

A Two-Phase Model Algorithm with Global Convergence for Nonlinear Programming ^{1 2}

J. M. MARTÍNEZ ³

November 7, 1996

¹This research was supported by FAPESP (Grant 90-3724-6), CNPq and FAEP-UNICAMP.

²The author is indebted to Sandra Santos, Zdeňek Dostál, Lúcio Santos and Cristina Maciel for their comments on the first draft of this paper.

³Full Professor, Department of Applied Mathematics, IMECC-UNICAMP, University of Campinas, CP 6065, 13081 Campinas SP, Brazil.

Abstract. The family of *feasible* methods for minimization with nonlinear constraints includes Rosen's Nonlinear Projected Gradient Method, the Generalized Reduced Gradient Method (GRG) and many variants of the Sequential Gradient Restoration Algorithm (SGRA). Generally speaking, a particular iteration of any of these methods proceeds in two phases. In the Restoration Phase, feasibility is restored by means of the resolution of an auxiliary nonlinear problem, generally a nonlinear system of equations. In the Minimization Phase, optimality is improved by means of the consideration of the objective function, or its Lagrangian, on the tangent subspace to the constraints. In this paper, minimal assumptions are stated on the Restoration Phase and the Minimization Phase that ensure that the resulting algorithm is globally convergent. The key point is the possibility of comparing two successive nonfeasible iterates by means of a suitable merit function that combines feasibility and optimality. The merit function allows one to work with a high degree of infeasibility at the first iterations of the algorithm. Global convergence is proved and a particular implementation of the model algorithm is described.

Key Words: Nonlinear programming, trust regions, GRG methods, SGRA methods, projected gradients, SQP methods, global convergence.

1 Introduction

We consider the problem

$$\begin{aligned} & \text{Minimize } f(x) \\ & \text{subject to } C(x) = 0, \quad x \in \Omega, \end{aligned} \tag{1}$$

where $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and $C : \mathbb{R}^n \rightarrow \mathbb{R}^m$ are continuously differentiable and $\Omega \subset \mathbb{R}^n$ is a closed and convex set. We denote $A(x) = (\nabla C_1(x), \dots, \nabla C_m(x))^T$ for all $x \in \mathbb{R}^n$.

Our objective is to define a model algorithm of iterative nature for solving (1), such that all the iterates x^k belong to Ω and each iteration consists, essentially, of two phases. In the first phase, a movement towards feasibility is performed, with a mild requirement concerning the reduction of $\|C(x)\|_2$. As a result, an intermediate point $y^{k,i} \in \Omega$ is obtained. In the second phase, optimality is improved by means of a partial minimization of the Lagrangian on the tangent subspace that passes through $y^{k,i}$. In this way, a trial point $z^{k,i} \in \Omega$ is obtained. However, although $y^{k,i}$ is more feasible than x^k , and $z^{k,i}$ is more optimal than $y^{k,i}$ (in the sense that the value of the Lagrangian approximation at $z^{k,i}$ is smaller than its value at $y^{k,i}$), the trial point $z^{k,i}$ could be worse than x^k (less feasible, less optimal, or both). In fact, if $z^{k,i}$ is detected to be worse than x^k , the iteration is restarted ($i \leftarrow i + 1$) allowing a smaller variation of $y^{k,i}$ and $z^{k,i}$ with respect to x^k .

The idea of decomposing the iteration in two phases, one related to optimality on the tangent subspace and the other one related to feasibility, is present in many successful nonlinear programming algorithms both for small and large-scale problems. Among these techniques, the best known are the Nonlinear Gradient Projection algorithm of Rosen (Refs. 1–2), the GRG method (Refs. 3–4) and the Sequential Gradient Restoration Algorithm (SGRA) developed by Miele and coworkers (Refs. 5–9). There exist excellent computational codes for solving large-scale practical optimization problems based on GRG and SGRA ideas. A common feature of GRG and SGRA is that both methods maintain near-feasibility of all the iterates with respect to the constraint $C(x) = 0$.

A common criticism against these *feasible methods* is that, when the constraint set is extremely curved, the strategy of getting close to it from the beginning tends to produce unnecessarily long trajectories to the solution. As a consequence, it is sensible to relax the feasibility requirement, especially at the first iterations of the process.

Our point of view is that feasibility and optimality are essentially different characteristics of the optimal points of (1), that justify a separated treatment at each iteration, as GRG and SGRA do. In most practical problems, a feasible non-optimal solution can be useful, but a non-feasible point satisfying optimality conditions on its level set is completely useless. However, as we mentioned above, the geometry of the feasible set often suggests that the feasibility objective should be abandoned far from the solution, in order to improve efficiency of the whole process. The model algorithm presented in this paper tries to take into account these observations in a natural way. On one hand, at each iteration we seek, first, a reduction of $\|C(x)\|_2$ and, second, a reduction of the Lagrangian on the tangent space. On the other hand, a fixed level of feasibility for the new point is not determined a priori, and the trial point arising from the iteration can be accepted or rejected, according to the value of a merit function that combines feasibility and optimality.

Convergence theories of Sequential Quadratic Programming methods based on penalty merit functions (or augmented Lagrangians) and trust regions have been developed in the last ten years. See, for example, Refs. 10–16. In this paper we also use augmented Lagrangians and trust regions for defining the globalization strategy. So, some ingredients of the theories developed in those papers (especially from Ref. 14) are present in our global convergence theory.

Specific implementations of the model algorithm presented here involve the choice of methods for minimizing $\|C(x)\|_2^2$ subject to $x \in \Omega, \|x - \bar{x}\| \leq \delta$ and for minimizing the Lagrangian with $x \in \Omega, \|x - \bar{x}\| \leq \delta$ and linear constraints. Available practical algorithms for performing efficiently those tasks exist, when Ω is a polyhedron (see Refs. 17–18, etc.) or, even, when Ω is a ball or an n -dimensional ellipsoid (see Ref. 19). However, we do not need that the chosen sub-algorithms should be convergent in any sense, so that many heuristics for specific situations can be incorporated. From the theoretical point of view, we do not use any special form of Ω and the choice of $\|\cdot\|$ is arbitrary. Clearly, when Ω is a polyhedron, the sup-norm should be preferred.

This paper is organized as follows. In Section 2 we give an overall description of the model algorithm, some definitions are introduced and a global motivation for the main steps is commented. In Section 3, a precise detailed description of the method is given. In Section 4 we prove that the model algorithm is well defined. This means that, when the stopping criterion does not hold at x^k , a new point x^{k+1} can be obtained in finite time. In Section 5 we prove that the search direction towards feasibility computed at each iteration of the algorithm tends to zero. In Section 6 it is proved that, for some subsequence, the search direction towards optimality defined by the algorithm also tends to zero. In Section 7 we discuss a specific choice for the feasibility step. A simple computer implementation is described and tested in Section 8. Conclusions are given in Section 9.

2 General Motivation

The algorithm that will be described in Section 3 produces a sequence of points x^k that are feasible with respect to Ω but not necessarily with respect to $C(x) = 0$. Moreover, given the current approximation x^k , the new point x^{k+1} will stand at a distance not larger than δ_k from x^k . More precisely, at the beginning of each iteration a trust region radius $\delta_{k,0} > 0$ is given, and the new approximation is obtained by the following scheme.

Procedure 2.1

Step 1. Set $i \leftarrow 0$.

Step 2. Find a trial point $z^{k,i}$ such that

$$z^{k,i} \in \Omega \quad \text{and} \quad \|z^{k,i} - x^k\| \leq \delta_{k,i}.$$

Step 3. If $z^{k,i}$ is satisfactory, set $x^{k+1} = z^{k,i}$ and terminate the iteration. Else, choose $\delta_{k,i+1} < \delta_{k,i}$, set $i \leftarrow i + 1$ and go to Step 2.

The computation of $z^{k,i}$ is performed in two phases. In the first phase, an intermediate point $y^{k,i} \in \Omega$ is computed such that

$$\|y^{k,i} - x^k\| \leq 0.8\delta_{k,i} \quad , \quad \varphi(x^k) - \varphi(y^{k,i}) \gg 0,$$

where

$$\varphi(x) = \frac{1}{2}\|C(x)\|_2^2$$

for all $x \in \mathbb{R}^n$. The symbol \gg means that the reduction $\varphi(x^k) - \varphi(y^{k,i})$ is at least as large as the reduction produced by a backtracking search along a Cauchy-like direction. So, $y^{k,i}$ is more feasible than x^k . Now, since $\|y^{k,i} - x^k\|$ is strictly smaller than $\delta_{k,i}$, there is place to improve optimality finding a trial point $z^{k,i}$ whose distance to x^k is smaller than $\delta_{k,i}$.

This is done in the second phase of the iteration. Assume that $\lambda^k \in \mathbb{R}^m$ is an arbitrary approximation of the Lagrange multiplier vector of the problem and define the Lagrangian function

$$\ell(x, \lambda) = f(x) + \langle C(x), \lambda \rangle$$

for all $x \in \mathbb{R}^n, \lambda \in \mathbb{R}^m$. Let us define the feasible tangent linear manifold that passes through $y^{k,i}$ as:

$$\pi_{k,i} = \{z \in \Omega \mid A(y^{k,i})[z - y^{k,i}] = 0\}. \quad (2)$$

In the second phase we seek $z^{k,i} \in \pi_{k,i}$ such that $\|z^{k,i} - x^k\| \leq \delta_{k,i}$ and

$$\ell(y^{k,i}, \lambda^k) - \ell(z^{k,i}, \lambda^k) \gg 0$$

where, again, the symbol \gg means that the involved reduction is as large as the reduction produced by a Cauchy direction and a backtracking search.

Therefore, $y^{k,i}$ is more feasible than x^k and $z^{k,i}$ is more optimal than $y^{k,i}$. However, $z^{k,i}$ could be less feasible and/or less optimal than x^k . In these conditions, we need a merit function for comparing the quality of $z^{k,i}$ with the quality of x^k . We decided to use the Augmented Lagrangian – like function introduced in Ref. 14, given by

$$\psi(x, \lambda) = \theta\ell(x, \lambda) + (1 - \theta)\varphi(x),$$

for some $\theta \in (0, 1]$. So, $\psi(x, \lambda)$ is a convex combination of $\ell(x, \lambda)$ and $\varphi(x)$ and our objective will be to find $z^{k,i}$ and $\lambda^{k,i}$ such that $\psi(z^{k,i}, \lambda^{k,i}) \ll \psi(x^k, \lambda^k)$. Unhappily, this is not always possible, even for arbitrary small $\delta_{k,i}$, unless we make a careful choice of the penalty parameter θ . However, let us show that a suitable choice of θ guarantees a sufficient decrease of the merit function at each iteration.

Observe that

$$\begin{aligned} \ell(z^{k,i}, \lambda^{k,i}) &= f(z^{k,i}) + \langle C(z^{k,i}), \lambda^{k,i} \rangle = f(z^{k,i}) + \langle C(z^{k,i}), \lambda^k \rangle + \langle C(z^{k,i}), \lambda^{k,i} - \lambda^k \rangle \\ &= \ell(z^{k,i}, \lambda^k) + \langle C(z^{k,i}), \lambda^{k,i} - \lambda^k \rangle. \end{aligned}$$

Since $z^{k,i} - y^{k,i}$ belongs to the null space of $A(y^{k,i})$, we have that $C(z^{k,i}) \approx C(y^{k,i})$, so

$$\ell(z^{k,i}, \lambda^{k,i}) \approx \ell(z^{k,i}, \lambda^k) + \langle C(y^{k,i}), \lambda^{k,i} - \lambda^k \rangle. \quad (3)$$

By the same reason, we have that

$$\varphi(z^{k,i}) \approx \varphi(y^{k,i}). \quad (4)$$

Combining (3) and (4), we obtain that, independently of the value of θ ,

$$\psi(z^{k,i}, \lambda^{k,i}) \approx \theta[\ell(z^{k,i}, \lambda^k) + \langle C(y^{k,i}), \lambda^{k,i} - \lambda^k \rangle] + (1 - \theta)\varphi(y^{k,i}). \quad (5)$$

Now, since $\varphi(y^{k,i}) \leq \varphi(x^k)$ (and the equality holds only when $y^{k,i} = x^k$), we can choose (Step 7 of Algorithm 3.1) θ sufficiently small so that

$$\theta[\ell(z^{k,i}, \lambda^k) + \langle C(y^{k,i}), \lambda^{k,i} - \lambda^k \rangle] + (1 - \theta)\varphi(y^{k,i}) \ll \psi(x^k, \lambda^k). \quad (6)$$

So, by (5) and (6), we have that, for small enough $\delta_{k,i}$,

$$\psi(z^{k,i}, \lambda^{k,i}) \ll \psi(x^k, \lambda^k). \quad (7)$$

The test (16) at Step 8 of Algorithm 3.1 states (7) in a rigorous way. The penalty parameters that are used at different iterations are, in general, different. So, (7) does not represent monotone decrease of a single merit function. In order to get convergence results based on the decrease of ψ , the definition of θ at each iteration will be made in order to satisfy a stability condition.

3 Description of the Algorithm

In this section, we define the model algorithm, the informal description of which was given in Section 2. Penalty parameters are computed using a strategy that ensures their convergence. However, the penalty parameter used at iteration $k + 1$ is not necessarily smaller than the one used at iteration k . This is an interesting feature, since excessively small penalty parameters can give premature emphasis to feasibility instead of optimality. Phase 1 of the algorithm goes from Step 2 to Step 3, while Phase 2 goes from Step 4 to Step 6. At Step 7 of the algorithm, the penalty parameter is computed in such a way that a sufficient decrease of the merit function is possible, if the trust region radius is small enough. We use the classical notation **Pred**, meaning Predicted Reduction of the merit function, and **Ared**, meaning Actual Reduction. The informal arguments used in Section 2 state that **Ared** \approx **Pred** if the trust region radius is small. At Step 8 we test if a sufficient decrease of the merit function took place from x^k to $z^{k,i}$. In the positive case, $z^{k,i}$ is accepted as new iterate. Otherwise, the trust region radius is reduced. Observe that the choice of the Lagrange multiplier estimates is completely arbitrary in this model algorithm, subject only to a boundedness condition. In specific implementations, the Lagrange multiplier estimators must be chosen in order to improve convergence properties and efficiency of the method.

