Ghost penalties in nonconvex constrained optimization: Diminishing stepsizes and iteration complexity

Francisco Facchinei
Sapienza University of Rome, Department of Computer, Control, and Management Engineering Antonio Ruberti, Via Ariosto 25, 00185 Rome, francisco.facchinei@uniroma1.it,

Vyacheslav Kungurtsev
Czech Technical University in Prague, Department of Computer Science, Faculty of Electrical Engineering, vyacheslav.kungurtsev@fel.cvut.cz,

Lorenzo Lampariello
Roma Tre University, Department of Business Studies, Via D’Amico 77, 00145 Rome, lorenzo.lampariello@uniroma3.it,

Gesualdo Scutari
Purdue University, School of Industrial Engineering, West-Lafayette, IN, USA, gscutari@purdue.edu,

We consider, for the first time, general diminishing stepsize methods for nonconvex, constrained optimization problems. We show that by using directions obtained in an SQP-like fashion convergence to generalized stationary points can be proved. In order to do so, we make use of classical penalty functions in an unconventional way. In particular, penalty functions only enter in the theoretical analysis of convergence while the algorithm itself is penalty-free. We then consider the iteration complexity of this method and some variants where the stepsize is either kept constant or decreased according to very simple rules. We establish convergence to $\delta$-approximate stationary points in at most $O(\delta^{-2})$, $O(\delta^{-3})$, or $O(\delta^{-4})$ iterations according to the assumptions made on the problem. These complexity results complement nicely the very few existing results in the field.

Key words: Constrained optimization, nonconvex problem, diminishing stepsize, generalized stationary point, iteration complexity
MSC2000 subject classification: 90C30, 90C60, 65K05
OR/MS subject classification: Primary: programming: nonlinear algorithms; secondary: analysis of algorithms: computational complexity

1. Introduction We consider Diminishing Stepsize Methods (DSMs), and some associated variants, for the solution of the general, nonconvex optimization problem

$$\min_{x} f(x)$$

s.t. \hspace{1cm} g(x) \leq 0, \hspace{1cm} \text{(P)}

where $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^m$ are LC$^1$ (i.e., continuously differentiable with locally Lipschitz gradients) functions. Equality constraints can easily be added, but we avoid this for the sake of notational simplicity. DSMs generate a sequence $\{x^\nu\}$ by setting

$$x^{\nu+1} = x^\nu + \gamma^\nu d^\nu, \hspace{1cm} \text{(1)}$$

where $d^\nu$ is a “suitable” direction and $\gamma^\nu$ is a positive stepsize such that

$$\lim_{\nu \to \infty} \gamma^\nu \downarrow 0 \hspace{1cm} \text{and} \hspace{1cm} \sum_{\nu=0}^{\infty} \gamma^\nu = \infty. \hspace{1cm} \text{(2)}$$

In order to study complexity issues, we also consider variants of such schemes, in that the stepsize is either kept constant or is progressively reduced according to the information gathered during the minimization process.
Advantages and disadvantages of DSMs are well-known. Their main disadvantage is their slow convergence rate. On the other hand, DSMs are attractive because of their simplicity and low computational cost and they play a crucial role when the problem data are affected by noise, see e.g. [3, 23]; they are also a key tool in distributed optimization, in stochastic optimization, in incremental methods, and whenever the computation of the objective function is very expensive or difficult, see e.g. [3, 4, 16, 18, 23, 27]. These positive aspects explain the continuing interest in this class of methods. DSMs were introduced in connection to nondifferentiable, unconstrained, convex optimization problems and can be readily extended to the convex, constrained case, see e.g. [3, 23, 28] as entry points to the literature. In this nondifferentiable, convex setting, convergence relies on the decrease of the iterates’ distance to the optimal solution set. If the diminishing stepsize method is applied to a differentiable problem, alternative convergence arguments based on the decrease of the objective function can be used, and this paves the way to the application of DSMs to differentiable, unconstrained, nonconvex problems. However, DSMs are much harder to analyze in a constrained, nonconvex setting, and to date only very partial results are available. In particular, to the best of our knowledge, convergence has been shown only if the feasible set is convex, see e.g., [2, 3] or in some specialized settings where feasibility of the iterates can be maintained throughout the optimization process [16]. The situation is summarized in Table 1. The main aim of this paper is:

**Table 1.**

|            | Convex | Nonconvex |
|------------|--------|-----------|
| Unconstrained | ✓      | ✓         |
| Constrained  | ✓      | ✗         |

(a) to give a general analysis for the bottom-right corner case in Table 1, thus completing the study of DSMs in the differentiable case,

(b) to perform an iteration complexity analysis, of some variants of the DSM.

Below we elaborate further on these two contributions. The results related to (a) considerably widen the scope of applicability of DSMs. We show how to use a Sequential Quadratic Programming (SQP)-like approach to systematically generate directions $d^\nu$ that, when plugged into (1), guarantee convergence, in the sense that every limit point of the sequence $\{x^\nu\}$ produced according to (1) and (2) is a generalized stationary solution for (P). By generalized stationary, we intended a point that can be: a stationary solution of the feasibility problem associated to (P), a Fritz-John or a KKT point point of (P). This is the natural target for an algorithm for constrained optimization when neither blanket assumptions about feasibility of (P) are made nor constraint qualifications are assumed to hold. Indeed, existence of a KKT point is not even guaranteed in our setting, and this well-established, broader approach is needed.

The directions $d^\nu$ used in our method are the solutions of suitable convex approximations of the original problem (P), constructed along the lines discussed in the seminal papers [6, 8] and also taking into account developments in [16, 27]. Specifically, given an iteration $x^\nu$, we consider SQP-like direction-finding subproblems of the type

$$\min_d \tilde{f}(d; x^\nu) \quad \text{s.t.} \quad \tilde{g}(d; x^\nu) \leq \kappa(x^\nu) e, \quad \|d\|_\infty \leq \beta$$

where $\tilde{f}(\bullet; x^\nu)$ is strongly convex, all $\tilde{g}_i(\bullet; x^\nu)$ are convex, $\kappa(x^\nu) \in \mathbb{R}$ is nonnegative and $e \in \mathbb{R}^m$ is the vector with all components being one (additional technical assumptions will be discussed later on). We show that subproblem (3) always has a unique solution by construction. The choice of these direction-generating subproblems includes, for example, SQP-type subproblems, when $\tilde{g}$ is just a
linearization of $g$ and $\hat{f}$ a positive definite quadratic approximation of $f$. But the conditions we impose on (3) leave room for much flexibility in tailoring (3) to the problem at hand and to exploit any available specific structure in (P). Of course, the choice of $\hat{f}$ and $\tilde{g}$ must aim at a balance that makes (3) as easy to solve as possible, while generating “good” directions $d^c$. We do not go into these practical details, since they are problem dependent, but we underline that the assumptions necessary for our analysis cover a sufficiently broad set of cases.

The main idea to prove convergence of our method is to show that a suitable exact penalty function exists for which the directions $d^c$ give “sufficient decrease”. In turn, the decrease property typically depends on the appropriate choice of a penalty parameter. The tuning of the penalty parameter, which must be carried out while the algorithm progresses, is a critical aspect of SQP methods, since it considerably influences the overall performance of such schemes. From this point of view, one of the main novelties of our approach is a “virtual” use of the penalty function. By this, we mean that we use the existence of a suitable value of the penalty parameter as a theoretical tool to establish convergence, but we never need to actually compute the penalty function nor do we have to know the appropriate value for the penalty parameter to perform the optimization process; whence the term ghost penalty.

Results indicated in (b) add to a thus-far sparse, but thriving literature that just recently began appearing on the topic of complexity analysis for nonconvex optimization problems. Disregarding classical results on the gradient method, see e.g. [21], this chapter was opened by Nesterov and Polyak’s paper [22], with their analysis of a cubic regularization method for the unconstrained minimization of a nonconvex, smooth function. An excellent review of results in this field is contained in [12], to which we refer the interested reader for a broader view on the subject. Here we only briefly discuss results on algorithms for nonconvex, inequality constrained problems aimed at locating generalized stationary points. The worst-case, first-order iteration complexity of an optimization algorithm is the maximal number of iterations needed to drive a suitably defined stationarity measure below a threshold $\delta$, i.e. to find a $\delta$–approximate generalized stationary point (see Section 5). There exist also iteration complexity results for algorithms using higher order derivatives, but in our method only first-order derivatives are necessary, and consequently we restrict our discussion to first-order complexity. Iteration complexity for nonconvex, constrained optimization problems has been the subject of [5, 9, 11, 12].

In this paper we establish iteration complexity results for some variants of our extremely simple DSM. At first, assuming that an extended Mangasarian-Fromovitz constraint qualification holds, we prove that the complexity to find a $\delta$–approximate KKT point is $O(\delta^{-2})$ if one uses a sufficiently small, constant stepsize. In addition, for this case we also establish some complexity bounds for DSMs. In the general case, where no constraint qualifications are assumed, we show that $O(\delta^{-4})$ iterations are needed at worst to find a $\delta$–approximate generalized stationary point. To perform this analysis we employ a stepsize that is kept constant until some test is satisfied triggering a reduction of the stepsize. Overall, this results in the use of “piecewise constant” stepsizes. If a feasible starting point is known, this complexity can be reduced to $O(\delta^{-3})$ if one is willing to accept that the algorithm depends (in a simple way) on the required precision. The above results require the knowledge of some problem constants (e.g. Lipschitz constants); in this case, the number of objective function/constraint functions evaluations coincides with the number of iterations. If we do not know the problem constants, we can define a further algorithm variant using a sort of line-search procedure which still guarantees an iteration complexity of $O(\delta^{-4})$, although a negligible (w.r.t. $O(\delta^{-2})$) amount of additional objective function/constraint functions evaluations may be needed.

It appears that our results are complementary and consistent with those in the literature. The algorithm in [5] takes $O(\delta^{-3})$ iterations to declare a point approximate generalized stationary. This complexity result of course lies in between our bounds of $O(\delta^{-2})$ and $O(\delta^{-4})$. When MFCQ holds, our $O(\delta^{-2})$ complexity matches those in [9, 11]. Regardless, be it $O(\delta^{-2})$, $O(\delta^{-3})$ or $O(\delta^{-4})$, we believe
that the bounds established in this paper, for our strikingly simple method, are very interesting and valuable, also in view of the paucity of results in the literature. We are not aware of any complexity results in the literature for nonconvex problems using an SQP-like scheme, nor do we know of any previous results for DSMs. Furthermore, as a major departure from the current proposals in the literature, our approach is not a two-phase algorithm: in fact, our method does not separate the search for an approximately feasible point from a subsequent feasibility improvement as done, e.g., in [5, 9, 11, 12].

The paper is organized as follows. In Section 2 we introduce some mathematical preliminaries and, in particular, the appropriate definition of generalized stationary point for a nonconvex, constrained problem. In Section 3 we discuss more in detail the direction finding subproblem (3) and introduce some assumptions that will be used to establish convergence. In Section 4 we show convergence to generalized stationary points, while in Section 5 we perform the iteration complexity analysis.

2. Stationary Points

We consider Problem (P), where \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) and \( g : \mathbb{R}^n \rightarrow \mathbb{R}^m \) are assumed to be LC\(^1\) on \( \mathbb{R}^n \) throughout the paper. Note that we do not assume that problem (P) is feasible, let alone that it has a solution. Therefore we aim at deriving convergence results for both feasible and infeasible problems, in some suitable sense.

A general constrained problem (P) can be viewed as a combination of two problems: (i) the feasibility one, i.e., the problem of finding a feasible point; (ii) the problem of finding a local minimum point of the objective function over the feasible set. Even the former problem is a hard one, since it essentially requires to compute a global minimum of the generally nonconvex function expressing the violation of the constraints. Consistently, we design our algorithm to converge to stationary solutions in a generalized sense, that is to points that either are stationary for (P) or are infeasible and stationary for the following violation-of-the-constraints optimization problem:

\[
\min_{\bar{x}} \max_i \{g_i(x)_+\},
\]

where \( \alpha_+ \triangleq \max\{0, \alpha\} \) for all \( \alpha \in \mathbb{R} \). Let

\[
M_1(x) \triangleq \left\{ \xi \mid \xi \in N_{\mathbb{R}_+^m}(g(x)), 0 = \nabla f(x) + \nabla g(x) \xi \right\}
\]

and

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

where \( N_{\mathbb{R}_+^m}(y) \) is the classical normal cone to the convex set \( R_+^m \) at \( y \), \( \nabla f \) is the gradient of \( f \) and \( \nabla g \) is the transposed Jacobian of \( g \). If \( x \) is feasible, i.e. if \( g(x) \leq 0 \), condition \( \xi \in N_{\mathbb{R}_+^m}(g(x)) \) can be more familiarly be rewritten as

\[
\xi_i \geq 0, \quad \xi_i g_i(x) = 0
\]

for all \( i \) (a similar reasoning applies to the normal cone expression in the definition of \( M_0(x) \)). We note explicitly that if \( x \) is not feasible, the set \( M_1(x) \) is empty. 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 point) or \( M_0(\hat{x}) \neq \{0\} \) (the point is a Fritz-John point), or both. 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 problem (4),

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

is equivalent to \( M_0(\hat{x}) \neq \{0\} \). Hence, the (generalized) stationarity criteria for the original problem (P) can naturally be specified by using the sets \( M_1 \) and \( M_0 \), as detailed in Definition 1.

**Definition 1.** A point \( \hat{x} \) is, for problem (P),

- a KKT solution if \( g(\hat{x}) \leq 0 \) and \( M_1(\hat{x}) \neq \emptyset \);
• a FJ solution if \( g(\hat{x}) \leq 0 \) and \( M_0(\hat{x}) \neq \{0\} \);
• an External Stationary (ES) solution if \( g_i(\hat{x}) > 0 \) for at least an index \( i \in \{1, \ldots, m\} \) and \( M_0(\hat{x}) \neq \{0\} \).

We call \( \hat{x} \) a stationary solution of \((P)\) if any of the previous cases occurs.

Since we did not make any regularity or feasibility assumption on problem \((P)\), finding a stationary solution in the sense just described is the appropriate requirement for a solution algorithm; we show that our method does converge to stationary points indeed. It also turns out that, under classical regularity conditions, our algorithm actually converges to KKT points. The constraint qualification (CQ) we use is the Mangasarian-Fromovitz one, suitably extended to (possibly) infeasible points.

**Definition 2.** We say that the extended Mangasarian-Fromovitz Constraint Qualification (eMFCQ) holds at \( \hat{x} \) if
\[
M_0(\hat{x}) = \{0\}.
\]

If \( \hat{x} \) is feasible, this condition reduces to the classical MFCQ and in turn, whenever the constraints are convex, it is well-known that the MFCQ is equivalent to Slater’s CQ, i.e. to the existence of a point \( \hat{x} \) such that \( g(\hat{x}) < 0 \). We state below a result that extends a standard property of the MFCQ for feasible points.

**Proposition 1.** If the eMFCQ holds at point \( \hat{x} \), then there exists a neighborhood of \( \hat{x} \) such that, for every point \( x \) belonging to it, the eMFCQ is satisfied.

