Parallel and Distributed Methods for Constrained Nonconvex Optimization—Part I: Theory

Gesualdo Scutari, *Senior Member, IEEE*, Francisco Facchinei, and Lorenzo Lampariello

***Abstract*—In this two-part paper, we propose a general algorithmic framework for the minimization of a nonconvex smooth function subject to *nonconvex* smooth constraints, and also consider extensions to some structured, nonsmooth problems. The algorithm solves a sequence of *(separable) strongly convex* problems and maintains feasibility at each iteration. Convergence to a stationary solution of the original nonconvex optimization is established. Our framework is very general and flexible and unifies several existing successive convex approximation (SCA)-based algorithms. More importantly, and differently from current SCA approaches, it naturally leads to *distributed and parallelizable* implementations for a large class of nonconvex problems. This Part I is devoted to the description of the framework in its generality. In Part II, we customize our general methods to several (multiagent) optimization problems in communications, networking, and machine learning; the result is a new class of centralized and *distributed* algorithms that compare favorably to existing ad-hoc (centralized) schemes.**

***Index Terms*—Distributed algorithms, nonconvex optimization, successive convex approximation.**

I. Introduction

THE minimization of a nonconvex objective function $U : \mathcal{K} \to \mathbb{R}$ subject to convex constraints $\mathcal{K}$ and nonconvex ones $g_j(\mathbf{x}) \leq 0$, with $g_j : \mathcal{K} \to \mathbb{R}$,

$$\left.\begin{array}{ll} \min\limits_{\mathbf{x}} & U(\mathbf{x}) \\ \text{s.t.} & g_j(\mathbf{x}) \leq 0,\ j = 1, \ldots, m \\ & \mathbf{x} \in \mathcal{K} \end{array}\right\} \triangleq \mathcal{X}, \qquad (\mathcal{P})$$

is an ubiquitous problem that arises in many fields, ranging from signal processing to communication, networking, machine learning, etc.

It is hardly possible here to even summarize the huge amount of solution methods that have been proposed for problem $\mathcal{P}$. Our focus in this paper is on *distributed* algorithms converging to stationary solutions of $\mathcal{P}$ while *preserving the feasibility of the iterates*. While the former feature needs no further comment, the latter is motivated by several reasons. First, in many cases the objective function U is not even defined outside the feasible set; second, in some applications one may have to interrupt calculations before a solution has been reached and it is then important that the current iterate is feasible; and third, in on-line implementations it is mandatory that some constraints are satisfied by every iterate (e.g., think of power budget or interference constraints). As far as we are aware of, there exists no method for the solution of $\mathcal{P}$ in its full generality that is both feasible *and* distributed.

Existing efforts pursuing the above design criteria include: 1) Feasible Interior Point (FIP) methods (e.g., [4], [5]), 2) Feasible Sequential Quadratic Programming (FSQP) methods (e.g., [6]); 3) Parallel Variable Distribution (PVD) schemes (e.g., [7]–[9]); 4) SCA algorithms (in the spirit of [10]–[15]); and some specialized algorithms with roots in the structural optimization field (e.g., [16]–[18]). FIP and FSQP methods maintain feasibility throughout the iterations but are centralized and computationally expensive. PVD schemes are suitable for implementation over parallel architectures but they require an amount of information exchange/knowledge that is often not compatible with a distributed architecture (for example they cannot be applied to the case study discussed in Part II of the paper [19]). Furthermore, when applied to problem $\mathcal{P}$, they call for the solution of possibly difficult nonconvex (smaller) subproblems; and convergence has been established only for convex [7], [9] or nonconvex but block separable g_js [8]. Standard SCA methods are centralized [10], [11], [15], with the exception of [13], [14] and some instances of [12] that lead instead to distributed schemes. However, convergence conditions have been established only in the case of *strongly convex* U [11] or *convex* and separable g_js [12]–[14]. Finally, methods developed in the structural engineering field, including [16]–[18], share some similarities with our approach, but in most cases they lack reliable mathematical foundations or do not prove convergence to stationary points of the original problem $\mathcal{P}$. We refer to Sec. III-B for a more detailed discussion on existing works.

Manuscript received January 14, 2016; revised September 25, 2016; accepted October 28, 2016. Date of publication December 7, 2016; date of current version February 7, 2017. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Gonzalo Mateos. The work of G. Scutari was supported in part by the USA National Science Foundation under Grants CIF 1632599, CIF 1564044, and CAREER Award 1555850; and in part by the Office of Naval Research under Grant N00014-16-1-2244. The work of F. Facchinei was partially supported by the MIUR project PLATINO, under Grant PON01_01007. This paper was presented in part at the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, May 2014 and in part at the IEEE ICASSP, Shanghai, China, March 2016.

G. Scutari is with the School of Industrial Engineering and the Cyber Center (Discovery Park), Purdue University, West-Lafayette, IN 47907 USA (e-mail: gscutari@purdue.edu).

F. Facchinei is with the Department of Computer, Control, and Management Engineering, University of Rome "La Sapienza," Rome 00185, Italy (e-mail: facchinei@diag.uniroma1.it).

L. Lampariello is with the Department of Business Studies, University of RomaTre, Roma 00154, Italy (e-mail: lorenzo.lampariello@uniroma3.it).

This paper has supplementary downloadable multimedia material available at http://ieeexplore.ieee.org provided by the authors. This includes a pdf file containing the proof of Theorem 8 in the paper. This material is 162 KB in size.

Digital Object Identifier 10.1109/TSP.2016.2637317

In this paper we propose a new framework for the general formulation $\mathcal{P}$ which, on one hand, maintains feasibility and, on the other hand, leads, under very mild additional assumptions, to parallel and distributed solution methods. The essential, natural idea underlying the proposed approach is to compute a solution of $\mathcal{P}$ by solving a sequence of (simpler) strongly convex subproblems whereby the nonconvex objective function and constraints are replaced by suitable convex approximations; the subproblems can be then solved (under some mild assumptions) in a distributed fashion using standard primal/dual decomposition techniques (e.g., [20], [21]). Additional key features of the proposed method are: i) it includes as special cases several classical SCA-based algorithms, such as (proximal) gradient or Newton type methods, block coordinate (parallel) descent schemes, Difference of Convex (DC) functions approaches, convex-concave approximation methods; ii) our convergence conditions unify and extend to the general class $\mathcal{P}$ those of current (centralized) SCA methods; iii) it offers much flexibility in the choice of the convex approximation functions: for instance, as a major departure from current SCA-based methods [applicable to special cases of $\mathcal{P}$] [10], [12] and DC programs [15], the proposed approximation of the objective function U need not be an *upper bound* of U at any feasible point, a fact that significantly enlarges the range of applicability of our framework; iv) by allowing alternative choices for the convex approximants, it encompasses a gamut of novel algorithms, offering great flexibility to control iteration complexity, communication overhead and convergence speed, and all converging under the *same* conditions; and v) it can deal with some kind of structured nonsmoothness, thus permitting to solve many problems of great practical interest. Quite interestingly, the proposed scheme leads to new efficient algorithms even when customized to solve well-researched problems. Examples include power control problems in cellular systems [22]–[25], MIMO relay optimization [26], dynamic spectrum management in DSL systems [27], [28], sum-rate maximization, proportional-fairness and max-min optimization of SISO/MISO/MIMO ad-hoc networks [13], [29]–[31], robust optimization of CR networks [32]–[34], transmit beamforming design for multiple co-channel multicast groups [35], [36], and cross-layer design of wireless networks [37], [38]. Part II of the paper [19] is devoted to the application of the proposed algorithmic framework to some of the aforementioned problems (and their generalizations). Numerical results show that our schemes compare favorably to existing ad-hoc ones (when they exist).

The rest of this two-part paper is organized as follows. Sec. II introduces the main assumptions underlying the study of the optimization problem $\mathcal{P}$ and provides an informal description of our new algorithms. Sec. III presents our novel framework based on SCA, whereas Sec. IV focuses on its distributed implementation in the primal and dual domain. Sec. V extends the proposed framework to some classes of nonsmooth problems; and, finally, Sec. VI draws some conclusions. In Part II of the paper [19] we apply our algorithmic framework to several resource allocation problems in wireless networks and machine learning and provide extensive numerical results showing that the proposed algorithms compare favorably to state-of-the-art schemes.

II. Technical Preliminaries and Main Idea

In this section we introduce the main assumptions underlying the study of the optimization problem $\mathcal{P}$ along with some technical results that will be instrumental to describe our approach. We also provide an informal description of our new algorithms that sheds light on the core idea of the proposed decomposition technique. The formal description of the framework is given in Sec. III.

Consider problem $\mathcal{P}$, whose feasible set is denoted by $\mathcal{X}$.

Assumption 1: We make the blanket assumptions:

A1) $\mathcal{K} \subseteq \mathbb{R}^n$ is closed and convex (and nonempty);
A2) U and each g_j are continuously differentiable on $\mathcal{K}$;
A3) $\nabla_{\mathbf{x}} U$ is Lipschitz continuous on $\mathcal{K}$ with constant $L_{\nabla U}$.
A4) U is coercive on $\mathcal{K}$.

The assumptions above are quite standard and are satisfied by a large class of problems of practical interest. In particular, A4 guarantees that the problem has a solution, even when the feasible set $\mathcal{X}$ is not bounded. Note that we do not assume convexity of U and $g_1, \ldots, g_m$; without loss of generality, convex constraints, if present, are accommodated in the set $\mathcal{K}$. For the sake of simplicity, we begin developing the algorithmic framework under the differentiability assumption A2; this assumption will be relaxed in Section V.

Our goal is to efficiently compute locally optimal solutions of $\mathcal{P}$, possibly in a distributed way, while preserving the feasibility of the iterates. Building on the idea of SCA methods, our approach consists in solving a sequence of *strongly convex inner* approximations of $\mathcal{P}$ in the form: given $\mathbf{x}^\nu \in \mathcal{X}$

$$\left.\begin{array}{ll} \min\limits_{\mathbf{x}} & \tilde{U}(\mathbf{x};\mathbf{x}^\nu) \\ \text{s.t.} & \tilde{g}_j(\mathbf{x};\mathbf{x}^\nu) \le 0,\ j=1,\ldots,m \\ & \mathbf{x} \in \mathcal{K} \end{array}\right\} \triangleq \mathcal{X}(\mathbf{x}^\nu), \qquad (\mathcal{P}_{\mathbf{x}^\nu})$$

where $\tilde{U}(\mathbf{x};\mathbf{x}^\nu)$ and $\tilde{g}_j(\mathbf{x};\mathbf{x}^\nu)$ represent approximations of $U(\mathbf{x})$ and $g_j(\mathbf{x})$ at the current iterate $\mathbf{x}^\nu$, respectively, and $\mathcal{X}(\mathbf{x}^\nu)$ denotes the feasible set of $\mathcal{P}_{\mathbf{x}^\nu}$.

We introduce next a number of assumptions that will be used throughout the paper.

Assumption 2 (On $\tilde{U}$): Let $\tilde{U}: \mathcal{K} \times \mathcal{X} \to \mathbb{R}$ be a function continuously differentiable with respect to the first argument and such that:

B1) $\tilde{U}(\bullet;\mathbf{y})$ is uniformly strongly convex on $\mathcal{K}$ with constant $c_{\tilde{U}} > 0$, i.e. $\forall \mathbf{x}, \mathbf{z} \in \mathcal{K},\ \forall \mathbf{y} \in \mathcal{X}$

$$(\mathbf{x}-\mathbf{z})^T \left(\nabla_{\mathbf{x}} \tilde{U}(\mathbf{x};\mathbf{y}) - \nabla_{\mathbf{x}} \tilde{U}(\mathbf{z};\mathbf{y})\right) \ge c_{\tilde{U}} \|\mathbf{x}-\mathbf{z}\|^2;$$

B2) $\nabla_{\mathbf{x}} \tilde{U}(\mathbf{y};\mathbf{y}) = \nabla_{\mathbf{x}} U(\mathbf{y})$, for all $\mathbf{y} \in \mathcal{X}$;
B3) $\nabla_{\mathbf{x}} \tilde{U}(\bullet;\bullet)$ is continuous on $\mathcal{K} \times \mathcal{X}$;

where $\nabla_{\mathbf{x}} \tilde{U}(\mathbf{u};\mathbf{w})$ denotes the partial gradient of $\tilde{U}$ with respect to the first argument evaluated at $(\mathbf{u};\mathbf{w})$.

Assumption 3 (On $\tilde{g}_j$s): Let each $\tilde{g}_j: \mathcal{K} \times \mathcal{X} \to \mathbb{R}$ satisfy the following:

C1) $\tilde{g}_j(\bullet;\mathbf{y})$ is convex on $\mathcal{K}$ for all $\mathbf{y} \in \mathcal{X}$;
C2) $\tilde{g}_j(\mathbf{y};\mathbf{y}) = g_j(\mathbf{y})$, for all $\mathbf{y} \in \mathcal{X}$;
C3) $g_j(\mathbf{x}) \le \tilde{g}_j(\mathbf{x};\mathbf{y})$ for all $\mathbf{x} \in \mathcal{K}$ and $\mathbf{y} \in \mathcal{X}$;
C4) $\tilde{g}_j(\bullet;\bullet)$ is continuous on $\mathcal{K} \times \mathcal{X}$;

C5) $\nabla_{\mathbf{x}} g_j(\mathbf{y}) = \nabla_{\mathbf{x}} \tilde{g}_j(\mathbf{y};\mathbf{y})$, for all $\mathbf{y} \in \mathcal{X}$;
C6) $\nabla_{\mathbf{x}} \tilde{g}_j(\bullet;\bullet)$ is continuous on $\mathcal{K} \times \mathcal{X}$;

where $\nabla_{\mathbf{x}} \tilde{g}_j(\mathbf{y};\mathbf{y})$ denotes the (partial) gradient of $\tilde{g}_j$ with respect to the first argument evaluated at $\mathbf{y}$ (the second argument is kept fixed at $\mathbf{y}$).

For some results we need stronger continuity properties of the (gradient of the) approximation functions.

Assumption 4:

B4) $\nabla_{\mathbf{x}} \tilde{U}(\mathbf{x};\bullet)$ is uniformly Lipschitz continuous on $\mathcal{X}$ with constant $\tilde{L}_{\nabla,2}$;
B5) $\nabla_{\mathbf{x}} \tilde{U}(\bullet;\mathbf{y})$ is uniformly Lipschitz continuous on $\mathcal{K}$ with constant $\tilde{L}_{\nabla,1}$;
C7) Each $\tilde{g}_j(\bullet;\bullet)$ is Lipschitz continuous on $\mathcal{K} \times \mathcal{X}$.

The key assumptions are B1, C1, and C3: B1 and C1 make $\mathcal{P}_{\mathbf{x}^\nu}$ strongly convex, whereas C2 and C3 guarantee $\mathbf{x}^\nu \in \mathcal{X}(\mathbf{x}^\nu) \subseteq \mathcal{X}$ (and, thus, nonemptiness of the feasible set of $\mathcal{P}_{\mathbf{x}^\nu}$ and iterate feasibility). The others are technical conditions (easy to be satisfied in practice) ensuring that the approximations have the same local first order behavior of the original functions. In the next section we provide some examples of approximate functions that automatically satisfy Assumptions 2–4. As a final remark, we point out that Assumptions 1–3 are in many ways similar *but generally weaker* than those used in the literature in order to solve special cases of problem $\mathcal{P}$ [10]–[14] (see Sec. III-B for a rather complete discussion on this topic).

Our weaker conditions on the approximations $\tilde{U}$ and $\tilde{g}$ along with a more general setting allow us to deal with a much larger class of problems than [10]–[14]; see Part II of the paper [19] for specific examples.

A. Regularity Conditions

We conclude this section mentioning certain standard regularity conditions on the stationary points of constrained optimization problems. These conditions are needed in the study of the convergence properties of our method.

Definition 1 (Regularity): A point $\bar{\mathbf{x}} \in \mathcal{X}$ is called *regular* for $\mathcal{P}$ if the Mangasarian-Fromovitz Constraint Qualification (MFCQ) holds at $\bar{\mathbf{x}}$, that is (see e.g. [39, Theorem 6.14]) if the following implication is satisfied:

$$\left.\begin{array}{c} \mathbf{0} \in \sum_{j\in\bar{J}} \mu_j \nabla_{\mathbf{x}} g_j(\bar{\mathbf{x}}) + N_{\mathcal{K}}(\bar{\mathbf{x}}) \\ \mu_j \geq 0,\ \forall j \in \bar{J} \end{array}\right\} \Rightarrow \mu_j = 0,\ \forall j \in \bar{J}, \quad (1)$$

where $N_{\mathcal{K}}(\bar{\mathbf{x}}) \triangleq \{\mathbf{d} \in \mathbb{R}^n : \mathbf{d}^T(\mathbf{y} - \bar{\mathbf{x}}) \leq 0,\ \forall \mathbf{y} \in \mathcal{K}\}$ is the normal cone to $\mathcal{K}$ at $\bar{\mathbf{x}}$, and $\bar{J} \triangleq \{j \in \{1,\ldots,m\} : g_j(\bar{\mathbf{x}}) = 0\}$ is the index set of those (nonconvex) constraints that are active at $\bar{\mathbf{x}}$.

A similar definition holds for problem $\mathcal{P}_{\mathbf{x}^\nu}$: a point $\bar{\mathbf{x}} \in \mathcal{X}(\mathbf{x}^\nu)$ is called *regular* for $\mathcal{P}_{\mathbf{x}^\nu}$ if

$$\left.\begin{array}{c} \mathbf{0} \in \sum_{j\in\bar{J}^\nu} \mu_j \nabla_{\mathbf{x}} \tilde{g}_j(\bar{\mathbf{x}};\mathbf{x}^\nu) + N_{\mathcal{K}}(\bar{\mathbf{x}}) \\ \mu_j \geq 0,\ \forall j \in \bar{J}^\nu \end{array}\right\} \Rightarrow \mu_j = 0,\ \forall j \in \bar{J}^\nu, \quad (2)$$

where $\bar{J}^\nu \triangleq \{j \in \{1,\ldots,m\} : \tilde{g}_j(\bar{\mathbf{x}};\mathbf{x}^\nu) = 0\}$. ■

We point out that the regularity of $\bar{\mathbf{x}}$ is implied by stronger but easier to be checked CQs, such as the Linear Independence CQ, see [40, Sec. 3.2] for more details. Note that if the feasible set is convex, as it is in $\mathcal{P}_{\mathbf{x}^\nu}$, the MFCQ can be substituted by the Slater's CQ; for a set like $\mathcal{X}(\mathbf{x}^\nu)$, Slater's CQ reads

$$\mathrm{ri}(\mathcal{K}) \cap \mathcal{X}_g^<(\mathbf{x}^\nu) \neq \emptyset,$$

where $\mathcal{X}_g^<(\mathbf{x}^\nu) \triangleq \{\mathbf{x} \in \mathcal{K} : \tilde{g}_j(\mathbf{x};\mathbf{x}^\nu) < 0,\ j = 1,\ldots,m\}$ and $\mathrm{ri}(\mathcal{K})$ is the relative interior of $\mathcal{K}$ (see, e.g., [41, Sec. 1.4]). In particular, this means that for problem $\mathcal{P}_{\mathbf{x}^\nu}$ either the MFCQ holds at all the feasible points or it does not hold at any point. Furthermore, because of C2 and C5, a point $\bar{\mathbf{x}}$ is regular for $\mathcal{P}$ if and only if $\bar{\mathbf{x}}$ is regular for $\mathcal{P}_{\bar{\mathbf{x}}}$ (and, therefore, if any feasible point of $\mathcal{P}_{\bar{\mathbf{x}}}$ is regular).

