MARYLAND

TECHNICAL RESEARCH REPORT

SRC TR 89-41

Avoiding the Maratos Effect by Means of a Nonmonotone Line Search: I. General Constrained Problems

By

E. R. Panier
and
A. L. Tits

SYSTEMS RESEARCH CENTER

UNIVERSITY OF MARYLAND

COLLEGE PARK, MARYLAND 20742

Avoiding the Maratos Effect by Means of a Nonmonotone Line Search
I. General Constrained Problems [1]

Eliane R. Panier[†] *and André L. Tits* [‡]

Abstract.

An essential condition for quasi-Newton optimization methods to converge superlinearly is that a full step of one be taken close to the solution. It is well known that, when dealing with constrained optimization problems, line search schemes ensuring global convergence of such methods may prevent this from occurring (the so called "Maratos effect"). Two types of techniques have been used to circumvent this difficulty. In the watchdog technique, the full step of one is occasionally accepted even when the line search criterion is violated; subsequent backtracking is used if global convergence appears to be lost. In a "bending" technique proposed by Mayne and Polak, backtracking is avoided by performing a search along an arc whose construction requires evaluation of constraint functions at an auxiliary point; along this arc, the full step of one is accepted close to a solution.

The main idea in the present paper is to combine Mayne and Polak's technique with a nonmonotone line search proposed by Grippo, Lampariello and Lucidi in the context of unconstrained optimization, in such a way that, asymptotically, function evaluations are no longer performed at auxiliary points. In a companion paper (part II), it is shown that a refinement of this scheme can be used in the context of recently proposed SQP-based methods generating feasible iterates.

Key words: constrained optimization, sequential quadratic programming, Maratos effect, superlinear convergence.

AMS(MOS) subject classifications: 90C30, 65K10.

[1] This research was supported in part by NSF's Engineering Research Centers Program No. NSFD-CDR-88-03012 and by NSF grant No. DMC-84-51515. The second author performed parts of this work while visiting INRIA (project META2), Rocquencourt, France, in the fall of 1988.

[†] Systems Research Center, University of Maryland, College Park, MD 20742

[‡] Systems Research Center and Electrical Engineering Department, University of Maryland, College Park, MD 20742

1. Introduction.

Consider the optimization problem

$$\min f(x) \text{ s.t. } h(x) = 0, \qquad g(x) \leq 0, \tag{P}$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $h : \mathbb{R}^n \to \mathbb{R}^{m_e}$, and $g : \mathbb{R}^n \to \mathbb{R}^{m_i}$, are smooth functions. Quasi-Newton methods have been extensively used for the solution of such problems in the framework of Sequential Quadratic Programming (SQP). Global convergence can be induced via a line search requiring at each step the decrease of a certain merit function. However, while superlinear convergence requires that a full step of one be taken close to a solution of (P), such line search may not accept a full step even arbitrarily close to a solution. This phenomenon has been first pointed out by Maratos in his Ph. D. Thesis [7]. Two types of techniques have been proposed to avoid this undesirable effect. In the watchdog technique [3] the step of one is tentatively accepted if sufficient decrease was achieved at the *previous* iteration, compared to the lowest value of the merit function obtained so far. If this lowest value is not improved upon within a given finite number of iterations, the algorithm restarts from the iterate at which this value was achieved. Under suitable conditions it is shown that a step of one is always accepted in the vicinity of a "strong" local solution of (P).[2] However, in the early iterations, numerous function and gradient evaluations may be wasted due to "backtracking". In an alternative technique proposed by Mayne and Polak [9], the search direction is "bent" using a correction based on the value of the constraints at an auxiliary point, and an arc search is performed. Again, it is shown that a step of one is eventually accepted. The price paid here is that of an additional constraint evaluation at each iteration.

The contribution of this paper is to propose and analyze yet another scheme for avoiding the Maratos effect. The new scheme combines Mayne and Polak's technique with a nonmonotone line search used by Grippo, Lampariello and Lucidi in the context of unconstrained optimization in such a way that, asymptotically, function evaluations are no longer performed at auxiliary points.

The proposed method is based on the following observation. Let $\{x_k\}$ be a sequence generated by the basic SQP iteration for (P), i.e., $x_{k+1} = x_k + d_k$ where d_k solves the quadratic program

$$\begin{aligned} \min_d \ & \frac{1}{2}\langle d, H_k d\rangle + \langle \nabla f(x_k), d\rangle \\ \text{s.t. } & h_j(x_k) + \langle \nabla h_j(x_k), d\rangle = 0, \quad j = 1, \ldots, m_e \\ & g_j(x_k) + \langle \nabla g_j(x_k), d\rangle \leq 0, \quad j = 1, \ldots, m_i \ , \end{aligned}$$

with H_k an approximation of the Hessian of the Lagrangian of (P) in the subspace tangent to the active constraints. It can be shown that, under suitable assumptions, if x_0 is close

[2] This assumes that, when full steps of one are taken, convergence is Q-superlinear. The watchdog technique could possibly be adapted to the case of mere two-step superlinear convergence (this is often all one can insure, see [18]) at the expense of possible much more extensive backtracking.

