# Ghost Penalties in Nonconvex Constrained Optimization: Diminishing Stepsizes and Iteration Complexity

Francisco Facchinei , Vyacheslav Kungurtsev , Lorenzo Lampariello , Gesualdo Scutari

informs
http://pubsonline.informs.org/journal/moor

MATHEMATICS OF OPERATIONS RESEARCH
Vol. 46, No. 2, May 2021, pp. 595–627
ISSN 0364-765X (print), ISSN 1526-5471 (online)

# Ghost Penalties in Nonconvex Constrained Optimization: Diminishing Stepsizes and Iteration Complexity

**Francisco Facchinei,[a] Vyacheslav Kungurtsev,[b] Lorenzo Lampariello,[c] Gesualdo Scutari[d]**

[a] Department of Computer, Control, and Management Engineering Antonio Ruberti, Sapienza University of Rome, 00185 Rome, Italy; [b] Department of Computer Science, Faculty of Electrical Engineering, Czech Technical University in Prague, 12135 Prague, Czech Republic; [c] Department of Business Studies, Roma Tre University, 00145 Rome, Italy; [d] School of Industrial Engineering, Purdue University, West Lafayette, Indiana 47097

**Contact:** francisco.facchinei@uniroma1.it, https://orcid.org/0000-0002-7714-1210 (FF); vyacheslav.kungurtsev@fel.cvut.cz, https://orcid.org/0000-0003-2229-8824 (VK); lorenzo.lampariello@uniroma3.it, https://orcid.org/0000-0003-4177-3598 (LL); gscutari@purdue.edu, https://orcid.org/0000-0002-6453-6870 (GS)

**Received:** October 12, 2017
**Revised:** January 16, 2019; July 19, 2019; December 20, 2019
**Accepted:** March 18, 2020
**Published Online in Articles in Advance:** February 2, 2021

**MSC2000 Subject Classification:** Primary: 90C30, 90C60, 65K05
**OR/MS Subject Classification:** Primary: Programming: nonlinear algorithms; secondary: analysis of algorithms: computational complexity

https://doi.org/10.1287/moor.2020.1079

**Abstract.** We consider nonconvex constrained optimization problems and propose a new approach to the convergence analysis based on penalty functions. We make use of classical penalty functions in an unconventional way, in that penalty functions only enter in the theoretical analysis of convergence while the algorithm itself is penalty free. Based on this idea, we are able to establish several new results, including the first general analysis for diminishing stepsize methods in nonconvex, constrained optimization, showing convergence to generalized stationary points, and a complexity study for sequential quadratic programming–type algorithms.

**Funding:** F. Facchinei was partially supported by the MIUR PLATINO (PLATform for INnOvative services in future interet) PON project [Grant PON01_01 007]. G. Scutari was supported by the National Science Foundation [Grants CIF 1564044 and CAREER 1555850], the Office of Naval Research [Grant N00014-16-1-2244], and Army Research Office [Grant W911NF1810238]. V. Kungurtsev was supported by Operational Program Research, Development, and Education, Project CZ.02.1.01/0.0/0.0/16_109/0000765 from the Research Center for Informatics.

**Keywords:** constrained optimization • nonconvex problem • diminishing stepsize • generalized stationary point • iteration complexity

## 1. Introduction

We consider the nonconvex constrained optimization problem

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f(x) \\ \text{s.t.} \quad & g(x) \le 0 \\ & x \in K, \end{aligned} \tag{P}$$

where $K \subseteq \mathbb{R}^n$ is a nonempty closed and convex set, and $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^m$ are $C^{1,1}$ (i.e., continuously differentiable with locally Lipschitz gradients) functions on an open set containing $K$.

Penalty functions (differentiable or nondifferentiable, exact or sequential) are part of the folklore in optimization and have been widely used in analyzing optimality conditions, stability and sensitivity properties, and in developing solution methods. In this paper, we put forward a new use of penalty functions in the design of algorithms for the solution of (P). In particular, we consider the classical nondifferentiable penalty function

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

where $\alpha_+ \triangleq \max\{0,\alpha\}$ and $\varepsilon$ is a positive penalty parameter. We propose a novel use of $W(x;\varepsilon)$ in that, contrary to usual penalty algorithms, the penalty function *only enters in the theoretical analysis* of convergence, whereas *the algorithm itself is penalty free*, hence the term *ghost* penalty. We establish (subsequential) convergence to *generalized stationary points* under essentially no assumptions beyond the $C^{1,1}$ requirement on the problem functions. In particular, we assume neither feasibility of the problem nor any constraint qualification and therefore (subsequential) convergence to generalized stationary point is the natural target for a well-behaved algorithm (Birgin et al. [8], Burke[11], Burke [12], Burke and Han [13], Burke and Hoheisel [14], Cartis et al. [15], Cartis et al. [16], Cartis et al. [17], Cartis et al. [18], Cartis et al. [19], El-Alem [29], Facchinei [30], Liu and Sun [40], Liu and Yuan [41], Martinez [44], Yuan [67]). Once our main convergence result has been established, the role played by further classical assumptions, like feasibility or constraint qualifications, for example, is easily understood and can be taken into account in a straightforward way.

Our framework is of a generalized sequential quadratic programming (SQP) type. At each iteration $x^\nu$, we generate a search direction $d(x^\nu)$ by solving a strongly convex optimization subproblem constructed along the lines discussed in the seminal papers (Burke [11], Burke and Han [13]) and also taking into account the developments in Facchinei et al. [32] and Scutari et al. [57]. The direction-finding subproblem reads as follows:

$$\underset{d}{\text{minimize}}\, \tilde{f}(d; x^\nu) \quad \text{s.t.} \quad \tilde{g}(d; x^\nu) \le \kappa(x^\nu)\, e,\ \|d\|_\infty \le \beta,\ d \in K - x^\nu, \tag{1}$$

where $\tilde{f}(\bullet; x^\nu)$ and $\tilde{g}_i(\bullet; x^\nu)$ are strongly convex and convex approximations of $f$ and $g$, respectively, $\kappa(x^\nu) \in \mathbb{R}$ is nonnegative and defined to make the subproblem feasible, $\beta$ is a user-chosen positive constant, and $e \in \mathbb{R}^m$ is the vector with all components being one. The classical SQP subproblem is a particular instance of (1), when $\tilde{g}$ is just a linearization of $g$ and $\tilde{f}$ a positive definite quadratic approximation of $f$. Our more general approach leaves room for much flexibility in tailoring the direction finding subproblem to the problem at hand and to exploit any available specific structure in (P) (see Section 3 for more details).

A step $\gamma^\nu$ is then taken along this direction so that

$$x^{\nu+1} = x^\nu + \gamma^\nu d(x^\nu). \tag{2}$$

We consider both classical diminishing stepsize methods (DSMs) wherein $\gamma^\nu$ is a positive stepsize such that

$$\lim_{\nu\to\infty} \gamma^\nu = 0 \qquad \text{and} \qquad \sum_{\nu=0}^{\infty} \gamma^\nu = \infty, \tag{3}$$

and stepsize selection rules where $\gamma^\nu$ is chosen in a more problem-tailored way, typically fixed for at least a subsequence of iterations if not for the entire sequence. Although the algorithms in this paper generate the search direction in a (generalized) SQP fashion, they differ markedly from classical SQP methods in the way they select stepsizes. Indeed, SQP methods traditionally have adopted effective globalization procedures based, for example, on line-search or trust-region strategies. Here, instead, we mostly study different techniques that may be useful, for example, in very large-scale or distributed settings and that, in addition, allow us to perform an iteration complexity analysis. Given the effectiveness of the SQP approach in handling nonconvex constraints, our results hopefully provide an alternative to expand the applicability of SQP-like methods.

It may be interesting to mention at this juncture the two papers (Auslender et al. [2], Bolte and Pauwels [9]), where, in the context of an extended SQP-like scheme, a stepsize of one is always taken, thus also foregoing line-search, trust-region, or other standard globalization techniques. The possibility to use a fixed (large) stepsize derives from the fact that the methods in Auslender et al. [2] and Bolte and Pauwels [9] are *feasible* methods, and the surrogate for the objective function is always an *upper* convex approximation; in fact, the algorithms analyzed in these two interesting works belong to the class of SQP-type majorization-minimization schemes. The essential ingredient in developing such methods is the ability to build approximations that majorize both the constraints and the objective function; we will discuss the consequences of this setting in the context of our scheme in Sections 5 and 6.

Some other related papers in the contemporary literature include Auslender [1], Cartis et al. [15], Fletcher [34], and Solodov [59], where penalty algorithms are analyzed. The methods discussed in these works aim at minimizing directly the penalty function by resorting to composite optimization approaches; this results in a double-loop structure whereby the penalty function is (approximatively) minimized for a certain value of the penalty parameter, and then the penalty parameter is possibly updated. Although there are some similarities in the analysis, the algorithms presented in our paper rely on a pure *SQP-like* subproblem, and penalty parameters enter only in the theoretical analysis and not in the subproblem definition.

Based on our ghost penalty approach, our main contributions are as follows:

a. the first (subsequential) convergence result for a wide class of DSMs for general nonconvex constrained optimization problems; and

b. *Iteration complexity results* for some suitable choices of the stepsize $\gamma^\nu$ leading, among other things, to the first iteration complexity analysis for SQP-type methods in a general setting.

The results related to (a) considerably widen the scope of applicability of DSMs. DSMs are part of the core techniques in optimization, and their advantages and disadvantages are well known (Bertsekas [6], Polyak [50], Shor [58]). However, DSMs are not yet fully understood when it comes to nonconvex problems. Indeed, the very recent paper (Davis et al. [24]) is the first study to establish convergence results for DSMs applied to unconstrained, nonsmooth, and nonconvex problems. The results in the present paper, therefore, close a surprising gap in the literature, because DSMs have never been proved to lead to convergence when addressing problems with nonconvex constraints except in some specialized settings where feasibility of the iterates can be maintained throughout the optimization process and constraint qualifications are assumed (Facchinei et al. [32], Scutari et al. [55]). We show that every limit point of the sequence $\{x^\nu\}$ produced according to (2) and (3) is a generalized stationary solution for (P). By generalized stationary, we intend a point that can be an infeasible stationary solution of the violation-of-the-constraint problem associated to (P), a Fritz-John (FJ), or a Karush-Kuhn-Tucker (KKT) point of (P). As mentioned previously, this is the natural target of an algorithm for constrained optimization when neither blanket assumptions about feasibility of (P) are made, nor constraint qualifications are assumed to hold. Many consider DSMs a *necessary evil* and nevertheless they are currently used in many settings, for example, in parallel and distributed optimization, in stochastic optimization, in multiagent settings, in incremental methods, and whenever the computation of the objective function is very expensive or the problem is affected by noise (Bertsekas[5], Bertsekas [6], Bertsekas and Tsitsiklis [7], Bottou et al. [10], Daneshmand et al. [21], Daneshmand et al. [22], Davis and Drusvyatskiy [23], Davis et al. [24], Dean et al. [25], Di Lorenzo and Scutari [26], Facchinei et al. [32], Facchinei et al. [33], Gupta et al. [35], Kingma and Ba [38], Nedic et al. [45], Nemirovski et al. [46], Ng and Yu [49], Polyak [50], Scutari et al. [55], Scutari et al. [56], Suctari et al. [57], Tatarenko and Touri [63], Wang et al. [65], Zeng and Yin [68]). In some cases, it might be hoped that soon more effective alternatives will be found; in other cases, alternatives are harder to anticipate. In any case, as DSMs evolve to deal with new classes of problems of contemporary interest, the need to tackle nonconvex constraints and to relax the conditions needed to analyze convergence emerges. The developments in this paper, dealing with a standard nonconvex optimization problem, are a first step in this direction and will hopefully pave the way for further advancements in the more challenging settings mentioned previously.

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 (Nesterov [47]), this chapter was opened by the largely ignored paper by Vavasis [64] but gained momentum only with the work of Nesterov and Polyak [48] on a cubic regularization method for the unconstrained minimization of a nonconvex, smooth function. An excellent review of results in this field is contained in Cartis et al. [18], to which we refer the interested reader for a broader view on the subject. Here we only discuss results on algorithms for nonconvex, inequality constrained problems aimed at locating *generalized* stationary points using first-order information, similarly to what we do in the present paper.

By using our ghost penalty approach, we are able to establish that $\mathcal{O}(\delta^{-4})$ iterations are needed at worst to find a $\delta$−approximate generalized stationary point (see Definition 3); this definition of $\delta$−approximate generalized stationarity relaxes the notion of an exact generalized stationary point and parallels similar developments in Birgin et al. [8], Cartis et al. [17], Cartis et al. [18], and Cartis et al. [19]. In line with what was discussed previously, we remark that our notion of (approximate) stationarity is naturally broader than the more standard (approximate) KKT conditions. The bound of $\mathcal{O}(\delta^{-4})$ can be reduced to $\mathcal{O}(\delta^{-3})$ if a feasible point is known in advance and to $\mathcal{O}(\delta^{-2})$ if some further conditions are met (see Section 5 for details). As far as we are aware of, these are the first iteration complexity results for SQP-type methods *in a general setting*, that is, without assuming, for example, feasibility of iterates (see the later discussion about Auslender et al. [2] and Bolte and Pauwels [9]). Indeed, with the exception of Cartis et al. [15], all other results for general nonconvex, constrained problems obtained so far in the literature are based on phase I–phase II type methods wherein an almost feasible point is first sought and then a second phase is started. More specifically, in Cartis et al. [15], a penalty approach is shown to take either $\mathcal{O}(\delta^{-2})$ or $\mathcal{O}(\delta^{-5})$ iterations to reach an approximate generalized stationary point, depending on the behavior of the penalty parameter during the minimization process. Cartis et al. [16] also describe a phase I–phase II cubic regularization method, possibly using Hessian information, for the solution of equality constrained problems and show that the number of iterations needed to reach an approximate generalized stationary point varies between $\mathcal{O}(\delta^{-3/2})$ and $\mathcal{O}(\delta^{-2})$ depending on certain algorithmic choices. Building on the algorithm in Cartis et al.[16], and using a different definition of approximate generalized stationary point, Birgin et al. [8] show that a phase I–phase II algorithm takes between $\mathcal{O}(\delta^{-3})$ and $\mathcal{O}(\delta^{-5})$ iterations to declare a point approximate generalized stationary, according to the choice of an algorithmic parameter. Finally, Cartis et al. [17] establish a bound of $\mathcal{O}(\delta^{-2})$, once again for a phase I–phase II method and using first-order information only. A detailed comparison of all these results is not straightforward, because of the many subtleties involved, and we defer a more detailed discussion on this issue to Remark 5. We conclude mentioning, once again, the SQP-like majorization-minimization methods proposed in Auslender et al. [2] and Bolte and Pauwels [9]. Differently to what we propose here, these two papers discuss only *feasible-type* methods and assume standard constraint qualifications. In this framework, interesting results are obtained concerning the *convergence rate* to zero of the *distance of the point generated by the method to a KKT solution* (as opposed to the more algorithmically oriented results in the papers discussed previously, where bounds are obtained on the number of iterations needed to satisfy a given stopping criterion). The distinction of iteration complexity and convergence rate is a subtle and sometimes blurred one. In a nutshell, the difference is that when we talk about complexity we are assuming that the constants appearing in the complexity bound are conceptually known a priori (e.g., Lipschitz constants), whereas in the case of a convergence rate, the bounds include constants that are possibly unknown in advance (e.g., the diameter of the region that contains all iterates). In Auslender et al. [2], linear convergence of the sequence of iterates to the optimal solution is obtained for strongly convex problems. The more general results in Bolte and Pauwels [9] dispense with convexity by assuming the Kurdyka-Łojasiewicz property and obtaining a convergence rate that depends on the Łojasiewicz exponent.

The paper is organized as follows. In Section 2, we introduce some mathematical preliminaries and, in particular, the appropriate definition of a generalized stationary point for nonconvex, constrained problems. In Section 3, we discuss in detail the direction finding subproblem and introduce some assumptions that will be used to establish convergence. In Section 4, we show convergence to generalized stationary points for DSMs, whereas in Section 5, we perform the iteration complexity analysis. We finish in Section 6 with a discussion on the boundedness of the sequence of iterates.

## 2. Generalized Stationary Points

We consider Problem (P), under the blanket assumptions indicated in the Introduction, and denote the feasible set of (P) by

$$\mathcal{X} \triangleq \{d \in \mathbb{R}^n \,:\, g(x) \leq 0,\, d \in K\}.$$

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, that is, the problem of finding a feasible point; and (ii) the problem of finding a local minimum point of the objective function over the feasible set. Even just the former problem is a hard one, because it essentially requires computing 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:

$$\begin{aligned} &\underset{x}{\text{minimize}}\ \max_i \{g_i(x)_+\}, \\ &x \in K, \end{aligned} \tag{4}$$

where, we recall, $\alpha_+ \triangleq \max\{0, \alpha\}$ for all $\alpha \in \mathbb{R}$. Let

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

and

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

where $N_{\mathbb{R}^m_-}(y)$ and $N_K(x)$ are the classical normal cones to the convex sets $\mathbb{R}^m_-$ and $K$ at $y$ and $x$, respectively, $\nabla f$ is the gradient of $f$ and $\nabla g$ is the transposed Jacobian of $g$. If $g(x) \leq 0$, the condition $\xi \in N_{\mathbb{R}^m_-}(g(x))$ can be more familiarly rewritten as $\xi_i \geq 0$, $\xi_i g_i(x) = 0$ 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 $\hat{x} \in K$ is not feasible; that is, if $g_i(\hat{x}) > 0$ for at least one index $i \in \{1, \ldots, m\}$, in view of the regularity of the functions involved, then the stationarity condition for problem (4),

$$0 \in \partial \max_i \{g_i(\hat{x})_+\} + N_K(\hat{x}), \tag{5}$$

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 $\hat{x} \in K$ is, for problem (P),

- a KKT solution if $g(\hat{x}) \leq 0$ and $M_1(\hat{x}) \neq \emptyset$;
- an 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 some $i \in \{1, \ldots, m\}$ and $M_0(\hat{x}) \neq \{0\}$.

We call $\hat{x} \in K$ a generalized stationary solution of (P) if any of these cases occurs.

Because we did not make any regularity or feasibility assumptions on problem (P), finding a generalized stationary solution in the sense just described is the appropriate requirement for a solution algorithm; we show that our method does converge to generalized stationary points as defined previously. 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 $\hat{x} \in K$ if

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

If $\hat{x}$ is feasible and $K = \mathbb{R}^n$, 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, that is, to the existence of a point $\tilde{x}$ such that $g(\tilde{x}) < 0$. The eMFCQ is rather standard and its definition goes back at least to Di Pillo and Grippo [27] and Di Pillo and Grippo[28], having its roots in Rockafellar [52]; since its introduction, it has been used rather often, especially in the analysis of penalty and SQP algorithms, because it arises quite naturally in these contexts. By using Craven [20, Motzkin's theorem of alternative 2.5.2], we see that the eMFCQ holds at $\hat{x} \in K$ if and only if

$$\exists \hat{d} \in T_K(\hat{x}): \ \nabla g_i(\hat{x})^T \hat{d} < 0, \quad \forall i : g_i(\hat{x}) = \max_j \{g_j(\hat{x})_+\}. \tag{6}$$

Because $K$ is convex, simple continuity arguments show that the latter condition is equivalent to

$$\exists \tilde{x} \in K: \ \nabla g_i(\hat{x})^T (\tilde{x} - \hat{x}) < 0, \quad \forall i : g_i(\hat{x}) = \max_j \{g_j(\hat{x})_+\}. \tag{7}$$

We state below a result that extends a standard property of the MFCQ for feasible points.

