Global convergence of a class of trust region algorithms for optimization using inexact projections on convex constraints

by A.R. Conn[1], N.I.M. Gould[2], A. Sartenaer[3]
and Ph.L. Toint[4]

September 12, 1995

Abstract. A class of trust region based algorithms is presented for the solution of nonlinear optimization problems with a convex feasible set. At variance with previously published analysis of this type, the theory presented allows for the use of general norms. Furthermore, the proposed algorithms do not require the explicit computation of the projected gradient, and can therefore be adapted to cases where the projection onto the feasible domain may be expensive to calculate. Strong global convergence results are derived for the class. It is also shown that the set of linear and nonlinear constraints that are binding at the solution are identified by the algorithms of the class in a finite number of iterations.

[1] IBM T.J. Watson Research Center,
Yorktown Heights, USA
[2] Rutherford Appleton Laboratory,
Chilton, Oxfordshire, England
[3] Belgian National Fund for Scientific Research, Facultés Universitaires ND de la Paix, Namur, Belgium
[4] Department of Mathematics, Facultés Universitaires ND de la Paix, Namur, Belgium

Keywords : Trust region methods, projected gradients, convex constraints.

GLOBAL CONVERGENCE OF A CLASS OF TRUST REGION ALGORITHMS FOR OPTIMIZATION USING INEXACT PROJECTIONS ON CONVEX CONSTRAINTS

by A.R. Conn[1], N.I.M. Gould[2], A. Sartenaer[3] and Ph.L. Toint[4]

Report 90/4 September 12, 1995

Abstract. A class of trust region based algorithms is presented for the solution of nonlinear optimization problems with a convex feasible set. At variance with previously published analysis of this type, the theory presented allows for the use of general norms. Furthermore, the proposed algorithms do not require the explicit computation of the projected gradient, and can therefore be adapted to cases where the projection onto the feasible domain may be expensive to calculate. Strong global convergence results are derived for the class. It is also shown that the set of linear and nonlinear constraints that are binding at the solution are identified by the algorithms of the class in a finite number of iterations.

[1] IBM T.J. Watson Research Center, Yorktown Heights, USA

[2] Rutherford Appleton Laboratory, Chilton, Oxfordshire, England

[3] Belgian National Fund for Scientific Research, Facultés Universitaires ND de la Paix, Namur, Belgium

[4] Department of Mathematics, Facultés Universitaires ND de la Paix, Namur, Belgium

Keywords : Trust region methods, projected gradients, convex constraints.

1 Introduction

Trust region methods for nonlinear optimization problems have become very popular over the last decade. One possible explanation of their success is their remarkable numerical reliability associated with the existence of a sound and complete convergence theory. The fact that they efficiently handle nonconvex problems has also been considered as an advantage.

As an integral part of this growing interest, research in convergence theory for this class of methods has been very active. First, a substantial body of theory was built for the unconstrained case (see [19] for an excellent survey). Problems involving bound constraints on the variables were then considered (see [1], [9] and [21]), as well as the more general case where the feasible region is a convex set on which the projection (with respect to the Euclidean norm) can be computed at a reasonable cost (see [4], [20] and [29]). The studied techniques are based on the use of the explicitly calculated projected gradient as a tool to predict which of the inequality constraints are binding at the problem's solution. Moreover, trust region methods for nonlinear equality constraints have also been studied by several authors (see [5], [8], [25] and [30], for instance).

This paper also considers the case where the feasible set is convex. It presents a convergence theory for a class of trust region algorithms with the following new features.

- The theory does not depend on the explicit use of the projection operator in the Euclidean norm, but allows for the use of a uniformly equivalent family of arbitrary norms.
- The gradient of the objective function can be approximated if its exact value is either impossible or too costly to compute at every iteration.
- The calculation of the "projected gradient" (with respect to the chosen norms) need not be carried out to full accuracy.
- When the feasible set is described by a system of linear and/or nonlinear (in)equalities, conditions are presented that guarantee that the algorithms of the class identify, in a finite number of iterations, the set of inequalities that are binding at the solution. We note that this description of the feasible set does not need its partition into faces.

In this sense, we see that our theory applies to problems similar to those considered in [4], [9], [20] and [29], although in a more general setting.

An attractive aspect of this theory is that it covers the case where a polyhedral norm is chosen to define an analog of the projection operator, allowing the use of linear (or convex) programming methods for the approximate calculation of the projected gradients. This type of algorithm should be especially efficient in the frequent situation where the feasible set is defined by a set of linear equalities and inequalities, and where a basis for the nullspace of the matrix of the active constraints is cheaply available. In network problems, for example, this can be very cheaply obtained and updated using a spanning tree of the problem's underlying graph (see [17] for a detailed presentation of the relevant algorithms). Other examples include multiperiodic operation research models resulting in staircase matrices.

The problem and notation are introduced in Section 2, together with a general class of algorithms. The convergence properties of this class are then analyzed in Section 3. A particular practical algorithm of the class is discussed in Section 4. The identification of the active constraints is presented in Section 5. Section 6 presents an analysis of the conditions under which the whole sequence of iterates can be shown to converge to a single limit point. Additional points and extensions of the theory are discussed in Section 7. A glossary of symbols can be found in Appendix B. All the assumptions used in the paper are finally summarized in Appendix C.

2 A class of trust region algorithms for problems with convex feasible domain

2.1 The problem

The problem we consider is that of finding a local solution of

$$\min f(x) \tag{2.1}$$

subject to the constraint

$$x \in X, \tag{2.2}$$

where x is a vector of $\mathbf{R}^n$, $f(\cdot)$ is a smooth function from $\mathbf{R}^n$ into $\mathbf{R}$ and X is a non-empty closed convex subset of $\mathbf{R}^n$, also called the *feasible set*. We assume that we can compute the function value $f(x)$ for any feasible point x. We are also given a feasible starting point x_0 and we wish to start the minimization procedure from this point.

If we define $\mathcal{L}$ by

$$\mathcal{L} \stackrel{\text{def}}{=} X \cap \{x \in \mathbf{R}^n \mid f(x) \leq f(x_0)\}, \tag{2.3}$$

we may formulate our assumptions on the problem as follows.

AS.1 The set $\mathcal{L}$ is compact.

AS.2 The objective function $f(x)$ is continuously differentiable and its gradient $\nabla f(x)$ is Lipschitz continuous in an open domain containing $\mathcal{L}$.

In particular, we allow for unbounded X, provided the set $\mathcal{L}$ remains bounded.

We will denote by $\langle \cdot, \cdot \rangle$ the Euclidean inner product on $\mathbf{R}^n$ and by $\|\cdot\|_2$ the associated ℓ_2-norm.

We recall that a subset K of $\mathbf{R}^n$ is a cone if it is closed under positive scalar multiplication, that is if $\lambda x \in K$ whenever $x \in K$ and $\lambda > 0$ (see [26, p. 13]). Given a cone K, one can define its *polar* (see [26, p. 121]) as

$$K^0 \stackrel{\text{def}}{=} \{y \in \mathbf{R}^n | \langle y, u \rangle \leq 0, \forall u \in K\} \tag{2.4}$$

and verify that K^0 is also a cone, and that $(K^0)^0 = K$ when K is a non-empty closed convex cone.

Given the convex set X, we can define $P_X(x)$, the *projection* of the vector $x \in \mathbf{R}^n$ onto X, as the unique minimizer of the problem

$$\min_{y \in X} \|y - x\|_2. \tag{2.5}$$

This projection operator is well known and has been much studied (see [33] for instance). We will also denote by $N(x)$ the *normal cone* of X at $x \in X$, that is

$$N(x) \stackrel{\text{def}}{=} \{y \in \mathbf{R}^n \mid \langle y, u - x \rangle \leq 0, \ \forall u \in X\}. \tag{2.6}$$

The *tangent cone* of X at $x \in X$ is the polar of the normal cone at the same point, that is

$$T(x) \stackrel{\text{def}}{=} N(x)^0 = \text{cl}\{\lambda(u - x) | \lambda \geq 0 \ \text{ and } \ u \in X\}, \tag{2.7}$$

where $\text{cl}\{S\}$ denotes the closure of the set S. We will also use the *Moreau decomposition* given by the identity

$$x = P_{T(y)}(x) + P_{N(y)}(x), \tag{2.8}$$

which is valid for all $x \in \mathbf{R}^n$ and all $y \in X$ (see [22]). This decomposition is illustrated in Figure 1. In this figure and all subsequent ones, the boundary of the feasible set X is drawn with a bold line.

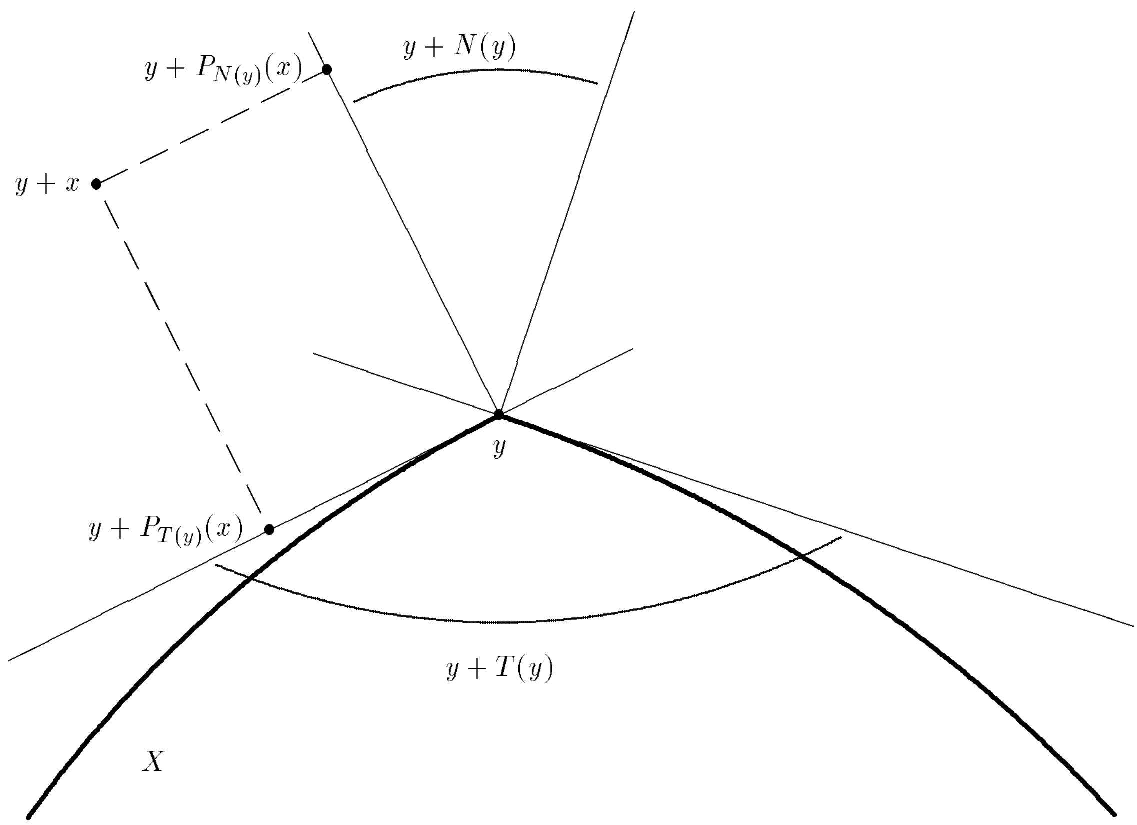

Figure 1: The normal and tangent cones at y, and the corresponding Moreau decomposition of x (translated to y)

We conclude this subsection with a result extracted from the classical perturbation theory of convex optimization problems. This result is well known and can be found in [14, p. 14–17] for instance.

Lemma 1 *Assume that D is a continuous point-to-set mapping from $S \subseteq \mathbf{R}^\ell$ into $\mathbf{R}^n$ such that the set $D(\epsilon)$ is convex and non-empty for each $\epsilon \in S$. Assume also that one is given a real-valued function $F(y,\epsilon)$ which is defined and continuous on the space $\mathbf{R}^n \times S$ and convex in y for each fixed ϵ. Then, the real-valued function F_* defined by*

$$F_*(\epsilon) \stackrel{\rm def}{=} \inf_{y \in D(\epsilon)} F(y,\epsilon) \tag{2.9}$$

and the solution set mapping y_* defined by

$$y_*(\epsilon) \stackrel{\rm def}{=} \{y \in D(\epsilon) | F(y,\epsilon) = F_*(\epsilon)\} \tag{2.10}$$

are both continuous on S.

2.2 Defining a local model of the objective function

The algorithm we propose for solving (2.1) subject to the constraint (2.2) is iterative and of trust region type. Indeed, at each iteration, we define a *model* of the objective function $f(x)$, and a region surrounding the current iterate, x_k say, where we believe this model to be adequate. The algorithm then finds, in this region, a candidate for the next iterate that sufficiently reduces the value of the model of the objective. If the function value calculated at this point matches its predicted value closely enough, the new point is then accepted as the next iterate and the trust region is possibly enlarged; otherwise the point is rejected and the trust region size decreased. With each iteration of our algorithm will be associated a norm: we will denote by $\|\cdot\|_{(k)}$ the norm associated with the kth iteration.

We now specify the conditions we impose on the model of the objective function. This model, defined in a neighbourhood of the kth iterate x_k, will be denoted by the symbol m_k and is meant to approximate the objective f in the *trust region*

$$B_k \stackrel{\rm def}{=} \{x \in \mathbf{R}^n | \, \|x - x_k\|_{(k)} \le \nu_1 \Delta_k\}, \tag{2.11}$$

where ν_1 is a positive constant and $\Delta_k > 0$ is the *trust region radius.* We will assume that m_k is differentiable and has Lipschitz continuous first derivatives in an open set containing B_k, that

$$m_k(x_k) = f(x_k) \tag{2.12}$$

and that $g_k \stackrel{\rm def}{=} \nabla m_k(x_k)$ approximates $\nabla f(x_k)$ in the following sense: there exists a nonnegative constant κ_1 such that the inequality

$$\|e_k\|_{[k]} \le \kappa_1 \Delta_k \tag{2.13}$$

holds for all k, where the error e_k is defined by $e_k \stackrel{\rm def}{=} g_k - \nabla f(x_k)$ and where the norm $\|\cdot\|_{[k]}$ is any norm that satisfies

$$|\langle x, y\rangle| \le \|x\|_{(k)} \|y\|_{[k]} \tag{2.14}$$

for all $x, y \in \mathbf{R}^n$. In particular, one can choose the *dual norm* of $\|\cdot\|_{(k)}$ defined by

$$\|y\|_{[k]} \stackrel{\text{def}}{=} \sup_{x \neq 0} \frac{|\langle x, y\rangle|}{\|x\|_{(k)}}. \tag{2.15}$$

Condition (2.13) is quite weak, as it merely requires that the first order information on the objective function be reasonably accurate whenever a short step must be taken. Indeed, one expects this first order behaviour to dominate for small steps.

Clearly, for the above conditions to be coherent from one iteration to the next, we need to assume some relationship between the various norms that we introduced. More precisely, we will assume that all these norms are *uniformly equivalent* in the following sense.

AS.3 There exist constants $\sigma_1, \sigma_3 \in (0,1]$ and $\sigma_2, \sigma_4 \geq 1$ such that, for all $k_1 \geq 0$ and $k_2 \geq 0$,

$$\sigma_1 \|x\|_{(k_1)} \leq \|x\|_{(k_2)} \leq \sigma_2 \|x\|_{(k_1)} \tag{2.16}$$

and

$$\sigma_3 \|x\|_{[k_1]} \leq \|x\|_{[k_2]} \leq \sigma_4 \|x\|_{[k_1]} \tag{2.17}$$

for all $x \in \mathbf{R}^n$.

If (2.15) is chosen, then (2.17) immediately results from (2.16) with $\sigma_3 = 1/\sigma_2$ and $\sigma_4 = 1/\sigma_1$.

We also note that (2.16) and (2.17) necessarily hold if the norms $\|\cdot\|_{(k_2)}$ and $\|\cdot\|_{[k_2]}$ are replaced by the ℓ_2-norm.

We finally introduce, for given k and for any nonnegative t, the quantity $\alpha_k(t) \geq 0$ given by

$$\alpha_k(t) \stackrel{\text{def}}{=} \Big| \min_{\substack{x_k + d \in X \\ \|d\|_{(k)} \leq t}} \langle g_k, d\rangle \Big|, \tag{2.18}$$

that is the magnitude of the maximum decrease of the linearized model achievable on the intersection of the feasible domain with a ball of radius t (in the norm $\|\cdot\|_{(k)}$) centered at x_k.

We note here that $\alpha_k(t)$ can be defined using the notion of support function of the convex set $\{d | x_k + d \in X \ \text{ and } \ \|d\|_{(k)} \leq t\}$. The properties that follow can then be derived in this framework. We have however chosen to use the more familiar vocabulary of classical optimization in order to avoid further prerequisites in convex analysis.

We then have the following simple properties.

Lemma 2 *For all $k \geq 0$,*

1. *the function $t \mapsto \alpha_k(t)$ is continuous and nondecreasing for $t \geq 0$,*

2. *the function $t \mapsto \frac{\alpha_k(t)}{t}$ is nonincreasing for $t > 0$,*

3. *the inequality*

$$\frac{\alpha_k(t)}{t} \leq \|P_{T(x_k)}(-g_k)\|_{[k]} \tag{2.19}$$

holds for all $t > 0$.

Proof. The first statement is an immediate consequence of the definition (2.18) and of Lemma 1 applied on the optimization problem of (2.18). In order to prove the second statement, consider $0 < t_1 < t_2$ and two vectors d_1 and d_2 such that

$$\alpha_k(t_1) = -\langle g_k, d_1 \rangle, \quad \|d_1\|_{(k)} \le t_1, \quad x_k + d_1 \in X, \tag{2.20}$$

and

$$\alpha_k(t_2) = -\langle g_k, d_2 \rangle, \quad \|d_2\|_{(k)} \le t_2, \quad x_k + d_2 \in X. \tag{2.21}$$

We observe that the point $x_k + (t_1/t_2)d_2$ lies between x_k and $x_k + d_2$, and therefore we have that $x_k + (t_1/t_2)d_2 \in X$. Furthermore,

$$\|\frac{t_1}{t_2} d_2\|_{(k)} = \frac{t_1}{t_2}\|d_2\|_{(k)} \le t_1 \tag{2.22}$$

and the point $x_k+(t_1/t_2)d_2$ thus lies in the feasible domain of the optimization problem associated with the definition of $\alpha_k(t_1)$ and d_1. As a consequence, we have that

$$\frac{\alpha_k(t_1)}{t_1} \ge \frac{1}{t_1}|\langle g_k, \frac{t_1}{t_2} d_2 \rangle| = \frac{\alpha_k(t_2)}{t_2}, \tag{2.23}$$

and the second statement of the lemma is proved.

The third statement is proved as follows. Applying the Moreau decomposition to $-g_k$, we obtain that, for any d such that $x_k + d \in X$ and $\langle g_k, d \rangle \le 0$,

$$\langle g_k, d \rangle = -\langle P_{T(x_k)}(-g_k), d \rangle - \langle P_{N(x_k)}(-g_k), P_{T(x_k)} d \rangle \ge -\langle P_{T(x_k)}(-g_k), d \rangle, \tag{2.24}$$

where we used the fact that $d \in T(x_k)$ and the fact that the tangent cone is the polar of the normal cone to derive the last inequality. Taking absolute values and applying (2.14) thus yields that

$$|\langle g_k, d \rangle| \le \|d\|_{(k)} \, \|P_{T(x_k)}(-g_k)\|_{[k]}. \tag{2.25}$$

We then obtain (2.19) by applying this inequality to any solution d of the optimization problem associated with the definition of $\alpha_k(t)$ in (2.18) and using the fact that $\|d\|_{(k)} \le t$. $\Box$

2.3 A class of trust region algorithms

We are now ready to define our first algorithm in more detail. Besides κ_1 as used in (2.13), it depends on the constants

$$0 < \mu_1 < \mu_2 < 1, \quad \mu_3 \in (0,1], \quad \mu_4 \in (0,1], \tag{2.26}$$

$$0 < \nu_3 < \nu_2 \le \nu_1, \quad \nu_4 \in (0,1], \tag{2.27}$$

$$0 < \eta_1 < \eta_2 < 1 \tag{2.28}$$

and

$$0 < \gamma_1 \le \gamma_2 < 1 < \gamma_3. \tag{2.29}$$

Algorithm 1

Step 0: initialization. The starting point x_0 is given, together with $f(x_0)$ and an initial trust region radius $\Delta_0 > 0$. Set $k = 0$.

Step 1: model choice. Choose m_k, a model of the objective function f in the trust region B_k centered at x_k, satisfying (2.12) and (2.13).

Step 2: determination of a Generalized Cauchy Point (GCP). If $\alpha_k \stackrel{\text{def}}{=} \alpha_k(1) = 0$, stop. Else, find a vector s_k^C such that, for some strictly positive $t_k \geq \|s_k^C\|_{(k)}$,

$$x_k + s_k^C \in X, \tag{2.30}$$

$$\|s_k^C\|_{(k)} \leq \nu_2 \Delta_k, \tag{2.31}$$

$$\langle g_k, s_k^C \rangle \leq -\mu_3 \alpha_k(t_k), \tag{2.32}$$

$$m_k(x_k + s_k^C) \leq m_k(x_k) + \mu_1 \langle g_k, s_k^C \rangle, \tag{2.33}$$

and, either

$$t_k \geq \min[\nu_3 \Delta_k, \nu_4] \tag{2.34}$$

or

$$m_k(x_k + s_k^C) \geq m_k(x_k) + \mu_2 \langle g_k, s_k^C \rangle. \tag{2.35}$$

Set the *Generalized Cauchy Point*

$$x_k^C = x_k + s_k^C. \tag{2.36}$$

Step 3: determination of the step. Find a vector s_k such that

$$x_k + s_k \in X \cap B_k \tag{2.37}$$

and

$$m_k(x_k) - m_k(x_k + s_k) \geq \mu_4 [m_k(x_k) - m_k(x_k^C)]. \tag{2.38}$$

Step 4: determination of the model accuracy. Compute $f(x_k + s_k)$ and

$$\rho_k = \frac{f(x_k) - f(x_k + s_k)}{m_k(x_k) - m_k(x_k + s_k)}. \tag{2.39}$$

Step 5: trust region radius updating. In the case where

$$\rho_k > \eta_1, \tag{2.40}$$

set

$$x_{k+1} = x_k + s_k \tag{2.41}$$

and

$$\Delta_{k+1} \in [\Delta_k, \gamma_3 \Delta_k], \text{ if } \rho_k \geq \eta_2, \tag{2.42}$$

or

$$\Delta_{k+1} \in [\gamma_2 \Delta_k, \Delta_k], \quad \text{if} \quad \rho_k < \eta_2. \tag{2.43}$$

Otherwise, set

$$x_{k+1} = x_k \tag{2.44}$$

and

$$\Delta_{k+1} \in [\gamma_1 \Delta_k, \gamma_2 \Delta_k]. \tag{2.45}$$

Step 6: loop. Increment k by one and go to Step 1.

