Achieving Global Optimality for Energy Efficiency Maximization in Wireless Networks

Alessio Zappone, *Member, IEEE*, Emil Björnson, *Member, IEEE*, Luca Sanguinetti, *Senior Member, IEEE*, and Eduard Jorswieck, *Senior Member, IEEE*

***Abstract*—The characterization of the global maximum of energy efficiency (EE) problems in wireless networks is a challenging problem due to its nonconvex nature in interference channels. The aim of this work is to develop a new and general framework to achieve globally optimal power control solutions. First, the hidden monotonic structure of the most common EE maximization problems is exploited jointly with fractional programming theory to obtain globally optimal solutions with exponential complexity in the number of network links. To overcome this issue, we also propose a framework to compute suboptimal power control strategies characterized by affordable complexity. This is achieved by merging fractional programming and sequential optimization. The proposed monotonic framework is used to shed light on the ultimate performance of wireless networks in terms of EE and also to benchmark the performance of the lower-complexity framework based on sequential programming. Numerical evidence is provided to show that the sequential fractional programming achieves global optimality.**

***Index Terms*—Energy efficiency, fractional programming, monotonic optimization, sequential programming, massive MIMO, relay networks, LTE.**

I. Introduction

The percentage of the global footprint (in terms of CO_2-equivalent emissions) due to the information and communications technology (ICT) was recently estimated to be 5% [1]. Although this is a small percentage, it is rapidly increasing, and the situation will escalate in the near future with the advent of 5G networks. Credited sources foresee the number of connected devices to reach 50 billions by 2020 [2] and that the data traffic will increase by $1000\times$ over the next 10 years [3]. If no countermeasures are taken, the energy demand to operate and provide such massive data rates to this massive number of devices will become unmanageable, and the resulting greenhouse gas emissions and electromagnetic pollution will exceed safety thresholds. While restricting the global ICT usage is unrealistic, a promising answer to this issue lies in optimizing the energy efficiency (EE) of ICT systems; that is, in maximizing the amount of information reliably transmitted per Joule of consumed energy. Moreover, the EE is of paramount importance for operators (e.g., to save on electricity bills) and for end-users (e.g., to prolong the lifetime of batteries). This has created a great interest in the design of new network architectures, as well as new beamforming and power control strategies taking into account the cost of energy so as to push the EE in communication systems towards new limits.

A. State-of-the-art

The EE of a wireless communication link is commonly defined as a benefit-cost ratio, where the data rate is compared with the associated energy consumption:

$$\text{EE [bit/Joule]} = \frac{\text{Data rate [bit/s]}}{\text{Energy Consumption [W]}}.$$

Given the fractional nature of the EE, the main mathematical tool for centralized optimization of EE-related metrics is fractional programming [4]—a branch of optimization theory that provides algorithms with polynomial complexity to globally maximize fractional functions with a concave numerator and a convex denominator [5]. However, even this powerful tool fails when interference-limited networks must be optimized. This is due to the fact that the presence of multi-user interference makes the numerator of the EE a non-concave function of the transmit power. A common way to circumvent this problem is to only consider suboptimal orthogonal or semi-orthogonal transmission schemes as well as interference cancellation techniques, to fall back into the noise-limited case. Contributions in this direction are given in [6]–[8]. In [6], [7] multi-carrier networks are considered and the global energy efficiency (GEE) of the system (defined as the ratio between the achievable sum rate and the total energy consumption) is optimized using orthogonal or semi-orthogonal subcarrier allocation schemes. In [8], the authors consider a multi-antenna system and aim to maximize the GEE when non-linear interference cancellation techniques are used. However, orthogonal interference suppression schemes inevitably result in a poor resource reutilization in multi-link networks and are thus not reasonable in large networks. The practically unavoidable channel estimation errors also break the orthogonality in many cases.

A. Zappone and E. Jorswieck are with TU Dresden, Faculty of Electrical and Computer Engineering, Communications Laboratory, Dresden, Germany ({alessio.zappone, eduard.jorswieck}@tu-dresden.de). E. Björnson is with the Department of Electrical Engineering (ISY), Linköping University, Linköping, Sweden (emil.bjornson@liu.se). L. Sanguinetti is with the University of Pisa, Dipartimento di Ingegneria dell'Informazione, Pisa, Italy (luca.sanguinetti@unipi.it) and also with the Large Systems and Networks Group (LANEAS), CentraleSupélec, Gif-sur-Yvette, France.

The work of A. Zappone was supported by the German Research Foundation (DFG), under grant CEMRIN - ZA 747/1-3.

E. Björnson is funded by ELLIIT and an Ingvar Carlsson Award from the Swedish Foundation for Strategic Research (SFF).

L. Sanguinetti was supported by the ERC Starting Grant 305123 MORE and by the research project 5GIOTTO funded by the University of Pisa.

The work of E. Jorswieck was supported by the German Research Foundation (DFG), within the Collaborative Research Center 912 HAEC.

A preliminary version of this paper will be presented at ICASSP 2016, Shanghai, China, 20-25 March 2016.

An alternative approach is to handle the interference by means of heuristic solutions, typically based on the use of alternating optimization techniques. Examples in this context can be found [9] wherein the minimum of the individual EEs is maximized and in [10], [11] where both the maximization of GEE and of the sum of the individual EEs are considered. While these approaches can be applied to interference-limited networks, they do not guarantee convergence and/or are not supported by strong optimality claims. Moreover, they are typically tailored to the maximization of specific EE metrics.

A first attempt to provide a unified framework to tackle EE optimization problems in a centralized way is based on the integration of traditional fractional programming methods with the tool of sequential optimization [12]–[16]. The basic idea of sequential optimization is to tackle a difficult problem by solving a sequence of easier approximate problems, which can be solved by standard methods. Provided that suitable approximations can be found, sequential optimization is able to obtain a solution which fulfills first-order optimality conditions for the original problem, while at the same time requiring only the solution of convex problems. Sequential optimization was first used for rate and signal-to-interference-plus-noise ratio (SINR) optimization in [12], [13], and it was more recently successfully integrated into fractional programming to tackle EE maximization problems. Contributions in this sense are [14]–[17], which consider both multi-antenna and multi-carrier systems, also addressing two-hop communications and full-duplex systems. However, sequential optimization can not guarantee to achieve global optimality, and indeed the above works do not provide any insight into the gap between sequential optimization algorithms and the global optimal solution. In general, no previous work provides efficient and provably convergent algorithms to obtain the global maximum of the EE in interference-limited networks.

A possible answer to close this gap is represented by monotonic optimization. This optimization framework can solve optimization problems where the utility and constraints are monotonically increasing/decreasing functions of the variables, which is a less restrictive assumption than convexity. The polyblock algorithm [18] and branch-reduce-and-bound (BRB) algorithm [19] are two key algorithms for solving these problems to global optimality. These algorithms have recently been used to solve different non-convex problems in communication and networking systems; for example, joint power control and scheduling [20] and sum rate maximization with multi-antenna transmitters [21]. The weighted sum-rate maximization problem is studied by monotonic optimization and the rate profile technique in [22] for Gaussian interference channels. In [23] monotonic optimization is used to characterize the Pareto boundary of rate optimization problems in wireless multi-cell networks. A good overview of monotonic optimization techniques in communication and networking is provided by [24] and [25]. However, to-date monotonic optimization has never been used for EE optimization, the main difficulty being that the EE is not a monotone function of the transmit powers, which is the case with more traditional performance metrics like rate or SINR.

B. Major contributions

The goal of this work is to shed light on the ultimate EE performance of interference-limited networks, by providing algorithms that provably converge to the global optimum of different network EE metrics, and algorithms that are able to approach the globally optimal solution, with polynomial computational complexity. This is achieved through the following major contributions:

- Fractional programming and monotonic optimization are jointly used to develop a new *monotonic fractional programming* framework that allows to compute globally optimal power control solutions for different EE maximization problems. Due to the use of monotonic optimization, the complexity of the proposed framework turns out to be exponential in the number of links, but still lower than standard global optimization algorithms. Moreover, convergence to the global optimum is theoretically ensured, whereas this is not always true for general global optimization methods.
- Fractional programming is used together with sequential convex optimization to develop a novel *sequential fractional programming* framework able to obtain candidate solutions fulfilling the Karush-Kuhn-Tucker (KKT) optimality conditions of EE maximization problems. The framework has a polynomial-time complexity and requires only to solve convex optimization problems. The effectiveness of such low-complexity solutions is validated by means of numerical results using the globally optimal solutions (computed by means of the monotonic fractional programming) as a benchmark.
- The monotonic fractional programming framework is used to get insights on the ultimate EE performance of wireless networks. This is achieved by characterizing the network energy-efficient Pareto boundary.
- Both frameworks are shown to be general enough to be applied to many relevant instances of contemporary and future wireless communication networks, such as massive MIMO networks, relay-assisted networks, LTE networks.

C. Outline and notation

The remainder of this paper is organized as follows. Section II defines the signal model and formulates the EE maximization problems. Section III introduces some useful mathematical definitions and results on fractional programming, monotonic optimization, and sequential programming. Sections IV and V develop the monotonic and sequential fractional programming frameworks, respectively. Numerical results for two notable case-studies of communication systems are illustrated in Section VI, whereas concluding remarks are made in Section VII.

The following notation is used throughout the paper. Scalars are denoted by lower case letters whereas boldface lower case letters are used for vectors. The superscripts ^{T} and ^{H} denote transpose and conjugate transpose, respectively. $\mathbf{1}_N$ and $\mathbf{0}_N$ are the N-dimensional all-one and all-zero vectors, respectively. $\mathbb{R}$ denotes the real number space and $\mathbb{C}$ is the complex number space. $\mathbb{R}^N$ stands for the $N \times 1$ real vector space and $\mathbb{R}^N_+$ denotes its non-negative orthant. $\mathbb{R}_{++}$ denotes the set of strictly positive real numbers. $\nabla_{\mathbf{x}} f$ denotes the

gradient vector of function $f(\mathbf{x})$ with respect to $\mathbf{x}$ and $|\mathbf{A}|$ stands for the determinant of a matrix $\mathbf{A}$. For $\mathbf{x} \in \mathbb{R}^N$ and $\mathbf{y} \in \mathbb{R}^N$, we use $\mathbf{x} \succeq \mathbf{y}$ to indicate that $\mathbf{x}$ is greater than or equal to $\mathbf{y}$ in a component-wise manner.

II. Signal model and Problem formulation

Consider a wireless network wherein K mutually interfering links are active over a communication bandwidth B [in Hz]. Each link includes a single-antenna transmitter node and a receiver node (possibly equipped with multiple antennas). Call p_k the transmit power level [in W] of link k (from the transmitter to its intended receiver) and assume that $0 \leq p_k \leq P_{\max,k}$ where $P_{\max,k}$ is the maximal transmit power.

Denote by $\gamma_k(\mathbf{p})$ the SINR of link k as a function of $\mathbf{p} = [p_1, \ldots, p_K] \in \mathbb{R}_+^K$. At this stage, no particular expression for the function $\gamma_k(\mathbf{p})$ will be specified while only the following general assumption is made:

Assumption 1. *The function $\gamma_k(\mathbf{p}) : \mathbb{R}_+^K \to \mathbb{R}_+$ is, for all k, such that the achievable rate $R_k(\mathbf{p})$ of link k can be expressed as the difference of two non-negative functions:*

$$R_k(\mathbf{p}) = B \log_2(1 + \gamma_k(\mathbf{p})) = q_k^+(\mathbf{p}) - q_k^-(\mathbf{p}) \tag{1}$$

with $q_k^+(\mathbf{p}), q_k^-(\mathbf{p}) : \mathbb{R}_+^K \to \mathbb{R}_+$.

