RAL-TR-2003-013

On the Convergence of Successive Linear Programming Algorithms

Richard H. Byrd[1,2], Nicholas I. M. Gould[3,4,5], Jorge Nocedal[6,7] and Richard A. Waltz[6,7]

ABSTRACT
We analyze the global convergence properties of a class of penalty methods for nonlinear programming. These methods include successive linear programming approaches, and more specifically the SLP-EQP approach presented in [1]. Every iteration requires the solution of two trust region subproblems involving linear and quadratic models, respectively. The interaction between the trust regions of these subproblems requires careful consideration. It is shown under mild assumptions that there exist an accumulation point which is a critical point for the penalty function.

[1] Department of Computer Science, University of Colorado, Boulder, CO 80309, USA. Email: richard@cs.colorado.edu .

[2] This work was supported in part by Air Force Office of Scientific Research grant F49620-00-1-0162 and Army Research Office grant DAAG55-98-1-0176.

[3] Computational Science and Engineering Department, Rutherford Appleton Laboratory, Chilton, Oxfordshire, OX11 0QX, England, EU. Email: n.gould@rl.ac.uk .

[4] Current reports available from "http://www.numerical.rl.ac.uk/reports/reports.html".

[5] This work was supported in part by the EPSRC grant GR/R46641.

[6] Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL 60208-3118, USA. Email: nocedal@ece.nwu.edu & rwaltz@ece.nwu.edu .

[7] This work was supported in part by National Science Foundation grants CCR-9987818, ATM-0086579 and CCR-0219438, and Department of Energy grant DE-FG02-87ER25047-A004.

Computational Science and Engineering Department
Atlas Centre
Rutherford Appleton Laboratory
Oxfordshire OX11 0QX
April 29, 2003.

1 Introduction

In this paper we study the global convergence properties of successive linear programming algorithms for nonlinear programming. The problem under consideration is

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f(x) && (1.1\text{a}) \\ \text{subject to} \quad & h_i(x) = 0, \quad i \in \mathcal{E} && (1.1\text{b}) \\ & g_i(x) \geq 0, \quad i \in \mathcal{I}, && (1.1\text{c}) \end{aligned}$$

where the objective function $f : I\!\!R^n \to I\!\!R$, and the constraint functions $h_i : I\!\!R^n \to I\!\!R$, $i \in \mathcal{E}$ $g_i : I\!\!R^n \to I\!\!R$, $i \in \mathcal{I}$, are assumed to be twice continuously differentiable. Our interest is in the case where there are a large number of unknowns.

The class of algorithms studied in this paper solve (1.1) by minimizing an exact penalty function [4, 8] of the form

$$\phi(x, \nu) = f(x) + \nu\|h(x)\| + \nu\|g^-(x)\|, \qquad (1.2)$$

where $\|\cdot\|$ is a (monotonic) norm,

$$g_i^-(x) = \min(g_i(x), 0),$$

and $\nu > 0$ is a parameter which is adaptively chosen so that critical points of (1.1) correspond to those of (1.2). For fixed ν, each iteration of a typical algorithm comprises two phases. First a piecewise linear model of the penalty function ϕ is minimized subject to a trust region bound. The constraints that are active at the solution of this problem determine the current working set. The second phase computes a step by minimizing a quadratic model of the penalty function subject to a set of equality constraints given by the working set, and subject to a trust region bound. A particular instance of this approach is the SLP-EQP algorithm proposed by Fletcher and Sainz de la Maza [7].

The main purpose of this article is to establish the global convergence of this class of penalty methods. The analysis will be phrased in the general context of composite nonsmooth optimization problems of the form

$$\min \phi(x) \equiv \omega(F(x)),$$

where F is a smooth function from $I\!\!R^n$ to $I\!\!R^p$, and ω is a convex function on $I\!\!R^p$. This includes (1.1)–(1.2) as a special case as well as several other important problems, such as linear and nonlinear fitting. Notice that, in the context of (1.1)–(1.2), this analysis presupposes that the penalty parameter ν has been fixed at a sufficiently large value such that critical points of (1.1) correspond to those of (1.2), and we will not consider here suitable mechanisms to ensure that this is so. In practice, ν will be adjusted a finite number of times as the iteration proceeds with this in mind.

This article is a companion to [1], which presents an actual implementation of this penalty approach in which the ℓ_1 norm is used to define the penalty function (1.2). As a result of this choice of norm (and by selecting an ℓ_∞-norm trust-region), the linear phase consists of solving the (linear programming) problem

$$\begin{array}{ll} \underset{d_{\text{LP}}}{\text{minimize}} & \ell(d^{\text{LP}}) \\ \text{subject to} & \|d^{\text{LP}}\|_\infty \le \Delta_{\text{LP}}. \end{array}$$

where

$$\ell(d) = \nabla f(x)^T d + \nu \sum_{i\in\mathcal{E}} |h_i(x) + \nabla h_i(x)^T d| + \nu \sum_{i\in\mathcal{I}} \max(0, -g_i(x) - \nabla g_i(x)^T d).$$

The working set $\mathcal{W}$ is subsequently defined as the set of constraints which are active at the solution of this problem if these constraints are linearly independent, or otherwise, some linearly independent subset of these.

