DOI: 10.1478/AAPP.98S2A8

AAPP | Atti della Accademia Peloritana dei Pericolanti
Classe di Scienze Fisiche, Matematiche e Naturali
ISSN 1825-1242

Vol. 98, No. S2, A8 (2020)

CONVERGENCE RATE FOR DIMINISHING STEPSIZE METHODS IN NONCONVEX CONSTRAINED OPTIMIZATION VIA GHOST PENALTIES

FRANCISCO FACCHINEI [a*], VYACHESLAV KUNGURTSEV [b], LORENZO LAMPARIELLO [c] AND GESUALDO SCUTARI [d]

ABSTRACT. This is a companion paper to "Ghost penalties in nonconvex constrained optimization: Diminishing stepsizes and iteration complexity" (to appear in Mathematics of Operations Research). We consider the ghost penalty scheme for nonconvex, constrained optimization introduced in that paper, coupled with a diminishing stepsize procedure. Under an extended Mangasarian-Fromovitz-type constraint qualification we give an expression for the maximum number of iterations needed to achieve a given solution accuracy according to a natural stationarity measure, thus establishing the first result of this kind for a diminishing stepsize method for nonconvex, constrained optimization problems.

1. Introduction

We consider the nonconvex constrained optimization problem

$$\begin{array}{ll} \underset{x\in\mathbb{R}^n}{\text{minimize}} & f(x) \\ \text{s.t.} & g(x) \leq 0 \end{array} \tag{P}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^m$ are $C^{1,1}$ (i.e., continuously differentiable with locally Lipschitz gradients) functions on $\mathbb{R}^n$. Equality constraints can be added, but we avoid this for the sake of simplicity.

The main aim of this paper is to study the convergence rate of a Diminishing Stepsize Method (DSM) proposed by Facchinei *et al.* (2020). The analysis of Facchinei *et al.* (2020) fills a gap in the literature in that, for the first time, it shows convergence of a DSM for a constrained optimization problem with nonconvex constraints. The results in this paper complete and complement the analysis of Facchinei *et al.* (2020) by providing a convergence rate study for that method. We refer the interested reader to (Facchinei *et al.* 2020) for a more detailed discussion of all of the background and formulation of the method and its motivation.

This paper is dedicated with affection to Antonino Maugeri on the occasion of his 75th birthday

Our framework is of a generalized Sequential Quadratic Programming (SQP)-type. At each iteration x^ν we generate a search direction $d(x^\nu)$ by solving a strongly convex optimization subproblem, which is described in Section **3**, constructed along the lines discussed in the seminal papers of Burke (1989) and Burke and Han (1989) and also taking into account the developments of Scutari *et al.* (2014) and Facchinei *et al.* (2017). A step γ^ν is then taken along $d(x^\nu)$ so that

$$x^{\nu+1} = x^\nu + \gamma^\nu d(x^\nu). \tag{1}$$

In (Facchinei *et al.* 2020), subsequential convergence to (generalized) stationary points of problem P is established for:

(a) the classical diminishing stepsize method, in which the stepsize policy is chosen according to the following rules:

$$\lim_{\nu\to\infty} \gamma^\nu \downarrow 0 \qquad \text{and} \qquad \sum_{\nu=0}^{\infty} \gamma^\nu = \infty;$$

(b) some more problem-oriented choices (including the case of a constant stepsize strategy) where the stepsize is held constant until some condition is met and then it is suitably reduced.

Iteration complexity for (b) is analyzed by Facchinei *et al.* (2020); in this paper we complete the analysis giving some corresponding complexity results for (a).

The paper is organized as follows. In the next section we recall some basic definitions. In Section **3** we define the search-direction subproblem and consider some related properties. In Section **4** we establish an upper bound on the number of iterations needed to satisfy a certain stopping criterion and also discuss the relationship of the stopping criterion to the KKT residual for problem (P).

2. KKT conditions and the eMFCQ

Consider Problem (P): our aim is to find a KKT point, i.e. a (feasible) point $x \in \mathbb{R}^n$ that, together with a vector of multipliers $\xi \in \mathbb{R}^m$, satisfies the KKT system

$$\begin{aligned} \nabla f(x) + \nabla g(x)\xi &= 0 \\ 0 \le \xi \perp g(x) &\le 0, \end{aligned} \tag{2}$$

where $\nabla f(x)$ is the gradient of f and $\nabla g(x)$ is the transposed Jacobian of g evaluated at x, and $\perp$ mean orthogonal, so that $\xi \perp g(x)$ stands for $\xi^T g(x) = 0$. We find it convenient to describe the set of multipliers associated to a KKT point x as

$$M_1(x) \triangleq \left\{ \xi \mid \xi \in N_{\mathbb{R}^m_-}(g(x)), 0 = \nabla f(x) + \nabla g(x)\xi \right\},$$

where $N_{\mathbb{R}^m_-}(g(x))$ denotes the normal cone to the non positive orthant at $g(x)$ (note that this implies that $M_1(x)$ is surely empty at infeasible points, so that $M_1(x) \neq \emptyset$ implies feasibility for x). Note that, if x is feasible, i.e. $g(x) \le 0$, condition $\xi \in N_{\mathbb{R}^m_-}(g(x))$ can be rewritten as

$$\xi_i \ge 0, \quad \xi_i g_i(x) = 0$$

for all i. Similarly we define the set of "abnormal" multipliers:

$$M_0(x) \triangleq \left\{ \xi \mid \xi \in N_{\mathbb{R}^m_-}(g(x) - \max_i\{g_i(x)_+\}e),\, 0 = \nabla g(x)\xi \right\}, \tag{3}$$

where $e \in \mathbb{R}^m$ is the vector with all components being one. Let $\hat{x}$ be a local minimum point of (P), then it is well-known that either $M_1(\hat{x}) \neq \emptyset$, (the point is a KKT solution) or $M_0(\hat{x}) \neq \{0\}$ (the point is a Fritz-John point), or both. However, there may be feasible points x, possibly not minimum points, for which $M_0(x) \neq \{0\}$, these are precisely the points where the Mangasarian-Fromovitz constraint qualification fails. On the contrary, it is classical to show that if $\hat{x}$ is not feasible, i.e. if $g_i(\hat{x}) > 0$ for at least an index $i \in \{1,\ldots,m\}$, then the stationarity condition for the problem $\underset{x}{\text{minimize}} \max_i\{g_i(\hat{x})_+\}$, that is

$$0 \in \partial \max_i\{g_i(\hat{x})_+\},$$

is equivalent to $M_0(\hat{x}) \neq \{0\}$.

The Constraint Qualification (CQ) we use in this paper is the Mangasarian-Fromovitz CQ, suitably extended to infeasible points.

Definition 2.1. *We say that the extended Mangasarian-Fromovitz Constraint Qualification (eMFCQ) holds at x if*

$$M_0(x) = \{0\}.$$

Indeed, note that if x is feasible, then the condition $M_0(x) = \{0\}$ is nothing else but the standard MFCQ. We make the following blanket assumption

Assumption CQ The eMFCQ holds at any point $x \in \mathbb{R}^n$.

3. Direction Finding Subproblem and its Properties

At each iteration of our algorithm we move from the current iteration x^ν along a direction $d(x^\nu)$ with a stepsize γ^ν, see (1). While the stepsize is chosen according to classic diminishing stepsize rules, the direction $d(x^\nu)$ is the solution of a suitable strongly convex subproblem that we describe next.