**Proof.** If \( \hat{x} \) is feasible, this is a classical result. If \( \hat{x} \) is not feasible, the condition \( M_0(\hat{x}) = \{0\} \) implies that \( \hat{x} \) is not a stationary point for the feasibility problem \((4)\), i.e. \( 0 \not\in \partial \max_i \{g_i(\hat{x})_+\} \). The assertion then easily follows from the outer semicontinuity of the subdifferential mapping (see \[25\] for the definition of outer semicontinuity).

### 3. Direction Finding Subproblem
At each iteration of our algorithm we move from the current iteration \( x^\nu \) along the direction \( d(x^\nu) \) with a stepsize \( \gamma^\nu \), see \((1)\). While the stepsize is chosen according to classic diminishing stepsizes rules, the direction \( d(x^b) \) is the solution of a suitable strongly convex subproblem that we describe next.

Given a (base) point \( x \) (which will actually be the current iterate \( x^\nu \) in the algorithm) \( d(x) \) is the unique solution of the following strongly convex optimization problem:

\[
\begin{align*}
\text{minimize} \quad & \tilde{f}(d; x) \\
\text{s.t.} \quad & \tilde{g}(d; x) \leq \kappa(x)e \\
& ||d||_\infty \leq \beta,
\end{align*}
\]

where \( e \in \mathbb{R}^m \) is the vector with all components being one, \( \kappa(x) \) a nonnegative quantity that will be defined shortly, and \( \beta \) is a positive constant. Moreover, \( \tilde{f} \) is a strongly convex surrogate of the original objective function \( f \), while \( \tilde{g} \) is a convex surrogate of the original constraints \( g \) (see Assumption A below for the conditions these surrogates must obey).

The r.h.s. \( \kappa(x)e \) in the surrogate constraints serves to suitably enlarge the feasible set of the subproblem in order to make it always nonempty. The additional constraint \( ||d||_\infty \leq \beta \) allows us to avoid issues with ever-increasing search directions. Overall, in the sequel we denote by \( \tilde{X}(x) \) the convex, feasible set of subproblem \((P_x)\), i.e.

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

and we equivalently refer to constraint \( ||d||_\infty \leq \beta \) as \( d \in \beta B_\infty^n \), where \( B_\infty^n \) 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 in [6], to which it reduces when the standard quadratic/linear approximations are used for \(\tilde{f}\) and \(\tilde{g}\):

\[
\tilde{f}(d; x) \triangleq \nabla f(x)^T d + \frac{1}{2}\|d\|^2;
\tilde{g}(d; x) \triangleq g(x) + \nabla g(x)^T d.
\]  

(5)

In this paper, \(\|\cdot\|\) denotes the Euclidean norm or the corresponding induced matrix norm. Note that if this approximations are employed and we set \(\kappa(x) = 0\) and \(\beta = +\infty\), \((P_x)\) boils down to the classical SQP-type subproblem. Here, on the one hand, we adopt the approach in [6] by taking \(\kappa(x)\) not necessarily zero and \(\beta < +\infty\) in order to guarantee the existence and some continuity properties of the solution mapping \(d(x)\); on the other hand, we introduce the use of general approximations \(\tilde{f}\) and \(\tilde{g}\), that may be very convenient in practice. 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 \(\tilde{f}\) and \(\tilde{g}\), which in turn is dictated by the original problem \((P)\).

But the use of models that go beyond the classical 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, sometimes even in closed form, see e.g. [16, 26], and, on the other hand, by the desire of faster convergence rates, see for example the discussion in Section 3 of [20].

For our approach to be legitimate and lead to useful convergence results, we obviously need to make assumptions on the surrogate functions \(\tilde{f}\) and \(\tilde{g}\).

**Assumption A**

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

A1) \(\tilde{f}(\bullet; x)\) is a strongly convex differentiable function on \(O_d\) for every \(x \in \mathbb{R}^n\) with modulus of strong convexity \(c > 0\) independent of \(x\);

A2) \(\tilde{f}(\bullet; \bullet)\) is continuous on \(O_d \times \mathbb{R}^n\);

A3) \(\nabla_1 \tilde{f}(\bullet; \bullet)\) is continuous \(O_d \times \mathbb{R}^n\);

A4) \(\nabla_1 \tilde{f}(0; x) = \nabla f(x)\) for every \(x \in \mathbb{R}^n\);

A5) \(\tilde{g}_i(\bullet; x)\) is a convex function on \(O_d\) for every \(x \in \mathbb{R}^n\);

A6) \(\tilde{g}_i(\bullet; \bullet)\) is continuous on \(O_d \times \mathbb{R}^n\);

A7) \(\tilde{g}_i(0; x) = g_i(x)\) for every \(x \in \mathbb{R}^n\);

A8) \(\nabla_1 \tilde{g}_i(\bullet; \bullet)\) is continuous \(O_d \times \mathbb{R}^n\);

A9) \(\nabla_1 \tilde{g}_i(0; x) = \nabla g_i(x)\), for every \(x \in \mathbb{R}^n\);

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

These conditions are easily satisfied in practice and have been employed in many recent papers; we refer the reader to [16, 27] as good sources of examples. Here we note that the classical quadratic/linear approximations (5) satisfy Assumption A, since \(f\) and \(g\) are smooth. To complete the description of subproblem \((P_x)\), following [6], we set

\[
\kappa(x) \triangleq (1 - \lambda) \max_i \{ g_i(x)_+ \} + \lambda \min_d \left\{ \max_i \{ \tilde{g}_i(d; x)_+ \} \mid \|d\|_{\infty} \leq \rho \right\},
\]

(6)

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

\[
\min_d \left\{ \max_i \{ \tilde{g}_i(d; x)_+ \} \mid \|d\|_{\infty} \leq \rho \right\}
\]

that always has an optimal solution because the feasible set is nonempty and compact. If \(\tilde{g}\) is linear, as in the classical choice given in (5), this problem reduces to a linear programming problem and
can be efficiently and easily solved. If $x$ is feasible for (P), i.e. $g(x) \leq 0$, we have $\kappa(x) = 0$, so that our subproblem $(P_x)$ is very similar to standard SQP ones. The term $\kappa(x)$ plays a key role when the SQP-type subproblems have an empty feasible set, a very well-known issue with SQP schemes. In fact, since $\kappa(x)$ is always nonnegative, being the sum of two nonnegative quantities, it restores feasibility by enlarging (with respect to the SQP choice $\tilde{g}(d; x) \leq 0$) the range of admissible values, see Figure 1. Moreover, the feasible set of problem $(P_x)$, for every $x$, is nonempty: choosing $\hat{d}$ at

\[ \hat{d} = \arg \min_d \left\{ \max_i \{ \tilde{g}_i(d; \hat{x}) \} \mid \|d\|_\infty \leq \rho \right\} e = \max_i \{ \tilde{g}_i(\hat{d}; x) \} e, \]

and, in turn,

\[ \tilde{g}(\hat{d}; x) = (1 - \lambda)\tilde{g}(\hat{d}; x) + \lambda \tilde{g}(\hat{d}; x) \]

\[ \leq [(1 - \lambda) \max_i \{ \tilde{g}_i(0; x) \} + \lambda \min_d \{ \max_i \{ \tilde{g}_i(d; x) \} \mid \|d\|_\infty \leq \rho \}] e = \kappa(x)e. \]

The function $\kappa(x)$ is obviously continuous and, under a very weak additional requirement, also locally Lipschitz continuous. This result was established in [6] whenever $\tilde{g}$ is the linear approximation in (5) and readily generalizes to the case of the surrogate $\tilde{g}$ we consider here.

**Proposition 2.** Under Assumption A, $\kappa(\bullet)$ is continuous on $\mathbb{R}^n$. If, in addition, $\tilde{g}(\bullet; \bullet)$ is locally Lipschitz continuous on $O_d \times \mathbb{R}^n$, then $\kappa(\bullet)$ is also locally Lipschitz continuous on $\mathbb{R}^n$.

**Proof.** The continuity of $\kappa(\bullet)$ is obvious, while the Lipschitz continuity under the additional condition derives immediately from, e.g., [24, Theorem 3.1].

Note that the local Lipschitz continuity of $\tilde{g}(\bullet; \bullet)$ is part of Assumption C to be introduced shortly.

Since the feasible set of $(P_x)$ is nonempty and because of the strong convexity of its objective, subproblem $(P_x)$ has one and only one solution, which we denote by $d(x)$.

The following technical lemma is very useful for the subsequent developments.

**Lemma 1.** The following results hold:

(i) if $\max_i \{ g_i(x) \}_{+} > 0$ and $\kappa(x) < \max_i \{ g_i(x) \}_{+}$, then, for all $\rho \in (0, \beta)$, there exists $d \in \rho \mathbb{B}^n_\infty$ such that $\tilde{g}(d; x) < \kappa(x)e$;

(ii) if $\max_i \{ g_i(x) \}_{+} > 0$ and $\kappa(x) = \max_i \{ g_i(x) \}_{+}$, then $\hat{x}$ is an ES point for (P);

(iii) if $\max_i \{ g_i(x) \}_{+} = 0$, then either $\hat{x}$ is a FJ point for (P) or, for all $\rho \in (0, \beta)$, there exists $d \in \rho \mathbb{B}^n_\infty$ such that $\tilde{g}(d; \hat{x}) < 0$.

**Proof.** (i) Choosing $\hat{d} = \arg \min_d \{ \max_i \{ \tilde{g}_i(d; \hat{x}) \} \mid \|d\|_\infty \leq \rho \}$, we can infer $\tilde{g}(\hat{d}; \hat{x}) \leq \min_d \{ \max_i \{ \tilde{g}_i(d; \hat{x}) \} \mid \|d\|_\infty \leq \rho \} e$, while $\tilde{g}(\hat{d}; \hat{x}) < \kappa(\hat{x})e < \max_i \{ g_i(\hat{x}) \}_{+}e$ and, thus,

\[ \tilde{g}(\hat{d}; \hat{x}) = \lambda \tilde{g}(\hat{d}; \hat{x}) + (1 - \lambda)\tilde{g}(\hat{d}; \hat{x}) < \kappa(\hat{x})e. \]
The thesis follows since $\rho < \beta$.

(ii) Equality $\kappa(\hat{x}) = \max_i\{g_i(\hat{x})_+\}$ holds if and only if $d = 0$ solves the minimization problem in the definition of $\kappa$ and, in turn, $M_0(\hat{x}) \neq \{0\}$ by A7 and A9.

(iii) With $\max_i\{g_i(\hat{x})_+\}$ being equal to zero, we have $\kappa(\hat{x}) = 0$ and $g(\hat{x}) \leq 0$.

If $M_0(\hat{x}) \neq \{0\}$, then, by definition, $\hat{x}$ is a FJ point for $(P)$ and the result holds.

Thus, let us suppose $M_0(\hat{x}) = \{0\}$. For those $j \in \{1, \ldots, m\}$ such that $g_j(\hat{x}) < 0$, we have $\tilde{g}_j(0; \hat{x}) = g_j(\hat{x}) < 0$; as for indices $k \in \{1, \ldots, m\}$ with $g_k(\hat{x}) = 0$, thanks to standard alternative theorems, there exists $d \in \mathbb{R}^n$ such that

$$0 > \nabla g_k(\hat{x})^T d = \nabla \tilde{g}_k(0; \hat{x})^T d = \lim_{\tau \downarrow 0} \frac{\tilde{g}_k(\tau d; \hat{x}) - \tilde{g}_k(0; \hat{x})}{\tau}.$$

Therefore, by continuity, we have $\tilde{g}_i(\tau d; \hat{x}) < 0$ for every $i$ and for any $\tau$ sufficiently small. □

Leveraging Lemma 1 and resorting to standard results in parametric optimization, we can establish a key continuity property for the solution mapping $d(\bullet)$ of subproblem $(P_x)$.

**Proposition 3.** Under Assumption A, let the eMFCQ hold at $\hat{x}$. Then,

(i) the MFCQ holds at every point of $\hat{X}(\hat{x})$;

(ii) a neighborhood of $\hat{x}$ exists such that, for every point $x$ belonging to it, function $d(\bullet)$ is continuous.

**Proof.** If the eMFCQ holds at $\hat{x}$, case (ii) in Lemma 1 cannot occur. On the other hand, as for both cases (i) and (iii) in Lemma 1, Slater’s constraint qualification holds for $\hat{X}(\hat{x})$ and, since $\hat{X}(\hat{x})$ is convex, this proves (i). Thanks to A6, the set-valued mapping $\hat{X}(\bullet) = \beta \mathbb{B}^n \cap \{d \in \mathbb{R}^n : \tilde{g}(d; \bullet) \leq \kappa(\bullet) e\}$ is outer semicontinuous at $\hat{x}$ by [1, Theorem 3.1.1], having taken into account that $\kappa(\bullet)$ is continuous by Proposition 2. Moreover, $\hat{X}(\bullet)$, by virtue of Slater’s constraint qualification, A5 and A6, is also inner semicontinuous (see [1, Theorem 3.1.6] and, for the definition of inner semicontinuity, [25]) at $\hat{x}$. Hence, thanks to A1, the continuity of $d(\bullet)$, leveraging [1, Theorems 3.1.1 and 4.3.3], follows from [25, Corollary 5.20]. □

To get reinforced convergence results in the next section, we need $d(\bullet)$ to be not only continuous, but also Hölder continuous on compact sets: for this reason, we introduce Assumption B.

**Assumption B**

For every compact set $K \subset \mathbb{R}^n$ two positive constants $\theta$ and $\alpha$ exist such that

$$\|d(y) - d(z)\| \leq \theta\|y - z\|^\alpha, \quad \forall y, z \in K.$$

Since it is not immediately obvious when this condition is satisfied, below we give a set of simple sufficient conditions on $\hat{f}$ and $\tilde{g}$ for Assumption B to hold.

**Assumption C**

C1) $\nabla \hat{f}(\bullet; \bullet)$ is locally Lipschitz continuous on $O_d \times \mathbb{R}^n$;

C2) each $\tilde{g}_j(\bullet; \bullet)$ is locally Lipschitz continuous on $O_d \times \mathbb{R}^n$.

The following proposition, which builds on the results in [29], shows the desired result.

**Proposition 4.** Assume that A1, A3, A5 and C hold. Suppose further that the MFCQ holds at $d(\hat{x}) \in \hat{X}(\hat{x})$ for every $\hat{x}$ in a compact set $K$. Then, there exists $\theta > 0$ such that, for every $y, z \in K$,

$$\|d(y) - d(z)\| \leq \theta\|y - z\|^\frac{1}{\alpha}. \quad (7)$$

**Proof.** Preliminarily, observe that by Proposition 2, $\kappa(\bullet)$ is locally Lipschitz continuous. The MFCQ at $d(\hat{x}) \in \hat{X}(\hat{x})$, for every $\hat{x} \in K$, implies, by [29, Lemma 3.1], that the set-valued mapping $\hat{X}$ has the Aubin property relative to $K$ at $\hat{x}$ for $d(\hat{x})$ for every $\hat{x} \in K$ (see [25] for the definition of...
the Aubin property). Therefore, in view of [29, Theorem 2.1], for every \( \hat{x} \in K \), there exist \( \theta' > 0 \), \( \theta'' > 0 \) and a neighborhood \( \mathcal{V} \) of \( \hat{x} \) such that, for every \( y, z \in \mathcal{V} \cap K \)

\[
\|d(y) - d(z)\| \leq \theta'\|y - z\| + \theta''\|y - z\|^2.
\]

By the previous relation and the compactness of set \( K \), (7) holds.

