Decomposition by Partial Linearization: Parallel Optimization of Multi-Agent Systems

Gesualdo Scutari, *Senior Member, IEEE*, Francisco Facchinei, Peiran Song, Daniel P. Palomar, *Fellow, IEEE*, and Jong-Shi Pang

***Abstract*—We propose a novel decomposition framework for the distributed optimization of *general* nonconvex sum-utility functions arising naturally in the system design of wireless multi-user interfering systems. Our main contributions are i) the development of the first class of (inexact) *Jacobi best-response* algorithms with provable convergence, where all the users simultaneously and iteratively solve a suitably convexified version of the original sum-utility optimization problem; ii) the derivation of a general dynamic pricing mechanism that provides a unified view of existing pricing schemes that are based, instead, on heuristics; and iii) a framework that can be easily particularized to well-known applications, giving rise to very efficient practical (Jacobi or Gauss–Seidel) algorithms that outperform existing *ad hoc* methods proposed for very specific problems. Interestingly, our framework contains as special cases well-known gradient algorithms for nonconvex sum-utility problems, and many block-coordinate descent schemes for convex functions.**

***Index Terms*—Nonconvex multi-agent problems, parallel and distributed optimization, successive convex approximation.**

I. Introduction

WIRELESS networks are composed of users that may have different objectives and generate interference, when no multiplexing scheme is imposed a priori to regulate the transmissions; examples are peer-to-peer, ad-hoc, and cognitive radio systems. A usual and convenient way of designing such multiuser systems is by optimizing the "social function", i.e., the (weighted) sum of the users' objective functions. Since centralized solution methods are too demanding in most applications, the main difficulty of this formulation lies in performing the optimization in a distributed manner with limited signaling among the users. When the social problem is a sum-separable *convex* programming, many distributed methods have been proposed, based on primal and dual decomposition techniques; see, e.g., [2]–[4] and references therein. In this paper we address the more frequent and difficult case in which the social function is nonconvex. It is well known that the problem of finding a global minimum of the social function is, in general, NP hard (see e.g., [5]), and centralized solution methods (e.g., based on combinatorial approaches) are too demanding in most applications. As a consequence, recent research efforts have been focused on finding efficiently high quality suboptimal solutions via easy-to-implement (possibly) distributed algorithms. A recent survey on nonconvex resource allocation problems in interfering networks modeled as Gaussian Interference Channels (ICs) is [6].

In an effort to obtain distributed albeit suboptimal algorithms a whole spectrum of approaches has been explored, trying to balance practical effectiveness and coordination requirements. At one end of the spectrum we find game-theoretical approaches, where users in the network are modeled as players that greedily optimize their own objective function. Game-theoretical models for power control problems over ICs have been proposed in [7]–[11] and [12]–[14] for SISO and MISO/MIMO systems, respectively. Two recent tutorials on the subject are [15], [16], while recent contributions using the more general mathematical theory of Variational Inequalities [17] are [18]–[20]. The advantage of game-theoretic methods is that they lead to distributed implementations (only local channel information is required at each user); however they converge to Nash equilibria that in general are not even stationary solutions of the nonconvex social problem. In contrast, other methods aim at reaching stationary solutions of the nonconvex social problem, at the cost of more signaling and coordination. *Sequential* decomposition algorithms were proposed in [21]–[24] for the sum-rate maximization problem over SISO/MIMO ICs, and in [25] for more general (nonconvex) functions. In these algorithms, only one agent at a time is allowed to update his optimization variables; a fact that in large scale networks may lead to excessive communication overhead and slow convergence.

The aim of this paper is instead the study of more appealing *simultaneous distributed* methods for *general* nonconvex sum-utility problems, where *all* users can update their variables at the same time. The design of such algorithms with provable

Manuscript received February 04, 2013; revised July 01, 2013 and September 18, 2013; accepted October 31, 2013. Date of publication November 26, 2013; date of current version January 14, 2014. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Anthony Man-Cho So. The research of Scutari and Song is supported by the grants NSF No. CNS-1218717 and NSF CAREER No. ECCS-1254739. The research of Palomar is supported by the Hong Kong RGC 617810 research grant. The research of Pang is supported by NSF grant No. CMMI 0969600 (awarded to the University of Illinois at Urbana-Champaign). This work has been presented in part at the Fifth International Conference on Network Games, Control and Optimization (NetGCooP), Paris, France, October 12–14, 2011 [1].

G. Scutari and P. Song are with the Department of Electrical Engineering, State University of New York at Buffalo, Buffalo 14260 USA (e-mail: gesualdo@buffalo.edu; peiranso@buffalo.edu).

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

D. P. Palomar is with the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong (e-mail: palomar@ust.hk).

J.-S. Pang is with the Department of Industrial and Systems Engineering, University of Southern California Viterbi School of Engineering, Los Angeles 90089, USA (e-mail: jongship@usc.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2013.2293126

convergence is much more difficult, as also witnessed by the scarcity of results available in the literature. Besides the application of the classical gradient projection algorithm to the sum-rate maximization problem over MIMO ICs [26], parallel iterative algorithms (with message passing) for DSL/ad-hoc SISO networks and MIMO broadcast interfering channels were proposed in [27]–[29] and [30], respectively. Unfortunately, the gradient schemes [26] suffer from slow convergence and do not exploit any degree of convexity that might be present in the sum-utility function; [27]–[29] hinge crucially on the special log-structure of the users' rate functions; and [30] is based on the connection with a weighted MMSE problem. This makes [27]–[30] not applicable to different classes of sum-utility problems.

Building on the idea first introduced in [1], the main contribution of this paper is to propose a new decomposition method that: *i)* converges to stationary points of a large class of (nonconvex) social problems, encompassing most sum-utility functions of practical interest (including functions of complex variables); *ii)* decomposes well across the users, resulting in the *parallel* solution of *convex* subproblems, one for each user; *iii)* converges also if the users' subproblems are solved in an inexact way; and *iv)* contains as special case the gradient algorithms for nonconvex sum-utility problems, and many block-coordinate descent schemes for convex functions. Moreover, the proposed framework can be easily particularized to well-known applications, such as [21]–[24], [29], [31], giving rise in a unified fashion to distributed simultaneous algorithms that outperform existing *ad-hoc* methods both theoretically and numerically. We remark that while we follow the seminal ideas put forward in [1], in this paper, besides providing full proofs of the results in [1], we i) consider a much wider class of social-problems and (possibly inexact) algorithms, including [1] as special cases, ii) discuss in detail the case of functions of complex variables, and iii) compare numerically to state-of-the-art alternative methods. To the best of our knowledge, this paper is the first attempt toward the development of decomposition techniques for *general* nonconvex sum-utility problems that allow *distributed simultaneous* (possibly *inexact*) *best-response*-based updates among the users.

On one hand, our approach draws on the Successive Convex Approximation (SCA) paradigm, but relaxes the key requirement that the convex approximation must be a tight global upper bound of the social function, as required instead in [27], [32], [33] (see Section VI for a detailed comparison with [27], [32], [33]). This represents a turning point in the design of distributed SCA-based methods, since up to date, finding such an upper bound convex approximation for sum-utility functions having no specific structure (as, e.g., [24], [26]–[30]) has been an elusive task.

On the other hand, our method also sheds new light on widely used pricing mechanisms: indeed, our scheme can be viewed as a dynamic pricing algorithm where the pricing rule derives from a deep understanding of the problem characteristics and is not obtained on an ad-hoc basis, as instead in [21]–[24], [31]. We conclude this review by mentioning the recent work [34], where the authors, developing ideas contained in [30], [33], proposed parallel schemes based of the SCA idea that are applicable (only) to the class of sum-utility problems for which a connection with a MMSE formulation can be established. Note that [33], [34], which share some ideas with our approach, appeared after [1].

The rest of the paper is organized as follows. Section II introduces the sum-utility optimization problem along with some motivating examples. Section III presents our novel decomposition mechanism based on partial linearizations; the algorithmic framework is described in Section IV. Section V extends our results to sum-utility problems in the complex domain; further generalizations are discussed in Section VI. In Section VII we apply our new algorithms to some resource allocation problems over SISO and MIMO ICs, and compare their performance with the state-of-the-art decomposition schemes. Finally, Section VIII draws some conclusions.

II. Problem Formulation

We consider the design of a multiuser system composed of I coupled users $\mathcal{I} \triangleq \{1,\ldots,I\}$. Each user i makes decisions on his own n_i-dimensional real strategy vector $\mathbf{x}_i$, which belongs to the feasible set $\mathcal{K}_i$; the vector variables of the other users is denoted by $\mathbf{x}_{-i} \triangleq (\mathbf{x}_j)_{j\neq i} \in \mathcal{K}_{-i} \triangleq \prod_{j\neq i}\mathcal{K}_j$; the users' strategy profile is $\mathbf{x} = (\mathbf{x}_i)_{i=1}^I$, and the joint strategy set of the users is $\mathcal{K} \triangleq \prod_{j\in\mathcal{I}}\mathcal{K}_j$. The system design is formulated as

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & U(\mathbf{x}) \triangleq \sum_{\ell\in\mathcal{I}_f} f_\ell(\mathbf{x}) \\ \text{subject to} \quad & \mathbf{x}_i \in \mathcal{K}_i, \quad \forall i \in \mathcal{I}, \end{aligned} \tag{1}$$

with $\mathcal{I}_f \triangleq \{1,\ldots,I_f\}$. Observe that, in principle, the set $\mathcal{I}_f$ of objective functions is different from the set $\mathcal{I}$ of users; we show shortly how to explore this extra degree of freedom to good effect. Of course, (1) contains the most common case where there is exactly one function for each user, i.e., $I = I_f$.

Assumptions: We make the following blanket assumptions:

A1) Each $\mathcal{K}_i$ is closed and convex;

A2) Each f_i is continuously differentiable on $\mathcal{K}$;

A3) Each $\nabla_{\mathbf{x}} f_i$ is Lipschitz continuous on $\mathcal{K}$, with constant $L_{\nabla f_i}$; let $L_{\nabla U} \triangleq \sum_i L_{\nabla f_i}$;

A4) U is coercive with respect to $\mathcal{K}$.

The assumptions above are quite standard and are satisfied by a large class of problems of practical interest. In particular, condition A4 guarantees that the social problem has a solution, even when the feasible $\mathcal{K}$ is not bounded; if $\mathcal{K}$ is bounded A4 is trivially satisfied. Note that, differently from classical Network Utility Maximization (NUM) problems, here we do not assume any convexity of the functions f_ℓ, thus, (1) is a nonconvex minimization problem. For the sake of simplicity, in (1) we assume that the users' strategies are real vectors; in Section V, we extend our framework to complex matrix strategies, to cover also the design of MIMO systems.

1) A Motivating Example: The social problem (1) is general enough to encompass many sum-utility problems of practical interest. It also includes well-known utility functions studied in the literature; an example is given next. Consider an N-parallel Gaussian IC composed of I active users, and let

$$r_i(\mathbf{p}_i,\mathbf{p}_{-i}) \triangleq \sum_{k=1}^{N} \log\left(1+\frac{|H_{ii}(k)|^2 p_{ik}}{\sigma_{ik}^2+\sum\limits_{j\neq i}|H_{ij}(k)|^2 p_{jk}}\right)$$

be the maximum achievable rate on link i, where $\mathbf{p}_i \triangleq (p_{ik})_{k=1}^N$ denotes the power allocation of user i over the N parallel channels, $\mathbf{p}_{-i} \triangleq (\mathbf{p}_j)_{j\neq i}$ is the power profile of all the other users $j \neq i$, $|H_{ij}(k)|^2$ is the gain of the channel between the j-th transmitter and the i-th receiver, σ_{ik}^2 is the variance of the thermal noise over carrier k at the receiver i, and $\sum_{j\neq i}|H_{ij}(k)|^2 p_{jk}$ represents the multiuser interference generated by the users $j \neq i$ at the receiver i. Each transmitter i is subject to the power constraints $\mathbf{p}_i \in \mathcal{P}_i$, with

$$\mathcal{P}_i \triangleq \left\{\mathbf{p}_i \in \mathbb{R}_+^N : \mathbf{W}_i\mathbf{p}_i \leq \mathbf{I}_i^{\max}\right\}, \tag{2}$$

where the inequality, with given $\mathbf{I}_i^{\max} \in \mathbb{R}_+^{m_i}$ and $\mathbf{W}_i \in \mathbb{R}_+^{m_i\times N}$ is intended to be component-wise. Note that the linear (vector) constraints in (2) are general enough to model classical power budget constraints and different interference constraints, such as spectral mask or interference-temperature limits. Finally, let $\theta_i : \mathbb{R}_+ \to \mathbb{R}$ be the utility functions of the users' rates. The system design can then be formulated as

$$\begin{aligned} \underset{\mathbf{p}_1,\ldots,\mathbf{p}_I}{\text{maximize}} \quad & \sum_{i\in\mathcal{I}} \theta_i\left(r_i(\mathbf{p}_i,\mathbf{p}_{-i})\right) \\ \text{subject to} \quad & \mathbf{p}_i \in \mathcal{P}_i, \quad \forall i \in \mathcal{I}. \end{aligned} \tag{3}$$

Note that (3) is an instance of (1), with $I_f = I$; moreover assumptions A1–A4 are satisfied if the utility functions $\theta_i(x)$ are i) concave and nondecreasing on $\mathbb{R}_+$, and ii) continuously differentiable with Lipschitz gradients. Interestingly, this class of functions $\theta_i(x)$ includes many well-known special cases studied in the literature, such as the weighted sum-rate function, the harmonic mean of the rates, the geometric mean of (one plus) the rates, etc.; see, e.g., [6], [21], [22], [35].

Since the class of problems (1) is in general nonconvex (generally NP hard [5]), the focus of this paper is on the design of *distributed* solution methods for computing stationary solutions (possibly local minima) of (1). Our major goal is to devise *simultaneous best-response* schemes fully decomposed across the users, meaning that all the users can solve *in parallel* a sequence of convex problems while converging to a stationary solution of the original nonconvex problem.

III. A New Decomposition Technique

We begin by introducing an informal description of our new algorithms that sheds light on the core idea of the novel decomposition technique and establishes the connection with classical descent gradient-based schemes. This will also explain why our scheme is expected to outperform current gradient methods. A formal description of the proposed algorithms along with their main properties is given in Section IV for the real case, and in Section V for the complex case.

A. What Do Conditional Gradient Methods Miss?

A classical approach to solve a nonconvex problem like (1) would be using some well-known gradient-based descent scheme. A simple way to generate a (feasible) descent direction is for example using the conditional gradient method (also called Frank-Wolfe method) [4]: given the current iterate $\mathbf{x}^n = (\mathbf{x}_i^n)_{i=1}^I$, the next feasible vector $\mathbf{x}^{n+1}$ is given by

$$\mathbf{x}^{n+1} = \mathbf{x}^n + \gamma^n\,\mathbf{d}^n \tag{4}$$

where $\mathbf{d}^n \triangleq \overline{\mathbf{x}}^n - \mathbf{x}^n$, $\overline{\mathbf{x}}^n = (\overline{\mathbf{x}}_i^n)_{i=1}^I$ is the solution of the following set of convex problems (one for each user):

$$\overline{\mathbf{x}}_i^n = \underset{\mathbf{x}_i\in\mathcal{K}_i}{\operatorname{argmin}} \left\{\nabla_{\mathbf{x}_i} U\left(\mathbf{x}^n\right)^T\left(\mathbf{x}_i - \mathbf{x}_i^n\right)\right\}, \tag{5}$$

for all $i \in \mathcal{I}$, and $\gamma^n \in (0,1]$ is the step-size of the algorithm that needs to be properly chosen to guarantee convergence.

Looking at (5) one infers that gradient methods are based on solving a sequence of parallel convex problems, one for each user, obtained by linearizing the *whole* utility function $U(\mathbf{x})$ around $\mathbf{x}^n$, a fact that does not exploit any "nice" structure that the original problem may potentially have.

At the basis of the proposed decomposition techniques, there is instead the attempt to properly exploit any degree of convexity that might be present in the social function. To capture this idea, for each user $i \in \mathcal{I}$, let $\mathcal{S}_i \subseteq \mathcal{I}_f$ be the set of indices of all the functions $f_j(\mathbf{x}_i, \mathbf{x}_{-i})$ that are convex in $\mathbf{x}_i$ on $\mathcal{K}_i$, for any given $\mathbf{x}_{-i} \in \mathcal{K}_{-i}$:

$$\mathcal{S}_i \triangleq \{j \in \mathcal{I}_f : f_j(\bullet, \mathbf{x}_{-i}) \text{ is convex on } \mathcal{K}_i, \ \ \forall \mathbf{x}_{-i} \in \mathcal{K}_{-i}\} \tag{6}$$

and let $\mathcal{C}_i \subseteq \mathcal{S}_i$ be a given subset of $\mathcal{S}_i$. The idea is to preserve the convex structure of the functions in $\mathcal{C}_i$ while linearizing the rest. Note that we allow the possibility that $\mathcal{S}_i = \emptyset$, even if we "hope" that $\mathcal{S}_i \neq \emptyset$, and actually this latter case occurs in most of the applications of interest, see Section VII. For each user $i \in \mathcal{I}$, we can introduce the following convex approximation of $U(\mathbf{x})$ around $\mathbf{x}^n \in \mathcal{K}$:

$$\begin{aligned} \widetilde{f}_{\mathcal{C}_i}(\mathbf{x}_i;\mathbf{x}^n) \triangleq & \sum_{j\in\mathcal{C}_i} f_j(\mathbf{x}_i,\, \mathbf{x}_{-i}^n) + \boldsymbol{\pi}_{\mathcal{C}_i}(\mathbf{x}^n)^T(\mathbf{x}_i - \mathbf{x}_i^n) \\ & + \frac{\tau_i}{2}\left(\mathbf{x}_i - \mathbf{x}_i^n\right)^T \mathbf{H}_i(\mathbf{x}^n)\left(\mathbf{x}_i - \mathbf{x}_i^n\right) \end{aligned} \tag{7}$$

with

$$\boldsymbol{\pi}_{\mathcal{C}_i}(\mathbf{x}^n) \triangleq \sum_{j\in\mathcal{C}_{-i}} \nabla_{\mathbf{x}_i} f_j(\mathbf{x})|_{\mathbf{x}=\mathbf{x}^n}, \tag{8}$$

where $\mathcal{C}_{-i} \triangleq \mathcal{I}_f\backslash\mathcal{C}_i$ is the complement of $\mathcal{C}_i$, τ_i is a given nonnegative constant, and $\mathbf{H}_i(\mathbf{x}^n)$ is an $n_i \times n_i$ uniformly positive definite matrix (possibly dependent on $\mathbf{x}^n$), i.e., $\mathbf{H}_i(\mathbf{x}^n) - c_{H_i}\mathbf{I} \succeq \mathbf{0}$, for some positive c_{H_i}. For notational simplicity, we omitted in $\widetilde{f}_{\mathcal{C}_i}(\mathbf{x}_i;\mathbf{x}^n)$ the dependence on τ_i and $\mathbf{H}_i(\mathbf{x}^n)$. Note that in (7), we added a proximal-like regularization term, in order to relax the convergence conditions of the resulting algorithm or enhance the convergence speed (cf. Section IV). A key feature of $\widetilde{f}_{\mathcal{C}_i}$ we will always require is that $\widetilde{f}_{\mathcal{C}_i}(\bullet;\mathbf{x})$ be uniformly strongly convex. By this we mean the following. Let $c_{\tau_i}(\mathbf{x})$ be the constant of strong convexity of $\widetilde{f}_{\mathcal{C}_i}(\bullet;\mathbf{x})$. We require that

$$c_{\tau_i} \triangleq \inf_{\mathbf{x}\in\mathcal{K}} c_{\tau_i}(\mathbf{x}) > 0. \tag{9}$$

Note that this *is not an additional assumption*, but just a requirement on the way τ_i is chosen. Under the uniformly positive definiteness of $\mathbf{H}_i(\mathbf{x}^n)$, condition (9) is always satisfied if $\tau_i > 0$; however it is also satisfied with $\tau_i = 0$ if $\sum_{j\in\mathcal{C}_i} f_j(\bullet, \mathbf{x}_{-i})$ is uniformly strongly convex on $\mathcal{K}_{-i}$; a fact that occurs in many applications, see, e.g., Section VII.

