Cross Layer Designs in WLAN Systems
N. Zorba, C. Skianis, and C. Verikoukis, Eds.
Leicester, UK: Troubador Publishing
ISBN: 9781848762275

Optimal Cross-Layer Design of Wireless Fading Multi-Hop Networks

Antonio G. Marques[(1)], Nikolaos Gatsis[(2)], and Georgios B. Giannakis[(2)]

[(1)] *Rey Juan Carlos University, Madrid, Spain*
[(2)] *University of Minnesota, Minneapolis, USA*

1 Introduction

The last decade has brought a rapid growth in demand for fast and error-resilient telecommunication services. In accordance with this growth, broadband wireless networks have become an integral part of the global communication infrastructure. Provisioning of quality-of-service (QoS) in broadband wireless networks requires coping with the challenges brought by the wireless interface, and allocating resources available at different layers among nodes, links, and end-to-end connections. Nonlinear optimization tools have been successfully adopted to analyze and design algorithms that fulfill such requirements; see e.g., [12, 18, 35, 53] and references therein. Optimal network designs are obtained by formulating a constrained optimization problem involving variables from different layers, and by exploiting information about the wireless channel. Solving such optimization problems dictates how resources are allocated across different layers, while network control protocols follow from the algorithms used for this solution. One of the most challenging issues to cope with in designing optimal cross-layer resource allocation schemes for wireless networks is the presence of fading. Fading renders wireless channels random, degrades the communication performance, and leads to location-dependent and time-varying link capacities. As a result, cross-layer schemes are required to account for the fading nature of the channel, and implement mechanisms to deal with it. Such schemes should be able to effectively exploit the diversity provided by the channel, and adapt the resource allocation to the channel state information (CSI) available.

Capitalizing on optimization theory and stochastic approximation tools, this chapter deals with channel-adaptive algorithms that allocate resources at transport, network, link, and physical layers. These algorithms emerge from the solution of constrained optimization problems that take into account the QoS, the interaction among layers, and the CSI available.

The model describing the multi-hop wireless network as well as its most relevant operating conditions are as follows. Nodes receive packets from the application layer intended for different destinations. Flow control and routing decisions respond to packet arrivals, and the long-term average end-to-end rates entail different utility levels. At the link layer, two different models are considered, whereby nodes access either orthogonally or non-orthogonally a set of parallel flat fading channels. Orthogonal here means that if a terminal is transmitting, no other link interfering with this transmission can be active;

see e.g., [18, Example 2.3], [54]. Constraints on the links that can be simultaneously activated are typically called *interference constraints*, and a resulting feasible set of links is called *schedule* or link activation set. In the non-orthogonal case, all link transmissions are allowed to use all channels, and interfering transmissions are treated as noise. The link rates are then functions of the signal-to-interference-plus-noise-ratio (SINR). At the physical layer, nodes can adapt their instantaneous power and rate loadings per fading realization, while also optimizing their average power consumption. The main difficulty to solve the resulting channel-adaptive optimization is due to the link layer. In the orthogonal case, link scheduling per fading realization may be a complex task, while in the non-orthogonal case the SINR dependence couples the power allocation decisions, and the problem is in general non-convex, and thus challenging.

There is a large body of works treating network optimization and control. The ones focusing on wireless multi-hop networks with the aforementioned media access types are briefly outlined next.

General cross-layer optimization problems are formulated in [10,11,13,17,21,30,31,33, 34,47,49,51–53,55,63,64,66]. Most rely on a dual approach to solve the problem, except for [11], [53, Sec. 3.4] which use a primal-dual method; [63,64] are based on scaled gradient projection; [55] utilizes a cross decomposition approach. Various layered architectures and network control algorithms result from the approach followed in each work. With regards to the physical and link layers, the aforementioned works can be classified as follows:

- An abstract (convex) link layer rate region is used in [33].
- Link activation sets are considered in [10,34,53], whereby links have fixed capacities.
- Link capacities as functions of resource allocation quantities local to each link are used in [31, 66]. This model emerges typically when there are enough orthogonal signaling dimensions allocated a priori.
- Interference constraints are introduced in [51], while adopting link capacity as a function of the power allocation over that link.
- High-SINR or related approximations of the link capacities as functions of the SINR are adopted in [11, 64].
- Low-SINR approximations for the link capacities are adopted in [13, 47].
- A staircase or a step function of the SINR and interference constraints are considered in [21, 55]. Moreover, half-duplex constraints are considered in [63]. The link capacity is a function of the SINR, because interference is still present under such constraints.
- Capacities are kept as generic functions of the SINR or the power allocations at all links in [30, 49, 52].
- The information-theoretic $\log(1 + \text{SINR})$ model is used in [17, 64].

