Multiobjective Signal Processing Optimization: The way to balance conflicting metrics in 5G systems

Emil Björnson, Eduard Jorswieck, Mérouane Debbah, Björn Ottersten

MULTI-OBJECTIVE SIGNAL PROCESSING OPTIMIZATION: THE WAY TO BALANCE CONFLICTING METRICS IN 5G SYSTEMS

Emil Björnson, Eduard Jorswieck, Mérouane Debbah, and Björn Ottersten

The evolution of cellular networks is driven by the dream of ubiquitous wireless connectivity: *Any data service is instantly accessible everywhere*. With each generation of cellular networks, we have moved closer to this wireless dream; first by delivering wireless access to voice communications, then by providing wireless data services, and recently by delivering a WiFi-like experience with wide-area coverage and user mobility management. The support for high data rates has been the main objective in recent years [1], as seen from the academic focus on sum-rate optimization and the efforts from standardization bodies to meet the peak rate requirements specified in IMT-Advanced. In contrast, a variety of metrics/objectives are put forward in the technological preparations for 5G networks: higher peak rates, improved coverage with uniform user experience, higher reliability and lower latency, better energy efficiency, lower-cost user devices and services, better scalability with number of devices, etc. These multiple objectives are coupled, often in a conflicting manner such that improvements in one objective lead to degradation in the other objectives. Hence, the design of future networks calls for new optimization tools that properly handle the existence and tradeoffs between multiple objectives.

In this article, we provide a review of multi-objective optimization (MOO), which is a mathematical framework to solve design problems with multiple conflicting objectives [2–6]. In contrast to conventional heuristic approaches where some objectives are converted into constraints, MOO enables a rigorous network design. MOO has been applied in many engineering and economic related fields, but has received little attention from the signal processing and wireless communication communities. We provide a survey of the basic definitions, properties, and algorithmic tools in MOO. This reveals how signal processing algorithms are used to visualize the inherent conflicts between 5G performance objectives, thereby allowing the network designer to understand the possible operating points and how to balance the objectives in an efficient and satisfactory way. For clarity, we provide a case study on massive multiple-input multiple-output (MIMO) systems, which is one of the key enablers of 5G cellular networks.

INTRODUCTION

We are currently at a point in time when many researchers in industry and academia are trying to formalize their expectations and requirements on the next generation wireless communication networks. These views are expressed in various magazine articles, white papers, and plenary talks. To get a sense of the range of expectations, one can take a look at the project *Mobile and wireless communications Enablers for the Twenty-twenty Information Society* (METIS), `http://www.metis2020.com/`, where telecommunications manufacturers, network operators, and academic partners are gathering their 5G requirements. The following summarizes their main objectives [7]:

- Higher user data rates: 10–100 times higher average user rates are expected, at least in urban scenarios.
- Higher area data rates: 1000 times higher average rates per unit area are anticipated.
- More connected devices: With the respective expected increases in user and area rates, 10–100 times more devices can be accommodated per unit area.
- Higher energy efficiency (EE): The throughput should be improved without increasing the operational cost or the energy consumption, thus greatly improving the EE. If EE is measured as area data rate per power expenditure, this requires a 1000 times EE improvement.

Furthermore, *heterogeneity* appears as a keyword that can be tied to a variety of network aspects:

- Heterogeneous networks: The combination of access points with different ranges, traffic loads, radio access technologies, licensed/unlicensed spectrum, and hardware capabilities makes the network highly heterogeneous. The same deployment strategy cannot be used everywhere and the same resource management scheme cannot be used throughout the day.
- Heterogeneous user conditions: As the performance requirements become tighter, the mobility and pathloss of a specific user determines its quality-of-service, unless the network is designed to counteract these effects.
- Heterogeneous devices: The differences in functionality and hardware capability of user devices are expected to grow. Large handheld devices can, for example, achieve high data rates by spatial multiplexing and advanced signal processing, while small sensors seek low data rates under extremely tight energy constraints.

- Heterogeneous service requirements: Some cyber-physical systems and public-safety applications require very fast and reliable response times, while best-effort delivery is fine for other types of data services. Similarly, certain multimedia applications have tight and continuous quality-of-service requirements, while other services are bursty in nature.

There are apparently many different requirements, or *objectives*, to keep in mind when designing future wireless networks. Unfortunately, these objectives cannot be treated separately because they are coupled; sometimes in a consistent fashion, but often in conflicting ways such that improvements in one objective lead to deterioration of other objectives. This is because the same network *resources* (e.g., time, frequency, space, power, and hardware) play key roles in all these requirements/objectives, but in incompatible ways. As a simple example, higher peak user rates can be achieved by using more power (which affects the EE), allocating more transmission resources to users with good channels (which means less uniform user experience and higher latencies), or making use of intricate signal processing algorithms (which increases the complexity and cost of user devices).

In order to achieve the ambitious 5G goals, efficient network operation with respect to *all* the conflicting 5G objectives is required. This calls for a design framework that handles multiple objectives and supports the search for the best attainable operating point. But can we really formulate and solve multi-objective problems rigorously or is heuristic trial-and-error the only option? Is there even any optimal solution? These are questions that we address in this article.

CONVENTIONAL SINGLE-OBJECTIVE OPTIMIZATION

The conventional approach to physical-layer system optimization is that of selecting a scalar *network utility function* that is maximized under a set of constraints [8, 9]. A common problem formulation is that of maximizing the weighted sum of the users' data rates under transmit power constraints [6, 10, 11]. Alternatively, one can minimize the transmitted power under the constraint of guaranteeing certain data rates to each user [12, 13]. In recent years, the EE (in bit/Joule) has also arisen as a utility function [14–16].

In essence, the conventional approach is to select one of the objectives listed above as the sole objective, while the other objectives are transformed into constraints. The inherent heuristic assumptions are: 1) one of the objectives is of dominating importance; and 2) it is known beforehand what are good values for the constraints related to the other objectives. Moreover, the short-term values of the different objectives are usually considered in these network utility problems and not the long-term values which are of main importance in the network design. Given the increased complexity due to heterogeneity, the need for long-term network optimization, and the diverse expectations on 5G networks, the conventional approach is no longer viable. However, we show later how to construct more appropriate single-objective problems.

NEW PARADIGM: MULTI-OBJECTIVE OPTIMIZATION

Instead of assuming that one of the objectives is the sole objective, the fundamental approach is to recognize the existence of multiple objectives [2]: $g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots, g_M(\mathbf{x})$ where M is the number of objectives. These objective functions can, for example, be area throughput, guaranteed rates for different classes of users, number of simultaneous users, energy efficiency, etc. Explicit examples are given later in this article, while the theory is applicable for any arbitrary functions. The notation $\mathbf{g}(\mathbf{x}) = [g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots, g_M(\mathbf{x})]^T$ is used to emphasize that the objective is vector-valued.

