

Feasible methods for nonconvex nonsmooth problems with applications in green communications

Francisco Facchinei¹  · Lorenzo Lampariello² · Gesualdo Scutari³

Received: 29 April 2015 / Accepted: 13 September 2016 / Published online: 1 October 2016
© Springer-Verlag Berlin Heidelberg and Mathematical Optimization Society 2016

Abstract We propose a general feasible method for nonsmooth, nonconvex constrained optimization problems. The algorithm is based on the (inexact) solution of a sequence of strongly convex optimization subproblems, followed by a step-size procedure. Key features of the scheme are: (i) it preserves feasibility of the iterates for nonconvex problems with nonconvex constraints, (ii) it can handle nonsmooth problems, and (iii) it naturally leads to parallel/distributed implementations. We illustrate the application of the method to an open problem in green communications whereby the energy consumption in MIMO multiuser interference networks is minimized, subject to nonconvex Quality-of-Service constraints.

Keywords Feasible method · Nonconvex problem · Nonsmooth optimization · Parallel and distributed implementation · Green communications

The work of Scutari was supported by the USA National Science Foundation (NSF) under Grants CIF 1564044, CCF 1632599 and CAREER Award No. 1555850, and the Office of Naval Research (ONR) Grant N00014-16-1-2244.

✉ Francisco Facchinei
francisco.facchinei@uniroma1.it

Lorenzo Lampariello
lorenzo.lampariello@uniroma3.it

Gesualdo Scutari
gscutari@purdue.edu

¹ Department of Computer, Control, and Management Engineering Antonio Ruberti, University of Rome La Sapienza, Via Ariosto 25, 00185 Rome, Italy

² Department of Business Studies, University of RomaTre, Via Silvio D'Amico 77, 00145 Rome, Italy

³ School of Industrial Engineering, Purdue University, West Lafayette, IN 47907-5400, USA

Mathematics Subject Classification 90C30 · 90C26 · 65K10 · 90C90

1 Introduction

We consider the solution of the problem

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & U(\mathbf{x}) \\ \text{s.t.} & \mathbf{x} \in \mathcal{X}, \end{array} \quad (\text{P})$$

where $\mathcal{X} \subseteq \mathbb{R}^n$ is a nonempty closed set and $U : O \rightarrow \mathbb{R}$ is locally Lipschitz on O , an open set containing \mathcal{X} . We are interested in a method for finding stationary points of the nonconvex problem (P) with the following properties: (i) feasibility is maintained throughout iterates, (ii) nonsmooth (objective or constraint) functions can be handled, and (iii) the algorithm is possibly amenable to parallel/distributed implementations.

Properties (i) and (iii) are of well-recognized interest and have been extensively studied in the literature, somewhat independently, both by optimizers and different engineering communities. Among these results, we list below the main approaches, paying attention to the ability to handle nonsmooth problems and with an emphasis on those methods that can be applied to nonconvex problems and, in particular, to problems with nonconvex constraints. Standard feasible approaches include the classical logarithmic barrier Interior-Point (IP) method by Fiacco and Mc Cormick [18], see also [10, 13, 21] for modern developments that try to overcome the numerical pitfalls of the original algorithm, and the Feasible Sequential Quadratic Programming (FSQP) method of Tits, see e.g. [27, 35–37] and [11, 14, 25, 38, 50] for further developments. Both IP and FSQP methods require smoothness and are not suited to parallel/distributed implementations. Parallel Variable Distribution (PVD) schemes, proposed in [17, 40, 46] for smooth problems, are suitable for implementation over parallel architectures. However, PVD methods require an amount of information exchange/knowledge that is often not compatible with a distributed architecture and they call for the solution of possibly difficult nonconvex subproblems; moreover, convergence has been established only for convex constraints [17, 46] or nonconvex constraints with a Cartesian product structure [40]. In [34] methods for DC programs are discussed. Among them, Algorithm II generates feasible iterates; parallel/distributed schemes are also considered, but only for problems with a convex feasible set. While the methods proposed in [34] enjoy very strong convergence properties, in that they converge to B-stationary points, they are potentially computationally very intensive and only apply to a specific subclass of DC programs. In [8] new results are proved on a proximal alternating linearized minimization method for nonconvex and nonsmooth problems. This method maintains feasibility but requires at each iteration the solution of (generally) nontrivial nonsmooth and nonconvex subproblems. The paper [48] is the rigorous culmination of a series of works in the structural engineering community such as the CONVex LINearization (CONLIN) method [19, 20] and the Method of Moving Asymptotes (MMA) [47]. It introduces a large class of feasible schemes for smooth problems, potentially suited to parallel/distributed implementations, called Conservative Convex Separable Approximation (CCSA). A serious

drawback of CCSA is that, besides requiring the difficult tuning of some unknown parameters, it finds stationary points of a reformulation that are not necessarily stationary also for the original problem. To some extent similar to CCSA methods is the INner Convex Approximation (INCA) algorithm, which was first proposed in 1978 by Marks and Wright [30] for problems with a smooth convex objective function and smooth nonconvex constraints. In the INCA approach a sequence of convex subproblems is solved where the original nonconvex constraints are replaced by *upper* convex approximations so that the algorithms generate only feasible iterates. INCA has been recently more rigorously studied in [3], in the case of a strongly convex objective function. In [1], the authors, unaware of [30], use a specific inner convex approximation for the feasible set and, in the spirit of majorization/minimization techniques, also approximate the possibly nonconvex objective function with an upper convex quadratic, so as to solve at each iteration a surrogate problem with a convex quadratic objective function and convex quadratic constraints. The resulting algorithm, which is potentially suited to parallel/distributed implementations, cannot deal with nonsmooth problems and the proposed approximation for the objective function, based on the global Lipschitz property of its gradient, is possibly hard to obtain and, more importantly, can lead to poor numerical performances (see Sect. 6).

As a major departure from the above approaches, in this work we consider general nonconvex problems and quite an ample class of nonsmooth functions and provide easily implementable solutions methods for (P) that, in many cases, with no difficulty can also lead to parallel and distributed implementations. Building on the INCA paradigm, along the lines put forward in [43, 44], our method is based on the (inexact) solution of a sequence of strongly convex optimization subproblems, followed by a step-size procedure. At each iteration \mathbf{x}^ν , a subproblem is formed of the type

$$\begin{aligned} & \underset{\mathbf{x}}{\text{minimize}} \quad \tilde{U}(\mathbf{x}; \mathbf{x}^\nu) \\ & \text{s.t.} \quad \mathbf{x} \in \tilde{\mathcal{X}}(\mathbf{x}^\nu), \end{aligned} \tag{P_{\mathbf{x}^\nu}}$$

where $\tilde{U}(\bullet; \mathbf{x}^\nu)$ is a strongly convex “approximation” of U and $\tilde{\mathcal{X}}(\mathbf{x}^\nu)$ is a closed, convex, *inner* approximation of \mathcal{X} , meaning that $\tilde{\mathcal{X}}(\mathbf{x}^\nu) \subseteq \mathcal{X}$. Let $\hat{\mathbf{x}}(\mathbf{x}^\nu)$ be the unique solution of problem (P _{\mathbf{x}^ν}); the underlying idea of the approach is that (the subproblems (P _{\mathbf{x}^ν}) are chosen so that) one is able to compute efficiently $\hat{\mathbf{x}}(\mathbf{x}^\nu)$. The new iteration $\mathbf{x}^{\nu+1}$ is then given by moving from \mathbf{x}^ν towards (an approximation of) $\hat{\mathbf{x}}(\mathbf{x}^\nu)$ according to a step-size γ^ν . We prove that, under appropriate assumptions, every limit point of the sequence $\mathbf{x}^{\nu+1}$ is stationary for the original problem (P).

Our main contributions are: (a) the definition and analysis of a general algorithmic framework enjoying properties (i), (ii), and (iii) and vastly improving over all the approaches described above; (b) the first practical feasible method for ample classes of *nonsmooth, nonconvex* problems requiring at each iteration only the solution of strongly convex subproblems (as opposite to [8, 34]); (c) a systematic approach to the development of parallel and distributed solution methods that can be applied to classes of problems for which parallel and distributed methods were not available; (d) the analysis of several step-size (Armijo-type, diminishing and fixed step-size) rules, and of corresponding inexactness criteria for the solution of problem (P _{\mathbf{x}^ν});

(e) the first numerically efficient solution method for an important open problem in *green communications*, paving the way to the application of our algorithmic framework to a host of unsolved resource allocation problems in interference networks. During the reviewing process, a referee brought to our attention the interesting, recent paper [7]. Also in [7] a sequence of strongly convex approximating subproblems is solved; in particular, following the INCA approach, the authors resort to inner convex approximations of the feasible set. The analysis in [7] is very elegant and shows full convergence to a single limit point; moreover, convergence rates aspects are investigated. There are however several differences with our approach: (i) [7] follows the classical majorization/minimization idea, as in [1], i.e., at each iteration the nonconvex objective function is approximated by an upper convex estimate; this feature, on the one hand simplifies the convergence analysis but, on the other hand, may be numerically impractical, see the eloquent numerical results in Sect. 6; (ii) the analysis is geared towards smooth problems and the assumptions made are strong, so that only simple and well-behaved forms of nonsmoothness, under (Clarke's) regularity, can actually be treated, see Sect. 5; (iii) no consideration is given to practical aspects like inexactness and parallel/distributed implementations.

The paper is organized as follows. In the next section we recall some basic background material on nonsmooth analysis and optimality conditions. In Sect. 3 we give an overall description of our approach and present some preliminary results, while in Sect. 4 we provide an in-depth convergence analysis. Section 5 is devoted to a detailed presentation of approximations that can be used in our framework and to their use in the development of parallel/distributed versions of our algorithm. Section 6 is dedicated to the application of our results to an open problem in green communications and shows how the flexibility of our approach easily permits to tailor the method to specific problems and obtain some impressive numerical results.

2 Background material

In this section we briefly recall some basic nonsmooth analysis notions, but we refer the reader to [31, 39] for more details and other results that, in some cases, will be freely invoked. We assume that all sets C and functions f considered in this section are closed and locally Lipschitz continuous, respectively.

2.1 Continuity properties of set-valued mappings

Let $M : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ be a set-valued mapping and C a subset of \mathbb{R}^n .

Definition 1 ([39, Definition 5.4]) The set-valued mapping M is outer semicontinuous relative to C at $\bar{\mathbf{x}} \in C$ if $\limsup_{\mathbf{x} \rightarrow \bar{\mathbf{x}}} M(\mathbf{x}) \subseteq M(\bar{\mathbf{x}})$, and inner semicontinuous at $\bar{\mathbf{x}}$ if $\liminf_{\mathbf{x} \rightarrow \bar{\mathbf{x}}} M(\mathbf{x}) \supseteq M(\bar{\mathbf{x}})$. M is called continuous (relative to C) at $\bar{\mathbf{x}}$ if both conditions hold.

The following definition extends to set-valued mappings the notion of locally bounded function.

Definition 2 ([39, Definition 5.14]) The set-valued mapping M is locally bounded at $\bar{x} \in \mathbb{R}^n$ if, for some neighborhood $\mathcal{V}_{\bar{x}}$ of \bar{x} , the set $M(\mathcal{V}_{\bar{x}}) \subseteq \mathbb{R}^m$ is bounded.

Finally, we recall the Aubin property, which is a Lipschitz-type condition that makes reference to specific points in the graph of M .

Definition 3 ([39, Definition 9.36]) The set-valued mapping M has the Aubin property relative to C at $\bar{x} \in C$ for $\bar{u} \in M(\bar{x})$, if the graph of M is locally closed at (\bar{x}, \bar{u}) and there are neighborhoods $\mathcal{V}_{\bar{x}}$ of \bar{x} , $\mathcal{W}_{\bar{u}}$ of \bar{u} , and a constant $\kappa \in \mathbb{R}_+ \triangleq \{t \in \mathbb{R} : t \geq 0\}$ such that, denoting by \mathbb{B} the closed unit ball,

$$M(\mathbf{y}) \cap \mathcal{W}_{\bar{u}} \subseteq M(\mathbf{x}) + \kappa \|\mathbf{y} - \mathbf{x}\| \mathbb{B} \quad \text{for all } \mathbf{x}, \mathbf{y} \in C \cap \mathcal{V}_{\bar{x}}.$$

2.2 Normal cones and subgradients

Definition 4 ([39, Definition 6.3, 6(19)]) Let a closed set $C \subseteq \mathbb{R}^n$ and a point $\bar{x} \in C$ be given. A vector \mathbf{v} is a regular normal to C at \bar{x} if $\mathbf{v}^T(\mathbf{y} - \bar{x}) \leq o(\|\mathbf{y} - \bar{x}\|)$, for $\mathbf{y} \in C$. The set of regular normals to C at \bar{x} is denoted by $\widehat{N}_C(\bar{x})$. Moreover, \mathbf{v} is a normal to C at \bar{x} if there are sequences $\mathbf{x}^\nu \xrightarrow{c} \bar{x}$ and $\mathbf{v}^\nu \rightarrow \mathbf{v}$ with $\mathbf{v}^\nu \in \widehat{N}_C(\mathbf{x}^\nu)$. The set of normals to C at \bar{x} is denoted by $N_C(\bar{x})$. Lastly, the convexified normal cone to C at \bar{x} is defined by $\overline{N}_C(\bar{x}) = \text{cl co } N_C(\bar{x})$, where cl and co denote the closure and the convex hull of a set.

The set C is termed regular at $\bar{x} \in C$ if $\widehat{N}_C(\bar{x}) = N_C(\bar{x})$; convex sets are regular at each of their points. With the aid of normal cones we can give the following definition of subgradients, see [31, Definition 1.77, 1.83 and (2.73)] or [39, Theorem 8.9 and 8(32)].

Definition 5 The set of regular subgradients of f at $\bar{x} \in C$ is $\widehat{\partial}f(\bar{x}) \triangleq \{\mathbf{v} | (\mathbf{v}, -1) \in \widehat{N}_{\text{epi}f}(\bar{x}, f(\bar{x}))\}$. The set of subgradients of f at $\bar{x} \in C$ is $\partial f(\bar{x}) \triangleq \{\mathbf{v} | (\mathbf{v}, -1) \in N_{\text{epi}f}(\bar{x}, f(\bar{x}))\}$. The set of Clarke subgradients of f at $\bar{x} \in C$ is $\bar{\partial}f(\bar{x}) \triangleq \{\mathbf{v} | (\mathbf{v}, -1) \in \overline{N}_{\text{epi}f}(\bar{x}, f(\bar{x}))\}$.

We recall [39, Theorems 8.6, 8.49] that, if $\bar{x} \in C$, $\widehat{\partial}f(\bar{x})$, $\partial f(\bar{x})$ and $\bar{\partial}f(\bar{x})$ are closed with $\widehat{\partial}f(\bar{x})$ convex, $\widehat{\partial}f(\bar{x}) \subseteq \partial f(\bar{x})$ and, by [39, Theorems 8.49 and 9.13], $\bar{\partial}f(\bar{x}) = \text{co } \partial f(\bar{x})$. Letting \bar{x} be a point in C , by [39, Theorem 8.7],

- ∂f is outer semicontinuous relative to C ,

and the following properties of the subgradient set characterizes locally Lipschitz continuous functions, see [39, Theorem 9.13]:

- $\partial f(\bar{x})$ is nonempty and compact;
- both $\widehat{\partial}f$ and ∂f are locally bounded as set-valued mappings.

The locally Lipschitz function f is said to be regular at $\bar{x} \in C$ if $\text{epi}f$ is regular at $(\bar{x}, f(\bar{x}))$. Furthermore, f is regular at $\bar{x} \in C$ if and only if $\widehat{\partial}f(\bar{x}) = \partial f(\bar{x})$ and, thus, the three subgradient sets coincide. In particular, if f is convex on C or continuously differentiable on an open set containing C , then it is regular at each point of C . Besides, in the former case, the set of subgradients $\partial f(\bar{x})$ coincides with the classical subdifferential from convex analysis while, in the latter, the set of subgradients collapses to the gradient.

2.3 Optimality conditions and constraint qualifications

Consider the constrained optimization problem

$$\begin{aligned} & \underset{\mathbf{x}}{\text{minimize}} && f(\mathbf{x}) \\ & \text{s.t.} && \mathbf{x} \in C, \end{aligned} \quad (1)$$

where $C \subseteq \mathbb{R}^n$ is a nonempty, closed set and f is locally Lipschitz on an open set containing C . The basic necessary optimality condition for a point $\bar{\mathbf{x}} \in C$ to be (locally) optimal for this problem is, see [39, Theorems 8.15 and 9.13],

$$\mathbf{0} \in \partial f(\bar{\mathbf{x}}) + N_C(\bar{\mathbf{x}}). \quad (2)$$

When f and C are convex, the previous condition is also sufficient for $\bar{\mathbf{x}}$ to be globally optimal. A key issue in applying this condition is the capability to estimate the normal cone $N_C(\bar{\mathbf{x}})$. To this end, as usual, we suppose that $C \triangleq \{\mathbf{x} \in S : h_j(\mathbf{x}) \leq 0, j = 1, \dots, m\}$, with $S \subseteq \mathbb{R}^n$ closed and, for every j , h_j locally Lipschitz continuous on an open set containing S . In this setting, if we suppose that a standard constraint qualification holds, we can obtain a “dual” description of normal vectors.

Definition 6 The (nonsmooth) Mangasarian–Fromovitz Constraint Qualification (MFCQ) holds at $\bar{\mathbf{x}} \in C$ if

$$\mathbf{0} \in \sum_{j=1}^m \mu_j \partial h_j(\bar{\mathbf{x}}) + N_S(\bar{\mathbf{x}}), \quad \boldsymbol{\mu} \in N_{\mathbb{R}_-^m}(\mathbf{h}(\bar{\mathbf{x}})) \Rightarrow \boldsymbol{\mu} = \mathbf{0}, \quad (3)$$

where $\mathbf{h}(\mathbf{x}) \triangleq (h_1(\mathbf{x}), \dots, h_m(\mathbf{x}))^T$ and $\mathbb{R}_-^m \triangleq \{\boldsymbol{\mu} \in \mathbb{R}^m \mid \boldsymbol{\mu} \leq \mathbf{0}\}$.

A somewhat weaker condition can be obtained by replacing $\sum_{j=1}^m \mu_j \partial h_j(\bar{\mathbf{x}})$ by $\partial(\sum_{j=1}^m \mu_j h_j(\bar{\mathbf{x}}))$; we refrain from doing this, since, for our purpose, the MFCQ (3) is what we can really expect to check and use in practice. The condition $\boldsymbol{\mu} \in N_{\mathbb{R}_-^m}(\mathbf{h}(\bar{\mathbf{x}}))$ can be rewritten, taking into account $\bar{\mathbf{x}} \in C$, as $\mu_j \geq 0$, $\mu_j h_j(\bar{\mathbf{x}}) = 0$ for all j . It is worth pointing out that, if $S = \mathbb{R}^n$ and all functions involved are smooth, (3) reduces to the classical MFCQ.

