \left[\left\| Z\left(x_i^{(k_i)}\right)^T \nabla f\left(x_i^{(k_i)}\right)\right\| + \left\| c\left(x_i^{(k_i)}\right)\right\|\right] = 0. \tag{3.40}$$

Proof: By Lemma 3.2, Algorithm 1 generates an infinite sequence of iterates satisfying (2.21) for values of $\mu = \mu_i$ converging to zero. But by our assumptions $\{|\Lambda_i^{(k)}|\}$ is bounded: the result follows. □

Before we show that all limit points are stationary points we establish a required boundedness result.

Lemma 3.4. *Suppose that the assumptions in Lemma* 3.3 *hold. Then*

$$\sum_{i=1}^{\infty}\sum_{k=0}^{k_i-1}\left[p_{\mu_i}\left(x_i^{(k)}\right) - p_{\mu_i}\left(x_i^{(k+1)}\right)\right] < +\infty. \tag{3.41}$$

Proof: It follows from Assumption 3.1 that there exists a constant $N_1 > 0$ such that for all integers $i > 0$ and $0 \le k \le k_i$, $|\Lambda_i^{(k)}| \le N_1$. Thus

$$\left\| c\left(x_i^{(0)}\right)\right\| \le \Lambda_i^{(k)} \mu_{i-1} \le N_1\, \mu_{i-1}.$$

Notice that since $x_i^{(k_i)} = x_{i+1}^{(0)}$, it follows that

$$\begin{aligned}
\sum_{i=1}^{\infty}\sum_{k=0}^{k_i-1}\left[p_{\mu_i}\left(x_i^{(k)}\right) - p_{\mu_i}\left(x_i^{(k+1)}\right)\right] &= \sum_{i=1}^{\infty}\left[p_{\mu_i}\left(x_i^{(0)}\right) - p_{\mu_i}\left(x_{i+1}^{(0)}\right)\right] \\
&= \sum_{i=1}^{\infty}\left[p_{\mu_i}\left(x_i^{(0)}\right) - p_{\mu_{i-1}}\left(x_i^{(0)}\right)\right] \\
&\quad + \sum_{i=1}^{\infty}\left[p_{\mu_{i-1}}\left(x_i^{(0)}\right) - p_{\mu_i}\left(x_{i+1}^{(0)}\right)\right] \\
&\le \sum_{i=2}^{\infty}\left[p_{\mu_i}\left(x_i^{(0)}\right) - p_{\mu_{i-1}}\left(x_i^{(0)}\right)\right] + N_2,
\end{aligned}$$

where $N_2 = p_{\mu_1}(x_1^{(0)}) - \inf\{p_{\mu_i}(x)\}$ is a constant since $p_{\mu_i}(x)$ is bounded below.

It follows from Algorithm 1 and (2.22) that

$$p_{\mu_i}(x_i^{(0)}) - p_{\mu_{i-1}}(x_i^{(0)}) = \left[\frac{1}{\mu_i} - \frac{1}{\mu_{i-1}}\right]\|c(x_i^{(0)})\|^2 \leq \frac{\mu_{i-1}^2}{\mu_i} N_1^2 \leq \mu_{i-1}^{4/5} N_1^2.$$

Therefore, (2.22) implies that

$$\sum_{i=1}^{\infty}\sum_{k=0}^{k_i-1}\left[p_{\mu_i}(x_i^{(k)}) - p_{\mu_i}(x_i^{(k+1)})\right] \leq N_1^2 \sum_{i=2}^{\infty} \mu_{i-1}^{4/5} + N_2 \leq \frac{N_1^2 \mu_1^{4/5}}{1-\rho^{4/5}} + N_2.$$

□

We can now prove that every limit point of the sequence $\{x_i^{(k)}\}$ is a stationary point of problem (1.1).