Assume that $\gamma, \eta > 0$ are given scaling parameters, $\delta_{min} > 0$ is a lower bound of the trust region radius used at the beginning of each iteration and $M_1 > 0$ is a given upper bound of the vector of Lagrange multiplier estimates. Recall that $\|\cdot\|$ denotes an arbitrary norm on

\mathbb{R}^n and its associated matricial norm.

Algorithm 3.1

Let $x^0 \in \Omega$ an arbitrary initial point, $\lambda^0 \in \mathbb{R}^m$, $\|\lambda^0\| \leq M_1$, an initial approximation of the Lagrange multipliers. Let $\{\omega_k\}$ be a sequence of nonnegative numbers such that $\sum_{k=0}^{\infty} \omega_k < \infty$ and $\theta_{-1} \in (0, 1]$. Assume that $x^k \in \Omega$ and $\lambda^k \in \mathbb{R}^m$ have been computed for some $k \in \{0, 1, 2, \dots\}$. Define

$$\theta_k^{min} = \min \{1, \theta_0, \dots, \theta_{k-1}\} \quad (8)$$

and

$$\theta_k^{large} = \min \{1, \theta_k^{min} + \omega_k\}. \quad (9)$$

The steps for computing x^{k+1} or for deciding to terminate the execution of the algorithm are described below.

Step 1. *Set the initial trust region radius*

Set

$$i \leftarrow 0, \quad \delta_{k,0} \geq \delta_{min}, \quad \theta_{k,-1} = \theta_k^{large}.$$

Step 2. *Compute the Cauchy direction for feasibility*

Compute

$$d_n^k = P(x^k - \gamma \nabla \varphi(x^k)) - x^k, \quad (10)$$

where, for all $z \in \mathbb{R}^n$, $P(z)$ is the orthogonal projection of z on Ω .

Step 3. *Compute a more feasible point*

If $d_n^k \neq 0$, compute, using *Algorithm 3.2*, $y^{k,i} \in \Omega$ such that

$$\|y^{k,i} - x^k\| \leq 0.8\delta_{k,i}, \quad \varphi(y^{k,i}) < \varphi(x^k).$$

If $d_n^k = 0$, define $y^{k,i} = x^k$.

Step 4. *Compute the Cauchy direction for optimality*

Compute

$$d_t^{k,i} = P_{k,i}[y^{k,i} - \eta \nabla f(y^{k,i})] - y^{k,i}, \quad (11)$$

where, for all $z \in \mathbb{R}^n$, $P_{k,i}(z)$ is the orthogonal projection of z on $\pi_{k,i}$. (If $d_n^k = 0$ we have that $y^{k,i} = x^k$ and, so, $d_t^{k,i}$ is independent of i . Therefore, in this case we can write $d_t^k = d_t^{k,i}$.)

Step 5. *Compute a more optimal point*

If $d_n^k = 0$ and $d_t^{k,i} = 0$, the algorithm terminates.

If $d_t^{k,i} \neq 0$, compute, using *Algorithm 3.3*, $z^{k,i} \in \pi_{k,i}$ such that

$$\|z^{k,i} - x^k\| \leq \delta_{k,i}, \quad \ell(z^{k,i}, \lambda^k) < \ell(y^{k,i}, \lambda^k).$$

If $d_t^{k,i} = 0$, define $z^{k,i} = y^{k,i}$.

Step 6. *Compute new estimates of the Lagrange multipliers*

Compute $\lambda^{k,i} \in \mathbb{R}^m$ such that $\|\lambda^{k,i}\| \leq M_1$. If $d_n^k = 0$ and $C(x^k) \neq 0$ choose $\lambda^{k,i} = \lambda^k$.

Step 7. *Compute the penalty parameter and the predicted reduction*

Define, for all $\theta \in [0, 1]$,

$$Pred_{k,i}(\theta) = \theta\{\ell(x^k, \lambda^k) - [\ell(z^{k,i}, \lambda^k) + \langle \lambda^{k,i} - \lambda^k, C(y^{k,i}) \rangle]\} + (1 - \theta)[\varphi(x^k) - \varphi(y^{k,i})]. \quad (12)$$

Define $\theta_{k,i}$ the supremum of the set of penalty parameters $\theta \in [0, \theta_{k,i-1}]$ that verify

$$Pred_{k,i}(\theta) \geq \frac{1}{2}[\varphi(x^k) - \varphi(y^{k,i})]. \quad (13)$$

Define

$$\mathbf{Pred}_{k,i} = Pred_{k,i}(\theta_{k,i}). \quad (14)$$

Step 8. *Test sufficient decrease of the merit function*

Compute

$$\mathbf{Ared}_{k,i} = \theta_{k,i}[\ell(x^k, \lambda^k) - \ell(z^{k,i}, \lambda^{k,i})] + (1 - \theta_{k,i})[\varphi(x^k) - \varphi(z^{k,i})]. \quad (15)$$

If

$$\mathbf{Ared}_{k,i} \geq 0.1\mathbf{Pred}_{k,i}, \quad (16)$$

define

$$x^{k+1} = z^{k,i}, \quad \lambda^{k+1} = \lambda^{k,i}, \quad \theta_k = \theta_{k,i}, \quad \delta_k = \delta_{k,i}, \quad j(k) = i,$$

$$\mathbf{Ared}_k = \mathbf{Ared}_{k,i}, \quad \mathbf{Pred}_k = \mathbf{Pred}_{k,i}$$

and finish iteration k .

If (16) does not hold, choose

$$\delta_{k,i+1} \in [0.1\delta_{k,i}, 0.9\delta_{k,i}], \quad (17)$$

set $i \leftarrow i + 1$ and go to Step 3.

Observe that, from the definition of $\theta_{k,i}$ at Step 7, it follows that $\theta_{k,i} > 0$ for all k, i . A simple proof of this fact can be made by induction.

Algorithm 3.2, which is described below, is called from Step 3 of Algorithm 3.1. Given the Cauchy-like direction d_n^k , a point along this direction where φ sufficiently decreases will be computed. Since d_n^k is a descent direction, a standard backtracking procedure can be used. Finally, at Step 3 of Algorithm 3.2, we find a new point $y^{k,i}$ which is as good as the one obtained by the backtracking procedure. We are free to use any methodology for finding this final point. Undoubtedly, much of the efficiency of the algorithm depends on that choice.

Algorithm 3.2

Define

$$\bar{t}(k, i, 1) = \min \left\{ 1, \frac{0.8\delta_{k,i}}{\|d_n^k\|} \right\}.$$

(Observe that $\bar{t}(k, i, 1)$ is the maximum value of $t \in [0, 1]$ such that $\|td_n^k\| \leq 0.8\delta_{k,i}$.) We compute $\underline{t}(k, i, 1)$ following the steps below.

Step 1. Set $t \leftarrow \bar{t}(k, i, 1)$.

Step 2. If

$$\varphi(x^k + td_n^k) \leq \varphi(x^k) + 0.1t \langle d_n^k, \nabla \varphi(x^k) \rangle, \quad (18)$$

set $\underline{t}(k, i, 1) \leftarrow t$ and go to Step 3.

If (18) does not hold, choose $t' \in [0.1t, 0.9t]$, replace $t \leftarrow t'$ and repeat Step 2. In the implementation of Section 8 we set $t' = t/2$.

Step 3. Choose $y^{k,i} \in \Omega$ such that $\|y^{k,i} - x^k\| \leq 0.8\delta_{k,i}$ and

$$\varphi(y^{k,i}) \leq \varphi(x^k + \underline{t}(k, i, 1)d_n^k).$$

Algorithm 3.3 is called from Step 5 of Algorithm 3.1. Since $d_t^{k,i}$ is a descent direction for $\ell(x, \lambda^k)$, a point that satisfies a sufficient descent condition along this direction is computed using backtracking. Finally, a new point $z^{k,i}$ is computed (being at least as good as the previously computed one) using an arbitrary algorithm for minimizing function with linear constraints, including the constraints $x \in \Omega$ and $\|x - y^{k,i}\| \leq \delta_{k,i}$. If Ω is a polyhedron and $\|\cdot\| = \|\cdot\|_\infty$ we can use any method (convergent or not) for linearly constrained optimization.

Algorithm 3.3

Define

$$\bar{t}(k, i, 2) = \max \{ \hat{t} \in [0, 1] \mid \|y^{k,i} + td_t^{k,i}\| \leq \delta_{k,i} \text{ for all } t \in [0, \hat{t}] \}.$$

Step 1. Set $t \leftarrow \bar{t}(k, i, 2)$.

Step 2. If

$$\ell(y^{k,i} + td_t^{k,i}, \lambda^k) \leq \ell(y^{k,i}, \lambda^k) + 0.1t \langle d_t^{k,i}, \nabla f(y^{k,i}) \rangle, \quad (19)$$

set $\underline{t}(k, i, 2) \leftarrow t$ and go to Step 3.

If (19) does not hold, choose $t' \in [0.1t, 0.9t]$, replace $t \leftarrow t'$ and repeat Step 2. As in the case of Algorithm 3.2, we set $t' = t/2$ in the implementation.

Step 3. Choose $z^{k,i} \in \pi_{k,i}$ such that $\|z^{k,i} - x^k\| \leq \delta_{k,i}$ and

$$\ell(z^{k,i}, \lambda^k) \leq \ell(y^{k,i} + \underline{t}(k, i, 2)d_t^{k,i}, \lambda^k).$$

Lemma 3.1 guarantees that, ultimately, we are using essentially the same merit function for measuring the progress of the algorithm. This property will be important in the convergence analysis.

Lemma 3.1. *If $\{x^k\}$ is an infinite sequence generated by Algorithm 3.1, then the sequence θ_k is convergent.*

Proof. By (8), (9) and the choice of $\theta_{k,i}$ at Algorithm 3.1, we have that, for all $k \in \{0, 1, 2, \dots\}$,

$$\theta_{k+1}^{min} \leq \theta_k \leq \theta_k^{min} + \omega_k.$$

Now, the sequence $\{\theta_k^{min}\}$ is nonincreasing and bounded below, therefore convergent. Thus, $\{\theta_k\}$ is enclosed between two convergent sequences and, so, $\{\theta_k\}$ is convergent. \square

4 The Algorithm is Well Defined

The objective of this section is to prove that, given an arbitrary approximation x^k , where Algorithm 3.1 does not terminate, we are able to find, in finite time, the new approximation x^{k+1} . The only assumption that is necessary for proving this property, stated below, is pretty minimal. We will maintain this Assumption, without restating it, all along the paper.

Assumption 4.1. *There exists an open set Ω' , $\Omega \subset \Omega' \subset \mathbb{R}^n$, such that $f(x)$, $\nabla f(x)$, $C(x)$ and $A(x)$ exist and are continuous for all $x \in \Omega'$.*

First, we prove a property that was announced in previous sections: d_n^k is a descent direction for φ .

Lemma 4.1. *If $d_n^k \neq 0$, then*

$$\langle d_n^k, \nabla \varphi(x^k) \rangle < 0.$$

Proof. The projection $P(x^k - \gamma \nabla \varphi(x^k))$ is the unique solution of the problem

$$\text{Minimize } \|x^k - \gamma \nabla \varphi(x^k) - z\|_2 \quad \text{subject to } z \in \Omega.$$

If $d_n^k \neq 0$, then $P(x^k - \gamma \nabla \varphi(x^k)) \neq x^k$, so

$$\|P(x^k - \gamma \nabla \varphi(x^k)) - (x^k - \gamma \nabla \varphi(x^k))\|_2^2 < \|x^k - (x^k - \gamma \nabla \varphi(x^k))\|_2^2.$$

Therefore,

$$\|P(x^k - \gamma \nabla \varphi(x^k)) - x^k\|_2^2 + 2\gamma \langle P(x^k - \gamma \nabla \varphi(x^k)) - x^k, \nabla \varphi(x^k) \rangle + \|\gamma \nabla \varphi(x^k)\|_2^2 < \|\gamma \nabla \varphi(x^k)\|_2^2.$$

So,

$$\|P(x^k - \gamma \nabla \varphi(x^k)) - x^k\|_2^2 + 2\gamma \langle d_n^k, \nabla \varphi(x^k) \rangle < 0. \quad (20)$$

This implies that $\langle d_n^k, \nabla \varphi(x^k) \rangle < 0$, as we wanted to prove. \square

An easy consequence of Lemma 4.1 is that Algorithm 3.2 does not cycle indefinitely. This is proved in Lemma 4.2.

Lemma 4.2. *Algorithm 3.2 is well defined.*

Proof. By Lemma 4.1, we have that $\langle d_n^k, \nabla \varphi(x^k) \rangle < 0$. Therefore, by the definition of directional derivative, (18) is obtained after a finite number of reductions of t . Since $x^k + \underline{t}(k, i, 1)d_n^k$ is an admissible choice for $y^{k,i}$ the proof is complete. \square

In the following lemma, the decrease of φ from x^k to $y^{k,i}$ is quantified.

Lemma 4.3. *Assume that $d_n^k \neq 0$. Then, there exists $\bar{t}_k > 0$ such that*

$$\varphi(y^{k,i}) \leq \varphi(x^k) + \min \{0.1, 0.08 \frac{\delta_{k,i}}{\|d_n^k\|}, 0.01\bar{t}_k\} \langle d_n^k, \nabla \varphi(x^k) \rangle.$$

Proof. By the definition of directional derivative, there exists $\bar{t}_k > 0$ such that (18) holds for all $t \in [0, \bar{t}_k]$. Let us define $\bar{t}(k, i, 1)$ as in Algorithm 3.2. We consider two possibilities: $\bar{t}(k, i, 1) \leq \bar{t}_k$ and $\bar{t}(k, i, 1) > \bar{t}_k$.

