Convergence Properties of an Augmented Lagrangian Algorithm for Optimization with a Combination of General Equality and Linear Constraints

by A.R. Conn[1], Nick Gould[2],
A. Sartenaer[3] and Ph.L. Toint[3]

September 11, 1995

[1] IBM T.J. Watson Research Center, P.O.Box 218, Yorktown Heights, NY 10598, USA
Email : arconn@watson.ibm.com

[2] Rutherford Appleton Laboratory, Chilton, Oxfordshire, England
Email : nimg@directory.rl.ac.uk
Current reports available by anonymous ftp from the directory
"pub/reports" on camelot.cc.rl.ac.uk (internet 130.246.8.61)

[3] Department of Mathematics, Facultés Universitaires ND de la Paix, 61, rue de Bruxelles,
B-5000 Namur, Belgium, EC
Email : as@math.fundp.ac.be and pht@math.fundp.ac.be
Current reports available by anonymous ftp from the directory
"pub/reports" on thales.math.fundp.ac.be (internet 138.48.4.14)

Keywords: Constrained optimization, augmented Lagrangian methods,
linear constraints, convergence theory.

Convergence Properties of an Augmented Lagrangian Algorithm for Optimization with a Combination of General Equality and Linear Constraints

A. R. Conn, Nick Gould, A. Sartenaer and Ph.L. Toint

September 11, 1995

Abstract

We consider the global and local convergence properties of a class of augmented Lagrangian methods for solving nonlinear programming problems. In these methods, linear and more general constraints are handled in different ways. The general constraints are combined with the objective function in an augmented Lagrangian. The iteration consists of solving a sequence of subproblems; in each subproblem the augmented Lagrangian is approximately minimized in the region defined by the linear constraints. A subproblem is terminated as soon as a stopping condition is satisfied. The stopping rules that we consider here encompass practical tests used in several existing packages for linearly constrained optimization. Our algorithm also allows different penalty parameters to be associated with disjoint subsets of the general constraints. In this paper, we analyze the convergence of the sequence of iterates generated by such an algorithm and prove global and fast linear convergence as well as showing that potentially troublesome penalty parameters remain bounded away from zero.

1 Introduction

In this paper, we consider the problem of calculating a local minimizer of the smooth function

$$f(x), \tag{1.1}$$

where x is required to satisfy the general equality constraints

$$c_i(x) = 0, \quad 1 \le i \le m \tag{1.2}$$

and the linear inequality constraints

$$Ax - b \ge 0. \tag{1.3}$$

Here f and c_i map $\Re^n$ into $\Re$, A is a p-by-n matrix and $b \in \Re^p$.

[0]This research was supported in part by the Advanced Research Projects Agency of the Department of Defense and was monitored by the Air Force Office of Scientific Research under Contract No F49620-91-C-0079. The United States Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon.

A classical technique for solving problem (**??**)–(**??**) is to minimize a suitable sequence of *augmented Lagrangian functions.* If we only consider the problem (**??**)–(**??**), these functions are defined by

$$\Phi(x,\lambda,\mu) = f(x) + \sum_{i=1}^{m} \lambda_i c_i(x) + \frac{1}{2\mu}\sum_{i=1}^{m} c_i(x)^2 \tag{1.4}$$

where the components λ_i of the vector λ are known as *Lagrange multiplier estimates* and μ is known as the *penalty parameter* (see, for instance, **?**, **?**, **?**, **?** and **?**, **?**). The question then arises how to deal with the additional linear inequality constraints (**??**). The case where A is the identity matrix (that is when (**??**) specifies bounds on the variables) has been considered by **?** (**?**) and **?** (**?**). They propose keeping these constraints explicitly outside the augmented Lagrangian formulation, handling them directly at the level of the augmented Lagrangian minimization. That is, a sequence of optimization problems, in which (**??**) is approximately minimized within the region defined by the simple bounds, is attempted. In this approach all linear inequalities other than bound constraints are converted to equations by introducing slack variables and incorporated in the augmented Lagrangian function. This strategy has been implemented and successfully applied within the LANCELOT package for large-scale nonlinear optimization (see **?**, **?**). However, such a method may be inefficient when linear constraints are present as there are a number of effective techniques specifically designed to handle such constraints directly (see **?**, **?**, **?**, **?**, **?**, **?**, or **?**, **?**, for instance). This is especially important for large-scale problems. The purpose of the present paper is therefore to define and analyze an algorithm where the constraints (**??**) are kept outside the augmented Lagrangian and handled at the level of the subproblem minimization, thus allowing the use of specialized packages to solve the subproblem.

Our proposal extends the method of **?** (**?**) in that not only bounds but general linear inequalities are treated separately. **?** (**?**, page 295) remarks on the potential advantages of this strategy.

Furthermore, it is often worthwhile from the practical point of view to associate different penalty parameters to subsets of the general constraints (**??**) to reflect different degrees of nonlinearity. This possibility has been considered by many authors, including **?** (**?**, page 292), **?**, **?** and **?** (**?**, page 124). In this case, the formulation of the augmented Lagrangian (**??**) can be refined: we partition the set of constraints (**??**) into q disjoint *subsets* $\{\mathcal{Q}_j\}_{j=1}^q$, and redefine the augmented Lagrangian as

$$\Phi(x,\lambda,\mu) = f(x) + \sum_{j=1}^{q}\sum_{i\in\mathcal{Q}_j}\left[\lambda_i c_i(x) + \frac{1}{2\mu_j}c_i(x)^2\right], \tag{1.5}$$

where μ is now a q-dimensional vector, whose j-th component is $\mu_j > 0$, the penalty parameter associated with subset $\mathcal{Q}_j$. Because of its potential usefulness, and because its analysis is difficult to directly infer from the single penalty parameter case, this refined formulation will be adopted in the present paper.

The theory presented below handles the linear inequality constraints in a purely geometric way. Hence the same theory applies without modifications if linear equality constraints are also imposed and all the iterates are assumed to stay feasible with respect to these new constraints.

It is indeed enough to apply the theory in the affine subspace corresponding to this feasible set. As a consequence, linear constraints need not be included in the augmented Lagrangian and thus have the desirable property that they have no impact on the structure of its Hessian matrix.

The paper is organized as follows. In Section **??**, we introduce our basic assumptions on the problem and the necessary terminology. Section **??** presents the proposed algorithm and the definition of a suitable stopping criterion for the subproblem. The global convergence analysis is developed in Section **??** while the rate of convergence is analyzed in Section **??**. Second order conditions are investigated in Section **??**. Section **??** considers some possible extensions of the theory. Finally, some conclusions and perspectives are outlined in Section **??**.

2 The problem and related terminology

We consider the problem stated in (**??**)–(**??**) and make the following assumptions.

AS1: The region $\mathcal{B} = \{x \mid Ax - b \geq 0\}$ is nonempty.

AS2: The functions $f(x)$ and $c_i(x)$, $(i = 1, \ldots, m)$, are twice continuously differentiable for all $x \in \mathcal{B}$.

Assumption AS1 is clearly necessary for the problem to make sense. We note that it does not prevent $\mathcal{B}$ from being unbounded.

We now introduce the notation that will be used throughout the paper.

Let $g(x)$ denote the gradient $\nabla_x f(x)$ of $f(x)$ and $H(x)$ denote its Hessian matrix $\nabla_{xx} f(x)$. We also define $J(x)$ to be the m-by-n Jacobian of $c(x)$, where

$$c(x) = [c_1(x), \cdots, c_m(x)]^T.$$

Hence

$$J(x)^T = [\nabla_x c_1(x), \ldots, \nabla_x c_m(x)].$$

Let $H_i(x)$ denote the Hessian matrix $\nabla_{xx} c_i(x)$ of $c_i(x)$. Finally, let $g^\ell(x, \lambda)$ and $H^\ell(x, \lambda)$ denote the gradient, $\nabla_x \ell(x, \lambda)$, and the Hessian matrix, $\nabla_{xx} \ell(x, \lambda)$, of the Lagrangian function

$$\ell(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i c_i(x).$$

We note that $\ell(x, \lambda)$ is the Lagrangian solely with respect to the c_i constraints. If we define *first-order Lagrange multiplier estimates* componentwise as

$$\bar{\lambda}(x, \lambda_{[\mathcal{Q}_j]}, \mu_j)_{[\mathcal{Q}_j]} = \lambda_{[\mathcal{Q}_j]} + c(x)_{[\mathcal{Q}_j]}/\mu_j \quad (j = 1, \ldots, q), \tag{2.1}$$

where $w_{[\mathcal{S}]}$ denotes the $|\mathcal{S}|$-dimensional subvector of w whose entries are indexed by the set $\mathcal{S}$, we shall use the identity

$$\begin{aligned} \nabla_x \Phi(x, \lambda, \mu) = & \ \nabla_x f(x) + \textstyle\sum_{j=1}^{q} \sum_{i \in \mathcal{Q}_j} \left[\lambda_i \nabla_x c_i(x) + \frac{1}{\mu_j} c_i(x) \nabla_x c_i(x)\right] \\ = & \ g^\ell(x, \bar{\lambda}(x, \lambda, \mu)). \end{aligned} \tag{2.2}$$

Now suppose that $\{x_k \in \mathcal{B}\}$, $\{\lambda_k\}$ and $\{\mu_k\}$ are infinite sequences of n-vectors, m-vectors and positive q-vectors, respectively. For any function F, we shall use the notation that F_k denotes F evaluated with arguments x_k, λ_k and/or μ_k as appropriate. So, for instance, using the identity (**??**), we have that

$$\nabla_x \Phi_k = \nabla_x \Phi(x_k, \lambda_k, \mu_k) = g^\ell(x_k, \bar{\lambda}_k), \tag{2.3}$$

where we have written (**??**) in the compact form

$$\bar{\lambda}_k = \bar{\lambda}(x_k, \lambda_k, \mu_k). \tag{2.4}$$

We denote the vector w at iteration k by w_k and its i-th component by $w_{k,i}$. We also use $w_{k,[\mathcal{S}]}$ to denote the $|\mathcal{S}|$-dimensional subvector of w_k whose entries are indexed by $\mathcal{S}$.

Now let $\{x_k\}, k \in \mathcal{K}$, for some subset $\mathcal{K}$ of the natural numbers $\mathbf{N}$, be a convergent subsequence with limit point x_*. Then we denote the matrix whose rows are those of A corresponding to active constraints at x_* — that is the constraints which are satisfied as equalities at x_* — by A_*. Furthermore, we choose Z_* to be a matrix whose columns form an orthonormal basis of the null space of A_*, that is

$$A_* Z_* = 0 \ \text{ and } \ Z_*^T Z_* = I.$$

We define the *least-squares Lagrange multiplier estimates* (corresponding to A_*)

$$\lambda(x) \stackrel{\text{def}}{=} -((J(x)Z_*)^+)^T Z_*^T g(x) \tag{2.5}$$

at all points where the right generalized inverse

$$(J(x)Z_*)^+ \stackrel{\text{def}}{=} Z_*^T J(x)^T (J(x)Z_* Z_*^T J(x)^T)^{-1}$$

of $J(x)Z_*$ is well defined. We note that, whenever $J(x)Z_*$ has full rank, $\lambda(x)$ is differentiable and its derivative is given in the following lemma

Lemma 2.1 *Suppose that AS2 holds. If $J(x)Z_*Z_*^T J(x)^T$ is nonsingular, $\lambda(x)$ is differentiable and its derivative is given by*

$$\nabla_x \lambda(x) = -((J(x)Z_*)^+)^T Z_*^T H^\ell(x, \lambda(x)) - (J(x)Z_*Z_*^T J(x)^T)^{-1} R(x) \tag{2.6}$$

where the i-th row of $R(x)$ is $(Z_*^T g(x) + Z_*^T J(x)^T \lambda(x))^T Z_*^T H_i(x)$.

Proof. The result follows by observing that (**??**) may be rewritten as

$$r(x) - Z_*^T J(x)^T \lambda(x) = Z_*^T g(x) \ \text{ and } \ J(x)Z_* r(x) = 0 \tag{2.7}$$

for some vector $r(x)$. Differentiating (**??**) and eliminating the derivative of $r(x)$ from the resulting equations gives the required result. □

We stress that, as stated, the Lagrange multiplier estimate (**??**) is not directly calculable as it requires a priori knowledge of x_*. It is merely introduced as an analytical device.

Finally, the symbol $\|\cdot\|$ will denote the ℓ_2-norm or the induced matrix norm. We are now in position to describe more precisely the algorithm that we propose to use.

3 Statement of the algorithm

We consider the algorithmic model we wish to use in order to solve the problem (**??**)–(**??**). This model proceeds at iteration k by computing an iterate x_k which satisfies (**??**) and approximately solves the subproblem

$$\min_{x \in \mathcal{B}} \Phi(x, \lambda_k, \mu_k), \tag{3.1}$$

where the values of the Lagrange multipliers λ_k and penalty parameters μ_k are fixed for the subproblem. Subsequently we update the Lagrange multipliers and/or decrease the penalty parameters, depending on how much the constraint violation for (**??**) has been reduced within each subset of the constraints. The motivation is simply to ensure global convergence by driving, in the worst case, the penalty parameters to zero, in which case the algorithms essentially reduce to the quadratic penalty function method (see, for example, **?**, **?**). The tests on the size of the general constraint violation are designed to allow the multiplier updates to take over in the neighbourhood of a stationary point.

The approximate minimization for problem (**??**) is performed in an *inner iteration* which is stopped as soon as its current iterate is "sufficiently critical". We propose to base this decision on the identification of the linear constraints that are "dominant" at x (even though they might not be active) and on a measure of criticality for the part of the problem where those constraints are irrelevant. Given $\omega \geq 0$, a criticality tolerance for the subproblem, we define, for a vector $x \in \mathcal{B}$, the set of *dominant constraints at* x as the constraints whose indices are in the set

$$D(x, \omega) \stackrel{\text{def}}{=} \{i \in \{1, \ldots, p\} \mid a_i^T x - b_i \leq \kappa_0 \omega\}, \tag{3.2}$$

for some $\kappa_0 > 0$. Here $a_i^T \in \Re^n$ is the i-th row of the matrix A and b_i is the corresponding component of the right-hand side vector b. Denoting by $A_{D(x,\omega)}$ the submatrix of A consisting of the row(s) whose index is in $D(x, \omega)$, we also define

$$N(x, \omega) = \{A_{D(x,\omega)}^T \xi \mid \xi \in \Re^{|D(x,\omega)|} \text{ and } \xi_i \leq 0, \quad (i = 1, \ldots, |D(x, \omega)|)\},$$

the cone spanned by the outwards normals of the dominant constraints. The associated polar cone is then

$$T(x, \omega) = N(x, \omega)^0 = \text{cl}\{d \mid d^T v \leq 0 \text{ for all } v \in N(x, \omega)\},$$

where $\text{cl}(V)$ denotes the closure of the set V. The cone $T(x, \omega)$ is the tangent cone with respect to the dominant constraints at x for the tolerance ω. Note that $D(x, \omega)$ might be empty, in which case $A_{D(x,\omega)}$ is assumed to be zero, $N(x, \omega)$ reduces to the origin and $T(x, \omega)$ is the full space.