**Proposition 1.** *If the eMFCQ holds at $\hat{x} \in K$, then there exists a neighborhood $\mathcal{V}$ of $\hat{x}$ such that, for every $x \in K \cap \mathcal{V}$, the eMFCQ is satisfied.*

**Proof.** If $\hat{x} \in K$ is feasible, this is a classical result. If $\hat{x} \in K$ is not feasible, the condition $M_0(\hat{x}) = \{0\}$ implies that $\hat{x}$ is not a stationary point for the feasibility problem (4), that is, $0 \notin \partial \max_i \{g_i(\hat{x})_+\} + N_K(\hat{x})$. The assertion then easily follows from the outer semicontinuity and local boundedness of the subdifferential mapping $\partial \max_i \{g_i(\bullet)_+\}$ and by (see Rockafellar and Wets [54, proposition 6.6]) the outer semicontinuity relative to $K$ of the set valued mapping $N_K$ (see Rockafellar and Wets [54] 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 (2). Although the stepsize is chosen according to several rules to be discussed in the following sections, the direction $d(x^\nu)$ is the solution of the strongly convex subproblem (1), briefly described in the Introduction, that we repeat here for the reader's convenience.

Given a (base) point $x \in K$ (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:

$$\begin{aligned} \underset{d}{\text{minimize}} \quad & \tilde{f}(d;x) \\ \text{s.t.} \quad & \tilde{g}(d;x) \le \kappa(x)e \\ & \|d\|_\infty \le \beta, \\ & d \in K - x, \end{aligned} \tag{P$_\text{x}$}$$

where $e \in \mathbb{R}^m$ is the vector with all components being one and $\beta$ is a user-chosen positive constant. Moreover, $\tilde{f}$ is a strongly convex surrogate of the original objective function $f$, whereas $\tilde{g}$ is a convex surrogate of the original constraints $g$ (see Assumption A for the conditions these surrogates must obey). Finally, following Burke [11], the quantity $\kappa(x)$ in the surrogate constraints, which serves to suitably enlarge the feasible set of the subproblem to ensure it is always nonempty, is defined, for every $x \in K$, as follows:

$$\kappa(x) \triangleq (1-\lambda)\max_i\{g_i(x)_+\} + \lambda \min_d \left\{ \max_i\{\tilde{g}_i(d;x)_+\} \,|\, \|d\|_\infty \le \rho,\, d \in K - x \right\}, \tag{8}$$

with $\lambda \in (0,1)$ and $\rho \in (0,\beta)$. Note that (8) requires the computation of the optimal value of the convex problem

$$\min_d \left\{ \max_i\{\tilde{g}_i(d;x)_+\} \,|\, \|d\|_\infty \le \rho,\, d \in K - x \right\} \tag{9}$$

that always has an optimal solution because the feasible set is nonempty and compact. If $x$ is feasible for (P), $\kappa(x) = 0$. The additional constraint $\|d\|_\infty \le \beta$ allows us to avoid issues of ever-increasing search directions. Overall, in the sequel we denote by $\widetilde{\mathcal{X}}(x)$ and $d(x)$ the convex feasible set and the unique solution of subproblem (P$_x$), respectively, that is,

$$\begin{aligned} \widetilde{\mathcal{X}}(x) &\triangleq \{d \in \mathbb{R}^n \,:\, \tilde{g}(d;x) \le \kappa(x)\,e,\, \|d\|_\infty \le \beta,\, d \in K - x\} \\ d(x) &\triangleq \arg\min_d \left\{ \tilde{f}(d;x) \,|\, d \in \widetilde{\mathcal{X}}(x) \right\}, \end{aligned}$$

and we equivalently refer to constraint $\|d\|_\infty \le \beta$ as $d \in \beta\mathbb{B}^n_\infty$, where $\mathbb{B}^n_\infty$ is the closed unit ball in $\mathbb{R}^n$ associated with the infinity norm.

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

A1. $\tilde{f}(\bullet;x)$ is a strongly convex function on $O_d$ for every $x \in K$ with modulus of strong convexity $c > 0$ independent of $x$;

A2. $\tilde{f}(\bullet;\bullet)$ is continuous on $O_d \times O_x$;

A3. $\nabla_1\tilde{f}(\bullet;\bullet)$ is continuous $O_d \times O_x$;

A4. $\nabla_1\tilde{f}(0;x) = \nabla f(x)$ for every $x \in K$;

A5. $\tilde{g}_i(\bullet;x)$ is a convex function on $O_d$ for every $x \in K$;

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

A7. $\tilde{g}_i(0;x) = g_i(x)$ for every $x \in K$;

A8. $\nabla_1\tilde{g}_i(\bullet;\bullet)$ is continuous on $O_d \times O_x$;

A9. $\nabla_1\tilde{g}_i(0;x) = \nabla g_i(x)$, for every $x \in K$;

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 used in many recent papers. While we refer the reader to Facchinei et al. [32] and Scutari et al. [57] as good sources of examples, we nevertheless pause to consider some possible choices for the surrogate functions $\tilde{f}$ and $\tilde{g}$ and to make a few general considerations.

### 3.1. On the Choice of $\tilde{f}$ and $\tilde{g}$

The direction finding subproblem $(\mathrm{P}_x)$ is a direct generalization of traditional SQP subproblems and, in particular, of the subproblems considered in Burke [11]. The most classical choice for $\tilde{f}$ and $\tilde{g}$ is

$$\tilde{f}(d;x) \triangleq \nabla f(x)^T d + \frac{1}{2} d^T H(x) d, \qquad \tilde{g}(d;x) \triangleq g(x) + \nabla g(x)^T d, \tag{10}$$

where $H(x)$ is a positive definite symmetric matrix. With this choice, Assumption A can be easily satisfied provided that, classically, the smallest eigenvalue of the positive definite matrix $H(x)$ is uniformly bounded away from zero. If we use the surrogates (10) in $(\mathrm{P}_x)$, and assuming $K = \mathbb{R}^n$, $(\mathrm{P}_x)$ becomes the more classical SQP-type subproblem

$$\begin{aligned} \underset{d}{\text{minimize}}\ & f(x) + \nabla f(x)^T d + \frac{1}{2} d^T H(x) d \\ \text{s.t.}\ & g(x) + \nabla g(x)^T d \le \kappa(x) \\ & \|d\|_\infty \le \beta. \end{aligned}$$

Regarding $\kappa$, we remark that, with the classical choice given in (10), problem (9) in its definition reduces to a linear program if $K$ is polyhedral and thus can be efficiently solved.

Although the previous discussion shows that we can cover classical SQP schemes using linear/quadratic approximations, it is interesting to at least hint at how the flexibility allowed by Assumption A can be exploited to define better approximations to the original problem (P). Suppose for example that $f(x) = f_1(x) + f_2(x)$ with both functions $C^{1,1}$, but with $f_1$ convex and $f_2$ not necessarily so. Instead of approximating the whole function with a quadratic model, we could well *preserve* the convex part and only approximate the nonconvex one, therefore setting

$$\tilde{f}(d;x) = f_1(x+d) + f_2(x) + \nabla f_2(x)^T d + \frac{1}{2} d^T H(x) d,$$

with $H(x)$ as before. It is clear that this $\tilde{f}$ satisfies Assumption A and is presumably a better approximation to $f$ than $\nabla f(x)^T d + \frac{1}{2} d^T H(x) d$ considered previously.

As a further example, assume that $f$ is the product of two functions $f_1(x) f_2(x)$ with $f_1$ and $f_2$ convex and (for simplicity of presentation) positive. This is a rather frequent case in applications (Scutari et al. [56]). Because we have $\nabla f(x) = f_2(x) \nabla f_1(x) + f_1(x) \nabla f_2(x)$, it seems rather natural to set

$$\tilde{f}(d;x) = f_2(x) f_1(x+d) + f_1(x) f_2(x+d) + \frac{1}{2} d^T H(x) d,$$

which, again, should result in a sensibly tailored approximation that preserves part of the structure of the objective function.

Of course, an underlying assumption of our approach is that subproblem $(\mathrm{P}_x)$ 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 because it permits one to exploit any potentially favorable structure in problem (P) and, in any case, to better tailor the subproblems to the original problem (Beck et al. [4], Facchinei et al. [32], Lipp and Boyd [39], Scutari et al. [57], Sun et al. [61], Svanberg [62]). This use is also motivated by the possibility to solve efficiently more complex subproblems than the classical quadratic ones, sometimes even in closed form, and by the desire for faster convergence behaviors (see Facchinei et al. [32], Hong et al. [36], Mairal [43], Martinez [44], Razaviyayn et al. [51], Scutari et al. [56], Sun et al. [61], and references therein).

Among all the possible choices for the approximating functions, the case where we take the $\tilde{g}_i$s to be Upper Convex Approximations (UCA) of $g_i$s is worth to be pointed out. More precisely, suppose that, in addition to Assumption A, for every $x \in K$, we have

$$\tilde{g}_i(d;x) \ \ge\ g_i(x+d), \quad \forall i = 1, \dots, m, \quad \forall d \in K - x. \tag{11}$$

The main consequence of this choice is that if $x \in \mathcal{X}$, then $0 \in \widetilde{\mathcal{X}}(x)$ and, by (11), $x + \widetilde{\mathcal{X}}(x) \subseteq \mathcal{X}$. This means that if $x^\nu \in \mathcal{X}$ and, according to (2), we set $x^{\nu+1} = x^\nu + \gamma^\nu d(x^\nu)$ with $\gamma^\nu \in (0,1]$ (a condition that will always be satisfied by all algorithms considered in this paper), also $x^{\nu+1}$ is feasible, that is, $x^{\nu+1} \in \mathcal{X}$. This simple observation has important algorithmic ramifications that will be explored further in the next three sections. The main issue if one wants to use UCAs is finding suitable majorants. It turns out this can be done in a host of situations; the interested reader can find a very rich array of examples in Facchinei et al. [32], Hong et al. [36], Hunder and

Lange [37], Razaviyayn et al. [51], Scutari et al. [55], Scutari et al. [56], and Sun et al. [61]. Here, we just consider two examples.

The simplest case is possibly the one in Auslender et al. [2], where a feasible SQP-like approach is developed that rests on the assumption (among others) that the $g_i$ have Lipschitz continuous gradients (with constant $L_{\nabla g_i}$) and the descent lemma is used for defining suitable majorants as

$$\tilde{g}_i(d;x) = g_i(x) + \nabla g_i(x)^T d + \frac{a}{2}\|d\|^2. \tag{12}$$

By taking $a \geq L_{\nabla g_i}$, (12) provides an UCA for $g_i$. A second example of a case in which we can very easily build majorants is when the function has a DC structure. Specifically, suppose that $g_i = g_i^+ - g_i^-$, with both $g_i^+$ and $g_i^-$ convex. In this case we can build an upper convex approximation by setting

$$\tilde{g}_i(d;x) = g_i^+(x+d) - \left(g_i^-(x) + \nabla g_i^-(x)^T d\right).$$

## 3.2. Main Properties of (P$_x$)

In this section, we state the main properties of subproblem (P$_x$). Starting from feasibility, we remark, as already mentioned, that the term $\kappa(x)$ plays a key role in guaranteeing that our subproblems (P$_x$) have a nonempty feasible set $\widetilde{\mathcal{X}}(x)$. Because $\kappa(x)$ 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 (Figure 1).

In fact, the feasible set of problem (P$_x$), for every $x \in K$, is nonempty: choosing $\hat{d}$ at which the minimum in the expression of $\kappa(x)$ is attained, we have

$$\tilde{g}\left(\hat{d};x\right) \leq \min_d\left\{\max_i\{\tilde{g}_i(d;x)_+\} \,|\, \|d\|_\infty \leq \rho,\ d \in K - x\right\}e = \max_i\left\{\tilde{g}_i\left(\hat{d};x\right)_+\right\}e,$$

and, in turn,

$$\begin{aligned}\tilde{g}\left(\hat{d};x\right) &= (1-\lambda)\tilde{g}\left(\hat{d};x\right) + \lambda\tilde{g}\left(\hat{d};x\right)\\ &\leq \left[(1-\lambda)\max_i\{\tilde{g}_i(0;x)_+\} + \lambda\min_d\left\{\max_i\{\tilde{g}_i(d;x)_+\} \,|\, \|d\|_\infty \leq \rho, d \in K - x\right\}\right]e = \kappa(x)e.\end{aligned}$$

In Lemma 1, we establish some preliminary properties concerning the feasible set of problem (P$_x$).

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

i. *for every $\hat{x} \in K$, and for every $\alpha > 0$ and $d \in \alpha\mathbb{B}^n_\infty \cap (K - \hat{x})$, the constraint qualification*

$$\left[-N_{\alpha\mathbb{B}^n_\infty}(d)\right] \cap N_{K-\hat{x}}(d) = \{0\} \tag{13}$$

*holds and, in turn, $N_{\alpha\mathbb{B}^n_\infty \cap (K-\hat{x})}(d) = N_{\alpha\mathbb{B}^n_\infty}(d) + N_{K-\hat{x}}(d)$;*

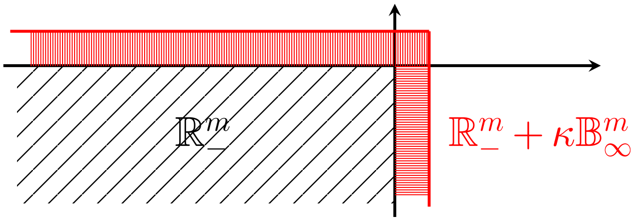
ii. *for every $\alpha > 0$, the set-valued mapping $\alpha\mathbb{B}^n_\infty \cap (K - \bullet)$ is continuous on $K$ relative to $K$;*

iii. *letting $C \triangleq \{(d,x) \in \beta\mathbb{B}^n_\infty \times K : d + x \in K\}$, the set-valued mapping $N_{\beta\mathbb{B}^n_\infty \cap (K-\bullet)}(\bullet)$ is outer semicontinuous on $C$ relative to $C$.*

**Proof.** (i) Let $0 \neq \eta \in [-N_{\alpha\mathbb{B}^n_\infty}(d)] \cap N_{K-\hat{x}}(d)$. Because of the convexity of the sets $\alpha\mathbb{B}^n_\infty$ and $K - \hat{x}$, we have $-\eta^T(v - d) \leq 0$ $\forall v \in \alpha\mathbb{B}^n_\infty$ and $\eta^T(y - d) \leq 0$ $\forall y \in (K - \hat{x})$. Choosing $y = 0 \in (K - \hat{x})$, one gets the following contradiction:

$$0 < \alpha \max_v\{-\eta^T v \,|\, v \in \mathbb{B}^n_\infty\} \leq -\eta^T d \leq 0,$$

**Figure 1.** (Color online) From $\mathbb{R}^m_-$ to $\mathbb{R}^m_- + \kappa\mathbb{B}^m_\infty$: The enlargement in the feasible region of (P$_x$).



thus proving relation (13). As a consequence, the other claim in (i) follows from Rockafellar and Wets [54, theorem 6.42].

ii. The property holds because of the continuity (relative to $K$) of the set-valued mapping $K - \bullet$ at every $x \in K$ and the fact that $\alpha\mathbb{B}^n_\infty \cap (K - x) \neq \emptyset$ for every $x \in K$.

iii. Suppose by contradiction that $(d^\nu, x^\nu) \xrightarrow[C]{} (\bar d, \bar x)$, $\eta^\nu \in N_{\beta\mathbb{B}^n_\infty \cap (K - x^\nu)}(d^\nu)$, $\eta^\nu \longrightarrow \bar\eta$ with $\bar\eta \notin N_{\beta\mathbb{B}^n_\infty \cap (K - \bar x)}(\bar d)$. Hence, $\bar z \in \beta\mathbb{B}^n_\infty \cap (K - \bar x)$ exists such that $\bar\eta^T(\bar z - \bar d) > 0$. By the inner semicontinuity relative to $K$ (see [54] for the definition of inner semicontinuity) of $\beta\mathbb{B}^n_\infty \cap (K - \bullet)$ at $\bar x$, $z^\nu$ exists such that $z^\nu \longrightarrow \bar z$ and $z^\nu \in \beta\mathbb{B}^n_\infty \cap (K - x^\nu)$. In turn, eventually we get $(\eta^\nu)^T(z^\nu - d^\nu) > 0$ in contradiction to the inclusion $\eta^\nu \in N_{\beta\mathbb{B}^n_\infty \cap (K - x^\nu)}(d^\nu)$. □

The function $\kappa(x)$ is obviously continuous and, under a very weak additional requirement, also locally Lipschitz continuous. This result has been shown in Burke [11] whenever $\tilde g$ is the linear approximation in (10) and readily generalizes to the case of the surrogate $\tilde g$ we consider here.

**Proposition 2.** *Under Assumption A, $\kappa(\bullet)$ is continuous on $K$ relative to $K$. If, in addition, $\tilde g(\bullet;\bullet)$ is locally Lipschitz continuous on $O_d \times O_x$, then $\kappa(\bullet)$ is also locally Lipschitz continuous on $K$.*

**Proof.** The continuity of $\kappa(\bullet)$ follows readily from the continuity (relative to $K$) of the set-valued mapping $\rho\mathbb{B}^n_\infty \cap (K - \bullet)$ at every $x \in K$: this in turn follows from (ii) in Lemma 1 with $\alpha = \rho$.

The Lipschitz continuity under the additional condition derives from Rockafellar [53, theorem 3.1]. Suffice it to observe that the constraint qualification (13) with $\alpha = \rho$ holds for every $x \in K$ and $d \in \rho\mathbb{B}^n_\infty \cap K - x$, and the problem in the definition of $\kappa$ is solvable for every $x$ in a neighborhood of $K$. □

The local Lipschitz continuity of $\tilde g(\bullet;\bullet)$ is part of Assumption C to be introduced shortly.

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

**Lemma 2.** *Under Assumption A, the following results hold for any $\hat x \in K$:*

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

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

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

**Proof.** (i) Choosing $\hat d \in \operatorname{argmin}_d\{\max_i\{\tilde g_i(d;\hat x)_+\} \mid \|d\|_\infty \le \rho,\ d \in K - \hat x\}$, we can infer $\tilde g(\hat d;\hat x) \le \min_d\{\max_i\{\tilde g_i(d;\hat x)_+\} \mid \|d\|_\infty \le \rho, d \in K - \hat x\}e$, while $\tilde g(\hat d;\hat x) \le \kappa(\hat x)e < \max_i\{g_i(\hat x)_+\}e$ and thus,

$$\tilde g\left(\hat d;\hat x\right) = \lambda \tilde g\left(\hat d;\hat x\right) + (1 - \lambda)\tilde g\left(\hat d;\hat x\right) < \kappa(\hat x)e,$$

with $\hat d \in \rho\mathbb{B}^n_\infty \cap (K - \hat x)$. The claim follows by continuity because $\rho < \beta$.

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 (13) with $\alpha = \rho$, A7, and A9.

iii. With $\max_i\{g_i(\hat x)_+\}$ being equal to zero, we have $\kappa(\hat x) = 0$ and $g(\hat x) \le 0$. 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 $j \in \{1, \dots, m\}$ such that $g_j(\hat x) < 0$, we have $\tilde g_j(0;\hat x) = g_j(\hat x) < 0$; as for indices $k \in \{1, \dots, m\}$ with $g_k(\hat x) = 0$, by (6), there exists $\hat d \in T_K(\hat x)$ such that

$$0 > \nabla g_k(\hat x)^T \hat d = \nabla_1 \tilde g_k(0;\hat x)^T \hat d = \lim_{\tau \downarrow 0} \frac{\tilde g_k\left(\tau \hat d;\hat x\right) - \tilde g_k(0;\hat x)}{\tau}.$$

Thus, there exists a sequence $\{d^\nu\}$ of feasible directions for $K$ at $\hat x$ such that $d^\nu \in T_K(\hat x)$ and $d^\nu \longrightarrow \hat d$. Choosing $\tau^\nu$ sufficiently small, we get $\hat x + \tau^\nu d^\nu \in K$ for every $\nu$ and the claim follows by continuity, observing that $\tilde g_i(\tau\hat d;\hat x) < 0$ for every $i$ and for any $\tau$ sufficiently small. □

The quantity

$$\theta(x) \triangleq \max_i\{g_i(x)_+\} - \kappa(x) = \lambda\left(\max_i\{g_i(x)_+\} - \min_d\left\{\max_i\{\tilde g_i(d;x)_+\} \mid \|d\|_\infty \le \rho,\ d \in K - x\right\}\right) \tag{14}$$

plays a key role in the previous lemma and in all the subsequent developments. As shown in the following proposition, $\theta$ turns out to be a stationarity measure for the violation-of-the-constraints problem (4).