Of course, this only describes a relatively abstract algorithmic class. In particular, we note the following:

1. We have not been very specific about the model m_k to be used in the trust region. In fact, we have merely stated that its value should coincide with that of the objective at the current iterate, and that its gradient at this point should approximate the gradient of the objective at the same point. We will also impose additional necessary assumptions on its curvature in order to derive the desired convergence results. This still remains very broad and requires further specification for any practical implementation of the algorithm.

   One very common model choice for a twice differentiable f is to use a quadratic of the form

   $$m_k(x_k + s) = f(x_k) + \langle \nabla f(x_k), s \rangle + \frac{1}{2} \langle s, H_k s \rangle, \tag{2.46}$$

   where H_k is a symmetric approximation to $\nabla^2 f(x_k)$. In particular, Newton's method corresponds to (2.46) with the choice of $H_k = \nabla^2 f(x_k)$.

   Another interesting choice is

   $$m_k(x_k + s) = f(x_k + s), \tag{2.47}$$

   that is the model and the objective are required to coincide on $X \cap B_k$. In that case, ρ_k will always be exactly one, and the trust region size Δ_k may be assumed to be very large. We then obtain a convergence theory of an algorithm which is no longer a trust region method in the classical sense. In particular, if the step s_k is determined by a linesearch procedure (see [1], [29]), the present theory then covers both linesearch and trust region algorithms in a single context.

2. When $k = 0$ or $x_k \neq x_{k-1}$ or $\Delta_k < \Delta_{k-1}$, the definition of the model m_k at Step 1 and the condition that (2.13) is satisfied may require the computation of a new sufficiently accurate approximate gradient g_k.

3. We now briefly motivate the conditions (2.30)–(2.35). Our main idea is to avoid the repeated computation of the projection onto the feasible set X within the GCP calculation, which is a convex *nonlinear* program. Instead, we allow the repeated solution of convex *linear* programs. Furthermore, these linear programs need not be solved to full accuracy. These two relaxations may indeed allow for a substantially reduced amount of calculation. We

have in mind the particular case where X is a polyhedral set and $\|\cdot\|_{(k)}$ is polyhedral for all k.

Condition (2.30) is imposed because we want our algorithm only to generate feasible points. This may be essential when some constraints are "hard", for instance when the objective function is undefined outside X.

Condition (2.31) simply requires the step to be inside a ball contained in the trust region defined by (2.11). This is intended to leave some freedom for the calculation of s_k in Step 3, even when the GCP is on the boundary of that smaller ball.

Condition (2.32) introduces the desired relaxations, while relating the definition of x_k^C to that of a point along the projected gradient path

$$x_k(\theta) = P_X(x_k - \theta g_k) \quad (\theta \geq 0). \tag{2.48}$$

Indeed, it can be shown that, if $\mu_3 = 1$ and $\|\cdot\|_{(k)} = \|\cdot\|_2$, then x_k^C achieves the same reduction in the linearized model as that obtained by the unique point $x_k(\theta_k)$ on the projected gradient path (2.48) having length t_k, if such a point exists. Condition (2.32) with $\mu_3 < 1$ can therefore be interpreted as a weakening of the condition (for example, required in [9], [21] and [29]) that x_k^C should be on the projected gradient path. This weakening is of great practical interest when the projection onto the feasible domain X is not readily computable.

An example is shown in Figure 2 using the ℓ_∞-norm, where the set of admissible steps s_k^C is represented by the shaded area, and where (2.32) with $\mu_3 = 1$ is achieved for the step $d_k(t_k)$.

Conditions (2.33) and (2.35) are in the spirit of the classical Goldstein conditions for a "projected search" on the model along the approximation of the projected gradient path implicitly defined by varying t_k. This projected search is similar to that introduced in [29] and modified in [20]. Condition (2.34) completes (2.33) and (2.35) by allowing the search to terminate with a point that sufficiently reduces the model m_k while having a length comparable to the trust region radius.

We note here that the value of t_k is never used by Algorithm 1 except in the definition of s_k^C. It is unnecessary to explicitly define its numerical value, provided its existence is guaranteed for the computed s_k^C. We note also that condition (2.32) implies that both s_k^C and the denominator of (2.39) are nonzero.

The vector x_k^C in (2.36) is called a *Generalized Cauchy Point*, or GCP, because it plays a role similar to that of the GCP in [4], [9], [20] and [29].

At this stage, it is far from obvious how a vector s_k^C satisfying the conditions of Step 2 can be computed. The existence and computation of a suitable step will be addressed in Sections 4 and 7.1.

4. Again, much freedom is left in the calculation of the step s_k in Step 3, but this fairly broad outline is sufficient for our analysis. However, this freedom is crucial in practical imple-

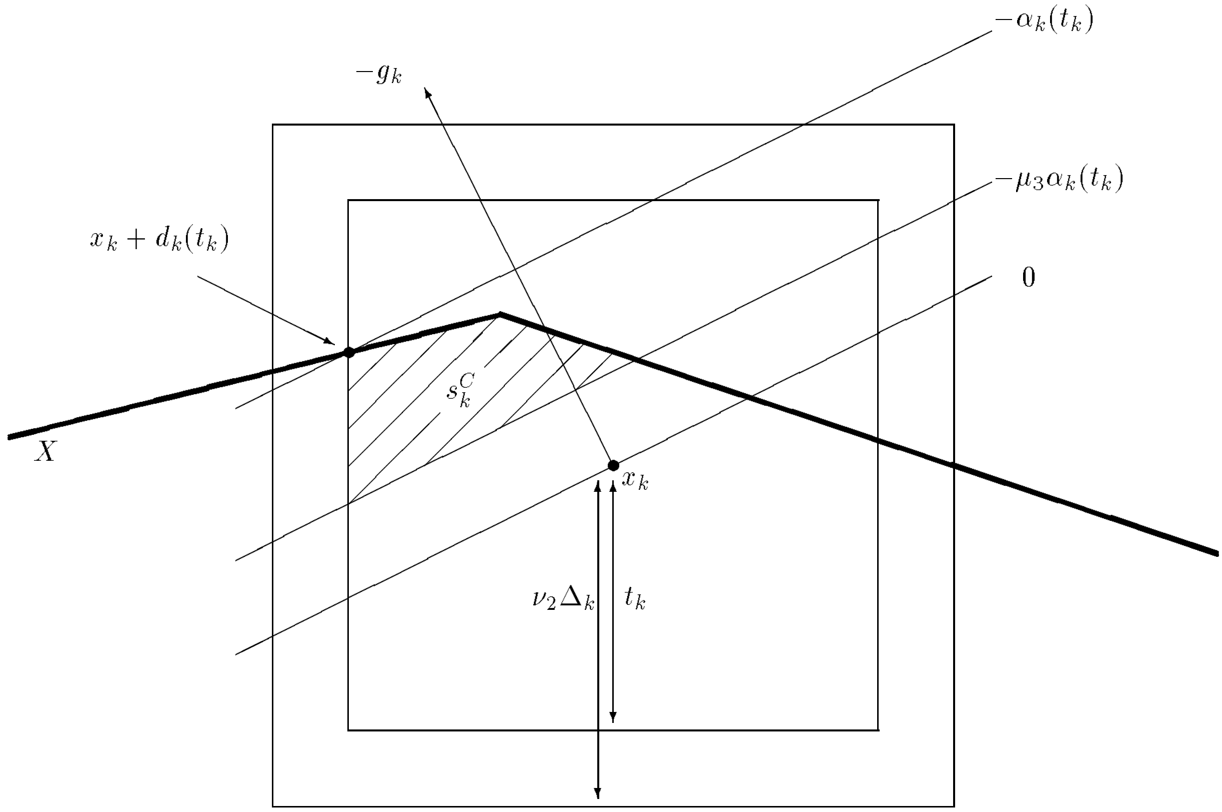

Figure 2: An illustration of condition (2.33) using the ℓ_∞-norm

mentations, as it allows a refinement of the GCP step based on second order information, hence providing a possibly fast ultimate rate of convergence.

5. Only a theoretical stopping rule has been specified at the beginning of Step 2. (This criterion will be justified in Section 3). Of course, any practical algorithm in our class must use a more practical test, which may depend on the particular class of models being used. The present hypothesis is however natural in our context, where we want to analyze the behaviour of the algorithm as k tends to infinity. We will therefore assume in the sequel that the test at the beginning of Step 2 is never triggered.

6. From the practical point of view, it may be unrealistic to let the trust region radius Δ_k grow to infinity, and most implementations do impose a uniform upper bound on these radii. This is coherent with (2.42), where a strict increase of Δ_k is not required.

7. The condition (2.45) may seem inappropriate when $\|s_k\|_{(k)}$ is small compared with the trust region radius Δ_k. Analogously to the observation in [29], this condition may be replaced

by the more practical

$$\Delta_{k+1} \in [\min(\gamma_0 \|s_k\|_{(k)}, \gamma_1 \Delta_k), \gamma_2 \Delta_k] \tag{2.49}$$

for some $\gamma_0 \in (0, 1]$ without modifying the theory presented below.

8. The algorithm necessarily depends on several constants. Typical values for some of them are $\mu_1 = 0.1$, $\mu_2 = 0.9$, $\mu_4 = 1$, $\nu_1 = 1$, $\nu_3 = 10^{-5}$, $\nu_4 = 0.01$, $\eta_1 = 0.25$, $\eta_2 = 0.75$, $\gamma_1 = 0.01$, $\gamma_2 = \frac{1}{2}$ and $\gamma_3 = 2$. Suitable values for the remaining constants will only become clear after extensive testing.

We call an iteration of the algorithm *successful* if the test (2.40) is satisfied, that is when the achieved objective reduction $f(x_k) - f(x_k + s_k)$ is large enough compared to the reduction $m_k(x_k) - m_k(x_k + s_k)$ predicted by the model. If (2.40) fails, the iteration is said to be *unsuccessful*. In what follows, the set of indices of successful iterations will be denoted by $\mathcal{S}$.

3 Global convergence for Algorithm 1

3.1 Criticality measures

If we are to prove that the iterates generated by Algorithm 1 converge to critical points for the problem (2.1)–(2.2), we clearly must specify how we will measure the "criticality" of a given feasible point. We say that a feasible point x_* is *critical* (or *stationary*) if and only if

$$-\nabla f(x_*) \in N(x_*). \tag{3.1}$$

We propose to use, as a measure of criticality, the quantity

$$\alpha_k[x] \stackrel{\text{def}}{=} |\min_{\substack{x+d \in X \\ \|d\|_{(k)} \leq 1}} \langle \nabla f(x), d \rangle|, \tag{3.2}$$

which can be interpreted as the magnitude of the maximum decrease of the *linearized objective function* achievable in the intersection of X with a ball of radius one (in the norm $\|\cdot\|_{(k)}$) centered at x. Observe that $\alpha_k[x]$ reduces to $\|\nabla f(x)\|_2$ when $X = \mathbf{R}^n$ and $\|\cdot\|_{(k)} = \|\cdot\|_2$.

Lemma 3 *Assume (AS.2) holds. Then, for all $k \geq 0$, $\alpha_k[\cdot]$ is continuous with respect to its argument.*

Proof. The continuity of $\alpha_k[\cdot]$ with respect to its argument is a direct consequence of Lemma 1 and of the continuity of ∇f. □

We now show that all the norms $\|\cdot\|_{(k)}$ are formally equivalent.

Theorem 4 *Assume (AS.2) and (AS.3) hold. Then there exists a positive constant $c_1 \geq 1$ such that*

$$\frac{1}{c_1} \alpha_{k_1}[x] \leq \alpha_{k_2}[x] \leq c_1 \alpha_{k_1}[x] \tag{3.3}$$

for all $x \in X$ and all $k_1 \geq 0$ and $k_2 \geq 0$.

Proof. We first observe that, using assumption (AS.3),

$$\|d\|_{(k)} = 1 \Longrightarrow \sigma_1 \leq \|d\|_2 \leq \sigma_2. \tag{3.4}$$

The lower (resp. upper) bound in this last inequality represents the smallest (resp. largest) possible distance (induced by $\|\cdot\|_2$) between x and the boundary of any ball, $\|d\|_{(k)} = 1$, for $k \geq 0$. The ball $\{x + d \mid \|d\|_2 \leq \sigma_2\}$ then contains all the balls of the form

$$\|d\|_{(k)} \leq 1, \tag{3.5}$$

while the ball $\{x + d \mid \|d\|_2 \leq \sigma_1\}$ is contained in them all. Consider now

$$\alpha_{\max} \stackrel{\text{def}}{=} |\min_{\substack{x+d \in X \\ \|d\|_2 \leq \sigma_2}} \langle \nabla f(x), d \rangle| \text{ and } \alpha_{\min} \stackrel{\text{def}}{=} |\min_{\substack{x+d \in X \\ \|d\|_2 \leq \sigma_1}} \langle \nabla f(x), d \rangle|. \tag{3.6}$$

Because of the second part of Lemma 2 (with $x_k = x$, $g_k = \nabla f(x)$ and $\|\cdot\|_{(k)} = \|\cdot\|_2$), we deduce that

$$\alpha_{\max} \leq \frac{\sigma_2}{\sigma_1} \alpha_{\min}. \tag{3.7}$$

Having established this property, we now return to the proof of Theorem 4 itself. If $\alpha_{k_1}[x] = \alpha_{k_2}[x]$, then (3.3) is trivially satisfied. We thus only consider the case where

$$\alpha_{k_1}[x] < \alpha_{k_2}[x], \tag{3.8}$$

say. In this situation, we will show that both d_1 and d_2, two vectors satisfying the relations

$$\alpha_{k_1}[x] = -\langle \nabla f(x), d_1 \rangle, \quad \|d_1\|_{(k_1)} \leq 1, \quad x + d_1 \in X, \tag{3.9}$$

and

$$\alpha_{k_2}[x] = -\langle \nabla f(x), d_2 \rangle, \quad \|d_2\|_{(k_2)} \leq 1, \quad x + d_2 \in X, \tag{3.10}$$

are such that

$$\sigma_1 \leq \|d_1\|_2 \leq \sigma_2 \text{ and } \sigma_1 \leq \|d_2\|_2 \leq \sigma_2. \tag{3.11}$$

We note that the two upper bounds in these inequalities immediately result from (AS.3) and (3.9)–(3.10). We therefore only consider the case where one or both lower bounds in (3.11) are violated. Assume, for instance, $\|d_1\|_2 < \sigma_1$. This solution of the minimization problem associated with $\alpha_{k_1}[x]$ is therefore in the interior of all the possible balls of the form (3.5). The only binding constraint at this point must be $x + d \in X$, and this is still true if the ball defined by $\|\cdot\|_{(k_1)}$ is replaced by that defined by $\|\cdot\|_{(k_2)}$. But this implies that (3.8) cannot hold, which is impossible. The case where $\|d_2\|_2 < \sigma_1$ is entirely similar. The inequalities (3.11) are therefore valid, and we obtain that

$$\alpha_{\min} \leq \alpha_{k_1}[x] \leq \alpha_{\max} \text{ and } \alpha_{\min} \leq \alpha_{k_2}[x] \leq \alpha_{\max}. \tag{3.12}$$

Combining these relations with (3.7) and (3.8), one deduces that

$$\alpha_{k_1}[x] < \alpha_{k_2}[x] \leq \alpha_{\max} \leq \frac{\sigma_2}{\sigma_1} \alpha_{\min} \leq \frac{\sigma_2}{\sigma_1} \alpha_{k_1}[x] \tag{3.13}$$

and (3.3) is proved with $c_1 \stackrel{\text{def}}{=} \frac{\sigma_2}{\sigma_1}$. □

The fact that $\alpha_k[x]$ can now be used as a criticality measure results from the following lemma.

Lemma 5 *Assume that (AS.1)–(AS.3) hold. Then, x_* is critical if and only if*

$$\alpha_k[x_*] = 0. \tag{3.14}$$

Proof. Consider first the minimization problem of (3.2) where we choose $\|\cdot\|_{(k)} = \|\cdot\|_2$, and let us denote the analog of (3.2) by $\alpha_2[x]$.

The criticality conditions for this problem can be expressed as

$$0 \in 2\zeta d + \nabla f(x) + N(x+d), \tag{3.15}$$

$$x + d \in X, \tag{3.16}$$

$$\|d\|_2 \leq 1 \tag{3.17}$$

and

$$\zeta\left(\|d\|_2^2 - 1\right) = 0. \tag{3.18}$$

Assume now that $\alpha_2[x_*] = 0$. Then the choice $d = 0$ is a solution of the minimization problem. The relation (3.1) then follows from (3.15).

Assume, on the other hand, that (3.1) holds. Then the conditions (3.15)–(3.18) are satisfied with $d = 0$ and $\zeta = 0$. It is then easy to verify that

$$\alpha_2[x_*] = 0 \tag{3.19}$$

follows.

As a consequence, x_* is critical if and only if (3.19) holds. But Theorem 4 and the fact that the ℓ_2-norm can be considered as one of the (k)-norms then yield the desired result. □

Lemmas 3 and 5 and Theorem 4 have the following important consequence.

Corollary 6 *Assume (AS.1)–(AS.3) hold and that the sequence $\{x_k\}$ is generated by Algorithm 1. Assume furthermore that there exists a subsequence of $\{x_k\}$, $\{x_{k_i}\}$ say, converging to x_* and that*

$$\lim_{i\to\infty} \alpha_{k_i}[x_{k_i}] = 0. \tag{3.20}$$

Then x_* is critical.

We note that, if formally equivalent, the criticality measures depending on k often differ from the practical point of view, when used in a stopping rule. If the problem's scaling is poor, a scaled measure is usually more appropriate. This scaling can be taken into account in the definition of the iteration dependent norms.

On the other hand, if the only first order information we can obtain is g_k (under the proviso (2.13)), then $\alpha_k[x]$ is unavailable, and one is naturally led to use

$$\alpha_k \stackrel{\rm def}{=} \alpha_k(1) = |\min_{\substack{x_k+d\in X\\ \|d\|_{(k)}\leq 1}} \langle g_k, d\rangle|, \tag{3.21}$$

which represents the amount of possible decrease for the *linearized model* in the intersection of the feasible domain with a ball of radius one. Clearly, $\alpha_k = \alpha_k[x_k]$ when $g_k = \nabla f(x_k)$, but this need not be the case in general. The value α_k was used in the "theoretical stopping rule" of Step 2 of Algorithm 1.

The replacement of $\alpha_k[x_k]$ by α_k has however a price. It may well happen indeed that an iterate x_k is a constrained critical point for the model m_k although x_k is not critical for the true problem. In that case, Algorithm 1 will stop at the beginning of Step 2. The model m_k should therefore reflect the noncriticality of x_k. The discrepancy between α_k and $\alpha_k[x_k]$ cannot be arbitrary large however, as is shown by the following result.

Lemma 7 *Let $x_k \in X$ be an iterate generated by Algorithm 1. Then*

$$|\alpha_k[x_k] - \alpha_k| \leq \|e_k\|_{[k]}. \tag{3.22}$$

Proof. Define d_k^* and d_k as two vectors satisfying

$$\alpha_k[x_k] = -\langle \nabla f(x_k), d_k^* \rangle, \quad \|d_k^*\|_{(k)} \leq 1, \quad x_k + d_k^* \in X, \tag{3.23}$$

and

$$\alpha_k = -\langle g_k, d_k \rangle, \quad \|d_k\|_{(k)} \leq 1, \quad x_k + d_k \in X. \tag{3.24}$$

Assume first that $\alpha_k[x_k] \geq \alpha_k$. Then, we can write that

$$\begin{aligned} 0 \leq \alpha_k[x_k] - \alpha_k &= \langle g_k, d_k \rangle - \langle \nabla f(x_k), d_k^* \rangle \\ &= \langle g_k, d_k - d_k^* \rangle + \langle e_k, d_k^* \rangle \\ &\leq \langle g_k, d_k - d_k^* \rangle + \|e_k\|_{[k]}, \end{aligned} \tag{3.25}$$

where we used the inequality (2.14). But the definitions of α_k, d_k and d_k^* imply that

$$\langle g_k, d_k \rangle = -\alpha_k \leq \langle g_k, d_k^* \rangle, \tag{3.26}$$

and hence (3.22) follows from (3.25). On the other hand, if $\alpha_k[x_k] < \alpha_k$, then a similar argument can be used to prove (3.22) with (3.25) replaced by

$$0 < \alpha_k - \alpha_k[x_k] \leq \langle \nabla f(x_k), d_k^* - d_k \rangle + \|e_k\|_{[k]} \tag{3.27}$$

and (3.26) by

$$\langle \nabla f(x_k), d_k^* \rangle = -\alpha_k[x_k] \leq \langle \nabla f(x_k), d_k \rangle. \tag{3.28}$$

□

The bound (3.22) will be used at the end of our global convergence analysis.

3.2 The model decrease

The traditional next step in a trust region oriented convergence analysis is to derive a lower bound on the reduction of the model value at an iteration where the current iterate x_k is noncritical. This lower bound usually involves the considered measure of criticality (α_k in our case), the trust region radius Δ_k and the inverse of the curvature of the model m_k (see [9], [19], [21], [23] and [29] for examples of such bounds). To define this notion of curvature more precisely, we follow

[29] and introduce, for an arbitrary continuously differentiable function q, the curvature at the point $x \in X$ along the step v, as defined by

$$\omega_k(q,x,v) \stackrel{\text{def}}{=} \frac{2}{\|v\|_{(k)}^2}\left[q(x+v) - q(x) - \langle \nabla q(x), v\rangle\right]. \tag{3.29}$$

If we assume that q is twice differentiable, the mean-value theorem (see [16, p. 11], for instance) implies that

$$\omega_k(q,x,v) = 2\int_0^1\int_0^1 \tau_2 \frac{\langle v, \nabla^2 q(x+\tau_1\tau_2 v)v\rangle}{\|v\|_{(k)}^2}\, d\tau_1\, d\tau_2. \tag{3.30}$$

It is also easy to verify that, if q is quadratic and $\|\cdot\|_{(k)} = \|\cdot\|_2$, then $\omega_k(q,x,v)$ is independent of x and of the norm of v, and reduces to the scaled Rayleigh quotient of $\nabla^2 q$ with respect to the direction v. We note that the Rayleigh quotient has already been used for similar purposes in the context of convergence analysis, namely in [7], [28] and [29].

We then obtain the following simple result.

Lemma 8 *If (AS.1)–(AS.3) hold, then there exists a finite constant $c_2 \geq 1$ such that*

$$\omega_k(f, x_k, s) \leq c_2 \tag{3.31}$$

for all $k \geq 0$ and all s such that $x_k + s \in \mathcal{L}$.

Proof. The Lipschitz continuity of $\nabla f(x)$ implies that

$$|f(x_k+s) - f(x_k) - \langle \nabla f(x_k), s\rangle| \leq \frac{1}{2} L_f \|s\|_2^2, \tag{3.32}$$

where L_f is the Lipschitz constant of $\nabla f(x)$ in the norm $\|\cdot\|_2$. We may then deduce from (3.29) that

$$\omega_k(f, x_k, s) \leq L_f \frac{\|s\|_2^2}{\|s\|_{(k)}^2}, \tag{3.33}$$

which gives (3.31) with $c_2 = \max[1, \sigma_2^2 L_f]$, by using (AS.3). $\square$

We are now in position to state our main result of this section.