Associated with each $\widetilde{f}_{\mathcal{C}_i}(\mathbf{x}_i;\mathbf{x}^n)$ we can define the following "best response" map that resembles (5):

$$\widehat{\mathbf{x}}_{\mathcal{C}_i}(\mathbf{x}^n,\tau_i) \triangleq \underset{\mathbf{x}_i\in\mathcal{K}_i}{\operatorname{argmin}}\ \widetilde{f}_{\mathcal{C}_i}(\mathbf{x}_i;\mathbf{x}^n). \tag{10}$$

Note that, in the setting above, $\widehat{\mathbf{x}}_{\mathcal{C}_i}(\mathbf{x}^n,\tau_i)$ is always well-defined, since the optimization problem in (10) is strongly convex and thus has a unique solution. Given (10), we can introduce the best-response mapping of the users, defined as

$$\mathcal{K}\ni\mathbf{y}\mapsto\widehat{\mathbf{x}}_{\mathcal{C}}(\mathbf{y},\boldsymbol{\tau}) \triangleq (\widehat{\mathbf{x}}_{\mathcal{C}_i}(\mathbf{y},\tau_i))_{i=1}^{I}; \tag{11}$$

and also set $\boldsymbol{\tau}\triangleq(\tau_i)_{i=1}^I$. The proposed search direction $\mathbf{d}^n$ at point $\mathbf{x}^n$ in (4) becomes then $\widehat{\mathbf{x}}_{\mathcal{C}}(\mathbf{x}^n,\boldsymbol{\tau})-\mathbf{x}^n$. The challenging question now is whether such direction is still a descent direction for the function U at $\mathbf{x}^n$ and how to choose the free parameters (such as τ_i's, γ^n's, and $\mathbf{H}_i(\mathbf{x}^n)$'s) in order to guarantee convergence to a stationary solution of the original nonconvex sum-utility problem. These issues are addressed in the next sections.

B. Properties of the Best-Response Mapping $\widehat{\mathbf{x}}_{\mathcal{C}}(\mathbf{y},\boldsymbol{\tau})$

Before introducing a formal description of the proposed algorithms, we derive next some key properties of the best-response map $\widehat{\mathbf{x}}_{\mathcal{C}}(\mathbf{y},\boldsymbol{\tau})$, which shed light on how to choose the free parameters in (10) and prove convergence.

Proposition 1: Given the social problem (1) under A1)–A4), suppose that each $\mathbf{H}_i(\mathbf{x})-c_{H_i}\mathbf{I}\succeq\mathbf{0}$ for all $\mathbf{x}\in\mathcal{K}$ and some $c_{H_i}>0$, and $(c_{\tau_i})_{i=1}^I>\mathbf{0}$. Then the mapping $\mathcal{K}\ni\mathbf{y}\mapsto\widehat{\mathbf{x}}(\mathbf{y},\boldsymbol{\tau})$ has the following properties:

(a) $\widehat{\mathbf{x}}_{\mathcal{C}}(\bullet,\boldsymbol{\tau})$ is Lipschitz continuous on $\mathcal{K}$, i.e., there exists a positive constant $\hat{L}$ such that

$$\|\widehat{\mathbf{x}}_{\mathcal{C}}(\mathbf{y},\boldsymbol{\tau})-\widehat{\mathbf{x}}_{\mathcal{C}}(\mathbf{z},\boldsymbol{\tau})\|\le\hat{L}\,\|\mathbf{y}-\mathbf{z}\|,\quad\forall\mathbf{y},\mathbf{z}\in\mathcal{K}; \tag{12}$$

(b) The set of the fixed-points of $\widehat{\mathbf{x}}_{\mathcal{C}}(\bullet,\boldsymbol{\tau})$ coincides with the set of stationary solutions of the social problem (1); therefore $\widehat{\mathbf{x}}_{\mathcal{C}}(\mathbf{y},\boldsymbol{\tau})$ has a fixed-point;

(c) For every given $\mathbf{y}\in\mathcal{K}$, the vector $\widehat{\mathbf{x}}_{\mathcal{C}}(\mathbf{y},\boldsymbol{\tau})-\mathbf{y}$ is a descent direction of the social function $U(\mathbf{x})$ at $\mathbf{y}$ such that

$$(\widehat{\mathbf{x}}_{\mathcal{C}}(\mathbf{y},\boldsymbol{\tau})-\mathbf{y})^T\,\nabla_{\mathbf{x}}U(\mathbf{y})\le-c\,\|\widehat{\mathbf{x}}_{\mathcal{C}}(\mathbf{y},\boldsymbol{\tau})-\mathbf{y}\|^2, \tag{13}$$

for some positive constant $c\ge c_{\boldsymbol{\tau}}$, with

$$c_{\boldsymbol{\tau}}\triangleq\min_{i\in\mathcal{I}}\{c_{\tau_i}\}. \tag{14}$$

(d) If $\nabla_{\mathbf{x}}U(\mathbf{x})$ is bounded on $\mathcal{K}$, then there exists a finite constant $\alpha>0$ such that

$$\|\widehat{\mathbf{x}}_{\mathcal{C}}(\mathbf{y},\boldsymbol{\tau})-\mathbf{y}\|\le\alpha,\quad\forall\mathbf{y}\in\mathcal{K}. \tag{15}$$

Proof: See Appendix A. ■

Proposition 1 makes formal the idea introduced in Section III-A and thus paves the way to the design of distributed best-response-like algorithms for (1) based on $\widehat{\mathbf{x}}_{\mathcal{C}}(\bullet,\boldsymbol{\tau})$. Indeed, the inequality (13) states that either $(\widehat{\mathbf{x}}_{\mathcal{C}}(\mathbf{x}^n)-\mathbf{x}^n)^T\,\nabla_{\mathbf{x}}U(\mathbf{x}^n)<0$ or $\widehat{\mathbf{x}}_{\mathcal{C}}(\mathbf{x}^n)=\mathbf{x}^n$. In the former case, $\widehat{\mathbf{x}}_{\mathcal{C}}(\mathbf{x}^n)-\mathbf{x}^n$ [the candidate $\mathbf{d}^n$ in (4)] is a descent direction of $U(\mathbf{x})$ at $\mathbf{x}^n$; in the latter case, $\mathbf{x}^n$ is a fixed-point of the mapping $\widehat{\mathbf{x}}_{\mathcal{C}}(\bullet,\boldsymbol{\tau})$ and thus a stationary solution of the original nonconvex problem (1) [Prop. 1 (b)].

Quite interestingly, we can also provide a characterization of the fixed-points of $\widehat{\mathbf{x}}_{\mathcal{C}}(\mathbf{y},\boldsymbol{\tau})$ [and thus the stationary solutions of (1)] in terms of Nash equilibria of a game with a proper pricing mechanism. Formally, we have the following.

Proposition 2: Any fixed-point $\mathbf{x}^\star$ of $\widehat{\mathbf{x}}_{\mathcal{C}}(\bullet,\boldsymbol{\tau})$ is a Nash equilibrium of the game where each user $i\in\mathcal{I}$ solves the following priced convex optimization problem: given $\mathbf{x}_{-i}$,

$$\min_{\mathbf{x}_i\in\mathcal{K}_i}\sum_{j\in\mathcal{C}_i}f_j(\mathbf{x}_i,\mathbf{x}_{-i})+\boldsymbol{\pi}_{\mathcal{C}_i}(\mathbf{x}^\star)^T\mathbf{x}_i. \tag{16}$$

According to the above proposition, the stationary solutions of (1) achievable as fixed-points of $\widehat{\mathbf{x}}_{\mathcal{C}_i}(\bullet,\boldsymbol{\tau})$ are *unilaterally* optimal for the objective functions in (16). This result is in agreement with those obtained in [22], [23] for the sum-rate maximization problem over SISO frequency selective-channels. Despite its theoretical interest, however, Prop. 2 does not help in practice to solve (1). Indeed, the computation of a Nash equilibrium of the game in (16) would require the a-priori knowledge of the prices $\boldsymbol{\pi}_{\mathcal{C}_i}(\mathbf{x}^\star)$ and thus the equilibrium itself, which of course is not available.

IV. Distributed Decomposition Algorithms

We are now ready to introduce our new algorithms, as a direct product of Prop. 1. We first focus on (inexact) Jacobi schemes (cf. Section IV-A); then we show that the same results hold also for (inexact) Gauss-Seidel updates (cf. Section IV-C).

A. Exact Jacobi Best-Response Schemes

The first algorithm we propose is a Jacobi scheme where all users update simultaneously their strategies based on the best-response $\widehat{\mathbf{x}}_{\mathcal{C}_i}(\bullet,\boldsymbol{\tau})$ (possibly with a memory); the formal description is given in Algorithm 1 below, and its convergence properties are given in Theorem 3.

Algorithm 1 : Exact Jacobi SCA Algorithm

Data : $\boldsymbol{\tau}\ge\mathbf{0}$, $\{\gamma^n\}>0$, $\mathbf{x}^0\in\mathcal{K}$. Set $n=0$.

(S.1): If $\mathbf{x}^n$ satisfies a termination criterion: STOP;
(S.2): For all $i\in\mathcal{I}$, compute $\widehat{\mathbf{x}}_{\mathcal{C}_i}(\mathbf{x}^n,\boldsymbol{\tau})$ [cf. (10)];
(S.3): Set $\mathbf{x}^{n+1}\triangleq\mathbf{x}^n+\gamma^n(\widehat{\mathbf{x}}_{\mathcal{C}}(\mathbf{x}^n,\boldsymbol{\tau})-\mathbf{x}^n)$;
(S.4): $n\leftarrow n+1$, and go to (S.1).

Theorem 3: Given the social problem (1) under A1–A4, suppose that one of the two following conditions is satisfied:

(a) For each i, $\mathbf{H}_i(\mathbf{x})$ is such that $\mathbf{H}_i(\mathbf{x})-c_{H_i}\mathbf{I}\succeq\mathbf{0}$ for all $\mathbf{x}\in\mathcal{K}$ and some $c_{H_i}>0$; furthermore $\{\gamma^n\}$ and $\boldsymbol{\tau}\ge\mathbf{0}$ are chosen so that

$$0<\inf_n\gamma^n\le\sup_n\gamma^n\le\gamma^{\max}\le1\ \text{ and }\ 2\,c_{\boldsymbol{\tau}}\ge\gamma^{\max}L_{\nabla U}, \tag{17}$$

with $c_{\boldsymbol{\tau}}$ defined in (14).

(b) For each i, $\mathbf{H}_i(\mathbf{x})$ is such that $\mathbf{H}_i(\mathbf{x})-c_{H_i}\mathbf{I}\succeq\mathbf{0}$ for all $\mathbf{x}\in\mathcal{K}$ and some $c_{H_i}>0$, $\boldsymbol{\tau}\ge\mathbf{0}$ is such that $c_{\boldsymbol{\tau}}>\mathbf{0}$, and furthermore $\{\gamma^n\}$ is chosen so that

$$\gamma^n\in(0,1],\quad\gamma^n\to0,\quad\text{and}\quad\sum_n\gamma^n=+\infty. \tag{18}$$

Then, either Algorithm 1 converges in a finite number of iterations to a stationary solution of (1) or every limit point of the sequence $\{\mathbf{x}^n\}_{n=1}^\infty$ (at least one such point exists) is a stationary

solution of (1). Moreover, none of such points is a local maximum of U.

Proof: See Appendix B. ■

Main features of Algorithm 1. The algorithm implements a novel *distributed* SCA decomposition: all the users solve *in parallel* a sequence of *decoupled* strongly convex optimization problems as in (10). The algorithm is expected to perform better than classical gradient-based schemes (at least in terms of convergence speed) at the cost of no extra signaling, because the structure of the objective functions is better preserved. It is guaranteed to converge under very mild assumptions (the weakest available in the literature) while offering some flexibility in the choice of the free parameters [conditions (a) or (b) of Theorem 3]. This degree of freedom can be exploited, e.g., to achieve the desired tradeoff between signaling, convergence speed, and computational effort, as discussed next.

As far as the computation of the best-response $\widehat{\mathbf{x}}_{\mathcal{C}_i}(\mathbf{x}^n, \boldsymbol{\tau})$ is concerned, at each iteration, every user needs to known $\sum_{j\in\mathcal{C}_i} f_j(\bullet, \mathbf{x}^n_{-i})$ and $\boldsymbol{\pi}_{\mathcal{C}_i}(\mathbf{x}^n)$. The signaling required to acquire this information is of course problem-dependent. If the problem under consideration does not have any specific structure, the most natural message-passing strategy is to communicate directly $\mathbf{x}^n_{-i}$ and $(\nabla_{\mathbf{x}_i} f_j(\mathbf{x}^n))_{j\notin\mathcal{C}_i}$. However, in many specific applications much less signaling may be needed; see Section VII for some examples.

On the choice of the free parameters. Convergence of Algorithm 1 is guaranteed either using a constant step-size rule [cf. (17)] or a diminishing step-size rule [cf. (18)]. Moreover, different choices of $\{\mathcal{C}_i\}$ are in general feasible for a given social function, resulting in different best-response functions and signaling among the users.

1) Constant Step-Size: In this case, $\gamma^n = \gamma \le \gamma^{\max}$ for all n, where $\gamma^{\max} \in (0, 1]$ needs to be chosen together with $\boldsymbol{\tau} \ge \mathbf{0}$ and $(\mathbf{H}_i(\mathbf{y}))_{i=1}^I$ so that the condition $2\, c_{\boldsymbol{\tau}} \ge \gamma^{\max} L_{\nabla U}$ is satisfied, with $c_{\boldsymbol{\tau}}$ defined in (14). This can be done in several ways. A simple (but conservative) choice satisfying that condition is, e.g., $\tau_i = \tau > 0$ for all $i \in \mathcal{I}$, $\gamma^{\max} \in (0, 1]$, and $\gamma/\tau \le 2/L_{\nabla U}$. Note that this condition imposes a constraint only on the ratio γ/τ, leaving free the choice of one of the two parameters.

An interesting special case worth mentioning is: $\gamma = \gamma^{\max} = 1$ for all n, $\mathbf{H}_i(\mathbf{y}) = \mathbf{I}$ for all $i \in \mathcal{I}$, and $\boldsymbol{\tau} > \mathbf{0}$ large enough so that $2\, c_{\boldsymbol{\tau}} \ge L_{\nabla U}$. This choice leads to the classical Jacobi best-response scheme (but with a proximal regularization), namely: at each iteration n,

$$\mathbf{x}_i^{n+1} = \widehat{\mathbf{x}}_{\mathcal{C}_i}(\mathbf{x}^n, \boldsymbol{\tau}), \qquad \forall i \in \mathcal{I}.$$

To the best of our knowledge, this algorithm along with its convergence conditions [Theorem 3a)] represents a new result in the optimization literature; indeed classic best-response nonlinear Jacobi schemes require much stronger (sufficient) conditions to converge (implying contraction) [4, Ch. 3.3.5]. Note that the choice of τ_i's to guarantee convergence [i.e., $2\, c_{\boldsymbol{\tau}} \ge L_{\nabla U}$] can be done locally by each user with no signaling exchange, once the Lipschitz constant $L_{\nabla U}$ is known.

As a final remark, we point out that in the case of constant and "sufficiently" small step-size γ^n, one can relax the synchronization requirements among the users allowing (partially) asynchronous updates of users' best-responses (in the sense of [4]); we omit the details because of space limitation.