**Proposition 3.** *Under Assumption A,*

i. *the nonnegative function $\theta(\bullet)$ is continuous on $K$ relative to $K$;*

ii. *$\theta(\hat{x}) = 0$ if and only if $\hat{x}$ is a stationary point for problem (4);*

iii. *we have, for every $x \in K$,*

$$\theta(x) \le \|\nabla g(x)\|_\infty \|d(x)\|. \tag{15}$$

**Proof.** (i) By the definition (8) of $\kappa$, $\kappa(\hat{x}) \le \max_i\{g_i(\hat{x})_+\}$ because $d = 0$ is feasible for the minimization problem in (8) and A7 holds. The continuity follows from Proposition 2.

ii. It is also clear that at any feasible point $\hat{x}$ for (P), $\theta(\hat{x}) = 0$; of course, every feasible point for (P) is stationary for problem (4). Consider now an infeasible point $\hat{x}$ for (P) and suppose that $\theta(\hat{x}) = 0$. By (ii) in Lemma 2, $\hat{x}$ turns out to be an ES point for (P). Hence, we are left to show that if $\hat{x}$ is an ES point for (P), then $\theta(\hat{x}) = 0$. For $\hat{x}$ to be ES, it is necessary and sufficient (see condition (5)) to have $M_0(\hat{x}) \ne \{0\}$ which in turn, by the Motzkin's alternative theorem (Craven [20, 2.5.2]), holds if and only if

$$\nexists\, d \in T_K(\hat{x}) \,:\, \nabla g_i(\hat{x})^T d < 0,\ \forall i \in I_+(\hat{x}) \triangleq \left\{ i \,:\, g_i(\hat{x}) = \max_j \{g_j(\hat{x})_+\} \right\}. \tag{16}$$

Suppose by contradiction that $\theta(\hat{x}) > 0$. Then, noting that $d \in K - \hat{x}$ implies $d \in T_K(\hat{x})$, Lemma 2(i) states that $d \in T_K(x)$ exists such that $\tilde{g}_i(d; \hat{x}) < \kappa(\hat{x})$ for all $i \in I_+(\hat{x})$. However, using A5, A7, and A9, we can write, for every $i \in I_+(\hat{x})$,

$$\max_i \{g_i(\hat{x})_+\} \;>\; \kappa(\hat{x}) \;>\; \tilde{g}_i(d; \hat{x}) \;\ge\; \tilde{g}_i(0; \hat{x}) + \nabla_1 \tilde{g}_i(0; \hat{x})^T (d - 0) \;\ge\; g_i(\hat{x}) + \nabla g_i(\hat{x})^T d.$$

Because $i \in I_+(\hat{x})$, this implies $\nabla g_i(\hat{x})^T d < 0$, contradicting (16).

(iii) Furthermore,

$$\begin{aligned}
0 \le \theta(x^\nu) &= \max_i \{g_i(x^\nu)_+\} - \kappa(x^\nu) \overset{(a)}{\le} \max_i \{g_i(x^\nu)_+\} - \max_i \{\tilde{g}_i(d(x^\nu); x^\nu)_+\} \\
&\overset{(b)}{\le} \max_i \{g_i(x^\nu)_+\} - \max_i \{(g_i(x^\nu) + \nabla g_i(x^\nu)^T d(x^\nu))_+\} \\
&\overset{(c)}{\le} \max_i \{(g_i(x^\nu) - g_i(x^\nu) - \nabla g_i(x^\nu)^T d(x^\nu))_+\} \le \|\nabla g(x^\nu)^T d(x^\nu)\|_\infty \le \|\nabla g(x^\nu)\|_\infty \|d(x^\nu)\|,
\end{aligned}$$

where (a) holds because $\tilde{g}(d(x^\nu); x^\nu) \le \kappa(x^\nu) e$, (b) is because of A5, A7, and A9, and (c) follows observing that $\max\{0, \alpha_1\} - \max\{0, \alpha_2\} \le \max\{0, \alpha_1 - \alpha_2\}$ for any $\alpha_1, \alpha_2 \in \mathbb{R}$. □

Leveraging Lemma 2 and resorting to standard results in parametric optimization, we can establish a key continuity property for the solution mapping $d(\bullet)$ of subproblem ($\mathrm{P}_x$).

**Proposition 4.** *Under Assumption A, let the eMFCQ hold at $\hat{x} \in K$ for problem (P). Then,*

i. *the MFCQ holds at every point of $\widetilde{\mathcal{X}}(\hat{x})$ for subproblem ($\mathrm{P}_{\hat{x}}$);*

ii. *a neighborhood $\mathcal{V}$ of $\hat{x}$ exists such that, for every point $x \in K \cap \mathcal{V}$, the mapping $d(\bullet)$ is continuous relative to $K$.*

**Proof.** If the eMFCQ holds at $\hat{x}$ for problem (P), case (ii) in Lemma 2 cannot occur. On the other hand, as for both cases (i) and (iii) in Lemma 2, Slater's constraint qualification holds for $\widetilde{\mathcal{X}}(\hat{x})$ and, since $\widetilde{\mathcal{X}}(\hat{x})$ is convex, this proves (i). Because of A6 and (ii) in Lemma 1, the set-valued mapping $\widetilde{\mathcal{X}}(\bullet) = [\beta \mathbb{B}^n_\infty \cap (K - \bullet)] \cap \{d \in \mathbb{R}^n \,:\, \tilde{g}(d; \bullet) \le \kappa(\bullet) e\}$ is outer semicontinuous at $\hat{x}$ relative to $K$ by Bank et al. [3, theorem 3.1.1], having taken into account that $\kappa(\bullet)$ is continuous by Proposition 2. Moreover, $\widetilde{\mathcal{X}}(\bullet)$, by virtue of the Slater's constraint qualification, A5, A6, and (ii) in Lemma 1, is also inner semicontinuous (Bank et al. [3, theorem 3.1.6]) at $\hat{x}$ relative to $K$. Hence, thanks to A1, the continuity (relative to $K$) of $d(\bullet)$, leveraging Bank et al. [3, theorems 3.1.1 and 4.3.3], follows from Rockafellar and Wets [54, corollary 5.20]. □

To prove some refinements of the 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 any compact set $S \subseteq K$, two positive constants $\theta$ and $\alpha$ exist such that

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

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

**Assumption C.** *The following results hold:*

C1. the partial gradient $\nabla_1 \tilde{f}(\bullet;\bullet)$ is locally Lipschitz continuous on $O_d \times O_x$;

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

The following proposition, which builds on the results in Yen [66], shows the desired result.

**Proposition 5.** *Under Assumptions A and C, let $S \subseteq K$ be compact. Suppose further that the eMFCQ holds at every $\hat{x} \in S$. Then, there exists $\theta > 0$ such that, for every $y, z \in S$,*

$$\|d(y) - d(z)\| \le \theta \|y - z\|^{\frac{1}{2}}. \tag{17}$$

**Proof.** Preliminarily, observe that by Proposition 2, $\kappa(\bullet)$ is locally Lipschitz continuous. Furthermore, by Proposition 4(i), we have that the MFCQ holds at every point in $\widetilde{\mathcal{X}}(\hat{x})$ and, in particular at $d(\hat{x})$. In turn, the MFCQ at $d(\hat{x}) \in \widetilde{\mathcal{X}}(\hat{x})$, for every $\hat{x} \in S$, implies, by Rockafellar [53, theorem 3.2], that the set-valued mapping $\widetilde{\mathcal{X}}$ has the Aubin property relative to $S$ at $\hat{x}$ for $d(\hat{x})$ for every $\hat{x} \in S$ (see Rockafellar and Wets [54] for the definition of the Aubin property). Therefore, in view of Yen [66, theorem 2.1], for every $\hat{x} \in S$, there exist $\theta' > 0$, $\theta'' > 0$ and a neighborhood $\mathcal{V}$ of $\hat{x}$ such that, for every $y$, $z \in \mathcal{V} \cap S$

$$\|d(y) - d(z)\| \le \theta' \|y - z\| + \theta'' \|y - z\|^{\frac{1}{2}}.$$

By the previous relation and the compactness of set $S$, (17) 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 (10) of standard SQP methods. Assumption C reinforces some of the requirements in Assumption A; we refer the reader to Facchinei et al. [32] 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 $\|d\|_\infty \le \beta$ corresponds to $2n$ bounds of the type $-\beta \le d_i \le \beta$. However, in what follows, we are interested only in the multipliers corresponding to the constraints $\tilde{g}(d;x) \le \kappa(x)e$, and therefore we find it expedient to write the KKT conditions as

$$0 \in \nabla_1 \tilde{f}(d(x);x) + \nabla_1 \tilde{g}(d(x);x)\xi + N_{\beta\mathbb{B}_\infty^n \cap (K-x)}(d(x)),$$

with the KKT multipliers $\xi$ satisfying the conditions $\xi \ge 0$ and $\xi^T(\tilde{g}(d(x);x) - \kappa(x)e) = 0$. We now establish the local boundedness of these KKT multipliers.

**Proposition 6.** *Under Assumption A, let $\hat{x}$ belong to $K$ and suppose that $\hat{d} \in \beta\mathbb{B}_\infty^n \cap (K - \hat{x})$ exists such that $\tilde{g}(\hat{d};\hat{x}) < \kappa(\hat{x})e$. Then, a neighborhood $\mathcal{V}$ of $\hat{x}$ exists such that, for every point $x \in K \cap \mathcal{V}$, the unique solution $d(x)$ of (P$_x$) is a KKT point for problem (P$_x$) and the set-valued mapping of the KKT multipliers is locally bounded at $\hat{x}$ relative to $K$.*

**Proof.** The condition $\tilde{g}(\hat{d};\hat{x}) < \kappa(\hat{x})e$ with $\hat{d} \in \beta\mathbb{B}_\infty^n \cap (K - \hat{x})$ is nothing else but the Slater's CQ for problem (P$_{\hat{x}}$), which obviously implies that the MFCQ holds at the unique solution of problem (P$_{\hat{x}}$). The derivation of the result is then rather classical and follows easily from Facchinei and Pang [31, proposition 5.4.3] taking into account Lemma 1(ii), Propositions 2 and 4, and the outer semicontinuity of $N_{\beta\mathbb{B}_\infty^n \cap (K-\bullet)}(\bullet)$ (see Lemma 1(iii)). □

## 4. Convergence of DSMs

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

**Algorithm 1.** **DSM Algorithm for (P)**

```
    Data: $\gamma^\nu \in (0,1]$ such that (3) holds, $x^0 \in K$, $\nu \longleftarrow 0$
    repeat
(S.1)   if $x^\nu$ is generalized stationary for (P) then
            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 $x^{\nu+1} = x^\nu + \gamma^\nu d(x^\nu)$, $\nu \longleftarrow \nu + 1$;
    end
```

The algorithm is always well defined if Assumption A, which guarantees existence and uniqueness of $d(x^\nu)$, holds. The main (and essentially only) computational burden is given by the computation of $\kappa(x^\nu)$ and the solution of the strongly convex subproblem $(\mathrm{P}_{x^\nu})$. This difficulty 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 establishes the main convergence properties of Algorithm 1. In a nutshell, the theorem shows that, unless $x^\nu$ is a generalized stationary point, $d(x^\nu)$ is a descent direction for $W(x^\nu;\varepsilon)$ if $\varepsilon$ is sufficiently small. Elaborating on this simple fact we can then show, without ever computing $W$ or actually determining a value for $\varepsilon$, that the sequence generated eventually lands on a generalized stationary point. The results in Theorem 1 do not exclude the possibility that Algorithm 1 generates an unbounded sequence. In Section 6, we discuss the meaning of this possible outcome and, more importantly, give several conditions under which we can guarantee that the sequence generated by Algorithm 1 (and also by the two algorithms we introduce in the next section) is bounded.

**Theorem 1.** *Consider the sequence $\{x^\nu\}$ generated by Algorithm 1 with $\tilde{f}$ and $\tilde{g}$ such that Assumption A holds. Then, the whole sequence $\{x^\nu\}$ is contained in $K$. Furthermore, either the sequence $\{x^\nu\}$ is unbounded or the following assertions hold:*

i. *at least one limit point $\hat{x}$ of $\{x^\nu\}$ is generalized 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 $\{x^\nu\}$, under Assumption B, every limit point of $\{x^\nu\}$ is a KKT solution for problem (P).*

**Proof.** Because the starting point $x^0$ belongs to the convex set $K$, the stepsize $\gamma^\nu \leq 1$ and, by the last constraint in $(\mathrm{P}_{x^\nu})$, $x^\nu + d(x^\nu) \in K$ for all $\nu$, it is easily seen that all points $x^\nu$ generated by the algorithm belong to $K$. We now assume, without loss of generality, that the sequence $\{x^\nu\}$ is bounded. Preliminarily, observe that, at each step, the solution $d(x^\nu)$ of subproblem $(\mathrm{P}_{x^\nu})$ is also a KKT point for $(\mathrm{P}_{x^\nu})$. In fact, suppose that at a certain iteration $\bar{\nu}$, $d(x^{\bar{\nu}})$ does not satisfy the KKT conditions for $(\mathrm{P}_{x^{\bar{\nu}}})$. The subproblem is always feasible by construction; let us analyze the three exhaustive cases considered in Lemma 2. In case (i), Slater's condition holds for $(\mathrm{P}_{x^{\bar{\nu}}})$ and $d(x^{\bar{\nu}})$ is a KKT point. In case (ii), $x^{\bar{\nu}}$ is an ES point of (P): hence, we would have stopped at step (S.1). In case (iii), either Slater's condition holds for $(\mathrm{P}_{x^{\bar{\nu}}})$ and $d(x^{\bar{\nu}})$ is a KKT point, or $x^{\bar{\nu}}$ is a FJ point for (P), in which case we would have stopped at step (S.1).

Thus, $d(x^\nu)$ is a KKT point for $(\mathrm{P}_{x^\nu})$ and multipliers $\{\xi^\nu\}$ exist such that $\xi^\nu \in N_{\mathbb{R}^m_-}(\tilde{g}(d(x^\nu);x^\nu) - \kappa(x^\nu)e)$ and

$$0 \in \nabla_1 \tilde{f}(d(x^\nu);x^\nu) + \nabla_1 \tilde{g}(d(x^\nu);x^\nu)\xi^\nu + N_{\beta\mathbb{B}^n_\infty \cap (K - x^\nu)}(d(x^\nu)). \tag{18}$$

Thanks to A1 and A4, we have

$$\nabla_1 \tilde{f}(d(x^\nu);x^\nu)^T d(x^\nu) = \left[\nabla_1 \tilde{f}(d(x^\nu);x^\nu) - \nabla_1 \tilde{f}(0;x^\nu) + \nabla_1 \tilde{f}(0;x^\nu)\right]^T d(x^\nu) \geq c\|d(x^\nu)\|^2 + \nabla f(x^\nu)^T d(x^\nu). \tag{19}$$

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

$$-\nabla_1 \tilde{g}_i(d(x^\nu);x^\nu)^T d(x^\nu) \leq \tilde{g}_i(0;x^\nu) - \tilde{g}_i(d(x^\nu);x^\nu), \tag{20}$$

and, by A7, because $\xi^\nu$ is nonnegative, in turn,

$$-\xi^\nu_i \nabla_1 \tilde{g}_i(d(x^\nu);x^\nu)^T d(x^\nu) \leq \xi^\nu_i\left[\tilde{g}_i(0;x^\nu) - \tilde{g}_i(d(x^\nu);x^\nu)\right] = \xi^\nu_i\left[g_i(x^\nu) - \kappa(x^\nu)\right], \tag{21}$$

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

Therefore, by (18), (19), and (21), we have, for some $\zeta^\nu \in N_{\beta\mathbb{B}^n_\infty \cap (K - x^\nu)}(d(x^\nu))$,

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

where the second inequality holds because of $0 \in \beta\mathbb{B}^n_\infty \cap (K - x^\nu)$. Therefore, recalling definition (14),

$$\nabla f(x^\nu)^T d(x^\nu) \leq -c\|d(x^\nu)\|^2 + \theta(x^\nu)\,\xi^{\nu T} e. \tag{22}$$

We also notice that, because $d(x^\nu)$ is feasible for problem $(\mathrm{P}_{x^\nu})$, by A5, A7, and A9,

$$\kappa(x^\nu) \geq \tilde{g}_i(d(x^\nu);x^\nu) \geq \tilde{g}_i(0;x^\nu) + \nabla \tilde{g}_i(0;x^\nu)^T d(x^\nu) = g_i(x^\nu) + \nabla g_i(x^\nu)^T d(x^\nu). \tag{23}$$

Let us now consider the nonsmooth (ghost) penalty function already described in the introduction

$$W(x;\varepsilon) \triangleq f(x) + \frac{1}{\varepsilon}\max_i\{g_i(x)_+\}, \tag{24}$$

with a positive penalty parameter $\varepsilon$. This function plays a key role in the subsequent convergence analysis, although it does not appear anywhere in the algorithm itself.

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

$$\begin{aligned}
&W(x^{\nu+1};\varepsilon) - W(x^\nu;\varepsilon)\\
&\quad = f\big(x^\nu + \gamma^\nu d(x^\nu)\big) - f(x^\nu) + \frac{1}{\varepsilon}\left[\max_i\{g_i\big(x^\nu+\gamma^\nu d(x^\nu)\big)_+\} - \max_i\{g_i(x^\nu)_+\}\right]\\
&\quad \overset{(a)}{\le} \gamma^\nu \nabla f(x^\nu)^T d(x^\nu) + \frac{(\gamma^\nu)^2 L_{\nabla f}}{2}\|d(x^\nu)\|^2 + \frac{1}{\varepsilon}\left[\max_i\{\big(g_i(x^\nu) + \gamma^\nu \nabla g_i(x^\nu)^T d(x^\nu)\big)_+\}\right.\\
&\qquad \left. - \max_i\{g_i(x^\nu)_+\} + \frac{(\gamma^\nu)^2 \max_i\{L_{\nabla g_i}\}}{2}\|d(x^\nu)\|^2\right]\\
&\quad \overset{(b)}{\le} \gamma^\nu \nabla f(x^\nu)^T d(x^\nu) + \frac{1}{\varepsilon}\left[\max_i\{(1-\gamma^\nu)g_i(x^\nu)_+ + \gamma^\nu \kappa(x^\nu)\} - \max_i\{g_i(x^\nu)_+\}\right]\\
&\qquad + \frac{(\gamma^\nu)^2}{2}\left(L_{\nabla f} + \frac{\max_i\{L_{\nabla g_i}\}}{\varepsilon}\right)\|d(x^\nu)\|^2\\
&\quad \le \gamma^\nu \nabla f(x^\nu)^T d(x^\nu) - \frac{\gamma^\nu}{\varepsilon}\left[\max_i\{g_i(x^\nu)_+\} - \kappa(x^\nu)\right] + \frac{(\gamma^\nu)^2}{2}\left(L_{\nabla f} + \frac{\max_i\{L_{\nabla g_i}\}}{\varepsilon}\right)\|d(x^\nu)\|^2\\
&\quad \le \gamma^\nu \nabla f(x^\nu)^T d(x^\nu) - \frac{\gamma^\nu}{\varepsilon}\,\theta(x^\nu) + \frac{(\gamma^\nu)^2}{2}\left(L_{\nabla f} + \frac{\max_i\{L_{\nabla g_i}\}}{\varepsilon}\right)\|d(x^\nu)\|^2,
\end{aligned} \tag{25}$$

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 $\gamma^\nu \le 1$ since, in view of (23), $\nabla g_i(x^\nu)^T d(x^\nu) \le \kappa(x^\nu) - g_i(x^\nu)$. Furthermore, we observe that

$$\nabla f(x^\nu)^T d(x^\nu) - \frac{1}{\varepsilon}\,\theta(x^\nu) \le -c\|d(x^\nu)\|^2 + \theta(x^\nu)\,\xi^{\nu T} e - \frac{1}{\varepsilon}\,\theta(x^\nu) \le -c\|d(x^\nu)\|^2 + \left(m\|\xi^\nu\|_\infty - \frac{1}{\varepsilon}\right)\theta(x^\nu), \tag{26}$$

where the first inequality is entailed by (22).