Definition 2 (Stationary solutions): A point $\bar{\mathbf{x}} \in \mathcal{X}$ is a stationary point of problem $\mathcal{P}$ if it satisfies the following KKT system:

$$\mathbf{0} \in \nabla_{\mathbf{x}} U(\bar{\mathbf{x}}) + \sum\nolimits_{j\in\bar{J}} \mu_j \nabla_{\mathbf{x}} g_j(\bar{\mathbf{x}}) + N_{\mathcal{K}}(\bar{\mathbf{x}})$$
$$\mu_j \geq 0,\ \forall j \in \bar{J}$$

for some suitable Lagrange multipliers μ_js.

It is well-known that a regular (local) minimum point of problem $\mathcal{P}$ is also stationary. Finding stationary points is actually the classical goal of solution algorithms for nonconvex problems.

In order to simplify the presentation, in the rest of this paper we assume the following regularity condition.

Assumption 5: All feasible points of problem $\mathcal{P}$ are regular.

One could relax this assumption and require regularity only at specific points, but at the cost of more convoluted statements; we leave this task to the reader. We remark, once again, that Assumption 5 implies that any feasible point of $\mathcal{P}_{\bar{\mathbf{x}}}$ is regular.

Algorithm 1: NOVA Algorithm for $\mathcal{P}$.

Data: $\gamma^\nu \in (0,1]$, $\mathbf{x}^0 \in \mathcal{X}$; set $\nu = 0$.
(S.1) If $\mathbf{x}^\nu$ is a stationary solution of $\mathcal{P}$: STOP.
(S.2) Compute $\hat{\mathbf{x}}(\mathbf{x}^\nu)$, the solution of $\mathcal{P}_{x^\nu}$ [cf. (3)].
(S.3) Set $\mathbf{x}^{\nu+1} = \mathbf{x}^\nu + \gamma^\nu(\hat{\mathbf{x}}(\mathbf{x}^\nu) - \mathbf{x}^\nu)$.
(S.4) $\nu \leftarrow \nu + 1$ and go to step (S.1).

III. Algorithmic Framework

We are now ready to formally introduce the proposed solution method for $\mathcal{P}$. Note first that, because of B1 and C1, each subproblem $\mathcal{P}_{\mathbf{x}^\nu}$ is strongly convex and thus has a unique solution, which is denoted by $\hat{\mathbf{x}}(\mathbf{x}^\nu)$ (a function of $\mathbf{x}^\nu$):

$$\hat{\mathbf{x}}(\mathbf{x}^\nu) \triangleq \operatorname*{argmin}_{\mathbf{x}\in\mathcal{X}(\mathbf{x}^\nu)} \tilde{U}(\mathbf{x};\mathbf{x}^\nu). \quad (3)$$

Starting from a feasible point $\mathbf{x}^0$, the proposed method consists in iteratively computing the solution $\hat{\mathbf{x}}(\mathbf{x}^\nu)$ to the surrogate problem (3), and then taking a step from $\mathbf{x}^\nu$ towards $\hat{\mathbf{x}}(\mathbf{x}^\nu)$; we named the new method iNner cOnVex Approximation (NOVA) algorithm. The formal description of the NOVA algorithm along with its convergence properties are given in Algorithm 1 and Theorem 2, respectively.

Theorem 2: Given the nonconvex problem $\mathcal{P}$ under Assumptions 1–3 and 5, let $\{\mathbf{x}^\nu\}$ be the sequence generated by Algorithm 1. The following hold.

(a) $\mathbf{x}^\nu \in \mathcal{X}(\mathbf{x}^\nu) \subseteq \mathcal{X}$ for all $\nu \geq 0$ (iterate feasibility);
(b) If the step-size γ^ν and $c_{\tilde{U}}$ are chosen so that

$$0 < \inf_\nu \gamma^\nu \leq \sup_\nu \gamma^\nu \leq \gamma^{\max} \leq 1 \text{ and } 2c_{\tilde{U}} > \gamma^{\max} L_{\nabla U}, \tag{4}$$

then $\{\mathbf{x}^\nu\}$ is bounded and each of its limit points is a stationary point of problem $\mathcal{P}$.
(c) If the step-size γ^ν is chosen so that

$$\gamma^\nu \in (0,1], \quad \gamma^\nu \to 0, \quad \text{and} \quad \sum_\nu \gamma^\nu = +\infty, \tag{5}$$

then $\{\mathbf{x}^\nu\}$ is bounded and at least one of its limit points is stationary. If, in addition, Assumption 4 holds and $\mathcal{X}$ is compact, every limit point of $\{\mathbf{x}^\nu\}$ is stationary.

Furthermore, if the algorithm does not stop after a finite number of steps, none of the stationary points in parts (b) and (c) is a local maximum of U.

Proof: See Appendix. ■

A. Discussions on Algorithm 1

Algorithm 1 describes a novel family of inner convex approximation methods for problem $\mathcal{P}$. Roughly speaking, it consists in solving the sequence of strongly convex problems $\mathcal{P}_{\mathbf{x}^\nu}$ wherein the original objective function U is replaced by the strongly convex (simple) approximation $\tilde{U}$, and the nonconvex constraints g_js with the convex upper estimates $\tilde{g}_j$ s; convex constraints, if any, are kept unaltered. A step-size in the update of the iterates $\mathbf{x}^\nu$ is also used, in the form of a convex combination via $\gamma^\nu \in (0,1]$ (cf. Step 3). Note that the iterates $\{\mathbf{x}^\nu\}$ generated by the algorithm are all feasible for the original problem $\mathcal{P}$. Convergence is guaranteed under mild assumptions that offer a lot of flexibility in the choice of the approximation functions and free parameters [cf. Theorem 2(b) and (c)], making the proposed scheme appealing for many applications. We provide next some examples of candidate approximants, covering a variety of situations and problems of practical interest.

1) On the Approximations $\tilde{g}_j$s: As already mentioned, while assumption C3 might look rather elusive, in many practical cases an upper approximate function for the nonconvex constraints g_js is close at hand. Some examples of $\tilde{g}_j$ satisfying Assumption 3 (and in particular C3) are given next; specific applications where such approximations are used are discussed in detail in Part II of the paper [19].

Example #1– Nonconvex constraints with Lipschitz gradients: If the nonconvex function g_j does not have a special structure but Lipschitz continuous gradient on $\mathcal{K}$ with constant $L_{\nabla g_j}$, by the Descent Lemma, see e.g. [21, Proposition A32], the following convex approximation function is a global upper bound of g_j: for all $\mathbf{x} \in \mathcal{K}$ and $\mathbf{y} \in \mathcal{X}$,

$$\tilde{g}_j(\mathbf{x};\mathbf{y}) \triangleq g_j(\mathbf{y}) + \nabla_{\mathbf{x}} g_j(\mathbf{y})^T(\mathbf{x}-\mathbf{y}) + \frac{L_{\nabla g_j}}{2}\|\mathbf{x}-\mathbf{y}\|^2 \geq g_j(\mathbf{x}). \tag{6}$$

Example #2– Nonconvex constraints with (uniformly) bounded Hessian matrix: Suppose that g_j is (nonconvex) $\mathcal{C}^2$ with second order bounded derivatives on $\mathcal{K}$. Then, one can find a matrix $\mathbf{G} \succ \mathbf{0}$ such that $\nabla^2_{\mathbf{x}} g_j(\mathbf{x}) + \mathbf{G} \succeq 0$ for all $\mathbf{x} \in \mathcal{K}$. For instance, one can set $\mathbf{G} = \left|\min_{\mathbf{x}\in\mathcal{K}} \lambda_{\min}(\nabla^2_{\mathbf{x}} g_j(\mathbf{x}))\right| \cdot \mathbf{I}$, with $\lambda_{\min}(\nabla^2_{\mathbf{x}} g_j(\mathbf{x}))$ denoting the minimum eigenvalue of $\nabla^2_{\mathbf{x}} g_j(\mathbf{x})$ (which is a negative quantity if g_j is nonconvex). Then, the unstructured nonconvex constraint g_j can be equivalently written as a DC function:

$$g_j(\mathbf{x}) = \underbrace{g_j(\mathbf{x}) + \frac{1}{2}\mathbf{x}^T\mathbf{G}\mathbf{x}}_{\triangleq g_j^+(\mathbf{x})} - \underbrace{\frac{1}{2}\mathbf{x}^T\mathbf{G}\mathbf{x}}_{\triangleq g_j^-(\mathbf{x})}, \tag{7}$$

where g_j^+ and g_j^- aretwo *convex*continuously differentiable functions. An approximant $\tilde{g}_j$ of g_j satisfying Assumption 3 can then readily be obtained by linearizing $g_j^-(\mathbf{x})$; see Example #3 below for details.

The two examples above cover successfully quite general *unstructured* functions g_j. However, in some cases, the function parameters involved in the approximations– the constants $L_{\nabla g_j}$ or $\left|\min_{\mathbf{x}\in\mathcal{K}} \lambda_{\min}(\nabla^2_{\mathbf{x}} g_j(\mathbf{x}))\right|$ – are not known exactly but need to be estimated; if the estimates are not tight, the resulting $\tilde{g}_j$ might be a loose overestimation of g_j, which may negatively affect the practical convergence of Algorithm 1. Other approximations can be obtained when g_j has further structure to exploit, as discussed in the next examples.

Example #3– Nonconvex constraints with DC structure: Suppose that g_j has a DC structure, that is,

$$g_j(\mathbf{x}) = g_j^+(\mathbf{x}) - g_j^-(\mathbf{x})$$

is the difference of two convex and continuously differentiable functions g_j^+ and g_j^-. By linearizing the concave part $-g_j^-$ and keeping the convex part g_j^+ unchanged, we obtain the following convex upper approximation of g_j: for all $\mathbf{x} \in \mathcal{K}$ and $\mathbf{y} \in \mathcal{X}$,

$$\tilde{g}_j(\mathbf{x};\mathbf{y}) \triangleq g_j^+(\mathbf{x}) - g_j^-(\mathbf{y}) - \nabla_{\mathbf{x}} g_j^-(\mathbf{y})^T(\mathbf{x}-\mathbf{y}) \geq g_j(\mathbf{x}). \tag{8}$$

Example #4– Bi-linear constraints: Suppose that g_j has a bilinear structure, that is,

$$g_j(x_1,x_2) = x_1 \cdot x_2. \tag{9}$$

Observe preliminarily that $g_j(x_1,x_2)$ can be rewritten as a DC function:

$$g_j(x_1,x_2) = \frac{1}{2}(x_1+x_2)^2 - \frac{1}{2}(x_1^2+x_2^2). \tag{10}$$

A valid $\tilde{g}_j$ can be then obtained linearizing the concave part in (10): for any given $(y_1,y_2) \in \mathbb{R}^2$,

$$\tilde{g}_j(x_1,x_2;\,y_1,y_2) \triangleq \frac{1}{2}(x_1+x_2)^2 - \frac{1}{2}(y_1^2+y_2^2) - y_1\cdot(x_1-y_1) - y_2\cdot(x_2-y_2).$$

In Part II of the paper [19] we show that the constraint functions of many resource allocation problems in wireless systems and networking fit naturally in Examples 1-4 above.

2) On the Approximation $\tilde{U}$: The function $\tilde{U}$ should be regarded as a (possibly simple) convex approximation that preserves the first order properties of U. Some instances of valid $\tilde{U}$s for a specific U occurring in practical applications are discussed next.

Example #5– Block-wise convex $U(\mathbf{x}_1,\ldots,\mathbf{x}_n)$: In many applications, the vector of variables $\mathbf{x}$ is partitioned in blocks

$\mathbf{x}=(\mathbf{x}_i)_{i=1}^I$ and the function U is convex in each block $\mathbf{x}_i$ separately, but not jointly. A natural approximation for such a U exploring its "partial" convexity is

$$\tilde{U}(\mathbf{x};\mathbf{y})=\sum_{i=1}^{I}\tilde{U}_i(\mathbf{x}_i;\mathbf{y}), \tag{11}$$

with each $\tilde{U}_i(\mathbf{x}_i;\mathbf{y})$ defined as

$$\tilde{U}_i(\mathbf{x}_i;\mathbf{y})\triangleq U(\mathbf{x}_i,\mathbf{y}_{-i})+\frac{\tau_i}{2}(\mathbf{x}_i-\mathbf{y}_i)^T\mathbf{H}_i(\mathbf{y})(\mathbf{x}_i-\mathbf{y}_i), \tag{12}$$

where $\mathbf{y}\triangleq(\mathbf{y}_i)_{i=1}^I$, $\mathbf{y}_{-i}\triangleq(\mathbf{y}_j)_{j\neq i}$, and $\mathbf{H}_i(\mathbf{y})$ is any uniformly positive definite matrix (possibly depending on $\mathbf{y}$). Note that the quadratic term in (12) can be set to zero if $U(\mathbf{x}_i,\mathbf{y}_{-i})$ is strongly convex in $\mathbf{x}_i$, uniformly for all feasible $\mathbf{y}_{-i}$. An alternative choice for $\tilde{U}_i(\mathbf{x}_i;\mathbf{y})$ is

$$\tilde{U}_i(\mathbf{x}_i;\mathbf{y})\triangleq\nabla_{\mathbf{x}_i}U(\mathbf{y})^T(\mathbf{x}_i-\mathbf{y}_i)+\frac{1}{2}(\mathbf{x}_i-\mathbf{y}_i)^T\nabla^2_{\mathbf{x}_i}U(\mathbf{y})(\mathbf{x}_i-\mathbf{y}_i)+\frac{\tau_i}{2}\|\mathbf{x}_i-\mathbf{y}_i\|^2,$$

where $\nabla^2_{\mathbf{x}_i}U(\mathbf{y})$ is the Hessian of U w.r.t. $\mathbf{x}_i$ evaluated in $\mathbf{y}$. One can also use any positive definite "approximation" of $\nabla^2_{\mathbf{x}_i}U(\mathbf{y})$. Needless to say, if $U(\mathbf{x}_1,\ldots,\mathbf{x}_n)$ is *jointly* convex in all the variables' blocks, then $\tilde{U}(\mathbf{x};\mathbf{y})$ can be chosen so that

$$\tilde{U}(\mathbf{x};\mathbf{y})\triangleq U(\mathbf{x})+\sum_i\frac{\tau_i}{2}\|\mathbf{x}_i-\mathbf{y}_i\|^2, \tag{13}$$

where $\frac{\tau_i}{2}\|\mathbf{x}_i-\mathbf{y}_i\|^2$ is not needed if $U(\mathbf{x}_i,\,\mathbf{x}_{-i})$ is strongly convex in $\mathbf{x}_i$, uniformly for all feasible $\mathbf{x}_{-i}$.

Example #6– (Proximal) gradient-like approximations: If no convexity whatsoever is present in U, mimicking proximal-gradient methods, a valid choice of $\tilde{U}$ is the first order approximation of U, that is, $\tilde{U}(\mathbf{x};\mathbf{y})=\sum_{i=1}^I\tilde{U}_i(\mathbf{x}_i;\mathbf{y})$, with each

$$\tilde{U}_i(\mathbf{x}_i;\mathbf{y})\triangleq\nabla_{\mathbf{x}_i}U(\mathbf{y})^T(\mathbf{x}_i-\mathbf{y}_i)+\frac{\tau_i}{2}\|\mathbf{x}_i-\mathbf{y}_i\|^2.$$

Note that even though classical (proximal) gradient descent methods (see, e.g., [21]) share the same approximation function, they are not applicable to problem $\mathcal{P}$, due to the nonconvexity of the feasible set.

Example #7– Sum-utility function: In multi-agent scenarios, the objective function U is generally written as $U(\mathbf{x})\triangleq\sum_{i=1}^I f_i(\mathbf{x}_1,\ldots,\mathbf{x}_I)$, that is, the sum of the utilities $f_i(\mathbf{x}_1,\ldots,\mathbf{x}_I)$ of I agents, each controlling the variables $\mathbf{x}_i$. A typical situation is when the f_is are convex in some agents' variables. To capture this property, let us define by

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

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 let $\mathcal{C}_i\subseteq\mathcal{S}_i$ be any subset of $\mathcal{S}_i$. Then, the following approximation function $\tilde{U}(\mathbf{x};\mathbf{y})$ satisfies Assumption 2 while exploiting the partial convexity of U (if any): $\tilde{U}(\mathbf{x};\mathbf{y})=\sum_{i=1}^I\tilde{U}_{\mathcal{C}_i}(\mathbf{x}_i;\mathbf{y})$, with each $\tilde{U}_{\mathcal{C}_i}$ defined as

$$\tilde{U}_{\mathcal{C}_i}(\mathbf{x}_i;\mathbf{y})\triangleq\sum_{j\in\mathcal{C}_i}f_j(\mathbf{x}_i,\mathbf{y}_{-i})+\sum_{k\notin\mathcal{C}_i}\nabla_{\mathbf{x}_i}f_k(\mathbf{y})^T(\mathbf{x}_i-\mathbf{y}_i)+\frac{\tau_i}{2}(\mathbf{x}_i-\mathbf{y}_i)^T\mathbf{H}_i(\mathbf{y})(\mathbf{x}_i-\mathbf{y}_i),$$

where $\mathbf{H}_i(\mathbf{y})$ is any uniformly positive definite matrix. Roughly speaking, for each agent i we built an approximation function such that the convex part of U w.r.t. $\mathbf{x}_i$ may be preserved while the nonconvex part is linearized.