2) Variable Step-Size: In scenarios where the knowledge of the system parameters, e.g., $L_{\nabla U}$, is not available, one can use the diminishing step-size rule (18). Under such a rule, convergence is guaranteed for *any* choice of $\mathbf{H}_i(\mathbf{x}) - c_{H_i}\mathbf{I} \succeq \mathbf{0}$ and $\boldsymbol{\tau} \ge \mathbf{0}$ such that $c_{\boldsymbol{\tau}} > 0$. Note that if $\sum_{j\in\mathcal{C}_i} f_j(\bullet, \mathbf{x}_{-i})$ is strongly convex on $\mathcal{K}_i$ for any $\mathbf{x}_{-i} \in \mathcal{K}_{-i}$, one can also set $\tau_i = 0$, otherwise any arbitrary but positive τ_i is necessary. We will show in the next section that a diminishing step-size rule is also useful to allow an inexact computation of the best-response $\widehat{\mathbf{x}}_{\mathcal{C}_i}(\mathbf{x}^n, \boldsymbol{\tau})$ while preserving convergence of the algorithm. Two classes of step-size rules satisfying (18) are: given $\gamma^0 = 1$,

$$\text{Rule\#1}: \gamma^n = \gamma^{n-1}\left(1 - \epsilon\,\gamma^{n-1}\right), \quad n = 1, \ldots, \tag{19}$$

$$\text{Rule\#2}: \gamma^n = \frac{\gamma^{n-1} + \alpha(n)}{1 + \beta(n)}, \qquad n = 1, \ldots, \tag{20}$$

where in (19) $\epsilon \in (0, 1)$ is a given constant, whereas in (20) $\alpha(n)$ and $\beta(n)$ are two nonnegative real functions of $n \ge 1$ such that: i) $0 \le \alpha(n) \le \beta(n)$; and ii) $\alpha(n)/\beta(n) \to 0$ as $n \to \infty$ while $\sum_n (\alpha(n)/\beta(n)) = \infty$. Examples of such $\alpha(n)$ and $\beta(n)$ are: $\alpha(n) = \alpha$ or $\alpha(n) = \log(n)^\alpha$, and $\beta(n) = \beta\, n$ or $\beta(n) = \beta\sqrt{n}$, where α, β are given constants satisfying $\alpha \in (0, 1)$, $\beta \in (0, 1)$, and $\alpha \le \beta$.

Another issue to discuss is the choice of the free positive definite matrices $\mathbf{H}_i(\mathbf{y})$. Mimicking (quasi-)Newton-like schemes [36], a possible choice is to consider for $\mathbf{H}_i(\mathbf{x}^n)$ a proper (diagonal) uniformly positive definite "approximation" of the Hessian matrix $\nabla^2_{\mathbf{x}_i} U(\mathbf{x}^n)$. The exact expression to consider depends on the amount of signaling and computational complexity required to compute such a $\mathbf{H}_i(\mathbf{x}^n)$, and thus varies with the specific problem under consideration.

3) On the Choice of $\mathcal{C}_i$'s: In general, more than one (feasible) choice of $\{\mathcal{C}_i\}$ is possible for a given social function, resulting in different decomposition schemes. Some illustrative examples are discussed next.

Example #1–(Proximal) Gradient/Newton Algorithms: If each $\mathcal{C}_i = \emptyset$ and $I = I_f$, $\widehat{\mathbf{x}}_{\mathcal{C}_i}(\mathbf{x}^n, \tau_i)$ reduces to the gradient response (5) (possibly with a proximal regularization). It turns out that (exact and inexact) gradient algorithms along with their convergence conditions are special cases of our framework. Note that if $\mathcal{S}_i = \emptyset$ for every i (i.e., no convexity whatsoever is present in U), this is the only possible choice, and indeed our approach reduces to a gradient-like method. On the other hand, as soon as at least some $\mathcal{S}_i \ne \emptyset$, we may depart from the gradient method and exploit the available convexity.

Note that our framework contains also Newtown-like updates. For instance, if $U(\mathbf{x}_i, \mathbf{x}^n_{-i})$ is convex in $\mathbf{x}_i \in \mathcal{K}_i$ for any $\mathbf{x}^n_{-i} \in \mathcal{K}_{-i}$, a feasible choice is $\mathcal{C}_i = \emptyset$ and $\mathbf{H}_i(\mathbf{x}^n) = \nabla^2_{\mathbf{x}_i} U(\mathbf{x}^n)$, resulting in:

$$\begin{aligned}\widehat{\mathbf{x}}_i(\mathbf{x}^n, \tau_i) \triangleq \operatorname*{argmin}_{\mathbf{x}_i\in\mathcal{K}_i} \Big\{ & \nabla_{\mathbf{x}_i} U(\mathbf{x}^n)^T(\mathbf{x}_i - \mathbf{x}_i^n) \\ & + \frac{1}{2}(\mathbf{x}_i - \mathbf{x}_i^n)^T \nabla^2_{\mathbf{x}_i} U(\mathbf{x}^n)(\mathbf{x}_i - \mathbf{x}_i^n) \\ & + \frac{\tau_i}{2}\|\mathbf{x}_i - \mathbf{x}_i^n\|^2 \Big\}. \end{aligned} \tag{21}$$

Essentially (21) corresponds to a Newton-like step of user i in minimizing the "reduced" problem $\min_{\mathbf{x}_i\in\mathcal{K}_i} U(\mathbf{x}_i, \mathbf{x}^n_{-i})$.

Example #2–Pricing Algorithms in [1]: Suppose that $I = I_f$, and each $\mathcal{S}_i = \{i\}$ (implying that $f_i(\bullet, \mathbf{x}_{-i})$ is convex on $\mathcal{K}_i$

for any $\mathbf{x}_{-i} \in \mathcal{K}_{-i}$). By taking each $\mathcal{C}_i = \{i\}$ and $\mathbf{H}_i(\mathbf{x}^n) = \mathbf{I}$, we obtain the pricing-based algorithms in [1]:

$$\widehat{\mathbf{x}}_i(\mathbf{x}^n, \tau_i) \triangleq \underset{\mathbf{x}_i \in \mathcal{K}_i}{\operatorname{argmin}} f_i(\mathbf{x}_i, \mathbf{x}_{-i}^n) + \boldsymbol{\pi}_i(\mathbf{x}^n)^T \mathbf{x}_i + \frac{\tau_i}{2} \|\mathbf{x}_i - \mathbf{x}_i^n\|^2,$$

where $\boldsymbol{\pi}_i(\mathbf{x}^n) \triangleq \sum_{j \neq i} \nabla_{\mathbf{x}_i} f_j(\mathbf{x}^n)$. Algorithm 1 based on the above best-response implements naturally a pricing mechanism; indeed, each $\boldsymbol{\pi}_i(\mathbf{x}^n)$ represents a dynamic pricing that measures somehow the marginal increase of the sum-utility of the other users due to a variation of the strategy of user i; roughly speaking, it works like a punishment imposed to each user for being too aggressive in choosing his own strategy and thus "hurting" the other users. Pricing algorithms based on heuristics have been proposed in a number of papers for the sum-rate maximization problem over SISO/SIMO/MIMO ICs [21]–[23], [31], [37]. However, on top of being *sequential* schemes, convergence of algorithms in the aforementioned papers is established under relatively strong assumptions (e.g., limited number of users, special classes of functions, specific channel models and transmission schemes, etc.), see [23]. The pricing in our framework is instead the natural consequence of the proposed SCA decomposition technique and leads to *simultaneous* algorithms that can be applied (with convergence guaranteed) to a very large class of problems, even when [21]–[23], [31], [37] fail.

Example #3–(Proximal) Jacobi Algorithms for a Single (Convex) Function: Suppose that the social function is a single function $f(\mathbf{x}_1, \ldots, \mathbf{x}_I)$ on $\mathcal{K} = \prod_i \mathcal{K}_i$, which is convex in each $\mathbf{x}_i$ (but not necessarily jointly). Of course, this optimization problem can be interpreted as a special case of the framework (1), with $\mathcal{C}_i = \mathcal{S}_i = \{1\} = \mathcal{I}_f$, for all $i \in \mathcal{I}$ and $f_1(\mathbf{x}) = f(\mathbf{x})$. Then, setting $\mathbf{H}_i(\mathbf{x}^n) = \mathbf{I}$, the best-response (10) of each user i reduces to

$$\widehat{\mathbf{x}}_{\mathcal{C}_i}(\mathbf{x}^n, \tau_i) \triangleq \underset{\mathbf{x}_i \in \mathcal{K}_i}{\operatorname{argmin}} f(\mathbf{x}_i, \mathbf{x}_{-i}^n) + \frac{\tau_i}{2} \|\mathbf{x}_i - \mathbf{x}_i^n\|^2. \quad (22)$$

Algorithm 1 based on (22) reads as a block-Jacobi schemes converging to the stationary solution of f over $\mathcal{K}$ (cf. Theorem 3). Note that if f is jointly convex in all variables, every stationary solution is a global minimizer of f on $\mathcal{K}$. To the best of our knowledge, these are new algorithms in the literature; moreover their convergence conditions enlarge current ones; see, e.g., [4, Sec. 3.2.4]. Quite interestingly, this new algorithm can be readily applied to solve the sum-rate maximization over MIMO *multiple access* channels [38], resulting in the first (inexact) *simultaneous* MIMO iterative waterfilling algorithm in the literature; we omit the details because of the space limitation.

Example #4–Algorithms for DC Programming: The proposed framework applies naturally to sum-utility problems where the users' functions are the difference of two convex functions, namely:

$$\begin{array}{ll} \underset{\mathbf{x}_1, \ldots, \mathbf{x}_I}{\text{minimize}} & \displaystyle\sum_{i \in \mathcal{I}} f_i^{\text{cvx}}(\mathbf{x}) + \sum_{i \in \mathcal{I}} f_i^{\text{ccv}}(\mathbf{x}) \\ \text{subject to} & \mathbf{x}_i \in \mathcal{K}_i, \ \ \forall i \in \mathcal{I} \end{array} \quad (23)$$

where $f_i^{\text{cvx}}(\mathbf{x})$ and $f_i^{\text{ccv}}(\mathbf{x})$ are convex and concave functions on $\mathcal{K}$, respectively. Letting

$$f_1(\mathbf{x}) \triangleq \sum_{i \in \mathcal{I}} f_i^{\text{cvx}}(\mathbf{x}) \quad \text{and} \quad f_2(\mathbf{x}) \triangleq \sum_{i \in \mathcal{I}} f_i^{\text{ccv}}(\mathbf{x}),$$

the optimization problem (23) can be interpreted as a special case of the framework (1), with $\mathcal{I}_f = \{1, 2\}$, $\mathcal{C}_i = \{1\}$ for all $i \in \mathcal{I}$. The best-response (10) of each user i reduces then to

$$\widehat{\mathbf{x}}_{\mathcal{C}_i}(\mathbf{x}^n, \tau_i) = \underset{\mathbf{x}_i \in \mathcal{K}_i}{\operatorname{argmin}} \left\{ f_1(\mathbf{x}_i, \mathbf{x}_{-i}^n) + \boldsymbol{\pi}_i(\mathbf{x}^n)^T \mathbf{x}_i + \frac{\tau_i}{2} \|\mathbf{x}_i - \mathbf{x}_i^n\|^2 \right\} \quad (24)$$

where $\boldsymbol{\pi}_i(\mathbf{x}^n) \triangleq \nabla_{\mathbf{x}_i} f_2(\mathbf{x}^n)$ and $\mathbf{H}_i(\mathbf{x}^n) = \mathbf{I}$. The above decomposition can be applied, e.g., to the sum-rate maximization (3), when all $\theta_i(x) = w_i\, x$, with $w_i > 0$; see Section VII.

B. Inexact Jacobi Best-Response Schemes

In many practical network settings, it can be useful to further reduce the computational effort needed to solve users' (convex) sub-problems (10) by allowing inexact computations of the best-response functions $\widehat{\mathbf{x}}_{\mathcal{C}_i}(\mathbf{x}^n, \boldsymbol{\tau})$. Algorithm 2 is a variant of Algorithm 1, in which suitable approximations of $\widehat{\mathbf{x}}_{\mathcal{C}_i}(\mathbf{x}^n, \boldsymbol{\tau})$ can be used.

Algorithm 2 : Inexact Jacobi SCA Algorithm

Data : $\{\varepsilon_i^n\}$ for $i \in \mathcal{I}$, $\boldsymbol{\tau} \geq \mathbf{0}$, $\{\gamma^n\} > 0$, $\mathbf{x}^0 \in \mathcal{K}$. Set $n = 0$.

(S.1): If $\mathbf{x}^n$ satisfies a termination criterion: STOP;
(S.2): For all $i \in \mathcal{I}$, solve (10) within the accuracy ε_i^n: Find $\mathbf{z}_i^n$ s.t. $\|\mathbf{z}_i^n - \widehat{\mathbf{x}}_{\mathcal{C}_i}(\mathbf{x}^n, \boldsymbol{\tau})\| \leq \varepsilon_i^n$;
(S.3): Set $\mathbf{x}^{n+1} \triangleq \mathbf{x}^n + \gamma^n (\mathbf{z}^n - \mathbf{x}^n)$;
(S.4): $n \leftarrow n + 1$, and go to (S.1).

The error term ε_i^n in Step 2 measures the accuracy used at iteration n in computing the solution $\widehat{\mathbf{x}}_{\mathcal{C}_i}(\mathbf{x}^n, \boldsymbol{\tau})$ of each problem (10). Note that if we set $\varepsilon_i^n = 0$ for all n and i, Algorithm 2 reduces to Algorithm 1. Obviously, the errors ε_i^n's and the step-size γ^n's must be chosen according to some suitable conditions, if one wants to guarantee convergence. These conditions are established in the following theorem.

Theorem 4: Let $\{\mathbf{x}^n\}_{n=1}^{\infty}$ be the sequence generated by Algorithm 2, under the setting of Theorem 3 with the addition assumption that $\nabla_{\mathbf{x}} U$ is bounded on $\mathcal{K}$. Suppose that $\{\gamma^n\}$ and $\{\varepsilon_i^n\}$ satisfy the following conditions: i) $\gamma^n \in (0, 1]$; ii) $\gamma^n \to 0$; iii) $\sum_n \gamma^n = +\infty$; iv) $\sum_n (\gamma^n)^2 < +\infty$; and v) $\sum_n \varepsilon_i^n \gamma^n < +\infty$ for all $i = 1, \ldots, I$. Then, either Algorithm 2 converges in a finite number of iterations to a stationary solution of (1) or every limit point of the sequence $\{\mathbf{x}^n\}_{n=1}^{\infty}$ (at least one such points exists) is a stationary solution of (1).

Proof: See Appendix B. ■

As expected, in the presence of errors, convergence of Algorithm 2 is guaranteed if the sequence of approximated problems (10) is solved with increasing accuracy. Note that, in addition to requiring $\varepsilon_i^n \to 0$, condition v) of Theorem 4 imposes also a constraint on the rate by which the ε_i^n go to zero, which depends on the rate of decrease of $\{\gamma^n\}$. Two instances of step-size rules satisfying the summability condition iv) are given by (19) and (some choices of) (20). An example of error sequence satisfying condition v) is $\varepsilon_i^n \leq c_i\, \gamma^n$, where c_i is any finite positive constant. Such a condition can be forced in Algorithm 2 in a distributed way, using classical error bound results in convex analysis; see, e.g., [17, Ch. 6, Prop. 6.3.7].

Finally, it is worth observing that Algorithm 2 (and 1) with a diminishing step-size rule satisfying i)-iv) of Theorem 4 can be made robust against (stochastic) errors on the price estimates,

due to an imperfect communication scenario (random link failures, noisy estimate, quantization, etc.). Because of the space limitation, we do not further elaborate on this here; see [39] for details.

C. (Inexact) Gauss-Seidel Best-Response Schemes

The Gauss-Seidel implementation of the proposed SCA decomposition is described in Algorithm 3, where the users solve sequentially, in an exact or inexact form, the convex subproblems (10). In the algorithm, we used the notation $\mathbf{x}_{i<}^{t+1} \triangleq (\mathbf{x}_1^{t+1}, \ldots, \mathbf{x}_{i-1}^{t+1})$ and $\mathbf{x}_{i\geq}^{t} \triangleq (\mathbf{x}_i^{t}, \ldots, \mathbf{x}_I^{t})$.

Algorithm 3 : Inexact Gauss-Seidel SCA Algorithm

Data : $\{\varepsilon_i^t\}$ for $i \in \mathcal{I}$, $\boldsymbol{\tau} \geq \mathbf{0}$, $\{\gamma^t\} > 0$, $\mathbf{x}^0 \in \mathcal{K}$. Set $t = 0$.

(S.1): If $\mathbf{x}^t$ satisfies a termination criterion: STOP;
(S.2): For $i = 1, \ldots, I$,
a) Find $\mathbf{z}_i^t$ s.t. $\|\mathbf{z}_i^t - \widehat{\mathbf{x}}_{\mathcal{C}_i}\left((\mathbf{x}_{i<}^{t+1}, \mathbf{x}_{i\geq}^{t}), \boldsymbol{\tau}\right)\| \leq \varepsilon_i^t$;
b) Set $\mathbf{x}_i^{t+1} \triangleq \mathbf{x}_i^t + \gamma^t\,(\mathbf{z}_i^t - \mathbf{x}_i^t)$
(S.3): $t \leftarrow t+1$, and go to (S.1).

Note that one round of Algorithm 3 (i.e., $t \leftarrow t+1$) wherein all users sequentially update their own strategies, corresponds to I consecutive iterations n of the Jacobi updates described in Algorithms 1 and 2. In Appendix C we prove that, quite interestingly, Algorithm 3 can be interpreted as an inexact Jacobi scheme based on the best-response $\widehat{\mathbf{x}}_{\mathcal{C}}(\bullet, \boldsymbol{\tau})$, satisfying Theorem 4. It turns out that convergence of Algorithm 3 follows readily from that of Algorithm 2, and is stated next.

Theorem 5: Let $\{\mathbf{x}^n\}_{n=1}^{\infty}$ be the sequence generated by Algorithm 3, under the setting of Theorem 4. Then, the conclusions of Theorem 4 holds.

Proof: See Appendix C. ■

V. The Complex Case

In this section we show how to extend our framework to sum-utility problems where the users' optimization variables are complex matrices. This will allow us to deal with the design of MIMO multiuser systems. Let us consider the following sum-utility optimization:

$$\begin{aligned} \underset{\mathbf{X}_1,\ldots,\mathbf{X}_I}{\text{minimize}} \quad & U(\mathbf{X}) \triangleq \sum_{\ell \in \mathcal{I}_f} f_\ell(\mathbf{X}) \\ \text{subject to} \quad & \mathbf{X}_i \in \mathcal{X}_i, \quad \forall i \in \mathcal{I}, \end{aligned} \tag{25}$$

where $\mathbf{X} \triangleq (\mathbf{X}_i)_{i\in\mathcal{I}}$, with $\mathbf{X}_i \in \mathbb{C}^{n_i \times m_i}$ being the (matrix) strategy of user i, $\mathcal{X}_i \subseteq \mathbb{C}^{n_i\times m_i}$, and $f_\ell : \mathcal{X} \to \mathbb{R}$, with $\mathcal{X} \triangleq \prod_{i\in\mathcal{I}} \mathcal{X}_i$; let define also $\mathcal{X}_{-i} \triangleq \prod_{j\neq i} \mathcal{X}_j$. We study (25) under the same assumptions A1–A4 stated for the real case, where in A2 the differentiability condition is now replaced by the $\mathbb{R}$-differentiability (see, e.g., [40], [41]), and in A3 $U(\mathbf{X})$ is required to have Lipschitz *conjugate-gradient* $\nabla_{\mathbf{X}^*} U(\mathbf{X})$ on $\mathcal{K}$, with constant $L_{\nabla U}^{\mathbb{C}}$, where $\mathbf{X}^*$ is the conjugate of $\mathbf{X}$.