In view of the description of normal vectors for Lipschitzian constraints given in [39, Corollary 10.50], if (3) holds at $\bar{\mathbf{x}} \in C$, thanks to [39, Theorem 9.13, Corollary 10.9] and by [39, Theorem 6.42], we have

$$N_C(\bar{\mathbf{x}}) \subseteq \bigcup \left\{ \sum_{j=1}^m \mu_j \partial h_j(\bar{\mathbf{x}}) \mid \boldsymbol{\mu} \in N_{\mathbb{R}_-^m}(\mathbf{h}(\bar{\mathbf{x}})) \right\} + N_S(\bar{\mathbf{x}}), \quad (4)$$

with equality holding if all functions h_j s and the set S are regular at $\bar{\mathbf{x}}$. A particularly important case is the convex one: whenever all functions h_j s and the set S are convex, and hence regular, equality holds in (4) and we get an exact representation of the

normal cone. Combining relations (2) and (4), it is straightforward to obtain KKT-like optimality conditions. Consider problem (1), with C having the constraint structure just discussed. Let $\bar{\mathbf{x}} \in C$ be a minimum point at which the MFCQ holds, then

$$\mathbf{0} \in \partial f(\bar{\mathbf{x}}) + \bigcup \left\{ \sum_{j=1}^m \mu_j \partial h_j(\bar{\mathbf{x}}) \mid \boldsymbol{\mu} \in N_{\mathbb{R}_-^m}(\mathbf{h}(\bar{\mathbf{x}})) \right\} + N_S(\bar{\mathbf{x}}). \tag{5}$$

The KKT-like conditions (5), although less stringent than the necessary condition for local optimality (2) (but the two are equivalent for regular data), are the stationarity conditions generally used when studying algorithms.

Remark 1 In this paper we express all optimality conditions and constraint qualifications by resorting to subgradients, *not* to Clarke subgradients. This choice allows us to get sharper optimality conditions and weaker constraint qualifications. The price we pay is a more complex technical analysis for some of our results. Nevertheless, all the conclusions in this paper still hold if Clarke subgradients are employed.

3 Outline of the method

Our aim is to compute stationary solutions of (P) while preserving feasibility of the iterates, by solving a sequence of “easier” convex subproblems. In this section we give a description of the overall method and present some preliminary technical results; the complete convergence properties are investigated in the next section.

Given the current feasible iterate \mathbf{x}^ν , we consider a strongly convex “approximation” to the original nonconvex problem (P). The convex feasible set $\tilde{\mathcal{X}}(\mathbf{x}^\nu)$ of this subproblem is contained in the feasible region \mathcal{X} of (P) and is such that $\mathbf{x}^\nu \in \tilde{\mathcal{X}}(\mathbf{x}^\nu)$. We then compute an (approximate) solution \mathbf{v}^ν of the subproblem and generate the new iteration $\mathbf{x}^{\nu+1}$ as a suitably chosen point on the segment $[\mathbf{x}^\nu, \mathbf{v}^\nu]$. By construction, $\mathbf{x}^{\nu+1}$ still belongs to \mathcal{X} since the whole segment $[\mathbf{x}^\nu, \mathbf{v}^\nu]$ is contained in $\tilde{\mathcal{X}}(\mathbf{x}^\nu) \subseteq \mathcal{X}$. Clearly, in order to prove convergence we need to make proper assumptions: these conditions are discussed in this and in the next section, while in Sect. 5 we present many examples of approximations that satisfy the required assumptions. In this section we only make minimal assumptions and, accordingly, prove a weak convergence result, see Proposition 1. Although Proposition 1 is of little interest *per se*, it constitutes the basis for the developments in the next section.

We introduce the following approximation of (P) at a feasible point $\mathbf{y} \in \mathcal{X}$:

$$\begin{aligned} & \underset{\mathbf{x}}{\text{minimize}} \quad \tilde{U}(\mathbf{x}; \mathbf{y}) \\ & \text{s.t.} \quad \mathbf{x} \in \tilde{\mathcal{X}}(\mathbf{y}) \end{aligned} \tag{P_y}$$

where $\tilde{U}(\mathbf{x}; \mathbf{y})$ and $\tilde{\mathcal{X}}(\mathbf{y})$ are convex approximations at \mathbf{y} of U and \mathcal{X} , respectively. The basic requirement on \tilde{U} is that it capture the first-order behavior of U at \mathbf{y} while being strongly convex. In order to make subproblem (P_y) convex, we also need $\tilde{\mathcal{X}}(\mathbf{y})$ to be convex; in addition, we also impose $\tilde{\mathcal{X}}(\mathbf{y})$ to contain \mathbf{y} and to be an inner approximation

of \mathcal{X} . In what follows $O_{\mathbf{x}}$ and $O_{\mathbf{y}}$ are an open, convex set and an open set containing \mathcal{X} , respectively.

Assumption U Let $\tilde{U} : O_{\mathbf{x}} \times O_{\mathbf{y}} \rightarrow \mathbb{R}$ satisfy the following properties:

- (U1) $\tilde{U}(\bullet; \mathbf{y})$ is a finite strongly convex function on $O_{\mathbf{x}}$ for every $\mathbf{y} \in \mathcal{X}$ with modulus of strong convexity independent of \mathbf{y} ;
- (U2) $\partial_1 \tilde{U}(\bullet; \bullet)$ is locally bounded relative to $\mathcal{X} \times \mathcal{X}$ at every (\mathbf{y}, \mathbf{y}) with $\mathbf{y} \in \mathcal{X}$;
- (U3) $\limsup_{(\mathbf{u}, \mathbf{w}) \xrightarrow{\mathcal{X} \times \mathcal{X}} (\mathbf{y}, \mathbf{y})} \partial_1 \tilde{U}(\mathbf{u}; \mathbf{w}) \subseteq \partial U(\mathbf{y})$ for every $\mathbf{y} \in \mathcal{X}$;

where $\partial_1 \tilde{U}(\mathbf{u}; \mathbf{w})$ denotes the set of subgradients of $\tilde{U}(\bullet; \mathbf{w})$ evaluated at \mathbf{u} .

Assumption X (X1) $\tilde{\mathcal{X}}(\mathbf{y})$ is closed and convex for every $\mathbf{y} \in \mathcal{X}$;
 (X2) $\mathbf{y} \in \tilde{\mathcal{X}}(\mathbf{y})$ and $\tilde{\mathcal{X}}(\mathbf{y}) \subseteq \mathcal{X}$ for every $\mathbf{y} \in \mathcal{X}$.

U1 and X1, together with X2, which ensures that the feasible set $\tilde{\mathcal{X}}(\mathbf{y})$ is nonempty, guarantee that all subproblems $(P_{\mathbf{y}})$ are strongly convex and have a unique solution, which we denote by $\hat{\mathbf{x}}(\mathbf{y})$:

$$\hat{\mathbf{x}}(\mathbf{y}) \triangleq \operatorname{argmin}_{\mathbf{x} \in \tilde{\mathcal{X}}(\mathbf{y})} \tilde{U}(\mathbf{x}; \mathbf{y}).$$

Our Algorithmic Framework is based on the successive (approximate) solution of subproblems $(P_{\mathbf{y}})$ followed by some kind of line-search procedure.

Algorithmic Framework (AF) 1

Data: $\varepsilon^v \geq 0$, $\mathbf{x}^0 \in \mathcal{X}$; set $v = 0$.

(S. 1) If \mathbf{x}^v satisfies a stationarity condition for (P): STOP.

(S. 2) Solve problem $(P_{\mathbf{x}^v})$ with accuracy ε^v :
 find $\mathbf{v}^v \in \tilde{\mathcal{X}}(\mathbf{x}^v)$ s.t. $\|\mathbf{v}^v - \hat{\mathbf{x}}(\mathbf{x}^v)\| \leq \varepsilon^v$.

(S. 3) Set $\mathbf{x}^{v+1} = \mathbf{x}^v + \gamma^v(\mathbf{v}^v - \mathbf{x}^v)$ for some $\gamma^v \in (0, 1]$.

(S. 4) $v \leftarrow v + 1$ and go to step (S. 1).

Assumptions U and X alone are not enough to establish the convergence of AF 1 to stationary points of problem (P), and yet some instructive aspects of AF 1 can be investigated under Assumptions U and X alone. U2 and U3 establish a link between the first-order properties of U and those of \tilde{U} . To illustrate their meaning, let us consider first the simple case of continuously differentiable U and \tilde{U} : U2 is then automatically satisfied, while U3 simply postulates that $\nabla_1 \tilde{U}(\mathbf{y}; \mathbf{y}) = \nabla U(\mathbf{y})$ for all $\mathbf{y} \in \mathcal{X}$, i.e. that $\tilde{U}(\bullet; \mathbf{y})$ and U have the same first-order behavior at the “base point” \mathbf{y} . Assume now that \tilde{U} is locally Lipschitz continuous: Lemma 1, in view of the local Lipschitz continuity of U , shows that, whenever \tilde{U} is (jointly) locally Lipschitz in both its variables and $\tilde{U}(\bullet; \mathbf{y})$ is convex for every $\mathbf{y} \in \mathcal{X}$ (as it must be by U1), U2 is automatically satisfied and, by virtue of the outer semicontinuity property of the partial subgradient set-valued mapping, U3 boils down to the simple condition $\partial_1 \tilde{U}(\mathbf{y}; \mathbf{y}) \subseteq \partial U(\mathbf{y})$ for all $\mathbf{y} \in \mathcal{X}$.

Lemma 1 *Let $\tilde{U} : O_{\mathbf{x}} \times O_{\mathbf{y}} \rightarrow \mathbb{R}$ be locally Lipschitz at $(\mathbf{y}, \mathbf{y}) \in \mathcal{X} \times \mathcal{X}$ and such that $\tilde{U}(\bullet; \mathbf{w})$ is convex on $O_{\mathbf{x}}$ for every $\mathbf{w} \in \mathcal{X}$ in a neighborhood of \mathbf{y} . Then, the set-valued mapping $\partial_1 \tilde{U}(\bullet; \bullet)$ is locally bounded and outer semicontinuous (relative to $\mathcal{X} \times \mathcal{X}$) at (\mathbf{y}, \mathbf{y}) .*

Proof Without loss of generality, suppose $(\mathbf{u}^v, \mathbf{w}^v) \xrightarrow{\mathcal{X} \times \mathcal{X}} (\mathbf{y}, \mathbf{y})$ and let $\{\xi^v\}$, with $\xi^v \in \partial_1 \tilde{U}(\mathbf{u}^v; \mathbf{w}^v)$, be a sequence of partial subgradients of $\tilde{U}(\bullet; \mathbf{w}^v)$ evaluated at \mathbf{u}^v . By [39, Corollary 10.11], we have $\xi^v \in \pi_{\mathbf{x}} \partial \tilde{U}(\mathbf{u}^v; \mathbf{w}^v)$, where $\pi_{\mathbf{x}}$ denotes the projection on the subspace of the first argument: therefore, there exists a sequence $\{\zeta^v\}$ such that $\{(\xi^v, \zeta^v)\} \in \partial \tilde{U}(\mathbf{u}^v; \mathbf{w}^v)$. By the local boundedness and outer semicontinuity (relative to $\mathcal{X} \times \mathcal{X}$) of the subgradient set of the locally Lipschitz function \tilde{U} , renumbering if necessary, we have $\{(\xi^v, \zeta^v)\} \rightarrow (\xi, \zeta) \in \partial \tilde{U}(\mathbf{y}; \mathbf{y})$. Since $\partial \tilde{U}(\mathbf{y}; \mathbf{y}) \subseteq \bar{\partial} \tilde{U}(\mathbf{y}; \mathbf{y})$, where $\bar{\partial} \tilde{U}(\mathbf{y}; \mathbf{y})$ is the set of Clarke subgradients of \tilde{U} at (\mathbf{y}, \mathbf{y}) , by [12, Proposition 2.5.3] and thanks to the regularity of $\tilde{U}(\bullet; \mathbf{y})$, we have $\xi \in \bar{\partial}_1 \tilde{U}(\mathbf{y}; \mathbf{y}) = \partial_1 \tilde{U}(\mathbf{y}; \mathbf{y})$; then, in view of [39, Proposition 5.15] and by Definition 1, the thesis follows. \square

In the general case, U2 and U3 suitably extend the previous considerations by requiring some consistency property between the subgradient set of U and the partial subgradient set of \tilde{U} at the “base point”.

Proposition 1 below essentially states that, if the sequence $\{\mathbf{x}^v\}$ produced by AF 1 is such that $\|\hat{\mathbf{x}}(\mathbf{x}^v) - \mathbf{x}^v\| \rightarrow 0$, then every limit point of $\{\mathbf{x}^v\}$ satisfies the “stationarity-looking” condition (7), which is formally similar to (2) applied to problem (P), but with the normal cone $N_{\mathcal{X}}(\bar{\mathbf{x}})$ replaced by the cone $\tilde{N}(\bar{\mathbf{x}})$, where the set $\tilde{N}(\bar{\mathbf{x}})$ is defined as

$$\begin{aligned} \tilde{N}(\mathbf{y}) &\triangleq \limsup_{\substack{\mathbf{u} \rightarrow \mathbf{y} \\ \mathbf{w} \rightarrow \mathbf{y}}} N_{\tilde{\mathcal{X}}(\mathbf{w})}(\mathbf{u}) \\ &= \{\eta \mid \exists \eta^v \rightarrow \eta, \mathbf{u}^v \rightarrow \mathbf{y}, \mathcal{X} \ni \mathbf{w}^v \rightarrow \mathbf{y} : \eta^v \in N_{\tilde{\mathcal{X}}(\mathbf{w}^v)}(\mathbf{u}^v)\}. \end{aligned}$$

Notice that, if $\mathbf{y} \in \mathcal{X}$, since $\mathbf{y} \in \tilde{\mathcal{X}}(\mathbf{y})$, then $\emptyset \neq N_{\tilde{\mathcal{X}}(\mathbf{y})}(\mathbf{y}) \subseteq \tilde{N}(\mathbf{y})$.

Proposition 1 *Let $\{\mathbf{x}^v\}$ be the sequence generated by AF 1 under Assumptions U and X.*

- (i) *All the iterates \mathbf{x}^v are feasible, i.e. they belong to \mathcal{X} .*
- (ii) *Suppose that $\{\mathbf{x}^v\}$ is bounded and such that*

$$\liminf_{v \rightarrow \infty} \|\hat{\mathbf{x}}(\mathbf{x}^v) - \mathbf{x}^v\| = 0. \tag{6}$$

Then, at least one limit point $\bar{\mathbf{x}}$ of $\{\mathbf{x}^v\}$ satisfies the condition

$$0 \in \partial U(\bar{\mathbf{x}}) + \tilde{N}(\bar{\mathbf{x}}). \tag{7}$$

- (iii) *Suppose that $\{\mathbf{x}^v\}$ is bounded and such that*

$$\lim_{v \rightarrow \infty} \|\hat{\mathbf{x}}(\mathbf{x}^v) - \mathbf{x}^v\| = 0. \tag{8}$$

Then, every limit point of $\{\mathbf{x}^v\}$ satisfies condition (7).

(iv) Whenever $\hat{\mathbf{x}}(\mathbf{x}^\nu) = \mathbf{x}^\nu$, the point \mathbf{x}^ν satisfies condition (7).

Proof (i) The thesis follows from Assumption X by induction and considering that $\mathbf{x}^{\nu+1}$ is a convex combination of $\mathbf{v}^\nu \in \tilde{\mathcal{X}}(\mathbf{x}^\nu)$ and $\mathbf{x}^\nu \in \tilde{\mathcal{X}}(\mathbf{x}^\nu)$, which is a convex set. We now prove (ii); (iii) follows applying (ii) to every convergent subsequence of $\{\mathbf{x}^\nu\}$. If (6) holds, by passing to subsequences, we have $(\hat{\mathbf{x}}(\mathbf{x}^\nu) - \mathbf{x}^\nu) \rightarrow 0$ and $\mathbf{x}^\nu \xrightarrow{\tilde{\mathcal{X}}} \bar{\mathbf{x}}$. By U1 and X1, $\hat{\mathbf{x}}(\mathbf{x}^\nu)$ is the unique global optimal solution of the strongly convex optimization problem $(P_{\mathbf{x}^\nu})$ for every ν ; this fact, in turn, is equivalent to the following condition:

$$0 \in \partial_1 \tilde{U}(\hat{\mathbf{x}}(\mathbf{x}^\nu); \mathbf{x}^\nu) + N_{\tilde{\mathcal{X}}(\mathbf{x}^\nu)}(\hat{\mathbf{x}}(\mathbf{x}^\nu)). \tag{9}$$

Thus, we can write $0 = \xi^\nu + \eta^\nu$, for some $\xi^\nu \in \partial_1 \tilde{U}(\hat{\mathbf{x}}(\mathbf{x}^\nu); \mathbf{x}^\nu)$ and $\eta^\nu \in N_{\tilde{\mathcal{X}}(\mathbf{x}^\nu)}(\hat{\mathbf{x}}(\mathbf{x}^\nu))$. By this and U2, we deduce that the sequence $\{\eta^\nu\}$ is bounded and we can assume, without loss of generality, that it converges to an element $\bar{\eta}$ belonging, by definition, to $\tilde{N}(\bar{\mathbf{x}})$. Thanks to U3, the thesis follows passing to the limit in $0 = \xi^\nu + \eta^\nu$. (iv) By (9) and U3, considering that $N_{\tilde{\mathcal{X}}(\mathbf{x}^\nu)}(\mathbf{x}^\nu) \subseteq \tilde{N}(\mathbf{x}^\nu)$, we have $0 \in \partial_1 \tilde{U}(\mathbf{x}^\nu; \mathbf{x}^\nu) + N_{\tilde{\mathcal{X}}(\mathbf{x}^\nu)}(\mathbf{x}^\nu) \subseteq \partial U(\mathbf{x}^\nu) + \tilde{N}(\mathbf{x}^\nu)$. \square

Proposition 1 sets the following agenda for the further study of AF 1.

1. What is the exact meaning of inclusion (7) and how does this condition relate to the classical “stationarity” condition (2)? The answer to this question revolves around the nature of cone $\tilde{N}(\bar{\mathbf{x}})$. As things stand now, since, except for X2, we made no assumptions linking \mathcal{X} and $\tilde{\mathcal{X}}(\bar{\mathbf{x}})$, (7) is not very informative. It is thus important to study concrete realizations of the inner approximation $\tilde{\mathcal{X}}$ that make (7) meaningful.
2. We will see that the validity of key conditions (6) and (8) depend critically on the choice of γ^ν (as well as on the characteristics of \tilde{U} and $\tilde{\mathcal{X}}$). We will consider fixed and diminishing step-size rules, and step-size determined by an Armijo-type condition.
3. In the subsequent analysis of the first two issues we will resort to additional conditions, beyond Assumptions U and X, on \tilde{U} and $\tilde{\mathcal{X}}$ (see Sect. 4.1). It is then essential to show that, in relevant cases, one can easily define suitable approximations \tilde{U} and $\tilde{\mathcal{X}}$ satisfying these assumptions, thus leading to practical solution methods.

Points 1 and 2 above will be discussed in the next section, while point 3 will be dealt with in Sect. 5.