Given a point x (which will actually be x^ν in the algorithm) $d(x)$ is the unique solution of the optimization problem:

$$\begin{array}{ll} \underset{d}{\text{minimize}} & \widetilde{f}(d;x) \\ \text{s.t.} & \widetilde{g}(d;x) \leq \kappa(x)e \\ & \|d\|_\infty \leq \beta, \end{array} \tag{P_x}$$

where $e \in \mathbb{R}^m$ is the vector with all components being one, $\kappa(x)$ a nonnegative quantity to be defined shortly, and β is a user-chosen positive constant. Moreover, $\widetilde{f}$ is a strongly convex surrogate of the original objective function f while $\widetilde{g}$ is a convex surrogate of the original constraint functions g (see Assumption A below for the conditions we impose on these surrogates).

The term $\kappa(x)e$ in the subproblem constraints serves to enlarge the feasible set of the subproblem in order to ensure it is always nonempty. The additional constraint $\|d\|_\infty \leq \beta$

allows one to avoid issues with search directions becoming too large. We denote by $\widetilde{\mathscr{X}}(x)$ the convex feasible set of subproblem (P_x), i.e.

$$\widetilde{\mathscr{X}}(x) \triangleq \{d \in \mathbb{R}^n : \widetilde{g}(d;x) \leq \kappa(x)\, e,\ \|d\|_\infty \leq \beta\},$$

and, when convenient, we rewrite the constraint $\|d\|_\infty \leq \beta$ as $d \in \beta \mathbb{B}^n_\infty$, where $\mathbb{B}^n_\infty$ is the closed unit ball in $\mathbb{R}^n$ associated with the infinity-norm.

The direction finding subproblem (P_x) is a direct generalization of the subproblems considered by Burke (1989), to which it reduces when the classical quadratic/linear approximations are used for $\widetilde{f}$:

$$\widetilde{f}(d;x) \triangleq \nabla f(x)^T d + \frac{1}{2}\|d\|^2; \qquad \widetilde{g}(d;x) \triangleq g(x) + \nabla g(x)^T d. \tag{4}$$

Note that if these approximations are employed and we set $\kappa(x) = 0$ and $\beta = +\infty$, subproblem (P_x) boils down to the classical SQP-type subproblem. Here we adopt the approach of Burke (1989) by taking $\kappa(x)$ not necessarily zero and $\beta < +\infty$ in order to guarantee the existence and continuity of the solution mapping $d(x)$. In addition we introduce the use of general approximations $\widetilde{f}$ and $\widetilde{g}$, this may be very convenient in practice by allowing flexibility in tailoring the direction finding subproblem to the problem at hand and to exploit any available specific structure in (P), see Scutari *et al.* (2014) and Facchinei *et al.* (2017). Of course, an underlying assumption of our approach is that subproblem (P_x) can be solved efficiently. We do not insist on this point because it is very dependent on the choice of $\widetilde{f}$ and $\widetilde{g}$ which in turn is dictated by the original problem (P). But we observe that the use of models that go beyond the standard quadratic/linear one in constrained optimization is emerging consistently in the literature, motivated, on the one hand, by the possibility to solve efficiently more complex subproblems than the classical quadratic ones and, on the other hand, by the desire of faster convergence rates, see for example the discussion in Section 3 of (Martínez 2017).

For this approach to be legitimate and lead to useful convergence results, we obviously need to make assumptions on the surrogate functions $\widetilde{f}$ and $\widetilde{g}$.

Assumption Approx

Let O_d be an open neighborhood of $\beta\mathbb{B}^n_\infty$ and $\widetilde{f} : O_d \times \mathbb{R}^n \to \mathbb{R}$ and $\widetilde{g}_i : O_d \times \mathbb{R}^n \to \mathbb{R}$ for every $i = 1,\ldots,m$ be continuously differentiable on O_d with respect to the first argument and such that

- ***A1):*** *$\widetilde{f}(\bullet;x)$ is a strongly convex function on O_d for every $x \in \mathbb{R}^n$ with modulus of strong convexity $c > 0$ independent of x;*
- ***A2):*** *$\widetilde{f}(\bullet;\bullet)$ is continuous on $O_d \times \mathbb{R}^n$;*
- ***A3):*** *$\nabla_1 \widetilde{f}(\bullet;\bullet)$ is continuous $O_d \times \mathbb{R}^n$;*
- ***A4):*** *$\nabla_1 \widetilde{f}(0;x) = \nabla f(x)$ for every $x \in \mathbb{R}^n$;*
- ***A5):*** *$\widetilde{g}_i(\bullet;x)$ is a convex function on O_d for every $x \in \mathbb{R}^n$;*
- ***A6):*** *$\widetilde{g}_i(\bullet;\bullet)$ is locally Lipschitz continuous on $O_d \times \mathbb{R}^n$;*
- ***A7):*** *$\widetilde{g}_i(0;x) = g_i(x)$ for every $x \in \mathbb{R}^n$;*
- ***A8):*** *$\nabla_1 \widetilde{g}_i(\bullet;\bullet)$ is locally Lipschitz continuous on $O_d \times \mathbb{R}^n$;*
- ***A9):*** *$\nabla_1 \widetilde{g}_i(0;x) = \nabla g_i(x)$, for every $x \in \mathbb{R}^n$;*

where $\nabla_1 \widetilde{f}(u;x)$ *and* $\nabla_1 \widetilde{g}_i(u;x)$ *denote the partial gradient of* $\widetilde{f}(\bullet;x)$ *and* $\widetilde{g}_i(\bullet;x)$ *evaluated at* u.

These conditions are easily satisfied in practice and have been used in many recent papers; we refer the reader to (Scutari *et al.* 2014; Facchinei *et al.* 2017) as a good source of examples. Here we only note that the classical quadratic/linear approximations (4) satisfy Assumption A, provided f and g are $C^{1,1}$.

To complete the description of subproblem (P_x), we must give the definition of $\kappa(x)$. Following Burke (1989), we set

$$\kappa(x) \triangleq (1-\lambda)\max_i\{g_i(x)_+\} + \lambda \min_d \left\{ \max_i\{\widetilde{g}_i(d;x)_+\} \mid \|d\|_\infty \le \rho \right\}, \tag{5}$$

with $\lambda \in (0,1)$ and $\rho \in (0,\beta)$. Note that the computation of $\kappa(x)$ requires the computation of the optimal value of the convex (see A5) optimization problem

$$\min_d \left\{ \max_i\{\widetilde{g}_i(d;x)_+\} \mid \|d\|_\infty \le \rho \right\}$$

that always has an optimal solution because the feasible set is nonempty and compact. This problem can easily be reformulated as a smooth optimization problem. In addition, if $\widetilde{g}$ is linear, as in the classical choice given in (4), this problem reduces to a linear program and can be efficiently solved. Note also that if x is feasible for (P), i.e. $g(x) \le 0$, we have $\kappa(x) = 0$, so that our subproblem (P_x) is very similar to standard SQP subproblems. The term $\kappa(x)$ plays a key role when the SQP-type subproblems have an empty feasible set, a well-known issue with SQP schemes. In fact, observing that $\kappa(x)$ is always nonnegative being the sum of two nonnegative quantities, it restores feasibility by enlarging (with respect to the SQP choice $\widetilde{g}(d;x) \le 0$) the range of admissible values. Hence, the feasible set of problem (P_x), for every x, is nonempty: choosing $\hat{d}$ at which the minimum in the expression of $\kappa(x)$ is attained, we have

$$\widetilde{g}(\hat{d};x) \le \min_d \left\{ \max_i\{\widetilde{g}_i(d;x)_+\} \mid \|d\|_\infty \le \rho \right\} e = \max_i\{g_i(\hat{d};x)_+\} e,$$

and, in turn,

$$\begin{aligned} \widetilde{g}(\hat{d};x) &= (1-\lambda)\widetilde{g}(\hat{d};x) + \lambda \widetilde{g}(\hat{d};x) \\ &\le [(1-\lambda)\max_i\{\widetilde{g}_i(0;x)_+\} + \lambda \min_d \{\max_i\{\widetilde{g}_i(d;x)_+\} \mid \|d\|_\infty \le \rho\}]\, e = \kappa(x) e. \end{aligned}$$