A Motivating Example. An instance of (25) is the MIMO version of (3):

$$\begin{aligned} \underset{\mathbf{Q}_1,\ldots,\mathbf{Q}_I}{\text{maximize}} \quad & \sum_{i\in\mathcal{I}} \theta_i\left(R_i(\mathbf{Q}_i, \mathbf{Q}_{-i})\right) \\ \text{subject to} \quad & \mathbf{Q}_i \in \mathcal{Q}_i, \quad \forall i \in \mathcal{I}. \end{aligned} \tag{26}$$

where $R_i(\mathbf{Q}_i, \mathbf{Q}_{-i})$ is the rate over the MIMO link i,

$$R_i(\mathbf{Q}_i, \mathbf{Q}_{-i}) \triangleq \log\det\left(\mathbf{I} + \mathbf{H}_{ii}^H \mathbf{R}_i(\mathbf{Q}_{-i})^{-1}\mathbf{H}_{ii}\mathbf{Q}_i\right), \tag{27}$$

$\mathbf{Q}_i$ is the covariance matrix of transmitter i, $\mathbf{R}_i(\mathbf{Q}_{-i}) \triangleq \mathbf{R}_{n_i} + \sum_{j\neq i} \mathbf{H}_{ij}\mathbf{Q}_j\mathbf{H}_{ij}^H$ is the covariance matrix of the multiuser interference plus the thermal noise $\mathbf{R}_{n_i}$ (assumed to be full-rank), with $\mathbf{Q}_{-i} \triangleq (\mathbf{Q}_j)_{j\neq i}$, $\mathbf{H}_{ij}$ is the channel matrix between the j-th transmitter and the i-th receiver, and $\mathcal{Q}_i$ is the set of constraints of user i,

$$\mathcal{Q}_i \triangleq \left\{\mathbf{Q}_i \in \mathbb{C}^{n_i\times n_i} : \mathbf{Q}_i \succeq \mathbf{0},\, \mathrm{tr}(\mathbf{Q}_i) \leq P_i,\, \mathbf{Q}_i \in \mathcal{Z}_i\right\}.$$

In $\mathcal{Q}_i$ we also included an arbitrary convex and closed set $\mathcal{Z}_i$, which allows us to add additional constraints, such as: i) null constraints $\mathbf{U}_i^H\mathbf{Q}_i = \mathbf{0}$, where $\mathbf{U}_i \in \mathbb{C}^{n_i\times r_i}$ is a full rank matrix with $r_i < n_i$; ii) soft-shaping constraints $\mathrm{tr}\left(\mathbf{G}_i^H\mathbf{Q}_i\mathbf{G}_i\right) \leq I_i^{\text{ave}}$, with $\mathbf{G}_i \in \mathbb{C}^{n_i\times m_{G_i}}$ for some $m_{G_i} > 0$; iii) peak-power constraints $\lambda_{\max}\left(\mathbf{F}_i^H\mathbf{Q}_i\mathbf{F}_i\right) \leq I_i^{\text{peak}}$, with $\mathbf{F}_i \in \mathbb{C}^{n_i\times m_{F_i}}$ for some $m_{F_i} > 0$; and iv) per-antenna constraints $[\mathbf{Q}_i]_{kk} \leq \alpha_{ik}$. Note that the optimization problems in [23], [24], [26] are special cases of (26).

A. Distributed Decomposition Algorithms

At the basis of the proposed decomposition techniques for (25) there is the (second order) Taylor expansion of a continuously $\mathbb{R}$-differentiable function $f : \mathbb{C}^{n\times m} \to \mathbb{R}$ [41]:

$$\begin{aligned} f(\mathbf{X} + \Delta\mathbf{X}) - f(\mathbf{X}) \approx\; & 2\left\langle \Delta\mathbf{X},\, \nabla_{\mathbf{X}^*} f(\mathbf{X})\right\rangle \\ & + \frac{1}{2}\,\mathrm{vec}([\boldsymbol{\Delta}\mathbf{X}, \boldsymbol{\Delta}\mathbf{X}^*])^H \mathcal{H}_{\mathbf{X}\mathbf{X}^*} f(\mathbf{X})\,\mathrm{vec}([\boldsymbol{\Delta}\mathbf{X}, \boldsymbol{\Delta}\mathbf{X}^*]), \end{aligned} \tag{28}$$

where $\langle \mathbf{A},\, \mathbf{B}\rangle \triangleq \mathrm{Re}\left\{\mathrm{tr}(\mathbf{A}^H\mathbf{B})\right\}$, $\mathrm{vec}(\bullet)$ denotes the "vec" operator, and $\mathcal{H}_{\mathbf{X}\mathbf{X}^*} f(\mathbf{X})$ is the so-called *augmented Hessian* of f, defined as [41]

$$\mathcal{H}_{\mathbf{X}\mathbf{X}^*} f(\mathbf{X}) \triangleq \frac{\partial}{\partial \mathrm{vec}([\mathbf{X}, \mathbf{X}^*])^T}\left(\frac{\partial f(\mathbf{X})}{\partial \mathrm{vec}([\mathbf{X}^*, \mathbf{X}])^T}\right)^T. \tag{29}$$

In [41], we proved that $\mathcal{H}_{\mathbf{X}\mathbf{X}^*} f(\mathbf{X})$ plays the role of the Hessian matrix for functions of real variables. In particular, f is strongly convex on $\mathbb{C}^{n\times m}$ if and only if there exists a $c_{f^{\mathbb{C}}} > 0$, the constant of strong convexity of f, such that

$$\mathrm{vec}([\mathbf{Y}, \mathbf{Y}^*])^H \mathcal{H}_{\mathbf{X}\mathbf{X}^*} f(\mathbf{X})\,\mathrm{vec}([\mathbf{Y}, \mathbf{Y}^*]) \geq c_{f^{\mathbb{C}}}\,\|\mathbf{Y}\|_F^2, \tag{30}$$

for all $\mathbf{X} \in \mathbb{C}^{n\times m}$ and $\mathbf{Y} \in \mathbb{C}^{n\times m}$, where $\|\bullet\|_F$ denotes the Frobenius norm. When (30) holds, we say that $\mathcal{H}_{\mathbf{X}\mathbf{X}^*} f(\mathbf{X})$ is *augmented* uniformly positive definite, and write $\mathcal{H}_{\mathbf{X}\mathbf{X}^*} f(\mathbf{X}) - c_{f^{\mathbb{C}}}\mathbf{I} \overset{\mathcal{A}}{\succeq} \mathbf{0}$ [41]. If f is only convex but not strongly convex, then $c_{f^{\mathbb{C}}}$ in (30) is zero.

Motivated by the Taylor expansion (28), and using the same symbols $\mathcal{S}_i$ and $\mathcal{C}_i$ to denote the complex counterparts of $\mathcal{S}_i$ and $\mathcal{C}_i$ introduced for the real case [cf. (6)], let us consider for each user i the following convex approximation of $U(\mathbf{X})$ at $\mathbf{X}^n$: denoting by $\Delta\mathbf{X}_i \triangleq \mathbf{X}_i - \mathbf{X}_i^n$,

$$\begin{aligned} \widetilde{f}_{\mathcal{C}_i}(\mathbf{X}_i; \mathbf{X}^n) \triangleq\; & \sum_{j\in\mathcal{C}_i} f_j(\mathbf{X}_i,\, \mathbf{X}_{-i}^n) + \left\langle \boldsymbol{\Pi}_{\mathcal{C}_i}(\mathbf{X}^n),\, \Delta\mathbf{X}_i\right\rangle \\ & + \frac{\tau_i}{2}\,\mathrm{vec}([\Delta\mathbf{X}_i, \Delta\mathbf{X}_i^*])^H \mathcal{H}_i(\mathbf{X}^n)\,\mathrm{vec}([\Delta\mathbf{X}_i, \Delta\mathbf{X}_i^*]) \end{aligned} \tag{31}$$

with

$$\mathbf{\Pi}_{\mathcal{C}_i}(\mathbf{X}^n) \triangleq \sum_{j\in\mathcal{C}_{-i}} \nabla_{\mathbf{X}_i^*} f_j(\mathbf{X})\big|_{\mathbf{X}=\mathbf{X}^n}, \tag{32}$$

where $\mathcal{H}_i(\mathbf{X}^n)$ is any given 2 nm $\times$ 2 nm matrix such that $\mathcal{H}_i(\mathbf{X}) - c_{\mathcal{H}_i}\mathbf{I} \overset{\mathcal{A}}{\succeq} \mathbf{0}$, for all $\mathbf{X} \in \mathcal{X}$ and some $c_{\mathcal{H}_i} > 0$. Note that if $\mathcal{H}_i(\mathbf{X}) = \mathbf{I}$, the quadratic term in (31) reduces to the standard proximal regularization $\tau_i \|\mathbf{X}_i - \mathbf{X}_i^n\|_F^2$. Then, the best-response matrix function of each user is

$$\widehat{\mathbf{X}}_{\mathcal{C}_i}(\mathbf{X}^n, \tau_i) \triangleq \operatorname*{argmin}_{\mathbf{X}_i\in\mathcal{X}_i} \widetilde{f}_{\mathcal{C}_i}(\mathbf{X}_i; \mathbf{X}^n). \tag{33}$$

Decomposition algorithms for (25) are formally the same as those proposed in Section IV for (1) [namely Algorithms 1–3], where the real-valued best-response map $\widehat{\mathbf{x}}_{\mathcal{C}}(\mathbf{x}^n, \boldsymbol{\tau})$ is replaced with the complex-valued counterpart $\widehat{\mathbf{X}}_{\mathcal{C}}(\mathbf{X}^n, \boldsymbol{\tau}) \triangleq (\widehat{\mathbf{X}}_{\mathcal{C}_i}(\mathbf{X}^n, \tau_i))_{i=1}^I$. Convergence conditions read as in Theorems 3–5, under the following natural changes: i) $L_{\nabla U}$ becomes $L_{\nabla U}^{\mathbb{C}}$; ii) the condition $\mathbf{H}_i(\mathbf{x}) - c_{H_i}\mathbf{I} \succeq \mathbf{0}$ for all $\mathbf{x} \in \mathcal{K}$ reads as $\mathcal{H}_i(\mathbf{X}) - c_{\mathcal{H}_i}\mathbf{I} \overset{\mathcal{A}}{\succeq} \mathbf{0}$, for all $\mathbf{X} \in \mathcal{X}$; and iii) in the constant $c_{\boldsymbol{\tau}}$ defined in (14) $c_{\tau_i}(\mathbf{x})$ is replaced with $c_{\tau_i}^{\mathbb{C}}(\mathbf{X})$, where $c_{\tau_i}^{\mathbb{C}}(\mathbf{X}) > 0$ is the constant of strong convexity of $\widetilde{f}_{\mathcal{C}_i}(\bullet; \mathbf{X})$ [41]:

$$\left\langle \mathbf{Z}_i - \mathbf{W}_i, \nabla_{\mathbf{X}_i^*}\widetilde{f}_{\mathcal{C}_i}(\mathbf{Z}_i; \mathbf{X}) - \nabla_{\mathbf{X}_i^*}\widetilde{f}_{\mathcal{C}_i}(\mathbf{W}_i; \mathbf{X}) \right\rangle \geq c_{\tau_i}^{\mathbb{C}}(\mathbf{X}) \|\mathbf{Z}_i - \mathbf{W}_i\|_F^2, \quad \forall \mathbf{Z}_i, \mathbf{W}_i \in \mathcal{X}_i.$$

VI. Extensions and Related Works

The key idea in the proposed SCA schemes, e.g., (33), is to convexify the nonconvex part of U via partial linearization of $\sum_{j\in\mathcal{C}_{-i}} f_j(\mathbf{X})$, resulting in the term $\langle \mathbf{\Pi}_{\mathcal{C}_i}(\mathbf{X}^n), \Delta\mathbf{X}_i\rangle$. In the same spirit of [27], [32], [33], it is not difficult to show that one can generalize this idea and replace the linear term $\langle \mathbf{\Pi}_{\mathcal{C}_i}(\mathbf{X}^n), \Delta\mathbf{X}_i\rangle$ in (31) with a nonlinear scalar function $\Pi_{\mathcal{C}_i}(\bullet; \mathbf{X}^n) : \mathcal{X}_i \ni \mathbf{X}_i \mapsto \Pi_{\mathcal{C}_i}(\mathbf{X}_i; \mathbf{X}^n)$. All the results presented so far are still valid provided that $\Pi_{\mathcal{C}_i}(\bullet; \mathbf{X}^n)$ enjoys the following properties: for all $\mathbf{X}^n \in \mathcal{X}$,

P1) $\Pi_{\mathcal{C}_i}(\bullet; \mathbf{X}^n)$ is $\mathbb{R}$-continuously differentiable on $\mathcal{X}_i$;
P2) $\nabla_{\mathbf{X}_i^*}\Pi_{\mathcal{C}_i}(\mathbf{X}_i^n; \mathbf{X}^n) = \sum_{j\in\mathcal{C}_{-i}} \nabla_{\mathbf{X}_i^*} f_j(\mathbf{X}^n)$;
P3) $\nabla_{\mathbf{X}_i^*}\Pi_{\mathcal{C}_i}(\mathbf{X}_i^n; \bullet)$ is uniformly Lipschitz on $\mathcal{X}$;
P4) $\Pi_{\mathcal{C}_i}(\mathbf{X}_i; \mathbf{X}^n)$ is continuous in $(\mathbf{X}_i; \mathbf{X}^n) \in \mathcal{X}_i \times \mathcal{X}$.

Similar conditions can be written in the real case for the nonlinear function $\pi_{\mathcal{C}_i}(\bullet; \mathbf{x}^n) : \mathcal{K}_i \ni \mathbf{x}_i \mapsto \pi_{\mathcal{C}_i}(\mathbf{x}_i; \mathbf{x}^n)$ replacing the linear pricing $\boldsymbol{\pi}_{\mathcal{C}_i}(\mathbf{x}^n)^T\mathbf{x}_i$. It is interesting to compare P1-P3 with conditions in [27], [32], [33]. First of all, our conditions do not require that the approximation function is a global upper bound of the original sum-utility function, a constraint that remains elusive for sum-utility problems with no special structure. Second, even when the aforementioned constraint can be met, it is not always guaranteed that the resulting convex subproblems are decomposable across the users, implying that a centralized implementation might be required. Third, SCA algorithms [27], [32], [33], even when distributed, are generally *sequential* schemes (unless the sum-utility has a special structure). On the contrary, the algorithms proposed in this paper do not suffer from any of the above drawbacks, which enlarges substantially the class of (large scale) nonconvex problems solvable using our framework.

VII. Applications and Numerical Results

In this section, we customize the proposed decomposition framework to the SISO and MIMO sum-rate maximization problems introduced in (3) and (26), respectively, and compare the resulting new algorithms with state-of-the-art schemes [23], [24], [29], [30], [33]. Quite interestingly, our algorithms are shown to outperform current schemes, in terms of convergence speed and computational effort, while reaching the same sum-rate. It is worth mentioning that this was not obvious at all, because algorithms in [23], [24], [29], [30], [33] are ad-hoc schemes for the sum-rate problem, whereas our framework has been introduced for general sum-utility problems.

A. Sum-rate maximization over SISO ICs

Consider the social problem (3), with $\theta_i(r_i(\mathbf{p}_i, \mathbf{p}_{-i})) = w_i r_i(\mathbf{p}_i, \mathbf{p}_{-i})$, where w_i are positive given weights; to avoid redundant constraints, let also assume w.l.o.g. that all the columns of $\mathbf{W}_i$ are linearly independent. We describe next two alternative decompositions for (3) corresponding to differ choices of $\mathcal{I}_f$ and the sets $\mathcal{C}_i$.