**Remark 1.** Assumptions A and C may look tediously detailed, but this is necessary to correctly identify the minimal conditions that make our method work. We emphasize that these conditions are trivially satisfied when one uses as \( \tilde{f} \) and \( \tilde{g} \) the classical quadratic/linear approximations (5) of standard SQP methods (if one supposes that \( f \) and \( g \) have not only continuous gradients, but locally Lipschitz gradients). Assumption C reinforces some of the requirements in Assumption A; we refer the reader to [16] for some examples of surrogate \( \tilde{g}s \) satisfying (Assumption A and) Assumption C beyond the obvious case of linear approximations.

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

\[
0 \in \nabla_1 \tilde{f}(d(x); x) + \nabla_1 \tilde{g}(d(x); x) \xi + N_{\beta \mathbb{B}_n^\infty}(d(x)),
\]

with the multipliers \( \xi \) satisfying the conditions \( \xi \geq 0 \) and \( \xi^T \tilde{g}(d(x); x) = 0 \), and where we denote by \( N_{\beta \mathbb{B}_n^\infty}(d(x)) \) the normal cone to \( \beta \mathbb{B}_n^\infty \) at \( d(x) \).

Finally, we establish the local boundedness of the KKT multipliers’ set of the subproblems \((P_x)\).

**Proposition 5.** Under Assumption A, suppose that \( \hat{d} \in \beta \mathbb{B}_n^\infty \) exists such that \( g(\hat{d}, \hat{x}) < 0 \). Then, a neighborhood of \( \hat{x} \) exists such that, for every point \( x \) belonging to it, the unique solution \( d(x) \) of \((P_x)\) is a KKT point of problem \((P_x)\) and the set-valued mapping of KKT multipliers is locally bounded at \( \hat{x} \).

**Proof.** The condition \( g(\hat{d}, \hat{x}) < 0 \) is nothing else but the Slater’s CQ for problem \((P_x)\), which obviously implies that the MFCQ holds at the unique solution of problem \((P_x)\). The derivation of the result is then rather classical and can easily be derived by, e.g., [17, Proposition 5.4.3].

### 4. Algorithm and Convergence

We are now ready to introduce the proposed algorithm, as given in Algorithm 1.

**Algorithm 1:** SCA Algorithm for \((P)\)

**Data:** \( \gamma^0 \in (0, 1); \ x^0, \nu \leftarrow 0; \)

**repeat**

(S.1) if \( x^\nu \) is stationary for \((P)\) then

\mid stop and return \( x^\nu \);

end

(S.2) compute \( \kappa(x^\nu) \) and the solution \( d(x^\nu) \) of problem \((P_{x^\nu})\);

(S.3) set \( x^{\nu+1} = x^\nu + \gamma^\nu d(x^\nu), \ \nu \leftarrow \nu + 1; \)

end

The algorithm is always well defined if Assumption A, which guarantees existence and uniqueness of \( d(x^\nu) \), holds. 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 the
solution of the strongly convex subproblem \((P_{x'})\). This complexity can range from that necessary to solve an LP and a strongly convex quadratic problem, whenever quadratic/linear approximations are used, to that of solving two convex optimization problems. Theorem 1 below establishes the main convergence properties of Algorithm 1.

**Theorem 1.** Consider the sequence \(\{x^\nu\}\) generated by Algorithm 1 with \(\tilde{f}\) and \(\tilde{g}\) such that Assumption A and (2) hold. Then, either the sequence \(\{x^\nu\}\) is unbounded or the following assertions hold:

(i) at least one limit limit point \(\hat{x}\) of \(\{x^\nu\}\) is stationary for problem \((P)\); in particular, if the eMFCQ holds at \(\hat{x}\), then \(\hat{x}\) is a KKT point for problem \((P)\);

(ii) if, in addition, the eMFCQ holds at every limit point of \(\{x^\nu\}\) and Assumption B holds, then every limit point of \(\{x^\nu\}\) is a KKT solution for problem \((P)\).

**Proof.** We assume, without loss of generality, that the sequence \(\{x^\nu\}\) is bounded. Preliminary, observe that, at each step, the solution \(d(x^\nu)\) of subproblem \((P_{x'})\) is also a KKT point for \((P_{x'})\). In fact, suppose that at a certain iteration \(\nu\), \(d(x^\nu)\) does not satisfy the KKT conditions for \((P_{x'})\). The subproblem is always feasible by construction; let us analyze the three, exhaustive cases considered in Lemma 1. In case (i), Slater’s condition holds for \((P_{x'})\) and \(d(x^\nu)\) is a KKT point. In case (ii), \(x^\nu\) is an ES point of \((P)\): hence, we would have stopped at step (S.1). In case (iii), either Slater’s condition holds for \((P_{x'})\) and \(d(x^\nu)\) is a KKT point, or \(x^\nu\) is an FJ point for \((P)\), in which case we would have stopped at step (S.1).

Thus, \(d(x^\nu)\) is a KKT point for \((P_{x'})\) and therefore (see the proof of Proposition 5), multipliers \(\{\xi^\nu\}\) exist such that \(\xi^\nu \in N_{R^m}(\tilde{g}(d(x^\nu); x^\nu) - \kappa(x^\nu)e)\) and

\[
0 \in \nabla_1 \tilde{f}(d(x^\nu); x^\nu) + \nabla_1 \tilde{g}(d(x^\nu); x^\nu)\xi^\nu + N_{R^m}(d(x^\nu)).
\] (8)

Thanks to A1 and A4, we have

\[
\nabla_1 \tilde{f}(d(x^\nu); x^\nu)^T d(x^\nu) = [\nabla_1 \tilde{f}(d(x^\nu); x^\nu) - \nabla_1 \tilde{f}(0; x^\nu) + \nabla_1 \tilde{f}(0; x^\nu)]^T d(x^\nu) \\
\geq c\|d(x^\nu)\|^2 + \nabla f(x^\nu)^T d(x^\nu).
\] (9)

Moreover, in view of A5, for every \(i = 1, \ldots, m\),

\[- \nabla_1 \tilde{g}_i(d(x^\nu); x^\nu)^T d(x^\nu) \leq \tilde{g}_i(0; x^\nu) - \tilde{g}_i(d(x^\nu); x^\nu)
\]

and, by A7, since \(\xi^\nu\) is nonnegative, in turn,

\[- \xi^\nu_i \nabla_1 \tilde{g}_i(d(x^\nu); x^\nu)^T d(x^\nu) \leq \xi^\nu_i [\tilde{g}_i(0; x^\nu) - \tilde{g}_i(d(x^\nu); x^\nu)] = \xi^\nu_i [g_i(x^\nu) - \kappa(x^\nu)]\]

(10)

where the equality follows observing that \(\xi^\nu\) belongs to \(N_{R^m}(\tilde{g}(d(x^\nu); x^\nu) - \kappa(x^\nu)e)\).

Therefore, by (8), (9) and (10), we have, for some \(\xi^\nu \in N_{R^m}(d(x^\nu))\),

\[
c\|d(x^\nu)\|^2 + \nabla f(x^\nu)^T d(x^\nu) \leq \nabla_1 \tilde{f}(d(x^\nu); x^\nu)^T d(x^\nu) = -\xi^\nu T \nabla_1 \tilde{g}(d(x^\nu); x^\nu)^T d(x^\nu) - \xi^\nu T d(x^\nu) \\
\leq \xi^\nu T [g(x^\nu) - \kappa(x^\nu)e] \leq \xi^\nu T \max_i \{g_i(x^\nu)\}_+ - \kappa(x^\nu)e,
\]

and, thus,

\[
\nabla f(x^\nu)^T d(x^\nu) \leq -c\|d(x^\nu)\|^2 + \xi^\nu T \max_i \{g_i(x^\nu)\}_+ - \kappa(x^\nu)e.
\] (11)

We also notice that, since \(d(x^\nu)\) is feasible for problem \((P_{x'})\), by A5, A7 and A9,

\[
\kappa(x^\nu) \geq \tilde{g}_i(d(x^\nu); x^\nu) \geq \tilde{g}_i(0; x^\nu) + \nabla \tilde{g}_i(0; x^\nu)^T d(x^\nu) = g_i(x^\nu) + \nabla g_i(x^\nu)^T d(x^\nu).
\] (12)
Let us now consider the nonsmooth (ghost) penalty function

\[ W(x; \varepsilon) \triangleq f(x) + \frac{1}{\varepsilon} \max_i \{g_i(x)_+\}. \]  

(13)

In the following analysis we will freely invoke some properties of function \((\bullet)_+ \triangleq \max \{0, \bullet\}\), namely \(\max \{0, \alpha_1\} \leq \max \{0, \alpha_2\}\) for any \(\alpha_1, \alpha_2 \in \mathbb{R}\) such that \(\alpha_1 \leq \alpha_2\), \(\max \{0, a \alpha\} = a \max \{0, \alpha\}\) for any \(\alpha \in \mathbb{R}\) and nonnegative scalar \(a\), and \(\max \{0, \alpha_1 + \alpha_2\} \leq \max \{0, \alpha_1\} + \max \{0, \alpha_2\}\) and \(\max \{0, \alpha_1\} - \max \{0, \alpha_2\} \leq \max \{0, \alpha_1 - \alpha_2\}\) for any \(\alpha_1, \alpha_2 \in \mathbb{R}\). We have

\[
W(x^{\nu+1}; \varepsilon) - W(x^{\nu}; \varepsilon)
\]

\[
= f(x^{\nu} + \gamma^{\nu} d(x^{\nu})) - f(x^{\nu}) + \frac{1}{\varepsilon} \left[ \max_i \{g_i(x^{\nu} + \gamma^{\nu} d(x^{\nu}))\}_+ - \max_i \{g_i(x^{\nu})\}_+ \right]
\]

\[
\leq \gamma^{\nu} \nabla f(x^{\nu})^T d(x^{\nu}) + \frac{\gamma^{nu}^T L_{\nabla g_i}}{2} \|d(x^{\nu})\|^2 + \frac{1}{\varepsilon} \left[ \max_i \{g_i(x^{\nu}) + \gamma^{\nu} \nabla g_i(x^{\nu})^T d(x^{\nu})\}_+ - \max_i \{g_i(x^{\nu})\}_+ \right]
\]

\[
- \max_i \{g_i(x^{\nu})\}_+ + \frac{(\gamma^{nu})^T L_{\nabla g_i}}{2\varepsilon} \|d(x^{\nu})\|^2
\]

\[
\leq \gamma^{\nu} \nabla f(x^{\nu})^T d(x^{\nu}) + \frac{\gamma^{nu}^T L_{\nabla g_i}}{2} \|d(x^{\nu})\|^2
\]

\[
+ \frac{(\gamma^{nu})^T L_{\nabla f}}{2} \max_i \{L_{\nabla g_i}\} \|d(x^{\nu})\|^2
\]

\[
\leq \gamma^{\nu} \nabla f(x^{\nu})^T d(x^{\nu}) - \frac{\gamma^{nu}}{\varepsilon} \left[ \max_i \{g_i(x^{\nu})\}_+ - \kappa(x^{\nu}) \right]
\]

\[
\leq -c \|d(x^{\nu})\|^2 + \frac{\gamma^{nu}}{\varepsilon} \left[ \max_i \{g_i(x^{\nu})\}_+ - \kappa(x^{\nu}) \right]
\]

\[
\leq -c \|d(x^{\nu})\|^2 + \frac{\gamma^{nu}}{\varepsilon} \left[ \max_i \{g_i(x^{\nu})\}_+ - \kappa(x^{\nu}) \right]
\]

\[
\leq -c \|d(x^{\nu})\|^2 + (m \|\xi^{\nu}\|_\infty - \frac{1}{\varepsilon}) \left[ \max_i \{g_i(x^{\nu})\}_+ - \kappa(x^{\nu}) \right].
\]

(14)

where (a) follows applying the descent lemma to \(f\) and \(g_i\) for every \(i = 1, \ldots, m\), with \(L_{\nabla f}\) and \(L_{\nabla g_i}\) being the Lipschitz moduli of \(\nabla f\) and \(\nabla g_i\) on the bounded set containing all iterates; (b) holds for any positive \(\gamma^{\nu} \leq 1\) since, in view of (12), \(\nabla g_i(x^{\nu})^T d(x^{\nu}) \leq \kappa(x^{\nu}) - g_i(x^{\nu})\). Furthermore, we observe that

\[
\nabla f(x^{\nu})^T d(x^{\nu}) - \frac{\gamma^{nu}}{\varepsilon} \left[ \max_i \{g_i(x^{\nu})\}_+ - \kappa(x^{\nu}) \right]
\]

\[
\leq -c \|d(x^{\nu})\|^2 + \frac{\gamma^{nu}}{\varepsilon} \left[ \max_i \{g_i(x^{\nu})\}_+ - \kappa(x^{\nu}) \right]
\]

\[
\leq -c \|d(x^{\nu})\|^2 + (\sum_{j} m \|\xi^{\nu}\|_\infty - \frac{1}{\varepsilon}) \left[ \max_i \{g_i(x^{\nu})\}_+ - \kappa(x^{\nu}) \right]
\]

\[
\leq -c \|d(x^{\nu})\|^2 + (m \|\xi^{\nu}\|_\infty - \frac{1}{\varepsilon}) \left[ \max_i \{g_i(x^{\nu})\}_+ - \kappa(x^{\nu}) \right].
\]

(15)

where the first and the last inequalities are entailed by (11) and \(\kappa(x^{\nu}) \leq \max_i \{g_i(x^{\nu})\}_+\), respectively.

By (15), for any fixed \(x^{\nu}\) and for any \(\eta \in (0, 1]\), there exists \(\varepsilon^{\nu} > 0\) such that

\[
\nabla f(x^{\nu})^T d(x^{\nu}) - \frac{1}{\varepsilon} \left[ \max_i \{g_i(x^{\nu})\}_+ - \kappa(x^{\nu}) \right] \leq -\eta c \|d(x^{\nu})\|^2 \ \forall \varepsilon \in (0, \varepsilon^{\nu}].
\]

(16)

We now distinguish two cases.