We then formulate our "sufficient criticality" criterion for the subproblem as follows: we require that

$$\|P_{T(x_k,\omega_k)}(-\nabla_x \Phi_k)\| \leq \omega_k, \tag{3.3}$$

where $P_V(\cdot)$ is the projection onto the convex set V and ω_k is a suitable tolerance at iteration k. Once x_k satisfying (**??**) has been determined by the inner iteration, we denote

$$D_k = D(x_k, \omega_k), \quad N_k = N(x_k, \omega_k) \text{ and } T_k = T(x_k, \omega_k). \tag{3.4}$$

For future reference, we define Z_k to be a matrix whose columns form an orthonormal basis of $\mathcal{V}_k$, the null space of A_{D_k}, and Y_k to be a matrix whose columns form an orthonormal basis of $\mathcal{W}_k = \mathcal{V}_k^{\perp}$. As above, we have that T_k is the full space and N_k reduces to the origin when D_k is empty. We note that, in this case, $Z_k = P_{T_k} = I$, the identity operator, and $Y_k = P_{N_k} = 0$. We also note that $\mathcal{V}_k \subseteq T_k$, and hence that

$$\|Z_k^T \nabla_x \Phi_k\| = \|Z_k Z_k^T \nabla_x \Phi_k\| \leq \|P_{T_k}(-\nabla_x \Phi_k)\|, \tag{3.5}$$

since $Z_k Z_k^T$ is the orthogonal projection onto $\mathcal{V}_k$.

It is important to note that the stopping rule (**??**) covers a number of more specific choices, including the rule used in much existing software for linearly constrained optimization (such as MINOS by **?**, **?**, LSNNO by **?**, **?**, or `VE14` and `VE19` from the Harwell Subroutine Library). The reader is referred to Section **??** for further details.

We are now in position to describe our algorithmic model more precisely. In this model, we define α_k to be the maximum penalty parameter at iteration k (see (**??**)). At this iteration, the parameters ω_k and η_k represent criticality and feasibility levels, respectively.

Algorithm 3.1

Step 0 [Initialization]. A partition of the set $\{1, \ldots, m\}$ into q disjoint subsets $\{\mathcal{Q}_j\}_{j=1}^q$ is given, as well as initial vectors of Lagrange multiplier estimates λ_0 and positive penalty parameters μ_0 such that

$$\mu_{0,j} < 1, \quad (j = 1, \ldots, q). \tag{3.6}$$

The strictly positive constants κ_0, $\omega_* \ll 1$, $\eta_* \ll 1$, $\tau < 1$, $\alpha_\eta < 1$, and $\beta_\eta < 1$ are specified. Set $\alpha_0 = \max_{j=1,\ldots,q} \mu_{0,j}$, $\omega_0 = \alpha_0$, $\eta_0 = \alpha_0^{\alpha_\eta}$ and $k = 0$.

Step 1 [Inner iteration]. Find $x_k \in \mathcal{B}$ that approximately solves (**??**), i.e. such that (**??**) holds.

Step 2 [Test for convergence]. If $\|P_{T_k}(-\nabla_x \Phi_k)\| \leq \omega_*$ and $\|c(x_k)\| \leq \eta_*$, stop.

Step 3 [Disaggregated updates]. For $j = 1, \ldots, q$, execute Step 3a if

$$\|c(x_k)_{[\mathcal{Q}_j]}\| \leq \eta_k, \tag{3.7}$$

or Step 3b otherwise.

Step 3a [Update Lagrange multiplier estimates]. Set

$$\begin{aligned} \lambda_{k+1,[\mathcal{Q}_j]} &= \bar{\lambda}(x_k, \lambda_{k,[\mathcal{Q}_j]}, \mu_{k,j})_{[\mathcal{Q}_j]}, \\ \mu_{k+1,j} &= \mu_{k,j}. \end{aligned}$$

Step 3b [Reduce the penalty parameter]. Set

$$\begin{aligned} \lambda_{k+1,[\mathcal{Q}_j]} &= \lambda_{k,[\mathcal{Q}_j]}, \\ \mu_{k+1,j} &= \tau_{k,j}\mu_{k,j}, \end{aligned} \tag{3.8}$$

where

$$\tau_{k,j} = \begin{cases} \tau & \text{if } \mu_{k,j} = \alpha_k, \\ \min(\tau, \alpha_k) & \text{otherwise.} \end{cases} \tag{3.9}$$

Step 4 [Aggregated updates]. Define

$$\alpha_{k+1} = \max_{j=1,\ldots,q} \mu_{k+1,j}. \tag{3.10}$$

If

$$\alpha_{k+1} < \alpha_k, \tag{3.11}$$

then set

$$\begin{aligned} \omega_{k+1} &= \alpha_{k+1}, \\ \eta_{k+1} &= \alpha_{k+1}^{\alpha_\eta}, \end{aligned} \tag{3.12}$$

otherwise set

$$\begin{aligned} \omega_{k+1} &= \omega_k \alpha_{k+1}, \\ \eta_{k+1} &= \eta_k \alpha_{k+1}^{\beta_\eta}. \end{aligned} \tag{3.13}$$

Increment k by one and go to Step 1.

Algorithm 3.1 is specifically designed for the first-order estimate (**??**), a formula with potential advantages for large-scale computations. We refer the reader to Section **??** for a further discussion of a more flexible choice of the multipliers, covering, among others, the choice of the least-squares estimates $\lambda(x)$ as defined in (**??**).

We immediately verify that our algorithm is coherent, in that

$$\lim_{k\to\infty} \omega_k = \lim_{k\to\infty} \eta_k = 0. \tag{3.14}$$

Indeed, we obtain from (**??**) that $\alpha_k < 1$ for all k, and (**??**) then follows from (**??**) and (**??**) if α_k tends to zero, or from (**??**) alone if α_k is bounded away from zero.

The restriction (**??**) is imposed in order to simplify the exposition. In a more practical setting, it may be ignored provided the definition of α_0 and (**??**) are replaced by

$$\alpha_0 = \min\left(\gamma_s, \max_{j=1,\ldots,q} \mu_{0,j}\right) \quad \text{and} \quad \alpha_{k+1} = \min\left(\gamma_s, \max_{j=1,\ldots,q} \mu_{k+1,j}\right),$$

respectively, for some constant $\gamma_s \in (0,1)$, and that (**??**) is replaced by

$$\max_{j=1,\ldots,q} \mu_{k+1,j} < \max_{j=1,\ldots,q} \mu_{k,j}.$$

Algorithm 3.1 may be extended in other ways. For instance, one may replace the definition of ω_0, the first equation in (**??**) and the first equation of (**??**) by

$$\omega_0 = \omega_s \alpha_0^{\alpha_\omega}, \quad \omega_{k+1} = \omega_s \alpha_{k+1}^{\alpha_\omega} \text{ and } \omega_{k+1} = \omega_k \alpha_{k+1}^{\beta_\omega},$$

for some $\omega_s > 0$, $\alpha_\omega > \alpha_\eta$ and $\beta_\omega > \beta_\eta$. The definition of η_0 and the second equation in (**??**) may then be replaced by

$$\eta_0 = \eta_s \alpha_0^{\beta_\eta} \text{ and } \eta_{k+1} = \eta_s \alpha_{k+1}^{\alpha_\eta}, \tag{3.15}$$

for some $\eta_s > 0$. None of these extensions alter the results of the convergence theory developed below. The values used in the LANCELOT package in a similar context are $\alpha_\eta = \tau = \gamma_s = 0.1$, and $\beta_\eta = 0.9$ (relation (**??**) is also used with $\eta_s = 0.12589$, ensuring that $\eta_0 = 0.01$). The values $\omega_s = \alpha_\omega = \beta_\omega = 1$ and $\mu_{0,j} = 0.1$ ($j = 1, \ldots, q$) also seem suitable. The parameters ω_* and η_* specify the final accuracy requested by the user.

Finally, the purpose of the update (**??**) is to put more emphasis on the feasibility of the constraints whose violation is proportionally higher, in order to achieve a "balance" amongst all constraint violations. This balance then allows the true asymptotic regime of the algorithm to be reached. The advantage of (**??**) is that this balancing effect is obtained gradually, and not enforced at every major iteration, as is the case in **?** (**?**). Furthermore Powell's approach increases the penalties corresponding to the constraints that are becoming too slowly feasible, based on the ℓ_∞-norm. Thus it is only when they have changed sufficiently so that they are all within the constraint violation tolerance that the Lagrange multiplier update is performed. By contrast, we update the multipliers of the well-behaved constraints (assuming they correspond to a particular partition — which is likely since that is, partly at least, why the partitions exist) independently of more badly behaved ones. In addition, by virtue of using the ℓ_2-norm, we do not give quite the same emphasis to the most violated constraint.

4 Global convergence analysis

We now proceed to show that Algorithm 3.1 is globally convergent under the following assumptions.

AS3: The iterates $\{x_k\}$ considered lie within a closed, bounded domain Ω.

AS4: The matrix $J(x_*)Z_*$ has column rank no smaller than m at any limit point, x_*, of the sequence $\{x_k\}$ considered in this paper.

We notice that AS3 implies that there exists at least a convergent subsequence of iterates, but does not, of course, guarantee that this subsequence converges to a stationary point, i.e. that "the algorithm works". We also note that it is always satisfied in practice because the linear constraints (**??**) includes lower and upper bounds on the variables, either actual or implied by the finite precision of computer arithmetic.

Assumption AS4 guarantees that the dimension of the null space of A_* is large enough to provide the number of degrees of freedom that are necessary to satisfy the nonlinear constraints and we requires that the gradients of these constraints (projected onto this null space) are linearly independent at every limit point of the sequence of iterates. This assumption is the direct generalization of AS3 used by **?** (**?**).

We shall analyse the convergence of our algorithm in the case where the convergence tolerances ω_* and η_* are both zero. We first need the following lemma, proving that (**??**) prevents both the reduced gradient of the augmented Lagrangian and its orthogonal complement from being arbitrarily large when ω_k is small.

Lemma 4.1 *Let $\{x_k\} \subset \mathcal{B}, k \in \mathcal{K}$, be a sequence which converges to the point x_* and suppose that*

$$\|P_{T_k}(-\nabla_x \Phi_k)\| \leq \omega_k,$$

where the ω_k are positive scalar parameters which converge to zero as $k \in \mathcal{K}$ increases. Then

$$\|Z_*^T \nabla_x \Phi_k\| \leq \|Z_k^T \nabla_x \Phi_k\| \leq \omega_k \ \text{ and } \ \|Y_k^T(x_k - x_*)\| \leq \kappa_1 \omega_k \tag{4.1}$$

for some $\kappa_1 > 0$ and for all $k \in \mathcal{K}$ sufficiently large.

Proof. Observe that, for $k \in \mathcal{K}$ sufficiently large, ω_k is sufficiently small and x_k sufficiently close to x_* to ensure that all the constraints in D_k are active at x_*. This implies that the subspace orthogonal to the normals of the dominant constraints at x_k, $\mathcal{V}_k$, contains the subspace orthogonal to the normals of the constraints active at x_*. Hence, we deduce that

$$\|Z_*^T \nabla_x \Phi_k\| \leq \|Z_k^T \nabla_x \Phi_k\| \leq \|P_{T_k}(-\nabla_x \Phi_k)\| \leq \omega_k,$$

where we have used (**??**) to obtain the second inequality and (**??**) to deduce the third. This proves the first part of (**??**).

We now turn to the second. If D_k is empty, then Y_k is the zero matrix and the second part of (**??**) immediately follows. Assume therefore that $D_k \neq \emptyset$. We first select a submatrix $\hat{A}_{D_k}$ of A_{D_k} that is of full row-rank and note that the orthogonal projection onto the subspace spanned by the $\{a_i\}_{i \in D_k}$ is nothing but

$$Y_k Y_k^T = \hat{A}_{D_k}^T [\hat{A}_{D_k} \hat{A}_{D_k}^T]^{-1} \hat{A}_{D_k}.$$

Hence we obtain from the orthogonality of Y_k, the bound $|D_k| \leq p$, (**??**) and (**??**) and the fact that all constraints in D_k are active at x_* for k sufficiently large, that

$$\begin{aligned} \|Y_k^T(x_k - x_*)\| &\leq \|\hat{A}_{D_k}^T [\hat{A}_{D_k} \hat{A}_{D_k}^T]^{-1}\| . \|\hat{A}_{D_k}(x_k - x_*)\| \\ &\leq \|\hat{A}_{D_k}^T [\hat{A}_{D_k} \hat{A}_{D_k}^T]^{-1}\| \, p \, \kappa_0 \omega_k. \end{aligned} \tag{4.2}$$

But there are only a finite number of nonempty sets D_k for all possible choices of x_k and ω_k, and we may thus deduce the second part of (**??**) from (**??**) by defining

$$\kappa_1 = p\kappa_0 \min \|\hat{A}_{D_k}^T [\hat{A}_{D_k} \hat{A}_{D_k}^T]^{-1}\|,$$

where the minimum is taken on all possible choices of D_k and $\hat{A}_{D_k}$. □

We now examine the behaviour of the sequence $\{\nabla_x \Phi_k\}$. We first recall a result extracted from the classical perturbation theory of convex optimization problems. This result is well known and can be found, for instance, on pp.14–17 of **?** (**?**).