Additional specific assumptions on q_k^+ and q_k^- will be introduced in Sections IV and V, when discussing the monotonic fractional programming and the sequential fractional programming frameworks, respectively. For now, it should be stressed that Assumption 1 is very general, and holds in at least following three notable cases.

(*i*) The typical SINR expression in interference networks takes the general form

$$\gamma_k(\mathbf{p}) = \frac{p_k \alpha_k}{\sigma^2 + \sum\limits_{i=1, i \neq k}^{K} p_i \beta_{i,k}} \tag{2}$$

where σ^2 is the power [in W] of the receiver noise (over the bandwidth B), α_k is the channel gain over link k, whereas $\{\beta_{i,k}\}$ account for the multi-user interference and depend on the other links' channel coefficients as well as on global system parameters. The particular expression of coefficients $\{\alpha_k, \{\beta_{i,k}\}\}$ is determined by the specific system under consideration. From (2), it follows that q_k^+ and q_k^- take the form:

$$q_k^+(\mathbf{p}) = \log_2 \left(\sigma^2 + p_k \alpha_k + \sum_{i=1, i \neq k}^{K} p_i \beta_{i,k} \right) \tag{3}$$

$$q_k^-(\mathbf{p}) = \log_2 \left(\sigma^2 + \sum_{i=1, i \neq k}^{K} p_i \beta_{i,k} \right). \tag{4}$$

(*ii*) A considerable extension of (2) is obtained by considering also a self-interference term in the denominator, proportional to the useful power:

$$\gamma_k(\mathbf{p}) = \frac{p_k \alpha_k}{\sigma^2 + p_k \phi_k + \sum\limits_{i=1, i \neq k}^{K} p_i \beta_{i,k}} \tag{5}$$

with coefficients $\{\phi_k\}$ also depending on propagation channels and system parameters. A non-zero coefficient ϕ_k arises in several relevant instances of communication systems, such as hardware-impaired networks, receivers with imperfect channel state information (CSI), relay-assisted communications, and systems affected by inter-symbol interference. A detailed discussion on the communication scenarios in which the SINR may take the form in (5) can be found in [16]. Given (5), the functions q_k^+ and q_k^- are easily found to be:

$$q_k^+(\mathbf{p}) = \log_2 \left(\sigma^2 + p_k(\alpha_k + \phi_k) + \sum_{i=1, i \neq k}^{K} p_i \beta_{i,k} \right) \tag{6}$$

$$q_k^-(\mathbf{p}) = \log_2 \left(\sigma^2 + p_k \phi_k + \sum_{i=1, i \neq k}^{K} p_i \beta_{i,k} \right). \tag{7}$$

(*iii*) A third notable SINR expression is that obtained in vector channels, when linear minimum mean square error (LMMSE) reception is used at the receiver, namely

$$\gamma_k(\mathbf{p}) = p_k \mathbf{v}_{kk}^H \left(\sigma^2 \mathbf{I}_r + p_k \mathbf{u}_k \mathbf{u}_k^H + \sum_{i=1, i \neq k}^{K} p_i \mathbf{v}_{ki} \mathbf{v}_{ki}^H \right)^{-1} \mathbf{v}_{kk} \tag{8}$$

where r denotes the dimension of the received signal, $\mathbf{v}_{ki}$ is the $r \times 1$ channel vector between transmitter i and receiver k, and the $r \times 1$ vector $\mathbf{u}_k$ accounts for self-interference terms (due to the same reasons as for the SINR expression in (5)). Given (8), the functions q_k^+ and q_k^- are expressed as

$$q_k^+(\mathbf{p}) = \log_2 \left| \sigma^2 \mathbf{I}_r + p_k (\mathbf{v}_{kk} \mathbf{v}_{kk}^H + \mathbf{u}_k \mathbf{u}_k^H) + \sum_{i=1, i \neq k}^{K} p_i \mathbf{v}_{ki} \mathbf{v}_{ki}^H \right| \tag{9}$$

$$q_k^-(\mathbf{p}) = \log_2 \left| \sigma^2 \mathbf{I}_r + p_k \mathbf{u}_k \mathbf{u}_k^H + \sum_{i=1, i \neq k}^{K} p_i \mathbf{v}_{ki} \mathbf{v}_{ki}^H \right|. \tag{10}$$

In the considered interference network scenario, the EE (measured in bit/Joule) of link k is defined as the ratio of the achievable rate and the total energy consumption

$$\mathrm{EE}_k(\mathbf{p}) = \frac{B \log_2(1 + \gamma_k(\mathbf{p}))}{\mu_k p_k + \Psi_k} \tag{11}$$

wherein $\mu_k \geq 1$ is the inverse of the power amplifier efficiency of transmitter node k and Ψ_k is the circuit power required to operate link k, accounting for the dissipation in analog hardware, digital signal processing, backhaul signaling, and other overhead costs (such as cooling and power supply losses) [5], [26], [27]. Clearly, (11) is a link-centric (or user-centric) performance metric. A network-centric definition of EE must combine the individual energy efficiencies of the different links. Although different approaches have been proposed in the literature, a single definition that unarguably best represents the EE of the whole network is not available, since the different EEs are typically conflicting objectives [5], [28]. Two of the

most well-established metrics to measure the network EE are the GEE defined as

$$\mathrm{GEE}(\mathbf{p}) = \frac{\sum_{k=1}^{K} B\log_2(1+\gamma_k(\mathbf{p}))}{\sum_{k=1}^{K} \mu_k p_k + \Psi_k} \tag{12}$$

and the weighted minimum energy efficiency (WMEE) given by

$$\mathrm{WMEE}(\mathbf{p}) = \min_{k=1,\ldots,K} w_k \frac{B\log_2(1+\gamma_k(\mathbf{p}))}{\mu_k p_k + \Psi_k} \tag{13}$$

where the coefficients $w_k \in \mathbb{R}_+$ are used to weigh the EEs of the individual links. Two other possible metrics are the weighted sum energy efficiency (WSEE), defined as

$$\mathrm{WSEE}(\mathbf{p}) = \sum_{k=1}^{K} w_k \frac{B\log_2(1+\gamma_k(\mathbf{p}))}{\mu_k p_k + \Psi_k} \tag{14}$$

and the weighted product energy efficiency (WPEE), defined as

$$\mathrm{WPEE}(\mathbf{p}) = \prod_{k=1}^{K} \left[\frac{B\log_2(1+\gamma_k(\mathbf{p}))}{\mu_k p_k + \Psi_k}\right]^{w_k}. \tag{15}$$

The GEE is the metric with the strongest physical interpretation, as it represents the benefit-cost ratio of the entire network, in terms of global amount of reliably transmitted data and global amount of consumed energy. However, it does not depend on the individual EEs, and therefore does not allow tuning the EE of the individual links according to specific needs. Instead, the WMEE, WSEE, and WPEE are more connected to a multi-objective approach, in which the objectives are the individual EEs [5], [28], which turns out to be quite useful in heterogeneous networks. By suitably choosing the weights it is possible to prioritize the links that require higher EE, choosing different operating points in the system energy-efficient Pareto region, defined as the region $\mathcal{E}$ containing all feasible $K \times 1$ vectors of the users' EEs:

$$\mathcal{E} = \left\{[\mathrm{EE}_1(\mathbf{p}), \ldots, \mathrm{EE}_K(\mathbf{p})]^T \ : \ \mathbf{p} \in \mathcal{P}\right\}. \tag{16}$$

A known results from multi-objective optimization theory ensures that by globally maximizing the WMEE for different choices of the weights it is possible to describe the complete Pareto boundary of the Pareto region (16). In general this is not possible by maximizing the WSEE or WPEE; for example, the WSEE maximization allows only to describe the convex hull of the Pareto region, and therefore the boundary of (16) can not be completely characterized unless the region is convex. In light of these considerations, the main focus of this article is on the maximization of GEE and WMEE. Nevertheless, it will be shown that most of the techniques developed in the sequel are general enough to apply also to the maximization of WSEE and WPEE.

Given this background, the problem to be solved can be mathematically stated as:

$$\underset{\mathbf{p}}{\text{maximize}} \quad u(\mathbf{p}) \quad \text{s.t.} \quad \mathbf{p} \in \mathcal{P} \tag{17}$$

wherein the objective $u(\mathbf{p})$ is chosen as either the GEE or WMEE (given by (12) or (13)), whereas $\mathcal{P}$ represents the feasible set of the problem given by

$$\mathcal{P} = \left\{\mathbf{p} \in \mathbb{R}_+^K;\ p_k \leq P_{\max,k}, c_k(\mathbf{p}) \geq 0 \ \ \forall k \in \{1,\ldots,K\}\right\} \tag{18}$$

where $c_k(\mathbf{p}) : \mathbb{R}_+^K \to \mathbb{R}_+$ accounts possible additional constraint functions. As for the SINRs, no particular expression is assumed here for $c_k(\mathbf{p})$, and only the following assumption is made:

Assumption 2. *The function $c_k(\mathbf{p})$ can be expressed $\forall k$ as the difference of two non-negative functions, namely:*

$$c_k(\mathbf{p}) = c_k^+(\mathbf{p}) - c_k^-(\mathbf{p}) \tag{19}$$

with $c_k^+(\mathbf{p}), c_k^-(\mathbf{p}) : \mathbb{R}_+^K \to \mathbb{R}_+$.

Additional specific assumptions on c_k^+ and c_k^- will be introduced in Sections IV and V, when discussing the monotonic fractional programming and the sequential fractional programming frameworks, respectively. Also, observe that the per-user power constraint in (18) well models uplink transmissions. However, the constraint functions $c_k(\mathbf{p})$ can be defined to also enforce a total power constraint $\sum_{k=1}^{K} p_k \leq P_{\max}$, which is particularly relevant in the downlink.

Regardless of the choice of the EE metric $u(\mathbf{p})$, the optimization problem in (17) belongs to the class of fractional programming problems [4], [5]. Such problems can be solved with a guaranteed polynomial-time complexity only if the numerator and denominator of the fraction to maximize are respectively concave and convex, and if the feasible set is also convex [5]. Unfortunately, this requirement is not fulfilled in interference-limited networks as it follows from the general SINR expressions given above. In fact, the functions q_k^- are always non-zero whenever multi-user interference is present, and this causes the links' achievable rates (i.e., the numerators of the individual EEs) to be non-concave functions of $\mathbf{p}$. As a result, fractional programs are in general NP-hard in interference-limited scenarios [4], [5], which calls for new tools to complement and extend the potentialities of fractional programming theory. In Section IV, we aim at solving (17) at the expense of computational complexity by combining fractional programming with monotonic optimization. Then, we look for local solutions of (17) that can be obtained with affordable complexity. This is accomplished in Section V by merging fractional programming with sequential optimization. Before turning to the development of these new optimization frameworks, next section will provide the necessary mathematical preliminaries and definitions.

III. Mathematical Preliminaries

This section provides a background on the optimization theories to be used in the remainder. Section III-A gives a short review of fractional programming theory [29] and also provides some background (see [30] for more details) to understand the complexity arguments mentioned at the end of Section II. Then, Section III-B gives an overview of monotonic optimization [18], [19]. Finally, Section III-C briefly discusses the framework of sequential optimization.