The available resources (e.g., time, frequency, space, power, and hardware) are modeled by a compact set $\mathcal{X} \subset \mathbb{R}^D$, which is called the *resource bundle* and has any finite dimension D. Each vector $\mathbf{x} \in \mathcal{X}$ represents a feasible way of utilizing the network resources. The satisfaction of this resource utilization equals $g_m(\mathbf{x}) \in \mathbb{R}$ with respect to the mth objective function. A larger value corresponds to higher satisfaction. For tractability we assume that $g_m(\mathbf{x})$ is a bounded continuous function of $\mathbf{x}$ and non-negative. We also assume that it exists a point $\mathbf{x}_0 \in \mathcal{X}$ such that $g_m(\mathbf{x}_0) = 0$ for all m. This operating point is the dissatisfaction of turning off the network and makes the satisfaction (for each objective) become a number from zero and upwards. Not all practical objectives satisfy these conditions by nature; for example, latency and error probability are typically to be minimized. However, there are standard transformations that reformulate such metrics into objective functions in our framework [3–6].

A key assumption is that the M objectives are *not* ordered and therefore studied without any preconceptions—all doors are kept open. In contrast to game theory, where each objective belongs to one of the competing agents, we assume that there is a *network designer* that would like to design the network to maximize *all* the M objectives simultaneously:

$$\begin{aligned} \underset{\mathbf{x}}{\text{maximize}} \quad & \mathbf{g}(\mathbf{x}) = [g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots, g_M(\mathbf{x})]^T \\ \text{subject to} \quad & \mathbf{x} \in \mathcal{X}. \end{aligned} \tag{1}$$

Note that (1) is the maximization of the vector $\mathbf{g}(\mathbf{x})$ containing the M objectives, which is defined as maximizing all elements simultaneously. This is known as a *multi-objective optimization problem (MOOP)* or, alternatively, as a multi-criteria or vector optimization problem [2–6]. These types of problems arise in many engineering fields because of the difficulty to find a scalar metric that exactly describes what we would like to achieve. We review the main concepts and properties related to MOOPs in this article. We provide the basic tools to understand the structure of MOOPs and how to solve these problems in practice. The properties are stated without proofs, while we recommend [3–5] for further details and [6] for a recent survey aimed at communication applications.

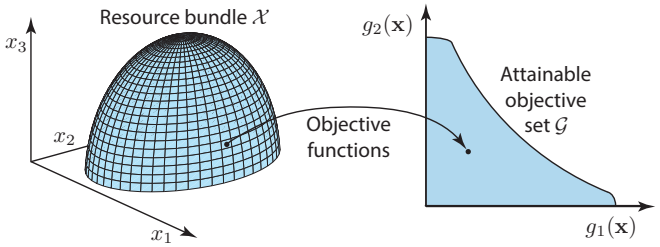

Fig. 1. Illustration of a MOOP with a three-dimensional resource bundle $\mathcal{X}$ and a two-dimensional attainable objective set. For each resource utilization $\mathbf{x} = [x_1\, x_2\, x_3]^T \in \mathcal{X}$, the objective functions $g_1(\mathbf{x})$ and $g_2(\mathbf{x})$ assign a vector $\mathbf{g}(\mathbf{x}) \in \mathcal{G}$.

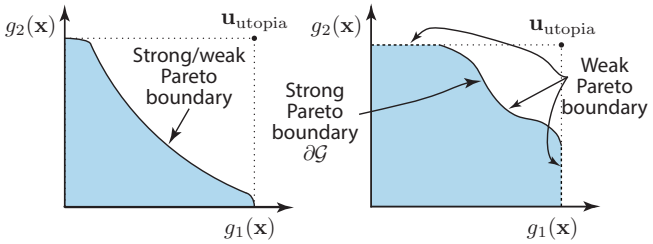

Fig. 2. Illustration of the Pareto boundary, which is either the complete upper boundary of $\mathcal{G}$ (left) or a subset of the upper boundary (right). The unattainable utopia point is also shown.

Property 1. The M objectives in (1) are conflicting and since there is no total order of vectors, there is (generally) no global optimum to the MOOP in (1).

This is the first important insight from the multi-objective framework; we cannot solve (1) in any globally optimal sense because there are *only subjectively optimal solutions*. Therefore, we turn the attention to the attainable objective set

$$\mathcal{G} = \{\mathbf{g}(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\} \tag{2}$$

which contains all the combinations of objective values $g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots, g_M(\mathbf{x})$ that are simultaneously attainable under the available resources. The relationship between the resource bundle $\mathcal{X}$ and the attainable objective set $\mathcal{G}$ is visualized in Fig. 1. Note that the origin is always in the objective set, $\mathbf{0} = [0\ \ldots\ 0]^T \in \mathcal{G}$, due to the assumptions above.

When formulating the MOOP, the resource bundle $\mathcal{X}$ is selected to minimize the preconditions made on the utilization of network resources. This keeps all the options open, because it is generally difficult to articulate the network requirements *a priori*—at least in a strict mathematical sense. Nevertheless, the resource bundle can include certain fundamental network performance constraints (e.g., that the M metrics should be better than in previous network generations).

PARETO OPTIMAL OPERATING POINTS

The shape of the attainable objective set $\mathcal{G}$ depends on the objective functions and the resource bundle $\mathcal{X}$, but it is usually a compact set with the property that $\mathbf{g} \in \mathcal{G}$ implies $c\mathbf{g} \in \mathcal{G}$ for all $c \in [0, 1]$ (i.e., the performance can be uniformly degraded). The set $\mathcal{G}$ can be convex or non-convex. Although Property 1 expresses that there is no global optimum, most points in $\mathcal{G}$ are strictly suboptimal. In fact, any point in the interior of $\mathcal{G}$ can be discarded because there exist other points in $\mathcal{G}$ that are more preferable with respect to all M objectives. The remaining points belong to the Pareto boundary.

Definition 1 (Pareto boundary). The strong Pareto boundary, $\partial\mathcal{G}$, consists of all points $\mathbf{g} \in \mathcal{G}$ for which there does not exist any $\mathbf{g}' \in \mathcal{G} \setminus \{\mathbf{g}\}$ with $g'_m \geq g_m$ for $m = 1, \ldots, M$.

The strong Pareto boundary consists of the attainable operating points that cannot be objectively dismissed, because none of the objectives can be improved without degrading other objectives. Evidently, any point that is not on the strong Pareto boundary is suboptimal because there exist other operating points that are better or at least as good for *every* objective. The strong Pareto boundary is as close to global optimality as one can get in multi-objective optimization; the operating points in $\partial\mathcal{G}$ are mutually unordered and can only be compared by subjective means. Each point $\mathbf{g} \in \partial\mathcal{G}$ describes a particular tradeoff between the M objectives. Hence, the Pareto boundary describes the set of (Pareto) efficient potential operating points from which we, as network designers, should select the one that is subjectively preferable to us.