If $\bar{t}(k, i, 1) \leq \bar{t}_k$, we will have $\underline{t}(k, i, 1) = \bar{t}(k, i, 1)$ and

$$\begin{aligned} \varphi(x^k + \bar{t}(k, i, 1)d_n^k) &\leq \varphi(x^k) + 0.1\bar{t}(k, i, 1) \langle d_n^k, \nabla \varphi(x^k) \rangle \\ &\leq \varphi(x^k) + 0.1 \min \left\{ 1, \frac{0.8\delta_{k,i}}{\|d_n^k\|} \right\} \langle d_n^k, \nabla \varphi(x^k) \rangle. \end{aligned} \quad (21)$$

Assume now that $\bar{t}(k, i, 1) > \bar{t}_k$. In this case, it follows from the definition of \bar{t}_k and the backtracking scheme of Algorithm 3.2, that $\underline{t}(k, i, 1) \geq \bar{t}_k/10$. Therefore,

$$\begin{aligned} \varphi(x^k + \underline{t}(k, i, 1)d_n^k) &\leq \varphi(x^k) + 0.1\underline{t}(k, i, 1) \langle d_n^k, \nabla \varphi(x^k) \rangle \\ &\leq \varphi(x^k) + \frac{0.1\bar{t}_k}{10} \langle d_n^k, \nabla \varphi(x^k) \rangle. \end{aligned} \quad (22)$$

By (21) and (22), we have that

$$\varphi(y^{k,i}) \leq \varphi(x^k + \underline{t}(k, i, 1)d_n^k) \leq \varphi(x^k) + \min \{0.1, 0.08 \frac{\delta_{k,i}}{\|d_n^k\|}, 0.01\bar{t}_k\} \langle d_n^k, \nabla \varphi(x^k) \rangle,$$

as we wanted to prove. \square

As so happened to be with d_n^k with respect to φ , we are able to prove that $d_t^{k,i}$ is a descent direction with respect to f .

Lemma 4.4. *If $d_t^{k,i} \neq 0$, then*

$$\langle d_t^{k,i}, \nabla f(y^{k,i}) \rangle < 0.$$

Proof. Since $d_t^{k,i} \neq 0$, the projection of $y^{k,i} - \eta \nabla f(y^{k,i})$ on $\pi_{k,i}$ is different from $y^{k,i}$. So, since $y^{k,i} \in \pi_{k,i}$,

$$\|P_{k,i}(y^{k,i} - \eta \nabla f(y^{k,i})) - (y^{k,i} - \eta \nabla f(y^{k,i}))\|_2^2 < \|y^{k,i} - (y^{k,i} - \eta \nabla f(y^{k,i}))\|_2^2.$$

So,

$$\|P_{k,i}(y^{k,i} - \eta \nabla f(y^{k,i})) - y^{k,i}\|_2^2 + 2\eta \langle P_{k,i}(y^{k,i} - \eta \nabla f(y^{k,i})) - y^{k,i}, \nabla f(y^{k,i}) \rangle + \|\eta \nabla f(y^{k,i})\|_2^2 < \|\eta \nabla f(y^{k,i})\|_2^2.$$

Therefore,

$$\|P_{k,i}(y^{k,i} - \eta \nabla f(y^{k,i})) - y^{k,i}\|_2^2 + 2\eta \langle P_{k,i}(y^{k,i} - \eta \nabla f(y^{k,i})) - y^{k,i}, \nabla f(y^{k,i}) \rangle < 0, \quad (23)$$

and the thesis follows trivially from this inequality. \square

Since $d_t^{k,i}$ belongs to the null-space of $A(y^{k,i})$, descent directions of f coincide with descent directions of the Lagrangian. This is the argument used in the proof of the following lemma.

Lemma 4.5. *Algorithm 3.3 is well defined.*

Proof. Since $A(y^{k,i})d_t^{k,i} = 0$ we have that $\langle d_t^{k,i}, \nabla f(y^{k,i}) \rangle = \langle d_t^{k,i}, \nabla \ell(y^{k,i}, \lambda^k) \rangle$. So, the fact that (19) holds for small enough t follows from the definition of directional derivative. \square

In Lemma 4.6, we quantify the amount of decrease of the Lagrangian, from x^k to $z^{k,i}$, when $d_n^k = 0$.

Lemma 4.6. *Assume that $d_n^k = 0$ and $d_t^k \neq 0$. Then, there exists $\tilde{t}_k > 0$ such that*

$$\ell(z^{k,i}, \lambda^k) \leq \ell(x^k, \lambda^k) + \min \{0.1, 0.1 \frac{\delta_{k,i}}{\|d_t^k\|}, 0.01\tilde{t}_k\} \langle d_t^k, \nabla f(x^k) \rangle$$

Proof. By the definition of directional derivative, there exists $\tilde{t}_k > 0$ such that (19) holds for all $t \in [0, \tilde{t}_k]$. Let us define $\bar{t}(k, i, 2)$ as in Algorithm 3.2. We consider two possibilities: $\bar{t}(k, i, 2) \leq \tilde{t}_k$ and $\bar{t}(k, i, 2) > \tilde{t}_k$.

If $\bar{t}(k, i, 2) \leq \tilde{t}_k$, we have that $\underline{t}(k, i, 2) = \bar{t}(k, i, 2)$. Moreover, since $y^{k,i} = x^k$ we have that $\bar{t}(k, i, 2) = \min \{1, \delta_{k,i} / \|d_t^k\|\}$ and

$$\begin{aligned} \ell(x^k + \bar{t}(k, i, 2)d_t^k, \lambda^k) &\leq \ell(x^k, \lambda^k) + 0.1\bar{t}(k, i, 2) \langle d_t^k, \nabla \ell(x^k, \lambda^k) \rangle \\ &\leq \ell(x^k, \lambda^k) + 0.1 \min \{1, \frac{\delta_{k,i}}{\|d_t^k\|}\} \langle d_t^k, \nabla \ell(x^k, \lambda^k) \rangle. \end{aligned} \quad (24)$$

Assume now that $\bar{t}(k, i, 2) > \tilde{t}_k$. In this case, it follows from the definition of \tilde{t}_k and the backtracking scheme of Algorithm 3.3, that $\underline{t}(k, i, 2) \geq \tilde{t}_k/10$. Therefore,

$$\begin{aligned} \ell(x^k + \underline{t}(k, i, 2)d_t^k, \lambda^k) &\leq \ell(x^k, \lambda^k) + 0.1\underline{t}(k, i, 2)\langle d_t^k, \nabla\ell(x^k, \lambda^k) \rangle \\ &\leq \ell(x^k, \lambda^k) + \frac{0.1\tilde{t}_k}{10}\langle d_t^k, \nabla\ell(x^k, \lambda^k) \rangle. \end{aligned} \quad (25)$$

By (24) and (25), we have that

$$\ell(z^{k,i}, \lambda^k) \leq \ell(x^k + \underline{t}(k, i, 2)d_t^k, \lambda^k) \leq \ell(x^k, \lambda^k) + \min\{0.1, 0.1\frac{\delta_{k,i}}{\|d_t^k\|}, 0.01\tilde{t}_k\}\langle d_t^k, \nabla\ell(x^k, \lambda^k) \rangle.$$

Since $\langle d_t^k, \nabla\ell(x^k, \lambda^k) \rangle = \langle d_t^k, \nabla f(x^k) \rangle$ the last inequality implies the desired result. \square

The following lemma guarantees that Algorithm 3.1 is well defined when $d_n^k \neq 0$. That is, if the feasibility condition does not hold, we can find, after a finite number of reductions of $\delta_{k,i}$, a new approximation x^{k+1} .

Lemma 4.7. *Assume that $d_n^k \neq 0$. Then, after a finite number of reductions of the trust region radius at (17), we obtain $i \geq 0$ such that (16) holds.*

Proof. By (12), (15) and the Mean Value Theorem, there exists $\xi_{k,i} \in [0, 1]$ such that

$$\begin{aligned} &\mathbf{Ared}_{k,i} - \mathbf{Pred}_{k,i} \\ &= \theta_{k,i}\langle C(y^{k,i}) - C(z^{k,i}), \lambda^{k,i} - \lambda^k \rangle + (1 - \theta_{k,i})[\varphi(y^{k,i}) - \varphi(z^{k,i})] \\ &= \theta_{k,i}\langle A(z^{k,i} + \xi_{k,i}(y^{k,i} - z^{k,i}))(y^{k,i} - z^{k,i}), \lambda^{k,i} - \lambda^k \rangle \\ &\quad + (1 - \theta_{k,i})\langle \nabla\varphi(z^{k,i} + \xi_{k,i}(z^{k,i} - y^{k,i})), y^{k,i} - z^{k,i} \rangle. \end{aligned}$$

But $A(y^{k,i})[y^{k,i} - z^{k,i}] = 0$, so $\langle \nabla\varphi(y^{k,i}), y^{k,i} - z^{k,i} \rangle = 0$. Therefore,

$$\begin{aligned} &\mathbf{Ared}_{k,i} - \mathbf{Pred}_{k,i} \\ &= \theta_{k,i}\langle [A(z^{k,i} + \xi_{k,i}(y^{k,i} - z^{k,i})) - A(y^{k,i})](y^{k,i} - z^{k,i}), \lambda^{k,i} - \lambda^k \rangle \\ &\quad + (1 - \theta_{k,i})\langle \nabla\varphi(z^{k,i} + \xi_{k,i}(z^{k,i} - y^{k,i})) - \nabla\varphi(y^{k,i}), y^{k,i} - z^{k,i} \rangle. \end{aligned}$$

This implies that

$$\begin{aligned} |\mathbf{Ared}_{k,i} - \mathbf{Pred}_{k,i}| &\leq 2M_1\|A(z^{k,i} + \xi_{k,i}(y^{k,i} - z^{k,i})) - A(y^{k,i})\|_2\|y^{k,i} - z^{k,i}\|_2 \\ &\quad + \|\nabla\varphi(z^{k,i} + \xi_{k,i}(z^{k,i} - y^{k,i})) - \nabla\varphi(y^{k,i})\|_2\|y^{k,i} - z^{k,i}\|_2 \\ &\leq [2cM_1\|A(z^{k,i} + \xi_{k,i}(y^{k,i} - z^{k,i})) - A(y^{k,i})\| + c\|\nabla\varphi(y^{k,i} + \xi_{k,i}(y^{k,i} - z^{k,i})) - \nabla\varphi(y^{k,i})\|]1.8\delta_{k,i}, \end{aligned} \quad (26)$$

where c is a constant that only depends on $\|\cdot\|$.

Moreover, by Lemma 4.3 and (13),

$$\mathbf{Pred}_{k,i} \geq \frac{\varphi(x^k) - \varphi(y^{k,i})}{2} \geq -\frac{1}{2} \min\{0.1, 0.08\frac{\delta_{k,i}}{\|d_n^k\|}, 0.01\bar{t}_k\}\langle d_n^k, \nabla\varphi(x^k) \rangle > 0.$$

So,

$$\begin{aligned} & \frac{|\mathbf{Ared}_{k,i} - \mathbf{Pred}_{k,i}|}{\mathbf{Pred}_{k,i}} \\ & \leq \frac{[2M_1 \|A(z^{k,i} + \xi_{k,i}(y^{k,i} - z^{k,i})) - A(y^{k,i})\| + \|\nabla\varphi(y^{k,i} + \xi_{k,i}(y^{k,i} - z^{k,i})) - \nabla\varphi(y^{k,i})\|] 1.8c\delta_{k,i}}{-\frac{1}{2} \min \{0.1, 0.08 \frac{\delta_{k,i}}{\|d_n^k\|}, 0.01\bar{t}_k\} \langle d_n^k, \nabla\varphi(x^k) \rangle} \end{aligned}$$

Therefore, if $\frac{0.08\delta_{k,i}}{\|d_n^k\|} \leq \min \{0.1, 0.01\bar{t}_k\}$ we have that

$$\begin{aligned} & \frac{|\mathbf{Ared}_{k,i} - \mathbf{Pred}_{k,i}|}{\mathbf{Pred}_{k,i}} \\ & \leq \frac{[2M_1 \|A(z^{k,i} + \xi_{k,i}(y^{k,i} - z^{k,i})) - A(y^{k,i})\| + \|\nabla\varphi(y^{k,i} + \xi_{k,i}(y^{k,i} - z^{k,i})) - \nabla\varphi(y^{k,i})\|] 1.8c}{-\frac{1}{2} 0.08 \frac{1}{\|d_n^k\|} \langle d_n^k, \nabla\varphi(x^k) \rangle} \end{aligned} \quad (27)$$

So, by the continuity of $A(x)$ and $\nabla\varphi(x)$, it follows that for small enough $\delta_{k,i}$, (16) takes place. \square

Finally, we prove that Algorithm 3.1 is well defined when $d_n^k = 0$ but $d_t^k \neq 0$. (In this case $x^k = y^{k,i}$ and $d_t^{k,i} = d_t^k$ for all i .) By Lemmas 4.7 and 4.8 it follows that Algorithm 3.1 is well defined.

Lemma 4.8. *Assume that $d_n^k = 0$ and $d_t^k \neq 0$. Then, after a finite number of reductions of the trust region radius at (17), we obtain $i \geq 0$ such that (16) holds.*

Proof. Since $d_n^k = 0$, we have, by Step 3 of Algorithm 3.1, that $x^k = y^{k,i}$ for all $i \in \{0, 1, 2, \dots\}$. So, by (12), we have that

$$\mathbf{Pred}_{k,i}(\theta) = \theta[\ell(x^k, \lambda^k) - \ell(z^{k,i}, \lambda^k) + \langle C(x^k), \lambda^k - \lambda^{k,i} \rangle]$$

By the choice of $\lambda^{k,i}$ at Step 6, we have that $\|C(x^k)\|_2 \|\lambda^k - \lambda^{k,i}\|_2 = 0$. Therefore,

$$\mathbf{Pred}_{k,i}(\theta) = \theta[\ell(x^k, \lambda^k) - \ell(z^{k,i}, \lambda^k)]$$

for all $i \in \{0, 1, 2, \dots\}$.

Now, by Lemma 4.6, we obtain

$$\ell(x^k, \lambda^k) - \ell(z^{k,i}, \lambda^k) \geq - \min \{0.1, 0.1 \frac{\delta_{k,i}}{\|d_t^k\|}, 0.01\tilde{t}_k\} \langle d_t^k, \nabla f(x^k) \rangle \quad (28)$$

for all $i \in \{0, 1, 2, \dots\}$. So,

$$\ell(x^k, \lambda^k) - \ell(z^{k,i}, \lambda^{k,i}) \geq 0 \quad (29)$$

for all $i \in \{0, 1, 2, \dots\}$. Since $d_n^k = 0$ implies that $y^{k,i} = x^k$ for all i , we have that $\theta_{k,i} = \theta_{k,-1}$ for all $i \geq 0$. Therefore, for all $i \geq 0$, we have, by Lemma 4.6, that

$$\mathbf{Pred}_{k,i} \geq \theta_{k,-1} [- \min \{0.1, 0.1 \frac{\delta_{k,i}}{\|d_t^k\|}, 0.01\tilde{t}_k\} \langle d_t^k, \nabla f(x^k) \rangle] \quad (30)$$

But, as in Lemma 4.7, we can deduce that

$$\begin{aligned} & |\mathbf{Ared}_{k,i} - \mathbf{Pred}_{k,i}| \\ & \leq [2M_1 \|A(z^{k,i} + \xi_{k,i}(y^{k,i} - z^{k,i})) - A(y^{k,i})\| + \|\nabla\varphi(y^{k,i} + \xi_{k,i}(y^{k,i} - z^{k,i})) - \nabla\varphi(y^{k,i})\|] 1.8c\delta_{k,i}. \end{aligned} \quad (31)$$

By (30), (31) and the continuity of $A(x)$ and $\nabla\varphi(x)$, the inequality

$$\left| \frac{\mathbf{Ared}_{k,i}}{\mathbf{Pred}_{k,i}} - 1 \right| \leq 0.9$$

holds if $\delta_{k,i}$ is small enough. This implies that, for some $i \in \{0, 1, 2, \dots\}$, the inequality (16) is satisfied. \square

5 The Feasibility Cauchy Direction Tends to Zero

In this section we prove that, if Algorithm 3.1 does not terminate in a finite number of iterations, then the Cauchy-like direction d_n^k tends to zero. Roughly speaking, this means that arbitrary close approximations to stationary points related to the minimization of $\varphi(x)$ subject to $x \in \Omega$ can be encountered. If stationary points coincide with feasible points (which is a possible characteristic of the problem), we can find points that are arbitrary close to feasibility.