4 The method in detail

In this section we present an in-depth analysis aimed at answering the questions of points 1-2 listed at the end of the previous section. In order to perform our analysis, we consider a more concrete realization of problem (P), that is,

$$\begin{aligned} & \underset{\mathbf{x}}{\text{minimize}} && F(\mathbf{x}) + H(\mathbf{x}) \\ & \text{s.t.} && g_j(\mathbf{x}) \leq 0, \quad j = 1, \dots, m \\ & && \mathbf{x} \in \mathcal{K}, \end{aligned} \tag{P1}$$

where $\mathcal{X} \triangleq \{\mathbf{x} \in \mathcal{K} : g_j(\mathbf{x}) \leq 0, j = 1, \dots, m\}$ is nonempty, with $\mathcal{K} \subseteq \mathbb{R}^n$ closed and convex and, letting $O \supseteq \mathcal{X}$ be open, $g_j : O \rightarrow \mathbb{R}$ locally Lipschitz continuous on O for every j ; moreover, $U \triangleq F + H$ is coercive on \mathcal{X} , with F continuously differentiable on O , its gradient Lipschitz continuous on O with constant $L_{\nabla F}$ and H locally Lipschitz continuous on O . We note that the above conditions guarantee the existence of a solution of problem (P1).

In this formulation the feasible set \mathcal{X} is described by the convex, closed set \mathcal{K} and m inequality constraints. The set \mathcal{K} is intended to represent the “easy” part of the constraints and is possibly defined by further convex inequalities. On the other hand, the explicit and “difficult” inequality constraints g_j s, which are possibly nonsmooth and nonconvex, need to be suitably approximated/simplified in order to build, starting from (P1) and at a generic feasible point \mathbf{y} , the better behaved subproblem (P1_y) below.

The objective function U is given by the sum $F + H$, with F smooth and H locally Lipschitz. Of course, this description of U is somewhat redundant since we could simply take $U = H$; however, it is useful to highlight, if present, a smooth component F , because, on this term, we can relax the assumptions that we need, instead, for the more general nonsmooth part H .

Taking into account that the general abstract problem (P) is actually represented by (P1), we are naturally led to consider the following approximation of (P1) at a feasible point \mathbf{y} :

$$\begin{aligned} & \underset{\mathbf{x}}{\text{minimize}} \quad \tilde{F}(\mathbf{x}; \mathbf{y}) + \tilde{H}(\mathbf{x}; \mathbf{y}) \\ & \text{s.t.} \quad \tilde{g}_j(\mathbf{x}; \mathbf{y}) \leq 0, \quad j = 1, \dots, m \\ & \quad \mathbf{x} \in \mathcal{K}, \end{aligned} \tag{P1_y}$$

where $\tilde{U} = \tilde{F} + \tilde{H}$ and $\tilde{\mathcal{X}}(\mathbf{y}) \triangleq \{\mathbf{x} \in \mathcal{K} : \tilde{g}_j(\mathbf{x}; \mathbf{y}) \leq 0, j = 1, \dots, m\}$.

In the next subsection we provide a detailed description of the assumptions that we make on problem (P1_y), while in the remaining part of the section we analyze the convergence properties of AF 1.

4.1 Assumptions

In order to strengthen the results of Proposition 1 and to take into account the specific structure of problems (P1) and (P1_y), below we state three assumptions about the approximations \tilde{F} , \tilde{H} , and \tilde{g}_j that complement Assumptions U and X. These conditions are somewhat technical, but in Sect. 5 we give many examples of approximations satisfying the assumptions, thus showing the breadth and wide practical applicability of our approach. The first two assumptions concern the objective function.

Assumption F $\tilde{F} : O_{\mathbf{x}} \times O_{\mathbf{y}} \rightarrow \mathbb{R}$ satisfies the following properties:

- (F1) $\tilde{F}(\bullet; \mathbf{y})$ is a finite strongly convex function on $O_{\mathbf{x}}$ for every $\mathbf{y} \in \mathcal{X}$ with modulus of strong convexity $c_{\tilde{F}} > 0$ independent of \mathbf{y} ;
- (F2) $\nabla_1 \tilde{F}(\bullet; \bullet)$ exists and is continuous on $O_{\mathbf{x}} \times O_{\mathbf{y}}$;
- (F3) $\nabla_1 \tilde{F}(\mathbf{y}; \mathbf{y}) = \nabla F(\mathbf{y})$, for every $\mathbf{y} \in \mathcal{X}$;

where $\nabla_1 \tilde{F}(\mathbf{u}; \mathbf{w})$ denotes the partial gradient of $\tilde{F}(\bullet; \mathbf{w})$ evaluated at \mathbf{u} .

Assumption H $\tilde{H} : O_{\mathbf{x}} \times O_{\mathbf{y}} \rightarrow \mathbb{R}$ satisfies the following properties:

- (H1) $\tilde{H}(\bullet; \mathbf{y})$ is a finite convex function on $O_{\mathbf{x}}$ for every $\mathbf{y} \in \mathcal{X}$;
- (H2) $\tilde{H}(\mathbf{y}; \mathbf{y}) = H(\mathbf{y})$, for every $\mathbf{y} \in \mathcal{X}$;
- (H3) $H(\mathbf{x}) \leq \tilde{H}(\mathbf{x}; \mathbf{y})$ for every $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{X}$;
- (H4) $\partial_1 \tilde{H}(\bullet; \bullet)$ is locally bounded relative to $\mathcal{X} \times \mathcal{X}$ at every (\mathbf{y}, \mathbf{y}) with $\mathbf{y} \in \mathcal{X}$;
- (H5) $\limsup_{(\mathbf{u}, \mathbf{w}) \xrightarrow{\mathcal{X} \times \mathcal{X}} (\mathbf{y}, \mathbf{y})} \partial_1 \tilde{H}(\mathbf{u}; \mathbf{w}) \subseteq \partial H(\mathbf{y})$ for every $\mathbf{y} \in \mathcal{X}$.

Assumption F on the smooth part of the objective function is just a verbatim repetition of Assumption U, see the discussion just before Lemma 1. Assumptions H1, H4 and H5 are again a plain restatement of Assumption U, the only difference being that condition H1 only requires convexity: indeed, in view of F1, it is enough to have \tilde{H} convex in order to make the overall approximation $\tilde{U} = \tilde{F} + \tilde{H}$ strongly convex for fixed \mathbf{y} . Conditions H2 and H3 are instead new. H2 is only made for convenience and can always be satisfied by a translation. The key assumption H3 requires $\tilde{H}(\mathbf{x}, \mathbf{y})$ to be an upper approximation of $H(\mathbf{x})$ making up, in some sense, for the lack of differentiability of H . Note that this condition is typical of majorization/minimizations schemes as, for example, [1, 7]. An important feature of AF 1 is that we do not require this majorization condition of \tilde{F} and therefore on \tilde{U} . This is very important from the practical point of view, since majorization schemes can be very slow, see for example Sect. 6, and also because, in general, it is easier to build an approximations which is not an upper estimate, see Sect. 5.

When considering the constraints, we assume that, at each given point $\mathbf{y} \in \mathcal{X}$ and for each function g_j , we can define a suitable approximation \tilde{g}_j that satisfies the following conditions.

Assumption G Each $\tilde{g}_j : O_{\mathbf{x}} \times O_{\mathbf{y}} \rightarrow \mathbb{R}$ satisfies the following properties:

- (G1) $\tilde{g}_j(\bullet; \mathbf{y})$ is a finite convex function on $O_{\mathbf{x}}$ for every $\mathbf{y} \in \mathcal{X}$;
- (G2) $\tilde{g}_j(\mathbf{y}; \mathbf{y}) = g_j(\mathbf{y})$, for every $\mathbf{y} \in \mathcal{X}$;
- (G3) $g_j(\mathbf{x}) \leq \tilde{g}_j(\mathbf{x}; \mathbf{y})$ for every $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{X}$;
- (G4) $\tilde{g}_j(\bullet; \bullet)$ is continuous relative to $\mathcal{X} \times \mathcal{X}$ at every (\mathbf{y}, \mathbf{y}) with $\mathbf{y} \in \mathcal{X}$;
- (G5) $\partial_1 \tilde{g}_j(\bullet; \bullet)$ is locally bounded relative to $\mathcal{X} \times \mathcal{X}$ at every (\mathbf{y}, \mathbf{y}) with $\mathbf{y} \in \mathcal{X}$;
- (G6) $\limsup_{(\mathbf{u}, \mathbf{w}) \xrightarrow{\mathcal{X} \times \mathcal{X}} (\mathbf{y}, \mathbf{y})} \partial_1 \tilde{g}_j(\mathbf{u}; \mathbf{w}) \subseteq \partial g_j(\mathbf{y})$ for every $\mathbf{y} \in \mathcal{X}$.

Conditions G1-G3 are just the translation of Assumption X in the setting of problems (P1) and (P1_y). While G4 is a fairly mild continuity assumption, the consistency condition G6, along with G5, is instrumental for building a bridge between the sets \tilde{N} and $N_{\mathcal{X}}$ at $\mathbf{y} \in \mathcal{X}$. In any case, G5 and G6 are identical to H4 and H5 and, if we are able to define suitable assumptions for the g_j s we can do the same for H .

For future reference, it may be useful to explicitly state the following fact that, in the light of the discussion so far, does not need any proof.

Proposition 2 Assumptions F, H, and G on problem (P1) imply that problem (P1_y) satisfies Assumptions U and X.

4.2 Main convergence result

In this subsection we study the main convergence properties of AF 1. We still have to specify how to choose the step-size γ^v at each iteration. We consider three options: (i) a sufficiently small fixed step-size; (ii) a diminishing step-size rule; (iii) a step-size that is chosen according to an Armijo-type line-search procedure. Each of these options has its pros and cons. Indeed, (i) is conceptually the simplest (provided that some Lipschitz and strong convexity constants are known), however it is expected to perform poorly in practice; (ii) is the easiest-to-implement rule and does not require any prior knowledge of any possibly-hard-to-know constant; (iii) is the most complex option and is expected to lead to a better practical behavior, since, in this case, the step-size is chosen taking into account the problem structure; rule (iii) is less suited to distributed environments, since performing a line-search is inherently a centralized task. The following theorem shows that, by choosing γ^v according to these three methods, we can ensure that (6) or (8) hold.

Proposition 3 *Given the nonconvex problem (P1) under F1-F3, H1-H3 and G1-G3, let $\{\mathbf{x}^v\}$ be the sequence generated by AF 1. The following hold:*

(i) *if the step-size γ^v and the error term ε^v are chosen so that*

$$0 < \inf_v \gamma^v \leq \sup_v \gamma^v \leq \gamma^{\max} \leq \min\{1, c_{\bar{F}}/L_{\nabla F}\} \tag{10}$$

$$\varepsilon^v \leq a_1 \min\{b^v / [\|\nabla F(\mathbf{x}^v) + \xi^v\|], c^v\}, \tag{11}$$

for some $\xi^v \in \partial_1 \tilde{H}(\mathbf{v}^v; \mathbf{x}^v)$ and for some non negative a_1, b^v and c^v with $\sum_v b^v < +\infty, \sum_v (c^v)^2 < +\infty$, then $\{\mathbf{x}^v\}$ is a bounded sequence such that

$$\lim_{v \rightarrow \infty} \|\hat{\mathbf{x}}(\mathbf{x}^v) - \mathbf{x}^v\| = 0; \tag{12}$$

(ii) *if the step-size γ^v and the error term ε^v are chosen so that*

$$\gamma^v \in (0, 1], \quad \lim_{v \rightarrow \infty} \gamma^v = 0, \quad \sum_v \gamma^v = +\infty, \quad \text{and} \quad \sum_v (\gamma^v)^2 < +\infty,$$

$$\varepsilon^v \leq \gamma^v a_1 \min\{a_2, 1/[\|\nabla F(\mathbf{x}^v) + \xi^v\|]\}, \tag{13}$$

for some $\xi^v \in \partial_1 \tilde{H}(\mathbf{v}^v; \mathbf{x}^v)$ and for some non negative a_1 and a_2 , then $\{\mathbf{x}^v\}$ is a bounded sequence such that

$$\liminf_{v \rightarrow \infty} \|\hat{\mathbf{x}}(\mathbf{x}^v) - \mathbf{x}^v\| = 0; \tag{14}$$

(iii) *if the step-size $\gamma^v \in (0, 1]$ is chosen by means of a backtracking rule so that $\gamma^v = 0.5^{i^v}$, where i^v is the smallest number in $0, 1, 2, \dots$ such that $\mathbf{x}(i^v) \triangleq \mathbf{x}^v + 0.5^{i^v}(\mathbf{v}^v - \mathbf{x}^v)$ satisfies the Armijo-type condition*

$$F(\mathbf{x}(i^v)) + H(\mathbf{x}(i^v)) \leq F(\mathbf{x}^v) + H(\mathbf{x}^v) + \alpha \gamma^v [\nabla F(\mathbf{x}^v)^T (\mathbf{v}^v - \mathbf{x}^v) + \tilde{H}(\mathbf{v}^v; \mathbf{x}^v) - \tilde{H}(\mathbf{x}^v; \mathbf{x}^v)], \tag{15}$$

for some $\alpha \in (0, 1)$, the error ε^v is such that $\lim_{v \rightarrow \infty} \varepsilon^v = 0$, and

$$\tilde{F}(\mathbf{v}^v; \mathbf{x}^v) + \tilde{H}(\mathbf{v}^v; \mathbf{x}^v) \leq \tilde{F}(\mathbf{x}^v; \mathbf{x}^v) + \tilde{H}(\mathbf{x}^v; \mathbf{x}^v), \quad (16)$$

then $\{\mathbf{x}^v\}$ is well-defined, bounded and condition (12) holds.

As stated above, according to AF 1, we consider the possibility of an inaccurate computation of $\hat{\mathbf{x}}(\mathbf{x}^v)$. The error must obey rather standard rules, with possibly the exception of (16), which, however, is a natural and easily enforced condition: indeed, (16) states that the approximation \mathbf{v}^v of the optimal solution $\hat{\mathbf{x}}(\mathbf{x}^v)$ of problem $(P1_{\mathbf{x}^v})$ has to at least improve on the current iteration \mathbf{x}^v . Rules (11) and (13) would require us to estimate $\|\mathbf{v}^v - \hat{\mathbf{x}}(\mathbf{x}^v)\|$. In principle, this is possible by resorting to appropriate error bounds, which are available for the strongly convex problem $(P1_{\mathbf{x}^v})$ [15]. However, from a practical point of view, taking into account that constants a_1, a_2 and sequences b^v and c^v are arbitrary, the meaning of (11) and (13) is simply that the error ε^v should go to zero, i.e. the same condition needed for the line-search option (iii).

To prove Proposition 3 we need two preliminary lemmas.

Lemma 2 Under Assumption F1-F3, H1 and G1-G3, for every \mathbf{x}^v ,

$$\nabla F(\mathbf{x}^v)^T (\hat{\mathbf{x}}(\mathbf{x}^v) - \mathbf{x}^v) + \tilde{H}(\hat{\mathbf{x}}(\mathbf{x}^v); \mathbf{x}^v) - \tilde{H}(\mathbf{x}^v; \mathbf{x}^v) \leq -c_{\tilde{F}} \|\hat{\mathbf{x}}(\mathbf{x}^v) - \mathbf{x}^v\|^2. \quad (17)$$

Proof For any given $\mathbf{x}^v \in \mathcal{X}$, by referring to problem $(P1_{\mathbf{y}})$, with $\mathbf{y} = \mathbf{x}^v$, thanks to F2, G1, and H1 we have

$$(\mathbf{z} - \hat{\mathbf{x}}(\mathbf{x}^v))^T \nabla_1 \tilde{F}(\hat{\mathbf{x}}(\mathbf{x}^v); \mathbf{x}^v) + \tilde{H}(\mathbf{z}; \mathbf{x}^v) - \tilde{H}(\hat{\mathbf{x}}(\mathbf{x}^v); \mathbf{x}^v) \geq 0, \quad \forall \mathbf{z} \in \tilde{\mathcal{X}}(\mathbf{x}^v). \quad (18)$$

By adding and subtracting $\nabla_1 \tilde{F}(\mathbf{x}^v; \mathbf{x}^v)$ in the LHS of (18), and choosing $\mathbf{z} = \mathbf{x}^v \in \tilde{\mathcal{X}}(\mathbf{x}^v)$, we get

$$0 \leq (\mathbf{x}^v - \hat{\mathbf{x}}(\mathbf{x}^v))^T (\nabla_1 \tilde{F}(\hat{\mathbf{x}}(\mathbf{x}^v); \mathbf{x}^v) - \nabla_1 \tilde{F}(\mathbf{x}^v; \mathbf{x}^v) + \nabla_1 \tilde{F}(\mathbf{x}^v; \mathbf{x}^v) + \tilde{H}(\mathbf{x}^v; \mathbf{x}^v) - \tilde{H}(\hat{\mathbf{x}}(\mathbf{x}^v); \mathbf{x}^v)),$$

which, using F1 and F3, gives (17). \square

Lemma 3 ([5, Lemma 3.4]) Let $\{a^v\}$, $\{b^v\}$, and $\{c^v\}$ be three sequences of numbers such that $\{b^v\} \geq 0$ for all v . Suppose that $a^{v+1} \leq a^v - b^v + c^v$, $\forall v = 0, 1, \dots$ and $\sum_{v=0}^{\infty} c^v < \infty$. Then, either $a^v \rightarrow -\infty$ or else $\{a^v\}$ converges to a finite value and $\sum_{v=0}^{\infty} b^v < \infty$.