The strong Pareto boundary is a subset of the upper boundary of $\mathcal{G}$. The complete upper boundary is referred to as the *weak* Pareto boundary and also contains points were some of the objectives (but not all) can be improved without degrading other objectives. This is illustrated in Fig. 2, where the strong Pareto boundary either equals the complete upper boundary (left set) or is a strict subset thereof (right set). Fig. 2 also shows the *utopia point*, which is defined as

$$\mathbf{u}_{\text{utopia}} = [u_1\ \ldots u_M]^T = \begin{bmatrix} \max_{\mathbf{x}\in\mathcal{X}} f_1(\mathbf{x}) \\ \vdots \\ \max_{\mathbf{x}\in\mathcal{X}} f_M(\mathbf{x}) \end{bmatrix}. \tag{3}$$

This is the ideal operating point that simultaneously maximizes all M objectives. If $\mathbf{u}_{\text{utopia}} \in \mathcal{G}$, the MOOP is trivial because the strong Pareto boundary consists of only the utopia point, $\partial\mathcal{G} = \{\mathbf{u}_{\text{utopia}}\}$, and it is the unique global optimum.

Property 2. Any MOOP with multiple conflicting objective functions is nontrivial in the sense that $\mathbf{u}_{\text{utopia}} \notin \mathcal{G}$ and, consequently, there is no global optimum.

Single-objective optimization problems are MOOPs with $M = 1$ and are thus trivial from the MOO perspective.

Since the Pareto boundary consists of all tentative effective operating points, we need to find the network parameters (i.e., the resource utilizations) that attain these points.

Definition 2 (Pareto optimal point). A point $\mathbf{x}^* \in \mathcal{X}$ in the resource bundle is a Pareto optimal point if $\mathbf{g}(\mathbf{x}^*) \in \partial\mathcal{G}$.

The mapping from a Pareto optimal point $\mathbf{x}^*$ to the Pareto boundary is given by the vector-valued multi-objective function $\mathbf{g}(\mathbf{x}^*)$ and is, hopefully, given in closed form. The inverse mapping is, on the other hand, hard to derive in most

cases. The multi-objective function might not be bijective, which means that multiple points in $\mathcal{X}$ can give exactly the same objective point. This happens frequently when transmitting from multi-antenna arrays, where the beamforming coefficients are only unique up to a common phase rotation [6].

SOLVING A MOOP BY VISUALIZATION

In practice, we would like to go beyond the Pareto boundary and actually *solve* the MOOP, in the sense of selecting a *single* Pareto optimal point $\mathbf{x}^*$ and its corresponding operating point $\mathbf{g}(\mathbf{x}^*) \in \partial\mathcal{G}$. To this end, we need to bring in the subjective preference of the network designer to compare different operating points at the Pareto boundary. This is not as simple as it might seem, because neither the Pareto boundary $\partial\mathcal{G}$ nor the objective set $\mathcal{G}$ are known beforehand. Simple closed-form expressions are seldom available. In fact, one needs to spend considerable computational resources on learning the objective set. For example, one can characterize $\mathcal{G}$ by computing a discrete set of sample points, which enables the network designer to visualize the different possibilities and make an informed decision. This is known as the *a posteriori method*, because the network designer formulates its subjective preference after the numerical computations have taken place [2].

We describe two approaches to compute sample points:

1. Traverse the resource bundle $\mathcal{X}$ by computing $\mathbf{g}(\mathbf{x})$ over a finite grid of $\mathbf{x} \in \mathcal{X}$. For example, if $0 \leq x_m \leq 1$ then we can limit ourselves to the 6 discrete values $x_m \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. If the same number of discrete values are taken for all D resource variables in $\mathcal{X}$, we have 6^D grid points to consider.

2. Traverse the strong Pareto boundary $\partial\mathcal{G}$ by searching for the outermost point in $\mathcal{G}$ in different directions. The search directions can, for example, be represented by vectors $\mathbf{v} = [v_1 \ \dots \ v_M]^T$ that point out (nonnegative) geometric directions from the origin (recall that $\mathbf{0} \in \mathcal{G}$ by definition). Each search corresponds to solving the single-objective optimization problem

$$\begin{aligned} \underset{\mathbf{x},\lambda}{\text{maximize}} \quad & \lambda \\ \text{subject to} \quad & g_m(\mathbf{x}) \geq \lambda v_m, \quad m = 1, \dots, M, \\ & \mathbf{x} \in \mathcal{X}, \end{aligned} \tag{4}$$

   which is a referred to as a weighted Chebyshev problem in the MOO literature [3–6] (in fact, it is the epigraph form of it [17]). If λ^* is the optimal value for a given $\mathbf{v}$, we can be sure that $\lambda^*\mathbf{v} \in \mathcal{G}$ and that this point lies on the weak Pareto boundary (upper boundary). If needed, one can guarantee to attain the strong Pareto boundary by slightly modifying (4); see [3]. By solving (4) for a finite set of search directions (e.g., equally spaced in the angular sense), one can obtain a set of sample points that characterizes the weak/strong Pareto boundary.

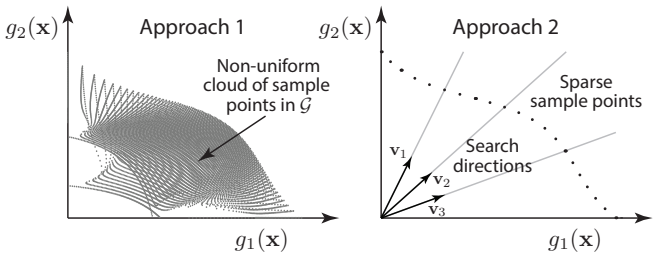

Fig. 3. Illustration of the two approaches to visualize the objective set $\mathcal{G}$ by computing sample points.

These two approaches have their respective pros and cons. The first approach is computationally efficient, assuming that the function values $\mathbf{g}(\mathbf{x})$ are easy to evaluate. The main limiting factor might be the memory storage, since the number of samples scales exponentially with D. Extensive postprocessing might also be required because most sample points will be in the interior of $\mathcal{G}$ and can be discarded since there are other samples that are better with respect to all M objectives. The resource bundle can sometimes be parameterized more efficiently by exploiting the objective functions. This can be used to improve the resolution of the objective set $\mathcal{G}$ using fewer samples. For example, transmit beamforming can be represented by one parameter per user [6, Section 3.2], which removes redundancy in multi-antenna wireless communications where the number of beamforming coefficients equals the number of users *times* the number of transmit antennas.

The second approach guarantees a high resolution because every sample point lies on the weak Pareto boundary. The downside is the computational complexity, which is proportional to the complexity of solving the search problem in (4). Indeed, this approach can only be utilized if there is a tractable way of solving (4). This is the case whenever there exists an efficient way to make a membership test; that is, to determine if a given point $\tilde{\mathbf{g}} \in \mathbb{R}^M$ belongs to the objective set or not. We elaborate on this in the box "Finding the Pareto Boundary by Bisection" below.

Fig. 3 illustrates the two approaches. The first approach gives a cloud of sample points that provides a sense of the shape of $\mathcal{G}$: is it convex, what are the numerical ranges, and are the objectives strongly/weakly conflicting? The density of points is non-uniform and it is not guaranteed that any sample point is exactly on the Pareto boundary. In contrast, the second approach gives a sparse set of sample points that are exactly on the Pareto boundary. Each point is found by searching in a certain direction (e.g., $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$) from the origin.