Example #8– Product of functions: The function U is often the product of functions (see Part II [19] for some examples); we consider here the product of two functions, but the proposed approach can be readily extended to the case of three or more functions or to the sum of such product terms. Suppose that $U(\mathbf{x})=f_1(\mathbf{x})f_2(\mathbf{x})$, with f_1 and f_2 convex and positive. In view of the expression of the gradient of U, $\nabla_{\mathbf{x}}U=f_2\nabla_{\mathbf{x}}f_1+f_1\nabla_{\mathbf{x}}f_2$, it seems natural to consider the approximation

$$\tilde{U}(\mathbf{x};\mathbf{y})=f_1(\mathbf{x})f_2(\mathbf{y})+f_1(\mathbf{y})f_2(\mathbf{x})+\frac{\tau_i}{2}(\mathbf{x}-\mathbf{y})^T\mathbf{H}(\mathbf{y})(\mathbf{x}-\mathbf{y}),$$

where, as usual, $\mathbf{H}(\mathbf{y})$ is a uniformly positive definite matrix; this term can be omitted if f_1 and f_2 are bounded away from zero on the feasible set and f_1+f_2 is strongly convex (for example if one of the two functions is strongly convex). It is clear that this $\tilde{U}$ satisfies Assumption 2. In case f_1 and f_2 are still positive but not necessarily convex, we can use the expression

$$\tilde{U}(\mathbf{x};\mathbf{y})=\tilde{f}_1(\mathbf{x};\mathbf{y})f_2(\mathbf{y})+f_1(\mathbf{y})\tilde{f}_2(\mathbf{x};\mathbf{y}),$$

where $\tilde{f}_1$ and $\tilde{f}_2$ are any legitimate approximations for f_1 and f_2, for example those considered in Examples 5-7 above. Finally, if f_1 and f_2 can take non-positive values, we can write

$$\tilde{U}(\mathbf{x};\mathbf{y})=\tilde{h}_1(\mathbf{x},\mathbf{y})+\tilde{h}_2(\mathbf{x},\mathbf{y}),$$

where $h_1(\mathbf{x},\mathbf{y})\triangleq\tilde{f}_1(\mathbf{x};\mathbf{y})f_2(\mathbf{y})$, $h_2(\mathbf{x},\mathbf{y})\triangleq f_1(\mathbf{y})\tilde{f}_2(\mathbf{x};\mathbf{y})$, and $\tilde{h}_1$ and $\tilde{h}_2$ are legitimate approximations for h_1 and h_2, for example, again, those considered in Examples 5-7. Note that in the last cases we no longer need the quadratic term because it is already included in the approximations $\tilde{f}_1$ and $\tilde{f}_2$, and $\tilde{h}_1$ and $\tilde{h}_2$, respectively. As a final remark, it is important to point out that the Us discussed above belong to a class of nonconvex functions for which it does not seem possible to find a global convex upper bound; therefore, all current SCA techniques (see, e.g., [10], [12], [15]) are not applicable.

Example #9– Composition of functions: Let $U(\mathbf{x})=h(\mathbf{f}(\mathbf{x}))$, where $h:\mathbb{R}^m\to\mathbb{R}$ is a finite convex smooth function such that $h(u_1,\ldots,u_m)$ is nondecreasing in each u_j, and $\mathbf{f}:\mathbb{R}^n\to\mathbb{R}^m$ is a smooth mapping, with $\mathbf{f}(\mathbf{x})=(f_1(\mathbf{x}),\ldots,f_m(\mathbf{x}))^T$ and f_i not necessarily convex. A widely class of functions $U(\mathbf{x})$ belonging to such a class are those arising from nonlinear least square-based problems, that is, $U(\mathbf{x})=\|\mathbf{f}(\mathbf{x})\|^2$, where $\mathbf{f}(\mathbf{x})$ is a smooth nonlinear (possibly) nonconvex map. A convex approximation satisfying Assumption 2 is

$$\tilde{U}(\mathbf{x};\mathbf{y})\triangleq h(\mathbf{f}(\mathbf{y})+\nabla\mathbf{f}(\mathbf{y})(\mathbf{x}-\mathbf{y}))+\frac{\tau}{2}\|\mathbf{x}-\mathbf{y}\|^2, \tag{14}$$

where $\nabla\mathbf{f}(\mathbf{y})$ is the Jacobian of $\mathbf{f}$ at $\mathbf{y}$.

We conclude the discussion on the approximation functions observing that, in Examples 5-7, the proposed $\tilde{U}(\mathbf{x};\mathbf{y})$s are all separable in the blocks $\mathbf{x}_i$ for any given $\mathbf{y}$; in Example 8, instead, the separability is problem dependent and should be examined on a case-by-case basis. The separability of the $\tilde{U}$s paves the

way to parallelizable and distributed versions of Algorithm 1; we discuss this topic more in detail in Sec. IV.

3) On the Choice of the Step-size Rule γ^ν: Theorem 2 states that Algorithm 1 converges either employing a constant step-size rule [cf. (4)] or a diminishing step-size rule [cf. (5)].

If a constant step-size is used, one can set in (4) $\gamma^\nu = \gamma \leq \gamma^{\max}$ for every ν, and choose any $\gamma^{\max} \in (0, 1]$ and $c_{\tilde{U}}$ so that $2c_{\tilde{U}} > \gamma^{\max} L_{\nabla U}$ (recall that $c_{\tilde{U}}$ is the constant of strong convexity of the approximation $\tilde{U}$ and, thus, is a degree of freedom). This can be done in several ways. For instance, if the chosen $\tilde{U}$ contains a proximal term with gain $\tau > 0$, i.e., a term of the type $(\tau/2)\|\mathbf{x} - \mathbf{y}\|^2$, then the inequality $2c_{\tilde{U}} > \gamma^{\max} L_{\nabla U}$ is readily satisfied setting $2\tau/\gamma^{\max} > L_{\nabla U}$ (we used $c_{\tilde{U}} \geq \tau$). Note that this simple (but conservative) condition imposes a constraint only on the ratio $\tau/\gamma^{\max}$, leaving free the choice of one of the two parameters. An interesting special case worth mentioning is when $\gamma^{\max} = 1$ and $2\tau > L_{\nabla U}$: the choice $\gamma^\nu = 1$ leads to an instance of Algorithm 1 with no memory, i.e., $\mathbf{x}^{\nu+1} = \hat{\mathbf{x}}(\mathbf{x}^\nu)$, for all ν.

When the Lipschitz constant $L_{\nabla U}$ cannot be estimated, one can use a diminishing step-size rule, satisfying the standard conditions (5). A rule that we found to work well in practice is, see [13]:

$$\gamma^\nu = \gamma^{\nu-1}(1 - \varepsilon\gamma^{\nu-1}), \quad \nu \geq 1, \tag{15}$$

with $\gamma^0 \in (0, 1]$ and $\varepsilon \in (0, 1)$. Other effective rules can be found in [13]. Notice that, while this rule may still require some tuning for optimal behavior, it is quite reliable, since in general we are not using a (sub)gradient direction, so that many of the well-known practical drawbacks associated with a (sub)gradient method with diminishing step-size are mitigated in our setting. Furthermore, this choice of step-size does not require any form of centralized coordination and, thus, provides a favorable feature in distributed environments.

We remark that it is possible to prove the convergence of Algorithm 1 also using other step-size rules, such as a standard Armijo-like line-search procedure. We omit the discussion of line-search based approaches because such options are not in line with our goal of developing distributed algorithms, see Sec. IV. In [11] it is shown that, in the specific case of a *strongly convex* U and, in our terminology, $\tilde{U} = U$ and $\mathcal{K} = \mathbb{R}^n$, by choosing $\gamma^\nu = 1$ at every iteration, one can prove the stationarity of every limit point of the sequence generated by Algorithm 1 (assuming regularity). We can easily derive this particular result from our general analysis, see Remark 15 in the Appendix. Here we only mention that, attractive as this result may be, the strong convexity of U is a very restrictive assumption, and forcing $\tilde{U} = U$ does not permit the development of distributed versions of Algorithm 1.

4) On Inexact Solutions: The main computational burden of Algorithm 1 is the calculation of $\hat{\mathbf{x}}(\mathbf{x}^\nu)$, i.e., the solution of $\mathcal{P}_{\mathbf{x}^\nu}$, in Step S.2. In some cases, see Part II [19], it is possible to obtain a closed-formula for $\hat{\mathbf{x}}(\mathbf{x}^\nu)$, but in general one must resort to iterative solvers that provide approximate solutions $\mathbf{z}^\nu$ only, with $\|\mathbf{z}^\nu - \hat{\mathbf{x}}(\mathbf{x}^\nu)\| \leq \varepsilon^\nu$. It is classical to study the behavior of schemes like Algorithm 1 also when using such approximate solutions. One can show that, under the standard assumption that ε^ν goes to zero "sufficiently fast", all results in Theorem 2 still hold for the variant of Algorithm 1 wherein inaccurate solutions are used in Step S.2. The developments are rather lengthy and tedious and we therefore omit them and refer the reader to similar results in [13] and [42].

5) On the Termination Criterion in Step S.1: Stationarity or, more practically, nearness to stationarity must be checked in Step S.1. This can be done in many classical ways; in our setting, it is very convenient to use $\|\hat{\mathbf{x}}(\mathbf{x}^\nu) - \mathbf{x}^\nu\|$. In fact, it can be shown that this quantity is a measure of stationarity: $\|\hat{\mathbf{x}}(\mathbf{x}) - \mathbf{x}\|$ is a continuous function that is zero if and only if $\mathbf{x}$ is a stationary point. A suitable termination criterion in Step S.1 is then $\|\hat{\mathbf{x}}(\mathbf{x}^\nu) - \mathbf{x}^\nu\| \leq \epsilon$, where ϵ is the desired accuracy.

Finally, as a technical note, it is interesting to contrast the different kinds of convergence that one can obtain by choosing a constant or a diminishing step-size rule. In the former case, every limit point of the sequence generated by Algorithm 1 is guaranteed to be a stationary solution of the original nonconvex problem $\mathcal{P}$, whereas, in the latter case, there exists at least a limit point being stationary, which thus is a weaker condition. On the other hand, a diminishing step-size has been observed to be numerically more efficient than a constant one. In order to obtain, also with a diminishing step-size rule, the strong convergence behavior that can be guaranteed when a constant step-size is employed, one needs extra conditions on the approximation functions $\tilde{U}$ and $\tilde{g}$ (cf. Assumptions 4); these conditions are quite standard and easy to be satisfied in most practical applications (as those studied in Part II [19]).

B. Related Works

Our approach draws on the SCA paradigm, which has been widely explored in the literature, see [10]–[15]. However, our framework and convergence conditions unify and extend current SCA methods in several directions. For instance, preliminarily, we observe that [12]–[14] address the simpler case of convex constraints. Furthermore, the following, more detailed observations are in order.

−On the approximation functions: Conditions on the approximation function $\tilde{U}$ as in Assumption 2 are relatively weak: this feature allows us to enlarge significantly the class of utility functions Us that can be successfully handled by Algorithm 1. A key difference with current SCA methods [applicable to special cases of $\mathcal{P}$] [10], [12] and DC programs [15] is that the approximation $\tilde{U}(\mathbf{x}; \mathbf{y})$ need not be an *upper bound* of U for every $\mathbf{x} \in \mathcal{K}$ and $\mathbf{y} \in \mathcal{X}$ [cf. Assumption 2]. This fact represents a big step forward in the literature of SCA methods; Part II of the paper [19] provides a solid evidence of the wide range of applicability of the proposed framework.

−Convergence conditions: There are only a few SCA-based methods in the literature handling nonconvex constraints, namely [10], [11], [15], and the existing convergence results are quite weak. In particular, the seminal paper [10] requires the objective function to be convex. In principle, one could think to

add an extra variable and use the "epigraph trick" to move the nonconvex objective function among the constraints. However, not only does this contrast with the compactness assumption on the feasible set required in [10], but one should also remark that [10, Th. 1] states that if *the whole sequence converges*, then the algorithm converges to a stationary point; in general, it is hardly possible to show that the sequence generated by the algorithms does converge. In [11], (subsequence) convergence to regular points is proved, but only for nonconvex problems with *strongly convex* objective functions; this fact restricts considerably the range of applicability of this result (for instance, none of the problems that we study in Part II [19] have strongly convex objective functions) and precludes the possibility to use the "epigraph trick". Finally, [15] can handle only (possibly nonsmooth) nonconvex problems whose objective functions and constraints have a DC form. To the best of our knowledge, this work is the first attempt towards the generalization of SCA methods to nonconvex problems having general nonconvex objective functions and constraints.

$-$*Distributed implementation:* A second key and unique feature of Algorithm 1, missing in current SCA schemes [10], [11], [15], is that it easily leads to distributed implementations, as we will discuss in Sec. IV. This feature, along with the feasibility of the iterates, represents a key difference also with classical techniques [6]–[9] that have been proposed in the literature to deal with nonconvex optimization problems.

IV. Distributed Implementation

In many applications, e.g., multi-agent optimization or distributed networking, it is desirable to keep users coordination and communication overhead at minimum level. In this section we discuss distributed versions of Algorithm 1. Of course, we need to assume that problem $\mathcal{P}$ has some suitable structure, and that consistent choices on $\tilde{U}$ and $\tilde{g}$ are made. Therefore, in this section we consider the following additional assumptions.

Assumption 6 (Decomposabilty): Given $\mathcal{P}$, suppose that:

- D1) the set $\mathcal{K}$ has a Cartesian structure, i.e., $\mathcal{K}=\mathcal{K}_1\times\cdots\times\mathcal{K}_I$, with each $\mathcal{K}_i\subset\mathbb{R}^{n_i}$, and $\sum_i n_i=n$; $\mathbf{x}\triangleq(\mathbf{x}_i)_{i=1}^I$ is partitioned accordingly, with each $\mathbf{x}_i\in\mathcal{K}_i$;
- D2) the approximate function $\tilde{U}(\mathbf{x};\mathbf{y})$ satisfying Assumption 2 is chosen so that $\tilde{U}(\mathbf{x};\mathbf{y})=\sum_i\tilde{U}_i(\mathbf{x}_i;\mathbf{y})$;
- D3) each approximate function $\tilde{g}_j(\mathbf{x};\mathbf{y})$ satisfying Assumption 3 is (block) separable in the $\mathbf{x}$-variables, for any given $\mathbf{y}$, that is, $\tilde{g}_j(\mathbf{x};\mathbf{y})=\sum_i\tilde{g}_j^i(\mathbf{x}_i;\mathbf{y})$, for some $\tilde{g}_j^i:\mathcal{K}_i\times\mathcal{X}\to\mathbb{R}$.

Condition D1 is a very natural assumption on problem $\mathcal{P}$ and is usually satisfied when a distributed implementation is called for. If problem $\mathcal{P}$ does not satisfy this assumption, it is not realistic to expect that efficient distributed solution methods can be devised; D2 and D3, instead, are assumptions on our algorithmic choices. In particular, condition D2 permits many choices for $\tilde{U}$. For instance, as already discussed at the end of the subsection "On the approximation $\tilde{U}$", essentially all $\tilde{U}$s introduced in Examples 5-7 (and possibly some of the $\tilde{U}$s in Example 8) satisfy D2. The critical condition in Assumption 6 is D3. Some examples of constraints functions g_j for which one can find a $\tilde{g}_j(\mathbf{x};\mathbf{y})$ satisfying D3 are:

$-$*Individual nonconvex constraints:* Each g_j (still nonconvex) depends only on one of the block variables $\mathbf{x}_1,\ldots,\mathbf{x}_I$, i.e, $g_j(\mathbf{x})=g_j^i(\mathbf{x}_i)$, for some $g_j^i:\mathcal{K}_i\to\mathbb{R}$ and i;

$-$*Separable nonconvex constraints:* Each g_j has the form $g_j(\mathbf{x})=\sum_i g_j^i(\mathbf{x}_i)$, with $g_j^i:\mathcal{K}_i\to\mathbb{R}$;

$-$*Nonconvex constraints with Lipschitz gradients:* Each g_j is not necessarily separable but has Lipschitz gradient on $\mathcal{K}$. In this case one can choose, e.g., the approximation $\tilde{g}_j$ as in (6).

It is important to remark that, even for problems $\mathcal{P}$ [or $\mathcal{P}_{\mathbf{x}^\nu}$] for which it looks hard to satisfy D3, the introduction of proper slack variables can help to decouple the constraint functions, making thus possible to find a $\tilde{g}_j$ that satisfies D3; we refer the reader to Part II of the paper [19] for some non trivial examples where this technique is applied.

For notational simplicity, let us introduce the vector function $\tilde{\mathbf{g}}^i(\mathbf{x}_i;\mathbf{x}^\nu)\triangleq(\tilde{g}_j^i(\mathbf{x}_i;\mathbf{x}^\nu))_{j=1}^m$, for $i=1,\ldots,I$. Under Assumption 6, each subproblem $\mathcal{P}_{\mathbf{x}^\nu}$ becomes

$$\begin{aligned}&\min_{\mathbf{x}}\ \sum_{i=1}^I\tilde{U}_i(\mathbf{x}_i;\mathbf{x}^\nu)\\ &\text{s.t. }\ \tilde{\mathbf{g}}(\mathbf{x};\mathbf{x}^\nu)\triangleq\sum_i\tilde{\mathbf{g}}^i(\mathbf{x}_i;\mathbf{x}^\nu)\le\mathbf{0},\qquad(\tilde{\mathcal{P}}_{\mathbf{x}^\nu})\\ &\quad\mathbf{x}_i\in\mathcal{K}_i,\quad i=1,\ldots,I.\end{aligned}$$

With a slight abuse of notation, we will still denote the feasible set of $\mathcal{P}_{\mathbf{x}^\nu}$ by $\mathcal{X}(\mathbf{x}^\nu)$.

The block separable structure of both the objective function and the constraints lends itself to a parallel decomposition of the subproblems $\mathcal{P}_{\mathbf{x}^\nu}$ in the primal or dual domain: hence, it allows the distributed implementation of Step 2 of Algorithm 1. In the next section we briefly show how to customize standard primal/dual decomposition techniques (see, e.g., [20], [21]) in order to solve subproblem $\mathcal{P}_{\mathbf{x}^\nu}$. We conclude this section observing that, if there are only individual constraints in $\mathcal{P}$, given $\mathbf{x}^\nu$, each $\mathcal{P}_{\mathbf{x}^\nu}$ can be split in I independent subproblems in the variables $\mathbf{x}_i$, even if the original nonconvex U is *not separable,* thus leading to a totally parallel implementation of Algorithm 1. To the best of our knowledge, this is the first attempt to obtain distributed algorithms for a nonconvex problem in the general form $\mathcal{P}$.