Proof of Proposition 3 First of all, we note that, by the descent lemma [4, Proposition A.24] and step (S. 3) of AF 1, we get

$$F(\mathbf{x}^{v+1}) \leq F(\mathbf{x}^v) + \gamma^v \nabla F(\mathbf{x}^v)^T (\mathbf{v}^v - \mathbf{x}^v) + \frac{(\gamma^v)^2 L_{\nabla F}}{2} \|\mathbf{v}^v - \mathbf{x}^v\|^2. \quad (19)$$

Furthermore,

$$\begin{aligned}
 H(\mathbf{x}^{\nu+1}) &\stackrel{(a)}{\leq} \tilde{H}(\mathbf{x}^{\nu+1}; \mathbf{x}^\nu) \stackrel{(b)}{\leq} \tilde{H}(\mathbf{x}^\nu; \mathbf{x}^\nu) + \gamma^\nu (\tilde{H}(\mathbf{v}^\nu; \mathbf{x}^\nu) - \tilde{H}(\mathbf{x}^\nu; \mathbf{x}^\nu)) \\
 &= \tilde{H}(\mathbf{x}^\nu; \mathbf{x}^\nu) + \gamma^\nu \left(\tilde{H}(\mathbf{v}^\nu; \mathbf{x}^\nu) - \tilde{H}(\hat{\mathbf{x}}(\mathbf{x}^\nu); \mathbf{x}^\nu) + \tilde{H}(\hat{\mathbf{x}}(\mathbf{x}^\nu); \mathbf{x}^\nu) - \tilde{H}(\mathbf{x}^\nu; \mathbf{x}^\nu) \right) \\
 &\stackrel{(c)}{\leq} H(\mathbf{x}^\nu) + \gamma^\nu (\boldsymbol{\xi}^\nu)^T (\mathbf{v}^\nu - \hat{\mathbf{x}}(\mathbf{x}^\nu)) + \gamma^\nu (\tilde{H}(\hat{\mathbf{x}}(\mathbf{x}^\nu); \mathbf{x}^\nu) - \tilde{H}(\mathbf{x}^\nu; \mathbf{x}^\nu)), \tag{20}
 \end{aligned}$$

where the last relation holds for any $\boldsymbol{\xi}^\nu \in \partial_1 \tilde{H}(\mathbf{v}^\nu; \mathbf{x}^\nu)$. The inequality (a) follows from H3, while (b) and (c) are due to H2 and to the convexity of $\tilde{H}(\bullet; \mathbf{y})$. Using

$$\begin{aligned}
 \nabla F(\mathbf{x}^\nu)^T (\mathbf{v}^\nu - \hat{\mathbf{x}}(\mathbf{x}^\nu) + \hat{\mathbf{x}}(\mathbf{x}^\nu) - \mathbf{x}^\nu) &\leq \nabla F(\mathbf{x}^\nu)^T (\mathbf{v}^\nu - \hat{\mathbf{x}}(\mathbf{x}^\nu)) \\
 &\quad - c_{\tilde{F}} \|\hat{\mathbf{x}}(\mathbf{x}^\nu) - \mathbf{x}^\nu\|^2 + \tilde{H}(\mathbf{x}^\nu; \mathbf{x}^\nu) - \tilde{H}(\hat{\mathbf{x}}(\mathbf{x}^\nu); \mathbf{x}^\nu),
 \end{aligned}$$

which follows from Lemma 2, and combining (19) and (20), we have

$$\begin{aligned}
 F(\mathbf{x}^{\nu+1}) + H(\mathbf{x}^{\nu+1}) &\leq F(\mathbf{x}^\nu) + H(\mathbf{x}^\nu) - \gamma^\nu c_{\tilde{F}} \|\hat{\mathbf{x}}(\mathbf{x}^\nu) - \mathbf{x}^\nu\|^2 \\
 &\quad + \gamma^\nu \varepsilon^\nu \|\nabla F(\mathbf{x}^\nu) + \boldsymbol{\xi}^\nu\| + \frac{(\gamma^\nu)^2 L_{\nabla F}}{2} \|\mathbf{v}^\nu - \mathbf{x}^\nu\|^2. \tag{21}
 \end{aligned}$$

Since $\|\mathbf{v}^\nu - \mathbf{x}^\nu\|^2 \leq 2\|\hat{\mathbf{x}}(\mathbf{x}^\nu) - \mathbf{x}^\nu\|^2 + 2\|\mathbf{v}^\nu - \hat{\mathbf{x}}(\mathbf{x}^\nu)\|^2 \leq 2\|\hat{\mathbf{x}}(\mathbf{x}^\nu) - \mathbf{x}^\nu\|^2 + 2(\varepsilon^\nu)^2$, from (21) we obtain

$$F(\mathbf{x}^{\nu+1}) + H(\mathbf{x}^{\nu+1}) \leq F(\mathbf{x}^\nu) + H(\mathbf{x}^\nu) - \gamma^\nu (c_{\tilde{F}} - \gamma^\nu L_{\nabla F}) \|\hat{\mathbf{x}}(\mathbf{x}^\nu) - \mathbf{x}^\nu\|^2 + T^\nu, \tag{22}$$

where $T^\nu \triangleq \gamma^\nu \varepsilon^\nu [\|\nabla F(\mathbf{x}^\nu) + \boldsymbol{\xi}^\nu\|] + L_{\nabla F} (\gamma^\nu \varepsilon^\nu)^2$. Note that in cases (i) and (ii), under the assumptions of the theorem, we have $\sum_{\nu=0}^\infty T^\nu < \infty$.

We now prove the thesis for cases (i), (ii) and (iii) separately.

(i) In view of relation (22), since $F + H$ is coercive, Lemma 3 implies that the sequence $\{F(\mathbf{x}^\nu) + H(\mathbf{x}^\nu)\}$ is convergent and, thus, that $\{\mathbf{x}^\nu\}$ is bounded. Furthermore, again by Lemma 3, $\lim_{\nu \rightarrow \infty} \sum_{t=\bar{\nu}}^\nu \gamma^t \|\hat{\mathbf{x}}(\mathbf{x}^t) - \mathbf{x}^t\|^2 < +\infty$, which, taking into account that γ^ν is bounded away from zero, readily entails (12).

(ii) Since $\lim_{\nu \rightarrow \infty} \gamma^\nu = 0$, there exists a positive constant ω such that by (22) we have, for $\nu \geq \bar{\nu}$ sufficiently large,

$$F(\mathbf{x}^{\nu+1}) + H(\mathbf{x}^{\nu+1}) \leq F(\mathbf{x}^\nu) + H(\mathbf{x}^\nu) - \gamma^\nu \omega \|\hat{\mathbf{x}}(\mathbf{x}^\nu) - \mathbf{x}^\nu\|^2 + T^\nu. \tag{23}$$

Since $F + H$ is coercive, Lemma 3 implies that the sequence $\{F(\mathbf{x}^\nu) + H(\mathbf{x}^\nu)\}$ is convergent and, thus, that $\{\mathbf{x}^\nu\}$ is bounded. Furthermore, again by Lemma 3, we have $\lim_{\nu \rightarrow \infty} \sum_{t=\bar{\nu}}^\nu \gamma^t \|\hat{\mathbf{x}}(\mathbf{x}^t) - \mathbf{x}^t\|^2 < +\infty$, from which, taking into account $\sum_{\nu=0}^\infty \gamma^\nu = +\infty$, (14) follows.

(iii) By (19) and relation (b) in (20), we have

$$F(\mathbf{x}^{\nu+1}) + H(\mathbf{x}^{\nu+1}) - [F(\mathbf{x}^{\nu}) + H(\mathbf{x}^{\nu})] \leq \gamma^{\nu} \left[\nabla F(\mathbf{x}^{\nu})^T (\mathbf{v}^{\nu} - \mathbf{x}^{\nu}) + \tilde{H}(\mathbf{v}^{\nu}; \mathbf{x}^{\nu}) - \tilde{H}(\mathbf{x}^{\nu}; \mathbf{x}^{\nu}) \right] + \frac{(\gamma^{\nu})^2 L_{\nabla F}}{2} \|\mathbf{v}^{\nu} - \mathbf{x}^{\nu}\|^2. \quad (24)$$

We also observe that, by F1, F3 and (16),

$$\nabla F(\mathbf{x}^{\nu})^T (\mathbf{v}^{\nu} - \mathbf{x}^{\nu}) + \tilde{H}(\mathbf{v}^{\nu}; \mathbf{x}^{\nu}) - \tilde{H}(\mathbf{x}^{\nu}; \mathbf{x}^{\nu}) \leq -\frac{c_{\tilde{F}}}{2} \|\mathbf{v}^{\nu} - \mathbf{x}^{\nu}\|^2. \quad (25)$$

Now, we are able to prove that there exist suitable values of the step-size γ^{ν} such that (15) holds. Indeed, by (24), the sufficient decrease condition is satisfied if

$$\gamma^{\nu} [\nabla F(\mathbf{x}^{\nu})^T (\mathbf{v}^{\nu} - \mathbf{x}^{\nu}) + \tilde{H}(\mathbf{v}^{\nu}; \mathbf{x}^{\nu}) - \tilde{H}(\mathbf{x}^{\nu}; \mathbf{x}^{\nu})] + \frac{(\gamma^{\nu})^2 L_{\nabla F}}{2} \|\mathbf{v}^{\nu} - \mathbf{x}^{\nu}\|^2 \leq \alpha \gamma^{\nu} [\nabla F(\mathbf{x}^{\nu})^T (\mathbf{v}^{\nu} - \mathbf{x}^{\nu}) + \tilde{H}(\mathbf{v}^{\nu}; \mathbf{x}^{\nu}) - \tilde{H}(\mathbf{x}^{\nu}; \mathbf{x}^{\nu})], \quad (26)$$

and thus, thanks to (25) and easy reasonings, for any γ^{ν} such that $\gamma^{\nu} \leq \min\{1, (1 - \alpha) \frac{c_{\tilde{F}}}{L_{\nabla F}}\}$. Therefore, we can always find a finite i^{ν} for which the Armijo condition (15) holds. By the same token, the step-size γ^{ν} is bounded away from zero and, more precisely, such that, for all ν ,

$$\gamma^{\nu} \geq \min \left\{ 1, \frac{1 - \alpha}{2} \frac{c_{\tilde{F}}}{L_{\nabla F}} \right\}. \quad (27)$$

Besides, (15) and (25) also imply

$$F(\mathbf{x}^{\nu+1}) + H(\mathbf{x}^{\nu+1}) - [F(\mathbf{x}^{\nu}) + H(\mathbf{x}^{\nu})] \leq -\alpha \gamma^{\nu} \frac{c_{\tilde{F}}}{2} \|\mathbf{v}^{\nu} - \mathbf{x}^{\nu}\|^2$$

and in turn, thanks to (27) and the coercivity of $F + H$, we deduce that $\{F(\mathbf{x}^{\nu}) + H(\mathbf{x}^{\nu})\}$ is convergent, $\{\mathbf{x}^{\nu}\}$ is bounded, and $\lim_{\nu \rightarrow \infty} \|\mathbf{v}^{\nu} - \mathbf{x}^{\nu}\| = 0$. By observing that $\|\hat{\mathbf{x}}(\mathbf{x}^{\nu}) - \mathbf{x}^{\nu}\| \leq \|\hat{\mathbf{x}}(\mathbf{x}^{\nu}) - \mathbf{v}^{\nu}\| + \|\mathbf{v}^{\nu} - \mathbf{x}^{\nu}\|$, we get (12). \square

Point (iii) in Proposition 3 and, in particular, (27) show that if we choose $c_{\tilde{F}}$ large enough (see the next section to appreciate how this is easily possible) we could always take $\gamma^{\nu} = 1$ and get (12). However, if the original objective function $U = F + H$ is itself strongly convex, then, without any additional requirement, a unit step-size leads to condition (12). The following proposition is a generalization of a similar result in [3] (see the first part of the proof of Proposition 3.2 therein) valid for smooth problems only and only when the exact solution $\hat{\mathbf{x}}(\mathbf{x}^{\nu})$ is used at each iteration (i.e. when $\varepsilon^{\nu} = 0$ at each iteration).

Proposition 4 *Given the nonconvex problem (P1) under G1-G3, suppose that $U = F + H$ is strongly convex with constant c_U and that we set $\tilde{U}(\mathbf{x}; \mathbf{y}) = U(\mathbf{x})$. Letting $\{\mathbf{x}^{\nu}\}$ be the sequence generated by AF 1,*

(iv) if the step-size γ^v and the error term ε^v are chosen so that

$$\gamma^v = 1, \quad \lim_{v \rightarrow \infty} \varepsilon^v = 0, \quad U(\mathbf{v}^v) \leq U(\mathbf{x}^v),$$

then $\{\mathbf{x}^v\}$ is a bounded sequence such that (12) holds.

Proof Preliminarily, we observe that, by choosing $\gamma^v = 1$, we have $\mathbf{x}^{v+1} = \mathbf{v}^v$. Then, by $U(\mathbf{v}^v) \leq U(\mathbf{x}^v)$, since $U = F + H$ is coercive, the sequence $\{U(\mathbf{x}^v)\}$ converges and $\{\mathbf{x}^v\}$ is bounded. Thus, by $\|\mathbf{v}^v - \hat{\mathbf{x}}(\mathbf{x}^v)\| \leq \varepsilon^v$, we have

$$\lim_{v \rightarrow \infty} \|U(\hat{\mathbf{x}}(\mathbf{x}^v)) - U(\mathbf{x}^v)\| = 0. \tag{28}$$

On the other hand, by the strong convexity of U , for every $\xi^v \in \partial U(\hat{\mathbf{x}}(\mathbf{x}^v))$,

$$U(\mathbf{x}^v) - U(\hat{\mathbf{x}}(\mathbf{x}^v)) \geq (\xi^v)^T (\mathbf{x}^v - \hat{\mathbf{x}}(\mathbf{x}^v)) + \frac{c_U}{2} \|\hat{\mathbf{x}}(\mathbf{x}^v) - \mathbf{x}^v\|^2. \tag{29}$$

Furthermore, since $\hat{\mathbf{x}}(\mathbf{x}^v)$ is the solution of problem (P1_{x^v}) and $\mathbf{x}^v \in \tilde{\mathcal{X}}(\mathbf{x}^v)$,

$$(\hat{\xi}^v)^T (\mathbf{x}^v - \hat{\mathbf{x}}(\mathbf{x}^v)) \geq 0 \tag{30}$$

for some $\hat{\xi}^v \in \partial U(\hat{\mathbf{x}}(\mathbf{x}^v))$. Thus, by (29) and (30), we get $U(\hat{\mathbf{x}}(\mathbf{x}^v)) - U(\mathbf{x}^v) \leq -\frac{c_U}{2} \|\hat{\mathbf{x}}(\mathbf{x}^v) - \mathbf{x}^v\|^2$ for every v and, in turn, by (28), $\lim_{v \rightarrow \infty} \|\hat{\mathbf{x}}(\mathbf{x}^v) - \mathbf{x}^v\| = 0$. \square

Combining Propositions 1, 3, and 4, we easily get convergence to points satisfying (7). By showing that, under the MFCQ, this condition is actually equivalent to the stationarity condition (5), the following theorem provides the main convergence properties of AF 1.

Theorem 1 *Let $\{\mathbf{x}^v\}$ be the sequence generated by AF 1. Under Assumptions F, H, and G, $\{\mathbf{x}^v\}$ is bounded and*

- (a) *if the step-size rules in (i) or (iii) of Proposition 3, or if U is strongly convex and the step-size in (iv) of Proposition 4 is adopted, every limit point of $\{\mathbf{x}^v\}$ satisfies (7);*
- (b) *if the step-size rule in (ii) of Proposition 3 is employed, at least one limit point of $\{\mathbf{x}^v\}$ satisfies (7).*

Furthermore, any limit point of $\{\mathbf{x}^v\}$ which satisfies (7) and the MFCQ with respect to the set \mathcal{X} , is also stationary for problem (P1) according to (5).

Proof In view of the previous comments and taking into account Proposition 2, (a) and (b) need no further proof. Then, we show that a limit point $\bar{\mathbf{x}}$ that satisfies (7) and the MFCQ also satisfies (5). It is enough to prove that if (3) holds at $\bar{\mathbf{x}} \in \mathcal{X}$, then

$$\tilde{N}(\bar{\mathbf{x}}) \subseteq \bigcup \left\{ \sum_{j=1}^m \mu_j \partial g_j(\bar{\mathbf{x}}) \mid \boldsymbol{\mu} \in N_{\mathbb{R}^m}(\mathbf{g}(\bar{\mathbf{x}})) \right\} + N_{\mathcal{K}}(\bar{\mathbf{x}}) \triangleq S(\bar{\mathbf{x}}). \tag{31}$$

We claim that (3) must hold for $\tilde{\mathcal{X}}(\mathbf{y})$ for all $\mathbf{y} \in \mathcal{X}$ and $\mathbf{u} \in \tilde{\mathcal{X}}(\mathbf{y})$ sufficiently close to $\bar{\mathbf{x}}$. If this were not true, sequences $\mathbf{y}^k \xrightarrow{\mathcal{X}} \bar{\mathbf{x}}$ and $\mathbf{u}^k \rightarrow \bar{\mathbf{x}}$ with $\mathbf{u}^k \in \tilde{\mathcal{X}}(\mathbf{y}^k)$ would exist such that

$$\mathbf{0} \in \sum_{j=1}^m \mu_j^k \partial_1 \tilde{g}_j(\mathbf{u}^k; \mathbf{y}^k) + N_{\mathcal{K}}(\mathbf{u}^k), \quad \mu^k \in N_{\mathbb{R}^m}(\tilde{\mathbf{g}}(\mathbf{u}^k; \mathbf{y}^k)),$$

with $\|\mu^k\| = 1$. Note that we use the iteration index k instead of ν not to create confusion with the algorithm iterates. Thus, we must have, for every k , $-\sum_{j=1}^m \mu_j^k \xi_j^k \in N_{\mathcal{K}}(\mathbf{u}^k)$, for some $\xi_j^k \in \partial_1 \tilde{g}_j(\mathbf{u}^k; \mathbf{y}^k)$ and $\mu^k \in N_{\mathbb{R}^m}(\tilde{\mathbf{g}}(\mathbf{u}^k; \mathbf{y}^k))$. Taking the limit and renumbering if necessary, by G5 and the outer semicontinuity (relative to \mathcal{K}) of the mapping $N_{\mathcal{K}}(\bullet)$ (see e.g. [39, Proposition 6.6]), we get $-\sum_{j=1}^m \bar{\mu}_j \bar{\xi}_j \in N_{\mathcal{K}}(\bar{\mathbf{x}})$, where $\bar{\xi}_j \in \partial g_j(\bar{\mathbf{x}})$ in view of G6, and $\bar{\mu} \in N_{\mathbb{R}^m}(\mathbf{g}(\bar{\mathbf{x}}))$ with $\|\bar{\mu}\| = 1$, by G4, G2 and the outer semicontinuity (relative to \mathbb{R}^m) of the mapping $N_{\mathbb{R}^m}(\bullet)$. This contradicts the assumed MFCQ at $\bar{\mathbf{x}} \in \mathcal{X}$. Therefore, for all $\mathbf{y} \in \mathcal{X}$ and $\mathbf{u} \in \tilde{\mathcal{X}}(\mathbf{y})$ sufficiently close to $\bar{\mathbf{x}}$,

$$\mathbf{0} \in \sum_{j=1}^m \mu_j \partial_1 \tilde{g}_j(\mathbf{u}; \mathbf{y}) + N_{\mathcal{K}}(\mathbf{u}), \quad \mu \in N_{\mathbb{R}^m}(\tilde{\mathbf{g}}(\mathbf{u}; \mathbf{y})) \Rightarrow \mu = \mathbf{0}. \tag{32}$$