Theorem 9 *Assume that (AS.1)–(AS.3) hold. Consider any sequence $\{x_k\}$ produced by Algorithm 1, and select a $k \geq 0$ such that x_k is not critical in the sense that $\alpha_k > 0$. Then, if one defines*

$$\omega_k^C \stackrel{\text{def}}{=} \begin{cases} \omega_k(m_k, x_k, s_k^C) & \text{if } s_k^C \text{ satisfies (2.35),} \\ 0 & \text{otherwise,} \end{cases} \tag{3.34}$$

one obtains that

$$\omega_k^C \geq 0. \tag{3.35}$$

Furthermore, there exists a constant $c_3 \in (0,1]$ such that

$$m_k(x_k) - m_k(x_k + s_k) \geq c_3 \alpha_k \min\left[1, \Delta_k, \frac{\alpha_k}{1+\omega_k^C}\right], \tag{3.36}$$

for all $k \geq 0$.

Proof. Let us first consider the case where $t_k \geq 1$. In this case, we obtain from (2.33), (2.32), the first statement of Lemma 2 and the definition (3.21) that

$$m_k(x_k) - m_k(x_k + s_k^C) \geq \mu_1\mu_3\alpha_k(t_k) \geq \mu_1\mu_3\alpha_k(1) = \mu_1\mu_3\alpha_k. \tag{3.37}$$

Assume now that $t_k < 1$. We first note that, because of (2.32) and the second part of Lemma 2, this last inequality and (3.21), we have that

$$\frac{|\langle g_k, s_k^C\rangle|}{t_k} \geq \mu_3\frac{\alpha_k(t_k)}{t_k} \geq \mu_3\frac{\alpha_k(1)}{1} = \mu_3\alpha_k. \tag{3.38}$$

Combining this inequality with (2.33), we obtain that

$$m_k(x_k) - m_k(x_k + s_k^C) \geq \mu_1\frac{|\langle g_k, s_k^C\rangle|}{t_k}t_k \geq \mu_1\mu_3\alpha_k t_k. \tag{3.39}$$

Now, if condition (2.34) is satisfied, we can deduce, by using (3.39), that

$$m_k(x_k) - m_k(x_k + s_k^C) \geq \mu_1\mu_3\alpha_k \min[\nu_3\Delta_k, \nu_4]. \tag{3.40}$$

On the other hand, if s_k^C satisfies (2.35), we observe that

$$\omega_k^C \geq \frac{2(1-\mu_2)}{\|s_k^C\|_{(k)}}\frac{|\langle g_k, s_k^C\rangle|}{\|s_k^C\|_{(k)}} \geq \frac{2(1-\mu_2)}{t_k}\frac{|\langle g_k, s_k^C\rangle|}{t_k}, \tag{3.41}$$

where we used the definition of ω_k^C and (2.35). Hence (3.35) is proved and, using (3.38), we have that

$$t_k \geq 2\mu_3(1-\mu_2)\frac{\alpha_k}{\omega_k^C} \geq 2\mu_3(1-\mu_2)\frac{\alpha_k}{1+\omega_k^C}. \tag{3.42}$$

Substituting this bound into (3.39) then yields that

$$m_k(x_k) - m_k(x_k + s_k^C) \geq 2\mu_1\mu_3^2(1-\mu_2)\frac{\alpha_k^2}{1+\omega_k^C}. \tag{3.43}$$

The inequality (3.36) now results from (3.37), (3.40), (3.43), (2.38) and $\nu_4 \leq 1$, with

$$c_3 = \mu_1\mu_3\mu_4 \min[\nu_3, \nu_4, 2\mu_3(1-\mu_2)] \leq 1. \tag{3.44}$$

□

We end this subsection by stating an easy corollary of Theorem 9, giving a lower bound on the decrease in the objective that is obtained on successful iterations.

Corollary 10 *Under the assumptions of Theorem 9, one obtains that*

$$f(x_k) - f(x_{k+1}) \geq \eta_1 c_3 \alpha_k \min\left[1, \Delta_k, \frac{\alpha_k}{1+\omega_k^C}\right], \tag{3.45}$$

for $k \in \mathcal{S}$.

Proof. The inequality (3.45) immediately results from (3.36), (2.39), (2.40) and (2.41). □

3.3 Convergence to critical points

This section will be devoted to the proof of global convergence of the iterates generated by Algorithm 1 to critical points.

For developing our convergence theory, we will need to introduce additional assumptions on the curvature of the models m_k. These assumptions, and the rest of our convergence analysis, will be phrased in terms of the quantity

$$\beta_k = 1 + \max_{i=0,\ldots,k} \left[\max[\omega_i^C, |\omega_i(m_i, x_i, s_i)|] \right]. \tag{3.46}$$

We note that β_k only measures curvature of the model along the s_k^C and s_k vectors. We also observe that the sequence $\{\beta_k\}$ is nondecreasing by definition.

We first recall two useful preliminary results in the spirit of [29].

Lemma 11 *Assume that (AS.1)–(AS.3) hold and consider a sequence $\{x_k\}$ of iterates generated by Algorithm 1. Then there exists a positive constant $c_4 \geq 1$ such that, for all $k \geq 0$,*

$$|f(x_k + s_k) - m_k(x_k + s_k)| \leq c_4 \beta_k \Delta_k^2. \tag{3.47}$$

Proof. We observe that

$$\begin{aligned} |f(x_k + s_k) - m_k(x_k + s_k)| &\leq |\langle \nabla f(x_k) - g_k, s_k \rangle| \\ &\quad + \tfrac{1}{2} \|s_k\|_{(k)}^2 |\omega_k(f, x_k, s_k) - \omega_k(m_k, x_k, s_k)| \\ &\leq \|e_k\|_{[k]} \, \|s_k\|_{(k)} \\ &\quad + \tfrac{1}{2} \|s_k\|_{(k)}^2 [|\omega_k(f, x_k, s_k)| + |\omega_k(m_k, x_k, s_k)|], \end{aligned} \tag{3.48}$$

where we used the definition (3.29), (2.12) and the inequality (2.14). But $\|s_k\|_{(k)} \leq \nu_1 \Delta_k$, and hence we obtain from (3.48), (2.13), (3.46) and Lemma 8 that

$$|f(x_k + s_k) - m_k(x_k + s_k)| \leq \kappa_1 \nu_1 \Delta_k^2 + \frac{1}{2} \nu_1^2 (c_2 + \beta_k) \Delta_k^2 \tag{3.49}$$

which then yields (3.47) with

$$c_4 = 2 \left(c_2 + \frac{\kappa_1}{\nu_1} \right) \max[1, \frac{1}{2} \nu_1^2]. \tag{3.50}$$

□

Lemma 12 *Assume that (AS.1)–(AS.3) hold and consider a sequence $\{x_k\}$ of iterates generated by Algorithm 1. Assume furthermore that there exists a constant $\epsilon \in (0, 1)$ such that*

$$\alpha_k \geq \epsilon \tag{3.51}$$

for all k. Then there exists a positive constant c_5 such that

$$\Delta_k \geq \frac{c_5}{\beta_k} \tag{3.52}$$

for all k.

Proof. Assume, without loss of generality, that

$$\epsilon < \frac{c_4 \beta_0 \Delta_0}{\gamma_1 c_3 (1 - \eta_2)}, \tag{3.53}$$

where γ_1 and η_2 are defined in the algorithm [(2.29) and (2.28)]. In order to derive a contradiction, assume also that there exists a k such that

$$\beta_k \Delta_k \leq \frac{\gamma_1 c_3 (1 - \eta_2)}{c_4} \epsilon \tag{3.54}$$

and define r as the first iteration number such that (3.54) holds. (Note that $r \geq 1$ because of (3.53).) The mechanism of Algorithm 1 then ensures that

$$\beta_{r-1} \Delta_{r-1} \leq \beta_r \frac{\Delta_r}{\gamma_1} \leq \frac{c_3 (1 - \eta_2)}{c_4} \epsilon \leq \epsilon \tag{3.55}$$

where we used the relations $\beta_{r-1} \leq \beta_r$, (2.45), (3.54) with $k = r$, $c_3 \leq 1$ and $c_4 \geq 1$. Combining the inequalities (3.51), (3.36), $\epsilon < 1$, $\beta_{r-1} \geq 1$ and (3.55), we now obtain that

$$m_{r-1}(x_{r-1}) - m_{r-1}(x_{r-1} + s_{r-1}) \geq c_3 \epsilon \min \left[1, \Delta_{r-1}, \frac{\epsilon}{\beta_{r-1}} \right] = c_3 \epsilon \Delta_{r-1}. \tag{3.56}$$

The relations (2.39), (3.47), (3.56) and the middle part of (3.55) together then imply that

$$|\rho_{r-1} - 1| = \frac{|f(x_{r-1} + s_{r-1}) - m_{r-1}(x_{r-1} + s_{r-1})|}{|m_{r-1}(x_{r-1}) - m_{r-1}(x_{r-1} + s_{r-1})|} \leq \frac{c_4 \beta_{r-1} \Delta_{r-1}}{c_3 \epsilon} \leq 1 - \eta_2. \tag{3.57}$$

Hence, $\rho_{r-1} \geq \eta_2$ and thus $\Delta_r \geq \Delta_{r-1}$. But we may deduce from this last inequality that

$$\beta_{r-1} \Delta_{r-1} \leq \beta_r \Delta_r \leq \frac{\gamma_1 c_3 (1 - \eta_2)}{c_4} \epsilon, \tag{3.58}$$

which contradicts the assumption that r is the first index with (3.54) satisfied. The inequality (3.54) therefore never holds and we obtain that, for all k,

$$\beta_k \Delta_k > \frac{\gamma_1 c_3 (1 - \eta_2)}{c_4} \epsilon. \tag{3.59}$$

The inequality (3.52) then follows from (3.59) by setting

$$c_5 = \frac{\gamma_1 c_3 (1 - \eta_2) \epsilon}{c_4}. \tag{3.60}$$

□

We now formulate our first assumption on the model's curvatures.

AS.4 The series

$$\sum_{k=0}^{\infty} \frac{1}{\beta_k} \tag{3.61}$$

is divergent.

As shown in [29], this condition is necessary for guaranteeing convergence to a stationary point. It is clearly satisfied in the common case where quadratic models of the form (2.46) are used, whose Hessian matrices H_k are uniformly bounded. This last assumption obviously holds when $f(x)$ is twice continuously differentiable over the compact set $\mathcal{L}$ and $H_k = \nabla^2 f(x_k)$.

Before proving one of the major results of this section, we recall the following technical lemma, due to Powell [24] (proofs can also be found in [9] or [32]).

Lemma 13 *Let $\{\Delta_k\}$ and $\{\beta_k\}$ be two sequences of positive numbers such that $\beta_k \Delta_k \geq c_5$ for all k, where c_5 is a positive constant. Let ϵ be a positive constant, $\mathcal{S}$ be a subset of $\{1, 2, \ldots\}$ and assume that, for some constants $\gamma_2 < 1$ and $\gamma_3 > 1$,*

$$\Delta_{k+1} \leq \gamma_3 \Delta_k \text{ for } k \in \mathcal{S}, \tag{3.62}$$

$$\Delta_{k+1} \leq \gamma_2 \Delta_k \text{ for } k \notin \mathcal{S}, \tag{3.63}$$

$$\beta_{k+1} \geq \beta_k \text{ for all } k \tag{3.64}$$

and

$$\sum_{k \in \mathcal{S}} \min\left[\Delta_k, \frac{\epsilon}{\beta_k}\right] < \infty. \tag{3.65}$$

Then

$$\sum_{k=1}^{\infty} \frac{1}{\beta_k} < \infty. \tag{3.66}$$

Using this lemma, we now show the following important result.

Theorem 14 *Assume (AS.1)–(AS.4) hold. Then, if $\{x_k\}$ is a sequence of iterates generated by Algorithm 1, one has that*

$$\liminf_{k \to \infty} \alpha_k = 0. \tag{3.67}$$

Proof. Assume, for the purpose of obtaining a contradiction, that there exists an $\epsilon \in (0, 1)$ such that (3.51) holds for all $k \geq 0$. Corollary 10 and the fact that the objective function is bounded below on $\mathcal{L}$ imply that

$$\eta_1 c_3 \epsilon \sum_{k \in \mathcal{S}} \min\left[1, \Delta_k, \frac{\epsilon}{\beta_k}\right] \leq \sum_{k \in \mathcal{S}} [f(x_k) - f(x_{k+1})] < \infty. \tag{3.68}$$

Thus, because of Lemma 12 and the inequalities $\epsilon < 1$ and $\beta_k \geq 1$, the sequences Δ_k and β_k then verify all the assumptions of Lemma 13, which then guarantees that

$$\sum_{k=0}^{\infty} \frac{1}{\beta_k} < \infty. \tag{3.69}$$

This last relation clearly contradicts (AS.4), and hence our initial assumption must be false, yielding (3.67). □

This theorem has the following interesting consequences.

Corollary 15 *Assume (AS.1)–(AS.4) hold. Assume furthermore that $\{x_k\}$ is a sequence of iterates generated by Algorithm 1 that converges to x_*, and that*

$$\lim_{k \to \infty} \|e_k\|_{[k]} = 0. \tag{3.70}$$

Then x_* is critical.

Proof. This result directly follows from (3.70), Lemma 7, Theorem 14 and Corollary 6. □

Corollary 16 *Assume (AS.1)–(AS.4) hold. If $\{x_k\}$ is a sequence of iterates generated by Algorithm 1 and if $\mathcal{S}$ is finite, then the iterates x_k are all equal to some x_* for k large enough, and x_* is critical.*

Proof. If $\mathcal{S}$ is finite, it results from (2.44) that x_k is unchanged for k large enough, and therefore that $x_k = x_* = x_{j+1}$ for k sufficiently large, where j is the largest index in $\mathcal{S}$. The relations (2.45) and (2.29) also imply that the sequence $\{\Delta_k\}$ converges to zero. Hence (2.13) ensures that (3.70) holds. We then apply Corollary 15 to deduce the criticality of x_*. □

If we now assume that $\mathcal{S}$ is infinite, we wish to replace the "lim inf" in (3.67) by a true limit, taken on all successful iterations, but this requires a slight strengthening of our assumption on the model curvature.

AS.5 We assume that

$$\lim_{k\to\infty} \beta_k[f(x_k) - f(x_{k+1})] = 0. \tag{3.71}$$

As discussed in [9], this assumption is not very severe, as we always have that (3.71) holds with the limit replaced by the limit inferior. Also (AS.5) is obviously satisfied when using a model with bounded curvature, as is assumed in [20] for example.

Theorem 17 *Assume (AS.1)–(AS.5) hold. Then, if $\{x_k\}$ is a sequence of iterates generated by Algorithm 1 and if the set $\mathcal{S}$ is infinite, one has that*

$$\lim_{\substack{k\to\infty\\ k\in\mathcal{S}}} \alpha_k = 0. \tag{3.72}$$

Proof. We proceed again by contradiction and assume that there exists an $\epsilon_1 \in (0,1)$ and a subsequence $\{m_i\}$ of successful iterates such that, for all m_i in this subsequence,

$$\alpha_{m_i} \geq \epsilon_1. \tag{3.73}$$

If we define

$$c_6 \stackrel{\rm def}{=} \max[1 - \frac{1}{c_1}, c_1 - 1], \tag{3.74}$$

where c_1 is given by Theorem 4, and if we choose

$$\epsilon_2 \in (0, \frac{\epsilon_1}{2(c_6+1)}), \tag{3.75}$$

Theorem 14 then ensures the existence of another subsequence $\{\ell_i\}$ such that

$$\alpha_k \geq \epsilon_2 \;\text{ for }\; m_i \leq k < \ell_i \;\text{ and }\; \alpha_{\ell_i} < \epsilon_2. \tag{3.76}$$

We now restrict our attention to the subsequence of successful iterations whose indices are in the set

$$\mathcal{K} \stackrel{\rm def}{=} \{k \in \mathcal{S} \mid m_i \leq k < \ell_i\}, \tag{3.77}$$

where m_i and ℓ_i belong respectively to the two subsequences defined above. Applying Corollary 10 for $k \in \mathcal{K}$, we obtain that

$$f(x_k) - f(x_{k+1}) \geq \eta_1 c_3 \epsilon_2 \min\left[\Delta_k, \frac{\epsilon_2}{\beta_k}\right], \tag{3.78}$$

where we used the inequalities $\epsilon_2 < 1$ and $\beta_k \geq 1$. But (AS.5) then implies that

$$\lim_{\substack{k\to\infty \\ k\in\mathcal{K}}} \beta_k \Delta_k = 0, \tag{3.79}$$

and hence,using (3.78), that

$$f(x_k) - f(x_{k+1}) \geq \eta_1 c_3 \epsilon_2 \Delta_k \tag{3.80}$$

for $k \in \mathcal{K}$ sufficiently large. As a consequence, we obtain, for i sufficiently large, that

$$\begin{aligned} \|x_{m_i} - x_{\ell_i}\|_2 &\leq \textstyle\sum_{k=m_i}^{\ell_i-1} \|x_{k+1} - x_k\|_2 \\ &\leq \sigma_2 \nu_1 \textstyle\sum_{k=m_i}^{\ell_i-1}{}^{(\mathcal{K})} \Delta_k \\ &\leq c_7 \textstyle\sum_{k=m_i}^{\ell_i-1}{}^{(\mathcal{K})} [f(x_k) - f(x_{k+1})] \\ &\leq c_7 [f(x_{m_i}) - f(x_{\ell_i})], \end{aligned} \tag{3.81}$$

where the sums with superscript $(\mathcal{K})$ are restricted to the indices in $\mathcal{K}$, and where

$$c_7 \stackrel{\text{def}}{=} \frac{\sigma_2 \nu_1}{\eta_1 c_3 \epsilon_2}. \tag{3.82}$$

Since the last right-hand side of (3.81) tends to zero as i tends to infinity and because of Lemma 3, we deduce that

$$|\alpha_{m_i}[x_{m_i}] - \alpha_{m_i}[x_{\ell_i}]| \leq \frac{\epsilon_1}{2(c_6+3)} \tag{3.83}$$

for i sufficiently large. We note now that (3.79), $\beta_k \geq 1$ and (2.13) imply that g_{m_i} is arbitrarily close to $\nabla f(x_{m_i})$, and hence Lemma 7 gives that

$$|\alpha_{m_i} - \alpha_{m_i}[x_{m_i}]| \leq \frac{\epsilon_1}{2(c_6+3)} \tag{3.84}$$

for i large enough. We observe also that, because of (2.13) and (2.42),

$$\|e_{\ell_i}\|_{[\ell_i]} \leq \kappa_1 \Delta_{\ell_i} \leq \kappa_1 \gamma_3 \Delta_{k_i}, \tag{3.85}$$

where k_i is the largest integer in $\mathcal{K}$ that is smaller than ℓ_i. As before, we now deduce from (3.79), $\beta_k \geq 1$, Lemma 7 and (3.85) that

$$|\alpha_{\ell_i} - \alpha_{\ell_i}[x_{\ell_i}]| \leq \frac{\epsilon_1}{2(c_6+3)} \tag{3.86}$$

for large i. Hence, using Theorem 4, we obtain that

$$|\alpha_{m_i}[x_{\ell_i}] - \alpha_{\ell_i}[x_{\ell_i}]| \leq c_6 \alpha_{\ell_i}[x_{\ell_i}] \leq c_6 \left[\alpha_{\ell_i} + \frac{\epsilon_1}{2(c_6+3)}\right] \tag{3.87}$$

for i sufficiently large. Using the triangular inequality together with (3.84), (3.83), (3.87) and (3.86), we obtain that, for large enough i,

$$\alpha_{m_i} - \alpha_{\ell_i} \leq |\alpha_{m_i} - \alpha_{\ell_i}| \leq c_6 \alpha_{\ell_i} + \frac{1}{2}\epsilon_1. \tag{3.88}$$

We then deduce from (3.76) and (3.75), that, for large enough i,

$$\alpha_{m_i} \leq \alpha_{\ell_i}(c_6+1) + \frac{1}{2}\epsilon_1 < \epsilon_1, \tag{3.89}$$

which contradicts (3.73) and proves the desired result. □

As above, we can obtain conclusions about convergent subsequences where the first order information is asymptotically correct. If $\mathcal{S}$ is finite, the convergence of the iterates to a critical point results from Corollary 16. Hence, we now restrict our attention to the case where $\mathcal{S}$ is infinite.

Corollary 18 *Assume (AS.1)–(AS.5) hold. Assume furthermore that $\mathcal{S}$ is infinite, that $\{x_{k_i}\}$ is a convergent subsequence of the successful iterates generated by Algorithm 1 and that*

$$\lim_{i\to\infty} \|e_{k_i}\|_{[k_i]} = 0. \tag{3.90}$$

Then x_*, the limit point of $\{x_{k_i}\}$, is critical.

Proof. The proof of this result is entirely similar to that of Corollary 15 except that we have to consider only the successful iterates. □

Finally, we are interested in what can be said on the criticality of limit points of $\{x_k\}$ if we do not assume (3.70).

Corollary 19 *Assume (AS.1)–(AS.5) hold, that $\{x_{k_i}\}$ is a subsequence of successful iterates generated by Algorithm 1 and that $\{x_{k_i}\}$ converges to x_*. Then*

$$\limsup_{i\to\infty} \alpha_{k_i}[x_*] \le \limsup_{i\to\infty} \|e_{k_i}\|_{[k_i]}. \tag{3.91}$$

Proof. If $\mathcal{S}$ is finite, then the result immediately follows from Corollary 16 and Lemma 5. Assume therefore that $\mathcal{S}$ is infinite. Because of Lemma 3, Lemma 7 and Theorem 17, we have that

$$\begin{aligned} \limsup_{i\to\infty} \alpha_{k_i}[x_*] &= \limsup_{i\to\infty} \alpha_{k_i}[x_{k_i}] \\ &\le \limsup_{i\to\infty} |\alpha_{k_i}[x_{k_i}] - \alpha_{k_i}| \\ &\le \limsup_{i\to\infty} \|e_{k_i}\|_{[k_i]}. \end{aligned} \tag{3.92}$$

□

Keeping in mind that the dependence of $\|\cdot\|_{[k_i]}$ on k_i, and hence on i, is irrelevant because of Theorem 4, Corollary 19 thus guarantees that all limit points are "as critical as the scaled accuracy of g_k as an approximation to $\nabla f(x_k)$ warrants".

4 A model algorithm for computing a Generalized Cauchy Point

A major difficulty in adapting the framework given by Algorithm 1 to a more practical setting is clearly the definition of a practical procedure to compute a GCP satisfying all the conditions of Step 2.

As indicated already, such procedures have been designed and implemented in the case where the projected gradient path defined by the classical ℓ_2-norm is explicitly available (see [1] and [29], for example).

We now consider the more general case presented in Sections 2 and 3, and we wish to find, at a given iteration, a GCP satisfying (2.30)–(2.35). The difficulty is then to produce a point that is not too far away from the *unavailable* projected gradient path. This cannot be done without considering the particular geometry of this path, which may closely follow the boundary of the feasible set. As a consequence, linear interpolation between two points on the projected gradient path is often unsuitable and a specialized procedure is presented in this section.

For the sake of clarity, in this section we will drop the subscript k, corresponding to the iteration number.

