Review   **Open Access**

Optimal resource allocation in wireless communication and networking

Alejandro Ribeiro

Abstract

Optimal design of wireless systems in the presence of fading involves the instantaneous allocation of resources such as power and frequency with the ultimate goal of maximizing long term system properties such as ergodic capacities and average power consumptions. This yields a distinctive problem structure where long term average variables are determined by the expectation of a not necessarily concave functional of the resource allocation functions. Despite their lack of concavity it can be proven that these problems have null duality gap under mild conditions permitting their solution in the dual domain. This affords a significant reduction in complexity due to the simpler structure of the dual function. The article discusses the problem simplifications that arise by working in the dual domain and reviews algorithms that can determine optimal operating points with relatively lightweight computations. Throughout the article concepts are illustrated with the optimal design of a frequency division broadcast channel.

Introduction

Operating variables of a wireless system can be separated in two types. Resource allocation variables $\mathbf{p}(\mathbf{h})$ determine instantaneous allocation of resources like frequencies and transmitted powers as a function of the fading coefficient $\mathbf{h}$. Average variables $\mathbf{x}$ capture system's performance over a large period of time and are related to instantaneous resource allocations via ergodic averages. A generic representation of the relationship between instantaneous and average variables is

$$\mathbf{x} \leq \mathbb{E}\left[\mathbf{f}_1\left(\mathbf{h}, \mathbf{p}(\mathbf{h})\right)\right], \tag{1}$$

where $\mathbf{f}_1(\mathbf{h}, \mathbf{p}(\mathbf{h}))$ is a vector function that maps channel $\mathbf{h}$ and resource allocation $\mathbf{p}(\mathbf{h})$ to instantaneous performance $\mathbf{f}_1(\mathbf{h}, \mathbf{p}(\mathbf{h}))$. The system's design goal is to select resource allocations $\mathbf{p}(\mathbf{h})$ to maximize ergodic variables $\mathbf{x}$ in some sense.

An example of a relationship having the form in (1) is a code division multiple access channel in which case $\mathbf{h}$ denotes the vector of channel coefficients, $\mathbf{p}(\mathbf{h})$ the instantaneous transmitted power, $\mathbf{f}_1(\mathbf{h}, \mathbf{p}(\mathbf{h}))$ the instantaneous communication rate determined by the signal to interference plus noise ratio, and $\mathbf{x}$ the ergodic rates determined by the expectation of the instantaneous rates. The design

Correspondence: aribeiro@seas.upenn.edu
Department of Electrical and Systems Engineering, University of Pennsylvania,
200 S. 33rd St., Philadelphia, PA, 19096, USA

goal is to allocate instantaneous power $\mathbf{p}(\mathbf{h})$ subject to a power constraint so as to maximize a utility of the ergodic rate vector $\mathbf{x}$. This interplay of instantaneous actions to optimize long term performance is pervasive in wireless systems. A brief list of examples includes optimization of orthogonal frequency division multiplexing [1], beamforming [2,3], cognitive radio [4,5], random access [6,7], communication with imperfect channel state information (CSI) [8,9], and various flavors of wireless network optimization [10-18].

In many cases of interest the functions $\mathbf{f}_1(\mathbf{h}, \mathbf{p}(\mathbf{h}))$ are nonconcave and as a consequence finding the resource allocation distribution $\mathbf{p}^*(\mathbf{h})$ that maximizes $\mathbf{x}$ requires solution of a nonconvex optimization problem. This is further complicated by the fact that since fading channels $\mathbf{h}$ take on a continuum of values there is an infinite number of $\mathbf{p}^*(\mathbf{h})$ variables to be determined. A simple escape to this problem is to allow for time sharing in order to make the range of $\mathbb{E}\left[\mathbf{f}_1(\mathbf{h}, \mathbf{p}(\mathbf{h}))\right]$ convex and permit solution in the dual domain without loss of optimality. While the nonconcave function $\mathbf{f}_1(\mathbf{h}, \mathbf{p}(\mathbf{h}))$ still complicates matters, working in the dual domain makes solution, if not necessarily simple, at least substantially simpler. However, time sharing is not easy to implement in fading channels.

In this article, we review a general methodology that can be used to solve optimal resource allocation problems in wireless communications and net-

working without resorting to time sharing [19,20]. The fundamental observation is that the range of $\mathbb{E}\left[\mathbf{f}_1(\mathbf{h}, \mathbf{p}(\mathbf{h}))\right]$ is convex if the probability distribution of the channel $\mathbf{h}$ contains no points of positive probability (Section "Duality in wireless systems optimization"). This observation can be leveraged to show lack of duality gap of general optimal resource allocation problems (Theorem 1) making primal and dual problems equivalent. The dual problem is simpler to solve and its solution can be used to recover primal variables (Section "Recovery of optimal primal variables") with reduced computational complexity due to the inherently separable structure of the problem Lagrangians (Section "Separability"). We emphasize that this reduction in complexity, as in the case of time sharing, just means that the problem becomes simpler to solve. In many cases it also becomes simple to solve, but this is not necessarily the case.

We also discuss a stochastic optimization algorithm to determine optimal dual variables that can operate without knowledge of the channel probability distribution (Section "Dual descent algorithms"). This algorithm is known to almost surely converge to optimal operating points in an ergodic sense (Theorem 5). Throughout the article concepts are illustrated with the optimal design of a frequency division broadcast channel (Section "Frequency division broadcast channel" in "Optimal wireless system design", Section "Frequency division broadcast channel" in "Recovery of optimal primal variables", and Section "Frequency division broadcast channel" in "Dual descent algorithms").

One of the best known resource allocation problems in wireless communications concerns the distribution of power on a block fading channel using capacity-achieving codes. The solution to this problem is easy to derive and is well known to reduce to waterfilling across the fading gain, e.g., [21, p. 245]. Since this article can be considered as an attempt to generalize this solution methodology to general wireless communication and networking problems it is instructive to close this introduction by reviewing the derivation of the waterfilling solution. This is pursued in the following section.

Power allocation in a point-to-point channel

Consider a transmitter having access to perfect CSI h that it uses to select a transmitted power $p(h)$ to convey information to a receiver. Using a capacity achieving code the instantaneous channel rate for fading realization h is $r(h) = \log(1 + hp(h)/N_0)$ where N_0 denotes the noise power at the receiver end. A common goal is to maximize the average rate $r := \mathbb{E}[r(h)]$ with respect to the probability distribution $m_h(h)$ of the channel gain h—which is an accurate approximation of the long term average rate—subject to an average power constraint P_0. We can formulate this problem as the optimization program

$$P = \max \mathbb{E}\left[\log\left(1 + \frac{hp(h)}{N_0}\right)\right]$$
$$\text{s.t.} \quad \mathbb{E}[p(h)] \leq q_0. \tag{2}$$

In most cases the fading channel h takes on a continuum of values. Therefore, solving (2) requires the determination of a power allocation function $p : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ that maps nonnegative fading coefficients to nonnegative power allocations. This means that (2) is an infinite dimensional optimization problem which in principle could be difficult to solve. Nevertheless, the solution to this program is easy to derive and given by waterfilling as we already mentioned. The widespread knowledge of the waterfilling solution masks the fact that is is rather remarkable that (2) is easy to solve and begs the question of what are the properties that make it so. Let us then review the derivation of the waterfilling solution in order to pinpoint these properties.

To solve (2) we work in the dual domain. To work in the dual domain we need to introduce the Lagrangian, the dual function, and the dual problem. Introduce then the nonnegative dual variable $\lambda \in \mathbb{R}^+$ and define the Lagrangian associated with the optimization problem in (2) as

$$\mathcal{L}(p, \lambda) = \mathbb{E}\left[\log\left(1 + \frac{hp(h)}{N_0}\right)\right] + \lambda\left[q_0 - \mathbb{E}[p(h)]\right]. \tag{3}$$

The dual function is defined as the maximum value of the Lagrangian across all functions $p : \mathbb{R}^+ \rightarrow \mathbb{R}^+$, which upon defining $\mathcal{P} := \{p : p(h) \geq 0\}$ as the set of nonnegative functions can be defined as

$$g(\lambda) := \max_{p \in \mathcal{P}} \mathcal{L}(p, \lambda)$$
$$= \max_{p \in \mathcal{P}} \mathbb{E}\left[\log\left(1 + \frac{hp(h)}{N_0}\right)\right] + \lambda\left[q_0 - \mathbb{E}[p(h)]\right]. \tag{4}$$

The dual problem corresponds to the minimization of the dual function with respect to all nonnegative multipliers λ,

$$D = \min_{\lambda \geq 0} g(\lambda) = \min_{\lambda \geq 0} \max_{p \in \mathcal{P}} \mathcal{L}(p, \lambda). \tag{5}$$

Since the objective in (2) is concave with respect to variables $p(h)$ and the constraint is linear in $p(h)$ the optimization problem in (2) is convex and as such it has null duality gap in the sense that $P = D$.

An entity that is important for the upcoming discussion is the primal Lagrangian maximizer function $p(\lambda) : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ whose values for given h are denoted as $p(h, \lambda)$.

This function is defined as the one that maximizes the Lagrangian for given dual variable λ

$$p(\lambda) = \underset{p \in \mathcal{P}}{\operatorname{argmax}} \, \mathcal{L}(p, \lambda). \qquad (6)$$

Using the definition of the Lagrangian maximizer function we can write the dual function as $g(\lambda) = \mathcal{L}(p(\lambda), \lambda)$.

Computing values $p(h, \lambda)$ of the Lagrangian maximizer function $p(\lambda)$ is easy. To see that this is true rewrite the Lagrangian in (3) so that there is only one expectation

$$\mathcal{L}(p, \lambda) = \mathbb{E}\left[\log\left(1 + \frac{hp(h)}{N_0}\right) - \lambda p(h)\right] + \lambda q_0. \qquad (7)$$

With the Lagrangian written in this form we can see that the maximization of $\mathcal{L}(p, \lambda)$ required by (6) can be decomposed in maximizations for each individual channel realization,

$$\max_{p \in \mathcal{P}} \mathcal{L}(p, \lambda) = \mathbb{E}\left[\max_{p(h)>0} \log\left(1 + \frac{hp(h)}{N_0}\right) - \lambda p(h)\right] + \lambda q_0. \qquad (8)$$

That the equality in (8) is true is a consequence of the fact that the expectation operator is linear and that there are no constraints coupling the selection of values $p(h_1)$ and $p(h_2)$ for different channel realizations $h_1 \neq h_2$. Functional values $p(h)$ in both sides of (8) are required to be nonnegative but other than that we can select $p(h_1)$ and $p(h_2)$ independently of each other as indicated in the right hand side of (8).

Since the right hand side of (8) states that to maximize the Lagrangian we can select functional values $p(h)$ independently of each other, values $p(h, \lambda)$ of the Lagrangian maximizer function $p(\lambda)$ defined in (6) are given by

$$p(h, \lambda) = \underset{p(h) \geq 0}{\operatorname{argmax}} \left[\log\left(1 + \frac{hp(h)}{N_0}\right) - \lambda p(h)\right]. \qquad (9)$$

The similarity between (6) and (9) is deceiving as the latter is a much easier problem to solve that involves a single variable. To find the Lagrangian maximizer value $p(h, \lambda)$ operating from (9) it suffices to solve for the null of the derivative with respect to $p(h)$. Doing this yields the Lagrangian maximizer

$$p(h, \lambda) = \left[\frac{1}{\lambda} - \frac{N_0}{h}\right]^+, \qquad (10)$$

where the operator $[x]^+ := \max(x, 0)$ denotes projection onto nonnegative reals, which is needed because of the constraint $p(h) \geq 0$.

Of particular interest is the Lagrangian maximizer function $p(\lambda^*)$ corresponding to the optimal Lagrange multiplier $\lambda^* := \operatorname{argmin}_{\lambda \geq 0} g(\lambda)$. Returning to the definition of the dual function in (4) we can bound $D = g(\lambda^*)$ as

$$D = \max_{p \in \mathcal{P}} \mathcal{L}(p, \lambda^*) \geq \mathcal{L}\left(p^*, \lambda^*\right). \qquad (11)$$

Indeed, since D is given by a Lagrangian maximization it must equal or exceed the values of the Lagrangian for any function p, and for the optimal function p^* in particular. Considering the explicit Lagrangian expression in (3) we can write $\mathcal{L}(p^*, \lambda^*)$ as

$$\mathcal{L}(p^*, \lambda^*) = \mathbb{E}\left[\log\left(1 + \frac{hp^*(h)}{N_0}\right)\right] + \lambda^*\left[q_0 - \mathbb{E}\left[p^*(h)\right]\right]. \qquad (12)$$

Observe that since p^* is the optimal power allocation function it must satisfy the power constraint implying that it must be $q_0 - \mathbb{E}\left[p^*(h)\right] \geq 0$. Since the optimal dual variable satisfies $\lambda^* \geq 0$ their product is also nonnegative allowing us to transform (12) into the bound

$$\mathcal{L}\left(p^*, \lambda^*\right) \geq \mathbb{E}\left[\log\left(1 + \frac{hp^*(h)}{N_0}\right)\right] = P, \qquad (13)$$

where the equality is true because P and p^* are the otpimal value and arguments of the primal optimization problem (2). Combining the bounds in (11) and (13) yields

$$D = \max_{p \in \mathcal{P}} \mathcal{L}\left(p, \lambda^*\right) \geq \mathcal{L}\left(p^*, \lambda^*\right) \geq P. \qquad (14)$$

Using the equivalence $P = D$ of primal and dual optimum values it follows that the inequalities in (13) must hold as equalities. The equality $\max_{p \in \mathcal{P}} \mathcal{L}(p, \lambda^*) = \mathcal{L}(p^*, \lambda^*)$, in particular, implies that the function p^* is the Lagrangian maximizing function corresponding to $\lambda = \lambda^*$, i.e.,

$$p^* = p\left(\lambda^*\right) = \underset{p}{\operatorname{argmax}} \, \mathcal{L}\left(p, \lambda^*\right). \qquad (15)$$

The important consequence of (15) follows from the fact that the Lagrangian maximizer function $p\left(\lambda^*\right)$ is easy to compute using the decomposition in (9). By extension, if the optimal multiplier λ^* is available, computation of the optimal power allocation function p^* also becomes easy. Indeed, making $\lambda = \lambda^*$ in (9) we can determine values $p^*(h)$ of p^* as

$$p^*(h) = p(h, \lambda^*) = \underset{p(h) \geq 0}{\operatorname{argmax}} \left[\log\left(1 + \frac{hp(h)}{N_0}\right) - \lambda^* p(h)\right], \qquad (16)$$

which are explicitly given by (10) with $\lambda = \lambda^*$.

To complete the problem solution we still need to determine the optimal multiplier λ^*. We show a method for doing so in Section "Dual descent algorithms", but it is important to recognize that this cannot be a difficult problem because the dual function is single-dimensional and convex—dual functions of maximization problems are always convex. This has to be contrasted with the infinite dimensionality of the primal problem. By working in the dual domain we reduce the problem of determining the infinite dimensional optimal power allocation function to the determination of the one dimensional optimal Lagrange multiplier λ^*.

Recapping the derivation of the optimal power allocation p^*, we see that there are three conditions that make (18) simple:

(C1) Since the optimization problem is convex it is equivalent to its Lagrangian dual implying that optimal primal variables can be determined as Lagrangian maximizers associated with the optimal multiplier λ^* [cf. (11)–(15)].

(C2) Due to the separable structure of the Lagrangian, determination of the optimal power allocation function is carried out through the solution of per-fading state subproblems [cf. (6)–(9)].

(C3) Because there is a finite number of constraints the dual function is finite dimensional even though there is an infinite number of primal variables.

Most optimization problems in wireless systems are separable in the sense of (C2) and have a finite number of constraints as [cf. (C3)] but are typically not convex [cf. (C1)].

To illustrate this latter point consider a simple variation of (2) where instead of using capacity achieving codes we use adaptive modulation and coding (AMC) relying on a set of L communication modes. The l-th mode supports a rate α_l and is used when the signal to noise ratio (SNR) at the receiver end is between β_{l-1} and β_l. Letting γ be the received SNR and $\mathbb{I}(\mathcal{E})$ denote the indicator function of the event $\mathcal{E}$, the communication rate function $C(\gamma)$ for AMC can be written as

$$C_{\text{AMC}}(\gamma) = \sum_{l=1}^{L} \alpha_l \mathbb{I}\left(\beta_l \leq \gamma < \beta_{l+1}\right). \quad (17)$$

The corresponding optimal power allocation problem subject to average power constraint q_0 can now be formulated as

$$P = \max \mathbb{E}\left[\sum_{l=1}^{L} \alpha_l \mathbb{I}\left(\beta_l \leq \frac{hp(h)}{N_0} < \beta_{l+1}\right)\right]$$

$$\text{s.t. } \mathbb{E}\left[p(h)\right] \leq q_0. \quad (18)$$

Similar to (2) the dual function of the optimization problem in (18) is one dimensional and its Lagrangian is separable in the sense that we can find Lagrangian maximizers by performing per-fading state maximizations. Alas, the problem is not convex because the AMC rate function $C(\gamma)$ is not concave, in fact, it is not even continuous.

Since it does not satisfy (C1), solving (18) in the dual domain is not possible in principle. Nevertheless the condition that allows determination of $p^*(h)$ as the Lagrangian maximizer $p(h, \lambda^*)$ [cf. (16)] is not the convexity of the primal problem but the lack of duality gap. We'll see in Section "Duality in wireless systems optimization" that this problem does have null duality gap as long as

the probability distribution of the channel h contains no points of strictly positive probability. Thus, the solution methodology in (3)-(16) can be applied to solve (18) despite the discontinuous AMC rate function $C(\gamma)$. This is actually a quite generic property that holds true for optimization problems where nonconcave functions appear inside expectations. We introduce this generic problem formulation in the next section.