In order to prove relation (31), take any $\bar{\eta}$ in $\tilde{N}(\bar{\mathbf{x}})$; we show that $\bar{\eta}$ belongs to $S(\bar{\mathbf{x}})$. By definition of $\tilde{N}(\bar{\mathbf{x}})$, there exist sequences $\mathcal{X} \ni \mathbf{y}^k \rightarrow \bar{\mathbf{x}}$ and $\mathbf{u}^k \rightarrow \bar{\mathbf{x}}$ such that $N_{\tilde{\mathcal{X}}(\mathbf{y}^k)}(\mathbf{u}^k) \ni \eta^k \rightarrow \bar{\eta}$. We also have

$$N_{\tilde{\mathcal{X}}(\mathbf{y}^k)}(\mathbf{u}^k) = \bigcup \left\{ \sum_{j=1}^m \mu_j^k \partial_1 \tilde{g}_j(\mathbf{u}^k; \mathbf{y}^k) \mid \mu^k \in N_{\mathbb{R}^m}(\tilde{\mathbf{g}}(\mathbf{u}^k; \mathbf{y}^k)) \right\} + N_{\mathcal{K}}(\mathbf{u}^k), \tag{33}$$

where (33) comes from G1, the constraint qualification (32), and rather standard facts in convex analysis (alternatively, see [39, Corollary 10.50] and [39, Theorem 9.13 along with Corollary 10.9 and Theorem 6.42]). Therefore

$$\eta^k = \sum_{j=1}^m \mu_j^k \xi_j^k + \vartheta^k, \tag{34}$$

for some $\xi_j^k \in \partial_1 \tilde{g}_j(\mathbf{u}^k; \mathbf{y}^k)$, $\mu^k \in N_{\mathbb{R}^m}(\tilde{\mathbf{g}}(\mathbf{u}^k; \mathbf{y}^k))$ and $\vartheta^k \in N_{\mathcal{K}}(\mathbf{u}^k)$. By passing to subsequences, we can distinguish two cases: either $(\mu^k, \vartheta^k) \rightarrow (\bar{\mu}, \bar{\vartheta})$ or $\lambda^k (\mu^k, \vartheta^k) \rightarrow (\bar{\mu}, \bar{\vartheta}) \neq (\mathbf{0}, \mathbf{0})$ for some sequence $\lambda^k \downarrow 0$. The latter case cannot actually occur; in fact, one would obtain $\lambda^k \eta^k = \sum_{j=1}^m \lambda^k \mu_j^k \xi_j^k + \lambda^k \vartheta^k$, which would entail $\mathbf{0} = \sum_{j=1}^m \bar{\mu}_j \bar{\xi}_j + \bar{\vartheta}$, with $\bar{\xi}_j \in \partial g_j(\bar{\mathbf{x}})$ (by G6), $\bar{\mu} \in N_{\mathbb{R}^m}(\mathbf{g}(\bar{\mathbf{x}}))$ (by G4, G2, and the outer semicontinuity relative to \mathbb{R}^m of $N_{\mathbb{R}^m}(\bullet)$), and $\bar{\vartheta} \in N_{\mathcal{K}}(\bar{\mathbf{x}})$ (by the outer semicontinuity relative to \mathcal{K} of $N_{\mathcal{K}}(\bullet)$). This contradicts the assumed MFCQ (3) at $\bar{\mathbf{x}}$. Hence, we are left with $(\mu^k, \vartheta^k) \rightarrow (\bar{\mu}, \bar{\vartheta})$. Recalling G5 and taking the limit

in (34), we have, without loss of generality, $\bar{\eta} = \sum_{j=1}^m \bar{\mu}_j \bar{\xi}_j + \bar{\vartheta}$, where, as before, $\bar{\xi}_j \in \partial g_j(\bar{\mathbf{x}})$, $\bar{\mu} \in N_{\mathbb{R}^m}(\mathbf{g}(\bar{\mathbf{x}}))$, and $\bar{\vartheta} \in N_{\mathcal{K}}(\bar{\mathbf{x}})$, thus concluding the proof. \square

4.3 From lim inf to lim when a diminishing step-size is adopted

When a diminishing step-size (case (ii) in Proposition 3) is employed, in view of Theorem 1, it appears that one can establish weaker convergence properties, compared to the result for cases (i), (iii), and (iv) in Proposition 4: indeed we were only able to show (14) instead of (12). Taking into account that the diminishing step-size rule is arguably the best option in case of nonconvex parallel/distributed implementations, see subsection 5.3, it is desirable to inquire whether we can improve the results for case (ii). It turns out that the Hölder behavior of the best-response mapping $\hat{\mathbf{x}}(\bullet)$ plays a key role.

Proposition 5 *Under the assumptions of Proposition 3, if, in addition,*

$$\|\hat{\mathbf{x}}(\mathbf{y}) - \hat{\mathbf{x}}(\mathbf{z})\| \leq \theta \|\mathbf{y} - \mathbf{z}\|^\beta, \tag{35}$$

for every $\mathbf{y}, \mathbf{z} \in \mathcal{X}$ and for positive scalars β and θ , then, also in case (ii) of Proposition 3, (12) holds, i.e. $\lim_{\nu \rightarrow \infty} \|\hat{\mathbf{x}}(\mathbf{x}^\nu) - \mathbf{x}^\nu\| = 0$.

Proof Denoting $\Delta \hat{\mathbf{x}}(\mathbf{x}^\nu) \triangleq \hat{\mathbf{x}}(\mathbf{x}^\nu) - \mathbf{x}^\nu$, by the statement (ii) of Proposition 3, we have $\liminf_{\nu \rightarrow \infty} \|\Delta \hat{\mathbf{x}}(\mathbf{x}^\nu)\| = 0$. Suppose by contradiction that $\limsup_{\nu \rightarrow \infty} \|\Delta \hat{\mathbf{x}}(\mathbf{x}^\nu)\| > 0$. Then, there exists $\delta > 0$ such that $\|\Delta \hat{\mathbf{x}}(\mathbf{x}^\nu)\| > \delta$ and $\|\Delta \hat{\mathbf{x}}(\mathbf{x}^\nu)\| < \delta/2$ for infinitely many ν 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:

$$\|\Delta \hat{\mathbf{x}}(\mathbf{x}^\nu)\| < \delta/2, \quad \|\Delta \hat{\mathbf{x}}(\mathbf{x}^{i_\nu})\| > \delta \tag{36}$$

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

$$\delta/2 \leq \|\Delta \hat{\mathbf{x}}(\mathbf{x}^j)\| \leq \delta, \quad \nu < j < i_\nu. \tag{37}$$

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

$$\begin{aligned} \delta/2 &< \|\Delta \hat{\mathbf{x}}(\mathbf{x}^{i_\nu})\| - \|\Delta \hat{\mathbf{x}}(\mathbf{x}^\nu)\| \leq \|\hat{\mathbf{x}}(\mathbf{x}^{i_\nu}) - \hat{\mathbf{x}}(\mathbf{x}^\nu)\| + \|\mathbf{x}^{i_\nu} - \mathbf{x}^\nu\| \\ &\stackrel{(a)}{\leq} \|\mathbf{x}^{i_\nu} - \mathbf{x}^\nu\| + \theta \|\mathbf{x}^{i_\nu} - \mathbf{x}^\nu\|^\beta \\ &\stackrel{(b)}{\leq} \sum_{t=\nu}^{i_\nu-1} \gamma^t (\|\Delta \hat{\mathbf{x}}(\mathbf{x}^t)\| + \|\mathbf{v}^t - \hat{\mathbf{x}}(\mathbf{x}^t)\|) \\ &\quad + \theta \left[\sum_{t=\nu}^{i_\nu-1} \gamma^t (\|\Delta \hat{\mathbf{x}}(\mathbf{x}^t)\| + \|\mathbf{v}^t - \hat{\mathbf{x}}(\mathbf{x}^t)\|) \right]^\beta \\ &\stackrel{(c)}{\leq} (\delta + \varepsilon^{\max}) \sum_{t=\nu}^{i_\nu-1} \gamma^t + \theta (\delta + \varepsilon^{\max})^\beta \left(\sum_{t=\nu}^{i_\nu-1} \gamma^t \right)^\beta, \end{aligned} \tag{38}$$

where (a) is due to (35), (b) comes from the triangle inequality and the updating rule of the algorithm and in (c) we used (37), $\|\mathbf{v}^t - \hat{\mathbf{x}}(\mathbf{x}^t)\| \leq \varepsilon^t$ and set $\varepsilon^{\max} \triangleq \max_{\nu} \varepsilon^{\nu}$, which is well-defined because by (13), $\varepsilon^{\nu} \rightarrow 0$. By (38) we have

$$\liminf_{\nu \rightarrow \infty} \left[(\delta + \varepsilon^{\max}) \sum_{t=\nu}^{i_{\nu}-1} \gamma^t + \theta (\delta + \varepsilon^{\max})^{\beta} \left(\sum_{t=\nu}^{i_{\nu}-1} \gamma^t \right)^{\beta} \right] > 0. \tag{39}$$

We prove next that (39) is in contradiction with the convergence of $\{F(\mathbf{x}^{\nu}) + H(\mathbf{x}^{\nu})\}$, which is established in Proposition 3. To this end, we first show that $\|\Delta\hat{\mathbf{x}}(\mathbf{x}^{\nu})\| \geq \delta/4$, for sufficiently large $\nu \in \mathcal{N}$. Reasoning as in (38), we have

$$\begin{aligned} \|\Delta\hat{\mathbf{x}}(\mathbf{x}^{\nu+1})\| - \|\Delta\hat{\mathbf{x}}(\mathbf{x}^{\nu})\| &\leq \|\mathbf{x}^{\nu+1} - \mathbf{x}^{\nu}\| + \theta \|\mathbf{x}^{\nu+1} - \mathbf{x}^{\nu}\|^{\beta} \\ &\leq \gamma^{\nu} (\|\Delta\hat{\mathbf{x}}(\mathbf{x}^{\nu})\| + \varepsilon^{\max}) + \theta (\gamma^{\nu})^{\beta} (\|\Delta\hat{\mathbf{x}}(\mathbf{x}^{\nu})\| + \varepsilon^{\max})^{\beta}, \end{aligned}$$

for any given ν . For $\nu \in \mathcal{N}$ large enough so that $\gamma^{\nu}(\delta/4 + \varepsilon^{\max}) + \theta(\gamma^{\nu})^{\beta}(\delta/4 + \varepsilon^{\max})^{\beta} < \delta/4$, suppose by contradiction that $\|\Delta\hat{\mathbf{x}}(\mathbf{x}^{\nu})\| < \delta/4$; this would give $\|\Delta\hat{\mathbf{x}}(\mathbf{x}^{\nu+1})\| < \delta/2$ and, thus, condition (37) (or (36)) would be violated. Then, it must be $\|\Delta\hat{\mathbf{x}}(\mathbf{x}^{\nu})\| \geq \delta/4$. From this, and using (23) we have, for sufficiently large $\nu \in \mathcal{N}$,

$$\begin{aligned} F(\mathbf{x}^{i_{\nu}}) + H(\mathbf{x}^{i_{\nu}}) &\leq F(\mathbf{x}^{\nu}) + H(\mathbf{x}^{\nu}) - \omega \sum_{t=\nu}^{i_{\nu}-1} \gamma^t \|\Delta\hat{\mathbf{x}}(\mathbf{x}^t)\|^2 + \sum_{t=\nu}^{i_{\nu}-1} T^t \\ &\leq F(\mathbf{x}^{\nu}) + H(\mathbf{x}^{\nu}) - \omega \frac{\delta^2}{16} \sum_{t=\nu}^{i_{\nu}-1} \gamma^t + \sum_{t=\nu}^{i_{\nu}-1} T^t. \end{aligned} \tag{40}$$

Since $\sum_{\nu=0}^{\infty} T^{\nu} < \infty$ and $T^{\nu} \geq 0$ for every ν , the series $\sum_{\nu=0}^{\infty} T^{\nu}$ is convergent, so that the Cauchy convergence criterion implies $\sum_{t=\nu}^{i_{\nu}-1} T^t \rightarrow 0$. Therefore, recalling that $\{F(\mathbf{x}^{\nu}) + H(\mathbf{x}^{\nu})\}$ converges, renumbering if necessary, (40) implies $\sum_{t=\nu}^{i_{\nu}-1} \gamma^t \rightarrow 0$, in contradiction with (39). \square

In view of Proposition 5, we aim at deriving suitable sufficient conditions for (35) to hold. Of course, one can draw on the vast literature on stability in optimization (see e.g. [9, 28], the recent [32, 33] and references therein). In particular, in [33] a second-order subdifferential characterization of Hölderian full stability (via the regular coderivative of the subdifferential) is established. However, these conditions are hard to verify in practice and, hence, below we prove the desired Hölderian behavior (35) under assumptions that, while not minimal, are easy to check and, in our context, not difficult to satisfy. We therefore require the following additional conditions; in the next section we show that they are easily met in many practical cases.

- Assumption B** (F4) $\nabla_1 \tilde{F}(\mathbf{x}; \bullet)$ is Lipschitz continuous on $O_{\mathbf{y}}$ for every $\mathbf{x} \in \mathcal{X}$ with modulus of Lipschitz continuity independent of \mathbf{x} ;
 (F5) $\nabla_1 \tilde{F}(\bullet; \mathbf{y})$ is Lipschitz continuous on $O_{\mathbf{x}}$ for every $\mathbf{y} \in \mathcal{X}$ with modulus of Lipschitz continuity independent of \mathbf{y} ;

(G7) each $\tilde{g}_j(\bullet; \bullet)$ is locally Lipschitz continuous on $O_x \times O_y$.

The following proposition, which builds on the results in [51], allows us to provide suitable sufficient conditions that guarantee the Hölder behavior (35) of $\hat{\mathbf{x}}(\bullet)$ in the case of a smooth objective function.

Proposition 6 *Assume that F1, F2, F4, F5, G1 and G7 hold and that $H \equiv 0$. Suppose further that \mathcal{X} is compact and the MFCQ holds at $\hat{\mathbf{x}}(\bar{\mathbf{x}}) \in \tilde{\mathcal{X}}(\bar{\mathbf{x}})$ for every $\bar{\mathbf{x}} \in \mathcal{X}$. Then, there exists $\theta > 0$ such that for every $\mathbf{y}, \mathbf{z} \in \mathcal{X}$:*

$$\|\hat{\mathbf{x}}(\mathbf{y}) - \hat{\mathbf{x}}(\mathbf{z})\| \leq \theta \|\mathbf{y} - \mathbf{z}\|^{\frac{1}{2}}. \tag{41}$$

Proof Observing that, in view of G1 and G7, by [12, Propositions 2.3.16 and 2.5.3], we have $\partial_1 g_j(\hat{\mathbf{x}}(\bar{\mathbf{x}}); \bar{\mathbf{x}}) = \partial_1 g_j(\hat{\mathbf{x}}(\bar{\mathbf{x}}); \bar{\mathbf{x}}) = \pi_{\bar{\mathbf{x}}} \partial g_j(\hat{\mathbf{x}}(\bar{\mathbf{x}}); \bar{\mathbf{x}})$, the MFCQ at $\hat{\mathbf{x}}(\bar{\mathbf{x}}) \in \tilde{\mathcal{X}}(\bar{\mathbf{x}})$, for every $\bar{\mathbf{x}} \in \mathcal{X}$, implies by [51, Lemma 3.1] that the set-valued mapping $\tilde{\mathcal{X}}$ has the Aubin property relative to \mathcal{X} at $\bar{\mathbf{x}}$ for $\hat{\mathbf{x}}(\bar{\mathbf{x}})$ for every $\bar{\mathbf{x}} \in \mathcal{X}$. Therefore, in view of [51, Theorem 2.1], for every $\bar{\mathbf{x}} \in \mathcal{X}$, there exist $\eta_{\bar{\mathbf{x}}} > 0, \theta_{\bar{\mathbf{x}}} > 0$ and a neighborhood $\mathcal{V}_{\bar{\mathbf{x}}}$ of $\bar{\mathbf{x}}$ such that, for every $\mathbf{y}, \mathbf{z} \in \mathcal{V}_{\bar{\mathbf{x}}} \cap \mathcal{X}$

$$\|\hat{\mathbf{x}}(\mathbf{y}) - \hat{\mathbf{x}}(\mathbf{z})\| \leq \eta_{\bar{\mathbf{x}}} \|\mathbf{y} - \mathbf{z}\| + \theta_{\bar{\mathbf{x}}} \|\mathbf{y} - \mathbf{z}\|^{\frac{1}{2}}.$$

By the previous relation and the compactness of set \mathcal{X} , (41) holds. □

5 Approximations and distributed implementation

In this section we show the wide applicability of our approach by defining functions \tilde{F}, \tilde{H} , and \tilde{g} satisfying Assumptions F, H, and G in an array of cases of practical interest. We also briefly discuss how a suitable selection of \tilde{F}, \tilde{H} , and \tilde{g} can lead to parallel and distributed versions of our scheme. We remark that, given a specific application, it is often possible to find *ad hoc* approximations that exploit the structure of the problem at hand; Sect. 6 provides examples in this sense. While the functions we consider are “common” and reflect what can be encountered in applications, the fact that these approximations, even if natural-looking, satisfy our assumptions is by no means obvious; on the contrary, it is evidence of the heed given to the choice of Assumptions F, H, and G. It is precisely because of these technical, but carefully selected assumptions that we are able to deal, for the first time in INCA-type methods, with non subdifferentially regular nonsmooth functions as those considered in Examples 2, 3, 4 and 7 in Sect. 5.2 below, which, for example, cannot be dealt with the method proposed in [7]. We believe that our treatment of problems involving this kind of functions probably provides in this context the most efficient solution method currently available.

5.1 Approximating F

Finding suitable approximations for F is in general easy, since F is differentiable and we do not need to impose any majorization property. Below we list a few possible choices for \tilde{F} . The verification of the validity of Assumption F, and F4 and F5 in

Assumption B is straightforward and, therefore, omitted. For the sake of simplicity, in all this sections we do not detail every time the sets $O_{\mathbf{x}}$ and $O_{\mathbf{y}}$, as they are easily determinable by the reader.

1. If F is convex, as assumed in [3,30], we can set

$$\tilde{F}(\mathbf{x}; \mathbf{y}) \triangleq F(\mathbf{x}) + \frac{\tau}{2} \|\mathbf{x} - \mathbf{y}\|^2,$$

where τ is a positive constant that can be taken to be 0 in case F is already strongly convex. This \tilde{F} is just a regularization of the original F .

2. In general, one can always resort to the first order approximation of F ,

$$\tilde{F}(\mathbf{x}; \mathbf{y}) \triangleq \nabla F(\mathbf{y})^T (\mathbf{x} - \mathbf{y}) + \frac{\tau}{2} \|\mathbf{x} - \mathbf{y}\|^2,$$

where τ is a positive constant. Classical (proximal) gradient methods [6] use this approximation but they are not applicable to (P1) because of the nonconvexity of the feasible set. Note that if $F \equiv 0$, the above approximation boils down to $\tilde{F}(\mathbf{x}; \mathbf{y}) \triangleq \frac{\tau}{2} \|\mathbf{x} - \mathbf{y}\|^2$. This means that even if $U = H$ we still have a “contribution” from F . Approximating $F \equiv 0$ with a proximal term is simply a way of requiring the overall $U \triangleq \tilde{F} + \tilde{H}$ to be strongly convex without making this assumption also on \tilde{H} .

3. In some cases, a certain degree of convexity may be present in F and it is useful to preserve this feature as much as possible. Suppose that the vector of variables \mathbf{x} is partitioned in blocks $\mathbf{x} = (\mathbf{x}_i)_{i=1}^I$ and the function F is convex in each \mathbf{x}_i but not jointly in \mathbf{x} . A natural approximation is then

$$\tilde{F}(\mathbf{x}; \mathbf{y}) = \sum_{i=1}^I \tilde{F}_i(\mathbf{x}_i; \mathbf{y}), \quad \text{with} \quad \tilde{F}_i(\mathbf{x}_i; \mathbf{y}) \triangleq F(\mathbf{x}_i, \mathbf{y}_{-i}) + \frac{\tau_i}{2} \|\mathbf{x}_i - \mathbf{y}_i\|^2, \quad (42)$$

where $\mathbf{y} \triangleq (\mathbf{y}_i)_{i=1}^I$, $\mathbf{y}_{-i} \triangleq (\mathbf{y}_j)_{j \neq i}$ and $\tau_i > 0$. If $F(\bullet, \mathbf{y}_{-i})$ is already uniformly strongly convex for all feasible \mathbf{y}_{-i} , the quadratic term in (42) can be dropped.