By (26), for any fixed $x^\nu$ and for any $\eta \in (0,1]$, there exists $\bar{\varepsilon}^\nu > 0$ such that

$$\nabla f(x^\nu)^T d(x^\nu) - \frac{1}{\varepsilon}\,\theta(x^\nu) \le -\eta c\|d(x^\nu)\|^2 \qquad \forall \varepsilon \in (0,\bar{\varepsilon}^\nu]. \tag{27}$$

We now distinguish two cases.

(I) Suppose that (27) does not hold uniformly for every $x^\nu$, that is $\eta \in (0,1]$, and a subsequence $\{x^\nu\}_{\mathcal{N}}$ exists, where $\mathcal{N} \subseteq \{0,1,2,\ldots\}$ such that we can construct a corresponding subsequence $\{\varepsilon^\nu\}_{\mathcal{N}} \in \mathbb{R}_+$ with $\varepsilon^\nu \downarrow 0$ on $\mathcal{N}$ and

$$\nabla f(x^\nu)^T d(x^\nu) - \frac{1}{\varepsilon^\nu}\,\theta(x^\nu) > -\eta c\|d(x^\nu)\|^2 \tag{28}$$

for every $\nu \in \mathcal{N}$. For (28) to hold, relying on (26), the multipliers' subsequence $\{\xi^\nu\}_{\mathcal{N}}$ must be unbounded. Combining (26) and (28), we get

$$0 \le c(1-\eta)\|d(x^\nu)\|^2 < \left(m\|\xi^\nu\|_\infty - \frac{1}{\varepsilon^\nu}\right)\theta(x^\nu),$$

and, thus, $\theta(x^\nu) > 0$ for every $\nu \in \mathcal{N}$. By the previous relation and (28), we also have

$$\frac{1}{\varepsilon^\nu} < \frac{\nabla f(x^\nu)^T d(x^\nu) + \eta c \|d(x^\nu)\|^2}{\theta(x^\nu)}. \tag{29}$$

As $\varepsilon^\nu \downarrow 0$ on $\mathcal{N}$, the right-hand side of (29) goes to infinity: by the boundedness of the numerator,

$$\theta(x^\nu) \underset{\mathcal{N}}{\longrightarrow} 0. \tag{30}$$

Let $\hat{x}$ be a cluster point of the subsequence $\{x^\nu\}_\mathcal{N}$. By (30), only cases (ii) and (iii) in Lemma 2 can occur at $\hat{x} \in K$. The existence of a $d$ as stipulated in Lemma 2 (iii) would entail, by Proposition 6, the boundedness of the KKT multipliers $\xi^\nu$ for $\nu \in \mathcal{N}$ large enough, thus giving a contradiction. Therefore, by Lemma 2 (ii), 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 (27) holds uniformly for every $x^\nu$: that is, for any $\eta \in (0, 1]$, there exists $\bar{\varepsilon} > 0$ such that

$$\nabla f(x^\nu)^T d(x^\nu) - \frac{1}{\varepsilon}\, \theta(x^\nu) \le -\eta c \|d(x^\nu)\|^2 \quad \forall \varepsilon \in (0, \bar{\varepsilon}], \;\; \forall \nu. \tag{31}$$

Combining relations (25) and (31), we get

$$\begin{aligned} W\big(x^{\nu+1}; \tilde{\varepsilon}\big) - W(x^\nu; \tilde{\varepsilon}) &\le -\gamma^\nu \eta c \|d(x^\nu)\|^2 + \frac{(\gamma^\nu)^2}{2}\left(L_{\nabla f} + \frac{\max_i\{L_{\nabla g_i}\}}{\tilde{\varepsilon}}\right)\|d(x^\nu)\|^2 \\ &= -\gamma^\nu\left[\eta c - \frac{\gamma^\nu}{2}\left(L_{\nabla f} + \frac{\max_i\{L_{\nabla g_i}\}}{\tilde{\varepsilon}}\right)\right]\|d(x^\nu)\|^2, \end{aligned} \tag{32}$$

for any $\tilde{\varepsilon} \in (0, \bar{\varepsilon}]$. Because $\lim_\nu \gamma^\nu = 0$, there exists a positive constant $\omega$ such that, by (32), for $\nu \ge \bar{\nu}$ sufficiently large,

$$W\big(x^{\nu+1}; \tilde{\varepsilon}\big) - W(x^\nu; \tilde{\varepsilon}) \le -\omega \gamma^\nu \|d(x^\nu)\|^2. \tag{33}$$

With $W$ being bounded from below, by (33), the sequence $\{W(x^\nu; \tilde{\varepsilon})\}$ converges and

$$\lim_\nu \sum_{t=\bar{\nu}}^{\nu} \gamma^t \|d(x^t)\|^2 < +\infty.$$

Therefore, because $\sum_{\nu=0}^\infty \gamma^\nu = +\infty$, we have

$$\liminf_{\nu\to\infty} \|d(x^\nu)\| = 0. \tag{34}$$

Recalling relation (15), taking the limit on a subsequence $\mathcal{N}$ such that $\|d(x^\nu)\| \underset{\mathcal{N}}{\longrightarrow} 0$, we have $\|\nabla g(x^\nu)\|_\infty \|d(x^\nu)\| \underset{\mathcal{N}}{\longrightarrow} 0$ and $\theta(x^\nu) \underset{\mathcal{N}}{\longrightarrow} 0$. Finally, let again $\hat{x}$ be a cluster point of subsequence $\{x^\nu\}_\mathcal{N}$. Because $\theta(x^\nu) \underset{\mathcal{N}}{\longrightarrow} 0$ implies $\kappa(\hat{x}) = \max_i\{g_i(\hat{x})_+\}$, cases (ii) or (iii) in Lemma 2 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 2 is ruled out and $\max_i\{g_i(\hat{x})_+\}$ cannot be strictly positive; then, $\kappa(\hat{x}) = \max_i\{g_i(\hat{x})_+\} = 0$. Furthermore, taking the limit in (18), we obtain, by A3, A4, A6–A9, KKT multipliers' boundedness, and the outer semicontinuity property (see Lemma 1 (iii)) of the normal cone mapping $N_{\beta\mathbb{B}_\infty^n \cap (K - \bullet)}(\bullet)$,

$$-\nabla f(\hat{x}) - \nabla g(\hat{x})\hat{\xi} \in N_{\beta\mathbb{B}_\infty^n \cap (K - \hat{x})}(0) = \{0\} + N_{K-\hat{x}}(0) = N_K(\hat{x}),$$

with $\hat{\xi} \in N_{\mathbb{R}_-^m}(g(\hat{x}) - \kappa(\hat{x})e) = N_{\mathbb{R}_-^m}(g(\hat{x}))$ and where the first equality follows from Lemma 1 (i). In turn, $\hat{x}$ is a KKT point for problem (P). This concludes the proof of case (i).

Consider now point (ii). If the eMFCQ holds at every limit point of $\{x^\nu\}$, then case (I) cannot occur because this would contradict the last sentence before (II); hence, we are in case (II). Observe that if, instead of the weaker (34),

$$\lim_{\nu\to\infty} \|d(x^\nu)\| = 0 \tag{35}$$

holds, we can reason similarly to what done above after (34) for any convergent subsequence of $\{x^\nu\}$ and conclude that (ii) holds. Therefore, it is enough to show that Assumption B entails (35).

Consider now the compact set containing all iterates $x^\nu$. Although $\liminf_{\nu\to\infty}\|d(x^\nu)\| = 0$, 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 $\nu$s. 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:

$$\|d(x^\nu)\| < \delta/2, \|d(x^{i_\nu})\| > \delta \tag{36}$$

and, if $i_\nu > \nu + 1$,

$$\delta/2 \le \|d(x^j)\| \le \delta, \nu < j < i_\nu. \tag{37}$$

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

$$\begin{aligned}\delta/2 < \|d(x^{i_\nu})\| - \|d(x^\nu)\| \;\le\; \|d(x^{i_\nu}) - d(x^\nu)\| \overset{(a)}{\le} \theta\|x^{i_\nu} - x^\nu\|^\alpha \\ \overset{(b)}{\le} \theta\left[\sum_{t=\nu}^{i_\nu-1}\gamma^t\|d(x^t)\|\right]^\alpha \overset{(c)}{\le} \theta\delta^\alpha\left(\sum_{t=\nu}^{i_\nu-1}\gamma^t\right)^\alpha,\end{aligned} \tag{38}$$

where (a) is because of 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 (37). By (38), we have

$$\liminf_{\nu\to\infty}\;\theta\delta^\alpha\left(\sum_{t=\nu}^{i_\nu-1}\gamma^t\right)^\alpha > 0. \tag{39}$$

We prove next that (39) is in contradiction with the convergence of $\{W(x^\nu;\tilde\varepsilon)\}$ for any $\tilde\varepsilon \in (0,\bar\varepsilon]$, where $\bar\varepsilon$ is defined around (31). To this end, we first show that $\|d(x^\nu)\| \ge \delta/4$, for sufficiently large $\nu \in \mathcal{N}$. Reasoning as in (38), we have

$$\|d(x^{\nu+1})\| - \|d(x^\nu)\| \le \theta\|x^{\nu+1} - x^\nu\|^\alpha \le \theta(\gamma^\nu)^\alpha\|d(x^\nu)\|^\alpha,$$

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 (37) (or (36)) would be violated. Then, it must be $\|d(x^\nu)\| \ge \delta/4$. From this, and using (33), we have, for sufficiently large $\nu \in \mathcal{N}$,

$$W(x^{i_\nu};\tilde\varepsilon) \le W(x^\nu;\tilde\varepsilon) - \omega\sum_{t=\nu}^{i_\nu-1}\gamma^t\|d(x^t)\|^2 \le W(x^\nu;\tilde\varepsilon) - \omega\frac{\delta^2}{16}\sum_{t=\nu}^{i_\nu-1}\gamma^t. \tag{40}$$

Because $\{W(x^\nu;\tilde\varepsilon)\}$ converges, as established immediately after (33), renumbering if necessary, relation (40) implies $\sum_{t=\nu}^{i_\nu-1}\gamma^t \to 0$, in contradiction with (39). This shows that (35) holds and concludes the proof of the theorem. □

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 (Burke [11], Burke [12], Burke and Han [13], Facchinei [30]). A key difference between our approach and those in Burke [11], Burke [12], Burke and Han [13], and Facchinei [30] is that we do not use any penalty parameter in the algorithm. Indeed, we use the penalty function and penalty parameter only in the proof of Theorem 1, as a tool of theoretical analysis, and thus we do not need to calculate any careful penalty parameter update, allowing for convergence for the conceptually simple procedure defined previously. We believe that this ghost approach is a novelty in the literature and represents an interesting use of penalty functions. Although our approach has some similarities to a classical Lyapunov function approach, it is different from it. Indeed, whereas, in case (II) considered in the proof, the penalty can be viewed as a Lyapunov function for the algorithm, the analysis of case (I) is rather different and more involved. Indeed the proof hinges on the behavior of the penalty function, of $\theta$, and of the penalty parameter and on how these quantities are connected.

**Remark 2.** All the developments in the proof of Theorem 1 up to Equation (32) 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 (3) are used.

**Remark 3.** Algorithm 1 can easily be made into a *feasible method*, that is, a method that only generates feasible iterates, if we take $\tilde{g}_i$s to be UCAs, see (11). As discussed in Section 3.1, in this case, if $x^\nu$ is feasible, then $x^{\nu+1}$ is also feasible, so that, if $x^0 \in \mathcal{X}$, Algorithm 1 generates only feasible iterates. This shows that Algorithm 1 contains as special cases some recent feasible methods that were shown to be rather effective (Facchinei [32], Scutari et al. [55], Scutari et al. [56]).

If, furthermore, we require $\tilde{f}$ to be an UCA for $f$, that is,

$$\tilde{f}(d;x) \geq f(x+d), \quad \forall d \in K - x, \tag{41}$$

we turn our scheme into a *majorization-minimization* (MM)-like method (Auslender et al. [2], Beck et al. [4], Bolte and Pauwels [9], Hong et al. [36], Hunter and Lange [37], Mairal [43], Razaviyayn et al. [51], Sun et al. [61]). In classical MM approaches the stepsize $\gamma^\nu$ is taken to be always one, whereas Algorithm 1 with UCAs for $f$ and $g_i$s gives a diminishing stepsize version of the method. In Section 5.3, we show that not only can we also guarantee convergence by setting the stepsize equal to one, but we can actually obtain, in this case, an iteration complexity result.

## 5. Iteration Complexity Analysis

We introduce some new rules for choosing the stepsize $\gamma^\nu$ at each iteration as an alternative to the diminishing one analyzed in the previous section. For these rules, we are able to perform a detailed iteration complexity analysis. Our analysis is in line with recent works on this topic (see Cartis et al. [18] for an up-to-date review). The purpose of the iteration complexity analysis is to give a bound on the number of iterations needed by an algorithm to reach a desired level of accuracy. This bound is expressed in terms of parameters of the algorithm and some *problem constants*, for example, Lipschitz moduli of the functions involved on a prescribed region or maximum or minimum values of the functions in the same region. This section is organized as follows. Theorem 2 gives our more general complexity result for Algorithm 2; Section 5.1 explores in detail the meaning of the stopping criteria used in Algorithm 2 and gives the definition of $\delta$-stationary point. The following three short sections examine some particular scenarios in which improved complexity bounds (or, in one case, global convergence rate) can be obtained. Finally, Section 5.5 describes a variant of Algorithm 2 that can be implemented and analyzed without any knowledge of any problem-related constants.

To perform our analysis in this section we make the following assumptions.

**Assumption D.** *The following conditions hold:*

D1. the set $K$ is bounded;

D2. the partial gradient $\nabla_1 \tilde{g}(\bullet;\bullet)$ is locally Lipschitz continuous on $O_d \times O_x$.

Assumption D1 serves to guarantee boundedness of the iterations and is made for simplicity of presentation. In Section 6, we shall discuss some alternative assumptions that make the iterates belong to a compact set defined by means of possibly known quantities, as required to perform a complexity analysis. As the discussion pertains to the algorithms presented in both the previous and current sections, we found that a detailed analysis of this condition is best deferred to not complicate the formal presentation of the results, as the insight involved is essentially modular, separate from the main ideas of analysis here and in Section 4.

Assumption D2, instead, depends essentially on the choice of $\tilde{g}$ and therefore is not an assumption on the problem itself, but a condition on our algorithmic choices. Clearly, because $g$ has a locally Lipschitz gradient, D2 is always satisfied if we take as $\tilde{g}$ the linearization of $g$.

From now on, we use some problem dependent constants: we collect their definitions in Table 1 for the reader's convenience.

We observe that if the eMFCQ holds everywhere in $K$, all generalized 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 (Theorem 3). However, if the eMFCQ is not valid at every point in $K$, we cannot rely solely on $\|d(x^\nu)\|$ to monitor progress toward stationarity, because the problem may admit KKT points but also FJ and ES solutions. For this reason, we use in combination $\|d(x^\nu)\|$ and $\theta(x^\nu)$ as measures of stationarity. We observe that, actually, $\|d(x^\nu)\|$ and $\theta(x^\nu)$ are linked to each other in view of the following relation, which is because of (15):

$$\theta(x^\nu) \leq \|\nabla g(x^\nu)\|_\infty \|d(x^\nu)\| \leq L\|d(x^\nu)\|, \tag{42}$$

**Table 1.** Problem-dependent constants.

| Constant | Definition |
|---|---|
| $\beta \in \mathbb{R}_+$ | User-set constant in the definition of $(\mathrm{P}_x)$ |
| $\eta \in (0,1]$ | User-set constant |
| $\lambda \in (0,1)$ | User-set constant in the definition of $\kappa$, see (8) |
| $c \in \mathbb{R}_+$ | Modulus of strong convexity of $\tilde{f}(\bullet; x)$, see Assumption A1 |
| $B$ | $\max_{x\in K} \|\nabla f(x)\|\beta + \eta c\beta^2$ |
| $f^m$ | $\min_{x\in K} f(x)$ |
| $g_+^M$ | $\max_{x\in K} \max_i\{g_i(x)_+\}$ |
| $L$ | $\max_{(d,x)\in\beta\mathbb{B}_\infty^n\times K} \|\nabla_1\tilde{g}(d;x)\|_\infty$ |
| $L_{\nabla f}$ | Lipschitz modulus of $\nabla f$ on $K$ |
| $L_{\nabla g_i}$ | Lipschitz modulus of $\nabla g_i$ on $K$ |
| $L_{\nabla \tilde{f}}$ | Lipschitz modulus of $\nabla_1\tilde{f}(\bullet;\bullet)$ on $\beta\mathbb{B}_\infty^n \times K$ |
| $L_{\nabla \tilde{g}}$ | Lipschitz modulus of $\nabla_1\tilde{g}(\bullet;\bullet)$ on $\beta\mathbb{B}_\infty^n \times K$ |

where $L \triangleq \max_{(d,x)}\{\|\nabla_1\tilde{g}(d;x)\|_\infty \,|\, (d,x) \in \beta\mathbb{B}_\infty^n \times K\}$. However, there is no reverse implication, and thus the two functions $\|d(x^\nu)\|$ and $\theta(x^\nu)$ must be suitably combined to provide reliable stopping criteria. The effect of monitoring both $\|d(x^\nu)\|$ and $\theta(x^\nu)$ on the outcome of the algorithm is analyzed in detail in Section 5.1.

To derive complexity results, we consider first 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, whereas everything between, from (S.2) to (S.6), is aimed at deciding whether to decrease the stepsize $\gamma^\nu$ and whether we should terminate (note that Algorithm 1, which is aimed at an asymptotic analysis, does not contain any practical stopping criterion).

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

```
      Data: $\delta > 0$, $\eta \in (0,1]$, $x^0 \in K$, $T^{-1} \in (0, \frac{2\max_i\{L_{\nabla g_i}\}}{\max\{L_{\nabla f}, \eta c\}}]$, $\gamma^{-1} = \frac{T^{-1}\eta c}{2\max_i\{L_{\nabla g_i}\}}$, $\nu \longleftarrow 0$
      repeat
(S.1)     compute $\kappa(x^\nu)$, the solution $d(x^\nu)$ of problem $(\mathrm{P}_{x^\nu})$ 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
(S.5)             set $\gamma^\nu = \frac{T^\nu \eta c}{2\max_i\{L_{\nabla g_i}\}}$, where $T^\nu = \frac{1}{2}\frac{\theta(x^\nu)}{\nabla f(x^\nu)^T d(x^\nu) + \eta c\|d(x^\nu)\|^2}$
              end
          else
(S.6)         set $T^\nu = T^{\nu-1}$ and $\gamma^\nu = \gamma^{\nu-1}$
          end
(S.7)     set $x^{\nu+1} = x^\nu + \gamma^\nu d(x^\nu)$, $\nu \longleftarrow \nu + 1$;
      end
```

We first note that the value of $T^{-1}$ guarantees that $\gamma^{-1} \leq 1$. Also, the variable $T^\nu$ is introduced just for notational purposes to make the statement of the algorithm and the proof of Theorem 2 easier to follow. The tests we must perform to decide whether to reduce the stepsize are very simple and involve quantities that are readily available once the direction finding subproblem $(\mathrm{P}_{x^\nu})$ has been solved. The following theorem provides the announced complexity result in this general case. For simplicity of presentation, we assume $\delta \leq 1$. This is by no means necessary but avoids the necessity of complicating the statement by considering uninteresting cases.

**Theorem 2.** *Let $\{x^\nu\}$ be the sequence generated by Algorithm 2 under Assumptions A, C1, and D and suppose that $\delta \leq 1$. Then, in at most $\mathcal{O}(\delta^{-4})$ iterations, Algorithm 2 stops either at step* (S.2) *or at step* (S.4)*; more precisely, the maximum number of iterations is given by the maximum between the expressions* (50) *and* (52).