Theorem 3.1. *Suppose that the conditions in Assumptions* 3.1 *and* 3.2 *are satisfied. Define the sequence* $\{x_k\}$ *to be the entire sequence, relabeled; i.e.,*

$$\{x_k\} = \{x_1^{(0)}, x_1^{(1)}, \ldots, x_1^{(k_1-1)}, x_1^{(k_1)} = x_2^{(0)}, x_2^{(1)}, \ldots\}.$$

Then

$$\lim_{k\to\infty} \|c(x_k)\| = 0, \tag{3.42}$$

and

$$\lim_{k\to\infty} \|Z(x_k)^T \nabla f(x_k)\| = 0. \tag{3.43}$$

Proof: To prove (3.42), we define

$$\gamma(x) = \begin{cases} \dfrac{\|c(x)\|^2}{\mu} - c(x)^T\lambda(x) & \text{if } \|c(x)\| > \Lambda\mu_i \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, $\gamma_k \geq 0$. It is obvious from Lemma 3.1 and Lemma 2.2 that

$$\sigma\beta\gamma_i^{(k)} \leq p_{\mu_i}(x_i^{(k)}) - p_{\mu_i}(x_i^{(k+)}) \leq p_{\mu_i}(x_i^{(k)}) - p_{\mu_i}(x_i^{(k+1)})$$

which, with Lemma 3.3, implies that

$$\sum_{k=k_0}^{\infty}\gamma_k \leq \sum_{i=1}^{\infty}\sum_{k=0}^{k_i-1}\gamma_i^{(k)} \leq \frac{1}{\sigma\beta}\sum_{i=1}^{\infty}\sum_{k=0}^{k_i-1}\left[p_{\mu_i}(x_i^{(k)}) - p_{\mu_i}(x_i^{(k+1)})\right] < +\infty.$$

Thus, $\lim_{k\to\infty}\gamma_k = 0$. Notice that since $\mu_i \to 0$ and $\|\lambda_i\|$ is bounded, it follows from the definition of $\{d_k\}$ that (3.42) holds.

To prove (3.43), note that from Lemma 3.3 it follows that for all $0 \le k \le k_i - 1$.

$$p_{\mu_i}\left(x_i^{(k)}\right) - p_{\mu_i}\left(x_i^{(k+1)}\right) \to 0 \quad \text{as } i \to \infty.$$

Similar to the proof of Lemma 3.2, it follows that for all $0 \le k \le k_i$

$$-\left[Z\left(x_i^{(k)}\right)^T \nabla f\left(x_i^{(k)}\right)\right]^T \left(\alpha_i^{(k)} h\right) \to 0 \quad \text{as } i \to \infty.$$

Assuming (3.43) does not hold, then there exist an $\epsilon > 0$ and a subsequence $k_i \in \mathcal{S}$ such that

$$\left\| Z\left(x_{k_i}\right)^T \nabla f\left(x_{k_i}\right) \right\| \ge \epsilon \quad \text{for } k_i \in \mathcal{S},$$

then

$$\left\| \alpha_{k_i} h_{k_i} \right\| \to 0 \quad \text{for } k_i \in \mathcal{S}.$$

And since $y_{k_i} = \alpha_{k_i} B_{k_i+1} h_{k_i}$, similar to the proof of Lemma 3.2, it follows that for $k_i \in \mathcal{S}$

$$-\frac{\left[Z\left(x_{k_i}\right)^T \nabla f\left(x_{k_i}\right)\right]^T \left(\alpha_{k_i} h_{k_i}\right)}{\left\| Z\left(x_{k_i}\right)^T \nabla f\left(x_{k_i}\right) \right\| \left\| \alpha_{k_i} h_{k_i} \right\|} \le \frac{\left\| \nabla_h p_{\mu_i}\left(u\left(\alpha_{k_i} h_{k_i}\right)\right) - \nabla_h p_{\mu_i}(u(0)) \right\|}{\epsilon(1-\omega)}$$

$$= \frac{\left\| y_{k_i} \right\|}{\epsilon(1-\omega)} \le \frac{M \left\| \alpha_{k_i} h_{k_i} \right\|}{\epsilon(1-\omega)} \to 0.$$

□

Finally, if there are only a finite number of limit points to problem (1.1), then the sequence $\{x_i^{(k)}\}$ converges.