The quadratic phase computes a step d by solving a (quadratic programming) problem of the form

$$\begin{array}{lll} \underset{d}{\text{minimize}} & \frac{1}{2}d^T H(x,\lambda)d + \nabla f(x)^T d & (1.3\text{a}) \\ \text{such that} & h_i(x) + \nabla h_i(x)^T d = 0, \quad i \in \mathcal{E}\cap\mathcal{W} & (1.3\text{b}) \\ & g_i(x) + \nabla g_i(x)^T d = 0, \quad i \in \mathcal{I}\cap\mathcal{W} & (1.3\text{c}) \\ & \|d\|_2 \le \Delta, & (1.3\text{d}) \end{array}$$

where H is the Hessian of the Lagrangian of the nonlinear program (1.1) (or some symmetric approximation of it) and λ is a vector of Lagrange multipliers. Notice that in this phase, an ℓ_2-norm trust region is used. The overall step taken by the algorithm is a linear combination of d_{LP} and the solution to (1.3) that guarantees a decrease in a quadratic approximation to (1.2). If this step does not decrease the objective, both trust regions are reduced.

Unlike Fletcher and Sainz de la Maza [7] our algorithm imposes a trust-region restriction on the second subproblem, and thus permits the use of second derivatives of the objective function and constraints. The two trust region radii, Δ_{LP} and Δ, operate quasi-independently. The update rules we propose are sufficiently weak to offer global convergence guarantees, but also to encourage accurate optimal active-set identification. The numerical results presented in our companion paper [1] suggest that a method of this type holds much promise.

In the next section we describe the algorithm to be analyzed, and in §3 we present the global convergence results. We note that the theory of non-smooth optimization developed by Yuan [12, 14] cannot be applied because in our algorithms the two trust regions influence each other, whereas Yuan assumes that a single trust region is used. The analysis presented here is significantly different from that in the literature due to the effects caused by the interactions between the two trust regions.

2 The Algorithm

We now study the global convergence of the exact penalty algorithms outlined above. As mentioned in the introduction, for greater generality we will state the problem as

$$\min \phi(x) \equiv \omega(F(x)), \tag{2.1}$$

where $F_i(x) = f_i(x)$, $i = 1, \ldots, p$ are smooth functions of x, and $\omega : I\!\!R^p \rightarrow I\!\!R$ is convex but may be nonsmooth. Such problems have been considered by a number of authors over the years [5, 6, 7, 9, 10, 11, 12, 13]. The penalty function (1.2) used to solve the nonlinear program is a special case of (2.1) obtained when $\mathcal{E} \cap \mathcal{I} = \emptyset$ and $\mathcal{E} \cup \mathcal{I} = \{2, \ldots, p\}$ by setting

$$f_1(x) = f(x), \quad f_i(x) = h_i(x),\ i \in \mathcal{E}, \quad f_i(x) = g_i(x),\ i \in \mathcal{I},$$

and defining ω appropriately. is convex but nonsmooth.

Let us describe our class of algorithms in this general setting. It consists of two phases based, respectively, on linear and quadratic models [2, 5, 7, 12] at the current estimate x_k of the minimizer. The linear model is

$$\ell_k(d) = \omega\left(F(x_k) + F'(x_k)d\right), \tag{2.2}$$

where $F'(x)$ is the Jacobian of $F(x)$. Notice that it is only the smooth component of the problem, $F(x)$, which is linearized. By including a second-order term to account for curvature, an appropriate quadratic model is

$$q_k(d) = \omega\left(F(x_k) + F'(x_k)d\right) + \tfrac{1}{2}\langle d, B_k d\rangle. \tag{2.3}$$

for some symmetric, not-necessarily positive definite, B_k.

We will impose trust-region bounds on our models. It is important in practice that we are allowed to use different norms to define the trust regions for the different models. For the linear model, we will use a (polyhedral) trust region of the form $\|\cdot\|_{\text{LP}} \leq \Delta^{\text{LP}}$, while for the quadratic model it will be $\|\cdot\| \leq \Delta$. Since all norms are equivalent in $I\!\!R^n$, there is a constant $\gamma \geq 1$ such that

$$\|d\| \leq \gamma \|d\|_{\text{LP}} \tag{2.4}$$

for all $d \in I\!\!R^n$.

Without further ado, we now define our algorithm.

Algorithm 2.1: Algorithm to minimize $\phi(\mathbf{x}) = \omega(\mathbf{F}(\mathbf{x}))$

Initial data: x_0, $\Delta_0 > 0$, $\Delta_0^{\mathrm{LP}} > 0$, $0 < \rho_u \le \rho_s < 1$, $0 < \kappa_l \le \kappa_u < 1$, $\eta > 0$, $0 < \tau < 1$, and $0 < \theta$.

For $k = 0, 1, \ldots$, until a stopping test is satisfied, perform the following steps.

1. Compute

$$d_k^{\mathrm{LP}} = \underset{\|d\|_{\mathrm{LP}} \le \Delta_k^{\mathrm{LP}}}{\arg\min} \; \ell_k(d)$$

2a. Cauchy point. Compute $\alpha_k \le 1$ as the first member of the sequence $\{\tau^i \min(1, \Delta_k / \|d_k^{\mathrm{LP}}\|\}_{i=0,1,\ldots}$ for which

$$\phi(x_k) - q_k(\alpha_k d_k^{\mathrm{LP}}) \ge \eta \left[\phi(x_k) - \ell_k(\alpha_k d_k^{\mathrm{LP}})\right]. \tag{2.5}$$

Set $d_k^{\mathrm{C}} = \alpha_k d_k^{\mathrm{LP}}$.