**Proof.** Suppose that Algorithm 2 performs $N$ iterations without stopping (we consider the iteration completed when we reach step (S.7)). We first count how many times $\gamma^\nu$ can be updated in step (S.5) of the algorithm: let

$$\mathcal{I} \triangleq \{0 < \nu_i \leq N \,|\, T^{\nu_i} \text{ and } \gamma^{\nu_i} \text{ are updated in S.5}\} \cup \{0\}$$

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 $\mathcal{I} = \{\nu_0 = 0, \nu_1 = 3, \nu_2 = 4, \nu_3 = 8\}$; we always have by definition $\nu_0 = 0$ 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

$$T^{\nu_i} = \frac{1}{2}\frac{\theta(x^{\nu_i})}{\nabla f(x^{\nu_i})^T d(x^{\nu_i}) + \eta c\|d(x^{\nu_i})\|^2}, \tag{43}$$

and the procedure did not stop at step (S.4): thus, $\theta(x^{\nu_i}) > \delta$ and (43) entails

$$T^{\nu_i} > \frac{\delta}{2B}, \tag{44}$$

with $B \triangleq \max_x\{\|\nabla f(x)\|\beta + \eta c\beta^2 \,|\, x \in K\} \geq \nabla f(x^\nu)^T d(x^\nu) + \eta c\|d(x^\nu)\|^2$. By the updating rule in (S.5), we also have $T^{\nu_i} \leq T^{-1}/2^i$; thus, in view of (44), $\delta/2B < T^{\nu_i} \leq T^{-1}/2^i$, so that

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

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

Let us set $I \triangleq |\mathcal{I}| - 1$; with this convention, the largest element in $\mathcal{I}$ is $\nu_I$. Counting from $\nu_i \in \mathcal{I} \setminus \{\text{last element in } \mathcal{I}\}$, 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 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 $N_I$ 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 previous example 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), because $\gamma^{-1} = T^{-1}\eta c/2\max_i\{L_{\nabla g_i}\}$, also $\gamma^\nu$ is nonincreasing. Moreover, by the definitions of $T^{-1}$ and $\gamma^{-1}$, on the one hand, $\eta c - \gamma^\nu/2L_{\nabla f} \geq \eta c - \gamma^{-1}/2L_{\nabla f} \geq \eta c - \eta c/2$, whereas, on the other hand, $-\gamma^\nu/2\,\max_i\{L_{\nabla g_i}\}/T^\nu = -\eta c/4$. Because of the previous relations, we have for every $\nu$

$$\eta c - \frac{\gamma^\nu}{2}\left(L_{\nabla f} + \frac{\max_i\{L_{\nabla g_i}\}}{T^\nu}\right) \geq \frac{\eta c}{4}, \tag{45}$$

and, in turn, by (25),

$$W(x^{\nu+1}; T^\nu) - W(x^\nu; T^\nu) \leq \gamma^\nu\left[\nabla f(x^\nu)^T d(x^\nu) - \frac{\theta(x^\nu)}{T^\nu} + \frac{3\eta c}{4}\|d(x^\nu)\|^2\right], \tag{46}$$

where we took $\varepsilon^\nu = T^\nu$. We now distinguish two cases. If the condition in step (S.3) is satisfied and $\gamma$ is updated in (S.5),

$$\nabla f(x^\nu)^T d(x^\nu) - \frac{\theta(x^\nu)}{T^\nu} = -\nabla f(x^\nu)^T d(x^\nu) - 2\eta c\|d(x^\nu)\|^2 \leq -\eta c\|d(x^\nu)\|^2, \tag{47}$$

where the inequality follows from the first condition in (S.3). If, on the contrary, $\gamma$ need not be reduced,

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

and, again, relation (47) is easily seen to hold. Therefore, in view of (46) and (47), we get for every $\nu$,

$$W(x^{\nu+1}; T^\nu) - W(x^\nu; T^\nu) \leq -\gamma^\nu \frac{\eta c}{4}\|d(x^\nu)\|^2. \tag{48}$$

Note that $N = \sum_{i \in \mathcal{I}}(N_i + 1)$, because the algorithm did not stop until iteration $N$, $\|d(x^\nu)\| > \delta$ for all iterates up to $N - 1$. Therefore, recalling definition (24) with $\varepsilon^\nu = T^\nu$, and observing that for every $\nu \in \{\nu_i, \ldots, \nu_i + N_i\}$, $\nu_i \in \mathcal{I}$, $\gamma^\nu$ is not reduced and $T^\nu = T^{\nu_i}$, we get

$$\begin{aligned}\delta^2 N &= \sum_{i=0}^{I} \delta^2(N_i + 1) < \sum_{i=0}^{I} \sum_{\nu=\nu_i}^{\nu_i+N_i} \|d(x^\nu)\|^2 \le \sum_{i=0}^{I} \frac{W(x^{\nu_i}; T^{\nu_i}) - W\big(x^{\nu_i+N_i+1}; T^{\nu_i}\big)}{\gamma^{\nu_i}\ \frac{\eta c}{4}} \\ &\le \frac{1}{\gamma^{\nu_I} \frac{\eta c}{4}} \Bigg[ f\big(x^0\big) - f\big(x^{\nu_I+N_I+1}\big) + \frac{1}{T^0} \max_i \{g_i\big(x^0\big)_+\} \\ &\quad - \frac{1}{T^{\nu_I+N_I+1}} \max_i \{g_i\big(x^{\nu_I+N_I+1}\big)_+\} + \sum_{i=1}^{I} \left(\frac{1}{T^{\nu_i}} - \frac{1}{T^{\nu_{i-1}}}\right) \max_j \{g_j(x^{\nu_i})_+\}\Bigg],\end{aligned} \tag{49}$$

where the second inequality is because of (48), whereas, observing that $\gamma^{\nu_I} \le \gamma^{\nu_i}$, the last inequality is valid as a result of a telescopic series argument because $\nu_i + N_i + 1 = \nu_{i+1}$. It is understood that if $I = 0$, the last summation in (49) has no terms. Letting $g_+^M \triangleq \max_x\{\max_i\{g_i(x)_+\} \,|\, x \in K\}$ and $f^m \triangleq \min_x\{f(x) \,|\, x \in K\}$, we distinguish two cases: (i) step (S.5) has never been reached, that is, $T$ has never been diminished; (ii) case (i) did not occur. In case (i), observing that $I = 0$, by (49), the algorithm stops after at most

$$\left\lceil \frac{8}{(\eta c)^2 T^{-1}} \max_i \{L_{\nabla g_i}\} \left[ f\big(x^0\big) - f^m + \frac{1}{T^{-1}} \max_i \{g_i\big(x^0\big)_+\}\right] \frac{1}{\delta^2} \right\rceil \tag{50}$$

iterations. In case (ii), by (49), we can write instead

$$\delta^2 N \overset{(a)}{<} \frac{1}{\gamma^{\nu_I} \frac{\eta c}{4}} \left( f\big(x^0\big) - f^m + \frac{1}{T^0} g_+^M - \frac{1}{T^0} g_+^M + \frac{1}{T^{\nu_I}} g_+^M \right) \overset{(b)}{=} \frac{8}{(\eta c)^2 T^{\nu_I}} \max_i \{L_{\nabla g_i}\} \left( f\big(x^0\big) - f^m + \frac{1}{T^{\nu_I}} g_+^M \right) \tag{51}$$

where (a), because $T^{\nu_i} \le T^{\nu_{i-1}}$, follows again from the summation of a telescopic series, and (b) is because of the updating rule for $\gamma^\nu$ in (S.5) at iteration $\nu_I$. In turn, taking into account that because we updated $T$ at least once, we have

$$T^{\nu_I} = \frac{1}{2} \frac{\theta(x^{\nu_I})}{\nabla f(x^{\nu_I})^T d(x^{\nu_I}) + \eta c \|d(x^{\nu_I})\|^2} > \frac{\delta}{2B},$$

and, in turn,

$$\delta^2 N < \frac{16B}{(\eta c)^2 \delta} \max_i \{L_{\nabla g_i}\} \left( f\big(x^0\big) - f^m + \frac{2B}{\delta} g_+^M \right),$$

thus, meaning that the procedure halts in at most

$$\left\lceil \frac{16B}{(\eta c)^2} \max_i \{L_{\nabla g_i}\} \left[ \frac{f\big(x^0\big) - f^m}{\delta^3} + \frac{2B\, g_+^M}{\delta^4} \right] \right\rceil \tag{52}$$

iterations. If $\delta \le 1$, this gives an overall complexity of $\mathcal{O}(\delta^{-4})$. □

### 5.1. On the Meaning of the Stopping Criteria at (S.2) and (S.4)

The following theorem elucidates the meaning of the stopping criteria in steps (S.2) and (S.4). This result and the ensuing discussion show that (S.2) and (S.4) guarantee that the algorithm stops in a finite number of iterations once a $\delta$−stationary point has been reached. To simplify the proof, we assume that $\delta < \min\{1, \beta\}$; this is very sensible because, on the one hand, we are mainly interested in what happens when $\delta$ is *small* and, on the other hand, $\beta$ is chosen by the user and is intended to be *large*, $\beta$ being simply a safeguard on the maximum length of the direction $d(x^\nu)$.

Preliminarily, we recall that the KKT conditions at a point $x^\nu \in K$ for problem (P) can be rewritten as

$$\left\|P_K\left(x^\nu - \frac{\nabla f(x^\nu) + \nabla g(x^\nu)\xi^\nu}{1 + \|\xi^\nu\|}\right) - x^\nu\right\| = 0, \quad \max_i \left| g_i(x^\nu) \frac{\xi_i^\nu}{1 + \|\xi^\nu\|} \right| = 0, \quad \max_i \{g_i(x^\nu)_+\} = 0, \tag{53}$$

where $P_K$ denotes the projection on the closed convex set $K$ and $\xi^\nu \geq 0$ are suitable multipliers. We also recall that $\theta$ is a stationarity measure for the violation-of-the-constraint problem (4): $\theta(x^\nu) = 0$ if and only if $x^\nu$ is stationary for (4), see Proposition 3 (ii).

**Theorem 3.** *Let Assumptions A, C1, and D hold, and consider $\delta < \min\{1, \beta\}$. If Algorithm 2*

i. *stops at step* (S.2), *$x^\nu$ is either infeasible almost stationary for the violation-of-the-constraints problem, that is,*

$$\max_i\{g_i(x^\nu)_+\} > \frac{L}{\lambda}\delta, \quad 0 < \theta(x^\nu) \leq L\,\delta, \tag{54}$$

*or it is a scaled-KKT point, that is,*

$$\begin{aligned}
&\max_i\{g_i(x^\nu)_+\} \leq \frac{L}{\lambda}\,\delta\\
&\left\| P_K\left(x^\nu - \left[\frac{1}{1+\|\xi^\nu\|}\nabla f(x^\nu) + \nabla g(x^\nu)\frac{\xi^\nu}{1+\|\xi^\nu\|}\right]\right) - x^\nu\right\| \leq \left(2 + L_{\nabla\tilde{f}} + L_{\nabla\tilde{g}}\right)\delta,\\
&\max_i\left| g_i(x^\nu)\,\frac{\xi_i^\nu}{1+\|\xi^\nu\|}\right| \leq \frac{1+\lambda}{\lambda}\,L\,\delta,
\end{aligned} \tag{55}$$

*for some $\xi^\nu \geq 0$, or either an ES or a FJ point;*

ii. *stops at step* (S.4), *$x^\nu$ is either an ES or a FJ point, or it is infeasible almost stationary for the violation-of-the-constraints problem, that is,*

$$\max_i\{g_i(x^\nu)_+\} > 0, \quad 0 < \theta(x^\nu) \leq \delta. \tag{56}$$

Before proving the theorem, some comments are in order. Theorem 3 shows that Algorithm 2 stops with a KKT, FJ, or ES solution, or a point that, at least, satisfies (54) or (55) or (56). This outcome is in line with many recent results in the literature. Relations (55), (54), and (56) are similar to classical conditions such as (i) and (ii) in Cartis et al. [17, theorem 2.9], (3.27) and (3.26), respectively, in Cartis et al. [19, theorem 3.8], or (10) and (9), respectively, in Cartis et al. [18, theorem 4.5]. Specifically, we obtain the scaled-type conditions (55): we refer the interested reader to Cartis et al. [17, section 2.1], but also Birgin [8], Cartis et al. [16], and Cartis et al. [19] for rather exhaustive discussions on this point. On the other hand, the *degenerate* cases (54) and (56) indicate, although in slightly different ways, that a stationarity condition for the violation-of-the-constraint problem is approximately satisfied at an infeasible point. To get more insight into the meaning of the stopping criteria, we discuss, in the same spirit as the analysis in Birgin et al. [8], what happens when $\delta$ goes to zero with fixed initial data. Thus, suppose we have a sequence $\{x_k\}$ each point of which satisfies at least one of (54), (55), or (56) for a sequence of values $\delta^k \downarrow 0$: in fact, we recall that, in view of Theorem 2, for every $k$ the algorithm stops, providing $x_k$, either at step (S.2) or at step (S.4) in a finite number of iterations $N = N^k$ (which is obviously nondecreasing with respect to $k$). Accordingly, let $I = I^k$ be the corresponding number of times $T^\nu$ and $\gamma^\nu$ have been reduced, apart from iteration 0. Moreover, because $\{x_k\}$ is contained in $K$, it is bounded and therefore we can assume, without loss of generality, that it converges to a point $\bar{x} \in K$.

Suppose first that $x_k$ satisfies the scaled-KKT condition (55) for every $k \in \mathcal{K} \subseteq \{0, 1, 2, \ldots\}$ for some $\mathcal{K}$. Passing to the limit in (55), if the corresponding sequence $\xi_k$ is bounded, $\bar{x}$ is a KKT point of problem (P). If, instead, $\xi_k$ is unbounded, $\bar{x}$ must be a FJ point, because it is feasible by the first inequality in (55) and the eMFCQ cannot hold there; otherwise, the sequence $\xi_k$ would be bounded by Proposition 4 (i) and Proposition 6.

Suppose now that one of the two degenerate cases (54) or (56) occurs at each $x_k$ with $k \in \mathcal{K}$. We show that, for both cases, $\bar{x}$ is a point where the eMFCQ does not hold and therefore it is either a FJ or an ES point. Let the algorithm stop at step (S.2) providing $x_k$ that satisfies (54) for all $k \in \mathcal{K}$, and assume by contradiction that the eMFCQ holds at $\bar{x}$. It follows that $d(x_k) \to d(\bar{x}) = 0$ and $\theta(x_k) \to \theta(\bar{x}) = 0$, because of the condition $d(x_k) \leq \delta^k$ for every $k \in \mathcal{K}$ and to the continuity relative to $K$ of function $d(\bullet)$ (on a neighborhood of $\bar{x}$, see Proposition 4) and $\theta(\bullet)$ (see Proposition 3), respectively. Besides, relying on Lemma 2 (iii), $\bar{d} \in \rho\mathbb{B}^n_\infty \cap (K - \bar{x})$ exists such that $\tilde{g}(\bar{d}; \bar{x}) < 0$. Thus, by the continuity relative to $K$ of the set-valued mapping $K - \bullet$ at $\bar{x}$ (see Lemma 1 (ii)), there exists $d_k \in \rho\mathbb{B}^n_\infty \cap (K - x_k)$ such that, for every $k \in \mathcal{K}$ sufficiently large, $\tilde{g}(d_k; x_k) < 0$. In turn, $\min_d\{\max_i\{\tilde{g}_i(d; x_k)_+\} \mid \|d\|_\infty \leq \rho,\ d \in K - x_k\} = 0$, $\kappa(x_k) = (1 - \lambda)\max_i\{g_i(x_k)_+\}$, and $\theta(x_k) = \lambda\max_i\{g_i(x_k)_+\} \leq L\delta^k$ in contradiction with $\max_i\{g_i(x_k)_+\} > L/\lambda\delta^k$. Therefore, the eMFCQ does not hold at $\bar{x}$.

Suppose now that the algorithm stops at step (S.4) for all $k \in \mathcal{K}$ and assume by contradiction that the eMFCQ holds at $\bar{x}$. We have, without loss of generality, $N^k > N^{k-1}$ for every $k \in \mathcal{K}$ and $\theta(x_k) \to \theta(\bar{x}) = 0$, because of the condition $\theta(x_k) \leq \delta^k$ for every $k \in \mathcal{K}$ and to the continuity relative to $K$ 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 $\mathcal{K}$, because $T^{N^k} = T^{\nu_{I^k}} \leq \frac{T^{-1}}{2^{I^k}}$ for every $k \in \mathcal{K}$. If the eMFCQ holds at $\bar{x}$, for any $k \in \mathcal{K}$ sufficiently large, $d(x_k)$ is a KKT point for $(\mathrm{P}_{x_k})$ by Proposition 1 and, in turn, by (26), we get

$$\nabla f(x_k)^T d(x_k) - \frac{1}{T^{N^k}} \theta(x_k) + \eta c \|d(x_k)\|^2 \leq \left( m \|\xi^{N^k}\|_\infty - \frac{1}{T^{N^k}} \right) \theta(x_k). \tag{57}$$

Because of 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 $\nabla f(x_k)^T d(x_k) + \eta c \|d(x_k)\|^2 > 0$ and $T^{N^k} > \theta(x_k) / \nabla f(x_k)^T d(x_k) + \eta c \|d(x_k)\|^2$ for every $k \in \mathcal{K}$ in (S.3). Therefore, the eMFCQ does not hold at $\bar{x}$. All this discussion motivates us to define a point at which Algorithm 2 stops a $\delta-$(generalized) stationary point.

**Definition 3.** A point generated by the algorithm is a $\delta-$(generalized) stationary point if it is either a scaled-KKT point satisfying (55) or an infeasible approximate stationary point for the violation-of-the-constraints-problem satisfying (54) or (56).

It may also be interesting to remark that if the eMFCQ holds at every point in $K$, (54) and (56) cannot occur if $\delta$ is small enough (see the previous discussion), and $\xi^\nu$ that, we shall see in the proof below, are the multipliers of the direction finding subproblems $(\mathrm{P}_{x^\nu})$, are bounded by Propositions 4 (i) and 6. In turn, this means that the algorithm stops at (S.2) with a point $x^\nu$ approximately satisfying the KKT conditions for (P) with $\xi^\nu$ being nothing else but approximate KKT multipliers (see Cartis et al. [17, section 2.1] for further details).

**Proof of Theorem 3.** (i) Suppose first that the algorithm stops because $\|d(x^\nu)\| \leq \delta$. Regardless of the validity of the constraint qualification, $d(x^\nu)$, which certainly satisfies the Fritz-John conditions, may satisfy or not the KKT conditions for the subproblem $(\mathrm{P}_{x^\nu})$. We now distinguish two cases, remarking that the following results hold whatever the choice of $\gamma^\nu$.

I. If $d(x^\nu)$ does not satisfy the KKT conditions for subproblem $(\mathrm{P}_{x^\nu})$, in view of Proposition 4, $x^\nu$ does not satisfy the eMFCQ and, thus, is either an ES or a FJ point.