Finally, we discuss briefly the KKT conditions for problem (P_x). Observe preliminarily that the constraint $\|d\|_\infty \le \beta$ corresponds to $2n$ bounds of the type $-\beta \le d_i \le \beta$. However, in what follows, we are interested only in the multipliers corresponding to the constraints $\widetilde{g}(d;x) \le 0$ and therefore we find it expedient to write the KKT conditions as

$$\begin{aligned} 0 &\in \nabla_1 \widetilde{f}(d(x);x) + \nabla_1 \widetilde{g}(d(x);x)\xi + N_{\beta \mathbb{B}^n_\infty}(d(x)) \\ 0 &\le \xi \perp [\widetilde{g}(d(x);x) - \kappa(x)] \le 0 \end{aligned} \tag{6}$$

with the multipliers ξ satisfying the conditions $\xi \ge 0$ and $\xi^T \widetilde{g}(d(x);x) = 0$, and where $N_{\beta \mathbb{B}^n_\infty}(d(x))$ is the normal cone to $\beta \mathbb{B}^n_\infty$ at $d(x)$.

The following proposition, collecting several useful facts about the objects considered so far, shows, among other things, that the KKT conditions are always satisfied at a solution of problem (P_x).

Proposition 3.1. *Under Assumptions CQ and Approx, the following properties hold:*

(a): $\kappa(\bullet)$ *is locally Lipschitz continuous on* $\mathbb{R}^n$*;*
(b): *for every x, the feasible set* $\widetilde{\mathscr{X}}(x)$ *of* (P_x) *is non empty and* (P_x) *has a unique solution* $d(x)$*;*
(c): *the MFCQ holds at every point of* $\widetilde{\mathscr{X}}(x)$*, for every x;*
(d): *the function* $d(\bullet)$ *is continuous on* $\mathbb{R}^n$ *and* $d(x) = 0$ *if and only if x is a KKT solution for* (P)*;*
(e): *the unique solution* $d(x)$ *of* (P_x) *is a KKT solution of problem* (P_x) *and the set of KKT multipliers is locally bounded at any* $x \in \mathbb{R}^n$.

Proof. Some of the statements are readily seen to hold and, in any case, with the exception of the fact that $d(x) = 0$ if and only if x is a KKT point of problem (P), all the properties can immediately be derived from Propositions 2, 4, and 6 in (Facchinei *et al.* 2020), taking into account that here the eMFCQ is assumed to hold at every point. The statement about $d(x) = 0$ can be proved by simply comparing (2) and (6), taking into account A4 and A9 and the fact that $N_{\beta\mathbb{B}^n_\infty}(0) = \{0\}$: it is enough to observe that

$$\begin{aligned} \max_i\{g_i(x)_+\} - \kappa(x) &\leq \max_i\{g_i(x)_+\} - \max_i\{\widetilde{g}_i(d(x);x)_+\} \\ &\leq \max_i\{g_i(x)_+\} - \max_i\{(g_i(x) + \nabla g_i(x)^T d(x))_+\} \qquad (7) \\ &\leq \|\nabla g(x)^T\|_\infty \|d(x)\|, \end{aligned}$$

which is due to

$$\kappa(x) \geq \widetilde{g}_i(d(x);x) \geq \widetilde{g}_i(0;x) + \nabla\widetilde{g}_i(0;x)^T d(x) = g_i(x) + \nabla g_i(x)^T d(x), \qquad (8)$$

for every i. In turn, with $d(x) = 0$, $\max_i\{g_i(x)_+\} = \kappa(x)$ and the claim follows from Lemma 2 of (Facchinei *et al.* 2020). □

4. Algorithm and Convergence Rate

Our scheme is based on the successive solution of subproblems (P_x) and the iterative process $x^{\nu+1} = x^\nu + \gamma^\nu d^\nu$, where the stepsizes γ^ν are non-increasing and such that

$$\lim_{\nu\to\infty} \gamma^\nu \downarrow 0 \qquad \text{and} \qquad \sum_{\nu=0}^{\infty} \gamma^\nu = \infty. \qquad (9)$$

Algorithm 1: DSM Algorithm for (P)

Data: $\gamma^0 \in (0,1]$, $\delta \geq 0$, x^0, $\nu \longleftarrow 0$;
repeat
(S.1) compute $\kappa(x^\nu)$ and the solution $d(x^\nu)$ of problem (P_{x^ν});
(S.2) **if** $\|d^\nu\| \leq \delta$ **then**
stop and **return** x^ν;
end
(S.3) set $x^{\nu+1} = x^\nu + \gamma^\nu d(x^\nu)$, $\nu \longleftarrow \nu + 1$;
end

The algorithm is always well defined if Assumption A, which guarantees existence and uniqueness of $d(x^\nu)$, holds. We also note that the stopping criterion at step S.2 is sound and sensible. In fact, in light of Proposition **3.1**, $\|d(x)\|$ is a valid stationarity measure, being a continuous function which is zero if and only if x is a KKT point of problem (P). In subsection **4.1** we complete the study of this stationarity measure showing its quantitative relationship with the approximate KKT conditions residual.

Remark 4.1. *Note that we included the possibility of taking $\delta = 0$ in Algorithm 1. This choice will generally lead the algorithm to produce an infinite sequence $\{x^\nu\}$: in fact, the algorithm will stop only in the exceptional case of $d(x^\nu) = 0$ for some ν, i.e. the algorithm will stop only if an exact KKT point is reached. This is the case considered by Facchinei* et al. *(2020): there, it is shown that, if $\delta = 0$, any limit point of the sequence $\{x^\nu\}$ is a KKT point for problem* (P)*. Of course, to have finite termination, we must assume $\delta > 0$; however, the case $\delta = 0$ will be considered in order to state a boundedness condition used in Theorem **4.1**, see Assumption Compact below.* □

The distinctive aspect of Algorithm 1 is its simplicity, a feature shared with all DSMs. The main (and essentially only) computational burden is given by the computation of $\kappa(x^\nu)$ and of the solution for the strongly convex subproblem (P_{x^ν}). Before giving the result on the convergence rate properties of Algorithm 1, we state the last assumption needed.

Assumption Compact The sequence $\{x^\nu\}$ generated by Algorithm 1 with $\delta = 0$ is bounded. Hence, the whole sequence $\{x^\nu\}$ is contained in some compact set $S \subseteq \mathbb{R}^n$.

We will comment further on this assumption in Remark **4.3**, after Theorem **4.1**.