A. Fractional programming

For a more comprehensive overview of fractional programming for EE maximation, the reader is referred to [5].

Definition 1 (Generalized fractional program)**.** *Let $\mathcal{D} \subseteq \mathbb{R}^N$ and consider the functions $f_k : \mathcal{D} \to \mathbb{R}$ and $g_k : \mathcal{D} \to \mathbb{R}_{++}$, with $k = 1, \ldots, K$. A generalized fractional program is the optimization problem defined as*

$$\underset{\mathbf{x}}{\text{maximize}} \ \min_{k=1,\ldots,K} \frac{f_k(\mathbf{x})}{g_k(\mathbf{x})} \quad \text{s.t.} \quad \mathbf{x} \in \mathcal{D}. \tag{20}$$

If $K = 1$, then the above problem reduces to the so-called single-ratio fractional program:

$$\underset{\mathbf{x}}{\text{maximize}} \ \frac{f_1(\mathbf{x})}{g_1(\mathbf{x})} \quad \text{s.t.} \quad \mathbf{x} \in \mathcal{D}. \tag{21}$$

Since the objective function in (20) is in general not concave, standard convex optimization algorithms are not guaranteed to solve (20) and specific algorithms are required. Towards this end, we have the following main result.

Proposition 1. *[31], [32]. A vector $\mathbf{x}^\star \in \mathcal{D}$ solves* (20) *if and only if*

$$\mathbf{x}^\star = \arg\max_{\mathbf{x}\in\mathcal{D}} \left\{ \min_{k=1,\ldots,K} \left[f_k(\mathbf{x}) - \lambda^\star g_k(\mathbf{x}) \right] \right\} \tag{22}$$

with $\lambda^\star$ being the unique zero of the auxiliary function $F(\lambda)$:

$$F(\lambda) = \max_{\mathbf{x}\in\mathcal{D}} \min_{k=1,\ldots,K} \left\{ f_k(\mathbf{x}) - \lambda g_k(\mathbf{x}) \right\}. \tag{23}$$

This result allows one to solve (20) by finding the unique zero of $F(\lambda)$. To this end, the most widely used algorithm is the (Generalized, if $K > 1$) Dinkelbach's algorithm [30], [32], reported in Algorithm 1.

Algorithm 1 Generalized Dinkelbach's algorithm

```
Initialize λ_0 with F(λ_0) ≥ 0, j = 0;
while F(λ_j) > ε do
    Solve the problem:
      x*_j = arg max_{x∈D} { min_{k=1,...,K} [ f_k(x) − λ_j g_k(x) ] };
    λ_{j+1} = min_{k=1,...,K} f_k(x*_j) / g_k(x*_j);
    j = j + 1;
end while
```

It can be shown that the update rule for λ follows Newton's method applied to the function $F(\lambda)$ [30]. Hence, Algorithm 1 exhibits a super-linear convergence rate, but converges to the global optimum of the corresponding instance of the fractional problem only provided that (22) can be *globally* solved at each iteration. If f is concave, g is convex, and all constraints are also convex, then this can be accomplished with polynomial-time complexity. Instead, if any of these requirements is not fulfilled, then (22) becomes a non-convex problem. As a consequence, the well-developed theory of convex optimization can not handle (22), which instead requires the use of global optimization algorithms. However, standard global optimization methods operate by exploring the whole feasible set [33], with a prohibitive computational complexity, even for small problem instances, and with a convergence that is only guaranteed if the functions have a limited variability (e.g., Lipschitz continuity [24]).

B. Monotonic optimization

Monotonic optimization is a relatively recent global optimization framework, which exploits monotonicity or hidden monotonicity structures in the objective and constraints to reduce computational complexity and provide a guaranteed convergence [18], [19]. The basic idea is that if the objective to maximize is increasing in all optimization variables, then it is not necessary to explore the complete feasible set of the problem, but only its outer boundary. This concept is made formal in the rest of this section.

Definition 2 (Monotonicity in $\mathbb{R}^N$)**.** *A function $f : \mathbb{R}^N \to \mathbb{R}$ is monotonically increasing if $f(\mathbf{y}) \geq f(\mathbf{x})$ when $\mathbf{y} \succeq \mathbf{x}$.*

Definition 3 (Hyper-rectangle in $\mathbb{R}^N$)**.** *Let $\mathbf{a},\ \mathbf{b} \in \mathbb{R}^N$ with $\mathbf{a} \preceq \mathbf{b}$. Then, the set of all $\mathbf{x} \in \mathbb{R}^N$ such that $\mathbf{a} \preceq \mathbf{x} \preceq \mathbf{b}$ is a hyper-rectangle in $\mathbb{R}^N$ and is denoted by $[\mathbf{a}, \mathbf{b}]$.*

Definition 4 (Normal and Co-normal sets)**.** *A set $\mathcal{S} \subset \mathbb{R}^N$ is normal if $\forall \mathbf{x} \in \mathcal{S}$, the hyper-rectangle $[\mathbf{0}, \mathbf{x}]$ belongs to $\mathcal{S}$. A set $\mathcal{S}_c \subset \mathbb{R}^N$ is co-normal in $[\mathbf{0}, \mathbf{b}]$ if $\forall \mathbf{x} \in \mathcal{S}_c$, then $[\mathbf{x}, \mathbf{b}] \subset \mathcal{S}_c$.*

A given function $h : \mathbb{R}^N \to \mathbb{R}$ defines a normal or a co-normal set if the following results hold true:

Proposition 2. *[18] The set $\mathcal{S} = \{\mathbf{x} \in \mathbb{R}^N : h(\mathbf{x}) \leq 0\}$ is normal and closed if h is lower semi-continuous and increasing. The set $\mathcal{S}_c = \{\mathbf{x} \in \mathbb{R}^N : h(\mathbf{x}) \geq 0\}$ is co-normal and closed if h is upper semi-continuous and increasing.*

Definition 5 (Monotonic optimization)**.** *A monotonic optimization problem in canonical form is defined as*

$$\underset{\mathbf{x}}{\text{maximize}} \ f(\mathbf{x}) \quad \text{s.t.} \quad \mathbf{x} \in \mathcal{S} \cap \mathcal{S}_c \tag{24}$$

where $f : \mathbb{R}^N \to \mathbb{R}$ is an increasing function, $\mathcal{S} \subset [\mathbf{0}, \mathbf{b}]$ is a compact, normal set with nonempty interior, and $\mathcal{S}_c$ is a closed co-normal set in $[\mathbf{0}, \mathbf{b}]$.

The main result of monotonic optimization theory states that the solution to (24) lies on the upper boundary of $\mathcal{S} \cap \mathcal{S}_c$ [18, Proposition 7]. Therefore, methods like the polyblock algorithm [18] and the BRB algorithm [19] can be used to globally solve (24) by searching only on the upper boundary of the feasible set, thus drastically simplifying the problem. Nevertheless, we remark that the complexity of monotonic optimization methods is still exponential in the number of variables. However, as already observed, it is much lower than general global optimization methods, which do not exploit any monotonicity structure [18]. This makes monotonic optimization attractive for the development of a framework to benchmark any suboptimal method for solving (24).

C. Sequential optimization

Sequential optimization is a powerful tool that provides the means to generate candidate solutions of non-convex

optimization problems with affordable complexity [34], while at the same time satisfying theoretical optimality claims. This statement is made precise in the following result:

Proposition 3. *[34] Let $\mathcal{F}$ be a maximization problem with continuous objective $f_0(\mathbf{x})$ and constraints $f_i(\mathbf{x}) \geq 0$ $\forall i \in \{1,\dots,I\}$ that define a compact set. Let $\mathcal{G}_j$ be a maximization problem with objective function $g_{0,j}(\mathbf{x})$, constraints $g_{i,j}(\mathbf{x}) \geq 0$ $\forall i \in \{1,\dots,I\}$, and optimal solution $\mathbf{x}_j^\star$. Assume that $\forall j$ and $\forall i \in \{1,\dots,I\}$ $g_{i,j}(\cdot)$ satisfies the following two properties:*

1) $g_{i,j}(\mathbf{x}) \leq f_i(\mathbf{x})$ $\forall \mathbf{x}$;
2) $g_{i,j}(\mathbf{x}_{j-1}^\star) = f_i(\mathbf{x}_{j-1}^\star)$.

Then, the sequence $\{\mathbf{x}_j^\star\}$ of solutions of $\{\mathcal{G}_j\}$ converges, and $\forall j$, $f_0(\mathbf{x}_j^\star) \geq f_0(\mathbf{x}_{j-1}^\star)$.
If the following third property is also satisfied $\forall j$ and $\forall i \in \{1,\dots,I\}$:

3) $\nabla g_{i,j}(\mathbf{x}_{j-1}^\star) = \nabla f_i(\mathbf{x}_{j-1}^\star)$

then the sequence $\{\mathbf{x}_j^\star\}$ converges to a point satisfying the KKT conditions of the original problem $\mathcal{F}$.

Proposition 3 shows that by solving the sequence of approximate problems $\{\mathcal{G}_j\}$, one can generate a sequence of feasible points $\mathbf{x}_j^\star$ that monotonically increases the value of the original objective f_0 and that converges to a point fulfilling the KKT optimality conditions of $\mathcal{F}$. The critical issue for this tool to be of practical use, is to find suitable approximate problems $\{\mathcal{G}_j\}$ fulfilling the assumptions of Proposition 3, while at the same being easier to solve than the original problem.

IV. Global optimality: Monotonic Fractional Programming

As mentioned above, among global optimization algorithms, monotonic optimization provides attractive complexity and convergence properties. However, it can not be directly employed to solve (17), because the EEs are not monotone functions of $\mathbf{p}$ in the sense of Definition 2. This section shows how this difficulty can be overcome by an interplay of fractional programming and monotonic optimization, provided that the following assumption holds:

Assumption 3. *The functions $q_k^+(\mathbf{p})$ and $q_k^-(\mathbf{p})$ in* (1)*, and the functions $c_k^+(\mathbf{p})$ and $c_k^-(\mathbf{p})$ in* (19) *are monotonic functions $\forall k \in \{1,\dots,K\}$ as stated in Definition 2.*

In other words, it is assumed that all achievable rates and constraint functions can be written as the difference of monotonic functions. Observe that no assumption on the concavity or convexity of $q_k^+(\mathbf{p})$, $q_k^-(\mathbf{p})$, $c_k^+(\mathbf{p})$, and $c_k^-(\mathbf{p})$ is made.

A. GEE maximization

GEE maximization belongs to the class of single-ratio fractional problems. Thus, finding its solution by the Dinkelbach's algorithm requires to solve the following auxiliary problem at iteration j:

$$\underset{\mathbf{p}}{\text{maximize}} \sum_{k=1}^{K} B\log_2(1+\gamma_k) - \lambda_j(\mu_k p_k + \Psi_k) \quad \text{s.t. } \mathbf{p}\in\mathcal{P} \tag{25}$$

for a given positive λ_j. Note that, at a first sight, the above problem is not a monotonic optimization problem in canonical form because:

- The objective function is not monotonic, since the achievable rates $\log_2(1+\gamma_k)$ are not increasing functions of $\mathbf{p}$, and since the negative term is in fact decreasing in the interfering powers.
- The constraint set is not guaranteed to be the intersection of a normal and a co-normal set, since the difference of two increasing functions is in general not increasing.