4.1 The RS Algorithm

We first define the following *restriction operator* associated with the feasible set X and a *centre* $x \in X$. This operator is defined as

$$R_x[y] \stackrel{\text{def}}{=} \arg \min_{z \in [x,y] \cap X} \|z - y\|_2 \tag{4.1}$$

for any $y \in \mathbf{R}^n$, where $[x, y]$ is the segment between x and y. The definition of $R_x[y]$ uses the ℓ_2-norm, but any other norm can be used because the associated minimization problem is unidimensional. The action of the restriction operator (4.1) is illustrated in Figure 3. It should be noted that computing $R_x[y]$ for a given y is often a very simple task.

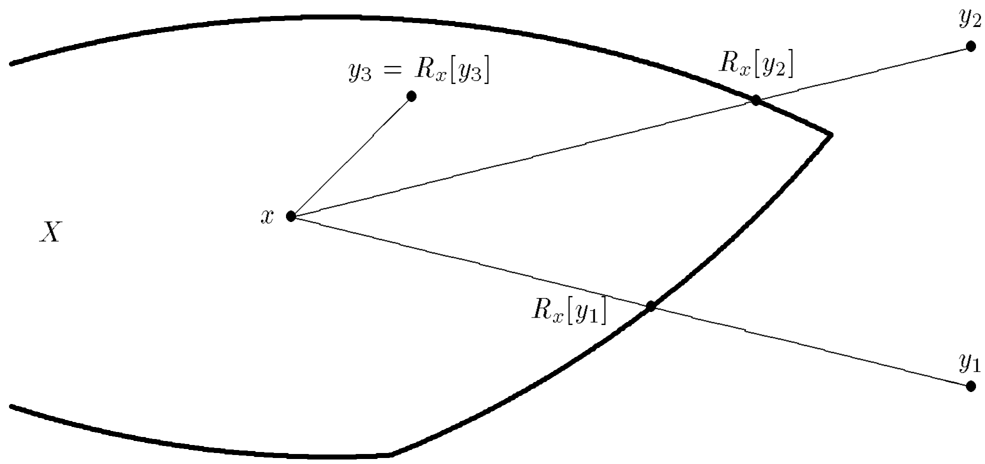

Figure 3: The restriction operator with centre x

The GCP Algorithm relies on a simple bisection linesearch algorithm on the restriction of a piecewise linear path with respect to a given center, called the *RS Algorithm* (which stands for Restricted Search Algorithm). Because of the definition of the restriction operator, this last algorithm closely follows the boundary of the feasible domain, as desired. It finds a point $x_* = x + z$ in $R_x[x^l, x^p, x^u]$, the restriction of a non-empty piecewise linear path consisting of the segment $[x^l, x^p]$ followed by $[x^p, x^u]$, where x^l, x^p and x^u are defined below. The restriction is computed with respect to the centre x and the resulting vector z is such that (2.33) and (2.35) hold with $s_k^C = z$. The RS Algorithm can be applied under the conditions that (2.35) is violated at $R_x[x^l]$ and that (2.33) is violated at $R_x[x^u]$. It therefore depends on the three points x^l, x^p and x^u defining the piecewise linear path, the centre x, and on the current model m (and hence on its gradient g). It also depends on an arbitrary bijective parametrisation of the path $[x^l, x^p, x^u]$. For example, one can choose the parameter to be the length of the arc along the path measured in the ℓ_2-norm. More formally, if

$$\delta_p = \|x^p - x^l\|_2 \text{ and } \delta_u = \delta_p + \|x^u - x^p\|_2, \tag{4.2}$$

we can define

$$x(\delta) \stackrel{\text{def}}{=} \begin{cases} \frac{\delta}{\delta_p} x^p + (1 - \frac{\delta}{\delta_p}) x^l & \text{if } \delta \leq \delta_p, \\ \frac{\delta - \delta_p}{\delta_u - \delta_p} x^u + (1 - \frac{\delta - \delta_p}{\delta_u - \delta_p}) x^p & \text{if } \delta \geq \delta_p \end{cases} \tag{4.3}$$

for any $\delta \in [0, \delta_u]$. The inner iterations of Algorithm RS will be denoted by the index j.

RS Algorithm

Step 0 : initialization. Set $l_0 = 0$, $u_0 = \delta_u$ and $j = 0$. Then define $\delta_0 = \frac{1}{2}(l_0 + u_0)$.

Step 1 : check the stopping conditions. Compute $x_j = R_x[x(\delta_j)]$ using (4.1) and (4.3). If

$$m(x_j) > m(x) + \mu_1 \langle g, x_j - x \rangle, \tag{4.4}$$

then set

$$l_{j+1} = l_j \text{ and } u_{j+1} = \delta_j, \tag{4.5}$$

and go to Step 2. Else, if

$$m(x_j) < m(x) + \mu_2 \langle g, x_j - x \rangle, \tag{4.6}$$

then set

$$l_{j+1} = \delta_j \text{ and } u_{j+1} = u_j, \tag{4.7}$$

and go to Step 2; else (that is if both (4.4) and (4.6) fail), set $x_* = x_j$ and STOP.

Step 2 : choose the next parameter value by bisection. Increment j by one, set

$$\delta_j = \frac{1}{2}(l_j + u_j) \tag{4.8}$$

and go to Step 1.

The fact that a vector x_* has been produced by the application of the RS Algorithm on the path $[x^l, x^p, x^u]$ with respect to the centre x and the model m will be denoted by

$$x_* = \text{RS}(x, m, x^l, x^p, x^u). \tag{4.9}$$

We have the following simple result.

Lemma 20 *Assume that the RS Algorithm is applied on a piecewise linear path $[x^l, x^p, x^u]$ satisfying the conditions stated in the paragraph preceding its description, with centre x and model m. Then this algorithm terminates with a suitable vector $x_* = x + z$ at which (2.33) and (2.35) hold in a finite number of iterations.*

Proof. We first note that (2.35) is violated at $R_x[x^l]$ and that (2.33) is violated at $R_x[x^u]$. As a consequence, the validity of the result directly follows from the inequality $\mu_1 < \mu_2$, the continuity of the model m on the restriction of the path $[x^l, x^p, x^u]$, and from the fact that the length of the interval $[l_j, u_j]$ tends geometrically to zero while its associated arc on the restricted path always contains a fixed connected set of acceptable points. □

4.2 The GCP Algorithm

We now describe the GCP Algorithm itself. It depends on the current iterate $x \in X$, on the current model m and its gradient g, on the current norm $\|\cdot\|$ and also on the current trust region radius, $\Delta > 0$. Its inner iterations will be identified by the index i. (Also recall that all subscripts k have been dropped, yielding, for instance, $\alpha(t)$ instead of $\alpha_k(t)$ and α instead of α_k.)

GCP Algorithm

Step 0: initialization. Set $i = 0$, $l_0 = 0$, $z_0^l = 0$ and $u_0 = \nu_2\Delta$. Also choose z_0^u an arbitrary vector such that $\|z_0^u\| > \nu_2\Delta$ and an initial parameter $t_0 \in (0, \nu_2\Delta]$.

Step 1: compute a candidate step. Compute a vector z_i such that

$$\|z_i\| \le t_i, \tag{4.10}$$

$$x + z_i \in X \tag{4.11}$$

and

$$\langle g, z_i\rangle \le -\mu_3\alpha(t_i). \tag{4.12}$$

Step 2: check the stopping rules on the model and step. If

$$m(x + z_i) > m(x) + \mu_1\langle g, z_i\rangle, \tag{4.13}$$

then set

$$u_{i+1} = t_i \quad z_{i+1}^u = z_i \tag{4.14}$$

and

$$l_{i+1} = l_i \quad z_{i+1}^l = z_i^l, \tag{4.15}$$

and go to Step 3. Else, if

$$m(x + z_i) < m(x) + \mu_2\langle g, z_i\rangle \tag{4.16}$$

and

$$t_i < \min[\nu_3\Delta, \nu_4], \tag{4.17}$$

then set

$$u_{i+1} = u_i \quad z_{i+1}^u = z_i^u \tag{4.18}$$

and

$$l_{i+1} = t_i \quad z_{i+1}^l = z_i, \tag{4.19}$$

and go to Step 3. Else (that is if (4.13) and either (4.16) or (4.17) fail), then set

$$x^C = x + z_i \tag{4.20}$$

and STOP.

Step 3: define a new trial step by bisection. We distinguish two mutually exclusive cases.

Case 1: $z^l_{i+1} = z^l_0$ or $z^u_{i+1} = z^u_0$. Set

$$t_{i+1} = \frac{1}{2}(l_{i+1} + u_{i+1}), \tag{4.21}$$

increment i by one and go to Step 1.

Case 2: $z^l_{i+1} \neq z^l_0$ and $z^u_{i+1} \neq z^u_0$. Define

$$z^p_{i+1} = \max\left[1, \frac{\|z^u_{i+1}\|}{\|z^l_{i+1}\|}\right] z^l_{i+1}, \tag{4.22}$$

set

$$x^C = \mathrm{RS}(x, m, x^l_{i+1}, x^p_{i+1}, x^u_{i+1}) \tag{4.23}$$

where

$$x^l_{i+1} = x + z^l_{i+1}, \quad x^p_{i+1} = x + z^p_{i+1} \text{ and } x^u_{i+1} = x + z^u_{i+1}, \tag{4.24}$$

and STOP.

The actual value of z^u_0 is irrelevant in practice: this quantity is merely used to detect if z^u_{i+1} has been updated in (4.14) at least once.

Figure 4 shows the situation at a given iteration of the GCP Algorithm in the case where $\|\cdot\|_{(k)} = \|\cdot\|_\infty$. In particular, the use of the point x^p as defined in Step 3 (Case 2) is illustrated. The symbols x^r, x^f, t^l, t^u, x_{t^l}, C_{t^l} and C_{t^u} are not yet defined, but will be introduced in the proof of Theorem 24 below.

We note that linear interpolation between $x^l_{i+1} = R_x[x^l_{i+1}]$ and $x^u_{i+1} = R_x[x^u_{i+1}]$ cannot be used in general in Step 3 (Case 2), because the geometry of the boundary of the feasible domain may imply that the (unknown) projected gradient path considerably departs from the segment $[x^l_{i+1}, x^u_{i+1}]$. This is the reason why a call is made to the RS Algorithm, which closely follows this boundary.

We emphasize that this GCP Algorithm is only a model, intended to show feasibility of our approach, but is not optimized from the point of view of efficiency. Many additional considerations are possible and indeed necessary before implementing the algorithm, including

- the details of the all important solver used to determine z_i in Step 1,
- a suitable choice of t_0,
- more efficient techniques for simple models (linear or quadratic, for instance), and also for specific choices of the norm $\|\cdot\|$.

The solver used in Step 1 obviously depends on X and the norm $\|\cdot\|$. For example, Step 1 reduces to a linear programming problem if X is polyhedral and a polyhedral norm is used; the classical projected gradient may also be obtained when the ℓ_2-norm is used and $\mu_3 = 1$.

If we denote by

$$x^C = \mathrm{GCP}(x, m, \|\cdot\|, \Delta) \tag{4.25}$$

the fact that the vector x^C has been obtained by the GCP Algorithm for the point x, the model m, the norm $\|\cdot\|$ and the radius Δ, we then replace Step 2 of Algorithm 1 by the simple call

$$x_k^C = \mathrm{GCP}(x_k, m_k, \|\cdot\|_{(k)}, \Delta_k). \tag{4.26}$$

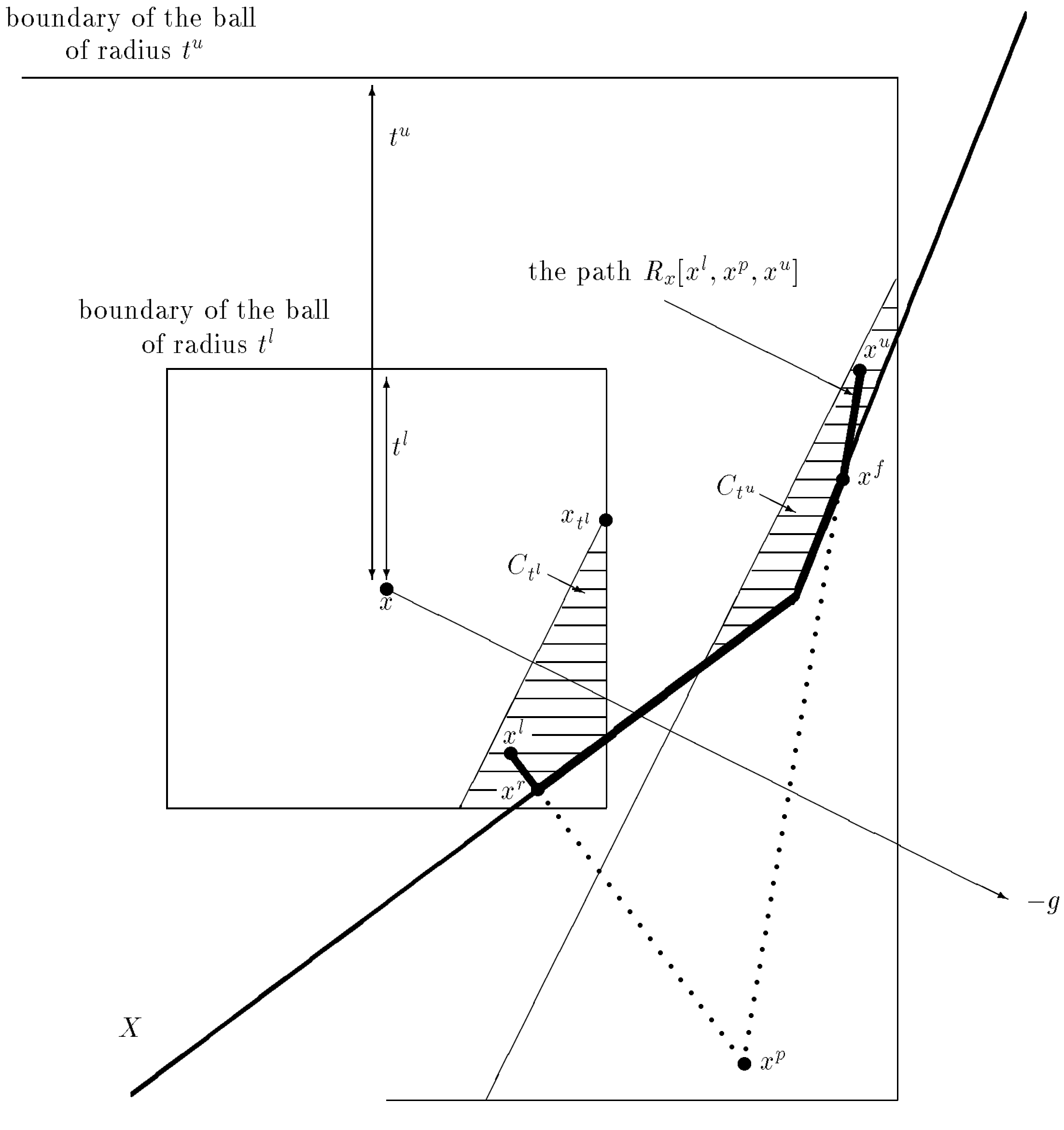

Figure 4: A "restricted path" with the ℓ_∞-norm

4.3 Properties of the GCP Algorithm

We now wish to show that the GCP Algorithm converges to a point satisfying (2.30)–(2.35) and terminates in a finite number of iterations.

The first result shows that, if a step z satisfies (2.32), then all *prolongations* of this step, that is all vectors of the form τz with $\tau \geq 1$, also satisfy the same condition.

Lemma 21 *Assume that there exists a $t \geq \|z\|$ such that*

$$\langle g, z\rangle \leq -\mu_3 \alpha(t) \tag{4.27}$$

for some $z \neq 0$. Then

$$\langle g, \tau z\rangle \leq -\mu_3 \alpha(\tau t) \tag{4.28}$$

for $\tau \geq 1$.

Proof. Using successively (4.27), the inequality $\tau \geq 1$ and the second part of Lemma 2, we obtain that

$$\langle g, \tau z\rangle \leq -\mu_3 \tau t \frac{\alpha(t)}{t} \leq -\mu_3 \tau t \frac{\alpha(\tau t)}{\tau t}, \tag{4.29}$$

yielding the desired bound. □

We are now in the position to prove that the GCP Algorithm is correctly stated, finite and coherent with the theoretical framework presented in Sections 2 and 3.

Lemma 22 *The GCP Algorithm has well-defined iterates.*

Proof. We have to verify that all the requested conditions for applying the RS Algorithm are fulfilled when a call to this algorithm is made. We first note that the RS Algorithm can only produce a feasible point because of the definition of the restriction operator. We also note that the mechanism of the GCP Algorithm ensures that the piecewise path to be restricted is non-empty, that (2.33) is always violated at $R_x[x^u_{i+1}] = x^u_{i+1}$ and, similarly, that (2.35) is always violated at $R_x[x^l_{i+1}] = x^l_{i+1}$. The RS Algorithm is therefore applied in the appropriate context. □

We now prove the desirable finiteness of the GCP Algorithm at noncritical points.

Theorem 23 *Assume that $\alpha > 0$. Then the GCP Algorithm terminates with a suitable x^C in a finite number of iterations.*

Proof. Assume that an infinite number of iterations are performed. We first consider the case where

$$z^l_i = z^l_0 \text{ for all } i \geq 0. \tag{4.30}$$

In this case, the mechanism of the GCP Algorithm implies that

$$t_i \leq (\frac{1}{2})^i \nu_2 \Delta. \tag{4.31}$$

Hence we obtain that

$$\|z_i\| \leq t_i \leq \min\left[1, \frac{2(1-\mu_1)\mu_3\alpha}{L_m}\right] \tag{4.32}$$

for all $i \geq i_1$, say, where L_m is the Lipschitz constant of the gradient of m with respect to the norm $\|\cdot\|$. For all $i \geq 0$, we have that

$$m(x+z_i) - m(x) - \mu_1 \langle g, z_i \rangle \leq (1-\mu_1)\langle g, z_i \rangle + \frac{1}{2} L_m \|z_i\|^2, \tag{4.33}$$

where we have used the Taylor's expansion of m around x and the definition of L_m. But the second part of Lemma 2 implies that

$$\frac{\alpha(t_i)}{t_i} \geq \frac{\alpha(1)}{1} = \alpha \tag{4.34}$$

for all $i \geq i_1$, and hence that

$$\alpha(t_i) \geq \alpha \|z_i\| \tag{4.35}$$

for $i \geq i_1$, because of the inequality $t_i \geq \|z_i\|$. Condition (4.12) then gives, for such i, that

$$\langle g, z_i \rangle \leq -\mu_3 \alpha(t_i) \leq -\mu_3 \alpha \|z_i\|. \tag{4.36}$$

Introducing this inequality in (4.33), we obtain that

$$m(x+z_i) - m(x) - \mu_1 \langle g, z_i \rangle \leq -(1-\mu_1)\mu_3 \alpha \|z_i\| + \frac{1}{2} L_m \|z_i\|^2 \tag{4.37}$$

for $i \geq i_1$. Using (4.32), we now deduce that

$$m(x+z_i) - m(x) - \mu_1 \langle g, z_i \rangle \leq 0 \tag{4.38}$$

for all $i \geq i_1$. As a consequence, (4.13) is always violated for sufficiently large i and (4.30) is therefore impossible.

We thus next consider the case where $z_i^u = z_0^u$ for all i. This implies that (4.13) is always false and that the algorithm either stops through (4.20) (in which case the convergence is clearly finite) or uses (4.19) at each iteration. But the effect of (4.19) is that l_i tends to $\nu_2 \Delta$ as i grows, and therefore (4.17) must fail for sufficiently large i because $\nu_3 < \nu_2$. The algorithm then terminates with (4.20) after finitely many iterations.

We conclude from these two arguments that, for the algorithm to be infinite, then one must have that $z_{i_1}^l \neq z_0^l$ for some $i_1 > 0$ and also that $z_{i_2}^u \neq z_0^u$ must be defined for some $i_2 > 0$. But, because the mechanism of the algorithm guarantees that the sequence $\{l_i\}$ is nondecreasing and that the sequence $\{u_i\}$ is nonincreasing, Case 2 in Step 3 therefore occurs for $i = \max(i_1, i_2)$. The RS Algorithm is thus used in (4.23) and Lemma 20 again ensures finite temination. □

Theorem 24 *The call (4.26) can be used as an implementation of Step 2 of Algorithm 1.*

Proof. We have to verify the compatibility of the GCP Algorithm with the conditions of Step 2 in Algorithm 1, that is we have to check that the step $s_k^C = x_k^C - x_k$ produced by (4.26) does indeed satisfy the conditions (2.30)–(2.35). All these conditions except (2.32) are clearly enforced by the mechanism of the GCP and RS Algorithms. We can therefore restrict our attention to the verification of (2.32) for the two different possible exits of the GCP Algorithm and their associated $s_k^C = x_k^C - x_k$. Dropping again the subscripts k, we have to verify that (4.27) holds with $z = x^C - x$.

The first case is when the GCP Algorithm terminates using (4.20). Then (4.12) ensures that (4.27) holds for $z = z_i$.

The second and last case is when the algorithm terminates through (4.23). The condition (4.12) again ensures that, in this case, (4.27) holds for $z = z_{i+1}^l$ for some $t_{i+1}^l \geq \|z_{i+1}^l\|$, and for $z = z_{i+1}^u$ for some $t_{i+1}^u \geq \|z_{i+1}^u\|$. For clarity of notations, we drop the subscript $i+1$ below.

We analyze the situation in the plane H containing x, x^l and x^u, and define, for $t > 0$, the convex sets

$$H_t \stackrel{\text{def}}{=} \{x + z \in H | \langle g, z \rangle \leq -\mu_3 \alpha(t)\}, \tag{4.39}$$

$$S_t \stackrel{\text{def}}{=} \{x + z \in H | x + z \in X \text{ and } \|z\| \leq t\} \tag{4.40}$$

and

$$C_t \stackrel{\text{def}}{=} H_t \cap S_t. \tag{4.41}$$

For a given $t > 0$, H_t is the half plane of all vectors $x + z \in H$ such that z satisfies (4.27), irrespective of the constraints $t \geq \|z\|$ and $x + z \in X$, while C_t is the subset of H_t for which these constraints hold.

We again distinguish two cases. The first case is when

$$\|z^l\| \geq \|z^u\|. \tag{4.42}$$

Using the first part of Lemma 2, we deduce that

$$\langle g, z^u \rangle \leq -\mu_3 \alpha(t^u) \leq -\mu_3 \alpha(t^l), \tag{4.43}$$

and therefore, using the inequality $t^l \geq \|z^l\| \geq \|z^u\|$, that the complete segment $[x^l, x^u]$ belongs to the convex set C_{t^l}. Hence (4.27) holds for t^l at every point of the segment $[x^l, x^u] = R_x[x^l, x^p, x^u]$.