The proof of Theorem **4.1** below is based on the nonsmooth penalty function

$$W(x;\varepsilon) \triangleq f(x) + \frac{1}{\varepsilon} \max_i \{g_i(x)_+\} \tag{10}$$

where, as usual, $a_+ = \max\{0, a\}$. Essentially, and rather classically, we will show that for a certain sufficiently small value $\bar{\varepsilon}$ of the penalty parameter, $d(x^\nu)$ is a direction of sufficient decrease for $W(x;\bar{\varepsilon})$ at x^ν. Based on this fact, we can derive the desired results. Note however that the penalty function itself is never used in the algorithm and, in particular, the user need not know the value of $\bar{\varepsilon}$, in contrast to standard penalty approaches in constrained optimization. For this reason, we use the term *ghost penalty*. We underline also the fact that we can see the penalty function as a Lyapunov function because we assume that the eMFCQ

holds; in (Facchinei *et al.* 2020), where this assumption is not made, the role played by the ghost penalty is more complicated.

Theorem 4.1. *Let $\{x^\nu\}$ be the sequence generated by Algorithm 1 with $\delta > 0$ under Assumptions CQ , Approx, and Compact. Then, Algorithm 1 finds a point for which $\|d(x^\nu)\| \le \delta$ and stops in at most N iterations, where*

(a): *if γ^0 is sufficiently small, N is the first iteration index for which*

$$\sum_{\nu=0}^{N-1} \gamma^\nu \ge \frac{[W^0 - W^m]}{\omega\,\delta^2}, \tag{11}$$

where, $\omega = c/2$ (see Assumption Approx for the definition of c), $\bar{\varepsilon}$ is a value of the penalty parameter to be specified in the proof, $W^0 \triangleq W(x^0;\bar{\varepsilon})$, and W^m is the minimum value attained by the continuous function $W(x;\bar{\varepsilon})$ on the compact set S;

(b): *if no assumptions are made on γ^0, N is the first iteration index for which*

$$\sum_{\nu=\bar{\nu}}^{\bar{\nu}+N-1} \gamma^\nu \ge \frac{[W^M - W^m]}{\omega\,\delta^2}, \tag{12}$$

for some suitable iteration $\bar{\nu}$ and where W^M is the maximum value attained by the continuous function $W(x;\bar{\varepsilon})$ on the compact set S.

Proof. Essentially, we aim at showing that a certain decrease in the penalty function is achieved at each iteration. The proof proceeds by a sequence of lemmas.

Lemma 4.1. *There is a positive constant Ξ such that for any $x \in S$ the multipliers ξ associated to the constraints $\widetilde{g}(d;x) \le \kappa(x)e$ at the solution $d(x)$ in subproblem (P_x) are bounded by Ξ, i.e. $\|\xi\|_\infty \le \Xi$ for any optimal multiplier ξ.*

Proof. The lemma follows from Proposition **3.1** (e) and the compactness of S. □

The following lemma gives a bound on the directional derivative of the objective function f in the direction $d(x^\nu)$ at every iteration ν.

Lemma 4.2. *At each iteration ν the following bound holds for the directional derivative of the objective function at x^ν in the direction $d(x^\nu)$:*

$$\nabla f(x^\nu)^T d(x^\nu) \le -c\|d(x^\nu)\|^2 + m\Xi\,[\max_i\{g_i(x^\nu)_+\} - \kappa(x^\nu)]. \tag{13}$$

Proof. Recall first that, given x^ν, $d(x^\nu)$ satisfies the KKT conditions (6) by Proposition **3.1**:

$$0 \in \nabla_1 \widetilde{f}(d(x^\nu);x^\nu) + \nabla_1 \widetilde{g}(d(x^\nu);x^\nu)\xi^\nu + N_{\beta\mathbb{B}^n_\infty}(d(x^\nu)), \tag{14}$$

for some multipliers $\xi^\nu \in N_{\mathbb{R}^m_-}(\widetilde{g}(d(x^\nu);x^\nu) - \kappa(x^\nu)e)$. Observe now that, thanks to A1 and A4,

$$\begin{aligned} \nabla_1 \widetilde{f}(d(x^\nu);x^\nu)^T d(x^\nu) &= [\nabla_1 \widetilde{f}(d(x^\nu);x^\nu) - \nabla_1 \widetilde{f}(0;x^\nu) + \nabla_1 \widetilde{f}(0;x^\nu)]^T d(x^\nu) \\ &\ge c\|d(x^\nu)\|^2 + \nabla f(x^\nu)^T d(x^\nu). \end{aligned} \tag{15}$$

Moreover, in view of A5,

$$-\nabla_1 \widetilde{g}_i(d(x^\nu);x^\nu)^T d(x^\nu) \le \widetilde{g}_i(0;x^\nu) - \widetilde{g}_i(d(x^\nu);x^\nu)$$

and, by A7, since ξ^ν is nonnegative, in turn,

$$-\xi_i^\nu \nabla_1 \widetilde{g}_i(d(x^\nu);x^\nu)^T d(x^\nu) \le \xi_i^\nu [\widetilde{g}_i(0;x^\nu) - \widetilde{g}_i(d(x^\nu);x^\nu)] = \xi_i^\nu [g_i(x^\nu) - \kappa(x^\nu)], \quad (16)$$

where the equality follows observing that ξ^ν belongs to $N_{\mathbb{R}^m_-}(\widetilde{g}(d(x^\nu);x^\nu) - \kappa(x^\nu)e)$.

Therefore, by (14), (15) and (16), we have, for some $\zeta^\nu \in N_{\beta\mathbb{B}^n_\infty}(d(x^\nu))$,

$$\begin{aligned} c\|d(x^\nu)\|^2 + \nabla f(x^\nu)^T d(x^\nu) &\le \nabla_1 \widetilde{f}(d(x^\nu);x^\nu)^T d(x^\nu) \\ &= -\xi^{\nu T} \nabla_1 \widetilde{g}(d(x^\nu);x^\nu)^T d(x^\nu) - \zeta^{\nu T} d(x^\nu) \\ &\le \xi^{\nu T}[g(x^\nu) - \kappa(x^\nu)e] \le \xi^{\nu T}[\max_i\{g_i(x^\nu)_+\} - \kappa(x^\nu)]e, \end{aligned}$$

and, thus, (13) easily follows taking into account Lemma **4.1** and the fact that, by definition, $\kappa(x^\nu) \le \max_i\{g_i(x^\nu)_+\}$. □

Lemma 4.3. *For any positive ε, we have*

$$\begin{aligned} & W(x^{\nu+1};\varepsilon) - W(x^\nu;\varepsilon) \\ & \quad \le \gamma^\nu \nabla f(x^\nu)^T d(x^\nu) - \frac{\gamma^\nu}{\varepsilon}\Big[\max_i\{g_i(x^\nu)_+\} - \kappa(x^\nu)\Big] \\ & \quad + \frac{(\gamma^\nu)^2}{2}(L_{\nabla f} + \frac{\max_i\{L_{\nabla g_i}\}}{\varepsilon})\|d(x^\nu)\|^2, \end{aligned} \quad (17)$$

$$\begin{aligned} & \nabla f(x^\nu)^T d(x^\nu) - \frac{1}{\varepsilon}\Big[\max_i\{g_i(x^\nu)_+\} - \kappa(x^\nu)\Big] \\ & \quad \le -c\|d(x^\nu)\|^2 + (m\Xi - \frac{1}{\varepsilon})\Big[\max_i\{g_i(x^\nu)_+\} - \kappa(x^\nu)\Big]. \end{aligned} \quad (18)$$