We need two additional general assumptions, 5.1 and 5.2, for proving $d_n^k \rightarrow 0$. Assumption 5.1 is a set of Lipschitz conditions on $A(x)$, $\nabla\varphi(x)$ and $\nabla f(x)$. Our convergence theory does not assume the existence of second derivatives.

Assumption 5.1. *There exist $L_1, L_2, L_3 > 0$ such that*

$$\|A(y) - A(x)\| \leq L_1 \|y - x\|, \quad (32)$$

$$\|\nabla\varphi(y) - \nabla\varphi(x)\| \leq L_2 \|y - x\| \quad (33)$$

and

$$\|\nabla f(y) - \nabla f(x)\| \leq L_3 \|y - x\|, \quad (34)$$

for all $x, y \in \Omega$.

Assumption 5.1 implies that

$$\|C(y) - C(x) - A(x)(y - x)\| \leq \frac{L_1}{2} \|y - x\|^2, \quad (35)$$

and

$$-\frac{L_2}{2} \|y - x\|^2 \leq \varphi(y) - \varphi(x) - \langle \nabla\varphi(x), y - x \rangle \leq \frac{L_2}{2} \|y - x\|^2 \quad (36)$$

and for all $x, y \in \Omega$.

Moreover, from Assumption 5.1 we also deduce that there exists $L_4 \geq 0$ such that for all $x, y \in \Omega$, $\|\lambda\| \leq M_1$,

$$\|\nabla \ell(y, \lambda) - \nabla \ell(x, \lambda)\| \leq L_4 \|y - x\| \quad (37)$$

This implies that for all $x, y \in \Omega$, $\|\lambda\| \leq M_1$,

$$-\frac{L_4}{2} \|y - x\|^2 \leq \ell(y, \lambda) - \ell(x, \lambda) - \langle \nabla \ell(x, \lambda), y - x \rangle \leq \frac{L_4}{2} \|y - x\|^2. \quad (38)$$

The second general Assumption 5.2 states that the quantities $\|C(x^k)\|$, $\|A(x^k)\|$, $|f(x^k)|$, $\|\nabla f(x)\|$ and $\|\nabla f(y^{k,i})\|$ are bounded. Clearly, this assumption holds, for example, if Ω is bounded.

Assumption 5.2. *There exist $M_2, M_3, M_4, M_5, M_6 > 0$ such that, for all $k \in \{0, 1, 2, \dots\}$, $i = 0, 1, \dots, j(k)$,*

$$\|C(x^k)\| \leq M_2, \|A(x^k)\| \leq M_3, |f(x^k)| \leq M_4, \|\nabla f(x^k)\| \leq M_5, \|\nabla f(y^{k,i})\| \leq M_6. \quad (39)$$

This implies that there exists $M_7 > 0$ such that

$$\|\nabla \varphi(x^k)\| \leq M_7 \quad (40)$$

for all $k \in \{0, 1, 2, \dots\}$.

In Section 2 we mentioned that, since $A(y^{k,i})(z^{k,i} - y^{k,i}) = 0$, we have that $C(z^{k,i}) \approx C(y^{k,i})$ and $\varphi(z^{k,i}) \approx \varphi(y^{k,i})$. From these approximations we deduced, informally, that **Ared** \approx **Pred**, which justifies that, for small $\delta_{k,i}$, sufficient decrease of the merit function can be obtained. In the following lemma, we state with precision in which sense the approximations take place.

Lemma 5.1. *Suppose that Assumption 5.1 holds. Then, there exist $c_1, c_2, c_3 > 0$ such that*

$$|\theta_{k,i} \langle \lambda^k - \lambda^{k,i}, C(z^{k,i}) - C(y^{k,i}) \rangle| \leq c_1 \delta_{k,i}^2, \quad (41)$$

$$|\varphi(z^{k,i}) - \varphi(y^{k,i})| \leq c_2 (\delta_{k,i}^4 + \|C(x_k)\| \delta_{k,i}^2) \quad (42)$$

and

$$|\mathbf{Ared}_{k,i} - \mathbf{Pred}_{k,i}| \leq c_3 (\theta_{k,i} \delta_{k,i}^2 + \delta_{k,i}^4 + \|C(x^k)\| \delta_{k,i}^2) \quad (43)$$

for all $k \in \{0, 1, 2, \dots\}$, $i = 0, 1, \dots, j(k)$.

Proof. By (12) and (15) we have that

$$\mathbf{Ared}_{k,i} - \mathbf{Pred}_{k,i} = \theta_{k,i} \langle \lambda^k - \lambda^{k,i}, C(z^{k,i}) - C(y^{k,i}) \rangle + (1 - \theta_{k,i}) [\varphi(y^{k,i}) - \varphi(z^{k,i})]. \quad (44)$$

Now, by Assumption 5.1, since $z^{k,i} \in \pi_{k,i}$, we have that

$$\|C(z^{k,i}) - C(y^{k,i})\| \leq \|A(y^{k,i})(z^{k,i} - y^{k,i})\| + \frac{L_1}{2} 1.8\delta_{k,i}^2 = 0.9L_1\delta_{k,i}^2. \quad (45)$$

Therefore,

$$|\theta_{k,i} \langle \lambda^k - \lambda^{k,i}, C(z^{k,i}) - C(y^{k,i}) \rangle| \leq 1.8cM_1L_1\theta_{k,i}\delta_{k,i}^2 = c_1\delta_{k,i}^2$$

for all $k \in \{0, 1, 2, \dots\}$, $i = 0, 1, \dots, j(k)$, where $c_1 = 1.8cM_1L_1$ and c is a constant that only depends on $\|\cdot\|$.

Now,

$$\begin{aligned} |\varphi(z^{k,i}) - \varphi(y^{k,i})| &= \left| \frac{1}{2} \|C(z^{k,i})\|_2^2 - \frac{1}{2} \|C(y^{k,i})\|_2^2 \right| \\ &= \left| \frac{1}{2} \|C(z^{k,i}) - C(y^{k,i}) - A(y^{k,i})(z^{k,i} - y^{k,i}) + C(y^{k,i})\|_2^2 - \frac{1}{2} \|C(y^{k,i})\|_2^2 \right| \\ &= \left| \frac{1}{2} [\|C(z^{k,i}) - C(y^{k,i}) - A(y^{k,i})(z^{k,i} - y^{k,i})\|_2^2 + \|C(y^{k,i})\|_2^2 \right. \\ &\quad \left. + 2\langle C(z^{k,i}) - C(y^{k,i}) - A(y^{k,i})(z^{k,i} - y^{k,i}), C(y^{k,i}) \rangle] - \frac{1}{2} \|C(y^{k,i})\|_2^2 \right| \\ &\leq \frac{1}{2} \|C(z^{k,i}) - C(y^{k,i}) - A(y^{k,i})(z^{k,i} - y^{k,i})\|_2^2 + \|C(z^{k,i}) - C(y^{k,i}) - A(y^{k,i})(z^{k,i} - y^{k,i})\|_2 \|C(y^{k,i})\|_2. \end{aligned}$$

So, by Assumption 5.1, since $\|C(y^{k,i})\|_2 \leq \|C(x^k)\|_2$,

$$|\varphi(z^{k,i}) - \varphi(y^{k,i})| \leq \frac{1}{2} \left(\frac{cL_1}{2} \|z^{k,i} - y^{k,i}\|^2 \right)^2 + \frac{cL_1}{2} \|z^{k,i} - y^{k,i}\|^2 \|C(x^k)\| \leq c_2(\delta_{k,i}^4 + \|C(x^k)\| \delta_{k,i}^2)$$

where $c_2 = 1.8 \max \{c^2L_1^2/8, cL_1/2\}$ and c is a norm-dependent constant.

By (41) and (42), defining $c_3 = \max \{c_1, c_2\}$, we obtain

$$|\mathbf{Ared}_{k,i} - \mathbf{Pred}_{k,i}| \leq c_3(\theta_{k,i}\delta_{k,i}^2 + \delta_{k,i}^4 + \|C(x^k)\| \delta_{k,i}^2)$$

for all $k \in \{0, 1, 2, \dots\}$, $i = 0, 1, \dots, j(k)$, as we wanted to prove. \square

In Section 4 we saw that Algorithm 3.1 terminates only when both the feasibility Cauchy-like direction d_n^k and the tangent Cauchy direction d_t^k are null. In this section and the following one we study the behavior of the algorithm when an infinite sequence is generated. As we mentioned before, our aim in this section is to prove that $d_n^k \rightarrow 0$. We will proceed showing that the negation of this property leads to a contradiction. Let us state that negation in the following hypothesis.

Hypothesis 5.1. *Let $\{x^k\}$ an infinite sequence generated by Algorithm 3.1. There exists K_1 , an infinite subset of $\{0, 1, 2, \dots\}$, and $\varepsilon > 0$ such that*

$$\|d_n^k\| \geq \varepsilon \quad \text{for all } k \in K_1.$$

Lemmas 5.2, 5.3 and 5.4 assume that Hypothesis 5.1 holds. In the first of these lemmas, we prove that the directional derivative of φ along d_n^k is bounded away from zero.

Lemma 5.2. *If Hypothesis 5.1 holds, there exists $c_4 > 0$ such that*

$$\langle d_n^k, \nabla\varphi(x^k) \rangle \leq -c_4. \quad (46)$$

for all $k \in K_1$.

Proof. By (20) we have that

$$\langle d_n^k, \nabla\varphi(x^k) \rangle \leq -\frac{\|d_n^k\|_2^2}{2\gamma} \leq -c\frac{\|d_n^k\|^2}{2\gamma} \leq -\frac{c\varepsilon^2}{2\gamma},$$

where $c > 0$ is a constant that depends only on $\|\cdot\|$. Therefore (46) holds with $c_4 = \frac{c\varepsilon^2}{2\gamma}$. \square

In Lemma 5.3 we prove that, under Hypothesis 5.1, the amount of decrease of φ , from x^k to $y^{k,i}$, is proportional to $\delta_{k,i}$.

Lemma 5.3. *Suppose that Hypothesis 5.1 and Assumptions 5.1 and 5.2 hold. Then there exists $\bar{t} > 0$ such that*

$$\varphi(y^{k,i}) \leq \varphi(x^k) - \min \{0.1, 0.08\frac{\delta_{k,i}}{\varepsilon}, 0.01\bar{t}\} c_4. \quad (47)$$

for all $k \in K_1$, $i = 0, 1, \dots, j(k)$, where $c_4 > 0$ is given in Lemma 5.2.

Proof. By (36) we have that

$$\varphi(x^k + td_n^k) \leq \varphi(x^k) + t\langle \nabla\varphi(x^k), d_n^k \rangle + \frac{t^2}{2}L_2\|d_{nor}^k\|^2$$

for all $t \in [0, 1]$. But, by Assumption 5.2,

$$\|d_n^k\| \leq M_7 \quad (48)$$

for all $k \in K_1$. Therefore,

$$\begin{aligned} \varphi(x^k + td_n^k) &\leq \varphi(x^k) + 0.1t\langle \nabla\varphi(x^k), d_n^k \rangle + 0.9t\langle \nabla\varphi(x^k), d_n^k \rangle + \frac{t^2}{2}L_2M_7^2 \\ &\leq \varphi(x^k) + 0.1t\langle \nabla\varphi(x^k), d_n^k \rangle - 0.9tc_4 + \frac{t^2}{2}L_2M_7^2 \\ &= \varphi(x^k) + 0.1t\langle \nabla\varphi(x^k), d_n^k \rangle - 0.9(tc_4 - \frac{t}{2}L_2M_7^2). \end{aligned}$$

Therefore, setting $\bar{t} = \frac{1.8c_4}{L_2M_7^2}$ we obtain that

$$\varphi(x^k + td_n^k) \leq \varphi(x^k) + 0.1t\langle \nabla\varphi(x^k), d_n^k \rangle$$

for all $t \in [0, \bar{t}]$. So, we can complete the proof as in Lemma 4.3 and (47) follows for all $k \in K_1$ using $\|d_n^k\| \geq \varepsilon$. \square

Lemma 5.4 is the last result in which we assume Hypothesis 5.1. In fact, by Lemma 5.3, the decrease of φ is large (proportional to $\delta_{k,i}$), by (13) the predicted reduction is of the same order as the decrease of φ and, by Lemma 5.1, the predicted reduction is a good approximation of the actual reduction, if $\delta_{k,i}$ is small. From these facts, we deduce, in Lemma 5.4, that it is not necessary to reduce excessively the trust region radius for obtaining sufficient decrease on the merit function.

Lemma 5.4. *Suppose that Hypothesis 5.1 and Assumptions 5.1 and 5.2 hold. Then, there exists $\bar{\delta} > 0$ such that $\delta_k \geq \bar{\delta}$ for all $k \in K_1$.*

Proof. By Lemma 5.3, (43), Hypothesis 5.1, Assumptions 5.1 and 5.2, we have that for all $k \in K_1$, if

$$\delta_{k,i} \leq \frac{\varepsilon}{0.08} \min \{0.1, 0.01\bar{t}\}, \quad (49)$$

the following inequality holds:

$$\frac{|\mathbf{Ared}_{k,i} - \mathbf{Pred}_{k,i}|}{\mathbf{Pred}_{k,i}} \leq \frac{12.5c_3\varepsilon(\delta_{k,i} + \delta_{k,i}^3 + M_2\delta_{k,i})}{c_4}$$

where $c_4 > 0$ is defined in Lemma 5.2.