The more complicated second case is when (4.42) fails. The proof proceeds by showing the existence of a continuous feasible path between x^l and x^u, depending on the parameter t, such that, for each point on this path, there is a $t \in [t^l, t^u]$ for which (4.27) holds at this point. To find this path, we first define, for all $t \in [t^l, t^u]$,

$$x_t \stackrel{\text{def}}{=} \arg \min_{y \in C_t} \|y - x^u\|_2, \tag{4.44}$$

that is the projection of x^u onto the convex set C_t. We note that both x^l and x_{t^l} belong to C_{t^l}, and hence that the segment $[x^l, x_{t^l}]$ lies in C_{t^l}. We also note that $x^u = x_{t^u} \in C_{t^u}$. Finally, x_t clearly belongs to C_t for all $t \in [t^l, t^u]$, because of (4.44). Furthermore, this set of x_t determines a continuous path, as can be seen by applying Lemma 1 to the minimization problem (4.44). The desired path from x^l to x^u then consists of the segment $[x^l, x_{t^l}]$ followed by the path determined by x_t for $t \in [t^l, t^u]$.

To complete the proof of the theorem for this second case, we use the path just obtained to show that (4.27) holds for some t at every point of $R_x[x^l, x^p, x^u]$. We observe here that this restriction belongs to the plane H. We successively consider three parts of the "restricted path", and show the desired property for each part in turn. This restricted path is that used by the GCP Algorithm. A case where $\|\cdot\| = \|\cdot\|_\infty$ is illustrated in Figure 4.

The first part of the restricted path consists of the segment $[x^l, x^r]$ (where $x^r = R_x[x^p]$) which is the restriction of the segment $[x^l, x^p]$. Using Lemma 21 and the fact that z^p is a multiple of z^l, we deduce that, for each point $y \in [x^l, x^r]$, there exists a t such that (4.27) is satisfied at this point for $z = y - x$. We also note that the same argument implies the existence of $t^p \geq \|z^p\| = \|z^u\|$ such that (4.27) also holds at z^p.

The second part of the restricted path consists of the segment $[x^f, x^u]$, where $x^f = R_x[x^f]$ is the first feasible point on the segment $[x^p, x^u]$. (Note that $[x^f, x^u]$ may be equal to $[x^p, x^u]$ when x^p is feasible or may be reduced to the point x^u if this is the only feasible point in $[x^p, x^u]$.) The segment $[x^f, x^u]$ is also contained in X and is therefore equal to its restriction. Because (4.27) holds with $t = \min[t^p, t^u]$ both for z^p and z^u, it must also hold, with the same t, for all z such that $z = y - x$ where $y \in [x^f, x^u] \subseteq [x^p, x^u]$.

The third part of the restricted path consists of the restriction of the segment $[x^p, x^f]$. If x^p is feasible, then the path reduces to $x^f = x^p$, and the desired property results from the analysis of the first part of the restricted path. Assume therefore that x^p is not feasible. Then the restriction of $[x^p, x^f]$ lies on the intersection of the boundary of X with H. It can therefore be viewed as the prolongation (as defined before Lemma 21) of a part of the path from x^l to x^u defined by the segment $[x^l, x_{tl}]$ followed by $\{x_t | t \in [t^l, t^u]\}$. Lemma 21 then guarantees the existence, for each point $y = x + z$ on the restriction of $[x^p, x^f]$, of a t such that (4.27) holds for z. This finally completes the proof. □

The proof of this last theorem also shows that the path used by the GCP Algorithm is not the only possible one. This can be seen, for example, by choosing $\|\cdot\| = \|\cdot\|_2$, in which case the *projected gradient path* (see [29]) is also acceptable (in the sense that each of its points satisfies (4.12)) and may be different from the restricted path used by the GCP Algorithm.

5 Identification of the correct active set

In this section, we consider the case where the convex set of feasible points X is defined as the intersection of a finite collection of larger convex sets X_i, that is

$$X = \bigcap_{i=1}^{m} X_i. \tag{5.1}$$

We will be interested in the behaviour of the class of algorithms presented in Section 2 as the iterates $\{x_k\}$ approach a limit point x_*. More precisely, if we denote the boundary of an arbitrary convex set Y by $\mathrm{bd}(Y)$, we can define the set of active boundaries, or *active set*, at the point $x \in X$ by

$$A(x) \stackrel{\text{def}}{=} \{i \in \{1, \ldots, m\} | x \in \mathrm{bd}(X_i)\}. \tag{5.2}$$

We note that $A(x)$ may be empty if X has a non-empty interior that contains x. The question we wish to analyze can then be phrased as "Is $A(x_k) = A(x_*)$ for k large enough?"

5.1 The assumptions

Clearly, our present assumptions are too general for such an analysis, and we need to strengthen them both from the algorithmic and the geometric point of view.

We first state precisely the additional conditions that are required in Algorithm 1. The idea is that the active constraints at the GCP x_k^C, indexed by $A(x_k^C)$, should be a good estimate of the constraints active at the limit point x_* when k is large enough, as in [4] and [9]. The test which ensures that the GCP asymptotically picks up the correct active constraints is motivated as follows. Assume that an iterative procedure is used to solve the linearized problem associated with $\alpha_k(t_k)$ in (2.18). When a step $\hat{s}_k^C$ satisfying condition (2.32) is obtained in the course of this iteration, we investigate if the correct active set has been found. If the current step $\hat{s}_k^C$ does not approximately minimize the linearized model *with respect to the constraints in* $A(x_k + \hat{s}_k^C)$, we anticipate that this is because the correct active set has not yet been determined. Consequently, additional constraints may need to be considered. For otherwise, the minimizer may be too far away — at infinity in the case of purely linear constraints. We may then choose to continue our iterative procedure. On the other hand, if $\hat{s}_k^C$ approximately minimizes the linearized model with respect to this restricted set of constraints, we may hope that the correct active set has been identified. In the worst case, this may result in finally solving the linearized problem exactly: at the solution $\hat{s}_k^C$, we know that (2.32) obviously holds, but also that this step solves the relaxed version of the same problem *where all constraints that are not in* $A(x_k + \hat{s}_k^C)$ *have been discarded.* This technique motivates our next assumption, in which we require that not only (2.32) holds at s_k^C, but also that this step approximately minimizes the linearized model with respect to the constraints in $A(x_k^C)$.

More precisely, if the quantity $\alpha_k^C(t)$ is defined, for a given x_k^C and for all $t \geq 0$, by

$$\alpha_k^C(t) \stackrel{\text{def}}{=} \Big| \min_{\substack{x_k + d \in X_k^C \\ \|d\|_{(k)} \leq t}} \langle g_k, d \rangle \Big|, \tag{5.3}$$

where

$$X_k^C \stackrel{\text{def}}{=} \bigcap_{i \in A(x_k^C)} X_i, \tag{5.4}$$

we can then formulate our assumption as follows.

AS.6 For all k sufficiently large, there exists a strictly positive $t_k \geq \|s_k^C\|_{(k)}$ such that

$$\langle g_k, s_k^C \rangle \leq -\mu_3 \alpha_k^C(t_k), \tag{5.5}$$

for some constant $\mu_3 \in (0, 1]$.

We note that, because $X \subseteq X_k^C$,

$$\alpha_k^C(t) \geq \alpha_k(t) \tag{5.6}$$

for all $t \geq 0$, and hence condition (5.5) is stronger than (2.32): it can therefore replace this condition, for large k, in the formulation of Algorithm 1. (This is the reason why the constant μ_3 has been re-used in (5.5).)

We also note that it is always possible to satisfy (AS.6) and (2.32) together because equality holds in condition (5.6) if x_k^C is chosen as the minimizer of the linearized problem associated with the definition of $\alpha_k(t)$ in (2.18) (see our motivation for (AS.6) above).

Once the correct active constraints have been identified by the GCP, one must then make sure they are not dropped at Step 3 of Algorithm 1. This is ensured by the following condition.

AS.7 For all k sufficiently large,

$$A(x_k^C) \subseteq A(x_k + s_k). \tag{5.7}$$

In a way entirely similar to that used in the proof of Lemma 2, one can deduce the following properties of $\alpha_k^C(t)$ as a function of t.

Lemma 25 *For all $k \geq 0$,*

1. *the function $t \mapsto \alpha_k^C(t)$ is continuous and nondecreasing for $t \geq 0$,*
2. *the function $t \mapsto \frac{\alpha_k^C(t)}{t}$ is nonincreasing for $t > 0$.*

By analogy with (3.21), we can also define

$$\alpha_k^C \stackrel{\text{def}}{=} \alpha_k^C(1). \tag{5.8}$$

Using this quantity, we obtain the following counterpart of Theorem 9 and Corollary 10.

Theorem 26 *Assume that (AS.1)–(AS.3) and (AS.6) hold. Consider any sequence $\{x_k\}$ produced by Algorithm 1, and assume that $\alpha_k^C > 0$ for a k sufficiently large. Then there exists a constant $c_8 \in (0,1]$ such that*

$$m_k(x_k) - m_k(x_k + s_k) \geq c_8 \alpha_k^C \min\left[1, \Delta_k, \frac{\alpha_k^C}{1 + \omega_k^C}\right], \tag{5.9}$$

for all k sufficiently large. Furthermore, one has that

$$f(x_k) - f(x_{k+1}) \geq \eta_1 c_8 \alpha_k^C \min\left[1, \Delta_k, \frac{\alpha_k^C}{1 + \omega_k^C}\right] \tag{5.10}$$

for all $k \in \mathcal{S}$ sufficiently large such that $\alpha_k^C > 0$.

Proof. The proof is entirely similar to those of Theorem 9 and Corollary 10, with all α_k being replaced by α_k^C, Lemma 2 replaced by Lemma 25 and the references to (2.32) by references to (5.5). $\square$

We note that we can then pursue the development of Section 3.3 using α_k^C instead of α_k, and deduce a counterpart of Theorem 14.

Theorem 27 *Assume (AS.1)–(AS.4) and (AS.6) hold. Then, if $\{x_k\}$ is a sequence of iterates generated by Algorithm 1, one has that*

$$\liminf_{k \to \infty} \alpha_k^C = 0. \tag{5.11}$$

Let us now examine the geometry of the feasible set. The relation (5.1) does not actually add any structure to X, because X_1 can obviously be chosen as X itself, and all other X_i ($i > 1$) can be chosen as identical to $\mathbf{R}^n$. We therefore need to specify further the nature of the description (5.1).

AS.8 For all $i \in \{1, \ldots, m\}$, the convex set X_i is defined by

$$X_i = \{x \in \mathbf{R}^n | h_i(x) \geq 0\}, \tag{5.12}$$

where the function h_i is from $\mathbf{R}^n$ into $\mathbf{R}$ and is continuously differentiable.

We note that the active set at $x \in X$ is now given by

$$A(x) = \{i \in \{1, \ldots, m\} | h_i(x) = 0\}. \tag{5.13}$$

We temporarily restrict ourselves to the case where only inequality constraints are present. This is indeed the case where the constraints identification problem is most apparent. We will discuss the introduction of linear equality constraints in Section 7.2.

We will use the strong constraint qualification based on the independence of the constraint normals at the limit points of the sequence of iterates $\{x_k\}$ generated by Algorithm 1. We first define L to be the set of all limit points of this sequence. Clearly, L is compact because of (AS.1).

AS.9 For all $x_* \in L$, the vectors $\{\nabla h_i(x_*)\}_{i \in A(x_*)}$ are linearly independent.

(AS.8) and (AS.9) imply that the normal cone at any $x_* \in L$ is polyhedral and of the form

$$N(x_*) = \{y \in \mathbf{R}^n | y = -\sum_{i \in A(x_*)} \lambda_i \nabla h_i(x_*), \lambda_i \geq 0\}. \tag{5.14}$$

We complete our assumptions by requiring Dunn's *nondegeneracy condition* [13] at every limit point $x_* \in L$. Before stating this condition, we recall that the relative interior of a convex set Y (denoted ri$[Y]$) is its interior when Y is regarded as a subset of its affine hull, that is the affine subspace with lowest dimensionality that contains Y (see [26, p. 44] for further details). Using this concept, we now express our condition as follows.

AS.10 For every limit point $x_* \in L$, one has that

$$-\nabla f(x_*) \in \text{ri}[N(x_*)]. \tag{5.15}$$

As discussed in [3], this last condition can be viewed as the generalization of the strict complementarity assumption used in [9] and [18]. It was also used in [2] and in [3] in a similar context. As in [2] and [3], we note that (AS.9), (AS.10) and (5.14) together imply the existence of a unique set of strictly positive multipliers. Thus, for every $x_* \in L$,

$$\nabla f(x_*) = \sum_{i \in A(x_*)} \lambda_i \nabla h_i(x_*), \tag{5.16}$$

for some uniquely defined $\lambda_i > 0$.

We finally assume that the gradient approximations are asymptotically exact.

AS.11

$$\lim_{k \to \infty} \|e_k\|_{[k]} = 0. \tag{5.17}$$

This assumption is not the weakest one for obtaining the results on constraint identification presented below, but its presence simplifies the exposition. A weaker requirement will be discussed in Section 7.

We note that none of the above assumptions require the feasible set to be polyhedral, or even that it has quasi-polyhedral faces (cf. [3]).

5.2 Connected sets of limit points

Using the assumptions presented in the preceding subsection, we examine the properties of the unique connected set of limit points of L containing a given $x_* \in L$, that we denote by L_*. We first show the following remarkable fact.

Lemma 28 *Assume that (AS.1)–(AS.10) hold. Then, for each connected set of limit points L_*, there exists a set $A(L_*) \subseteq \{1, \ldots, m\}$ such that*

$$A(x_*) = A(L_*) \tag{5.18}$$

for all $x_* \in L_*$.

Proof. Consider two limit points $x_*, y_* \in L_*$ such that

$$A(x_*) \neq A(y_*) \tag{5.19}$$

and assume, without loss of generality, that there exists $j \in \{1, \ldots, m\}$ such that $j \in A(y_*)$ but $j \notin A(x_*)$. Because of the path-connectivity of L_*, we know that there exists a continuous path $z(t)$ such that

$$z(0) = x_*, \quad z(1) = y_* \text{ and } z(t) \in L_*, \; \forall t \in [0, 1]. \tag{5.20}$$

The condition (5.19) and the definition of j also ensure the existence of $t_+ \in (0, 1]$ such that

$$j \notin A(z(t)), \; \forall t \in [0, t_+) \text{ and } j \in A(z(t_+)). \tag{5.21}$$

Let us also consider $t_- \in [0, t_+)$ such that $A(z(t))$ is constant, and equal to A_- say, on the interval $[t_-, t_+)$. We now choose a sequence $\{t_j\}$ in the interval $[t_-, t_+)$ and converging to t_+. Equation (5.16) implies that

$$\nabla f(z(t_j)) = \sum_{i \in A_-} \lambda_i^-(t_j) \nabla h_i(z(t_j)) \tag{5.22}$$

for all t_j and for some uniquely defined $\lambda_i^-(t_j) > 0$. We now wish to show by contradiction that the sequences $\{\lambda_i^-(t_j)\}$ are bounded for all $i \in A_-$. Assume indeed that the sequence of vectors $\{\lambda^-(t_j)\}$ is unbounded, where these vectors have $\{\lambda_i^-(t_j)\}_{i \in A_-}$ for fixed j as components. In this case, we can select a subsequence $\{t_\ell\} \subseteq \{t_j\}$ such that

$$\|\lambda^-(t_\ell)\|_2 \longrightarrow \infty \text{ and } \frac{\lambda^-(t_\ell)}{\|\lambda^-(t_\ell)\|_2} \longrightarrow \lambda^\diamond, \tag{5.23}$$

where $\lambda^\diamond$ is normalized and has at least one strictly positive component. We then obtain from (5.22) that

$$\frac{\nabla f(z(t_\ell))}{\|\lambda^-(t_\ell)\|_2} = \sum_{i \in A_-} \frac{\lambda_i^-(t_\ell)}{\|\lambda^-(t_\ell)\|_2} \nabla h_i(z(t_\ell)), \tag{5.24}$$

which gives in the limit that

$$0 = \sum_{i \in A_-} \lambda_i^\diamond \nabla h_i(z(t_+)), \tag{5.25}$$

using the continuity of $z(\cdot)$, $\nabla f(\cdot)$ and $\nabla h_i(\cdot)$. If we now define

$$A_+ \stackrel{\rm def}{=} A(z(t_+)), \tag{5.26}$$

we note that the fact that the set $\{x \in \mathbf{R}^n | A(x) \supseteq A_-\}$ is closed and (5.21) ensure that $A_- \subset A_+$. Therefore, because of (AS.9) and the fact that $z(t_+) \in L$, we may deduce from (5.25) that all the components of $\lambda^\diamond$ are zero, which we just saw is impossible. Hence the sequence $\{\lambda^-(t_j)\}$ must be bounded, as well as the sequences of its components. From each of these component's sequences, we may thus extract converging subsequences with limit points λ_i^-. Using the continuity of $z(\cdot)$, $\nabla f(\cdot)$ and $\nabla h_i(\cdot)$ and taking again the limit in (5.22) for these subsequences, we obtain that

$$\nabla f(z(t_+)) = \sum_{i \in A_-} \lambda_i^- \nabla h_i(z(t_+)). \tag{5.27}$$

On the other hand, (5.16) implies that

$$\nabla f(z(t_+)) = \sum_{i \in A_+} \lambda_i^+ \nabla h_i(z(t_+)) \tag{5.28}$$

for some uniquely defined set of $\lambda_i^+ > 0$. But the fact that $A_- \subset A_+$, ensures that (5.27) and (5.28) cannot hold together. Our initial assumption (5.19) is thus impossible, which proves the lemma. $\Box$

We now define the distance from any vector x to any compact set Y by

$$\mathrm{dist}(x, Y) \stackrel{\rm def}{=} \min_{y \in Y} \|x - y\|_2, \tag{5.29}$$

and the neighbourhood of any compact set Y of radius δ by

$$\mathcal{N}(Y, \delta) \stackrel{\rm def}{=} \{x \in \mathbf{R}^n | \mathrm{dist}(x, Y) \le \delta\}. \tag{5.30}$$

After showing that different active sets cannot appear in a single connected set of limit points, we now show that connected sets of limit points corresponding to different active sets are "well separated".

Lemma 29 *Assume (AS.1)–(AS.10) hold. Then there exists a* $\psi \in (0,1)$ *such that*

$$\mathrm{dist}(x_*, L_*') \ge \psi \tag{5.31}$$

for every $x_* \in L$ *and each compact connected set of limit points* L_*' *such that* $A(L_*') \ne A(x_*)$.

Proof. Consider any $x_* \in L$. To this x_*, we can associate the sets

$$D_i \stackrel{\rm def}{=} \{x \in \mathcal{L} | i \in A(x)\} \tag{5.32}$$

for $i \notin A(x_*)$. For each $x_* \in L_*$, there is only a finite number of such sets, and each of them is compact. Because of Lemma 28, the sets D_i and L_* are disjoint for all $i \notin A(x_*)$. From the compactness of L, we then deduce the existence of $\psi > 0$ such that

$$\min_{x_* \in L} \min_{i \notin A(x_*)} \min_{x \in D_i} \|x_* - x\|_2 \ge \psi. \tag{5.33}$$

(Without loss of generality, we may assume that $\psi < 1$.) Hence the distance from x_* to any $L'_* \subset L$ such that $A(L'_*)$ contains some index $j \notin A(x_*)$ is bounded below by ψ, which then implies the desired result. □

We next show that, for k large enough, every iterate x_k lies in the neighbourhood of a well defined connected set of limit points, and also that all constraints that are not binding for this set are also inactive at x_k.

Lemma 30 *Assume (AS.1)–(AS.10) hold. Assume also that the sequence $\{x_k\}$ is generated by Algorithm 1. Then there exist a $\delta \in (0, \frac{1}{4}\psi)$, $\psi \in (0,1)$, and a $k_1 \geq 0$ such that, for all $k \geq k_1$, there exists a compact connected set of limit points $L_{*k} \subseteq L$ such that*

$$x_k \in \mathcal{N}(L_{*k}, \delta) \tag{5.34}$$

and

$$A(x) \subseteq A(L_{*k}) \text{ for all } x \in \mathcal{N}(L_{*k}, \delta) \cap \mathcal{L} \tag{5.35}$$

Proof. Because of the bounded nature of the sequence $\{x_k\}$ (ensured by (AS.1)), we may divide the complete sequence into a number of subsequences, each of which converges to a given connected set of limit points. For k large enough, x_k therefore lies in the neighbourhood of one such connected set, L_{*k} say. The inclusion (5.34) then follows for δ small enough and for k sufficiently large. We then obtain (5.35) by using (5.33) and imposing the additional requirement that $\delta < \psi/4$. □

We now prove that, if an iterate x_k is close to its associated set of limit points but x_k^C has an incomplete set of active bounds, then α_k^C is bounded away from zero by a small constant independent of k.

Lemma 31 *Assume (AS.1)–(AS.11) hold. Then there exists $k_2 \geq k_1$ (where k_1 is as defined in Lemma 30 with $\delta < \frac{1}{2}$) such that, if there exists $j \in \{1, \ldots, m\}$ with*

$$j \in A(L_{*k}) \text{ and } j \notin A(x_k^C) \tag{5.36}$$

for some $k \geq k_2$, then

$$\alpha_k^C \geq \epsilon_* \tag{5.37}$$

for some $\epsilon_* \in (0,1)$ independent of k and j.

Proof. Consider, for a given $x_* \in L$ with $A(x_*) \neq \emptyset$ and a given $i \in A(x_*)$, the quantity

$$\alpha_{*i}(x_*) \stackrel{\text{def}}{=} |\min_{\substack{x_*+d \in X_{\{i\}} \\ \|d\|_{(k)} \leq 1/2}} \langle \nabla f(x_*), d \rangle|, \tag{5.38}$$

where $X_{\{i\}}$ is defined by

$$X_{\{i\}} \stackrel{\text{def}}{=} \bigcap_{j \in \{1,\ldots,m\} \setminus \{i\}} X_j. \tag{5.39}$$

$\alpha_{*i}(x_*)$ is the magnitude of the decrease obtained by minimizing the linearized objective from x_* in a ball of radius $1/2$ (in the norm $\|\cdot\|_{(k)}$) when dropping the ith (active) constraint. Because of (AS.9) and (AS.10), one has that

$$\alpha_{*i}(x_*) > 0 \tag{5.40}$$

for all choices of $x_* \in L$ and $i \in A(x_*)$. Lemma 1 and the continuity of ∇f also ensure that $\alpha_{*i}(x_*)$ is a continous function of x_*. We first minimize $\alpha_{*i}(x_*)$ on the compact set of all $x_* \in L$ such that $i \in A(x_*)$. For each such set, this produces a strictly positive result. We next take the smallest of these results on all i such that $i \in A(x_*)$ for some $x_* \in L$, yielding a strictly positive lower bound $2\epsilon_*$. In short,

$$\min_i \min_{x_*} \alpha_{*i}(x_*) \geq 2\epsilon_* \tag{5.41}$$

for some $\epsilon_* > 0$.

Consider now $k \geq k_1$. Then, by Lemma 30, we know that we can associate with x_k a unique connected set of limit points L_{*k} such that (5.34) holds. We then choose a particular $x_{*k} \in L_{*k} \cap \mathcal{N}(x_k, \delta)$, for which we have that

$$\{x_{*k} + d \in X_{\{i\}} | \|d\|_{(k)} \leq \frac{1}{2}\} \subset \{x_k + d \in X_{\{i\}} | \|d\|_{(k)} \leq 1\} \tag{5.42}$$

for all $i \in \{1, \ldots, m\}$, where we used the inequality $\delta < \frac{1}{2}$. Observe also that (5.39) imply that

$$X_{\{i\}} \subseteq X_k^C \tag{5.43}$$

for all $i \notin A(x_k^C)$.