Theorem 3.2. *Suppose that the conditions in Assumptions* 3.1 *and* 3.2 *are satisfied and that the sequence* $\{x_k\}$ *is the one described in Theorem* 3.1. *Moreover, suppose that every stationary point of* (1.1) *is isolated. Then*

$$\lim_{k \to \infty} x_k = x^* \tag{3.44}$$

holds, where x^* *is a stationary point of* (1.1).

Proof: Since $\{x_k\}$ is bounded, there exists a subsequence x_{k_j} of $\{x_k\}$ such that

$$\lim_{j \to \infty} x_{k_j} = x^*,$$

where $x^* \in D$ is an accumulation point of $\{x_k\}$. But by Lemma 3.5, (3.28) holds at x^*. That is, x^* is a stationary point of (1.1). Therefore, x^* is an isolated accumulation point of $\{x_k\}$.

Now we prove (3.44) by contradiction. Suppose $\{x_k\}$ does not converge. Since x^* is an isolated accumulation point of $\{x_k\}$, there exists a subsequence $\{x_{k_j}\}$ of $\{x_k\}$ and an $\epsilon > 0$ such that $\|x_{k_j+1} - x_{k_j}\| \ge \epsilon$ (Lemma 4.10, [16]). But using Theorem 3.1 it follows from (3.41) that $\lim_{k\to\infty} \|h_k\| = 0$ and $\lim_{k\to\infty} \|v_k\| = 0$. Hence, $\lim_{k\to\infty} \|x_{k+1} - x_k\| = 0$. □

3.1. *Remarks*

In [4] Byrd and Nocedal propose algorithms based on reduced Hessian methods. Byrd and Nocedal prove that, for their algorithms,

$$\lim_{k\to\infty} [\|Z(x_k)^T \nabla f(x_k)\| + \|c(x_k)\|] = 0 \tag{3.45}$$

under an assumption stronger than condition (3.34). In particular, Byrd and Nocedal assume that there exists a $\gamma > 0$ such that

$$\text{eig}_{\min}\left(Z_k^T \nabla^2 L(x, \lambda_k) Z_k\right) \geq \gamma, \quad \forall x \text{ in the line search segment.} \tag{3.46}$$

Moreover, algorithms in [4] cannot preserve the positive definiteness of B_k without assumption (3.46). However, assumption (3.46) is rarely satisfied when x_k is far away from the solution. Therefore, in contrast to Algorithm 1, algorithms in [4] may fail when applied to general nonlinear functions.

4. Numerical results

In this section we present-results of numerical experiments illustrating the performance of Algorithm 1. The problem set consists of a number of nonlinear equality constrained problems selected from the CUTE collection [3] and two problems generated by the authors. All numerical experiments discussed in this section were performed in MATLAB Version 4.1 on a Sun 4/670 workstation.

All test problems are briefly described in Table 1. Most problems in Table 1 (all except TEST1 and TEST2) are from the CUTE collection [3]. Problem TEST1 is minimization of

Table 1. Description of problems.

Problems	n	m	$nnz(A)$	Constraints
BT6	5	2	5	Nonlinear
BT11	5	3	8	Nonlinear
DIPIGRI	7	4	19	Nonlinear
DTOC2	58	36	144	Nonlinear
DTOC4	29	18	65	Nonlinear
DTOC6	21	10	31	Nonlinear
GENHS28	300	298	894	Linear
HS100	7	4	19	Nonlinear
MWRIGHT	5	3	8	Nonlinear
ORTHREGA	517	256	1792	Nonlinear
ORTHREGC	505	250	1750	Nonlinear
ORTHREGD	203	100	500	Nonlinear
TEST1	200	160	dense	Quadratic
TEST2	200	160	dense	Nonlinear

a Rosenbrock function [9] with quadratic equality constraints, i.e.,

$$\begin{aligned} &\text{minimize} \quad \sum_{i=1}^{n-1}\left[(1-x_i)^2 + 100\left(x_{i+1}-x_i^2\right)^2\right] \\ &\text{subject to} \quad a_i^T x + .5x^T M_i x = 0, \quad i = 1, 1, \ldots, m, \end{aligned}$$