However, (25) exhibits a hidden monotonicity structure as shown in the following proposition:

Proposition 4. *If Assumption 3 holds true, then* (25) *can be expressed a monotonic optimization problem in canonical form.*

Proof: Observe that (25) can be equivalently written as

$$\underset{\mathbf{p}}{\text{maximize}} \quad q^+(\mathbf{p}) - q^-(\mathbf{p},\lambda_j) \quad \text{s.t.} \quad \mathbf{p}\in\mathcal{P} \tag{26}$$

wherein $q^+(\mathbf{p})$ and $q^-(\mathbf{p},\lambda_j)$ are increasing in $\mathbf{p}$ and given by

$$q^+(\mathbf{p}) = \sum_{k=1}^{K} q_k^+(\mathbf{p}) \tag{27}$$

$$q^-(\mathbf{p},\lambda_j) = \sum_{k=1}^{K} q_k^-(\mathbf{p}) + \lambda_j\big(\mu_k p_k + \Psi_k\big). \tag{28}$$

Next, define $\mathbf{p}_{\max} = [P_{\max,1},\dots,P_{\max,K}]$ and introduce the auxiliary variable $t = q^-(\mathbf{p}_{\max},\lambda_j) - q^-(\mathbf{p},\lambda_j)$. Then, for any given λ_j, (26) can be rewritten as

$$\underset{(t,\,\mathbf{p})}{\text{maximize}} \quad q^+(\mathbf{p}) + t \tag{29}$$

$$\text{s.t.} \quad (t,\,\mathbf{p}) \in \mathcal{P}\cap\mathcal{Q} \tag{30}$$

with

$$\mathcal{Q} = \left\{ (t,\mathbf{p}) : \begin{array}{c} 0 \leq t + q^-(\mathbf{p},\lambda_j) \leq q^-(\mathbf{p}_{\max},\lambda_j) \\ 0 \leq t \leq q^-(\mathbf{p}_{\max},\lambda_j) - q^-(\mathbf{0}_K,\lambda_j) \end{array} \right\}.$$

Problem (29) is not a monotonic problem yet, because the constraint functions $c_k(\mathbf{p})$ are expressed as the difference of increasing functions. To overcome this problem, observe that the set of constraints $c_k(\mathbf{p}) \geq 0$ with $k = 1,\dots,K$, can be equivalently rewritten as the following single constraint:

$$\min_{k=1,\dots,K} \big[c_k^+(\mathbf{p}) - c_k^-(\mathbf{p})\big] \geq 0 \iff \tag{31}$$

$$\min_{k=1,\dots,K} \left[c_k^+(\mathbf{p}) - \left(\sum_{i=1}^{K} c_i^-(\mathbf{p}) - \sum_{i=1,i\neq k}^{K} c_i^-(\mathbf{p}) \right) \right] = \tag{32}$$

$$\underbrace{\min_{k=1,\dots,K} \left[c_k^+(\mathbf{p}) + \sum_{i=1,i\neq k}^{K} c_i^-(\mathbf{p}) \right]}_{c^+(\mathbf{p})} - \underbrace{\sum_{i=1}^{K} c_i^-(\mathbf{p})}_{c^-(\mathbf{p})} \geq 0 \tag{33}$$

which is the difference of the two increasing functions $c^+(\mathbf{p})$ and $c^-(\mathbf{p})$. Similarly as above, we can thus introduce the

auxiliary variable s and reformulate (29) as

$$\begin{aligned}\underset{(t,s,\mathbf{p})}{\text{maximize}} \quad & q^+(\mathbf{p})+t \\ \text{s.t. } & (t,\,\mathbf{p})\in\mathcal{Q}\ ,\ 0\le s\le c^-(\mathbf{p}_{\max})-c^-(\mathbf{0}_K) \\ & c^-(\mathbf{p})+s\le c^-(\mathbf{p}_{\max})\ , c^+(\mathbf{p})+s\ge c^-(\mathbf{p}_{\max}).\end{aligned} \tag{34}$$

In order to complete the proof, it remains to verify that (34) fulfills Definition 5, thus being a monotonic problem in canonical form. To this end, let us first observe that the objective of (34) is monotonic in $(t,s,\mathbf{p})$.

Next, to show that the feasible set of (34) is the intersection of a normal and a co-normal set, let us observe that, for any $\mathbf{p}$ in the feasible set, we have that

$$q^-(\mathbf{0}_K,\lambda_j)\le q^-(\mathbf{p},\lambda_j) \tag{35}$$

$$c^-(\mathbf{0}_K)\le c^-(\mathbf{p}). \tag{36}$$

As a consequence, the feasible set of (34) can be written as the intersection of the following two sets:

$$\begin{aligned}\mathcal{S}=\Big\{(t,s,\mathbf{p}):\,&\mathbf{p}\preceq\mathbf{p}_{\max}, t+q^-(\mathbf{p},\lambda_j)\le q^-(\mathbf{p}_{\max},\lambda_j),\\ &s+c^-(\mathbf{p})\le c^-(\mathbf{p}_{\max})\Big\}\end{aligned} \tag{37}$$

$$\mathcal{S}_c=\left\{(t,s,\mathbf{p}):\mathbf{p}\succeq\mathbf{0}_K, t\ge 0, s+c^+(\mathbf{p})\ge c^-(\mathbf{p}_{\max})\right\}. \tag{38}$$

Then, since all the constraint functions in (37) and (38) are monotonic and continuous, by virtue of Proposition 2 and employing again (35), it follows that $\mathcal{S}$ and $\mathcal{S}_c$ are normal and co-normal sets in the hyper-rectangle given by:

$$\begin{aligned}[0,q^-(\mathbf{p}_{\max},\lambda_j)-q^-(\mathbf{0}_K,\lambda_j)]\ &\times\ [c^-(\mathbf{p}_{\max})-c^-(\mathbf{0}_K)]\\ &\times\ [\mathbf{0}_K,\mathbf{p}_{\max}].\end{aligned} \tag{39}$$

This completes the proof. ∎

B. WMEE maximization

WMEE maximization belongs to the class of generalized fractional programs and requires to solve the following auxiliary problem at iteration j:

$$\begin{aligned}\underset{\mathbf{p}}{\text{maximize}} \ & \min_{k=1,\ldots,K} q_k^+(\mathbf{p})-q_k^-(\mathbf{p})-\lambda_j(\mu_k p_k+\Psi_k)\\ \text{s.t.} \quad & \mathbf{p}\in\mathcal{P}.\end{aligned} \tag{40}$$

As in the case of GEE maximization, the objective function is not monotonic. However, the following result can be proved:

Proposition 5. *If Assumption 3 holds true, then* (40) *can be expressed as a monotonic problem in canonical form.*

Proof: Let $\nu_k(\mathbf{p},\lambda_j)=q_k^-(\mathbf{p})+\lambda_j(\mu_k p_k+\Psi_k)$ such that we may rewrite the objective function as

$$\begin{aligned}&q_k^+(\mathbf{p})-q_k^-(\mathbf{p})-\lambda_j(\mu_k p_k+\Psi_k)\\ &=q_k^+(\mathbf{p})-\left(\sum_{i=1}^K\nu_i(\mathbf{p},\lambda_j)-\sum_{i=1,i\neq k}^K\nu_i(\mathbf{p},\lambda_j)\right)\\ &=\left(q_k^+(\mathbf{p})+\sum_{i=1,i\neq k}^K\nu_i(\mathbf{p},\lambda_j)\right)-\sum_{i=1}^K\nu_i(\mathbf{p},\lambda_j).\end{aligned} \tag{41}$$

Then, introduce $t=\sum_{i=1}^K\nu_i(\mathbf{p}_{\max},\lambda_j)-\sum_{i=1}^K\nu_i(\mathbf{p},\lambda_j)$ and reformulate (40) as

$$\begin{aligned}\underset{(t,\mathbf{p})}{\text{maximize}} \ & \min_{k=1,\ldots,K} q_k^+(\mathbf{p})+\sum_{i=1,i\neq k}^K\nu_i(\mathbf{p},\lambda_j)+t\\ \text{s.t.} \quad & (t,\,\mathbf{p})\in\mathcal{P}\cap\mathcal{Q}'\end{aligned} \tag{42}$$

with

$$\mathcal{Q}'=\left\{(t,\mathbf{p}):\begin{array}{l}0\le t\le\sum_{i=1}^K\nu_i(\mathbf{p}_{\max},\lambda_j)-\nu_i(\mathbf{p},\lambda_j)\\ 0\le t\le\sum_{i=1}^K\nu_i(\mathbf{p}_{\max},\lambda_j)-\nu_i(\mathbf{0}_K,\lambda_j)\end{array}\right\}.$$

Reformulating the constraints $c_k(\mathbf{p})\ge 0\ \forall k$ as the single constraint (33), (42) is shown to be a monotonic problem in canonical form by using the same arguments adopted in the proof of Proposition 4. ∎

Observe that the results of Propositions 4 and 5 do not specifically require an affine power consumption model, as the one $\mu_k p_k+\Psi_k$ adopted in this article. Indeed, the above results can be extended to any power consumption model such that $q^-(\mathbf{p},\lambda_j)$ (for GEE maximization) and $\nu_k(\mathbf{p},\lambda_j)$ (for WMEE maximization) are monotonic in $\mathbf{p}$. Moreover, the above result applies also to scenarios in which each user has $N_c>1$ constraint functions $\{c_{k,i}(\mathbf{p})\}_{i=1}^{N_c}$, provided $c_{k,i}$ can be still expressed as the difference of two monotonic functions.

C. WSEE and WPEE maximization

We now look at the maximization of WSEE and WPEE, which belong to the classes of sum of ratios and product of ratios problems, respectively. Both are hard to solve even if all numerators are concave, all denominators are convex, and the feasible set is convex [5], [35]. Nevertheless, the proposed monotonic fractional programming framework can also be used to solve WSEE and WPEE maximization problems. To begin with, observe that the WSEE can be expressed as a single ratio:

$$\mathrm{WSEE}(\mathbf{p})=\sum_{k=1}^K\frac{q_k^+(\mathbf{p})-q_k^-(\mathbf{p})}{\mu_k p_k+\Psi_k} \tag{43}$$

$$=\frac{\sum_{k=1}^K\left(q_k^+(\mathbf{p})-q_k(\mathbf{p})^-\right)\prod_{i\neq k}(\mu_i p_i+\Psi_i)}{\prod_{k=1}^K(\mu_k p_k+\Psi_k)}. \tag{44}$$

Since the product of increasing functions is still an increasing function, (44) turns out to be a single ratio whose numerator is the difference of increasing functions, and the denominator is an increasing function. Hence, the method adopted in Section IV-A can also be used to globally maximize the WSEE.

The same property holds for WPEE maximization since (15) can be reformulated as

$$\mathrm{WPEE}(\mathbf{p})=\frac{\prod_{k=1}^K\left(q_k^+(\mathbf{p})-q_k^-(\mathbf{p})\right)}{\prod_{k=1}^K(\mu_k p_k+\Psi_k)}. \tag{45}$$

Expanding the products in the numerator of the above function, we obtain again the difference of two increasing functions, while the denominator is clearly increasing with respect to $\mathbf{p}$.