to a strong local minimizer x^* for (P), the exact penalty function w used in [6], defined by

$$w(x) = f(x) + r\sum_{j=1}^{m_e} |h_j(x)| + r\sum_{j=1}^{m_i} \max\{g_j(x), 0\}$$

for some suitable $r > 0$, eventually satisfies the condition

$$w(x_{k+1}) \leq w(x_{k-3}) + \alpha\{\hat{w}(x_k; x_{k+1} - x_k) - w(x_k)\} \tag{1.1}$$

where α is any fixed number in $(0,1)$ and where $\hat{w}(x;d)$ is obtained by replacing in the expression of $w(x+d)$ the functions f, h_j, $j = 1, \ldots, m_e$ and g_j, $j = 1, \ldots, m_i$ by their first order approximation about x (this will be proved in Theorem 3.8 below). Consider now instead the iteration $x_{k+1} = x_k + t_k d_k$ where t_k is computed so as to satisfy[3]

$$w(x_k + t_k d_k) \leq W_k + \alpha\{\hat{w}(x_k; t_k d_k) - w(x_k)\}, \tag{1.2}$$

with

$$W_k = \max_{\ell=0,\ldots,3} w(x_{k-\ell}),$$

A "nonmonotone" line search, based on this type of criterion, has recently been proposed and analyzed in the unconstrained case by Grippo, Lampariello and Lucidi [5] who proved global convergence of the resulting algorithm. In view of (1.1), line search criterion (1.2) will always accept the step $t_k = 1$, close to x^*, *provided* $t_k = 1$ *has just been taken three times in a row.* This idea is taken up in this paper. To ensure that three consecutive steps of one will eventually be taken, we make use of Mayne and Polak's correction and arc search whenever (1.2) does not accept the step of one. Global convergence and local two-step superlinear convergence of the resulting algorithm are proven. Compared to the technique used in [9], the proposed approach has the advantage of not requiring evaluation of w at auxiliary points except in the early iterations. On the other hand, in contrast to the watchdog technique, it does not resort to any backtracking. In a companion paper[2], it is shown that a refinement of the scheme described in this paper can be used in the context of recently proposed SQP-based methods generating feasible iterates [10], [11]. In this case an additional challenge is that of achieving feasibility of the successive iterates, asymptotically with a full step of one, without resorting to evaluation of the constraints at auxiliary points.

The balance of the paper is organized as follows. The algorithm is presented in Section 2. To better highlight the main issues, the convergence analysis carried out in Section 3 is limited to the case when only equality constraints are present. Some examples are discussed in Section 4. Finally, Section 5 is devoted to concluding remarks.

---

[3] Note that a line search test similar to (1.2) but with W_k replaced by $w(x_{k-3})$ would often be impossible to satisfy.

2. An algorithm.

A point x^* is said to be a *Karush-Kuhn-Tucker (KKT) point* for (P) if $h_j(x^*) = 0,\ j = 1,\dots,m_e,\ g_j(x^*) \le 0,\ j = 1,\dots,m_i$ and there exist some multipliers $\lambda_j^*,\ j = 1,\dots,m_e,\ \mu_j^*,\ j = 1,\dots,m_i$ with $\mu_j^* \ge 0$, such that

$$\nabla f(x^*) + \sum_{j=1}^{m_e} \lambda_j^* \nabla h_j(x^*) + \sum_{j=1}^{m_i} \mu_j^* \nabla g_j(x^*) = 0$$

and

$$\mu_j^* g_j(x^*) = 0, \quad j = 1,\dots,m_i\ .$$

We present below an algorithm for finding a KKT point for problem (P). For the computation of a search direction, it makes use of quadratic programs $QP(x,H)$ defined for $x \in \mathbb{R}^n$ and $H \in \mathbb{R}^{n\times n}$ symmetric positive definite by

$$\begin{aligned} \min_d \quad & \frac{1}{2}\langle d, Hd\rangle + \langle \nabla f(x), d\rangle \\ \text{s.t.} \quad & h_j(x) + \langle \nabla h_j(x), d\rangle = 0, \quad j = 1,\dots,m_e \\ & g_j(x) + \langle \nabla g_j(x), d\rangle \le 0, \quad j = 1,\dots,m_i. \end{aligned} \tag{2.1}$$

For simplicity, it is assumed in this paper that the feasible sets of the problems $QP(x,H)$ encountered by the algorithm are always nonempty.[4] In the early iterations, an arc search is performed based on a correction $\tilde{d}$ obtained by solving the quadratic program $\widetilde{QP}\,(x,y,H)$ defined for $x \in \mathbb{R}^n$, $y \in \mathbb{R}^n$ and $H \in \mathbb{R}^{n\times n}$ symmetric positive definite by

$$\begin{aligned} \min_d \quad & \frac{1}{2}\langle y+d, H(y+d)\rangle + \langle \nabla f(x), y+d\rangle \\ \text{s.t.} \quad & h_j(x+y) + \langle \nabla h_j(x), d\rangle = 0, \quad j = 1,\dots,m_e \\ & g_j(x+y) + \langle \nabla g_j(x), d\rangle \le 0, \quad j = 1,\dots,m_i. \end{aligned} \tag{2.2}$$