A. Dual Decomposition Methods

Subproblem $\mathcal{P}_{\mathbf{x}^\nu}$ is convex and can be solved in a distributed way if the constraints $\tilde{\mathbf{g}}(\mathbf{x};\mathbf{x}^\nu)\le\mathbf{0}$ are dualized. The dual problem associated with each $\mathcal{P}_{\mathbf{x}^\nu}$ is: given $\mathbf{x}^\nu\in\mathcal{X}(\mathbf{x}^\nu)$,

$$\max_{\boldsymbol{\lambda}\ge\mathbf{0}}\ d(\boldsymbol{\lambda};\mathbf{x}^\nu)\tag{16}$$

with

$$d(\boldsymbol{\lambda};\mathbf{x}^\nu)\triangleq\min_{\mathbf{x}\in\mathcal{K}}\left\{\sum_{i=1}^I\left(\tilde{U}_i(\mathbf{x}_i;\mathbf{x}^\nu)+\boldsymbol{\lambda}^T\tilde{\mathbf{g}}^i(\mathbf{x}_i;\mathbf{x}^\nu)\right)\right\}\tag{17}$$

Note that, for $\boldsymbol{\lambda}\ge0$, by Assumptions 2 and 3, the minimization in (17) has a unique solution, which will be denoted by $\widehat{\mathbf{x}}(\boldsymbol{\lambda};\mathbf{x}^\nu)\triangleq(\widehat{\mathbf{x}}_i(\boldsymbol{\lambda};\mathbf{x}^\nu))_{i=1}^I$, with

$$\widehat{\mathbf{x}}_i(\boldsymbol{\lambda};\mathbf{x}^\nu)\triangleq\operatorname*{argmin}_{\mathbf{x}_i\in\mathcal{K}_i}\left\{\tilde{U}_i(\mathbf{x}_i;\mathbf{x}^\nu)+\boldsymbol{\lambda}^T\tilde{\mathbf{g}}^i(\mathbf{x}_i;\mathbf{x}^\nu)\right\}.\tag{18}$$

Before proceeding, let us mention the following standard condition.

D4) $\tilde{\mathbf{g}}(\bullet;\mathbf{x}^\nu)$ is uniformly Lipschitz continuous on $\mathcal{K}$ with constant $L_{\tilde{g}}$;

We remind that D4 is implied by condition C7 of Assumption 4; therefore we do not need to assume it if Assumption 4 is invoked. But in order to connect our results below to classical ones, it is good to highlight it as a separate assumption.

The next lemma summarizes some desirable properties of the dual function $d(\boldsymbol{\lambda};\mathbf{x}^\nu)$, which are instrumental to prove the convergence of dual schemes. The Lemma is similar to [14, Lemma 2].

Lemma 3: Given $\mathcal{P}_{\mathbf{x}^\nu}$, under Assumptions 1–3, 5, and 6, the following hold.

(a) $d(\boldsymbol{\lambda};\mathbf{x}^\nu)$ is differentiable with respect to $\boldsymbol{\lambda}$ on $\mathbb{R}^m_+$, with gradient

$$\nabla_{\boldsymbol{\lambda}} d(\boldsymbol{\lambda};\mathbf{x}^\nu) = \sum_i \tilde{\mathbf{g}}^i\left(\widehat{\mathbf{x}}_i(\boldsymbol{\lambda};\mathbf{x}^\nu);\mathbf{x}^\nu\right). \tag{19}$$

(b) If in addition D4 holds, then $\nabla_{\boldsymbol{\lambda}} d(\boldsymbol{\lambda};\mathbf{x}^\nu)$ is Lipschitz continuous on $\mathbb{R}^m_+$ with constant $L_{\nabla d} \triangleq L^2_{\tilde{g}}\sqrt{m}/c_{\tilde{U}}$.

Proof: See Appendix C. ■

The dual-problem can be solved, e.g., using well-known gradient algorithms [41]; an instance is given in Algorithm 2, whose convergence is stated in Theorem 4. The proof of the theorem follows from Lemma 3 and standard convergence results for gradient projection algorithms (e.g., see [43, Th. 3.2] and [41, Prop. 8.2.6] for the theorem statement under assumptions (a) and (b), respectively). We remark that an assumption made in the aforementioned references is that subproblem $\mathcal{P}_{\mathbf{x}^\nu}$ has a zero-duality gap and the dual problem (16) has a non-empty solution set. In our setting, this is guaranteed by Assumption 5, that ensures that $\mathcal{X}(\mathbf{x}^\nu)$ satisfies Slater's CQ (see the discussion around Assumption 5).

In (20) $[\bullet]_+$ denotes the Euclidean projection onto $\mathbb{R}_+$, i.e., $[x]_+ \triangleq \max(0,x)$.

Theorem 4: Given $\mathcal{P}$, under Assumptions 1–3, 5, and 6, suppose that one of the two following conditions is satisfied:

(a) D4 holds true and $\{\alpha^n\}$ is chosen such that $0 < \inf_n \alpha^n \le \sup_n \alpha^n < 2/L_{\nabla d}$, for all $n \ge 0$;

(b) $\nabla_{\boldsymbol{\lambda}} d(\bullet;\mathbf{x}^\nu)$ is uniformly bounded on $\mathbb{R}^m_+$, and α^n is chosen such that $\alpha^n > 0$, $\alpha^n \to 0$, $\sum_n \alpha^n = \infty$, and $\sum_n (\alpha^n)^2 < \infty$.

Then, the sequence $\{\boldsymbol{\lambda}^n\}$ generated by Algorithm 2 converges to a solution of (16), and the sequence $\{\widehat{\mathbf{x}}(\boldsymbol{\lambda}^n;\mathbf{x}^\nu)\}$ converges to the unique solution of $\mathcal{P}_{\mathbf{x}^\nu}$.

Remark 5 (On the distributed implementation): The implementation of Algorithm 1 based on Algorithm 2 leads to a double-loop scheme: given the current value of the multipliers $\boldsymbol{\lambda}^n$, the subproblems (18) can be solved in parallel across the blocks; once the new values $\widehat{\mathbf{x}}_i(\boldsymbol{\lambda}^n;\mathbf{x}^\nu)$ are available, the multipliers are updated according to (20). Note that when $m = 1$ (i.e., there is only one shared constraint), the update in (20) can be replaced by a bisection search, which generally converges quite fast. When $m > 1$, the potential slow convergence of gradient updates (20) can be alleviated using accelerated gradient-based (proximal) schemes; see, e.g., [44], [45].

Algorithm 2: Dual-based Distributed Implementation of Step 2 of NOVA Algorithm (Algorithm 1).

Data: $\boldsymbol{\lambda}^0 \ge \mathbf{0}$, $\mathbf{x}^\nu$, $\{\alpha^n\} > 0$; set $n = 0$.

(S.2a) If $\boldsymbol{\lambda}^n$ satisfies a suitable termination criterion: STOP.

(S.2b) Solve in parallel (18): for all $i = 1,\dots,I$, compute $\widehat{\mathbf{x}}_i(\boldsymbol{\lambda}^n;\mathbf{x}^\nu)$.

(S.2c) Update $\boldsymbol{\lambda}$ according to

$$\boldsymbol{\lambda}^{n+1} \triangleq \left[\boldsymbol{\lambda}^n + \alpha^n \sum_{i=1}^I \tilde{\mathbf{g}}^i\left(\widehat{\mathbf{x}}_i(\boldsymbol{\lambda}^n;\mathbf{x}^\nu);\mathbf{x}^\nu\right)\right]_+. \tag{20}$$

(S.2d) $n \leftarrow n+1$ and go back to (S.2a).

As far as the communication overhead is concerned, the required signaling is in the form of message passing and of course is problem dependent; see Part II of the paper [19] for specific examples. For instance, in networking applications where there is a cluster-head, the update of the multipliers can be performed at the cluster, and, then, it can be broadcasted to the users. In fully decentralized networks instead, the update of $\boldsymbol{\lambda}$ can be done by the users themselves, by running consensus based algorithms to locally estimate $\sum_{i=1}^I \tilde{\mathbf{g}}^i\left(\widehat{\mathbf{x}}_i(\boldsymbol{\lambda}^n;\mathbf{x}^\nu);\mathbf{x}^\nu\right)$. This, in general, requires a limited signaling exchange among neighboring nodes only. Note also that the size of the dual problem (the dimension of $\boldsymbol{\lambda}$) is equal to m (the number of shared constraints), which makes Algorithm 2 scalable in the number of blocks (users).

B. Primal Decomposition Methods

Algorithm 2 is based on the relaxation of the shared constraints into the Lagrangian, resulting, in general, in a violation of these constraints during the intermediate iterates. In some applications this fact may prevent the on-line implementation of the algorithm. In this section we propose a distributed scheme that does not suffer from this issue: we cope with the shared constraints using a primal decomposition technique.

Introducing the slack variables $\mathbf{t} \triangleq (\mathbf{t}_i)_{i=1}^I$, with each $\mathbf{t}_i \in \mathbb{R}^m$, $\mathcal{P}_{\mathbf{x}^\nu}$ can be rewritten as

$$\begin{array}{ll} \min\limits_{(\mathbf{x}_i,\mathbf{t}_i)_{i=1}^I} & \sum_{i=1}^I \tilde{U}_i(\mathbf{x}_i;\mathbf{x}^\nu), \\ \text{s.t.} & \mathbf{x}_i \in \mathcal{K}_i, \quad \forall i = 1,\dots,I, \\ & \tilde{\mathbf{g}}^i\left(\mathbf{x}_i;\mathbf{x}^\nu\right) \le \mathbf{t}_i, \quad \forall i = 1,\dots,I, \\ & \sum_{i=1}^I \mathbf{t}_i \le \mathbf{0}. \end{array} \tag{21}$$

When $\mathbf{t} = (\boldsymbol{t}_i)_{i=1}^I$ is fixed, (21) can be decoupled across the users: for each $i = 1,\dots,I$, solve

$$\begin{array}{ll} \min\limits_{\mathbf{x}_i} & \tilde{U}_i(\mathbf{x}_i;\mathbf{x}^\nu), \\ \text{s.t.} & \mathbf{x}_i \in \mathcal{K}_i, \\ & \tilde{\mathbf{g}}^i\left(\mathbf{x}_i;\mathbf{x}^\nu\right) \overset{\boldsymbol{\mu}_i(\mathbf{t}_i;\mathbf{x}^\nu)}{\le} \mathbf{t}_i, \end{array} \tag{22}$$

where $\boldsymbol{\mu}_i(\mathbf{t}_i;\mathbf{x}^\nu)$ is the optimal Lagrange multiplier associated with the inequality constraint $\tilde{\mathbf{g}}^i\left(\mathbf{x}_i;\mathbf{x}^\nu\right) \le \mathbf{t}_i$. Note that the existence of $\boldsymbol{\mu}_i(\mathbf{t}_i;\mathbf{x}^\nu)$ is guaranteed if (22) has zero-duality gap [46, Prop. 6.5.8] (e.g., when some CQ hold), but $\boldsymbol{\mu}_i(\mathbf{t}_i;\mathbf{x}^\nu)$ may not be unique.Let us denote by $\mathbf{x}_i^\star(\mathbf{t}_i;\mathbf{x}^\nu)$ the unique solution of (22), given $\mathbf{t} = (\boldsymbol{t}_i)_{i=1}^I$. The optimal partition $\mathbf{t}^\star \triangleq (\mathbf{t}_i^\star)_{i=1}^I$ of

the shared constraints can be found solving the so-called *master* (convex) problem (see, e.g., [20]):

$$\begin{array}{ll}\min\limits_{\boldsymbol{t}} & P(\boldsymbol{t};\mathbf{x}^\nu) \triangleq \sum_{i=1}^I \tilde{U}_i(\mathbf{x}_i^\star(\mathbf{t}_i;\mathbf{x}^\nu);\mathbf{x}^\nu) \\ \text{s.t.} & \sum_{i=1}^I \mathbf{t}_i \leq \mathbf{0}.\end{array} \tag{23}$$

Due to the non-uniqueness of $\boldsymbol{\mu}_i(\mathbf{t}_i;\mathbf{x}^\nu)$, the objective function in (23) is nondifferentiable; problem (23) can be solved by subgradient methods. The partial subgradient of $P(\boldsymbol{t};\mathbf{x}^\nu)$ with respect to the first argument evaluated at $(\mathbf{t};\mathbf{x}^\nu)$ is

$$\partial_{\mathbf{t}_i} P(\mathbf{t};\mathbf{x}^\nu) = -\boldsymbol{\mu}_i(\mathbf{t}_i;\mathbf{x}^\nu), \quad i=1,\ldots,I.$$

We refer to [41, Prop. 8.2.6] for standard convergence results for subgradient projection algorithms. We also remark that solving (23) requires a *master-node* to be dedicated to this task, and thus calls for some additional, limited degree of coordination. The necessary signaling and how the coordination can be implemented in practice is very much problem-dependent; therefore we do not go further into these details.

V. Extension to Nonsmooth Problems

The results described so far were obtained under smoothness conditions on all the problem functions. In this section, motivated also by some important applications in machine learning (see Part II of this paper, [19]), we extend our approach to cope with some structured type of nonsmoothness. In particular, with respect to the class of nonsmooth problems we describe below, it is possible to extend (almost) all results we have already established for Problem $\mathcal{P}$, under assumptions that are the natural generalization to the nonsmooth domain of those used in the smooth case.

Consider the following extension of problem $\mathcal{P}$:

$$\begin{array}{ll}\min\limits_{\mathbf{x}} & U(\mathbf{x}) + R(\mathbf{x}) \\ \text{s.t.} & \left.\begin{array}{l} g_j(\mathbf{x}) \leq 0,\ j=1,\ldots,m \\ h_l(\mathbf{x}) \triangleq h_l^+(\mathbf{x}) - h_l^-(\mathbf{x}) \leq 0,\ l=1,\ldots,p \\ \mathbf{x}\in\mathcal{K},\end{array}\right\} \triangleq \mathcal{X}\end{array} \tag{24}$$

where R is a (possibly nondifferentiable) convex, finite regularization term, and the DC constraints h_l are such that h_l^+ and h_l^- are (possibly nondifferentiable) convex and finite, with at most one of them being nonsmooth, while U, g, and $\mathcal{K}$ are as in problem $\mathcal{P}$. Typical examples for R are $\|\cdot\|_1$ and $\|\cdot\|_2$. In particular, regularization by the ℓ_1 norm has been intensely studied in recent years, in order to promote sparsity of the solution, see [42], [47]–[50] as entry points to the literature. On the other hand, DC-type constraints have been used in several applications, ranging from communications to machine learning; the interested reader can refer to [51], [52] and references therein for recent methodological developments in the field of DC programming. We remark that, unlike standard DC programming algorithms, the proposed framework can deal also with objective functions and further constraints that are not expressed as DC functions but nevertheless nonconvex (see Part II [19] of the paper for several motivating applications falling within this category). Moreover, we maintain feasibility of iterates, a feature that, to the best of our knowledge, is lacking in DC programming methods.

A. Technical Preliminaries and Assumptions

In order to analyze Problem (24), we need suitable assumptions: we modify Assumption 1 quite naturally by replacing A4 with the requirement that $U+R$ be coercive on $\mathcal{K}$; we refer to this modified condition as **Assumption 1′**. Furthermore, we must extend to the new nonsmooth setting the concept of regularity and of KKT stationary point.

Definition 6 (Regularity): A point $\bar{\mathbf{x}}\in\mathcal{X}$ is called *regular* for (24) if the nonsmooth Mangasarian-Fromovitz Constraint Qualification (nMFCQ) holds at $\bar{\mathbf{x}}$, that is if the following implication is satisfied:

$$\left.\begin{array}{rcl}\mathbf{0} & \in & \sum_{j\in\bar{J}}\mu_j\nabla_{\mathbf{x}} g_j(\bar{\mathbf{x}}) \\ & & +\sum_{l\in\bar{L}}\lambda_l\partial_{\mathbf{x}} h_l(\bar{\mathbf{x}}) + N_{\mathcal{K}}(\bar{\mathbf{x}}) \\ & & \mu_j\geq 0,\ \forall j\in\bar{J} \\ & & \lambda_l\geq 0,\ \forall l\in\bar{L}\end{array}\right\} \Rightarrow \begin{array}{l}\mu_j=0,\ \forall j\in\bar{J} \\ \lambda_l=0,\ \forall l\in\bar{L}\end{array} \tag{25}$$

where $N_{\mathcal{K}}(\bar{\mathbf{x}}) \triangleq \{\mathbf{d}\in\mathbb{R}^n : \mathbf{d}^T(\mathbf{y}-\bar{\mathbf{x}})\leq 0,\ \forall\mathbf{y}\in\mathcal{K}\}$ is the normal cone to $\mathcal{K}$ at $\bar{\mathbf{x}}$, $\bar{J}\triangleq\{j\in\{1,\ldots,m\} : g_j(\bar{\mathbf{x}})=0\}$ and $\bar{L}\triangleq\{j\in\{1,\ldots,p\} : h_l(\bar{\mathbf{x}})=0\}$ are the indices sets of those (nonconvex) constraints that are active at $\bar{\mathbf{x}}$, and $\partial_{\mathbf{x}} h_l(\bar{\mathbf{x}})$ denotes the set of Clarke subgradients of h_l at $\bar{\mathbf{x}}$; see [53]. Note that ∂h_l^-, the set of Clarke subgradients of the convex function h_l^- coincides with the classical subdifferential from convex analysis.

Definition 7 (Nonsmooth KKT): We recall that, if $\bar{\mathbf{x}}\in\mathcal{X}$ is a minimum of problem (24) at which the nMFCQ holds, then the following nonsmooth KKT (nKKT) system must be satisfied:

$$\begin{aligned}\mathbf{0} \in \nabla_{\mathbf{x}} U(\bar{\mathbf{x}}) + \partial_{\mathbf{x}} R(\bar{\mathbf{x}}) + \sum_{j\in\bar{J}}\mu_j\nabla_{\mathbf{x}} g_j(\bar{\mathbf{x}}) \\ + \sum_{l\in\bar{L}}\lambda_l\partial_{\mathbf{x}} h_l(\bar{\mathbf{x}}) + N_{\mathcal{K}}(\bar{\mathbf{x}}) \\ \mu_j\geq 0,\ \forall j\in\bar{J}, \quad \lambda_l\geq 0,\ \forall l\in\bar{L},\end{aligned}$$

for some suitable Lagrange multipliers μ_j, λ_ls.
These stationarity conditions are those generally used when one studies algorithms in nonsmooth contexts.

B. Algorithmic Framework

We are now in the position to extend our method to Problem (24). The proposed scheme has essentially the same form of Algorithm 1, with the only difference that $\hat{\mathbf{x}}(\mathbf{x}^\nu)$ in Step S.2 of Algorithm 1 is defined as solution of the following strongly convex approximating subproblem: given the current iterate $\mathbf{x}^\nu$,

$$\begin{array}{ll}\min\limits_{\mathbf{x}} & \tilde{U}(\mathbf{x};\mathbf{x}^\nu) + R(\mathbf{x}) \\ \text{s.t.} & \tilde{g}_j(\mathbf{x};\mathbf{x}^\nu)\leq 0,\ j=1,\ldots,m \\ & \tilde{h}_l(\mathbf{x};\mathbf{x}^\nu)\leq 0,\ l=1,\ldots,p \\ & \mathbf{x}\in\mathcal{K},\end{array} \tag{26}$$

where $\tilde{U}$ and $\tilde{g}_j$ are approximations of U and g_j satisfying Assumptions 2 and 3, respectively; and $\tilde{h}_l$ is a convex approximation of the nonsmooth DC constraints $h_l(\mathbf{x}) \leq 0$, which is chosen as follows.