Given a $k \geq k_1$ and such that x_k satisfies (5.36), we now distinguish two cases. The first is when $\alpha_k^C \geq \alpha_{*j}(x_{*k})$, in which case (5.37) immediately follows from (5.41). The second is when $\alpha_k^C < \alpha_{*j}(x_{*k})$. If we define d_k^C and d_* as two vectors satisfying

$$\alpha_k^C = -\langle g_k, d_k^C \rangle, \quad \|d_k^C\|_{(k)} \leq 1, \quad x_k + d_k^C \in X_k^C, \tag{5.44}$$

and

$$\alpha_{*j}(x_{*k}) = -\langle \nabla f(x_{*k}), d_* \rangle, \quad \|d_*\|_{(k)} \leq \frac{1}{2}, \quad x_{*k} + d_* \in X_{\{i\}}, \tag{5.45}$$

we can write that

$$\begin{aligned} 0 < \alpha_{*j}(x_{*k}) - \alpha_k^C &= \langle g_k, d_k^C \rangle - \langle \nabla f(x_{*k}), d_* \rangle \\ &= \langle g_k, d_k^C - d_* \rangle + \langle g_k - \nabla f(x_{*k}), d_* \rangle \\ &\leq \langle g_k, d_k^C - d_* \rangle + \tfrac{1}{2} \|g_k - \nabla f(x_{*k})\|_{[k]}, \end{aligned} \tag{5.46}$$

where we used the inequality (2.14). Combining now (5.42), (5.43) and the definitions of α_k^C, d_k^C and d_*, we obtain that

$$\langle g_k, d_k^C \rangle = -\alpha_k^C \leq \langle g_k, d_* \rangle. \tag{5.47}$$

Substituting this last inequality in (5.46), using (AS.11) and the Lipschitz continuity of ∇f (reducing δ if necessary), we can find $k_2 \geq k_1$ sufficiently large such that

$$0 < \alpha_{*j}(x_{*k}) - \alpha_k^C \leq \epsilon_* \tag{5.48}$$

when $k \geq k_2$. The inequality (5.37) then follows again from (5.41). □

5.3 Active constraints identification

We now wish to show that, given a limit point x_*, the set of active contraints at x_*, that is $A(L_*)$, is identified by Algorithm 1 in a finite number of iterations.

We first show that, if the trust region radius is small and the correct active set is not identified at x_k^C (k large enough), which implies, by Lemma 31, that (5.37) holds, then the kth iterate is successful.

Lemma 32 *Assume (AS.1)–(AS.9) hold. Assume furthermore that (5.37) holds and*

$$\beta_k \Delta_k \leq \frac{c_8 \epsilon_* (1-\eta_2)}{c_4} \tag{5.49}$$

for some $k \geq k_2$. Then iteration k is successful ($k \in \mathcal{S}$) and $\Delta_{k+1} \geq \Delta_k$.

Proof. We first observe that (2.28) and the inequalities $c_4 \geq 1$ and $c_8 \leq 1$ imply that

$$\frac{c_8(1-\eta_2)}{c_4} \leq 1. \tag{5.50}$$

Using Theorem 26, (5.37), (5.49), (5.50) and the inequalities $\epsilon_* < 1$ and $\beta_k \geq 1$, one then deduces that

$$f(x_k) - m_k(x_k + s_k) \geq c_8 \epsilon_* \Delta_k. \tag{5.51}$$

But this last inequality, Lemma 11 and (5.49) then ensure that

$$|\rho_k - 1| \leq \frac{c_4 \beta_k \Delta_k}{c_8 \epsilon_*} \leq 1 - \eta_2. \tag{5.52}$$

Hence $\rho_k \geq \eta_2$ and the conclusion of the lemma follows. □

We also need the result that the gradient projected onto the tangent cone at a point y having the correct active set goes to zero as both this point and the iterates tend to a set of limit points.

Lemma 33 *Assume (AS.1)–(AS.11) hold. Consider any subsequence whose indices form $K \subseteq \mathbf{N}$ such that*

$$\lim_{\substack{k \in K \\ k \to \infty}} \mathrm{dist}(x_k, L_*) = 0 \tag{5.53}$$

for some connected set of limit points L_*,

$$\lim_{\substack{k \in K \\ k \to \infty}} \|y_k - x_k\|_{(k)} = 0 \tag{5.54}$$

for some sequence $\{y_k\}_{k \in K}$ such that $y_k \in X$ and

$$A(y_k) = A(L_*) \tag{5.55}$$

for all $k \in K$. Then one has that

$$\lim_{\substack{k \in K \\ k \to \infty}} P_{T(y_k)}(-g_k) = 0. \tag{5.56}$$

Proof. We first note that (5.55), Lemma 1 and the continuity of the constraints' normals imply the continuity of the operators $P_{T(\cdot)}$ and $P_{N(\cdot)}$ as functions of $\{y|A(y) = A(L_*)\}$ in a sufficiently small neighbourhood of L_*. We also observe that the Moreau decomposition of $-g_k$ gives that

$$-g_k = P_{T(y_k)}(-g_k) + P_{N(y_k)}(-g_k). \tag{5.57}$$

This last equation, the limits (5.53), (5.54), (AS.10) and (AS.11) then give (5.56) by continuity. □

Amongst the finitely many active sets $\{A(x_*)\}_{x_*\in L}$, we now consider a maximal one and denote it by A_*. This is to say that $A_* = A(x_*)$ for some $x_* \in L$ and that

$$A_* \not\subset A(y_*) \tag{5.58}$$

for any $y_* \in L$. We are now in position to prove that A_* is identified at least on a subsequence of successful iterations.

Lemma 34 *Assume (AS.1)–(AS.11) hold and that the sequence $\{x_k\}$ is generated by Algorithm 1. Then there exists a subsequence $\{k_i\}$ of successful iterations such that, for i large enough,*

$$A(x_{k_i}) = A_*. \tag{5.59}$$

Proof. We define the subsequence $\{k_j\}$ as the sequence of successful iterations whose iterates approach limit points with active set equal to A_*, that is

$$\{k_j\} \stackrel{\text{def}}{=} \{k \in \mathcal{S} | A(L_{*k}) = A_*\}, \tag{5.60}$$

and assume, for the purpose of obtaining a contradiction, that

$$A(x_{k_j+1}) \neq A_* \tag{5.61}$$

for all j large enough. Assume now, again for the purpose of contradiction, that

$$A_* \subseteq A(x^C_{k_j}) \tag{5.62}$$

for such a j. Using successively (AS.7), (5.61) and Lemma 30, we then deduce that, for j sufficiently large,

$$A_* \subset A(L_{*k_j+1}), \tag{5.63}$$

which is impossible because of (5.58). Hence (5.62) cannot hold, and there must exist a $p_j \in A_* = A(L_{*k_j})$ such that $p_j \notin A(x^C_{k_j})$ for j large enough. From Lemma 31, we then deduce that (5.37) holds for all j sufficiently large. But Theorem 26 and the inequalities $\epsilon_* < 1$ and $\beta_{k_j} \geq 1$ then give that

$$\beta_{k_j}[f(x_{k_j}) - f(x_{k_j+1})] \geq \eta_1 c_8 \epsilon_* \min[\beta_{k_j}\Delta_{k_j}, \epsilon_*], \tag{5.64}$$

for j large enough, and thus, using (AS.5), that

$$\lim_{j\to\infty} \beta_{k_j}\Delta_{k_j} = 0. \tag{5.65}$$

The inequality $\beta_{k_j} \geq 1$ and (2.11) then give that

$$\|s_{k_j}\|_{(k_j)} \leq \nu_1 \Delta_{k_j} \leq \frac{1}{2}\delta < \frac{\psi}{4} \tag{5.66}$$

for j larger than $j_1 \geq 1$, say. But this last inequality, Lemma 29 and Lemma 30 imply that x_{k_j+1} cannot jump to the neighbourhood of any other connected set of limit points with a different active set , and hence x_{k_j+1} belongs to $\mathcal{N}(L_*, \delta)$ again for some L_* such that $A(L_*) = A_*$. The same property also holds for the next successful iterate, x_{k_j+q}, say, and we have that $A(L_{*k_j+q}) = A_*$. Therefore, the subsequence $\{k_j\}$ is identical to the complete sequence of successful iterations with $k \geq k_{j_1}$. Hence we may deduce from (5.65) that

$$\lim_{\substack{k\to\infty \\ k\in\mathcal{S}}} \beta_k \Delta_k = 0. \tag{5.67}$$

In particular, we have that

$$\beta_k \Delta_k \leq \frac{c_8 \gamma_1^2 \epsilon_* (1-\eta_2)}{2c_4} \tag{5.68}$$

for all $k \in \mathcal{S}$ sufficiently large. But the mechanism of the algorithm and (5.67) also give the limit

$$\lim_{k\to\infty} \Delta_k = 0. \tag{5.69}$$

As a consequence, we note that, for k large enough, x_k, x_k^C and $x_k + s_k$ all belong to $\mathcal{N}(L_*, \delta)$ for a single connected set of limit points L_*.

We also note that Lemma 32, the fact that (5.37) now holds for $k \in \mathcal{S}$ and (5.67) together imply that

$$k \in \mathcal{S} \Longrightarrow \Delta_{k+1} \geq \Delta_k \tag{5.70}$$

for k large enough.

We can therefore deduce the desired contradiction from (5.70) and (5.69) if we can prove that all iterations are eventually successful.

Assume therefore that this is not the case. It is then possible to find a subsequence K of sufficiently large k such that

$$k \notin \mathcal{S} \;\text{ and }\; k+1 \in \mathcal{S}. \tag{5.71}$$

Note that, because of (2.45) and of the nondecreasing nature of the sequence $\{\beta_k\}$, one has that

$$\beta_k \Delta_k \leq \frac{1}{\gamma_1} \beta_{k+1} \Delta_{k+1} \leq \frac{c_8 \gamma_1 \epsilon_* (1-\eta_2)}{2c_4} \tag{5.72}$$

for $k \in K$ sufficiently large, where we used (5.68) to deduce the last inequality. Now, if one has that

$$A(x_k^C) \subset A(L_*), \tag{5.73}$$

then Lemmas 31 and 32 together with (5.72) and (2.29) imply that $k \in \mathcal{S}$, which contradicts (5.71). Hence (5.73) cannot hold, and (AS.7) together with Lemma 30 give that

$$A(x_k + s_k) = A(x_k^C) = A(L_*) \tag{5.74}$$

for all $k \in K$ sufficiently large. Observe now that, since $k \notin \mathcal{S}$, one has that $x_{k+1} = x_k$ because of (2.44), and hence, using (2.12), that

$$\begin{aligned} m_{k+1}(x_{k+1}+s_{k+1}) &- m_k(x_k+s_k) = m_{k+1}(x_k+s_{k+1}) - m_k(x_k+s_k) \\ &= \langle g_{k+1}, s_{k+1}\rangle - \langle g_k, s_k\rangle + \tfrac{1}{2}[\|s_{k+1}\|^2_{(k+1)}\omega_{k+1}(m_{k+1}, x_k, s_{k+1}) \\ &\qquad -\|s_k\|^2_{(k)}\omega_k(m_k, x_k, s_k)] \\ &\geq \langle g_{k+1}-g_k, s_{k+1}\rangle + \langle -g_k, s_k - s_{k+1}\rangle - \tfrac{1}{2}\nu_1^2\beta_k\Delta_k^2 - \tfrac{1}{2}\nu_1^2\beta_{k+1}\Delta_{k+1}^2. \end{aligned} \tag{5.75}$$

But, using successively the identity $x_k = x_{k+1}$, the Cauchy-Schwarz inequality, (AS.3), (2.11), (2.13) and (2.45), we have that

$$\begin{aligned} \langle g_{k+1}-g_k, s_{k+1}\rangle &= \langle g_{k+1}-\nabla f(x_k), s_{k+1}\rangle + \langle \nabla f(x_k) - g_k, s_{k+1}\rangle \\ &= \langle e_{k+1}, s_{k+1}\rangle - \langle e_k, s_{k+1}\rangle \\ &\geq -\|e_{k+1}\|_{[k+1]}\|s_{k+1}\|_{(k+1)} - \|e_k\|_{[k+1]}\|s_{k+1}\|_{(k+1)} \\ &\geq -\|s_{k+1}\|_{(k+1)}\left[\|e_{k+1}\|_{[k+1]} + \sigma_4\|e_k\|_{[k]}\right] \\ &\geq -\nu_1\Delta_{k+1}\left[\kappa_1\Delta_{k+1} + \sigma_4\kappa_1\Delta_k\right] \\ &\geq -\nu_1\kappa_1\Delta_{k+1}^2\left[1+\tfrac{\sigma_4}{\gamma_1}\right] \end{aligned} \tag{5.76}$$

for all $k \in K$, and also that

$$\begin{aligned} \langle -g_k, s_k - s_{k+1}\rangle &= \langle P_{T(x_k+s_k)}(-g_k), s_k - s_{k+1}\rangle + \langle P_{N(x_k+s_k)}(-g_k), s_k - s_{k+1}\rangle \\ &\geq -\|P_{T(x_k+s_k)}(-g_k)\|_{[k]}\|s_k - s_{k+1}\|_{(k)} \\ &\qquad -\langle P_{N(x_k+s_k)}(-g_k), P_{T(x_k+s_k)}(s_{k+1}-s_k)\rangle \\ &\geq -\|P_{T(x_k+s_k)}(-g_k)\|_{[k]}\|s_k - s_{k+1}\|_{(k)} \\ &\geq -(\sigma_2 + \tfrac{1}{\gamma_1})\|P_{T(x_k+s_k)}(-g_k)\|_{[k]}\nu_1\Delta_{k+1} \end{aligned} \tag{5.77}$$

for all $k \in K$, where we have used the Moreau decomposition of $-g_k$, the fact that $s_{k+1} - s_k \in T(x_k+s_k)$, (2.14), the fact that the cone $T(x_k+s_k)$ is the polar of $N(x_k+s_k)$, (2.11), (AS.3) and (2.45). Using (2.45) again, (5.75), (5.76), (5.77) and the nondecreasing nature of $\{\beta_k\}$, we also deduce that, for such k,

$$\begin{aligned} &m_{k+1}(x_{k+1}+s_{k+1}) - m_k(x_k+s_k) \\ &\qquad \geq -\nu_1\Delta_{k+1}\left[\kappa_1(1+\tfrac{\sigma_4}{\gamma_1})\Delta_{k+1} + (\sigma_2+\tfrac{1}{\gamma_1})\|P_{T(x_k+s_k)}(-g_k)\|_{[k]} + \tfrac{\nu_1}{2}(1+\tfrac{1}{\gamma_1^2})\beta_{k+1}\Delta_{k+1}\right]. \end{aligned} \tag{5.78}$$

We now observe that, because of (2.37) and (5.69), we have that $\|s_k\|_{(k)}$ tends to zero when k tends to infinity. Applying now Lemma 33 using (5.74) (with $y_k = x_k + s_k$) to the subsequence $k \in K$, we deduce from (5.78), (5.56), (5.69) and (5.67) that

$$m_{k+1}(x_{k+1}+s_{k+1}) - m_k(x_k+s_k) \geq -\frac{1}{2}c_8\epsilon_*\Delta_{k+1} \tag{5.79}$$

for k large enough in K. On the other hand, we can also apply Theorem 26 to iteration $k+1$ and obtain

$$f(x_{k+1}) - m_{k+1}(x_{k+1}+s_{k+1}) \geq c_8\epsilon_*\Delta_{k+1}, \tag{5.80}$$

where we used (5.67), the inequalities $\epsilon_* < 1$ and $\beta k + 1 \geq 1$, and the fact that (5.37) holds for all sufficiently large $k \in \mathcal{S}$. Hence we obtain that

$$\begin{aligned} f(x_k) - m_k(x_k + s_k) &= f(x_{k+1}) - m_{k+1}(x_{k+1} + s_{k+1}) + m_{k+1}(x_{k+1} + s_{k+1}) - m_k(x_k + s_k) \\ &\geq \tfrac{1}{2} c_8 \epsilon_* \Delta_{k+1} \\ &\geq \tfrac{1}{2} c_8 \gamma_1 \epsilon_* \Delta_k \end{aligned} \tag{5.81}$$

for all $k \in K$ sufficiently large. But then, using the definition of ρ_k, Lemma 11 and (5.72), one obtains that

$$|\rho_k - 1| \leq \frac{2c_4}{c_8 \gamma_1 \epsilon_*} \beta_k \Delta_k \leq 1 - \eta_2 \tag{5.82}$$

and hence that $\rho_k \geq \eta_2$ for all $k \in K$ large enough. But this last inequality implies that $k \in \mathcal{S}$, which contradicts (5.71). The condition (5.71) is thus impossible for k sufficiently large. All iterates are eventually successful, which produces the desired contradiction.

As a consequence, (5.61) cannot hold for all j, and we obtain that there exists a subsequence $\{k_p\} \subset \{k_j\}$ such that, for all p,

$$A_* = A(x_{k_p+1}) = A(x_{k_p+q}), \tag{5.83}$$

where $k_p + q$ is the first successful iteration after iteration k_p. The lemma is thus proved if we choose $\{k_i\} = \{k_p + q\}$. $\square$

The last step in our analysis of the active set identification is to show that, once detected, the maximal active set A_* cannot be abandoned for sufficiently large k. This is the essence of the final theorem of this section.

Theorem 35 *Assume that (AS.1)–(AS.11) hold and that the sequence $\{x_k\}$ is generated by Algorithm 1. Then one has that*

$$A(x_*) = A_* \tag{5.84}$$

for all $x_* \in L$, and

$$A(x_k) = A_* \tag{5.85}$$

for all k sufficiently large.

Proof. Consider $\{k_i\}$, the subsequence of successful iterates such that (5.59) holds, as given by Lemma 34. Assume furthermore that this subsequence is restricted to sufficiently large indices, that is $k_i \geq k_2$ for all i. Assume finally that there exists a subsequence of $\{k_i\}$, $\{k_p\}$ say, such that, for each p, there is a j_p with

$$j_p \in A(x_{k_p}) = A_* \text{ and } j_p \notin A(x_{k_p+1}). \tag{5.86}$$

Now Lemma 30, (5.58) and (5.59) give that $A(L_{*k_p}) = A_*$. Using this observation and (AS.7), we obtain that

$$j_p \in A(L_{*k_p}) \text{ and } j_p \notin A(x_{k_p}^C) \tag{5.87}$$

for all p. But Lemma 31 then ensures that

$$\alpha_{k_p}^C \geq \epsilon_* \tag{5.88}$$

for all p. Combining this inequality with Theorem 26 and the relations $\epsilon_* < 1$ and $\beta_{k_p} \geq 1$, one obtains that, for all p,

$$\beta_{k_p}[f(x_{k_p}) - f(x_{k_p+1})] \geq \eta_1 c_8 \epsilon_* \min[\beta_{k_p}\Delta_{k_p}, \epsilon_*]. \tag{5.89}$$

Using (AS.5), we then deduce that

$$\lim_{p\to\infty} \beta_{k_p}\Delta_{k_p} = 0. \tag{5.90}$$

Theorem 26 and the inequalities $\epsilon_* < 1$ and $\beta_{k_p} \geq 1$ then imply that

$$f(x_{k_p}) - m_{k_p}(x_{k_p} + s_{k_p}) \geq c_8 \epsilon_* \Delta_{k_p} \tag{5.91}$$

for all p sufficiently large. On the other hand, we have that, for all k,

$$\begin{array}{rcl} f(x_k) - m_k(x_k + s_k) & \leq & |\langle g_k, s_k\rangle| + \beta_k \|s_k\|^2_{(k)} \\ & \leq & \alpha_k(\|s_k\|_{(k)}) + \beta_k \nu_1^2 \Delta_k^2 \\ & \leq & \frac{\alpha_k(\|s_k\|_{(k)})}{\|s_k\|_{(k)}}\nu_1 \Delta_k + \beta_k \nu_1^2 \Delta_k^2, \end{array} \tag{5.92}$$

where we used (3.29), (3.46), (2.18) and (2.11). Combining (5.91) with (5.92) taken at $k = k_p$, applying the third statement of Lemma 2 and dividing both sides by Δ_{k_p}, we obtain that

$$c_8 \epsilon_* \leq \nu_1 \|P_{T(x_{k_p})}(-g_{k_p})\|_{[k_p]} + \beta_{k_p} \nu_1^2 \Delta_{k_p}. \tag{5.93}$$

Assuming that the sequence $\{x_{k_p}\}$ converges to some x_* in some L_* (or taking a further subsequence if necessary), using (5.90) and Lemma 33 (with $K = \{k_p\}$, $y_k = x_k$ and $A(L_*) = A_*$), we deduce that (5.93) is impossible for p large enough. As a consequence, no such subsequence $\{k_p\}$ exists and we have that, for large i,

$$A_* \subseteq A(x_{k_i+1}) \subseteq A(L_{*k_i+1}), \tag{5.94}$$

where we used Lemma 30 to deduce the last inclusion. But (5.94) and the maximality of A_* impose that

$$A_* = A(x_{k_i+1}) = A(L_{*k_i+1}) \tag{5.95}$$

for i large enough. Hence we deduce that, for sufficiently large i,

$$A(x_{k_i+q}) = A_*, \tag{5.96}$$

where $k_i + q$ is the index of the first successful iteration after iteration k_i. Hence $k_i + q \in \{k_i\}$. We can therefore repeatedly apply (5.96) and deduce that

$$\{k_i\} = \{k \in \mathcal{S} | k \text{ is sufficiently large }\} \tag{5.97}$$

and also that $A(x_k) = A_*$ for all $k \in \mathcal{S}$ large enough, hence proving (5.85). Moreover, A_* is then the only possible active set for the limit points, which proves (5.84). □

6 Convergence to a minimizer

The purpose of this section is to analyse conditions under which the complete sequence of iterates produced by Algorithm 1 can be shown to converge to a single limit point. By Corollary 18 and (AS.11), this limit point is of course critical. We will assume in this section that there are infinitely many successful iterations. Indeed, the convergence of the sequence of iterates is trivial if all iterations are unsuccessful for sufficiently large k.

We define C_*, the set of feasible points whose active set is the same as that of all the limit points, that is

$$C_* \stackrel{\rm def}{=} \{x \in X | A(x) = A_*\}. \tag{6.1}$$

We also define $V(x)$ to be the plane tangent to the constraints indexed by A_*, that is

$$V(x) \stackrel{\rm def}{=} \{z \in \mathbf{R}^n | J_*(x)z = 0\}, \tag{6.2}$$

where $J_*(x)$ is the Jacobian matrix whose rows are equal to $\{\nabla h_i(x)^T\}_{i\in A_*}$.

As we wish to use the second order information associated with the objective function, we must clearly assume that it exists.

AS.12 The objective function $f(\cdot)$ is twice continuously differentiable in an open domain containing X.

We can now prove that if the model curvature along successful steps is asymptotically uniformly positive and if a limit point is an isolated local minimizer, then the complete sequence of iterates converges to this single limit point. In the statement of this result we use the second order sufficiency condition that the Hessian of the objective is positive definite on the tangent plane to the constraints at the solution (see Theorems 6.1 and 6.2 in [4], for instance), which guarantees the isolated character of the minimizer.