A nondifferentiable penalty function $w : \mathbb{R}^n \to \mathbb{R}$ defined by

$$w(x) = f(x) + r\sum_{j=1}^{m_e} |h_j(x)| + r\sum_{j=1}^{m_i} \max\{g_j(x), 0\}\ ,$$

where r is a positive scalar, is used in the line search. In order to have a consistent line search, it is assumed that, at each iteration k, $r \ge |\lambda_{k,j}|,\ j = 1,\dots,m_e$ and $r \ge \mu_{k,j},\ \forall j = 1,\dots,m_i$, for multiplier vectors λ_k and μ_k associated with the constraints in $QP(x_k, H_k)$, where x_k is the current iterate and H_k the current estimate of the Hessian of

[4] Mechanisms for dealing with the case where (2.1) has inconsistent constraints have been proposed by several authors, see, e.g., [4], [12], [15].

the Lagrangian.[5] The line search test involves a maximum past value W_k defined over the last four iterates by

$$W_k = \max_{\ell=0,\dots,3} w(x_{k-\ell})$$

where negative indices that may appear in the early iterations are discarded, as well as a function $\hat{w}(\cdot;\cdot)$ defined for x and $d \in \mathbb{R}^n$ by

$$\hat{w}(x;d) = f(x) + \langle \nabla f(x), d\rangle + r\sum_{j=1}^{m_e} |h_j(x) + \langle \nabla h_j(x), d\rangle| + r\sum_{j=1}^{m_i} \max\{g_j(x) + \langle \nabla g_j(x), d\rangle, 0\}.$$

Algorithm 2.1.

Parameters. $\alpha \in (0, \frac{1}{2})$, $\beta \in (0,1)$.

Data. $x_0 \in X$, $H_0 \in \mathbb{R}^{n\times n}$, symmetric positive definite.

Step 0. Initialization. Set $k = 0$.

Step 1. Computation of a new iterate.

i. Compute d_k solution of the quadratic program $QP(x_k, H_k)$

ii. If $d_k = 0$ stop.

iii. If

$$w(x_k + d_k) \le W_k + \alpha\{\hat{w}(x_k; d_k) - w(x_k)\}, \tag{2.3}$$

set $t_k = 1$ and $x_{k+1} = x_k + d_k$.
Otherwise,

- compute $\tilde{d}_k$ by solving the quadratic program $\widetilde{QP}(x_k, d_k, H_k)$. If there is no solution or if $||\tilde{d}_k|| > ||d_k||$, set $\tilde{d}_k = 0$.
- Compute t_k, the first number t in the sequence $\{1, \beta, \beta^2, \cdots\}$ satisfying

$$w(x_k + td_k + t^2\tilde{d}_k) \le W_k + \alpha\{\hat{w}(x_k; td_k) - w(x_k)\} \tag{2.4}$$

and set $x_{k+1} = x_k + t_k d_k + t_k^2 \tilde{d}_k$.

Step 2. Updates. Compute a new symmetric positive definite approximation H_{k+1} to the Hessian of the Lagrangian. Increase k by 1. Go back to Step 1. □

Remarks

(i) Away from a solution of (P), the correction $\tilde{d}_k$ can always be taken as zero, resulting in some savings in function evaluations.

(ii) The same theoretical properties would be achieved if the Grippo-Lampariello-Lucidi line search (2.4) were replaced by, say, an Armijo search. Yet, in conjunction with the "full step" test (2.3) which is essential, it is natural to use the former.

---

[5] The question of identifying a suitable value of r is not addressed here, see, e.g., [9], [14].

3. Convergence analysis in the equality constrained case.

As indicated in the introduction, to better highlight the main issues, we consider here the case when only equality constraints are present. There is no conceptual difficulty in extending the results to the general case.

Thus let $m_i = 0$ and, for simplicity, denote by m the number of equality constraints, i.e., consider the problem

$$\min f(x) \text{ s.t. } h(x) = 0, \qquad (P')$$

with $h : \mathbb{R}^n \to \mathbb{R}^m$.

3.1. Global convergence.

In this section, global convergence of Algorithm 2.1 is proven under the following standard assumptions.

A1. The functions f, h_j, $j = 1, \ldots, m$ are continuously differentiable.
A2. The set $\{x \in \mathbb{R}^n \mid w(x) \leq w(x_0)\}$ is compact.
A3. For any $x \in \mathbb{R}^n$, $\frac{\partial h}{\partial x}(x)$ has full row rank.

We also assume that there exist $\sigma_1,\ \sigma_2 > 0$ such that

$$\sigma_1 \|x\|^2 \leq \langle x, H_k x \rangle \leq \sigma_2 \|x\|^2, \quad \forall x \in \mathbb{R}^n,\ \forall k \in \mathbb{N}. \tag{3.1}$$

Note that Assumption A3 implies that for any vector x and symmetric positive definite matrix H, the quadratic program $QP(x, H)$ has a nonempty feasible set, and hence a unique solution d and a unique multiplier vector $\lambda \in \mathbb{R}^m$ satisfying

$$\begin{aligned} &Hd + \nabla f(x) + \sum_{j=1}^{m} \lambda_j \nabla h_j(x) = 0 \\ &h_j(x) + \langle \nabla h_j(x), d \rangle = 0,\ j = 1, \ldots, m. \end{aligned} \tag{3.2}$$

Finally, we assume that, for every $k \in \mathbb{N}$, the multiplier vector λ_k associated with the solution of $QP(x_k, H_k)$ satisfies

$$|\lambda_{k,j}| \leq r, \quad j = 1, \ldots, m. \tag{3.3}$$

Proposition 3.1.