2b. Compute d_k so that $\|d_k\| \le \Delta_k$ and

$$q_k(d_k) \le q_k(d_k^{\mathrm{C}}).$$

3. Compute

$$\rho_k = \frac{\phi(x_k) - \phi(x_k + d_k)}{\phi(x_k) - q_k(d_k)}.$$

4a. If $\rho_k \ge \rho_s$, choose

$$\Delta_{k+1} \ge \Delta_k,$$

otherwise set

$$\Delta_{k+1} \in \left[\kappa_l \|d_k\|, \kappa_u \Delta_k\right]. \tag{2.6}$$

4b. If $\rho_k \ge \rho_u$, set

$$x_{k+1} = x_k + d_k,$$

otherwise set

$$x_{k+1} = x_k.$$

5. If $\rho_k \ge \rho_u$ pick

$$\Delta_{k+1}^{\mathrm{LP}} \ge \|d_k^{\mathrm{C}}\|_{\mathrm{LP}} \quad \text{such that} \quad \Delta_{k+1}^{\mathrm{LP}} \le \Delta_k^{\mathrm{LP}} \quad \text{if} \quad \alpha_k < 1. \tag{2.7}$$

otherwise pick

$$\Delta_{k+1}^{\mathrm{LP}} \in [\min(\theta \|d_k\|_{\mathrm{LP}}, \Delta_k^{\mathrm{LP}}), \Delta_k^{\mathrm{LP}}]. \tag{2.8}$$

Step 1 aims to find the largest reduction in the linearized model within its trust region—we refer to this as the *linearized* problem, and attach the suffix LP to quantities associated with it. The intentions here are twofold. First, the aim for certain classes of problem, such as those for which ω is polyhedral, is ultimately to be able to identify which polyhedral components define the minimizer to (2.1). This is not the issue under consideration here, but it does have some ramifications on the design of our algorithm since we hope that our algorithm class is broad enough to permit correct component identification in the polyhedral case.

Secondly, the direction given by d_{LP} is also used to define the Cauchy step which, as in many trust-region methods, is used to guarantee convergence and to ensure that a step will ultimately be taken from a non-critical iterate. This happens since the linear model and true objective may be made arbitrarily close in the event that the trust-region radius shrinks to zero. The convergence analysis in Section 3 confirms this intuition. Because the quadratic model q_k is used in defining the final step d in Step 2b, we define the Cauchy point in Step 2a to give decrease on the quadratic model. Notice that we are simply requiring that the decrease in the quadratic model should be no less than a fraction of that achieved by the linear model for steps of the same length. Since, as we shall see in Section 3, this step is sufficient to ensure convergence, Step 2b simply allows us to pick a step that gives at least as much reduction in the quadratic model as at its Cauchy point, but also allows the step to expand into a possibly enlarged *master* trust region.

Steps 3 and 4 are standard trust-region acceptance rules [3]. The ratio ρ_k of the actual to the predicted reduction of ϕ is used as a step acceptance criterion. If this ratio is negative, or close to zero, the step is rejected and the overall trust-region radius reduced. Otherwise the step will be accepted and, if ρ_k is close to one, the radius may be enlarged. We say that iteration k is *successful* if $\rho_k \geq \rho_u$. It is *very successful* if $\rho_k \geq \rho_s$.

Step 5 indicates how we plan to manage the radius for the linear model. If the master radius has been reduced, we require that the trust region for the linear model be related to the norm of the overall step (and thus Δ^{LP} will ultimately also be reduced), but no larger than its previous value. For successful iterations, if α_k has not been unduly restricted at the Cauchy point, we attribute some of this success to the linear model and increase the linear-model radius. Conversely, if the iteration was successful, but α_k is small, we have no reason to attribute this success to the step from the linear model, so we ensure that the radius does not increase but it may decrease if it is clearly too large. [1]

3 Convergence Results

In this section, we investigate the global convergence properties of Algorithm 2.1. In order to proceed, we need to make the following assumptions on the problem and the algorithm:

P1. F is continuously differentiable and its derivatives are Lipschitz continuous, with Lipschitz constant λ^{F}, throughout a convex region containing the iterates $\{x_k\}$ generated by Algorithm 2.1.

P2. ω is convex and Lipschitz continuous, with Lipschitz constant λ^{ω}, throughout a compact region containing the values $\{F(x_k)\}$ generated by Algorithm 2.1.

[1] The upper bound of one on α_k in (2.7) is used for simplicity. However this bound can be generalized.

P3. The Hessian matrices B_k in (2.3) are bounded; thus there exists $\beta > 0$ such that $d^T B_k d \leq \beta \|d\|^2$ for all k and all $d \in I\!R^n$.

Assumption P3 is made to simplify the analysis; see [12] for an analysis of a composite nonsmooth optimization algorithm in which B_k is computed by quasi-Newton updating.

Under assumptions P1-P2 it follows immediately that both $\phi(x)$ and $\ell_k(d)$ are Lipschitz continuous, and in particular that

$$|\ell_k(d) - \ell_k(0)| \leq \lambda \|d\|_{\text{LP}} \tag{3.1}$$

for some Lipschitz constant $\lambda > 0$.

The goal of our analysis is to prove that Algorithm 2.1 will find a critical point, i.e., a point where the directional derivative of ϕ is nonegative in all directions. To measure criticality, we follow Yuan [12] and define

$$\Psi_k(\Delta) = \phi(x_k) - \min_{\|d\|_{\text{LP}} \leq \Delta} \ell_k(d), \tag{3.2}$$

which is the optimal decrease in the linear model ℓ_k for a radius of size Δ. For future reference we note that, from assumption P2 and the subsequent convexity of $\ell_k(d)$, we have

$$\phi(x_k) - \ell_k(\alpha d) \geq \alpha[\phi(x_k) - \ell_k(d)] \tag{3.3}$$

for any $\alpha \in [0, 1]$.