Theorem 36 *Assume that (AS.1)–(AS.12) hold, that the sequence $\{x_k\}$ is generated by Algorithm 1 and that the set $\mathcal{S}$ is infinite. Assume also that there is an $\epsilon > 0$ such that*

$$\liminf_{\substack{k\in\mathcal{S}\\ k\to\infty}} \omega_k(m_k, x_k, s_k) \geq \epsilon \tag{6.3}$$

and that, for some $x_* \in L$, $\nabla^2 f(x_*)$ is positive definite on the corresponding tangent plane $V(x_*)$. Then

$$\lim_{k\to\infty} x_k = x_*. \tag{6.4}$$

Proof. We first observe that x_* is a critical point because of (AS.11) and Corollary 18. We consider $\{x_{k_i}\}$, a subsequence of successful iterates converging to x_*. We now choose $\delta_1 > 0$ small enough to ensure the following two conditions. The first is that we can define $Z(x)$, a matrix whose columns form a continuous basis for the tangent plane $V(x)$. The existence of such a basis is ensured in a sufficiently small neighbourhood $\mathcal{N}(x_*, \delta_1)$ of x_* by assumptions (AS.8) and (AS.9). The second condition is that $Z(x)^T\nabla^2 f(x)Z(x)$ (that is $\nabla^2 f(x)$ restricted to the subspace $V(x)$) is uniformly positive definite in $\mathcal{N}(x_*, \delta_1) \cap C_*$.

We now introduce

$$\delta_* \stackrel{\rm def}{=} \frac{\epsilon\delta_1}{4\sigma_2 + \epsilon} < \delta_1 \tag{6.5}$$

and define $f_{\mathcal{P}}$ to be the largest value of the objective such that the level set

$$\mathcal{P} \stackrel{\text{def}}{=} \{x \in \mathcal{N}(x_*, \delta_1) \cap C_* | f(x) \leq f_{\mathcal{P}}\} \subset \mathcal{N}(x_*, \delta_*), \tag{6.6}$$

which is possible because the positive definiteness of $Z(x)^T \nabla^2 f(x) Z(x)$ in $\mathcal{N}(x_*, \delta_1) \cap C_*$ guarantees the strict convexity of $f(x)$ in this set.

We then use Theorem 35 and choose i_1 such that $k_{i_1} \geq 0$ is sufficiently large to guarantee that, for all $i \geq i_1$,

$$x_{k_i} \in \mathcal{P}, \tag{6.7}$$

and also, for all $k \in \mathcal{S}$ with $k \geq k_{i_1}$,

$$x_k \in C_* \tag{6.8}$$

and

$$\omega_k(m_k, x_k, s_k) \geq \frac{1}{2}\epsilon. \tag{6.9}$$

We note that, for $k \geq 0$,

$$s_k \in T(x_k). \tag{6.10}$$

Because of (6.8) and Lemma 33 with $y_k = x_k$, we deduce that

$$\|P_{T(x_k)}(-g_k)\|_{[k]} \leq \delta_* \tag{6.11}$$

for all $k \in \mathcal{S}$ larger than $k_{i_2} \geq k_{i_1}$, say.

Consider now

$$0 > m_{k_i}(x_{k_i} + s_{k_i}) - m_{k_i}(x_{k_i}) = \langle g_{k_i}, s_{k_i} \rangle + \frac{1}{2}\|s_{k_i}\|_{(k_i)}^2 \omega_{k_i}(m_{k_i}, x_{k_i}, s_{k_i}), \tag{6.12}$$

where the equality results from (3.29) and the inequality from the definition of the step s_{k_i}. Using successively (6.12), (6.9), the Moreau decomposition of $-g_{k_i}$ and (6.10), we then deduce that

$$\|s_{k_i}\|_{(k_i)} < \frac{-2}{\omega_{k_i}(m_{k_i}, x_{k_i}, s_{k_i})} \frac{\langle g_{k_i}, s_{k_i} \rangle}{\|s_{k_i}\|_{(k_i)}} \leq \frac{4}{\epsilon} \frac{|\langle P_{T(x_{k_i})}(-g_{k_i}), s_{k_i} \rangle|}{\|s_{k_i}\|_{(k_i)}}, \tag{6.13}$$

for $i \geq i_2$. Hence, using (2.14) and (6.11),

$$\|s_{k_i}\|_{(k_i)} \leq \frac{4}{\epsilon}\|P_{T(x_{k_i})}(-g_{k_i})\|_{[k_i]} \leq \frac{4\delta_*}{\epsilon}, \tag{6.14}$$

for $i \geq i_2$. Using this last relation, the equivalence of norms and the triangle inequality, we obtain that, for such i,

$$\|x_{k_i+1} - x_*\|_2 \leq \|s_{k_i}\|_2 + \|x_{k_i} - x_*\|_2 \leq \left[\frac{4\sigma_2}{\epsilon} + 1\right] \delta_* = \delta_1. \tag{6.15}$$

We now observe that, $k_i \in \mathcal{S}$ implies $f(x_{k_i+1}) < f(x_{k_i}) \leq f_{\mathcal{P}}$. Hence, $x_{k_i+1} \in \mathcal{P}$ and all conditions that were satisfied at x_{k_i} are again satisfied at the next successful iteration after k_i. The argument can therefore be applied recursively to show that

$$x_{k_i+j} \in \mathcal{P} \subset \mathcal{N}(x_*, \delta_1) \tag{6.16}$$

for all $j \geq 1$. Since δ_1 is arbitrarily small, this proves the convergence of the complete sequence $\{x_k\}$ to x_*. □

7 Discussion and extensions

The purpose of this section is to discuss further aspects of the theory presented above, both from the point of view of practical implementation and of interesting theoretical extensions.

7.1 Simple relaxation based tests for inexact projections

A computational difficulty in the framework of Algorithm 1 is the practical enforcement of condition (4.12) in the GCP calculation. Indeed, although the left-hand-side can be readily calculated for any vector z, the right-hand-side contains the quantity $\alpha(t_i)$ which may not be available. However, an upper bound on $\alpha(t_i)$ can often be derived in the following way.

Assume, for example, that we have computed a candidate for the GCP step, z_i say, such that

$$\|z_i\| \leq t_i \quad \text{and} \quad |\langle g, z_i\rangle| = \alpha(\|z_i\|). \tag{7.1}$$

The last of these conditions merely says that z_i minimizes the linearized model in a "ball" of radius $\|z_i\|$. The aim is then to verify that z_i satisfies (4.12), i.e. that z_i gives a large enough reduction of this linearized model compared to that obtained by the minimizer in a ball of radius $t_i \geq \|z_i\|$. Using the definition of $\alpha(t_i)$ and the second part of Lemma 2, it is easy to see that

$$\alpha(t_i) \leq t_i \frac{|\langle g, z_i\rangle|}{\|z_i\|}, \tag{7.2}$$

and (4.12) can be thus guaranteed by checking the stronger condition

$$\langle g, z_i\rangle \leq -\mu_3 t_i \frac{|\langle g, z_i\rangle|}{\|z_i\|}, \tag{7.3}$$

which is equivalent to verifying that

$$\|z_i\| \geq \mu_3 t_i. \tag{7.4}$$

The situation described by (7.1) is far from being unrealistic. It may arise, for example, if $\alpha(t_i)$ is computed by an iterative method starting from x and ensuring (7.1) at each of its iterations.

Another interesting case is when X is polyhedral and $\|\cdot\|_{(k)}$ is the infinity norm for all k. We then find a vector z_i satisfying (4.12) by applying a simplex-like method to the linear programming problem (2.18). Using the fact that the current iterate is feasible and adding slack variables if necessary, this problem can then be rewritten (again dropping the k's) as

$$\min\langle g, d\rangle \tag{7.5}$$

subject to the constraint

$$Ad = 0 \tag{7.6}$$

and the componentwise inequalities

$$l \leq d \leq u \tag{7.7}$$

for some constraint matrix A and some vectors of lower and upper bounds l and u depending on the value of t in (2.18) (or, equivalently, of t_i in (4.12)). If we use a simplex-based method for solving this problem, we calculate, at each iteration of this method, an admissible iterate d_ℓ and an associated admissible basis B_ℓ. It is then easy to compute

$$\pi_\ell = g_{B_\ell}^T B_\ell^{-1} \text{ and } \mu_{\ell j} = \max(0, \pi_\ell A e_j - g_j) \quad (j = 1, \ldots, n), \tag{7.8}$$

where g_{B_ℓ} is the basic part of g and e_j is the j-th vector of the canonical basis of $\mathbf{R}^n$. Remarkably, π_ℓ and the vector μ_ℓ (whose components are the $\mu_{\ell j}$) provide an admissible point for the problem

$$\max -\langle Al, \pi\rangle - \langle u - l, \mu\rangle + \langle g, l\rangle \tag{7.9}$$

subject to

$$\pi A - \mu \leq g \tag{7.10}$$

and

$$\mu \geq 0. \tag{7.11}$$

But this problem is the dual of problem (7.5)–(7.7) after the change of variables $d' = d - l$. As a consequence, we can use the weak duality theorem for linear programming (see [17, p. 40], for instance) and deduce that $\langle Al, \pi_\ell\rangle + \langle u - l, \mu_\ell\rangle - \langle g, l\rangle$ is an upper bound on the value of $\alpha(t_i)$ in (4.12). We may then stop our simplex-based algorithm as soon as

$$|\langle g, d_\ell\rangle| \geq \mu_3 \min_{r=1,\ldots,\ell} [\langle Al, \pi_r\rangle + \langle u - l, \mu_r\rangle - \langle g, l\rangle] \tag{7.12}$$

since this condition implies

$$|\langle g, d_\ell\rangle| \geq \mu_3 \alpha(t_i), \tag{7.13}$$

thus ensuring (4.12) for $z_i = d_\ell$. This technique therefore allows for the inexact solution of the linear program implicit in (2.18).

We also note that the use of interior point methods for linear programming (see [27], for instance) seems quite attractive for solving the same problem in the case where $\|\cdot\|$ is a polyhedral norm and X is polyhedral. These algorithms indeed provide a sequence of feasible approximate solutions together with an estimate of the corresponding duality gaps, which can then be used to stop the process as soon as condition (4.12) is satisfied.

7.2 Constraint identification in the presence of linear equations

We now consider the case where the feasible domain X is defined not only by a set of convex inequalities (as in (AS.8)) but also by a set of independent linear equations of the form

$$p_i(x) = 0, \quad i = 1, \ldots, q, \tag{7.14}$$

where each of the p_i is an affine function from $\mathbf{R}^n$ into $\mathbf{R}$.

We first observe that identifying the active p_i at the solution is trivial: they are all active by definition. The only remaining question is then to examine if their very presence can upset the theory developed in Section 5. We also note that representing an equation by two inequalities of opposite sign does not fit with this theory, because (AS.9) is then automatically violated. We therefore need to discuss this case separately.

The simplest way to exploit the identification theory for inequalities is to "eliminate" the linear equations and view Algorithm 1 as restricted to the affine subspace, W say, where the equations (7.14) hold. We therefore consider the reduction of the original problem to W as follows. Assume that Z is a $n \times n - q$ matrix whose columns form an orthonormal basis of the linear subspace parallel to W. The problem can now be rewritten as

$$\min \hat{f}(y) \stackrel{\text{def}}{=} f(Zy) \tag{7.15}$$

subject to the constraints

$$\hat{h}_i(y) \stackrel{\text{def}}{=} h_i(Zy) \geq 0 \quad (i = 1, \ldots, m), \tag{7.16}$$

where $y \in \mathbf{R}^{n-q}$ (see [15, p. 156] for an introduction to the variable reduction technique). The idea is to show that, if an adapted version of (AS.6)–(AS.11) holds for the problem including the constraints (7.14), then (AS.6)–(AS.11) hold for problem (7.15)–(7.16). The theory of Section 5 then applies without any modification.

(AS.6)–(AS.8) and (AS.11) need not be modified for handling the constraints (7.14). Therefore they also hold for problem (7.15)–(7.16). (AS.9) however requires the following modification.

AS.9b For all $x_* \in L$, the vectors $\{\nabla h_i(x_*)\}_{i \in A(x_*)}$ and $\{\nabla p_i(x_*)\}_{i=1}^{q}$ are linearly independent.

The formal expression of (AS.10) is unchanged, but (AS.8) and (AS.9b) imply that the normal cone $N(x_*)$ is now defined by

$$N(x_*) = \{y \in \mathbf{R}^n | y = -\sum_{i \in A(x_*)} \lambda_i \nabla h_i(x_*) - \sum_{i=1}^{q} \xi_i \nabla p_i(x_*), \lambda_i \geq 0\} \tag{7.17}$$

instead of (5.14).

Defining $x_* \stackrel{\text{def}}{=} Zy_*$ and $\hat{A}(y_*) \stackrel{\text{def}}{=} A(x_*)$, we first note that (AS.9) holds for problem (7.15)–(7.16) as a consequence of (AS.9b).

Theorem 37 *Assume that (AS.9b) holds. Then the vectors $\{\nabla \hat{h}_i(y_*)\}_{i \in \hat{A}(y_*)}$ are linearly independent.*

The proof of this result belongs to the folklore of mathematical programming, and an easy proof is given in the Appendix A.

Similarly, (AS.9b) and (AS.10) with (7.17) imply that (AS.10) holds for problem (7.15)–(7.16), as expressed in the following proposition.

Theorem 38 *Assume that (AS.9b) and (AS.10) hold with (7.17). Then*

$$-\nabla \hat{f}(y_*) \in \text{ri}[\hat{N}(y_*)], \tag{7.18}$$

where

$$\hat{N}(y_*) \stackrel{\text{def}}{=} \{z \in \mathbf{R}^{n-q} | z = -\sum_{i \in \hat{A}(y_*)} \lambda_i \nabla \hat{h}_i(y_*), \lambda_i \geq 0\}. \tag{7.19}$$

The proof of this result can also be found in the Appendix A.

The conclusion of this simple reduction exercise is that all the conditions required for the theory of Section 5 to hold are satisfied for problem (7.15)–(7.16). The presence of equality constraints therefore does not affect the identification of active inequality constraints in a finite number of iterations of Algorithm 1.

7.3 Constraint identification without linear independence of constraint's normals

One may note that (AS.9) is a rather strong constraint qualification, and wonder if it can be weakened without affecting the result that "the correct active set" is identified in a finite number of iterations.

In order to answer this question, we first note that Algorithm 1 and the GCP and RS Algorithms do not depend in any way on the particular parametrization (description) of the feasible set X that is used. The constraints functions h_i were indeed introduced only in (AS.8) and play no role in the theoretical algorithm. As a consequence, one can clearly add redundant constraints of the form

$$r_i(x) \geq 0 \quad (i = 1, \ldots, m_r) \tag{7.20}$$

to the set $\{h_i\}_{i=1}^{m}$ without modifying the result that the algorithm will identify the correct active constraints in the set $\{1, \ldots, m\}$.

Identification of the active redundant constraints in $\{r_i\}_{i=1}^{m_r}$ will then depend on the existence, for each of these constraints, of a set $A_i \subseteq \{1, \ldots, m\}$ such that

$$\{x \in X | A(x) = A_i\} \subseteq \{x \in X | r_i(x) = 0\}. \tag{7.21}$$

If this property holds for r_i and if $A_i = A_*$, then the activity of r_i will clearly be detected in a finite number of iterations.

For example, if $r_i(x)$ is a multiple of $h_j(x)$, say, and if $j \in A_*$, then r_i is identified as an active constraint in a finite number of iterations. Another example is given by the problem

$$\min x + y \tag{7.22}$$

subject to

$$h_1(x, y) = x \geq 0, \quad h_2(x, y) = y \geq 0 \text{ and } r_1(x, y) = x + 4y \geq 0. \tag{7.23}$$

In this case, the constraint r_1 is active if and only if both h_1 and h_2 are active ($A_1 = \{1, 2\}$). It is therefore detected as an active constraint in a finite number of iterations because the activity of h_1 and h_2 is.

On the other hand, if we consider the problem

$$\min y \tag{7.24}$$

subject to

$$h_1(x, y) = y - x^2 \geq 0 \text{ and } r_1(x, y) = y \geq 0, \tag{7.25}$$

we note that the activity of r_1 at the solution may not be detected in a finite number of iterations. This is because there is no subset $A_1 \subseteq \{1, \ldots, m\} = \{1\}$ such that (7.21) holds.

The above arguments show that a "weak" active constraint identification is possible without the assumption of linear independence of the constraints' normals. In order to avoid this assumption and to obtain this identification property more directly, several researchers have used a purely geometrical description of the feasible domain for some less general cases (see [3], [4] and [31]). It would be quite interesting to develop such a geometric theory in our framework. This approach seems indeed possible, because a specialization of our identification results to linear inequalities shows that the "correct active face" of the corresponding convex polytope is identified by Algorithm 1 in a finite number of iterations. This geometric rephrasing of nonlinear constraint identification results is the subject of ongoing research.

7.4 A further discussion on the use of approximate gradients

The technique for handling inexact gradient information, as proposed in Section 2.2, is identical to that analyzed by Toint in [29], but is quite different from that proposed by Carter in [6] for the unconstrained case, where he only requires that, for all $k \geq 0$,

$$\|D_k^{-T} e_k\|_2 \leq \tau \|D_k^{-T} g_k\|_2 \tag{7.26}$$

for some $\tau \in [0, 1-\eta_2)$ and some symmetric positive definite scaling matrices D_k such that the norms $\|D_k^{-T}(\cdot)\|_2$ do satisfy AS.3. Convergence is proved under this remarkably weak condition by using the property that

$$\lim_{\Delta_k \to 0} (1-\rho_k) \leq \lim_{\Delta_k \to 0} \frac{\|D_k^{-T} e_k\|_2}{\|D_k^{-T} g_k\|_2 \cos \vartheta_k} \leq \lim_{\Delta_k \to 0} \frac{\tau}{\cos \vartheta_k}, \tag{7.27}$$

where ϑ_k is the angle between $D_k s_k$ and $-D_k^{-T} g_k$. The next step in Carter's development is to show that ϑ_k tends to zero when the trust region radius Δ_k tends to zero, for a large class of trust region schemes applied on unconstrained problems. The relation (7.27) then implies that $\rho_k \geq \eta_2$ for small enough Δ_k, and hence the kth iteration is successful, the trust region radius increases and the algorithm can proceed.

This line of reasoning unfortunately does not apply to constrained problems, where it may well happen that the negative gradient and its approximation both point outside the feasible domain. As a consequence, if x_k lies on the boundary of X, the accuracy level τ requested for e_k may depend on ϑ_k, which can be bounded away from zero as it depends on the angle of $D_k^{-T} g_k$ with the plane tangent to the constraint boundary at x_k. For example, if one considers the problem

$$\min -2x_1 - 2x_2 \tag{7.28}$$

with the constraints

$$x_1 \leq 0 \text{ and } x_2 \leq 3, \tag{7.29}$$

and if one assumes that $D_k = I$, x_k is the origin and that $m_k(s) = -2s_1 - \beta s_2$ for some $\beta > 0$, it is not difficult to verify that

$$\tau \leq (1-\eta_2) \cos \vartheta_k \leq (1-\eta_2)\beta / \sqrt{4+\beta^2} \tag{7.30}$$

is required in (7.26) for the iteration to be successful with $\Delta_{k+1} \geq \Delta_k$, and this value depends on the geometry of the feasible set at x_k (see Figure 5, where the shaded area corresponds to all steps that produce a model decrease).

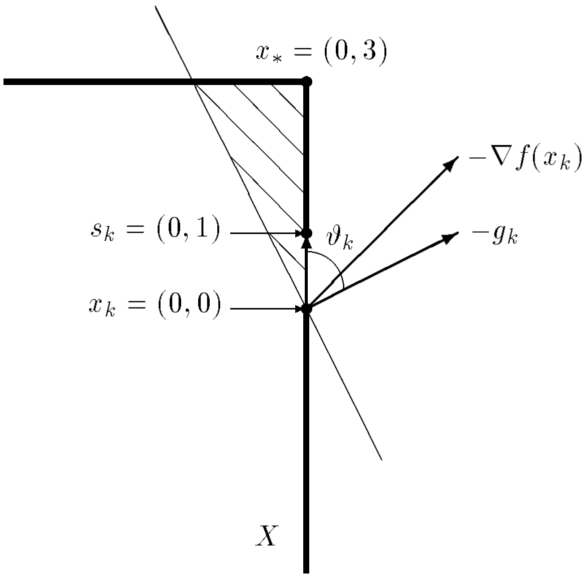

Figure 5: The impact of the feasible set geometry on the angle ϑ_k.

A fixed value, as used in [6], is therefore insufficient to cope with a possibly complex geometry of the feasible set X, and an adaptive scheme, as that suggested by (2.13), is necessary. Furthermore, our purposely broad assumptions (2.37) and (2.38) are too loose to guarantee a well-defined (isotonic, for example) behaviour of ϑ_k as Δ_k tends to zero. Finally, Carter also exploits in his theory the fact that the problem is unconstrained, and thus that $\|D_k^{-T} g_k\|_2$ can be viewed as a criticality measure for the problem at hand. When constraints are present, this is not the case anymore, and the lack of relation between a criticality measure and the right-hand-side of (7.26) makes the direct adaptation of this criterion to the constrained framework quite difficult.

Condition (2.13) also differs from the more abstract condition used by Moré in [19], namely that e_k should tend to zero for a converging sequence of iterates. This condition is related to (3.70) and (3.90) in our analysis.

One attractive feature of Carter's condition (7.26) is the fact that the accuracy requirement is relative to the size of the approximating vector g_k, and hence also to the size of the true gradient $\nabla f(x_k)$, as can be seen as follows. From (7.26), we have that

$$\frac{\|D_k^{-T} g_k\|_2}{\|D_k^{-T} \nabla f(x_k)\|_2} \leq 1 + \frac{\|D_k^{-T} e_k\|_2}{\|D_k^{-T} \nabla f(x_k)\|_2} \leq 1 + \frac{\tau \|D_k^{-T} g_k\|_2}{\|D_k^{-T} \nabla f(x_k)\|_2}, \tag{7.31}$$

and hence, using the fact that $\tau \in [0,1)$,

$$\|D_k^{-T} g_k\|_2 \leq \frac{1}{1-\tau} \|D_k^{-T} \nabla f(x_k)\|_2, \tag{7.32}$$

yielding the desired inequality.

It is important to note that our condition (2.13) can be made relative as well, in the form of the criterion

$$\|e_k\|_{[k]} \leq \min[\kappa_1 \Delta_k, \kappa_2] \, \|g_k\|_{[k]}, \tag{7.33}$$

where $\kappa_2 \in [0,1)$. This relative criterion does in fact imply (2.13). This implication is based on the following simple result.

Lemma 39 *Assume that (AS.3) and (7.33) hold. Then there exists a constant $c_9 > 0$ such that*

$$\|g_k\|_{[k]} \leq c_9 \tag{7.34}$$

for all $k \geq 0$.