4. In some practical problems, for example in multi-agent scenarios, the objective function F is given by the sum of the utilities $f_i(\mathbf{x}_1, \dots, \mathbf{x}_I)$ of I agents, each of them controlling the variable block \mathbf{x}_i : thus, $F(\mathbf{x}) \triangleq \sum_{i=1}^I f_i(\mathbf{x}_1, \dots, \mathbf{x}_I)$. Typically, it may happen that some of the f_i s are convex in some agents' variables: our purpose, following [16,45], is to employ an approximation such that, for each agent i , the convex part of F w.r.t. \mathbf{x}_i is kept unaltered, while the nonconvex part is linearized. Let

$$\mathcal{S}_i \triangleq \{j = 1, \dots, I : f_j(\bullet, \mathbf{x}_{-i}) \text{ is convex, } \forall (\mathbf{x}_i, \mathbf{x}_{-i}) \in \mathcal{K}\}$$

be the set of indices of all the functions $f_j(\mathbf{x}_i, \mathbf{x}_{-i})$ that are convex in \mathbf{x}_i , for any feasible \mathbf{x}_{-i} , and \mathcal{C}_i be any subset of \mathcal{S}_i ; one can set

$$\tilde{F}(\mathbf{x}; \mathbf{y}) = \sum_{i=1}^I \tilde{F}_{C_i}(\mathbf{x}_i; \mathbf{y}),$$

with

$$\tilde{F}_{C_i}(\mathbf{x}_i; \mathbf{y}) \triangleq \sum_{j \in C_i} f_j(\mathbf{x}_i, \mathbf{y}_{-i}) + \sum_{k \notin C_i} [f_k(\mathbf{y}) + \nabla_i f_k(\mathbf{y})^T (\mathbf{x}_i - \mathbf{y}_i)] + \frac{\tau_i}{2} \|\mathbf{x}_i - \mathbf{y}_i\|^2,$$

where τ_i is a positive constant.

5. Assume that F is a saddle function, i.e., F depends on two groups of variables, \mathbf{x}_1 and \mathbf{x}_2 , with $F(\bullet, \mathbf{x}_2)$ convex for every fixed \mathbf{x}_2 and $F(\mathbf{x}_1, \bullet)$ concave for every fixed \mathbf{x}_1 . It is easy to check that

$$\begin{aligned} \tilde{F}(\mathbf{x}_1, \mathbf{x}_2; \mathbf{y}_1, \mathbf{y}_2) &\triangleq F(\mathbf{x}_1, \mathbf{y}_2) + \nabla_2 F(\mathbf{y}_1, \mathbf{y}_2)^T (\mathbf{x}_2 - \mathbf{y}_2) \\ &\quad + \frac{\tau}{2} \|(\mathbf{x}_1, \mathbf{x}_2) - (\mathbf{y}_1, \mathbf{y}_2)\|^2 \end{aligned}$$

satisfies Assumption F for any positive τ . Needless to say, if $F(\bullet, \mathbf{x}_2)$ is strongly convex in \mathbf{x}_1 , uniformly relative to \mathbf{x}_2 , we can employ $\|\mathbf{x}_2 - \mathbf{y}_2\|^2$ instead of $\|(\mathbf{x}_1, \mathbf{x}_2) - (\mathbf{y}_1, \mathbf{y}_2)\|^2$.

6. If the set \mathcal{K} is contained in $\mathbb{R}_{++}^n \triangleq \{x \in \mathbb{R}^n : x_i > 0, \forall i\}$, we can set

$$\begin{aligned} \tilde{F}(\mathbf{x}; \mathbf{y}) &\triangleq F(\mathbf{y}) + \sum_{i: \nabla_i F(\mathbf{y}) \geq 0} \nabla_i F(\mathbf{y})(x_i - y_i) \\ &\quad - \sum_{i: \nabla_i F(\mathbf{y}) < 0} \nabla_i F(\mathbf{y}) \left(\frac{y_i^2}{x_i} - y_i \right) + \frac{\tau}{2} \|\mathbf{x} - \mathbf{y}\|^2, \end{aligned}$$

where τ is a positive constant. This kind of approximation is used in structural optimization problems [19,47,48] for which it proves to be extremely effective.

5.2 Approximating H or g_j

The approximations of the nondifferentiable component H of the objective function and of the constraint g_j must obey almost the same assumptions. In fact, Assumptions H and G differ only in the additional condition G4. Since all our examples do satisfy G4, we deal with the approximations for H and g_j together, referring to a generic locally Lipschitz function h . In Assumption B we have also condition G7, which is slightly harder to satisfy. We do not require the approximations of H or g_j to be differentiable; in fact, what matters is that the resulting subproblem $(P1_{\mathbf{y}})$, can be efficiently solved, and this does not necessarily imply that \tilde{H} or \tilde{g}_j should be differentiable, see e.g. the discussion after (46). Below we provide several examples, discussing only those properties (among G1-G6 and G7) that are not immediately obvious.

1. If h is convex, we can simply set $\tilde{h}(\mathbf{x}; \mathbf{y}) \triangleq h(\mathbf{x})$. In particular, if h is a constraint function, this means that such constraint can be incorporated in the convex set \mathcal{K} .

2. If h is concave, the linear approximation

$$\tilde{h}(\mathbf{x}; \mathbf{y}) \triangleq h(\mathbf{y}) + \boldsymbol{\xi}^T(\mathbf{x} - \mathbf{y}),$$

for any $\boldsymbol{\xi} \in \partial h(\mathbf{y})$, satisfies G4-G6 thanks to the local boundedness and outer semicontinuity of the subgradient set of the locally Lipschitz function h . In order for G7 to hold, we must assume h to be continuously differentiable with locally Lipschitz gradient.

3. Consider $h(\mathbf{x}) = \min_i \{h_i(\mathbf{x})\}$, with h_i smooth and convex for every $i = 1, \dots, p$. The constraint $\min_i \{h_i(\mathbf{x})\} \leq 0$ is a difficult *disjunctive-type* constraint requiring at least one of the $h_i(\mathbf{x}) \leq 0$ to be satisfied. In order to define a suitable approximation, we derive first an exact formula for the subdifferential of the min function. To this end, let $I(\mathbf{y}) \triangleq \{i \in \{1, \dots, p\} | h(\mathbf{y}) = h_i(\mathbf{y})\}$ be the set of active indices at \mathbf{y} and $I^e(\mathbf{y}) \triangleq \{i \in \{1, \dots, p\} | \mathbf{y} \in \text{cl int}\{\mathbf{x} | h(\mathbf{x}) = h_i(\mathbf{x})\} \subseteq I(\mathbf{y})$ be the set of essentially active indices at \mathbf{y} (cl and int indicate the closure and the interior of a set).

Proposition 7 *Let all h_i be continuously differentiable (not necessarily convex). Then,*

$$\partial h(\mathbf{y}) = \bigcup_{i \in I^e(\mathbf{y})} \{\nabla h_i(\mathbf{y})\}. \quad (43)$$

Proof By [31, Proposition 1.113] and [41, Proposition 4.1.1], we have:

$$\partial h(\mathbf{y}) \subseteq \bigcup_{i \in I^e(\mathbf{y})} \{\nabla h_i(\mathbf{y})\}.$$

We prove now that $\nabla h_i(\mathbf{y}) \in \partial h(\mathbf{y})$ for every $i \in I^e(\mathbf{y})$. Indeed, for any $i \in I^e(\mathbf{y})$, we have $\mathbf{y} \in \text{cl int}\{\mathbf{x} | h(\mathbf{x}) = h_i(\mathbf{x})\}$. Thus, there exists a sequence \mathbf{w}^v converging to \mathbf{y} such that $\mathbf{w}^v \in \text{int}\{\mathbf{x} | h(\mathbf{x}) = h_i(\mathbf{x})\}$ and $\partial h(\mathbf{w}^v) = \{\nabla h(\mathbf{w}^v)\} = \{\nabla h_i(\mathbf{w}^v)\}$ for every v . Therefore, thanks to [39, Theorem 9.61], $\nabla h_i(\mathbf{y}) \in \partial h(\mathbf{y})$ for every $i \in I^e(\mathbf{y})$ and the thesis follows. \square

In view of the previous considerations, we propose the approximation

$$\tilde{h}(\mathbf{x}; \mathbf{y}) \triangleq h_{i_y}(\mathbf{x}), \quad \text{for any } i_y \in I^e(\mathbf{y}).$$

As regards G4, it is enough to recall that $I(\mathbf{w}) \subseteq I(\mathbf{y})$ for every \mathbf{w} in some neighborhood of \mathbf{y} . Observing that $\partial_1 \tilde{h}(\mathbf{x}; \mathbf{y}) = \{\nabla h_{i_y}(\mathbf{x})\}$ with $i_y \in I^e(\mathbf{y})$, G5 holds trivially. Turning to G6, we preliminarily check that $I^e(\mathbf{w}) \subseteq I^e(\mathbf{y})$ for every \mathbf{w} sufficiently close to \mathbf{y} . We show that if $i \notin I^e(\mathbf{y})$, then, for every \mathbf{w} in some neighborhood of \mathbf{y} , we have $i \notin I^e(\mathbf{w})$. Let $i \notin I^e(\mathbf{y})$. Then, there exists a neighborhood \mathcal{V} of \mathbf{y} with $\mathcal{V} \cap \text{int}\{\mathbf{x} | h(\mathbf{x}) = h_i(\mathbf{x})\} = \emptyset$. If there were $\mathbf{w} \in \text{int}\mathcal{V}$ such that $i \in I^e(\mathbf{w})$, we would also have $\mathbf{w} \in \text{cl int}\{\mathbf{x} | h(\mathbf{x}) = h_i(\mathbf{x})\}$. Hence, there would be a sequence $\{\mathbf{v}^v\}$ converging to \mathbf{w} with $\mathbf{v}^v \in \text{int}\{\mathbf{x} | h(\mathbf{x}) = h_i(\mathbf{x})\}$ and we would get $\mathbf{v}^v \notin \mathcal{V}$ for every v , a contradiction. Therefore, $I^e(\mathbf{w}) \subseteq I^e(\mathbf{y})$ and this fact, together with Proposition 7, shows that also G6 holds. G7 does not hold for this approximation.

More in general, if each h_i is only smooth (i.e., not necessarily convex) and, for every h_i , there exists a differentiable convex approximation \tilde{h}_i such that Assumption G holds, the convex function given by

$$\tilde{h}(\mathbf{x}; \mathbf{y}) \triangleq \tilde{h}_{i_{\mathbf{y}}}(\mathbf{x}; \mathbf{y}), \quad \text{for any } i_{\mathbf{y}} \in I^e(\mathbf{y}),$$

satisfies Assumption G.

We conclude the analysis of this case by providing a very simple procedure that, in most practical cases, provides an index in $I^e(\mathbf{y})$. Given \mathbf{y} , we can easily compute $I(\mathbf{y})$. Now, generate a random nonzero direction d and set $i_{\mathbf{y}} \triangleq \operatorname{argmin}_{i \in I(\mathbf{y})} \{\nabla h_i(\mathbf{y})^T d\}$. By using standard continuity arguments on Taylor expansions, it is easy to see that if the minimum in the definition of $i_{\mathbf{y}}$ is reached for a single index, then $i_{\mathbf{y}} \in I^e(\mathbf{y})$. If the minimum is not unique, we can simply generate a different direction or use a second order expansion if the h_i s are twice continuously differentiable; we omit the details.

- Suppose that h is the difference of a smooth convex function h^+ and a possibly nonsmooth, convex function h^- , having thus the following nonsmooth DC structure: $h(\mathbf{x}) = h^+(\mathbf{x}) - h^-(\mathbf{x})$. By linearizing the concave part $-h^-$ and keeping the convex (smooth) part h^+ unchanged, we obtain the following convex upper approximation of h :

$$\tilde{h}(\mathbf{x}; \mathbf{y}) \triangleq h^+(\mathbf{x}) - h^-(\mathbf{y}) + \xi^T(\mathbf{x} - \mathbf{y}),$$

- for any $\xi \in \partial(-h^-(\mathbf{y}))$. Note that, by [39, Exercise 10.10], we have $\partial h(\mathbf{y}) = \nabla h^+(\mathbf{y}) + \partial(-h^-(\mathbf{y}))$ and $\partial_1 \tilde{h}(\mathbf{x}; \mathbf{y}) = \nabla h^+(\mathbf{x}) + \xi$, with $\xi \in \partial(-h^-(\mathbf{y}))$. In view of the local boundedness and outer semicontinuity of the subgradient set of the locally Lipschitz function $-h^-$, G4-G6 are valid. In order for G7 to hold, we must assume h^- to be continuously differentiable with locally Lipschitz gradient. To make a concrete example, suppose that $h^-(\mathbf{x}) \triangleq \max_i \{h_i^-(\mathbf{x})\}$, with $h_i^-(\mathbf{x})$ convex and smooth for every $i = 1, \dots, p$. We can then write $h(\mathbf{x}) = h^+(\mathbf{x}) + \min_i \{-h_i^-(\mathbf{x})\}$, and thus we can resort to Proposition 3 to compute $\partial \min_i \{-h_i^-(\mathbf{x})\}$.
- On the other hand, when h is the difference of a convex nonsmooth function h^+ and a convex smooth function h^- , by linearizing the concave part h^- and keeping the convex (nonsmooth) part h^+ unchanged, we obtain

$$\tilde{h}(\mathbf{x}; \mathbf{y}) \triangleq h^+(\mathbf{x}) - h^-(\mathbf{y}) - \nabla h^-(\mathbf{y})^T(\mathbf{x} - \mathbf{y}).$$

- We remark, again by [39, Exercise 10.10], that $\partial h(\mathbf{y}) = \partial h^+(\mathbf{y}) - \nabla h^-(\mathbf{y})$ and $\partial_1 \tilde{h}(\mathbf{x}; \mathbf{y}) = \partial h^+(\mathbf{x}) - \nabla h^-(\mathbf{y})$. Therefore, G4-G6 hold. In this case G7 holds if the gradient of h^- is locally Lipschitz.
- Let $h(\mathbf{x}) = s(\mathbf{f}(\mathbf{x}))$, where $s : \mathbb{R}^m \rightarrow \mathbb{R}$ is a finite convex function such that $s(u_1, \dots, u_m)$ is nondecreasing in each u_j , and $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is a smooth mapping, with $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \dots, f_m(\mathbf{x}))^T$ and f_i concave for every i . Consider the convex approximation

$$\tilde{h}(\mathbf{x}; \mathbf{y}) \triangleq s(\mathbf{f}(\mathbf{y}) + \nabla \mathbf{f}(\mathbf{y})(\mathbf{x} - \mathbf{y})), \tag{44}$$

where $\nabla \mathbf{f}(\mathbf{y})$ is the Jacobian of \mathbf{f} at \mathbf{y} . G1 holds by, e.g., [39, Exercise 2.20], and so obviously do G2–G4. Furthermore, since, in view of [39, Theorem 10.6], $\partial h(\mathbf{y}) = \nabla \mathbf{f}(\mathbf{y})^T \partial s(\mathbf{f}(\mathbf{y}))$ and $\partial_1 \tilde{h}(\mathbf{x}; \mathbf{y}) = \nabla \mathbf{f}(\mathbf{y})^T \partial s(\mathbf{f}(\mathbf{y})) + \nabla \mathbf{f}(\mathbf{y})(\mathbf{x} - \mathbf{y})$, also G5 and G6 are valid. Finally, G7 is clearly satisfied. This example can be expanded assuming that the f_i are smooth mappings, not necessarily concave, for which approximations $\tilde{f}_i(\mathbf{x}; \mathbf{y})$ satisfying Assumption G exist and such that each $\tilde{f}_i(\bullet, \mathbf{y})$ is C^1 . In this case, following the same steps above, it is easy to see that the approximation $\tilde{h}(\mathbf{x}; \mathbf{y}) \triangleq s(\tilde{f}_1(\mathbf{x}; \mathbf{y}), \dots, \tilde{f}_m(\mathbf{x}; \mathbf{y}))$ satisfies Assumption G. If in addition all \tilde{f}_i satisfy G7, also $\tilde{h}(\mathbf{x}; \mathbf{y})$ satisfies G7.

7. Consider the widely used soft thresholding function $h(x) = \min\{x+a, \max\{0, x-a\}\}$ with $a > 0$. We can resort to the following approximation

$$\tilde{h}(x; y) \triangleq \begin{cases} \max\{0, x-a\}, & y \geq -a \\ x+a, & y < -a. \end{cases}$$

Conditions G4–G6 trivially hold at any point different from $(-a, -a)$. The same happens also in this “problematic” point: indeed, $\limsup_{x, y \rightarrow -a} \partial_1 \tilde{h}(x; y) = \{0, 1\} = \partial h(-a)$. Also G7 holds quite clearly. Note that, as in the previous case, we can easily extend this approximation to the composition of the soft thresholding function with a smooth function f for which a suitable approximation \tilde{f} is available; we leave the details to the reader.

8. In the case of a nonconvex function h with a Lipschitz continuous (with constant $L_{\nabla h}$) gradient, the following convex approximation is such that conditions G1–G7 hold:

$$\tilde{h}(\mathbf{x}; \mathbf{y}) \triangleq h(\mathbf{y}) + \nabla h(\mathbf{y})^T (\mathbf{x} - \mathbf{y}) + \frac{L_{\nabla h}}{2} \|\mathbf{x} - \mathbf{y}\|^2.$$

9. Whenever $h : \mathbb{R} \rightarrow \mathbb{R}$ is polynomial, thus,

$$h(x) = a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0,$$

relying on its Taylor series expansion around y , the approximation

$$\begin{aligned} \tilde{h}(x; y) \triangleq & h(y) + h'(y)(x - y) + (n - 1) \max_{i=2, \dots, n} \left\{ \frac{|h^{(i)}(y)|}{i!} \right\} \\ & \times [(x - y)^2 + (x - y)^{\bar{n}}], \end{aligned}$$

where $\bar{n} = n$ if n is even and $\bar{n} = n + 1$ otherwise, satisfies G1–G7. The procedure adopted here for unidimensional polynomials can easily be extended to the multidimensional case, but the required computations can become very expensive for dimensions greater than two.

5.3 Parallel/distributed implementations

AF 1 can be viewed as a double loop iterative scheme. Given an *outer* iteration \mathbf{x}^v , to generate \mathbf{x}^{v+1} one needs to (a) (approximatively) solve problem $(P1_{\mathbf{x}^v})$ and (b) choose

a suitable γ^ν . In this subsection, we make some high-level considerations on how to perform in a parallel/distributed fashion operations (a) and (b) mentioned above, leading to a parallel/distributed instance of AF 1; more detailed discussions could only be made with reference to specific problems.

As far as task (b) is concerned, one can choose any of the rules (i)-(iii) in Proposition 3 or (iv) in Proposition 4. Options (i), (ii) and (iv) are more suitable for a fully parallel/distributed implementation than the line-search procedure (iii), which instead requires a high degree of coordination among cores/nodes in a parallel/distributed environment. Also, experience shows that, although theoretically attractive, option (i) could perform poorly, since practical estimates of $c_{\tilde{F}}/L_{\nabla F}$ are usually very small, leading at each iteration to little progress towards stationarity. On the other hand, option (iv) can be used only in the presence of strongly convex objective functions. Therefore, the diminishing step-size rule (ii) seems to be the more appropriate choice for parallel/distributed environments. This is the main reason why in Sect. 4 special attention was devoted to the study of (ii).