V. Local optimality: Sequential Fractional Programming

As observed in Section IV, despite enjoying a lower complexity than standard global optimization algorithms, the complexity of the proposed monotonic fractional programming framework is still exponential. Motivated by the need of providing also a practical optimization framework for large networks, in this section fractional programming is combined with the sequential optimization results from Proposition 3. This results in a novel sequential fractional programming framework able to compute candidate solutions of EE problems with polynomial-time complexity, while at the same time fulfilling theoretical optimality claims. In particular, the proposed algorithm will be guaranteed to converge to points fulfilling the KKT first-order optimality conditions of the GEE and WMEE maximization problems. Section VI will provide numerical evidence that the sequential fractional programming approach actually attains global optimality, as it finds the same solution as the globally optimal solution computed by means of the monotonic fractional programming framework in Section IV.

This is not the first time that fractional programming and sequential optimization are used together to solve EE optimization problems [14], [16]. In [14], the two theories are employed to maximize the GEE as well as the WPEE under the assumption that the SINR takes the form in (2) whereas (5) is adopted in [16]. In these works, the approximate problems $\mathcal{G}_j$ required by Proposition 3 are obtained by means of a logarithmic approximation of the achievable rate plus a change of variable. Compared to [14], [16], a different approach is pursued here that has the following main advantages: *i*) no change of variables is required, thereby being more direct than previous approaches; *ii*) it is more general since it can handle SINRs in the form of (8), besides those given as (2) and (5), whereas the approaches from [14], [16] do not apply to the SINR expression in (8). In particular, we only require the following additional assumption:

Assumption 4. *The functions $q_k^+(\mathbf{p})$ and $q_k^-(\mathbf{p})$ in* (1)*, and the functions $c_k^+(\mathbf{p})$ and $c_k^-(\mathbf{p})$ in* (19) *are concave functions of $\mathbf{p}$ $\forall k$.*

In other words, it is only required that all the achievable rates and constraint functions can be expressed as the difference of concave functions. Observe that Assumption 3 is not required in the sequel.[1] As in Section IV, we consider the GEE and WMEE maximizations separately.

A. GEE maximization

By virtue of Assumptions 1 and 2, the GEE maximization problem can be cast as:

$$\underset{\mathbf{p}}{\text{maximize}} \quad \frac{\sum_{k=1}^K q_k^+(\mathbf{p}) - q_k^-(\mathbf{p})}{\sum_{k=1}^K \mu_k p_k + \Psi_k} \tag{46a}$$

$$\text{s.t.} \quad 0 \leq p_k \leq P_{\max,k}\ \forall k \tag{46b}$$

$$c_k^+(\mathbf{p}) - c_k^-(\mathbf{p}) \geq 0\ \forall k. \tag{46c}$$

[1]Recall that in general no relation exists between concavity and monotonicity. A function can be concave but not monotonic, or vice versa.

If Assumption 4 holds true, then the numerator of (46a) and the constraint functions in (46c) are expressed as the difference of concave functions, and therefore are not concave in general. As pointed out in Section III-A, this prevents from directly using fractional programming to solve (46). To circumvent this issue, we exploit its hidden structure and obtain the following main result:

Proposition 6. *For any given $\mathbf{p}_j$, denote by $\mathcal{G}_j$ the optimization problem*

$$\underset{\mathbf{p}}{\text{maximize}} \ \frac{\sum_{k=1}^K q_k^+(\mathbf{p}) - \left[q_k^-(\mathbf{p}_j) + \left(\nabla_{\mathbf{p}} q_k^-|_{\mathbf{p}=\mathbf{p}_j}\right)^T(\mathbf{p}-\mathbf{p}_j)\right]}{\sum_{k=1}^K \mu_k p_k + \Psi_k} \tag{47a}$$

$$\text{s.t. } 0 \leq p_k \leq P_{\max,k}\ \forall k \tag{47b}$$

$$c_k^+(\mathbf{p}) - \left[c_k^-(\mathbf{p}_j) + \left(\nabla_{\mathbf{p}} c_k^-|_{\mathbf{p}=\mathbf{p}_j}\right)^T(\mathbf{p}-\mathbf{p}_j)\right] \geq 0\ \forall k \tag{47c}$$

and call $\mathbf{p}_j^\star$ its optimal solution. If $\mathbf{p}_j = \mathbf{p}_{j-1}^\star$ $\forall j \geq 1$, and $\mathbf{p}_0$ is any feasible power vector, then $\{\mathbf{p}_j^\star\}$ converges to a point fulfilling the KKT optimality conditions of (46)*. Moreover, $\forall j \geq 1$, $\text{GEE}(\mathbf{p}_j^\star) \geq \text{GEE}(\mathbf{p}_{j-1}^\star)$.*

Proof: Recall that any concave function is upper-bounded by its first-order Taylor expansion at any point. Since $q_k^-(\mathbf{p})$ and $c_k^-(\mathbf{p})$ are concave functions, for any power vector $\mathbf{p}_j$ we thus have that

$$\begin{aligned} q_k^+(\mathbf{p}) - q_k^-(\mathbf{p}) &\geq q_k^+(\mathbf{p}) - \left[q_k^-(\mathbf{p}_j) + \left(\nabla_{\mathbf{p}} q_k^-|_{\mathbf{p}=\mathbf{p}_j}\right)^T(\mathbf{p}-\mathbf{p}_j)\right] \\ c_k^+(\mathbf{p}) - c_k^-(\mathbf{p}) &\geq c_k^+(\mathbf{p}) - \left[c_k^-(\mathbf{p}_j) + \left(\nabla_{\mathbf{p}} c_k^-|_{\mathbf{p}=\mathbf{p}_j}\right)^T(\mathbf{p}-\mathbf{p}_j)\right] \end{aligned} \tag{48}$$

Hence, (47a) and (47c) are lower bounds of (46a) and (46c), respectively. Next, since the lower bounds in (48) are tight when evaluated in $\mathbf{p}_j$, it immediately follows that (47a) and (47c) are equal to (46a) and (46c), respectively, for $\mathbf{p} = \mathbf{p}_j$. Similarly, it can be shown that the gradients of (47a) and (47c) are equal to those of (46a) and (46c), for $\mathbf{p} = \mathbf{p}_j$. Thus, (47) fulfills all the assumptions of Proposition 3 with respect to (46), which completes the proof of this proposition. ■

From Proposition 7, it follows that we can monotonically increase the GEE value, eventually fulfilling the KKT first-order optimality conditions of (46), by solving a sequence of problems like (47). To this end, observe that for any $\mathbf{p}_j$, (47a) has a concave numerator, and affine denominator, while the constraint functions in (47b) and (47c) are all affine or concave. As a result, (47) is a single-ratio problem, which can be globally solved with polynomial-time complexity by means of fractional programming theory and in particular using Algorithm 1.

Remark 1. *Observe that the gradients in* (48) *can be computed in closed form, once the SINR expression is specified.*

If the SINRs are expressed as in (5)*, then we have*

$$\frac{\partial q_k^-}{\partial p_\ell}=\begin{cases}\dfrac{\phi_k}{\ln(2)\left(\sigma^2+\phi_k p_k+\sum_{i=1,i\neq k}^K \beta_{i,k}p_i\right)}, & \ell=k,\\ \dfrac{\beta_{\ell,k}}{\ln(2)\left(\sigma^2+\phi_k p_k+\sum_{i=1,i\neq k}^K \beta_{i,k}p_i\right)} & \ell\neq k.\end{cases} \tag{49}$$

The formulas for the SINR (2) *can be obtained by setting* $\phi_k = 0$ *in* (49)*. On the other hand, if the SINR takes the expression in* (8)*, then we have*[2]

$$\frac{\partial q_k^-}{\partial p_\ell}=\begin{cases}\dfrac{1}{\ln(2)}\mathbf{u}_k^H\left(\sigma^2\mathbf{I}_r+p_k\mathbf{u}_k\mathbf{u}_k^H+\sum\limits_{i=1,i\neq k}^K p_i\mathbf{v}_{ki}\mathbf{v}_{ki}^H\right)^{-1}\mathbf{u}_k\ , \ell=k,\\ \dfrac{1}{\ln(2)}\mathbf{v}_{k\ell}^H\left(\sigma^2\mathbf{I}_r+p_k\mathbf{u}_k\mathbf{u}_k^H+\sum\limits_{i=1,i\neq k}^K p_i\mathbf{v}_{ki}\mathbf{v}_{ki}^H\right)^{-1}\mathbf{v}_{k\ell}\ , \ell\neq k.\end{cases} \tag{50}$$

B. WMEE maximization

If Assumptions 1 and 2 are satisfied, the WMEE maximization problem can be cast as

$$\underset{\mathbf{p}}{\text{maximize}} \quad \min_{k=1,\ldots,K}\frac{q_k^+(\mathbf{p})-q_k^-(\mathbf{p})}{\mu_k p_k+\Psi_k} \tag{51a}$$

$$\text{s.t.} \quad 0\leq p_k\leq P_{\max,k}\ \forall k \tag{51b}$$

$$c_k^+(\mathbf{p})-c_k^-(\mathbf{p})\geq 0\ \forall k. \tag{51c}$$

By virtue of Assumption 4, each numerator in (51a) and the constraint functions in (51c) are the difference of concave functions, and therefore are not concave in general. In principle, the same approach used for the GEE could be used here to obtain a polynomial-time power allocation algorithm. However, this does not yield a KKT point of (51), for the simple reason that, unlike the GEE, the WMEE is not differentiable, thus implying that (51) admits no KKT conditions. Instead, it is possible to obtain a KKT solution of the epigraph-form representation of (51), expressed as [37]:

$$\underset{(t,\mathbf{p})}{\text{maximize}} \quad t \tag{52a}$$

$$\text{s.t.} \quad 0\leq p_k\leq P_{\max,k}\ \forall k \tag{52b}$$

$$c_k^+(\mathbf{p})-c_k^-(\mathbf{p})\geq 0\ \forall k \tag{52c}$$

$$q_k^+(\mathbf{p})-q_k^-(\mathbf{p})-t(\mu_k p_k+\Psi_k)\geq 0\ \forall k. \tag{52d}$$

Then, we have the following main result:

Proposition 7. *For any given* $\mathbf{p}_j$*, denote by* $\mathcal{G}_j$ *the optimization problem*

$$\max_{(t,\mathbf{p})}\ \min_{k=1,\ldots,K}\frac{q_k^+(\mathbf{p})-\left[q_k^-(\mathbf{p}_j)+\left(\nabla_{\mathbf{p}}q_k^-|_{\mathbf{p}=\mathbf{p}_j}\right)^T(\mathbf{p}-\mathbf{p}_j)\right]}{\mu_k p_k+\Psi_k} \tag{53a}$$

$$\text{s.t. } 0\leq p_k\leq P_{\max,k}\ \forall k \tag{53b}$$

$$c_k^+(\mathbf{p})-\left[c_k^-(\mathbf{p}_j)+\left(\nabla_{\mathbf{p}}c_k^-|_{\mathbf{p}=\mathbf{p}_j}\right)^T(\mathbf{p}-\mathbf{p}_j)\right]\geq 0\ \forall k \tag{53c}$$

[2]Recall that $\dfrac{\partial \log_2|\mathbf{A}+x\mathbf{B}|}{\partial x}=\mathrm{tr}\left((\mathbf{A}+x\mathbf{B})^{-1}\mathbf{B}\right)$, for a scalar x and $\mathbf{A},\mathbf{B}$ being square matrices of proper dimensions [36].

and call $\mathbf{p}_j^\star$ *its optimal solution. If* $\mathbf{p}_j=\mathbf{p}_{j-1}^\star\ \forall j\geq 1$*, and* $\mathbf{p}_0$ *is any feasible power vector, then* $\{\mathbf{p}_j^\star\}$ *converges to a point fulfilling the KKT optimality conditions of* (52)*. Moreover,* $\forall j\geq 1$*,* $\mathrm{WMEE}(\mathbf{p}_j^\star)\geq\mathrm{WMEE}(\mathbf{p}_{j-1}^\star)$.