Proof. Because of (7.33), we have that

$$\|g_k\|_{[k]} \leq \|\nabla f(x_k)\|_{[k]} + \|e_k\|_{[k]} \leq \frac{1}{\sigma_3} \|\nabla f(x_k)\|_2 + \kappa_2 \|g_k\|_{[k]} \tag{7.35}$$

and hence the compactness of $\mathcal{L}$ implies that (7.34) holds with

$$c_9 = \frac{1}{\sigma_3(1-\kappa_2)} \max_{x \in \mathcal{L}} \|\nabla f(x)\|_2. \tag{7.36}$$

□

As a result of this lemma, we obtain from (7.33) that

$$\|e_k\|_{[k]} \leq c_9 \min[\kappa_1 \Delta_k, \kappa_2] \leq c_9 \kappa_1 \Delta_k, \tag{7.37}$$

and (2.13) therefore holds with κ_1 replaced by $c_9\kappa_1$. The theory developed in this paper is therefore also valid when condition (7.33) is imposed instead of (2.13).

We end this subsection by noting that (AS.11) can be omitted without altering the constraint identification result of Theorem 35 in the case where the complete sequence of iterates converges to a single limit point, x_*, and where the model's gradients, g_k, converge themselves to a well defined limit g_* such that $-g_*$ belongs to the relative interior of the normal cone at x_*. This amounts to replacing (AS.11) by the following.

AS.11b

$$\lim_{k \to \infty} x_k = x_* \tag{7.38}$$

and

$$\lim_{k \to \infty} g_k = g_* \;\text{ and }\; -g_* \in \text{ri}[N(x_*)]. \tag{7.39}$$

The theory of Section 5 must then be adapted accordingly. In particular, the proof of Lemma 31 is modified by replacing $\nabla f(x_*)$ by g_* in (5.38); the minimum over x_* then disappears from (5.41) and the rest of the proof follows.

The second crucial adaptation is the observation that Lemma 33 merely requires that

$$\lim_{\substack{k \in K \\ k \to \infty}} \|e_k\|_{[k]} = 0, \tag{7.40}$$

which is weaker than (AS.11). Condition (7.40) fortunately holds whenever Lemma 33 is used: it is ensured by (5.69) and (2.13) in the proof of Lemma 34, and by (5.90) and (2.13) in the proof of Theorem 35 since $\beta_k \geq 1$ for all k.

Assumption (AS.11b) seems natural if the correct active set is to be identified at all, since the vectors g_k should clearly provide some consistent first order information for this property to hold.

7.5 An extension to noisy objective function values

We note that equation (2.12) (specifying that the model and function values should coincide at the current iterate) is not used anywhere in the convergence theory of Section 3, except in Lemma 11. This leaves some room for a further generalization of Algorithm 1 where not only gradient vectors are allowed to be inexact but also where the objective function values themselves are not known exactly.

Indeed define the quantity E_k by

$$E_k \stackrel{\rm def}{=} \frac{f(x_k) - m_k(x_k)}{m_k(x_k) - m_k(x_k + s_k)}. \tag{7.41}$$

E_k is therefore a measure of the uncertainty of the objective function value relative to the predicted model decrease for the current step s_k. Clearly, if $|E_k|$ is of the order of one or larger, then the predicted model reduction is comparable to the uncertainty in the objective, and the step s_k is then likely to be completely useless: the algorithm might as well stop at x_k. Conversely, if $|E_k|$ is small, then the predicted model reduction is significant compared to the uncertainty in the objective value, and the algorithm may proceed.

This argument is very nicely supported by the theory, as can be seen as follows. We first note that the term $|f(x_k) - m_k(x_k)|$ now appears in the right-hand-side of (3.48) and (3.49), so that (3.47) becomes

$$|f(x_k + s_k) - m_k(x_k + s_k)| \leq |f(x_k) - m_k(x_k)| + c_4 \beta_k \Delta_k^2. \tag{7.42}$$

We then use this inequality instead of (3.47) to obtain that

$$|\rho_{r-1} - 1| \leq 2|E_{r-1}| + \frac{c_4 \beta_{r-1} \Delta_{r-1}}{c_3 \epsilon} \tag{7.43}$$

instead of (3.57), and the right-hand-side of this inequality is smaller than $1 - \eta_2$ provided that we assume the bound

$$|E_k| \leq \frac{1}{2} \phi (1 - \eta_2) \tag{7.44}$$

for all k and for some $\phi \in [0, 1)$, and provided that (3.53) is replaced by

$$\epsilon < \frac{c_4 \beta_0 \Delta_0}{\gamma_1 c_3 (1 - \eta_2)(1 - \phi)} \tag{7.45}$$

and (3.54) by

$$\beta_k \Delta_k \leq \frac{\gamma_1 c_3 (1 - \eta_2)(1 - \phi)}{c_4} \epsilon. \tag{7.46}$$

One then can deduce (3.52) with

$$c_5 = \frac{\gamma_1 c_3 (1-\eta_2)(1-\phi)}{c_4} \epsilon. \tag{7.47}$$

The rest of the global convergence theory of Section 3 then follows as before. Hence we conclude that, provided the relative uncertainty on the objective value E_k satisfies the typically very modest bound (7.44) ($|E_k| \leq 0.1$ for $\phi = 0.8$ and $\eta_2 = 0.75$), the Theorems 14 and 17 still hold.

8 Conclusions and perspectives

In this paper, we have presented a class of trust region algorithms for problems with convex constraints that uses general norms, approximate gradients and inexact projections onto the feasible domain. We have proved global convergence of the iterates generated by this class to critical points. Identification of the final set of active inequality constraints in a finite number of iterations is also shown under slightly stronger assumptions. Interestingly, this theory does not assume the locally polyhedral character of the constrained set.

We have also considered practical implementation issues, including an explicit procedure for computing an approximate Generalized Cauchy Point. Application of these ideas to problems whose linear constraints represent the flow conservation laws in a network is presently under study.

References

[1] M. Bierlaire, Ph. L. Toint and D. Tuyttens, "On iterative algorithms for linear least squares problems with bound constraints" Linear Algebra and Applications (to appear), 1990.

[2] J.V. Burke, "On the identification of active constraints II: the nonconvex case", SIAM Journal on Numerical Analysis (to appear), 1989.

[3] J.V. Burke and J.J. Moré, "On the identification of active constraints", SIAM Journal on Numerical Analysis, vol. 25, pp. 1197–1211, 1988.

[4] J.V. Burke, J.J. Moré and G. Toraldo, "Convergence properties of trust region methods for linear and convex constraints", Mathematical Programming, vol. 47, pp. 305–336, 1990.

[5] R.H. Byrd, R.B. Schnabel and G.A. Schultz, "A trust region algorithm for nonlinearly constrained optimization", SIAM Journal on Numerical Analysis, vol. 24, pp. 1152–1170, 1987.

[6] R.G. Carter, "On the global convergence of trust region algorithms using inexact gradient information", (submitted to SIAM Journal on Numerical Analysis), 1987.

[7] R.G. Carter, "Safeguarding Hessian approximations in trust region algorithms", (submitted to SIAM Journal on Numerical Analysis), 1988.

[8] M.R. Celis, J.E. Dennis and R.A. Tapia, "A trust region strategy for nonlinear equality constrained optimization", in "Numerical Optimization 1984" (P.T. Boggs, R.H. Byrd and R.B. Schnabel, eds.), pp. 71–82, 1985.

[9] A.R. Conn, N.I.M. Gould and Ph.L. Toint, "Global convergence of a class of trust region algorithms for optimization with simple bounds", SIAM Journal on Numerical Analysis, vol. 25, pp. 433–460, 1988. Correction, same journal, vol.26, pp. 764–767, 1989

[10] A.R. Conn, N.I.M. Gould and Ph.L. Toint, "Testing a class of methods for solving minimization problems with simple bounds on the variables", Mathematics of Computation, vol. 50(182), pp. 399–430, 1988.

[11] A.R. Conn, N.I.M. Gould, M. Lescrenier and Ph.L. Toint, "Performance of a multifrontal scheme for partially separable optimization", Report 88/4, Dept. of Mathematics, FUNDP Namur (B), 1988.

[12] J.E. Dennis and R.B. Schnabel, "Numerical methods for unconstrained optimization and nonlinear equations", Prentice-Hall, Englewood Cliffs, 1983.

[13] J.C. Dunn, "On the convergence of projected gradient processes to singular critical points", Journal of Optimization Theory and Applications, vol. 25, pp. 203–216, 1987.

[14] A.V. Fiacco, "Introduction to sensitivity and stability analysis in nonlinear programming", Academic Press, New York, 1983.

[15] P.E. Gill, W. Murray and M.H. Wright, "Practical Optimization", Academic Press, New York, 1981.

[16] W.A. Gruver and E. Sachs, "Algorithmic methods in optimal control", Pitman, Boston, 1980.

[17] J.L. Kennington and R.V. Helgason, "Algorithms for Network Programming", John Wiley and Sons, New York, 1980.

[18] M. Lescrenier, "Partially separable optimization and parallel computing", Report 86/5, Dept. of Mathematics, FUNDP Namur (B), 1986.

[19] J.J. Moré, "Recent developments in algorithms and software for trust region methods", in "Mathematical Programming: The State of the Art" (A. Bachem, M. Grötschel and B. Korte, eds.), pp. 258–287, Springer Verlag, Berlin, 1983.

[20] J.J. Moré, "Trust regions and projected gradients", in "System Modelling and Optimization" (M. Iri and K. Yajima, eds.), Proceedings of the 13th IFIP Conference on System Modelling and Optimization, Tokyo (J), August 31–September 4, 1987, Lecture Notes in Control and Information Sciences, vol. 113, Springer Verlag, Berlin, pp. 1–13, 1988.

[21] J.J. Moré and G. Toraldo, "Algorithms for bound constrained quadratic programming problems", Numerische Mathematik, vol. 55, pp. 377–400, 1989.

[22] J.J. Moreau, "Décomposition orthogonale d'un espace hilbertien selon deux cônes mutuellement polaires", Comptes-Rendus Académie des Sciences (Paris), vol. 255, pp. 238–240, 1962.

[23] M.J.D. Powell, "A New Algorithm for Unconstrained Optimization", in "Nonlinear Programming" (J.B. Rosen, O.L. Mangasarian and K. Ritter, eds.), Academic Press, New York, 1970.

[24] M.J.D. Powell, "On the global convergence of trust region algorithms for unconstrained minimization", Mathematical Programming, vol. 29(3), pp. 297–303, 1984.

[25] M.J.D. Powell and Y. Yuan, "A trust region algorithm for equality constrained optimization", Report DAMTP1986–NA2, Dept. of Applied Mathematics and Theoretical Physics, University of Cambridge (UK), 1986.

[26] R.T. Rockafellar, "Convex Analysis", Princeton University Press, Princeton, 1970.

[27] M.J. Todd, "Recent Developments and New Directions in Linear Programming", in "Mathematical Programming: Recent Developments and Applications", M. Iri and K. Tanabe (eds.), Kluwer Academic Publishers, 1989.

[28] Ph.L. Toint, "Convergence properties of a class of minimization algorithms that use a possibly unbounded sequence of quadratic approximations", Report 81/1, Dept. of Mathematics, FUNDP Namur (B), 1981.

[29] Ph.L. Toint, "Global convergence of a class of trust region methods for nonconvex minimization in Hilbert space", IMA Journal of Numerical Analysis, vol. 8, pp. 231–252, 1988.

[30] A. Vardi, "A trust region algorithm for equality constrained minimization: convergence properties and implementation", SIAM Journal on Numerical Analysis, vol. 22(3), pp. 575–591, 1985.

[31] S. Wright, "Convergence of SQP-like methods for constrained optimization", SIAM Journal on Control and Optimization, vol. 27(1), pp. 13–26, 1989.

[32] Y. Yuan, "Conditions for convergence of trust region algorithms for nonsmooth optimization", Mathematical Programming, vol. 31(2), pp. 220–228, 1985.

[33] E.H. Zarantonello, "Projections on convex sets in Hilbert space and spectral theory", in "Contributions to Nonlinear Functional Analysis" (E.H. Zarantonello, ed.), Academic Press, New York, 1971.

Appendix

A Proof of Theorems 37 and 38

Considering the variable reduction introduced in Section 7.2, we first note that

$$\nabla \hat{f}(y) = Z^T \nabla f(x) \text{ and } \nabla \hat{h}_i(y) = Z^T \nabla h_i(x). \tag{A.1}$$

A.1 Proof of Theorem 38

(AS.10) with (7.17) yields that

$$\nabla f(x_*) = \sum_{i \in A(x_*)} \lambda_i \nabla h_i(x_*) + \sum_{i=1}^{q} \xi_i \nabla p_i(x_*) \tag{A.2}$$

for some $\lambda_i > 0$ and $\xi_i \neq 0$. Applying Z^T to both sides of this relation and noting that $Z^T \nabla p_i(x_*) = 0$ by definition, we obtain the desired conclusion. □

A.2 Proof of Theorem 37

Assume that

$$\sum_{i \in \hat{A}(y_*)} \phi_i \nabla \hat{h}_i(y_*) = 0. \tag{A.3}$$

Premultiplying by Z and using (A.1), we obtain that

$$\sum_{i \in A(x_*)} \phi_i Z Z^T \nabla h_i(x_*) = 0. \tag{A.4}$$

Assume furthermore, for the purpose of contradiction, that

$$\sum_{i \in A(x_*)} \phi_i (I - ZZ^T) \nabla h_i(x_*) \neq 0. \tag{A.5}$$

Since $I - ZZ^T$ is the orthogonal projection onto the subspace spanned by the vectors $\{\nabla p_i(x_*)\}$, we can write that

$$\sum_{i \in A(x_*)} \phi_i (I - ZZ^T) \nabla h_i(x_*) = \sum_{i=1}^{q} \chi_i \nabla p_i(x_*) \tag{A.6}$$

for some χ_i, not all χ_i being zero. Adding (A.4) to (A.6), we obtain

$$\sum_{i \in A(x_*)} \phi_i \nabla h_i(x_*) - \sum_{i=1}^{q} \chi_i \nabla p_i(x_*) = 0, \tag{A.7}$$

which contradicts (AS.9b). Hence (A.5) does not hold, and

$$\sum_{i \in A(x_*)} \phi_i (I - ZZ^T) \nabla h_i(x_*) = 0. \tag{A.8}$$

Summing (A.4) and (A.8), and using (AS.9b), we deduce that $\phi_i = 0$ for all $i \in A(x_*)$, which yields the desired conclusion. □

B Glossary

Symbol	Definition	Purpose
$\|\cdot\|_{(k)}, \|\cdot\|_{[k]}$	Section 2.2	iteration dependent norm and its dual
$\alpha_k(t)$	(2.18)	the magnitude of the maximum linearized model decrease achievable in the intersection of X and a ball of radius t centered at x_k
α_k	(3.21)	$\alpha_k(1)$
$\alpha_k^C(t)$	(5.3)	the magnitude of the maximum linearized model decrease achievable in the intersection of X_k^C and a ball of radius t centered at x_k
α_k^C	(5.8)	$\alpha_k^C(1)$
$\alpha_k[x]$	(3.2)	the magnitude of the maximum linearized objective decrease achievable in the intersection of X and a ball of radius 1 centered at x
β_k	(3.46)	monotonically increasing upper bound on the model's curvature along relevant directions (at iteration k)
$\gamma_1, \gamma_2, \gamma_3$	(2.42), (2.43), (2.45)	contraction/expansion factors for trust region updating
δ	Lemma 30	
Δ_k	(2.11)	the trust region radius
η_1, η_2	(2.40), (2.42), (2.43)	model accuracy levels
κ_1	(2.13)	the model's gradient accuracy relative to the trust region radius
μ_1, μ_2	(2.33), (2.35)	Goldstein-like constants for the projected search
μ_3	(2.32)	the relative projection accuracy
μ_4	(2.38)	model value relaxation w.r.t. value at the GCP
ν_1	(2.11)	outer trust region radius definition parameter
ν_2	(2.31)	inner trust region radius definition parameter
ν_3, ν_4	(2.34)	minimum steplength condition parameter
ρ_k	(2.39)	ratio of actual (function) to predicted (model) decrease
$\sigma_1, \sigma_2, \sigma_3, \sigma_4$	(2.16), (2.17)	constants in the uniform equivalence of the norms $\|\cdot\|_{(k)}$ and $\|\cdot\|_{[k]}$
ψ	Lemma 29	lower bound on the distance between connected sets of limit points
$\omega_k(q, x, v)$	(3.29)	the curvature of the function q from x along v
ω_k^C	(3.34)	$= \omega_k(m_k, x_k, s_k^C)$

Symbol	Definition	Purpose
$A(x)$	(5.2), (5.13)	the active set at x
A_*	(5.58)	the maximal active set at limit points
$\mathrm{bd}(Y)$	Section 5	the boundary of the convex set Y
B_k	(2.11)	the trust region at iteration k
c_1	Theorem 4, (3.3)	uniform equivalence constant for $\alpha_k[x]$
c_2	Lemma 8, (3.31)	uniform upper bound on $\omega_k(f, x_k, s)$
c_3	Theorem 9, (3.44)	model decrease parameter
c_4	Lemma 11, (3.50)	
c_5	Lemma 12, (3.60)	
c_6	(3.74)	
c_7	(3.82)	
c_8	Theorem 26, (5.9)	
c_9	Lemma 39, (7.36)	upper bound on the model's gradient norm
C_t	(4.41)	set of admissible GCP steps of length at most t
C_*	(6.1)	set of feasible points with active set equal to A_*
$\mathrm{dist}(x, Y)$	(5.29)	the distance from x to the compact set Y
D_k	after (7.26)	symmetric positive definite scaling matrix at iteration k
e_k	after (2.13)	difference between the model's and the objective's gradients
E_k	(7.41)	uncertainty of the objective value relative to the predicted model decrease
f	after (2.1)	the objective function
g_k	after (2.12)	the gradient of the model at iteration k, taken at x_k
h_i	(AS.8), (5.12)	inequality constraint functions
H_k	after (2.46)	symmetric approximation to the objective's Hessian at x_k
$J_*(x)$	after (6.2)	the Jacobian matrix of the h_i restricted to rows whose index is in A_* taken at x
k_1	Lemma 30	
k_2	Lemma 31	
K, K^0	(2.4)	cone and its polar

Symbol	Definition	Purpose
L	before (AS.9)	set of all limit points
$\mathcal{L}$	(2.3)	the intersection of the feasible domain with the level set associated with $f(x_0)$
L_f	after (3.32)	the Lipschitz constant of the objective's gradient
L_m	after (4.32)	the Lipschitz constant of the model's gradient
L_*	Section 5.2	the connected set of limit points containing x_*
L_{*k}	in Lemma (30)	the (maximal) connected component of limit points associated with x_k
L'_*	Lemma 29, (5.31)	connected set of limit points *not* containing x_*
m_k	Section 2.2	the model of the objective at iteration k
$N(x)$	(2.6)	the normal cone to X at the feasible point x
$\mathcal{N}(Y, \delta)$	(5.30)	neighbourhood of a compact set Y of radius δ
p_i	(7.14)	linear equality constraint functions
P_X	before (2.5)	the orthogonal projection onto X
r_i	(7.20)	redundant inequality constraint functions
$\mathrm{ri}(Y)$	before (5.15)	relative interior of the convex set Y
$R_x[\cdot]$	(4.1)	the restriction operator
$R_x[x^l, x^p, x^u]$	Section 4.1	restriction of the path $[x^l, x^p, x^u]$
s_k^C	(2.30)–(2.35)	the step from x_k to the Generalized Cauchy Point
s_k	(2.37)–(2.38)	the step at iteration k
$\mathcal{S}$	end of Section 2.3	the set of indices of sucessful iterations
t_k	before (2.30)	upper bound on the length of the GCP step
$T(x)$	(2.7)	the tangent cone to X at the feasible point x
$V(x)$	(6.2)	the linear subspace such that $x + V(x)$ is the tangent plane at x to the constraints indexed by A_*
W	Section 7.2	affine subspace determined by the linear equality constraints p_i
x^f	Section 4.2	
x^l	Section 4.2	
x^p	Section 4.2	
x^r	Section 4.2	

Symbol	Definition	Purpose
x^u	Section 4.2	
x_k	Section 2.2	the iterate of Algorithm 1 at iteration k
$x_k(\theta)$	(2.48)	the projected gradient path starting from x_k
x_k^C	(2.36)	the Generalized Cauchy Point
x_t	(4.44)	the projection of x^u on the convex set C_t
x_*	(3.1)	a critical point
X	after (2.2)	the convex feasible domain
X_i	(5.1), (5.12)	convex sets whose intersection is the feasible domain
X_k^C	(5.4)	relaxation of the feasible domain determined by the constraints active at the GCP
Z	Section 7.2	matrix whose columns form an orthonormal basis of the linear subspace parallel to W
$Z(x)$	before (6.5)	matrix whose columns form a continuous basis for $V(x)$

C Summary of the assumptions

AS.1 The set $\mathcal{L}$ is compact.

AS.2 The objective function $f(x)$ is continuously differentiable and its gradient $\nabla f(x)$ is Lipschitz continuous in an open domain containing $\mathcal{L}$.

AS.3 There exist constants $\sigma_1, \sigma_3 \in (0,1]$ and $\sigma_2, \sigma_4 \geq 1$ such that, for all $k_1 \geq 0$ and $k_2 \geq 0$,

$$\sigma_1 \|x\|_{(k_1)} \leq \|x\|_{(k_2)} \leq \sigma_2 \|x\|_{(k_1)} \quad \text{and} \quad \sigma_3 \|x\|_{[k_1]} \leq \|x\|_{[k_2]} \leq \sigma_4 \|x\|_{[k_1]}$$

for all $x \in \mathbf{R}^n$.

AS.4 The series

$$\sum_{k=0}^{\infty} \frac{1}{\beta_k}$$

is divergent.

AS.5 The limit

$$\lim_{k \to \infty} \beta_k [f(x_k) - f(x_{k+1})] = 0$$

holds.

AS.6 For all k sufficiently large,

$$\langle g_k, s_k^C \rangle \leq -\mu_3 \alpha_k^C(t_k),$$

for some strictly positive $t_k \geq \|s_k^C\|_{(k)}$ and some constant $\mu_3 \in (0,1]$.

AS.7 For all k sufficiently large,

$$A(x_k^C) \subseteq A(x_k + s_k).$$

AS.8 For all $i \in \{1, \ldots, m\}$, the convex set X_i is defined by

$$X_i = \{x \in \mathbf{R}^n | h_i(x) \geq 0\},$$

where the function h_i is from $\mathbf{R}^n$ into $\mathbf{R}$ and is continuously differentiable.

AS.9 For all $x_* \in L$, the vectors $\{\nabla h_i(x_*)\}_{i \in A(x_*)}$ are linearly independent.

AS.10 For every limit point $x_* \in L$,

$$-\nabla f(x_*) \in \text{ri}[N(x_*)].$$

AS.11

$$\lim_{k \to \infty} \|e_k\|_{[k]} = 0.$$

AS.12 The objective function $f(\cdot)$ is twice continuously differentiable in an open domain containing X.

AS.9b For all $x_* \in L$, the vectors $\{\nabla h_i(x_*)\}_{i \in A(x_*)}$ and $\{\nabla p_i(x_*)\}_{i=1}^{q}$ are linearly independent.

AS.11b

$$\lim_{k \to \infty} x_k = x_*, \quad \lim_{k \to \infty} g_k = g_* \text{ and } -g_* \in \text{ri}[N(x_*)].$$