Lemma 4.2 *Assume that U is a continuous point-to-set mapping from $S \subseteq \Re^\ell$ into $\Re^n$ such that the set $U(\theta)$ is convex and non-empty for each $\theta \in S$. Assume that the real-valued function $F(y, \theta)$ is defined and continuous on the space $\Re^n \times S$ and convex in y for each fixed θ. Then, the real-valued function F_* defined by*

$$F_*(\theta) \stackrel{\text{def}}{=} \inf_{y \in U(\theta)} F(y, \theta)$$

is continuous on S.

We now show that, if it converges, the sequence $\{\nabla_x \Phi_k\}$ tends to a vector which is a linear combination of the rows of A_* with non-negative coefficients.

Lemma 4.3 *Let $\{x_k\} \subset \mathcal{B}, k \in \mathcal{K}$, be a sequence which converges to the point x_* and suppose that the gradients $\nabla_x \Phi_k$, $k \in \mathcal{K}$, converge to some limit $\nabla_x \Phi_*$. Assume furthermore that (**??**) holds for $k \in \mathcal{K}$ and that ω_k tends to zero as $k \in \mathcal{K}$ increases. Then,*

$$\nabla_x \Phi_* = A_*^T \pi_*$$

for some vector $\pi_* \geq 0$, where A_* is the matrix whose rows are those of A corresponding to active constraints at x_*.

Proof. We first define

$$\sigma_k \stackrel{\text{def}}{=} \max_{\substack{A(x_k+d)-b\geq 0\\ \|d\|\leq 1}} \left(-\nabla_x {\Phi_k}^T d\right) \tag{4.3}$$

with the aim to show that this quantity tends to zero when $k \in \mathcal{K}$ increases. We obtain from (**??**), the Moreau decomposition (see **?**, **?**) of $\nabla_x \Phi_k$ and the Cauchy-Schwarz inequality, that

$$\begin{array}{rcl} \sigma_k & \leq & \max_{\substack{A(x_k+d)-b\geq 0\\ \|d\|\leq 1}} P_{T_k}(-\nabla_x \Phi_k)^T d + \max_{\substack{A(x_k+d)-b\geq 0\\ \|d\|\leq 1}} P_{N_k}(-\nabla_x \Phi_k)^T d \\ & \leq & \|P_{T_k}(-\nabla_x \Phi_k)\| + \max_{d\in B_k} P_{N_k}(-\nabla_x \Phi_k)^T d, \end{array} \tag{4.4}$$

where $B_k \stackrel{\text{def}}{=} \{d \in \Re^n \mid a_i^T(x_k + d) - b_i \geq 0 \quad (i \in D_k) \quad \text{and} \quad \|d\| \leq 1\}$. As, for x_k sufficiently close to x_* and ω_k sufficiently small, all the constraints in D_k must be active at x_*, we have that N_k is included in the normal cone $N(x_*, 0)$ and therefore the vector $P_{N_k}(-\nabla_x \Phi_k)$ belongs to this normal cone. Moreover, since the maximization problem of the last right-hand side of (**??**) is a concave program, since x_* is feasible for (**??**), and since $\|x_* - x_k\| \leq 1$ for $k \in \mathcal{K}$ large enough, we thus deduce that $d = x_* - x_k$ is a global solution of this problem. Observing that

$$P_{N_k}(-\nabla_x \Phi_k)^T d = [Y_k Y_k^T P_{N_k}(-\nabla_x \Phi_k)]^T d = P_{N_k}(-\nabla_x \Phi_k)^T Y_k Y_k^T d,$$

we obtain that

$$\max_{d\in B_k} P_{N_k}(-\nabla_x \Phi_k)^T d = \max_{d\in B_k} P_{N_k}(-\nabla_x \Phi_k)^T Y_k Y_k^T d \leq \|P_{N_k}(-\nabla_x \Phi_k)\|.\|Y_k^T(x_k - x_*)\|, \tag{4.5}$$

where we used the Cauchy-Schwartz inequality to deduce the last inequality. We may now apply Lemma **??** and deduce from the second part of (**??**), (**??**) and the contractive character of the projection onto a convex set containing the origin that

$$\max_{d\in B_k} P_{N_k}(-\nabla_x \Phi_k)^T d \leq \kappa_1 \omega_k \|\nabla_x \Phi_k\|,$$

and thus, from (**??**) and our assumptions, that

$$\sigma_k \leq \omega_k + \kappa_1 \omega_k \|\nabla_x \Phi_k\|.$$

Our assumption on the ω_k sequence then implies that σ_k converges to zero as k increases in $\mathcal{K}$.

Consider now the minimization problem

$$\begin{array}{ll} \min_{d\in\Re^n} & \nabla_x\Phi_*^T d, \\ \text{subject to} & A(x_*+d)-b\geq 0, \\ & \|d\|\leq 1. \end{array} \tag{4.6}$$

Since the sequences $\{\nabla_x\Phi_k\}$ and $\{x_k\}$ converge to $\nabla_x\Phi_*$ and x_* respectively, we deduce from Lemma **??** applied to the optimization problem (**??**) (with the choices $\theta^T = (\nabla_x\Phi^T, x^T)$, $U(\theta) = \{d \mid A(x+d)-b\geq 0,\ \|d\|\leq 1\}$, $y=d$, $F(y,\theta)=\nabla_x\Phi^T d$), and the convergence of the sequence σ_k to zero that the optimal value for problem (**??**) is zero. The vector $d=0$ is thus a solution for problem (**??**) and satisfies

$$\nabla_x\Phi_* = A_*^T\pi_* - 2\zeta d = A_*^T\pi_*$$

for some vector $\pi_*\geq 0$, which ends the proof. □

The important part of our convergence analysis is the next lemma.

Lemma 4.4 *Suppose that AS1 and AS2 hold. Let $\{x_k\}\subset\mathcal{B}, k\in\mathcal{K}$, be a sequence satisfying AS3 which converges to the point x_* for which AS4 holds and let $\lambda_*=\lambda(x_*)$, where λ satisfies (**??**). Assume that $\{\lambda_k\}$, $k\in\mathcal{K}$, is any sequence of vectors and that $\{\mu_k\}$, $k\in\mathcal{K}$, form a nonincreasing sequence of q-dimensional vectors. Suppose further that (**??**) holds where the ω_k are positive scalar parameters which converge to zero as $k\in\mathcal{K}$ increases. Then*

(i) There are positive constants κ_2 and κ_3 such that

$$\|\bar\lambda(x_k,\lambda_k,\mu_k)-\lambda_*\|\leq\kappa_2\omega_k+\kappa_3\|x_k-x_*\|, \tag{4.7}$$

$$\|\lambda(x_k)-\lambda_*\|\leq\kappa_3\|x_k-x_*\|, \tag{4.8}$$

and

$$\|c(x_k)_{[\mathcal{Q}_j]}\|\leq\kappa_2\omega_k\mu_{k,j}+\mu_{k,j}\|(\lambda_k-\lambda_*)_{[\mathcal{Q}_j]}\|+\kappa_3\mu_{k,j}\|x_k-x_*\|, \tag{4.9}$$

for all $j=1,\ldots,q$ and all $k\in\mathcal{K}$ sufficiently large.

Suppose, in addition, that $c(x_*)=0$. Then

*(ii) x_* is a Kuhn-Tucker point (first-order stationary point) for the problem (**??**)– (**??**), λ_* is the corresponding vector of Lagrange multipliers, and the sequences $\{\bar\lambda(x_k,\lambda_k,\mu_k)\}$ and $\{\lambda(x_k)\}$ converge to λ_* for $k\in\mathcal{K}$;*

(iii) The gradients $\nabla_x\Phi_k$ converge to $g^\ell(x_*,\lambda_*)$ for $k\in\mathcal{K}$.

Proof. As a consequence of AS2–AS4, we have that for $k\in\mathcal{K}$ sufficiently large, $(J_kZ_*)^+$ exists, is bounded and converges to $(J(x_*)Z_*)^+$. Thus, we may write

$$\|((J_kZ_*)^+)^T\|\leq\kappa_2 \tag{4.10}$$

for some constant $\kappa_2 > 0$. Equations (**??**) and (**??**), the inner iteration termination criterion (**??**) and Lemma **??** give that

$$\|Z_*^T(g_k + J_k^T\bar{\lambda}_k)\| \leq \omega_k \tag{4.11}$$

for all $k \in \mathcal{K}$ large enough. By assumptions AS2, AS3, AS4 and (**??**), $\lambda(x)$ is bounded for all x in a neighbourhood of x_*. Thus we may deduce from (**??**), (**??**) and (**??**) that

$$\begin{array}{rcl} \|\bar{\lambda}_k - \lambda(x_k)\| & = & \|((J_kZ_*)^+)^T Z_*^T g_k + \bar{\lambda}_k\| \\ & = & \|((J_kZ_*)^+)^T(Z_*^T g_k + (J_kZ_*)^T\bar{\lambda}_k)\| \\ & \leq & \|((J_kZ_*)^+)^T\|\omega_k \\ & \leq & \kappa_2\omega_k. \end{array} \tag{4.12}$$

Moreover, from the integral mean value theorem and Lemma **??** we have that

$$\lambda(x_k) - \lambda(x_*) = \int_0^1 \nabla_x\lambda(x(s))ds \cdot (x_k - x_*), \tag{4.13}$$

where $\nabla_x\lambda(x)$ is given by equation (**??**), and where $x(s) = x_k + s(x_* - x_k)$. Now the terms within the integral sign are bounded for all x sufficiently close to x_* and hence (**??**) gives

$$\|\lambda(x_k) - \lambda_*\| \leq \kappa_3\|x_k - x_*\| \tag{4.14}$$

for all $k \in \mathcal{K}$ sufficiently large and for some constant $\kappa_3 > 0$, which implies the inequality (**??**). We then have that $\lambda(x_k)$ converges to λ_*. Combining (**??**) and (**??**) we obtain

$$\|\bar{\lambda}_k - \lambda_*\| \leq \|\bar{\lambda}_k - \lambda(x_k)\| + \|\lambda(x_k) - \lambda_*\| \leq \kappa_2\omega_k + \kappa_3\|x_k - x_*\|, \tag{4.15}$$

which gives the required inequality (**??**). Then, since by assumption ω_k tends to zero as k increases, (**??**) implies that $\bar{\lambda}_k$ converges to λ_* and therefore, from the identity (**??**), $\nabla_x\Phi_k$ converges to $g^\ell(x_*, \lambda_*)$. Furthermore, multiplying (**??**) by $\mu_{k,j}$, we obtain

$$c(x_k)_{[\mathcal{Q}_j]} = \mu_{k,j}((\bar{\lambda}_k - \lambda_*)_{[\mathcal{Q}_j]} + (\lambda_* - \lambda_k)_{[\mathcal{Q}_j]}). \tag{4.16}$$

Taking norms of (**??**) and using (**??**), we derive (**??**).

Now suppose that

$$c(x_*) = 0. \tag{4.17}$$

Lemma **??** and the convergence of $\nabla_x\Phi_k$ to $g^\ell(x_*, \lambda_*)$ give that

$$g(x_*) + J(x_*)^T\lambda_* = A_*^T\pi_*$$

for some vector $\pi_* \geq 0$. This last equation and (**??**) show that x_* is a Kuhn-Tucker point and λ_* is the corresponding set of Lagrange multipliers. Moreover (**??**) and (**??**) ensure the convergence of the sequences $\{\bar{\lambda}(x_k, \lambda_k, \mu_k)\}$ and $\{\lambda(x_k)\}$ to λ_* for $k \in \mathcal{K}$. Hence the lemma is proved. □

We finally require the following lemma in the proof of global convergence, which shows that the Lagrange multiplier estimates cannot behave too badly.

Lemma 4.5 *Suppose that, for some j $(1 \leq j \leq q)$, $\mu_{k,j}$ converges to zero as k increases when Algorithm 3.1 is executed. Then the product $\mu_{k,j}\|\lambda_{k,[\mathcal{Q}_j]}\|$ converges to zero.*

Proof. As $\mu_{k,j}$ converges to zero, Step 3b must be executed infinitely often for the j-th subset. Let $\mathcal{K}_j = \{k_0, k_1, k_2, ...\}$ be the set of indices of the iterations in which Step 3b is executed.

We consider how the j-th subset of Lagrange multiplier estimates changes between two successive iterations indexed in the set $\mathcal{K}_j$. Firstly note that $\lambda_{k_v+1,[\mathcal{Q}_j]} = \lambda_{k_v,[\mathcal{Q}_j]}$. At iteration k_v+t, for $k_v < k_v + t \le k_{v+1}$, we have

$$\lambda_{k_v+t,[\mathcal{Q}_j]} = \lambda_{k_v,[\mathcal{Q}_j]} + \sum_{l=1}^{t-1} \frac{c(x_{k_v+l})_{[\mathcal{Q}_j]}}{\mu_{k_v+l,j}} \tag{4.18}$$

where the summation is null if $t = 1$, and

$$\mu_{k_{v+1},j} = \mu_{k_v+t,j} = \mu_{k_v+1,j} = \tau_{k_v,j}\, \mu_{k_v,j}. \tag{4.19}$$

Substituting (**??**) into (**??**), multiplying both sides by $\mu_{k_v+t,j}$, taking norms and using (**??**), yields

$$\mu_{k_v+t,j}\|\lambda_{k_v+t,[\mathcal{Q}_j]}\| \le \tau\mu_{k_v,j}\|\lambda_{k_v,[\mathcal{Q}_j]}\| + \sum_{l=1}^{t-1}\|c(x_{k_v+l})_{[\mathcal{Q}_j]}\|$$

and hence

$$\mu_{k_v+t,j}\|\lambda_{k_v+t,[\mathcal{Q}_j]}\| \le \tau\mu_{k_v,j}\|\lambda_{k_v,[\mathcal{Q}_j]}\| + \sum_{l=1}^{k_{v+1}-k_v-1}\|c(x_{k_v+l})_{[\mathcal{Q}_j]}\|.$$

Using the fact that (**??**) holds for $k_v + 1 \le k_v + l \le k_{v+1} - 1$, we deduce that

$$\begin{aligned}
\mu_{k_v+t,j}\|\lambda_{k_v+t,[\mathcal{Q}_j]}\| &\le \tau\mu_{k_v,j}\|\lambda_{k_v,[\mathcal{Q}_j]}\| + \sum_{l=1}^{k_{v+1}-k_v-1}\eta_{k_v+l} \\
&\le \tau\mu_{k_v,j}\|\lambda_{k_v,[\mathcal{Q}_j]}\| + \sum_{l=1}^{\infty}\eta_{k_v+l}.
\end{aligned}$$

Now defining

$$\delta_v \stackrel{\text{def}}{=} \mu_{k_v,j}\|\lambda_{k_v,[\mathcal{Q}_j]}\| \;\text{ and }\; \rho_v \stackrel{\text{def}}{=} \sum_{l=1}^{\infty}\eta_{k_v+l}, \tag{4.20}$$

we obtain that

$$\mu_{k_v+t,j}\|\lambda_{k_v+t,[\mathcal{Q}_j]}\| \le \tau\delta_v + \rho_v \tag{4.21}$$

for all t such that $k_v < k_v + t \le k_{v+1}$, and, in particular,

$$\delta_{v+1} \le \tau\delta_v + \rho_v. \tag{4.22}$$

Thus, from (**??**) and the inequality $\tau < 1$, if ρ_v converges to zero, then δ_v and hence, from (**??**), $\mu_{k_v+t,j}\|\lambda_{k_v+t,[\mathcal{Q}_j]}\|$ both converge to zero. To complete the proof it therefore suffices to show that ρ_v converges to zero as v tends to infinity.