Referring to task (a), the design of parallel/distributed solution methods for $(P1_{\mathbf{x}^\nu})$ depends on the specific structure of the original problem (P1) as well as on the choice of the approximations \tilde{F} , \tilde{H} , and \tilde{g} . The flexibility offered by our approach in the definition of $(P1_{\mathbf{x}^\nu})$ is of great help to obtain alternative distributed solution methods. Suppose that the feasible set of (P1) is block-separable, i.e., $\mathcal{K} \triangleq \mathcal{K}_1 \times \dots \times \mathcal{K}_p$, with each $\mathcal{K}_i \subseteq \mathbb{R}^{n_i}$, and $\mathcal{X} \triangleq \{\mathbf{x} \triangleq (\mathbf{x}_i)_{i=1}^p, \mathbf{x}_i \in \mathcal{K}_i, i = 1, \dots, p : \sum_{i=1}^p g_j^i(\mathbf{x}_i) \leq \mathbf{0}, j = 1, \dots, m\}$. Then, problem (P1) becomes

$$\begin{aligned} & \underset{\mathbf{x}}{\text{minimize}} && F(\mathbf{x}) + H(\mathbf{x}) \\ & \text{s.t.} && \sum_{i=1}^p g_j^i(\mathbf{x}_i) \leq 0, \quad j = 1, \dots, m \\ & && \mathbf{x}_i \in \mathcal{K}_i, \quad i = 1, \dots, p, \end{aligned} \tag{45}$$

The structure of (45) naturally suggests approximations \tilde{F} , \tilde{H} , and \tilde{g}_j^i that are additively block-separable, like the examples given in case 2 and 3 in Sect. 5.1 and in case 2 and 8 in Sect. 5.2. This results in convex subproblems of the form

$$\begin{aligned} & \underset{\mathbf{x}}{\text{minimize}} && \sum_{i=1}^p \tilde{F}_i(\mathbf{x}_i; \mathbf{x}^\nu) + \sum_{i=1}^p \tilde{H}_i(\mathbf{x}_i; \mathbf{x}^\nu) \\ & \text{s.t.} && \sum_{i=1}^p \tilde{g}_j^i(\mathbf{x}_i; \mathbf{x}^\nu) \leq 0, \quad j = 1, \dots, m \\ & && \mathbf{x}_i \in \mathcal{K}_i, \quad i = 1, \dots, p, \end{aligned} \tag{46}$$

that can be solved in a parallel/distributed way by standard techniques, see, e.g., [6]. It is worth mentioning a special instance of (45), where the feasible set has a Cartesian structure, i.e.,

$$\begin{aligned} & \underset{\mathbf{x}}{\text{minimize}} && F(\mathbf{x}) + H(\mathbf{x}) \\ & \text{s.t.} && \mathbf{g}_i(\mathbf{x}_i) \leq \mathbf{0}, \quad i = 1, \dots, p \\ & && \mathbf{x}_i \in \mathcal{K}_i, \quad i = 1, \dots, p. \end{aligned} \tag{47}$$

Choosing separable approximations for F and H , in the form $\tilde{F}(\mathbf{x}; \mathbf{x}^\nu) \triangleq \sum_{i=1}^p \tilde{F}_i(\mathbf{x}_i; \mathbf{x}^\nu)$ and $\tilde{H}(\mathbf{x}; \mathbf{x}^\nu) \triangleq \sum_{i=1}^p \tilde{H}_i(\mathbf{x}_i; \mathbf{x}^\nu)$, lends itself to *totally separable* convex subproblems that can be solved by tackling *in parallel* p strongly convex problems. We point out that, even though the constraint functions g_j s in (P1) are not additively block-separable as in (45), it might still be possible to choose appropriate approximations so that the resulting subproblems (P1 $_{\mathbf{x}^\nu}$) can be solved in a parallel/distributed fashion, see for example cases 2 and 8 in Sect. 5.2. It is also interesting to observe that, when (P1) involves nondifferentiable functions, our approach provides a systematic way to possibly develop parallel and distributed solution methods for classes of problems for which such option is not currently available.

6 Green communications: energy minimization in wireless systems

In this section we present an application of our algorithmic framework to an important open problem in *green communications*, resulting in an efficient *ad-hoc* solution method for this class of problems. Our numerical results show superior performance with respect to recent proposals.

Energy efficiency and Quality-of-Service (QoS) have been two key aspects in the design of modern multiuser communication systems. Energy efficiency is a growing concern due to energy costs and associated environmental issues. It has been reported that ICT infrastructures account for more than 3% of the world energy consumptions [22].

Moreover, given the rate at which wireless connected devices are increasing, as well as the mass deployment of 5G systems, mobile communications are expected to consume significantly more energy, if no countermeasures are taken. On the other hand, the explosive growth of data and multimedia services in wireless networks (video conferencing, streaming, mobile TV, 3D services) entails strict guarantees on QoS, such as delay and data rates, which contrast with the need of energy saving. These issues have motivated in the past few years a growing interest in energy-aware optimization design of wireless communication networks, under general QoS constraints [24, 29, 53]. This emerging research area is called green communications; see [52] for a recent overview of the state of the art. Although there are substantial differences between the many, often specialized, formulations of problems in this field, all approaches share the definition of the energy efficiency as fractional functions and leverage fractional programming as a key tool for the analysis and design of solution methods.

In this section, we consider a very general and challenging energy-efficient design in green communications, namely: the minimization of the sum-energy in MIMO multiuser interference networks, subject to rate profile constraints. The resulting optimization problem is nonconvex, with nonconvexities occurring both in the objective function and in the constraints. The multiuser nature of the system, along with concurrent interfering communications, prevent one to apply fractional programming-based methods, leaving the design of efficient solution methods an open problem [52]. Furthermore, the proposed formulation uses complex variables and calls for a solution method that preserves feasibility of the iterates, thus ruling out off-the-shelf solvers.

We remark that, although in this section we focus on a specific setup (MIMO ad-hoc interference networks) and formulation (the energy minimization subject to QoS con-

straints), the proposed methodology and framework open the way to the solution of a variety of currently open resource allocation problems in (green) communications [52]. Examples include (i) energy efficient designs in relay MIMO interference networks, multi-cell cellular systems, or densely deployed small cells; (ii) max-min multicast (capacity) multi-stream beamforming-like problems; and (iii) *joint* optimization of precoders, receive filters, and power allocation in the scenarios described in (i)-(ii).

6.1 Problem formulation

In this subsection, we freely use standard facts in communications (see, e.g., [49]). The reader interested mainly in the mathematical formulation of the problem can take (50) as starting point of the subsequent discussion.

Consider a wireless *ad-hoc* network composed of I multiple antenna pairs transmitter-receiver, communicating over MIMO channels; we denote by $\mathcal{I} \triangleq \{1, \dots, I\}$ the set of active pairs (also termed users) in the system. Each transmitter and receiver i is equipped with T_i and R_i antennas, respectively. The optimization variables of each transceiver i are given by the transmit covariance matrix $\mathbf{Q}_i \in \mathbb{C}^{T_i \times T_i}$, which is a positive semi-definite matrix. Each transmitter i is subject to the power constraint $\text{tr}(\mathbf{Q}_i) \leq P_i$, where P_i is the maximum average transmit power, and $\text{tr}(\mathbf{Q}_i)$ denotes the trace of \mathbf{Q}_i . Since no *a-priori* multiple access scheme is assumed for the users (like, e.g., OFDMA, TDMA, or CDMA), multiuser interference is experienced at each receiver. Under standard information theoretical assumptions, the maximum achievable rate r_i on each MIMO link i (measured in bit/s/Hz) can be written as follows [49]: given $\mathbf{Q} \triangleq (\mathbf{Q}_1, \dots, \mathbf{Q}_I)$, with each $\mathbf{Q}_i \geq \mathbf{0}$,

$$r_i(\mathbf{Q}) \triangleq \log_2 \det \left(\mathbf{I} + \mathbf{H}_{ii} \mathbf{Q}_i \mathbf{H}_{ii}^H \mathbf{R}_i (\mathbf{Q}_{-i})^{-1} \right), \tag{48}$$

where $\mathbf{H}_{ij} \in \mathbb{C}^{R_i \times T_j}$ is the channel matrix between the transmitter of user j and the receiver of link i ; $\mathbf{R}_i (\mathbf{Q}_{-i}) \triangleq \sigma_i^2 \cdot \mathbf{I} + \sum_{j \neq i} \mathbf{H}_{ij} \mathbf{Q}_j \mathbf{H}_{ij}^H$ is the covariance matrix of the Gaussian noise $\sigma_i^2 \cdot \mathbf{I}$ (assumed to be proportional to the identity matrix, without loss of generality, otherwise one can always pre-whiten the channel matrices) plus the interference at the receiver i due to the other transmitters; and $\mathbf{Q}_{-i} \triangleq (\mathbf{Q}_1, \dots, \mathbf{Q}_{i-1}, \mathbf{Q}_{i+1}, \dots, \mathbf{Q}_I)$ is the tuple of the covariance matrices of all the transmitters except the i -th one. If the transmission of each user i takes T_i seconds to complete (implying thus $r_i(\mathbf{Q}) > 0$), the energy consumed by user i per bit that can be reliably transmitted is

$$E_i(\mathbf{Q}) \triangleq \frac{T_i (\mu_i \text{tr}(\mathbf{Q}_i) + P_{c,i})}{T_i W_i r_i(\mathbf{Q})} = \frac{(\mu_i \text{tr}(\mathbf{Q}_i) + P_{c,i})}{W_i r_i(\mathbf{Q})} \text{ [J/bit]}, \tag{49}$$

where W_i is the communication bandwidth of user i , $1/\mu_i$ is the efficiency of the power amplifier at transmitter i , while $P_{c,i}$ includes the power dissipated in all other circuit blocks of the transmitter and receiver i to operate the terminals (assumed to be constant [2]). The QoS of each communication is measured in terms of a given

minimum achievable rate $\bar{r}_i > 0$, leading to constraints of the type $r_i(\mathbf{Q}) \geq \bar{r}_i$. The energy-efficient design of the wireless system consists then in minimizing the sum of the users' energy consumption while guaranteeing the rate QoS:

$$\begin{aligned} \underset{\mathbf{Q}}{\text{minimize}} \quad & E(\mathbf{Q}) \triangleq \sum_{i=1}^I E_i(\mathbf{Q}) \\ \text{s.t.} \quad & g_i(\mathbf{Q}) \triangleq \bar{r}_i - r_i(\mathbf{Q}) \leq 0, \quad \forall i = 1, \dots, I, \\ & \mathbf{Q}_i \geq \mathbf{0}, \quad \text{tr}(\mathbf{Q}_i) \leq P_i, \quad \forall i = 1, \dots, I. \end{aligned} \quad (50)$$

Note that in the above formulation, the g_i s define the nonconvex constraints whereas $\mathbf{Q}_i \geq \mathbf{0}$ and $\text{tr}(\mathbf{Q}_i) \leq P_i$ are convex constraints and define the set \mathcal{K} in (P1). Hereafter, we denote by \mathcal{Q} the feasible set of problem (50). We also assume that P_i s and \bar{r}_i s are set so that \mathcal{Q} is nonempty.

Remark 2 Problem (50) contains complex variables. One could reformulate the problem into the real domain by using separate variables for the real and imaginary parts of the complex variables, but this would lead to awkward reformulations wherein all the desirable structure of the original functions gets lost. Following a well-established path in the signal processing community, we work directly with complex variables by means of “Wirtinger derivatives”. The main advantage of this approach is that we can use “Wirtinger calculus” to easily compute in practice derivatives of the rate (and energy) functions in (50) directly in the complex domain. It can be shown that all results in this paper extend to the complex domain when using Wirtinger derivatives instead of classical gradients. In what follows we freely use the Wirtinger calculus and refer the reader to [23,26,42] for more information on this topic.

Problem (50) represents an open problem in the wireless communication community, see [52]. Furthermore, the constraints $\mathbf{Q}_i \geq \mathbf{0}$ and $g_i(\mathbf{Q}) \leq 0$ cannot be violated, since the objective function could be not defined outside the feasible set \mathcal{Q} , calling thus for the use of a feasible method. Since (50) is a special case of (P1), one can readily resort to our approach to efficiently compute stationary solutions. We propose next a novel nontrivial *ad-hoc* approximation of the objective function, opening the way to an efficient INCA-based solution method, which exhibits very good numerical performances.

6.2 Ad-hoc approximations

In this section we define suitable approximations \tilde{E} and \tilde{g}_i for (50). We remark that, while the proposed approximations satisfy Assumptions F, G and B, the one of the objective function is substantially different from those in Sect. 5 and exploits the partial convexity of E in each of the users' variables \mathbf{Q}_i s.

Given $\mathbf{Q}^v \triangleq (\mathbf{Q}_i^v)_{i=1}^I$ such that each $r_i(\mathbf{Q}^v) > 0$, observe that: (i) setting $\mathbf{Q}_{-i} = \mathbf{Q}_{-i}^v$, each term $E_i(\mathbf{Q}_i, \mathbf{Q}_{-i}^v)$ of the sum in $E(\mathbf{Q})$ is the product of a linear and a convex function in \mathbf{Q}_i , namely $\mu_i \text{tr}(\mathbf{Q}_i) + P_{c,i}$ and $1/W_i r_i(\mathbf{Q}_i, \mathbf{Q}_{-i}^v)$; and (ii) the other summands, $\sum_{j \neq i} E_j(\mathbf{Q}_i, \mathbf{Q}_{-j}^v)$, are nonconvex in \mathbf{Q}_i . Exploiting such a

structure, a convex approximation of E can be obtained for each user i by convexifying $(\mu_i \text{tr}(\mathbf{Q}_i) + P_{c,i}) / (W_i r_i(\mathbf{Q}_i, \mathbf{Q}_{-i}^v))$, while retaining the partial convexity in \mathbf{Q}_i and linearizing the nonconvex part $\sum_{j \neq i} E_j(\mathbf{Q}_i, \mathbf{Q}_{-j}^v)$. More formally, let us define

$$\begin{aligned} \tilde{E}_i(\mathbf{Q}_i; \mathbf{Q}^v) \triangleq & \frac{\mu_i \text{tr}(\mathbf{Q}_i) + P_{c,i}}{W_i r_i(\mathbf{Q}^v)} + \frac{\mu_i \text{tr}(\mathbf{Q}_i^v) + P_{c,i}}{W_i r_i(\mathbf{Q}_i, \mathbf{Q}_{-i}^v)} \\ & + \sum_{j \neq i} \left\langle \nabla_{\mathbf{Q}_i^*} E_j(\mathbf{Q}^v), \mathbf{Q}_i - \mathbf{Q}_i^v \right\rangle + \tau_i \|\mathbf{Q}_i - \mathbf{Q}_i^v\|_F^2, \end{aligned} \tag{51}$$

where the first two terms on the right-hand-side are the convexification of $(\mu_i \text{tr}(\mathbf{Q}_i) + P_{c,i}) / (W_i r_i(\mathbf{Q}_i, \mathbf{Q}_{-i}^v))$; the third one comes from the linearization of $\sum_{j \neq i} E_j(\mathbf{Q}_i, \mathbf{Q}_{-j}^v)$, with $\langle \mathbf{A}, \mathbf{B} \rangle \triangleq \text{Re}\{\text{tr}(\mathbf{A}^H \mathbf{B})\}$ and $\nabla_{\mathbf{Q}_i^*} E_j(\mathbf{Q}^v)$ denoting the conjugate gradient of E_j w.r.t. \mathbf{Q}_i evaluated at \mathbf{Q}^v (see [42]), given by

$$\frac{(\mu_j \text{tr}(\mathbf{Q}_j) + P_{c,j})}{W_j r_j^2(\mathbf{Q}^v)} \mathbf{H}_{ji}^H \left(\mathbf{R}_j(\mathbf{Q}_{-j}^v)^{-1} - \left(\mathbf{R}_j(\mathbf{Q}_{-j}^v) + \mathbf{H}_{jj} \mathbf{Q}_j^v \mathbf{H}_{jj}^H \right)^{-1} \right) \mathbf{H}_{ji};$$

and the fourth term is added to make $\tilde{E}_i(\mathbf{Q}_i; \mathbf{Q}^v)$ uniformly strongly convex in \mathbf{Q}_i . Thus, the sum-energy surrogate function $\tilde{E}(\mathbf{Q}; \mathbf{Q}^v)$ is defined as:

$$\tilde{E}(\mathbf{Q}; \mathbf{Q}^v) = \sum_{i=1}^I \tilde{E}_i(\mathbf{Q}_i; \mathbf{Q}^v). \tag{52}$$

In order to build an upper convex approximation of each (nonconvex) g_i , one can readily exploit the DC structure of the rate function $r_i(\mathbf{Q})$, that is,

$$\begin{aligned} r_i(\mathbf{Q}) &= r_i^+(\mathbf{Q}) - r_i^-(\mathbf{Q}_{-i}), \\ r_i^+(\mathbf{Q}) &\triangleq \log_2 \det \left(\mathbf{R}_i(\mathbf{Q}_{-i}) + \mathbf{H}_{ii} \mathbf{Q}_i \mathbf{H}_{ii}^H \right) \text{ and } r_i^-(\mathbf{Q}_{-i}) \triangleq \log_2 \det(\mathbf{R}_i(\mathbf{Q}_{-i})). \end{aligned} \tag{53}$$

Following Example 4 in Sec. 5.2 and adapting it to the complex case, a tight lower bound of $r_i(\mathbf{Q})$ [and thus upper bound of $g_i(\mathbf{Q})$] can be obtained by retaining the concave part in (53) and linearizing the convex function $-r_i^-$, which leads to the following rate approximation function:

$$r_i(\mathbf{Q}) \geq \tilde{r}_i(\mathbf{Q}; \mathbf{Q}^v) \triangleq r_i^+(\mathbf{Q}) - \tilde{r}_i^-(\mathbf{Q}_{-i}; \mathbf{Q}^v), \tag{54}$$

$$\tilde{r}_i^-(\mathbf{Q}_{-i}; \mathbf{Q}^v) \triangleq r_i^-(\mathbf{Q}_{-i}^v) + \sum_{j \neq i} \left\langle \nabla_{\mathbf{Q}_j^*} r_i^-(\mathbf{Q}_{-i}^v), \mathbf{Q}_j - \mathbf{Q}_j^v \right\rangle, \tag{55}$$

where $\nabla_{\mathbf{Q}_j^*} r_i^-(\mathbf{Q}_{-i}^v) = \mathbf{H}_{ij}^H \mathbf{R}_i(\mathbf{Q}_{-i}^v)^{-1} \mathbf{H}_{ij}$ is the conjugate gradient of r_i^- w.r.t. \mathbf{Q}_j . Using (52) and (54), the convex approximation of the nonconvex problem (50) reads: given the feasible point \mathbf{Q}^v , let

$$\begin{aligned}
& \underset{\mathbf{Q}}{\text{minimize}} \quad \tilde{E}(\mathbf{Q}; \mathbf{Q}^v) \\
& \text{s.t.} \quad \tilde{g}_i(\mathbf{Q}; \mathbf{Q}^v) \triangleq \bar{r}_i - \tilde{r}_i(\mathbf{Q}; \mathbf{Q}^v) \leq 0, \quad \forall i = 1, \dots, I, \\
& \quad \mathbf{Q}_i \geq \mathbf{0}, \quad \text{tr}(\mathbf{Q}_i) \leq P_i, \quad \forall i = 1, \dots, I.
\end{aligned} \tag{56}$$

The following proposition, whose proof is omitted because of space limitations, summarize the main properties of problems (50) and (56).