Therefore, if $\frac{12.5c_3\varepsilon(\delta_{k,i} + \delta_{k,i}^3 + M_2\delta_{k,i})}{c_4} \leq 0.9$, the inequality (16) holds. The desired result follows from this fact. \square

Theorem 5.5 is the main result of this section. Using the previous results, which were obtained with Hypothesis 5.1, we will arrive to a contradiction. This implies that Hypothesis 5.1 cannot be true. The idea of the proof is that, since under Hypothesis 5.1 $\delta_{k,i}$ is bounded away from zero, the decrease of the merit function is bounded away from zero. Although the merit function changes from one iteration to another, the stability properties of the choice of θ and the boundedness assumption 5.2 lead us to a contradiction.

Theorem 5.1. *Suppose that Assumptions 5.1 and 5.2 hold. Then, if $\{x^k\}$ is an infinite sequence generated by Algorithm 3.1, we have that*

$$\lim_{k \rightarrow \infty} d_n^k = 0. \quad (50)$$

Proof. We proceed by contradiction. If (50) does not hold, it follows that Hypothesis 5.1 takes place. Therefore, by (13), Lemmas 5.2 and 5.4 we have that

$$\mathbf{Pred}_k \geq c'_1 \equiv \frac{1}{2} \min \left\{ 0.1, 0.08 \frac{\bar{\delta}}{\varepsilon}, 0.01\bar{t} \right\} > 0$$

for all $k \in K_1$. Therefore, $\mathbf{Ared}_k \geq 0.1c'_1 > 0$ for all $k \in K_1$. Let us write, for all $k \in \{0, 1, 2, \dots\}$,

$$\ell_k = \ell(x^k, \lambda^k), \quad \varphi_k = \varphi(x^k), \quad \psi_k = \theta_k \ell_k + (1 - \theta_k) \varphi_k.$$

Then, for all $k \in \{0, 1, 2, \dots\}$ we have that

$$\begin{aligned} \psi_{k+1} &= \theta_{k+1} \ell_{k+1} + (1 - \theta_{k+1}) \varphi_{k+1} \\ &= \theta_{k+1} \ell_{k+1} + (1 - \theta_{k+1}) \varphi_{k+1} - [\theta_k \ell_{k+1} + (1 - \theta_k) \varphi_{k+1}] + [\theta_k \ell_{k+1} + (1 - \theta_k) \varphi_{k+1}] \\ &= (\theta_{k+1} - \theta_k) \ell_{k+1} + (\theta_k - \theta_{k+1}) \varphi_{k+1} + [\theta_k \ell_{k+1} + (1 - \theta_k) \varphi_{k+1}] \\ &= (\theta_k - \theta_{k+1}) (\varphi_{k+1} - \ell_{k+1}) + [\theta_k \ell_k + (1 - \theta_k) \varphi_k] - \beta_k \\ &= (\theta_k - \theta_{k+1}) (\varphi_{k+1} - \ell_{k+1}) + \psi_k - \beta_k, \end{aligned} \tag{51}$$

where $\beta_k \geq 0$ for all $k \in \{0, 1, 2, \dots\}$ and $\beta_k \geq c_2 = 0.1c'_1 > 0$ for an infinite set of indices. Now, by the choice of θ at Algorithm 3.1, we have that

$$\theta_k - \theta_{k+1} + \omega_k \geq 0. \tag{52}$$

for all $k \in \{0, 1, 2, \dots\}$. By Assumption 5.2, there exists an upper bound $c > 0$ such that

$$|\varphi_k - \ell_k| \leq c$$

for all $k \in \{0, 1, 2, \dots\}$. Therefore, by (51) and (52), we have that

$$\begin{aligned} \psi_{j+1} &= (\theta_j - \theta_{j+1} + \omega_j (\varphi_{j+1} - \ell_{j+1}) + \psi_j - \beta_j - \omega_j) (\varphi_{j+1} - \ell_{j+1}) \\ &\leq (\theta_j - \theta_{j+1} + \omega_j) c + \psi_j - \beta_j + \omega_j c \\ &= (\theta_j - \theta_{j+1}) c + \psi_j - \beta_j + 2\omega_j c \end{aligned}$$

for $j = 0, 1, \dots, k-1$. Adding these k inequalities, we obtain

$$\psi_k \leq \psi_0 + (\theta_0 - \theta_k) c + \sum_{j=0}^{k-1} 2c\omega_j - \sum_{j=0}^{k-1} \beta_j \leq \psi_0 + 2c + \sum_{j=0}^{k-1} 2c\omega_j - \sum_{j=0}^{k-1} \beta_j \tag{53}$$

for all $k \geq 1$. Since the series $\sum_{j=0}^{\infty} \omega_j$ is convergent, and β_k is bounded away from 0 for $k \in K_1$, (53) implies that ψ_k is unbounded below. This contradicts Assumption 5.2. \square

6 The Tangent Cauchy Direction Tends to Zero

In Section 5 we proved that d_n^k , the Cauchy-like direction related to feasibility, tends to zero. In many circumstances, for example, when Ω is bounded and all the stationary points corresponding to the minimization of $\varphi(x)$ on Ω are regular, this property implies that $\|C(x^k)\| \rightarrow 0$. In other situations the property $\|d_n^k\| \rightarrow 0$ is not sufficient to guarantee that $\|C(x^k)\|$ tends to 0. In those unfortunate cases, the algorithm generally converges to a non-feasible local minimizer of $\varphi(x)$ on Ω . This possibility is present in all known practical algorithms for Nonlinear Programming. Many times, global convergence theories include

assumptions that guarantee that stationary points of φ are necessarily feasible points. Here we are going to state the property $\|C(x^k)\| \rightarrow 0$ as an additional assumption, observing that the results of Section 5 imply that, very likely, it will be satisfied if the feasible region is nonempty and, certainly, it will hold if the assumption on the regularity of feasible points is fulfilled. Needless to say, the results on Section 5 imply that $\|d_n^k\|$ tends to zero also when the feasible region is empty. Obviously, in this case $\|C(x^k)\|$ does not tend to zero.

Assumption 6.1. *Algorithm 3.1 generates an infinite sequence $\{x^k\}$ such that*

$$\lim_{k \rightarrow \infty} \|C(x^k)\| = 0. \quad (54)$$

As we discussed in Section 2, optimality is related to annihilation of the tangent direction $d_t^{k,i}$. Therefore, it will be natural to prove that the sequence defined by our model algorithm satisfies, in some sense, $d_t^{k,i} \rightarrow 0$. This will be the goal of the present section. We will proceed, as in Section 5, by contradiction. In Hypothesis 6.1 it will be assumed that $\|d_t^{k,i}\|$ is bounded away from zero. A number of results will be deduced from this hypothesis and, finally, we will get a contradiction.

Hypothesis 6.1. *Let $\{x^k\}$ be an infinite sequence generated by Algorithm 3.1. There exists $k_0 \in \{0, 1, 2, \dots\}$, $\varepsilon > 0$ such that*

$$\|d_t^{k,i}\| \geq \varepsilon \quad (55)$$

for all $k \geq k_0$, $i = 0, 1, \dots, j(k)$.

The first result deduced from Hypothesis 6.1 is that the directional derivative of f along $d_t^{k,i}$ is bounded away from zero.

Lemma 6.1. *Suppose that Assumption 6.1 and Hypothesis 6.1 hold. Then, there exists $c_5 > 0$ such that*

$$\langle d_t^{k,i}, \nabla f(y^{k,i}) \rangle \leq -c_5 \quad (56)$$

for all $k \geq k_0$, $i = 0, 1, \dots, j(k)$

Proof. We deduce (23) as in Lemma 4.4. Then, (56) follows defining $c_5 = \frac{\varepsilon^2}{2\eta}$. \square

Since directional derivatives of f in the tangent space coincide with directional derivatives of the Lagrangian, it follows, by Lemma 6.1, that the decrease of the Lagrangian (with fixed multipliers) from $y^{k,i}$ to $z^{k,i}$ is proportional to $\delta_{k,i}$. This is proved in Lemma 6.2 and Corollary 6.1.

Lemma 6.2. *Suppose that Hypothesis 6.1 and Assumptions 5.1, 5.2 and 6.1 hold. Then, there exists $\tilde{t} > 0$ such that*

$$\ell(z^{k,i}, \lambda^k) \leq \ell(y^{k,i}, \lambda^k) - \min \{0.1, 0.01\tilde{t}, 0.02 \frac{\delta_{k,i}}{\varepsilon}\} c_5. \quad (57)$$

for all $k \geq k_0$, $i = 0, 1, \dots, j(k)$.

Proof. By Assumptions 5.1, 5.2 and 6.1 and Hypothesis 6.1, we have, for all $k \geq k_0$, $i = 0, 1, \dots$, $t \geq 0$,

$$\begin{aligned}
\ell(y^{k,i} + td_t^{k,i}, \lambda^k) &\leq \ell(y^{k,i}, \lambda^k) + t\langle \nabla \ell(y^{k,i}, \lambda^k), d_t^{k,i} \rangle + \frac{t^2}{2}L_4\|d_t^{k,i}\|^2 \\
&\leq \ell(y^{k,i}, \lambda^k) + 0.1t\langle \nabla f(y^{k,i}), d_t^{k,i} \rangle + 0.9t\langle \nabla f(y^{k,i}), d_t^{k,i} \rangle + \frac{t^2}{2}L_4M_6^2 \\
&\leq \ell(y^{k,i}, \lambda^k) + 0.1t\langle \nabla f(y^{k,i}), d_t^{k,i} \rangle + 0.9t(-c_5) + \frac{t^2}{2}L_4M_6^2 \\
&\leq \ell(y^{k,i}, \lambda^k) + 0.1t\langle \nabla f(y^{k,i}), d_t^{k,i} \rangle - t(0.9c_5 - \frac{t}{2}L_4M_6^2).
\end{aligned}$$

Therefore, defining $\tilde{t} = \frac{1.8c_5}{L_4M_6^2}$, we have that

$$\ell(y^{k,i} + td_t^{k,i}, \lambda^k) \leq \ell(y^{k,i}, \lambda^k) + 0.1t\langle \nabla f(y^{k,i}), d_t^{k,i} \rangle \quad (58)$$

for all $t \in [0, \tilde{t}]$, $k \geq k_0$, $i = 0, 1, \dots$

Now, fix k and i and, as in Lemma 4.3, let us consider the cases $\bar{t}(k, i, 2) \leq \tilde{t}$ and $\bar{t}(k, i, 2) > \tilde{t}$. If $\bar{t}(k, i, 2) \leq \tilde{t}$ we have, by (58) and (19), that $\underline{t}(k, i, 2) = \bar{t}(k, i, 2)$. Moreover, if $\bar{t}(k, i, 2) < 1$ we have that

$$\|y^{k,i} + \bar{t}(k, i, 2)d_t^{k,i}\| = \delta_{k,i}, \quad \text{so,} \quad \|\bar{t}(k, i, 2)d_t^{k,i}\| \geq \|y^{k,i} + \bar{t}(k, i, 2)d_t^{k,i}\| - \|y^{k,i}\| \geq 0.2\delta_{k,i}.$$

Therefore, $\bar{t}(k, i, 2) \leq \tilde{t}$ implies that

$$\underline{t}(k, i, 2) = \bar{t}(k, i, 2) \geq \min \left\{ 1, 0.2 \frac{\delta_{k,i}}{\varepsilon} \right\}.$$

From this inequality we deduce that

$$\begin{aligned}
\ell(y^{k,i} + \underline{t}(k, i, 2)d_t^{k,i}, \lambda^k) &\leq \ell(y^{k,i}, \lambda^k) + 0.1\underline{t}(k, i, 2)\langle \nabla f(y^{k,i}), d_t^{k,i} \rangle \\
&\leq \ell(y^{k,i}, \lambda^k) + 0.1 \min \left\{ 1, 0.2 \frac{\delta_{k,i}}{\varepsilon} \right\} \langle \nabla f(y^{k,i}), d_t^{k,i} \rangle \\
&\leq \ell(y^{k,i}, \lambda^k) - \min \left\{ 0.1, 0.02 \frac{\delta_{k,i}}{\varepsilon} \right\} c_5.
\end{aligned} \quad (59)$$

Now, we analyze the case $\bar{t}(k, i, 2) > \tilde{t}$. By the definition of \tilde{t} and the backtracking procedure at Algorithm 3.3, we have that $\underline{t}(k, i, 2) \geq \tilde{t}/10$. Therefore,

$$\begin{aligned}
\ell(y^{k,i} + \underline{t}(k, i, 2)d_t^{k,i}, \lambda^k) &\leq \ell(y^{k,i}, \lambda^k) + 0.1\underline{t}(k, i, 2)\langle \nabla f(y^{k,i}), d_t^{k,i} \rangle \\
&\leq \ell(y^{k,i}, \lambda^k) + 0.01\tilde{t}\langle \nabla f(y^{k,i}), d_t^{k,i} \rangle \\
&\leq \ell(y^{k,i}, \lambda^k) - 0.01\tilde{t}c_5.
\end{aligned} \quad (60)$$

Clearly, (57) follows from (59) and (60). \square

Corollary 6.1. *Suppose that Hypothesis 6.1 and Assumptions 5.1, 5.2 and 6.1 hold. Then, there exist $c_6 > 0$, $\tilde{\delta} > 0$ such that*

$$\ell(y^{k,i}, \lambda^k) - \ell(z^{k,i}, \lambda^k) \geq c_6 \min \{\tilde{\delta}, \delta_{k,i}\} \quad (61)$$

for all $k \geq k_0, i = 0, 1, \dots, j(k)$.

Proof. The inequality (61) follows from Lemma 6.2 defining

$$c_6 = \frac{0.02c_5}{\varepsilon} \quad \text{and} \quad \tilde{\delta} = \varepsilon \min \{5, 0.5\tilde{t}\}.$$

□

The choice of d_n^k and $\underline{t}(k, i, 1)$ and Assumption 5.2 guarantee that $\|\underline{t}(k, i, 1)d_n^k\|$ is not greater than a multiple of $\|C(x^k)\|$. Therefore, if we decided to choose $y^{k,i} = x^k + \underline{t}(k, i, 1)d_n^k$ (which is admissible but not recommendable) the property $\|y^{k,i} - x^k\| \leq c\|C(x^k)\|$ (for some $c > 0$) should hold. In fact, we need that property for the convergence proof, since it tells that, when x^k is close to feasibility, the displacement towards feasibility must be small. Therefore, we will assume that the choice of $y^{k,i}$ satisfies that property and we will show, in Section 7, that there exist reasonable choices of $y^{k,i}$ for which that assumption holds.