On the choice of $\tilde{h}_l$: We distinguish the following three cases:

1) h_l^+ and h_l^- are continuously differentiable: one can simply rely on the approximation in Example 3, i.e., $\tilde{h}_l(\mathbf{x};\mathbf{y}) \triangleq h_l^+(\mathbf{x}) - h_l^-(\mathbf{y}) - \nabla_{\mathbf{x}} h_l^-(\mathbf{y})^T(\mathbf{x}-\mathbf{y})$;
2) h_l^+ is continuously differentiable and h_l^- nonsmooth: one can resort to the upper approximation $\tilde{h}_l(\mathbf{x};\mathbf{y}) \triangleq h_l^+(\mathbf{x}) - h_l^-(\mathbf{y}) - \xi_l^{-T}(\mathbf{x}-\mathbf{y})$, for any $\xi_l^- \in \partial_{\mathbf{x}} h_l^-(\mathbf{y})$, where $\partial_{\mathbf{x}} h_l^-(\mathbf{y})$ denotes the set of Clarke subgradients of h_l^- at $\mathbf{y}$;
3) h_l^+ is nonsmooth and h_l^- continuously differentiable: one can use the upper approximation $\tilde{h}_l(\mathbf{x};\mathbf{y}) \triangleq h_l^+(\mathbf{x}) - h_l^-(\mathbf{y}) - \nabla h_l^+(\mathbf{y})^T(\mathbf{x}-\mathbf{y})$.

We term the modification of Algorithm 1 described above *Algorithm 1′*. A few remarks are in order.

−If $R \equiv 0$ and there are no DC constraints ($p = 0$) Algorithm 1′ reduces to Algorithm 1, so that Algorithm 1′ is actually a proper extension of Algorithm 1;

−Subproblem (26) is nonsmooth when $R \not\equiv 0$ or when at least one DC constraint is such that h_l^+ is nonsmooth; this means that the approach we propose is sensible only when we are able to solve efficiently subproblem (26). This is often the case: for example, if $R = \|\cdot\|_1$ (and all $\tilde{h}_l$ are differentiable), very efficient solution methods exist, see e.g. [42], [47]–[50] and references therein.

The following theorem summarizes the convergence properties of Algorithm 1′ when applied to problem (24).

Theorem 8: Given the nonsmooth nonconvex problem (24) under Assumptions 1′, 2, 3, and 5, let $\{\mathbf{x}^\nu\}$ be the sequence generated by Algorithm 1′. The following hold.

(a) $\mathbf{x}^\nu \in \mathcal{X}$ for all $\nu \geq 0$ (iterate feasibility);
(b) If the step-size γ^ν and $c_{\tilde{U}}$ are chosen so that

$$0 < \inf_\nu \gamma^\nu \leq \sup_\nu \gamma^\nu \leq \gamma^{\max} \leq 1 \text{ and } 2c_{\tilde{U}} > \gamma^{\max} L_{\nabla U}, \tag{27}$$

then $\{\mathbf{x}^\nu\}$ is bounded and each of its limit points is a stationary point of problem (24).
(c) If the step-size γ^ν is chosen so that

$$\gamma^\nu \in (0,1], \quad \gamma^\nu \to 0, \quad \text{and} \quad \sum_\nu \gamma^\nu = +\infty, \tag{28}$$

then $\{\mathbf{x}^\nu\}$ is bounded and at least one of its limit points is stationary.

Furthermore, if the algorithm does not stop after a finite number of steps, none of the stationary points in parts (b) and (c) is a local maximum of $U + R$.

Proof: Because of space limitation, the proof is omitted; it can be found in the on-line technical report [54] and the on-line supplementing material of this paper. ■

Theorem 8 mimics quite faithfully the results in Theorem 2 for Algorithm 1; the only difference is that [see (c)], when a diminishing step-size is adopted in nonsmooth cases, we can only prove that at *least one* of the limit points of the sequence is stationary, while, in the smooth case, we could show that *every* limit point of the sequence generated by the algorithm is stationary. While this theoretical difference is likely to have no significant consequences in practice, see also the results in Part II [19], it testifies to the higher complexity of the nonsmooth setting.

VI. Conclusions

In this Part I of the two-part paper, we proposed a novel general algorithmic framework based on convex approximation techniques for the solution of nonconvex, possibly nonsmooth optimization problems: we point out that the nonconvexity may occur both in the objective function and in the constraints. Some key novel features of our scheme are: i) it maintains feasibility *and* leads to *parallel and distributed* solution methods for a very general class of nonconvex, possibly nonsmooth problems; ii) it offers a lot of flexibility in choosing the approximation functions, enlarging significantly the class of problems that can be solved with provable convergence; iii) by choosing different approximation functions, different (distributed) schemes can be obtained: they are all convergent, but differ for (practical) convergence speed, complexity, communication overhead, and a priori knowledge of the system parameters; iv) it includes as special cases several classical SCA-based algorithms and improves on their convergence properties; and v) it provides new efficient algorithms also for old problems. In Part II [19] we customize the developed algorithmic framework to a variety of new (and old) multi-agent optimization problems in signal processing, communications and networking, providing a solid evidence of its good performance. Quite interestingly, even when compared with existing schemes that have been designed for very specific problems, our algorithms are shown to outperform them.

Appendix

We first introduce some intermediate technical results that are instrumental to prove Theorem 2. The proof of Theorem 2 is given in Appendix B.

A. Intermediate Results

We first prove Lemma 9–Lemma 13, providing some key properties of the sequence $\{\mathbf{x}^\nu\}$ generated by Algorithm 1 and of the best-response function $\hat{\mathbf{x}}(\bullet)$ defined in (3). Finally, with Theorem 14 we establish some technical conditions under which (at least one) regular limit point of the sequence generated by Algorithm 1 is a stationary solution of the original nonconvex problem $\mathcal{P}$; the proof of Theorem 2 will rely on such conditions. We recall that, for the sake of simplicity, throughout this section we tacitly assume that Assumptions 1–3 and 5 are satisfied.

Lemma 9. The first lemma shows, among other things, that Algorithm 1 produces a sequence of points that are feasible for the original problem $\mathcal{P}$.

Lemma 9: The following properties hold.

(i) $\mathbf{y} \in \mathcal{X}(\mathbf{y}) \subseteq \mathcal{X}$ for all $\mathbf{y} \in \mathcal{X}$;
(ii) $\hat{\mathbf{x}}(\mathbf{y}) \in \mathcal{X}(\mathbf{y}) \subseteq \mathcal{X}$ for all $\mathbf{y} \in \mathcal{X}$.

Moreover, the sequence $\{\mathbf{x}^\nu\}$ generated by Algorithm 1 is such that:

(iii) $\mathbf{x}^\nu \in \mathcal{X}$;
(iv) $\mathbf{x}^{\nu+1} \in \mathcal{X}(\mathbf{x}^\nu) \cap \mathcal{X}(\mathbf{x}^{\nu+1})$.

Proof: (i) the first implication $\mathbf{y} \in \mathcal{X}(\mathbf{y})$ follows from $\tilde{g}_j(\mathbf{y};\mathbf{y}) = g_j(\mathbf{y}) \le 0$, for all $j = 1,\dots,m$ [due to C2]. For the inclusion $\mathcal{X}(\mathbf{y}) \subseteq \mathcal{X}$, it suffices to recall that, by C3, we have $g_j(\mathbf{x}) \le \tilde{g}_j(\mathbf{x};\mathbf{y})$ for all $\mathbf{x} \in \mathcal{K}$, $\mathbf{y} \in \mathcal{X}$, and $j = 1,\dots,m$, implying that, if $\mathbf{x} \in \mathcal{X}(\mathbf{y})$, then $\mathbf{x} \in \mathcal{X}$.

(ii) $\hat{\mathbf{x}}(\mathbf{y}) \in \mathcal{X}(\mathbf{y})$ since it is the optimal solution of $\mathcal{P}_\mathbf{y}$ (and thus also feasible).

(iii) In view of (i) and (ii), it follows by induction and the fact that $\mathbf{x}^{\nu+1}$ is a convex combination of $\mathbf{x}^\nu \in \mathcal{X}(\mathbf{x}^\nu)$ and $\hat{\mathbf{x}}(\mathbf{x}^\nu) \in \mathcal{X}(\mathbf{x}^\nu)$, which is a convex subset of $\mathcal{X}$.

(iv) By (iii), $\mathbf{x}^{\nu+1} \in \mathcal{X}(\mathbf{x}^\nu)$. Furthermore, we have $\tilde{g}_j(\mathbf{x}^{\nu+1};\mathbf{x}^{\nu+1}) = g_j(\mathbf{x}^{\nu+1}) \le 0$, for all $j = 1,\dots,m$, where the equality follows from C2 and the inequality is due to $\mathbf{x}^{\nu+1} \in \mathcal{X}$; thus, $\mathbf{x}^{\nu+1} \in \mathcal{X}(\mathbf{x}^{\nu+1})$. ■

Lemma 10. With Lemma 10, we establish some key properties of the best-response function $\hat{\mathbf{x}}(\bullet)$. We will use the following definitions. Given $\mathbf{y}, \mathbf{z} \in \mathcal{X}$ and $\rho > 0$, let

$$\mathbf{w}_\rho(\hat{\mathbf{x}}(\mathbf{y}), \mathbf{z}) \triangleq \hat{\mathbf{x}}(\mathbf{y}) - \rho \nabla_\mathbf{x} \tilde{U}(\hat{\mathbf{x}}(\mathbf{y}); \mathbf{z}); \tag{29}$$

and let $P_{\mathcal{X}(\mathbf{y})}(\mathbf{u})$ denote the Euclidean projection of $\mathbf{u} \in \mathbb{R}^n$ onto the closed convex set $\mathcal{X}(\mathbf{y})$:

$$P_{\mathcal{X}(\mathbf{y})}(\mathbf{u}) = \operatorname*{argmin}_{\mathbf{x} \in \mathcal{X}(\mathbf{y})} \|\mathbf{x} - \mathbf{u}\|. \tag{30}$$

Lemma 10: The best-response function $\mathcal{X} \ni \mathbf{y} \mapsto \hat{\mathbf{x}}(\mathbf{y})$ satisfies the following:

(i) For every $\mathbf{y} \in \mathcal{X}$, $\hat{\mathbf{x}}(\mathbf{y}) - \mathbf{y}$ is a descent direction for U at $\mathbf{y}$ such that

$$\nabla U(\mathbf{y})^T (\hat{\mathbf{x}}(\mathbf{y}) - \mathbf{y}) \le -c_{\tilde{U}} \|\hat{\mathbf{x}}(\mathbf{y}) - \mathbf{y}\|^2, \tag{31}$$

where $c_{\tilde{U}} > 0$ is the constant of uniform strong convexity of $\tilde{U}$ (cf. B1);

(ii) For every $\mathbf{y} \in \mathcal{X}$, it holds that

$$\hat{\mathbf{x}}(\mathbf{y}) = P_{\mathcal{X}(\mathbf{y})}\left(\mathbf{w}_\rho\left(\hat{\mathbf{x}}(\mathbf{y}), \mathbf{y}\right)\right), \tag{32}$$

for every fixed $\rho > 0$.

(iii) Suppose that also B5 holds true. Then, $\hat{\mathbf{x}}(\bullet)$ is continuous at every $\bar{\mathbf{x}} \in \mathcal{X}$ such that $\hat{\mathbf{x}}(\bar{\mathbf{x}}) \in \mathcal{X}(\bar{\mathbf{x}})$ is regular.

Proof: (i) By Assumption 2, for any given $\mathbf{y} \in \mathcal{X}$, $\hat{\mathbf{x}}(\mathbf{y})$ is the solution of the strongly convex optimization problem $\mathcal{P}_{\mathbf{x}^\nu}$; therefore,

$$(\mathbf{z} - \hat{\mathbf{x}}(\mathbf{y}))^T \nabla_\mathbf{x} \tilde{U}\left(\hat{\mathbf{x}}(\mathbf{y}); \mathbf{y}\right) \ge 0, \quad \forall \mathbf{z} \in \mathcal{X}(\mathbf{y}). \tag{33}$$

By choosing $\mathbf{z} = \mathbf{y}$ [recall by Lemma 9(i) that $\mathbf{y} \in \mathcal{X}(\mathbf{y})$], we get

$$(\mathbf{y} - \hat{\mathbf{x}}(\mathbf{y}))^T \Big(\nabla_\mathbf{x} \tilde{U}(\hat{\mathbf{x}}(\mathbf{y}); \mathbf{y}) - \nabla_\mathbf{x} \tilde{U}(\mathbf{y}; \mathbf{y}) + \nabla_\mathbf{x} \tilde{U}(\mathbf{y}; \mathbf{y})\Big) \ge 0,$$

which, using B1 and B2, leads to

$$(\mathbf{y} - \hat{\mathbf{x}}(\mathbf{y}))^T \nabla_\mathbf{x} U(\mathbf{y}) \ge c_{\tilde{U}} \|\hat{\mathbf{x}}(\mathbf{y}) - \mathbf{y}\|^2.$$

(ii) It follows readily from the fixed-point characterization of the solution $\hat{\mathbf{x}}(\mathbf{y})$ of the strongly convex subproblem $\mathcal{P}_\mathbf{y}$: see, e.g., [55, Prop. 1.5.8].

(iii) We first observe that, under the assumed regularity of all the points in $\mathcal{X}(\bar{\mathbf{x}})$, $\mathcal{X}(\bullet)$ is continuous at $\bar{\mathbf{x}}$ [39, Example 5.10]. It follows from [39, Proposition 4.9] (see also [39, Example 5.57]) that, for every fixed $\mathbf{u} \in \mathbb{R}^n$, the map $\mathbf{x} \mapsto P_{\mathcal{X}(\mathbf{x})}(\mathbf{u})$ is continuous at $\mathbf{x} = \bar{\mathbf{x}}$. This, together with B1, B3 and B5 is sufficient for $\hat{\mathbf{x}}(\bullet)$ to be continuous at $\bar{\mathbf{x}}$ [56, Theorem 2.1]. ■

Lemma 11. Under the extra conditions B4-B5, with the following lemma, which is reminiscent of similar results in [56] and [57], we can establish a suitable sensitivity property of the best-response function $\hat{\mathbf{x}}(\bullet)$; Lemma 11 will play a key role in the proof of statement (c) of Theorem 2.

Lemma 11: Suppose that B4-B5 hold and there exist $\bar{\rho} > 0$ and $\beta > 0$ such that

$$\|P_{\mathcal{X}(\mathbf{y})}(\mathbf{w}_\rho(\hat{\mathbf{x}}(\mathbf{z}), \mathbf{z})) - P_{\mathcal{X}(\mathbf{z})}(\mathbf{w}_\rho(\hat{\mathbf{x}}(\mathbf{z}), \mathbf{z}))\| \le \beta \|\mathbf{y} - \mathbf{z}\|^{\frac{1}{2}}, \tag{34}$$

for all $\rho \in (0, \bar{\rho}]$ and $\mathbf{y}, \mathbf{z} \in \mathcal{X}$. Then there exists $\tilde{\rho} \in (0, \bar{\rho}]$ such that

$$\|\hat{\mathbf{x}}(\mathbf{y}) - \hat{\mathbf{x}}(\mathbf{z})\| \le \eta_\rho \|\mathbf{y} - \mathbf{z}\| + \theta_\rho \|\mathbf{y} - \mathbf{z}\|^{\frac{1}{2}}, \tag{35}$$

for all $\mathbf{y}, \mathbf{z} \in \mathcal{X}$ and $\rho \in (0, \tilde{\rho}]$, with

$$\begin{aligned} \eta_\rho &\triangleq \frac{\rho\, \tilde{L}_{\nabla,2}}{1 - \sqrt{1 + \rho^2 \tilde{L}_{\nabla,1}^2 - 2\rho c_{\tilde{U}}}} \\ \theta_\rho &\triangleq \frac{\beta}{1 - \sqrt{1 + \rho^2 \tilde{L}_{\nabla,1}^2 - 2\rho c_{\tilde{U}}}}, \end{aligned} \tag{36}$$

where $\tilde{L}_{\nabla,1}$ and $\tilde{L}_{\nabla,2}$ are the Lipschitz constants of $\nabla_\mathbf{x} \tilde{U}(\bullet;\mathbf{y})$ and $\nabla_\mathbf{x} \tilde{U}(\mathbf{x};\bullet)$, respectively (cf. B4 and B5); $\tilde{L}_{\nabla,1}$ is assumed to be such that $\tilde{L}_{\nabla,1} \ge c_{\tilde{U}}$ without loss of generality.

Proof: Using (32) we have, for every $\rho > 0$,

$$\begin{aligned} &\|\hat{\mathbf{x}}(\mathbf{y}) - \hat{\mathbf{x}}(\mathbf{z})\| \\ &\quad = \|P_{\mathcal{X}(\mathbf{y})}(\mathbf{w}_\rho(\hat{\mathbf{x}}(\mathbf{y}), \mathbf{y})) - P_{\mathcal{X}(\mathbf{z})}(\mathbf{w}_\rho(\hat{\mathbf{x}}(\mathbf{z}), \mathbf{z}))\| \\ &\quad \le \|P_{\mathcal{X}(\mathbf{y})}(\mathbf{w}_\rho(\hat{\mathbf{x}}(\mathbf{y}), \mathbf{y})) - P_{\mathcal{X}(\mathbf{y})}(\mathbf{w}_\rho(\hat{\mathbf{x}}(\mathbf{z}), \mathbf{y}))\| \\ &\qquad + \|P_{\mathcal{X}(\mathbf{y})}(\mathbf{w}_\rho(\hat{\mathbf{x}}(\mathbf{z}), \mathbf{y})) - P_{\mathcal{X}(\mathbf{z})}(\mathbf{w}_\rho(\hat{\mathbf{x}}(\mathbf{z}), \mathbf{z}))\|. \end{aligned} \tag{37}$$

We bound next the two terms on the RHS of (37).