Suboptimal low-complexity approaches where the link-layer rate region is substituted by an achievable inner bound have also been pursued [5, 9, 14, 32, 67, 68]; such approaches are also termed "layered." The premise for this substitution is that methods applicable to wireline networks can then be used for routing and congestion control. In general, it is important to stress that the vast majority of the aforementioned prior art assumes that wireless channels are deterministic. Exceptions include [10, 30, 34], which also deal with random channels taking values from a finite set; [31] where the random channels are modeled as stationary and ergodic processes; and [17,49,51,52] which consider continuous fading.

The works mentioned in the previous paragraph rely on nonlinear optimization tools

to solve the resource allocation task. Differently, [15, 16, 19, 41–43, 65] develop resource allocation schemes by using a Lyapunov stability approach. Those are built upon dynamic backpressure policies, first introduced in the seminal work of [58]. Specifically, the routing and scheduling components here are based on differential backlogs, capturing the differences between queue lengths of neighboring nodes in the network. An important feature in these works is the introduction of virtual queues in order to ensure that constraints on long-term average quantities (e.g., power) are satisfied [41]. Moreover, congestion control is added based on utility maximization, following a dual [15, 19, 42], or, a primal-dual approach [16]. The overall framework is also referred to as *stochastic network optimization*; see [18] for a tutorial treatment. Recent extensions explicitly deal with wireless links that are not reliable [44, 50]. A related approach to backpressure policies also uses queue lengths as a basis for generic network utility maximization problems [57].

Accounting for wireless fading effects represents the main difference of this chapter's themes from most of the state-of-the-art works in the literature. In the cross-layer optimization of orthogonal-access networks, instantaneous constraints involving link capacities are considered, whereby an instant corresponds to a fading realization. Such constraints differ from those found elsewhere in the cross-layer optimization literature. In the non-orthogonal case, continuous fading induces a favorable hidden convexity structure [52], which in turn can be used for efficient algorithmic solutions to the cross-layer optimization problem.

The remainder of the chapter is structured as follows. Starting with the orthogonal formulation, the model of the multi-hop network and the operation of the network, link, and physical layers are described in Section 2. The cross-layer design problem is formulated in Section 3, having as optimization variables the average end-to-end rates, instantaneous network layer flows, link schedules, average power consumption, and instantaneous power allocations across tones. Higher average end-to-end rates and lower average powers are promoted by considering rate-utility and power-cost functions whose inputs are variables averaged over all possible states of the fading channel. Section 4 shows how the optimal cross-layer resource allocation can be expressed as a function of the *instantaneous* CSI and the optimal Lagrange multipliers associated with the optimization problem. An offline scheme based on smooth subgradients is developed in Section 5.1 in order to obtain optimal Lagrange multipliers, which are subsequently used in an online fashion for network control in Section 5.2. Using stochastic approximation tools [24], online schemes are proposed in Section 6.1 to estimate the multipliers. Convergence and optimality of the stochastic schemes is characterized. By establishing relationships between the Lagrange multipliers and the queue lengths in the network, as in e.g., [18, Sec. 4.10], [38], [34], queue stability and average queue delay of the developed schemes are discussed in Section 6.2. Next, the focus is placed on networks with non-orthogonal access, where the link capacity becomes a function of the SINR, giving rise to *non-convexity* in general. The cross-layer optimization problem and its duality properties are the subjects of Section 7.1. A subgradient descent algorithm along with weighted running averages of the primal iterates are then developed in Section 7.2. The overall scheme yields a near-optimal solution to the cross-layer resource allocation problem, which is subsequently employed for network control. Finally, Section 8 concludes this chapter.