(I) Suppose that (16) does not hold uniformly for every \(x^{\nu}\), that is \(\eta \in (0, 1]\), and subsequence \(\{\varepsilon^{\nu}\}_N \in \mathbb{R}_+ \) and \(\{x^{\nu}\}_N\) exist such that \(\varepsilon^{\nu} \downarrow 0\) on \(N\) and

\[
\nabla f(x^{\nu})^T d(x^{\nu}) - \frac{1}{\varepsilon} \left[ \max_i \{g_i(x^{\nu})\}_+ - \kappa(x^{\nu}) \right] \geq -\eta c \|d(x^{\nu})\|^2
\]

(17)

for every \(\nu \in N\). For (17) to hold, relying on (15), the multipliers’ subsequence \(\{\xi^{\nu}\}_N\) must be unbounded. Combining (15) and (17), we get

\[
0 \leq c(1 - \eta) \|d(x^{\nu})\| < \left( m \|\xi^{\nu}\|_\infty - \frac{1}{\varepsilon} \right) \left[ \max_i \{g_i(x^{\nu})\}_+ - \kappa(x^{\nu}) \right],
\]
and, thus, \(\max_i \{g_i(x^\nu) - \kappa(x^\nu)\} > 0\) for every \(\nu \in \mathcal{N}\). By the previous relation and (17), we also have
\[
\frac{1}{\varepsilon^\nu} < \frac{\nabla f(x^\nu)^T d(x^\nu) + \eta \varepsilon \|d(x^\nu)\|^2}{\max_i \{g_i(x^\nu)\} - \kappa(x^\nu)}.
\]
\(18\)
As \(\varepsilon^\nu \downarrow 0\) on \(\mathcal{N}\), the right hand side of (18) goes to infinity: by the boundedness of the numerator,
\[
\max_i \{g_i(x^\nu)\} - \kappa(x^\nu) \xrightarrow{\mathcal{N}} 0.
\]
\(19\)
Let \(\hat{x}\) be a cluster point of the subsequence \(\{x^\nu\}^\infty_{\nu=0}\). By (19), only cases (ii) and (iii) in Lemma 1 can occur. But in case (iii) it is not possible that \(\hat{d}\) exists such that \(\tilde{g}(\hat{d}, \hat{x}) < 0\). Indeed, by Proposition 5 this would entail the boundedness of the multipliers \(\xi^\nu\) for \(\nu \in \mathcal{N}\) large enough, thus giving a contradiction. Therefore, by Lemma 1 we conclude that \(\hat{x}\) is either an ES or FJ point for (P).

(II) As opposed to (I), consider the case in which relation (16) holds uniformly for every \(x^\nu\): that is, for any \(\eta \in (0,1]\), there exists \(\bar{\varepsilon} > 0\) such that
\[
\nabla f(x^\nu)^T d(x^\nu) - \frac{1}{\varepsilon^\nu} \left[ \max_i \{g_i(x^\nu)\} - \kappa(x^\nu) \right] \leq -\eta \varepsilon \|d(x^\nu)\|^2 \quad \forall \varepsilon \in (0, \bar{\varepsilon}], \forall x^\nu.
\]
\(20\)
Combining relations (14) and (20), we get
\[
W(x^{\nu+1}; \bar{\varepsilon}) - W(x^\nu; \bar{\varepsilon}) \leq -\gamma^\nu \eta \|d(x^\nu)\|^2 + \frac{(\gamma^\nu)^2}{2} \left( L_{\nabla f} + \frac{\max_i \{L_{\nabla g_i}\}}{\bar{\varepsilon}} \right) \|d(x^\nu)\|^2
\]
\[
= -\gamma^\nu \left[ \eta \gamma^\nu \left( L_{\nabla f} + \frac{\max_i \{L_{\nabla g_i}\}}{\bar{\varepsilon}} \right) \right] \|d(x^\nu)\|^2,
\]
\(21\)
for any \(\bar{\varepsilon} \in (0, \bar{\varepsilon}]\). Since \(\lim_{\nu \uparrow \infty} \gamma^\nu = 0\), there exists a positive constant \(\omega\) such that, by (21), for \(\nu \geq \bar{\nu}\) sufficiently large,
\[
W(x^{\nu+1}; \bar{\varepsilon}) - W(x^\nu; \bar{\varepsilon}) \leq -\omega \gamma^\nu \|d(x^\nu)\|^2.
\]
\(22\)
With \(W\) being bounded from below, by (22), the sequence \(\{W(x^\nu; \bar{\varepsilon})\}\) converges and
\[
\lim_{\nu \uparrow \infty} \sum_{t=\nu}^{\nu}\gamma^t \|d(x^t)\|^2 < +\infty.
\]
Therefore, since \(\sum_{\nu=0}^{\infty} \gamma^\nu = +\infty\), we have
\[
\lim_{\nu \uparrow \infty} \|d(x^\nu)\| = 0.
\]
\(23\)
Furthermore,
\[
0 \leq \max_i \{g_i(x^\nu)\} - \kappa(x^\nu) \leq \max_i \{g_i(x^\nu)\} - \max_i \{\tilde{g}_i(d(x^\nu); x^\nu)\}
\]
\[
\leq \max_i \{g_i(x^\nu)\} - \max_i \{(g_i(x^\nu) - \nabla g_i(x^\nu)^T d(x^\nu))\}
\]
\[
\leq \max_i \{(g_i(x^\nu) - g_i(x^\nu) - \nabla g_i(x^\nu)^T d(x^\nu))\} \leq \|\nabla g(x^\nu)^T d(x^\nu)\|_{\infty},
\]
where (a) holds since \(\tilde{g}(d(x^\nu); x^\nu) \leq \kappa(x^\nu)e\), and (b) is due to A5, A7 and A9. Taking the limit on a subsequence \(\mathcal{N}\) such that \(\|d(x^\nu)\| \xrightarrow{\mathcal{N}} 0\), we have \(\|\nabla g(x^\nu)^T d(x^\nu)\|_{\infty} \xrightarrow{\mathcal{N}} 0\) and
\[
\max_i \{g_i(x^\nu)\} - \kappa(x^\nu) \xrightarrow{\mathcal{N}} 0.
\]
\(24\)
Finally, let again \(\hat{x}\) be a cluster point of subsequence \(\{x^\nu\}^\infty_{\nu=0}\). Since (24) implies \(\kappa(\hat{x}) = \max_i \{g_i(\hat{x})\}\), cases (ii) or (iii) in Lemma 1 may occur: specifically, \(\hat{x}\) is either an ES, or a FJ, or a KKT point for
(P). In particular, if the eMFCQ holds at \( \hat{x} \), case (ii) in Lemma 1 is ruled out and \( \max \{ g_i(\hat{x})_+ \} \) cannot be strictly positive; then, \( \kappa(\hat{x}) = \max \{ g_i(\hat{x})_+ \} = 0 \). Furthermore, taking the limit in (8), we obtain, by A3, A4, A6-A9, multipliers’ boundedness and outer semicontinuity (relative to \( \theta \mathbb{R}_\infty^m \) and \( \mathbb{R}_\infty^m \), respectively) of the normal cone mapping (see e.g. [25, Proposition 6.6]),

\[
-\nabla f(\hat{x}) - \nabla g(\hat{x}) \hat{x} \in \mathbb{N}_{\theta \mathbb{R}_\infty^m}(0),
\]

with \( \hat{x} \in N_{\mathbb{R}_\infty^m}(g(\hat{x}) - \kappa(\hat{x})e) = N_{\mathbb{R}_\infty^m}(g(\hat{x})) \), and \( \hat{x} \) is a KKT point for problem (P). This concludes the proof of case (i).

Consider now point (ii). Note that if the eMFCQ holds at every limit point of \( \{ x^\nu \} \), then case (I) above cannot occur since this would contradict the last sentence before (II); hence, we are in case (II). Observe that if, instead of the weaker (23),

\[
\lim_{\nu \to \infty} \| d(x^\nu) \| = 0
\]

holds, we can reason similarly to what done above after (23) for any convergent subsequence of \( \{ x^\nu \} \), and conclude that (ii) holds. Therefore, it is enough to show that Assumption B entails (25).

Let \( K \) be a compact set containing all iterates \( x^\nu \). While \( \liminf_{\nu \to \infty} \| d(x^\nu) \| = 0 \), suppose by contradiction that \( \limsup_{\nu \to \infty} \| d(x^\nu) \| > 0 \). Then, there exists \( \delta > 0 \) such that \( \| d(x^\nu) \| > \delta \) and \( \| d(x^\nu) \| < \delta/2 \) for infinitely many \( \nu \) vs. Therefore, there is an infinite subset of indices \( \mathcal{N} \) such that, for each \( \nu \in \mathcal{N} \), and some \( i_\nu > \nu \), the following relations hold:

\[
\| d(x^\nu) \| < \delta/2, \quad \| d(x^{i_\nu}) \| > \delta
\]

and, if \( i_\nu > \nu + 1 \),

\[
\delta/2 < \| d(x^j) \| \leq \delta, \quad \nu < j < i_\nu.
\]

Hence, for all \( \nu \in \mathcal{N} \), we can write

\[
\frac{\delta}{2} < \| d(x^{i_\nu}) \| - \| d(x^\nu) \| \leq \| d(x^{i_\nu}) - d(x^\nu) \| \overset{(a)}{\leq} \| x^{i_\nu} - x^\nu \| = \theta \| x^{i_\nu} - x^\nu \|^\alpha
\]

\[
\leq \theta \left[ \sum_{t=\nu}^{i_\nu-1} \gamma^t \| d(x^t) \| \right]^{\alpha} \overset{(b)}{\leq} \theta \delta^\alpha \left( \sum_{t=\nu}^{i_\nu-1} \gamma^t \right)^\alpha, \tag{28}
\]

where (a) is due to Assumption B with \( \alpha \) and \( \theta \) positive scalars, (b) comes from the triangle inequality and the updating rule of the algorithm and in (c) we used (27). By (28) we have

\[
\liminf_{\nu \to \infty} \theta \delta^\alpha \left( \sum_{t=\nu}^{i_\nu-1} \gamma^t \right)^\alpha > 0. \tag{29}
\]

We prove next that (29) is in contradiction with the convergence of \( \{ W(x^\nu; \bar{\varepsilon}) \} \) for any suitable \( \bar{\varepsilon} \in (0, \bar{\varepsilon}] \). To this end, we first show that \( \| d(x^\nu) \| \geq \delta/4 \), for sufficiently large \( \nu \in \mathcal{N} \). Reasoning as in (28), we have

\[
\| d(x^{\nu+1}) \| - \| d(x^\nu) \| \leq \theta \| x^{\nu+1} - x^\nu \| = \theta (\gamma^\nu)^\alpha \| d(x^\nu) \|^\alpha,
\]

for any given \( \nu \). For \( \nu \in \mathcal{N} \) large enough so that \( \theta (\gamma^\nu)^\alpha (\delta/4)^\alpha < \delta/4 \), suppose by contradiction that \( \| d(x^\nu) \| < \delta/4 \); this would give \( \| d(x^{\nu+1}) \| < \delta/2 \) and, thus, condition (27) (or (26)) would be violated. Then, it must be \( \| d(x^\nu) \| \geq \delta/4 \). From this, and using (22), we have, for sufficiently large \( \nu \in \mathcal{N} \),

\[
W(x^{i_\nu}; \bar{\varepsilon}) \leq W(x^\nu; \bar{\varepsilon}) - \omega \sum_{t=\nu}^{i_\nu-1} \gamma^t \| d(x^t) \|^2 \leq W(x^\nu; \bar{\varepsilon}) - \omega \frac{\delta^2}{16} \sum_{t=\nu}^{i_\nu-1} \gamma^t. \tag{30}
\]

Since that \( \{ W(x^\nu; \bar{\varepsilon}) \} \) converges, renumbering if necessary, relation (30) implies \( \sum_{t=\nu}^{i_\nu-1} \gamma^t \to 0 \), in contradiction with (29). This shows that (25) holds and concludes the proof of the theorem. \( \square \)
The convergence properties in Theorem 1 (i) are very much in the spirit of analogous results for constrained optimization where no regularity conditions are made, see for example [6, 7, 8, 15]. A key difference between our approach and those in, e.g., [6, 7, 8, 15] is that we do not use any penalty parameter in the algorithm, and this is certainly advantageous both theoretically and numerically. Indeed, we use penalty function and penalty parameter only in the proof of Theorem 1, as a tool of theoretical analysis. We believe that this ghost penalty approach is a novelty in the literature and represents a new interesting use of penalty functions.

Remark 2. Algorithm 1 requires the exact solution of subproblem (Pν). However, it is standard and rather easy to show that approximate solutions of subproblems (Pν) could be used while still maintaining the convergence results in Theorem 1, provided that the accuracy of the solution increases as iterations progress.

Remark 3. Meaningful results are obtained in Theorem 1 if the sequence \(\{x^\nu\}\) is bounded. Although this is practically rather sensible, the question arises on when the sequence generated by the algorithm is bounded; can we give a priori conditions that guarantee the boundedness of the iterations? It is possible to give a satisfactory answer to these questions, at the price of a much more convoluted analysis; we eschewed this for the sake of simplicity of presentation. But let us at least hint here at two approaches that can be used to get iterates bounded.

A first easy case is when the original problem (P) includes a constraint of the type \(x \in X\), where X is a (typically simple) compact, convex set. For example, X could define upper and lower bounds an all variables. It is easy to show that all results in this and in the next section still hold if we redefine subproblem (3) by adding the constraint X:

\[
\min_d \tilde{f}(d; x^\nu) \quad \text{s.t.} \quad \tilde{g}(d; x^\nu) \leq \kappa(x^\nu) \varepsilon, \quad \|d\|_\infty \leq \beta, \quad x^\nu + d \in X.
\]

If one requires the algorithm to start from an initial point \(x^0 \in X\), this simple strategy obviously guarantees the boundedness of the iterations, which all belong to the compact set X, since \(\gamma^\nu \in (0, 1]\) and X is convex; see [6] for results in this vein.

Another possibility is to use, as ghost penalty, not W as defined in (13), but a nondifferentiable penalty function with barriers, see e.g. [13, 14, 15, 19]. Let \(\mu\) be a positive constant and consider the set \(S \triangleq \{x : \|g(x)\|_3^3 < \mu\}\), which is an enlargement of the original feasible set of Problem (P). On this set we can define the constraints \(\tilde{g}_i(x) \triangleq \frac{1}{\mu - \|g_i(x)\|}g_i(x)\). It is clear that

\[
\min_x f(x) \quad \text{s.t.} \quad \tilde{g}(x) \leq 0,
\]

is equivalent to the original problem in the sense that it has the same feasible region (although defined through different constraints) and objective function. Furthermore, it can be checked that this problem has exactly the same stationary points of the original one. The penalty function with barriers \(P(x; \varepsilon)\) is the usual penalty function considered so far, but for this modified problem:

\[
\tilde{P}(x; \varepsilon) = f(x) + \frac{1}{\varepsilon}\|\tilde{g}(x)\|.
\]

Clearly, \(\tilde{P}(x; \varepsilon)\) goes to infinity when x approaches the boundary of S. Therefore, if one assumes that S is bounded, the level sets of \(\tilde{P}\) are guaranteed to be compact (on S) for every \(\varepsilon\), and this fact can be used to show boundedness of the iterates of (a simple modification of) Algorithm 1.

Remark 4. We finally note that all the developments in the proof of Theorem 1 up to equation (21) are valid independent of the updating rule for the stepsize \(\gamma^\nu \in (0, 1]\). In the light of this observation, in the next section we invoke some of the relations in the proof of Theorem 1 even when stepsizes not satisfying (2) are employed.
If the eMFCQ holds on a compact set containing all the iterates produced by the algorithm, we can use a sufficiently small, but constant stepsize instead of a diminishing one, and still guarantee that all limit points are KKT solutions for problem (P).

**Corollary 1.** Let $f$ and $g$ be LC$^4$ and suppose that Assumption A holds. Consider the sequence \( \{x^\nu\} \) generated by Algorithm 1 where a fixed stepsize: \( \gamma^\nu = \gamma > 0 \) is used at all iterations $\nu$. Suppose that the eMFCQ holds at any limit point of \( \{x^\nu\} \). Then, there exists a positive $\bar{\gamma}$ such that, if $\gamma \leq \bar{\gamma}$, either \( \{x^\nu\} \) is unbounded or every limit point of \( \{x^\nu\} \) is a KKT point of problem (P).