The first result, which is well known, shows that $\Psi_k(1)$ may be used to measure criticality.

Lemma 3.1 [12, Lemma 2.1] *Suppose that P1-P2 hold, that*

$$\lim_{\substack{k \to \infty \\ k \in \mathcal{K}}} \Psi_k(1) = 0$$

and that $\{x_k\}_{\mathcal{K}}$ converges to x_*. Then x_* is a critical point of $\phi(x)$.

Our next result provides a lower bound on the achievable reduction in the linearized model for a radius of size Δ relative to that achieved with a radius of one. For the proof of this and the following lemma we define d_Δ to be a minimizer of

$$\min_{\|d\|_{\text{LP}} \leq \Delta} \ell_k(d). \tag{3.4}$$

Lemma 3.2 *Suppose that assumptions P1-P2 hold. Then*

$$\Psi_k(\Delta) \geq \min(\Delta, 1)\Psi_k(1) \tag{3.5}$$

for any scalar $\Delta > 0$.

Proof. Let d_1 be a minimizer of (3.4) when $\Delta = 1$, so that

$$\Psi_k(1) = \phi(x_k) - \ell_k(d_1).$$

There are two cases to consider. First consider the case $\Delta \geq 1$. Since $\|d_1\|_{\mathrm{LP}} \leq \Delta$, the definition (3.2) implies that

$$\Psi_k(\Delta) \geq \phi(x_k) - \ell_k(d_1) = \Psi_k(1),$$

which gives (3.5) in this case.

In the second case, $\Delta < 1$, and we need to show that $\Psi_k(\Delta) \geq \Delta\Psi_k(1)$. By definition of d_1 we have that $\|\Delta d_1\|_{\mathrm{LP}} \leq \Delta$, and so by (3.2) and (3.3),

$$\begin{aligned}\Psi_k(\Delta) &\geq \phi(x_k) - \ell_k(\Delta d_1)\\ &\geq \Delta(\phi(x_k) - \ell_k(d_1))\\ &= \Delta\Psi_k(1).\end{aligned}$$

□

We shall also need the following result which states that, at a non-critical point of ϕ, the trust-region bound for the linearized problem, $\|d_\Delta\|_{\mathrm{LP}} \leq \Delta$, is active whenever the radius Δ is small enough.

Lemma 3.3 *Suppose that assumptions P1-P2 hold (and thus that there is a Lipschitz constant λ for which (3.1) holds) and that $\Psi_k(1) \neq 0$. Then if d_Δ is a minimizer of (3.4),*

$$\|d_\Delta\|_{LP} \geq \min(\Delta, \frac{\Psi_k(1)}{\lambda}). \tag{3.6}$$

Proof. As before, let d_1 denote a minimizer of (3.4) when $\Delta = 1$. Suppose that $\|d_\Delta\|_{\mathrm{LP}} < \Psi_k(1)/\lambda$. Then (3.1) gives that

$$\ell_k(d_\Delta) \geq \ell_k(0) - \lambda\|d_\Delta\|_{\mathrm{LP}} > \ell_k(0) - \Psi_k(1) = \ell_k(d_1). \tag{3.7}$$

If $\Delta \geq 1$ this contradicts our definition of d_Δ as a minimizer of (3.4), so we must have $\|d_\Delta\|_{\mathrm{LP}} \geq \Psi_k(1)/\lambda$ and thus (3.6) in this case. If $\Delta < 1$ then (3.7) and the convexity of ℓ_k imply that ℓ_k is strictly decreasing along a line from d_Δ to d_1 (at least initially). Therefore, since d_Δ minimizes ℓ_k, it cannot lie in the strict interior of the trust region $\|d\|_{\mathrm{LP}} \leq \Delta$, and hence $\|d_\Delta\|_{\mathrm{LP}} = \Delta$. □

The next result provides a lower bound on the achievable reduction in the quadratic model in terms of the stepsize, the trust-region radius for the linearized problem and our criticality measure.

Lemma 3.4 *Suppose that assumptions P1-P2 hold. Then the model decrease satisfies*

$$\phi(x_k) - q_k(d_k) \geq \phi(x_k) - q_k(d_k^C) \geq \eta\alpha_k\Psi_k(\Delta_k^{LP}) \geq \eta\alpha_k \min(\Delta_k^{LP}, 1)\Psi_k(1).$$

Proof. The first inequality follows directly from the requirement in step 2b of Algorithm 2.1. To prove the second, note that inequality (3.3) and the requirement in step 2a give that

$$\begin{aligned} \phi(x_k) - q_k(d_k^{\mathrm{C}}) = \;& \phi(x_k) - q_k(\alpha_k d_k^{\mathrm{LP}}) \;\geq \eta\,[\phi(x_k) - \ell_k(\alpha_k d_k^{\mathrm{LP}})] \\ \geq\;& \eta\alpha_k\,[\phi(x_k) - \ell_k(d_k^{\mathrm{LP}})] \;=\; \eta\alpha_k\Psi_k(\Delta_k^{\mathrm{LP}}). \end{aligned}$$

The third inequality follows immediately from Lemma 3.2. □

We will also require an upper bound on the achievable reduction in the objective function, ϕ.

Lemma 3.5 *Suppose that assumptions P1-P3 hold. Then*

$$|q_k(d_k) - \phi(x_k + d_k)| \leq M\|d_k\|^2$$

for some positive constant M.