Let $\tilde{d}$ and x be some vectors in $\mathbb{R}^n$, and H be a symmetric positive definite matrix in $\mathbb{R}^{n \times n}$. Then (i) if $d = 0$ solves $QP(x, H)$, then x is a KKT point for (P'). Moreover if $d \neq 0$ solves $QP(x, H)$, if λ is the associated multiplier vector, then (ii)

$$\hat{w}(x; td) - w(x) \leq -t\{\langle d, Hd \rangle + r \sum_{j=1}^{m} |h_j(x)| - \sum_{j=1}^{m} \lambda_j h_j(x)\} \quad \forall t \in [0, 1],$$

and, if r satisfies $|\lambda_j| \leq r,\ j = 1, \ldots, m$, then (iii)

$$\frac{w(x + td + t^2 \tilde{d}) - w(x)}{\hat{w}(x; td) - w(x)} \to 1, \text{ as } t \to 0,\ t > 0.$$

Proof.

The first claim follows directly from the optimality conditions (3.2) associated with $QP(x, H)$. Next, as a function of t, $\hat{w}(x; td)$ is convex. Thus, for $t \in [0, 1]$,

$$\hat{w}(x; td) - w(x) \leq t\{\hat{w}(x; d) - w(x)\}.$$

Using the optimality conditions (3.2) we then obtain

$$\begin{aligned}
\hat{w}(x; d) - w(x) &= \langle \nabla f(x), d \rangle + r \sum_{j=1}^{m} (|h_j(x) + \langle \nabla h_j(x), d \rangle| - |h_j(x)|) \\
&= -\langle d, Hd \rangle - \sum_{j=1}^{m} \lambda_j \langle \nabla h_j(x), d \rangle - r \sum_{j=1}^{m} |h_j(x)| \\
&= - \left\{ \langle d, Hd \rangle + r \sum_{j=1}^{m} |h_j(x)| - \sum_{j=1}^{m} \lambda_j h_j(x) \right\}.
\end{aligned}$$

thus proving (ii). That (iii) holds follows directly from the definition of $\hat{w}$ and the fact that under the current assumptions, in view of (ii), $\hat{w}(x; td) - w(x) \neq 0 \ \forall t > 0$. □

It follows readily from Proposition 3.1 (ii) and (3.3) that, whenever $d_k \neq 0$, the directional derivative of w at x_k in direction d_k is negative. Proposition 3.1 (iii) then implies that (2.4) holds for all t small enough so that the line search in Step 1 *iii*, whenever it is performed, is well defined. Thus Algorithm 2.1 is well defined. In view of Proposition 3.1 (i), if the algorithm stops at Step 1 *ii*, the last iterate x_k is a KKT point of (P'). From now on, we will assume that stop at Step 1 *ii* never occurs so that an infinite sequence $\{x_k\}$ is generated.

The following property, which holds true even though monotone decrease is not enforced by the line search rule, is a key to global convergence.

Proposition 3.2.

The sequence $\{x_k\}$ is bounded and the sequences $\{t_k d_k\}$ and $\{\|x_{k+1} - x_k\|\}$ both converge to zero.

Proof.

This statement can be proven similarly to what is done in the first part of the proof of the Theorem in [5], provided one first shows that, if $\hat{w}(x_k; t_k d_k) - w(x_k)$ converges to zero on a subsequence, then, on that same subsequence, $t_k d_k$ and $\|x_{k+1} - x_k\|$ also converge to zero. Thus, to complete the proof, assume that $\hat{w}(x_k; t_k d_k) - w(x_k)$ goes to zero on a subsequence. Then, on the same subsequence, the following holds. First, in view of Proposition 3.1 (ii) and of (3.1) and (3.3), $t_k \|d_k\|^2$ goes to zero. Next, boundedness of t_k yields that $t_k d_k$ goes to zero. Finally the fact that, whenever it is defined, $\tilde{d}_k$ satisfies $\|\tilde{d}_k\| \leq \|d_k\|$, implies that $\|x_{k+1} - x_k\|$ converges to zero. □

Theorem 3.3.

Let x^* be an accumulation point of the sequence generated by the algorithm and $\{x_k\}_{k\in K}$ be any subsequence converging to x^*. Then, x^* is a KKT point of (P) and the subsequence $\{d_k\}_{k\in K}$ converges to zero.

Proof.

In view of (3.1), we may assume, without loss of generality that the subsequence $\{H_k\}_{k\in K}$ converges to some symmetric positive definite matrix H^*. In view of Assumption A3, $QP(x^*, H^*)$ has a unique solution d^* and $\{d_k\}_{k\in K}$ goes to d^*. If $d^* = 0$ the claim follows from Proposition 3.1 (i). Proceeding by contradiction, suppose now that $d^* \neq 0$. In view of (3.3) and of the fact that, by construction $\|\tilde{d}_k\| < \|d_k\|$ for all k, there is no loss of generality in assuming that $\{\lambda_k\}_{k\in K}$ converges to the unique multiplier λ^* associated with $QP(x^*, H^*)$ which thus satisfies $|\lambda_j^*| \le r$, $j = 1, \ldots, m$, and that $\{\tilde{d}_k\}_{k\in K}$ converges to some $\tilde{d}^*$. In view of Proposition 3.1 $(ii) - (iii)$ it follows that for some $\underline{t} > 0, t_k > \underline{t}$ for all $k \in K$ large enough. Proposition 3.2 then implies that $\{d_k\}_{k\in K}$ goes to zero, a contradiction. □

3.2. Superlinear convergence.

In order to prove superlinear convergence, we assume some more regularity on the functions involved. Assumption A1 is replaced by

A1′. The functions f, h_j, $j = 1, \ldots, m$ are three times continuously differentiable.