Proof. Inequality (17) is due to the following chain on relations:

$$\begin{aligned} W(x^{\nu+1};\varepsilon) \;-\; & W(x^\nu;\varepsilon) \\ = \; & f(x^\nu + \gamma^\nu d(x^\nu)) - f(x^\nu) + \tfrac{1}{\varepsilon}\Big[\max_i\{g_i(x^\nu + \gamma^\nu d(x^\nu))_+\} - \max_i\{g_i(x^\nu)_+\}\Big] \\ \overset{(a)}{\le} \; & \gamma^\nu \nabla f(x^\nu)^T d(x^\nu) + \tfrac{(\gamma^\nu)^2 L_{\nabla f}}{2}\|d(x^\nu)\|^2 \\ & + \tfrac{1}{\varepsilon}\Big[\max_i\{(g_i(x^\nu) + \gamma^\nu \nabla g_i(x^\nu)^T d(x^\nu))_+\} \\ & - \max_i\{g_i(x^\nu)_+\} + \frac{(\gamma^\nu)^2 \max_i\{L_{\nabla g_i}\}}{2}\|d(x^\nu)\|^2\Big] \\ \overset{(b)}{\le} \; & \gamma^\nu \nabla f(x^\nu)^T d(x^\nu) + \tfrac{1}{\varepsilon}\Big[\max_i\{(1-\gamma^\nu)g_i(x^\nu)_+ + \gamma^\nu \kappa(x^\nu)\} - \max_i\{g_i(x^\nu)_+\}\Big] \\ & + \tfrac{(\gamma^\nu)^2}{2}(L_{\nabla f} + \tfrac{\max_i\{L_{\nabla g_i}\}}{\varepsilon})\|d(x^\nu)\|^2 \\ \le \; & \gamma^\nu \nabla f(x^\nu)^T d(x^\nu) - \frac{\gamma^\nu}{\varepsilon}\Big[\max_i\{g_i(x^\nu)_+\} - \kappa(x^\nu)\Big] \\ & + \tfrac{(\gamma^\nu)^2}{2}(L_{\nabla f} + \tfrac{\max_i\{L_{\nabla g_i}\}}{\varepsilon})\|d(x^\nu)\|^2, \end{aligned}$$

where (a) follows applying the descent lemma to f and g_i for every $i = 1,\dots,m$, with $L_{\nabla f}$ and $L_{\nabla g_i}$ being the Lipschitz moduli of ∇f and ∇g_i; (b) holds for $\gamma^\nu \le 1$ since, in view of (8), $\nabla g_i(x^\nu)^T d(x^\nu) \le \kappa(x^\nu) - g_i(x^\nu)$. Inequality (18) is obtained by Lemma **4.2** by subtracting $\frac{1}{\bar\varepsilon}\Big[\max_i\{g_i(x^\nu)_+\} - \kappa(x^\nu)\Big]$ from both sides of (13). □

Proof of Theorem 4.1 (a). Set $\bar\varepsilon \triangleq \frac{1}{m\Xi}$. We immediately get from (18)

$$\nabla f(x^\nu)^T d(x^\nu) - \frac{1}{\bar\varepsilon}\Big[\max_i\{g_i(x^\nu)_+\} - \kappa(x^\nu)\Big] \le -c\|d(x^\nu)\|^2.$$

Plugging this expression in (17) we obtain

$$\begin{aligned} W(x^{\nu+1};\bar\varepsilon) - W(x^\nu;\bar\varepsilon) &\le -\gamma^\nu c\|d(x^\nu)\|^2 + \tfrac{(\gamma^\nu)^2}{2}(L_{\nabla f} + \tfrac{\max_i\{L_{\nabla g_i}\}}{\bar\varepsilon})\|d(x^\nu)\|^2 \\ &= -\gamma^\nu\Big[c - \tfrac{\gamma^\nu}{2}(L_{\nabla f} + \tfrac{\max_i\{L_{\nabla g_i}\}}{\bar\varepsilon})\Big]\|d(x^\nu)\|^2. \end{aligned} \tag{19}$$

If

$$\gamma^\nu \le \frac{c}{L_{\nabla f} + m\Xi\max_i\{L_{\nabla g_i}\}} \triangleq \bar\gamma, \tag{20}$$

the quantity in square brackets in the formula (19) is greater than or equal to $1/2$ and we can write

$$W(x^{\nu+1};\bar\varepsilon) - W(x^\nu;\bar\varepsilon) \le -\omega\gamma^\nu\|d(x^\nu)\|^2, \qquad \omega \triangleq \frac{c}{2},$$

and hence

$$\gamma^\nu\|d(x^\nu)\|^2 \le \frac{[W(x^\nu;\bar\varepsilon) - W(x^{\nu+1};\bar\varepsilon)]}{\omega}.$$

Assume now that $\gamma^0 \le \bar\gamma$ and recall that the sequence $\{\gamma^\nu\}$ is non-increasing. Supposing that $\|d(x^\nu)\| > \delta$ for every $\nu \in \{0,\dots,N-1\}$ and taking the sum up to $N-1$, we have

$$\delta^2\sum_{\nu=0}^{N-1}\gamma^\nu < \sum_{\nu=0}^{N-1}\gamma^\nu\|d(x^\nu)\|^2 \le \frac{W(x^0;\bar\varepsilon) - W(x^N;\bar\varepsilon)}{\omega} \le \frac{W^0 - W^m}{\omega}.$$

We therefore see that the maximum number of iterations required in order to make $d(x^\nu)$ smaller than δ is N, with N such that

$$\sum_{\nu=0}^{N-1}\gamma^\nu \ge \frac{[W^0 - W^m]}{\omega\,\delta^2}$$

Proof of Theorem 4.1 (b). When considering a diminishing stepsize procedure, a finite $\bar\nu$ exists such that, for every $\nu \ge \bar\nu$, (20) holds; clearly, the number of iterations $\bar\nu$ is problem-dependent and relies on initial algorithmic choices such as the updating rule for the diminishing stepsize.

Supposing $\|d(x^\nu)\| > \delta$ for every $\nu \in \{0,\dots,\bar\nu+N\}$ and summing $\gamma^\nu\|d(x^\nu)\|^2 \le [W(x^\nu;\bar\varepsilon) - W(x^{\nu+1};\bar\varepsilon)]/\omega$ from $\bar\nu$ up to $\bar\nu+N-1$, we have

$$\delta^2\sum_{\nu=\bar\nu}^{\bar\nu+N}\gamma^\nu < \sum_{\nu=\bar\nu}^{\bar\nu+N}\gamma^\nu\|d(x^\nu)\|^2 \le \frac{W(x^{\bar\nu};\bar\varepsilon) - W(x^{\bar\nu+N};\bar\varepsilon)}{\omega} \le \frac{W^M - W^m}{\omega},$$

Reasoning as in (a), we get (12). □

Remark 4.2. *In order for case (a) to hold, γ^0 must be "sufficiently small", as stated in the theorem; the proof shows, more precisely, that actually γ^0 must be smaller than $\bar{\gamma}$, with $\bar{\gamma}$ defined in* (20). □