Optimal wireless system design

Let us return to the relationship in (1) where $\mathbf{h}$ denotes the random fading state that ranges on a continuous space $\mathcal{H}$, $\mathbf{p}(\mathbf{h})$ the instantaneous resource allocation, $\mathbf{f}_1(\mathbf{h}, \mathbf{p}(\mathbf{h}))$ a vector function mapping $\mathbf{h}$ and $\mathbf{p}(\mathbf{h})$ to instantaneous system performance, and $\mathbf{x}$ an ergodic average. The expectation in (1) is with respect to the joint probability distribution $m_{\mathbf{h}}(\mathbf{h})$ of the vector channel $\mathbf{h}$.

It is convenient to think of $\mathbf{f}_1(\mathbf{h}, \mathbf{p}(\mathbf{h}))$ as a family of functions indexed by the parameter $\mathbf{h}$ with $\mathbf{p}(\mathbf{h})$ denoting the corresponding variables. Notice that there is one vector $\mathbf{p}(\mathbf{h})$ per fading state $\mathbf{h}$, which translates into an infinite number of resource allocation variables if $\mathbf{h}$ takes on an infinite number of values. Consequently, it is adequate to refer to the set $\mathbf{p} := \{\mathbf{p}(\mathbf{h})\}_{\mathbf{h} \in \mathcal{H}}$ of all resource allocations as the resource allocation function. The number of ergodic limits $\mathbf{x}$ of interest, on the other hand, is assumed finite.

Instantaneous resource allocations $\mathbf{p}(\mathbf{h})$ are further constrained to a given bounded set $\mathcal{P}(\mathbf{h})$. These restrictions define a set of admissible resource allocation functions that we denote as

$$\mathcal{P} := \{\mathbf{p} : \mathbf{p}(\mathbf{h}) \in \mathcal{P}(\mathbf{h}), \quad \text{for all } \mathbf{h} \in \mathcal{H}\}. \quad (19)$$

Variables $\mathbf{p}(\mathbf{h})$ determine system performance over short periods of times. As such, they are of interest to the system designer but transparent to the end user except to the extent that they determine the value of ergodic variables $\mathbf{x}$. Therefore, we adopt as our design objective the maximization of a concave utility function $f_0(\mathbf{x})$ of the ergodic average $\mathbf{x}$. Putting these preliminaries together we write the following program as an abstract formulation of optimal resource allocation problems in wireless systems

$$P = \max \ f_0(\mathbf{x})$$

$$\text{s.t. } \mathbf{x} \leq \mathbb{E}\left[\mathbf{f}_1(\mathbf{h}, \mathbf{p}(\mathbf{h}))\right], \quad \mathbf{f}_2(\mathbf{x}) \geq \mathbf{0},$$

$$\mathbf{x} \in \mathcal{X}, \quad \mathbf{p} \in \mathcal{P}, \quad (20)$$

where we added further constraints on the set of ergodic averages $\mathbf{x}$. These constraints are in the form of a bounded convex set inclusion $\mathbf{x} \in \mathcal{X}$ and a concave function inequality $\mathbf{f}_2(\mathbf{x}) \geq \mathbf{0}$. In the problem formulation in (20) the set $\mathcal{X}$ is convex and the functions $f_0(\mathbf{x})$ and $\mathbf{f}_2(\mathbf{x})$ are concave. The family of functions $\mathbf{f}_1(\mathbf{h}, \mathbf{p}(\mathbf{h}))$ is *not* necessarily concave with respect to $\mathbf{p}(\mathbf{h})$ and the set $\mathcal{P}$ is

not necessarily convex. The sets $\mathcal{X}$ and $\mathcal{P}$ are assumed compact to guarantee that $\mathbf{x}$ and $\mathbf{p(h)}$ are finite. For the expectation in (20) to exist we need to have $\mathbf{f}_1(\mathbf{h}, \mathbf{p(h)})$ integrable with respect to the probability distribution $m_{\mathbf{h}}(\mathbf{h})$ of the vector channel $\mathbf{h}$. This imposes a (mild) restriction on the functions $\mathbf{f}_1(\mathbf{h}, \mathbf{p(h)})$ and the power allocation function $\mathbf{p}$. Integrability is weaker than continuity.

For future reference define $\mathbf{x}^*$ and $\mathbf{p}^*$ as the arguments that solve (20)

$$(\mathbf{x}^*, \mathbf{p}^*) = \underset{}{\text{argmax}} \ f_0(\mathbf{x})$$
$$\text{s.t} \quad \mathbf{x} \leq \mathbb{E}\left[\mathbf{f}_1(\mathbf{h}, \mathbf{p(h)})\right], \quad \mathbf{f}_2(\mathbf{x}) \geq \mathbf{0},$$
$$\mathbf{x} \in \mathcal{X}, \quad \mathbf{p} \in \mathcal{P}. \qquad (21)$$

The configuration pair $(\mathbf{x}^*, \mathbf{p}^*)$ attains the optimum value $P = f_0(\mathbf{x}^*)$ and satisfies the constraints $\mathbf{x}^* \leq \mathbb{E}\left[\mathbf{f}_1(\mathbf{h}, \mathbf{p}^*(\mathbf{h}))\right]$ and $\mathbf{f}_2(\mathbf{x}^*) \geq \mathbf{0}$ as well as the set constraints $\mathbf{x}^* \in \mathcal{X}$, and $\mathbf{p}^* \in \mathcal{P}$. Observe that the pair $(\mathbf{x}^*, \mathbf{p}^*)$ need not be unique. It may be, and it is actually a common occurrence in practice, that more than one configuration is optimal. Thus, (21) does not define a pair of variables but a set of pairs of optimal configurations. As it does not lead to confusion we use $(\mathbf{x}^*, \mathbf{p}^*)$ to represent both, the set of optimal configurations and an arbitrary element of this set.

To write the Lagrangian we introduce a nonnegative Lagrange multiplier $\mathbf{\Lambda} = \left[\boldsymbol{\lambda}_1^T, \boldsymbol{\lambda}_2^T\right]^T$ where $\boldsymbol{\lambda}_1 \geq \mathbf{0}$ is associated with the constraint $\mathbf{x} \leq \mathbb{E}\left[\mathbf{f}_1(\mathbf{h}, \mathbf{p(h)})\right]$ and $\boldsymbol{\lambda}_2 \geq \mathbf{0}$ with the constraint $\mathbf{f}_2(\mathbf{x})$. The Lagrangian of the primal optimization problem in (20) is then defined as

$$\mathcal{L}(\mathbf{x}, \mathbf{p}, \mathbf{\Lambda}) := f_0(\mathbf{x}) + \boldsymbol{\lambda}_1^T \left[\mathbb{E}\left(\mathbf{f}_1(\mathbf{h}, \mathbf{p(h)})\right) - \mathbf{x}\right] + \boldsymbol{\lambda}_2^T \mathbf{f}_2(\mathbf{x}).$$
$$(22)$$

The corresponding dual function is the maximum of the Lagrangian with respect to primal variables $\mathbf{x} \in \mathcal{X}$ and $\mathbf{p} \in \mathcal{P}$

$$g(\mathbf{\Lambda}) := \max_{\mathbf{x} \in \mathcal{X}, \mathbf{p} \in \mathcal{P}} \mathcal{L}(\mathbf{x}, \mathbf{p}, \mathbf{\Lambda}). \qquad (23)$$

The dual problem is defined as the minimum value of the dual function over all nonnegative dual variables

$$D = \min_{\mathbf{\Lambda} \geq \mathbf{0}} g(\mathbf{\Lambda}), \qquad (24)$$

and the optimal dual variables are defined as the arguments that achieve the minimum in (24)

$$\mathbf{\Lambda}^* := \underset{\mathbf{\Lambda} \geq \mathbf{0}}{\text{argmin}} \ g(\mathbf{\Lambda}). \qquad (25)$$

Notice that the optimal dual argument $\mathbf{\Lambda}^*$ is a set as in the case of the primal optimal arguments because there may be more than one vector that achieves the minimum in (24). As we do with the optimal primal variables $(\mathbf{x}^*, \mathbf{p}^*)$, we use $\mathbf{\Lambda}^*$ to denote the set of optimal dual variables and an arbitrary element of this set. A particular example of this generic problem formulation is presented next.

Frequency division broadcast channel

A common access point (AP) administers a power budget q_0 to communicate with a group of J nodes. The physical layer uses frequency division so that at most one terminal can be active at any given point in time. The goal is to design an algorithm that allocates power and frequency to maximize a given ergodic rate utility metric while ensuring that rates are at least $r_{\min}$ and not more than $r_{\max}$.

Denote as h_i the channel to terminal i and define the vector $\mathbf{h} := \left[h_1 \ldots, h_J\right]^T$ grouping all channel realizations. In any time slot the AP observes the realization $\mathbf{h}$ of the fading channel and decides on suitable power allocations $p_i(\mathbf{h})$ and frequency assignments $\alpha_i(\mathbf{h})$. Frequency assignments $\alpha_i(\mathbf{h})$ are indicator variables $\alpha_i(\mathbf{h}) \in \{0, 1\}$ that take value $\alpha_i(\mathbf{h}) = 1$ when information is transmitted to node i and $\alpha_i(\mathbf{h}) = 0$ otherwise. If $\alpha_i(\mathbf{h}) = 1$ communication towards i ensues at power $p_i(\mathbf{h})$ resulting in a communication rate $C(h_i p_i(\mathbf{h})/N_0)$ determined by the SNR $h_i p_i(\mathbf{h})/N_0$. The specific form of the function $C(h_i p_i(\mathbf{h})/N_0)$ mapping channels and powers to transmission rates depends on the type of modulation and codes used. One possibility is to select capacity achieving codes leading to $C(h_i p_i(\mathbf{h})/N_0) = \log(1 + h_i p_i(\mathbf{h})/N_0)$. A more practical choice is to use AMC in which case $C(h_i p_i(\mathbf{h})/N_0) = C_{\text{AMC}}(h_i p_i(\mathbf{h})/N_0)$ with $C_{\text{AMC}}(h_i p_i(\mathbf{h})/N_0)$ as given in (17).

Regardless of the specific form of $C(h_i p_i(\mathbf{h})/N_0)$ we can write the ergodic rate of terminal i as

$$r_i = \mathbb{E}\left[\alpha_i(\mathbf{h}) C\left(\frac{h_i p_i(\mathbf{h})}{N_0}\right)\right]. \qquad (26)$$

The factor $C(h_i p_i(\mathbf{h})/N_0)$ is the instantaneous rate achieved if information is conveyed. The factor $\alpha_i(\mathbf{h})$ indicates wether this information is indeed conveyed or not. The expectation weights the instantaneous capacity across fading states and is equivalent to the consideration of an infinite horizon time average.

Similarly, $p_i(\mathbf{h})$ denotes the power allocated for communication with node i, but this communication is attempted only if $\alpha_i(\mathbf{h}) = 1$. Thus, the instantaneous power utilized to communicate with i for channel realization $\mathbf{h}$ is $\alpha_i(\mathbf{h}) p_i(\mathbf{h})$. The total instantaneous power is the sum of these products for all i and the long term average power consumption can be approximated as the expectation

$$\mathbb{E}\left[\sum_{i=1}^{J} \alpha_i(\mathbf{h}) p_i(\mathbf{h})\right] \leq q_0, \qquad (27)$$

that according to the problem statement cannot exceed the budget q_0.

To avoid collisions between communication attempts the indicator variables $\alpha_i(\mathbf{h})$ are restricted so that at most one of them is 1. Define the vector $\boldsymbol{\alpha}(\mathbf{h}) :=$

$\left[\alpha_1(\mathbf{h}), \ldots, \alpha_J(\mathbf{h})\right]^T$ corresponding to values of the function $\boldsymbol{\alpha} := \left[\alpha_1, \ldots, \alpha_J\right]^T$ and introduce the set of vector functions

$$\mathcal{A} := \left\{ \boldsymbol{\alpha} : \boldsymbol{\alpha}(\mathbf{h}) \in \{0, 1\}, \quad \mathbf{1}^T \boldsymbol{\alpha}(\mathbf{h}) \le 1 \right\}. \qquad (28)$$

We can now express the frequency exclusion constraints as $\boldsymbol{\alpha} \in \mathcal{A}$.

We still need to model the restriction that the achieved capacity r_i needs to be between $r_{\min}$ and $r_{\max}$ but this is easily modeled as the constraint $r_{\min} \le r_i \le r_{\max}$. Defining the vector $\mathbf{r} = \left[r_1, \ldots, r_J\right]^T$ this constraint can be written as $\mathbf{r} \in \mathcal{R}$ with the set $\mathcal{R}$ defined as

$$\mathcal{R} := \left\{ \mathbf{r} = \left[r_1, \ldots, r_J\right]^T : r_{\min} \le r_i \le r_{\max} \right\}. \qquad (29)$$

We finally introduce a monotonic nondecreasing utility function $U(r_i)$ to measure the value of rate r_i and formally state our design goal as the maximization of the aggregate utility $\sum_{i=1}^{J} U(r_i)$. Using the definitions in (26)–(29) the operating point that maximizes this aggregate utility for a frequency division broadcast channel follows from the solution of the optimization problem

$$P = \max \sum_{i=1}^{J} U(r_i)$$
$$\text{s.t} \quad r_i \le \mathbb{E}\left[\alpha_i(\mathbf{h}) C\left(\frac{h_i p_i(\mathbf{h})}{N_0}\right)\right],$$
$$q_0 \ge \mathbb{E}\left[\sum_{i=1}^{J} \alpha_i(\mathbf{h}) p_i(\mathbf{h})\right],$$
$$\mathbf{r} \in \mathcal{R}, \quad \boldsymbol{\alpha} \in \mathcal{A}, \qquad (30)$$

where we relaxed the rate expression in (26) to an inequality constraint, which we can do without loss of optimality because the utility $U(r_i)$ is monotonic nondecreasing.

The problem formulation in (30) is of the form in (20). The ergodic rates $\mathbf{r}$ in (30) are represented by the ergodic variables $\mathbf{x}$ in (20) whereas the power and frequency allocation functions $\mathbf{p}$ and $\boldsymbol{\alpha}$ of (30) correspond to the resource allocation function $\mathbf{p}$ of (20). The set $\mathcal{R}$ maps to the set $\mathcal{X}$ and the set $\mathcal{A}$ to the set $\mathcal{P}$. There are no functions in (30) taking the place of the function $\mathbf{f}_2(\mathbf{x})$ of (20). The function $\mathbf{f}_1(\mathbf{h}, \mathbf{p}(h))$ in (20) is a placeholder for the stacking of the functions $\alpha_i(\mathbf{h}) C(h_i p_i(\mathbf{h})/N_0)$ for different i and the negative of the power consumptions $-\sum_{i=1}^{J} \alpha_i(\mathbf{h}) p_i(\mathbf{h})$ in (30). The power constraint $q_0 \ge \mathbb{E}\left[\sum_{i=1}^{J} \alpha_i(\mathbf{h}) p_i(\mathbf{h})\right]$ is not exactly of the form in (1) because q_0 is a constant not a variable but this doesn't alter the fundamentals of the problem. The functions $\alpha_i(\mathbf{h}) C(h_i p_i(\mathbf{h})/N_0)$ are not concave and the set $\mathcal{A}$ is not convex. This makes the program in (30) nonconvex but is consistent with the restrictions imposed in (20).

To write the Lagrangian corresponding to this optimization problem introduce multipliers $\boldsymbol{\lambda} := \left[\lambda_1, \ldots, \lambda_J\right]^T$ associated with the capacity constraints and μ associated with the power constraint. The Lagrangian is then given by

$$\mathcal{L}(\mathbf{r}, \mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\lambda}, \mu) = \sum_{i=1}^{J} U(r_i) + \lambda_i \left[\mathbb{E}\left[\alpha_i(\mathbf{h}) C\left(\frac{h_i p_i(\mathbf{h})}{N_0}\right)\right] - r_i \right]$$
$$+ \mu \left[q_0 - \mathbb{E}\left[\sum_{i=1}^{J} \alpha_i(\mathbf{h}) p_i(\mathbf{h})\right]\right]. \qquad (31)$$

The dual function, dual problem, and optimal dual arguments are defined as in (22)–(25) by replacing $\mathcal{L}(\mathbf{r}, \mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\lambda}, \mu)$ for $\mathcal{L}(\mathbf{x}, \mathbf{p}, \boldsymbol{\Lambda})$. Since (30) is a nonconvex program we do not know if the dual problem is equivalent to the primal problem. We explore this issue in the following section.