2 System Description

In this section, the architecture and operation of the wireless multi-hop network model is described. Consider a multi-hop wireless network with I nodes. Two nodes i, j in $\{1,\ldots,I\}$ are physically linked if they can communicate with each other. The set of nodes that a node i can communicate with constitutes the neighborhood of i, and is denoted by $\mathcal{N}(i)$. Node connectivity is captured by a *directed* graph $\mathcal{G}$ where vertices correspond to wireless nodes and an edge connecting two vertices is present only if the nodes represented by the vertices are close enough to be physically linked. Hence, nodes i and j are connected through two directed edges: the link (i,j) from i to j; and the link (j,i) from j to i.

Nodes can transmit over a set of K flat fading parallel channels, indexed by $k \in \{1,\ldots,K\}$. The terms channels, tones, and bands are used interchangeably throughout this chapter. Zero-mean additive white Gaussian noise (AWGN) with variance σ_j^k is assumed added at the receiver j over channel k. The kth channel's instantaneous power gain from node i to node j is denoted by $h_{i,j}^k$. Specifically, $h_{i,j}^k$ is the squared magnitude of the fading coefficient. The overall channel is described by vector $\mathbf{h}$, which collects all $h_{i,j}^k$ gains. Channels are assumed stationary and ergodic, and are allowed to be correlated across links, tones, and time.

When both receivers and transmitters have access to an accurate estimate of $\mathbf{h}$, the system can be designed using perfect CSI (P-CSI). Since P-CSI may not be realistic in some practical scenarios [25], one is also interested in designing the system based of Q-CSI. For such a case, the range of values each $h_{i,j}^k$ takes is divided into non-overlapping regions; and instead of the analog-amplitude $h_{i,j}^k$, receivers and transmitters have available only its quantized version $h_{i,j}^{kQ}$ (equivalently, only a binary codeword is available indexing the region $h_{i,j}^k$ falls into). Since $h_{i,j}^k$ is random, $h_{i,j}^{kQ}$ is also a discrete random variable; and likewise $\mathbf{h}^Q$ is random, taking vector values from a set with finite cardinality. Although $\mathbf{h}$ will be used as the default notation for the CSI, the problem formulation and derivations up to Section 6 are valid both for the P-CSI and the Q-CSI cases. P-CSI is assumed throughout Section 7. Differences will be stressed wherever needed.

The goal is to develop adaptive algorithms that use the instantaneous CSI to allocate resources at the network, link, and physical layers, so that pre-specified QoS metrics are optimized. The QoS metrics considered are introduced in Section 3. Each layer's operation is described next.

2.1 Network Layer Operation

Packets generated exogenously at each node correspond to possibly different applications (such as HTTP or file transfer), and are destined for different nodes. Packet streams are referred to as flows, and are indexed by f. Each node serves flows that have other nodes as destination. The destination node associated with each flow f is denoted by $d(f)$, and the average arrival rate of exogenous packets of flow f to node i is denoted by $\bar{a}_i^f$. Correspondingly, the instantaneous arrival rate of exogenous packets of flow f to node i is denoted by a_i^f. The instantaneous rate of flow f that during the channel realization

$\mathbf{h}$ is sent from node i to node j is denoted by $r_{i,j}^f(\mathbf{h})$.[1] Packets of flow f arrive to node i from two different sources: (a) packets arriving from the neighbors of i (*endogenous* network traffic); and (b) packets coming from the transport layer of node i (*exogenous* network traffic). Note that rates $r_{i,j}^f(\mathbf{h})$ are in fact *routing variables*, because they dictate how packets of various flows are forwarded to the outgoing links of node i. For brevity, the set of flows that can be generated exogenously at node i, i.e., flows not having i as destination, is denoted by $\mathcal{F}(i) := \{f | i \neq d(f)\}$. Variables a_i^f and $r_{i,j}^f$ are defined for $i = 1, ..., I$, $j \in \mathcal{N}(i)$, and $f \in \mathcal{F}(i)$.