Assumption 6.2. *There exists $c_2 > 0$ such that, if $\{x^k\}$ is generated by Algorithm 3.1, we have*

$$\|y^{k,i} - x^k\| \leq c_2\|C(x^k)\| \quad (62)$$

for all $k \in \{0, 1, 2, \dots\}, i = 0, 1, \dots, j(k)$.

The following lemma has two objectives. On one hand, by Corollary 6.1 and the definition of *Pred*, we could expect that the predicted decrease of the merit function should be proportional to $\theta\delta_{k,i}$. However, a deterioration of optimality can occur from x^k to $y^{k,i}$ and a (hopefully small) deterioration of feasibility can occur from $y^{k,i}$ to $z^{k,i}$. Therefore, we can only expect that the predicted decrease should be proportional to $\theta\delta_{k,i}$ minus a multiple of $\|C(x^k)\|$. See (63) below. This will imply that, if $\|C(x^k)\|$ is less than some multiple of $\delta_{k,i}$, the predicted decrease will be positive and large enough. In particular, it will not be necessary to decrease the penalty parameter for those values of $C(x^k)$ and $\delta_{k,i}$.

Lemma 6.3. *Suppose that Hypothesis 6.1 and Assumptions 5.1, 5.2, 6.1 and 6.2 hold. Then, there exist $c_8, c_9, c_{10}, \alpha > 0$ such that for all $k \geq k_0, i = 0, 1, \dots, j(k), \theta \in [0, 1]$, we have that*

$$Pred_{k,i}(\theta) - \frac{1}{2}[\varphi(x^k) - \varphi(y^{k,i})] \geq \theta[c_6 \min \{\tilde{\delta}, \delta_{k,i}\} - c_8\|C(x^k)\|] - c_9\|C(x^k)\|$$

and

$$Pred_{k,i}(\theta) \geq \theta c_6 \min \{\tilde{\delta}, \delta_{k,i}\} - c_{10}\|C(x^k)\| \quad (63)$$

where $\tilde{\delta}$ and c_6 are defined in Corollary 6.1.

Moreover, there exists $k_1 \geq k_0$ such that, whenever $\|C(x^k)\| \leq \alpha\delta_{k,i}$, we have

$$Pred_{k,i}(\theta) \geq \frac{1}{2}[\varphi(x^k) - \varphi(y^{k,i})]$$

for all $\theta \in [0, 1]$ and

$$\theta_{k,i} = \theta_{k,j} = \theta_{k,-1}$$

for all $k \geq k_1$, $i = 0, 1, \dots, j(k)$, $j = 0, 1, \dots, i - 1$.

Proof. For all $k \geq k_0$, $i = 0, 1, \dots, j(k)$ we have that

$$\begin{aligned} & \ell(x^k, \lambda^k) - \ell(z^{k,i}, \lambda^k) + \langle \lambda^k - \lambda^{k,i}, C(y^{k,i}) \rangle \\ & \geq \ell(y^{k,i}, \lambda^k) - \ell(z^{k,i}, \lambda^k) - |\ell(y^{k,i}, \lambda^k) - \ell(x^k, \lambda^k)| - \|\lambda^k - \lambda^{k,i}\|_2 \|C(y^{k,i})\|_2 \end{aligned}$$

But, by Assumptions 5.1, 5.2 and 6.2,

$$\begin{aligned} |\ell(y^{k,i}, \lambda^k) - \ell(x^k, \lambda^k)| & \leq \frac{L_4}{2} \|x^k - y^{k,i}\|^2 + |\langle \nabla \ell(x^k, \lambda^k), y^{k,i} - x^k \rangle| \\ & \leq (M_5 + M_1 M_3) \|y^{k,i} - x^k\| + \frac{L_4}{2} \|y^{k,i} - x^k\|^2 \\ & \leq [(M_5 + M_1 M_3) c_7 + \frac{L_4}{2} \|C(x^k)\|] \|C(x^k)\| \end{aligned}$$

By Assumption 6.1, $\{\|C(x^k)\|\}$ is bounded. So, by Corollary 6.1, since $\|C(y^{k,i})\|_2 \leq \|C(x^k)\|_2$, there exists $c_8 > 0$ such that

$$\ell(x^k, \lambda^k) - \ell(z^{k,i}, \lambda^k) + \langle \lambda^k - \lambda^{k,i}, C(y^{k,i}) \rangle \geq c_6 \min \{\tilde{\delta}, \delta_{k,i}\} - c_8 \|C(x^k)\|$$

for all $k \geq k_0$, $i = 0, 1, \dots, j(k)$. This implies that

$$Pred_{k,i}(\theta) \geq \theta [c_6 \min \{\tilde{\delta}, \delta_{k,i}\} - c_8 \|C(x^k)\|]$$

for all $k \geq k_0$, $i = 0, 1, \dots, j(k)$, $\theta \in [0, 1]$.

Now, by Assumptions 5.1 and 5.2, there exist a norm-dependent constant $c > 0$ such that

$$\varphi(x^k) - \varphi(y^{k,i}) \leq \frac{L_2}{2} \|y^{k,i} - x^k\|^2 + c \|\nabla \varphi(x^k)\| \|y^{k,i} - x^k\| \leq \frac{L_2}{2} c_7^2 \|C(x^k)\|^2 + c M_7 c_7 \|C(x^k)\|.$$

So, there exists $c_9 > 0$ such that

$$\varphi(x^k) - \varphi(y^{k,i}) \leq c_9 \|C(x^k)\|$$

for all $k \geq k_0$, $i = 0, 1, \dots, j(k)$. Therefore,

$$\begin{aligned} Pred_{k,i}(\theta) - \frac{1}{2} [\varphi(x^k) - \varphi(y^{k,i})] & \geq \theta [c_6 \min \{\tilde{\delta}, \delta_{k,i}\} - c_8 \|C(x^k)\|] - c_9 \|C(x^k)\|. \\ & = \theta c_6 \min \{\tilde{\delta}, \delta_{k,i}\} - c_{10} \|C(x^k)\| \end{aligned}$$

for all $k \geq k_0$, $i = 0, 1, \dots, j(k)$, where $c_{10} = c_8 + c_9$.

Let $k_1 \geq k_0$ be such that $c_6 \tilde{\delta} - c_{10} \|C(x^k)\| \geq 0$ for all $k \geq k_1$. Therefore, if $k \geq k_1$ and

$$\|C(x^k)\| \leq \frac{c_6}{c_{10}} \tilde{\delta}_{k,i}$$

we have that

$$\text{Pred}(k, 1) \geq \frac{1}{2}[\varphi(x^k) - \varphi(y^{k,i})].$$

This implies that, defining $\alpha = c_6/c_{10}$, whenever $k \geq k_1$, $\|C(x^k)\| \leq \alpha\delta_{k,i}$, $i = 0, 1, \dots, j(k)$, we have that

$$\theta_{k,i} = \theta_{k,i-1}.$$

Since $\delta_{k,i-1} \geq \delta_{k,i}$, this implies that

$$\theta_{k,i} = \theta_{k,j} = \theta_{k,-1} = \theta_k^{\text{large}}$$

for all $j = 0, 1, \dots, i-1$. This completes the proof of the Lemma. \square

In the following lemma we prove that, under Hypothesis 6.1, the penalty parameter θ_k tends to zero. This is important for the final convergence proof because, by (43), it implies that **Pred** is a high order approximation of **Ared** ($|\mathbf{Ared} - \mathbf{Pred}| = o(\delta_{k,i}^2)$). Recall that $\theta_k \rightarrow 0$ is not a property of all sequences generated by Algorithm 3.1, but only of sequences that satisfy Hypothesis 6.1.

Lemma 6.4. *Suppose that Hypothesis 6.1 and Assumptions 5.1, 5.2, 6.1 and 6.2 hold. Then*

$$\lim_{k \rightarrow \infty} \theta_k = 0.$$

Proof. By Lemma 5.1 and Assumption 6.1, there exists $c_{11} > 0$ such that

$$|\mathbf{Ared}_{k,i} - \mathbf{Pred}_{k,i}| \leq c_{11}(\delta_{k,i}^2 + \delta_{k,i}^4) \quad (64)$$

for all $k \geq k_0$, $i = 0, 1, \dots, j(k)$.

Suppose that $\{\theta_k\}$ does not converge to 0. Since, by Lemma 3.1, $\{\theta_k\}$ is convergent, there exists $k_2 \geq k_1$, $\bar{\theta} > 0$ such that $\theta_k \geq \bar{\theta} > 0$ for all $k \geq k_2$. This implies that

$$0 < \bar{\theta} \leq \theta_k = \theta_{k,j(k)} \leq \theta_{k,i}$$

for all $k \geq k_2$, $i = 0, 1, \dots, j(k)$. Therefore, by (63),

$$\mathbf{Pred}_{k,i} \geq \bar{\theta}c_6 \min\{\tilde{\delta}, \delta_{k,i}\} - c_{10}\|C(x^k)\|$$

for all $k \geq k_2$, $i = 0, 1, \dots, j(k)$. In particular, using (16), we obtain that

$$\mathbf{Ared}_k \geq 0.1\mathbf{Pred}_k \geq 0.1[\bar{\theta}c_6 \min\{\tilde{\delta}, \delta_k\} - c_{10}\|C(x^k)\|].$$

Suppose, for a moment, that there exists $\hat{\delta} > 0$ such that $\delta_k \geq \hat{\delta}$ for some subsequence. In that case, taking k large enough, we obtain that

$$\bar{\theta}c_6 \min\{\tilde{\delta}, \delta_k\} - c_{10}\|C(x^k)\| \geq \frac{1}{2}\bar{\theta}c_6 \min\{\tilde{\delta}, \hat{\delta}\}.$$

This implies that \mathbf{Ared}_k is bounded away from zero for some subsequence. So, defining ψ_k as in Theorem 5.1, we obtain that $\psi_k \rightarrow -\infty$, which leads, as in Theorem 5.1, to a contradiction.

So, we can assume that

$$\lim_{k \rightarrow \infty} \delta_k = \lim_{k \rightarrow \infty} \delta_{k,j(k)} = 0.$$

Let $\bar{\delta} \leq \min \{\tilde{\delta}, \delta_{min}\}$ be such that

$$|\mathbf{Ared}_{k,i} - \mathbf{Pred}_{k,i}| \leq \frac{\bar{\theta} c_6 \bar{\delta}}{40} \quad (65)$$

whenever $\delta_{k,i} \leq \bar{\delta}$.

Let $k_3 \geq k_2$ be such that

$$c_{10} \|C(x_k)\| \leq \frac{\bar{\theta} c_6 \bar{\delta}}{20}$$

for $k \geq k_3$.

Let $k_4 \geq k_3$ be such that $\delta_k = \delta_{k,j(k)} \leq \bar{\delta}/10$ for all $k \geq k_4$. This implies that for all $k \geq k_4$, there exists a rejected step $i < j(k)$ such that $\delta_{k,i} \in [\bar{\delta}/10, \bar{\delta}]$.

However,

$$\mathbf{Pred}_{k,i} \geq \bar{\theta} c_6 \frac{\bar{\delta}}{10} - \bar{\theta} c_6 \frac{\bar{\delta}}{20} = \bar{\theta} c_6 \frac{\bar{\delta}}{20}$$

while $|\mathbf{Ared}_{k,i} - \mathbf{Pred}_{k,i}| \leq \frac{\bar{\theta} c_6 \bar{\delta}}{40}$. Therefore, $\frac{|\mathbf{Ared}_{k,i} - \mathbf{Pred}_{k,i}|}{\mathbf{Pred}_{k,i}} \leq 0.5$. Therefore, $\mathbf{Ared}_{k,i} \geq 0.1 \mathbf{Pred}_{k,i}$, contradicting the fact that $i < j(k)$. \square

Assumption 6.3 is the final one needed for proving that $d_t^{k,i} \rightarrow 0$. It says that when $\|C(x^k)\| \geq \alpha \delta_{k,i}/10$, the decrease of φ from x^k to $y^{k,i}$ is proportional to $\|C(x^k)\| \delta_{k,i}$. Unlike Assumption 6.2, this assumption is not necessarily satisfied by the choice $x^k + \underline{t}(k, i, 1) d_n^k$. However, we will see in Section 7 that it holds for quite reasonable choices of the feasibility step.

Assumption 6.3. *If Hypothesis 6.1 and Assumptions 5.1, 5.2, 6.1 and 6.2 hold, there exist $\varepsilon_1 > 0$, $c_{12} > 0$ such that for all $k \in \{0, 1, 2, \dots\}$, $i = 0, 1, \dots, j(k)$, if $\|C(x^k)\| \leq \varepsilon_1$, and $\|C(x^k)\| \geq \frac{\alpha}{10} \delta_{k,i}$, the following inequality holds:*

$$\varphi(x^k) - \varphi(y^{k,i}) \geq c_{12} \|C(x^k)\| \delta_{k,i}.$$

Theorem 6.6 is the main result of this section. Essentially, it states that Hypothesis 6.1 cannot be true. Therefore, the algorithm produces points that are arbitrary close to optimality.