Let x^* be an accumulation point of the sequence generated by the algorithm (known to exist in view of Proposition 3.2). In view of Theorem 3.3, x^* is a KKT point for (P'). We denote by λ^* the optimal multipliers vector at x^* and we suppose that

$$|\lambda_j^*| < r, \quad j = 1, \ldots, m$$

(strict inequality). Finally, we assume that the *second order sufficiency conditions* are satisfied at x^*, i.e., that $\nabla_{xx} L(x^*, \lambda^*)$ is positive definite on the subspace

$$\{p | \langle \nabla h_j(x^*), p\rangle = 0, \quad j = 1, \ldots, m\} ,$$

where $L(x, \lambda)$ denotes the Lagrangian function

$$L(x, \lambda) = f(x) + \sum_{j=1}^{m} \lambda_j h_j(x) .$$

Proposition 3.4.

The entire sequence $\{x_k\}$ converges to x^*.

Proof.

From the assumptions on x^*, there exists a neighborhood $N(x^*)$ of x^* such that $\forall x \in N(x^*)$, there exists no KKT point other than x^* in $N(x^*)$ [16]. Since d_k becomes arbitrarily small close to x^* (Theorem 3.3), the entire sequence $\{x_k\}$ converges to x^*. □

It follows from (3.2) that the sequence of multipliers $\{\lambda_k\}$ converges to λ^*.

Assume now that the approximations H_k to the Hessian of the Lagrangian at x^* satisfy

$$\frac{\|P_k(H_k - \nabla^2_{xx}L(x^*,\lambda^*))P_k d_k\|}{\|d_k\|} \to 0 \quad \text{as } k \to \infty$$

where the matrices P_k are defined by

$$P_k = I - \frac{\partial h}{\partial x}(x_k)^T \left(\frac{\partial h}{\partial x}(x_k)\frac{\partial h}{\partial x}(x_k)^T\right)^{-1} \frac{\partial h}{\partial x}(x_k)\,.$$

(Note that, in view of Assumption A3, $\frac{\partial h}{\partial x}(x_k)\frac{\partial h}{\partial x}(x_k)^T$ is invertible.) We will show that a step of one is always accepted for k large enough and that two-step superlinear convergence occurs. The following result will be instrumental.

Lemma 3.5.

Direction d_k computed in Step 1 i of Algorithm 2.1 can be decomposed into $d_k = P_k d_k + \hat{d}_k$ with

$$\|\hat{d}_k\| = O(\|h(x_k)\|)$$

Proof.

Optimality condition (3.2) for $QP(x_k, H_k)$ yields

$$h(x_k) = -\frac{\partial h}{\partial x}(x_k)d_k$$

Thus

$$\begin{aligned}\hat{d}_k = (I - P_k)d_k &= \frac{\partial h}{\partial x}(x_k)^T \left(\frac{\partial h}{\partial x}(x_k)\frac{\partial h}{\partial x}(x_k)^T\right)^{-1} \frac{\partial h}{\partial x}(x_k)d_k \\ &= -\frac{\partial h}{\partial x}(x_k)^T \left(\frac{\partial h}{\partial x}(x_k)\frac{\partial h}{\partial x}(x_k)^T\right)^{-1} h(x_k)\end{aligned}$$

The result then follows from Assumption A3 and boundedness of $\{x_k\}$. □

In order for superlinear convergence to take place, it is necessary that the line search eventually accept the step of one whenever it is performed. The next proposition asserts that it is indeed the case here. The proof is inspired from that of Proposition 15 in [8].

Proposition 3.6.

(i) The unique solution $\tilde{d}_k$ of $\widetilde{QP}(x_k, d_k, H_k)$ (known to exist in view of Assumption A3) satisfies

$$\tilde{d}_k = O(\|d_k\|^2) \tag{3.4}$$

and (ii) for k large enough, a step of one is accepted whenever the line search is performed.

Proof.

The vector $D_k = d_k + \tilde{d}_k$ is solution of the quadratic program

$$\begin{aligned}&\min_D \quad \frac{1}{2}\langle D, H_k D\rangle + \langle \nabla f(x_k), D\rangle \\ &\text{s.t.} \quad h_j(x_k) + \langle \nabla h_j(x_k), D\rangle = h_j(x_k) + \langle \nabla h_j(x_k), d_k\rangle - h_j(x_k + d_k), \quad j = 1, \ldots, m.\end{aligned}$$