By looking at visualizations of the Pareto boundary, such as the ones in Fig. 3, the network designer can understand the fundamental properties and tradeoffs between conflicting objectives. Visualization is a powerful tool that supports the network designer in making an informed decision. This is the essence of the *a posteriori* method. Since it is difficult to visualize more than three dimensions at a time, one needs to limit the granularity to a few objectives at a time. This issue can be treated in an iterative fashion where the network

designer makes preliminary decisions (e.g., regarding the preferred minimal level for different objectives) which replaces the current resource bundle $\mathcal{X}$ with a smaller set $\tilde{\mathcal{X}} \subset \mathcal{X}$. This interactive process continues until the network designer is satisfied—a type of psychological convergence [4].

FINDING THE PARETO BOUNDARY BY BISECTION

The single-objective optimization problem in (4) finds the weak Pareto boundary in the direction $\mathbf{v}$ from the origin. This problem can be solved by checking if a series of points, each denoted $\boldsymbol{\mu} = [\mu_1 \ \dots \ \mu_M]^T \in \mathbb{R}^M$, belong to the attainable objective set $\mathcal{G}$ or not. This is determined by the *membership test*

$$\begin{aligned} \text{find} \quad & \mathbf{x} \in \mathcal{X} \\ \text{subject to} \quad & g_m(\mathbf{x}^*) = \mu_m. \end{aligned} \tag{5}$$

The complexity of this feasibility problem is a baseline for other optimization problems that involve the same resource bundle and objective functions—if the membership test is computationally intractable, there is little chance that any meaningful problem formulation is practically solvable. Fortunately, there are many cases when the membership test is efficiently solvable; for example, it is a convex problem in many beamforming design problems for cellular networks [6].

Equipped with a tractable membership test, we can solve (4) by first defining a range $[\lambda_{\min}, \lambda_{\max}]$ of values for λ, such that $\lambda_{\min}\mathbf{v} \in \mathcal{G}$ and $\lambda_{\max}\mathbf{v} \notin \mathcal{G}$. The lower limit can be $\lambda_{\min} = 0$, since the origin is always attainable. The upper limit is selected for the MOOP at hand, for example, by exploiting the utopia point (if it is known) or by relaxing the problem to find other unattainable points. The following algorithm solves (4):

Input: Range $[\lambda_{\min}, \lambda_{\max}]$ and accuracy $\epsilon > 0$
while $\lambda_{\max} - \lambda_{\min} > \epsilon$ **do**
  Make membership test (5) for $\boldsymbol{\mu} = \frac{\lambda_{\max}+\lambda_{\min}}{2}\mathbf{v}$
  if $\boldsymbol{\mu} \in \mathcal{G}$ **then**
    $\lambda_{\min} \leftarrow \frac{\lambda_{\max}+\lambda_{\min}}{2}$
  else
    $\lambda_{\max} \leftarrow \frac{\lambda_{\max}+\lambda_{\min}}{2}$
  end if
end while
Output: Attainable point $\mathbf{a} = \lambda_{\min}\mathbf{v}$

This is a classical bisection algorithm that cuts the range $[\lambda_{\min}, \lambda_{\max}]$ in half in each iteration [17]. Bisection has fast convergence and the distance between $\mathbf{a}$ and the Pareto boundary is below $\epsilon\|\mathbf{v}\|$ for given $\epsilon > 0$.

SOLVING A MOOP BY SCALARIZATION

An alternative way to solve MOOPs in practice is the *a priori method* where the network designer articulates preferences before any computations take place. The purpose is to find

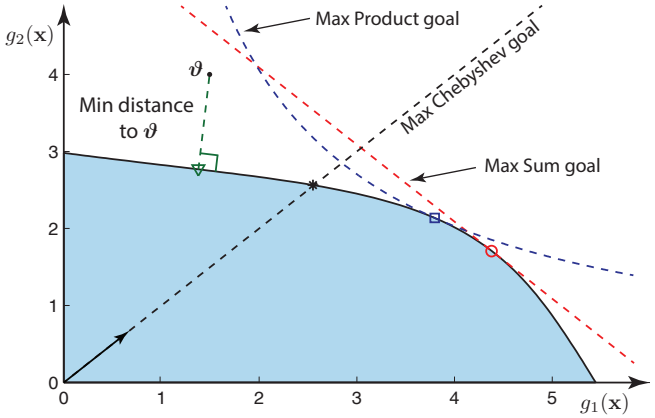

Fig. 4. Illustration of the Pareto optimal operating points achieved by scalarization using common goal functions.

the operating point $\mathbf{g} \in \mathcal{G}$ that satisfies these preferences as well as possible. In particular, the designer can specify a *goal function* $f : \mathbb{R}^M \to \mathbb{R}$ that for any conceivable operating point $\mathbf{g}$ (attainable or not) produces a scalar describing how preferable that point is (large value means high preference). The goal function describes a certain subjective tradeoff between the objectives and thus imposes an order on the vectors in the objective set $\mathcal{G}$. Consequently, the MOOP in (1) is converted into the single-objective optimization problem

$$\begin{aligned} \underset{\mathbf{x}}{\text{maximize}} \quad & f\big(g_1(\mathbf{x}), g_2(\mathbf{x}), \dots, g_M(\mathbf{x})\big) \\ \text{subject to} \quad & \mathbf{x} \in \mathcal{X}. \end{aligned} \tag{6}$$

This conversion is called *scalarization* and the solution is a weak, and usually also strong, Pareto boundary point. In contrast to the conventional approach of having a sole performance objective and expressing other potential objectives as constraints, (6) combines the M objectives into a scalar goal function and has no additional constraints. It is indeed possible to impose constraints on the acceptable values for certain objectives also in the scalarization case, but it is not required.

The goal function can take many forms and a variety of classes of functions can be found in the literature; see [3–6]. We describe four important goal function classes. The most common goal function might be the weighted sum

$$f_{\text{sum}}(\cdot) = \sum_{m=1}^{M} w_m g_m(\mathbf{x}) \tag{7}$$

where $w_1, \dots, w_M$ are positive weights that specify the priority of each objective; the priority of the mth objective grows by increasing the corresponding weight w_m. One should be careful when interpreting the relative priorities, because the objectives can have different scales, units, and couplings.

Similarly, one can consider the weighted product

$$f_{\text{product}}(\cdot) = \prod_{m=1}^{M} \big(g_m(\mathbf{x})\big)^{w_m} \tag{8}$$

where the weights are defined as before but act differently. Note that (8) is the (weighted) geometric mean, while (7) is

the (weighted) arithmetic mean. Generally speaking, the geometric mean is better at comparing objectives with different numerical ranges, because the relative scaling has no impact.