Suppose first that α_k is bounded away from zero. Then we must have that (**??**) is used for all k sufficiently large, with $\alpha_{k+1} = \alpha_{\min}$ for some $\alpha_{\min} \in (0, 1)$. This and the definition of ρ_v in (**??**) imply that

$$\rho_v = \frac{\eta_{k_v+1}}{1 - \alpha_{\min}^{\beta_\eta}}$$

for sufficiently large v. As (**??**) also guarantees that η_k tends to zero, we deduce that ρ_v converges to zero. This completes the proof for the first case.

Suppose now that α_k converges to zero. This implies that each of the q independent penalty parameters is reduced an infinite number of times. Consider the progress of α_k over the course of q successive decreases (**??**). As (**??**) only happens when the currently largest penalty parameter, $\mu_{k,j} say$, is reduced, as (**??**) requires that this penalty parameter is reduced by τ, and because there can only possibly be at most $q-1$ other penalty parameters in the interval $(\tau\mu_{k,j}, \mu_{k,j}]$, it follows that α_k must be reduced by at least τ over q successive decreases (**??**). Thus, considering the possible outcomes (**??**) and (**??**), each η_{k_v+l} must be bounded by a quantity of the form $(\tau^i\alpha_{k_v})^{\alpha_\eta+t\beta_\eta}$ for some indices i and t. Furthermore, at most q such terms can involve any particular i and t. Therefore, since $\tau\alpha_{k_v}<1$, we obtain that

$$\begin{aligned}\rho_v &\leq q\sum_{i=0}^{\infty}\sum_{t=0}^{\infty}(\tau^i\alpha_{k_v})^{\alpha_\eta+t\beta_\eta}\\ &= q\sum_{i=0}^{\infty}\frac{(\tau^i\alpha_{k_v})^{\alpha_\eta}}{1-(\tau^i\alpha_{k_v})^{\beta_\eta}}\\ &\leq q\sum_{i=0}^{\infty}\frac{(\tau^i\alpha_{k_v})^{\alpha_\eta}}{1-(\tau\alpha_{k_v})^{\beta_\eta}}\\ &= q\frac{\alpha_{k_v}^{\alpha_\eta}}{(1-\tau^{\alpha_\eta})(1-(\tau\alpha_{k_v})^{\beta_\eta})}.\end{aligned}$$

Thus we see that, as α_{k_v} converges to zero, so does ρ_v, completing the proof for the second case. □

We can now derive the desired global convergence property of Algorithm 3.1, which is analogous to Theorem 4.4 in **?** (**?**).

Theorem 4.6 *Assume that AS1 and AS2 hold. Let x_* be any limit point of the sequence $\{x_k\}$ generated by Algorithm 3.1 of Section* **??** *for which AS3 and AS4 hold and let $\mathcal{K}$ be the set of indices of an infinite subsequence of the x_k whose limit is x_*. Finally, let $\lambda_*=\lambda(x_*)$. Then conclusions (i), (ii) and (iii) of Lemma* **??** *hold.*

Proof. Our assumptions are sufficient to reach the conclusions of part (i) of Lemma **??**. We now show that $c(x_*)_{[\mathcal{Q}_j]}=0$ for $j=1,\ldots,q$, and therefore that $c(x_*)=0$. To see this, we consider a j $(1\leq j\leq q)$ and analyze two separate cases.

The first case is when $\mu_{k,j}$ is bounded away from zero. Hence Step 3a must be executed every iteration for k sufficiently large, implying that (**??**) is always satisfied for k large enough. We then deduce from (**??**) that $c(x_k)_{[\mathcal{Q}_j]}$ converge to zero.

The second case is when $\mu_{k,j}$ converges to zero. Then Lemma **??** shows that $\mu_{k,j}\|(\lambda_k-\lambda_*)_{[\mathcal{Q}_j]}\|$ tends to zero. Using this limit and (**??**) in (**??**), we obtain that $c(x_k)_{[\mathcal{Q}_j]}$ tends to zero, as desired.

As a consequence, conclusions (ii) and (iii) of Lemma **??** hold. □

We finally note that global convergence of Algorithm 3.1 can be proved under much weaker assumptions on $\tau_{k,j}$ and ω_k. The reader is again referred to **?** (**?**) for further details.

5 Asymptotic convergence analysis

The distinction between dominant and non-dominant (*floating*) linear inequality constraints has some implications in terms of the identification of those constraints that are active at a limit point of the sequence of iterates generated by the algorithm. Given such a point x_* we know from Theorem **??** that it is critical, i.e. that $-g^\ell(x_*, \lambda_*) \in N_* = N(x_*, 0)$ for the corresponding Lagrange multipliers λ_*. If we now consider a linear constraint with index $i \in \{1, \ldots, p\}$ that is active at x_*, we may define the normal cone $N_*^{[i]}$ to be the cone spanned by the outwards normals to all linear inequality constraints active at x_*, except the i-th one. We then say that the i-th linear inequality constraint is *strongly active* at x_* if $-g^\ell(x_*, \lambda_*) \notin N_*^{[i]}$. In other words, the i-th constraint is strongly active at a critical point if this point ceases to be critical when this constraint is ignored. Let us denote by $S(x_*)$ the set of strongly active constraints at x_*. All non-strongly active constraints at x_* are called *weakly active* at x_*. We now prove the reasonable result that all strongly active constraints at a limit point x_* are dominant for k large enough.

Theorem 5.1 *Assume that AS1–AS3 hold. Let $\{x_k\}$, $k \in \mathcal{K}$, be a convergent subsequence of iterates produced by Algorithm 3.1, whose limit point is x_* with corresponding Lagrange multipliers λ_*. Assume furthermore that AS4 holds at x_*. Then*

$$S(x_*) \subseteq D_k$$

for all k sufficiently large.

Proof. Consider a linear inequality constraint $i \in S(x_*)$. Then, by definition of this latter set, we have that $-g^\ell(x_*, \lambda_*) \notin N_*^{[i]}$. Since Theorem **??** guarantees that $\nabla_x \Phi_k$ converges to $g^\ell(x_*, \lambda_*)$ and as $N_*^{[i]}$ is closed, we have that $-\nabla_x \Phi_k \notin N_*^{[i]}$ for $k \in \mathcal{K}$ large enough. Therefore, one obtains from the Moreau decomposition (see **?**, **?**) of $-\nabla_x \Phi_k$ that

$$\|P_{T_*^{[i]}}(-\nabla_x \Phi_k)\| \geq \epsilon \tag{5.1}$$

for some $\epsilon > 0$ and for all sufficiently large $k \in \mathcal{K}$, where $T_*^{[i]} = \left[N_*^{[i]}\right]^0$. We have also from (**??**) that $\|P_{T_k}(-\nabla_x \Phi_k)\|$ is arbitrarily small, because ω_k tends to zero (see (**??**)). Assume now that, for some arbitrarily large $k \in \mathcal{K}$, we have that $i \notin D_k$. This implies that $T_*^{[i]} \subseteq T_k$, and hence that (**??**) is impossible. We therefore deduce that i must belong to D_k, which proves the theorem. □

This result is important and is the generalization of Theorem 5.4 by **?** (**?**). It can also be interpreted as a means of active constraint identification, as is clear from the following easy corollary.

Corollary 5.2 *Suppose that the conditions of Theorem **??** hold. Assume furthermore that all linear inequality constraints active at x_* have linearly independent normals and are non-degenerate, in the sense that*

$$-g^\ell(x_*, \lambda_*) \in \mathrm{ri}[N_*], \tag{5.2}$$

where $\mathrm{ri}[V]$ denotes the relative interior of a convex set V. Then D_k is identical to the set of active linear inequality constraints at x_* for all $k \in \mathcal{K}$ sufficiently large.

Proof. The non-degeneracy assumption and the linear independence of the active constraints normals imply that λ_* is unique and only has strictly negative components. Therefore each of the active linear inequality constraints at x_* is strongly active at x_*, and the desired conclusion follows from Theorem **??**. $\Box$

We note here that the non-degeneracy assumption corresponds to strict complementarity slackness in our context (see, for instance, **?**, **?**, or **?**, **?**).

We now make some additional assumptions before pursuing our local convergence analysis. We intend to show that all penalty parameters are bounded away from zero.

AS5: The second derivatives of the functions $f(x)$ and $c_i(x)$ $(1 \le i \le m)$ are Lipschitz continuous at any limit point x_* of the sequence of iterates $\{x_k\}$.

AS6: Suppose that (x_*, λ_*) is a Kuhn-Tucker point for problem (**??**)–(**??**) and let $\mathcal{I}$ be any subset of the linear inequality constraints which are active at x_* that contains all strongly active constraints ($S(x_*) \subseteq \mathcal{I}$) plus an arbitrary subset of weakly active constraints at x_*. Then, if the columns of the matrix Z form an orthonormal basis of the subspace orthogonal to the normals of the constraints in $\mathcal{I}$, we assume that the matrix

$$\begin{pmatrix} Z^T H^\ell(x_*, \lambda_*) Z & Z^T J(x_*)^T \\ J(x_*) Z & 0 \end{pmatrix}$$

is nonsingular for all possible choices of the weakly active constraints in the set $\mathcal{I}$.

We note that AS6 implies AS4 and seems reasonable in that the definition of strongly and weakly active constraints may vary with small perturbations in the problem, for instance when $g^\ell(x_*, \lambda_*)$ lies in one of the extreme faces of the cone N_*. Our assumption might be seen as a safeguard against the possible effect of all such perturbations.

We now make the distinction between the subsets for which the penalty parameter converges to zero and those for which it stays bounded away from zero. We define

$$\mathcal{Z} \stackrel{\text{def}}{=} \{j \in \{1, \ldots, q\} \mid \lim_{k \to \infty} \mu_{k,j} = 0\} \text{ and } \mathcal{P} \stackrel{\text{def}}{=} \{1, \ldots, q\} \setminus \mathcal{Z}.$$

We also denote

$$\mu_{k,\mathcal{Z}} \stackrel{\text{def}}{=} \max_{j \in \mathcal{Z}} \mu_{k,j} \tag{5.3}$$

and

$$\rho_k \stackrel{\text{def}}{=} \sum_{j \in \mathcal{Z}} \mu_{k,j} \|(\lambda_k - \lambda_*)_{[\mathcal{Q}_j]}\|.$$

We now prove an analog to Lemma 5.1 by **?** (**?**) which is suitable for our more general framework.

Lemma 5.3 *Assume that AS1–AS3 hold. Let $\{x_k\}$, $k \in \mathcal{K}$, be a convergent subsequence of iterates produced by Algorithm 3.1, whose limit point is x_* with corresponding Lagrange multipliers λ_*. Assume that AS5 and AS6 hold at x_*. Assume furthermore that $\mathcal{Z} \neq \emptyset$.*

(i) If $\mathcal{P} = \emptyset$, there are positive constants $\bar{\alpha} < 1$, κ_4, κ_5, κ_6, κ_7 and an integer k_1 such that, if $\alpha_{k_1} \leq \bar{\alpha}$, then

$$\|x_k - x_*\| \leq \kappa_4 \omega_k + \kappa_5 \alpha_k \|\lambda_k - \lambda_*\|, \tag{5.4}$$

$$\|\bar{\lambda}(x_k, \lambda_k, \mu_k) - \lambda_*\| \leq \kappa_6 \omega_k + \kappa_7 \alpha_k \|\lambda_k - \lambda_*\|, \tag{5.5}$$

and

$$\|c(x_k)\| \leq q\kappa_6 \omega_k \alpha_k + q\alpha_k (1 + \kappa_7 \alpha_k) \|\lambda_k - \lambda_*\|, \tag{5.6}$$

for all $k \geq k_1$, $k \in \mathcal{K}$.

(ii) If, on the other hand, $\mathcal{P} \neq \emptyset$, there are positive constants $\bar{\alpha} < 1$, κ_4, κ_5, κ_6, κ_7 and an integer k_1 such that, if $\mu_{k_1, \mathcal{Z}} \leq \bar{\alpha}$, then

$$\|(\bar{\lambda}(x_k, \lambda_k, \mu_k) - \lambda_*)_{[\mathcal{Q}_j]}\| \leq \kappa_6 \eta_k + \kappa_7 \rho_k, \tag{5.7}$$

and

$$\|c(x_k)_{[\mathcal{Q}_j]}\| \leq \kappa_6 \eta_k \mu_{k,\mathcal{Z}} + (1 + \kappa_7 \mu_{k,\mathcal{Z}}) \rho_k, \tag{5.8}$$

for all $k \geq k_1$, $k \in \mathcal{K}$, and all $j \in \mathcal{Z}$.

Proof. We will denote the gradient and Hessian of the Lagrangian function, taken with respect to x, at the limit point (x_*, λ_*) by g_*^ℓ and H_*^ℓ, respectively. Similarly, J_* will denote $J(x_*)$. We also define $\delta_k = x_k - x_*$. We observe that the assumptions of the lemma guarantee that Theorem **??** can be used.