For every $\rho > 0$, it holds

$$\begin{aligned} &\|P_{\mathcal{X}(\mathbf{y})}(\mathbf{w}_\rho(\hat{\mathbf{x}}(\mathbf{y}), \mathbf{y})) - P_{\mathcal{X}(\mathbf{y})}(\mathbf{w}_\rho(\hat{\mathbf{x}}(\mathbf{z}), \mathbf{y}))\|^2 \\ &\quad \overset{(a)}{\le} \|\mathbf{w}_\rho(\hat{\mathbf{x}}(\mathbf{y}), \mathbf{y}) - \mathbf{w}_\rho(\hat{\mathbf{x}}(\mathbf{z}), \mathbf{y})\|^2 \\ &\quad = \|\hat{\mathbf{x}}(\mathbf{y}) - \hat{\mathbf{x}}(\mathbf{z})\|^2 \\ &\qquad + \rho^2 \|\nabla_\mathbf{x} \tilde{U}(\hat{\mathbf{x}}(\mathbf{z}); \mathbf{y}) - \nabla_\mathbf{x} \tilde{U}(\hat{\mathbf{x}}(\mathbf{y}); \mathbf{y})\|^2 \\ &\qquad - 2\rho\, (\hat{\mathbf{x}}(\mathbf{y}) - \hat{\mathbf{x}}(\mathbf{z}))^T \Big(\nabla_\mathbf{x} \tilde{U}(\hat{\mathbf{x}}(\mathbf{y}); \mathbf{y}) - \nabla_\mathbf{x} \tilde{U}(\hat{\mathbf{x}}(\mathbf{z}); \mathbf{y})\Big) \\ &\quad \overset{(b)}{\le} \|\hat{\mathbf{x}}(\mathbf{y}) - \hat{\mathbf{x}}(\mathbf{z})\|^2 + \rho^2 \tilde{L}_{\nabla,1}^2 \|\hat{\mathbf{x}}(\mathbf{y}) - \hat{\mathbf{x}}(\mathbf{z})\|^2 \\ &\qquad - 2\,\rho\, c_{\tilde{U}} \|\hat{\mathbf{x}}(\mathbf{y}) - \hat{\mathbf{x}}(\mathbf{z})\|^2 \\ &\quad = (1 + \rho^2 \tilde{L}_{\nabla,1}^2 - 2\,\rho c_{\tilde{U}}) \|\hat{\mathbf{x}}(\mathbf{y}) - \hat{\mathbf{x}}(\mathbf{z})\|^2, \end{aligned} \tag{38}$$

where (a) is due to the non-expansive property of the projection operator $P_{\mathcal{X}(\mathbf{y})}(\bullet)$ and (b) follows from B1 and B5. Note that $1 + \rho^2 \tilde{L}_{\nabla,1}^2 - 2\rho c_{\tilde{U}} > 0$ since we assumed $\tilde{L}_{\nabla,1} \ge c_{\tilde{U}}$.

Let us bound now the second term on the RHS of (37). For every $\rho \in (0, \bar{\rho}]$, we have

$$\begin{aligned}&\|P_{\mathcal{X}(\mathbf{y})}(\mathbf{w}_\rho(\hat{\mathbf{x}}(\mathbf{z}),\mathbf{y})) - P_{\mathcal{X}(\mathbf{z})}(\mathbf{w}_\rho(\hat{\mathbf{x}}(\mathbf{z}),\mathbf{z}))\| \\ &\quad\le \|P_{\mathcal{X}(\mathbf{y})}(\mathbf{w}_\rho(\hat{\mathbf{x}}(\mathbf{z}),\mathbf{y})) - P_{\mathcal{X}(\mathbf{y})}(\mathbf{w}_\rho(\hat{\mathbf{x}}(\mathbf{z}),\mathbf{z}))\| \\ &\qquad + \|P_{\mathcal{X}(\mathbf{y})}(\mathbf{w}_\rho(\hat{\mathbf{x}}(\mathbf{z}),\mathbf{z})) - P_{\mathcal{X}(\mathbf{z})}(\mathbf{w}_\rho(\hat{\mathbf{x}}(\mathbf{z}),\mathbf{z}))\| \\ &\quad\overset{(a)}{\le} \|\mathbf{w}_\rho(\hat{\mathbf{x}}(\mathbf{z}),\mathbf{y}) - \mathbf{w}_\rho(\hat{\mathbf{x}}(\mathbf{z}),\mathbf{z})\| + \beta\|\mathbf{y}-\mathbf{z}\|^{\frac{1}{2}} \\ &\quad\overset{(b)}{\le} \rho\,\tilde{L}_{\nabla,2}\,\|\mathbf{y}-\mathbf{z}\| + \beta\|\mathbf{y}-\mathbf{z}\|^{\frac{1}{2}},\end{aligned} \tag{39}$$

where (a) is due to the non-expansive property of the projection $P_{\mathcal{X}(\mathbf{y})}(\bullet)$ and (34), and (b) follows from B4.

Combining (37), (38) and (39) we obtain the desired result (35) with $\tilde{\rho} = \min\{2c_{\tilde{U}}/\tilde{L}^2_{\nabla,1}, \bar{\rho}\}$ (so that $0 < 1 + \rho^2\tilde{L}^2_{\nabla,1} - 2\rho c_{\tilde{U}} < 1$ for every $\rho \in (0, \tilde{\rho}]$). ■

Lemmas 12 and 13. While Assumptions 1–3 and B4-B5 in Lemma 11 are quite standard, condition (34) is less trivial and not easy to be checked. The following Lemma 13 provides some easier to be checked sufficient conditions that imply (34). To prove Lemma 13 we need first Lemma 12, as stated next.

Lemma 12: Consider $\bar{\mathbf{x}} \in \mathcal{X}$. By assuming C7, the following hold:

(i) If $\tilde{\mathbf{x}} \in \mathcal{X}(\bar{\mathbf{x}})$ is regular, then $\mathcal{X}(\bullet)$ enjoys the Aubin property at $(\bar{\mathbf{x}}, \tilde{\mathbf{x}})$;[1]

(ii) If in addition $\mathcal{X}$ is compact, then a neighborhood $\mathcal{V}_{\tilde{\mathbf{x}}}$ of $\bar{\mathbf{x}}$, a neighborhood $\mathcal{W}_{\tilde{\mathbf{x}}}$ of $\tilde{\mathbf{x}}$, and a constant $\hat{\beta} > 0$ exist such that

$$\|P_{\mathcal{X}(\mathbf{y})}(\mathbf{u}) - P_{\mathcal{X}(\mathbf{z})}(\mathbf{u})\| \le \hat{\beta}\|\mathbf{y}-\mathbf{z}\|^{\frac{1}{2}} \tag{40}$$

for all $\mathbf{y}, \mathbf{z} \in \mathcal{X} \cap \mathcal{V}_{\tilde{\mathbf{x}}}$, and $\mathbf{u} \in \mathcal{W}_{\tilde{\mathbf{x}}}$.

Proof: (i) Under Assumptions 1–3 and C7, the statement follows readily from [58, Theorem 3.2] in view of the regularity of $\tilde{\mathbf{x}}$.

(ii) Since $\mathcal{X}(\bullet)$ has the Aubin property at $(\bar{\mathbf{x}}, \tilde{\mathbf{x}})$, there exist a neighborhood $\mathcal{V}_{\tilde{\mathbf{x}}}$ of $\bar{\mathbf{x}}$, a neighborhood $\mathcal{W}_{\tilde{\mathbf{x}}}$ of $\tilde{\mathbf{x}}$, and a constant $\hat{\beta} > 0$ such that [57, Lemma 1.1]:

$$\|P_{\mathcal{X}(\mathbf{y})\cap\mathcal{B}_{\tilde{\mathbf{x}}}}(\mathbf{u}) - P_{\mathcal{X}(\mathbf{z})\cap\mathcal{B}_{\tilde{\mathbf{x}}}}(\mathbf{u})\| \le \hat{\beta}\|\mathbf{y}-\mathbf{z}\|^{\frac{1}{2}}, \tag{41}$$

for all $\mathbf{y}, \mathbf{z} \in \mathcal{X} \cap \mathcal{V}_{\tilde{\mathbf{x}}}$, and $\mathbf{u} \in \mathcal{W}_{\tilde{\mathbf{x}}}$, where $\mathcal{B}_{\tilde{\mathbf{x}}}$ denotes a closed convex neighborhood of $\tilde{\mathbf{x}}$. Since $\mathcal{X}$ is compact, one can always choose $\mathcal{B}_{\tilde{\mathbf{x}}}$ such that $\mathcal{X}(\bar{\mathbf{x}}) \subset \mathcal{B}_{\tilde{\mathbf{x}}}$ for every $\bar{\mathbf{x}} \in \mathcal{X}$ and, thus,

$$\|P_{\mathcal{X}(\mathbf{y})\cap\mathcal{B}_{\tilde{\mathbf{x}}}}(\mathbf{u}) - P_{\mathcal{X}(\mathbf{z})\cap\mathcal{B}_{\tilde{\mathbf{x}}}}(\mathbf{u})\| = \|P_{\mathcal{X}(\mathbf{y})}(\mathbf{u}) - P_{\mathcal{X}(\mathbf{z})}(\mathbf{u})\|,$$

which proves the desired result.

We can now derive sufficient conditions for (34) to hold.

Lemma 13: Suppose that C7 holds true, $\mathcal{X}$ is compact and $\hat{\mathbf{x}}(\bar{\mathbf{x}}) \in \mathcal{X}(\bar{\mathbf{x}})$ is regular for every $\bar{\mathbf{x}} \in \mathcal{X}$. Then, property (34) holds.

Proof: It follows from Lemma 12(ii) that, for every $\bar{\mathbf{x}} \in \mathcal{X}$, there exist a neighborhood $\mathcal{V}_{\bar{\mathbf{x}}}$ of $\bar{\mathbf{x}}$, a neighborhood $\mathcal{W}_{\hat{\mathbf{x}}(\bar{\mathbf{x}})}$ of $\hat{\mathbf{x}}(\bar{\mathbf{x}})$, and a constant $\hat{\beta} > 0$ such that:

$$\|P_{\mathcal{X}(\mathbf{y})}(\mathbf{u}) - P_{\mathcal{X}(\mathbf{z})}(\mathbf{u})\| \le \hat{\beta}\|\mathbf{y}-\mathbf{z}\|^{\frac{1}{2}} \tag{42}$$

for every $\mathbf{y}, \mathbf{z} \in \mathcal{X} \cap \mathcal{V}_{\bar{\mathbf{x}}}$, $\mathbf{u} \in \mathcal{W}_{\hat{\mathbf{x}}(\bar{\mathbf{x}})}$.

[1]See [39, Def. 9.36] for the definition of the Aubin property. Note also that we use some results from [57] where a point-to-set map that has the Aubin property is called pseudo-Lipschitz [57, Def. 1.1].

Suppose now by contradiction that (34) does not hold. Then, for all $\bar{\rho}^\nu > 0$ and $\beta^\nu > 0$ there exist $\rho^\nu \in (0, \bar{\rho}^\nu]$, $\bar{\mathbf{x}}^\nu$, and $\bar{\mathbf{y}}^\nu \in \mathcal{X}$ such that:

$$\begin{aligned}&\|P_{\mathcal{X}(\bar{\mathbf{y}}^\nu)}(\mathbf{w}_{\rho^\nu}(\hat{\mathbf{x}}(\bar{\mathbf{x}}^\nu),\bar{\mathbf{x}}^\nu)) - P_{\mathcal{X}(\bar{\mathbf{x}}^\nu)}(\mathbf{w}_{\rho^\nu}(\hat{\mathbf{x}}(\bar{\mathbf{x}}^\nu),\bar{\mathbf{x}}^\nu))\| \\ &\qquad > \beta^\nu\|\bar{\mathbf{y}}^\nu - \bar{\mathbf{x}}^\nu\|^{\frac{1}{2}}.\end{aligned} \tag{43}$$

Furthermore, in view of the compactness of $\mathcal{X}$, denoting by $D_{\mathcal{X}}$ the (finite) diameter of $\mathcal{X}$, the LHS of (43) can be bounded by

$$\begin{aligned}D_{\mathcal{X}} \ge\ &\|P_{\mathcal{X}(\bar{\mathbf{y}}^\nu)}(\mathbf{w}_{\rho^\nu}(\hat{\mathbf{x}}(\bar{\mathbf{x}}^\nu),\bar{\mathbf{x}}^\nu)) \\ &- P_{\mathcal{X}(\bar{\mathbf{x}}^\nu)}(\mathbf{w}_{\rho^\nu}(\hat{\mathbf{x}}(\bar{\mathbf{x}}^\nu),\bar{\mathbf{x}}^\nu))\|.\end{aligned} \tag{44}$$

Suppose without loss of generality that $\beta^\nu \to +\infty$, $\bar{\rho}^\nu \downarrow 0$, and $\bar{\mathbf{x}}^\nu \underset{\mathcal{N}}{\to} \bar{\mathbf{x}} \in \mathcal{X}(\bar{\mathbf{x}}) \subseteq \mathcal{X}$ and $\bar{\mathbf{y}}^\nu \underset{\mathcal{N}}{\to} \bar{\mathbf{y}} \in \mathcal{X}(\bar{\mathbf{y}}) \subseteq \mathcal{X}$, possibly on a suitable subsequence $\mathcal{N}$ [recall that $\bar{\mathbf{x}}^\nu \in \mathcal{X}(\bar{\mathbf{x}}^\nu)$ and $\bar{\mathbf{y}}^\nu \in \mathcal{X}(\bar{\mathbf{y}}^\nu)$]. From (43) and (44), we obtain

$$D_{\mathcal{X}} \ge \limsup_{\nu\to+\infty} \beta^\nu\|\bar{\mathbf{y}}^\nu - \bar{\mathbf{x}}^\nu\|^{\frac{1}{2}},$$

which, in turn, considering that $\beta^\nu \to \infty$ and $\|\bar{\mathbf{y}}^\nu - \bar{\mathbf{x}}^\nu\|^{\frac{1}{2}} \ge 0$, implies

$$\lim_{\nu\to+\infty} \|\bar{\mathbf{y}}^\nu - \bar{\mathbf{x}}^\nu\|^{\frac{1}{2}} = 0. \tag{45}$$

Then, it must be $\bar{\mathbf{x}} = \bar{\mathbf{y}}$.

Invoking now the continuity of $\hat{\mathbf{x}}(\bullet)$ at $\bar{\mathbf{x}}$ [cf. Lemma 10(iii)] and $\nabla_{\mathbf{x}}\tilde{U}(\bullet;\bullet)$ on $\mathcal{K} \times \mathcal{X}$ [cf. B3], we have

$$\mathbf{w}_{\rho^\nu}(\hat{\mathbf{x}}(\bar{\mathbf{x}}^\nu),\bar{\mathbf{x}}^\nu) = \hat{\mathbf{x}}(\bar{\mathbf{x}}^\nu) - \rho^\nu\nabla_{\mathbf{x}}\tilde{U}(\hat{\mathbf{x}}(\bar{\mathbf{x}}^\nu),\bar{\mathbf{x}}^\nu) \underset{\mathcal{N}}{\to} \hat{\mathbf{x}}(\bar{\mathbf{x}}). \tag{46}$$

Therefore, for every $\hat{\beta} > 0$ and neighborhoods $\mathcal{V}_{\bar{\mathbf{x}}}$ and $\mathcal{W}_{\hat{\mathbf{x}}(\bar{\mathbf{z}})}$, there exists a sufficiently large ν such that (43) holds with $\beta^\nu > \hat{\beta}$ (recall that $\beta^\nu \to +\infty$), $\bar{\mathbf{x}}^\nu$, $\bar{\mathbf{y}}^\nu \in \mathcal{V}_{\bar{\mathbf{x}}} \cap \mathcal{X}$ [due to (45)], and $\mathbf{w}_{\rho^\nu}(\hat{\mathbf{x}}(\bar{\mathbf{x}}^\nu),\bar{\mathbf{x}}^\nu) \in \mathcal{W}_{\hat{\mathbf{x}}(\bar{\mathbf{z}})}$ [due to (46)]; this is in contradiction with (42). ■

We recall that the assumption on the regularity of $\hat{\mathbf{x}}(\bar{\mathbf{x}}) \in \mathcal{X}(\bar{\mathbf{x}})$ for every $\bar{\mathbf{x}} \in \mathcal{X}$, as required in Lemma 13, is implied by Assumption 5.

Theorem 14. The last theorem of this section provides technical conditions under which a(t least one) regular limit point of the sequence generated by Algorithm 1 is a stationary solution of the original nonconvex problem $\mathcal{P}$.

Theorem 14: Let $\{\mathbf{x}^\nu\}$ be the sequence generated by Algorithm 1 under Assumptions 1–3 and 5. The following hold.

(a) Suppose

$$\liminf_{\nu\to\infty} \|\hat{\mathbf{x}}(\mathbf{x}^\nu) - \mathbf{x}^\nu\| = 0. \tag{47}$$

Then, at least one regular limit point of $\{\mathbf{x}^\nu\}$ is a stationary solution of $\mathcal{P}$.

(b) Suppose

$$\lim_{\nu\to\infty} \|\hat{\mathbf{x}}(\mathbf{x}^\nu) - \mathbf{x}^\nu\| = 0. \tag{48}$$

Then, every regular limit point of $\{\mathbf{x}^\nu\}$ is a stationary solution of $\mathcal{P}$.

Proof: We prove only (a); (b) follows applying the result in (a) to every convergent subsequence of $\{\mathbf{x}^\nu\}$.