The nodes are equipped with queues (buffers) that can store the incoming packets. Packets in the queue will be transmitted as soon as the conditions of the physical layer allow. In this work, we will consider that queues are stable if the limit of the running average of the expected queue lengths as the time goes to infinity is finite [18]. Based on these operating conditions, the following *necessary* average flow conservation condition needs to be satisfied for such queues to be stable (all expectations hereafter are with respect to the stationary distribution of $\mathbf{h}$, unless mentioned otherwise):

$$\bar{a}_i^f + \sum_{j \in \mathcal{N}(i)} \mathbb{E}\left[r_{j,i}^f(\mathbf{h})\right] \leq \sum_{j \in \mathcal{N}(i)} \mathbb{E}\left[r_{i,j}^f(\mathbf{h})\right], \quad \forall i, f \in \mathcal{F}(i). \tag{1}$$

Clearly, equation (1) is necessary because if it is not satisfied, the size of the queues will grow arbitrarily large. It is also assumed that the queues never became empty (this is known as full buffer assumption). Again, it is easy to prove that if that is not the case and (1) is satisfied with equality, the queues will grow arbitrarily large. The full buffer assumption is reasonable because the present formulation will aim at maximizing the average arrival rate. These issues will be discussed in more detail in Section 6.

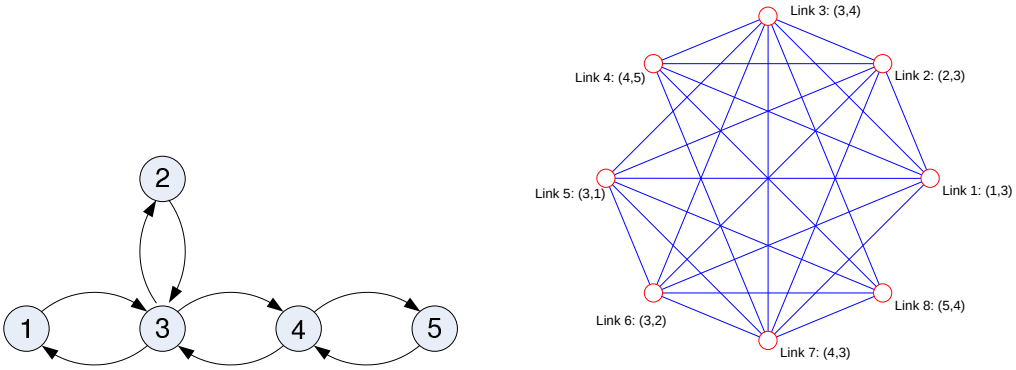
Lastly, it is important to remark that in the subsequent formulation, variables $\bar{a}_i^f$ will not be fixed, but optimally found as the solution of an optimization problem. From a practical point of view, this implies that nodes implement flow control at the transport layer. Further details are given in Section 3.

2.2 Link Layer Operation: Orthogonal Transmissions

As in e.g., [10, 29, 51, 62], links at the outset are allowed to access simultaneously but orthogonally (in time or frequency) any of the channels. Consideration of orthogonal access is well motivated from an operational perspective, since it decreases the complexity of the system and many deployed systems implement it. Moreover, from an optimality perspective, orthogonal access is (nearly) optimal when the interference is strong.

Since we consider orthogonal transmissions, the topology of the connectivity graph $\mathcal{G}$ plays a fundamental role in defining the sets of feasible scheduling policies. Intuitively, orthogonal access implies that most nodes in the neighborhoods of the origin and destination of an active link remain silent. More rigourously, if the link from node i to node j is active, the following links cannot be activated: (a) links whose origin or destination is i, (b) links whose origin or destination is j, (c) links whose destination is a neighbor of i, and (d) links whose origin is a neighbor of j. Those four conditions constitute the interference constraints of the present orthogonal access network model. Note that these include

[1] Since resources will be adapted every time the channel $\mathbf{h}$ changes, instantaneous variables are written as a function of $\mathbf{h}$.



(a) (b)

Figure 1: Example of connectivity graph $\mathcal{G}$ (left) and associated contention graph $\mathcal{W}$ (right). There are six maximal independent sets in $\mathcal{W}$: $\{(1,3),(5,4)\}$, $\{(4,5),(3,1)\}$, $\{(2,3),(5,4)\}$, $\{(4,5),(3,2)\}$, $\{(3,4)\}$, $\{(4,3)\}$