The weighted Chebyshev formulation, also known as the weighted max-min formulation, played a key role when we computed sample points on the Pareto boundary in the *a posteriori* method. The weighted Chebyshev goal function is

$$f_{\text{chebyshev}}(\cdot) = \min_{1 \leq m \leq M} \frac{g_m(\mathbf{x})}{w_m}. \tag{9}$$

This scalarization is equivalent to (4) if we write it on epigraph form [17] and select the weights $w_1, \ldots, w_M$ as $w_m = v_m$ for all m. Hence, this scalarization searches for the Pareto boundary in the direction $[w_1 \ \ldots \ w_M]^T$ from the origin.

Alternatively, the network designer can specify a preferable operating point $\boldsymbol{\vartheta} \in \mathbb{R}^M$ (e.g., the utopia point $\boldsymbol{\vartheta} = \mathbf{u}_{\text{utopia}}$). The *distance goal function* is defined as

$$f_{\text{distance}}(\cdot) = -\|\boldsymbol{\vartheta} - \mathbf{g}(\mathbf{x})\| \tag{10}$$

and measures the distance from the preferable point in some appropriately selected norm $\|\cdot\|$. The norm $\|\boldsymbol{\vartheta} - \mathbf{g}(\mathbf{x})\|$ should be small (preferably zero), thus the negative sign is used to achieve a goal function that is to be maximized.

The final operating point is determined by the choice of goal function. Interestingly, the computational complexity also varies with the goal function; the scalarized problem in (6) may be convex (i.e., solvable in polynomial time) for some classes of functions, while other classes give non-convex problems with exponential complexity—or even worse. For example, [11] proved that transmit beamforming optimization in cellular networks is (quasi-)convex for the weighted Chebyshev goal function and strongly NP-hard for most other goal functions. This result has general implications.

Property 3. The weighted Chebyshev goal function is the safest choice in terms of computational complexity; if there exists a tractable membership test, it can be solved efficiently as described in "Finding the Pareto Boundary by Bisection".

Since goal functions are inherently subjective, no choice is better than the others in terms of optimality. Property 3 inspired [6] to propose what is known as *the pragmatic approach to resource allocation*: select the weighted Chebyshev goal function (due to its tractable complexity) and exploit the weights to adapt to the needs of the network designer.

The operating points attained by different scalarizations are illustrated in Fig. 4, for a scenario where the attainable ranges are different for the two objectives. The goal functions in (7)–(10) are considered for $w_1 = w_2 = 1$. Let f^* denotes the optimal function value in (6), which of course is different for each goal function. The optimal operating point with the sum goal function lies on the level curve $f_{\text{sum}}(\cdot) = f^*$, which is the red line in Fig. 4. Similarly, $f_{\text{product}}(\cdot) = f^*$ gives the blue parabolic level curve of the product goal function. These level curves touch the objective set $\mathcal{G}$ in unique Pareto boundary points, which are the optimal operating points for the respective scalarized problems. As described earlier, the Chebyshev goal function searches on a line from the origin. For $w_1 = w_2 = 1$ this is the line where the two objectives have equal values. If there is a preferable operating point $\boldsymbol{\vartheta} \notin \mathcal{G}$ as in Fig. 4, (6) provides the operating point that minimizes the distance to $\mathcal{G}$ (the Euclidean distance is used in Fig. 4).

The function classes in (7)–(9) are parameterized by the weights $\mathbf{w} = [w_1, \ldots, w_M]^T$. Different weight selections give different Pareto optimal points when solving (6). By varying $\mathbf{w}$ over the set $\mathcal{W} = \{\mathbf{w} : w_m \geq 0 \ \forall m, \ \sum_m w_m = 1\} \in \mathbb{R}^M$ of positive weights that sum up to one, we can attain the whole Pareto boundary or a subset thereof, depending on the function class [4]. Since each scalarization in (6) is a single-objective optimization problem, it is equipped with conventional Karush-Kuhn-Tucker (KKT) optimality conditions [17]. By considering all $\mathbf{w} \in \mathcal{W}$, these can be extended to a joint set of optimality conditions for all points achieved by the function class [5]. These optimality conditions describe the structure of the resource utilizations that achieve the Pareto boundary; for example, it was utilized in [6, Section 3.2] to parameterize any efficient transmit beamforming.

Finally, we note that game theory provides an alternative way to select operating points from the Pareto boundary, by specifying the rules of a game instead of a goal function [18]. These techniques are mainly for systems with separate objectives that compete for shared resources, while single-operator networks typically have dedicated resources.

CASE STUDY: DESIGNING MASSIVE MIMO SYSTEMS

We exemplify the usefulness of MOO by a case study. The goal is to visualize tradeoffs between conflicting 5G objectives and describe how the framework can be used to acquire new insights and prove old heuristic observations. In recent years, coordinated multipoint (CoMP) techniques have shown the potential to greatly improve the area rates in cellular networks. This is achieved by deploying antenna arrays at base stations (BSs) and apply a coordinated space division multiple access (SDMA) scheme across the network [6, 19–21]. Unfortunately, CoMP is difficult to implement since the coordination signaling is limited [22], the signal processing complexity increases drastically [11], and the performance gains are not robust to the inter-user interference caused by having imperfect channel state information (CSI) [20].

The concept of massive MIMO has gained traction since it might eliminate the CoMP issues listed above [23–26]. Massive MIMO is based on the idea of deploying large arrays with unconventionally many active antennas at the BSs and serve a much smaller number of users; for example, hundreds of antennas that serve several tens of users. One would imagine that adding more antennas and users into a system would make CoMP even more difficult to implement, but the beauty of massive MIMO is that this is not the case [23]. The excessive number of antennas brings robustness to imperfect

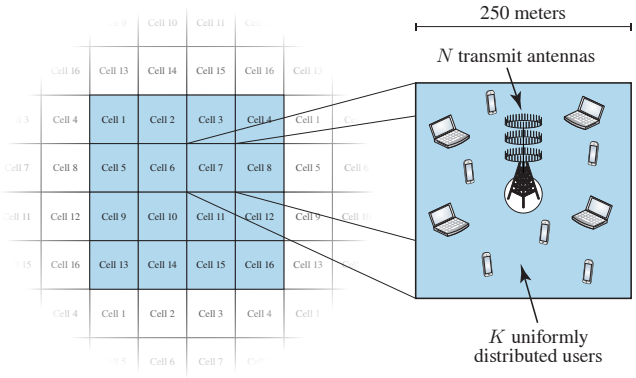

Fig. 5. Illustration of the scenario in the case study: A cellular network with N antennas per BS and K users per cell.

CSI, makes low-complexity signal processing close to optimal [24], and allows for simple implicit intercell coordination [25]. Massive MIMO systems are even robust to the distortions caused by hardware imperfections [26].

In this case study, we strive to optimize the downlink transmission of a massive MIMO system to balance $M = 3$ conflicting objectives: high average user rates, high average area rates, and high energy efficiency. The cellular network that we consider has 16 cells, each consisting of a BS with N antennas and K single-antenna users. The bandwidth is $B = 10\,\text{MHz}$, the emitted power per BS is denoted P Watt (W), and $\sigma^2 = 10^{-13}\,\text{W}$ is the average noise power.