Proof: The proof is articulated into two parts. The first one follows along the same line of reasoning used for Proposition 7 and aims at showing that (53) fulfills the first two properties of Proposition 3 with respect to (51). As a result, we have that $\forall j\geq 1$, $\mathrm{WMEE}(\mathbf{p}_j^\star)\geq\mathrm{WMEE}(\mathbf{p}_{j-1}^\star)$. The second part considers the epigraph-form of (53) given by:

$$\max_{(t,\mathbf{p})}\ t \tag{54a}$$

$$\text{s.t. } 0\leq p_k\leq P_{\max,k}\ \forall k \tag{54b}$$

$$c_k^+(\mathbf{p})-\left[c_k^-(\mathbf{p}_j)+\left(\nabla_{\mathbf{p}}c_k^-|_{\mathbf{p}=\mathbf{p}_j}\right)^T(\mathbf{p}-\mathbf{p}_j)\right]\geq 0\ \forall k \tag{54c}$$

$$q_k^+(\mathbf{p})-\left[q_k^-(\mathbf{p}_j)+\left(\nabla_{\mathbf{p}}q_k^-|_{\mathbf{p}=\mathbf{p}_j}\right)^T(\mathbf{p}-\mathbf{p}_j)\right]-t(\mu_k p_k+\Psi_k)\geq 0\ \forall k\ . \tag{54d}$$

By similar arguments as those used to prove Proposition 7, it follows that (54) fulfills the three properties in Proposition 3 with respect to (52). Also, for any $j\geq 1$, $\mathbf{p}_j^\star$ is a solution of both (53) and (54). These two facts together imply that upon convergence the KKT conditions of (52) are fulfilled. ■

Observe that the solution of (53) can be computed with polynomial complexity by means of Algorithm 1, since the numerator and denominator of (53a) are concave and convex, respectively, and all constraint functions are concave or affine.

VI. Numerical examples

In this section we analyze the performance of the proposed monotonic fractional programming framework from Section IV and of the sequential fractional programming framework from Section V. Among the many possible scenarios under the umbrella of the proposed frameworks, we focus on the two case-studies of a multi-antenna LTE network, the leading standard in 4G systems, and of a massive multiple-input multiple-output (MIMO) network, one of the strongest candidate technologies for 5G networks.

A. MIMO LTE network with LMMSE detection

Consider the uplink of a multi-cell LTE system in which the same resource block is used by multiple user equipments (UEs). Each UE and each base station (BS) are equipped with N_T and N_R antennas, respectively. Let us denote by $\mathbf{H}_{k,\ell}\in\mathbb{C}^{N_R\times N_T}$ the channel matrix between UE k and BS ℓ, while $\mathbf{q}_k\in\mathbb{C}^{N_T}$ is the unit-norm precoding vector, and s_k is the unit-modulus information symbol sent by UE k. UE k is associated with BS $a(k)$. In order to perform data detection, the signal received at BS $a(k)$ is linearly filtered by an LMMSE detector, which is well-known to be the optimal linear receive structure [38]. For the case at hand, the signal received at BS $a(k)$ is

$$\mathbf{y}_{a(k)}=\mathbf{c}_k^H\left(\sqrt{p_k}\mathbf{H}_{k,a(k)}\mathbf{q}_k s_k+\sum_{i=1,i\neq k}^K\sqrt{p_i}\mathbf{H}_{i,a(k)}\mathbf{q}_i s_i+\mathbf{n}_{a(k)}\right) \tag{55}$$

with $\mathbf{n}_{a(k)} \sim \mathcal{CN}(0, \sigma^2)$ modeling the receiver noise and the LMMSE receiver $\mathbf{c}_k$ given by

$$\mathbf{c}_k = \sqrt{p_k}\left(\sigma^2 \mathbf{I}_{N_R} + \sum_{i=1, i\neq k} p_i \mathbf{H}_{i,a(k)} \mathbf{q}_i \mathbf{q}_i^H \mathbf{H}_{i,a(k)}^H\right)^{-1} \mathbf{H}_{k,a(k)} \mathbf{q}_k. \tag{56}$$

Plugging (56) into (55), the SINR enjoyed by UE k turns out to be

$$\gamma_k = p_k \mathbf{q}_k^H \mathbf{H}_{k,a(k)}^H \left(\sigma^2 \mathbf{I}_{N_R} + \sum_{i=1, i\neq k} p_i \mathbf{H}_{i,a(k)} \mathbf{q}_i \mathbf{q}_i^H \mathbf{H}_{i,a(k)}^H\right)^{-1} \times \mathbf{H}_{k,a(k)} \mathbf{q}_k, \tag{57}$$

which is formally equivalent to (8), with $\mathbf{v}_{ki} = \mathbf{H}_{i,a(k)} \mathbf{q}_i$, $\mathbf{u}_k = \mathbf{0}$, and $r = N_R$. Thus, the functions $q_k^+(\mathbf{p})$ and $q_k^-(\mathbf{p})$ are expressed as in (9) and (10).

In our numerical simulations, we considered a square area of $2\,\text{km} \times 2\,\text{km}$, wherein UEs are randomly placed and equipped with $N_T = 2$ antennas each. The area is served by $L = 3$ BSs placed at coordinates $(0.5; 0.5)$ km, $(0.5; -0.5)$ km, $(-0.5; 0)$ km, with respect to a reference system with the origin at the center of the square and UEs are associated to the nearest BS. All propagation channels are generated as realizations of uncorrelated Rayleigh fading, using the path-loss model in [39] with power decay factor equal to 3.5. All mobiles have the same maximum feasible power $P_{\max} = -20\,\text{dBW}$ and hardware-dissipated power $\Psi_k = -20\,\text{dBW}$. The receiver noise power is generated as $\sigma^2 = FB\mathcal{N}_0$, wherein $F = 3\,\text{dB}$ is the noise figure, $B = 180\,\text{Hz}$ is the communication bandwidth, and $\mathcal{N}_0 = -174\,\text{dBm/Hz}$ is the thermal noise power spectral density.

Fig. 1 shows the energy-efficient Pareto region of the system for $K = 2$ UEs. 200 sample points at the Pareto boundary are obtained by solving the WMEE maximization problem for 200 different choices of the weights, each corresponding to finding the outmost point in a certain search direction. The maximization problem was solved using the proposed monotonic fractional programming framework. As a comparison, we show the non-uniform grid of operating points that are achieved by a grid search over 40000 equally-spaced feasible transmit power points. We note that the monotonic fractional programming framework is able to characterize the complete region, while the 40000 points from the grid search fail to explore all parts of the region. Two particular operating points are shown as reference. The point where both UEs use full power is in the interior of the region and thus inefficient, for any choice of EE metric. The maximum GEE points found by sequential fractional programming and monotonic fractional programming coincide, and are also in the interior of the region. This is an interesting result, as it shows that GEE maximization does not necessarily yield a point on the boundary of the energy-efficient Pareto region, thus showing how the GEE metric might fail to capture the efficiency of the individual links. A similar scenario is illustrated in Fig. 2, which considers the case with $K = 3$ UEs. In this case, the Pareto region is illustrated using a number of lines that lie on the boundary. All star-marked points on each line were computed using the proposed monotonic fractional programming framework. It is interesting to observe that the obtained region does not define a convex set, which implies that in general only WMEE maximization can guarantee to characterize the energy-efficient Pareto region of wireless networks.

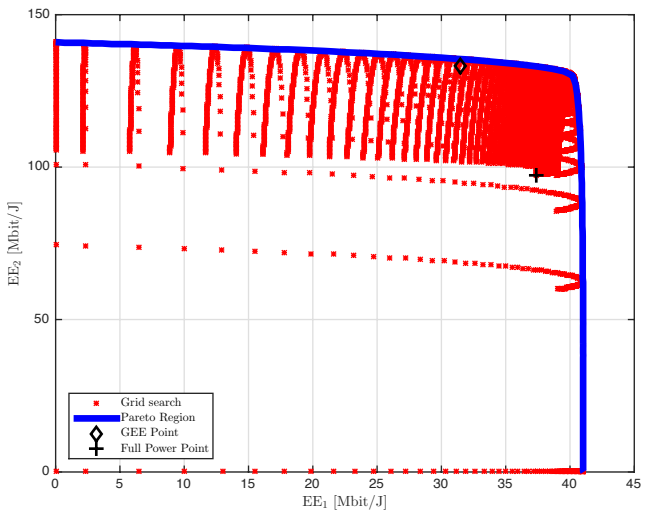

Fig. 1. Energy-efficient Pareto region for $K = 2$, generated by: 1) Monotonic Fractional Programming; 2) Grid Search. The points corresponding to the maximum GEE and to Full Power Allocation are also reported, which are both inside the Pareto region.

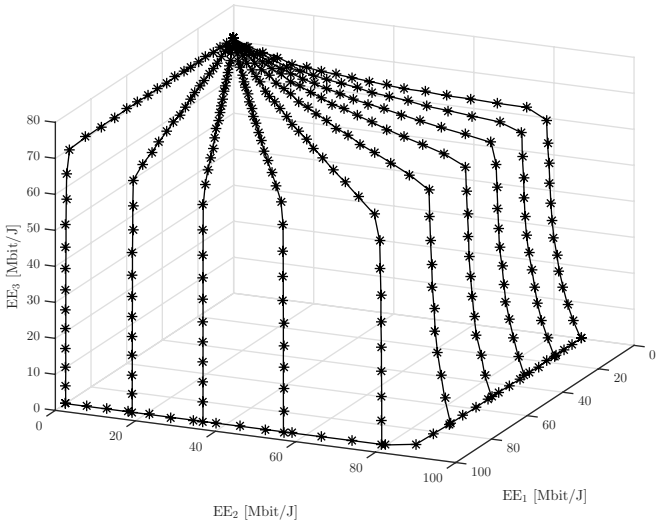

Fig. 2. Energy-efficient Pareto region for $K = 3$, generated by Monotonic Fractional Programming. The region can be seen to define a non-convex set.

B. Massive MIMO networks

Consider the uplink[3] of a massive MIMO network composed of K single-antenna UEs and L BSs (which can represent either macro cells or small cells), each equipped with N_R antennas. Each UE k is associated to a specific serving

[3]Similar results can be obtained for the downlink, but we selected the uplink to reduce the amount of notation.