Proof. Under the eMFCQ, case (I) in the proof of Theorem 1 does not occur (see comments at the beginning of point (ii)). Hence, (20) holds for every $x^\nu$ and some fixed $\bar{\varepsilon}$. It is then easy to see that, if we take $\gamma^\nu = \gamma$, with

$$
\gamma \leq \bar{\gamma} < \frac{2\eta c}{\left( L_{\nabla f} + \max_{\nu} \{L_{\nabla g}\} \right)},
$$

relation (22) still holds, the sequence \( \{W(x^\nu; \bar{\varepsilon})\} \) converges and $\lim_{\nu \rightarrow \infty} \|d(x^\nu)\| = 0$. Hence, along the same line of reasoning in the proof of case (II), every limit point is a KKT solution for problem (P). \( \square \)

**5. Complexity Analysis** In this section we derive iteration complexity results for Algorithm 1 under different updating rules for $\gamma^\nu$. As the problem is nonconvex and constrained, we follow [10] and [11] in defining the iteration complexity to be the maximum number of iterations required to find a $\delta$–stationary point, i.e. a point where a suitable stationarity measure is below a certain desired threshold $\delta$. In subsection 5.1 we consider the “good” case in which the eMFCQ holds. In this circumstance, all stationary solutions are KKT points for problem (P) and, as in classical SQP methods, the norm of the direction $d(x^\nu)$ is a natural stationarity measure, see Lemma 2. In subsection 5.2 we study what happens when we drop the eMFCQ. In this case a more complex approach must be adopted since the problem may admit KKT points but also FJ and external stationary solutions and we use in combination $\|d(x^\nu)\|$ and a measure of stationarity for the feasibility problem (4).

**5.1. The eMFCQ holds** We make the following assumptions.

**Assumption D**

D1) Each iterate $x^\nu$ belongs to a compact set $K$;

D2) $\nabla_1 \tilde{g}(\bullet; \bullet)$ is locally Lipschitz continuous on $O_d \times K$;

D3) eMFCQ holds at every $x^\nu \in K$.

Assumption D2 depends essentially on the choice of $\tilde{g}$. Clearly, if $g$ has a locally Lipschitz gradient, this assumption is always satisfied if we take as $\tilde{g}$ the linearization of $g$. Assumption D1 is made, once again, for simplicity of presentation; we refer the reader to Remark 3 for comments on how this assumption can be enforced. Condition D3 is the key requirement. In fact, it implies that the MFCQ holds at every point in $\tilde{X}(x^\nu)$, see Proposition 3 (i). In turn, by Proposition 5, the set-valued mapping of KKT multipliers of the subproblems $(P_\nu)$ is locally bounded and therefore, thanks to D1, also globally bounded on the sequence $\{x^\nu\}$ generated by Algorithm 1. In this section we denote by M this global bound. From now on, we employ problem dependent constants, some new, some already introduced; we collect their definitions in Table 2 for reader convenience.

Lemma 2 allows us to relate the KKT conditions of (P) to $\|d(x^\nu)\|$, showing that $\|d(x^\nu)\|$ can be used as stationarity measure. This can be viewed as a generalization of classical results in SQP methods. Note that the lemma holds whatever the choice of $\gamma^\nu$.

**Lemma 2.** Let $\{x^\nu\}$ be the sequence generated by Algorithm 1 under Assumptions A, C and D. Then, for every $x^\nu$, we have

$$
\|\nabla f(x^\nu) + \nabla g(x^\nu)x^\nu\| \leq \left[ L_{\nabla f} + \left( L_{\nabla g} + \frac{1}{\beta} \right) M \right] \|d(x^\nu)\|, \tag{32}
$$
Regarding feasibility, we observe that, letting $\xi^\nu \in N_{\partial g_i}(d(x^\nu); x^\nu) - \kappa(x^\nu)e$, we have

$$
\|\nabla f(x^\nu) + \nabla g(x^\nu)\xi^\nu\| = \|\nabla \tilde{f}(0; x^\nu) - \nabla \tilde{f}(d(x^\nu); x^\nu) + \nabla \tilde{f}(d(x^\nu); x^\nu)
+ \nabla \tilde{g}(d(x^\nu); x^\nu)\xi^\nu + \nabla \tilde{g}(0; x^\nu)\xi^\nu - \nabla \tilde{g}(d(x^\nu); x^\nu)\xi^\nu\|
\leq L_{\nabla \tilde{f}}\|d(x^\nu)\| + L_{\nabla \tilde{g}}\|\xi^\nu\||\|d(x^\nu)\| + \|\xi^\nu\|,
$$

(35)

for some $\xi^\nu \in N_{\partial g_i}(d(x^\nu))$, where the equality is due to A4 and A9 and the inequality follows from (8) (that of course still holds in the setting of this theorem). Since $\xi^\nu = 0$ whenever $\|d(x^\nu)\|_\infty < \beta$, consider $d(x^\nu)$ such that $\|d(x^\nu)\|_\infty = \beta$: by (35),

$$
\|\nabla f(x^\nu) + \nabla g(x^\nu)\xi^\nu\| \leq L_{\nabla \tilde{f}}\|d(x^\nu)\| + L_{\nabla \tilde{g}}\|\xi^\nu\||\|d(x^\nu)\| + \|\xi^\nu\| \leq L_{\nabla \tilde{f}}\|d(x^\nu)\| + L_{\nabla \tilde{g}}\|\xi^\nu\||\|d(x^\nu)\| + \frac{1}{\beta}\|\xi^\nu\||\|d(x^\nu)\|.
$$

(36)

Regarding feasibility, we observe that, letting $i \in \{1, \ldots, m\}$ such that $0 < g_i(x^\nu) = \max_i\{g_i(x^\nu)_+\}$,

$$
\max_i\{g_i(x^\nu)_+\} - \kappa(x^\nu) \leq \max_i\{g_i(x^\nu)_+\} - \max_i\{\tilde{g}_i(d(x^\nu); x^\nu)_+\}
\leq \tilde{g}_i(0; x^\nu) - \tilde{g}_i(d(x^\nu); x^\nu) \leq L_{\tilde{g}}\|d(x^\nu)\|,
$$

where the inequalities are due to the feasibility of $d(x^\nu)$ and to A7, respectively. Then,

$$
\max_i\{g_i(x^\nu)_+\} - \kappa(x^\nu) = \lambda \left[ \max_i\{g_i(x^\nu)_+\} - \min_i \left\{ \max_i\{\tilde{g}_i(d; x^\nu)_+\} | \|d\|_\infty \leq \rho \right\} \right] \leq L_{\tilde{g}}\|d(x^\nu)\|.
$$

(37)
We now show that, for any $x^r$, a positive constant $a$ exists such that
\[
\min_d \left\{ \max_i \{\bar{g}_i(d; x^r)_+\} \mid \|d\|_\infty \leq \rho \right\} \leq a \|d(x^r)\|. \tag{38}
\]
Suppose on the contrary that subsequences $\{a^r\}_N \in \mathbb{R}_+$ and $\{x^r\}_N$ exist such that $a^r \to +\infty$ and
\[
\min_d \left\{ \max_i \{\bar{g}_i(d; x^r)_+\} \mid \|d\|_\infty \leq \rho \right\} > a^r \|d(x^r)\|. \tag{39}
\]
Relation (39) implies
\[
\min_d \left\{ \max_i \{\bar{g}_i(d; x^r)_+\} \mid \|d\|_\infty \leq \rho \right\} > 0. \tag{40}
\]
Observing that, by A7, $\max_i \{\bar{g}_i(x^r)_+\} \geq \min_i \{\bar{g}_i(d; x^r)_+\} \|d\|_\infty \leq \rho > 0$, thanks to assumption D3 and in view of Lemma 1, we have $\kappa(x^r) < \max_i \{\bar{g}_i(x^r)_+\}$ and $\|d(x^r)\| \neq 0$ for every $\nu \in N$. Moreover, since $\min_i \{\bar{g}_i(d; \bullet)_+\} \|d\|_\infty \leq \rho$ and $d(\bullet)$ are continuous on $\mathbb{R}^n$ and $K$, respectively, we have, renumbering if necessary, $\|d(x^r)\| \to ||d(\hat{x})|| = 0$, with $\hat{x}$ cluster point of subsequence $\{x^r\}_N$. Resorting again to Lemma 1, $\kappa(\hat{x}) = \max_i \{\bar{g}_i(\hat{x})\}_+$ and $\hat{d} \in \rho \mathbb{B}^n_\infty$ exists such that $\bar{g}_i(\hat{d}; \hat{x}) < 0$. 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 $\nu \in N$ sufficiently large, we have $\bar{g}_i(d; x^r) < 0$, in contradiction to (40). Hence, (38) holds.

Combining (37) and (38), we get
\[
\max_i \{g_i(x^r)_+\} \leq \frac{L_\beta}{\lambda} \|d(x^r)\| + \min_d \left\{ \max_i \{\bar{g}_i(d; x^r)_+\} \mid \|d\|_\infty \leq \rho \right\} \leq \left( \frac{L_\beta}{\lambda} + a \right) \|d(x^r)\|. \tag{41}
\]
Concerning the complementarity condition, letting $\max_i |g_i(x^r)\xi^r_\nu| = |g_i(x^r)\xi^r_\nu| > 0$ for some $i \in \{1, \ldots, m\}$ such that $\bar{g}_i(d(x^r); x^r) = \kappa(x^r)$, we have, in view of A7 and (41),
\[
0 \leq \kappa(x^r) = \bar{g}_i(d(x^r); x^r) - \bar{g}_i(0; x^r) + \bar{g}_i(0; x^r) = L_\beta \|d(x^r)\| + g_i(x^r) \leq L_\beta \|d(x^r)\| + \max_i \{g_i(x^r)_+\} \leq \left( L_\beta \frac{1+a}{\lambda} + a \right) \|d(x^r)\|
\]
and, thus,
\[
|g_i(x^r)\xi^r_\nu| = |\bar{g}_i(0; x^r) - \bar{g}_i(d(x^r); x^r) + \bar{g}_i(d(x^r); x^r)| |\xi^r_\nu| \leq (L_\beta \|d(x^r)\| + |g_i(d(x^r); x^r)|) |\xi^r_\nu| \leq \left( L_\beta \frac{1+a}{\lambda} + a \right) |\xi^r_\nu| \|d(x^r)\|. \tag{42}
\]
We can now invoke the local boundedness of the set of KKT multipliers and, thus, in addition to (41), by (36) and (42), we get
\[
\|\nabla f(x^r) + \nabla g(x^r)\xi^r_\nu\| \leq \left( L_{\nu f} + L_{\nu g} M + \frac{M}{\lambda} \right) \|d(x^r)\|\max_i |g_i(x^r)\xi^r_\nu| \leq \left( L_{\nu g} \frac{1+a}{\lambda} + a \right) M \|d(x^r)\| \tag{43}
\]
where $M$ is the bound on the multipliers defined just before the statement of the Lemma.

With bounds (41) and (43) in mind, we can give the main complexity result of this subsection. Preliminarily, we remark that, by (32)-(34),
\[
\|d(x^r)\| \leq \delta \implies \max \left\{ \|\nabla f(x^r) + \nabla g(x^r)^T \xi^r_\nu\|, \max_i |g_i(x^r)_+|, \max_i |g_i(x^r)\xi^r_\nu| \right\} \leq b\delta, \tag{44}
\]
where $b$ is defined in Table 2. Note that the condition $\|d(x^r)\| \leq \delta$ can easily be checked at each iteration and can be very naturally used as stopping criterion in the algorithm; this requires no knowledge of the constant $b$. Of course, if we do not know the constant $b$, we cannot say exactly how much the KKT conditions are violated when $\|d(x^r)\| \leq \delta$, but the meaning of (44) is to show that the smaller the $\delta$ the less the KKT conditions are violated and that, in the limit, when $\delta$ goes to zero the violation of the KKT conditions also goes to zero. Consistent with these observations and with the analysis of classical SQP methods, in the theorem below we analyze how many iterations are needed to drive $\|d(x^r)\|$ below $\delta$. We consider both the cases of constant and diminishing stepsizes.
Theorem 2. Let \( \{x^\nu\} \) be the sequence generated by Algorithm 1 under Assumptions A, C and D. Then, in at most \( N \) iterations, Algorithm 1 drives the criticality measure \( \|d(x^\nu)\| \) below \( \delta \), where

(i) if \( \gamma^\nu = \gamma \), with \( \gamma \) satisfying (31),

\[
N = \left\lceil \frac{\|W^0 - W^m\|}{\gamma \omega \delta^2} \right\rceil; \tag{45}
\]

(ii) if a diminishing stepsize with sufficiently small \( \gamma^0 \) is employed, \( N \) is the first iteration index for which

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

(iii) whenever any diminishing stepsize is employed, \( N \triangleq \bar{\nu} + \bar{N} \) is the first iteration index for which

\[
\sum_{\nu=\bar{\nu}}^{\bar{\nu} + \bar{N} - 1} \gamma^\nu \geq \frac{\|W^M - W^m\|}{\omega \delta^2}, \tag{46}
\]

for some suitable \( \bar{\nu} \), independent of \( \delta \), and \( \bar{N} \) (we assume without loss of generality that \( N > \bar{\nu} \)).

Proof. We preliminarily recall that by Remark 4, since the eMFCQ is assumed to hold, we can freely invoke (20), (21), and (22) (by Corollary 1). Let \( \bar{\varepsilon} \) be the value of the penalty parameter in (20); note that, under assumptions D1 and D3, one can take \( \bar{\varepsilon} = \frac{1}{M} \).

(i) If one employs a sufficiently small positive stepsize \( \gamma^\nu = \gamma \), more precisely if \( \gamma \) satisfies (31), we have \( \|d(x^\nu)\|^2 \leq [W(x^\nu; \bar{\varepsilon}) - W(x^{\nu+1}; \bar{\varepsilon})]/\omega \) for every \( \nu \). Taking the sum of iterations up to \( N - 1 \), and supposing that \( \|d(x^\nu)\| > \delta \) for all \( \nu \in \{0, \ldots, N - 1\} \), we have

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

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 \( K \). Therefore, our procedure drives the criticality measure \( \|d(x^\nu)\| \) below \( \delta \) in a number of iterations equal to \( \left\lceil \frac{\|W^0 - W^m\|}{\gamma \omega \delta^2} \right\rceil \), at most.

(ii) Whenever the diminishing stepsize procedure is such that \( \gamma^0 \) is sufficiently small, more precisely if it satisfies (31), we have \( \gamma^\nu \|d(x^\nu)\|^2 \leq [W(x^\nu; \bar{\varepsilon}) - W(x^{\nu+1}; \bar{\varepsilon})]/\omega \) for every \( \nu \); supposing that \( \|d(x^\nu)\| > \delta \) for all iterates up to \( N - 1 \), we have

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

In this case the maximum number of iterations required to have \( \|d(x^\nu)\| \) below \( \delta \) is \( N \) such that

\[
\sum_{\nu=0}^{N-1} \gamma^\nu \geq \frac{\|W^0 - W^m\|}{\omega \delta^2}.
\]

(iii) When considering a generic diminishing stepsize procedure, a finite \( \bar{\nu} \) exists such that, for every \( \nu \geq \bar{\nu} \), the descent condition (22) holds; the number of iterations \( \bar{\nu} \) is problem-dependent and relies on initial algorithmic choices such as the updating rule for the diminishing stepsize.