Using second order expansions about x_k of $h_j(x_k + d_k)$, $j = 1, \ldots, m$, one can see that D_k is solution a quadratic program similar to $QP(x_k, H_k)$ with right hand side perturbed by $O(\|d_k\|^2)$. It follows that $D_k = d_k + O(\|d_k\|^2)$, so that $\tilde{d}_k = O(\|d_k\|^2)$. Next, in view of the above, for k large enough, if a line search is performed the correction $\tilde{d}_k$ solution of $\widetilde{QP}(x_k, d_k, H_k)$ satisfies $\|\tilde{d}_k\| \leq \|d_k\|$ and thus is used in the line search. Proceeding as in the proof of Proposition 15 in [8], using (3.4), one obtains,

$$w(x_k + d_k + \tilde{d}_k) - w(x_k + d_k) - \alpha(\hat{w}(x_k; d_k) - w(x_k)) \leq \left\{\frac{1}{2} - \alpha\right\} \{\hat{w}(x_k; d_k) - w(x_k)\}$$

$$+\frac{1}{2}\langle d_k, \{\frac{\partial^2 L}{\partial x^2}(x_k, \lambda_k) - H_k\} d_k\rangle + O(\|d_k\|^3).$$

Decomposing d_k as in Lemma 3.5 and making use of Proposition 3.1 and of the fact that $|\lambda_j^*| < r$, $j = 1, \ldots, m$, we obtain

$$w(x_k + d_k + \tilde{d}_k) - w(x_k + d_k) - \alpha(\hat{w}(x_k; d_k) - w(x_k)) \leq$$

$$\left\{\alpha - \frac{1}{2}\right\} \langle d_k, H_k d_k\rangle + \frac{1}{2}\langle d_k, P_k\{\frac{\partial^2 L}{\partial x^2}(x_k, \lambda_k) - H_k\} P_k d_k\rangle + O(\|d_k\|^3).$$

In view of (3.1), of the fact that $\alpha < \frac{1}{2}$ and of the convergence of the projection of H_k on the subspace orthogonal to the gradients of the active constraints at x^* to the corresponding projection of the Hessian of the Lagrangian at x^*, for k large enough the right hand side of the last inequality is negative. □

Theorem 3.7.

Under the stated assumptions, the convergence is two-step superlinear, i.e.,

$$\lim_{k \to \infty} \frac{\|x_{k+2} - x^*\|}{\|x_k - x^*\|} = 0.$$

Moreover,

$$\|x_{k+1} - x^*\| = O(\|x_k - x^*\|) \, .$$

Proof.

This result can be proven via a slight modification of Theorem 1 in [13] or Theorem 4 in [17], taking into account the facts established in Proposition 3.6 that, a step of one is taken close to x^* and that whenever it is needed the correction $\tilde{d}_k$ satisfies $\tilde{d}_k = O(\|d_k\|^2)$ so that the good properties of the direction d_k are also enjoyed by $d_k + \tilde{d}_k$. □

Finally, we show that Algorithm 2.1 improves on the anti-Maratos effect scheme of [9] in that the constraints are evaluated at auxiliary points in the early iterations only.

Theorem 3.8.

For k large enough, correction $\tilde{d}_k$ is not computed.

Proof.

We show that, for k big enough, we always have

$$w(x_k + d_k) - \alpha(\hat{w}(x_k; d_k) - w(x_k)) \leq w(x_{k-3}). \tag{3.5}$$

From the definition of w and $\hat{w}$, since

$$h_j(x_k) + \langle \nabla h_j(x_k), d_k \rangle = 0, \; j = 1, \ldots, m \tag{3.6}$$

we have

$$w(x_k + d_k) - \alpha(\hat{w}(x_k; d_k) - w(x_k)) = f(x_k + d_k) + r \sum_{j=1}^{m} |h_j(x_k + d_k)|$$

$$-\alpha(\langle \nabla f(x_k), d_k \rangle - r \sum_{j=1}^{m} |h_j(x_k)|) .$$

From (3.6) again, we have

$$h_j(x_k + d_k) = O(\|d_k\|^2) \tag{3.7}$$

and, since for given k either $x_k = x_{k-1} + d_{k-1} + \tilde{d}_{k-1}$ or $x_k = x_{k-1} + d_{k-1}$ (Proposition 3.6),

$$h_j(x_k) = O(\|d_{k-1}\|^2). \tag{3.8}$$

Therefore,

$$w(x_k + d_k) - \alpha(\hat{w}(x_k; d_k) - w(x_k)) = f(x_k + d_k) - \alpha\langle \nabla f(x_k), d_k \rangle + O(\|d_{k-1}\|^2) + O(\|d_k\|^2). \tag{3.9}$$

Expanding $f(x_k + d_k)$ to first order about x^* yields

$$f(x_k + d_k) = f(x^*) + \langle \nabla f(x^*), x_k + d_k - x^* \rangle + O(\|x_k + d_k - x^*\|^2)$$

which gives, from the KKT conditions associated with x^*,

$$f(x_k + d_k) = f(x^*) - \sum_{j=1}^{m} \lambda_j^* \langle \nabla h_j(x^*), x_k + d_k - x^* \rangle + O(\|x_k + d_k - x^*\|^2). \tag{3.10}$$

Since $h(x^*) = 0$, expansion of h_j, $j = 1, \ldots, m$, to first order about x^* yields

$$\langle \nabla h_j(x^*), x_k + d_k - x^* \rangle = h_j(x_k + d_k) + O(\|x_k + d_k - x^*\|^2), \; j = 1, \ldots, m.$$

Substituting in (3.10) we obtain

$$f(x_k + d_k) = f(x^*) - \sum_{j=1}^{m} \lambda_j^* h_j(x_k + d_k) + O(\|x_k + d_k - x^*\|^2). \tag{3.11}$$