BS $a(k)$, while interfering with all other UEs. Let us define by $\mathbf{h}_{k,m} \in \mathbb{C}^{N_R}$ the channel vector from UE k to BS m. Denoting by $\mathbf{c}_k \in \mathbb{C}^{N_R}$ the linear detector that BS $a(k)$ applies to its received signal to detect the signal from UE k, a lower bound[4] on the uplink SINR of UE k takes the following form [40]:

$$\gamma_k = \frac{p_k \left| \mathrm{E}\{\mathbf{c}_k^H \mathbf{h}_{k,a(k)}\} \right|^2}{p_k \mathrm{var}\{\mathbf{c}_k^H \mathbf{h}_{k,a(k)}\} + \mathbf{i}_k} \tag{58}$$

with

$$\mathbf{i}_k = \sigma^2 \mathrm{E}\{\|\mathbf{c}_k\|^2\} + \sum_{i=1, i\neq k}^{K} p_i \mathrm{E}\{|\mathbf{c}_k^H \mathbf{h}_{i,a(k)}|^2\}. \tag{59}$$

Assume that a maximum ratio combining (MRC) detector is employed. This amounts to setting $\mathbf{c}_k = \hat{\mathbf{h}}_{k,a(k)}$ where $\hat{\mathbf{h}}_{k,a(k)}$ denotes the estimate of $\mathbf{h}_{k,a(k)}$ given by

$$\mathbf{h}_{k,a(k)} = \hat{\mathbf{h}}_{k,a(k)} + \tilde{\mathbf{h}}_{k,a(k)} \tag{60}$$

with $\tilde{\mathbf{h}}_{k,a(k)}$ being the estimation error statistically independent of $\hat{\mathbf{h}}_{k,a(k)}$. We consider Rayleigh fading channels $\mathbf{h}_{k,m} \sim \mathcal{CN}(0, d_{k,m}\mathbf{I}_{N_R})$ where the variance $d_{k,m}$ accounts for the large-scale channel fading and attenuation from UE k to BS m. If a minimum mean square error (MMSE)-based channel estimation scheme is used at the BS (with full pilot reuse across cells) [40], then we have that $\hat{\mathbf{h}}_{k,a(k)} \sim \mathcal{CN}(\mathbf{0}, \rho_{k,a(k)}\mathbf{I}_{N_R})$ and $\tilde{\mathbf{h}}_{k,a(k)} \sim \mathcal{CN}(\mathbf{0}, (d_{k,a(k)} - \rho_{k,a(k)})\mathbf{I}_{N_R})$ where

$$\rho_{k,a(k)} = \frac{d_{k,a(k)}}{\tau + \sum_m d_{k,m}} \tag{61}$$

with τ being a given parameter that depends on the total pilot transmit power over the pilot sequence. Under the above assumptions, we have that

$$\gamma_k = \frac{p_k \alpha_k}{\sigma_k^2 + p_k \phi_k + \sum_{i=1, i\neq k}^{K} p_i \beta_{i,k}} \tag{62}$$

with $\alpha_k = \rho_{k,a(k)}$, $\phi_k = d_{k,a(k)} + \sum_{m \neq a(k)} \rho_{k,m}^2 / \rho_{k,a(k)}$, and $\beta_{i,k} = d_{i,a(k)}\rho_{i,a(k)} / \rho_{k,a(k)}$, which is formally equivalent to (5). Thus, the functions $q_k^+(\mathbf{p})$ and $q_k^-(\mathbf{p})$ are expressed as in (6) and (7) for all $k = 1, \ldots, K$.

Similar results can be obtained when the system is affected by hardware impairments [41], [42]; for example, unavoidable clock drifts in local oscillators, finite-precision digital-to-analog converters, amplifier non-linearities, non-ideal analog filters, and so forth. For the sake of simplicity, let us assume that the hardware impairments are only present at the UEs.[5] Following [41], hardware impairments result in a reduction of the uplink signals by a factor $1 - \epsilon^2$, with ϵ being the error vector magnitude, and in an additive Gaussian distortion noise which carries the removed useful power. In these circumstances, a lower bound of the SINR can be computed as

$$\gamma_k = \frac{p_k (1-\epsilon^2) \left| \mathrm{E}\{\mathbf{c}_k^H \mathbf{h}_{k,a(k)}\} \right|^2}{p_k (1-\epsilon^2) \mathrm{var}\{\mathbf{c}_k^H \mathbf{h}_{k,a(k)}\} + p_k \epsilon^2 \mathrm{E}\{|\mathbf{c}_k^H \mathbf{h}_{k,a(k)}|^2\} + \mathbf{i}_k} \tag{63}$$

with $\mathbf{i}_k$ given in (59). Plugging $\mathbf{c}_k = \hat{\mathbf{h}}_{k,a(k)}$ into the above equation and taking into account that in the presence of hardware impairments $\hat{\mathbf{h}}_{k,a(k)} \sim \mathcal{CN}(\mathbf{0}, \sqrt{1-\epsilon^2}\rho_{k,a(k)}\mathbf{I}_{N_R})$ and $\tilde{\mathbf{h}}_{k,a(k)} \sim \mathcal{CN}(\mathbf{0}, (d_{k,a(k)} - \sqrt{1-\epsilon^2}\rho_{k,a(k)})\mathbf{I}_{N_R})$, the SINR is easily found to be in the same form of (62).

In our numerical simulations, we consider a similar cellular setup as in Section VI-A, but with a massive antenna deployment at the BS. In particular, each BS is equipped with $N_R = 50$ antennas, while the mobiles have a single antenna. We consider the presence of hardware impairments with $\epsilon = 10^{-1}$, and of channel estimation errors with $\tau = 0.3$.

Our numerical experiments confirm that the sequential fractional programming framework performs as the monotonic fractional programming framework and thus achieves the global GEE maximum, and not a suboptimal solution. Recall that the latter framework is guaranteed to find the GEE maximum, but with a computational complexity that grows exponentially in the number of UEs. If the algorithm is initiated at $\lambda_0 = 0$, corresponding to zero GEE, it finds the sum-rate maximizing solution in the first iteration of Dinkelbach's algorithm. In our experiments, only one or two further iterations are required to converge to the global GEE optimum, which is line with the super-linear convergence rate of Dinkelbach's algorithm. At convergence, the difference between the current GEE value λ_j and the next obtained GEE value λ_{j+1} is negligible. Fig. 3 shows the behavior when λ_j equals the GEE value obtained by the sequential fractional programming framework with $K = 2$. Fig. 4 shows the corresponding behavior for $K = 3$. In both cases, the lower and upper bounds in the BRB algorithm that solves the monotonic subproblem converge to the same GEE value ($\lambda_{j+1} = \lambda_j$), which validates that the algorithm has already converged to the global maximum. The bounding behavior is very typical for the BRB algorithm [24]; a relatively small difference between the lower and upper bounds is obtained quickly, while many more iterations in the BRB algorithm are required to push the difference down to zero. Notice that 40 iterations are sufficient for $K = 2$, while the number of iterations for $K = 3$ is of the order of 10^4, which shows the exponential complexity scaling with the number of UEs.

Finally, Figs. 5 and 6 compare the GEE achieved by the monotonic fractional programming and the sequential fractional programming frameworks, versus $P_{\max}$, for $K = 2$ and $K = 3$, respectively. The GEE obtained when the sequential framework is used to maximize the sum-rate, and when all users transmit with full power (i.e. $p_k = P_{\max}$ for all k) are shown, too. The results again confirm the optimality of the sequential approach, which performs as the monotonic approach. In addition, both schemes saturate at large $P_{\max}$, because once $P_{\max}$ is large enough to allow attaining the GEE global maximum, the excess transmit power is no longer

[4]This refers to the standard worst-case lower bound on the mutual information where the uncorrelated interference is treated as Gaussian noise, which is information-theoretic optimal for small interference powers.

[5]The impact of BS hardware impairments vanishes as N_R increases [42], thus UE hardware impairments are expected to dominate in massive MIMO networks.

used. Using full power at all UEs is globally optimal up to $P_{\max} = -34\,\text{dBW}$, whereas using all the excess power at larger $P_{\max}$ will degrade the GEE. The GEE obtained by sum-rate maximization is globally optimal up to $P_{\max} = -28\,\text{dBW}$, while it decreases for larger $P_{\max}$. This shows that at least one UE should transmit at $-28\,\text{dBW}$ at the globally optimal GEE point, but not all of them.

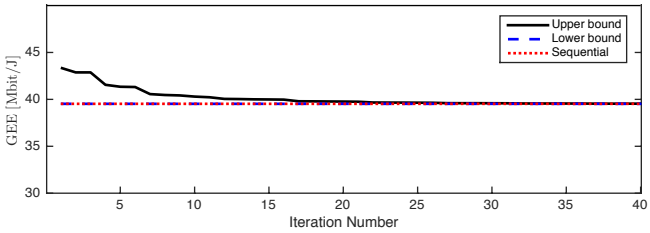

Fig. 3. Convergence behavior of the BRB algorithm, with $K = 2$, in the last iteration of Dinkelbach's algorithm. The behavior shows the convergence to the global GEE optimal solution and illustrates the optimality of the sequential fractional programming framework.

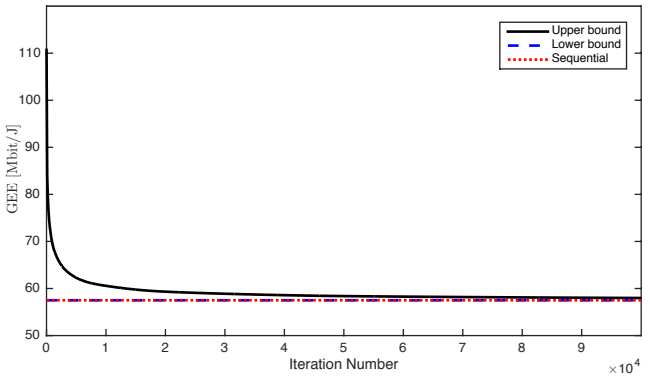

Fig. 4. Convergence behavior of the BRB algorithm, with $K = 3$, in the last iteration of Dinkelbach's algorithm. The behavior shows the convergence to the global GEE optimal solution and illustrates the optimality of the sequential fractional programming framework.

VII. Conclusion

This work has dealt with the characterization of the global solution of EE maximization problems in wireless networks. Two optimization frameworks for energy-efficient power control have been developed. The former merges the tool of monotonic optimization with the theory of fractional programming, and is guaranteed to yield the global solution of the EE maximization problems. Although still exponentially increasing with the number of links, the complexity of this approach is significantly lower than that of standard global optimization methods, because the proposed approach requires to explore only the boundary of the feasible set, while conventional global optimization methods require to search the whole feasible set of the problem. Thus, the developed monotonic fractional programming framework provides an effective way to characterize the energy-efficient Pareto region of multi-user wireless networks, as well as to benchmark practical algorithms which are not guaranteed to achieve global optimality. The latter optimization framework combines the tool

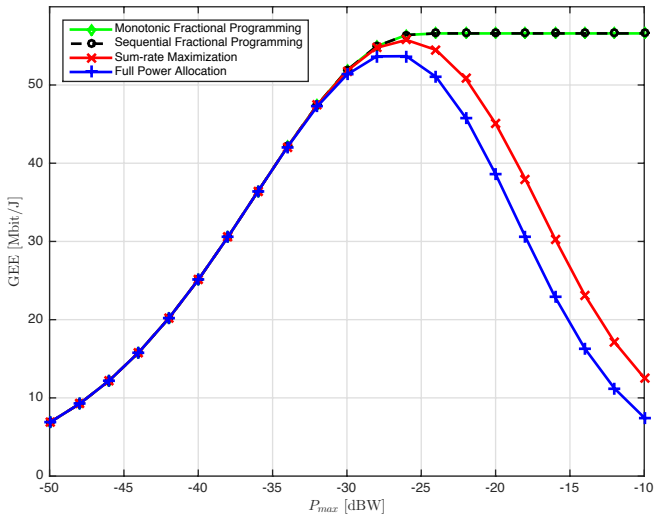

Fig. 5. Achieved GEE versus $P_{\max}$ for $K = 2$, using: 1) Monotonic fractional programming; 2) Sequential Fractional Programming, 3) Sum-rate maximization by sequential programming; 4) Full Power Allocation.