We first note that there is only a finite number of possible D_k, and we may thus consider subsequences of $\mathcal{K}$ such that D_k is constant in each subsequence. We also note that each $k \in \mathcal{K}$ belongs to a unique such subsequence. In order to prove our result, it is thus sufficient to consider an arbitrary infinite subsequence $\bar{\mathcal{K}}$ such that, for $k \in \bar{\mathcal{K}}$, D_k is independent of k. This "constant" index set will be denoted by D. As a consequence, the cones N_k and T_k, the subspaces $\mathcal{V}_k$ and $\mathcal{W}_k$ and the orthogonal matrices Z_k and Y_k are also independent of k; they are denoted by N, T, $\mathcal{V}$, $\mathcal{W}$, Z and Y, respectively.

Using (**??**) and Taylor's expansion around x_*, we obtain that

$$\begin{array}{rcl} \nabla_x \Phi_k & = & g_k + J_k^T \bar{\lambda}_k \\ & = & g(x_*) + H(x_*)\delta_k + J_*^T \bar{\lambda}_k + \sum_{i=1}^m \bar{\lambda}_{k,i} H_i(x_*)\delta_k + r_1(x_k, x_*, \bar{\lambda}_k) \\ & = & g_*^\ell + H_*^\ell \delta_k + J_*^T(\bar{\lambda}_k - \lambda_*) + r_1(x_k, x_*, \bar{\lambda}_k) + r_2(x_k, x_*, \bar{\lambda}_k, \lambda_*), \end{array} \tag{5.9}$$

where

$$r_1(x_k, x_*, \bar{\lambda}_k) \stackrel{\text{def}}{=} \int_0^1 [H^\ell(x_* + s\delta_k, \bar{\lambda}_k) - H^\ell(x_*, \bar{\lambda}_k)]\delta_k \, ds$$

and

$$r_2(x_k, x_*, \bar{\lambda}_k, \lambda_*) \stackrel{\text{def}}{=} \sum_{i=1}^m (\bar{\lambda}_{k,i} - \lambda_{*,i}) H_i(x_*)\delta_k.$$

The boundedness and Lipschitz continuity of the Hessian matrices of f and c_i in a neighbourhood of x_*, together with the convergence of $\bar{\lambda}_k$ to λ_* then imply that

$$\|r_1(x_k, x_*, \bar{\lambda}_k)\| \leq \kappa_8 \|\delta_k\|^2, \tag{5.10}$$

and

$$\|r_2(x_k, x_*, \bar{\lambda}_k, \lambda_*)\| \leq \kappa_9 \|\delta_k\| \, \|\bar{\lambda}_k - \lambda_*\| \tag{5.11}$$

for some positive constants κ_8 and κ_9. Moreover, using Taylor's expansion again, along with the fact that Theorem **??** ensures the equality $c(x_*) = 0$, we obtain that

$$c(x_k) = J_* \delta_k + r_3(x_k, x_*), \tag{5.12}$$

where

$$[r_3(x_k, x_*)]_i = \int_0^1 s \int_0^1 \delta_k^T H_i(x_* + ts\delta_k)\delta_k \, dt \, ds$$

(see **?**, **?**, p.11). The boundedness of the Hessian matrices of the c_i in a neighbourhood of x_* then gives that

$$\|r_3(x_k, x_*)\| \leq \kappa_{10} \|\delta_k\|^2 \tag{5.13}$$

for some positive constant κ_{10}. Combining (**??**) and (**??**), we obtain

$$\begin{pmatrix} H_*^\ell & J_*^T \\ J_* & 0 \end{pmatrix} \begin{pmatrix} \delta_k \\ \bar{\lambda}_k - \lambda_* \end{pmatrix} = \begin{pmatrix} \nabla_x \Phi_k - g_*^\ell \\ c(x_k) \end{pmatrix} - \begin{pmatrix} r_1 + r_2 \\ r_3 \end{pmatrix}, \tag{5.14}$$

where we have suppressed the arguments of the residuals r_1, r_2 and r_3 for brevity. Using the orthogonal decomposition of $\Re^n$ into $\mathcal{V} \oplus \mathcal{W}$ and defining

$$Q = \begin{pmatrix} Z & Y & 0 \\ 0 & 0 & I \end{pmatrix},$$

we may rewrite (**??**) as

$$Q^T \begin{pmatrix} H_*^\ell & J_*^T \\ J_* & 0 \end{pmatrix} Q Q^T \begin{pmatrix} \delta_k \\ \bar{\lambda}_k - \lambda_* \end{pmatrix} = Q^T \begin{pmatrix} \nabla_x \Phi_k - g_*^\ell \\ c(x_k) \end{pmatrix} - Q^T \begin{pmatrix} r_4 \\ r_3 \end{pmatrix},$$

where $r_4 \stackrel{\text{def}}{=} r_1 + r_2$. Expanding this last equation gives that

$$\begin{pmatrix} Z^T H_*^\ell Z & Z^T H_*^\ell Y & Z^T J_*^T \\ Y^T H_*^\ell Z & Y^T H_*^\ell Y & Y^T J_*^T \\ J_* Z & J_* Y & 0 \end{pmatrix} \begin{pmatrix} Z^T \delta_k \\ Y^T \delta_k \\ \bar{\lambda}_k - \lambda_* \end{pmatrix} = \begin{pmatrix} Z^T(\nabla_x \Phi_k - g_*^\ell) \\ Y^T(\nabla_x \Phi_k - g_*^\ell) \\ c(x_k) \end{pmatrix} - \begin{pmatrix} Z^T r_4 \\ Y^T r_4 \\ r_3 \end{pmatrix}. \tag{5.15}$$

We now observe that (**??**), the inclusion $\mathcal{V} \subseteq T$ and the fact that ω_k tends to zero imply that

$$Z^T g_*^\ell = 0. \tag{5.16}$$

Substituting (**??**) in (**??**), removing the middle horizontal block and rearranging the terms of this latter equation then yields that

$$\begin{pmatrix} Z^T H_*^\ell Z & Z^T J_*^T \\ J_* Z & 0 \end{pmatrix} \begin{pmatrix} Z^T \delta_k \\ \bar{\lambda}_k - \lambda_* \end{pmatrix} = \begin{pmatrix} Z^T(\nabla_x \Phi_k - H_*^\ell Y Y^T \delta_k) \\ c(x_k) - J_* Y Y^T \delta_k \end{pmatrix} - \begin{pmatrix} Z^T r_4 \\ r_3 \end{pmatrix}. \tag{5.17}$$

Roughly speaking, we now proceed by showing that the right-hand side of this relation is of the order of $\theta_k + \rho_k$, where

$$\theta_k \stackrel{\text{def}}{=} \begin{cases} \omega_k & \text{if } \mathcal{P} = \emptyset, \\ \eta_k & \text{if } \mathcal{P} \neq \emptyset. \end{cases} \tag{5.18}$$

We will then ensure that the vector on the left-hand side is of the same size, which is essentially the result we aim to prove. We first observe that

$$\|\delta_k\| = \|ZZ^T\delta_k + YY^T\delta_k\| \le \|Z^T\delta_k\| + \kappa_1\omega_k \tag{5.19}$$

from (??). We then obtain from (??) and (??) that

$$\|\bar{\lambda}_k - \lambda_*\| \le \kappa_{11}\omega_k + \kappa_3\|Z^T\delta_k\|, \tag{5.20}$$

where $\kappa_{11} = \kappa_2 + \kappa_3\kappa_1$. Furthermore, from (??), (??), (??), (??) and (??),

$$\left\| \begin{pmatrix} Z^T r_4 \\ r_3 \end{pmatrix} \right\| \le \kappa_{12}\|Z^T\delta_k\|^2 + \kappa_{13}\|Z^T\delta_k\|\omega_k + \kappa_{14}{\omega_k}^2, \tag{5.21}$$

where $\kappa_{12} = \kappa_8 + \kappa_3\kappa_9 + \kappa_{10}$, $\kappa_{13} = 2\kappa_1(\kappa_8 + \kappa_{10}) + \kappa_9(\kappa_{11} + \kappa_3\kappa_1)$, and $\kappa_{14} = \kappa_1^2(\kappa_8 + \kappa_{10}) + \kappa_1\kappa_9\kappa_{11}$. We now bound $c(x_k)$ by distinguishing components from $\mathcal{Z}$ and $\mathcal{P}$. We first note that, since the penalty parameters for each subset in $\mathcal{P}$ are bounded away from zero, the test (??) is satisfied for all k sufficiently large. Moreover, the remaining components of $c(x_k)$ satisfy the bound

$$\|c(x_k)_{[\mathcal{Q}_j]}\| \le \kappa_2\omega_k\mu_{k,j} + \mu_{k,j}\|(\lambda_k - \lambda_*)_{[\mathcal{Q}_j]}\| + \kappa_3\mu_{k,j}\|x_k - x_*\|, \tag{5.22}$$

for all $j \in \mathcal{Z}$ and all k sufficiently large, using (??). Hence, using (??), (??) and (??), we deduce that

$$\begin{aligned} \|c(x_k)\| &\le \sum_{j\in\mathcal{P}} \|c(x_k)_{[\mathcal{Q}_j]}\| + \sum_{j\in\mathcal{Z}} \|c(x_k)_{[\mathcal{Q}_j]}\| \\ &\le q\eta_k + q\kappa_2\omega_k\mu_{k,\mathcal{Z}} + \rho_k + q\kappa_3\mu_{k,\mathcal{Z}}\|\delta_k\|. \end{aligned} \tag{5.23}$$

Note that the first term of the last right-hand side only appears if $\mathcal{P}$ is not empty. Since the algorithm ensures that

$$\omega_k \le \eta_k \tag{5.24}$$

because $\alpha_\eta < 1$ and $\beta_\eta < 1$, we may obtain from (??), (??) and (??) that

$$\left\| \begin{pmatrix} Z^T(\nabla_x\Phi_k - H_*^\ell YY^T\delta_k) \\ c(x_k) - J_*YY^T\delta_k \end{pmatrix} \right\| \le \kappa_{15}\theta_k + \rho_k + q\kappa_{11}\mu_{k,\mathcal{Z}}\theta_k + q\kappa_3\mu_{k,\mathcal{Z}}\|Z^T\delta_k\|, \tag{5.25}$$

where $\kappa_{15} = q + 1 + \kappa_1(\|Z^T H_*^\ell Y\| + \|J_*Y\|)$. By assumption AS6, the coefficient matrix on the left-hand side of (??) is nonsingular. Let M be the norm of its inverse. Multiplying both sides of the equation by this inverse and taking norms, we obtain from (??), (??), (??) and (??) that

$$\begin{aligned} \left\| \begin{pmatrix} Z^T\delta_k \\ \bar{\lambda}_k - \lambda_* \end{pmatrix} \right\| \le\ & M[\kappa_{12}\|Z^T\delta_k\|^2 + \kappa_{13}\|Z^T\delta_k\|\theta_k + \kappa_{14}\theta_k^2 + \kappa_{15}\theta_k \\ & + \rho_k + q\kappa_{11}\mu_{k,\mathcal{Z}}\theta_k + q\kappa_3\mu_{k,\mathcal{Z}}\|Z^T\delta_k\|]. \end{aligned} \tag{5.26}$$

Suppose now that k is sufficiently large to ensure that

$$\theta_k \le \frac{1}{4M\kappa_{13}} \tag{5.27}$$

and let

$$\bar{\alpha} \stackrel{\text{def}}{=} \min\left[\alpha_0, \frac{1}{4Mq\kappa_3}\right]. \tag{5.28}$$

Recall that α_0 and hence $\bar{\alpha} < 1$. Then, if $\mu_{k,\mathcal{Z}} \le \bar{\alpha}$, the relations (**??**)–(**??**) give

$$\|Z^T\delta_k\| \le \frac{1}{2}\|Z^T\delta_k\| + M[\kappa_{16}\theta_k + \rho_k + \kappa_{12}\|Z^T\delta_k\|^2], \tag{5.29}$$

where $\kappa_{16} = q\kappa_{11} + \kappa_{14} + \kappa_{15}$. As δ_k, and hence $\|Z^T\delta_k\|$ converge to zero, we have that

$$\|Z^T\delta_k\| \le \frac{1}{4M\kappa_{12}} \tag{5.30}$$

for k large enough. Hence inequalities (**??**) and (**??**) yield that

$$\|Z^T\delta_k\| \le 4M(\kappa_{16}\theta_k + \rho_k). \tag{5.31}$$

If $\mathcal{P}$ is empty, we use (**??**), (**??**) and (**??**), the fact that $\mu_{k,\mathcal{Z}} = \alpha_k$ and the inequality

$$\rho_k \le q\alpha_k\|\lambda_k - \lambda_*\|$$

to deduce (**??**), where $\kappa_4 \stackrel{\text{def}}{=} 4M\kappa_{16} + \kappa_1$ and $\kappa_5 = 4Mq$. Defining $\kappa_6 \stackrel{\text{def}}{=} \kappa_2 + \kappa_3\kappa_4$ and $\kappa_7 \stackrel{\text{def}}{=} \kappa_3\kappa_5$, we deduce (**??**) from (**??**) and (**??**). Now, using (**??**),

$$\|c(x_k)\| \le \sum_{j=1}^{q}\|c(x_k)_{[\mathcal{Q}_j]}\| = \sum_{j=1}^{q}\mu_{k,j}\|(\bar{\lambda}_k - \lambda_k)_{[\mathcal{Q}_j]}\| \le q\alpha_k(\|\bar{\lambda}_k - \lambda_*\| + \|\lambda_k - \lambda_*\|) \tag{5.32}$$

and (**??**) then follows from (**??**) and (**??**).

If, on the other hand, $\mathcal{P}$ is not empty, (**??**) results from (**??**), (**??**), (**??**) with $\theta_k = \eta_k$ and (**??**), with $\kappa_6 \stackrel{\text{def}}{=} 4M\kappa_3\kappa_{16} + \kappa_2 + \kappa_3\kappa_1$ and $\kappa_7 \stackrel{\text{def}}{=} 4M\kappa_3$. Finally, (**??**) results from (**??**) and (**??**). □

For the remaining of this section, we will restrict our attention to the case where the sequence of iterates converges to a single limit point. Obviously, this makes AS3 unnecessary. We briefly comment at the end of the section on why this additional assumption cannot be relaxed.

We now show that, if the maximum penalty parameter α_k converges to zero, then the Lagrange multiplier estimates λ_k converge to their true values λ_*.