Duality in wireless systems optimization

For any optimization problem the dual minimum D provides an upper bound for the primal optimum value P. This is easy to see by combining the definitions of the dual function in (23) and the Lagrangian in (22) to write

$$g(\boldsymbol{\Lambda}) = \max_{\mathbf{x} \in \mathcal{X}, \mathbf{p} \in \mathcal{P}} f_0(\mathbf{x}) + \boldsymbol{\lambda}_1^T \left[\mathbb{E}\left(\mathbf{f}_1\left(\mathbf{h}, \mathbf{p}(\mathbf{h})\right)\right) - \mathbf{x}\right] + \boldsymbol{\lambda}_2^T \mathbf{f}_2(\mathbf{x}). \qquad (32)$$

Because the dual function value $g(\boldsymbol{\Lambda})$ is obtained by maximizing the right hand side of (32), evaluating this expression for arbitrary primal variables yields an upper bound on $g(\boldsymbol{\Lambda})$. Using a pair $(\mathbf{x}^*, \mathbf{p}^*)$ of optimal primal arguments as this arbitrary selection yields the inequality

$$g(\boldsymbol{\Lambda}) \ge f_0(\mathbf{x}^*) + \boldsymbol{\lambda}_1^T \left[\mathbb{E}\left(\mathbf{f}_1\left(\mathbf{h}, \mathbf{p}^*(\mathbf{h})\right)\right) - \mathbf{x}^*\right] + \boldsymbol{\lambda}_2^T \mathbf{f}_2(\mathbf{x}^*). \qquad (33)$$

Since the pair $\mathbf{x}^*, \mathbf{p}^*$ is optimal, it is feasible, which means that we must have $\mathbb{E}\left[\mathbf{f}_1\left(\mathbf{h}, \mathbf{p}^*(\mathbf{h})\right)\right] - \mathbf{x}^* \ge \mathbf{0}$ and $\mathbf{f}_2(\mathbf{x}^*) \ge \mathbf{0}$. Lagrange multipliers are also nonnegative according to definition. Therefore, the last two summands in the right hand side of (33) are nonnegative from which it follows that

$$g(\boldsymbol{\Lambda}) \ge f_0(\mathbf{x}^*) = P. \qquad (34)$$

The inequality in (34) is true for any $\boldsymbol{\Lambda}$ and therefore true for $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}^*$ in particular. It then follows that the dual optimum D upper bounds the primal optimum P,

$$D \ge P, \qquad (35)$$

as we had claimed. The difference D-P is called the duality gap and provides an indication of the loss of optimality incurred by working in the dual domain.

For the problem in (20) the duality gap is null as long as the channel probability distribution $m_{\mathbf{h}}(\mathbf{h})$ contains no point of positive probability as we claim in the following

theorem which is a simple generalization of a similar result in [20].

Theorem 1. *Let P denote the optimum value of the primal problem* (20) *and D that of its dual in* (24) *and assume there exists a strictly feasible point* $(\mathbf{x}_0, \mathbf{p}_0)$ *that satisfies the constraints in* (20) *with strict inequality. If the channel probability distribution* $m_{\mathbf{h}}(\mathbf{h})$ *contains no point of positive probability the duality gap is null, i.e.,*

$$P = D. \tag{36}$$

The condition on the channel distribution not having points of positive probability is a mild requirement satisfied by practical fading channel models including Rayleigh, Rice, and Nakagami. The existence of a strictly feasible point $(\mathbf{x}_0, \mathbf{p}_0)$ is a standard constraint qualification requirement which is also not stringent in practice.

In order to prove Theorem 1 we take a detour in Section "Lyapunov's convexity theorem" to define atomic and nonatomic measures along with the presentation of Lyapunov's Convexity Theorem. The proof itself is presented in Section "Proof of Theorem 1". The implications of Theorem 1 are discussed in Sections "Recovery of optimal primal variables" and "Dual descent algorithms".
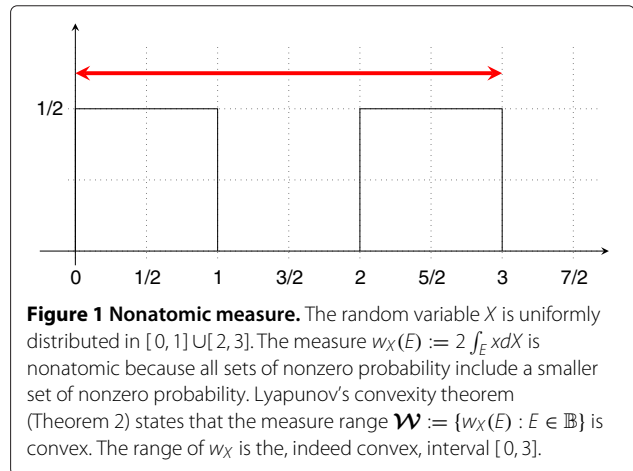
Lyapunov's convexity theorem

The proof of Theorem 1 uses a theorem by Lyapunov concerning the range of nonatomic measures [22]. Measures assign strictly positive values to sets of a Borel field. When all points have zero measure the measure is called nonatomic as we formally define next.

Definition 1 (**Nonatomic measure**). *Let w be a measure defined on the Borel field* $\mathbb{B}$ *of subsets of a space* $\mathcal{X}$*. Measure w is nonatomic if for any measurable set* $E_0 \in \mathbb{B}$ *with* $w(E_0) > 0$*, there exist a subset E of* E_0*; i.e.,* $E \subset E_0$*, such that* $w(E_0) > w(E) > 0$*.*

Familiar measures are probability related, e.g., the probability of a set for a given channel distribution. To build intuition on the notion of nonatomic measure consider a random variable X taking values in $[0, 1]$ and $[2, 3]$. The probability of landing in each of these intervals is $1/2$ and X is uniformly distributed inside each of them; see Figure 1. The space $\mathcal{X}$ is the real line, and the Borel field $\mathbb{B}$ comprises all subsets of real numbers. For every subset $E \in \mathbb{B}$ define the measure of E as twice the integral of x, weighted by the probability distribution of X on the set E, i.e.,

$$w_X(E) := 2 \int_E x dX. \tag{37}$$

Note that, except for the factor 2, the value of $w_X(E)$ represents the contribution of the set E to the expected value of X and that when E is the whole space $\mathcal{X}$, it holds

Figure 1 Nonatomic measure. The random variable X is uniformly distributed in $[0, 1] \cup [2, 3]$. The measure $w_X(E) := 2 \int_E x dX$ is nonatomic because all sets of nonzero probability include a smaller set of nonzero probability. Lyapunov's convexity theorem (Theorem 2) states that the measure range $\mathcal{W} := \{w_X(E) : E \in \mathbb{B}\}$ is convex. The range of w_X is the, indeed convex, interval $[0, 3]$.
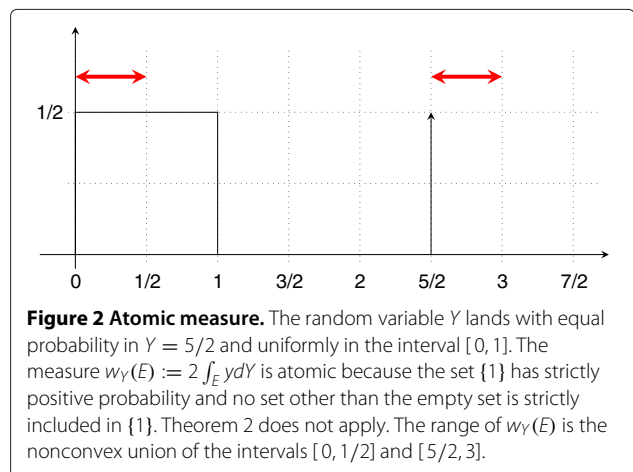
$w_X(\mathcal{X}) = 2\mathbb{E}_X(x)$. According to Definition 1, $w_X(E)$ is a nonatomic measure of elements of $\mathbb{B}$. Every subset E_0 with $w_X(E_0) > 0$ includes at least an interval (a, b). The measure of the set $E := E_0 - ((a + b)/2, b)$ formed by removing the upper half of (a, b) from E_0 is $w_X(E) = w_X(E_0) - (b - a)/2$. The measure of E satisfies $w_X(E) > 0$ as required for $w_X(E)$ to be nonatomic.

To contrast this with an example of an atomic measure consider a random variable Y landing equiprobably in $[0, 1]$ or $5/2$; see Figure 2. In this case, the measure $w_Y(E) := 2 \int_E y dY$ is atomic because the set $E_0 = \{5/2\}$ has positive measure $w_Y(E) = 1$. The only set $E \subset E_0$ is the empty set whose measure is null.

The difference between the distributions of X and Y is that Y contains a point of strictly positive probability, i.e., an atom. This implies presence of delta functions in the probability density function of Y. Or, in a rather cleaner statement the cumulative distribution function (cdf) of X is continuous whereas the cdf of Y is not.

Lyapunov's convexity theorem introduced next refers to the range of values taken by (vector) nonatomic measures.

Figure 2 Atomic measure. The random variable Y lands with equal probability in $Y = 5/2$ and uniformly in the interval $[0, 1]$. The measure $w_Y(E) := 2 \int_E y dY$ is atomic because the set $\{1\}$ has strictly positive probability and no set other than the empty set is strictly included in $\{1\}$. Theorem 2 does not apply. The range of $w_Y(E)$ is the nonconvex union of the intervals $[0, 1/2]$ and $[5/2, 3]$.

Theorem 2 (**Lyapunov's convexity theorem [22]**). *Consider nonatomic measures $w_1, \ldots, w_n$ on the Borel field $\mathbb{B}$ of subsets of a space $\mathcal{X}$ and define the vector measure $\mathbf{w}(E) := [w_1(E), \ldots, w_n(E)]^T$. The range $\mathcal{W} := \{\mathbf{w}(E) : E \in \mathbb{B}\}$ of the measure $\mathbf{w}$ is convex. I.e., if $\mathbf{w}(E_1) = \mathbf{w}_1$ and $\mathbf{w}(E_2) = \mathbf{w}_2$, then for any $\alpha \in [0, 1]$ there exists $E_0 \in \mathbb{B}$ such that $\mathbf{w}(E_0) = \alpha\mathbf{w}_1 + (1-\alpha)\mathbf{w}_2$.*

Returning to the probability measures defined in terms of the probability distributions of the random variables X and Y, Theorem 2 asserts that the range of $w_X(E)$, i.e., the set of all possible values taken by w_X is convex. In fact, it is not difficult to verify that the range of w_X is the convex interval $[0, 3]$ as shown in Figure 1. Theorem 2 does not claim anything about w_Y. In this case, it is easy to see that the range of w_Y is the (non-convex) union of the intervals $[0, 1/2]$ and $[5/2, 3]$; see Figure 2.

Proof of Theorem 1
To establish zero duality gap we will consider a perturbed version of (20) obtained by perturbing the constraints used to define the Lagrangian in (22). The perturbation function $P(\boldsymbol{\Delta})$ assigns to each (perturbation) parameter set $\boldsymbol{\Delta} := \left[\delta_1^T, \delta_2^T\right]^T$ the solution of the (perturbed) optimization problem

$$P(\boldsymbol{\Delta}) = \max \ f_0(\mathbf{x})$$
$$\text{s.t.} \quad \mathbb{E}\left[\mathbf{f}_1\left(\mathbf{h}, \mathbf{p}(\mathbf{h})\right)\right] - \mathbf{x} \geq \delta_1, \quad \mathbf{f}_2(\mathbf{x}) \geq \delta_2,$$
$$\mathbf{x} \in \mathcal{X}, \quad \mathbf{p} \in \mathcal{P}. \tag{38}$$

The perturbed problem in (38) can be interpreted as a modified version of (20), where we allow the constraints to be violated by $\boldsymbol{\Delta}$ amounts. To prove that the duality gap is zero, it suffices to show that $P(\boldsymbol{\Delta})$ is a concave function of $\boldsymbol{\Delta}$; see, e.g., [23].

Let $\boldsymbol{\Delta} = [\delta_1^T, \delta_2^T]^T$ and $\boldsymbol{\Delta}' = [\delta_1'^T, \delta_2'^T]^T$ be an arbitrary given pair of perturbations. Let $(\mathbf{x}, \mathbf{p})$ be a pair of ergodic limits and resource allocation variables achieving the optimum value $P(\boldsymbol{\Delta})$ corresponding to perturbation $\boldsymbol{\Delta}$. Likewise, denote as $(\mathbf{x}', \mathbf{p}')$ a pair that achieves the optimum value $P(\boldsymbol{\Delta}')$ corresponding to perturbation $\boldsymbol{\Delta}'$. For arbitrary $\alpha \in [0, 1]$, we are interested in the solution of (38) under perturbation $\boldsymbol{\Delta}_\alpha := \alpha\boldsymbol{\Delta} + (1-\alpha)\boldsymbol{\Delta}'$. In particular, to show that the perturbation function $P(\boldsymbol{\Delta})$ is concave we need to establish that

$$P(\boldsymbol{\Delta}_\alpha) = P\left[\alpha\boldsymbol{\Delta} + (1-\alpha)\boldsymbol{\Delta}'\right]$$
$$\geq \alpha P(\boldsymbol{\Delta}) + (1-\alpha)P(\boldsymbol{\Delta}'). \tag{39}$$

The roadblock to establish concavity of the perturbation functions is the constraint $\mathbf{x} \leq \mathbb{E}\left[\mathbf{f}_1(\mathbf{h}, \mathbf{p}(\mathbf{h}))\right]$. More specifically, the difficulty with this constraint is the ergodic limit $\mathbb{E}\left[\mathbf{f}_1(\mathbf{h}, \mathbf{p}(\mathbf{h}))\right]$. Let us then isolate the challenge by defining the ergodic limit span

$$\mathcal{Y} := \left\{\mathbf{y} \ \exists \mathbf{p} \in \mathcal{P} \text{ for which } \mathbf{y} = \mathbb{E}\left[\mathbf{f}_1\left(\mathbf{h}, \mathbf{p}(\mathbf{h})\right)\right]\right\}. \tag{40}$$

The set $\mathcal{Y}$ contains all the possible values that the expectation $\mathbb{E}\left[\mathbf{f}_1(\mathbf{h}, \mathbf{p}(\mathbf{h}))\right]$ can take as the resource allocation function $\mathbf{p}$ varies over the admissible set $\mathcal{P}$. When the channel distribution $m_\mathbf{h}(\mathbf{h})$ contains no points of positive probability, the set $\mathcal{Y}$ is convex as we claim in the following theorem.

Theorem 3. *Let $\mathcal{Y}$ be ergodic limit span set in* (40). *If the channel probability distribution $m_\mathbf{h}(\mathbf{h})$ contains no point of positive probability the set $\mathcal{Y}$ is convex.*

Before proving Theorem 3 let us apply it to complete the proof of Theorem 1. For doing that consider two arbitrary points $\mathbf{y} \in \mathcal{Y}$ and $\mathbf{y}' \in \mathcal{Y}$. If these points belong to $\mathcal{Y}$ there exist respective resource allocation functions $\mathbf{p} \in \mathcal{P}$ and $\mathbf{p}' \in \mathcal{P}$ such that

$$\mathbf{y} = \mathbb{E}\left[\mathbf{f}_1(\mathbf{h}, \mathbf{p}(\mathbf{h}))\right], \qquad \mathbf{y}' = \mathbb{E}\left[\mathbf{f}_1(\mathbf{h}, \mathbf{p}'(\mathbf{h}))\right]. \tag{41}$$

If $\mathcal{Y}$ is a convex set as claimed by Theorem 3, we must have that for any given α the point $\mathbf{y}_\alpha := \alpha\mathbf{y} + (1-\alpha)\mathbf{y}'$ also belongs to $\mathcal{Y}$. In turn, this means there exists a resource allocation function $\mathbf{p}_\alpha \in \mathcal{P}$ for which

$$\mathbf{y}_\alpha := \alpha\mathbb{E}\left[\mathbf{f}_1(\mathbf{h}, \mathbf{p}(\mathbf{h}))\right] + (1-\alpha)\mathbb{E}\left[\mathbf{f}_1(\mathbf{h}, \mathbf{p}'(\mathbf{h}))\right]$$
$$= \mathbb{E}\left[\mathbf{f}_1(\mathbf{h}, \mathbf{p}_\alpha(\mathbf{h}))\right]. \tag{42}$$

Further define the ergodic limit convex combination $\mathbf{x}_\alpha := \alpha\mathbf{x} + (1-\alpha)\mathbf{x}'$. We first show that the pair $(\mathbf{x}_\alpha, \mathbf{p}_\alpha)$ is feasible for the problem for perturbation $\boldsymbol{\Delta}_\alpha$.

The convex combination satisfies the constraint $\mathbf{x}_\alpha \in \mathcal{X}$ because the set $\mathcal{X}$ is convex. We also have $\mathbf{f}_2(\mathbf{x}_\alpha) \geq \delta_{2,\alpha}$ because the function $\mathbf{f}_2(\mathbf{x})$ is concave. To see that this is true simply notice that concavity of $\mathbf{f}_2(\mathbf{x})$ implies that $\mathbf{f}_2(\mathbf{x}_\alpha) = \mathbf{f}_2[\alpha\mathbf{x} + (1-\alpha)\mathbf{x}'] \geq \alpha\mathbf{f}_2(\mathbf{x}) + (1-\alpha)\mathbf{f}_2(\mathbf{x}')$. But $\mathbf{f}_2(\mathbf{x}) \geq \delta_2$ and $\mathbf{f}_2(\mathbf{x}') \geq \delta_2'$ because $\mathbf{x}$ and $\mathbf{x}'$ are feasible in (38) with perturbations $\boldsymbol{\Delta}$ and $\boldsymbol{\Delta}'$. Substituting these latter two inequalities into the previous one yields $\mathbf{f}_2(\mathbf{x}_\alpha) \geq \alpha\delta_2 + (1-\alpha)\delta_2' = \delta_{2,\alpha}$ according to the definition of $\boldsymbol{\Delta}_\alpha$.