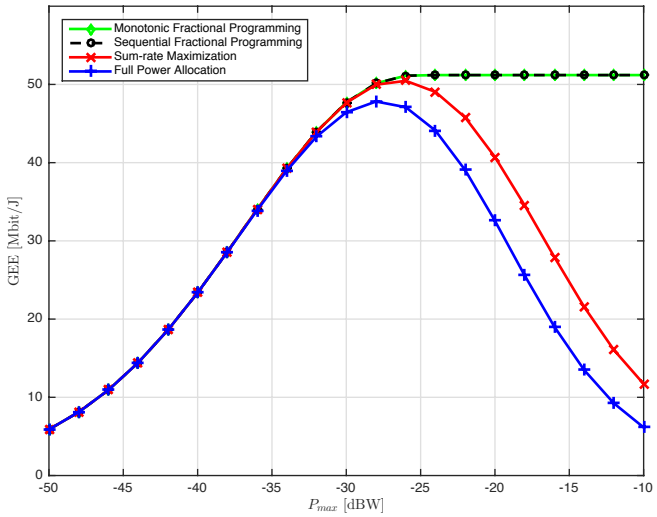

Fig. 6. Achieved GEE versus $P_{\max}$ for $K = 3$, using: 1) Monotonic fractional programming; 2) Sequential Fractional Programming, 3) Sum-rate maximization by sequential programming; 4) Full Power Allocation.

of sequential optimization with fractional programming theory. The resulting sequential fractional programming framework is guaranteed to be first-order optimal, and exhibits a polynomial-time complexity. A numerical analysis has been provided to compare the global optimal solution obtained by monotonic fractional programming, and the EE obtained by the sequential fractional programming. The results indicate that the latter virtually performs as the global optimal procedure, and that the proposed monotonic fractional programming framework indeed provides an effective way of computing the energy-efficient Pareto boundary of a communication system.

References

[1] A. Fehske, J. Malmodin, G. Biczók, and G. Fettweis, "The Global Footprint of Mobile Communications–The Ecological and Economic Perspective," *IEEE Communications Magazine, issue on Green Communications*, pp. 55–62, Aug. 2011.

[2] Ericsson White Paper, "More than 50 billion connected devices," Ericsson, Tech. Rep. 284 23-3149 Uen, Feb. 2011.

[3] "The 1000x data challenge," Qualcomm, Tech. Rep. [Online]. Available: http://www.qualcomm.com/1000x

[4] C. Isheden, Z. Chong, E. Jorswieck, and G. Fettweis, "Framework for link-level energy efficiency optimization with informed transmitter," *IEEE Transactions on Wireless Communications*, vol. 11, no. 8, pp. 2946–2957, Aug. 2012.

[5] A. Zappone and E. Jorswieck, "Energy efficiency in wireless networks via fractional programming theory," *Foundations and Trends in Communications and Information Theory*, vol. 11, no. 3-4, pp. 185–396, 2015.

[6] D. W. K. Ng, E. S. Lo, and R. Schober, "Energy-efficient resource allocation in multi-cell OFDMA systems with limited backhaul capacity," *IEEE Transactions on Wireless Communications*, vol. 11, no. 10, pp. 3618–3631, Oct. 2012.

[7] Q. Xu, X. Li, H. Ji, and X. Du, "Energy-efficient resource allocation for heterogeneous services in OFDMA downlink networks: Systematic perspective," *IEEE Transactions on Vehicular Technology*, vol. 63, no. 5, pp. 2071–2082, June 2014.

[8] J. Xu and L. Qiu, "Energy efficiency optimization for MIMO broadcast channels," *IEEE Transactions on Wireless Communications*, vol. 12, no. 2, pp. 690–701, Feb. 2013.

[9] B. Du, C. Pan, W. Zhang, and M. Chen, "Distributed energy-efficient power optimization for CoMP systems with max-min fairness," *IEEE Communications Letters*, vol. 18, no. 6, pp. 999–1002, June 2014.

[10] S. He, Y. Huang, S. Jin, and L. Yang, "Coordinated beamforming for energy efficient transmission in multicell multiuser systems," *IEEE Transactions on Communications*, vol. 61, no. 12, pp. 4961–4971, Dec. 2013.

[11] S. He, Y. Huang, L. Yang, and B. Ottersten, "Coordinated multicell multiuser precoding for maximizing weighted sum energy efficiency," *IEEE Transactions on Signal Processing*, vol. 62, no. 3, pp. 741–751, Feb. 2014.

[12] M. Chiang, C. Wei, D. P. Palomar, D. O'Neill, and D. Julian, "Power control by geometric programming," *IEEE Transactions on Wireless Communications*, vol. 6, no. 7, pp. 2640–2651, July 2007.

[13] L. Venturino, X. Wang, and M. Lops, "Multiuser detection for cooperative networks and performance analysis," *IEEE Transactions on Signal Processing*, vol. 54, no. 9, pp. 3315–3329, Sept. 2006.

[14] L. Venturino, A. Zappone, C. Risi, and S. Buzzi, "Energy-efficient scheduling and power allocation in downlink OFDMA networks with base station coordination," *IEEE Transactions on Wireless Communications*, vol. 14, no. 1, pp. 1–14, Jan. 2015.

[15] A. Zappone, E. A. Jorswieck, and S. Buzzi, "Energy efficiency and interference neutralization in two-hop MIMO interference channels," *IEEE Transactions on Signal Processing*, vol. 62, no. 24, pp. 6481–6495, Dec. 2014.

[16] A. Zappone, L. Sanguinetti, G. Bacci, E. A. Jorswieck, and M. Debbah, "Energy-efficient power control: A look at 5G wireless technologies," *IEEE Transactions on Signal Processing, to appear, http://arxiv.org/abs/1503.04609*, 2015.

[17] D. Nguyen, L.-N. Tran, P. Pirinen, and M. Latva-aho, "Precoding for full duplex multiuser MIMO systems: Spectral and energy efficiency maximization," *IEEE Transactions on Signal Processing*, vol. 61, no. 16, pp. 4038–4050, August 2013.

[18] H. Tuy, "Monotonic optimization," *SIAM Journal on Optimization*, vol. 11, no. 2, pp. 464–494, 2000.

[19] H. Tuy, F. Al-Khayyal, and P. Thach, "Monotonic optimization: Branch and cut methods," in *Essays and Surveys in Global Optimization*, C. Audet, P. Hansen, and G. Savard, Eds. Springer US, 2005.

[20] L. Qian and Y. Zhang, "S-MAPEL: Monotonic optimization for non-convex joint power control and scheduling problems," *IEEE Transactions on Wireless Communications*, vol. 9, no. 5, pp. 1708–1719, May 2010.

[21] E. Björnson, G. Zheng, M. Bengtsson, and B. Ottersten, "Robust monotonic optimization framework for multicell MISO systems," *IEEE Transactions on Signal Processing*, vol. 60, no. 5, pp. 2508–2523, May 2012.

[22] L. Liu, R. Zhang, and K. C. Chua, "Achieving global optimality for weighted sum-rate maximization in the K-User Gaussian interference channel with multiple antennas," *IEEE Transactions on Wireless Communications*, vol. 11, no. 5, pp. 1933–1945, May 2012.

[23] W. Utschick and J. Brehmer, "Monotonic optimization framework for coordinated beamforming in multicell networks," *IEEE Transactions on Signal Processing*, vol. 60, no. 4, pp. 1899–1909, April 2012.

[24] E. Björnson and E. A. Jorswieck, "Optimal resource allocation in coordinated multi-cell systems," *Now Publishers: Foundations and Trends in Communications and Information Theory*, vol. 9, no. 2-3, pp. 113–381, Jan. 2013.

[25] Y. J. Zhang, L. Qian, and J. Huang, "Monotonic optimization in communication and networking systems," *Now Publishers: Foundations and Trends in Networking*, vol. 7, no. 1, pp. 1–75, 2012.

[26] E. Björnson, L. Sanguinetti, J. Hoydis, and M. Debbah, "Optimal design of energy-efficient multi-user MIMO systems: Is massive MIMO the answer?" *IEEE Transactions on Wireless Communications*, vol. 14, no. 6, pp. 3059–3075, June 2015.

[27] E. Björnson, L. Sanguinetti, and M. Kountouris, "Deploying dense networks for maximal energy efficiency: Small cells meet massive MIMO," *IEEE J. Sel. Areas Commun.*, 2016, to appear. [Online]. Available: http://arxiv.org/abs/1505.01181

[28] E. Björnson, E. Jorswieck, M. Debbah, and B. Ottersten, "Multiobjective signal processing optimization: The way to balance conflicting metrics in 5G systems," *IEEE Signal Processing Magazine*, vol. 31, no. 6, pp. 14–23, 2014.

[29] S. Schaible, "Fractional programming," *Zeitschrift für Operations Research*, vol. 27, no. 1, pp. 39–54, 1983. [Online]. Available: http://dx.doi.org/10.1007/BF01916898

[30] J. P. Crouzeix and J. A. Ferland, "Algorithms for generalized fractional programming," *Mathematical Programming*, vol. 52, pp. 191–207, 1991.

[31] R. Jagannathan, "On some properties of programming problems in parametric form pertaining to fractional programming," *Management Science*, vol. 12, no. 7, Mar. 1966.

[32] W. Dinkelbach, "On nonlinear fractional programming," *Management Science*, vol. 13, no. 7, pp. 492–498, Mar. 1967.

[33] H. Tuy, *Convex Analysis and Global Optimization (Nonconvex Optimization and Its Applications)*. Kluwer Academic Publishers, Dordrecht, 1998.

[34] B. R. Marks and G. P. Wright, "A general inner approximation algorithm for non-convex mathematical programs," *Operations Research*, vol. 26, no. 4, pp. 681–683, 1978.

[35] S. Schaible, "Fractional programming," *Zeitschrift für Operations Theory and Applications*, vol. 27, no. 1, pp. 347–352, 1983.

[36] S. A. Jafar and A. Goldsmith, "Transmitter optimization and optimality of beamforming for multiple antenna systems," *IEEE Transactions on Wireless Communications*, vol. 3, no. 4, pp. 1165–1175, July 2004.

[37] S. P. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge Univ Press, 2004.

[38] S. Verdú, *Multiuser detection*. Cambridge Univ Press, 1998.

[39] G. Calcev, D. Chizhik, B. Goransson, S. Howard, H. Huanga, A. Kogiantis, A. Molisch, A. Moustakas, D. Reed, and H. Xu, "A wideband spatial channel model for system-wide simulations," *IEEE Transactions on Vehicular Technology*, vol. 56, no. 2, Mar. 2007.

[40] J. Hoydis, S. ten Brink, and M. Debbah, "Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?" *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 2, pp. 160–171, Feb. 2013.

[41] E. Björnson, E. G. Larsson, and M. Debbah, "Massive MIMO for maximal spectral efficiency: How many users and pilots should be allocated?" *IEEE Trans. Wireless Commun.*, 2016, to appear. [Online]. Available: http://arxiv.org/pdf/1412.7102.pdf

[42] E. Björnson, J. Hoydis, M. Kountouris, and M. Debbah, "Massive MIMO systems with non-ideal hardware: Energy efficiency, estimation, and capacity limits," *IEEE Transaction Information Theory*, vol. 60, no. 11, pp. 7112–7139, Nov. 2014.