Lemma 5.4 *Assume AS1 and AS2 hold. Assume that $\{x_k\}$, the sequence of iterates generated by Algorithm 3.1, converges to the single limit point x_* at which AS6 holds, and with corresponding Lagrange multipliers λ_*. Then, if α_k tends to zero, the sequence λ_k converges to λ_*.*

Proof. Recall that AS6 implies AS4 and therefore that our assumptions are sufficient to apply Theorem **??**.

We observe that the desired convergence holds if $\lambda_{k,[\mathcal{Q}_j]}$ converges to $\lambda_{*,[\mathcal{Q}_j]}$, for all $j = 1, \ldots, q$. It is thus sufficient to show this latter result for an arbitrary j between 1 and q. The result is obvious if Step 3a is executed infinitely often for the j-th subset. Indeed, each time this step is executed, $\lambda_{k+1,[\mathcal{Q}_j]} = \bar{\lambda}_{k,[\mathcal{Q}_j]}$ and the inequality (**??**) guarantees that $\bar{\lambda}_{k,[\mathcal{Q}_j]}$ converges to

$\lambda_{*,[\mathcal{Q}_j]}$. Suppose therefore that Step 3a is not executed infinitely often for this subset. Then $\|(\lambda_k - \lambda_*)_{[\mathcal{Q}_j]}\|$ will remain fixed for all $k \geq k_2$, for some $k_2 > 0$, as Step 3b is executed for each remaining iteration. But (**??**) then implies that $\|c(x_k)_{[\mathcal{Q}_j]}\| \leq \kappa_{17}\mu_{k,j}$, for some constant $\kappa_{17} > 0$ and for all $k \geq k_3 \geq k_2$. As α_k tends to zero and $\alpha_\eta < 1$, $\kappa_{17}\mu_{k,j} \leq \kappa_{17}\alpha_k \leq \alpha_k^{\alpha_\eta} = \eta_k$ for all k sufficiently large for which α_k strictly decreases. But then inequality (**??**) must be satisfied for some $k \geq k_3$, which is impossible, as this would imply that Step 3a is again executed for the j-th subset. Hence Step 3a must be executed infinitely often. □

We now consider the behaviour of the maximum penalty parameter α_k and show the important result that, under stated assumptions, it is bounded away from zero. The proof of this result is inspired by the technique developed by **?** (**?**). When the single penalty parameter definition of the augmented Lagrangian (**??**) is used (or, equivalently, when $q = 1$), one then avoids a steadily increasing ill-conditioning of the Hessian of the augmented Lagrangian. Note that this ill-conditioning is also avoided when $q > 1$, as we show below in Theorem **??**.

Theorem 5.5 *Assume AS1 and AS2 hold and suppose that the sequence of iterates $\{x_k\}$ of Algorithm 3.1 converges to a single limit point x_* with corresponding Lagrange multipliers λ_*, at which AS5 and AS6 hold. Then there is a constant $\alpha_{\min} \in (0,1)$ such that $\alpha_k = \alpha_{\min}$ for all k.*

Proof. Suppose otherwise that α_k tends to zero (that is $\mathcal{P} = \emptyset$), and hence that $\mu_{k,j}$ tends to zero for each j between 1 and q. Then Step 3b must be executed infinitely often for each subset. We aim to obtain a contradiction to this statement by showing that Step 3a is always executed for each subset for sufficiently large k. We note that our assumptions are sufficient to apply Theorem **??**. Furthermore, we may apply Lemma **??** to the complete sequence of iterates.

First observe that

$$\alpha_k \leq \bar{\alpha} < 1 \tag{5.33}$$

for all $k \geq k_1$, where $\bar{\alpha}$ and k_1 are those of Lemma **??**. Note that

$$\omega_k \leq \alpha_k$$

for all $k \geq k_1$. This follows by definition if (**??**) is executed. Otherwise it is a consequence of the fact that α_k is unchanged while ω_k is reduced, when (**??**) occurs. Let k_4 be the smallest integer such that

$$\alpha_k^{1-\alpha_\eta} \leq \frac{1}{q(2+\kappa_6)}, \tag{5.34}$$

and

$$\alpha_k^{1-\beta_\eta} \leq \min\left[\frac{1}{\kappa_{18}}, \frac{1}{q(2\kappa_{18}+\kappa_6)}\right], \tag{5.35}$$

where $\kappa_{18} = \max(1, \kappa_6 + \kappa_7)$. Note that (**??**) and (**??**) imply that

$$\alpha_k \leq \alpha_k^{1-\beta_\eta} \leq \frac{1}{\kappa_{18}} \leq \frac{1}{\kappa_7} \tag{5.36}$$

for all $k \geq \max(k_1, k_4)$. Furthermore, let k_5 be such that

$$\|\lambda_k - \lambda_*\| \leq 1 \tag{5.37}$$

for all $k \geq k_5$, which is possible because of Lemma **??**. Now define $k_6 = \max(k_1, k_4, k_5)$, let Γ be the set $\{k \mid$ (**??**) is executed at iteration $k-1$ and $k \geq k_6\}$ and let k_0 be the smallest element of Γ. By the assumption that α_k tends to zero, Γ has an infinite number of elements.

By definition of Γ, for iteration k_0, $\omega_{k_0} = \alpha_{k_0}$ and $\eta_{k_0} = \alpha_{k_0}^{\alpha_\eta}$. Then inequality (**??**) gives that, for each j,

$$\begin{aligned}
\|c(x_{k_0})_{[\mathcal{Q}_j]}\| &\leq \|c(x_{k_0})\| \\
&\leq q(\alpha_{k_0} + \kappa_7 \alpha_{k_0}^2)\|\lambda_{k_0} - \lambda_*\| + q\kappa_6 \omega_{k_0} \alpha_{k_0} \\
&\leq 2q\alpha_{k_0}\|\lambda_{k_0} - \lambda_*\| + q\kappa_6 \omega_{k_0} \alpha_{k_0} && \text{(from (??))} \\
&\leq q\alpha_{k_0}(2 + \kappa_6 \alpha_{k_0}) && \text{(from (??))} \\
&\leq q(2 + \kappa_6)\alpha_{k_0} && \text{(from (??))} \\
&\leq \alpha_{k_0}^{\alpha_\eta} = \eta_{k_0} && \text{(from (??)).}
\end{aligned} \tag{5.38}$$

As a consequence of this inequality, Step 3a will be executed for each j with $\lambda_{k_0+1,[\mathcal{Q}_j]} = \bar{\lambda}(x_{k_0}, \lambda_{k_0,[\mathcal{Q}_j]}, \mu_{k_0,j})_{[\mathcal{Q}_j]}$. Inequality (**??**) together with (**??**) guarantee that

$$\|\lambda_{k_0+1} - \lambda_*\| \leq \kappa_6 \omega_{k_0} + \kappa_7 \alpha_{k_0}\|\lambda_{k_0} - \lambda_*\| \leq \kappa_{18}\alpha_{k_0}. \tag{5.39}$$

We shall now make use of an inductive proof. Assume that, for each j, Step 3a is executed for iterations $k_0 + i$, $(0 \leq i \leq t)$, and that

$$\|\lambda_{k_0+i+1} - \lambda_*\| \leq \kappa_{18}\alpha_{k_0}^{1+\beta_\eta i}. \tag{5.40}$$

Inequalities (**??**) and (**??**) show that this is true for $t = 0$. We aim to show that the same is true for $i = t+1$. Our assumption that Step 3a is executed gives that, for iteration $k_0 + t + 1$, $\alpha_{k_0+t+1} = \alpha_{k_0}$, $\omega_{k_0+t+1} = \alpha_{k_0}^{t+2}$, and $\eta_{k_0+t+1} = \alpha_{k_0}^{\beta_\eta(t+1)+\alpha_\eta}$. Then, inequality (**??**) yields that, for each j,

$$\begin{aligned}
\|c(x_{k_0+t+1})_{[\mathcal{Q}_j]}\| &\leq \|c(x_{k_0+t+1})\| \\
&\leq q(\alpha_{k_0+t+1} + \kappa_7 \alpha_{k_0+t+1}^2)\|\lambda_{k_0+t+1} - \lambda_*\| \\
&\quad + q\kappa_6 \omega_{k_0+t+1}\alpha_{k_0+t+1} \\
&\leq 2q\alpha_{k_0+t+1}\|\lambda_{k_0+t+1} - \lambda_*\| + q\kappa_6 \omega_{k_0+t+1}\alpha_{k_0+t+1} && \text{(from (??))} \\
&\leq 2q\kappa_{18}\alpha_{k_0}\alpha_{k_0}^{1+\beta_\eta t} + q\kappa_6 \alpha_{k_0}^{t+3} && \text{(from (??))} \\
&\leq 2q\kappa_{18}\alpha_{k_0}\alpha_{k_0}^{\alpha_\eta+\beta_\eta t} + q\kappa_6 \alpha_{k_0}^{\alpha_\eta+\beta_\eta(t+1)+1} \\
&\leq q(2\kappa_{18} + \kappa_6)\alpha_{k_0}^{1-\beta_\eta}\alpha_{k_0}^{\beta_\eta(t+1)+\alpha_\eta} && \text{(from (??))} \\
&\leq \alpha_{k_0}^{\beta_\eta(t+1)+\alpha_\eta} = \eta_{k_0+t+1} && \text{(from (??)).}
\end{aligned}$$

Hence Step 3a will again be executed for each j with

$$\lambda_{k_0+t+2,[\mathcal{Q}_j]} = \bar{\lambda}(x_{k_0+t+1}, \lambda_{k_0+t+1,[\mathcal{Q}_j]}, \mu_{k_0+t+1,j})_{[\mathcal{Q}_j]}.$$

Inequality (**??**) then implies that

$$
\begin{aligned}
\|\lambda_{k_0+t+2}-\lambda_*\| &\le \kappa_6\omega_{k_0+t+1}+\kappa_7\alpha_{k_0+t+1}\|\lambda_{k_0+t+1}-\lambda_*\| \\
&\le \kappa_6\alpha_{k_0}^{t+2}+\kappa_7\kappa_{18}\alpha_{k_0}\alpha_{k_0}^{1+\beta_\eta t} && \text{(from (??))} \\
&\le \kappa_6\alpha_{k_0}^{1+\beta_\eta(t+1)}+\kappa_7\kappa_{18}\alpha_{k_0}\alpha_{k_0}^{1+\beta_\eta t} \\
&= (\kappa_6+\kappa_7\kappa_{18}\alpha_{k_0}^{1-\beta_\eta})\alpha_{k_0}^{1+\beta_\eta(t+1)} \\
&\le (\kappa_6+\kappa_7)\alpha_{k_0}^{1+\beta_\eta(t+1)} && \text{(from (??))} \\
&\le \kappa_{18}\alpha_{k_0}^{1+\beta_\eta(t+1)}
\end{aligned}
$$

which establishes (**??**) for $i=t+1$. Thus Step 3a is executed for each $j=1,\ldots,q$ for all iterations $k\ge k_0$. But this implies that Γ is finite, which contradicts the assumption that Step 3b is executed infinitely often for each subset. Hence the theorem is proved. □

This theorem was all that was needed in **?** (**?**). However, the situation is more complex here because q may be larger than one. If the ill-conditioning of the Hessian is to be avoided, we must now prove the stronger result that *all* penalty parameters stay bounded away from zero.

Theorem 5.6 *Assume AS1 and AS2 hold and suppose that the sequence of iterates $\{x_k\}$ of Algorithm 3.1 converges to a single limit point x_* with corresponding Lagrange multipliers λ_*, at which AS5 and AS6 hold. Then there is a constant $\mu>0$ such that $\mu_{k,j}\ge\mu$ for all k and all $j=1,\ldots,q$.*

Proof. Assume otherwise that $\mathcal{Z}$ is not empty, and hence that $\mu_{k,\mathcal{Z}}$ converges to zero. Then Step 3b must be executed infinitely often for $j\in\mathcal{Z}$. We aim to obtain a contradiction to this statement by showing that, for any $j\in\mathcal{Z}$, Step 3a is always executed for sufficiently large k. We may deduce from Theorem **??** that α_k attains its minimum value $\alpha_{\min}\in(0,1)$ at iteration $k_{\max}$, say. Hence, $\mathcal{P}\ne\emptyset$. Furthermore, we may apply Lemma **??** to the complete sequence of iterates. Let $k_7\ge k_{\max}$ be the smallest integer for which

$$
\mu_{k,\mathcal{Z}}\le\min\left[\bar\alpha,\frac{1}{2\kappa_6+\kappa_7},\frac{\alpha_{\min}^{\beta_\eta+\epsilon}-\alpha_{\min}}{q(2\kappa_6+\kappa_7)}\right] \tag{5.41}
$$

for all $k\ge k_7\ge k_1$, where $\bar\alpha$ and k_1 are those of Lemma **??**, and where $\epsilon=\frac{1}{2}(1-\beta_\eta)$. Note that $\alpha_{\min}^{\beta_\eta+\epsilon}>\alpha_{\min}$ as $\beta_\eta<1$.