II. If, on the contrary, $d(x^\nu)$ satisfies the KKT conditions for subproblem $(\mathrm{P}_{x^\nu})$, letting $\xi^\nu \in N_{\mathbb{R}^m_-}(\tilde{g}(d(x^\nu); x^\nu) - \kappa(x^\nu) e)$, we get the following relation, which is equivalent to (18) that still holds with $N_{\beta \mathbb{B}^n_\infty} = \{0\}$ because $\|d(x^\nu)\| \leq \delta$:

$$x^\nu + d(x^\nu) = P_K \left( x^\nu + d(x^\nu) - \frac{\nabla_1 \tilde{f}(d(x^\nu); x^\nu) + \nabla_1 \tilde{g}(d(x^\nu); x^\nu) \xi^\nu}{1 + \|\xi^\nu\|} \right). \tag{58}$$

Let us bound now the terms in relations (53). As for the gradient of the Lagrangian-related condition for problem (P), we have the following bound:

$$\begin{aligned}
\left\| P_K \left( x^\nu - \frac{\nabla f(x^\nu) + \nabla g(x^\nu) \xi^\nu}{1 + \|\xi^\nu\|} \right) - x^\nu \right\| &= \left\| P_K \left( x^\nu - \frac{\nabla f(x^\nu) + \nabla g(x^\nu) \xi^\nu}{1 + \|\xi^\nu\|} \right) - [x^\nu + d(x^\nu)] + d(x^\nu) \right\| \\
&\overset{(a)}{=} \Bigg\| d(x^\nu) + P_K \left( x^\nu - \frac{\nabla f(x^\nu) + \nabla g(x^\nu) \xi^\nu}{1 + \|\xi^\nu\|} \right) \\
&\quad - P_K \left( x^\nu + d(x^\nu) - \frac{\nabla_1 \tilde{f}(d(x^\nu); x^\nu) + \nabla_1 \tilde{g}(d(x^\nu); x^\nu) \xi^\nu}{1 + \|\xi^\nu\|} \right) \Bigg\| \\
&\overset{(b)}{\leq} \|d(x^\nu)\| + \Bigg\| - d(x^\nu) + \frac{\nabla_1 \tilde{f}(d(x^\nu); x^\nu) - \nabla_1 \tilde{f}(0; x^\nu)}{1 + \|\xi^\nu\|} \\
&\quad + \frac{\nabla_1 \tilde{g}(d(x^\nu); x^\nu) \xi^\nu - \nabla_1 \tilde{g}(0; x^\nu) \xi^\nu}{1 + \|\xi^\nu\|} \Bigg\| \\
&\overset{(c)}{\leq} \left( 2 + L_{\nabla \tilde{f}} + L_{\nabla \tilde{g}} \right) \|d(x^\nu)\|,
\end{aligned}$$

where (a) follows from (58), (b) holds because of A4 and A9 and because the projection mapping is nonexpansive, and (c) is because of C1 and D2. As for the complementarity conditions, consider $\bar{\imath} \in \{1, \ldots, m\}$ such that $|g_{\bar{\imath}}(x^\nu) \xi^\nu_{\bar{\imath}}| = \max_i |g_i(x^\nu) \xi^\nu_i|$ with $\tilde{g}_{\bar{\imath}}(d(x^\nu); x^\nu) = \kappa(x^\nu)$; otherwise, $\xi^\nu_{\bar{\imath}} = 0$. If $g_{\bar{\imath}}(x^\nu) \geq 0$, $g_{\bar{\imath}}(x^\nu) \leq \max_i \{g_i(x^\nu)_+\}$, whereas if $g_{\bar{\imath}}(x^\nu) < 0$,

$$|g_{\bar{\imath}}(x^\nu)| = -g_{\bar{\imath}}(x^\nu) - \tilde{g}_{\bar{\imath}}(d(x^\nu); x^\nu) + \tilde{g}_{\bar{\imath}}(d(x^\nu); x^\nu) \leq \nabla_1 \tilde{g}_{\bar{\imath}}(d(x^\nu); x^\nu)^T d(x^\nu),$$

where the inequality is because of (20) and $-\tilde{g}_{\bar{i}}(d(x^\nu); x^\nu) = -\kappa(x^\nu) \leq 0$. Overall,

$$\max_i |g_i(x^\nu)\, \xi_i^\nu| \leq \left(\max_i\{g_i(x^\nu)_+\} + L\|d(x^\nu)\|\right)\|\xi^\nu\|.$$

In turn, if $\max_i\{g_i(x^\nu)_+\} \leq \frac{L}{\lambda}\delta$, then

$$\max_i \left| g_i(x^\nu)\, \frac{\xi^\nu}{1 + \|\xi^\nu\|}\right| \leq \frac{1+\lambda}{\lambda}\, L\,\delta.$$

If, on the contrary, $\max_i\{g_i(x^\nu)_+\} \leq (L/\lambda)\delta$, nonetheless, by (42), we have $\theta(x^\nu) \leq L\,\delta$.

ii. To exit at step (S.4), either $x^\nu$ is an ES or a FJ point, or $\theta(x^\nu)$ must be strictly positive. In fact, under the eMFCQ, by (22), if $\theta(x^\nu) = 0$, then $\nabla f(x^\nu)^T d(x^\nu) \leq -\eta c\|d(x^\nu)\|^2$ and the first condition in step (S.3) does not hold. In turn, for $\theta(x^\nu)$ to be strictly positive, we must have $\max_i\{g_i(x^\nu)\} > 0$. □

## 5.2. Complexity of $\mathcal{O}(\delta^{-3})$ with Constant Stepsize if a Feasible Starting Point Is Known

If a feasible starting point is available, then by choosing a sufficiently small initial $T^{-1}$ or, correspondingly, a sufficiently small initial stepsize $\gamma^{-1}$, the iteration complexity of Algorithm 2 can be reduced to $\mathcal{O}(\delta^{-3})$. Actually, it turns out that in this case, as well as in the cases analyzed in the next two sections, the stepsize is never reduced, so that the updating step of Algorithm 2 actually becomes a *fixed stepsize iteration*

$$x^{\nu+1} = x^\nu + \bar{\gamma} d(x^\nu). \tag{59}$$

A reduction of the iteration complexity when a feasible point is available seems rather sensible because if we start with a feasible point, we have already solved the feasibility problem that is a part of the constrained optimization. Nevertheless, it was, in principle, not clear that our algorithm could take advantage of this fact, because the search for feasibility and that for optimality are combined in a single step, unlike typical methods designed for strong complexity results for constrained nonconvex problems that use two distinct phases.

**Corollary 1.** *Assume the same setting of Theorem 2, fix a prescribed tolerance $\delta$ and set, according to this value, $T^{-1} = \min\{\frac{\delta}{B}, \frac{2\max_i\{L_{\nabla g_i}\}}{\max\{L_{\nabla f}, \eta c\}}\}$. If the starting point $x^0$ is feasible, then, in at most*

$$\left\lceil \frac{8}{(\eta c)^2} \max_i\{L_{\nabla g_i}\} \max\left\{B, \frac{\max\{L_{\nabla f}, \eta c\}}{2\max_i\{L_{\nabla g_i}\}}\right\}\left(f(x^0) - f^m\right)\frac{1}{\delta^3}\right\rceil$$

*iterations, Algorithm 2 stops either at step (S.2) or at step (S.4). Furthermore, the stepsize is never updated and is constant throughout the algorithm.*

**Proof.** We use the same notation and terminology introduced in the proof of Theorem 2. 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

$$\frac{\delta}{B} \geq T^{-1} = T^{\nu-1} > \frac{\theta(x^\nu)}{\nabla f(x^\nu)^T d(x^\nu) + \eta c\|d(x^\nu)^2\|} \geq \frac{\theta(x^\nu)}{B},$$

so that $\theta(x^\nu) \leq \delta$ and the algorithm stops. Hence, step (S.5) is never reached and the stepsize is never updated. Looking back at the corresponding case (i) in Theorem 2, in view of (50) and recalling that $\delta \leq 1$, the procedure is shown to stop after the claimed number of iterations, at worst. □

Note the somewhat unusual feature that algorithmic choices, that is, $T^{-1}$, are linked to the desired accuracy.

## 5.3. Complexity of $\mathcal{O}(\delta^{-2})$ with Constant Stepsize if a Feasible Starting Point Is Known and Upper Approximations Are Used

Suppose again that a feasible starting point is available and, in addition, assume that UCAs for the $g_i$s are used, (see Section 3.1 and (11) in particular). Then, not only can we get $\mathcal{O}(\delta^{-2})$ complexity, but, differently from the previous section, there is no dependence of $T^{-1}$ on $\delta$. Also, in this case, it turns out that the stepsize need not be updated, and the algorithm reduces to the fixed stepsize scheme (59). Furthermore, if we are in an

MM setting, that is, if we choose an UCA also for $f$ (see Remark 3 and (41)), we can take the fixed stepsize to be one, provided some minimal assumptions on $\tilde{f}$ are satisfied.

**Corollary 2.** *Assume the same setting of Theorem 2. If the starting point $x^0$ is feasible and the $\tilde{g}_i$s are upper convex approximations for the $g_i$s, then, in at most*

$$\left\lceil \frac{8}{(\eta c)^2 T^{-1}} \max_i \{L_{\nabla g_i}\}[f(x^0) - f^m] \frac{1}{\delta^2} \right\rceil$$

*iterations, Algorithm 2 stops either at step* (S.2) *or at step* (S.4). *Furthermore, the stepsize is never updated, is constant throughout the algorithm progress, and can be set equal to one provided that $\eta c \geq L_{\nabla f}$.*

**Proof.** The algorithm only produces feasible iterates (see the discussion after (11)). Therefore, we have $\theta(x^\nu) = 0$ for all $\nu$. As a consequence, step (S.5) is never reached, and the stepsize is never updated: in fact, if the test in (S.3) is met, then the algorithm immediately stops at (S.4) because $\theta(x^\nu) = 0$. Then, reasoning again as in case (i) in Theorem 2, because of (50), we see that the algorithm stops after at most the claimed number of iterations. Suppose further that $\eta c \geq L_{\nabla f}$, then it is easy to see from the instructions in Data that we can choose $\gamma^{-1} = 1$. □

We remark that it is easy to show that the condition $\eta c \geq L_{\nabla f}$ implies that $\tilde{f}$ is an UCA and therefore the requirement in the corollary imposes that we use not any arbitrary UCA, but only UCAs that additionally satisfy $\eta c \geq L_{\nabla f}$. At the same time, in standard MM algorithms, it is usually possible to show convergence with a unitary stepsize without requiring $\eta c \geq L_{\nabla f}$, or similar assumptions. However, we must observe that the constants $\eta$ and $c$ are algorithmic choices and therefore the condition $\eta c \geq L_{\nabla f}$ can always be enforced. For example, if analogously to what done in (12), we set

$$\tilde{f}(d; x) = f(x) + \nabla f(x)^T d + \frac{c}{2}\|d\|^2, \tag{60}$$

it is enough to choose $c$ so that $\eta c \geq L_{\nabla f}$. Additionally, and more importantly, the condition $\eta c \geq L_{\nabla f}$ is needed here to establish for the first time, as far as we are aware of, the iteration complexity for an MM method. Our iteration complexity complements the convergence rate obtained in Bolte and Pauwels [9]. In that paper, assuming a Kurdyka-Łojasiewicz property plus other technical conditions, the authors show that, under suitable constraint qualifications, that we do not require, the whole sequence produced by an MM method converges to a KKT point $x^\infty$ and give expressions for the convergence rate of $\|x^\nu - x^\infty\|$.

### 5.4. Rate of $\mathcal{O}(\delta^{-2})$ Global Convergence When eMFCQ Holds

If the eMFCQ holds at every point in $K$, then we can prove that Algorithm 2 has a *global convergence rate* of $\mathcal{O}(\delta^{-2})$. Once again, under suitable assumptions, one can show that in Algorithm 2 the stepsize is never updated, so that the algorithm reduces to the fixed stepsize iteration (59).

**Corollary 3.** *Assume the same setting of Theorem 2 and, in addition, suppose that the eMFCQ holds at every point in K. If we choose $T^{-1}$ and, correspondingly, $\gamma^{-1}$ sufficiently small (as will be specified in the proof, see also the later comments), being M an upper bound on the norm of multipliers for the subproblems ($\mathrm{P}_{x^\nu}$), in at most*

$$\left\lceil \frac{8mM}{(\eta c)^2} \max_i \{L_{\nabla g_i}\}\left[f(x^0) - f^m + mM \max_i \{g_i(x^0)_+\}\right] \frac{1}{\delta^2} \right\rceil \tag{61}$$

*iterations, Algorithm 2 stops at step* (S.2). *Furthermore, the stepsize is never updated and is constant throughout the algorithm.*

Because we never reach (S.4), the only stopping criterion actually used is the one based on $\|d(x^\nu)\|$ in (S.2), in accordance with what happens in classical SQP-type methods when constraint qualifications are assumed to hold everywhere.

**Proof of Corollary 3.** We first recall that, because of the eMFCQ, by Propositions 4 (i) and 6, taking into account the compactness of $K$, the norm of multipliers $\xi^\nu$ of the subproblems ($\mathrm{P}_{x^\nu}$) is bounded from above by some constant $M$.

By (26), which, in view of the eMFCQ, is still valid because it is derived from the optimality conditions for subproblem (P$_{x^\nu}$), we have

$$\nabla f(x^\nu)^T d(x^\nu) - \frac{1}{T^\nu}\theta(x^\nu) + \eta c\|d(x^\nu)\|^2 \le \left(m\|\xi^\nu\|_\infty - \frac{1}{T^\nu}\right)\theta(x^\nu) \le \left(mM - \frac{1}{T^{-1}}\right)\theta(x^\nu) \le 0 \qquad (62)$$

if $T^{-1} \le 1/mM$, and, in turn, for all $\nu$ it never happens that $\nabla f(x^\nu)^T d(x^\nu) + \eta c\|d(x^\nu)\|^2 > 0$ and $T^\nu > \theta(x^\nu)/(\nabla f(x^\nu)^T d(x^\nu) + \eta c\|d(x^\nu)\|^2)$ in (S.3), and therefore, (S.4) and (S.5) are never reached. Looking back at the proof of Theorem 2, we then see that only case (i) therein can occur and, setting, for example, $T^{-1} = 1/mM$ in (50), the algorithm stops at worst after the claimed number of iterations. □

The fixed stepsize iteration (59) is valid provided that $\bar{\gamma} \le \eta c/(2mM\max_i\{L_{\nabla g_i}\})$. Finally, we remark that the bound given by (61) is different from those seen thus far, in that it depends on the usually unknown quantity $M$. The bound given by (61) should therefore be regarded as a *global convergence rate* (see the Introduction).

### 5.5. Problem Constants Are Not Used

The implementation of Algorithm 2 requires the use 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 difficult to compute constants while retaining complexity results similar to those in Theorem 2. The answer is positive, at the price of a *small amount* of additional function evaluations. Moreover, differently from all previous developments, we must make a numerical, although simple, 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 2, the updating of $\gamma^\nu$ in (S.5) guarantees condition (48), that is, the sufficient decrease of the (ghost) penalty function. However, at a more basic level, this sufficient decrease condition can always be reached if the step $\gamma^\nu$ is sufficiently small. Therefore, one could choose at each iteration the stepsize $\gamma^\nu$ to guarantee that the sufficient decrease condition (48) 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.

**Algorithm 3. Algorithm for (P) Without Constants**

```
       Data: $\delta > 0$, $\eta \in (0,1]$, $x^0 \in K$, $T^{-1} > 0$, $\gamma^{-1} = 1$, $\nu \longleftarrow 0$
       repeat
(S.1)      compute $\kappa(x^\nu)$, the solution $d(x^\nu)$ of problem (P$_{x^\nu}$) and $\theta(x^\nu)$;
(S.2)      if $\|d(x^\nu)\| \le \delta$ 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 $\theta(x^\nu) \le \delta$ then
                   stop and return $x_\delta = x^\nu$
               else
(S.5)              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
(S.6)          set $T^\nu = T^{\nu-1}$
           end
(S.7)      while $W(x^\nu + \gamma^\nu d(x^\nu); T^\nu) - W(x^\nu; T^\nu) > -\gamma^\nu \frac{\eta c}{4}\|d(x^\nu)\|^2$ do
               set $\gamma^\nu \longleftarrow \frac{1}{2}\gamma^\nu$
           end
(S.8)      set $x^{\nu+1} = x^\nu + \gamma^\nu d(x^\nu)$, $\nu \longleftarrow \nu + 1$;
       end
```

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 (48) 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 following theorem shows that Algorithm 3 needs an amount of iterations which is similar (likely smaller, see the comments after the proof) to that required by Algorithm 2.

However, although for Algorithm 3 this quantity is also equal to the number of function and constraints evaluations, we now may need some extra objective and constraint function evaluations, as detailed next.

**Theorem 4.** *Let $\{x^\nu\}$ be the sequence generated by Algorithm 3 under Assumptions A, C1, and D and suppose that $\delta \le 1$. Then, in at most $\mathcal{O}(\delta^{-4})$ iterations, Algorithm 3 stops either at step (S.2) or at step (S.4); more precisely, the maximum number of iterations is given by the maximum between the expressions (69) and (71). Moreover, the algorithm needs a number of objective and constraint function evaluations that is equal to the number of iterations plus at most $\mathcal{O}(\log_2(\delta^{-1}))$ further evaluations, with the precise expression of this additional number of evaluations given by (66).*

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

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

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 2, 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 2. Clearly $T^\nu = T^{\nu_i}$ for every $\nu \in \{\nu_i, \dots, \nu_i + N_i\}$. Following the same line of reasoning as in the proof of Theorem 2, one can readily show that

$$\nabla f(x^\nu)^T d(x^\nu) - \frac{\theta(x^\nu)}{T^\nu} \le -\eta c \|d(x^\nu)\|^2, \tag{63}$$

for every $\nu$, which, in turn, by (25), implies

$$W(x^{\nu+1}; T^\nu) - W(x^\nu; T^\nu) \le -\gamma^\nu \left[\eta c - \frac{\gamma^\nu}{2}\left(L_{\nabla f} + \frac{\max_i\{L_{\nabla g_i}\}}{T^\nu}\right)\right] \|d(x^\nu)\|^2, \tag{64}$$

where we took $\varepsilon^\nu = T^\nu$. We now note that for every $\nu \in \{\nu_i, \dots, \nu_i + N_i\}$, $\nu_i \in \mathcal{I}$, we have $\gamma^\nu \ge G(\delta)$, with

$$G(\delta) \triangleq \min\left\{\frac{1}{2}, \frac{3\eta c}{4} \frac{\delta}{L_{\nabla f}\delta + 2B \max_i\{L_{\nabla g_i}\}}\right\}. \tag{65}$$

Indeed, this is trivial for $\nu_0 = 0$, because we assumed $\gamma^{-1} = 1$ and $G(\delta) \le 1/2$. Suppose by contradiction that $0 \ne \nu_i \in \mathcal{I}$ and $\gamma^\nu < G(\delta)$. By the definition (65) of $G(\delta)$, if we set $\gamma^\nu < 2G(\delta)$, we get, recalling (64) and $T^{\nu_i} > \delta/2B$, $W(x^{\nu+1}; T^{\nu_i}) - W(x^\nu; T^{\nu_i}) \le -\gamma^\nu(\eta c/4)\|d(x^\nu)\|^2$, that is, the test at (S.7) is surely not satisfied if $\gamma^\nu < 2G(\delta)$. This, in turn, contradicts $\gamma^\nu < G(\delta)$, because it shows that in the loop (S.7) we should have stopped at the previous iterate of the cycle. Therefore, taking into account that $\gamma^\nu$ is obtained at (S.7) after a certain number (possibly zero) of halvings of the current value of the stepsize, at each iteration $\gamma^\nu \ge G(\delta)$ for every $\delta$. We conclude that $\gamma^\nu$, globally, needs to be halved no more than $\log_2(\gamma^{-1}/G(\delta))$ times: hence, after at most

$$\log_2\left(\gamma^{-1} \max\left\{2, \frac{4}{3\eta c}\left(L_{\nabla f}\delta + 2B \max_i\{L_{\nabla g_i}\}\right)\right\} \frac{1}{\delta}\right) \tag{66}$$

halvings, one achieves the sought decrease condition

$$W(x^{\nu+1}; T^\nu) - W(x^\nu; T^\nu) \le -\gamma^\nu \frac{\eta c}{4} \|d(x^\nu)\|^2 \le -G(\delta) \frac{\eta c}{4} \|d(x^\nu)\|^2, \tag{67}$$

for every $\nu$. Recalling definition (24), and similarly to (49),

$$\begin{aligned}
\delta^2 N &= \sum_{i=0}^{I} \delta^2(N_i + 1) < \sum_{i=0}^{I} \sum_{\nu=\nu_i}^{\nu_i+N_i} \|d(x^\nu)\|^2 \le \sum_{i=0}^{I} \frac{W(x^{\nu_i}; T^{\nu_i}) - W(x^{\nu_i+N_i+1}; T^{\nu_i})}{G(\delta)\,\frac{\eta c}{4}} \\
&\le \frac{1}{G(\delta)\,\frac{\eta c}{4}} \Bigg[ f(x^{\nu_0}) - f(x^{\nu_I+N_I+1}) + \frac{1}{T^{\nu_0}} \max_i\{g_i(x^{\nu_0})_+\} \\
&\quad - \frac{1}{T^{\nu_I+N_I+1}} \max_i\{g_i(x^{\nu_I+N_I+1})_+\} + \sum_{i=1}^{I}\left(\frac{1}{T^{\nu_i}} - \frac{1}{T^{\nu_{i-1}}}\right) \max_j\{g_j(x^{\nu_i})_+\} \Bigg],
\end{aligned} \tag{68}$$

where the second inequality is because of (67), whereas the last inequality is valid as a result of a telescopic series argument because $\nu_i + N_i + 1 = \nu_{i+1}$. Setting $g_+^M \triangleq \max_x\{\max_i\{g_i(x)_+\} \,|\, x \in K\}$ and $f^m \triangleq \min_x\{f(x) \,|\, x \in K\}$, as in the proof of Theorem 2, we distinguish two cases: (i) step (S.5) has never been reached, thus $I = 0$ and, by (68), the algorithms stops after at most

$$\left\lceil \frac{4}{\eta c}\max\left\{2,\ \frac{4}{3\eta c}\left(L_{\nabla f}\delta + 2B\max_i\{L_{\nabla g_i}\}\right)\right\}\left(f(x^0) - f^m + \frac{1}{T^{-1}}\max_i\{g_i(x^0)_+\}\right)\frac{1}{\delta^3}\right\rceil \tag{69}$$

iterations, in view of the definition of $G$ and $\delta \le 1$. If case (i) did not occur, by (68) we can write

$$\delta^2 N < \frac{1}{G(\delta)\frac{\eta c}{4}}\left(f(x^0) - f^m + \frac{1}{T^{\nu_0}}g_+^M - \frac{1}{T^{\nu_0}}g_+^M + \frac{1}{T^{\nu_I}}g_+^M\right), \tag{70}$$

where the inequality follows, recalling that $T^{\nu_i} \le T^{\nu_{i-1}}$, from the summation of a telescopic series. In turn, because

$$T^{\nu_I} = \frac{1}{2}\frac{\theta(x^{\nu_I})}{\nabla f(x^{\nu_I})^T d(x^{\nu_I}) + \eta c\|d(x^{\nu_I})\|^2} > \frac{\delta}{2B},$$

by (70), the procedure halts in at most

$$\left\lceil \frac{4}{\eta c}\max\left\{2,\ \frac{4}{3\eta c}\left(L_{\nabla f}\delta + 2B\max_i\{L_{\nabla g_i}\}\right)\right\}\left[\frac{f(x^0) - f^m}{\delta^3} + \frac{2B\,g_+^M}{\delta^4}\right]\right\rceil \tag{71}$$

iterations. Because $\delta \le 1$, one can take the overall complexity to be of $\mathcal{O}(\delta^{-4})$. □