Proof. Since $F \in C^1$ with Lipschitz derivatives, assumption P1 implies that

$$\|F(x_k + d_k) - F(x_k) - F'(x_k)d_k\| \leq \lambda^{\mathrm{F}}\|d_k\|^2.$$

Using this, the Lipschitz continuity of ω from assumption P2, and assumption P3 we have

$$\begin{aligned} |q_k(d_k) - \phi(x_k + d_k)| &= |\omega(F(x_k) + F'(x_k)d_k) + \tfrac{1}{2}\langle d_k, B_k d_k\rangle - \omega(F(x_k + d_k))| \\ &\leq \lambda^{\omega}\|F(x_k + d_k) - F(x_k) - F'(x_k)d_k\| + \tfrac{1}{2}\beta\|d_k\|^2 \\ &\leq (\lambda^{\omega}\lambda^{\mathrm{F}} + \tfrac{1}{2}\beta)\|d_k\|^2 \\ &= M\|d_k\|^2 \end{aligned}$$

where $M = \lambda^{\omega}\lambda^{\mathrm{F}} + \frac{1}{2}\beta$. □

The following technical result essentially says that either the Cauchy step is on the boundary of a trust region, or it has a lower bound proportional to the optimality criterion.

Lemma 3.6 *Suppose that assumptions P1-P3 hold. Then at any iteration of Algorithm 2.1*

$$\alpha_k\Delta_k^{LP} \geq \|d_k^{C}\|_{LP} \geq \min\left(\frac{\Delta_k}{\gamma}, \Delta_k^{LP}, \frac{\Psi_k(1)}{\lambda}, \min\left(1, \frac{1}{\Delta_k^{LP}}\right)\frac{2(1-\eta)\tau\Psi_k(1)}{\beta\gamma^2}\right). \tag{3.8}$$

Proof. The first inequality in (3.8) follows immediately since

$$\|d_k^{\mathrm{C}}\|_{\mathrm{LP}} = \alpha_k\|d_k^{\mathrm{LP}}\|_{\mathrm{LP}} \leq \alpha_k\Delta_k^{\mathrm{LP}}.$$

To establish the second inequality, suppose first that the decrease condition (2.5) in step 2a of Algorithm 2.1 is immediately satisfied for $\alpha_k = \min(1, \Delta_k/\|d_k^{\mathrm{LP}}\|)$. Then, using (2.4) and Lemma 3.3,

$$\begin{aligned} \|d_k^{\mathrm{C}}\|_{\mathrm{LP}} = \|\alpha_k d_k^{\mathrm{LP}}\| \;&=\; \min\left(\frac{\Delta_k}{\|d_k^{\mathrm{LP}}\|}, 1\right)\|d_k^{\mathrm{LP}}\|_{\mathrm{LP}} \\ &\geq\; \min\left(\frac{\Delta_k}{\gamma}, \Delta_k^{\mathrm{LP}}, \frac{\Psi_k(1)}{\lambda}\right), \end{aligned} \tag{3.9}$$

which gives the first three terms in (3.8). On the other hand if $\alpha_k < \min(1, \Delta_k/\|d_k^{\rm LP}\|)$, then the decrease condition (2.5) must have been violated for α_k/τ, and so

$$\phi(x_k)-q_k(\alpha_k d_k^{\rm LP}/\tau) = \phi(x_k)-\ell_k(\alpha_k d_k^{\rm LP}/\tau)-\tfrac{1}{2}(\alpha_k/\tau)^2\langle d_k^{\rm LP}, B_k d_k^{\rm LP}\rangle \leq \eta\,[\phi(x_k) - \ell_k(\alpha_k d_k^{\rm LP}/\tau)]\,.$$

Now using Assumption P3, (2.4), (3.3) and Lemma 3.2, this inequality implies that

$$\begin{array}{rcl} \tfrac{1}{2}(\alpha_k/\tau)^2\langle d_k^{\rm LP}, B_k d_k^{\rm LP}\rangle & \geq & (1-\eta)\,[\phi(x_k) - \ell_k(\alpha_k d_k^{\rm LP}/\tau)] \\ \tfrac{1}{2}(\alpha_k/\tau)^2\beta\gamma^2\|d_k^{\rm LP}\|_{\rm LP}^2 & \geq & (1-\eta)(\alpha_k/\tau)\Psi_k(\Delta_k^{\rm LP}) \\ \tfrac{1}{2}(\alpha_k/\tau)\beta\gamma^2\|d_k^{\rm LP}\|_{\rm LP}\,\Delta_k^{\rm LP} & \geq & (1-\eta)\min(\Delta_k^{\rm LP},1)\Psi_k(1) \\ \alpha_k\|d_k^{\rm LP}\|_{\rm LP} & \geq & \dfrac{2(1-\eta)\tau}{\beta\gamma^2}\min\left(1, \dfrac{1}{\Delta_k^{\rm LP}}\right)\Psi_k(1). \end{array} \qquad (3.10)$$

Since $\alpha_k d_k^{\rm LP} = d_k^{\rm C}$, this inequality combined with (3.9) gives the second inequality in (3.8).
□

Our next result is crucial. It provides lower bounds on both the master trust-region radius Δ_k and the length of the Cauchy step at a non-critical iterate in the case where the trust-region radius for the linearized problem stays bounded.