Remark 4.3. *Meaningful results are obtained in Theorem **4.1** if the sequence $\{x^\nu\}$ generated by the algorithm with $\delta = 0$ is bounded. Although this is practically rather sensible, the question naturally arises if one can give* a priori *conditions guaranteeing the boundedness of the iterations. It is possible to give a satisfactory answer to these questions only for the price of a much more convoluted analysis; we eschew this for the sake of simplicity of presentation. The topic is however treated in great detail by Facchinei* et al. *(2020) to which we refer the reader for further information. All results on this aspect exposed by Facchinei* et al. *(2020) can be adapted to our setting. Here we only mention that there are two main avenues. One is that of giving explicit* a priori *coercivity-type conditions on problem* (P) *that guarantee boudedness of the iterates. The other approach is more direct and simply amounts to assuming that the original problem* (P) *includes a constraint of the type $x \in K$, where $K \subseteq \mathbb{R}^n$ is a (typically simple) compact, convex set. For example K could define upper and lower bounds an all variables. It is easy, although formally complicated, to show that all the results in this section still hold if we redefine subproblem* (P_x)*, by appending the constraint K to it, as*

$$\underset{d}{\text{minimize}}\ \widetilde{f}(d;x^\nu) \quad s.t. \quad \widetilde{g}(d;x^\nu) \leq \kappa(x^\nu)\, e,\ \ \|d\|_\infty \leq \beta,\ \ x^\nu + d \in K, \tag{21}$$

and by similarly modifying the feasible set of the minimization problem in the definition of κ. If one, then, requires the algorithm to start from an initial point $x^0 \in K$, this simple strategy obviously guarantees the boundedness of the iterations, which all belong to the compact set K, since $\gamma^\nu \in (0,1]$ and K is convex. □

The practical meaning of Theorem **4.1** depends on the actual rule according to which we choose γ^ν. To make a concrete example, suppose that one relies on the classical generic term of the harmonic series and sets $\gamma^\nu = \frac{\gamma^0}{\nu+1}$ and that we are considering case (a) in Theorem **4.1**. Observing that $\sum_{\nu=0}^{N-1} \gamma^\nu = \sum_{\nu=0}^{N-1} \frac{\gamma^0}{\nu+1} = \sum_{\nu=1}^{N} \frac{\gamma^0}{\nu}$ and $\gamma^0 \ln(N+1) < \sum_{\nu=1}^{N} \frac{\gamma^0}{\nu} < \gamma^0[\ln N + 1]$,

$$\left\lfloor \exp^{\frac{[W^0 - W^m]}{\gamma^0 \omega \delta^2} - 1} \right\rfloor < N < \left\lceil \exp^{\frac{[W^0 - W^m]}{\gamma^0 \omega \delta^2}} \right\rceil. \tag{22}$$

Case (b) is a variant of case (a). It is clear that if we use a diminishing stepsize rule, then, sooner or later, an iteration $\bar{\nu}$ occurs for which $\gamma^{\bar{\nu}}$ satisfies the condition in point (a) and we can apply the results in (a) starting from that iteration. This case is worth considering because in practice it is the most realistic one, since in general it is difficult to establish the "sufficiently smal" value $\bar{\gamma}$ that should be used in case (a). It is easy to see that when the generic term of the harmonic series $\frac{\gamma^0}{\nu+1}$ is employed, we can take

$$\bar{\nu} = \left\lceil \gamma^0 \frac{L_{\nabla f} + m\Xi \max_i\{L_{\nabla g_i}\}}{2c} \right\rceil - 1$$

and, since $\sum_{\nu=\bar{\nu}}^{\bar{\nu}+N-1}\gamma^{\nu}=\sum_{\nu=\bar{\nu}}^{\bar{\nu}+N-1}\frac{\gamma^0}{\nu+1}=\sum_{\nu=1}^{N}\frac{\gamma^0}{\bar{\nu}+\nu}$, we have

$$\gamma^0\ln\left(\frac{\bar{\nu}+N}{\bar{\nu}+1}+\frac{1}{\bar{\nu}+1}\right)<\sum_{\nu=1}^{N}\frac{\gamma^0}{\bar{\nu}+\nu}<\gamma^0\left[\ln\left(\frac{\bar{\nu}+N}{\bar{\nu}+1}\right)+\frac{1}{\bar{\nu}+1}\right],$$

and, in turn,

$$\left\lfloor(\bar{\nu}+1)\exp^{\frac{[W^0-W^m]}{\gamma^0\omega\delta^2}-\frac{1}{\bar{\nu}+1}}\right\rfloor-\bar{\nu}<N<\left\lceil(\bar{\nu}+1)\exp^{\frac{[W^0-W^m]}{\gamma^0\omega\delta^2}}\right\rceil-\bar{\nu}. \tag{23}$$

The discussion above hints at the fact that the use of a diminishing stepszize is, in a sense, a necessary evil, that one has to accept because the critical value $\bar{\gamma}$ is not known in advance. It is clear that, if $\gamma^0\leq\bar{\gamma}$, it is better to have a sequence $\{\gamma^{\nu}\}$ that decreases "slowly". From this point of view, the best possible choice would be the limiting case of a *constant stepsize* $\gamma^{\nu}=\gamma^0\leq\bar{\gamma}$ for all ν. Of course, this choice does not give rise to a DSM since the conditions (9) are not satisfied. Nevertheless, it is easy to check that this choice would still give termination in a finite number of steps. Indeed, since (20) is still satisfied, we can follow the proof of Theorem **4.1** from there and we still have $\|d(x^{\nu})\|^2\leq[W(x^{\nu};\bar{\varepsilon})-W(x^{\nu+1};\bar{\varepsilon})]/\gamma^0\omega$ for every ν. Supposing that $\|d(x^{\nu})\|>\delta$ for every $\nu\in\{0,\ldots,N-1\}$ and taking the sum of iterations up to N, we have

$$N\gamma^0\delta^2<\sum_{\nu=0}^{N-1}\gamma^0\|d(x^{\nu})\|^2\leq\frac{W(x^0;\bar{\varepsilon})-W(x^N;\bar{\varepsilon})}{\omega}\leq\frac{W^0-W^m}{\omega},$$

where $W^0\triangleq W(x^0;\bar{\varepsilon})$ and W^m is the minimum value attained by the continuous function $W(x;\bar{\varepsilon})$ on the compact set S. Therefore, our procedure finds a point satisfying $\|d(x^{\nu})\|\leq\delta$ in N iterations, with

$$N=\left\lceil\frac{[W^0-W^m]}{\gamma^0\,\omega\,\delta^2}\right\rceil. \tag{24}$$

This bound on the number of iterations is clearly much better in general than bounds like (22) or (23). Informally speaking, the bound (24) should be regarded as an ideal, limiting case for DSMs, a case obtained by taking the slowest possibly decreasing sequence past the necessary value $\bar{\gamma}$.

4.1. Approximate KKT Conditions. We already discussed how $\|d(x)\|$ is a valid stationarity measure that can be resorted to in order to devise a sensible stopping criterion. However, it is of interest to give quantitative relations between the stopping criterion $\|d(x)\|\leq\delta$ and the fulfillment of $\delta-$*approximate KKT conditions* (see below for a precise definition). It turns out that the condition $\|d(x)\|\leq\delta$ entails the $\hat{\delta}-$approximate KKT conditions, where $\hat{\delta}$ is equal to δ up to a fixed multiplicative factor. Lemma **4.4** below is the key tool in establishing the connection between the $\|d(x)\|\leq\delta$ stopping criterion and the approximate KKT conditions. In what follows, $L_{\nabla\widetilde{f}}$, $L_{\nabla\widetilde{g}}$ and $L_{\widetilde{g}}$ denote the Lipschitz moduli of $\nabla_1\widetilde{f}(\bullet;\bullet)$, $\nabla_1\widetilde{g}(\bullet;\bullet)$ and $\widetilde{g}(\bullet;\bullet)$ on $\beta\mathbb{B}_{\infty}^n\times S$.