Consider the j-th subset, for some $j\in\mathcal{Z}$. At iteration $k\ge k_7$, the algorithm ensures that

$$
\mu_{k+1,j}\|(\lambda_{k+1}-\lambda_*)_{[Q_j]}\|\le\alpha_{\min}\mu_{k,j}\|(\lambda_k-\lambda_*)_{[Q_j]}\|
$$

if Step 3b is executed for the j-th subset, while (**??**) ensures that

$$
\mu_{k+1,j}\|(\lambda_{k+1}-\lambda_*)_{[Q_j]}\|\le\mu_{k,j}(\kappa_6\eta_k+\kappa_7\rho_k)
$$

if Step 3a is executed for the same subset. Summing on all $j\in\mathcal{Z}$, and defining

$$
\begin{aligned}
\mathcal{Z}_{k,a} &= \{j\in\mathcal{Z}\mid \text{Step 3a is executed for the } j\text{-th subset at iteration } k\} \\
\mathcal{Z}_{k,b} &= \{j\in\mathcal{Z}\mid \text{Step 3b is executed for the } j\text{-th subset at iteration } k\},
\end{aligned}
$$

we obtain that

$$\begin{aligned} \rho_{k+1} &\leq \alpha_{\min} \sum_{j\in\mathcal{Z}_{k,b}} \mu_{k,j} \|(\lambda_k - \lambda_*)_{[Q_j]}\| + \sum_{j\in\mathcal{Z}_{k,a}} \mu_{k,j}(\kappa_6 \eta_k + \kappa_7 \rho_k) \\ &\leq (\alpha_{\min} + \kappa_7 q \mu_{k,\mathcal{Z}})\rho_k + \kappa_6 q \mu_{k,\mathcal{Z}} \eta_k. \end{aligned} \tag{5.42}$$

For the purpose of obtaining a contradiction, assume now that

$$\rho_k \geq \frac{1}{2}\eta_k \tag{5.43}$$

for all $k \geq k_7$. Then (**??**) gives that, for all $k \geq k_7$,

$$\frac{\rho_{k+1}}{\rho_k} \leq \alpha_{\min} + \kappa_7 q \mu_{k,\mathcal{Z}} + 2\kappa_6 q \mu_{k,\mathcal{Z}} \leq \alpha_{\min}^{\beta_\eta+\epsilon} < 1 \tag{5.44}$$

because of (**??**). Hence we obtain from (**??**) that

$$\rho_{k+1} \leq \rho_{k_7} \alpha_{\min}^{(k-k_7+1)(\beta_\eta+\epsilon)}.$$

Therefore, since $\rho_{k_7}\alpha_{\min}^{(k-k_7+1)\epsilon}$ tends to zero, we obtain that

$$\rho_{k+1} < \frac{1}{2}\alpha_{\min}^{\alpha_\eta+(k_7-k_{\max})\beta_\eta}\alpha_{\min}^{(k-k_7+1)\beta_\eta} = \frac{1}{2}\alpha_{\min}^{\alpha_\eta+(k-k_{\max}+1)\beta_\eta} = \frac{1}{2}\eta_{k+1}$$

for all sufficiently large k, where the last equality results from the definition of $k_{\max}$ and (**??**). But this contradicts (**??**), which implies that (**??**) does not hold for all k sufficiently large. As a consequence, there exists a subsequence $\mathcal{K}$ such that

$$\rho_k < \frac{1}{2}\eta_k \tag{5.45}$$

for all $k \in \mathcal{K}$. Consider such a k. Then, using (**??**) and (**??**), we deduce that

$$\rho_{k+1} < \frac{1}{2}\eta_k(\alpha_{\min} + q\kappa_7\mu_{k,\mathcal{Z}} + 2q\kappa_6\mu_{k,\mathcal{Z}}) \leq \frac{1}{2}\alpha_{\min}^{\beta_\eta+\epsilon}\eta_k \leq \frac{1}{2}\eta_{k+1},$$

where we have used (**??**) to obtain the second inequality. As a consequence, $k+1 \in \mathcal{K}$ and (**??**) holds for all k sufficiently large. Returning to subset $j \in \mathcal{Z}$, we now obtain from (**??**) and (**??**) that

$$\|c(x_k)_{[\mathcal{Q}_j]}\| \leq \eta_k(\kappa_6\mu_{k,\mathcal{Z}} + \frac{1}{2}(1+\kappa_7\mu_{k,\mathcal{Z}})) \leq \eta_k,$$

for all k sufficiently large, because of (**??**). Hence Step 3a is executed for the subset j and for all sufficiently large k, which implies that j does not belong to $\mathcal{Z}$. Therefore $\mathcal{Z}$ is empty and the proof of the theorem is completed. □

As in **?** (**?**), we examine the rate of convergence of our algorithms.

Theorem 5.7 *Under the assumptions of Theorem* **??**, *the iterates* x_k *and the Lagrange multipliers* $\bar{\lambda}_k$ *of Algorithm 3.1 are at least R-linearly convergent with R-factor at most* $\alpha_{\min}^{\beta_\eta}$, *where* $\alpha_{\min}$ *is the smallest value of the maximum penalty parameter generated by the algorithm.*

Proof. The proof parallels that of Lemma **??**. First, Theorem **??** shows that the maximum penalty parameter α_k stays bounded away from zero, and thus remains fixed at some value $\alpha_{\min} > 0$, for $k \geq k_{\max}$. For all subsequent iterations,

$$\omega_{k+1} = \alpha_{\min}\omega_k \text{ and } \eta_{k+1} = \alpha_{\min}^{\beta_\eta}\eta_k \tag{5.46}$$

hold. Moreover, Theorem **??** implies that, for all $j = 1, \ldots, q$, (**??**) hold for all $k \geq k_{\max}$ sufficiently large. Hence and because of (**??**), the bound on the right-hand side of (**??**) may be replaced by $\kappa_{15}\omega_k + q\eta_k$, and thus

$$\|Z^T\delta_k\| \leq M[\kappa_{15}\omega_k + q\eta_k + \kappa_{12}\|Z^T\delta_k\|^2 + \kappa_{13}\|Z^T\delta_k\|\omega_k + \kappa_{14}\omega_k^2]. \tag{5.47}$$

Therefore, if k is sufficiently large that

$$\omega_k \leq \frac{1}{2M\kappa_{13}} \tag{5.48}$$

and

$$\|Z^T\delta_k\| \leq \frac{1}{4M\kappa_{12}}, \tag{5.49}$$

inequalities (**??**)–(**??**) can be rearranged to yield

$$\|Z^T\delta_k\| \leq 4M((\kappa_{14} + \kappa_{15})\omega_k + q\eta_k).$$

But then (**??**) gives that

$$\|\delta_k\| \leq \kappa_{19}\omega_k + \kappa_{20}\eta_k \tag{5.50}$$

where $\kappa_{19} = \kappa_1 + 4M(\kappa_{14} + \kappa_{15})$ and $\kappa_{20} = 4Mq$. As $\beta_\eta < 1$ and $\alpha_{\min} < 1$, (**??**) and (**??**) show that x_k converges to x_* at least R-linearly, with R-factor $\alpha_{\min}^{\beta_\eta}$. Inequalities (**??**) and (**??**) then guarantee the same property for $\bar{\lambda}_k$. $\square$

To conclude this section, we note that the conclusions of Theorems **??**, **??** and **??** require that the complete sequence of iterates converges to a unique limit point. As indicated above, this assumption cannot be relaxed. The counterexample presented by **?** (**?**) (where the linear inequality constraints are simple bound constraints on the problem's variables) shows that the sequence of penalty parameters may indeed converge to zero, if there is more than a single limit point.

6 Second order conditions

If we further strengthen the stopping test for the inner iteration beyond (**??**) to include second-order conditions, we can then guarantee that our algorithms converge to an isolated local solution. More specifically, we require the following additional assumption.

AS7: Suppose that x_k satisfies (**??**), converges to x_* for $k \in \mathcal{K}$, such that Z_* has a rank strictly greater than m. Then, if Z is defined as in AS6, we assume that $Z^T\nabla_{xx}\Phi_k Z$ is uniformly positive definite (that is, its smallest eigenvalue is uniformly bounded away from zero) for all $k \in \mathcal{K}$ sufficiently large.

We can then prove the following result.

Theorem 6.1 *Under assumptions AS1–AS3, AS5–AS7, the iterates x_k, $k \in \mathcal{K}$, generated by Algorithm 3.1 converge to an isolated local solution of (**??**)–(**??**).*

Proof. By definition of Φ,

$$\nabla_{xx}\Phi_k = H^\ell(x_k, \bar{\lambda}_k) + \sum_{j=1}^{q} \frac{1}{\mu_{k,j}} J_{\mathcal{Q}_j}(x_k)^T J_{\mathcal{Q}_j}(x_k), \tag{6.1}$$

where $J_{\mathcal{Q}_j}(x)$ is the Jacobian of $c(x)_{[\mathcal{Q}_j]}$. Note that the rank of Z is at least that of Z_*. AS7 then implies that there exists a nonzero vector s such that

$$J(x_k)Zs = 0$$

and hence

$$J_{\mathcal{Q}_j}(x_k)Zs = 0 \tag{6.2}$$

for each j. For any such vector, AS7 further implies that

$$s^T Z^T \nabla_{xx}\Phi_k Zs \geq \kappa_{21}\|s\|^2$$

for some $\kappa_{21} > 0$, which in turn gives that

$$s^T Z^T H^\ell(x_k, \bar{\lambda}_k)Zs \geq \kappa_{21}\|s\|^2,$$

because of (**??**) and (**??**). By continuity of H^ℓ as x_k and $\bar{\lambda}_k$ approach their limits, this ensures that

$$s^T Z^T H^\ell(x_*, \lambda_*)Zs \geq \kappa_{21}\|s\|^2$$

for all nonzero s satisfying

$$J(x_*)Zs = 0,$$

which implies that x_* is an isolated local solution of (**??**)–(**??**) (see, for instance, **?** (**?**, Thm. 3.11). □

If we assume that the inner iteration stopping test is tightened so that $\nabla_{xx}\Phi_k$ is required to be uniformly positive definite in the null space of the dominant constraints, and if we assume that the non-degeneracy condition (**??**) holds, then Corollary **??** ensures that $Z_k = Z = Z_*$ for sufficiently large k and Theorem **??** holds. A weaker version of this result also holds, where only positive semi-definiteness of the augmented Lagrangian's Hessian is required, yielding then that x_* is a (possibly not isolated) minimizer of the problem.

7 Extensions

7.1 Flexible Lagrange multiplier updates

The formula (**??**) has definite advantages for large-scale computations, but may otherwise appear unduly restrictive. The purpose of the first extension we consider is to introduce more freedom in our algorithmic framework, by replacing this formula by a more general condition, allowing a much larger class of Lagrange multiplier updates to be used. More specifically, we consider modifying Algorithm 3.1 as follows.

Algorithm 7.1

This algorithm is identical to Algorithm 3.1, except that Step 3 is replaced by the following, where γ is a constant in $(0,1)$.

Step 3 [Disaggregated updates]. Compute a new vector of Lagrange multiplier estimates $\hat{\lambda}_{k+1}$. For $j = 1, \ldots, q$, execute Step 3a if

$$\|c(x_k)_{[\mathcal{Q}_j]}\| \leq \eta_k,$$

or Step 3b otherwise.

Step 3a [Update Lagrange multiplier estimates]. Set

$$\begin{aligned} \lambda_{k+1,[\mathcal{Q}_j]} &= \begin{cases} \hat{\lambda}_{k+1,[\mathcal{Q}_j]} & \text{if } \|\hat{\lambda}_{k+1,[\mathcal{Q}_j]}\| \leq \mu_{k+1,j}^{-\gamma}, \\ \lambda_{k,[\mathcal{Q}_j]} & \text{otherwise,} \end{cases} \\ \mu_{k+1,j} &= \mu_{k,j}. \end{aligned}$$

Step 3b [Reduce the penalty parameter]. Set

$$\begin{aligned} \lambda_{k+1,[\mathcal{Q}_j]} &= \begin{cases} \hat{\lambda}_{k+1,[\mathcal{Q}_j]} & \text{if } \|\hat{\lambda}_{k+1,[\mathcal{Q}_j]}\| \leq \mu_{k+1,j}^{-\gamma}, \\ \lambda_{k,[\mathcal{Q}_j]} & \text{otherwise,} \end{cases} \\ \mu_{k+1,j} &= \tau_{k,j}\mu_{k,j}, \end{aligned}$$

where $\tau_{k,j}$ is defined by (**??**) in Algorithm 3.1.

Algorithm 7.1 allows a more flexible choice of the multipliers than Algorithm 3.1, but requires that some control is enforced to prevent their growth at an unacceptably fast rate. It covers, among others, the choice of the least-squares estimates $\lambda(x)$ as defined in (**??**).

The global convergence theory presented in Section **??** for Algorithm 3.1 can be extended to cover Algorithm 7.1. This extension is detailed in **?** (**?**). **?** (**?**) extend the local convergence analysis of Section **??** to Algorithm 7.1, under the additional condition that

$$\|\hat{\lambda}_{k+1} - \lambda_*\| \leq \kappa_{22}\|x_k - x_*\| + \kappa_{23}\omega_k,$$

holds for some positive constants κ_{22} and κ_{23} and all $k \in \mathcal{K}$ sufficiently large, where $\mathcal{K}$ is the index set of a subsequence of iterates (generated by Algorithm 7.1) converging to the critical point x_* with corresponding Lagrange multipliers λ_*. Both (**??**) and (**??**) satisfy this condition because of Theorem **??**.

We also note that Corollary **??** ensures that the least-squares multiplier estimates (**??**) are implementable when the non-degeneracy condition (**??**) holds. By this we mean that the estimates

$$\hat{\lambda}_k = -((J_k Z_k)^+)^T Z_k^T g_k$$

are identical to those defined in (**??**) for all k sufficiently large, and, unlike (**??**), are well defined when x_* is unknown.

7.2 Alternative criticality measures

In Algorithms 3.1 and 7.1 we used the criticality measure $\|P_{T_k}(-\nabla_x \Phi_k)\|$ in order to define the stopping criterion of the inner iteration (see (**??**)), because it is general. However, this quantity might not be easily computed in the course of the numerical method used to calculate x_k, especially when the dimension of the problem is high. It is therefore of interest to examine other criticality measures that might be easier to calculate. It is the purpose of this section to analyze such alternative proposals.

Given D_k, N_k, and A_{D_k} as above, we first claim that (**??**) can be replaced by the requirement that there exists a set of non-positive "dominant multipliers" $\{\xi_{ik}\}_{i \in M_k}$ ($M_k \subseteq D_k$, $\xi_{ik} \leq 0$) such that

$$\|\nabla_x \Phi_k + A_{D_k}^T \xi_k\| \leq \omega_k, \tag{7.1}$$

where ξ_k is the $|D_k|$-dimensional vector whose i-th component is ξ_{ik} if $i \in M_k$ or zero otherwise. We prove this claim.