Each cell is a square of size 250×250 meters (i.e., the area is $A = 0.25^2\,\text{km}^2$) and we apply classic wrap-around to avoid edge effects; the scenario is shown in Fig. 5. The K users are uniformly distributed in the cell, with a minimum distance of 35 meters. For a randomly picked user, let λ_{serving} be the channel variance from the serving BS and $P\lambda_{\text{intercell}}$ be the average intercell interference power. We are concerned with average behaviors and define the expectations $\Lambda_1 = \mathbb{E}\{\frac{1}{\lambda_{\text{serving}}}\}$ and $\Lambda_2 = \mathbb{E}\{\frac{\lambda_{\text{intercell}}}{\lambda_{\text{serving}}}\}$ for later use. Using the same 3GPP pathloss model as in [16], we get $\Lambda_1 = 1.72 \cdot 10^9$ and $\Lambda_2 = 0.54$.

The optimization/resource variables in this case study are the number of BS antennas N, the number of users K, and the transmit power P per cell. The resource bundle is

$$\mathcal{X} = \left\{ [K\ N\ P]^T : \begin{array}{r} 1 \leq K \leq \frac{N}{2}, \\ 2 \leq N \leq N_{\max}, \\ 0 \leq P \leq NP_{\max} \end{array} \right\} \qquad (11)$$

where $N_{\max} = 500$ is the maximal number of antennas that can fit at each BS, $P_{\max} = 20\,\text{W}$ is the maximal emitted power per BS antenna, and the constraint $K \leq \frac{N}{2}$ makes sure that we have many more BS antennas than active users.

Next, we define the average user rate and the total power consumption per cell. For simplicity, we assume that each BS has obtained perfect CSI for its users and applies zero-forcing precoding, which is a signal processing technique that cancels out intracell interference by beamforming and adapts the power allocation to guarantee the same rate to each user. Similar to [16], the average user rate can be shown to be

$$R_{\text{average}} = B\left(1 - \frac{K}{\Upsilon}\right)\log_2\left(1 + \frac{\frac{P}{K}(N-K)}{\sigma^2\Lambda_1 + P\Lambda_2}\right), \qquad (12)$$

under the assumption that each user knows its useful channel and treats intercell interference as noise. The prelog-factor $(1 - \frac{K}{\Upsilon})$ accounts for the necessary overhead for channel acquisition and $\Upsilon = 1000$ is the number of channel uses that the channel stays fixed. It is selected as $\Upsilon = B_{\text{coherence}}\tau_{\text{coherence}}$, where $B_{\text{coherence}} = 200\,\text{kHz}$ is the coherence bandwidth and $\tau_{\text{coherence}} = 5\,\text{ms}$ is the coherence time. Looking inside the logarithm of (12), $\frac{P}{K}$ is the average transmit power per user, $N - K$ is the effective array gain, and $\sigma^2\Lambda_1 + P\Lambda_2$ is the average degradation from noise and intercell interference.

Based on the models and the practical numbers in [16, 27, 28], the total power consumption per cell is given by

$$P_{\text{total}} = \frac{P}{\eta} + NC_N + KC_K + \frac{C_{\text{precoding}}}{L} + C_0 \qquad (13)$$

where $\eta = 0.31$ is the efficiency of the power amplifiers at the BS, $C_N = 1\,\text{W}$ is the hardware power consumed per transmit antenna, $C_K = 0.3\,\text{W}$ is the hardware power per user, and $C_0 = 10\,\text{W}$ is the static hardware power. In addition, $C_{\text{precoding}} = 3K^2N\frac{B}{T}$ is the floating-point operations per second (flops) required to compute zero-forcing precoding, while $L = 12.8\,\text{Gflops/W}$ is a typical computational efficiency.

We are now ready to define our three objective functions:

$$g_1(\mathbf{x}) = R_{\text{average}} \qquad [\text{bit/s/user}] \qquad (14)$$

$$g_2(\mathbf{x}) = \frac{K}{A}R_{\text{average}} \qquad [\text{bit/s/km}^2] \qquad (15)$$

$$g_3(\mathbf{x}) = \frac{KR_{\text{average}}}{P_{\text{total}}} \qquad [\text{bit/J}] \qquad (16)$$

where $\mathbf{x} = [K\ N\ P]^T$ are the optimization/resource variables. The objective $g_1(\mathbf{x})$ is the average user rate, $g_2(\mathbf{x})$ is the average area rate, and $g_3(\mathbf{x})$ is the energy efficiency.

DESIGNING MASSIVE MIMO BY MOO FRAMEWORK

We have now defined a MOOP of the type in (1). The resource bundle is given by (11) and the three objectives are defined in (14)–(16). We now describe how the MOO framework can be used to study tradeoffs between these objectives, with the purpose of deriving new insights and confirming old beliefs.

The tradeoff between the average user rate and the EE is shown in Fig. 6. The objective set with respect to these two objectives was generated by the second approach described earlier (i.e., searching for the Pareto boundary in different directions). Fig. 6 shows that these two objectives are aligned up to the point $g_1 = 20.4\,\text{Mbit/s/user}$ and $g_3 = 11.1\,\text{Mbit/J}$, where the maximal EE is achieved. The objectives are then conflicting, because the user rates can only be further increased by making drastic sacrifices in the EE.

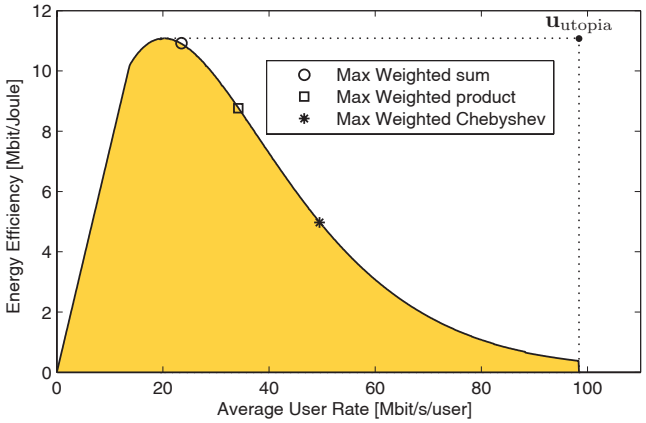

Fig. 6. Visualization of the tradeoff between two objectives in the case study: average user rate and energy efficiency.

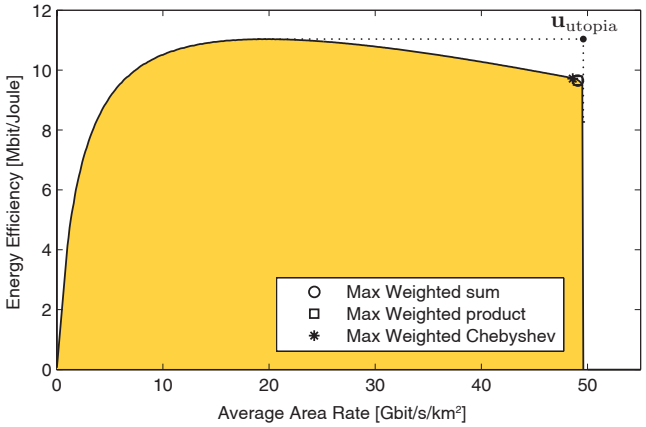

Fig. 7. Visualization of the tradeoff between two objectives in the case study: average area rate and energy efficiency.