Lemma 4.4. *Let $\{x^{\nu}\}$ be the sequence generated by Algorithm 1 with $\delta=0$ under Assumptions CQ, Approx and Compact. Then, a common positive constant a exists such that, for*

every x^ν, we have

$$\|\nabla f(x^\nu)+\nabla g(x^\nu)\xi^\nu\| \le \left[L_{\nabla\widetilde{f}}+\left(L_{\nabla\widetilde{g}}+\frac{1}{\beta}\right)\Xi\right]\|d(x^\nu)\|, \tag{25}$$

$$\max_i\{g_i(x^\nu)_+\} \le \left(\frac{L_{\widetilde{g}}}{\lambda}+a\right)\|d(x^\nu)\|, \tag{26}$$

$$\max_i|g_i(x^\nu)\xi_i^\nu| \le \max\left\{2L_{\widetilde{g}},\left(\frac{L_{\widetilde{g}}}{\lambda}+a\right)\right\}\Xi\|d(x^\nu)\|, \tag{27}$$

for some $0\le\xi^\nu\in\mathbb{R}^m$.

Proof. As for the perturbed gradient of the Lagrangian-related condition, letting $\xi^\nu \in N_{\mathbb{R}^m_-}(\widetilde{g}(d(x^\nu);x^\nu)-\kappa(x^\nu)e)$, which implies $\xi\ge 0$, we have

$$\begin{aligned}\|\nabla f(x^\nu)+\nabla g(x^\nu)\xi^\nu\| &= \|\nabla_1\widetilde{f}(0;x^\nu)-\nabla_1\widetilde{f}(d(x^\nu);x^\nu)+\nabla_1\widetilde{f}(d(x^\nu);x^\nu)\\ &\quad+\nabla_1\widetilde{g}(d(x^\nu);x^\nu)\xi^\nu+\nabla_1\widetilde{g}(0;x^\nu)\xi^\nu-\nabla_1\widetilde{g}(d(x^\nu);x^\nu)\xi^\nu\|\\ &\le L_{\nabla\widetilde{f}}\|d(x^\nu)\|+L_{\nabla\widetilde{g}}\|\xi^\nu\|\|d(x^\nu)\|+\|\zeta^\nu\|,\end{aligned} \tag{28}$$

for some $\zeta^\nu\in N_{\beta\mathbb{B}^n_\infty}(d(x^\nu))$, where the equality is due to A4 and A9 and the inequality follows from (6). Since $\zeta^\nu=0$ whenever $\|d(x^\nu)\|_\infty<\beta$, consider $d(x^\nu)$ such that $\|d(x^\nu)\|_\infty=\beta$: by (28),

$$\begin{aligned}\|\nabla f(x^\nu)+\nabla g(x^\nu)\xi^\nu\| &\le L_{\nabla\widetilde{f}}\|d(x^\nu)\|+L_{\nabla\widetilde{g}}\|\xi^\nu\|\|d(x^\nu)\|+\|\zeta^\nu\|\tfrac{\|d(x^\nu)\|_\infty}{\beta}\\ &\le L_{\nabla\widetilde{f}}\|d(x^\nu)\|+L_{\nabla\widetilde{g}}\|\xi^\nu\|\|d(x^\nu)\|+\tfrac{1}{\beta}\|\zeta^\nu\|\|d(x^\nu)\|\\ &\le \left[L_{\nabla\widetilde{f}}+\left(L_{\nabla\widetilde{g}}+\tfrac{1}{\beta}\right)\Xi\right]\|d(x^\nu)\|,\end{aligned} \tag{29}$$

where the last inequality follows observing that, in view of Proposition **3.1** and Lemma **4.1** and by continuity, ζ^ν is bounded on S, and we can take, without loss of generality, Ξ to be an upper bound also for $\|\zeta^\nu\|$. Regarding feasibility, by (7), we get $\max_i\{g_i(x^\nu)_+\}-\kappa(x^\nu)\le L_{\widetilde{g}}\|d(x^\nu)\|$, and, in turn,

$$\begin{aligned}\max_i\{g_i(x^\nu)_+\}-\kappa(x^\nu) &= \lambda\left[\max_i\{g_i(x^\nu)_+\}-\min_d\{\max_i\{\widetilde{g}_i(d;x^\nu)_+\}\,|\,\|d\|_\infty\le\rho\}\right]\\ &\le L_{\widetilde{g}}\|d(x^\nu)\|.\end{aligned} \tag{30}$$

We now show that, for any x^ν,

$$\min_d\left\{\max_i\{\widetilde{g}_i(d;x^\nu)_+\}\,|\,\|d\|_\infty\le\rho\right\}\le a\|d(x^\nu)\| \tag{31}$$

for some positive constant a.

Suppose on the contrary that subsequences $\{a^\nu\}_{\mathscr{N}}\in\mathbb{R}_+$ and $\{x^\nu\}_{\mathscr{N}}$ exist such that $a^\nu\underset{\mathscr{N}}{\to}+\infty$ and

$$\min_d\left\{\max_i\{\widetilde{g}_i(d;x^\nu)_+\}\,|\,\|d\|_\infty\le\rho\right\}>a^\nu\|d(x^\nu)\|. \tag{32}$$

Relation (32) implies

$$\min_d \left\{ \max_i \{\widetilde{g}_i(d;x^\nu)_+\} \,|\, \|d\|_\infty \le \rho \right\} > 0. \tag{33}$$

Observing that, by A7, $\max_i\{g_i(x^\nu)_+\} \ge \min_d \{\max_i\{\widetilde{g}_i(d;x^\nu)_+\} \,|\, \|d\|_\infty \le \rho\} > 0$, thanks to Assumptions CQ and Compact, and in view of Lemma 2 of Facchinei *et al.* (2020), we have $\kappa(x^\nu) < \max_i\{g_i(x^\nu)_+\}$ and $\|d(x^\nu)\| \neq 0$ for every $\nu \in \mathcal{N}$. Moreover, since $\min_d \{\max_i\{\widetilde{g}_i(d;\bullet)_+\} \,|\, \|d\|_\infty \le \rho\}$ and $d(\bullet)$ are continuous on $\mathbb{R}^n$ and S, respectively, see Proposition **3.1**, we have, invoking Theorem 1 of Facchinei *et al.* (2020), $\|d(x^\nu)\| \underset{\mathcal{N}}{\to} \|d(\hat{x})\| = 0$, with $\hat{x}$ cluster point of subsequence $\{x^\nu\}_{\mathcal{N}}$. Resorting again to Lemma 2 of Facchinei *et al.* (2020), $\kappa(\hat{x}) = \max_i\{g_i(\hat{x})\}_+$ and $\hat{d} \in \rho\mathbb{B}^n_\infty$ exist such that $\widetilde{g}(\hat{d};\hat{x}) < 0$. Hence, by continuity (A6), a neighborhood of $\hat{d}$ exists such that, for any $d \in \rho\mathbb{B}^n_\infty$ belonging to it, and for $\nu \in \mathcal{N}$ sufficiently large, we have $\widetilde{g}(d;x^\nu) < 0$, in contradiction to (33). Hence, (31) holds.