Summing \( \gamma^\nu \|d(x^\nu)\|^2 \leq [W(x^\nu; \bar{\varepsilon}) - W(x^{\nu+\bar{N}}; \bar{\varepsilon})]/\omega \) from \( \bar{\nu} \) up to \( N = \bar{\nu} + \bar{N} - 1 \), and considering \( \|d(x^\nu)\| > \delta \) all iterates up to \( \bar{\nu} + \bar{N} - 1 \), we have

\[
\delta^2 \sum_{\nu=\bar{\nu}}^{\bar{\nu} + \bar{N} - 1} \gamma^\nu < \sum_{\nu=\bar{\nu}}^{\bar{\nu} + \bar{N} - 1} \gamma^\nu \|d(x^\nu)\|^2 \leq \frac{W(x^{\bar{\nu}}; \bar{\varepsilon}) - W(x^{\bar{\nu} + \bar{N}}; \bar{\varepsilon})}{\omega} \leq \frac{W^M - W^m}{\omega},
\]
where $W^M$ is the maximum value attained by the continuous function $W(x; \tilde{\varepsilon})$ on the compact set $K$. Reasoning as above, we get (46).

In case (i), in view of relation (45), the smaller the constant stepsize $\gamma$ taken, the bigger the number of iterations needed to find a $\delta$-approximate KKT solution of problem (P). Hence, in view of relation (31), the best one can do is to set in (31) $\tilde{\varepsilon} = \frac{1}{M}$, thus obtaining the (possibly over)estimate

$$N = \left[ \frac{W^0 - W^m}{2\eta c \left( L_{\nabla f} + M \max_i \{ L_{\nabla g_i} \} \right) \omega \delta^2} \right].$$

The analysis for case (ii) depends on the particular updating rule that is adopted. Suppose, for example, that one relies on the classical generic term of the harmonic series and sets $\gamma^\nu = \frac{0}{\nu + 1}$. Observing that $\sum_{\nu=0}^{N-1} \gamma^\nu = \sum_{\nu=0}^{N-1} \frac{0}{\nu+1} = \sum_{\nu=1}^{N} \frac{0}{\nu}$ and $\gamma^0 \ln(N + 1) < \sum_{\nu=1}^{N} \frac{0}{\nu} < \gamma^0 [\ln N + 1]$, we get

$$\left[ \exp \left( \frac{[W^0 - W^m]}{\gamma^\nu \omega \delta^2} - 1 \right) \right] < N < \left[ \exp \left( \frac{[W^0 - W^m]}{\gamma^\nu \omega \delta^2} \right) \right],$$

which clearly entails a deterioration in the worst case iteration performance with respect to a constant stepsizes approach. On the other hand, the practical behavior of a diminishing strategy is often much better than the one with a constant stepsize. Most of the times, choosing a theoretical sound constant stepsize leads to extremely small stepsizes, making progress to the solution painfully slow, while a diminishing approach permits the use of much larger stepsizes in the early stages, often leading to faster algorithms, see e.g. [16, 26].

Case (iii) is a trivial variant of case (ii). It is clear that if we use a diminishing stepsize rule then, sooner or later an iteration $\tilde{\nu}$ occurs for which $\gamma^{\tilde{\nu}}$ satisfies the condition in point (ii) and we can apply the results in (ii) 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 small” value, given by the r.h.s. of (31), that should be used in (i) and (ii). For example, it is easy to see that when the generic term of the harmonic series $\frac{0}{\nu + 1}$ is employed, we can take

$$\tilde{\nu} = \left[ \frac{\gamma^0 L_{\nabla f} + M \max_i \{ L_{\nabla g_i} \} }{2\eta c} \right] - 1.$$

5.2. The eMFCQ need not hold
If we drop assumption D3, we cannot rely solely on $\|d(x^\nu)\|$ to monitor progressions towards stationarity, since we could be converging to an ES or a FJ point. We then resort also to the nonnegative continuous function $\theta(x^\nu) \triangleq \max_i \{ g_i(x^\nu)_+ \} - \kappa(x^\nu)$, for which we established the bound

$$\theta(x^\nu) \leq L_3 \|d(x^\nu)\|$$

in the proof of Lemma 2, see (37). However, there is no reverse implication and thus the two functions $\|d(x^\nu)\|$ and $\theta(x^\nu)$ must be suitably combined to provide a reliable stopping criterion.

We remark that for relation (47) to hold, assumptions D2 and D3 are not invoked at all. Also, $\theta(x^\nu)$ is a stationarity measure for the feasibility problem (4) since it is equal to zero if and only if $x^\nu$ is stationary for (4) (see the proof of point (ii) in Lemma 1). The effect of suitably monitoring both $\|d(x^\nu)\|$ and $\theta(x^\nu)$ on the outcome of the algorithm is analyzed after Theorem 3.

To derive complexity results in this hard case, we consider Algorithm 2 with a piecewise constant choice of stepsizes. By this we mean that Algorithm 2 starts with a certain $\gamma^{-1}$ and keeps it fixed until a certain test is met; when this happens, the stepsize is reduced to a new, prescribed value and then kept fixed until possibly the test is met again, and so on. We underline that the only difference between this scheme and Algorithm 1 is in the rules for choosing $\gamma^\nu$ at each iteration and, of course, in the presence of suitable stopping criteria: specifically, the steps (S.1) and (S.7) correspond to the
previous Algorithm 1, while everything in between, from (S.2) to (S.6), is aimed at deciding whether

to decrease the stepsizes \( \gamma^* \) and whether we should stop (note that Algorithm 1, which was aimed

at an asymptotic analysis, did not contain any practical stopping criterion).

**Algorithm 2:** Modified Algorithm for (P)

| Data: \( \delta > 0, x^0, T^{-1} \in \left(0, \frac{2\max(L, \sqrt{\nu})}{\max(L, \nu)}\right], \gamma^{-1} = \frac{\nu}{2\max(L, \nu)}, \nu \leftarrow 0; \) |
| repeat |
| (S.1) compute \( \kappa(x^\nu) \), the solution \( d(x^\nu) \) of problem \( (P_{x^\nu}) \) and \( \theta(x^\nu) \); |
| (S.2) if \( \|d(x^\nu)\| < \delta \) then |
| \( x_\delta = x^\nu \); |
| end |
| (S.3) if \( \nabla f(x^\nu)^T d(x^\nu) + \eta \|d(x^\nu)\|^2 > 0 \) and \( T^{\nu-1} > \frac{\theta(x^\nu)}{\nabla f(x^\nu)^T d(x^\nu) + \eta \|d(x^\nu)\|^2} \) then |
| \( x_\delta = x^\nu \); |
| end |
| (S.4) if \( \theta(x^\nu) \leq \delta \) then |
| \( x_\delta = x^\nu \); |
| else |
| set \( \gamma^\nu = \frac{\nu}{\max(L, \sqrt{\nu})} \), where \( T^{\nu} = \frac{1}{\eta} \frac{\theta(x^\nu)}{\nabla f(x^\nu)^T d(x^\nu) + \eta \|d(x^\nu)\|^2} \); |
| end |
| (S.5) set \( x^{\nu+1} = x^\nu + \gamma^\nu d(x^\nu) \), \( \nu \leftarrow \nu + 1; \) |
| end |

We first note that the value of \( T^{-1} \) guarantees that \( \gamma^{-1} \leq 1 \). A second observation is about the

stopping tests: we have two of them, in steps (S.2) and (S.4). The test at (S.2) is the same used in

the previous subsection. Here, to take into account the lack of the eMFCQ, we also resort to the

additional test in (S.4). We discuss in detail the meaning of these stopping criteria after Theorem

3. The variable \( T^{\nu} \) is introduced just for notational purposes, in order to make the statement of

the algorithm and the proof of Theorem 3 easier to follow. The tests we must perform to decide

whether to reduce the stepsizes are very simple and involve quantities that are readily available once

the direction finding subproblem \( (P_{x^\nu}) \) has been solved.

The following theorem provides the announced complexity result in this general case. For sim-

plicity of presentation we assume \( \delta \leq 1 \). This is by no means necessary but avoids the necessity to

complicate the statement by considering uninteresting cases; in any event, see also Remark 5 for

further discussion on this point.

**Theorem 3.** Let \( \{x^\nu\} \) be the sequence generated by Algorithm 2 under Assumptions A, C2 and

d1. Then, letting \( \delta \leq 1 \), in at most \( \mathcal{O}(\delta^{-4}) \) iterations, Algorithm 2 stops either at step (S.2) or at

step (S.4).

**Proof.** Suppose that Algorithm 2 performs \( N \) iterations without stopping\(^1\). We first count how

many times \( \gamma^\nu \) can be updated in step (S.5) of the algorithm: let

\[
\mathcal{I} \triangleq \{0 < \nu_1 \leq N \mid T^{\nu_1} \text{ and } \gamma^{\nu_1} \text{ are updated in (S.5)} \} \cup \{0\}
\]

be the set of iterations’ indices \( \nu \) (in increasing order) at which the need to modify \( \gamma^\nu \) and \( T^{\nu} \)

emerges, union iteration 0. Therefore, for example, if we update \( T \) and \( \gamma \) in (S.5) at iterations 3, 4 and 8, we have \( \mathcal{I} = \{0, 1, 3, 4, 5, 8\} \); note that we always have by definition \( \nu_0 = 0 \)

\(^1\) We consider an iteration completed when we reach (S.7).
and that the set $\mathcal{I}$ does not include repeated indices. We show that $\mathcal{I}$ has finite cardinality. If $\nu_i \neq 0$ belongs to $\mathcal{I}$, we have

$$T^{\nu_i} = \frac{1}{2} \nabla f(x^{\nu_i})^T d(x^{\nu_i}) + \eta c \|d(x^{\nu_i})\|^2,$$

(48)

and the procedure did not stop at step (S.4); thus, $\theta(x^{\nu_i}) > \delta$ and (48) entail

$$T^{\nu_i} > \frac{\delta}{2B},$$

(49)

where $B \triangleq \max_x \{\|\nabla f(x)\|\beta + \eta c \beta^2 \mid x \in K\} \geq \nabla f(x^{\nu})^T d(x^{\nu}) + \eta c \|d(x^{\nu})\|^2$. By the updating rule in (S.5), we also have $T^{\nu_i} \leq \frac{T^{\nu_i} - \max_{\nu} \{L_{v_{\nu_i}}\}}{2} = \frac{T^{\nu_i} - \max_{\nu} \{L_{v_{\nu_i}}\}}{2}$; thus, in view of (49), $\frac{\delta}{2B} < T^{\nu_i} \leq \frac{T^{\nu_i} - \max_{\nu} \{L_{v_{\nu_i}}\}}{2}$, so that

$$i < \log_2 \frac{T^{-1}2B}{\delta}.$$

Therefore, if we do not stop, i.e., if $\theta(x^{\nu_i}) > \delta$ for all iterations up to $N - 1$, the cardinality of $\mathcal{I}$, i.e., the times $\gamma^{\nu}$ is reduced, is at most $\left\lceil \log_2 \frac{T^{-1}2B}{\delta} \right\rceil$.

Let us set $I \triangleq |\mathcal{I}| - 1$; with this convention note that the largest element in $\mathcal{I}$ is $\nu_i$. Counting from $\nu_i \in I \setminus \{\text{last element in } I\}$, let now $N_i$ be the number of iterations in which $\gamma^{\nu}$ remains unchanged: $T^{\nu} = T^{\nu_i}$ and $\gamma^{\nu} = \gamma^{\nu_i}$ for every $\nu \in \{\nu_i, \ldots, \nu_i + N_i\}$. In other words, $N_i$ is the number of iterations after $\nu_i$ in which step (S.5) is not reached; in the example given above where $\mathcal{I} = \{\nu_0 = 0, \nu_1 = 3, \nu_2 = 4, \nu_3 = 8\}$, we have $N_0 = 2, N_1 = 0, N_2 = 3$. Therefore $\nu_i + N_i$ is simply the last iteration after $\nu_i$ before $\gamma$ and $T$ are updated. The last index $N_1$ is defined, with the same rationale, as the number of iterations performed after $\nu_i$, before we reach the iteration where we stop. Considering the example above, and supposing that we stop at iteration 11, we have $N_3 = 2$.

We observe that, by virtue of the condition in step (S.3) and the updating rule in step (S.5) or (S.6), $T^{\nu}$ is non increasing. Hence, again by the updating rule in (S.5) or (S.6), since $\gamma^{-1} = \frac{T^{-1}2B}{\max_{\nu} \{L_{v_{\nu_i}}\}}$, also $\gamma^{\nu}$ is non increasing. Moreover, by the definitions of $T^{-1}$ and $\gamma^{-1}$, on the one hand,

$$\gamma c - \frac{\gamma^{\nu}}{2} L_{v_{\nu_i}} \geq \gamma c - \frac{\gamma^{\nu}}{2} L_{v_{\nu_i}} \geq \gamma c - \frac{\gamma^{\nu}}{2} \gamma c,$$

while, on the other hand,

$$\gamma c - \frac{\gamma^{\nu}}{2} L_{v_{\nu_i}} \geq \gamma c - \frac{\gamma^{\nu}}{2} \gamma c - \frac{\gamma^{\nu}}{2} \gamma c.$$

Thanks to the previous relations, we have for every $\nu$

$$\eta c - \frac{\gamma^{\nu}}{2} \left(L_{v_{\nu_i}} + \frac{\max_{\nu} \{L_{v_{\nu_i}}\}}{T^{\nu_i}}\right) \geq \frac{\eta c}{4}.$$

(50)

For every $\nu \in \{\nu_i, \ldots, \nu_i + N_i\}$, $\nu_i \in I$, $\gamma^{\nu}$ is not reduced and therefore

$$\nabla f(x^{\nu})^T d(x^{\nu}) + \eta c \|d(x^{\nu})\|^2 \leq 0 \quad \text{or} \quad T^{\nu} = T^{\nu_i} \leq \frac{\theta(x^{\nu})}{\nabla f(x^{\nu})^T d(x^{\nu}) + \eta c \|d(x^{\nu})\|^2}.$$

(51)

Thus, by (51), we also have

$$\nabla f(x^{\nu})^T d(x^{\nu}) - \frac{\theta(x^{\nu})}{T^{\nu_i}} \leq -\eta c \|d(x^{\nu})\|^2.$$

In turn, (14), by (50), entails

$$W(x^{\nu+1}; T^{\nu_i}) - W(x^{\nu}; T^{\nu_i}) \leq -\gamma^{\nu} \frac{\eta c}{4} \|d(x^{\nu})\|^2,$$

(52)

for every $\nu \in \{\nu_i, \ldots, \nu_i + N_i\}$, $\nu_i \in I$ and where we took $\epsilon^{\nu} = T^{\nu_i}$. 