Proposition 8 *Given problems (50) and (56), the following hold:*

- (i) any feasible point of (50) satisfies the MFCQ;
- (ii) convex surrogates \tilde{E} and \tilde{g}_i s in (56) satisfy Assumptions F, G and B.

Hence, (56) is an instance of problem (P1_y) and AF 1 can be readily applied.

6.3 Numerical results

In this section we present numerical results to assess the effectiveness of our solution method, which we term *Partial Linearization INCA* (PL-INCA). We implemented PL-INCA, based on AF 1 and convex subproblems (56), in MATLAB. We used the diminishing step-size rule $\gamma^{v+1} = \gamma^v(1 - \alpha\gamma^v)$, for all $v \geq 0$, with $\gamma^0 = 1$ and $\alpha = 1e^{-3}$. The proximal coefficients τ_i in (51) are set equal to 10^{-2} . As we mentioned earlier, standard methods cannot be used to solve problem (50), since feasibility is to be maintained throughout the iterations. This fact restricts drastically the possible choices for a comparison; we decided to settle for the very recent approach proposed in [1, 7], which we adapted to deal with complex variables. The method in [7] is based on the solution of a sequence of convex subproblems wherein the nonconvex objective function and constraints are replaced by *convex global upper* approximations. While the nonconvex constraints $g_i(\mathbf{Q}) \leq 0$ can be approximated as in our method [see \tilde{g}_i in (56)], obtaining a convex upper bound of E is less immediate. The only available option seems to set

$$\tilde{E}^{\text{up}}(\mathbf{Q}; \mathbf{Q}^v) \triangleq \sum_{i=1}^I \sum_{j=1}^I \left\langle \nabla_{\mathbf{Q}_j^*} E_i(\mathbf{Q}^v), \mathbf{Q}_j - \mathbf{Q}_j^v \right\rangle + \sum_{i=1}^I L_i \|\mathbf{Q}_i - \mathbf{Q}_i^v\|_F^2, \tag{57}$$

where L_i is the Lipschitz constant of $\nabla_{\mathbf{Q}^*} E_i(\bullet)$ on \mathcal{Q} . An exact expression of L_i is not easy to compute. The following is an upper bound of L_i over \mathcal{Q} (the proof is quite tedious and is omitted):

$$L_i^{\text{up}} = \left(\sum_{i=1}^I \frac{\mu_i P_i + P_{c,i}}{W_i \bar{r}_i} \right) \cdot \lambda_{\max} \left(\sum_{j=1}^I \left(\frac{4}{\bar{r}_j^2} + \frac{1}{\bar{r}_j} \right) \mathbf{H}_{j,i}^H \mathbf{H}_{j,i} \otimes \mathbf{H}_{j,i}^H \mathbf{H}_{j,i} \right), \tag{58}$$

where $\lambda(\mathbf{A})$ is the maximum eigenvalue of matrix \mathbf{A} , and \otimes denotes the Kronecker product. We refer to the scheme in [7] tailored to the complex problem (56) and based on (57)–(58), as *Linearized Upper INCA* (L-Up-INCA). In L-Up-INCA, given an iteration \mathbf{Q}^v , the new iteration is the solution of subproblem

$$\begin{aligned}
& \underset{\mathbf{Q}}{\text{minimize}} && \tilde{E}^{\text{up}}(\mathbf{Q}; \mathbf{Q}^v) \\
& \text{s.t.} && \tilde{g}_i(\mathbf{Q}; \mathbf{Q}^v) \leq 0, \quad \forall i = 1, \dots, I, \\
& && \mathbf{Q}_i \succeq \mathbf{0}, \quad \text{tr}(\mathbf{Q}_i) \leq P_i, \quad \forall i = 1, \dots, I,
\end{aligned} \tag{59}$$

where $\tilde{g}_i(\mathbf{Q}; \mathbf{Q}^v)$ is defined as in (56).

Both methods solve a convex optimization problem at each iteration: PL-INCA solves (56) and then takes a step-size γ^v in the direction given by the solution of the subproblem, while L-Up-INCA solves (59) and takes its solution as the new iteration (or, equivalently, takes a step-size of 1). In our implementations, the two algorithms differ in (i) the choice of the objective function in the subproblems and (ii) the step-size. While $L_i = L_i^{\text{up}}$ guarantees theoretical convergence of L-Up-INCA, in principle it might not be a tight bound of L_i , a fact that could hamper progresses of L-Up-INCA; and indeed L-Up-INCA with $L_i = L_i^{\text{up}}$ performed very poorly in all problem instances we tested, as we discuss shortly. Therefore, we also considered other instances of L-Up-INCA, obtained by progressively reducing the value of L_i . Note that if L_i falls below the Lipschitz constant of $\nabla_{\mathbf{Q}^*} E_i(\bullet)$, theoretical convergence of L-Up-INCA is no longer guaranteed.

We simulated a network composed of $I = 10$ randomly deployed users (pairs), whose transceivers are equipped with $T_i = R_i = 2$ antennas for each i . The MIMO channels are simulated according to the Rayleigh channel model [49]; the transmit powers P_i s and the rate thresholds \bar{r}_i s are chosen to guarantee that the feasible set of (50) is nonempty. Both algorithms, PL-INCA and L-Up-INCA, are initialized from the same randomly chosen feasible point. In this setting we generated 10 independent channel realizations and initial feasible points \mathbf{Q}^0 and run both algorithms on the resulting 10 problems.¹ We used *cvx* in MATLAB to solve, at each iteration, the strongly convex subproblems. Since the original problem is nonconvex, we measured the progresses of the algorithms toward convergence using the “stationarity measure” $\|\hat{\mathbf{Q}}(\mathbf{Q}^v) - \mathbf{Q}^v\|_\infty$, where $\hat{\mathbf{Q}}(\mathbf{Q}^v)$ is the unique solution of subproblem (56) and $\|\bullet\|_\infty$ denotes the infinity norm. Note that $\|\hat{\mathbf{Q}}(\mathbf{Q}^v) - \mathbf{Q}^v\|_\infty$ is a continuous function, which is zero if and only if \mathbf{Q}^v is a stationary point.

The results of our experiments are summarized in Fig. 1. In Fig. 1a, b we plot the average (over the 10 realizations) of the normalized stationarity measure $\|\hat{\mathbf{Q}}(\mathbf{Q}^v) - \mathbf{Q}^v\|_\infty / S^{\text{max}}$ and of the normalized sum-energy $E(\mathbf{Q}^v) / E^{\text{max}}$, respectively, achieved by PL-INCA and L-Up-INCA versus the number of iterations (note that we use the log-scale for the x-axis), where S^{max} and E^{max} are the maximum of the stationarity measure and of E over all the randomly chosen initial points, respectively. The normalizations are done just for the purpose of a simpler graphical representation. We did not terminate the algorithms but let them run for 5000 iterations. The figures clearly show that the performance of PL-INCA is vastly superior to that of L-Up-INCA: the stationarity measure (as well as the objective function) associated with L-Up-INCA does not decrease significantly over 5000 iterations, remaining thus much higher than that of PL-INCA. Finally, note the behavior of L-Up-INCA when L_i is set to L_i^{min} , the largest value empirically found to make the algorithm divergent on all 10 instances: for this

¹ The data of the problems are available from the authors on request.

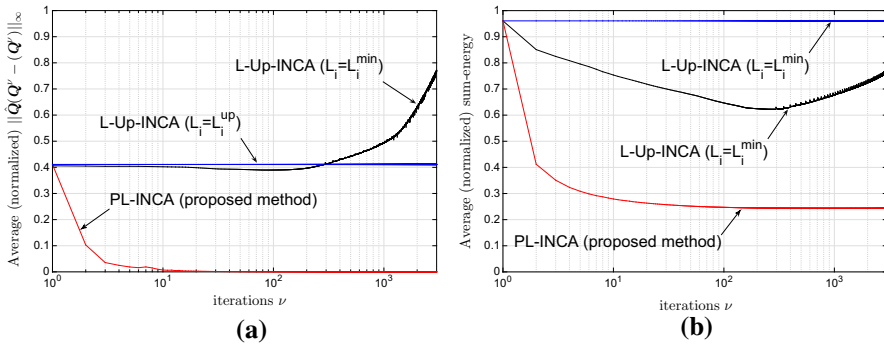


Fig. 1 PL-INCA and two instances of L-Up-INCA (the x -axis is in log scale). **a** Normalized average stationarity measure $\|\hat{Q}(\mathbf{Q}^\nu) - \mathbf{Q}^\nu\|_\infty / S^{\max}$ versus iterations ν . **b** Normalized sum-energy $E(\mathbf{Q}^\nu) / E^{\max}$ versus iterations ν

value of L_i , the stationarity measure and the objective function progressively increase, see Fig. 1. We also tested L-Up-INCA for intermediate values of L_i , between L_i^{up} and L_i^{\min} , but in all cases where clear divergence did not occur, the resulting behavior is still very poor and is described by lines very close to the ones of L-Up-INCA ($L_1 = L_1^{\text{up}}$) (in Fig. 1a, b), with a slightly better diminishing slope. While we reported only averages, we remark that the average curves are representative of the behavior of the single realizations and in no case we observed large deviations from the average behavior. With regard to the cpu times, the main computational burden per iteration is given, for both algorithms, by the solution of the respective convex subproblems, (56) for PL-INCA and (59) for L-Up-INCA. Although (59) is marginally a simpler problem than (56), we found that, on average, the time needed by cvx to solve an instance of (56) is less than twice the time to solve one of (59). In practice, this means that the huge gap between PL-INCA over L-Up-INCA in terms of iterations translates in a similar superiority in terms of computational times.

References

1. Auslender, A., Shefi, R., Teboulle, M.: A moving balls approximation method for a class of smooth constrained minimization problems. *SIAM J. Optim.* **20**(6), 3232–3259 (2010)
2. Auer, G., Giannini, V., Desset, C., Godor, I., Skillermark, P., Olsson, M., Imran, M.A., Sabella, D., Gonzalez, M.J., Blume, O., Fehske, A.: How much energy is needed to run a wireless network? *IEEE Wirel. Commun.* **18**(5), 40–49 (2011)
3. Beck, A., Ben-Tal, A., Tetrushvili, L.: A sequential parametric convex approximation method with applications to nonconvex truss topology design problems. *J. Glob. Optim.* **47**(1), 29–51 (2010)
4. Bertsekas, D.P.: *Nonlinear Programming*. Athena Scientific, Belmont (1992)
5. Bertsekas, D.P., Tsitsiklis, J.N.: *Neuro-Dynamic Programming*. Athena Scientific, Belmont (1996)
6. Bertsekas, D.P., Tsitsiklis, J.N.: *Parallel and Distributed Computation: Numerical Methods*. Athena Scientific, Belmont (1997)
7. Bolte, J., Pauwels, E.: Majorization-minimization procedures and convergence of sqp methods for semi-algebraic and tame programs. *Math. Oper. Res.* (2016)
8. Bolte, J., Sabach, S., Teboulle, M.: Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Math. Program.* **146**(1–2), 459–494 (2014)
9. Bonnans, J.F., Shapiro, A.: *Perturbation Analysis of Optimization Problems*. Springer Science & Business Media, Berlin (2000)

10. Byrd, R.H., Nocedal, J., Waltz, R.A.: Feasible interior methods using slacks for nonlinear optimization. *Comput. Optim. Appl.* **26**(1), 35–61 (2003)
11. Chen, L., Wang, Y., He, G.: A feasible active set QP-free method for nonlinear programming. *SIAM J. Optim.* **17**(2), 401–429 (2006)
12. Clarke, F.H.: *Optimization and Nonsmooth Analysis*. Wiley, New York (1983)
13. Conn, A.R., Gould, N.I.M., Orban, D., Toint, P.L.: A primal-dual trust-region algorithm for non-convex nonlinear programming. *Math. Program.* **87**(2), 215–249 (2000)
14. Facchinei, F., Lazzari, C.: Local feasible QP-free algorithms for the constrained minimization of SC1 functions. *J. Optim. Theory Appl.* **119**(2), 281–316 (2003)
15. Facchinei, F., Pang, J.-S.: *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer, Berlin (2003)
16. Facchinei, F., Scutari, G., Sagratella, S.: Parallel selective algorithms for nonconvex big data optimization. *IEEE Trans. Signal Process.* **63**(7), 1874–1889 (2015)
17. Ferris, M., Mangasarian, O.: Parallel variable distribution. *SIAM J. Optim.* **4**(4), 815–832 (1994)
18. Fiacco, A.V., Mc Cormick, G.P.: *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*. Wiley, New York (1968)
19. Fleury, C.: CONLIN: an efficient dual optimizer based on convex approximation concepts. *Struct. Opt.* **1**(2), 81–89 (1989)
20. Fleury, C., Braibant, V.: Structural optimization: a new dual method using mixed variables. *Int. J. Numer. Methods Eng.* **23**(3), 409–428 (1986)
21. Forsgren, A., Gill, P.E.: Primal-dual interior methods for nonconvex nonlinear programming. *SIAM J. Optim.* **8**(4), 1132–1152 (1998)
22. Herault, L. Holistic approach for future energy efficient cellular networks. In: 11th Leti Annual Review, CEA-LETI (2009)
23. Hjørungnes, A.: *Complex-Valued Matrix Derivatives With Applications in Signal Processing and Communications*. Cambridge University Press, London (2011)
24. Ho, P.-H., Shen, G., Subramaniam, S., Mouftah, H.T., Qiao, C., Wosinska, L.: Special issue on energy-efficiency in optical networks. *IEEE J. Sel. Areas Commun.* **32**(8) (2014)
25. Kanzow, C., Qi, H.-D.: A QP-free constrained Newton-type method for variational inequality problems. *Math. Program.* **85**(1), 81–106 (1999)
26. Kreuz-Delgado, K.: The complex gradient operator and the CR-calculus. Technical Report, Electrical and Computer Engineering, Jacobs School of Engineering, University of California, San Diego (2009)
27. Lawrence, C.T., Tits, A.L.: A computationally efficient feasible sequential quadratic programming algorithm. *SIAM J. Optim.* **11**(4), 1092–1118 (2001)
28. Levy, A.B., Poliquin, R.A., Rockafellar, R.T.: Stability of locally optimal solutions. *SIAM J. Optim.* **10**(2), 580–604 (2000)
29. Li, G.Y., Xu, S., Swami, A., Himayat, N., Fettweis, G.: Special issue on energy-efficient wireless communications. *IEEE J. Sel. Areas Commun.* **29**(8) (2011)
30. Marks, B.R., Wright, G.P.: A general inner approximation algorithm for nonconvex mathematical programs. *Oper. Res.* **26**(4), 681–683 (1978)
31. Mordukhovich, B.S.: *Variational Analysis and Generalized Differentiation I: Basic Theory*. Springer, Berlin (2006)
32. Mordukhovich, B.S., Nghia, T.T.A.: Full Lipschitzian and Hölderian stability in optimization with applications to mathematical programming and optimal control. *SIAM J. Optim.* **24**(3), 1344–1381 (2014)
33. Mordukhovich, B.S., Nghia, T.T.A., Rockafellar, R.T.: Full stability in finite-dimensional optimization. *Math. Oper. Res.* **40**, 226–252 (2014)
34. Pang, J.S., Razaviyayn, M., Alvarado, A.: Computing B-stationary points of nonsmooth DC programs. *Math. Oper. Res.* [arXiv:1511.01796](https://arxiv.org/abs/1511.01796) (2014) (submitted)
35. Panier, E.R., Tits, A.L.: A superlinearly convergent feasible method for the solution of inequality constrained optimization problems. *SIAM J. Control Optim.* **25**(4), 934–950 (1987)
36. Panier, E.R., Tits, A.L.: On combining feasibility, descent and superlinear convergence in inequality constrained optimization. *Math. Program.* **59**(1–3), 261–276 (1993)
37. Panier, E.R., Tits, A.L., Harskovits, J.N.: A QP-free, globally convergent, locally superlinearly convergent algorithm for inequality constrained optimization. *SIAM J. Control Optim.* **26**(4), 788–811 (1988)

38. Qi, H.-D., Qi, L.: A new QP-free, globally convergent, locally superlinearly convergent algorithm for inequality constrained optimization. *SIAM J. Optim.* **11**(1), 113–132 (2000)
39. Rockafellar, R.T., Wets, J.B.: *Variational Analysis*. Springer, Berlin (1998)
40. Sagastizábal, C.A., Solodov, M.V.: Parallel variable distribution for constrained optimization. *Comput. Optim. Appl.* **22**(1), 111–131 (2002)
41. Scholtes, S.: *Introduction to Piecewise Differentiable Equations*. Springer, Berlin (2012)
42. Scutari, G., Facchinei, F., Pang, J.-S., Palomar, D.P.: Real and complex monotone communication games. *IEEE Trans. Inf. Theory* **60**(7), 4197–4231 (2014)
43. Scutari, G., Facchinei, F., Lampariello, L., Song, P.: Distributed methods for constrained nonconvex multi-agent optimization-part I: Theory. [arXiv:1410.4754](https://arxiv.org/abs/1410.4754) (2014)
44. Scutari, G., Facchinei, F., Lampariello, L., Song, P.: Parallel and distributed methods for nonconvex optimization. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 840–844. IEEE (2014)
45. Scutari, G., Facchinei, F., Song, P., Palomar, D.P., Pang, J.-S.: Decomposition by partial linearization: parallel optimization of multi-agent systems. *IEEE Trans. Signal Process.* **62**(3), 641–656 (2014)
46. Solodov, M.: On the convergence of constrained parallel variable distribution algorithms. *SIAM J. Optim.* **8**(1), 187–196 (1998)
47. Svanberg, K.: The method of moving asymptotes—a new method for structural optimization. *Int. Jour. for Numerical Methods in Engineering* **24**(2), 359–373 (1987)
48. Svanberg, K.: A class of globally convergent optimization methods based on conservative convex separable approximations. *SIAM J. Optim.* **12**(2), 555–573 (2002)
49. Tse, D., Viswanath, P.: *Fundamentals of Wireless Communication*. Cambridge University Press, Cambridge (2005)
50. Yang, Y.-F., Li, D.-H., Qi, L.: A feasible sequential linear equation method for inequality constrained optimization. *SIAM J. Optim.* **13**(4), 1222–1244 (2003)
51. Yen, N.D.: Hölder continuity of solutions to a parametric variational inequality. *Appl. Math. Optim.* **31**(3), 245–255 (1995)
52. Zappone, A., Jorswieck, E.A.: *Energy Efficiency in Wireless Networks via Fractional Programming Theory*. Foundations and Trends® in Communications and Information Theory, (2015)
53. Zhang, H., Gladisch, A., Pickavet, M., Tao, Z., Mohr, W.: Special issue on energy efficiency in communications. *IEEE Commun. Mag.* **48**(11) (2010)