1) Decomposition #1–Pricing Algorithms: Since each user's rate $r_i(\mathbf{p}_i, \mathbf{p}_{-i})$ is concave in $\mathbf{p}_i \in \mathcal{P}_i$, a natural choice is $\mathcal{I}_f = \mathcal{I}$ and $\mathcal{C}_i = \{i\}$, which leads to the following class of strongly concave subproblems [cf. (7)]: given $\mathbf{p}^n = (\mathbf{p}_i^n)_{i=1}^I$ and choosing $\mathbf{H}_i(\mathbf{p}^n) = \mathbf{I}$, the best-response of user i is

$$\hat{\mathbf{p}}_i(\mathbf{p}^n) \triangleq \operatorname*{argmax}_{\mathbf{p}_i\in\mathcal{P}_i} \left\{ w_i\, r_i(\mathbf{p}_i, \mathbf{p}_{-i}^n) + \boldsymbol{\pi}_i(\mathbf{p}^n)^T\mathbf{p}_i - \frac{\tau_i}{2}\|\mathbf{p}_i - \mathbf{p}_i^n\|^2 \right\},$$

where $\boldsymbol{\pi}_i(\mathbf{p}^n) \triangleq (\pi_{ik}(\mathbf{p}^n))_{k=1}^N$ is the pricing factor, given by

$$\pi_{i,k}(\mathbf{p}^n) \triangleq -\sum_{j\in\mathcal{N}_i} w_j\, |H_{ji}(k)|^2 \frac{\mathtt{snr}_{jk}^n}{(1+\mathtt{snr}_{jk}^n)\cdot \mathtt{mui}_{jk}^n}; \tag{34}$$

$\mathcal{N}_i$ denotes the set of neighbors of user i, i.e., the set of users j's which user i interferers with; and $\mathtt{snr}_{jk}^n$ and $\mathtt{mui}_{jk}^n$ are the SINR and the multiuser interference-plus-noise power experienced by user j, generated by the power profile $\mathbf{p}^n$:

$$\mathtt{snr}_{jk}^n \triangleq \frac{|H_{jj}(k)|^2 p_{jk}^n}{\mathtt{mui}_{jk}^n},\ \mathtt{mui}_{jk}^n \triangleq \sigma_{jk}^2 + \sum_{i\neq j} |H_{ji}(k)|^2 p_{ik}^n.$$

The best-response $\hat{\mathbf{p}}_i(\mathbf{p}^n)$ can be computed in closed form (up to the multipliers associated with the inequality constraints in $\mathcal{P}_i$) according to the following multi-level waterfilling-like expression [41]:

$$\hat{\mathbf{p}}_i(\mathbf{p}^n) \triangleq \left[\frac{1}{2}\,\mathbf{p}_i^n \circ \left(\mathbf{1} - (\mathbf{snr}_i^n)^{-1}\right) + -\frac{1}{2\,\tau_i}\left(\tilde{\boldsymbol{\mu}}_i - \sqrt{\left[\tilde{\boldsymbol{\mu}}_i - \tau_i\,\mathbf{p}_i^n \circ (\mathbf{1} + (\mathbf{snr}_i^n)^{-1})\right]^2 + 4\tau_i w_i \mathbf{1}}\right)\right]^+ \tag{35}$$

where $\circ$ denotes the Hadamard product, $(\mathbf{snr}_i^n)^{-1} \triangleq (1/\mathtt{snr}_{ik}^n)_{k=1}^N$ and $\tilde{\boldsymbol{\mu}}_i \triangleq \boldsymbol{\pi}_i(\mathbf{p}^n) + \mathbf{W}_i^T\boldsymbol{\mu}_i$, with the multiplier vector $\boldsymbol{\mu}_i$ chosen to satisfy the nonlinear complementarity condition (CC) $\mathbf{0} \leq \boldsymbol{\mu}_i \perp \mathbf{I}_i^{\max} - \mathbf{W}_i\hat{\mathbf{p}}_i(\mathbf{p}^n) \geq \mathbf{0}$. The optimal

$\boldsymbol{\mu}_i$ satisfying the CC can be efficiently computed (in a finite number of steps) using a multiple nested bisection method as described in [41, Alg. 6]; we omit the details because of the space limitation. Note that, in the presence of the power budget constraint only (as in [23], [29], [30]), $\boldsymbol{\mu}_i$ reduces to a scalar quantity μ_i such that $0 \leq \mu_i \perp P_i - \mathbf{1}^T \hat{\mathbf{p}}_i(\mathbf{p}^n) \geq 0$, whose solution can be obtained using the classical bisection algorithms (or the methods in [42]).

Given $\hat{\mathbf{p}}_i(\mathbf{p}^n)$, one can now use any of the algorithms introduced in Section IV. For instance, a good candidate is the exact Jacobi scheme with diminishing step-size (Algorithm 1), whose convergence is guaranteed if, e.g., rules in (19) or (20) are used for the sequence $\{\gamma^n\}$ (Theorem 3). Note that the proposed algorithm is fairly distributed. Indeed, given the interference generated by the other users [and thus the MUI coefficients $\mathtt{mui}_{jk}^n$] and the current interference price $\boldsymbol{\pi}_i(\mathbf{p}^n)$, each user can efficiently and locally compute the optimal power allocation $\hat{\mathbf{p}}_i(\mathbf{p}^n)$ via the waterfilling-like expression (35). The estimation of the prices $\pi_{ik}(\mathbf{p}^n)$ requires however some signaling among nearby users. Interestingly, the pricing expression in (34) as well as the signaling overhead necessary to compute it coincides with that in [23]. But, because of their sequential nature, algorithms in [23] require more CSI exchange in the network than our *simultaneous* schemes.

2) Decomposition #2–DC Algorithms: An alternative class of algorithms for the sum-rate maximization problem under consideration can be obtained exploring the D.C. nature of the rate functions (cf. Example #4 in Section IV-A). The sum-rate can indeed be decomposed as the sum of a concave and convex function, namely $U(\mathbf{p}) = f_1(\mathbf{p}) + f_2(\mathbf{p})$, where

$$f_1(\mathbf{p}) \triangleq \sum_i w_i \sum_k \log\left(\sigma_{i,k}^2 + \sum_j |H_{ij}(k)|^2 p_{jk}\right)$$

$$f_2(\mathbf{p}) \triangleq -\sum_i w_i \sum_k \log\left(\sigma_{i,k}^2 + \sum_{j\neq i} |H_{ij}(k)|^2 p_{jk}\right),$$

which is an instance of (23) with $\mathcal{I}_f = \{1, 2\}$. A natural choice of $\mathcal{C}_i$ is then $\mathcal{C}_i = \{1\}$ for all $i \in \mathcal{I}$, resulting in the best-response:

$$\widetilde{\mathbf{p}}_i(\mathbf{p}^n) \triangleq \operatorname*{argmax}_{\mathbf{p}_i \in \mathcal{P}_i} \left\{ f_1(\mathbf{p}_i, \mathbf{p}_{-i}^n) + \boldsymbol{\pi}_i(\mathbf{p}^n)^T \mathbf{p}_i - \frac{\tau_i}{2} \|\mathbf{p}_i - \mathbf{p}_i^n\|^2 \right\},$$

where $\boldsymbol{\pi}_i(\mathbf{p}^n) \triangleq (\pi_{ik}(\mathbf{p}^n))_{k=1}^N$, with

$$\pi_{i,k}(\mathbf{p}^n) \triangleq -\sum_{j\in\mathcal{N}_i} w_j |H_{ji}(k)|^2 \frac{1}{\mathtt{mui}_{jk}^n}. \tag{36}$$

We remark that the best-response $\widetilde{\mathbf{p}}_i(\mathbf{p}^n)$ can be efficiently computed by a fixed-point iterate, in the same spirit of [29]; we omit the details because of the space limitation. Note that the communication overhead to compute the prices (34) and (36) is the same, but the computation of $\widetilde{\mathbf{p}}_i(\mathbf{p}^n)$ requires more CSI exchange in the network than that of $\hat{\mathbf{p}}_i(\mathbf{p}^n)$, since each user i also needs to estimate the cross-channels $\{|H_{ji}(k)|^2\}_{j\in\mathcal{N}_i}$.

Numerical Example #1. We compare now Algorithm 1 based on the best-response $\hat{\mathbf{p}}_i(\mathbf{p}^n)$ in (35) (termed SJBR), with those proposed in [29] [termed SCALE and SCALE one-step, the latter being a simplified version of SCALE where instead of solving the fixed-point (16) in [29], only one iteration of (16) is performed],[23] (termed MDP), [30] (termed WMMSE). Since in the aforementioned papers only power budget constraints can be dealt with, to allow the comparison, we simplified the sum-rate maximization problem described above and considered only power budget constraints (and all $w_i = 1$). We assume the same power budget $P_i = P$, noise variances $\sigma_{ik}^2 = \sigma^2$, and $\mathtt{snr} = P/\sigma^2 = 3$ dB for all the users. We simulated SISO frequency channels with $N = 64$ subcarriers; the channels are generated as FIR filters of order $L = 10$, whose taps are i.i.d. Gaussian random variables with zero mean and variance $1/(d_{ij}^3(L+1)^2)$, where d_{ij} is the distance between the transmitter j and the receiver i. All the algorithms are initialized by choosing the uniform power allocation, and are terminated when (the absolute value) of the sum-utility error in two consecutive rounds becomes smaller than $1e-6$. The accuracy in the bisection loops (required by all methods) is set to $1e-7$. In our algorithm, we used rule (19) with $\epsilon = 1e-2$ and set all $\tau_i = 0$. In Fig. 1, we plot the average number of iterations required by the aforementioned algorithms to converge versus the number of users; the average is taken over 100 independent channel realizations; we set $d_{ij}/d_{ii} = 3$ and $d_{ij} = d_{ji}$ and $d_{ii} = d_{jj}$ for all i and $j \neq i$. As benchmark, we also plot two instances of proximal conditional gradient algorithms [4], which can be interpreted as special cases of our SJBR with $\mathcal{C}_i = \emptyset$ for all $i \in \mathcal{I}$ (cf. Ex. #1 in Section IV-A). In one instance [termed Gradient (SJBR tuning)] we set the free parameters τ_i and ϵ as in SJBR, whereas in the other one [termed Gradient (opt. tuning)] we chose $\tau_i = 50$ for all $i \in \mathcal{I}$ and $\epsilon = 1e-2$, which leads experimentally to the fastest behavior of the gradient algorithm.

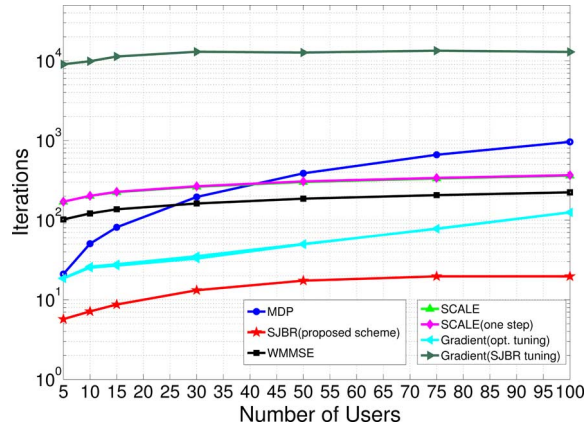

Fig. 1. Average number of iterations versus number of users in SISO frequency-selective ICs. Note that all algorithms are simultaneous except MDP; this means that, at each iteration, in MDP there is only one user updating his strategy, whereas in the other algorithms all users do so).

All the algorithms reach the same average sum-rate (that thus is not reported here, see [43]), but their convergence behavior is quite different. The figure clearly shows that our SJBR outperforms all the others (note that SCALE, WMMSE, and the proximal gradient are also simultaneous-based schemes). For instance, the gap with the WMMSE is about one order of magnitude, for all the network sizes considered in the experiment, while the gap with MDP is up to three orders of magnitude. The good behavior of our scheme has been observed also for other choices of d_{ij}/d_{ii}, termination tolerances, and step-size rules; we cannot present here more experiments because of space limitation; we refer the interested reader to the technical report [43]

for more numerical results. Note that SJBR, SCALE one-step, WMMSE, MDP, and gradient schemes have similar per-user computational complexity, whereas SCALE is much more demanding and is not appealing for a real-time implementation. Therefore, Fig. 1 provides also a rough indication of the per-user cpu time of SJBR, SCALE one-step, WMMSE, and gradient algorithms.

It is also interesting to compare the proposed algorithm with gradient schemes. A first natural question is whether the partial linearization (as performed in SJBR) really improves the convergence speed of the algorithm. The answer is given by the comparison in Fig. 1 between SJBR and "Gradient (SJBR tuning)". One can see that, under the same choice of $\{\gamma^n\}$ and $(\tau_i)_{i=1}^I$, the former is almost three order of magnitude faster then the latter, for all the network sizes considered in the experiment. If an independent, ad hoc tuning of $\{\gamma^n\}$ and $(\tau_i)_{i=1}^I$ is performed for the gradient algorithm, the gap reduces up to one order of magnitude, still in favor of SJBR. This result supports the intuition motivating this work: preserving the structure of the problem via a partial linearization can significantly improve the convergence speed of the algorithm.

The comparison with gradient algorithms also reveals a well-known issue of these schemes: the convergence behavior strongly depends on the choice of the step-size sequence $\{\gamma^n\}$ and the proximal gains τ_i. It is then natural to ask whether also the proposed algorithms suffer from the same drawback. To answer this question, in Fig. 2 we compare the convergence behavior of the proximal condition gradient algorithm with that of SJBR, using the step-size rule (19), but changing the free parameter $\epsilon \in (0,1)$ by several orders of magnitude. For gradient schemes, we considered two choices of τ_i, namely: $\tau_i = 0$ and $\tau_i = 50$ (as in Fig. 1); the latter resulting in the experimentally fastest behavior of gradient schemes (see Fig. 1). More specifically, in Fig. 2, we plot the average number of iterations needed to reach convergence within the accuracy of $1e-6$ versus $\epsilon \in (0,1)$, for different number of users (the rest of the setting is as in Fig. 1). The figure clearly shows that, differently from gradient algorithms, the convergence behavior of our scheme appears to be almost independent of the choice of ϵ. This is a very desirable feature that lets one avoid the expensive and difficult tuning of the step-size, thus making the proposed algorithms a very good candidate in many applications. We remark one more time that the gradient method is very sensitive to the choice of parameters; indeed, based on further simulations that we do not report here for lack of space, the behavior of the gradient method is very sensitive to the number of users and characteristics of the network (SNR, pair distances, etc...) and its optimal behavior requires different tunings of parameters each time.

B. Sum-rate maximization over MIMO ICs

Let us focus now on the MIMO formulation (26), assuming $\theta_i(R_i(\mathbf{Q}_i,\mathbf{Q}_{-i})) = w_i\, R_i(\mathbf{Q}_i,\mathbf{Q}_{-i})$, with $w_i > 0$.

1) Decomposition #1: Pricing Algorithms: Choosing $I_f = I$, $\mathcal{C}_i = \{i\}$, and $\mathcal{H}_i(\mathbf{Q}^n) = \mathbf{I}$, the best-response of user i is

$$\hat{\mathbf{Q}}_i(\mathbf{Q}^n,\tau_i) \triangleq \underset{\mathbf{Q}_i\in\mathcal{Q}_i}{\operatorname{argmax}}\big\{w_i\, r_i(\mathbf{Q}_i,\mathbf{Q}_{-i}^n)+\langle \mathbf{\Pi}_i(\mathbf{Q}^n),\mathbf{Q}_i-\mathbf{Q}_i^n\rangle -\tau_i\,\|\mathbf{Q}_i-\mathbf{Q}_i^n\|_F^2\big\} \quad (37)$$

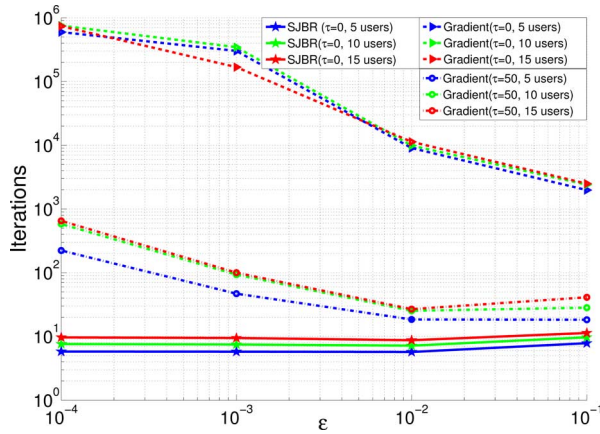

Fig. 2. Proximal conditional gradient algorithms versus SJBR: Average number of iterations versus $\epsilon \in (0,1)$ [cf. (19)].

with

$$\mathbf{\Pi}_i(\mathbf{Q}^n) \triangleq -\sum_{j\in\mathcal{N}_i} w_j\, \mathbf{H}_{ji}^H \widetilde{\mathbf{R}}_j(\mathbf{Q}_{-j}^n)\mathbf{H}_{ji},$$

where $\mathcal{N}_i$ is defined as in the SISO case, and

$$\widetilde{\mathbf{R}}_j(\mathbf{Q}_{-j}^n) \triangleq \mathbf{R}_j(\mathbf{Q}_{-j}^n)^{-1} - (\mathbf{R}_j(\mathbf{Q}_{-j}^n) + \mathbf{H}_{jj}\mathbf{Q}_j^n\mathbf{H}_{jj}^H)^{-1}.$$

Note that, once the price matrix $\mathbf{\Pi}_i(\mathbf{Q}^n)$ is given, the best-response $\hat{\mathbf{Q}}_i(\mathbf{Q}^n,\tau_i)$ can be computed locally by each user solving a convex optimization problem. Moreover, for some specific structures of the feasible sets $\mathcal{Q}_i$, the case of full-column rank channel matrices $\mathbf{H}_i$, and $\tau_i = 0$, a solution in closed form (up to the multipliers associated with the power budget constraints) is also available [24]. Given $\hat{\mathbf{Q}}_i(\mathbf{Q}^n,\tau_i)$, one can now use any of the algorithms introduced in Section V. To the best of our knowledge, our schemes are the first class of *best-response Jacobi* (inexact) algorithms for MIMO IC systems based on *pricing* with provable convergence.

Complexity Analysis and Message Exchange. It is interesting to compare the computational complexity and signaling (i.e., message exchange) of our algorithms, e.g., Algorithm 1 based on the best-response $\hat{\mathbf{Q}}_i(\mathbf{Q}^n,\tau_i)$ (termed MIMO-SJBR) with those of the schemes proposed in the literature for a similar problem, namely the MIMO-MDP [23], [24], and the MIMO-WMMSE [30]. We assume that all channel matrices $\mathbf{H}_{ii}$'s are full-column rank, and set $\tau_i = 0$ in (37). For the purpose of complexity analysis, since all algorithms include a similar bisection step which generally takes few iterations, we will ignore this step in the computation of the complexity (as in [30]). Also, WMMSE and SJBR are simultaneous schemes, while MDP is sequential; we then compare the algorithms by given the *per-round complexity*, where one round means one update of all users. Denoting by n_T (resp. n_R) the number of antennas at each transmitter (resp. receiver), the computational complexity of the algorithms is:

- MIMO-MDP: $\mathcal{O}\left(I^2(n_T n_R^2 + n_T^2 n_R + n_R^3) + I\, n_T^3\right)$
- MIMO-WMMSE: $\mathcal{O}\left(I^2(n_T n_R^2 + n_T^2 n_R + n_T^3) + I\, n_R^3\right)$ [30]
- MIMO-SJBR: $\mathcal{O}\left(I^2(n_T n_R^2 + n_T^2 n_R) + I(n_T^3 + n_R^3)\right)$.

It is clear that the complexity of the three algorithms is very similar, and same in order in the case in which $n_T = n_R (\triangleq n)$, given by $\mathcal{O}(I^2 n^3)$.