all constraints considered in the primary interference model (also called node-exclusive interference model)—see e.g., [10, 34, 54]—while they are included in the secondary interference model (2-hop interference model) [15, 54]. These interference constraints can be used to create what is oftentimes referred to as *contention graph* or *conflict graph* of $\mathcal{G}$; see e.g., [10]. The contention graph $\mathcal{W}$ is an undirected graph whose vertices are the directed links of the wireless network, and an edge connects two vertices of $\mathcal{W}$ if the corresponding directed links of $\mathcal{G}$ cannot be activated simultaneously. The contention graph $\mathcal{W}$ can be easily built in a systematic way using the connectivity graph $\mathcal{G}$ and the four orthogonality requirements (a)–(d). Fig. 1 illustrates an example of $\mathcal{G}$ and $\mathcal{W}$. Once $\mathcal{W}$ is known, one can rely on graph-theoretic tools to find the feasible scheduling policies. Specifically, an *independent set* is a set of vertices in a graph no two of which are adjacent; i.e, it is a set of vertices such that for every two vertices, there is no edge connecting the two. Moreover, a *maximal* independent set is an independent set that is not a subset of any other independent set; i.e., it is an independent set such that adding any other node to the set forces the set to contain an edge. Based on these definitions, it is clear that two physical links can be simultaneously activated only if they belong to the same independent set of $\mathcal{W}$. Therefore, finding feasible scheduling policies amounts to finding maximal independent sets of $\mathcal{W}$.

To rigorously formulate the orthogonality constraints, let s be a maximal independent set of the contention graph $\mathcal{W}$, and $\mathcal{S}$ be the collection of all maximal independent sets of $\mathcal{W}$. Since links are allowed to share orthogonally the access to the channel, $w_s^k(\mathbf{h})$ in $[0,1]$ will denote the nonnegative fraction of time the links in set s are scheduled per realization $\mathbf{h}$.[2] Clearly, the following relationship holds:

$$\sum_{s\in\mathcal{S}} w_s^k(\mathbf{h}) \leq 1, \quad \forall \mathbf{h}, k. \tag{2}$$

[2]Orthogonal sharing can also be implemented using e.g., orthogonal codes.

This way, if $w_s^k(\mathbf{h}) = 1$, the links in s can transmit during the entire duration of realization $\mathbf{h}$, while links that do not belong to s have to remain silent. On the other hand, if $w_s^k(\mathbf{h}) = 0.7$ and $w_{s'}^k(\mathbf{h}) = 0.3$, links in s can transmit during the 70% of the duration of realization $\mathbf{h}$, and links in s' can transmit during the remaining 30%. Time sharing is not necessarily difficult to implement in practice. For example, in existing OFDMA systems, typical bounds on the channel coherence and symbol intervals are 5–100 ms and 5–500 μs, respectively. This means that during a coherence interval several hundreds of symbols are transmitted; hence, those symbols can be assigned to different links. Even so, the ensuing analysis will show that in most situations time sharing is not needed because the optimal scheduling will assign the channel to a single independent set.

To account for the fact that a link can belong to different maximal independent sets, let $\mathcal{S}(i,j)$ denote the collection of maximal independent sets that contain link (i,j), i.e., $\mathcal{S}(i,j) := \{s \in \mathcal{S} : (i,j) \in s\}$, and let $w_{i,j}^k(\mathbf{h})$ in $[0,1]$ denote the nonnegative fraction of time the directed link (i,j) is scheduled during the current channel realization. Then it must hold that

$$w_{i,j}^k(\mathbf{h}) \leq \sum_{s \in \mathcal{S}(i,j)} w_s^k(\mathbf{h}), \quad \forall \mathbf{h}, k, i, j \in \mathcal{N}(i). \tag{3}$$

Clearly, optimal allocation requires active links to satisfy (3) with equality. Note also that (2) and (3) need to hold for each and every channel realization. Differently, (1) involves averages over the channel distribution, and therefore it does not depend on a specific $\mathbf{h}$. In the following, constraints that need to hold for all $\mathbf{h}$ will be referred to as *instantaneous* constraints, while constraints that do not depend on the specific realization of $\mathbf{h}$ will be referred to as *average* constraints.