We are left to show that the pair $(\mathbf{x}_\alpha, \mathbf{p}_\alpha)$ satisfies the constraint $\mathbb{E}\left[\mathbf{f}_1(\mathbf{h}, \mathbf{p}(\mathbf{h}))\right] - \mathbf{x} \leq \delta_{1,\alpha}$. For doing so recall that since $(\mathbf{x}, \mathbf{p})$ is feasible for perturbation Δ and $(\mathbf{x}', \mathbf{p}')$ is feasible for perturbation Δ' we must have

$$\mathbb{E}\left[\mathbf{f}_1(\mathbf{h}, \mathbf{p}(\mathbf{h}))\right] - \mathbf{x} \geq \delta_1,$$
$$\mathbb{E}\left[\mathbf{f}_1(\mathbf{h}, \mathbf{p}'(\mathbf{h}))\right] - \mathbf{x}' \geq \delta_1'. \tag{43}$$

Perform a convex combination of the inequalities in (43) and use the definitions of $\mathbf{x}_\alpha := \alpha\mathbf{x} + (1-\alpha)\mathbf{x}'$ and $\delta_{1,\alpha} := \alpha\delta_1 + (1-\alpha)\delta_1'$ to write

$$\alpha\mathbb{E}\left[\mathbf{f}_1(\mathbf{h}, \mathbf{p}(\mathbf{h}))\right] + (1-\alpha)\mathbb{E}\left[\mathbf{f}_1(\mathbf{h}, \mathbf{p}'(\mathbf{h}))\right] - \mathbf{x}_\alpha \geq \delta_{1,\alpha}. \tag{44}$$

Combining (42) and (44) it follows that $\mathbb{E}\left[\mathbf{f}_1(\mathbf{h}, \mathbf{p}_\alpha(\mathbf{h}))\right] - \mathbf{x}_\alpha \leq \delta_{1,\alpha}$ completing the proof that the pair $(\mathbf{x}_\alpha, \mathbf{p}_\alpha)$ is feasible for the problem for perturbation $\boldsymbol{\Delta}_\alpha$.

The utility yield of this feasible pair is $f_0(\mathbf{x}_\delta)$ which we can bound as

$$f_0(\mathbf{x}_\alpha) \geq \alpha f_0(\mathbf{x}) + (1 - \alpha)f_0(\mathbf{x}'). \qquad (45)$$

Since $(\mathbf{x}, \mathbf{p})$ is optimal for perturbation $\boldsymbol{\Delta}$ we have $f_0(\mathbf{x}) = P(\boldsymbol{\Delta})$ and, likewise, $f_0(\mathbf{x}') = P(\boldsymbol{\Delta}')$. Further noting that the optimal yield $P(\boldsymbol{\Delta}_\alpha)$ for perturbation $\boldsymbol{\Delta}_\alpha$ must exceed the yield $f_0(\mathbf{x}_\alpha)$ of feasible point $\mathbf{x}_\alpha$ we conclude that

$$P(\boldsymbol{\Delta}_\alpha) \geq \alpha P(\boldsymbol{\Delta}) + (1 - \alpha)P(\boldsymbol{\Delta}'). \qquad (46)$$

The expression in (46) implies that $P(\boldsymbol{\Delta})$ is a concave function of the perturbation vector $\boldsymbol{\Delta}$. The duality gap is therefore null as stated in (36).

We proceed now to the proof of Theorem 3.

Proof. (Theorem 3) Consider two arbitrary points $\mathbf{y} \in \mathcal{Y}$ and $\mathbf{y}' \in \mathcal{Y}$. If these points belong tho $\mathcal{Y}$ there exist respective resource allocation functions $\mathbf{p} \in \mathcal{P}$ and $\mathbf{p}' \in \mathcal{P}$ such that

$$\mathbf{y} = \mathbb{E}\left[\mathbf{f}_1(\mathbf{h}, \mathbf{p}(\mathbf{h}))\right], \qquad \mathbf{y}' = \mathbb{E}\left[\mathbf{f}_1(\mathbf{h}, \mathbf{p}'(\mathbf{h}))\right]. \qquad (47)$$

To prove that $\mathcal{Y}$ is a convex set we need to show that for any given α the point $\mathbf{y}_\alpha := \alpha\mathbf{y} + (1 - \alpha)\mathbf{y}'$ also belongs to $\mathcal{Y}$. In turn, this means we need to find a resource allocation function $\mathbf{p}_\alpha \in \mathcal{P}$ for which

$$\begin{aligned} \mathbf{y}_\alpha &:= \alpha\mathbb{E}\left[\mathbf{f}_1\left(\mathbf{h}, \mathbf{p}(\mathbf{h})\right)\right] + (1-\alpha)\mathbb{E}\left[\mathbf{f}_1\left(\mathbf{h}, \mathbf{p}'(\mathbf{h})\right)\right] \\ &= \mathbb{E}\left[\mathbf{f}_1(\mathbf{h}, \mathbf{p}_\alpha(\mathbf{h}))\right]. \end{aligned} \qquad (48)$$

For this we will use Theorem 2 (Lyapunov's convexity theorem). Consider the space of all possible channel realizations $\mathcal{H}$, and the Borel field $\mathbb{B}$ of all possible subsets of $\mathcal{H}$. For every s et $E \in \mathbb{B}$ define the vector measure

$$\mathbf{w}(E) := \left[\int_E \mathbf{f}_1(\mathbf{h}, \mathbf{p}(\mathbf{h}))dm_\mathbf{h}, \int_E \mathbf{f}_1(\mathbf{h}, \mathbf{p}'(\mathbf{h}))dm_\mathbf{h}\right], \qquad (49)$$

where the integrals are over the set E with respect to the channel distribution $m_\mathbf{h}(\mathbf{h})$. A vector of channel realizations $\mathbf{h} \in \mathcal{H}$ is a point in the space $\mathcal{H}$. The set E is a collection of vectors $\mathbf{h}$. Each of these sets is assigned vector measure $\mathbf{w}(E)$ defined in terms of the power distributions $\mathbf{p}(\mathbf{h})$ and $\mathbf{p}'(\mathbf{h})$. The entries of $\mathbf{w}(E)$ represent the contribution of realizations $\mathbf{h} \in E$ to the ergodic limits in (48). The first group of entries measure such contributions when the resource allocation is $\mathbf{p}(\mathbf{h})$. Likewise, the second group of entries of $\mathbf{w}(E)$ denote the contributions to the ergodic limits of the resource distribution $\mathbf{p}'(\mathbf{h})$.

Two particular sets that are important for this proof are the empty set, $E = \emptyset$, and the entire space $E = \mathcal{H}$. For $E = \mathcal{H}$, the integrals $\int_E (\cdot)dm_\mathbf{h} = \int_\mathcal{H} (\cdot)dm_\mathbf{h}$ in (49) coincide

with the expected value operators $\mathbb{E}(\cdot)$ in (47). We write this explicitly as

$$\mathbf{w}(\mathcal{H}) = \left[\mathbb{E}\left[\mathbf{f}_1(\mathbf{h}, \mathbf{p}(\mathbf{h}))\right], \mathbb{E}\left[\mathbf{f}_1(\mathbf{h}, \mathbf{p}'(\mathbf{h}))\right]\right] = [\mathbf{y}, \mathbf{y}']. \qquad (50)$$

For $E = \emptyset$, or any other zero-measure set for that matter, we have $\mathbf{w}(\emptyset) = \mathbf{0}$.

The measure $\mathbf{w}(E)$ is nonatomic. This follows from the fact that the channel distribution contains no points of positive probability combined with the requirement that the resource allocation values $\mathbf{p}(\mathbf{h})$ belong to the bounded sets $\mathcal{P}(\mathbf{h})$. Combining these two properties it is easy to see that that there are no channel realizations with positive measure, i.e., $\mathbf{w}(\mathbf{h}) = 0$ for all $\mathbf{h} \in \mathcal{H}$. This is sufficient to ensure that $\mathbf{w}(E)$ is a nonatomic measure.

Being $\mathbf{w}(E)$ nonatomic it follows from Theorem 2 that the range of $\mathbf{w}$ is convex. Hence, the vector

$$\mathbf{w}_0 = \alpha\mathbf{w}(\mathcal{H}) + (1 - \alpha)\mathbf{w}(\emptyset) = \alpha\mathbf{w}(\mathcal{H}), \qquad (51)$$

belongs to the range of possible measures. Therefore, there must exist a set E_0 such that $\mathbf{w}(E_0) = \mathbf{w}_0 = \alpha\mathbf{w}(\mathcal{H})$. Focusing on the entries of $\mathbf{w}(E_0)$ that correspond to the resource allocation function $\mathbf{p}$ it follows that

$$\int_{E_0} \mathbf{f}_1(\mathbf{h}, \mathbf{p}(\mathbf{h}))dm_\mathbf{h} = \alpha\mathbb{E}\left[\mathbf{f}_1(\mathbf{h}, \mathbf{p}(\mathbf{h}))\right]. \qquad (52)$$

The analogous relation holds for the entries of $\mathbf{w}(E_0)$ corresponding to $\mathbf{p}'$, i.e., (52) is valid if $\mathbf{p}(\mathbf{h})$ is replaced by $\mathbf{p}'(\mathbf{h})$ but this fact is inconsequential for this proof.

Consider now the complement set E_0^c defined as the set for which $E_0 \cup E_0^c = \mathcal{H}$ and $E_0 \cap E_0^c = \emptyset$. Given this definition and the additivity property of measures, we arrive at $\mathbf{w}(E_0) + \mathbf{w}(E_0^c) = \mathbf{w}(\mathcal{H})$. Combining the latter with (51), yields

$$\mathbf{w}(E_0^c) = \mathbf{w}(\mathcal{H}) - \mathbf{w}(E_0) = (1 - \alpha)\mathbf{w}(\mathcal{H}). \qquad (53)$$

We mimic the reasoning leading from (51) to (52), but now we restrict the focus of (53) to the *second* entries of $\mathbf{w}(E_0^c)$. It therefore follows that

$$\int_{E_0^c} \mathbf{f}_1(\mathbf{h}, \mathbf{p}'(\mathbf{h}))dm_\mathbf{h} = (1 - \alpha)\mathbb{E}\left[\mathbf{f}_1(\mathbf{h}, \mathbf{p}'(\mathbf{h}))\right]. \qquad (54)$$

Define now power distributions $\mathbf{p}_\alpha(\mathbf{h})$ coinciding with $\mathbf{p}(\mathbf{h})$ for channel realization $\mathbf{h} \in E_0$ and with $\mathbf{p}'(\mathbf{h})$ when $\mathbf{h} \in E_0^c$, i.e.,

$$\mathbf{p}_\alpha(\mathbf{h}) = \begin{cases} \mathbf{p}(\mathbf{h}) & \mathbf{h} \in E_0, \\ \mathbf{p}'(\mathbf{h}) & \mathbf{h} \in E_0^c. \end{cases} \qquad (55)$$

The resource distribution $\mathbf{p}_\alpha$ satisfies the set constraint $\mathbf{p}_\alpha \in \mathcal{P}$ in (19). Indeed, to see that $\mathbf{p}_\alpha(\mathbf{h}) \in \mathcal{P}(\mathbf{h})$ for all $\mathbf{h} \in \mathcal{H}$ note that $\mathbf{p}(\mathbf{h})$ and $\mathbf{p}'(\mathbf{h})$ are feasible in their respective problems and as such $\mathbf{p}(\mathbf{h}) \in \mathcal{P}(\mathbf{h})$ and $\mathbf{p}'(\mathbf{h}) \in \mathcal{P}(\mathbf{h})$

for all channels $\mathbf{h} \in \mathcal{H}$. Because for given channel realization $\mathbf{h}$ it holds that either $\mathbf{p}_\alpha(\mathbf{h}) = \mathbf{p}(\mathbf{h})$ when $\mathbf{h} \in E_0$ or $\mathbf{p}_\alpha(\mathbf{h}) = \mathbf{p}'(\mathbf{h})$ when $\mathbf{h} \in E_0^c$ it follows that $\mathbf{p}_\alpha(\mathbf{h}) \in \mathcal{P}(\mathbf{h})$ for all channel realizations $\mathbf{h} \in E_0 \cup E_0^c = \mathcal{H}$. According to the definition in (19) this implies that $\mathbf{p}_\alpha \in \mathcal{P}$.

Let us now ponder the ergodic limit $\mathbb{E}\left[\mathbf{f}_1(\mathbf{h}, \mathbf{p}_\alpha(\mathbf{h}))\right]$ associated with power allocation $\mathbf{p}_\alpha(\mathbf{h})$.

Using (52) and (54), the average link capacities for power allocation $\mathbf{p}_\alpha(\mathbf{h})$ can be expressed in terms of $\mathbf{p}(\mathbf{h})$, $\mathbf{p}'(\mathbf{h})$ as

$$
\begin{aligned}
& \mathbb{E}\left[\mathbf{f}_1\ (\mathbf{h}, \mathbf{p}_\alpha(\mathbf{h}))\right] \\
& = \int_{E_0} \mathbf{f}_1(\mathbf{h}, \mathbf{p}_\alpha(\mathbf{h})) dm_{\mathbf{h}} + \int_{E_0^c} \mathbf{f}_1(\mathbf{h}, \mathbf{p}_\alpha(\mathbf{h})) dm_{\mathbf{h}} \\
& = \int_{E_0} \mathbf{f}_1(\mathbf{h}, \mathbf{p}(\mathbf{h})) dm_{\mathbf{h}} + \int_{E_0^c} \mathbf{f}_1(\mathbf{h}, \mathbf{p}'(\mathbf{h})) dm_{\mathbf{h}} \\
& = \alpha \mathbb{E}\left[\mathbf{f}_1(\mathbf{h}, \mathbf{p}(\mathbf{h}))\right] + (1 - \alpha) \mathbb{E}\left[\mathbf{f}_1(\mathbf{h}, \mathbf{p}'(\mathbf{h}))\right].
\end{aligned}
\tag{56}
$$

The first equality in (56) holds becauset the space $\mathcal{H}$ is divided into E_0 and its complement E_0^c. The second equality is true because when restricted to E_0, $\mathbf{p}_\alpha(\mathbf{h}) = \mathbf{p}(\mathbf{h})$; and when restricted to E_0^c, $\mathbf{p}_\alpha(\mathbf{h}) = \mathbf{p}'(\mathbf{h})$. The third equality follows from (52) and (54).

Comparing (56) with (48) we see that the power allocation $\mathbf{p}_\alpha$ yields ergodic limit $\mathbf{y}_\alpha$. Therefore $\mathbf{y}_\alpha \in \mathcal{Y}$ implying convexity of $\mathcal{Y}$ as we wanted to show. □

Recovery of optimal primal variables

Having null duality gap means that we can work in the dual domain without loss of optimality. In particular, instead of finding the primal maximum P we can find the dual minimum D, which we know are the same according to Theorem 1. A not so simple matter is how to recover an optimal primal pair $(\mathbf{x}^*, \mathbf{p}^*)$ given an optimal dual vector $\mathbf{\Lambda}^*$. Observe that recovering optimal variables is more important than knowing the maximum yield because optimal variables are needed for system implementation. In this section we study the recovery of optimum primal variables $(\mathbf{x}^*, \mathbf{p}^*)$ from a given optimal multiplier $\mathbf{\Lambda}^*$.

Start with an arbitrary, not necessarily optimal, multiplier $\mathbf{\Lambda}$ and define the primal Lagrangian maximizer set as

$$
(\mathbf{x}(\mathbf{\Lambda}), \mathbf{p}(\mathbf{\Lambda})) = \underset{x \in \mathcal{X}, \mathbf{p} \in \mathcal{P}}{\text{argmax}}\ \mathcal{L}(\mathbf{x}, \mathbf{p}, \mathbf{\Lambda}).
\tag{57}
$$

The elements of $(\mathbf{x}(\mathbf{\Lambda}), \mathbf{p}(\mathbf{\Lambda}))$ yield the maximum possible Lagrangian value for given dual variable. Comparing the definition of the dual function in (23) and that of the

Lagrangian maximizers in (57) it follows that we can write the dual function $g(\mathbf{\Lambda})$ as

$$
g(\mathbf{\Lambda}) = \mathcal{L}(\mathbf{x}(\mathbf{\Lambda}), \mathbf{p}(\mathbf{\Lambda}), \mathbf{\Lambda}).
\tag{58}
$$

Particular important pairs of Lagrangian maximizers are those associated with a given optimal dual variable $\mathbf{\Lambda}^*$. As we show in the following theorem, these variables are related with the optimal primal variables $(\mathbf{x}^*, \mathbf{p}^*)$.

Theorem 4. *For an optimization problem of the form in* (20) *let* $\left(\mathbf{x}(\mathbf{\Lambda}^*), \mathbf{p}(\mathbf{\Lambda}^*)\right)$ *denote the Lagrangian maximizer set as defined in* (57) *corresponding to a given optimal multiplier* $\mathbf{\Lambda}^*$. *The optimal argument set* $(\mathbf{x}^*, \mathbf{p}^*)$ *is included in this set of Lagrangian maximizers, i.e.,*

$$
(\mathbf{x}^*, \mathbf{p}^*) \subseteq \left(\mathbf{x}(\mathbf{\Lambda}^*), \mathbf{p}(\mathbf{\Lambda}^*)\right).
\tag{59}
$$

Proof. Reinterpret $(\mathbf{x}^*, \mathbf{p}^*)$ as a particular pair of optimal primal arguments. Start by noting that the value of the dual function $g(\mathbf{\Lambda}^*)$ can be upper bounded by the Lagrangian evaluated at $(\mathbf{x}, \mathbf{p}) = (\mathbf{x}^*, \mathbf{p}^*)$

$$
g(\mathbf{\Lambda}^*) = \max_{x \in \mathcal{X}, p \in \mathcal{P}} \mathcal{L}(\mathbf{x}, \mathbf{p}, \mathbf{\Lambda}^*) \geq \mathcal{L}(\mathbf{x}^*, \mathbf{p}^*, \mathbf{\Lambda}^*).
\tag{60}
$$

Indeed, the inequality would be true for any $x \in \mathcal{X}$ and $\mathbf{p} \in \mathcal{P}$ because that is what being the maximum means.