Let $\bar{\mathbf{x}}$ be a regular accumulation point of the subsequence $\{\mathbf{x}^\nu\}_{\mathcal{N}}$ of $\{\mathbf{x}^\nu\}$ satisfying (47); thus, there exists $\mathcal{N}' \subseteq \mathcal{N}$ such

that $\lim_{\mathcal{N}'\ni\nu\to\infty}\mathbf{x}^\nu=\bar{\mathbf{x}}$. We show next that $\bar{\mathbf{x}}$ is a KKT point of the original problem. Let $\bar{J}$ and J^ν be the following sets:

$$\bar{J}\triangleq\{j\in[1,\ldots,m]:\ g_j(\bar{\mathbf{x}})=0\},$$
$$J^\nu\triangleq\{j\in[1,\ldots,m]:\ \tilde{g}_j(\hat{\mathbf{x}}(\mathbf{x}^\nu);\mathbf{x}^\nu)=0\}$$

with $\nu\in\mathcal{N}'$. Using $\lim_{\mathcal{N}'\ni\nu\to\infty}\|\hat{\mathbf{x}}(\mathbf{x}^\nu)-\mathbf{x}^\nu\|=0$ [cf. (47)] along with the continuity of $\tilde{g}_j$, by C2, we have

$$\lim_{\mathcal{N}'\ni\nu\to\infty}\tilde{g}_j(\hat{\mathbf{x}}(\mathbf{x}^\nu);\mathbf{x}^\nu)=\tilde{g}_j(\bar{\mathbf{x}};\bar{\mathbf{x}})=g_j(\bar{\mathbf{x}}),\ j=1,\ldots,m. \tag{49}$$

The limit above implies that there exists a positive integer $\tilde{\nu}\in\mathcal{N}'$ such that

$$J^\nu\subseteq\bar{J},\ \forall\nu\geq\tilde{\nu}\text{ and }\nu\in\mathcal{N}'. \tag{50}$$

Since the functions $\nabla_{\mathbf{x}}\tilde{U}$ and $\nabla_{\mathbf{x}}\tilde{g}_j$ are continuous, we get, by B2,

$$\lim_{\mathcal{N}'\ni\nu\to\infty}\nabla_{\mathbf{x}}\tilde{U}(\hat{\mathbf{x}}(\mathbf{x}^\nu);\mathbf{x}^\nu)=\nabla_{\mathbf{x}}\tilde{U}(\bar{\mathbf{x}};\bar{\mathbf{x}})=\nabla U(\bar{\mathbf{x}}), \tag{51}$$

and, for $j=1,\ldots,m$, by C5,

$$\lim_{\mathcal{N}'\ni\nu\to\infty}\nabla_{\mathbf{x}}\tilde{g}_j(\hat{\mathbf{x}}(\mathbf{x}^\nu);\mathbf{x}^\nu)=\nabla_{\mathbf{x}}\tilde{g}_j(\bar{\mathbf{x}};\bar{\mathbf{x}})=\nabla g_j(\bar{\mathbf{x}}). \tag{52}$$

We claim now that for sufficiently large $\nu\in\mathcal{N}'$, the MFCQ holds at $\hat{\mathbf{x}}(\mathbf{x}^\nu)\in\mathcal{X}(\mathbf{x}^\nu)$. Assume by contradiction that the following implication does not hold for infinitely many $\nu\in\mathcal{N}'$:

$$\underbrace{\begin{array}{c}-\sum_{j\in J^\nu}\mu_j^\nu\nabla_{\mathbf{x}}\tilde{g}_j(\hat{\mathbf{x}}(\mathbf{x}^\nu);\mathbf{x}^\nu)\in N_{\mathcal{K}}(\hat{\mathbf{x}}(\mathbf{x}^\nu))\\ \mu_j^\nu\geq 0,\ \forall j\in J^\nu,\\ \Downarrow\end{array}}_{}\quad \mu_j^\nu=0,\ \forall j\in J^\nu. \tag{53}$$

It follows that a nonempty index set $\bar{\bar{J}}\subseteq\bar{J}$ exists such that, after a suitable renumeration, for every $\nu\in\mathcal{N}'$, we must have

$$\begin{array}{c}-\sum_{j\in\bar{\bar{J}}}\mu_j^\nu\nabla_{\mathbf{x}}\tilde{g}_j(\hat{\mathbf{x}}(\mathbf{x}^\nu);\mathbf{x}^\nu)\in N_{\mathcal{K}}\left(\hat{\mathbf{x}}(\mathbf{x}^\nu)\right)\\ \mu_j^\nu\geq 0,\ \forall j\in\bar{\bar{J}}\\ \sum_{j\in\bar{\bar{J}}}\mu_j^\nu=1.\end{array} \tag{54}$$

We may assume without loss of generality that, for each $j\in\bar{\bar{J}}$, the sequence $\{\mu_j^\nu\}$ converges to a limit $\bar{\mu}_j$ such that $\sum_{j\in\bar{\bar{J}}}\bar{\mu}_j=1$. In view of the inclusion $\bar{\bar{J}}\subseteq\bar{J}$, by taking the limit $\mathcal{N}'\ni\nu\to\infty$ in (54), and invoking (52) along with the outer semicontinuity of the mapping $N_{\mathcal{K}}(\bullet)$ [39, Prop. 6.6], we get

$$\begin{array}{c}-\sum_{j\in\bar{\bar{J}}}\bar{\mu}_j\nabla_{\mathbf{x}}g_j(\bar{\mathbf{x}})\in N_{\mathcal{K}}\left(\bar{\mathbf{x}}\right)\\ \bar{\mu}_j\geq 0,\ \forall j\in\bar{\bar{J}}\\ \sum_{j\in\bar{\bar{J}}}\bar{\mu}_j=1,\end{array} \tag{55}$$

in contradiction with the regularity of $\bar{\mathbf{x}}$ [the MFCQ holds at $\bar{\mathbf{x}}$, see (1)]. Therefore, (53) must hold for sufficiently large $\nu\in\mathcal{N}'$, implying that the KKT system of problem $\mathcal{P}_{\mathbf{x}^\nu}$ has a solution for every sufficiently large $\nu\in\mathcal{N}'$: thus, there exist $(\mu_j^\nu)_{j=1}^m$ such that

$$\begin{array}{c}-\left[\nabla_{\mathbf{x}}\tilde{U}(\hat{\mathbf{x}}(\mathbf{x}^\nu);\mathbf{x}^\nu)+\sum_{j=1}^m\mu_j^\nu\nabla_{\mathbf{x}}\tilde{g}_j(\hat{\mathbf{x}}(\mathbf{x}^\nu);\mathbf{x}^\nu)\right]\\ \in N_{\mathcal{K}}(\hat{\mathbf{x}}(\mathbf{x}^\nu))\\ 0\leq\mu_j^\nu\perp\tilde{g}_j(\hat{\mathbf{x}}(\mathbf{x}^\nu);\mathbf{x}^\nu)\leq 0,\ \ j=1,\ldots,m.\end{array} \tag{56}$$

Note that by (50) and the complementarity slackness in (56), $\mu_j^\nu=0$ for all $j\notin\bar{J}$ and large $\nu\in\mathcal{N}'$. Moreover, the sequence of nonnegative multipliers $\{\boldsymbol{\mu}^\nu\triangleq(\mu_j^\nu)_{j\in\bar{J}}\}_{\nu\in\mathcal{N}'}$ must be bounded, as shown next. Suppose by contradiction that $\lim_{\mathcal{N}'\ni\nu\to\infty}\|\boldsymbol{\mu}^\nu\|=+\infty$ for some $\{\hat{\mathbf{x}}(\mathbf{x}^\nu)\}_{\mathcal{N}'}$ (possibly over a subsequence). Dividing both sides of (56) by $\|\boldsymbol{\mu}^\nu\|$ and taking the limit $\mathcal{N}'\ni\nu\to\infty$, one would get

$$\begin{array}{c}-\sum_{j\in\bar{J}}\bar{\bar{\mu}}_j\nabla g_j(\bar{\mathbf{x}})\in N_{\mathcal{K}}(\bar{\mathbf{x}})\\ 0\leq\bar{\bar{\mu}}_j\perp g_j(\bar{\mathbf{x}})\leq 0,\ \ j\in\bar{J},\end{array} \tag{57}$$

for some $\bar{\bar{\boldsymbol{\mu}}}\triangleq(\bar{\bar{\mu}}_j)_{j\in\bar{J}}\neq\mathbf{0}$, in contradiction with (1).

Therefore, $\{\boldsymbol{\mu}^\nu\triangleq(\mu_j^\nu)_{j\in\bar{J}}\}_{\nu\in\mathcal{N}'}$ must have a limit; let us denote by $(\bar{\mu}_j)_{j\in\bar{J}}$ such a limit (after a suitable renumeration). Taking the limit $\mathcal{N}'\ni\nu\to\infty$ in (56), and using (51) and (52) along with the outer semicontinuity of the mapping $N_{\mathcal{K}}(\bullet)$, we get

$$\begin{array}{c}-\left[\nabla U(\bar{\mathbf{x}})+\sum_{j\in\bar{J}}\bar{\mu}_j\nabla g_j(\bar{\mathbf{x}})\right]\in N_{\mathcal{K}}(\bar{\mathbf{x}})\\ 0\leq\bar{\mu}_j\perp g_j(\bar{\mathbf{x}})\leq 0,\quad j\in\bar{J}.\end{array} \tag{58}$$

It follows from (58) that $\bar{\mathbf{x}}$ is a stationary solution of the original problem $\mathcal{P}$. ■

B. Proof of Theorem 2

Proof of statement (a). It follows from Lemma 9.

Proof of statement (b). Invoking Theorem 14(b), it is sufficient to show that (48) in Theorem 14 is satisfied.

By the descent lemma [21, Propo. A.24] and Step 3 of Algorithm 1, we get:

$$\begin{aligned}U(\mathbf{x}^{\nu+1})\leq\ & U(\mathbf{x}^\nu)+\gamma^\nu\nabla U(\mathbf{x}^\nu)^T\left(\hat{\mathbf{x}}(\mathbf{x}^\nu)-\mathbf{x}^\nu\right)\\ &+\tfrac{(\gamma^\nu)^2L_{\nabla U}}{2}\|\hat{\mathbf{x}}(\mathbf{x}^\nu)-\mathbf{x}^\nu\|^2.\end{aligned}$$

Invoking (31) in Lemma 10, we obtain

$$U(\mathbf{x}^{\nu+1})\leq U(\mathbf{x}^\nu)-\gamma^\nu\left(c_{\tilde{U}}-\frac{\gamma^\nu L_{\nabla U}}{2}\right)\|\hat{\mathbf{x}}(\mathbf{x}^\nu)-\mathbf{x}^\nu\|^2. \tag{59}$$

Since $0<\inf_\nu\gamma^\nu\leq\sup_\nu\gamma^\nu\leq\gamma^{\max}\leq 1$ and $2c_{\tilde{U}}>\gamma^{\max}L_{\nabla U}$, we deduce from (59) that either $U(\mathbf{x}^\nu)\to-\infty$ or $\{U(\mathbf{x}^\nu)\}$ converges to a finite value and

$$\lim_{\nu\to\infty}\|\hat{\mathbf{x}}(\mathbf{x}^\nu)-\mathbf{x}^\nu\|=0. \tag{60}$$

By assumption A4, $\{U(\mathbf{x}^\nu)\}$ is convergent and the sequence $\{\mathbf{x}^\nu\}\subseteq\mathcal{X}$ [Lemma 9(iii)] is bounded. Therefore, (60) holds true and $\{\mathbf{x}^\nu\}$ has a limit point in $\mathcal{X}$. By Theorem 14(b) and (60), statement (b) of the theorem follows readily. Finally, by (59), $U(\mathbf{x}^\nu)$ is a decreasing sequence: hence, no limit point of $\{\mathbf{x}^\nu\}$ can be a local maximum of U.

Proof of statement (c). Invoking Theorem 14(a), it is sufficient to show that (47) in Theorem 14 is satisfied. Following

the same steps as in the proof of statement (b), by (59) and $\gamma^\nu \to 0$, for $\nu \geq \bar{\nu}$ sufficiently large, there exists a positive constant ζ such that:

$$U(\mathbf{x}^{\nu+1}) \leq U(\mathbf{x}^\nu) - \gamma^\nu \zeta \|\hat{\mathbf{x}}(\mathbf{x}^\nu) - \mathbf{x}^\nu\|^2, \tag{61}$$

which, again, by A4, leads to

$$\lim_{\nu\to\infty} \sum_{t=\bar{\nu}}^{\nu} \gamma^t \|\hat{\mathbf{x}}(\mathbf{x}^t) - \mathbf{x}^t\|^2 < +\infty. \tag{62}$$

The desired result (47) follows from (62) and $\sum_{\nu=0}^{\infty} \gamma^\nu = +\infty$. Similarly to the previous case, by (62), eventually $U(\mathbf{x}^\nu)$ is a decreasing sequence: thus, no limit point of $\{\mathbf{x}^\nu\}$ can be a local maximum of U.

Suppose now that Assumption 4 holds. By Theorem 14(b) it is sufficient to prove that (48) holds true. For notational simplicity, we set $\Delta\hat{\mathbf{x}}(\mathbf{x}^\nu) \triangleq \hat{\mathbf{x}}(\mathbf{x}^\nu) - \mathbf{x}^\nu$. We already proved that $\liminf_\nu \|\Delta\hat{\mathbf{x}}(\mathbf{x}^\nu)\| = 0$; therefore, (48) holds if $\limsup_\nu \|\Delta\hat{\mathbf{x}}(\mathbf{x}^\nu)\| = 0$, as stated next.

First of all, note that, by Assumption 4, Lemma 13 and, by consequence, Lemma 11 hold true; therefore, there exists $\tilde{\rho} > 0$ such that (cf. Lemma 11)

$$\|\hat{\mathbf{x}}(\mathbf{x}^\nu) - \hat{\mathbf{x}}(\mathbf{x}^t)\| \leq \eta_\rho \|\mathbf{x}^\nu - \mathbf{x}^t\| + \theta_\rho \|\mathbf{x}^\nu - \mathbf{x}^t\|^{\frac{1}{2}}, \tag{63}$$

for any $\nu, t \geq 1$ and $\rho \in (0, \tilde{\rho}]$, with η_ρ and θ_ρ defined in (36) (cf. Lemma 11).

Suppose by contradiction that $\limsup_\nu \|\Delta\hat{\mathbf{x}}(\mathbf{x}^\nu)\| > 0$. Then, there exists $\delta > 0$ such that $\|\Delta\hat{\mathbf{x}}(\mathbf{x}^\nu)\| > 2\delta + \sqrt{\delta/2}$ for infinitely many ν, and also $\|\Delta\hat{\mathbf{x}}(\mathbf{x}^\nu)\| < \delta + \sqrt{\delta/2}$ for infinitely many ν. Thus, there exists an infinite subset of indices $\mathcal{N}$ such that, for each $\nu \in \mathcal{N}$ and some $i_\nu > \nu$, the following hold:

$$\|\Delta\hat{\mathbf{x}}(\mathbf{x}^\nu)\| < \delta + \sqrt{\delta/2}, \quad \|\Delta\hat{\mathbf{x}}(\mathbf{x}^{i_\nu})\| > 2\delta + \sqrt{\delta/2} \tag{64}$$

and, in case $i_\nu > \nu + 1$,

$$\delta + \sqrt{\delta/2} \leq \|\Delta\hat{\mathbf{x}}(\mathbf{x}^j)\| \leq 2\delta + \sqrt{\delta/2}, \quad \nu < j < i_\nu. \tag{65}$$

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

$$\begin{aligned} \delta < & \|\Delta\hat{\mathbf{x}}(\mathbf{x}^{i_\nu})\| - \|\Delta\hat{\mathbf{x}}(\mathbf{x}^\nu)\| \\ \leq & \|\hat{\mathbf{x}}(\mathbf{x}^{i_\nu}) - \hat{\mathbf{x}}(\mathbf{x}^\nu)\| + \|\mathbf{x}^{i_\nu} - \mathbf{x}^\nu\| \\ \overset{(a)}{\leq} & (1+\eta_\rho)\|\mathbf{x}^{i_\nu} - \mathbf{x}^\nu\| + \theta_\rho \|\mathbf{x}^{i_\nu} - \mathbf{x}^\nu\|^{\frac{1}{2}} \\ \overset{(b)}{\leq} & (1+\eta_\rho)\left(2\delta + \sqrt{\delta/2}\right) \sum_{t=\nu}^{i_\nu - 1} \gamma^t \\ & + \theta_\rho \left(2\delta + \sqrt{\delta/2}\right)^{\frac{1}{2}} \left(\sum_{t=\nu}^{i_\nu - 1} \gamma^t\right)^{\frac{1}{2}}, \end{aligned} \tag{66}$$

where (a) is due to (63) and (b) comes from the triangle inequality and the updating rule of the algorithm. It follows from (64) and (66) that

$$\begin{aligned} \liminf_\nu \Big[& (1+\eta_\rho)\left(2\delta + \sqrt{\delta/2}\right) \sum_{t=\nu}^{i_\nu - 1} \gamma^t \\ & + \theta_\rho \left(2\delta + \sqrt{\delta/2}\right)^{\frac{1}{2}} \left(\sum_{t=\nu}^{i_\nu - 1} \gamma^t\right)^{\frac{1}{2}} \Big] > 0. \end{aligned} \tag{67}$$

We now prove that $\|\Delta\hat{\mathbf{x}}(\mathbf{x}^\nu)\| \geq \delta/2$ for sufficiently large $\nu \in \mathcal{N}$. Reasoning as in (66), we have

$$\begin{aligned} \|\Delta\hat{\mathbf{x}}\,(\mathbf{x}^{\nu+1})\| - \|\Delta\hat{\mathbf{x}}(\mathbf{x}^\nu)\| \leq & \\ (1+\eta_\rho)\|\mathbf{x}^{\nu+1} - \mathbf{x}^\nu\| + \theta_\rho \|\mathbf{x}^{\nu+1} - \mathbf{x}^\nu\|^{\frac{1}{2}} & \\ \leq (1+\eta_\rho)\gamma^\nu \|\Delta\hat{\mathbf{x}}(\mathbf{x}^\nu)\| + \theta_\rho (\gamma^\nu)^{1/2} \|\Delta\hat{\mathbf{x}}(\mathbf{x}^\nu)\|^{\frac{1}{2}}, & \end{aligned} \tag{68}$$

for any given ν. For large $\nu \in \mathcal{N}$, so that $(1+\eta_\rho)\gamma^\nu \delta/2 + \theta_\rho(\gamma^\nu \delta/2)^{\frac{1}{2}} < \delta/2 + \sqrt{\delta/2}$, suppose by contradiction that $\|\Delta\hat{\mathbf{x}}(\mathbf{x}^\nu)\| < \delta/2$; this would give $\|\Delta\hat{\mathbf{x}}(\mathbf{x}^{\nu+1})\| < \delta + \sqrt{\delta/2}$ and condition (65) (or, in case, (64)) would be violated. Then, it must be

$$\|\Delta\hat{\mathbf{x}}(\mathbf{x}^\nu)\| \geq \delta/2. \tag{69}$$

Using (69), we can show now that (67) is in contradiction with the convergence of $\{U(\mathbf{x}^\nu)\}$. By (61), (possibly over a subsequence) for sufficiently large $\nu \in \mathcal{N}$, we have

$$\begin{aligned} U(\mathbf{x}^{i_\nu}) \leq & \; U(\mathbf{x}^\nu) - \zeta \textstyle\sum_{t=\nu}^{i_\nu - 1} \gamma^t \|\Delta\hat{\mathbf{x}}(\mathbf{x}^t)\|^2 \\ < & \; U(\mathbf{x}^\nu) - \zeta \tfrac{\delta^2}{4} \textstyle\sum_{t=\nu}^{i_\nu - 1} \gamma^t, \end{aligned} \tag{70}$$

where, in the last inequality, we have used (65) (or, in case, (64)) and (69). Thus, since $U(\mathbf{x}^\nu)$ converges, (70) implies $\lim_{\nu\in\mathcal{N}} \sum_{t=\nu}^{i_\nu - 1} \gamma^t = 0$, in contradiction with (67).

Remark 15: As we already mentioned in Subsection III-A3, in [11] it is shown that, in the specific case of a *strongly convex* U, $\tilde{U} = U$, and $\mathcal{K} = \mathbb{R}^n$, one can choose $\gamma^\nu = 1$ at every iteration and prove the stationarity of every limit point of the sequence generated by Algorithm 1 (assuming regularity). For completeness we sketch how this result can be readily obtained using our framework (and actually slightly improved on by also considering the case in which $\mathcal{K}$ is not necessarily $\mathbb{R}^n$). The proof is based on Theorem 14(b) and a result in [11]. By Theorem 14(b), it is enough to show that (48) holds. But (48) does indeed hold because of the strong convexity of U, as shown at the beginning of Proposition 3.2 in [11]. Note that the strong convexity of U plays here a fundamental role and that, once we remove this restrictive assumption, things get considerably more difficult, as clearly shown by the complexity of the proof of Theorem 2.