It is interesting to compare the worst-case bounds (52) (for Algorithm 2) and (71) (for Algorithm 3). It is clear that, at least for a small $\delta$ (see the definition of $G$), the bound (71) is approximatively $\frac{2}{3}$ of the bound (52). This better behavior of Algorithm 3 has a simple explanation. The steps used in Algorithm 3 are generally larger than those used in Algorithm 2, where problem constants 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 result in that the algorithm is marginally more complex, requires a numerical use of the penalty function, and calls for additional objective function and constraint evaluations that may increase the computational effort. However, this increase is negligible when $\delta$ is small, because the additional number of function evaluations is $\mathcal{O}(\log_2(\delta^{-1}))$, implying that the overall order of function evaluations is maintained to be $\mathcal{O}(\delta^{-4})$.

**Remark 4.** It is easy to see that the results in Sections 5.2–5.4 for Algorithm 2 can be extended to Algorithm 3, but we do not pursue this for lack of space.

**Remark 5.** Although the analysis in this section is about SQP-type approaches, a few other complexity bounds are available in the literature for different methods using first-order information in the context of nonconvex, constrained optimization (see the Introduction). A direct comparison of all these results is difficult, because different assumptions and, in some cases, different concepts of (approximate) generalized stationary points are called for; furthermore, the various methods differ markedly in the overall structure (penalty versus phase I–phase II versus SQP schemes) and in the algorithmic computational effort required at each iteration, not to mention that in some cases results are given for equality constraints only. With this in mind, here we try to briefly highlight the main features of Birgin et al. [8], Cartis et al. [15], Cartis et al. [16], and Cartis et al. [17]. The analysis provided in Cartis et al. [15, section 3.2] is for a penalty-based approach and gives the "first worst-case global evaluation bounds for constrained optimization when both the objective and the constraints are allowed to be nonconvex." The main similarity with our analysis is in the use of a nondifferentiable penalty function, which, however, is used according to a classical double-loop scheme: at each outer iteration a penalty parameter is chosen and then the penalty function is minimized inexactly using a trust-region–like method for which complexity results are provided. This should be contrasted with our *ghost* use of penalties where the penalty function and the penalty parameter are not used in the algorithm itself. Another difference is in the subproblems to be solved at each iteration. Our method follows the standard SQP approach and the subproblems should be regarded as (simple) approximations of the original problem and as such do not involve any penalty parameter. The subproblems in Cartis et al. [15], instead, aim at approximating the penalty function and, as such, necessarily include the penalty parameter. On the one hand, avoiding the use of the penalty parameter in the subproblems is a favorable numerical feature, we believe; on the other hand approximating directly the penalty function, as done in Cartis et al. [15], nicely avoids the issue of the feasibility of subproblems, because the minimization of the penalty function is unconstrained. Putting together a judicious analysis of the parameter-updating scheme and of the inner penalty minimization, Cartis et al. [15] can then give estimates for the number of iterations necessary to reach an approximate generalized stationary point. If, during the minimization process, the penalty parameter grows unbounded, a complexity of $\mathcal{O}(\delta^{-5})$ is obtained. If, on the other hand, an upper bound, which in principle is unknown in advance, for the penalty parameter exists, a condition that should be interpreted as a constraint qualification, then, using the terminology of the present paper, a *convergence rate* of $\mathcal{O}(\delta^{-2})$ is obtained. Note that (a) the issue of the boundedness of the iterates is not dealt with (for our algorithm, see also next section), and (b) the objective function $f$ is assumed to be bounded from below on $\mathbb{R}^n$.

The phase I–phase II approach in Cartis et al. [16, sections 4 and 5] is for equality constrained problems and relies on target-following and (inner) trust-region techniques. Both phases are based on a cubic regularization method involving the use of second-order derivatives, with subproblems to be solved at each iterations that are potentially expensive. However, a remarkable complexity bound of $\mathcal{O}(\delta^{-3/2})$, matching the one for the cubic regularization method in the unconstrained case, is achieved.

The high-level scheme put forward in Birgin et al. [8] for nonconvex problems with (equality and) inequality constraints falls also within a phase I–phase II (of target-following–type) framework: the unconstrained nonlinear minimization problems to be solved in each phase are assumed to be dealt with by some minimization algorithm with known complexity guarantees. Once this algorithm is given, the analysis derives bounds that range between $\mathcal{O}(\delta^{-3})$ and $\mathcal{O}(\delta^{-5})$ according to the choice of an algorithmic parameter, with the better complexity corresponding to *weaker* notions of almost generalized stationary points. It has to be remarked that, as a major departure from all other works, in Birgin et al. [8], an emphasis is given to *unscaled* KKT conditions, that is, to an approximate notion of stationarity that does not depend on the magnitude of the multipliers involved.

Finally, in Cartis et al. [17], a phase I–phase II (of target-following–type) method is presented, which resorts again to an inner trust-region approach in both phases. Assuming the objective function to be upper and lower bounded on the feasible set, and the gradient of the objective and the Jacobian of the constraint functions to be Lipschitz continuous on $\mathbb{R}^n$ and on a suitable extended neighborhood of the feasible region, respectively, the algorithm is proven to reach an approximate generalized stationary point in at most $\mathcal{O}(\delta^{-2})$ iterations. These results are currently the most advanced for a first-order phase I–phase II method, and, remarkably, the bounds obtained there match the best result for first-order unconstrained minimization methods.

## 6. Boundedness of Iterates

Boundedness of the sequence generated by an SQP-type method is a difficult issue. With a few earlier exceptions (Facchinei [30]), this topic probably came to a wider attention only with the important paper (Solodov [60]) that motivated researchers to look better into this issue (Auslender [1], Auslender et al. [2], Bolte and Pauwels [9], Liu and Yuan [42]; with the latter reference dealing only with equality constraints). In our framework, generating an unbounded sequence is a natural possibility that cannot and should not be excluded in principle, because we do not make any standard assumption such as feasibility, existence of an optimal solution, or regularity of the constraints; quoting from Bolte and Pauwels [9, p. 11], where a similar possibility is considered, "The divergence property…is a positive result, a convergence result, which does not correspond to a failure of the method but rather to the absence of minimizers in a given zone." To clarify this point consider

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & x^2 \\ \text{s.t.} \quad & e^x \leq 0, \end{aligned}$$

which is an infeasible convex problem and has no ES, FJ, or KKT solutions. Nevertheless, we can apply one of the algorithms studied in this paper to it, and the only sensible outcome is *an attempt to minimize infeasibility* with the generation of an unbounded sequence. Indeed, if the sequence generated by the algorithm were bounded, every limit point should be critical, but because there are no critical points, the sequence must necessarily be unbounded. Yet, also in the spirit of the works mentioned at the beginning of this section, it is of course of great interest to see under what conditions we can guarantee the boundedness of the sequence $\{x^\nu\}$ for the algorithms presented in this paper. We analyze this point, identifying some settings where boundedness can be guaranteed. This discussion, which in no way tries to be exhaustive, is also useful to illustrate some of the characteristics of our methods. We remark that the properties of Algorithms 2 and 3 have been studied in the previous section under the assumption that $K$ is bounded. However, it is easy to see from the proofs that this condition can be substituted by any of the assumptions studied in this section that still guarantee boundedness of the sequence generated by Algorithms 2 and 3. The only adjustment that needs to be made is that the Lipschitz constants used in the proofs of Theorems 2 and 4 are no longer the Lipschitz constants on $K$ but rather the ones on the compact set $S$, which is shown to contain the sequence $\{x^\nu\}$ and that any reference to the boundedness of $K$ should be substituted by a reference to the compactness of $S$. Of course, in order for this approach to be sensible as a complexity bound, the set $S$ must be determined a priori and should not depend on the sequence generated by the algorithm.

1. Valid for Algorithms 1–3: The boundedness of $K$ obviously guarantees the boundedness of $\{x^\nu\}$ for all the algorithms we considered. We already used this fact for Algorithms 2 and 3, but the same result holds also for Algorithm 1: we report this case here for completeness and uniformity of presentation. In fact, for all the three algorithms, we have that $\gamma^\nu \in (0,1]$ and therefore the constraint $d \in K - x^\nu$ in ($\mathrm{P}_x$) and the convexity of $K$ guarantee that if $x^\nu \in K$ then also $x^\nu + \gamma^\nu d(x^\nu)$ belongs to $K$. This case covers most instances of practical interest because, in basically all real-world problems, variables are naturally limited by lower and upper bounds, and we can take $K$ to be the rectangle defined by these quantities.

2. Valid for Algorithms 1–3: Another setting in which we can guarantee the boundedness of the iterations is when we turn our schemes into feasible methods by choosing $\tilde{g}_i$s that are UCAs of the $g_i$s and a feasible starting point $x^0$ (see Section 3.1).

2a. In this setting, assume that the following classical condition holds:

$$\mathcal{L}_1 \triangleq \left\{x \in K : g(x) \le 0, f(x) \le f\left(x^0\right)\right\} \text{ is bounded,} \tag{72}$$

that is, the level set of value $f(x^0)$ for the objective function intersected with the feasible set is bounded. Then, if we also assume that $\tilde{f}$ is an UCA of $f$ (see Remark 3), we can show that the sequence $\{x^\nu\}$ generated by any of the Algorithms 1–3 is contained in the bounded set $\mathcal{L}_1$. Because the sequence $\{x^\nu\}$ belongs to $\mathcal{X}$, it is enough to show that, at each iteration, $f(x^{\nu+1}) \le f(x^\nu)$. To this end, observe that, because each $x^\nu$ is feasible, we always have $\max_i\{g_i(x^\nu)_+\} = \kappa(x^\nu) = \theta(x^\nu) = 0$, and $d = 0$ is feasible for ($\mathrm{P}_{x^\nu}$). Therefore, applying the minimum principle to ($\mathrm{P}_{x^\nu}$), we have $\nabla_1\tilde{f}(d(x^\nu); x^\nu)^T(0 - d(x^\nu)) \ge 0$, and, in turn,

$$\nabla_1\tilde{f}(d(x^\nu); x^\nu)^T d(x^\nu) \le 0. \tag{73}$$

Because $\tilde{f}$ is (strongly) convex, we get

$$\begin{aligned} f(x^\nu) = \tilde{f}(0; x^\nu) &\ge \tilde{f}\left(\gamma^\nu d(x^\nu); x^\nu\right) + \nabla_1\tilde{f}\left(\gamma^\nu d(x^\nu); x^\nu\right)^T\left(0 - \gamma^\nu d(x^\nu)\right) \\ &\ge f\left(x^\nu + \gamma^\nu d(x^\nu)\right) - \gamma^\nu\nabla\tilde{f}\left(\gamma^\nu d(x^\nu); x^\nu\right)^T d(x^\nu), \end{aligned} \tag{74}$$

where the second inequality follows from (41). By the strong convexity with modulus $c$ of $\tilde{f}$ (see A1), we can also write

$$\left(\nabla_1\tilde{f}(d(x^\nu); x^\nu) - \nabla_1\tilde{f}\left(\gamma^\nu d(x^\nu); x^\nu\right)\right)^T\left(d(x^\nu) - \gamma^\nu d(x^\nu)\right) \ge c\left(1 - \gamma^\nu\right)^2\|d(x^\nu)\|^2,$$

which, with simple manipulations, yields

$$-\gamma^\nu\nabla_1\tilde{f}\left(\gamma^\nu d(x^\nu); x^\nu\right)^T d(x^\nu) \ge -\gamma^\nu\nabla_1\tilde{f}(d(x^\nu); x^\nu)^T d(x^\nu) + c\gamma^\nu\left(1 - \gamma^\nu\right)\|d(x^\nu)\|^2.$$

Plugging this inequality in (74) we get

$$f\left(x^\nu + \gamma^\nu d(x^\nu)\right) \le f(x^\nu) + \gamma^\nu\nabla_1\tilde{f}(d(x^\nu); x^\nu)^T d(x^\nu),$$

which, in view of (73), shows that $f(x^{\nu+1}) \le f(x^\nu)$ as desired for any choice of $\gamma^\nu$.

2b. The requirement that $\tilde{f}$ be an upper approximation of $f$ is actually not needed for Algorithm 3. In fact, noting that $W(x^\nu; \varepsilon) = f(x^\nu)$ for any feasible $x^\nu$ and for any positive $\varepsilon$, step (S.7) in Algorithm 3 guarantees $f(x^{\nu+1}) \le f(x^\nu)$. In fact, in the current setting, a suitable stepsize can be found in step (S.7) because of (63) (that still holds even if $K$ is not bounded, see the conditions in the if-block at step (S.3)), recalling that $\theta(x^\nu) = 0$ for any feasible $x^\nu$. It is therefore clear that the whole sequence $\{x^\nu\}$ is contained in $\mathcal{L}_1$.

2c. We can avoid the UCA requirement on $\tilde{f}$ also for Algorithm 2, provided we assume that $\nabla f$ is Lipschitz continuous on $K$ (with modulus $L_{\nabla f}$). By the descent lemma we can write

$$f\big(x^\nu + \gamma^\nu d(x^\nu)\big) - f(x^\nu) \le \gamma^\nu \nabla f(x^\nu)^T d(x^\nu) + (\gamma^\nu)^2 \frac{L_{\nabla f}}{2} \|d(x^\nu)\|^2.$$

In turn, we get from (47) (which, again, is easily seen to hold in the current setting), taking into account that $\theta(x^\nu) = 0$ because $x^\nu$ is feasible,

$$f\big(x^\nu + \gamma^\nu d(x^\nu)\big) - f(x^\nu) \le -\gamma^\nu \left(\eta c - \frac{\gamma^\nu}{2} L_{\nabla f}\right) \|d(x^\nu)\|^2, \tag{75}$$

for every $\nu$. The instructions in Data of Algorithm 2 are easily seen to entail $\gamma^\nu \in (0, \min\{1, 2\eta c / L_{\nabla f}\}]$, so that (75) implies that $\{x^\nu\}$ is all contained in $\mathcal{L}_1$, and therefore that $\{x^\nu\}$ is bounded if $\mathcal{L}_1$ is bounded.

2d. If we want to eliminate the UCA property of $\tilde{f}$ also for Algorithm 1, we need again $\nabla f$ to be Lipschitz continuous on $K$ and to strengthen condition (72), requiring that

$$\mathcal{L}_2^\alpha \triangleq \{x \in K : g(x) \le 0, f(x) \le \alpha\} \text{ is bounded for every } \alpha \in \mathbb{R}. \tag{76}$$

Then, invoking (22), as soon as $\gamma^\nu$ becomes smaller than $2c/L_{\nabla f}$, we stay in the set $\mathcal{L}_2^\alpha$ for some value of $\alpha$; by (76), this implies the boundedness of $\{x^\nu\}$.

2e. Finally, it is worth observing that if the feasible set $\mathcal{X}$ is bounded, the use of UCAs for the constraints $g$ is enough to guarantee the boundedness of $\{x^k\}$ because, if we start wih a feasible point, $\{x^\nu\}$ remains feasible whatever the algorithm we use (Section 3.1).

3. Valid for Algorithms 1–3: Knowing a feasible point $x^0$ to start the algorithm from can be difficult in some applications. However, fortunately, the results in point 2 can be generalized to avoid the feasibility requirement.

3a. Suppose that we start the algorithm with a possibly infeasible point $x^0 \in K$. Assume that

$$\mathcal{L}_3 \triangleq \left\{x \in K : g(x) \le \max_i \{g_i(x^0)_+\}\right\} \text{ is bounded.} \tag{77}$$

If we use $\tilde{g}_i$s that are UCAs for the $g_i$s, we can show by induction that the whole sequence $\{x^\nu\}$ generated by Algorithms 1–3 is contained in $\mathcal{L}_3$. In fact, the starting point $x^0$ of course belongs to $\mathcal{L}_3$. Suppose now that $x^\nu \in \mathcal{L}_3$, meaning that

$$\tilde{g}_i(0; x^\nu) = g_i(x^\nu) \le \max_i \{g_i(x^0)_+\}.$$

We also have

$$\tilde{g}_i(d(x^\nu); x^\nu) \le \kappa(x^\nu) \le \max_i \{g_i(x^\nu)_+\},$$

where the first inequality is just feasibility for subproblem $(\mathrm{P}_{x^\nu})$, and the second one follows by the definition of $\kappa(x^\nu)$. The last two displayed formulas show that, for any $\gamma^\nu \in [0,1]$,

$$g_i\big(x^\nu + \gamma^\nu d(x^\nu)\big) \le \tilde{g}_i\big(\gamma^\nu d(x^\nu); x^\nu\big) \le \max_i \{g_i(x^\nu)_+\},$$

where the first inequality is because of the UCA property (11), whereas the second relation derives from the convexity of $g_i(\cdot; x^\nu)$.

3b. When the constraints $g_i$s are convex, it is well known that the boundedness of $\mathcal{L}_3$ holds if and only if the feasible set is bounded. Then, in principle we can set $\tilde{g}(d; x) = g(x + d)$ and only approximate the objective function. This particular $\tilde{g}$ is a UCA, indeed. This approach seems particularly well suited to the case in which the $g_i$s are linear because the resulting subproblem then has simple linear constraints. Keeping the original (convex) constraints in the subproblems is something routinely done in most MM methods.