Consider now the Lagrangian $\mathcal{L}(\mathbf{x}^*, \mathbf{p}^*, \mathbf{\Lambda}^*)$ that according to (22) we can explicitly write as

$$
\begin{aligned}
& \mathcal{L}(\mathbf{x}^*, \mathbf{p}^*, \mathbf{\Lambda}^*) \quad\quad\quad\quad\quad\quad\quad\quad\quad\quad (61) \\
& = f_0(\mathbf{x}^*) + \lambda_1^{*T}\left[\mathbb{E}\left(\mathbf{f}_1\left(\mathbf{h}, \mathbf{p}^*(\mathbf{h})\right)\right) - \mathbf{x}^*\right] + \lambda_2^{*T}\mathbf{f}_2(\mathbf{x}^*).
\end{aligned}
$$

Since the pair $(\mathbf{x}^*, \mathbf{p}^*)$ is an optimal argument of (20) we must have $\mathbb{E}\left(\mathbf{f}_1\left(\mathbf{h}, \mathbf{p}^*(\mathbf{h})\right)\right) - \mathbf{x}^* \geq \mathbf{0}$ and $\mathbf{f}_2(\mathbf{x}^*) \geq \mathbf{0}$. The multipliers also satisfy $\lambda_1^* \geq \mathbf{0}$ and $\lambda_2^* \geq \mathbf{0}$ because they are required to be nonnegative. Thus, the last two summands in (61) are nonnegative from where it follows that

$$
\mathcal{L}(\mathbf{x}^*, \mathbf{p}^*, \mathbf{\Lambda}^*) \geq f_0(\mathbf{x}^*).
\tag{62}
$$

Combining (60) and (62) and using the definitions $g(\mathbf{\Lambda}^*)$ and $P = f_0(\mathbf{x}^*)$ yields

$$
D = g(\mathbf{\Lambda}^*) \geq \mathcal{L}(\mathbf{x}^*, \mathbf{p}^*, \mathbf{\Lambda}^*) \geq f_0(\mathbf{x}^*) = P.
\tag{63}
$$

But since according to Theorem 1 the duality gap is null $D = P$ and the inequalities must hold hold as equalities. In particular, we have

$$
g(\mathbf{\Lambda}^*) = \mathcal{L}(\mathbf{x}^*, \mathbf{p}^*, \mathbf{\Lambda}^*),
\tag{64}
$$

which according to (58) means the pair $(\mathbf{p}^*, \mathbf{\Lambda}^*)$ is a Lagrangian maximizer associated with $\mathbf{\Lambda} = \mathbf{\Lambda}^*$. Since this is true for any pair of optimal variables $(\mathbf{p}^*, \mathbf{\Lambda}^*)$ it follows that the set $(\mathbf{p}^*, \mathbf{\Lambda}^*)$ is included in the set of Lagrangian maximizers $\left(\mathbf{x}(\mathbf{\Lambda}^*), \mathbf{p}(\mathbf{\Lambda}^*)\right)$ as stated in (59). □

According to (59) optimal arguments $(\mathbf{x}^*, \mathbf{p}^*)$ can be recovered from Lagrangian maximizers $\left(\mathbf{x}(\mathbf{\Lambda}^*), \mathbf{p}(\mathbf{\Lambda}^*)\right)$ associated with optimal multipliers. One has to take care

to interpret this set inclusion properly. Equation (59) does not mean that we can always compute $(\mathbf{x}^*, \mathbf{p}^*)$ by finding Lagrangian maximizers and as such it may or may not be a useful result. If the Lagrangian maximizer pair is unique then the set $\left(\mathbf{x}(\boldsymbol{\Lambda}^*), \mathbf{p}(\boldsymbol{\Lambda}^*)\right)$ is a singleton and by extension so is the (included) set $(\mathbf{x}^*, \mathbf{p}^*)$ of optimal primal variables. In this case Lagrangian maximizers can be used as proxies for optimal arguments. When the set $\left(\mathbf{x}(\boldsymbol{\Lambda}^*), \mathbf{p}(\boldsymbol{\Lambda}^*)\right)$ is not a singleton this is not possible and recovering optimal primal variables from optimal multipliers $\boldsymbol{\Lambda}^*$ is somewhat more difficult.

In general, problems in optimal wireless networking are such that the Lagrangian maximizer resource allocation functions $\mathbf{p}(\boldsymbol{\Lambda}^*)$ are unique to within a set of zero measure. The ergodic limit Lagrangian maximizers $\mathbf{x}(\boldsymbol{\Lambda}^*)$, however, are not unique in many cases. This is more a nuisance than a fundamental problem. Algorithms that find primal optimal operating points regardless of the characteristics of the set of Lagrangian maximizers are studied in Section "Dual descent algorithms".

Separability

To determine the primal Lagrangian maximizers in (57) it is convenient to reorder terms in the definition in (22) to write

$$
\begin{aligned}
\mathcal{L}(\mathbf{x}, \mathbf{p}, \boldsymbol{\Lambda}) & \qquad (65) \\
:= f_0(\mathbf{x}) - \boldsymbol{\lambda}_1^T \mathbf{x} + \boldsymbol{\lambda}_2^T \mathbf{f}_2(\mathbf{x}) & + \left[\mathbb{E}\left(\boldsymbol{\lambda}_1^T \mathbf{f}_1\left(\mathbf{h}, \mathbf{p}(\mathbf{h})\right) \right) \right].
\end{aligned}
$$

We say that in (22) the Lagrangian is grouped by dual variables because each dual variable appears in only one summand. The expression in (65) is grouped by primal variables because each summand contains a single primal variable if we interpret $f_0(\mathbf{x}) - \boldsymbol{\lambda}_1^T \mathbf{x} + \boldsymbol{\lambda}_2^T \mathbf{f}_2(\mathbf{x})$ as a single term and the expectation as a weighted sum—which is not true, but close enough for an intuitive interpretation.

Writing the Lagrangian as in (65) simplifies computation of Lagrangian maximizers. Since there are no constraints coupling the selection of optimal $\mathbf{x}$ and $\mathbf{p}$ in (65) we can separate the determination of the pair $(\mathbf{x}(\boldsymbol{\Lambda}), \mathbf{p}(\boldsymbol{\Lambda}))$ in (57) into the determination of the ergodic limits

$$
\mathbf{x}(\boldsymbol{\Lambda}) = \operatorname*{argmax}_{x \in \mathcal{X}} f_0(\mathbf{x}) - \boldsymbol{\lambda}_1^T \mathbf{x} + \boldsymbol{\lambda}_2^T \mathbf{f}_2(\mathbf{x}) \qquad (66)
$$

and the resource allocation function

$$
\mathbf{p}(\boldsymbol{\Lambda}) = \operatorname*{argmax}_{\mathbf{p} \in \mathcal{P}} \mathbb{E}\left(\boldsymbol{\lambda}_1^T \mathbf{f}_1\left(\mathbf{h}, \mathbf{p}(\mathbf{h})\right) \right). \qquad (67)
$$

The computation of the resource allocation function in (67) can be further separated. The set $\mathcal{P}$ as defined in (19) constrains separate values $\mathbf{p}(\mathbf{h})$ of the function $\mathbf{p}$ but doesn't couple the selection of $\mathbf{p}(\mathbf{h}_1)$ and $\mathbf{p}(\mathbf{h}_2)$ for different channel realizations $\mathbf{h}_1 \neq \mathbf{h}_2$. Further observing that

expectation is a linear operation we can separate (67) into per-fading state subproblems of the form

$$
\mathbf{p}(\mathbf{h}, \boldsymbol{\Lambda}) = \operatorname*{argmax}_{\mathbf{p}(\mathbf{h}) \in \mathcal{P}(\mathbf{h})} \boldsymbol{\lambda}_1^T \mathbf{f}_1\left(\mathbf{h}, \mathbf{p}(\mathbf{h})\right). \qquad (68)
$$

The absence of coupling constraints in the Lagrangian maximization, which permits separation into the ergodic limit maximization in (66) and the per-fading maximizations in (68), is the fundamental difference between (57) and the primal problem in (20). In the latter, the selection of optimal $\mathbf{x}^*$ and $\mathbf{p}^*$ as well as the selection of $\mathbf{p}(\mathbf{h}_1)$ and $\mathbf{p}(\mathbf{h}_2)$ for different channel realizations $\mathbf{h}_1 \neq \mathbf{h}_2$ are coupled by the constraint $\mathbf{x} \leq \mathbb{E}\left[\mathbf{f}_1\left(\mathbf{h}, \mathbf{p}(\mathbf{h})\right)\right]$.

The decomposition in (68) is akin to the decomposition in (16) for the particular case of a point to point channel with capacity achieving codes. It implies that to determine the optimal power allocation $\mathbf{p}^*$ it is convenient to first determine the optimal multiplier $\boldsymbol{\Lambda}^*$. We then proceed to exploit the lack of duality gap and the separable structure of the Lagrangian to compute values $\mathbf{p}^*(\mathbf{h})$ of the optimal resource allocation independently of each other. The computation of optimal ergodic averages is also separated as stated in (66). This separation reduces computational complexity because of the reduced dimensionality of (66) and (68) with respect to that of (20).

For the separation in (66) and (68) to be possible we just need to have a nonatomic channel distribution and ensure existence of a strictly feasible point as stated in the hypotheses of Theorem 1. These two properties are true for most optimal wireless communication and networking problems. A particular example is discussed in the following section.

Frequency division broadcast channel

Consider the optimal frequency division broadcast channel problem in (30) whose Lagrangian $\mathcal{L}(\mathbf{r}, \mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\lambda}, \mu)$ is given by the expression in (31). Terms in $\mathcal{L}(\mathbf{r}, \mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\lambda}, \mu)$ can be rearranged to uncover the separable structure of the Lagrangian

$$
\begin{aligned}
\mathcal{L}(\mathbf{r}, \mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\lambda}, \mu) = & \sum_{i=1}^{J}\left(U(r_i) - \lambda_i r_i\right) + \mu q_0 \\
& - \mathbb{E}\left[\sum_{i=1}^{J} \alpha_i(\mathbf{h})\left[\lambda_i C\left(\frac{h_i p_i(\mathbf{h})}{N_0} \right) - \mu p_i(\mathbf{h}) \right] \right].
\end{aligned} \qquad (69)
$$

This rearrangement is equivalent to the generic transformation leading from (22) to (65). As we observed after (65) the computation of Lagrangian maximizing ergodic limits and resource allocation functions can be separated as can the computation of resource allocation values corresponding to different fading channel realizations. This is the case in this particular example. In fact, there is more separability to be exploited in (69).

With regards to primal ergodic variables $\mathbf{r}$ notice that each Lagrangian maximizing rate $r_i(\boldsymbol{\lambda}, \boldsymbol{\mu})$ depends only on λ_i and that we can compute each $r_i(\boldsymbol{\lambda}, \boldsymbol{\mu}) = r_i(\lambda_i)$ separately as

$$r_i(\lambda_i) = \operatorname*{argmax}_{r_i \in [r_{\min}, r_{\max}]} U(r_i) - \lambda_i r_i. \tag{70}$$

This can be easily computed as $U(r_i)$ is a one dimensional concave function. As a particular case consider the identity utility $U(r_i) = r_i$. Since the Lagrangian becomes a linear function of r_i, the maximum occurs at either $r_{\max}$ or $r_{\min}$ depending on the sign of $1 - \lambda_i$. When $\lambda_i = 1$ the Lagrangian becomes independent of c_i. In this case any value in the interval $[r_{\min}, r_{\max}]$ is a Lagrangian maximizer. Putting these observations together we have

$$r_i(\lambda_i) = \begin{cases} r_{\max} & \text{if } \lambda_i < 1, \\ [r_{\min}, r_{\max}] & \text{if } \lambda_i = 1, \\ r_{\min} & \text{if } \lambda_i > 1. \end{cases} \tag{71}$$

Notice that the Lagrangian maximizer $r_i(\lambda_i)$ is not unique if $\lambda_i = 1$. Therefore, if $\lambda_i^* = 1$ it is not possible to recover the optimal rate r_i^* from the Lagrangian maximizer $r_i(\lambda_i^*)$ corresponding to the optimal multiplier. In fact, an optimal multiplier $\lambda_i^* = 1$ is uninformative with regards to the optimal ergodic rate as it just implies that $r_i^* \in [r_{\min}, r_{\max}]$ which we know is true because this is the feasible range of r_i. If you think this is an unlikely scenario because it is too much of a coincidence to have $\lambda_i^* = 1$, think again. Having $\lambda_i^* = 1$ is the most likely situation. If $\lambda_i^* \neq 1$ the optimal rate r_i^* is either $r_i^* = r_{\max}$ or $r_i^* = r_{\min}$. However, the capacity bounds $r_{\min}$ and $r_{\max}$ are selected independently of the remaining system parameters. It is quite unlikely—indeed, not true in most cases—that the optimal power and frequency allocation yields a rate r_i^* determined by these arbitrarily selected parameters.

As we observed in going from (67) to (68) determination of the optimal power and frequency allocation functions requires maximization of the terms inside the expectation. These implies solution of the optimization problems

$$\mathbf{p}(\mathbf{h}, \boldsymbol{\lambda}, \boldsymbol{\mu}), \boldsymbol{\alpha}(\mathbf{h}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \operatorname{argmax}$$

$$\sum_{i=1}^{J} \alpha_i(\mathbf{h}) \left[\lambda_i C \left(\frac{h_i p_i(\mathbf{h})}{N_0} \right) - \mu p_i(\mathbf{h}) \right].$$

$$\text{s.t. } \alpha_i(\mathbf{h}) \in \{0, 1\}, \quad \mathbf{1}^T \boldsymbol{\alpha}(\mathbf{h}) \leq 1, \tag{72}$$

where we opened up the constraint $\boldsymbol{\alpha} \in \mathcal{A}$ into its per-fading state components.

The maximization in (72) can be further simplified. Begin by noting that irrespectively of the value of $\boldsymbol{\alpha}(\mathbf{h})$ the best possible power allocation $p_i(\mathbf{h}, \boldsymbol{\lambda}, \boldsymbol{\mu})$ of terminal i is

the one that maximizes its potential contribution to the sum in (72), i.e.,

$$p_i(\mathbf{h}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \operatorname*{argmax}_{p_i(\mathbf{h})} \left[\lambda_i C \left(\frac{h_i p_i(\mathbf{h})}{N_0} \right) - \mu p_i(\mathbf{h}) \right]. \tag{73}$$

If $\alpha_i(\mathbf{h}) = 1$ this contribution *is* added to the sum in (72). If it is multiplied by $\alpha_i(\mathbf{h}) = 0$ it *is not* added to the sum in (72). Either way $p_i(\mathbf{h}, \boldsymbol{\lambda}, \boldsymbol{\mu})$ as given by (73) is the optimal power allocation for terminal i.

To determine the frequency allocation $\boldsymbol{\alpha}(\mathbf{h}, \boldsymbol{\lambda}, \boldsymbol{\mu})$ define the discriminants

$$d_i(\mathbf{h}, \boldsymbol{\lambda}, \boldsymbol{\mu}) := \max_{p_i(\mathbf{h})} \left[\lambda_i C \left(\frac{h_i p_i(\mathbf{h})}{N_0} \right) - \mu p_i(\mathbf{h}) \right], \tag{74}$$

which we can use to rewrite (72) as

$$\boldsymbol{\alpha}(\mathbf{h}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \operatorname{argmax} \sum_{i=1}^{J} \alpha_i(\mathbf{h}) d_i(\mathbf{h}, \boldsymbol{\lambda}, \boldsymbol{\mu}).$$

$$\text{s.t.} \quad \alpha_i(\mathbf{h}) \in \{0, 1\}, \quad \mathbf{1}^T \boldsymbol{\alpha}(\mathbf{h}) \leq 1. \tag{75}$$

Since at most one $\alpha_i(\mathbf{h}) = 1$ in (75), the best we can do is to select the terminal with the largest discriminant when that discriminant is positive. If all discriminants are negative the best we can do is to make $\alpha_i(\mathbf{h}) = 0$ for all i.

The Lagrangian maximizers $p_i(\mathbf{h}, \boldsymbol{\lambda}, \boldsymbol{\mu})$ and $\boldsymbol{\alpha}(\mathbf{h}, \boldsymbol{\lambda}, \boldsymbol{\mu})$ in (73) and (75) are almost surely unique for all values of $\boldsymbol{\lambda}$ and μ. In particular, optimal allocations $p_i^*(\mathbf{h})$ and $\boldsymbol{\alpha}^*(\mathbf{h})$ can be obtained by making $\boldsymbol{\lambda} = \boldsymbol{\lambda}^*$ and $\boldsymbol{\mu} = \boldsymbol{\mu}^*$ in (73)–(75).

Dual descent algorithms

Determining optimal dual variables $\boldsymbol{\Lambda}^*$ is easier than determining optimal primal pairs $(\mathbf{p}^*, \mathbf{x}^*)$ because there are a finite number of multipliers and the dual function is convex. If the dual function is convex descent algorithms are guaranteed to converge towards the optimum value, which implies we just need to determine descent directions for the dual function.