Another tradeoff is illustrated in Fig. 7, where the average area rate and the EE are compared. These objectives are also aligned until the EE reaches its maximum value. However, one can increase the area rate beyond this point with only minor losses in EE. By noting that $g_2(\mathbf{x}) = \frac{K}{A} g_1(\mathbf{x})$ and comparing with the previous figure, this obviously means that the area rate is improved by transmitting to more users (i.e., having a larger K) and not by increasing the rate per user. This conclusion is supported by Fig. 8, which shows the three-dimensional objective set with respect to all objectives.

Fig. 8 reveals that high area rates are only achievable when the rate per active user is low, which means that we serve many user devices in parallel. In contrast, high rates per user is only achievable by having fewer active users. High energy efficiency is possible when the rate per user is small. These different operating points are achieved by different resource utilizations $\mathbf{x} \in \mathcal{X}$; thus, the number of antennas/users are different and the signal processing related to precoding changes. This proves the otherwise heuristic belief that the network architecture must be flexible (e.g., in terms of switching off antennas and precoding adaptation)

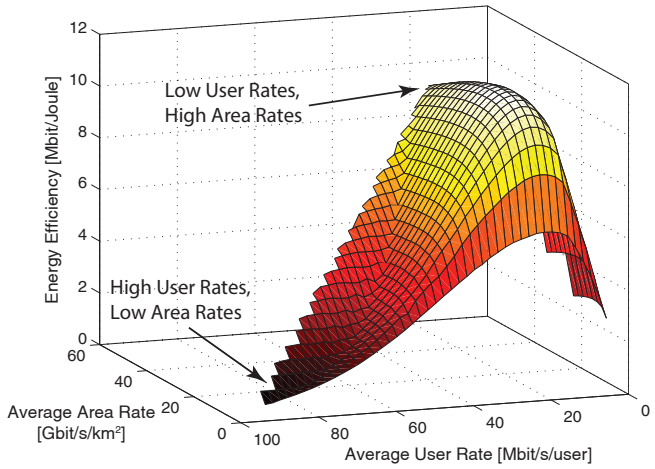

Fig. 8. Visualization of the tradeoff between all three objectives in the case study: average user rate, average area rate, and energy efficiency.

if different operating points should be attainable in different traffic cases.

The discussion above is typical for the *a posteriori* method; we analyzed the shape of the Pareto boundary and drew conclusions on which operating points that are preferable to us. If we would instead utilize the *a priori* method, then we need to specify a goal function. This can be done by picking any of the function classes described in (7)–(10) and selecting the corresponding parameters (e.g., weights) to describe our subjective goals. To aid us in this process, suppose we know the utopia point $\mathbf{u}_{\text{utopia}}$ (defined in (3)) in advance. This point contains the maximal value for each objective, if we would focus completely on it. If the three objectives are equally important to us, it makes sense to normalize their numerical ranges. This is achieved by setting $\mathbf{w} = [\frac{1}{u_1} \; \frac{1}{u_2} \; \frac{1}{u_3}]^T$ in the weighted sum goal, $\mathbf{w} = [1 \; 1 \; 1]^T$ in the weight product goal, and $\mathbf{w} = \mathbf{u}_{\text{utopia}}$ in the weighted Chebyshev goal formulation. The corresponding operating points when solving the scalarized problem in (6) are shown in Fig. 6 and Fig. 7. The shape of the region has a great impact on the spread of the operating points, but different weights still give different operating points (as discussed earlier). The utopia point $\mathbf{u}_{\text{utopia}} = [u_1 \; u_2 \; u_3]^T$ is also shown in these figures. We observe that it is far outside the attainable objective set in Fig. 6, since the two objectives are strongly conflicting. On the contrary, the utopia point is quite close to the objective set in Fig. 7, where the conflict is rather mild.

Finally, we remark that the *a posteriori* and *a priori* methods can be combined. The network architecture can, for example, be designed by studying the shape of the attainable objective set and making sure that the network can adapt and achieve different operating points at the Pareto boundary at different times. The system designer can then formulate multiple goal functions that are exploited for efficient real-time network adaptation, based on current traffic load, service requirements, and capability of the user devices.

CONCLUSIONS AND FUTURE DIRECTIONS

The design expectations on 5G wireless networks cannot be properly articulated by a single performance objective. There are many conflicting objectives, such as improving the peak user rates, average area rates, and energy efficiency. The network design thus calls for multi-objective optimization, which is rigorous framework for studying and solving design problems with multiple objectives. This article provided a survey on this topic. There is no objectively optimal solution to this type of problems, but there are two main methods to find subjectively optimal solutions that fit the needs of the network designer. The *a posterior method* computes sample points on the Pareto boundary—the set of tentative operating points where no objective can be improved without degrading another objective. The sample points are used to visualize the Pareto boundary for the network designer, who can then make well-informed design decisions. Alternatively, the network designer can specify a goal function that describes the acceptable tradeoffs between objectives and infers an order on the attainable operating points. One can then maximize this tradeoff by solving a conventional optimization problem and thereby obtain the most suitable Pareto boundary point.

We also provided a case study on network dimensioning of cellular networks that allows for massive MIMO deployment. This example illustrates our vision of how the MOO framework can be utilized to balance conflicting performance objectives when designing future wireless communication networks. While the analytic tools provided by MOO are well-established, the applications to communication networks are greatly unexplored. A particular research challenge is to formulate MOOPs with a modeling granularity that allows us to answer fundamental design questions related to how the system can efficiently manage the heterogeneous 5G characteristics described in the introduction. To this end, the models must capture the main practical propagation characteristics, be robust to hardware imperfections and uncertain model parameters, and allow for optimization of the signal processing techniques. All of this is to be done while making the basic optimization operations (e.g., the membership test described above) computationally tractable.

ACKNOWLEDGMENTS

This article has been supported by the International Postdoc Grant 2012-228 from the Swedish Research Council and the ERC Starting Grant 305123 MORE (Advanced Mathematical Tools for Complex Network Engineering).

AUTHORS

Emil Björnson (emil.bjornson@liu.se) received the M.S. degree from Lund University, Sweden, in 2007 and the Ph.D. degree in 2011 from KTH Royal Institute of Technology, Sweden. From 2012 to 2014, he was a joint postdoctoral researcher at Supélec, France, and KTH Royal Institute of Technology, Sweden, sponsored by a personal International Postdoc Grant from the Swedish Research Council. He is the first author of the book "Optimal Resource Allocation in Coordinated Multi-Cell Systems" and received best conference paper awards in 2009, 2011, and 2014. He is now a Research Fellow in the tenure-track at Linköping University, Sweden.