Now, since d_k satisfies the optimality conditions (3.2) associated with $QP(x_k, H_k)$, we have, in view of (3.1)

$$\langle \nabla f(x_k), d_k \rangle = \sum_{j=1}^{m} \lambda_{k,j} h_j(x_k) + O(\|d_k\|^2). \tag{3.12}$$

From (3.11) and (3.12), we have,

$$f(x_k + d_k) - \alpha \langle \nabla f(x_k), d_k \rangle = f(x^*) - \sum_{j=1}^{m} \lambda_j^* h_j(x_k + d_k) - \alpha \sum_{j=1}^{m} \lambda_{k,j} h_j(x_k)$$

$$+O(\|x_k + d_k - x^*\|^2) + O(\|d_k\|^2).$$

so that, using (3.7) and (3.8) and the boundedness of $\{\lambda_k\}$ (since $\{\lambda_k\}$ converges to λ^*), we get

$$f(x_k + d_k) - \alpha \langle \nabla f(x_k), d_k \rangle = f(x^*) + O(\|x_k + d_k - x^*\|^2) + O(\|d_{k-1}\|^2) + O(\|d_k\|^2).$$

Substituting in (3.9) and using Theorem 3.7, we obtain,

$$w(x_k + d_k) - \alpha(\hat{w}(x_k; d_k) - w(x_k)) \leq f(x^*) + o(\|x_{k-3} - x^*\|^2). \tag{3.13}$$

Finally it is shown in [3, Lemma 1] with assumptions equivalent to ours that, there exists a positive scalar C such that, for x close enough to x^*,

$$w(x) \geq f(x^*) + C\|x - x^*\|^2.$$

This, together with (3.13), implies (3.5). □

4. Some numerical examples.

To obtain a preliminary assessment of the practical value of the new algorithm, numerical tests were performed on two small size problems previously used in related literature. The first problem was considered in [3] to illustrate the fact that the Maratos effect can potentially be very damaging. The second one was produced in [18] as a case where mere two-step superlinear convergence occurs.

In all the numerical tests, we used the parameter values $\alpha = 0.3$, $\beta = 0.8$ and $r = 10$, the matrices H_k were updated according to Powell's modified BFGS formula [12] and execution was terminated when the condition $\|d_k^0\| \leq 10^{-6}$ was satisfied. The examples were run on a SUN$^{\text{TM}}$ 3/110.

Example 4.1 [3]

$$\begin{aligned} \min \quad & -x_1 + 10(x_1^2 + x_2^2 - 1) \\ \text{s.t.} \quad & x_1^2 + x_2^2 - 1 = 0, \end{aligned}$$

with solution $(1,0)^T$.

An analysis in [3] indicates that the Maratos effect can very seriously affect the performance of an SQP algorithm on this problem. Numerical results with VF02AD [1] reported by Powell [14] with initial points $(0.8, 0.6)^T$ confirm this diagnostic.

Algorithm 2.1 terminated after 5 iterations at the point $(1.000000,\ 1.550916 \cdot 10^{-7})^T$, with a total of 11 evaluations of the objective and 11 evaluations of the constraint. The $\tilde{d}$ correction was used only in the first two iterations. In this example both nonmonotone line search and initialization via the $\tilde{d}$ correction were instrumental in the good behavior of the algorithm. First while the full step of one was taken at all but the first iteration, this did require an increase of the merit function at the second iteration. Second, when we reran the test without making use of the $\tilde{d}$ correction, the Maratos effect was clearly visible: the total number of iterations increased significantly, though less dramatically than when in addition the nonmonotone line search was replaced by an Armijo-type line search.

Example 4.2 [18]

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & g(x) = 0 \end{aligned}$$

where

$$f(x) = \frac{1}{2}x_2^2 - x_1x_2 + \frac{1}{6(1-x_2)^3}\{-4(x_2-x_1)^3 - 6(x_2-x_1)^2(x_1-x_2^2)$$

$$-12(x_2-x_1)(x_1-x_2^2)^2 - 17(x_1-x_2^2)^3 + 3\frac{(x_1-x_2^2)^4}{1-x_2}\}$$

and

$$g(x) = x_1 + \frac{1}{(1-x_2^2)}\{(x_2-x_1)^2 + (x_2-x_1)(x_1-x_2^2) + 2(x_1-x_2^2)^2\},$$

with solution $(0,0)^T$.

On this problem, starting at the point $(0.01, 0.1)^T$, Algorithm 2.1 terminated after 5 iterations at the point $(1.012154 \cdot 10^{-8},\ -7.667375 \cdot 10^{-8})^T$. A full step of one was taken at each iteration and $\tilde{d}_k$ was never used. Here again, occasional increase of the merit function was observed indicating a beneficial effect of the nonmonotone line search.

From the initial point $(0.01, 0.5)^T$ Newton's method does not converge [18] and the stabilizing role of the line search becomes essential. Algorithm 2.1 converged in 14 iterations and 31 evaluations of both the objective and the constraint function to the point $(8.506928 \cdot 10^{-10},\ -4.578437 \cdot 10^{-8})^T$. A full step of one was taken at the last 9 iterations (again with occasional increase of the merit function). Correction $\tilde{d}_k$ was computed at iterations 1, 3 and 5. At iterations 3 and 5, however, the condition $\|\tilde{d}_k\| \leq \|d_k\|$ was violated, so that $\tilde{d}_k$ was not used and no additional function evaluation was performed.