Descent directions for the dual function can be constructed from the constraint slacks associated with the Lagrangian maximizers. To do so, consider a given $\boldsymbol{\Lambda}$ and use the definition of the Lagrangian maximizer pair $(\mathbf{x}(\boldsymbol{\Lambda}), \mathbf{p}(\boldsymbol{\Lambda}))$ in (57) to write the dual function as

$$g(\boldsymbol{\Lambda}) = f_0(\mathbf{x}(\boldsymbol{\Lambda})) \tag{76}$$
$$+ \boldsymbol{\lambda}_1^T \left[\mathbb{E} \left(\mathbf{f}_1 \left(\mathbf{h}, \mathbf{p}(\mathbf{h}, \boldsymbol{\Lambda}) \right) \right) - \mathbf{x}(\boldsymbol{\Lambda}) \right] + \boldsymbol{\lambda}_2^T \mathbf{f}_2(\mathbf{x}(\boldsymbol{\Lambda})).$$

Further consider the Lagrangian $\mathcal{L}(\mathbf{x}(\boldsymbol{\Lambda}), \mathbf{p}(\boldsymbol{\Lambda}), \mathbf{M})$ evaluated at an arbitrary multiplier $\mathbf{M} = [\boldsymbol{\mu}_1^T, \boldsymbol{\mu}_2^T]^T$ and primal Lagrangian multipliers corresponding to the given $\boldsymbol{\Lambda}$. This

Lagrangian lower bounds the dual function value $g(\mathbf{M}) = \max \mathcal{L}(\mathbf{x}, \mathbf{p}, \mathbf{M})$ which allows us to write

$$g(\mathbf{M}) \geq f_0(\mathbf{x}(\boldsymbol{\Lambda})) \tag{77}$$
$$+ \boldsymbol{\mu}_1^T \left[\mathbb{E}\left(\mathbf{f}_1\left(\mathbf{h}, \mathbf{p}(\mathbf{h}, \boldsymbol{\Lambda})\right)\right) - \mathbf{x}(\boldsymbol{\Lambda})\right] + \boldsymbol{\mu}_2^T \mathbf{f}_2(\mathbf{x}(\boldsymbol{\Lambda})).$$

Subtracting (76) from (77) yields

$$(\boldsymbol{\mu}_1 - \boldsymbol{\lambda}_1)^T \left[\mathbb{E}\left(\mathbf{f}_1\left(\mathbf{h}, \mathbf{p}(\mathbf{h}, \boldsymbol{\Lambda})\right)\right) - \mathbf{x}(\boldsymbol{\Lambda})\right]$$
$$+ (\boldsymbol{\mu}_2 - \boldsymbol{\lambda}_2)^T \mathbf{f}_2(\mathbf{x}(\boldsymbol{\Lambda})) \leq g(\mathbf{M}) - g(\boldsymbol{\Lambda}). \tag{78}$$

Defining the vector $\tilde{s}(\boldsymbol{\Lambda}) = [\tilde{s}_1^T(\boldsymbol{\Lambda}), \tilde{s}_2^T(\boldsymbol{\Lambda})]^T$ with components

$$\tilde{s}_1(\boldsymbol{\Lambda}) := \mathbb{E}\left(\mathbf{f}_1\left(\mathbf{h}, \mathbf{p}(\mathbf{h}, \boldsymbol{\Lambda})\right)\right) - \mathbf{x}(\boldsymbol{\Lambda}),$$
$$\tilde{s}_2(\boldsymbol{\Lambda}) := \mathbf{f}_2(\mathbf{x}(\boldsymbol{\Lambda})), \tag{79}$$

and recalling the multiplier definitions $\mathbf{M} = [\boldsymbol{\mu}_1^T, \boldsymbol{\mu}_2^T]^T$ and $\boldsymbol{\Lambda} = [\boldsymbol{\lambda}_1^T, \boldsymbol{\lambda}_2^T]^T$ we can write

$$\tilde{s}^T(\boldsymbol{\Lambda})(\mathbf{M} - \boldsymbol{\Lambda}) \leq g(\mathbf{M}) - g(\boldsymbol{\Lambda}). \tag{80}$$

If the dual function $g(\boldsymbol{\Lambda})$ is differentiable, the expression in (80) implies that $\tilde{s}(\boldsymbol{\Lambda}) = \nabla g(\boldsymbol{\Lambda})$ is its gradient. If the dual function is nondifferentiable $\tilde{s}(\boldsymbol{\Lambda})$ defines a subgradient of the dual function. In either case $-\tilde{s}(\boldsymbol{\Lambda})$ is a descent direction of the dual function. This can be verified by substituting $\mathbf{M} = \boldsymbol{\Lambda}^*$ in (80) to conclude that for any $\boldsymbol{\Lambda} \notin \boldsymbol{\Lambda}^*$ it must be

$$\tilde{s}^T(\boldsymbol{\Lambda})(\boldsymbol{\Lambda}^* - \boldsymbol{\Lambda}) \leq g(\boldsymbol{\Lambda}^*) - g(\boldsymbol{\Lambda}) < 0. \tag{81}$$

Since the inner product of $\tilde{s}(\boldsymbol{\Lambda})$ and $(\boldsymbol{\Lambda}^* - \boldsymbol{\Lambda})$ is negative the vectors $-\tilde{s}(\boldsymbol{\Lambda})$ and $(\boldsymbol{\Lambda}^* - \boldsymbol{\Lambda})$ form an angle smaller than $\pi/2$. This can be interpreted as meaning that standing at $\boldsymbol{\Lambda}$, the vector $-\tilde{s}(\boldsymbol{\Lambda})$ points in the direction of $\boldsymbol{\Lambda}^*$.

Having a descent direction available we can introduce a time index t and a stepsize ϵ_t to define the dual subgradient descent algorithm as the one with iterates $\boldsymbol{\lambda}(t)$ obtained through recursive application of

$$\boldsymbol{\lambda}(t+1) = [\boldsymbol{\lambda}(t) - \epsilon_t \tilde{s}(\boldsymbol{\Lambda}(t))]^+. \tag{82}$$

This algorithm is known to converge to optimal dual variables if the stepsize vanishes at a nonsummable rate and to approach $\boldsymbol{\Lambda}^*$ if the stepsize is constant; see e.g., [20, Section 6].

A problem in implementing (82) is that computing the subgradient component $\tilde{s}_1(\boldsymbol{\Lambda})$ in (79) is costly. To compute $\tilde{s}_1(\boldsymbol{\Lambda})$, we need to evaluate the expectation $\mathbb{E}\left[\mathbf{f}_1\left(\mathbf{h}, \mathbf{p}(\mathbf{h}, \boldsymbol{\Lambda})\right)\right]$ where each of the resource allocations $\mathbf{p}(\mathbf{h}, \boldsymbol{\Lambda})$ follows from the solution of the optimization problem in (68). Therefore, to approximate the expectation $\mathbb{E}\left[\mathbf{f}_1\left(\mathbf{h}, \mathbf{p}(\mathbf{h}, \boldsymbol{\Lambda})\right)\right]$ we need to determine $\mathbf{p}(\mathbf{h}, \boldsymbol{\Lambda})$ for a grid of channel values, which gets impractical if $\mathbf{h}$ has large dimension. A Montecarlo approximation of $\mathbb{E}\left[\mathbf{f}_1\left(\mathbf{h}, \mathbf{p}(\mathbf{h}, \boldsymbol{\Lambda})\right)\right]$ could be computed but that is also costly. Furthermore, to compute $\mathbb{E}\left[\mathbf{f}_1\left(\mathbf{h}, \mathbf{p}(\mathbf{h}, \boldsymbol{\Lambda})\right)\right]$ we

need to know the probability distribution $m_{\mathbf{h}}(\mathbf{h})$ which needs to be estimated from channel observations. To overcome these difficulties we replace the gradient $\tilde{s}(\boldsymbol{\Lambda}(t))$ in (82) by a stochastic subgradient as we discuss in the following section.

Stochastic subgradient descent

Consider a given channel realization $\mathbf{h}$ and given multiplier $\boldsymbol{\Lambda}$ and define the vector

$$\mathbf{s}_1(\mathbf{h}, \boldsymbol{\Lambda}) := \mathbf{f}_1\left(\mathbf{h}, \mathbf{p}(\mathbf{h}, \boldsymbol{\Lambda})\right) - \mathbf{x}(\boldsymbol{\Lambda}). \tag{83}$$

This definition is made such that the expected value of $\mathbf{s}_1(\mathbf{h}, \boldsymbol{\Lambda})$ with respect to the channel distribution is the subgradient component $\tilde{s}_1(\boldsymbol{\Lambda})$ in (79). Thus, if we define the vector $\mathbf{s}(\mathbf{h}, \boldsymbol{\Lambda}) = [\mathbf{s}_1^T(\mathbf{h}, \boldsymbol{\Lambda}), \mathbf{s}_2^T(\boldsymbol{\Lambda})]^T$ with $\mathbf{s}_1(\mathbf{h}, \boldsymbol{\Lambda})$ as in (83) and $\mathbf{s}_2(\boldsymbol{\Lambda}) = \tilde{s}_2(\boldsymbol{\Lambda}) = \mathbf{f}_2(\mathbf{x}(\boldsymbol{\Lambda}))$ we have

$$\mathbb{E}\left[\mathbf{s}(\mathbf{h}, \boldsymbol{\Lambda})\right] = \tilde{s}(\boldsymbol{\Lambda}). \tag{84}$$

Formally, (84) implies that $\mathbf{s}(\mathbf{h}, \boldsymbol{\Lambda})$ is a stochastic subgradient of the dual function. Intuitively, (84) implies that $\mathbf{s}(\mathbf{h}, \boldsymbol{\Lambda})$ is an average descent direction of the dual function because its expectation is a descent direction. Thus, if we draw independent channel realizations $\mathbf{h}(t)$ and replace $\tilde{s}(\boldsymbol{\Lambda}(t))$ for $\mathbf{s}(\mathbf{h}(t), \boldsymbol{\Lambda}(t))$ in (82) we expect to observe some sort of convergence towards optimum multipliers.

The advantage of this substitution is that to compute the stochastic subgradient $\mathbf{s}(\mathbf{h}, \boldsymbol{\Lambda})$, we do not need to evaluate an expectation as in the case of the subgradient $\tilde{s}(\boldsymbol{\Lambda})$. As a consequence, using stochastic subgradients as descent directions results in an algorithm that is computationally lighter. Perhaps more important, we can operate without knowledge of the channel probability distribution if we use the current channel realization $\mathbf{h}(t)$ as our channel sample. These observations motivate the introduction of the following dual stochastic subgradient descent algorithm.

(S1) Primal iteration. Given multipliers $\boldsymbol{\Lambda}(t)$ observe current channel realization $\mathbf{h}(t)$ and determine primal Lagrangian maximizers [cf. (66) and (68)]

$$\mathbf{x}(t) := \mathbf{x}(\boldsymbol{\Lambda}(t))$$
$$= \operatorname*{argmax}_{\mathbf{x} \in \mathcal{X}} f_0(\mathbf{x}) - \boldsymbol{\lambda}_1^T(t)\mathbf{x} + \boldsymbol{\lambda}_2^T(t)\mathbf{f}_2(\mathbf{x}),$$
$$\mathbf{p}(t) := \mathbf{p}(\mathbf{h}(t), \boldsymbol{\Lambda}(t))$$
$$= \operatorname*{argmax}_{\mathbf{p}(\mathbf{h}(t)) \in \mathcal{P}(\mathbf{h}(t))} \boldsymbol{\lambda}_1^T(t)\mathbf{f}_1\left(\mathbf{h}(t), \mathbf{p}(\mathbf{h}(t))\right). \tag{85}$$

(S2) Dual stochastic subgradient. With the Lagrangian maximizers determined by (85) compute the stochastic subgradient $\mathbf{s}(\mathbf{h}(t), \boldsymbol{\Lambda}(t)) = \left[\mathbf{s}_1^T(\mathbf{h}(t), \boldsymbol{\Lambda}(t)), \mathbf{s}_2^T(\boldsymbol{\Lambda}(t))\right]^T$ of the dual function with components [cf. (83) and (79)]

$$\mathbf{s}_1(\mathbf{h}(t), \boldsymbol{\Lambda}(t)) = \mathbf{f}_1\left[\mathbf{h}(t), \mathbf{p}(\mathbf{h}(t), \boldsymbol{\Lambda}(t))\right] - \mathbf{x}(\boldsymbol{\Lambda}(t)),$$
$$\mathbf{s}_2(\boldsymbol{\Lambda}(t)) = \mathbf{f}_2(\mathbf{x}(\boldsymbol{\Lambda}(t))). \tag{86}$$

(S3) Dual iteration. With stochastic subgradients as in (86) and given step size ϵ descend in the dual domain along the direction $-\mathbf{s}(t)$ [cf. (82)]

$$\mathbf{\Lambda}(t+1) = [\mathbf{\Lambda}(t) - \epsilon_t \, \mathbf{s}\,(\mathbf{h}(t), \mathbf{\Lambda}(t))]^+$$

$$= \begin{bmatrix} \mathbf{\Lambda}_1(t) - \epsilon_t \, [\mathbf{f}_1\,[\mathbf{h}(t), \mathbf{p}\,(\mathbf{h}(t), \mathbf{\Lambda}(t))] - \mathbf{x}(\mathbf{\Lambda}(t))] \\ \mathbf{\Lambda}_2(t) - \epsilon_t \, \mathbf{f}_2(\mathbf{x}(\mathbf{\Lambda}(t))) \end{bmatrix}^+ .$$

(87)

The core of the dual stochastic subgradient descent algorithm is the dual iteration (S3). The purpose of the primal iteration (S1) is to compute the stochastic subgradients in (S2) that are needed to implement the dual descent update in (S3). We can think of the primal variables $\mathbf{x}(\mathbf{\Lambda}(t))$ and $\mathbf{p}\,(\mathbf{h}(t), \mathbf{\Lambda}(t))$ as a byproduct of the descent implementation.

Convergence properties depend on whether constant or time varying step sizes are used. If the stepsizes ϵ_t form a nonsummable but square summable series, i.e., $\sum_{t=0}^{\infty} \epsilon(t) = \infty$ and $\sum_{t=0}^{\infty} \epsilon^2(t) < \infty$, then using a simple supermartingale argument it can be shown that $\mathbf{\Lambda}(t)$ converges to $\mathbf{\Lambda}^*$ almost surely [24]. If constant stepsizes $\epsilon_t = \epsilon$ for all t are used, $\mathbf{\Lambda}(t)$ does not converge to $\mathbf{\Lambda}^*$ but it can be shown that $\mathbf{\Lambda}(t)$ visits a neighborhood of the optimal multiplier set $\mathbf{\Lambda}^*$ [19, Appendix]. Excursions away from this set are possible, but the set is visited infinitely often. The suboptimality of this set is controlled by the step size ϵ.

If $\mathbf{\Lambda}(t)$ approaches or converges to $\mathbf{\Lambda}^*$ it follows as a consequence of Theorem 4 that an optimal primal pair $(\mathbf{x}^*, \mathbf{p}^*)$ can be computed from the Lagrangian maximizers if the latter are unique. Observe that this does not require a separate computation because the Lagrangian maximizers are computed in the primal iteration (S1). One may question that at time t we do not compute the Lagrangian maximizer function $\mathbf{p}(\mathbf{\Lambda}(t))$ but just the single value $\mathbf{p}(\mathbf{h}(t), \mathbf{\Lambda}(t))$. However, $\mathbf{h}(t)$ is the channel realization at time t which means that $\mathbf{p}\,(\mathbf{h}(t), \mathbf{\Lambda}(t))$ is the value we need to compute to adapt to the current channel realization.

This permits reinterpretation of (S1)–(S3) as a policy to determine wireless systems' operating points. At time t we observe current channel realization $\mathbf{h}(t)$ and determine resource allocation $\mathbf{p}\,(\mathbf{h}(t), \mathbf{\Lambda}(t))$ which we proceed to implement in the current time slot. In this case the core of the algorithm is the primal iteration (S1) and the dual variable $\mathbf{\Lambda}(t)$ is an internal state that determines the operating point. Steps (S2) and (S3) are implemented to update this internal state so that as time progresses $\mathbf{\Lambda}(t)$ approaches $\mathbf{\Lambda}^*$ and the policy becomes optimal because it chooses the best possible resource allocation adapted to the current channel realization $\mathbf{h}(t)$.

The reinterpretation of (S1)–(S3) as a policy to determine resource allocations $\mathbf{p}(t) = \mathbf{p}\,(\mathbf{h}(t), \mathbf{\Lambda}(t))$ associated with observed channel realizations $\mathbf{h}(t)$ motivates a redefinition of the concept of solution to the wireless optimization problem in (20). In principle, solving (20) entails finding the optimal resource allocation $\mathbf{p}^*$ and the optimal ergodic average $\mathbf{x}^*$ such that the problem constraints are satisfied and $P = f_0(\mathbf{x}^*)$ [cf. (21)]. Heeding the interpretation of dual stochastic subgradient descent as a policy, we are interested in the optimality of the *sequences* of power allocations $\mathbf{p}(\mathbb{N}) := \{\mathbf{p}(t)\}_{t \in \mathbb{N}}$ and average variables $\mathbf{x}(\mathbb{N}) := \{\mathbf{x}(t)\}_{t \in \mathbb{N}}$ generated by (S1)–(S3). Further notice that since (S1)–(S3) is a stochastic algorithm the sequences $\mathbf{p}(\mathbb{N})$ and $\mathbf{x}(\mathbb{N})$ generated in a particular run are instantiations of respective random processes $\mathbf{P}(\mathbb{N})$ and $\mathbf{X}(\mathbb{N})$. We are therefore interested in the optimality of the processes $\mathbf{P}(\mathbb{N})$ and $\mathbf{X}(\mathbb{N})$.

To be more specific consider the channel stochastic process $\mathbf{H}(\mathbb{N})$ whose instances are sequences of channel realizations $\mathbf{h}(\mathbb{N}) := \{\mathbf{h}(t)\}_{t \in \mathbb{N}}$ drawn independently from the channel probability distribution $m_{\mathbf{h}}(\mathbf{h})$. Suppose we are also given a sequence of variables $\mathbf{x}(\mathbb{N}) := \{\mathbf{x}(t)\}_{t \in \mathbb{N}}$ drawn from a stochastic process $\mathbf{X}(\mathbb{N})$ and a resource allocation function $\mathbf{p}(\mathbf{h})$ that dictates allocation of resources $\mathbf{p}(\mathbf{h}(t))$ for channel realization $\mathbf{h}(t)$. Assuming that the process $\mathbf{X}(\mathbb{N})$ is ergodic, the constraint $\mathbf{x} \leq \mathbb{E}\,[\mathbf{f}_1\,(\mathbf{h}, \mathbf{p}(\mathbf{h}))]$ is equivalent to

$$\lim_{t \to \infty} \frac{1}{t} \sum_{u=1}^{t} \mathbf{x}(u) \leq \lim_{t \to \infty} \frac{1}{t} \sum_{u=1}^{t} \mathbf{f}_1\,(\mathbf{h}(u), \mathbf{p}(\mathbf{h}(u))) \quad \text{a.s.}$$

(88)

Indeed, since $\mathbf{X}(\mathbb{N})$ and $\mathbf{H}(\mathbb{N})$ are ergodic processes the limit $\lim_{t \to \infty} \frac{1}{t} \sum_{u=1}^{t} \mathbf{x}(u)$ is a constant that we could denote by $\mathbf{x}$ and the limit $\lim_{t \to \infty} \frac{1}{t} \sum_{u=1}^{t} \mathbf{f}_1\,(\mathbf{h}(u), \mathbf{p}(\mathbf{h}(u)))$ is equivalent to the expectation $\mathbb{E}\,[\mathbf{f}_1\,(\mathbf{h}, \mathbf{p}(\mathbf{h}))]$.