Theorem 6.6. *Suppose that Assumptions 5.1, 5.2, 6.1, 6.2 and 6.3 hold. Then, there exists K_1 an infinite subset of $\{0, 1, 2, \dots\}$ such that for all $k \in K_1$ there exists $i(k) \in \{0, 1, \dots, j(k)\}$ such that*

$$\lim_{k \in K_1} \|d_t^{k,i(k)}\| = 0.$$

Proof. Suppose that Hypothesis 6.1 holds. By Assumption 6.3, (13) and (43), there exists $k_5 \in \{0, 1, 2, \dots\}$ such that whenever $\|C(x^k)\| \geq \frac{\alpha}{10}\delta_{k,i}$, $k \geq k_5$, $i = 0, 1, \dots, j(k)$, we have

$$\begin{aligned} \frac{|\mathbf{Ared}_{k,i} - \mathbf{Pred}_{k,i}|}{\mathbf{Pred}_{k,i}} &\leq \frac{c_3(\theta_{k,i}\delta_{k,i}10\|C(x^k)\|/\alpha + \delta_{k,i}^3 10\|C(x^k)\|/\alpha + \delta_{k,i}^2\|C(x^k)\|)}{0.5c_{12}\|C(x^k)\|\delta_{k,i}} \\ &\leq \frac{2c_3}{\alpha c_{12}}(10\theta_{k,i} + 10\delta_{k,i}^2 + \alpha\delta_{k,i}) \leq \frac{2c_3}{\alpha c_{12}}(10\theta_{k,i} + 1000\|C(x^k)\|^2/\alpha^2 + 10\|C(x^k)\|/\alpha). \end{aligned} \quad (66)$$

By (66), since $\|C(x^k)\| \rightarrow 0$ and $\theta_{k,i} \rightarrow 0$, there exists $k_6 \geq k_5$ such that whenever $k \geq k_6$, and $\delta_{k,i} \leq 10\|C(x^k)\|/\alpha$, the quotient $\frac{|\mathbf{Ared}_{k,i} - \mathbf{Pred}_{k,i}|}{\mathbf{Pred}_{k,i}}$ is less than 0.9 so that (16) holds and the trust region radius $\delta_{k,i}$ is accepted ($j(k) = i$). Let $k_7 \geq k_6$ be such that $10\|C(x^k)\|/\alpha \leq \delta_{min}$ for all $k \geq k_7$. If, for some $k \geq k_7$ we have that $i \leq j(k)$ and $\delta_{k,i} \leq \|C(x^k)\|/\alpha$, we necessarily have that $i \geq 1$ and $\delta_{k,i-1} \leq 10\|C(x^k)\|/\alpha$, which implies that $i - 1 = j(k)$. Therefore, for all $k \geq k_7$, we have that $\delta_k = \delta_{k,j(k)} > \|C(x^k)\|/\alpha$. Thus, for all $k \geq k_7$, $i = 0, 1, \dots, j(k)$ we have that

$$\delta_{k,i} > \|C(x^k)\|/\alpha.$$

By Hypothesis 6.1, Assumption 6.1 and Lemma 6.3 this implies that $\theta_{k,i} > \theta_{k-1}$ for all $k \geq k_7$. This fact contradicts that $\theta_k \rightarrow 0$. It follows that Hypothesis 6.1 cannot hold. Clearly, the negation of this hypothesis implies the thesis of the theorem. \square

7 Choice of the Feasibility Step

The objective of this section is to show that Assumption 6.3 is reasonable. In fact, we prove that, if an inexact-Newton type step (see Ref. 20) for the problem $C(x) = 0, x \in \Omega$ can be defined from any point $x \in \Omega$, then the choice of $y^{k,i}$ at Step 3 of Algorithm 2.2 can be made in a natural way that imply both Assumptions 6.2 and 6.3.

Assumption 7.1. *There exists $\varepsilon_2 > 0$, $r \in (0, 1)$, $c_{13} > 0$ such that for all $x \in \Omega$ such that $\|C(x)\| \leq \varepsilon_2$, there exists $\bar{y} \in \Omega$ such that*

$$\|A(x)(\bar{y} - x) + C(x)\|_2 \leq r\|C(x)\|_2, \quad (67)$$

$$\|\bar{y} - x\| \leq c_{13}\|C(x)\|. \quad (68)$$

Suppose that r is close to 1 and c_{13} is large. Then, Assumption 7.1 says that, if $\|C(x)\|$ is small enough, it is possible to obtain a (small but proportional to $\|C(x)\|$) decrease of the norm of the linear approximation of C , at a point \bar{y} which, in turn, cannot be very far from x . The increment $\bar{y} - x$ is an inexact-Newton step in the classical sense of Ref. 20. If such an increment does not exist (for $r \approx 1$ and c_{13} large), then the point x must be close to a

stationary point of φ . The absence of stationary nonfeasible points in a compact environment certainly imply Assumption 7.1. From a practical point of view, the increment $\bar{y} - x$ can be computed using quadratic solvers. Algorithm 7.1 uses this increment if it turns out to be better than the Cauchy-like increment computed by Algorithm 3.2.

Algorithm 7.1 – A particular choice for $y^{k,i}$

At Step 3 of Algorithm 3.2, if $\|C(x^k)\| \leq \varepsilon_2$, compute $s^k \in \mathbb{R}^n$, $\bar{y}^k = x^k + s^k$, such that

$$\|A(x^k)s^k + C(x^k)\|_2 \leq r\|C(x^k)\|_2, \quad (69)$$

$$\bar{y}^k \in \Omega \quad \text{and} \quad \|s^k\| \leq c_{13}\|C(x^k)\|. \quad (70)$$

If $\|s^k\| \leq 0.8\delta_{k,i}$, set $\sigma_{k,i} = 1$, else set

$$\sigma_{k,i} = \frac{0.8\delta_{k,i}}{\|s^k\|}. \quad (71)$$

Finally, set

$$y^{k,i} = \text{Argmin} \{\varphi(x^k + \sigma_{k,i}s^k), \varphi(x^k + \underline{t}(k, i, 1)d_n^k)\}. \quad (72)$$

Clearly, if Assumption 7.1 holds, Algorithm 7.1 is well defined.

Lemma 7.1. *Suppose that Assumptions 5.1, 5.2 and 7.1 hold and that $y^{k,i}$ is chosen using Algorithm 7.1. Then, Assumptions 6.2 and 6.3 also hold.*

Proof. By Assumption 5.2, there exists $c > 0$, a norm-dependent constant, such that

$$\|d_n^k\| \leq c\|C(x^k)\|$$

for all $k \in \{0, 1, 2, \dots\}$. Since $\underline{t}(k, i, 1) \leq 1$, it follows that $y^{k,i}$ satisfies Assumption 6.2, when $y^{k,i} = x^k + \underline{t}(k, i, 1)d_n^k$. Moreover, since $\sigma_{k,i} \leq 1$, Assumption 6.2 is also satisfied when $y^{k,i} = x^k + \sigma_{k,i}s^k$. Therefore, it is only necessary to prove that Assumption 6.3 holds.

By (35) and (70), there exists $c_{14} > 0$ such that, if $\|C(x^k)\| \leq \varepsilon_2$,

$$\|C(\bar{y}^k)\|_2 \leq \|C(x^k) + A(x^k)s^k\|_2 + c_{14}\|C(x^k)\|_2^2.$$

So, by (67),

$$\|C(\bar{y}^k)\|_2 \leq r\|C(x^k)\|_2 + c_{14}\|C(x^k)\|_2^2.$$

Therefore,

$$\|C(\bar{y}^k)\|_2^2 \leq r^2\|C(x^k)\|_2^2 + 2rc_{14}\|C(x^k)\|_2^3 + c_{14}^2\|C(x^k)\|_2^4.$$

So,

$$\|C(x^k)\|_2^2 - \|C(\bar{y}^k)\|_2^2 \geq [1 - r^2 - 2rc_{14}\|C(x^k)\|_2 - c_{14}^2\|C(x^k)\|_2^2]\|C(x^k)\|_2^2.$$

So, when $\|C(x^k)\| \leq \varepsilon_2$ and $2rc_{14}\|C(x^k)\|_2 + c_{14}^2\|C(x^k)\|_2^2 \leq r^2/2$, we have that

$$\|C(x^k)\|_2^2 - \|C(\bar{y}^k)\|_2^2 \geq (1 - \frac{r^2}{2})\|C(x^k)\|_2^2.$$

Therefore, there exists $\varepsilon_3 > 0$ such that, whenever $\frac{\alpha}{10}\delta_{k,i} \leq \|C(x^k)\| \leq \varepsilon_3$,

$$\varphi(x^k) - \varphi(\bar{y}^k) \geq \frac{1}{20}\left(1 - \frac{r^2}{2}\right)c'\alpha\|C(x^k)\|\delta_{k,i}, \quad (73)$$

where $c' > 0$ is a norm-dependent constant.

On the other hand, by (67), if $\sigma_{k,i} < 1$, we have that

$$\begin{aligned} \|C(x^k) + \sigma_{k,i}A(x^k)s^k\|_2^2 &\leq (1 - \sigma_{k,i})\|C(x^k)\|_2^2 + \sigma_{k,i}\|C(x^k) + A(x^k)s^k\|_2^2 \\ &\leq (1 - \sigma_{k,i})\|C(x^k)\|_2^2 + \sigma_{k,i}r^2\|C(x^k)\|_2^2 = [1 - \sigma_{k,i}(1 - r^2)]\|C(x^k)\|_2^2. \end{aligned}$$

Therefore, by (71),

$$\|C(x^k)\|_2^2 - \|C(x^k) + \sigma_{k,i}A(x^k)s^k\|_2^2 \geq \sigma_{k,i}(1 - r^2)\|C(x^k)\|_2^2 = \frac{0.8\delta_{k,i}}{\|s^k\|}(1 - r^2)\|C(x^k)\|_2^2.$$

Thus, by (70),

$$\|C(x^k)\|_2^2 - \|C(x^k) + \sigma_{k,i}A(x^k)s^k\|_2^2 \geq \frac{0.8(1 - r^2)c''}{c_{13}}\|C(x^k)\|\delta_{k,i}, \quad (74)$$

where $c'' > 0$ is a norm-dependent constant.

Now,

$$\begin{aligned} &\|C(x^k)\|_2^2 - \|C(x^k + \sigma_{k,i}s^k)\|_2^2 \\ &= \|C(x^k)\|_2^2 - \|C(x^k + \sigma_{k,i}s^k) - [C(x^k) + \sigma_{k,i}A(x^k)s^k] + [C(x^k) + \sigma_{k,i}A(x^k)s^k]\|_2^2 \\ &\geq \|C(x^k)\|_2^2 - \|C(x^k) + \sigma_{k,i}A(x^k)s^k\|_2^2 \\ &\quad - \|C(x^k + \sigma_{k,i}s^k) - [C(x^k) + \sigma_{k,i}A(x^k)s^k]\|_2^2 \\ &\quad - 2\|C(x^k + \sigma_{k,i}s^k) - [C(x^k) + \sigma_{k,i}A(x^k)s^k]\|_2\|C(x^k) + \sigma_{k,i}A(x^k)s^k\|_2. \end{aligned} \quad (75)$$

But, by (35), (70) and (71),

$$\begin{aligned} \|C(x^k + \sigma_{k,i}s^k) - [C(x^k) + \sigma_{k,i}A(x^k)s^k]\|_2 &\leq \frac{c'''L_1}{2}\sigma_{k,i}^2\|s^k\|^2 \\ &= 0.32c'''L_1\delta_{k,i}^2, \end{aligned} \quad (76)$$

whenever $\|C(x^k)\| \leq \varepsilon_2$, where $c''' > 0$ is a norm-dependent constant.

Moreover, by (74),

$$\|C(x^k)\|_2^2 - \|C(x^k) + \sigma_{k,i}A(x^k)s^k\|_2^2 \geq c_{14}\|C(x^k)\|\delta_{k,i}, \quad (77)$$

where $c_{14} = \frac{0.8(1-r^2)c''}{c_{13}}$.

Finally, by Assumption 5.2,

$$\|C(x^k) + \sigma_{k,i}A(x^k)s^k\|_2 \leq c''''[\|C(x^k)\| + M_3\|C(x^k)\|], \quad (78)$$

where c'''' is a norm-dependent constant.

So, by (75)–(78), we have that, if $\|C(x^k)\| \leq \varepsilon_2$,

$$\|C(x^k)\|_2^2 - \|C(x^k + \sigma_{k,i}s^k)\|_2^2 \geq c_{14}\|C(x^k)\|\delta_{k,i} - c_{15}^2\delta_{k,i}^4 - c_{16}\delta_{k,i}^2\|C(x^k)\|,$$

where $c_{15} = 0.32c'''L_1$ and $c_{16} = 2c_{15}c''''(1 + M_3)$.

Therefore, if $\delta_{k,i} \leq 10\|C(x^k)\|/\alpha$ and $\|C(x^k)\| \leq \varepsilon_2$,

$$\begin{aligned} & \|C(x^k)\|_2^2 - \|C(x^k + \sigma_{k,i}s^k)\|_2^2 \\ & \geq c_{14}\|C(x^k)\|\delta_{k,i} - c_{15}^2\delta_{k,i}(10\|C(x^k)\|/\alpha)^3 - c_{16}\|C(x^k)\|\delta_{k,i}10\|C(x^k)\|/\alpha \\ & = c_{14}\|C(x^k)\|\delta_{k,i} - c_{15}^2\|C(x^k)\|\delta_{k,i}(10/\alpha)^2\|C(x^k)\|^2 - c_{16}\|C(x^k)\|\delta_{k,i}10\|C(x^k)\|/\alpha \\ & = \|C(x^k)\|\delta_{k,i}[c_{14} - c_{15}^2(10/\alpha)^2\|C(x^k)\|^2 - c_{16}(10/\alpha)\|C(x^k)\|] \end{aligned}$$

So, if and ε_2 is small enough, we obtain that

$$\varphi(x^k) - \varphi(x^k + \sigma_{k,i}s^k) \geq c_{17}\delta_{k,i}\|C(x^k)\|, \quad (79)$$

where $c_{17} = c_{14}/2$. By (73), (79) and (72), it follows that Assumption 6.3 holds. \square

8 A Particular Implementation

With the aim of testing the reliability of the Model Algorithm 3.1, we implemented a particular version of it, the description of which is displayed below. This implementation deals with problems in which Ω is a bounded n -dimensional box ($\Omega = \{x \in \mathbb{R}^n \mid l \leq x \leq u\}$), so that the boundedness assumption is automatically satisfied. The general term of the convergent series $\sum \omega_k$ was chosen as $\omega_k = \omega/(1.1)^{k+1}$, where $\omega \geq 0$ is a parameter that reflects the level of nonmonotonicity that we wish for the penalty parameter. The values chosen for the scale-dependent parameters $\delta_{0,0}$, δ_{min} , γ and η were $\delta_{0,0} = 1$, $\delta_{min} = \gamma = \delta = 10^{-3}$. The initial penalty parameter θ_{-1} was 0.9.