where $a_i \in \Re^n$, $i = 1, 2, \ldots, m$, are vectors, and $M_i \in \Re^{n\times n}$, $i = 1, 2, \ldots, m$, are symmetric. The nonlinearity of problem TEST1 is high if the symmetric matrices M_i, $i = 1, 2, \ldots, m$, are not extremely sparse. In problem TEST2 we add perturbation functions to both the objective and constraint functions in TEST1. Namely, problem TEST2 is

$$\begin{aligned} &\text{minimize} \quad \sum_{i=1}^{n-1}\left[(1-x_i)^2 + 100\left(x_{i+1}-x_i^2\right)^2\right] + \delta_0(x) \\ &\text{subject to} \quad a_i^T x + .5x^T M_i x + \delta_i(x) = 0, \quad i = 1, 1, \ldots, m, \end{aligned}$$

where $\delta_0(x)$, $\delta_i(x)$, $i = 1, 2, \ldots, m$, are perturbation functions. The perturbation functions are generated randomly to be linear combinations of polynomials, trigonometric functions, logarithmic functions and exponential functions. For example, $\delta_0(x)$ could be

$$\begin{aligned} \delta_0(x) = &\left(x_1^2 + x_4\right)^2 + 1.0 + \log\left(1 + x_2^2 + x_3^2\right) \\ &+ 10\sin(2\pi x_5)\cos(2\pi x_6) - e^{-(x_5-x_3)^2} + \ldots . \end{aligned}$$

In problems TEST1 and TEST2, the matrices $A = [a_1, a_2, \ldots, a_m]$ and M_i, $i = 1, 2, \ldots, m$ are created randomly.

When solving problems in Table 1 using Algorithm 1, we take $\mu_0 = 1$, $\rho = 0.1$ and $\sigma = 0.0001$. We set the stopping criterion to be $\mu < 10^{-8}$. For problems TEST1 and TEST2, the starting point is $x_0 = [.5, \ .5, \ \ldots, \ .5]^T$. For the testing problems drawn from the CUTE collection, we take the default values.

Table 2 illustrates the results of our numerical experiments for problems in Table 1. The first column gives the name of the problems we solved. The second column shows the number of iterations (the total number of "inner" iterations, i.e., the sum of k for every μ_i) taken to satisfy the stopping criterion for different problems. Column "function evaluations" presents the number of function evaluations needed for problems in Table 1. Sub-column "f, c" ("g, A") indicates the number of function (gradient) evaluations required. The last column shows how accurate Algorithm 1 reaches when applied to the testing problems, where the quantity "*error*" is defined as

$$error = \sqrt{\|Z(x)^T \nabla f(x)\|_2^2 + \|c(x)\|_2^2}.$$

The numerical experiment results indicate that the proposed method is quite robust for different kinds of constrained problems. The number of function evaluations is generally low in the test. Extensive numerical experiment is under way and a thorough comparison with some well-known existing algorithms will be presented in a future report.

Table 2. Results using Algorithm 1.

Problems	Number of iterations	Function evaluations		*error*
		f, c	g, A	
BT6	12	37	21	$O(10^{-6})$
BT11	9	25	18	$O(10^{-7})$
DIPIGRI	16	70	27	$O(10^{-6})$
DTOC2	12	17	17	$O(10^{-5})$
DTOC4	4	8	8	$O(10^{-5})$
DTOC6	11	18	17	$O(10^{-6})$
GENHS28	6	9	8	$O(10^{-6})$
HS100	17	75	29	$O(10^{-7})$
MWRIGHT	14	36	22	$O(10^{-5})$
ORTHREGA	83	883	91	$O(10^{-5})$
ORTHREGC	24	61	31	$O(10^{-5})$
ORTHREGD	21	90	38	$O(10^{-5})$
TEST1	104	443	116	$O(10^{-5})$
TEST2	149	526	161	$O(10^{-5})$

5. Discussion and concluding remarks

We have presented a quasi-Newton quadratic penalty method for solving equality constrained minimization problems. When quasi-Newton methods are applied to nonlinear equality constrained minimization problems, one of the major difficulties is preserving positive definiteness of the approximating matrices in a reasonable and robust way. In addition, due to the effect of the penalty term, the quadratic penalty function often forces steps to be short when far from the solution. In this paper we have proposed a new approach which not only maintains positive definite Hessian approximations, but also avoids unacceptably small steps when far from the solution.