5. Concluding remarks

The key ideas behind Algorithm 2.1 are that, in conjunction with a two-step superlinearly convergent SQP iteration, (i) a nonmonotone (more precisely, "four-step monotone") line search can prevent occurrence of the Maratos effect, provided the process is suitably initialized, and (ii) such initialization can be performed at the expense of a few additional function evaluations in the early iterations. A crucial result, uncovered by Grippo, Lampariello and Lucidi in the context of unconstrained minimization, is that global convergence is ensured despite occasional increase of the merit function. While, for the sake of clarity, the exposition has focused on a simple-minded SQP iteration, it should be clear that similar consideration apply to more sophisticated schemes.

Besides resulting in avoidance of the Maratos effect, the nonmonotone line search often speeds up convergence by allowing a full step of one to be taken early on. This aspect is stressed in [5] and is clearly apparent from our numerical tests.

Acknowledgements. The authors wish to thank Dr. J.F. Bonnans for his helpful comments.

References

[1] *Harwell Subroutine Library*, Library Reference Manual, Harwell, England, 1982.

[2] J. F. Bonnans, E. R. Panier and A. L. Tits, *Avoiding the Maratos Effect by Means of a Nonmonotone Line Search. II. Inequality Constrained Problems – Feasible Iterates*, Systems Research Center, University of Maryland, Technical Report SRC-TR-89-42, College Park, MD 20742, 1989.

[3] R. M. Chamberlain, M. J. D. Powell, C. Lemarechal and H. C. Pedersen, *The Watchdog Technique for Forcing Convergence in Algorithms for Constrained Optimization*, Math. Programming Stud., 16 (1982), pp. 1–17.

[4] R. Fletcher, *Numerical Experiments with an Exact L_1 Penalty Function Method*, in Nonlinear Programming 4, O. L. Mangasarian, R. R. Meyer and S. M. Robinson, eds., Academic Press, New York, 1981, pp. 99–129.

[5] L. Grippo, F. Lampariello and S. Lucidi, *A Nonmonotone Line Search Technique for Newton's Method*, SIAM J. Numer. Anal., 23 (1986), pp. 707–716.

[6] S. P. Han, *A Globally Convergent Method for Nonlinear Programming*, J. Optim. Theory Appl., 22 (1977), pp. 297–309.

[7] N. Maratos, *Exact Penalty Function Algorithms for Finite Dimensional and Optimization Problems*, Ph.D. Thesis, Imperial College of Science and Technology, London, U.K., 1978.

[8] D. Q. Mayne and E. Polak, *A Superlinearly Convergent Algorithm for Constrained Optimization Problems*, Imperial College of Science and Technology, Computing and Control Publication 78/52, London, 1978.

[9] ———, *A Superlinearly Convergent Algorithm for Constrained Optimization Problems*, Math. Programming Stud., 16 (1982), pp. 45–61.

[10] E. R. Panier and A. L. Tits, *A Superlinearly Convergent Feasible Method for the Solution of Inequality Constrained Optimization Problems*, SIAM J. Control Optim., 25 (1987), pp. 934–950.

[11] ———, *On Feasibility, Descent and Superlinear Convergence in Inequality Constrained Optimization*, Systems Research Center, University of Maryland, Technical Report SRC-TR-89-27, College Park, MD 20742, 1989.

[12] M. J. D. Powell, *A Fast Algorithm for Nonlinearly Constrained Optimization Calculations*, in Numerical Analysis, Dundee, 1977, Lecture Notes in Mathematics 630, G. A. Watson, ed., Springer-Verlag, 1978, pp. 144–157.

[13] ———, *The Convergence of Variable Metric Methods for Nonlinearly Constrained Optimization Calculations*, in Nonlinear Programming 3, O. L. Mangasarian, R. R. Meyer and S. M. Robinson, eds., Academic Press, New York, 1978, pp. 27–63.

[14] ———, *Extensions to subroutine VF02AD*, in System Modeling and Optimization, R. F. Drenick and F. Kozin, eds., Lecture Notes in Control and Information Sciences, 38, Springer-Verlag, New York–Heidelberg–Berlin, 1982, pp. 529–538.

[15] ———, *Variable Metric Methods for Constrained Optimization*, in Mathematical Programming, The State of the Art, Bonn 1982, A. Bachem, M. Grötschel and B. Korte, eds., Springer-Verlag, New York–Heidelberg–Berlin, 1983, pp. 288–311.

[16] S. M. Robinson, *Perturbed Kuhn-Tucker Points and Rates of Convergence for a Class of Nonlinear-Programming Algorithms*, Mathematical Programming, 7 (1974), pp. 1–16.

[17] J. Stoer, *The Convergence of Sequential Quadratic Programming Methods for Solving Nonlinear Programs*, in Recent Advances in Communication and Control Theory, R. E. Kalman, G. I. Marchuk, A. E. Ruberti and A. J. Viterbi, eds., Optimization Software, Inc., New York, N.Y., 1987, pp. 412–421.

[18] Y. X. Yuan, *An Only 2-Step Q-Superlinear Convergence Example for Some Algorithms that use Reduced Hessian Approximations*, Math. Programming, 32 (1985), pp. 224–231.