Lemma 3.7 *Suppose Algorithm 2.1 is applied to the problem* (2.1) *and that assumptions P1-P3 hold. Suppose that* $\{\Delta_k^{LP}\}$ *is bounded above, and that* $\Psi_k(1) \geq \delta > 0,\ \forall\ k$. *Then there exists a constant* $\Delta_{min} > 0$ *such that*

$$\Delta_k \geq \Delta_{min} \ \ \textit{and} \ \ \alpha_k\Delta_k^{LP} \geq \frac{\Delta_{min}}{\gamma} \qquad (3.11)$$

for all k.

Proof. By assumption, there exists $\Delta_{\rm max} \geq 1$ such that

$$\Delta_k^{\rm LP} \leq \Delta_{\rm max} \text{ for all } k. \qquad (3.12)$$

This inequality, the assumption $\Psi_k(1) \geq \delta$ and Lemma 3.6 imply

$$\|d_k^{\rm C}\|_{\rm LP} \geq \min\left(\frac{\Delta_k}{\gamma}, \Delta_k^{\rm LP}, \Delta_{\rm crit}\right), \qquad (3.13)$$

where

$$\Delta_{\rm crit} = \min\left(\frac{1}{\lambda}, \frac{2(1-\eta)\tau}{\beta\gamma^2\Delta_{\rm max}}\right)\delta. \qquad (3.14)$$

If the iteration is successful ($\rho_k \geq \rho_u$), the rule (2.7) for choosing $\Delta_k^{\rm LP}$ in Step 5 of the algorithm ensures that $\Delta_{k+1}^{\rm LP} \geq \|d_k^{\rm C}\|$ and therefore

$$\Delta_{k+1}^{\rm LP} \geq \min\left(\frac{\Delta_k}{\gamma}, \Delta_k^{\rm LP}, \Delta_{\rm crit}\right). \qquad (3.15)$$

Let us now consider the case when the iteration is unsuccessful. Using Lemma 3.4 and equation (3.13) we have that

$$\begin{aligned}\phi(x_k) - q_k(d_k) &\geq \phi(x_k) - q_k(d_k^{\mathrm{C}}) &&\geq \eta\alpha_k \min\left(\Delta_k^{\mathrm{LP}}, 1\right)\delta = \eta\alpha_k\Delta_k^{\mathrm{LP}} \min\left(\frac{1}{\Delta_k^{\mathrm{LP}}}, 1\right)\delta \\ &\geq \frac{\eta\delta}{\Delta_{\max}}\alpha_k\Delta_k^{\mathrm{LP}} &&\geq \frac{\eta\delta}{\Delta_{\max}} \min\left(\frac{\Delta_k}{\gamma}, \Delta_k^{\mathrm{LP}}, \Delta_{\mathrm{crit}}\right).\end{aligned} \tag{3.16}$$

From Lemma 3.5 and (3.16) we have that

$$1 - \rho_k \leq \frac{|\phi(x_k + d_k) - q_k(d_k)|}{\phi(x_k) - q_k(d_k)} \leq \frac{M\|d_k\|^2\Delta_{\max}}{\eta\delta \min\left(\frac{\Delta_k}{\gamma}, \Delta_k^{\mathrm{LP}}, \Delta_{\mathrm{crit}}\right)}. \tag{3.17}$$

This implies that $\|d_k\|$ and $(1 - \rho_k)$ are related by the inequality

$$\|d_k\|^2 \geq \frac{(1-\rho_k)\eta\delta}{M\Delta_{\max}} \min\left(\frac{\Delta_k}{\gamma}, \Delta_k^{\mathrm{LP}}, \Delta_{\mathrm{crit}}\right). \tag{3.18}$$

at each step. Now since the iteration is unsuccessful, $\rho_k < \rho_u$ and $1 - \rho_k > 1 - \rho_u$, which, using (2.4) and (3.18), implies

$$\begin{aligned}\theta^2\|d_k\|_{\mathrm{LP}}^2 \geq \frac{\theta^2}{\gamma^2}\|d_k\|^2 &\geq \theta^2\frac{(1-\rho_u)\eta\delta}{\gamma^2 M\Delta_{\max}} \min\left(\frac{\Delta_k}{\gamma}, \Delta_k^{\mathrm{LP}}, \Delta_{\mathrm{crit}}\right) \\ &\geq \min\left(\frac{\Delta_k}{\gamma}, \Delta_k^{\mathrm{LP}}, \Delta_{\mathrm{crit}}, \frac{(1-\rho_u)\eta\theta^2\delta}{\gamma^2 M\Delta_{\max}}\right)^2.\end{aligned}$$

Using this fact and the lower bound in (2.8) we have that, if the step is unsuccessful

$$\Delta_{k+1}^{\mathrm{LP}} \geq \min\left(\frac{\Delta_k}{\gamma}, \Delta_k^{\mathrm{LP}}, \Delta_{\mathrm{crit}}, \frac{(1-\rho_u)\eta\theta^2\delta}{\gamma^2 M\Delta_{\max}}\right). \tag{3.19}$$

Since the right side of (3.19) is clearly less than or equal to the right side of (3.15), which holds when the step is accepted, then (3.19) must hold at each iteration.