The pivotal feature of our approach is a local transformation, defined at the current iterate, that leads to a *curved* correction path. The curved path allows for long steps; moreover, the quasi-Newton update, defined by the current and next point along the curve, naturally yields a well-behaved approximation to the reduced Hessian of the Lagrangian function.

We have established global convergence properties; superlinear convergence is conjectured and will be studied in a subsequent paper. Numerical results of preliminary computational experiments indicate practical potential. Indeed, the theoretical properties along with our numerical results indicate that our algorithm has considerable potential for efficiently solving nonlinear equality constrained minimization problems.

Acknowledgments

We thank the referees for several useful references and for many helpful comments that improved the presentation of this paper.

References

1. M.C. Bartholomew-Biggs, "Constrained minimization using recursive quadratic programming," in Numerical Methods for Nonlinear Optimization, F.A. Lootsma (Ed.), Academic Press: London, 1972, pp. 411–428.
2. P.T. Boggs, J.W. Tolls, and P. Wang, "On the local convergence of quasi-Newton methods for constrained optimization," SIAM Journal of Control and Optimization, vol. 20, pp. 161–171, 1982.
3. I. Bongartz, A.R. Conn, N.I.M. Could, and Ph.L. Toint, "CUTE: Constrained and unconstrained testing environment," Research Report, IBM T.J. Watson Research Center, Yorktown Heights, New York, 1993.
4. R.H. Byrd and J. Nocedal, "An analysis of reduced Hessian methods for constrained optimization," Mathematical Programming, vol. 49, pp. 285–323, 1991.
5. R.H. Byrd and R.B. Schnabel, "Continuity of the null space basis and constrained optimization," Mathematical Programming, vol. 35, pp. 32–41, 1986.
6. T.F. Coleman, "On characterizations of superlinear convergence for constrained optimization," Lectures in Applied Mathematics, vol. 26, pp. 113–133, 1990.
7. T.F. Coleman and A.R. Conn, "On the local convergence of a quasi-Newton method for the nonlinear programming problem," SIAM Journal on Numerical Analysis, vol. 21, pp. 755–769, 1984.
8. T.F. Coleman and C. Hempel, "Computing a trust region step for a penalty function," SIAM Journal on Scientific Computing, vol. 11, pp. 180–201, 1990.
9. J.E. Dennis, Jr. and R.B. Schnabel, Numerical Methods for Unconstrained Optimization and Nonlinear Equations, Prentice-Hall: Englewood Cliffs, NJ, 1983.
10. A.V. Fiacco and G.P. McCormick, Nonlinear Programming: Sequential Unconstrained Minimization Techniques, John Wiley and Sons, 1968.
11. R. Fontecilla, "Local convergence of secant methods for nonlinear constrained optimization," SIAM Journal on Numerical Analysis, vol. 25, pp. 692–712, 1988.
12. D. Gabay, "Reduced quasi-Newton methods with feasibility improvement for nonlinearly constrained optimization," Mathematical Programming Study, vol. 16, pp. 18–44, 1982.
13. N.I.M. Gould, "On the accurate determination of search directions for simple differentiable penalty functions," I.M.A. Journal on Numerical Analysis, vol. 6, pp. 357–372, 1986.
14. N.I.M. Gould, "On the convergence of a sequential penalty function method for constrained minimization," SIAM J. on Numerical Analysis, vol. 26, pp. 107–128, 1989.
15. W.W. Hager, "Analysis and implementation of a dual algorithm for constrained optimization," Journal of Optimization Theory and Applications, vol. 79, pp. 427–462, 1993.
16. J.J. Moré and D.C. Sorensen, "Computing a trust region step," SIAM Journal on Scientific and Statistical Computing, vol. 4, pp. 553–572, 1983.
17. W. Murray, "An algorithm for constrained minimization," in Optimization, R. Fletcher (Ed.), Academic Press, London, 1969, pp. 189–196.
18. J. Nocedal and M. Overton, "Projected Hessian updating algorithms for nonlinearly constrained optimization," SIAM Journal on Numerical Analysis, vol. 22, pp. 821–850, 1985.