Writing the constraint $\mathbf{x} \leq \mathbb{E}\,[\mathbf{f}_1\,(\mathbf{h}, \mathbf{p}(\mathbf{h}))]$ in the more cumbersome form shown in (88) has the advantage that the latter can be generalized to cases in which we are given stochastic processes $\mathbf{P}(\mathbb{N})$ and $\mathbf{X}(\mathbb{N})$ in which $\mathbf{X}(\mathbb{N})$ is not necessarily ergodic and realizations $\mathbf{p}(t)$ of the process $\mathbf{P}(\mathbb{N})$ are more general than just functions of the channel state $\mathbf{h}(t)$. This concept of solution is formally defined next.

Definition 2. *Consider the channel stochastic process* $\mathbf{H}(\mathbb{N})$ *whose instances are sequences* $\mathbf{h}(\mathbb{N}) := \{\mathbf{h}(t)\}_{t \in \mathbb{N}}$ *drawn independently from the channel probability distribution* $m_{\mathbf{h}}(\mathbf{h})$. *We say the stochastic processes* $\mathbf{P}^\dagger(\mathbb{N})$ *and* $\mathbf{X}^\dagger(\mathbb{N})$ *with realizations* $\mathbf{p}^\dagger(\mathbb{N}) := \{\mathbf{p}^\dagger(t)\}_{t \in \mathbb{N}}$ *and* $\mathbf{x}^\dagger(\mathbb{N}) := \{\mathbf{x}^\dagger(t)\}_{t \in \mathbb{N}}$ *solve the optimization problem in* (20) *if*

(i) Instantaneous feasibility. *Sequence values satisfy the set constraints* $\mathbf{x}^\dagger(t) \in \mathcal{X}$, $\mathbf{p}^\dagger(t) \in \mathcal{P}(\mathbf{h}(t))$ *for all times t.*

(ii) Almost sure average feasibility. *Ergodic limits of sequences $\mathbf{x}^{\dagger}(\mathbb{N})$ and $\mathbf{p}^{\dagger}(\mathbb{N})$ are feasible with probability 1, i.e.,*

$$\lim_{t\to\infty} \frac{1}{t}\sum_{u=1}^{t}\mathbf{x}^{\dagger}(u) \leq \lim_{t\to\infty}\frac{1}{t}\sum_{u=1}^{t}\mathbf{f}_1\left(\mathbf{h}(u),\mathbf{p}^{\dagger}(u)\right) \quad \text{a.s.,}$$

(89)

$$\mathbf{f}_2\left[\lim_{t\to\infty}\frac{1}{t}\sum_{u=1}^{t}\mathbf{x}^{\dagger}(u)\right] \geq \mathbf{0} \qquad \text{a.s.}$$

(90)

(iii) Almost sure optimality. *The yield of the ergodic limit of $\mathbf{x}(\mathbb{N})$ is almost surely optimal, i.e,*

$$P = f_0\left[\lim_{t\to\infty}\frac{1}{t}\sum_{u=1}^{t}\mathbf{x}^{\dagger}(u)\right] \quad \text{a.s.}$$

(91)

If the stochastic process $\mathbf{X}^{\dagger}(\mathbb{N})$ is ergodic and the process $\mathbf{P}^{\dagger}(\mathbb{N})$ is such that realizations $\mathbf{p}^{\dagger}(t) = \mathbf{p}^{\dagger}(\mathbf{h}(t))$ are functions of current channel states, Definition 2 is equivalent to (21) with $\mathbf{x}^{*} = \lim_{t\to\infty}\frac{1}{t}\sum_{u=1}^{t}\mathbf{x}^{\dagger}(u)$ and $\mathbf{p}^{*}(\mathbf{h}(t)) = \mathbf{p}^{\dagger}(\mathbf{h}(t))$. Definition 2 is more general because it allows correlation between values of $\mathbf{X}(\mathbb{N})$ and lets $\mathbf{p}(t)$ be more complex than just a function of the current channel realization $\mathbf{h}(t)$. This added generality is needed because processes $\mathbf{P}(\mathbb{N})$ and $\mathbf{X}(\mathbb{N})$ defined as per (S1)–(S3) yield correlated processes $\mathbf{P}(\mathbb{N})$ and $\mathbf{X}(\mathbb{N})$ in which $\mathbf{p}(t)$ is a function of the current channel realization $\mathbf{h}(t)$ and the current Lagrange multiplier $\mathbf{\Lambda}(t)$. Processes $\mathbf{P}(\mathbb{N})$ and $\mathbf{X}(\mathbb{N})$ are close to optimal in the sense of Definition 2 as we describe in the following theorem.

Theorem 5 (**Ergodic stochastic optimization [19]**). *Consider the optimization problem in* (20) *as well as processes $\mathbf{P}(\mathbb{N})$ and $\mathbf{X}(\mathbb{N})$ generated by the stochastic dual descent algorithm* (S1)–(S3). *Let $\hat{S}^2 \geq \mathbb{E}\left[\|\mathbf{s}\left(\mathbf{h},\mathbf{\Lambda}\right)\|^2\right]$ be a bound on the second moment of the norm of the stochastic subgradients $\mathbf{s}\left(\mathbf{h},\mathbf{\Lambda}\right)$ and assume the same hypotheses of Theorem* 1. *Sequences $\mathbf{P}(\mathbb{N})$ and $\mathbf{X}(\mathbb{N})$ are such that:*

(i) Feasibility. *Items (i) and (ii) of Definition* 2 *hold true.*
(ii) Near optimality. *The ergodic average of $\mathbf{x}(t)$ almost surely converges to a value with optimality gap smaller than $\epsilon\hat{S}^2/2$, i.e,*

$$P - f_0\left[\lim_{t\to\infty}\frac{1}{t}\sum_{u=1}^{t}\mathbf{x}(u)\right] \leq \frac{\epsilon\hat{S}^2}{2} \quad \text{a.s.}$$

(92)

The sequences $\mathbf{p}(\mathbb{N})$ and $\mathbf{x}(\mathbb{N})$ satisfy the constraints in (89) and (90) almost surely and the objective function evaluated at the ergodic limit $\bar{\mathbf{x}} := (1/t)\sum_{u=1}^{t}\mathbf{x}(u)$ is within $\epsilon\hat{S}^2/2$ of optimal. Since $\mathcal{X}$ and $\mathcal{P}$ are compact sets it follows that the bound $\hat{S}^2$ is finite. Therefore, reducing ϵ it is possible to make $\mathbf{f}_0(\bar{\mathbf{x}})$ arbitrarily close to P and as a consequence the sequences $\mathbf{p}(\mathbb{N})$ and $\mathbf{x}(\mathbb{N})$ arbitrarily close to

optimal. It follows that the processes $\mathbf{P}(\mathbb{N})$ and $\mathbf{X}(\mathbb{N})$ generated by (S1)–(S3) are arbitrarily close to processes $\mathbf{P}^{\dagger}(\mathbb{N})$ and $\mathbf{X}^{\dagger}(\mathbb{N})$ that are optimal in the sense of Definition 2.

Variables $\mathbf{p}^*$ and $\mathbf{x}^*$ optimal in the sense of (21) are not computed by (S1)–(S3). Rather, (89) implies that, asymptotically, (S1)–(S3) is drawing resource allocation realizations $\mathbf{p}(t) = \mathbf{p}(\mathbf{h}(t),\mathbf{\Lambda}(t))$ and variables $\mathbf{x}(t) := \mathbf{x}(\mathbf{\Lambda}(t))$ that are close to optimal as per Definition 2. The important point here is that having a procedure to generate stochastic processes close to optimal in the sense of Definition 2 is sufficient for practical implementation.

An example application of the dual stochastic subgradient descent algorithm (S1)–(S3) is discussed in the next section.

Frequency division broadcast channel
To implement dual stochastic descent for the frequency division broadcast channel we need to specify the primal iteration (S1) and the dual iteration (S2). To specify the primal iteration (S1) we need to compute Lagrangian maximizers for which it suffices to recall the expressions in Section "Frequency division broadcast channel" of "Recovery of optimal primal variables". For the ergodic rate r_i we make $\lambda_i = \lambda_i(t)$ in (70) to conclude that the primal iterate $r_i(t) = r_i(\lambda_i(t))$ is

$$r_i(t) = \underset{r_i\in[r_{\min},r_{\max}]}{\arg\max} U(r_i) - \lambda_i(t)r_i.$$

(93)

For the power allocations $p_i(t) = p_i(\mathbf{h}(t),\lambda(t),\mu(t))$ and the frequency assignments $\boldsymbol{\alpha}(t) = \boldsymbol{\alpha}(\mathbf{h}(t),\lambda(t),\mu(t))$ we need to set the multipliers to $\lambda = \lambda(t)$ and $\mu = \mu(t)$ and also set the value of the channel to its current state $\mathbf{h} = \mathbf{h}(t)$. This substitution in (73) yields the power allocation

$$p_i(t) = \underset{p_i(\mathbf{h}(t))}{\arg\max}\left[\lambda_i(t)C\left(\frac{h_i(t)p_i(\mathbf{h}(t))}{N_0}\right) - \mu p_i(\mathbf{h}(t))\right].$$

(94)

To determine the frequency assignments $\boldsymbol{\alpha}(t)$ we first substitute $\lambda = \lambda(t)$, $\mu = \mu(t)$, and $\mathbf{h} = \mathbf{h}(t)$ in (74) to compute the discriminants $d_i(t) = d_i(\mathbf{h}(t),\lambda(t),\mu(t))$

$$d_i(t) := \underset{p_i(\mathbf{h}(t))}{\max}\left[\lambda_i(t)C\left(\frac{h_i(t)p_i(\mathbf{h}(t))}{N_0}\right) - \mu p_i(\mathbf{h}(t))\right],$$

(95)

from where we conclude that the frequency assignment $\boldsymbol{\alpha}(t) = \boldsymbol{\alpha}(\mathbf{h}(t),\lambda(t),\mu(t))$ is given by the solution of [cf. (75)]

$$\boldsymbol{\alpha}(t) = \arg\max \sum_{i=1}^{J}\alpha_i(\mathbf{h}(t))d_i(t).$$

$$s.t. \qquad \alpha_i(\mathbf{h}(t)) \in \{0,1\}, \quad \mathbf{1}^{T}\boldsymbol{\alpha}(\mathbf{h}(t)) \leq 1.$$

(96)

Recall that since at most one $\alpha_i(\mathbf{h}) = 1$ in (96), the optimal frequency allocation is to make $\alpha_i(\mathbf{h}) = 1$ for the terminal with the largest discriminant when that discriminant is positive. If all discriminants are negative we make $\alpha_i(\mathbf{h}) = 0$ for all i.

The ESO algorithm for optimal resource allocation in broadcast channels is completed with an iteration in the dual domain [cf. (83) and (87)]

$$\lambda_i(t+1) = \left[\lambda_i(t) - \epsilon\left[\alpha_i(t)C\left(\frac{h_i p_i(t)}{N_0}\right) - r_i(t)\right]\right]^+,$$

$$\mu(t+1) = \left[\mu(t) - \epsilon\left[q_0 - \sum_{i=1}^{J}\alpha_i(t)p_i(t)\right]\right]^+. \quad (97)$$

As per Theorem 5 iterative application of (93)–(97) yields sequences $r_i(t)$, $\alpha_i(t)$ and $p_i(t)$ such that: (i) The sum utility for the ergodic limits of $r_i(t)$ is almost surely within a small constant of optimal; (ii) The power constraint in (27) and the rate constraints in (26) are almost surely satisfied in an ergodic sense. This result is true despite the presence of the non-convex integer constraint $\boldsymbol{\alpha} \in \mathcal{A}$, the non-concave function $C(h_i p_i(t)/N_0)$, the lack of access to the channel's probability distribution, and the infinite dimensionality of the optimization problem.

Numerical results

The dual stochastic subgradient descent algorithm for optimal resource allocation in frequency division broadcast channels defined by (93)–(97) is simulated for a system with $J = 16$ nodes. Three AMC modes corresponding to capacities 1, 2 and 3 bits/s/Hz are used with transitions at SINR 1, 3 and 7. Fading channels are generated as i.i.d. Rayleigh with average powers 1 for the first four nodes, i.e., $j = 1, \ldots, 4$, and 2, 3 and 4 for subsequent groups of 4 nodes. Noise power is $N_0 = 1$ and average power available is $q_0 = 3$. Rate of packet acceptance is constrained to be $0 \leq r_i(t) \leq 2$ bits/s/Hz. The optimality criteria is proportional fair scheduling, i.e., $U_i(r_i) = \log(r_i)$ for all i. Steps size is $\epsilon = 0.1$.

Figure 3 shows evolution of dual variables $\lambda_i(t)$ and corresponding rates $r_i(t)$ for representative nodes $i = 1$ with average channels $\mathbb{E}h_1(t) = 1$ and $i = 9$ with $\mathbb{E}h_9(t) = 3$. The time average rate $\bar{r}_i(t) := (1/t)\sum_{u=1}^{t} r_i(u)$ is also shown. Neither multipliers $\lambda_i(t)$ nor rates $r_i(t)$ converge, but ergodic rates $\bar{r}_i(t)$ do converge. Multiplier $\lambda_1(t)$ associated with node 1 is larger than multiplier $\lambda_9(t)$ of node 9. This improves fairness of resource allocation by increasing the chances of allocating user 1 even when the channel $h_1(t)$ is smaller than $h_9(t)$—recall that channel $h_9(t)$ is stronger on average. Convergence of the algorithm is ratified by Figures 4 and 5. Figure 4 shows evolution of the objective $\sum_{i=1}^{J} U_i(\bar{r}_i(t))$ and the dual function value $g(t) := g(\boldsymbol{\lambda}(t), \mu(t))$. Notice that the objective value is decreasing towards the maximum objective. This is not
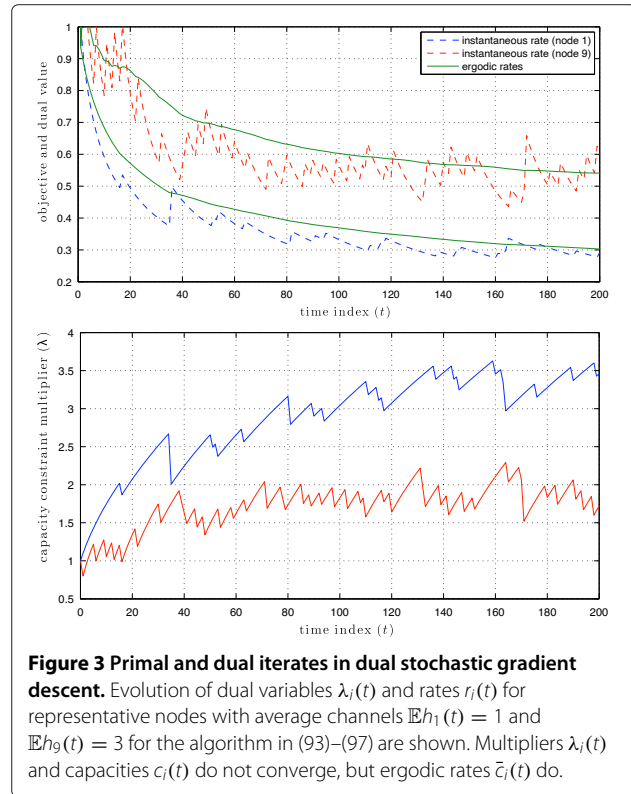

Figure 3 Primal and dual iterates in dual stochastic gradient descent. Evolution of dual variables $\lambda_i(t)$ and rates $r_i(t)$ for representative nodes with average channels $\mathbb{E}h_1(t) = 1$ and $\mathbb{E}h_9(t) = 3$ for the algorithm in (93)–(97) are shown. Multipliers $\lambda_i(t)$ and capacities $c_i(t)$ do not converge, but ergodic rates $\bar{c}_i(t)$ do.

a contradiction, because variables $\bar{r}_i(t)$ are infeasible but approach feasibility as t grows. The dual function's value is an upper bound on the maximum utility and it can be observed to approach the objective as t grows. Eventually, the objective value becomes smaller than the dual value as expected. Figure 5 corroborates satisfaction of the power constraint in (27) and the rate constraints in (26). The amount by which the power constraint (27) is violated is shown in the top. In the bottom we show the corresponding figure for the rate constraint in (26). Since there are
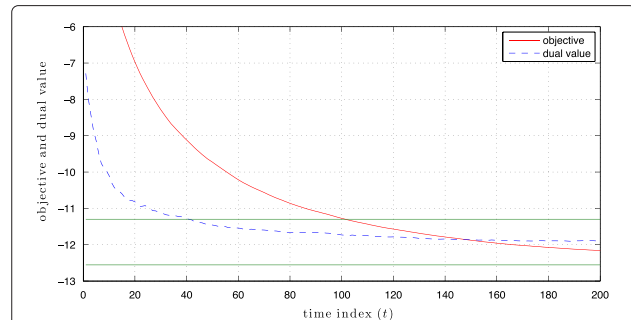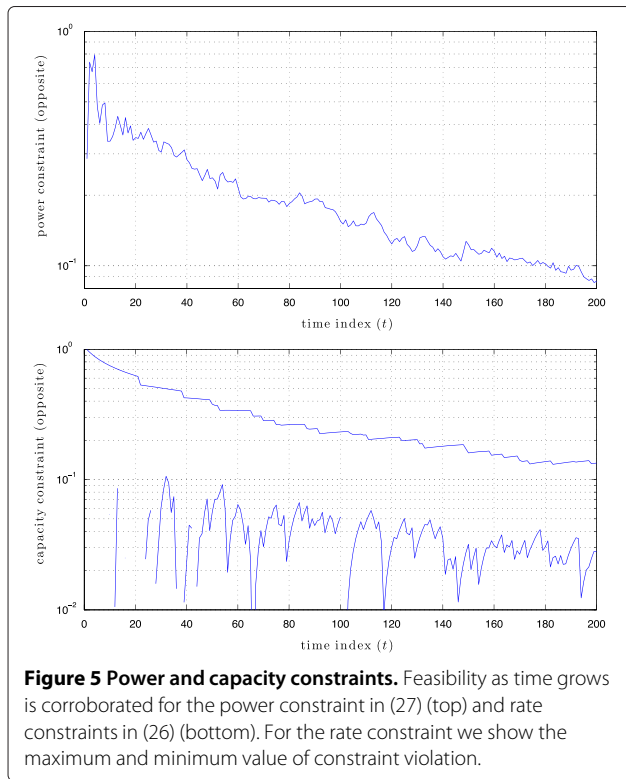

Figure 4 Optimal frequency division broadcast channel. Objective value $\sum_{i=1}^{J} U_i(\bar{c}_i(t))$ and dual function's value $g(t) := g(\boldsymbol{\lambda}(t), \mu(t))$ for the algorithm in (93)–(97) are shown along with lines marking optimal utility and 90% of optimal yield. Utility yield becomes optimal as time grows.