We remark that, in a real system, the MUI covariance matrices $\mathbf{R}_i(\mathbf{Q}_{-i})$ come from an estimation process. It is thus interesting to understand how the complexity changes when the computation of $\mathbf{R}_i(\mathbf{Q}_{-i})$ from $\mathbf{R}_{n_i} + \sum_{j \neq i} \mathbf{H}_{ij}\mathbf{Q}_j\mathbf{H}_{ij}^H$ is not included in the analysis. We obtain the following:

- MIMO-MDP: $\mathcal{O}\left(I^2(n_T n_R^2 + n_T^2 n_R + n_R^3) + I\, n_T^3\right)$
- MIMO-WMMSE: $\mathcal{O}\left(I^2(n_T^2 n_R + n_T^3) + I(n_R^3 + n_T n_R^2)\right)$
- MIMO-SJBR: $\mathcal{O}\left(I^2(n_T n_R^2 + n_T^2 n_R) + I(n_T^3 + n_R^3)\right)$.

Finally, if one is interested in the time necessary to complete one iteration, it can be shown that it is proportional to the above complexity divided by I.

As far as the communication overhead is concerned, the same remarks we made about the schemes described in the SISO setting, apply also here for the MIMO case. The only difference is that now the users need to exchange a (pricing) matrix rather than a vector, resulting in $\mathcal{O}(I^2 n_T^2)$ amount of message exchange per-iteration for all the algorithms.

2) Decomposition #2–WMMSE Algorithms: In [30], the authors showed that the MIMO problem (26) (under power constraints only) is equivalent to the sum-MSE minimization shown in (38) at the bottom of the page, where $\mathbf{Q}_i = \mathbf{V}_i\mathbf{V}_i^H$, $\mathbf{V} \triangleq (\mathbf{V}_i)_{i=1}^I$, $\mathbf{U} \triangleq (\mathbf{U}_i)_{i=1}^I$ and $\mathbf{W} \triangleq (\mathbf{W})_{i=1}^I$ are two auxiliary matrix variables, and $\mathbf{E}_i(\mathbf{U}, \mathbf{V})$ is the MSE matrix at the receiver i (see (3) in [30]). The formulation (38) has some desirable properties, namely: i) $f(\mathbf{W}, \mathbf{U}, \mathbf{V})$ is continuously ($\mathbb{R}$-)differentiable with Lipschitz continuous (conjugate) gradient on the feasible set; ii) $f(\mathbf{W}, \mathbf{U}, \mathbf{V})$ is convex in each variables $\mathbf{W}$, $\mathbf{U}$, $\mathbf{V}$; iii) the minimization of $f(\mathbf{W}, \mathbf{U}, \mathbf{V})$ w.r.t. to each $\mathbf{W}$, $\mathbf{U}$, $\mathbf{V}$ can be performed in parallel by the users; and iv) the optimal solutions of the individual minimizations are available in closed form, see [30] for details. We will denote such optimal solutions as $\hat{\mathbf{W}}_i(\mathbf{U}, \mathbf{V})$, $\hat{\mathbf{U}}_i(\mathbf{U}, \mathbf{V})$, and $\hat{\mathbf{V}}_i(\mathbf{U}, \mathbf{W})$, for all $i \in \mathcal{I}$, where we made explicit the dependence on the variables that are kept fixed. In [30] the authors proposed to use the (Gauss-Seidel) block coordinate descent method to solve (38), resulting in the so-called MIMO-WMMSE algorithm.

It is not difficult to see that (38) can be cast into our framework, resulting in the following best-response mapping for each user i: $\hat{\mathbf{X}}_i(\mathbf{W}^n, \mathbf{U}^n, \mathbf{V}^n) \triangleq (\hat{\mathbf{W}}_i(\mathbf{U}^n, \mathbf{V}^n), \hat{\mathbf{U}}_i(\mathbf{U}^n, \mathbf{V}^n), \hat{\mathbf{V}}_i(\mathbf{U}^n, \mathbf{W}^n))$. We can then compute a stationary solution of (38) and thus (26) using any of the *Jacobi* algorithms introduced in the previous sections based on $\hat{\mathbf{X}}_i(\mathbf{W}^n, \mathbf{U}^n, \mathbf{V}^n)$ (or its inexact computation). Note that the computational complexity as well as the communication overhead of such algorithms are roughly the same of those of the MIMO-WMMSE [30].

TABLE I
AVERAGE NUMBER OF ITERATIONS (termination accuracy $= 1e-6$)

	# of users $= 10$			# of users $= 50$			# of users $= 100$		
	d=1	d=2	d=3	d=1	d=2	d=3	d=1	d=2	d=3
MDP	1370.5	187	54.4	4148.5	1148	348	8818	1904	704
WMMSE	169.2	68.8	53.3	138.5	115.2	76.7	154.3	126.9	103.2
JSBR	169.2	24.3	6.9	115.2	34.3	9.3	114.3	28.4	9.7

TABLE II
AVERAGE NUMBER OF ITERATIONS (termination accuracy $= 1e-3$)

	# of users $= 10$			# of users $= 50$			# of users $= 100$		
	d=1	d=2	d=3	d=1	d=2	d=3	d=1	d=2	d=3
MDP	429.4	74.3	32.8	1739.5	465.5	202	3733	882	442.6
WMMSE	51.6	19.2	14.7	59.6	24.9	16.3	69.8	26.0	19.2
JSBR	48.6	9.4	4.0	46.9	12.6	5.1	49.7	12	5.5

Numerical Example #2. In Tables I and II we compare the MIMO-SJBR, the MIMO-MDP [23], [24], and the MIMO-WMMSE [23], [24], in terms of average number of iterations required to reach convergence, for different number of users, normalized distances $d \triangleq d_{ij}/d_{ii}$ (with $d_{ij} = d_{ji}$ and $d_{ii} = d_{jj}$ for all i and $j \neq i$), and termination accuracy (namely: $1e-3$ and $1e-6$). We considered the following setup. All the transmitters/receivers are equipped with 4 antennas; we simulated uncorrelated fading channel model, where the coefficients are Gaussian distributed with zero mean and variance $1/d_{ij}^3$; and we set $\mathbf{R}_{n_i} = \sigma^2\mathbf{I}$ for all i, and $\mathtt{snr} \triangleq P/\sigma^2 = 3$ dB. We used the step-size rule (19) with $\epsilon = 1e-5$ and $\tau_i = 0$. We computed the best-response (37) using the closed form solution [24].

In our simulations all the algorithms reached the same average sum-rate. Given the results in Tables I and II, the following comments are in order. The proposed SJBR outperforms all the others schemes in terms of iterations, while having similar (or even better) computational complexity. Interestingly, the iteration gap with the other schemes reduces with the distance and the termination accuracy. More specifically: i) SJBR seems to be much faster than all the other schemes (about one order of magnitude) when $d_{ij}/d_{ii} = 3$ [say low interference scenarios], and just a bit faster (or comparable to MIMO-WMMSE) when $d_{ij}/d_{ii} = 1$ [say high interference scenarios]; and ii) SJBR is much faster than all the others, if an high termination accuracy is set (see Table I). Also, the convergence speed of SJBR is not affected too much by the number of users. Finally, in our experiments, we also observed that the performance of SJBR are not affected too much by the choice of the parameter ϵ in the (19): a change of ϵ of many orders of magnitude leads to a difference in the average number of iterations which is within 5%; we refer the reader to [43] for details, where one can also find a comparison of several other step-size rules. We must stress however that MIMO-MDP and MIMO-WMMSE do not need any tuning, which is an advantage with respect to our method.

$$\begin{aligned} \min_{\mathbf{W},\mathbf{U},\mathbf{V}} \quad & f(\mathbf{W}, \mathbf{U}, \mathbf{V}) \triangleq \sum_{i \in \mathcal{I}} w_i \left(\mathrm{tr}\left(\mathbf{W}_i\, \mathbf{E}_i(\mathbf{U}, \mathbf{V})\right) - \log\det(\mathbf{W}_i)\right) \\ \text{s.t.} \quad & \mathrm{tr}(\mathbf{V}_i\mathbf{V}_i^H) \leq P_i, \ \ \mathbf{W}_i \succeq \mathbf{0}, \quad \forall i \in \mathcal{I}, \end{aligned} \tag{38}$$

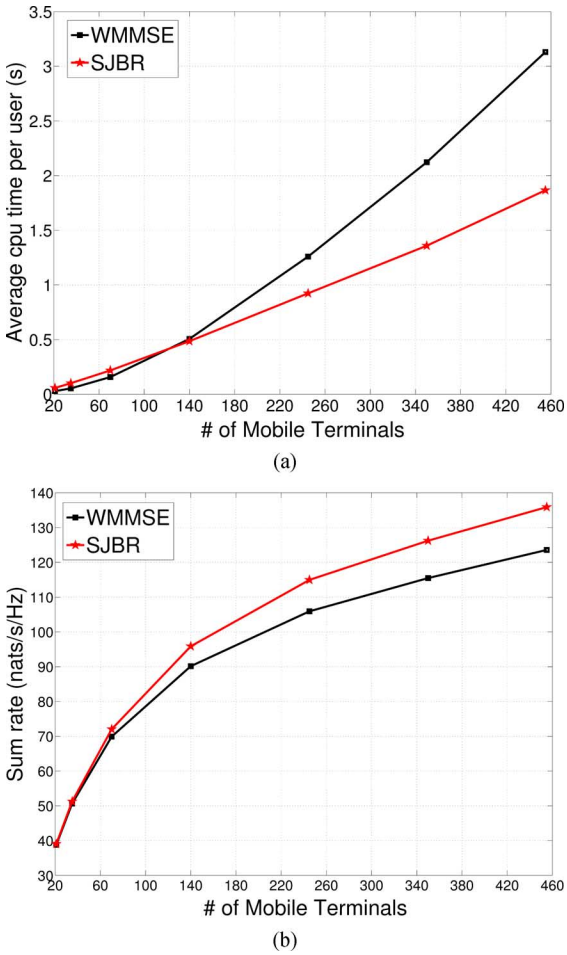

Fig. 3. WMMSE and SJBR over Interference Broadcast Channels. (a) Average cpu time versus # of mobile terminals. (b) Average sum-rate versus # of mobile terminals.

Numerical Example #3. Finally we compare WMMSE [30] and our algorithm (still termed SJBR) in a more realistic scenario, namely a MIMO broadcast cellular system composed of multiple cells, with one Base Station (BS) and multiple randomly generated Mobile Terminals (MTs) in each cell. No orthogonal transmissions are imposed a priori among the BSs (and the MTs inside each cell); therefore each MT experiences both intra-cell and inter-cell interference. We refer to [30] for a detailed description of the system model, the explicit expressions of the BS-MT downlink rates, and the corresponding sum-rate maximization problem.

The setup of our experiments is the following. We simulated seven cells with multiple randomly generated MTs; each BS and MT is equipped with four transmit and receive antennas. Channels are Rayleigh fading, whose path-loss are generated using the 3 GPP(TR 36.814) methodology [44]. We assume white zero-mean Gaussian noise at each mobile receiver, with variance σ^2, and same power budget P for all the BSs; the SNR is set to $\mathtt{snr} \triangleq P/\sigma^2 = 3$ dB. Both algorithms WMMSE and SJBR are initialized by choosing the same feasible randomly generated point, and are terminated when (the absolute value) of the sum-rate error in two consecutive rounds becomes smaller than $1e-2$. In our algorithm, we used the step-size rule (19), with $\gamma^0 = 1$ and $\epsilon = 1e-3$, and we set all $\tau_i = 0$; the unique solution of each (strongly convex) subproblem [cf. (33)] is computed in closed form (up to the multiplier associated with the power constraints that can be efficiently computed using bisection) adapting the procedure in [24], we omit further details because of space limitation. The accuracy in the bisection loops (required by both algorithms) is set to $1e-5$. Our experiments were run using Matlab R2012a on a 12×2.40 GHz Intel Xeon E5645 Processor Cores machine, equipped with 48 GB of memory and 24576 Kbytes of data cache; the operation system is Linux (RedHat Enterprise Linux 6.1 2.6.32 Kernel). In Fig. 3(a) we plot the average cpu time versus the total number of MTs for the two algorithms under the same termination criterion, whereas in Fig. 3(b) we reported the final achieved average sum-rate. The curves are averaged over 1500 channel/topology realizations. It can be observed that SJBR significantly outperforms WMMSE in terms of cpu time when the number of active users is large; moreover SJBR also yields better sum-rates. We observed similar results also under different settings (e.g., SNR, number of cells/BSs, etc.); we refer to the technical report [43] for more experiments.

VIII. Conclusion

In this paper, we proposed a novel decomposition framework, based on SCA, to compute stationary solutions of general nonconvex sum-utility problems (including social functions of complex variables). The main result is a new class of convergent *distributed Jacobi* (inexact) best-response algorithms, where all users *simultaneously* solve (inexactly) a suitably convexified version of the original social problem. Our framework contains as special cases many decomposition methods already proposed in the literature, such as gradient algorithms, and many block-coordinate descent schemes for convex functions. Finally, we tested our methodology on some sum-rate maximization problems over SISO/MIMO ICs; our experiments show that our algorithms are faster than ad-hoc state-of-the-art methods while having the same (user) computational complexity in the SISO case and similar (or better) complexity in the MIMO case. Some interesting future directions of this work are under investigation, e.g., how to adaptively choose the step-size rule (so that no a-priori tuning is needed), and how to generalize our framework to scenarios when only long-term channel statistics are available.

Appendix

For notational simplicity, in the following we will omit in each $\widehat{\mathbf{x}}_{\mathcal{C}_i}(\mathbf{y},\tau_i)$ [and $\widehat{\mathbf{x}}_{\mathcal{C}}(\mathbf{y},\boldsymbol{\tau})$] the dependence on $\mathcal{C}_i$ and τ_i, and write $\widehat{\mathbf{x}}_i(\mathbf{y})$ [and $\widehat{\mathbf{x}}(\mathbf{y})$]; also, we introduce $f_{\mathcal{C}_i}(\mathbf{x}_i,\mathbf{x}_{-i}) \triangleq \sum_{j\in\mathcal{C}_i} f_j(\mathbf{x}_i,\mathbf{x}_{-i})$ and $f_{\mathcal{C}_{-i}}(\mathbf{x}_i,\mathbf{x}_{-i}) \triangleq \sum_{j\in\mathcal{C}_{-i}} f_j(\mathbf{x}_i,\mathbf{x}_{-i})$.

A. Proof of Proposition 1

Before proving the proposition, let us introduce the following intermediate result whose proof is a consequence of assumptions A1–A3 and thus is omitted.

Lemma 6: Let $\tilde{f}(\mathbf{x};\mathbf{y}) \triangleq \sum_i \tilde{f}_{\mathcal{C}_i}(\mathbf{x}_i;\mathbf{y})$, with $\tilde{f}_{\mathcal{C}_i}(\mathbf{x}_i;\mathbf{y})$ defined in (7). Then the following hold:

(i) $\tilde{f}(\bullet;\mathbf{y})$ is uniformly strongly convex on $\mathcal{K}$ with constant $c_{\boldsymbol{\tau}} > 0$, i.e.,

$$(\mathbf{x}-\mathbf{w})^T\left(\nabla_{\mathbf{x}}\tilde{f}(\mathbf{x};\mathbf{y}) - \nabla_{\mathbf{x}}\tilde{f}(\mathbf{w};\mathbf{y})\right) \geq c_{\boldsymbol{\tau}}\|\mathbf{x}-\mathbf{w}\|^2, \tag{39}$$

for all $\mathbf{x},\mathbf{w}\in\mathcal{K}$ and given $\mathbf{y}\in\mathcal{K}$;

(ii) $\nabla_{\mathbf{x}}\tilde{f}(\mathbf{x};\bullet)$ is uniformly Lipschitz continuous on $\mathcal{K}$, i.e., there exists a $0 < L_{\nabla\tilde{f}} < \infty$ independent on $\mathbf{x}$ such that

$$\left\|\nabla_{\mathbf{x}}\tilde{f}\left(\mathbf{x};\mathbf{y}\right) - \nabla_{\mathbf{x}}\tilde{f}\left(\mathbf{x};\mathbf{w}\right)\right\| \leq L_{\nabla\tilde{f}}\left\|\mathbf{y}-\mathbf{w}\right\|, \qquad (40)$$

for all $\mathbf{y},\mathbf{w} \in \mathcal{K}$ and given $\mathbf{x} \in \mathcal{K}$.

We prove now the statements of Proposition 1 in the following order (c)-(a)-(b)-(d).

(a): Given $\mathbf{y} \in \mathcal{K}$, by definition, each $\widehat{\mathbf{x}}_i(\mathbf{y})$ is the unique solution of the problem (10) and thus satisfies the minimum principle: for all $\mathbf{z}_i \in \mathcal{K}_i$,

$$\begin{aligned}\left(\mathbf{z}_i - \widehat{\mathbf{x}}_i(\mathbf{y})\right)^T &\left(\nabla_{\mathbf{x}_i} f_{\mathcal{C}_i}\left(\widehat{\mathbf{x}}_i(\mathbf{y}),\, \mathbf{y}_{-i}\right)\right.\\ &\left.+\boldsymbol{\pi}_{\mathcal{C}_i}(\mathbf{y}) + \tau_i\,\mathbf{H}_i(\mathbf{y})\left(\widehat{\mathbf{x}}_i(\mathbf{y}) - \mathbf{y}_i\right)\right) \geq 0. \qquad (41)\end{aligned}$$

Summing and subtracting $\nabla_{\mathbf{x}_i} f_{\mathcal{C}_i}\left(\mathbf{y}_i,\, \mathbf{y}_{-i}\right)$ in (41), choosing $\mathbf{z}_i = \mathbf{y}_i$, and using $\boldsymbol{\pi}_{\mathcal{C}_i}(\mathbf{y}) \triangleq \nabla_{\mathbf{x}_i} f_{\mathcal{C}_{-i}}(\mathbf{y})$, we get

$$\begin{aligned}\left(\mathbf{y}_i - \widehat{\mathbf{x}}_i(\mathbf{y})\right)^T &\left(\nabla_{\mathbf{x}_i} f_{\mathcal{C}_i}\left(\widehat{\mathbf{x}}_i(\mathbf{y}),\, \mathbf{y}_{-i}\right) - \nabla_{\mathbf{x}_i} f_{\mathcal{C}_i}\left(\mathbf{y}_i,\, \mathbf{y}_{-i}\right)\right)\\ &+ \left(\mathbf{y}_i - \widehat{\mathbf{x}}_i(\mathbf{y})\right)^T \nabla_{\mathbf{x}_i} U(\mathbf{y})\\ &- \tau_i \left(\widehat{\mathbf{x}}_i(\mathbf{y}) - \mathbf{y}_i\right)^T \mathbf{H}_i(\mathbf{y})\left(\widehat{\mathbf{x}}_i(\mathbf{y}) - \mathbf{y}_i\right) \geq 0, \qquad (42)\end{aligned}$$

for all $i \in \mathcal{I}$. Recalling the definition of $c_{\boldsymbol{\tau}}$ [cf. (14)] and using (42), we obtain

$$\left(\mathbf{y}_i - \widehat{\mathbf{x}}_i(\mathbf{y})\right)^T \nabla_{\mathbf{x}_i} U(\mathbf{y}) \geq c_{\boldsymbol{\tau}}\left\|\widehat{\mathbf{x}}_i(\mathbf{y}) - \mathbf{y}_i\right\|^2, \qquad (43)$$

for all $i \in \mathcal{I}$. Summing (43) over i we obtain (13).