C. Proof of Lemma 3

Lemma 3 is similar to [14, Lemma 2]. Since the proof of [14, Lemma 2] is not available, for completeness, we report next the proof of Lemma 3, which follows the same steps as that of [14, Lemma 2].

(a) It is a consequence of Danskin's theorem [21, Prop. A.43].

(b) The statement follows from the uniform Lipschitz continuity of $\widehat{\mathbf{x}}(\bullet; \mathbf{x}^\nu)$ on $\mathbb{R}^m_+$ with constant $L_{\nabla d}$, which is proved next. For notational simplicity, let us write $\widehat{\mathbf{x}}_{\boldsymbol{\lambda}} \triangleq \widehat{\mathbf{x}}(\boldsymbol{\lambda}; \mathbf{x}^\nu)$ and $\widehat{\mathbf{x}}_{\boldsymbol{\lambda}'} \triangleq \widehat{\mathbf{x}}(\boldsymbol{\lambda}'; \mathbf{x}^\nu)$. Defining $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) \triangleq \sum_{i=1}^I \left(\tilde{U}_i(\mathbf{x}_i; \mathbf{x}^\nu) + \boldsymbol{\lambda}^T \tilde{\mathbf{g}}^i(\mathbf{x}_i; \mathbf{x}^\nu)\right)$, we

have, by the minimum principle,

$$\begin{aligned}\left(\widehat{\mathbf{x}}_{\boldsymbol{\lambda}'}-\widehat{\mathbf{x}}_{\boldsymbol{\lambda}}\right)^T\nabla_{\mathbf{x}}\mathcal{L}\left(\widehat{\mathbf{x}}_{\boldsymbol{\lambda}},\boldsymbol{\lambda}\right) &\geq 0\\ \left(\widehat{\mathbf{x}}_{\boldsymbol{\lambda}}-\widehat{\mathbf{x}}_{\boldsymbol{\lambda}'}\right)^T\nabla_{\mathbf{x}}\mathcal{L}\left(\widehat{\mathbf{x}}_{\boldsymbol{\lambda}'},\boldsymbol{\lambda}'\right) &\geq 0.\end{aligned}$$

Adding the two inequalities above and summing and subtracting $\nabla_{\mathbf{x}}\mathcal{L}\left(\widehat{\mathbf{x}}_{\boldsymbol{\lambda}},\boldsymbol{\lambda}'\right)$, we obtain

$$\begin{aligned}&c_{\tilde{U}}\cdot\|\widehat{\mathbf{x}}_{\boldsymbol{\lambda}}-\widehat{\mathbf{x}}_{\boldsymbol{\lambda}'}\|^2\\ &\quad\leq\left(\widehat{\mathbf{x}}_{\boldsymbol{\lambda}'}-\widehat{\mathbf{x}}_{\boldsymbol{\lambda}}\right)^T\left[\nabla_{\mathbf{x}}\mathcal{L}\left(\widehat{\mathbf{x}}_{\boldsymbol{\lambda}},\boldsymbol{\lambda}\right)-\nabla_{\mathbf{x}}\mathcal{L}\left(\widehat{\mathbf{x}}_{\boldsymbol{\lambda}},\boldsymbol{\lambda}'\right)\right]\\ &\quad=\left(\widehat{\mathbf{x}}_{\boldsymbol{\lambda}}-\widehat{\mathbf{x}}_{\boldsymbol{\lambda}'}\right)^T\left[\nabla_{\mathbf{x}}\mathcal{L}\left(\widehat{\mathbf{x}}_{\boldsymbol{\lambda}},\boldsymbol{\lambda}'\right)-\nabla_{\mathbf{x}}\mathcal{L}\left(\widehat{\mathbf{x}}_{\boldsymbol{\lambda}},\boldsymbol{\lambda}\right)\right],\end{aligned}\tag{71}$$

where, in the first inequality, we used the uniform strong convexity of $\mathcal{L}\left(\bullet,\boldsymbol{\lambda}'\right)$. Hence, we have

$$\begin{aligned}c_{\tilde{U}}\cdot\|\widehat{\mathbf{x}}_{\boldsymbol{\lambda}}-\widehat{\mathbf{x}}_{\boldsymbol{\lambda}'}\| &\leq \sum_{j=1}^m\left|\lambda_j'-\lambda_j\right|\|\nabla_{\mathbf{x}}\tilde{g}_j\left(\widehat{\mathbf{x}}_{\boldsymbol{\lambda}};\mathbf{x}^\nu\right)\|\\ &\overset{(a)}{\leq}\sum_{j=1}^m\left|\lambda_j'-\lambda_j\right|L_{\tilde{g}}=\left\|\boldsymbol{\lambda}'-\boldsymbol{\lambda}\right\|_1\cdot L_{\tilde{g}}\\ &\leq L_{\tilde{g}}\sqrt{m}\left\|\boldsymbol{\lambda}'-\boldsymbol{\lambda}\right\|_2,\end{aligned}\tag{72}$$

where (a) follows from the uniform Lipschitz continuity of $\tilde{\mathbf{g}}$. The inequality above proves the Lipschitz property of $\widehat{\mathbf{x}}(\bullet;\mathbf{x}^\nu)$.

REFERENCES

[1] G. Scutari, F. Facchinei, L. Lampariello, and P. Song, "Parallel and distributed methods for nonconvex optimization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Florence, Italy, May 4–9, 2014, pp. 840–844.

[2] P. Song, G. Scutari, F. Facchinei, and L. Lampariello, "D3m: Distributed multi-cell multigroup multicasting," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, Shanghai, China, Mar. 20–25, 2016, pp. 3741–3745.

[3] G. Scutari, F. Facchinei, L. Lampariello, and P. Song, "Parallel and distributed methods for nonconvex optimization—Part I: Theory," Oct. 2014, arXiv:1410.4754.

[4] R. Byrd, J. Nocedal, and R. Waltz, "Feasible interior methods using slacks for nonlinear optimization," *Comput. Opt. Appl.*, vol. 26, no. 1, pp. 35–61, 2003.

[5] A. Fiacco and G. M. Cormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*. Hoboken, NJ, USA: Wiley, 1968.

[6] C. Lawrence and A. Tits, "A computationally efficient feasible sequential quadratic programming algorithm," *SIAM J. Optim.*, vol. 11, no. 4, pp. 1092–1118, 2001.

[7] M. Ferris and O. Mangasarian, "Parallel variable distribution," *SIAM J. Optim.*, vol. 4, no. 4, pp. 815–832, 1994.

[8] C. Sagastizábal and M. Solodov, "Parallel variable distribution for constrained optimization," *Comput. Optim. Appl.*, vol. 22, no. 1, pp. 111–131, 2002.

[9] M. Solodov, "On the convergence of constrained parallel variable distribution algorithms," *SIAM J. Optim.*, vol. 8, no. 1, pp. 187–196, 1998.

[10] B. Marks and G. Wright, "A general inner approximation algorithm for nonconvex mathematical programs," *Oper. Res.*, vol. 26, no. 4, pp. 681–683, 1978.

[11] A. Beck, A. Ben-Tal, and L. Tetruashvili, "A sequential parametric convex approximation method with applications to nonconvex truss topology design problems," *J. Global Optim.*, vol. 47, no. 1, pp. 29–51, 2010. doi: 10.1007/s10898-009-9456-5.

[12] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM J. Optim.*, vol. 23, no. 2, pp. 1126–1153, 2013.

[13] G. Scutari, F. Facchinei, P. Song, D. P. Palomar, and J.-S. Pang, "Decomposition by partial linearization: Parallel optimization of multiuser systems," *IEEE Trans. Signal Process.*, vol. 62, no. 3, pp. 641–656, Feb. 2014.

[14] A. Alvarado, G. Scutari, and J.-S. Pang, "A new decomposition method for multiuser dc-programming and its applications," *IEEE Trans. Signal Process.*, vol. 62, no. 11, pp. 2984–2998, Mar. 2014.

[15] T. Quoc and M. Diehl, "Sequential convex programming methods for solving nonlinear optimization problems with dc constraints," 2011, arXiv:1107.5841.

[16] C. Fleury, "CONLIN: An efficient dual optimizer based on convex approximation concepts," *Struct. Opt.*, vol. 1, no. 2, pp. 81–89, 1989.

[17] K. Svanberg, "The method of moving asymptotes-A new method for structural optimization," *Int. J. Numer. Methods Eng.*, vol. 24, no. 2, pp. 359–373, 1987.

[18] K. Svanberg, "A class of globally convergent optimization methods based on conservative convex separable approximations," *SIAM J. Opt.*, vol. 12, no. 2, pp. 555–573, 2002.

[19] G. Scutari, F. Facchinei, L. Lampariello, S. Sardellitti, and P. Song, "Parallel and distributed methods for constrained nonconvex optimization—Part II: Applications in communications and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 8, pp. 1945–1960, 2017.

[20] D. P. Palomar and M. Chiang, "Alternative distributed algorithms for network utility maximization: Framework and applications," *IEEE Trans. Autom. Control*, vol. 52, no. 12, pp. 2254–2269, Dec. 2007.

[21] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, 2nd ed. Belmont, MA, USA: Athena Scientific, 1989.

[22] K. Phan, S. Vorobyov, C. Telambura, and T. Le-Ngoc, "Power control for wireless cellular systems via D.C. programming," in *Proc. IEEE/SP 14th Workshop Stat. Signal Process.*, 2007, pp. 507–511.

[23] H. Al-Shatri and T. Weber, "Achieving the maximum sum rate using D.C. programming in cellular networks," *IEEE Trans. Signal Process.*, vol. 60, no. 3, pp. 1331–1341, Mar. 2012.

[24] N. Vucic, S. Shi, and M. Schubert, "DC programming approach for resource allocation in wireless networks," in *Proc. 8th Int. Symp. Model. Optim. Mobile, Ad Hoc Wireless Netw.*, 2010, pp. 380–386.

[25] M. Chiang, C. Tan, D. Palomar, D. O. Neill, and D. Julian, "Power control by geometric programming," *IEEE Trans. Wireless Commun.*, vol. 6, no. 7, pp. 2640–2651, Jul. 2007.

[26] A. Khabbazibasmenj, F. Roemer, S. Vorobyov, and M. Haardt, "Sum-rate maximization in two-way AF MIMO relaying: Polynomial time solutions to a class of DC programming problems," *IEEE Trans. Signal Process.*, vol. 60, no. 10, pp. 5478–5493, Oct. 2012.

[27] Y. Xu, T. Le-Ngoc, and S. Panigrahi, "Global concave minimization for optimal spectrum balancing in multi-user DSL networks," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 2875–2885, Jul. 2008.

[28] P. Tsiaflakis, M. Diehl, and M. Moonen, "Distributed spectrum management algorithms for multiuser DSL networks," *IEEE Trans. Signal Process.*, vol. 56, no. 10, pp. 4825–4843, Oct. 2008.

[29] D. Schmidt, C. Shi, R. Berry, M. Honig, and W. Utschick, "Distributed resource allocation schemes: Pricing algorithms for power control and beamformer design in interference networks," *IEEE Signal Process. Mag.*, vol. 26, no. 5, pp. 53–63, Sep. 2009.

[30] R. Mochaourab, P. Cao, and E. Jorswieck, "Alternating rate profile optimization in single stream MIMO interference channels," *IEEE Signal Process. Lett.*, vol. 21, no. 2, pp. 221–224, Feb. 2014.

[31] J. Qiu, R. Zhang, Z.-Q. Luo, and S. Cui, "Optimal distributed beamforming for MISO interference channels," *IEEE Trans. Signal Process.*, vol. 59, no. 11, pp. 5638–5643, Nov. 2011.

[32] S.-J. Kim and G. B. Giannakis, "Optimal resource allocation for MIMO ad hoc cognitive radio networks," *IEEE Trans. Inf. Theory*, vol. 57, no. 5, pp. 3117–3131, May 2011.

[33] Y. Zhang, E. Dall'Anese, and G. B. Giannakis, "Distributed optimal beamformers for cognitive radios robust to channel uncertainties," *IEEE Trans. Signal Process.*, vol. 60, no. 12, pp. 6495–6508, Dec. 2012.

[34] Y. Yang, G. Scutari, and D. Palomar, "Robust MIMO cognitive radio under interference temperature constraints," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 11, pp. 2465–2482, Nov. 2013.

[35] N. D. Sidiropoulos, T. N. Davidson, and Z.-Q. Luo, "Transmit beamforming for physical-layer multicasting," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 2239–2251, Jun. 2006.

[36] E. Karipidis, N. Sidiropoulos, and Z.-Q. Luo, "Quality of service and max-min fair transmit beamforming to multiple cochannel multicast groups," *IEEE Trans. Signal Process.*, vol. 56, no. 3, pp. 1268–1279, Mar. 2008.

[37] M. Chiang, S. Low, A. Calderbank, and J. Doyle, "Layering as optimization decomposition: A mathematical theory of network architectures," *Proc. IEEE*, vol. 95, no. 1, pp. 255–312, Jan. 2007.
[38] M. Chiang, P. Hande, T. Lan, and C. W. Tan, *Power Control in Wireless Cellular Networks*. Delft, The Netherlands: Now Publishers, Inc., 2008.
[39] R. Rockafellar and J. Wets, *Variational Analysis*. Berlin, Germany: Springer-Verlag, 1998.
[40] F. Facchinei and J.-S. Pang, "Exact penalty functions for generalized Nash problems," in *Large Scale Nonlinear Optimization*, G. D. Pillo and M. Roma, Eds. Berlin, Germany: Springer-Verlag, 2006, pp. 115–126.
[41] D. Bertsekas, A. Nedic, and A. Ozdaglar, *Convex Analysis and Optimization*. Belmont, MA, USA: Athena Scientific, 2003.
[42] A. Daneshmand, F. Facchinei, V. Kungurtsev, and G. Scutari, "Hybrid random/deterministic parallel algorithms for convex and nonconvex big data optimization," *IEEE Trans. Signal Process.*, vol. 63, no. 15, pp. 3914–3929, Aug. 2015.
[43] M. Su and H.-K. Xu, "Remarks on the gradient-projection algorithm," *J. Nonlinear Anal. Optim.*, vol. 1, pp. 35–43, Jul. 2010.
[44] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course (Applied Optimization)*. Berlin, Germany: Springer-Verlag, 2004.
[45] Y. Nesterov, "Smooth minimization of non-smooth functions," *Math. Program.*, vol. 103, no. 1, pp. 127–152, 2005. doi: 10.1007/s10107-004-0552-5.
[46] D. Bertsekas, A. Nedic, and A. Ozdaglar, *Convex Analysis and Optimization*. Belmont, MA, USA: Athena Scientific, 2003.
[47] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, *Optimization with Sparsity-inducing Penalties. Foundations and Trends in Machine Learning*. Delft, The Netherlands: Now Publishers, Inc., Dec. 2011.
[48] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, Jan. 2009.
[49] F. Facchinei, S. Sagratella, and G. Scutari, "Flexible parallel algorithms for big data optimization," *IEEE Trans. Signal Process.*, vol. 63, no. 7, pp. 1874–1889, Apr. 2015.
[50] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc. Series B (Methodological)*, vol. 58, pp. 267–288, 1996.
[51] H. A. Le Thi and T. Pham Dinh, "Recent advances in dc programming and DCA," in *Trans. Comput. Intell. XIII*, vol. 8342, 2014, pp. 1–37.
[52] H. A. Le Thi, V. N. Huynh, and T. Pham Dinh, "DC programming and DCA for general dc programs," in *Proc. Adv. Comput. Methods Knowl. Eng.*, 2014, pp. 15–35.
[53] F. Clarke, *Optimization and Nonsmooth Analysis*. Hoboken, NJ, USA: Wiley, 1983.
[54] G. Scutari, F. Facchinei, and L. Lampariello, "Parallel and distributed methods for constrained nonconvex optimization—Part I: Theory," School Ind. Eng., Purdue Univ., West Lafayette, IN, USA, *Tech. Rep.*, Nov. 2016. Available at arXiv:1410.4754.
[55] F. Facchinei and J.-S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Berlin, Germany: Springer-Verlag, 2003.
[56] S. Dafermos, "Sensitivity analysis in variational inequalities," *Math. Oper. Res.*, vol. 13, no. 3, pp. 421–434, 1988.
[57] N. Yen, "Hölder continuity of solutions to a parametric variational inequality," *Appl. Math. Optim.*, vol. 31, no. 3, pp. 245–255, 1995.
[58] R. Rockafellar, "Lipschitzian properties of multifunctions," *Nonlinear Anal.: Theory, Methods Appl.*, vol. 9, no. 8, pp. 867–885, 1985.

Gesualdo Scutari (S'05–M'06–SM'11) received the degree in electrical engineering and the Ph.D. degree (both with honors) from the University of Rome "La Sapienza," Rome, Italy, in 2001 and 2005, respectively. He is currently an Associate Professor in the School of Industrial Engineering, Purdue University, West Lafayette, IN, USA, and is the Scientific Director for the area of big-data analytics at the Cyber Center (Discovery Park), Purdue University. He had previously held several research appointments, namely, at the University of California at Berkeley, Berkeley, CA, USA; Hong Kong University of Science and Technology, Hong Kong; University of Rome, "La Sapienza," Rome, Italy; and University of Illinois at Urbana-Champaign, Urbana, IL, USA. His research interests include theoretical and algorithmic issues related to big-data optimization, equilibrium programming, and their applications to signal processing, medical imaging, machine learning, and networking. He is currently an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING and was an Associate Editor of the IEEE SIGNAL PROCESSING LETTERS. He also served with the IEEE Signal Processing Society Technical Committee on Signal Processing for Communications. He received the 2006 Best Student Paper Award at the International Conference on Acoustics, Speech, and Signal Processing 2006, the 2013 NSF Faculty Early Career Development (CAREER) Award, the 2013 UB Young Investigator Award, the 2015 AnnaMaria Molteni Award for Mathematics and Physics (from ISSNAF), and the 2015 IEEE Signal Processing Society Young Author Best Paper Award.

Francisco Facchinei received the Ph.D. degree in system engineering from the University of Rome, "La Sapienza," Rome, Italy. He is currently a Full Professor of operations research, Engineering Faculty, University of Rome, "La Sapienza." His research interests include theoretical and algorithmic issues related to nonlinear optimization, variational inequalities, complementarity problems, equilibrium programming, and computational game theory.

Lorenzo Lampariello received the M.Sc. degree in electrical engineering and the Ph.D. degree in system engineering from Sapienza University of Rome, Rome, Italy. He is currently a Research Fellow in the Department of Business Studies, Roma Tre University, Rome, Italy. His research interests include the areas of multiagent and multiobjective optimization, nonlinear programming, and equilibrium problems. In particular, his primary interests include computational game theory, variational inequalities, nonsmooth nonconvex optimization, and bilevel programming, with an emphasis on distributed algorithms and on applications in networking and finance.