We can consider Δ_k in a similar fashion. If Δ_k was decreased because $\rho_k < \rho_s$ then $1 - \rho_k > 1 - \rho_s$ and (3.18) implies

$$\begin{aligned}\frac{\kappa_l^2}{\gamma^2}\|d_k\|^2 &\geq \frac{(1-\rho_s)\eta\kappa_l^2\delta}{\gamma^2 M\Delta_{\max}} \min\left(\frac{\Delta_k}{\gamma}, \Delta_k^{\mathrm{LP}}, \Delta_{\mathrm{crit}}\right) \\ &\geq \min\left(\frac{\Delta_k}{\gamma}, \Delta_k^{\mathrm{LP}}, \Delta_{\mathrm{crit}}, \frac{(1-\rho_s)\eta\kappa_l^2\delta}{\gamma^2 M\Delta_{\max}}\right)^2.\end{aligned}$$

Together with (2.6) this implies

$$\frac{\Delta_{k+1}}{\gamma} \geq \min\left(\frac{\Delta_k}{\gamma}, \Delta_k^{\mathrm{LP}}, \Delta_{\mathrm{crit}}, \frac{(1-\rho_s)\eta\kappa_l^2\delta}{\gamma^2 M\Delta_{\max}}\right). \tag{3.20}$$

Since Δ_k is not reduced when $\rho_k \geq \rho_s$, (3.20) must then hold at each iteration.

Now we can combine the recursions (3.19) and (3.20) to yield

$$\min\left(\frac{\Delta_{k+1}}{\gamma}, \Delta_{k+1}^{\text{LP}}\right) \geq \min\left(\frac{\Delta_k}{\gamma}, \Delta_k^{\text{LP}}, \Delta_{\text{crit}}, \frac{(1-\rho_s)\eta\kappa_l^2\delta}{\gamma^2 M\Delta_{\max}}, \frac{(1-\rho_u)\eta\theta^2\delta}{\gamma^2 M\Delta_{\max}}\right). \tag{3.21}$$

which holds at every iteration. Applying this recursion over the entire sequence implies that for all k

$$\begin{aligned}\min\left(\frac{\Delta_k}{\gamma}, \Delta_k^{\text{LP}}\right) &\geq \min\left(\frac{\Delta_0}{\gamma}, \Delta_0^{\text{LP}}, \Delta_{\text{crit}}, \frac{(1-\rho_s)\eta\kappa_l^2\delta}{\gamma^2 M\Delta_{\max}}, \frac{(1-\rho_u)\eta\theta^2\delta}{\gamma^2 M\Delta_{\max}}\right) \\ &\equiv \Delta_{\text{low}},\end{aligned}$$

Thus we can conclude that $\Delta_k \geq \Delta_{\min} \equiv \gamma\Delta_{\text{low}}$ for all k. It then follows from Lemma 3.6 that

$$\alpha_k \Delta_k^{\text{LP}} \geq \Delta_{\text{low}} = \frac{\Delta_{\min}}{\gamma}$$

for all k. □

This immediately enables us to deduce that if the algorithm is unable to make progress, it must be because it has reached a critical point.

Corollary 3.8 *Suppose that there are only finitely number of iterations for which $\rho_k \geq \rho_u$. Then $x_k = x_*$ for all sufficiently large k, and x_* is a critical point of $\phi(x)$.*

Proof. Step 4 of the algorithm ensures that if there are only a finite number of (successful) iterations for which $\rho_k \geq \rho_u$, then $x_k = x_*$ for all $k > k_0$ for some $k_0 \geq 0$. Moreover, $\Psi_k(1) = \Psi_{k_0}(1)$ for all $k \geq k_0$. Furthermore, as $\rho_k < \rho_u$ for all $k \geq k_0$, the update rules for the trust regions imply that Δ_k converges to zero and Δ_k^{LP} is bounded above for all k. But then $\Psi_k(1) = 0$ for all $k \geq k_0$, since otherwise Lemma 3.7 contradicts the fact that Δ_k converge to zero. It thus follows from Lemma 3.1 that x_* is a critical point of ϕ. □

Finally we are able to state our main global-convergence result.

Theorem 3.9 *Suppose Algorithm 2.1 is applied to the problem* (2.1) *and that P1-P3 hold. Then either*

$$\Psi_l(1) = 0 \ \textit{for some} \ l \geq 0$$

or

$$\lim_{k\to\infty} \phi(x_k) = -\infty$$

or

$$\liminf_{k\to\infty} \Psi_k(1) = 0.$$

Proof. If there are only a finite number of successful iterations, the first of the stated possibilities follows immediately from Corollary 3.8. The second of these possibilities might also occur. Consequently, we need only consider the remaining case where there is an infinite subsequence $\mathcal{K}$ of successful iterations, that is that $\rho_k \geq \rho_u$ for all $k \in \mathcal{K}$, for which $\{\phi(x_k)\}$ is bounded from below.

The proof proceeds by contradiction. Assume there is a constant δ such that $\Psi_k(1) \geq \delta > 0$, $\forall\ k$. We will consider separately the two cases, when the LP trust region radius $\{\Delta_k^{\text{LP}}\}$ is bounded above, and the case when $\{\Delta_k^{\text{LP}}\}$ is unbounded.

Case 1. If $\{\Delta_k^{\text{LP}}\}$ is bounded above, it follows from Lemma 3.7 that $\Delta_k \geq \Delta_{\min} > 0$. For our infinite subsequence $\mathcal{K}$ of successful iterations, Lemmas 3.2, 3.4 and 3.7 give

$$\begin{aligned}\phi(x_k) - \phi(x_{k+1}) &\geq \rho_u(\phi(x_k) - q_k(d_k)) \\ &\geq \eta\alpha_k \min(\Delta_k^{\text{LP}}, 1)\delta \\ &\geq \eta\alpha_k\Delta_k^{\text{LP}} \min(1, 1/\Delta_k^{\text{LP}})\delta \\ &\geq \eta\Delta_{\min}\delta/(\gamma\Delta_{\max}) > 0\end{aligned}$$

for all $k \in \mathcal{K}$, where $\Delta_{\max} > 1$ is the upper bound for Δ_k^{LP}. But then summing this inequality over all $k \in \mathcal{K}$ contradicts the fact that the sequence $\{\phi(x_k)\}$ is bounded from below. Thus Case 1 does not occur.