\]
Note that \( N = \sum_{i \in I} (N_i + 1) \), since the algorithm did not stop until iteration \( N \), \( \|d(x^\nu)\| > \delta \) for all iterates up to \( N - 1 \). Therefore, recalling definition (13) with \( \varepsilon = T^\nu \), we get
\[
\delta^2 N = \sum_{i=0}^I \delta^2 (N_i + 1) < \sum_{i=0}^I \sum_{\nu_i + N_i} \|d(x^\nu)\|^2 \leq \frac{1}{\gamma^\nu} \sum_{i=0}^I \frac{W(x^\nu; T^\nu_i) - W(x^\nu + N_i; T^\nu_i)}{\gamma^\nu} \\
\leq \frac{1}{\gamma^\nu} \left[ f(x^\nu) - f(x^\nu + N_i^\nu) + \frac{1}{2\sigma} \max_i \{g_i(x^\nu)\}_+ \right] \\
- \frac{1}{\nu^\nu} \left[ g_i(x^\nu + N_i^\nu) + \frac{1}{2\sigma} \max_j \{g_j(x^\nu)\}_+ \right] + \frac{1}{T^\nu_i} - \frac{1}{T^\nu_{i-1}} \max_j \{g_j(x^\nu)\}_+ \right]}
\]
where the second inequality is due to (52) while, observing that \( \gamma^\nu_i \leq \gamma^\nu_i \), the last inequality is valid as a result of a telescopic series argument since \( \nu_i + N_i + 1 = \nu_i + 1 \). Note that it is understood that if \( I = 0 \) the last summation in (53) has no terms. Letting \( f^M \triangleq \max_x \{f(x) \mid x \in K\} \), \( g^M_+ \triangleq \max_x \{\max_i \{g_i(x)\}_+ \mid x \in K\} \) and \( f^m \triangleq \min_x \{f(x) \mid x \in K\} \), by (53) we can write
\[
\delta^2 N < \frac{1}{\gamma^\nu} \left[ f^M - f^m + \frac{1}{2\sigma} g^M_+ - \frac{1}{2\sigma} g^M_+ + \frac{1}{2\sigma} g^M_+ \right] \leq \frac{8}{(\eta e)^2 T^{-1}} \max_i \{L_{\nabla g_i}(f^M - f^m + \frac{1}{T^{-1}} g^M_+) \}
\]
where (a), since \( T^\nu_i \leq T^\nu_{i-1} \), follows again from the summation of a telescopic series, (b) is due to the updating rule for \( \gamma^\nu \) in (S.5) at iteration \( \nu_i \). We can now distinguish two cases: (i) step (S.5) has never been reached, i.e. \( T \) has never been diminished; (ii) case (i) did not occur. Note that we generally expect (ii) to happen and (i) is a somewhat residual case. If case (i) occurs, by (54), the algorithm stops after at most
\[
\left[ \frac{8}{(\eta e)^2 T^{-1}} \max_i \{L_{\nabla g_i}(f^M - f^m + \frac{1}{T^{-1}} g^M_+) \} \right] \delta^2
\]
iterations. In case (ii), instead, again from (54) and taking into account that since we updated \( T \) at least once we have
\[
T^\nu_i = \frac{1}{2 \nabla f(x^\nu_i)^2 d(x^\nu) + \eta e \|d(x^\nu)\|^2} > \frac{\delta}{2B}.
\]
we see the procedure halts in at most
\[
\left[ \frac{16 B}{(\eta e)^2 \max_i \{L_{\nabla g_i}(f^M - f^m + \frac{2 B g^M_+}{\delta^4}) \} \right] \delta^4
\]
iterations. If \( \delta \leq 1 \), this gives an overall complexity of \( \mathcal{O}(\delta^{-4}) \). \( \square \)

**Remark 5.** If \( \delta \) is not assumed to be less or equal to 1, the number of iterations needed will still be given either by (56) or by (55). Correspondingly the complexity becomes \( \mathcal{O}(\max \{\delta^{-4}, \delta^{-2}\}) \). We did not state the theorem in this form because it seemed to us slightly misleading, since the \( \delta^{-2} \) could only possibly occur when \( \delta \) is larger than 1 and therefore \( \delta^{-2} > \delta^{-4} \).

In order to understand the meaning of the stopping criteria in steps (S.2) and (S.4), we reason asymptotically, in the same spirit of the analysis in [5]. Note that when we exit the algorithm either \( \|d(x^\nu)\| \) or \( \theta(x^\nu) \) is below \( \delta \). The analysis below shows that, in the limit, for \( \delta \) going to zero, this is enough to guarantee that Algorithm 2 finds stationary points, i.e. KKT, FJ or ES solutions.

Let us run Algorithm 2 for values of the threshold \( \delta = \delta^k \), with \( \delta^k \downarrow 0 \) and with fixed initial data. We recall that, in view of Theorem 3, for every \( k \), the scheme stops either at step (S.2) or at step (S.4) in a finite number of iterations \( N = N^k \), which is nondecreasing with respect to \( k \). Accordingly, let \( I = I^k \) be the corresponding number of times in which \( T^\nu \) and \( \gamma^\nu \) have been reduced, apart from the iteration 0. We distinguish two cases.
(i) The algorithm stops at step (S.4) an infinite number of times for every $k$ belonging to a subsequence $\mathcal{K}$, returning $x_{\delta k}$. We further consider two cases: (a) $N^k$ remains constant for every $k$ sufficiently large and (b) $N^k > N^{k-1}$ an infinite number of times.

(a) If an index $\bar{k} \in \mathcal{K}$ exists such that the number of iterations $N^k$ remains constant for every $k \geq \bar{k}$ belonging to $\mathcal{K}$, we have $\theta(x_{\delta \bar{k}}) = 0$. We recall that, for $d(x_{\delta \bar{k}})$ to be a KKT point for $(P_{x_{\delta \bar{k}}})$, relation (11) must hold; hence, since $\nabla f(x_{\delta \bar{k}})^T d(x_{\delta \bar{k}}) > -\eta c||d(x_{\delta \bar{k}})||^2 \geq -c||d(x_{\delta \bar{k}})||^2$ by (S.3) and $\theta(x_{\delta \bar{k}}) = 0$, (11) is not valid, and $d(x_{\delta \bar{k}})$ is not a KKT point for $(P_{x_{\delta \bar{k}}})$. Finally, leveraging Lemma 1, $x_{\delta \bar{k}}$ is either an ES or a FJ point for $(P)$.

(b) Since the sequence $\{x_k\}$ is bounded we can assume without loss of generality that $x_{\delta \bar{k}} \to \bar{x}$, and $N^k > N^{k-1}$ for every $k \in \mathcal{K}$. We have $\theta(x_{\delta \bar{k}}) \to \theta(\bar{x}) = 0$, due to the condition $\theta(x_{\delta \bar{k}}) \leq \delta^k$ for every $k \in \mathcal{K}$ and to the continuity of function $\theta(\bullet)$. Furthermore, it holds $I^k \geq I^{k-1} + 1$ for every $k \in \mathcal{K}$ and, in turn, $T^{N_k} \downarrow 0$ on $\mathcal{K}$, since $T^{N_k} = T^{\nu_{\delta \bar{k}}} \leq \frac{1}{\delta^k}$ for every $k$. Assume by contradiction that the eMFCQ holds at $\bar{x}$: by Proposition 1, for any $k \in \mathcal{K}$ sufficiently large, $d(x_{\delta \bar{k}})$ is a KKT point for $(P_{x_{\delta \bar{k}}})$ and, in turn, by (15),

$$
\nabla f(x_{\delta \bar{k}})^T d(x_{\delta \bar{k}}) - \frac{1}{T^{N_k}} \theta(x_{\delta \bar{k}}) + \eta c||d(x_{\delta \bar{k}})||^2 \leq \left(m\|\xi^{N_k}\|_{\infty} - \frac{1}{T^{N_k}}\right) \theta(x_{\delta \bar{k}}).
$$

(57)

Thanks to the local boundedness of the set of KKT multipliers and because $T^{N_k} \downarrow 0$ on $\mathcal{K}$, eventually the right hand side of (57) is nonpositive, in contradiction to the condition $\nabla f(x_{\delta \bar{k}})^T d(x_{\delta \bar{k}}) + \eta c||d(x_{\delta \bar{k}})||^2 > 0$ and $T^{N_k} \nabla f(x_{\delta \bar{k}})^T d(x_{\delta \bar{k}} + \eta c||d(x_{\delta \bar{k}})||^2$ for every $k \in \mathcal{K}$ in (S.3). Therefore, $\bar{x}$ is either an ES or a FJ point for $(P)$.

(ii) The algorithm stops at step (S.4) only a finite number of times. Thus, eventually, the algorithm stops at step (S.2), returning $x_{\delta k}$. We consider the two cases (a) and (b) as before.

(a) If a sufficiently large index $\bar{k}$ exists such that the number of iterations $N^k$ remains constant for every $k \geq \bar{k}$, we have $d(x_{\delta \bar{k}}) = 0$ and, in turn, by (47), $\theta(x_{\delta \bar{k}}) = 0$. In view of Lemma 1, if the eMFCQ holds at $x_{\delta \bar{k}}$, then $\kappa(x_{\delta \bar{k}}) = \max\{g_i(x_{\delta \bar{k}})\} = 0$ and, thanks to (8), by A3, A4, A6-A9, $x_{\delta \bar{k}}$ is a KKT point for $(P)$; otherwise, it is either an ES or a FJ point for $(P)$.

(b) Since the sequence $\{x_k\}$ is bounded we can assume without loss of generality that $x_{\delta \bar{k}} \to \bar{x}$, and $N^k > N^{k-1}$ for every $k \in \mathcal{K}$. It follows that $d(x_{\delta \bar{k}}) \to 0$ and, again by (47), $\theta(x_{\delta \bar{k}}) \to \theta(\bar{x}) = 0$, due to the condition $d(x_{\delta \bar{k}}) \leq \delta^k$ for every $k \in \mathcal{K}$ and to the continuity of function $\theta(\bullet)$. If the eMFCQ holds at $\bar{x}$, then function $d(\bullet)$ is continuous on a neighborhood of $\bar{x}$ (see Proposition 3); besides, relying on Lemma 1, $\kappa(\bar{x}) = \max\{g_i(\bar{x})\} = 0$. Taking the limit in (8), by A3, A4, A6-A9, and observing that $d(x_{\delta \bar{k}}) \to d(\bar{x}) = 0$, $\bar{x}$ is shown to be a KKT point for $(P)$; differently, whenever the eMFCQ is not satisfied, $\bar{x}$ is either an ES or a FJ point for $(P)$.

We conclude this section by observing that if a feasible starting point is available, then by choosing a sufficiently small initial step $\gamma^{-1}$, the iteration complexity of Algorithm 2 can be reduced to $O(\delta^{-3})$. This result seems rather sensible because if we start with a feasible point we have already solved the feasibility problem which is a part of the constrained optimization. Nevertheless, it was in principle not clear that our algorithm could take advantage of this fact, since the search for feasibility and that for optimality are combined in a single step.

**Corollary 2.** Assume the same setting of Theorem 3, fix a prescribed tolerance $\delta$ and set, according to this value, $T = \min\{\frac{\delta}{B^2 \max\{L_{\nu}, \eta c\}}\}$. If the starting point $x^0$ is feasible, then, in at most $O(\delta^{-3})$ iterations, Algorithm 2 stops either at step (S.2) or at step (S.4).
Proof. We use the same notation and terminology introduced in the proof of Theorem 3. We first observe that Algorithm 2 never updates $\gamma^\nu$ and $T^\nu$. Indeed, suppose that the test in (S.3) is met for the first time at iteration $\nu$. The claim follows noting that if the condition in (S.3) is verified, then
\[
\frac{\delta}{B} \geq T^{-1} = T^{\nu-1} > \frac{\theta(x^\nu)}{\nabla f(x^\nu)^2d(x^\nu) + \eta \|d(x^\nu)\|^2} \geq \frac{\theta(x^\nu)}{B},
\]
so that $\theta(x^\nu) \leq \delta$ and the algorithm stops. Because of this, we can write $N = N_0 + 1$. Hence, assuming that the stopping test at (S.2) is not satisfied, we have $\|d(x^\nu)\| > \delta$, for all $\nu \leq N_0 + 1$ and
\[
\delta^2(N_0 + 1) < \sum_{\nu=0}^{N_0} \|d(x^\nu)\|^2 \leq \frac{W(x^0;T^{-1}) - W(x^{N_0+1};T^{-1})}{8 \max \{L_{y_1}\}} \leq \frac{\max \{g_i(x^0)\}}{\max \{g_i(x^{N_0+1})\}} - f(x^{N_0+1}) - \frac{1}{T-1} \max \{g_i(x^{N_0+1})\} \]
where we have used (52) and, in the last inequality, the feasibility of $x^0$. We can write $T^{-1} = e(\delta) \frac{\delta}{B}$, where
\[
e(\delta) = \begin{cases} 1 & \text{if } \delta \leq B \frac{\max \{L_{y_1}\}}{\max \{L_{x_i}\}} \\ B \frac{\delta}{\max \{L_{x_i}\}} & \text{otherwise} \end{cases}
\]
Therefore we have
\[
N_0 + 1 < \frac{B8 \max \{L_{y_1}\}}{(\eta \epsilon)^2 e(\delta) \delta^3} [f(x^0) - f^m]
\]
and, recalling that $\delta \leq 1$, the assertion follows easily.

The proof above essentially shows that, if we start with the given “small” $T^{-1}$ and, correspondingly, with a “small” $\gamma^{-1}$, we never update the stepsize and the algorithm terminates in $O(\delta^{-3})$ iterations, at most. This result is somewhat unusual, since it links algorithmic choices to the desired accuracy. However, in spite of this, it seems perfectly practical and easily implementable.

5.3. Problem constants are not known The implementation of Algorithm 2 requires the knowledge of some of the problem constants in Table 1. Hence the question arises whether we can modify the algorithm to avoid the use of potentially unknown constants, while retaining complexity results similar to those in Theorem 3. The answer is positive, at the price of a “small amount” of additional function evaluations. Additionally, differently from all previous developments, we must make a numerical use of the penalty function $W$. Observe that in Algorithm 2 the problem constants are used to set some initial values in Data and, more critically, in (S.5). Referring to the proof of Theorem 3, the updating of $\gamma^\nu$ in (S.5) guarantees condition (52), i.e. the sufficient decrease of the (ghost) penalty function. But, at a more basic level, this sufficient decrease condition can always be reached if the step $\gamma^\nu$ is sufficiently small. So, one could choose at each iteration the stepsize $\gamma^\nu$ so as to guarantee that the sufficient decrease condition (52) is satisfied. This can be accomplished without any knowledge of the problem constants; we only need to know the user-set quantities $c$ and $\eta$ as shown in Algorithm 3.