Steps 1 and 2 of Algorithm 3.1 were implemented as described in Section 3. For the implementation of Step 3 we considered the auxiliary box-constrained problem

$$\text{Minimize } \varphi(x) \quad \text{subject to} \quad \|x - x^k\|_\infty \leq \min \{0.8\delta_{k,i}, 10^8\|C(x^k)\|_2\}. \quad (80)$$

Problem (80) was solved using the algorithm described in Ref. 21 (called BOX from here on), with $x^k + \underline{t}(k, i, 1)d_n^k$ as initial approximation. According to Ref. 22, the practical behavior of BOX is similar to the box-constraint solver used in the package LANCELOT Ref. 23. Obviously, we do not need to solve (80) exactly, so we use a stopping criterion based on the projected gradient of φ . Namely, the execution of BOX is stopped when the projected gradient at the current iterate is less than one tenth the projected gradient at the initial point. The approximate solution of (80) is accepted as feasible trial point $y^{k,i}$ if it satisfies Assumption 6.3 with $c_{12} = 10^{-5}$. Otherwise Algorithm 7.1 is used for computing the feasible step. However, such a failure was never detected in the experiments.

The computation of $d_t^{k,i}$ involves a well-conditioned quadratic programming subproblem, the dual of which is easily solved using a quadratic solver subroutine of BOX.

For the implementation of Step 5 we consider the auxiliary linearly constrained problem

$$\text{Minimize } \ell(z, \lambda^k) \quad \text{subject to} \quad A(y^{k,i})(z - y^{k,i}) = 0, \|z - x^k\|_\infty \leq \delta_{k,i}. \quad (81)$$

A rough approximation of the solution of (81) was computed using an iterative procedure described below. Analogously to the case of (80), we used $y^{k,i} + \underline{t}(k, i, 2)d_t^{k,i}$ as initial point.

The choice of $\delta_{k,0}$ at the beginning of iteration k ($k > 0$) obeyed the following rule: If $\mathbf{Ared}_k \geq \mathbf{Pred}_k$, then $\delta_{k,0} = \max\{\delta_{min}, 2\delta_k\}$, otherwise $\delta_{k,0} = \max\{\delta_{min}, \delta_k\}$. The choice of $\delta_{k,i+1}$ after a failure of the sufficient decrease test is as follows: $\delta_{k,i+1} = \|z^{k,i} - x^k\|_\infty/2$ if $\mathbf{Ared}_{k,i} > 0$ and $\delta_{k,i+1} = \|z^{k,i} - x^k\|_\infty/10$ otherwise. In these experiments we set $\lambda^{k,i} = 0$ for all k, i .

The procedure used for dealing with (81) is a very simple algorithm devised only with the purpose of testing the reliability of Algorithm 3.1. At each iteration of this algorithm, given the current iterate z , a search direction is computed as the difference between the projection of $z - \nabla f(z)$ on the feasible region of (81) and z . Along this direction we perform a cubic interpolation backtracking search with a single descent condition. The procedure stops when the 2-norm of the search direction is less than 10^{-3} or when a maximum of 100 points have been tried.

A FORTRAN 77 double precision code was written according to the implementation described above. The test problems are taken from a comprehensive study concerning several Sphere Packing Problems (see Ref. 24). All the tests were run using a Pentium-166 MHz. The same problems were solved using an Augmented Lagrangian method that uses the strategy of (Ref. 23) with BOX as underlying box-constraint solver.

The problems are:

Problem 1: Given $\nu, m \in \{0, 1, 2, \dots\}$, $n = \nu m$,

$$\text{Minimize } \sum_{i=1}^{m-1} \sum_{j=i+1}^m \sum_{k=1}^{\nu} x_{(i-1)\nu+k} x_{(j-1)\nu+k}$$

$$\text{subject to } \sum_{k=1}^{\nu} x_{(i-1)\nu+k}^2 - 1 = 0, i = 1, \dots, m, -10 \leq x_i \leq 10, i = 1, \dots, n.$$

The initial points 1, 2 and 3 are random with $x_i \in [-10, 10]$ for all $i = 1, \dots, n$. The fourth initial point is (1, 2, 3, 4, 5, 6, 7, 1, 2, ...).

Problem 2: Given $\nu, m, p \in \{0, 1, 2, \dots\}$, $n = \nu m$,

$$\text{Minimize } \sum_{i=1}^{m-1} \sum_{j=i+1}^m \frac{1}{(\sum_{k=1}^{\nu} (x_{(i-1)\nu+k} - x_{(j-1)\nu+k})^2 + 1)^p}$$

$$\text{subject to } \sum_{k=1}^{\nu} x_{(i-1)\nu+k}^2 - 1 = 0, i = 1, \dots, m, -10 \leq x_i \leq 10, i = 1, \dots, n.$$

The initial point is random with $x_i \in [-10, 10]$ for all $i = 1, \dots, n$.

Problem 3: Given $\nu, m \in \{0, 1, 2, \dots\}$, $n = \nu m$,

$$\begin{aligned} \text{Minimize } & \sum_{i=1}^{m-1} \sum_{j=i+1}^m \left(\sum_{k=1}^{\nu} x_{(i-1)\nu+k} - x_{(j-1)\nu+k} \right)^2)^{-6} - 2 \left(\sum_{k=1}^{\nu} x_{(i-1)\nu+k} - x_{(j-1)\nu+k} \right)^2)^{-3} \\ \text{subject to } & \sum_{k=1}^{\nu} x_{(i-1)\nu+k}^2 - 1 = 0, i = 1, \dots, m, -10 \leq x_i \leq 10, i = 1, \dots, n. \end{aligned}$$

The initial point is random with $x_i \in [-10, 10]$ for all $i = 1, \dots, n$.

The results are given in Table 1. Under the column ALM we report the results of the Augmented Lagrangian Method (Number of subproblems solved and CPU time, in seconds). Under the column TPM we report the results of the Two-Phase Algorithm described above (iterations and CPU time, in seconds). The execution of this algorithm was stopped when the current iteration arrived to the same level of feasibility and functional value as the solution obtained by ALM. In two cases (marked with (*)), 100 iterations of TPM were not sufficient to obtain the required precision.

Problem	ν, m, p	n	Initial point	ALM	ω	TPM
1	$\nu = 4, m = 500$	2000	1	8, 1420	10	3, 74
			2	8, 1399	10	4, 91
			3	8, 1792	10	4, 81
			4	7, 1849	10	9, 446
2	$\nu = 4, m = 25, p = 1$	100	1	9, 527	0	4, 20
					10	7, 104
					1000	7, 103
	$\nu = 4, m = 25, p = 2$	100	1	8, 280	0	4, 12
					10	8, 134
					1000	8, 134
	$\nu = 4, m = 25, p = 4$	100	1	8, 253	0	5, 29
					10	4, 79
					1000	4, 79
	$\nu = 4, m = 25, p = 10$	100	1	7, 527	0	58, 67
					10	100, 177 (*)
					1000	100, 177 (*)
3	$\nu = 3, m = 60$	180	1	14, 988	0	37, 1412
	$\nu = 4, m = 25$	100	1	11, 105	0	8, 53
					1000	8, 62

Table 1: Numerical Experiments

These numerical examples seem to indicate that, at least for constraints with some particular nonlinear structure (by the way, easy to evaluate), the algorithmic ideas presented in this work are reliable. In fact, in most cases the overall computer time of the method was less than that of the Augmented Lagrangian Algorithm. Needless to say, this is not a numerical study but only a set of examples that encourage the investment in more elaborate implementations.

9 Final Remarks

Feasible methods for minimization with nonlinear constraints, like projected gradient variations, GRG and SGRA play an important role in practical optimization and are still widely used in technological applications. The most frequent criticism against these methods is that they tend to keep prematurely close to the feasible region, so they should be slow if the feasible set is highly nonlinear. The theoretical results presented in this paper suggest that new implementations can be formulated for this type of methods, in which feasibility is controlled by means of a suitable merit function so that close adherence to the constraints is not necessary in the first iterations.

We think that the idea which underlies GRG, SGRA and Nonlinear Projected Gradient algorithms, that feasibility must be controlled independently of optimality in a separate algorithmic phase, is well supported by the very nature of the constrained optimization problem. Undoubtedly, this separate control involves additional constraint evaluations. In well behaved problems, a Sequential Quadratic Programming step (essentially a Newton step for the optimality conditions) will generally produce an improvement of feasibility, in the sense that the point on the linear tangent manifold $A(x^k)(y - x^k) + C(x^k) = 0$ which is closest to x^k should be more feasible than x^k . In those cases, testing feasibility improvement should be superfluous. However, if such property does not hold, there seems to be little reason to seek a point on the tangent manifold for which optimality is improved. Therefore, in many cases it is worthwhile paying for controlling the decrease of φ in an intermediate phase.

The main conclusion of this work is that, at least from a theoretical point of view, GRG, SGRA and other feasible methods admit variations that are globally convergent and are not subject to the common criticism of premature adherence to the constraints. A lot of work is still necessary involving practical implementation of these ideas. Related works along the lines of this paper are being subject of current intensive research. For example, Bielschowsky Ref. 25 developed a procedure for controlling feasibility and optimality that does not use of unifying merit functions at all. We hope that comprehensive comparative studies concerning these variations and the classical methods should be completed in the next few years.

References

1. ROSEN J. B., *The Gradient Projection Method for Nonlinear Programming, Part 1, Linear Constraints*, SIAM Journal on Applied Mathematics, Vol. 8, pp. 181-217, 1960.
2. ROSEN J. B., *The Gradient Projection Method for Nonlinear Programming, Part 2, Nonlinear Constraints*, SIAM Journal on Applied Mathematics, Vol. 9, pp. 514-532, 1961.
3. ABADIE J. AND CARPENTIER J., *Generalization of the Wolfe Reduced-Gradient Method to the Case of Nonlinear Constraints*, Optimization, Edited by R. Fletcher, Academic Press, New York, pp. 37-47, 1968.
4. LASDON L. S., *Reduced Gradient Methods*, in Nonlinear Optimization 1981. Edited by M. J. D. Powell, Academic Press, New York, pp. 235-242, 1982.

5. MIELE A., HUANG H. Y. AND HEIDEMAN J. C., *Sequential Gradient-Restoration Algorithm for the Minimization of Constrained Functions, Ordinary and Conjugate Gradient Version*, Journal of Optimization Theory and Applications, Vol. 4, pp. 213-246, 1969.
6. MIELE A., LEVY A. V. AND CRAGG E. E., *Modifications and Extensions of the Conjugate-Gradient Restoration Algorithm for Mathematical Programming Problems*, Journal of Optimization Theory and Applications, Vol. 7, pp. 450-472, 1971.
7. MIELE A., SIMS E. M. AND BASAPUR V. K., *Sequential Gradient-Restoration Algorithm for Mathematical Programming Problems with Inequality Constraints, Part 1, Theory*, Rice University, Aero-Astronautics Report No. 168, 1983.
8. ROM M. AND AVRIEL M., *Properties of the Sequential Gradient-Restoration Algorithm (SGRA), Part 1: Introduction and Comparison with Related Methods*, Journal of Optimization Theory and Applications, Vol. 62, pp. 77-98, 1989.
9. ROM M. AND AVRIEL M., *Properties of the Sequential Gradient-Restoration Algorithm (SGRA), Part 2: Convergence Analysis*, Journal of Optimization Theory and Applications, Vol. 62, pp. 99-126, 1989.
10. BYRD R. AND OMOJOKUN E., *Robust Trust-Region Methods for Nonlinearly Constrained Optimization*, First SIAM Conference on Optimization, Houston, Texas, 1987.
11. EL-ALEM, M. M., *A Global Convergence Theory for the Celis-Dennis-Tapia Trust Region Algorithm for Constrained Optimization*, SIAM Journal on Numerical Analysis, Vol. 28, pp. 266-290, 1991.
12. DENNIS J. E., EL-ALEM M. M. AND MACIEL M. C., *A Global Convergence Theory for General Trust-Region-Based Algorithms for Equality Constrained Optimization*, SIAM Journal on Optimization (to appear).
13. EL-ALEM M. M., *A robust trust region algorithm with a nonmonotonic penalty parameter scheme for constrained optimization*, SIAM Journal on Optimization, Vol. 5, pp. 348-378, 1995.
14. GOMES F. M., MACIEL M. C. AND MARTÍNEZ J. M., *Nonlinear Programming Algorithms Using Trust Regions and Augmented Lagrangians with Nonmonotone Penalty Parameters*, Technical Report, Institute of Mathematics, University of Campinas, Campinas, SP, Brazil, 1995.
15. OMOJOKUN E., *Trust-Region Strategies for Optimization with Nonlinear Equality and Inequality Constraints*, PhD Thesis, Department of Computer Science, University of Colorado, Boulder, Colorado, 1989.
16. POWELL M. J. D. AND YUAN Y., *A Trust-Region Algorithm for Equality Constrained Optimization*, Mathematical Programming, Vol. 49, pp. 190-211, 1991.
17. GILL P. E., MURRAY W. AND WRIGHT M. H., *Practical Optimization*, Academic Press, London, England, 1981.

18. MURTAGH R. B. AND SAUNDERS M. A., *Large-Scale Linearly Constrained Optimization*, Mathematical Programming, Vol. 14 , pp. 41-72, 1978.
19. MARTÍNEZ J. M. AND SANTOS S. A., *A trust region strategy for minimization on arbitrary domains*, Mathematical Programming, Vol. 68, pp. 267-302, 1995.
20. DEMBO R. S., EISENSTAT S. C. AND STEIHAUG T., *Inexact Newton Methods*, SIAM Journal on Numerical Analysis, Vol. 19, pp. 400-408, 1982.
21. FRIEDLANDER A., MARTÍNEZ J. M. AND SANTOS S. A., *A New Trust Region Algorithm for Bound Constrained Minimization*, Applied Mathematics and Optimization 30, pp. 235-266, 1994.
22. DINIZ-EHRHARDT M. A., GOMES-RUGGIERO M. A. AND SANTOS S. A., *Comparing the Numerical Performance of Two Trust-Region Algorithms for Large-Scale Bound-Constrained Minimization*. Fifth SIAM Conference on Optimization, Victoria, Canada, 1996.
23. CONN A. R., GOULD N. I. M. AND TOINT PH. L., *A Globally Convergent Augmented Lagrangian Algorithm for Optimization with General Constraints and Simple Bounds*, SIAM Journal on Numerical Analysis, Vol. 28, pp. 545 - 572, 1991.
24. CONWAY J. H. AND SLOANE N. J. C., *Sphere Packings, Lattices and Groups*, Springer-Verlag, New York, 1988.
25. BIELSCHOWSKY R. H., *Nonlinear Programming Algorithms with Dynamic Definition of Near-Feasibility: Theory and Implementations*, Doctoral Dissertation, Institute of Mathematics, University of Campinas, Campinas, SP, Brazil, 1996.