Case 2. Suppose that the LP trust region radius $\{\Delta_k^{\text{LP}}\}$ is unbounded. Then, since the radius is only incresed in step 5 of Algorithm 2.1 when $\alpha_k \geq 1$, there is an infinite sequence $\mathcal{K}$ such that $\Delta_k^{\text{LP}} > 1$, $\alpha_k \geq 1$ and $\rho_k \geq \rho_u$, for all $k \in \mathcal{K}$. Then from Lemmas 3.2 and 3.4 we have

$$\begin{aligned}\phi(x_k) - \phi(x_{k+1}) &\geq \rho_u(\phi(x_k) - q_k(d_k)) \\ &\geq \rho_u\eta\alpha_k \min(\Delta_k^{\text{LP}}, 1)\Psi_k(1) \\ &\geq \rho_u\eta\Psi_k(1) \\ &\geq \rho_u\eta\delta,\end{aligned}$$

for all $k \in \mathcal{K}$. This again contradicts the assumption that $\{\phi(x)\}$ is bounded from below, and Case 2 cannot occur.

Cases 1 and 2 therefore imply that the assumption $\Psi_k(1) \geq \delta > 0$, $\forall\ k$ must be false which proves the desired result

$$\liminf_{k\to\infty} \Psi_k(1) = 0.$$

□

This result guarantees that, if $\phi(x)$ is bounded below, the criticality criterion $\Psi_k(1)$ eventually becomes arbitrarily small. This implies that if the sequence $\{x_k\}$ is bounded there exists an accumulation point of Algorithm 2.1 which is a critical point for (2.1). However, it does not imply that the optimal polyhedral components are identified in a finite number of steps for polyhedral ω – an important result which will not be covered.

4 Conclusions and Perspectives

In this paper we have proposed a trust-region algorithm for composite non-smooth optimization that uses a combination of linear and quadratic model steps and has separate quasi-autonomous trust-regions to control these. At least one subsequence generated by the algorithm is shown to be globally convergent to a critical point of the problem under modest assumptions.

Our framework for trust-region radius updates is deliberately general. This is because we wished it to apply both in the case of the current implementation of our evolving nonlinear programming code SLIQUE [1] as well as to cover its future evolution.

We have not considered the ultimate convergence rate of the algorithm, nor its ability to identify the optimal active polyhedral components in a finite number of iterations (these two aspects are most likely strongly linked [7]), although we have strong numerical evidence to suggest that the latter does occur and that the convergence rate may thereafter be made to be superlinear. The study of these and other issues is ongoing.

References

[1] R. H. Byrd, N. I. M. Gould, J. Nocedal, and R. A. Waltz. An active set algorithm for nonlinear programming using linear programming and equality constrained subproblems. Technical Report OTC 2002/4, Optimization Technology Center, Northwestern University, Evanston, IL, USA, 2002.

[2] C. M. Chin and R. Fletcher. On the global convergence of an SLP-filter algorithm that takes EQP steps. Numerical Analysis Report NA/199, Department of Mathematics, University of Dundee, Dundee, Scotland, 1999.

[3] A. R. Conn, N. I. M. Gould, and Ph. L. Toint. *Trust-region methods.* SIAM, Philadelphia, 2000.

[4] A. R. Conn and T. Pietrzykowski. A penalty function method converging directly to a constrained optimum. *SIAM Journal on Numerical Analysis*, 14(2):348–375, 1977.

[5] R. Fletcher. *Practical Methods of Optimization: Constrained Optimization*, volume 2, chapter 14: Non-differentiable optimization. J. Wiley and Sons, Chicester and New York, 1981.

[6] R. Fletcher. A model algorithm for composite nondifferentiable optimization problems. *Mathematical Programming Studies*, 17:67–76, 1982.

[7] R. Fletcher and E. Sainz de la Maza. Nonlinear programming and nonsmooth optimization by successive linear programming. *Mathematical Programming*, 43(3):235–256, 1989.

[8] T. Pietrzykowski. An exact potential method for constrained maxima. *SIAM Journal on Numerical Analysis*, 6(2):299–304, 1969.

[9] M.J.D. Powell. General algorithms for discrete nonlinear approximation calculations. In C. K. Chui, L. L. Schumaker, and J. D. Ward, editors, *Approximation Theory IV*, pages 187–218, London, 1983. Academic Press.

[10] S. J. Wright. An inexact algorithm for composite nondifferentiable optimization. *Mathematical Programming*, 44(2):221–234, 1989.

[11] Y. Yuan. An example of only linear convergence of trust region algorithms for nonsmooth optimization. *IMA Journal of Numerical Analysis*, 4(3):327–335, 1984.

[12] Y. Yuan. Conditions for convergence of trust region algorithms for nonsmooth optimization. *Mathematical Programming*, 31(2):220–228, 1985.

[13] Y. Yuan. On the superlinear convergence of a trust region algorithm for nonsmooth optimization. *Mathematical Programming*, 31(3):269–285, 1985.

[14] Y. Yuan. On the convergence of a new trust region algorithm. *Numerische Mathematik*, 70(4):515–539, 1995.