4. Valid for Algorithms 1–3: Another interesting case arises if we suppose that the eMFCQ holds and

$$\mathcal{L}_4^\alpha \triangleq \left\{x \in K : \max_i \{g_i(x)_+\} \le \alpha\right\} \text{ is bounded for every } \alpha \in \mathbb{R}_+. \tag{78}$$

Note that (78) simply states that the function $\max_i\{g_i(x)_+\}$ is coercive. We can therefore find positive $\alpha_1$ and $\alpha_2$ such that if $x^\nu \in \mathcal{L}_4^{\alpha_1}$ then $x^\nu + \gamma^\nu d(x^\nu) \in \mathcal{L}_4^{\alpha_2}$ for all $\gamma^\nu \in (0, 1]$. Now, following the same line of reasoning as relations (25), we have

$$\max_i\{g_i\big(x^\nu + \gamma^\nu d(x^\nu)\big)_+\} - \max_i\{g_i(x^\nu)_+\} \leq -\gamma^\nu \left(\theta(x^\nu) - \frac{\gamma^\nu}{2}\max_i\{L_{\nabla g_i}\}\|d(x^\nu)\|^2\right) \tag{79}$$

for every $x^\nu \in \mathcal{L}_4^{\alpha_2}$, where $L_{\nabla g_i}$ are Lipschitz constants of the gradients of $g_i$ on $\mathcal{L}_4^{\alpha_2}$; we remark that because $\mathcal{L}_4^{\alpha_2}$ is bounded, existence of these constants is a very mild requirement. Denote by $\bar\theta > 0$ a positive constant such that $\theta(x) \geq \bar\theta$ for all points in the set $\Delta \triangleq \mathcal{L}_4^{\alpha_2} \setminus \operatorname{int} \mathcal{L}_4^{\alpha_1}$; note that this set is compact by (78). Such $\bar\theta$ surely exists because the eMFCQ implies there are no ES in the set $\Delta$ and therefore the continuous function $\theta(x)$ is positive on $\Delta$. By (79), we can then write, for any $x^\nu \in \Delta$,

$$\max_i\{g_i\big(x^\nu + \gamma^\nu d(x^\nu)\big)_+\} - \max_i\{g_i(x^\nu)_+\} \leq -\gamma^\nu \left(\bar\theta - \frac{\gamma^\nu}{2}\max_i\{L_{\nabla g_i}\}\beta^2\right). \tag{80}$$

It is then clear that a threshold value $\bar\gamma > 0$ exists such that, if $\gamma^\nu \leq \bar\gamma$, then $\max_i\{g_i(x^\nu + \gamma^\nu d(x^\nu))_+\} \leq \max_i\{g_i(x^\nu)_+\}$. Now, two cases can occur. If $x^\nu$ belongs to int $\mathcal{L}_4^{\alpha_1}$, then, by how we have chosen $\alpha_2$, $x^{\nu+1}$ belongs to $\mathcal{L}_4^{\alpha_2}$. If instead $x^\nu$ belongs to $\Delta$, by taking $\gamma^\nu \leq \bar\gamma$, we are again sure that $x^{\nu+1}$ still belongs to $\mathcal{L}_4^{\alpha_2}$. We can so conclude that by using stepsizes smaller that $\bar\gamma$, iterations never leave the set $\mathcal{L}_4^{\alpha_2}$ and therefore stay bounded.

5. Valid for Algorithms 2 and 3: The eMFCQ assumption in case 4 can be replaced by the requirement that $f$ be bounded from below on $K$ if one uses Algorithm 3, a condition to which the Lipschitz continuity of $\nabla f$ and $\nabla g_i$ on $K$ has to be added when Algorithm 2 is resorted to. For both cases, the proof of the claim reduces to showing that, even without requiring $K$ to be bounded, we still have the sufficient descent condition

$$W\big(x^{\nu+1}; T^\nu\big) - W(x^\nu; T^\nu) \leq -\gamma^\nu \frac{\eta c}{4}\|d(x^\nu)\|^2 \tag{81}$$

for every $\nu$. In fact, once relation (81) has been proven to be valid, in turn we get

$$\begin{aligned} T^{\nu+1}\big(f\big(x^{\nu+1}\big) - \bar f\big) + \max_i\{g_i\big(x^{\nu+1}\big)_+\} &\leq T^\nu\big(f\big(x^{\nu+1}\big) - \bar f\big) + \max_i\{g_i\big(x^{\nu+1}\big)_+\} \\ &\leq T^\nu\big(f(x^\nu) - \bar f\big) + \max_i\{g_i(x^\nu)_+\} - T^\nu\gamma^\nu\frac{\eta c}{4}\|d(x^\nu)\|^2, \end{aligned}$$

where the first relation follows from observing that $T^\nu$ is nonincreasing. The sequence generated by the algorithms is now easily shown to be bounded. Indeed, the inequality shows that the nonnegative sequence $\{T^\nu(f(x^\nu) - \bar f) + \max_i\{g_i(x^\nu)_+\}\}$ is nonincreasing and therefore convergent. Suppose now by contradiction that $\{x^k\}$ is unbounded. By (78) this implies that $\max_i\{g_i(x^\nu)_+\}$ goes to infinity; because $f(x^\nu) - \bar f$ is nonnegative, this contradicts the convergence of the sequence $\{T^\nu(f(x^\nu) - \bar f) + \max_i\{g_i(x^\nu)_+\}\}$.

Let us now show why, in the current setting, (81) is still satisfied for Algorithms 2 and 3.

5a. Concerning Algorithm 3, (81) is enforced as the algorithm progresses (see step (S.7)). We remark that, even in the present setting, given an iterate $x^\nu$, in step (S.7) a sufficiently small stepsize $\gamma^\nu$ still exists such that (81) is verified: this follows by standard reasoning ab absurdo in view of (63), which still holds even if $K$ is not bounded (see the conditions in the if-block at step (S.3)) and observing that the directional derivative of $\max_i\{g_i(x^\nu)_+\}$ is bounded from above by $-\theta(x^\nu)$, because of (23).

5b. As for Algorithm 2, with $\nabla f$ and $\nabla g_i$ assumed to be Lipschitz continuous on $K$, let $L_{\nabla f}$ and $L_{\nabla g_i}$ be the corresponding Lipschitz moduli. Again, without requiring $K$ to be bounded as done in Section 5, condition (81) is clearly satisfied, because (46) and (47) remain valid following the same line of reasoning as relation (25) (here with $K$ not assumed to be bounded, but under the Lipschitz continuity of $\nabla f$ and $\nabla g$), and as a straightforward consequence of the conditions in the if-block at step (S.3), respectively.

In both cases 5a and 5b, although $\{x^\nu\}$ is contained in $\mathcal{L}_4^\alpha$ for some $\alpha$, this quantity is possibly unknown in advance.

We summarize these conditions implying boundedness in Table 2. We also clarify (see the last column that only applies to Algorithms 2 and 3) on a case-by-case basis if these assumptions make it possible to perform an iteration complexity (IC) or a global convergence rate (GCR) analysis. In fact, when the bounded set to which

**Table 2.** Summary of conditions for boundedness of iterates.

| Case | $K$ | $\tilde{f}$ | $\tilde{g}$ | $x^0$ | Other assumptions | Algorithm | IC or GCR |
|---|---|---|---|---|---|---|---|
| 1 | Bounded | — | — | — | — | 1, 2, 3 | IC |
| 2a | — | UCA (41) | UCA (11) | Feasible | $\mathcal{L}_1$ bounded, see (72) | 1, 2, 3 | IC |
| 2b | — | — | UCA (11) | Feasible | $\mathcal{L}_1$ bounded, see (72) | 3 | IC |
| 2c | — | — | UCA (11) | Feasible | $\mathcal{L}_1$ bounded, see (72), $\nabla f$ Lipschitz | 2, 3 | IC |
| 2d | — | — | UCA (11) | Feasible | $\mathcal{L}_2^\alpha$ bounded, see (76), $\nabla f$ Lipschitz | 1, 2, 3 | IC |
| 2e | — | — | UCA (11) | Feasible | $\mathcal{X}$ bounded | 1, 2, 3 | IC |
| 3a | — | — | UCA (11) | — | $\mathcal{L}_3$ bounded, see (77) | 1, 2, 3 | IC |
| 3b | — | — | $\tilde{g} = g$ | — | $g_i$s convex, $\mathcal{X}$ bounded | 1, 2, 3 | IC |
| 4 | — | — | — | — | $\mathcal{L}_4^\alpha$ bounded, see (78), eMFCQ | 1, 2, 3 | GCR |
| 5a | — | — | — | — | $\mathcal{L}_4^\alpha$ bounded, see (78), $f$ low. bounded | 3 | GCR |
| 5b | — | — | — | — | $\mathcal{L}_4^\alpha$ bounded, see (78), $f$ low. bounded, $\nabla f$ and $\nabla g$ Lipschitz | 2, 3 | GCR |

the sequence $\{x^\nu\}$ belongs is defined by means of quantities that are known in advance, for example, if it is $\mathcal{L}_1$, we can still speak of iteration complexity results derived for Algorithms 2 and 3, as done in Theorems 2 and 4; when the set is known to exist, but is not known beforehand (as for example in cases 5a and 5b), we obtain instead a global convergence rate for the corresponding algorithms, because the constants involved in the big $\mathcal{O}$ bounds cannot be determined a priori.

## Acknowledgments

The authors thank Philippe Toint, whose detailed comments on an earlier version of this paper helped shape it. The constructive criticisms of two referees were also useful in improving the quality of the paper.

## References

[1] Auslender A (2013) An extended sequential quadratically constrained quadratic programming algorithm for nonlinear, semidefinite, and second-order cone programming. *J. Optim. Theory Appl.* 156(2):183–212.

[2] Auslender A, Shefi R, Teboulle M (2010) A moving balls approximation method for a class of smooth constrained minimization problems. *SIAM J. Optim.* 20(6):3232–3259.

[3] Bank B, Guddat J, Klatte D, Kummer B, Tammer K (1982) *Non-Linear Parametric Optimization* (Akademie-Verlag, Berlin).

[4] Beck A, Ben-Tal A, Tetruashvili L (2010) A sequential parametric convex approximation method with applications to nonconvex truss topology design problems. *J. Global Optim.* 47(1):29–51.

[5] Bertsekas D (2012) Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. Sra S, Nowozin S, Wright SJ, eds. *Optimization for Machine Learning* (MIT Press, Cambridge, MA), 85–119.

[6] Bertsekas D (2015) *Convex Optimization Algorithms*. (Athena Scientific, Belmont, MA).

[7] Bertsekas D, Tsitsiklis J (1989) *Parallel and Distributed Computation: Numerical Methods*, vol. 23 (Prentice Hall, Englewood Cliffs, NJ).

[8] Birgin E, Gardenghi J, Martìnez J, Santos S, Toint P (2016) Evaluation complexity for nonlinear constrained optimization using unscaled KKT conditions and high-order models. *SIAM J. Optim.* 26(2):951–967.

[9] Bolte J, Pauwels E (2016) Majorization-minimization procedures and convergence of SQP methods for semi-algebraic and tame programs. *Math. Oper. Res.* 41(2):442–465.

[10] Bottou L, Curtis F, Nocedal J (2018) Optimization methods for large-scale machine learning. *SIAM Rev.* 60(2):223–311.

[11] Burke J (1989) A sequential quadratic programming method for potentially infeasible mathematical programs. *J. Math. Anal. Appl.* 139(2):319–351.

[12] Burke J (1992) A robust trust region method for constrained nonlinear programming problems. *SIAM J. Optim.* 2(2):325–347.

[13] Burke J, Han SP (1989) A robust sequential quadratic programming method. *Math. Programming* 43(1):277–303.

[14] Burke J, Hoheisel T (2013) Epi-convergent smoothing with applications to convex composite functions. *SIAM J. Optim.* 23(3):1457–1479.

[15] Cartis C, Gould N, Toint P (2011) On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming. *SIAM J. Optim.* 21(4):1721–1739.

[16] Cartis C, Gould N, Toint P (2013) On the evaluation complexity of cubic regularization methods for potentially rank-deficient nonlinear least-squares problems and its relevance to constrained nonlinear optimization. *SIAM J. Optim.* 23(3):1553–1574.

[17] Cartis C, Gould N, Toint P (2017) Corrigendum: On the complexity of finding first-order critical points in constrained nonlinear optimization. *Math. Programming* 161(1–2):611–626.

[18] Cartis C, Gould NI, Toint PL (2018) Optimality of orders one to three and beyond: characterization and evaluation complexity in constrained nonconvex optimization. *J. Complexity* 53:8–94.

[19] Cartis C, Gould NI, Toint PL (2019) *Evaluation Complexity Bounds for Smooth Constrained Nonlinear Optimization Using Scaled KKT Conditions and High-Order Models Approximation and Optimization* (Springer, New York).

[20] Craven B (1978) *Mathematical Programming and Control Theory* (Springer Science & Business Media, New York).
[21] Daneshmand A, Facchinei F, Kungurtsev V, Scutari G (2015) Hybrid random/deterministic parallel algorithms for convex and nonconvex big data optimization. *IEEE Trans. Signal Processing* 63(15):3914–3929.
[22] Cannelli L, Facchinei F, Scutari G, Kungurtsev V (2020). Asynchronous optimization over graphs: Linear convergence under error bound conditions. *IEEE Transactions on Automatic, Control*, Early access online publication, doi: 10.1109/TAC.2020.3033490.
[23] Davis D, Drusvyatskiy D (2019) Stochastic model-based minimization of weakly convex functions. *SIAM J. Optim.* 29(1):207–239.
[24] Davis D, Drusvyatskiy D, Kakade S, Lee JD (2019) Stochastic subgradient method converges on tame functions. *Foundations Comput. Math.* 20:119–154.
[25] Dean J, Corrado G, Monga R, Chen K, Devin M, Mao M, Senior A, Tucker P, Yang K, Le Q (2012) Large scale distributed deep networks. Pereira F, Burges CJC, Bottou L, Weinberger KQ, eds, *Advances in Neural Information Processing Systems* vol. XX (Curran Associates, Red Hook, NY), 1223–1231.
[26] Di Lorenzo P, Scutari G (2016) Next: In-network nonconvex optimization. *IEEE Trans. Signal Inform. Processing Networks* 2(2):120–136.
[27] Di Pillo G, Grippo L (1988) On the exactness of a class of nondifferentiable penalty functions. *J. Optim. Theory Appl.* 57(3):399–410.
[28] Di Pillo G, Grippo L (1989) Exact penalty functions in constrained optimization. *SIAM J. Control Optim.* 27(6):1333–1360.
[29] El-Alem M (1999) A global convergence theory for dennis, el-alem, and maciel's class of trust-region algorithms for constrained optimization without assuming regularity. *SIAM J. Optim.* 9(4):965–990.
[30] Facchinei F (1997) Robust recursive quadratic programming algorithm model with global and superlinear convergence properties. *J. Optim. Theory Appl.* 92(3):543–579.
[31] Facchinei F, Pang JS (2003) *Finite-Dimensional Variational Inequalities and Complementarity Problems* (Springer, New York).
[32] Facchinei F, Lampariello L, Scutari G (2017) Feasible methods for nonconvex nonsmooth problems with applications in green communications. *Math. Program.* 164(1–2):55–90.
[33] Facchinei F, Scutari G, Sagratella S (2015) Parallel selective algorithms for nonconvex big data optimization. *IEEE Trans. Signal Processing* 63(7):1874–1889.
[34] Fletcher R (1985) An $\ell_1$ penalty method for nonlinear constraints. *Proc. SIAM Conf. Numerical Optim.*, vol, 20, 26–40.
[35] Gupta M, Bengio S, Weston J (2014) Training highly multiclass classifiers. *J. Machine Learning Res.* 15(1):1461–1492.
[36] Hong M, Razaviyayn M, Luo ZQ, Pang JS (2016) A unified algorithmic framework for block-structured optimization involving big data: With applications in machine learning and signal processing. *IEEE Signal Processing Magazine* 33(1):57–77.
[37] Hunter D, Lange K (2004) A tutorial on MM algorithms. *Amer. Statist.* 58(1):30–37.
[38] Kingma D, Ba J (2014) Adam: A method for stochastic optimization. Preprint, submitted December 22, 2014, https://arxiv.org/abs/1412.6980.
[39] Lipp T, Boyd S (2016) Variations and extension of the convex–concave procedure. *Optim. Engrg.* 17(2):263–287.
[40] Liu X, Sun J (2004) A robust primal-dual interior-point algorithm for nonlinear programs. *SIAM J. Optim.* 14(4):1163–1186.
[41] Liu X, Yuan Y (2000) A robust algorithm for optimization with general equality and inequality constraints. *SIAM J. Sci. Comput.* 22(2):517–534.
[42] Liu X, Yuan Y (2011) A sequential quadratic programming method without a penalty function or a filter for nonlinear equality constrained optimization. *SIAM J. Optim.* 21(2):545–571.
[43] Mairal J (2013) Optimization with first-order surrogate functions. Dasgupta S, McAllester D, eds. *Proc. 30th International Conf. on Machine Learning*, PMLR 28(3):783–791.
[44] Martınez J (2017) On high-order model regularization for constrained optimization. *SIAM J. Optim.* 27(4):2447–2458.
[45] Nedic A, Ozdaglar A, Parrilo P (2010) Constrained consensus and optimization in multi-agent networks. *IEEE Trans. Automatic Control* 55(4):922–938.
[46] Nemirovski A, Juditsky A, Lan G, Shapiro A (2009) Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.* 19(4):1574–1609.
[47] Nesterov Y (2013) *Introductory Lectures on Convex Optimization: A Basic Course*, vol. 87 (Springer Science & Business Media, New York).
[48] Nesterov Y, Polyak B (2006) Cubic regularization of Newton method and its global performance. *Math. Programming* 108(1):177–205.
[49] Ng TY, Yu W (2007) Joint optimization of relay strategies and resource allocations in cooperative cellular networks. *IEEE J. Selected Areas Comm.* 25(2):328–339.
[50] Polyak B (1987) *Introduction to Optimization. Translations Series in Mathematics and Engineering* (Optimization Software, New York).
[51] Razaviyayn M, Hong M, Luo ZQ (2013) A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM J. Optim.* 23(2):1126–1153.
[52] Rockafellar R (1982) *Lagrange Multipliers and Subderivatives of Optimal Value Functions in Nonlinear Programming. Nondifferential and Variational Techniques in Optimization* (Springer, New York).
[53] Rockafellar R (1985) Lipschitzian properties of multifunctions. *Nonlinear Anal. Theory Methods Appl.* 9(8):867–885.
[54] Rockafellar R, Wets J (1998) *Variational Analysis* (Springer, New York).
[55] Scutari G, Facchinei F, Lampariello L (2017) Parallel and distributed methods for constrained nonconvex optimization—Part I: Theory. *IEEE Trans. Signal Processing* 65(8):1929–1944.
[56] 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 Trans. Signal Processing* 65(8):1945–1960.
[57] Scutari G, Facchinei F, Song P, Palomar D, Pang JS (2014) Decomposition by partial linearization: Parallel optimization of multi-agent systems. *IEEE Trans. Signal Processing* 62(3):641–656.
[58] Shor N (1985) *Minimization Methods for Non-Differentiable Functions*, vol. 3 (Springer Science & Business Media, New York).
[59] Solodov M (2004) On the sequential quadratically constrained quadratic programming methods. *Math. Oper. Res.* 29(1):64–79.

[60] Solodov M (2009) Global convergence of an SQP method without boundedness assumptions on any of the iterative sequences. *Math. Programming* 118(1):1–12.
[61] Sun Y, Babu P, Palomar D (2017) Majorization-minimization algorithms in signal processing, communications, and machine learning. *IEEE Trans. Signal Processing* 65(3):794–816.
[62] Svanberg K (2002) A class of globally convergent optimization methods based on conservative convex separable approximations. *SIAM J. Optim.* 12(2):555–573.
[63] Tatarenko T, Touri B (2017) Non-convex distributed optimization. *IEEE Trans. Automatic Control* 62(8):3744–3757.
[64] Vavasis S (1993) Black-box complexity of local minimization. *SIAM J. Optim.* 3(1):60–80.
[65] Wang X, Ma S, Goldfarb D, Liu W (2017) Stochastic quasi-Newton methods for nonconvex stochastic optimization. *SIAM J. Optim.* 27(2):927–956.
[66] Yen N (1995) Hölder continuity of solutions to a parametric variational inequality. *Appl. Math. Optim.* 31(3):245–255.
[67] Yuan Y (1995) On the convergence of a new trust region algorithm. *Numerical Math.* 70(4):515–539.
[68] Zeng J, Yin W (2018) On nonconvex decentralized gradient descent. *IEEE Trans. Signal Processing* 66(11):2834–2848.