Combining (30) and (31), we get

$$\max_i\{g_i(x^\nu)_+\} \le \frac{L_{\widetilde{g}}}{\lambda}\|d(x^\nu)\| + \min_d \left\{ \max_i \{\widetilde{g}_i(d;x^\nu)_+\} \,|\, \|d\|_\infty \le \rho \right\} \le \left(\frac{L_{\widetilde{g}}}{\lambda} + a\right) \|d(x^\nu)\|. \tag{34}$$

Concerning the complementarity condition, again without loss of generality, letting $\max_i |g_i(x^\nu)\xi_i^\nu| = |g_{\bar{\imath}}(x^\nu)\xi_{\bar{\imath}}^\nu| > 0$ for some $\bar{\imath} \in \{1,\dots,m\}$ such that $\widetilde{g}_{\bar{\imath}}(d(x^\nu);x^\nu) = \kappa(x^\nu)$, we distinguish two cases.

If $g_{\bar{\imath}}(x^\nu) > 0$, thanks to the local boundedness of the set of KKT multipliers (recall Proposition **3.1** and Lemma **4.1**), we have

$$|g_{\bar{\imath}}(x^\nu)\xi_{\bar{\imath}}^\nu| \le \max_i\{g_i(x^\nu)_+\}|\xi_{\bar{\imath}}^\nu| \le \left(\frac{L_{\widetilde{g}}}{\lambda} + a\right) \Xi \|d(x^\nu)\|. \tag{35}$$

If, on the contrary, $g_{\bar{\imath}}(x^\nu) < 0$, in view of A7,

$$0 \le \kappa(x^\nu) \;=\; \widetilde{g}_{\bar{\imath}}(d(x^\nu);x^\nu) - \widetilde{g}_{\bar{\imath}}(0;x^\nu) + \widetilde{g}_{\bar{\imath}}(0;x^\nu) \le L_{\widetilde{g}}\|d(x^\nu)\| + g_{\bar{\imath}}(x^\nu) < L_{\widetilde{g}}\|d(x^\nu)\|,$$

and, thus,

$$\begin{aligned} |g_{\bar{\imath}}(x^\nu)\xi_{\bar{\imath}}^\nu| &= |\widetilde{g}_{\bar{\imath}}(0;x^\nu) - \widetilde{g}_{\bar{\imath}}(d(x^\nu);x^\nu) + \widetilde{g}_{\bar{\imath}}(d(x^\nu);x^\nu)|\,|\xi_{\bar{\imath}}^\nu| \\ &\le (L_{\widetilde{g}}\|d(x^\nu)\| + |g_{\bar{\imath}}(d(x^\nu);x^\nu)|)\,|\xi_{\bar{\imath}}^\nu| \\ &\le 2L_{\widetilde{g}}|\xi_{\bar{\imath}}^\nu|\,\|d(x^\nu)\| \le 2L_{\widetilde{g}}\Xi\|d(x^\nu)\|. \end{aligned} \tag{36}$$

□

We remark that the results in Lemma **4.4**, which are obtained considering $\delta = 0$, are still valid for any x^ν that is generated by Algorithm 1 with $\delta > 0$, since the sequence produced in the latter case results in nothing else but a truncation of the one generated whenever δ is set to zero.

We now define x^ν to be a $\delta-$approximate KKT point for problem (P) if each of the conditions in the KKT system is satisfied within a tolerance δ: more precisely,

$$\max\left\{ \|\nabla f(x^\nu) + \nabla g(x^\nu)^T \xi^\nu\|, |\max_i g_i(x^\nu)_+|, \max_i |g_i(x^\nu)\xi_i^\nu| \right\} \le \delta.$$

Note that in principle we should also include the term $\|\xi_-^\nu\|$ in the max above, but this is not necessary here, because we already know that $\xi^\nu \geq 0$. It is interesting to highlight that this definition of approximate KKT point meets some important criteria for an approximate, or sequential, optimality condition (Andreani *et al.* 2010). The following theorem establishes the desired connection between the stopping criterion of Algorithm 1 and approximate KKT points; its proof follows easily from (25)-(27).

Theorem 4.2. *Let $\{x^\nu\}$ be the sequence generated by Algorithm 1 under Assumptions CQ, Approx and Compact. If, at iteration ν of Algorithm 1, we have $\|d(x^\nu)\| \leq \delta$, then x^ν is a $\delta'-$approximate KKT point for problem P with*

$$\delta' \triangleq \delta \max\left\{\left[L_{\nabla\widetilde{f}} + \left(L_{\nabla\widetilde{g}} + \frac{1}{\beta}\right)\Xi\right], \left[\frac{L_{\widetilde{g}}}{\lambda} + a\right], \left[\max\left\{2L_{\widetilde{g}}, \left(\frac{L_{\widetilde{g}}}{\lambda} + a\right)\right\}\Xi\right]\right\}.$$

Acknowledgments

Francisco Facchinei was partially supported by Progetto di Ateneo Distributed optimization algorithms for Big Data. Vyacheslav Kungurtsev was supported by the OP VVV project CZ.02.1.01/0.0/0.0/16_019/0000765 “Research Center for Informatics". Lorenzo Lampariello was partially supported by the MIUR PRIN 2017 (grant 20177WC4KE). Gesualdo Scutari was partially supported by the NSF Grants CIF 1564044, CIF 1719205, and CMMI 1832688; and the ARO under Grant W911NF1810238.

References

Andreani, R., Martínez, J. M., and Svaiter, B. F. (2010). “A new sequential optimality condition for constrained optimization and algorithmic consequences”. *SIAM Journal on Optimization* **20**(6), 3533–3554. DOI: 10.1137/090777189.

Burke, J. V. (1989). “A sequential quadratic programming method for potentially infeasible mathematical programs”. *Journal of Mathematical Analysis and Applications* **139**(2), 319–351. DOI: 10.1016/0022-247X(89)90111-X.

Burke, J. V. and Han, S.-P. (1989). “A robust sequential quadratic programming method”. *Mathematical Programming* **43**(1), 277–303. DOI: 10.1007/BF01582294.

Facchinei, F., Kungurtsev, V., Lampariello, L., and Scutari, G. (2020). “Ghost penalties in nonconvex constrained optimization: Diminishing stepsizes and iteration complexity”. *Mathematics of Operations Research*, to appear.

Facchinei, F., Lampariello, L., and Scutari, G. (2017). “Feasible methods for nonconvex nonsmooth problems with applications in green communications”. *Mathematical Programming* **164**, 55–90. DOI: 10.1007/s10107-016-1072-9.

Martínez, J. M. (2017). “On high-order model regularization for constrained optimization”. *SIAM Journal on Optimization* **27**(4), 2447–2458. DOI: 10.1137/17M1115472.

Scutari, G., Facchinei, F., Song, P., Palomar, D. P., and Pang, J.-S. (2014). “Decomposition by partial linearization: parallel optimization of multi-agent systems”. *IEEE Transactions on Signal Processing* **62**(3), 641–656. DOI: 10.1109/TSP.2013.2293126.

[a] University of Rome La Sapienza,
Department of Computer, Control, and System Engineering Antonio Ruberti,
Via Ariosto 25, 00185 Rome, Italy

[b] Czech Technical University in Prague,
Department of Computer Science, Faculty of Electrical Engineering,
Prague, Czech Republic

[c] Roma Tre University,
Department of Business Studies,
Via D'Amico 77, 00145 Rome, Italy

[d] Purdue University,
School of Industrial Engineering,
West-Lafayette, IN, USA

* To whom correspondence should be addressed | email: francisco.facchinei@uniroma1.it

Paper contributed to the meeting on "Variational Analysis, PDEs and Mathematical Economics",
held in Messina, Italy (19–20 September 2019), on the occasion of Prof. Antonino Maugeri's 75th birthday,
under the patronage of the *Accademia Peloritana dei Pericolanti*

Manuscript received 31 March 2020; published online 13 December 2020