(b): Let us use the notation as in Lemma 6. Given $\mathbf{y},\mathbf{z} \in \mathcal{K}$, by the minimum principle, we have

$$\begin{aligned}\left(\mathbf{v} - \widehat{\mathbf{x}}(\mathbf{y})\right)^T \nabla_{\mathbf{x}}\tilde{f}\left(\widehat{\mathbf{x}}(\mathbf{y});\mathbf{y}\right) \geq 0 \qquad &\forall \mathbf{v} \in \mathcal{K}\\ \left(\mathbf{w} - \widehat{\mathbf{x}}(\mathbf{z})\right)^T \nabla_{\mathbf{x}}\tilde{f}\left(\widehat{\mathbf{x}}(\mathbf{z});\mathbf{z}\right) \geq 0 \qquad &\forall \mathbf{w} \in \mathcal{K}. \qquad (44)\end{aligned}$$

Setting $\mathbf{v} = \widehat{\mathbf{x}}(\mathbf{z})$ and $\mathbf{w} = \widehat{\mathbf{x}}(\mathbf{y})$, summing the two inequalities above, and adding and subtracting $\nabla_{\mathbf{x}}\tilde{f}\left(\widehat{\mathbf{x}}(\mathbf{y});\mathbf{z}\right)$, we obtain:

$$\begin{aligned}&\left(\widehat{\mathbf{x}}(\mathbf{z}) - \widehat{\mathbf{x}}(\mathbf{y})\right)^T \left(\nabla_{\mathbf{x}}\tilde{f}\left(\widehat{\mathbf{x}}(\mathbf{z});\mathbf{z}\right) - \nabla_{\mathbf{x}}\tilde{f}\left(\widehat{\mathbf{x}}(\mathbf{y});\mathbf{z}\right)\right)\\ &\leq \left(\widehat{\mathbf{x}}(\mathbf{y}) - \widehat{\mathbf{x}}(\mathbf{z})\right)^T \left(\nabla_{\mathbf{x}}\tilde{f}\left(\widehat{\mathbf{x}}(\mathbf{y});\mathbf{z}\right) - \nabla_{\mathbf{x}}\tilde{f}\left(\widehat{\mathbf{x}}(\mathbf{y});\mathbf{y}\right)\right). \qquad (45)\end{aligned}$$

Using (39) we can now lower bound the left-hand-side of (45) as

$$\begin{aligned}&\left(\widehat{\mathbf{x}}(\mathbf{z}) - \widehat{\mathbf{x}}(\mathbf{y})\right)^T \left(\nabla_{\mathbf{x}}\tilde{f}\left(\widehat{\mathbf{x}}(\mathbf{z});\mathbf{z}\right) - \nabla_{\mathbf{x}}\tilde{f}\left(\widehat{\mathbf{x}}(\mathbf{y});\mathbf{z}\right)\right)\\ &\qquad \geq c_{\boldsymbol{\tau}}\left\|\widehat{\mathbf{x}}(\mathbf{z}) - \widehat{\mathbf{x}}(\mathbf{y})\right\|^2, \qquad (46)\end{aligned}$$

whereas the right-hand side of (45) can be upper bounded as

$$\begin{aligned}&\left(\widehat{\mathbf{x}}(\mathbf{y}) - \widehat{\mathbf{x}}(\mathbf{z})\right)^T \left(\nabla_{\mathbf{x}}\tilde{f}\left(\widehat{\mathbf{x}}(\mathbf{y});\mathbf{z}\right) - \nabla_{\mathbf{x}}\tilde{f}\left(\widehat{\mathbf{x}}(\mathbf{y});\mathbf{y}\right)\right)\\ &\qquad \leq L_{\nabla\tilde{f}}\left\|\widehat{\mathbf{x}}(\mathbf{y}) - \widehat{\mathbf{x}}(\mathbf{z})\right\|\left\|\mathbf{y}-\mathbf{z}\right\|, \qquad (47)\end{aligned}$$

where the inequality follows from the Cauchy-Schwartz inequality and (40). Combining (45), (46), and (47), we obtain the desired Lipschitz property of $\widehat{\mathbf{x}}(\bullet)$.

(c): Let $\mathbf{x}^\star \in \mathcal{K}$ be a fixed point of $\widehat{\mathbf{x}}(\mathbf{y})$, that is $\mathbf{x}^\star = \widehat{\mathbf{x}}(\mathbf{x}^\star)$. By definition, each $\widehat{\mathbf{x}}_i(\mathbf{y})$ satisfies (41), for any given $\mathbf{y} \in \mathcal{K}$. Setting $\mathbf{y} = \mathbf{x}^\star$ and using $\mathbf{x}^\star = \widehat{\mathbf{x}}(\mathbf{x}^\star)$, (41) reduces to

$$\left(\mathbf{z}_i - \mathbf{x}_i^\star\right)^T \nabla_{\mathbf{x}_i} U(\mathbf{x}^\star) \geq 0, \qquad (48)$$

for all $\mathbf{z}_i \in \mathcal{K}_i$ and $i \in \mathcal{I}$. Taking into account the Cartesian structure of $\mathcal{K}$ and summing (48) over $i \in \mathcal{I}$ we obtain $\left(\mathbf{z}-\mathbf{x}^\star\right)^T \nabla_{\mathbf{x}} U(\mathbf{x}^\star) \geq 0$, for all $\mathbf{z} \in \mathcal{K}$, with $\mathbf{z} \triangleq (\mathbf{z}_i)_{i=1}^I$; therefore $\mathbf{x}^\star$ is a stationary solution of (1). The converse holds because i) $\widehat{\mathbf{x}}(\mathbf{x}^\star)$ is the unique optimal solution of (10) with $\mathbf{y} = \mathbf{x}^\star$, and ii) $\mathbf{x}^\star$ is also an optimal solution of (10), since it satisfies the minimum principle.

(d): It follows readily from (43). □

B. Proof of Theorems 3 and 4

We prove Theorem 4; Theorem 3(b) is a special case; the proof of simpler Theorem 3(a) is omitted and can be obtained following similar steps. The line of the proof is based on standard descent arguments, but suitably combined with the properties of $\widehat{\mathbf{x}}(\mathbf{y})$ (cf. Prop. 1), and the presence of errors $\{\epsilon_i^n\}$. We will also use the following lemma, which is the deterministic version of the Robbins-Siegmund result for random sequences [45, Lemma 11] (but without requiring the nonnegativity of X^n and Z^n as instead in [45, Lemma 11]).

Lemma 7: Let $\{X^n\}$, $\{Y^n\}$, and $\{Z^n\}$ be three sequences of numbers such that $Y^n \geq 0$ for all n. Suppose that

$$X^{n+1} \leq X^n - Y^n + Z^n, \quad \forall n = 0,1,\ldots$$

and $\sum_n Z^n < \infty$. Then either $X^n \to -\infty$ or else $\{X^n\}$ converges to a finite value and $\sum_n Y^n < \infty$. □

We are now ready to prove Theorem 4. For any given $n \geq 0$, the Descent Lemma [36] yields

$$\begin{aligned}U\left(\mathbf{x}^{n+1}\right) \leq U\left(\mathbf{x}^n\right) &+ \gamma^n\,\nabla_{\mathbf{x}} U\left(\mathbf{x}^n\right)^T\left(\mathbf{z}^n - \mathbf{x}^n\right)\\ &+ \frac{\left(\gamma^n\right)^2 L_{\nabla U}}{2}\left\|\mathbf{z}^n - \mathbf{x}^n\right\|^2, \qquad (49)\end{aligned}$$

with $\mathbf{z}^n \triangleq (\mathbf{z}_i^n)_{i=1}^I$, and $\mathbf{z}_i^n$ defined in Step 2 (Algorithm 2). Using

$$\begin{aligned}\left\|\mathbf{z}^n - \mathbf{x}^n\right\|^2 &\leq 2\left\|\widehat{\mathbf{x}}(\mathbf{x}^n) - \mathbf{x}^n\right\|^2 + 2\sum_i \left\|\mathbf{z}_i^n - \widehat{\mathbf{x}}_i(\mathbf{x}^n)\right\|^2\\ &\leq 2\left\|\widehat{\mathbf{x}}(\mathbf{x}^n) - \mathbf{x}^n\right\|^2 + 2\sum_i (\varepsilon_i^n)^2,\end{aligned}$$

where in the last inequality we used $\left\|\mathbf{z}_i^n - \widehat{\mathbf{x}}_i(\mathbf{x}^n)\right\| \leq \varepsilon_i^n$, and

$$\begin{aligned}&\nabla_{\mathbf{x}} U\left(\mathbf{x}^n\right)^T\left(\mathbf{z}^n - \widehat{\mathbf{x}}(\mathbf{x}^n) + \widehat{\mathbf{x}}(\mathbf{x}^n) - \mathbf{x}^n\right)\\ &\qquad \leq -c_{\boldsymbol{\tau}}\left\|\widehat{\mathbf{x}}(\mathbf{x}^n) - \mathbf{x}^n\right\|^2 + \sum_i \varepsilon_i^n\left\|\nabla_{\mathbf{x}_i} U(\mathbf{x}^n)\right\|, \qquad (50)\end{aligned}$$

which follows from Prop. 1(c), (49) yields: for all $n \geq 0$,

$$U\left(\mathbf{x}^{n+1}\right) \leq U\left(\mathbf{x}^n\right) - \gamma^n\left(c_{\boldsymbol{\tau}} - \gamma^n L_{\nabla U}\right)\left\|\widehat{\mathbf{x}}(\mathbf{x}^n) - \mathbf{x}^n\right\|^2 + T_n, \qquad (51)$$

where $T_n \triangleq \gamma^n \sum_i \varepsilon_i^n \left\|\nabla_{\mathbf{x}_i} U(\mathbf{x}^n)\right\| + \left(\gamma^n\right)^2 L_{\nabla U} \sum_i (\varepsilon_i^n)^2$. Note that, under the assumptions of the theorem, $\sum_{n=0}^{\infty} T_n < \infty$. Since $\gamma^n \to 0$, we have for some positive constant β_1 and sufficiently large n, say $n \geq \bar{n}$,

$$U\left(\mathbf{x}^{n+1}\right) \leq U\left(\mathbf{x}^n\right) - \gamma^n\beta_1\left\|\widehat{\mathbf{x}}(\mathbf{x}^n) - \mathbf{x}^n\right\|^2 + T_n. \qquad (52)$$

Invoking Lemma 7 with the identifications $X^n = U\left(\mathbf{x}^{n+1}\right)$, $Y^n = \gamma^n\beta_1\left\|\widehat{\mathbf{x}}(\mathbf{x}^n) - \mathbf{x}^n\right\|^2$ and $Z^n = T_n$ while

using $\sum_n T_n < \infty$, we deduce from (52) that either $\{U(\mathbf{x}^n)\} \to -\infty$ or else $\{U(\mathbf{x}^n)\}$ converges to a finite value and

$$\lim_{n\to\infty} \sum_{t=\bar{n}}^{n} \gamma^t \left\|\widehat{\mathbf{x}}(\mathbf{x}^t) - \mathbf{x}^t\right\|^2 < +\infty. \tag{53}$$

Since $U(\mathbf{x})$ is coercive, $U(\mathbf{x}) \geq \min_{\mathbf{y}\in\mathcal{K}} U(\mathbf{y}) > -\infty$, implying that $\{U(\mathbf{x}^n)\}_n$ is convergent; it follows from (53) and $\sum_{n=0}^{\infty} \gamma^n = \infty$ that $\liminf_{n\to\infty} \|\widehat{\mathbf{x}}(\mathbf{x}^n) - \mathbf{x}^n\| = 0$.

Using Prop. 1, we show next that $\lim_{n\to\infty} \|\widehat{\mathbf{x}}(\mathbf{x}^n) - \mathbf{x}^n\| = 0$; for notational simplicity we will write $\triangle\widehat{\mathbf{x}}(\mathbf{x}^n) \triangleq \widehat{\mathbf{x}}(\mathbf{x}^n) - \mathbf{x}^n$. Suppose, by contradiction, that $\limsup_{n\to\infty} \|\triangle\widehat{\mathbf{x}}(\mathbf{x}^n)\| > 0$. Then, there exists a $\delta > 0$ such that $\|\triangle\widehat{\mathbf{x}}(\mathbf{x}^n)\| > 2\delta$ for infinitely many n and also $\|\triangle\widehat{\mathbf{x}}(\mathbf{x}^n)\| < \delta$ for infinitely many n. Therefore, one can always find an infinite set of indexes, say $\mathcal{N}$, having the following properties: for any $n \in \mathcal{N}$, there exists an integer $i_n > n$ such that

$$\|\triangle\widehat{\mathbf{x}}(\mathbf{x}^n)\| < \delta, \qquad \left\|\triangle\widehat{\mathbf{x}}(\mathbf{x}^{i_n})\right\| > 2\delta \tag{54}$$

$$\delta \leq \left\|\triangle\widehat{\mathbf{x}}(\mathbf{x}^j)\right\| \leq 2\delta \qquad n < j < i_n. \tag{55}$$

Given the above bounds, the following holds: for all $n \in \mathcal{N}$,

$$\begin{aligned}
\delta &\overset{(a)}{<} \left\|\triangle\widehat{\mathbf{x}}(\mathbf{x}^{i_n})\right\| - \|\triangle\widehat{\mathbf{x}}(\mathbf{x}^n)\| \\
&\leq \left\|\widehat{\mathbf{x}}(\mathbf{x}^{i_n}) - \widehat{\mathbf{x}}(\mathbf{x}^n)\right\| + \left\|\mathbf{x}^{i_n} - \mathbf{x}^n\right\| \qquad (56)\\
&\overset{(b)}{\leq} (1+\hat{L}) \left\|\mathbf{x}^{i_n} - \mathbf{x}^n\right\| \qquad (57)\\
&\overset{(c)}{\leq} (1+\hat{L}) \sum_{t=n}^{i_n-1} \gamma^t \left(\left\|\triangle\widehat{\mathbf{x}}(\mathbf{x}^t)\right\| + \left\|\mathbf{z}^t - \widehat{\mathbf{x}}(\mathbf{x}^t)\right\|\right)\\
&\overset{(d)}{\leq} (1+\hat{L})\,(2\delta + \varepsilon^{\max}) \sum_{t=n}^{i_n-1} \gamma^t, \qquad (58)
\end{aligned}$$

where (a) follows from (54) and (55); (b) is due to Prop. 1(a); (c) comes from the triangle inequality and the updating rule of the algorithm; and in (d) we used (54), (55), and $\|\mathbf{z}^t - \widehat{\mathbf{x}}(\mathbf{x}^t)\| \leq \sum_i \varepsilon_i^t$, where $\varepsilon^{\max} \triangleq \max_n \sum_i \varepsilon_i^n < \infty$. It follows from (58) that

$$\liminf_{n\to\infty} \sum_{t=n}^{i_n-1} \gamma^t \geq \frac{\delta}{(1+\hat{L})(2\delta + \varepsilon^{\max})} > 0. \tag{59}$$

We show next that (59) is in contradiction with the convergence of $\{U(\mathbf{x}^n)\}_n$. To do that, we preliminary prove that, for sufficiently large $n \in \mathcal{N}$, it must be $\|\triangle\widehat{\mathbf{x}}(\mathbf{x}^n)\| \geq \delta/2$. Proceeding as in (58), we have: for any given $n \in \mathcal{N}$,

$$\begin{aligned}
\left\|\triangle\widehat{\mathbf{x}}(\mathbf{x}^{n+1})\right\| - \|\triangle\widehat{\mathbf{x}}(\mathbf{x}^n)\| &\leq (1+\hat{L}) \left\|\mathbf{x}^{n+1} - \mathbf{x}^n\right\| \\
&\leq (1+\hat{L})\gamma^n \left(\|\triangle\widehat{\mathbf{x}}(\mathbf{x}^n)\| + \varepsilon^{\max}\right).
\end{aligned}$$

It turns out that for sufficiently large $n \in \mathcal{N}$ so that $(1+\hat{L})\gamma^n < \delta/(\delta + 2\varepsilon^{\max})$, it must be

$$\|\triangle\widehat{\mathbf{x}}(\mathbf{x}^n)\| \geq \delta/2; \tag{60}$$

otherwise the condition $\left\|\triangle\widehat{\mathbf{x}}(\mathbf{x}^{n+1})\right\| \geq \delta$ would be violated [cf. (55)]. Hereafter we assume w.l.o.g. that (60) holds for all $n \in \mathcal{N}$ (in fact, one can alway restrict $\{\mathbf{x}^n\}_{n\in\mathcal{N}}$ to a proper subsequence).

We can show now that (59) is in contradiction with the convergence of $\{U(\mathbf{x}^n)\}_n$. Using (52) (possibly over a subsequence), we have: for sufficiently large $n \in \mathcal{N}$,

$$\begin{aligned}
U\left(\mathbf{x}^{i_n}\right) &\leq U\left(\mathbf{x}^n\right) - \beta_2 \sum_{t=n}^{i_n-1} \gamma^t \left\|\triangle\widehat{\mathbf{x}}(\mathbf{x}^t)\right\|^2 + \sum_{t=n}^{i_n-1} T_t\\
&\overset{(a)}{<} U\left(\mathbf{x}^n\right) - \beta_2(\delta^2/4) \sum_{t=n}^{i_n-1} \gamma^t + \sum_{t=n}^{i_n-1} T_t \qquad (61)
\end{aligned}$$

where in (a) we used (55) and (60), and β_2 is some positive constant. Since $\{U(\mathbf{x}^n)\}_n$ converges and $\sum_{n=0}^{\infty} T_n < \infty$, (61) implies $\lim_{\mathcal{N}\ni n\to\infty} \sum_{t=n}^{i_n-1} \gamma^t = 0$, which contradicts (59).