In Data we no longer need to set the initial $T$ and $\gamma$ to some small values that depend on problem constants. Indeed, whatever the initial values, it is the algorithm itself that sets them to the appropriate quantities. In Algorithm 2, updating the stepsize at (S.5) makes (52) satisfied at all subsequent iterations, until the if section at (S.3) is possibly re-entered. In Algorithm 3 instead, we do not have such a guarantee, and thus we perform the “line-search” in (S.7) at each iteration. The
Algorithm 3: Algorithm for (P) without constants

Data: $\delta > 0$, $x^0$, $T^{-1} > 0$, $\gamma^{-1} = 1$, $\nu \leftarrow 0$;

repeat

(S.1) compute $\kappa(x^\nu)$, the solution $d(x^\nu)$ of problem (P$_{x^\nu}$) and $\theta(x^\nu)$;

(S.2) if $\|d(x^\nu)\| \leq \delta$ then
   stop and return $x_\delta = x^\nu$;
end

(S.3) if $\nabla f(x^\nu)^T d(x^\nu) + \eta c\|d(x^\nu)\|^2 > 0$ and $T^{\nu-1} > \frac{\theta(x^\nu)}{\nabla f(x^\nu)^T d(x^\nu) + \eta c\|d(x^\nu)\|^2}$ then

(S.4) if $\theta(x^\nu) \leq \delta$ then
   stop and return $x_\delta = x^\nu$;
else
   set $T^\nu = \frac{1}{2} \frac{\theta(x^\nu)}{\nabla f(x^\nu)^T d(x^\nu) + \eta c\|d(x^\nu)\|^2}$;
end
else
   set $T^\nu = T^{\nu-1}$;
end

(S.7) while $W(x^\nu + \gamma^\nu d(x^\nu); T^\nu) - W(x^\nu; T^\nu) > -\gamma^\nu \frac{\theta}{4} \|d(x^\nu)\|^2$ do
   set $\gamma^\nu \leftarrow \frac{1}{2} \gamma^\nu$;
end

(S.8) set $x^{\nu+1} = x^\nu + \gamma^\nu d(x^\nu)$, $\nu \leftarrow \nu + 1$;
end

The following theorem shows that Algorithm 3 needs an amount of iterations which is similar (likely smaller, see comments after the proof) to that required by Algorithm 2. However, while for Algorithm 3 this quantity is also equal to the number of function and constraints evaluations, we now have an extra cost, in that we may need $\log_2 \frac{G-1}{\gamma}$ extra function and constraints evaluations, where

$$G \triangleq \frac{3\eta c}{4} \left( \frac{L_{\nabla f} + 2B \max_i \{L_{\nabla g_i}\}}{\delta} \right)^{-1}. \quad (58)$$

In what follows we assume, without loss of generality, that $G \leq \frac{1}{2}$. If the expression of $G$ given by (58) is larger than $\frac{1}{2}$, we simply redefine $G$ to be $\frac{1}{2}$.

**Theorem 4.** Let \( \{x^\nu\} \) be the sequence generated by Algorithm 3 under Assumptions A, C2 and D1 and suppose that $\delta \leq 1$. Then, in at most a number $\text{Iter}$ iterations, with $\text{Iter} \triangleq \mathcal{O}(\delta^{-4})$ and $\text{Iter} + \log_2 \frac{\gamma^{-1}}{\delta}$ function evaluations, Algorithm 3 stops either at step (S.2) or at step (S.4).

**Proof.** The proof is a variant of that of Theorem 3 to which we refer for notation and terminology. Suppose that Algorithm 2 performs $N$ iterations without stopping. We first count how many times $T^\nu$ can be updated in step (S.5) of the algorithm: let

$$\mathcal{I} \triangleq \{0 < \nu_i \leq N \mid T^{\nu_i} \text{ is updated in (S.5)} \} \cup \{0\}$$

be the set of iterations’ indices $\nu$ (in increasing order) at which we need to modify $T^\nu$, union iteration 0. Repeating verbatim the first part in the proof of Theorem 3, one can show that $\mathcal{I}$ has finite cardinality and, if $\nu_i \in \mathcal{I}$ then

$$i < \log_2 \frac{T^{-1}2B}{\delta}.$$
Define now $I$ and $N_i$ as in the proof of Theorem 3. Clearly $T' = T^{\nu_i}$ for every $\nu \in \{\nu_1, \ldots, \nu_i + N_i\}$. For every $\nu \in \{\nu_i, \ldots, \nu_i + N_i\}$, $\nu_i \in \mathcal{I}$, $T' \nu$ is not reduced and therefore

$$
\nabla f(x^{\nu_i})^T d(x^{\nu_i}) + \eta \gamma ||d(x^{\nu_i})||^2 \leq 0 \quad \text{or} \quad T' \nu = T^{\nu_i} \leq \frac{\theta(x^{\nu_i})}{\nabla f(x^{\nu_i})^T d(x^{\nu_i}) + \eta \gamma ||d(x^{\nu_i})||^2}. \tag{59}
$$

By (59),

$$
\nabla f(x^{\nu_i})^T d(x^{\nu_i}) - \frac{\theta(x^{\nu_i})}{T^{\nu_i}} \leq -\eta \gamma ||d(x^{\nu_i})||^2,
$$

for every $\nu \in \{\nu_i, \ldots, \nu_i + N_i\}$, $\nu_i \in \mathcal{I}$ which, in turn, by (14), implies

$$
W(x^{\nu_i+1}; T^{\nu_i}) - W(x^{\nu_i}; T^{\nu_i}) \leq -\gamma' \left[ \eta \gamma - \frac{\gamma'}{2} \left( L_{\nabla f} + \frac{\max_i \{L_{\nabla g_i}\}}{T^{\nu_i}} \right) \right] ||d(x^{\nu_i})||^2, \tag{60}
$$

where we took $\epsilon' = T^{\nu_i}$. We now note that for every $\nu \in \{\nu_i, \ldots, \nu_i + N_i\}$, $\nu_i \in \mathcal{I}$, we have $\gamma' \geq G$. Indeed, this is trivial for $\nu = 0$, since we assumed $\gamma^{-1} = 1$ and $G \leq \frac{1}{2}$. Suppose by contradiction that $0 \neq \nu_i \in \mathcal{I}$ and $\gamma' < G$. Note that the definition (58) of $G$ shows that, if we set $\gamma' < 2G$, we get, recalling (60) and $T^{\nu_i} > \frac{\delta}{2\eta}$, that $W(x^{\nu_i+1}; T^{\nu_i}) - W(x^{\nu_i}; T^{\nu_i}) \leq -\gamma' \eta \gamma ||d(x^{\nu_i})||^2$, i.e. the test at (S.7) is surely not satisfied if $\gamma' < 2G$. This, in turn, contradicts $\gamma' < G$, since it shows that in the loop (S.7) we should have stopped at the previous iterate of the cycle. Therefore, taking into account that $\gamma'$ is obtained at (S.7) after a certain number (possibly zero) of halvings of the current value of the stepsize, at each iteration $\gamma' \geq G$. We conclude that $\gamma'$, globally, needs to be halved no more than $\log_2 \frac{1}{\gamma'}$ times in order to achieve the sought decrease condition

$$
W(x^{\nu_i+1}; T^{\nu_i}) - W(x^{\nu_i}; T^{\nu_i}) \leq -\gamma' \eta \gamma ||d(x^{\nu_i})||^2 \leq -G \eta \gamma ||d(x^{\nu_i})||^2, \tag{61}
$$

for every $\nu \in \{\nu_i, \ldots, \nu_i + N_i\}$, $\nu_i \in \mathcal{I}$. Recalling definition (13), and similarly to (53),

$$
\delta^2 N = \sum_{i=0}^{I} \delta^2 (N_i + 1) < \sum_{i=0}^{I} \sum_{\nu=\nu_i}^{\nu_i+N_i} ||d(x^{\nu_i})||^2 \leq \sum_{i=0}^{I} W(x^{\nu_i}; T^{\nu_i}) - W(x^{\nu_i+N_i+1}; T^{\nu_i}) \frac{G \eta \gamma}{4} \tag{62}
$$

where the second inequality is due to (61), while the last inequality is valid as a result of a telescopic series argument since $\nu_i + N_i + 1 = \nu_i+1$. Setting $f^M \triangleq \max_x \{ f(x) \mid x \in K \}$, $g^M_+ \triangleq \max_x \{ \max_i \{g_i(x)\} \mid x \in K \}$ and $f^m \triangleq \min_x \{ f(x) \mid x \in K \}$, by (62) we can write

$$
\delta^2 N < \frac{1}{G \eta \gamma} \left( f^M - f^m + \frac{1}{T_0} g^M_+ - \frac{1}{T_0} g^M_+ + \frac{1}{T_0} g^M_+ \right), \tag{63}
$$

where the inequality follows, recalling that $T^{\nu_i} \leq T^{\nu_i-1}$, from the summation of a telescopic series. We now have two cases: (i) $I = 0$, i.e. $\nu_i = \nu_0$ and (ii) $I > 0$, i.e. $\nu_i > \nu_0$. In case (i), (63) and the definition of $G$ immediately give

$$
N < \frac{16}{3(\eta \gamma)^2} \left( L_{\nabla f} + 2B \max_i \{L_{\nabla g_i}\} \right) \left( f^M - f^m + \frac{g^M_+}{T_0} \right) \frac{1}{\delta^3}. \tag{64}
$$

In case (ii), since

$$
T^{\nu_i} = \frac{1}{2} \frac{\theta(x^{\nu_i})}{\nabla f(x^{\nu_i})^T d(x^{\nu_i}) + \eta \gamma ||d(x^{\nu_i})||^2} > \frac{\delta}{2B},
$$
and using the definition of $G$, (63) shows that the procedure halts in at most
\[
\left\lceil \frac{16}{3 \eta \epsilon^2} \left( L_{\nabla f} \delta + 2 B_{\max} \{ L_{\nabla g_i} \} \right) \left[ \frac{f^m-f^m}{\delta^3} + \frac{2 B g^M}{\delta^4} \right] \right\rceil
\]
iterations. Recalling that $\delta \leq 1$, (64) and (65) show the overall complexity of $O(\epsilon^{-1})$.
\[\square\]

It is interesting to compare the worst-case bounds (56) (for Algorithm 2) and (65) (for Algorithm 3). It is clear that, at least for a small $\delta$, the bound (65) is approximatively $\frac{2}{3}$ of the bound (56). This better behaviour of Algorithm 3 has a simple explanation. The steps used in Algorithm 3 are generally larger than those used in Algorithm 2, where the (theoretical) constants of the problem are used to define a “pessimistic” step-length. In Algorithm 3, instead, local information is gathered through the line-search in (S.7) that permits the definition of a stepsize better adapted to the problem.

Algorithm 3 also has the additional merit of not requiring the knowledge of the problem constants. We pay a price for this better results in that the algorithm is marginally more complex and requires additional objective and constraint evaluations that may increase the computational effort. However, note that this increase is negligible when $\delta$ is small, since Iter is of order $O(\epsilon^{-1})$ while the additional number of function evaluations is (see the definition of $G$) $O(\log_2(\epsilon^{-1}))$, implying that the overall order of function evaluations is maintained to be $O(\epsilon^{-1})$.

Acknowledgments. The authors are very thankful to Philippe Toint, whose detailed comments on an earlier version of this paper helped improve it.

Facchinei was partially supported by MIUR PLATINO (PLATform for InNuOvative services in future internet) PON project, under Grant Agreement no. PON01_01007. Scutari was supported by the USA National Science Foundation (NSF) under Grants CIF 1564044 and CAREER Award No. 1555850, and the Office of Naval Research (ONR) Grant N00014-16-1-2244. Kungurtsev was supported by the Czech Science Foundation project 17-26999S

References
[1] Bank B, Guddat J, Klatte D, Kummer B, Tammer K (1982) Non-Linear Parametric Optimization (Akademie-Verlag Berlin).
[2] Bertsekas D (1999) Nonlinear programming (Athena scientific Belmont).
[3] Bertsekas D (2015) Convex optimization algorithms (Athena Scientific Belmont).
[4] Bertsekas D, Tsitsiklis J (1989) Parallel and distributed computation: numerical methods, volume 23 (Prentice hall Englewood Cliffs, NJ).
[5] Birgin E, Gardenghi J, Martínez J, Santos S, Toint P (2016) Evaluation complexity for nonlinear constrained optimization using unscaled KKT conditions and high-order models. SIAM Journal on Optimization 26(2):951–967.
[6] Burke J (1989) A sequential quadratic programming method for potentially infeasible mathematical programs. Journal of Mathematical Analysis and Applications 139(2):319–351.
[7] Burke J (1992) A robust trust region method for constrained nonlinear programming problems. SIAM Journal on Optimization 2(2):325–347.
[8] Burke J, Han SP (1989) A robust sequential quadratic programming method. Mathematical Programming 43(1):277–303.
[9] Cartis C, Gould N, Toint P (2011) On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming. SIAM Journal on Optimization 21(4):1721–1739.
[10] Cartis C, Gould N, Toint P (2014) On the complexity of finding first-order critical points in constrained nonlinear optimization. Mathematical Programming 144(1-2):93–106.
[11] Cartis C, Gould N, Toint P (2017) Corrigendum: On the complexity of finding first-order critical points in constrained nonlinear optimization. Mathematical Programming 161(1-2):611–626.
[12] Cartis C, Gould N, Toint PL (2017) Optimality of orders one to three and beyond: characterization and evaluation complexity in constrained nonconvex optimization. arXiv preprint arXiv:1705.07285.

[13] Di Pillo G, Grippo L (1986) An exact penalty function method with global convergence properties for nonlinear programming problems. Mathematical Programming 36(1):1–18.

[14] Di Pillo G, Grippo L (1989) Exact penalty functions in constrained optimization. SIAM Journal on control and optimization 27(6):1333–1360.

[15] Facchinei F (1997) Robust recursive quadratic programming algorithm model with global and superlinear convergence properties. Journal of Optimization Theory and Applications 92(3):543–579.

[16] Facchinei F, Lampariello L, Scutari G (2016) Feasible methods for nonconvex nonsmooth problems with applications in green communications. Mathematical Programming 1–36.

[17] Facchinei F, Pang JS (2003) Finite-Dimensional Variational Inequalities and Complementarity Problems (Springer).

[18] Facchinei F, Scutari G, Sagratella S (2015) Parallel selective algorithms for nonconvex big data optimization. IEEE Trans. on Signal Processing 63(7):1874–1889.

[19] Lucidi S (1992) New results on a continuously differentiable exact penalty function. SIAM Journal on optimization 2(4):558–574.

[20] Martínez J (2017) On high-order model regularization for constrained optimization. Technical report, tech. report, 891 http://www. optimization-online. org/DB HTML/2017/02/5850. html. 892 [28].

[21] Nesterov Y (2013) Introductory lectures on convex optimization: A basic course, volume 87 (Springer Science & Business Media).

[22] Nesterov Y, Polyak B (2006) Cubic regularization of newton method and its global performance. Mathematical Programming 108(1):177–205.

[23] Polyak B (1987) Introduction to optimization. translations series in mathematics and engineering. Optimization Software.

[24] Rockafellar R (1985) Lipschitzian properties of multifunctions. Nonlinear Analysis: Theory, Methods & Appl. 9(8):867–885.

[25] Rockafellar R, Wets J (1998) Variational Analysis (Springer).

[26] Scutari G, Facchinei F, Lampariello L, Sardellitti S, Song P (2017) Parallel and distributed methods for constrained nonconvex optimization part ii: Applications in communications and machine learning. IEEE Transactions on Signal Processing 65(8):1945–1960.

[27] Scutari G, Facchinei F, Song P, Peiran, Palomar D, Pang JS (2014) Decomposition by partial linearization: Parallel optimization of multi-agent systems. IEEE Transactions on Signal Processing 62(3):641–656.

[28] Shor N (1985) Minimization methods for non-differentiable functions, volume 3 (Springer Science & Business Media).

[29] Yen N (1995) Hölder continuity of solutions to a parametric variational inequality. Applied Mathematics & Optimization 31(3):245–255.