Lemma 7.1 *Assume that there exists a non-positive ξ_k such that (**??**) holds at x_k. Then (**??**) also holds at x_k.*

Proof. Since the vector $A_{D_k}^T \xi_k$ belongs, by construction, to the cone N_k defined in (**??**), we can immediately deduce from the definition of the othogonal projection and (**??**) that

$$\|P_{T_k}(-\nabla_x \Phi_k)\| = \| - \nabla_x \Phi_k - P_{N_k}(-\nabla_x \Phi_k)\| \leq \| - \nabla_x \Phi_k - A_{D_k}^T \xi_k\| \leq \omega_k,$$

which is the desired inequality. □

Condition (**??**) is appealing for two reasons. Firstly, a set of (possibly approximate) multipliers is available in many numerical procedures that might be used to perform the inner iteration and to compute a suitable x_k; one can then select those multipliers which correspond to the dominant constraints, further restrict this choice to the non-positive ones and finally check (**??**). Such a scheme is implicitly used by both the **?** (**?**) barrier-function quadratic programming codes `VE14` and `VE19` and the **?** (**?**) general linearly constrained minimization package `LCONG`.

Alternatively, suitable multipliers can be computed, for instance by (approximately) solving the least-squares problem

$$\min_{\xi} \|\nabla_x \Phi_k + A_{D_k}^T \xi\|$$

and selecting the non-positive components of the resulting vector ξ, or by (approximately) solving the constrained least-squares problem

$$\min_{\xi \leq 0} \|\nabla_x \Phi_k + A_{D_k}^T \xi\|. \tag{7.2}$$

Condition (**??**) is also appealing as it provides, in a single condition, both a stopping condition on the inner iteration and a measure of the tolerated "inexactness" in solving the associated least-squares problem, if this is the procedure chosen to obtain the dominant multipliers.

We may therefore deduce from Lemma **??** that the convergence theory holds for Algorithms 3.1 and 7.1 whenever (**??**) is used instead of (**??**).

Condition (**??**) can be further specialized. For instance, one might choose to impose the familiar "reduced gradient" criterion

$$\|\hat{Z}(x_k)^T \nabla_x \Phi_k\| \leq \omega_k,$$

where $\hat{Z}(x_k)$ is an orthogonal matrix whose columns span the null space of the constraints active at x_k, provided that the multipliers associated with these linear constraints are all non-positive. In this case, we have that

$$\|P_{T_k}(-\nabla_x \Phi_k)\| \leq \|P_{T(x_k,0)}(-\nabla_x \Phi_k)\| = \|\hat{Z}(x_k)^T \nabla_x \Phi_k\| \leq \omega_k, \tag{7.3}$$

because $T(x_k, 0)$, the tangent cone to the set determined by the linear inequality constraints active at x_k, contains T_k. As a consequence, the convergence theory still holds when this criterion, which has been implemented by several subroutines for minimizing a general objective function subject to linear constraints (for example, the **?**, **?**, quadratic programming code `E04NFF` and the more general package `E04UCF`), is used as an inner-iteration stopping rule within Algorithms 3.1 and 7.1. This is also true for reduced gradient methods (e.g. MINOS **?**, **?**, or LSNNO by **?**, **?**) which compute a full column rank matrix $\check{Z}(x_k)$ whose columns are generally non-orthonormal but depend upon a subset of the (finite number) of coefficients for the linear constraints. Indeed, the norm of $\check{Z}(x_k)$ is then bounded above and away from zero, and a relationship that is a weighted form of (**??**) thus also holds in these cases.

In order to preserve coherence with the framework presented in **?** (**?**), we finally note that σ_k as defined in (**??**) may also be viewed as a criticality measure. Hence we might decide to stop the inner iteration when

$$\sigma_k \leq \omega_k. \tag{7.4}$$

The reader is referred to **?** (**?**) for a proof that global convergence is still obtained for this modification of Algorithms 3.1 and 7.1. However, the authors have not been able to prove the desired local convergence properties with only (**??**). Instead, the local convergence theory is covered for Algorithms 3.1 and 7.1 for the stronger condition

$$\sigma_k \leq \omega_k^2 \tag{7.5}$$

(see **?**, **?** for details). This condition is theoretically interesting, but might be practically too strong. Note, as we now show, that it implies a variant of (**??**).

Theorem 7.2 *Assume that $\{x_k\}$, $k \in \mathcal{K}$, is a convergent subsequence of vectors of $\mathcal{B}$ such that (**??**) holds for each $k \in \mathcal{K}$, where the ω_k converge to zero as k increases in $\mathcal{K}$. Then the inequality*

$$\|P_{T_k}(-\nabla_x \Phi_k)\| \leq \kappa_{24} \omega_k \tag{7.6}$$

also holds for each $k \in \mathcal{K}$ sufficiently large and for some $\kappa_{24} \geq 1$.

Proof. We first consider the simple case where $p = 0$, that is when no linear inequality is present. In this case, it is easy to check from (**??**) that $\sigma_k = \|\nabla_x \Phi_k\|$. But we must have that

$D_k = \emptyset$. Thus $\sigma_k = \|P_{T_k}(-\nabla_x \Phi_k)\|$. We therefore obtain that (**??**) holds with $\kappa_{24} = 1$ and k large enough to ensure that $\omega_k \leq 1$.

Assume now that $p > 0$. The Moreau decomposition of $-\nabla_x \Phi_k$ (see **?**, **?**) is given by

$$-\nabla_x \Phi_k = P_{T_k}(-\nabla_x \Phi_k) + P_{N_k}(-\nabla_x \Phi_k).$$

If $P_{T_k}(-\nabla_x \Phi_k)$ is zero, then (**??**) obviously holds for any choice of κ_{24}. Assume therefore that $P_{T_k}(-\nabla_x \Phi_k)$ is nonzero. We now show that $x_k + d_k \in \mathcal{B}$, where we define

$$d_k \stackrel{\text{def}}{=} \epsilon_k \frac{P_{T_k}(-\nabla_x \Phi_k)}{\|P_{T_k}(-\nabla_x \Phi_k)\|}, \quad \text{with} \quad \epsilon_k \stackrel{\text{def}}{=} \min\left[1, \frac{\kappa_0 \omega_k}{\|A\|_\infty}\right]. \tag{7.7}$$

Assume first that $i \in D_k$. Then $-a_i \in N_k$ and $a_i^T d_k \geq 0$ have rephrased the relevant paragraph and we hope that it is now clearer.have rephrased the relevant paragraph and we hope that it is now clearer.because of the polarity of N_k and T_k. Since $x_k \in \mathcal{B}$, we obtain that

$$a_i^T(x_k + d_k) - b_i = (a_i^T x_k - b_i) + a_i^T d_k \geq 0. \tag{7.8}$$

On the other hand, if $i \notin D_k$, we have that $a_i^T x_k - b_i > \kappa_0 \omega_k$ and hence

$$(a_i^T x_k - b_i) + a_i^T d_k > \kappa_0 \omega_k - \|a_i\|.\|d_k\| = \kappa_0 \omega_k - \epsilon_k \|a_i\| \geq \kappa_0 \omega_k - \kappa_0 \omega_k = 0. \tag{7.9}$$

Gathering (**??**) and (**??**), we obtain that $x_k + d_k \in \mathcal{B}$, as desired. Furthermore, since $\|d_k\| \leq 1$ by definition, we have verified that d_k is feasible for the minimization problem (**??**) associated with the definition of σ_k. Hence,

$$\begin{array}{rcl} \sigma_k & \geq & -\nabla_x \Phi_k{}^T d_k \\ & = & P_{T_k}(-\nabla_x \Phi_k)^T d_k + P_{N_k}(-\nabla_x \Phi_k)^T d_k \\ & = & \|P_{T_k}(-\nabla_x \Phi_k)\| \, \|d_k\| \\ & = & \epsilon_k \|P_{T_k}(-\nabla_x \Phi_k)\|, \end{array} \tag{7.10}$$

where we have used successively the Moreau decomposition of $-\nabla_x \Phi_k$, the definition of d_k and the orthogonality of the terms in the Moreau decomposition. If $\epsilon_k = 1$, then (**??**) and (**??**) imply that

$$\|P_{T_k}(-\nabla_x \Phi_k)\| \leq \omega_k{}^2 \leq \omega_k \tag{7.11}$$

for $k \in \mathcal{K}$ sufficiently large. Otherwise, we deduce from (**??**), (**??**) and (**??**) that

$$\|P_{T_k}(-\nabla_x \Phi_k)\| \leq \frac{\|A\|_\infty}{\kappa_0} \omega_k. \tag{7.12}$$

As a consequence of (**??**) and (**??**), we therefore obtain that (**??**) holds with

$$\kappa_{24} = \max\left[1, \frac{\|A\|_\infty}{\kappa_0}\right].$$

Combining all cases, we conclude that (**??**) holds with this last value of κ_{24}. □

We finally note that Lemma **??** and Theorem **??** do not depend on the actual form of the augmented Lagrangian (**??**), but are valid independently of the function minimized in the inner iteration. This observation could be useful if alternative techniques for augmenting the Lagrangian are considered for a merit function.

8 Conclusion

We have considered a class of augmented Lagrangian algorithms for constrained nonlinear optimization, where the linear constraints present in the problem are handled directly and where multiple penalty parameters are allowed. The algorithms in this class have the advantage that efficient techniques for handling linear constraints may be used at the inner iteration level, and also that the sparsity pattern of the Hessian of the augmented Lagrangian is independent of that of the linear constraints. The global and local convergence results available for the specific case where linear constraints reduce to simple bounds have been extended to the more general and useful context where any form of linear constraint is permitted.

We finally note that the theory presented is directly relevant to practical computation, as the inner iteration stopping rule (**??**) covers the type of optimality tests used in available packages for linearly constrained problems. This means that these packages can be applied to obtain an (approximate) solution of the subproblem, and constitutes a realistic and attractive algorithmic development.

It is now the author's intention to perform extensive numerical experiments on large-scale problems. This development requires considerable care and sophistication if an efficient solver for the subproblem is to be integrated with the class of algorithms described here.

Acknowledgements

The authors wish to acknowledge funding provided by a NATO travel grant. They are also grateful to J. Nocedal and the anonymous referees for their constructive comments.

References

M. Arioli, T.F. Chan, I.S. Duff, N.I.M. Gould, and J.K. Reid. Computing a search direction for large-scale linearly constrained nonlinear optimization calculations. Technical Report TR/PA/93/34, CERFACS, Toulouse, France, 1993.

M. Avriel. *Nonlinear Programming: Analysis and Methods.* Prentice-Hall, Englewood Cliffs, N.J., 1976.

D. P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods.* Academic Press, London, 1982.

J. V. Burke, J. J. Moré, and G. Toraldo. Convergence properties of trust region methods for linear and convex constraints. *Mathematical Programming, Series A*, **47**(3), 305–336, 1990.

A. R. Conn, N. I. M. Gould, and Ph. L. Toint. A globally convergent augmented Lagrangian algorithm for optimization with general constraints and simple bounds. *SIAM Journal on Numerical Analysis*, **28**(2), 545–572, 1991.

A. R. Conn, N. I. M. Gould, and Ph. L. Toint. LANCELOT*: a Fortran package for large-scale nonlinear optimization (Release A).* Number 17 *in* 'Springer Series in Computational Mathematics'. Springer Verlag, Heidelberg, Berlin, New York, 1992*a*.

A. R. Conn, N. I. M. Gould, and Ph. L. Toint. On the number of inner iterations per outer iteration of a globally convergent algorithm for optimization with general nonlinear equality constraints and simple bounds. *in* D. Griffiths and G. Watson, eds, 'Proceedings of the 14th Biennial Numerical Analysis Conference Dundee 1991', pp. 49–68. Longmans, 1992*b*.

A. R. Conn, Nick Gould, A. Sartenaer, and Ph. L. Toint. Global convergence of a class of trust region algorithms for optimization using inexact projections on convex constraints. *SIAM Journal on Optimization*, **3**(1), 164–221, 1993*a*.

A. R. Conn, Nick Gould, A. Sartenaer, and Ph. L. Toint. Global convergence of two augmented Lagrangian algorithms for optimization with a combination of general equality and linear constraints. Technical Report 93/01, Department of Mathematics, FUNDP, Namur, Belgium, 1993*b*.

A. R. Conn, Nick Gould, A. Sartenaer, and Ph. L. Toint. Local convergence properties of two augmented Lagrangian algorithms for optimization with a combination of general equality and linear constraints. Technical Report 93/20, Department of Mathematics, FUNDP, Namur, Belgium, 1993*c*.

J. C. Dunn. On the convergence of projected gradient processes to singular critical points. *Journal of Optimization Theory and Applications*, **55**, 203–216, 1987.

A. V. Fiacco. *Introduction to sensitivity and stability analysis in nonlinear programming.* Academic Press, New York, 1983.

R. Fletcher. *Practical Methods of Optimization.* J. Wiley and Sons, Chichester, second edn, 1987.

A. L. Forsgren and W. Murray. Newton methods for large-scale linear equality-constrained minimization. *SIAM Journal on Matrix Analysis and Applications*, **14**(2), 560–587, 1993.

N. I. M. Gould. On the convergence of a sequential penalty function method for constrained minimization. *SIAM Journal on Numerical Analysis*, **26**, 107–128, 1989.

W. A. Gruver and E. Sachs. *Algorithmic Methods in Optimal Control.* Pitman, Boston, USA, 1980.

Harwell Subroutine Library. *A catalogue of subroutines (release 11).* Advanced Computing Department, Harwell Laboratory, Harwell, UK, 1993.

M. R. Hestenes. Multiplier and gradient methods. *Journal of Optimization Theory and Applications*, **4**, 303–320, 1969.

IMSL. *MATH/LIBRARY User's Manual, Vol 1-3.* IMSL Inc., Houston, Texas, USA, 1987.

J. J. Moreau. Décomposition orthogonale d'un espace hilbertien selon deux cônes mutuellement polaires. *Comptes-Rendus de l'Académie des Sciences (Paris)*, **255**, 238–240, 1962.

B. A. Murtagh and M. A. Saunders. Large-scale linearly constrained optimization. *Mathematical Programming*, **14**, 41–72, 1978.

NAG. *Fortran Manual, Mark 16.* NAG Ltd, Oxford, UK, 1993.

M. J. D. Powell. A method for nonlinear constraints in minimization problems. *in* R. Fletcher, ed., 'Optimization', London and New York, 1969. Academic Press.

Ph. L. Toint and D. Tuyttens. LSNNO: a Fortran subroutine for solving large scale nonlinear network optimization problems. *ACM Transactions on Mathematical Software*, **18**(3), 308–328, 1992.

R. J. Vanderbei and T. J. Carpenter. Indefinite systems for interior point methods. *Mathematical Programming*, **58**(1), 1–32, 1993.