Eduard Jorswieck (eduard.jorswieck@tu-dresden.de) received the M.S. and Ph.D. degrees from the Technische Universität Berlin, Germany, in 2000 and 2004, respectively. He was with the Fraunhofer Institute for Telecommunications, Heinrich-Hertz-Institut (HHI) Berlin, from 2000 to 2008. From 2005 to 2008, he was a lecturer at the Technische Universität Berlin. From 2006 to 2008 he was a post-doc and assistant professor at the KTH Royal Institute of Technology, Sweden. Since 2008, he has been the head of the Chair of Communications Theory and Full Professor at Dresden University of Technology (TUD), Germany. In 2006, he received the IEEE Signal Processing Society Best Paper Award.

Mérouane Debbah (merouane.debbah@supelec.fr) received the M.S. and Ph.D. degrees from École Normale Supérieure de Cachan, France, in 1999 and 2002, respectively. From 1999 to 2002, he worked for Motorola Labs. From 2002 to 2003, he was appointed senior researcher at the Vienna Research Center for Telecommunications (FTW) (Austria). From 2003 to 2007, he joined the mobile communications department of the Institut Eurecom, France, as an assistant professor. He is currently a professor at Supélec, France, and holds the Alcatel-Lucent Chair on Flexible Radio. He received the 2005 Mario Boella Prize Award, the 2007 GLOBECOM Best Paper Award, the 2009 Wi-Opt Best Paper Award, the 2010 Newcom++ Best Paper Award, as well as the 2007 Valuetools, 2008 Valuetools, and 2009 CrownCom Best Student Paper Awards. He is a WWRF fellow.

Björn Ottersten (bjorn.ottersten@uni.lu) received the M.S. degree from Linköping University, Sweden, in 1986 and the Ph.D. degree in 1989 from Stanford University, California. In 1991 he was appointed Professor of Signal Processing at KTH Royal Institute of Technology, Sweden. During 96/97 Dr. Ottersten was Director of Research at ArrayComm Inc, a start-up based on Ottersten's patented technology. Currently, Dr. Ottersten is Director for the Interdisciplinary Centre for Security, Reliability and Trust at the University of Luxembourg. He coauthored articles that received the IEEE Signal Processing Society Best Paper Award in 1993, 2001, 2006, and 2013, and several IEEE conference papers receiving Best Paper Awards. In 2011 he received the IEEE Signal Processing Society Technical Achievement Award. He is editor-in-chief of EURASIP Signal Processing Journal. He is a Fellow of the IEEE and EURASIP.

REFERENCES

[1] S. Tombaz, A. Västberg, and J. Zander, "Energy- and cost-efficient ultra-high-capacity wireless access," *IEEE Wireless Commun. Mag.*, vol. 18, no. 5, pp. 18–24, 2011.

[2] L. Zadeh, "Optimality and non-scalar-valued performance criteria," *IEEE Trans. Autom. Control*, vol. 8, no. 1, pp. 59–60, 1963.

[3] R.T. Marler and J.S. Arora, "Survey of multi-objective optimization methods for engineering," *Struct Multidisc Optim*, vol. 26, pp. 369–395, 2004.

[4] J. Branke, K. Deb, K. Miettinen, and R. Slowinski (Eds.), *Multiobjective Optimization: Interactive and Evolutionary Approaches*, Springer, 2008.

[5] R. Bot, S.-M. Grad, and G. Wanka, *Duality in Vector Optimization*, Springer, 2009.

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

[7] A. Osseiran, "The 5G future scenarios identified by METIS - the first step toward a 5G mobile and wireless communications system," *Press Release*, Sept. 2013.

[8] F. Kelly, "Charging and rate control for elastic traffic," *European Trans. Telecom.*, vol. 8, pp. 33–37, 1997.

[9] D.P. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1439–1451, 2006.

[10] H. Weingarten, Y. Steinberg, and S. Shamai, "The capacity region of the Gaussian multiple-input multiple-output broadcast channel," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 3936–3964, 2006.

[11] Y.-F. Liu, Y.-H. Dai, and Z.-Q. Luo, "Coordinated beamforming for MISO interference channel: Complexity analysis and efficient algorithms," *IEEE Trans. Signal Process.*, vol. 59, no. 3, pp. 1142–1157, 2011.

[12] F. Rashid-Farrokhi, K.J.R. Liu, and L. Tassiulas, "Transmit beamforming and power control for cellular wireless systems," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 8, pp. 1437–1450, 1998.

[13] M. Chiang, P. Hande, T. Lan, and C.W. Tan, "Power control in wireless cellular networks," *Foundations and Trends in Networking*, vol. 2, no. 4, pp. 355–580, 2008.

[14] Y. Chen, S. Zhang, S. Xu, and G.Y Li, "Fundamental trade-offs on green wireless networks," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 30–37, 2011.

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

[16] E. Björnson, L. Sanguinetti, J. Hoydis, and M. Debbah, "Optimal design of energy-efficient multi-user MIMO systems: Is massive MIMO the answer?," *IEEE Trans. Wireless Commun.*, Submitted.

[17] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[18] E. Jorswieck, L. Badia, T. Fahldieck, E. Karipidis, and J. Luo, "Spectrum sharing improves the network efficiency for cellular operators," *IEEE Commun. Mag.*, vol. 52, no. 3, pp. 129–136, 2014.

[19] R.H. Roy and B. Ottersten, "Spatial division multiple access wireless communication systems," European Patent EP0616742, 1994.

[20] D. Gesbert, M. Kountouris, R.W. Heath, C.-B. Chae, and T. Sälzer, "Shifting the MIMO paradigm," *IEEE Signal Process. Mag.*, vol. 24, no. 5, pp. 36–46, 2007.

[21] D. Gesbert, S. Hanly, H. Huang, S. Shamai, O. Simeone, and W. Yu, "Multi-cell MIMO cooperative networks: A new look at interference," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1380–1408, 2010.

[22] P. Marsch and G. Fettweis, "On multicell cooperative transmission in backhaul-constrained cellular systems," *Ann. Telecommun.*, vol. 63, pp. 253–269, 2008.

[23] T.L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, 2010.

[24] F. Rusek, D. Persson, B.K. Lau, E.G. Larsson, T.L. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, 2013.

[25] J. Hoydis, K. Hosseini, S. ten Brink, and M. Debbah, "Making smart use of excess antennas: Massive MIMO, small cells, and TDD," *Bell Labs Technical Journal*, vol. 18, no. 2, pp. 5–21, 2013.

[26] E. Björnson, J. Hoydis, M. Kountouris, and M. Debbah, "Massive MIMO systems with non-ideal hardware: Energy efficiency, estimation, and capacity limits," *IEEE Trans. Inf. Theory*, Submitted.

[27] H. Yang and T.L. Marzetta, "Total energy efficiency of cellular large scale antenna system multiple access mobile networks," in *Proc. OnlineGreenComm*, 2013.

[28] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson, M.A. Imran, D. Sabella, M.J. Gonzalez, O. Blume, and A. Fehske, "How much energy is needed to run a wireless network?," *IEEE Wireless Commun. Mag.*, vol. 18, no. 5, pp. 40–49, 2011.