Figure 5 Power and capacity constraints. Feasibility as time grows is corroborated for the power constraint in (27) (top) and rate constraints in (26) (bottom). For the rate constraint we show the maximum and minimum value of constraint violation.

J of these constraints we show the minimum and maximum violation. All constraints are satisfied as t grows. Resulting power allocations appear in Figure 6 for a channel with average power $\mathbb{E}h_{if}(t) = 1$ and for a channel with $\mathbb{E}h_{if}(t) = 2$. Power allocation is opportunistic in that power is allocated only when channel realizations are above average.

Conclusions

This article reviews recent results which state that problems of the form in (20) in which nonconcave functions appear inside expectations have null duality gap as long as the probability distribution of the fading coefficient $\mathbf{h}$ contains no points of strictly positive probability. Lack of duality gap permits solution in the dual domain leading to a substantial reduction in the computational cost of determining optimal operating points of the wireless system. Working in the dual domain leads to a solution methodology that can be interpreted as a generalization of the derivation of the waterfilling power allocation in point to point channels reviewed in Section "Power allocation in a point-to-point channel".

Specifically, the problem of determining the optimal resource allocation function $\mathbf{p}^*$ in (20) is challenging due to its infinite dimensionality and lack of convexity. However, in the dual domain we need to determine the optimal multiplier $\mathbf{\Lambda}^*$ that minimizes the dual function in (23).

This is simpler because the dual function is convex and finite dimensional. Once we have found an optimal dual variable we can determine optimal operating points as Lagrangian maximizers. In doing so we can exploit the separable structure of the Lagrangian to decompose the optimization problem into the per fading state subproblems in (68). We emphasize that solving the optimization programs in (68) is not necessarily easy if the dimensionality of $\mathbf{h}$ is large. Nevertheless solving (68) is always simpler than solving (20) and in some cases plain simple. Lack of duality gap and Lagrangian separability are further exploited to propose the dual stochastic subgradient descent algorithm (S1)–(S3) which converges to an optimal operating point with probability 1 in an ergodic sense.

There are three key points that permit the development of the solution methodology outlined in the previous paragraph:

Nonatomic fading distribution. A nonatomic fading distribution leads to the lack of duality gap. The fact that $P = D$, i.e., that primal and dual optimal values are the same, is what allows us to work in the dual domain without loss of optimality. In formal terms, lack of duality gap is the tool that we used to recover the optimal primal variables $(\mathbf{x}^*, \mathbf{p}^*)$ from the optimal dual variable $\mathbf{\Lambda}^*$ by determining
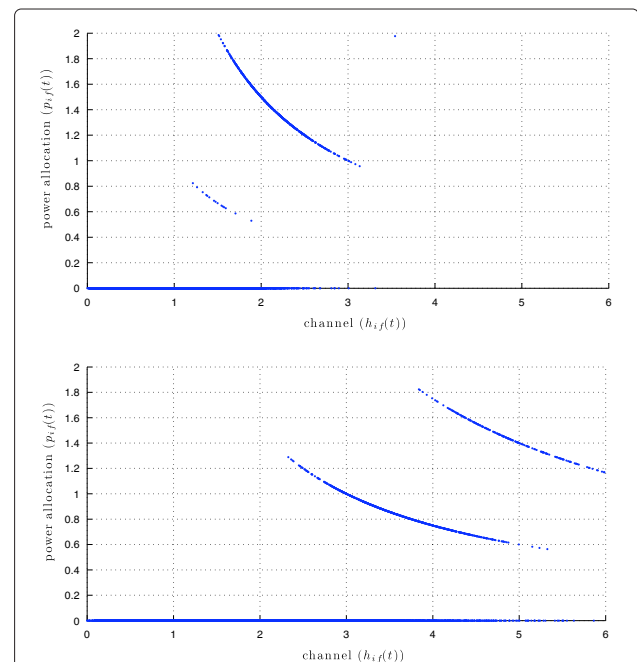

Figure 6 Power allocations. Power allocated as a function of channel realization is shown for channels with average power $\mathbb{E}h_{if}(t) = 1$ (top) and $\mathbb{E}h_{if}(t) = 2$ (bottom). The resulting power allocation is opportunistic in that power is allocated only when channel realizations are above average.

the primal Lagrangian maximizers for $\mathcal{L}(\mathbf{x}, \mathbf{p}, \mathbf{\Lambda}^*)$ [cf. Theorem 4]. It is important to distinguish between convexity of the optimization problem and lack of duality gap. Null duality gap may follow from convexity, but convexity is rare in wireless communications systems. Lack of duality gap can also follow from a nonatomic fading distribution, which is a common occurrence in wireless systems.

Lagrangian Separability. According to Theorem 4 null duality gap permits computation of the optimal pair $(\mathbf{x}^*, \mathbf{p}^*)$ as the Lagrangian maximizers $(\mathbf{x}(\mathbf{\Lambda}^*), \mathbf{p}(\mathbf{\Lambda}^*))$. This is not a simplification per se but leads to a simplification because the computation of the Lagrangian maximizer function $\mathbf{p}(\mathbf{\Lambda}^*)$ can be separated into per fading state problems whose solution determines values $\mathbf{p}(\mathbf{h}, \mathbf{\Lambda}^*)$ of this function [cf. (66)–(68)]. The Lagrangian is separable in this sense because neither the constraints nor the objective function involve a nonlinear function coupling the selection of values $\mathbf{p}(\mathbf{h}_1)$ and $\mathbf{p}(\mathbf{h}_2)$ for different channel realizations $\mathbf{h}_1 \neq \mathbf{h}_2$. Whenever $\mathbf{p}(\mathbf{h}_1)$ and $\mathbf{p}(\mathbf{h}_2)$ appear as part of the same constraint they appear as different terms of an expectation operation. This absence of coupling is what permits exchanging the order of maximization and expectation in going from (67) to (68).

Finite number of constraints. Working in the dual domain is simpler than working in the primal domain because the dual function is finite dimensional whereas the primal problem is infinite dimensional. We have a finite dimensional dual function as long as the original optimization problem has a finite number of constraints.

Nonatomic fading distributions, Lagrangian separability, and having a finite number of constraints are properties that appear in many, indeed most, problems in optimal design of wireless systems. In such cases the methodology described in this article can be applied to their solution.

Further reading

The use of dual problems as a shortcut to solve optimization problems in communications has a rich history [25-28]; see also [29] for a comprehensive treatment. Lack of duality gap in non-convex optimization problems has also been observed in the context of asymmetric digital subscriber lines [30,31]. In network optimization lack of duality gap leads to the optimality of layered architectures which renders the complexity of wireless networking essentially identical to the complexity of physical layer optimization [32-35]. For the use of techniques discussed here in the solution of specific problems we refer the reader to [36-42]. For further details on dual stochastic sub gradient descent, the literature on convergence of subgradient descent algorithms [43-45], and stochastic subgradient descent [46-49] is of interest.

References
1. X Wang, GB Giannakis, Resource allocation for wireless multiuser OFDM networks, IEEE Trans. Inf. Theory. **57**(7), 4359–4372 (2011)
2. V Ntranos, N Sidiropoulos, L Tassiulas, On multicast beamforming for minimum outage, IEEE Trans. Wirel. Commun. **8**(6), 3172–3181 (2009)
3. ND Sidiropoulos, TN Davidson, ZQ Luo, Transmit beamforming for physical-layer multicasting, IEEE Trans. Signal Process. **54**(6), 2239–2251 (2006)
4. JA Bazerque, GB Giannakis, Distributed scheduling and resource allocation for cognitive OFDMA radios, Mobile Nets. Apps. **13**(5), 452–462 (2008)
5. Z Quan, S Cui, AH Sayed, Optimal linear cooperation for spectrum sensing in cognitive radio networks, IEEE J. Sel. Topics Signal Process. **2**, 28–40 (2008)
6. Y Hu, A Ribeiro, Optimal wireless networks based on local channel state information, IEEE Trans. Signal Process. **60**(9), 4913–4929 (September 2012)
7. Y Hu, A Ribeiro, Adaptive distributed algorithms for optimal random access channels, IEEE Trans. Wirel. Commun. **10**(8), 2703–2715 (2011)
8. Y Hu, A Ribeiro, in *Proc. Int. Conf. Acoustics Speech Signal Process*, vol. 1, A Optimal wireless multiuser channels with imperfect channel state information. (Kyoto Japan, 2012), pp. 1–4
9. Y Hu, A Ribeiro, in *Global Telecommun. Conf.*, vol. 1, Optimal transmission over a fading channel with imperfect channel state information. (Houston, TX, 2011), pp. 1–5
10. L Chen, SH Low, M Chiang, JC Doyle, in *Proc. IEEE INFOCOM,* vol. 1, Cross-layer congestion control, routing and scheduling design in ad hoc wireless networks. (Barcelona, Spain, 23–29 April 2005), pp. 1–13
11. M Chiang, SH Low, RA Calderbank, JC Doyle, Layering as optimization decomposition, Proc IEEE. **95**, 255–312 (2007)
12. A Eryilmaz, R Srikant, Joint congestion control, routing, and MAC for stability and fairness in wireless networks, IEEE J. Sel. Areas Commun. **24**(8), 1514–1524 (2006)
13. L Georgiadis, MJ Neely, Resource allocation and cross-layer control in wireless networks, Found Trends Netw. **1**, 1–144 (2006)
14. JW Lee, RR Mazumdar, NB Shroff, Opportunistic power scheduling for dynamic multi-server wireless systems, IEEE Trans. Wirel. Commun. **5**(6), 1506–1515 (2006)
15. X Lin, NB Shroff, R Srikant, A tutorial on cross-layer optimization in wireless networks, IEEE J. Sel. Areas Commun. **24**(8), 1452–1463 (2006)
16. MJ Neely, E Modiano, CE Rohrs, Dynamic power allocation and routing for time-varying wireless networks, IEEE J. Sel. Areas Commun. **23**, 89–103 (2005)
17. X Wang, K Kar, Cross-layer rate optimization for proportional fairness in multihop wireless networks with random access, IEEE J. Sel. Areas Commun. **24**(8), 1548–1559 (2006)
18. Y Yi, S Shakkottai, Hop-by-hop congestion control over a wireless multi-hop network, IEEE/ACM Trans. Netw. **15**(133–144), 1548–1559 (2007)
19. A Ribeiro, Ergodic stochastic optimization algorithms for wireless communication and networking, IEEE Trans. Signal Process. **58**(12), 6369–6386 (2010)

20. A Ribeiro, G Giannakis, Separation principles in wireless networking, IEEE Trans. Inf. Theory. **56**(9), 4488–4505 (2010)
21. S Boyd, L Vandenberghe, *Convex Optimization*. (Cambridge University Press, Cambridge, 2004)
22. AA Lyapunov, Complètement, Sur les, Fonctions-vecteur, URSS, Additives, Bull. Acad. Sci. Sèr. Math. **4**, 465–478 (1940)
23. RT Rockafellar, *Convex Analysis*. (Princeton University Press, Princeton, NJ, 1970)
24. NZ Shor, *Minimization Methods for Non-Differentiable Functions*. (Springer, Berlin, 1985)
25. FP Kelly, A Maulloo, D Tan, Rate control for communication networks: shadow prices, proportional fairness and stability, J. Oper. Res. Soc. **49**(3), 237–252 (1998)
26. SH Low, DE Lapsley, Optimization flow control, I: basic algorithm and convergence, IEEE/ACM Trans. Netw. **7**(6), 861–874 (1998)
27. SH Low, A duality model of TCP and queue management algorithms, IEEE/ACM Trans. Netw. **11**(4), 525–536 (2003)
28. SH Low, F Paganini, JC Doyle, Internet congestion control, IEEE Control Syst. Mag. **22**, 28–43 (2002)
29. R Srikant, *The Mathematics of Internet Congestion Control*, 1st edn. (Birkhauser, 2004)
30. ZQ Luo, S Zhang, Dynamic spectrum management: complexity and duality, IEEE J. Sel. Topics Signal Process. **1**(2), 57–73 (2008)
31. W Yu, R Lui, Dual methods for nonconvex spectrum optimization of multicarrier systems, IEEE Trans. Commun. **54**(7), 1310–1322 (2006)
32. RA Berry, EM Yeh, Cross-layer wireless resource allocation, IEEE Signal Process. Mag. **21**(5), 59–68 (2004)
33. MJ Neely, Energy optimal control for time-varying wireless networks, IEEE Trans. Inf. Theory. **52**(7), 2915–2934 (2006)
34. MJ Neely, E Modiano, CP Li, in *Proc. IEEE INFOCOM,* vol 3. Fairness and optimal stochastic control for heterogeneous networks. (Miami, FL 13–17, March 2005), pp. 1723–1734
35. A Ribeiro, G Giannakis, in *Proc. Conf. on Info. Sciences and Systems,* vol. 1. Layer separability of wireless networks. (Princeton Univ. Princeton, NJ, 2008), pp. 821–826
36. N Gatsis, A Ribeiro, G Giannakis, A class of convergent algorithms for resource allocation in wireless fading networks, IEEE Trans. Wirel. Commun. **9**(5), 1808–1823 (2010)
37. N Gatsis, A Ribeiro, G Giannakis, in *Proc. Int. Conf. Acoustics Speech Signal Process,* vol. 1. Cross-layer optimization of wireless fading ad-hoc networks. (Taipei, Taiwan, 2009), pp. 2353–2356
38. Y Hu, A Ribeiro, in *Proc. Int. Conf. Acoustics Speech Signal Process,* vol. 1. Optimal wireless networks based on local channel state information. (Prague Czech Republic, 2011), pp. 3124–3127
39. Y Hu, A Ribeiro, in *Proc. Allerton Conf. on Commun. Control Computing,* vol. 1. Adaptive distributed algorithms for optimal random access channels. (Monticello, 2010), pp. 1474–1481
40. A Ribeiro, G Giannakis, in *Proc. Int. Conf. Acoustics Speech Signal Process,* vol. 1. Optimal FDMA over wireless fading mobile ad-hoc networks. (Las Vegas, NV, 2008), pp. 2765–2768
41. A Ribeiro, T Luo, N Sidiropoulos, G Giannakis, in *Proc. IEEE Int. Conf. on Computer Commun,* vol. 1. Modelling and optimization of stochastic routing for wireless multihop networks. (Anchorage, AK, 2007), pp. 1748–1756
42. A Ribeiro, N Sidiropoulos, G Giannakis, Optimal distributed stochastic routing algorithms for wireless multihop networks, IEEE Trans. Wirel. Commun. **7**(11), 4261–4272 (2008)
43. A Juditsky, G Lan, A Nemirovski, A Shapiro, Stochastic approximation approach to stochastic programming, SIAM J. Optim. **19**(4), 1574–1609 (2009)
44. T Larsson, M Patriksson, A Str omberg, Ergodic primal convergence in dual subgradient schemes for convex programming, Math. Progr. **86**(2), 283–312 (1999)
45. A Nedic, A Ozdaglar, Approximate primal solutions and rate analysis for dual subgradient methods, SIAM J. Optim. **19**(4), 1757–1780 (2009)
46. BT Polyak, Newstochasticapproximationtypeprocedures Autom, Remote Control. **51**, 937–946 (1990)
47. BT Polyak, AB Juditsky, Acceleration of stochastic approximation by averaging, SIAM J. Control Optim. **30**(4), 838–855 (1992)
48. A Ribeiro, in *Proc. IEEE Workshop on Signal Process. Advances in Wireless Commun* vol. 1. Stochastic learning algorithms for optimal design of wireless fading networks. (Marakech, Morocco, 2010), pp. 1–5
49. A Ribeiro, in *Proc. Int. Conf. Acoustics Speech Signal Process,* vol. 1, Ergodic stochastic optimization algorithms for wireless communication and networking. (Dallas, TX, 2010), pp. 3326–3329