Finally, since the sequence $\{\mathbf{x}^n\}$ is bounded [due to the coercivity of $U(\mathbf{x})$ and the convergence of $\{U(\mathbf{x}^n)\}_n$], it has at least one limit point $\bar{\mathbf{x}}$ that must belong to $\mathcal{K}$. By the continuity of $\widehat{\mathbf{x}}(\bullet)$ [Prop. 1(a)] and $\lim_{n\to\infty} \|\widehat{\mathbf{x}}(\mathbf{x}^n) - \mathbf{x}^n\| = 0$, it must be $\widehat{\mathbf{x}}(\bar{\mathbf{x}}) = \bar{\mathbf{x}}$. By Prop. 1(b) $\bar{\mathbf{x}}$ is also a stationary solution of the social problem (1).

Note that, in the setting of Theorem 3, $\varepsilon_i^n = 0$ for all i and n; therefore $T_n = 0$ for all n. It follows from (52) that $U(\mathbf{x}^n)$ is a decreasing sequence, which entails that no limit point of $\{\mathbf{x}^n\}$ can be a local maximum. □

C. Proof of Theorem 5

The main idea of the proof is to interpret Algorithm 3 as an instance of the inexact Jacobi scheme described in Algorithm 2, and show that Theorem 4 is satisfied. It is not difficult to show that this reduces to prove that, for all $i = 1, \ldots, I$, the sequence $\mathbf{z}_i^t$ in Step 2a) of Algorithm 3 satisfies

$$\|\mathbf{z}_i^t - \widehat{\mathbf{x}}_i(\mathbf{x}^t)\| \leq \tilde{\varepsilon}_i^t, \tag{62}$$

for some $\{\tilde{\varepsilon}_i^t\}$ such that $\sum_t \tilde{\varepsilon}_i^t \gamma^t < \infty$. The following holds for the LHS of (62):

$$\begin{aligned}
&\|\mathbf{z}_i^t - \widehat{\mathbf{x}}_i(\mathbf{x}^t)\|\\
&\leq \|\widehat{\mathbf{x}}_i(\mathbf{x}_{i<}^{t+1}, \mathbf{x}_{i\geq}^t) - \widehat{\mathbf{x}}_i(\mathbf{x}^t)\| + \|\mathbf{z}_i^t - \widehat{\mathbf{x}}_i(\mathbf{x}_{i<}^{t+1}, \mathbf{x}_{i\geq}^t)\|\\
&\overset{(a)}{\leq} \|\widehat{\mathbf{x}}_i(\mathbf{x}_{i<}^{t+1}, \mathbf{x}_{i\geq}^t) - \widehat{\mathbf{x}}_i(\mathbf{x}^t)\| + \varepsilon_i^t\\
&\overset{(b)}{\leq} \hat{L} \|\mathbf{x}_{i<}^{t+1} - \mathbf{x}_{<i}^t\| + \varepsilon_i^t\\
&\overset{(c)}{\leq} \hat{L}\gamma^t \left(\left\| \left(\widehat{\mathbf{x}}_j(\mathbf{x}_{j<}^{t+1}, \mathbf{x}_{j\geq}^t) - \mathbf{x}_j^t\right)_{j=1}^{i-1} \right\| + \sum_{j<i} \varepsilon_j^t \right) + \varepsilon_i^t\\
&\overset{(d)}{\leq} \hat{L}\gamma^t \beta_i + \hat{L}\gamma^t \sum_{j<i} \varepsilon_j^t + \varepsilon_i^t,
\end{aligned}$$

where (a) follows from the error bound in Step 2a) of Algorithm 3; in (b) we used Prop. 1a); (c) follows from Step 2b); and in (d) we used Prop. 1d), with $\beta_i < \infty$ being a positive constant. It turns out that (62) is satisfied choosing $\tilde{\varepsilon}_i^t \triangleq \hat{L}\gamma^t\beta_i + \hat{L}\gamma^t \sum_{j<i} \varepsilon_j^t + \varepsilon_i^t$. □

ACKNOWLEDGMENT

The authors wish to thank the Associate Editor, Prof. A. So, and the anonymous reviewers for their valuable comments. The authors are also deeply grateful to Prof. T. Luo, W.-C. Liao, and Y. Yang whose comments contributed to improve the quality of the paper.

REFERENCES

[1] G. Scutari, D. P. Palomar, F. Facchinei, and J.-S. Pang, "Distributed dynamic pricing for MIMO interfering multiuser systems: A unified approach," in *Proc. Int. Conf. NETwork Games, COntrol, OPtimiz. (NetGCooP 2011)*, Paris, France, Oct. 2011, pp. 12–14.

[2] 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.

[3] M. Chiang, S. H. Low, A. R. Calderbank, and J. C. Doyle, "Layering as optimization decomposition: A mathematical theory of network architectures," *Proc. IEEE*, vol. 95, no. 1, pp. 255–312, Jan. 2007.

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

[5] Z.-Q. Luo and S. Zhang, "Spectrum management: Complexity and duality," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 1, pp. 57–72, Feb. 2008.

[6] M. Hong and Z.-Q. Luo, "Signal processing and optimal resource allocation for the interference channel," in *Library in Signal Processing*. New York: Academic Press, 2013, vol. 2, Communications and Radar Signal Processing, ch. 8, pp. 409–462.

[7] W. Yu, G. Ginis, and J. M. Cioffi, "Distributed multiuser power control for digital subscriber lines," *IEEE J. Sel. Areas Commun.*, vol. 20, no. 5, pp. 1105–1115, Jun. 2002.

[8] Z.-Q. Luo and J.-S. Pang, "Analysis of iterative waterfilling algorithm for multiuser power control in digital subscriber lines," *EURASIP J. Appl. Signal Process.*, vol. 2006, pp. 1–10, May 2006.

[9] G. Scutari, D. P. Palomar, and S. Barbarossa, "Optimal linear precoding strategies for wideband noncooperative systems based on game theory—Part I&II: Nash equilibria and distributed algorithms," *IEEE Trans. Signal Process.*, vol. 56, no. 3, pp. 1230–1267, Mar. 2008.

[10] G. Scutari, D. P. Palomar, and S. Barbarossa, "Asynchronous iterative water-filling for Gaussian frequency-selective interference channels," *IEEE Trans. Inf. Theory*, vol. 54, no. 7, pp. 2868–2878, Jul. 2008.

[11] R. Cendrillon, J. Huang, M. Chiang, and M. Moonen, "Autonomous spectrum balancing for digital subscriber lines," *IEEE Trans. Signal Process.*, vol. 55, no. 8, pp. 4241–4257, Aug. 2007.

[12] G. Scutari, D. P. Palomar, and S. Barbarossa, "Competitive design of multiuser MIMO systems based on game theory: A unified view," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 7, pp. 1089–1103, Sep. 2008.

[13] G. Scutari, D. P. Palomar, and S. Barbarossa, "The MIMO iterative waterfilling algorithm," *IEEE Trans. Signal Process.*, vol. 57, no. 5, May 2009.

[14] G. Scutari and D. P. Palomar, "MIMO cognitive radio: A game theoretical approach," *IEEE Trans. Signal Process.*, vol. 58, no. 2, pp. 761–780, Feb. 2010.

[15] E. Larsson, E. Jorswieck, J. Lindblom, and R. Mochaourab, "Game theory and the flat-fading Gaussian interference channel," *IEEE Signal Process. Mag.*, vol. 26, no. 5, pp. 18–27, Sep. 2009.

[16] A. Leshem and E. Zehavi, "Game theory and the frequency selective interference channel," *IEEE Signal Process. Mag.*, vol. 26, no. 5, pp. 28–40, Sep. 2009.

[17] F. Facchinei and J.-S. Pang*, Finite-Dimensional Variational Inequalities and Complementarity Problem*. New York, NY, USA: Springer-Verlag, 2003.

[18] J.-S. Pang, G. Scutari, D. P. Palomar, and F. Facchinei, "Design of cognitive radio systems under temperature-interference constraints: A variational inequality approach," *IEEE Trans. Signal Process.*, vol. 58, no. 6, pp. 3251–3271, Jun. 2010.

[19] G. Scutari, D. Palomar, F. Facchinei, and J.-S. Pang, "Flexible design of cognitive radio wireless systems: From game theory to variational inequality theory," *IEEE Signal Process. Mag.*, vol. 26, no. 5, pp. 107–123, Sep. 2009.

[20] G. Scutari, D. Palomar, F. Facchinei, and J.-S. Pang, "Convex optimization, game theory, and variational inequality theory in multiuser communication systems," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 35–49, May 2010.

[21] J. Huang, R. Berry, and M. L. Honig, "Distributed interference compensation for wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 5, pp. 1074–1084, May 2006.

[22] F. Wang, M. Krunz, and S. Cui, "Price-based spectrum management in cognitive radio networks," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 1, pp. 74–87, Feb. 2008.

[23] 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.

[24] 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.

[25] L. Grippo and M. Sciandrone, "On the convergence of the block nonlinear Gauss-Seidel method under convex constraints," *Oper. Res. Lett.*, vol. 26, no. 3, pp. 127–136, Apr. 2000.

[26] S. Ye and R. S. Blum, "Optimized signaling for MIMO interference systems with feedback," *IEEE Trans. Signal Process.*, vol. 51, no. 11, pp. 2839–2848, Nov. 2003.

[27] M. Chiang, C. W. Tan, D. P. 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.

[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] J. Papandriopoulos and J. S. Evans, "Scale: A low-complexity distributed protocol for spectrum balancing in multiuser DSL networks," *IEEE Trans. Inf. Theory*, vol. 55, no. 8, pp. 3711–3724, Aug. 2009.

[30] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4331–4340, Sep. 2011.

[31] C. Shi, D. A. Schmidt, R. A. Berry, M. L. Honig, and W. Utschick, "Distributed interference pricing for the MIMO interference channel," presented at the Int. Conf. Commun. (ICC), Princeton, NJ, USA, Jun. 14–18, 2009.

[32] B. R. Marks and G. P. Wright, "A general inner approximation algorithm for nonconvex mathematical programs," *Oper. Res.*, vol. 26, no. 2, pp. 681–683, 1978.

[33] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," Arxiv.org, 2012 [Online]. Available: http://arxiv.org/abs/1209.2385v1

[34] M. Hong, Q. Li, Y.-F. Liu, and Z.-Q. Luo, "Decomposition by successive convex approximation: a unifying approach for linear transceiver design in interfering heterogenous networks," Arxiv.org, 2012 [Online]. Available: http://arxiv.org/abs/1210.1507v1

[35] R. Cendrillon, W. Yu, M. Moonen, J. Verlinden, and T. Bostoen, "Optimal multiuser spectrum balancing for digital subscriber lines," *IEEE Trans. Signal Process.*, vol. 54, no. 5, pp. 922–933, May 2006.

[36] D. Bertsekas*, Nonlinear Programming*, 2nd ed. Belmont, MA, USA: Athena Scientific, 1999.

[37] C. Shi, R. Berry, and M. Honig, "Distributed interference pricing for ofdm wireless networks with non-separable utilities," presented at the 42nd Annu. Conf. Inf. Sci. Syst., Princeton, NJ, USA, Mar. 2008.

[38] W. Yu, W. Rhee, S. Boyd, and J. Cioffi, "Iterative water-filling for Gaussian vector multiple access channels," *IEEE Trans. Inf. Theory*, vol. 50, no. 1, pp. 145–151, Jan. 2004.

[39] Y. Yang, G. Scutari, and D. Palomar, "Parallel stochastic decomposition algorithms for multi-agent systems," presented at the IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC), Darmstadt, Germany, Jun. 16–19, 2013.

[40] A. Hjorungnes*, Complex-Valued Matrix Derivatives With Applications in Signal Processing and Communications*. London, U.K.: Cambridge Univ. Press, 2011.

[41] G. Scutari, F. Facchieni, J.-S. Pang, and D. P. Palomar, "Real and complex monotone communication games," *IEEE Trans. Inf. Theory*, 2013 [Online]. Available: http://arxiv.org/abs/1212.6235, accepted for publication

[42] D. P. Palomar and J. Fonollosa, "Practical algorithms for a family of waterfilling solutions," *IEEE Trans. Signal Process.*, vol. 53, no. 2, pp. 686–695, Feb. 2005.

[43] P. Song, "Numerical comparison of decomposition algorithms for nonconvex sum-utility problems," Dept. of Elect. Eng., Univ. at Buffalo, the State Univ. of New York, Tech. Rep., 2013 [Online]. Available: http://www.eng.buffalo.edu/~peiranso/SongTechRepNumCompJan13.pdf

[44] *3GPP TR 36.814*, [Online]. Available: http://www.3gpp.org/ftp/Specs/archive/36_serier/36.814

[45] B. T. Polyak*, Introduction to Optimization*. New York, NY, USA: Optimization Software, Inc., Publications Div., 1987.

Gesualdo Scutari (S'05–M'06–SM'11) received the Electrical Engineering and Ph.D. degrees (both with honors) from the University of Rome "La Sapienza", Rome, Italy, in 2001 and 2005, respectively.

He is an Assistant Professor in the Department of Electrical Engineering at State University of New York (SUNY) at Buffalo, Buffalo, NY. He had previously held several research appointments, namely, at University of California at Berkeley, Berkeley, CA; Hong Kong University of Science and Technology, Hong Kong; University of Rome, "La Sapienza", Rome, Italy; University of Illinois at Urbana-Champaign, Urbana, IL. His primary research interests focus on theoretical and algorithmic issues related to continuous optimization, equilibrium programming and their applications to signal processing, communications and networking; big data analytics; and distributed decisions.

Dr. Scutari is an Associate Editor of IEEE TRANSACTIONS ON SIGNAL PROCESSING, and he served as an Associate Editor of IEEE SIGNAL PROCESSING LETTERS. He serves on the IEEE Signal Processing Society Technical Committee on Signal Processing for Communications (SPCOM).

Dr. Scutari received the 2006 Best Student Paper Award at the International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2006, the 2013 NSF Faculty Early Career Development (CAREER) Award, and the 2013 UB Young Investigator Award.

Francisco Facchinei received the Ph.D. degree in system engineering from the University of Rome, "La Sapienza," Rome, Italy.

He is full professor of Operations Research, Engineering Faculty, University of Rome, "La Sapienza." His research interests focus on theoretical and algorithmic issues related to nonlinear optimization, variational inequalities, complementarity problems, equilibrium programming, and computational game theory.

Peiran Song received the B.S. degree in information and computing science and the M.S. degree in computational mathematics form Beijing Jiaotong University, Beijing, China, in 2008 and 2010, respectively. She is currently a Ph.D. student in the Department of Electrical Engineering at the State University of New York at Buffalo, NY.

Her research interests include applications of convex optimization theory, cooperative communications and distributed designs.

Daniel P. Palomar (S'99–M'03–SM'08–F'12) received the Electrical Engineering and Ph.D. degrees from the Technical University of Catalonia (UPC), Barcelona, Spain, in 1998 and 2003, respectively.

He is an Associate Professor in the Department of Electronic and Computer Engineering at the Hong Kong University of Science and Technology (HKUST), Hong Kong, which he joined in 2006. Since 2013 he is a Fellow of the Institute for Advance Study (IAS) at HKUST. He had previously held several research appointments, namely, at King's College London (KCL), London, UK; Technical University of Catalonia (UPC), Barcelona; Stanford University, Stanford, CA; Telecommunications Technological Center of Catalonia (CTTC), Barcelona; Royal Institute of Technology (KTH), Stockholm, Sweden; University of Rome "La Sapienza", Rome, Italy; and Princeton University, Princeton, NJ. His current research interests include applications of convex optimization theory, game theory, and variational inequality theory to financial systems and communication systems.

Dr. Palomar is an IEEE Fellow, a recipient of a 2004/06 Fulbright Research Fellowship, the 2004 Young Author Best Paper Award by the IEEE Signal Processing Society, the 2002/03 best Ph.D. prize in Information Technologies and Communications by the Technical University of Catalonia (UPC), the 2002/03 Rosina Ribalta first prize for the Best Doctoral Thesis in Information Technologies and Communications by the Epson Foundation, and the 2004 prize for the best Doctoral Thesis in Advanced Mobile Communications by the Vodafone Foundation and COIT.

He serves as an Associate Editor of IEEE TRANSACTIONS ON INFORMATION THEORY, and has been an Associate Editor of IEEE TRANSACTIONS ON SIGNAL PROCESSING, a Guest Editor of the *IEEE Signal Processing Magazine* 2010 Special Issue on "Convex Optimization for Signal Processing," the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS 2008 Special Issue on "Game Theory in Communication Systems," and the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS 2007 Special Issue on "Optimization of MIMO Transceivers for Realistic Communication Networks."

Jong-Shi Pang received the Ph.D. degree in operations research from Stanford University, Stanford, CA.

He is the Epstein Family Chair Professor of Industrial and Systems Engineering Department at the University of Souther California. Prior this position, he was the Caterpillar Professor and Head of the Department of Industrial and Enterprise Systems Engineering at the University of Illinois at Urbana-Champaign. He had taught at the Rensselaer Polytechnic Institute, Troy, NY, The Johns Hopkins University, Baltimore, MD, The University of Texas at Dallas, and Carnegie-Mellon University, Pittsburgh, PA. His research interests are in continuous optimization and equilibrium programming and their applications in engineering, economics, and finance. Among the current projects, he is studying various extensions of the basic Nash equilibrium problem, including the Stackelberg game and its multi-leader generalization, and the dynamic version of the Nash problem. The mathematical tool for the latter problem is a new class of dynamical systems known as differential variational inequalities, which provides a powerful framework for dealing with applications that involve dynamics, unilateral constraints, and disjunction.

Dr. Pang has received several awards and honors, most notably the George B. Dantzig Prize in 2003 jointly awarded by the Mathematical Programming Society and the Society for Industrial and Applied Mathematics and the 1994 Lanchester Prize by the Institute for Operations Research and Management Science. He was recently selected as a member in the inaugural 2009 class of the SIAM Fellows. He is an ISI highly cited author in the mathematics category.