2.3 Physical Layer Operation

The resources adapted at the physical layer are power and rate per link, per channel, and per CSI realization. Specifically, $p_{i,j}^k(\mathbf{h})$ denotes the instantaneous nominal power transmitted over channel k from node i to node j during the channel realization $\mathbf{h}$. Nominal here means that $p_{i,j}^k(\mathbf{h})$ is the power that would be transmitted if channel k were allocated to link (i,j) during the entire duration of $\mathbf{h}$. For the general case, where $w_{i,j}^k(\mathbf{h}) \leq 1$, the power *effectively* transmitted over channel k from node i to node j during channel realization $\mathbf{h}$ is $w_{i,j}^k(\mathbf{h})p_{i,j}^k(\mathbf{h})$. Furthermore, two power constraints are considered. On the one hand, the instantaneous power $p_{i,j}^k(\mathbf{h})$ is bounded by a maximum pre-specified level $\check{p}_{i,j}^k$ (spectral mask). On the other hand, the average transmitted power $\bar{p}_i := \mathbb{E}[\sum_k \sum_{j \in \mathcal{N}(i)} w_{i,j}^k(\mathbf{h})p_{i,j}^k(\mathbf{h})]$ cannot exceed a maximum average power budget $\check{p}_i$.

Under BER or capacity constraints, rate and power variables are coupled. This rate-power coupling is represented by the function $C_{i,j}^k(\mathbf{h}, p_{i,j}^k(\mathbf{h}))$. Similar to the power case, $C_{i,j}^k(\mathbf{h}, p_{i,j}^k(\mathbf{h}))$ represents the nominal transmitted rate, while $C_{i,j}^k(\mathbf{h}, p_{i,j}^k(\mathbf{h}))w_{i,j}^k(\mathbf{h})$ is the effective rate transmitted over channel k from node i to node j for the duration of channel realization $\mathbf{h}$. It is assumed throughout that the rate-power function $C_{i,j}^k(\mathbf{h}, p_{i,j}^k(\mathbf{h}))$ is increasing and strictly concave in $p_{i,j}^k(\mathbf{h})$. This holds generally for orthogonal access but, for example, not when multiuser interference is present (see Section 7 for details). For instance, if sufficiently strong error control coding is employed, $C_{i,j}^k(\mathbf{h}, p_{i,j}^k(\mathbf{h}))$ is given by Shannon's capacity formula $\log(1 + h_{i,j}^k p_{i,j}^k(\mathbf{h})/\sigma_j^k)$ [20], which is certainly increasing

and strictly concave. Additional examples of concave rate-power functions can be found in [37].

3 Problem Formulation

This section formulates the optimal resource allocation problem. Resource allocation algorithms will be designed so that lower average power consumption and higher exogenous average arrival rates are promoted. To this end, the rate utility functions $U_i^f(\cdot)$ are selected to be increasing (so that higher rates are promoted) and strictly concave (so that fairness among users and flows is enforced). Similarly, the power cost functions $J_i(\cdot)$ are chosen to be increasing and strictly convex. Note that different utilities may be chosen for different flows f. For example, utility functions that are almost linear are appropriate for best-effort traffic because user satisfaction increases as rate increases. On the other hand, applications such as video streaming with buffering may require the average rate to exceed a minimum prescribed value, and do not experience any significant improvement once that value has been achieved. For this kind of services, concave utility functions can be adopted to yield very high reward for rate increments when the minimum rate has not been achieved, but almost zero reward for rate increments above the minimum rate. Note however that the present formulation does not accommodate real-time traffic with hard delay constraints. Design trade-offs between power cost and rate utility can be accounted for by proper weighting of functions $U_i^f(\cdot)$ and $J_i(\cdot)$.

Taking into account the previous considerations, the optimal channel-adaptive cross-layer resource allocation is obtained as the solution of the following constrained optimization problem:

$$\mathsf{P} = \min_{\substack{\bar{a}_i^f, \bar{p}_i, r_{i,j}^f(\mathbf{h}), \\ w_{i,j}^k(\mathbf{h}), w_s^k(\mathbf{h}), p_{i,j}^k(\mathbf{h})}} \quad - \sum_{i, f \in \mathcal{F}(i)} U_i^f\left(\bar{a}_i^f\right) + \sum_i J_i\left(\bar{p}_i\right) \tag{4a}$$

$$\text{subj. to} \quad \bar{a}_i^f + \sum_{j \in \mathcal{N}(i)} \mathbb{E}[r_{j,i}^f(\mathbf{h})] \leq \sum_{j \in \mathcal{N}(i)} \mathbb{E}[r_{i,j}^f(\mathbf{h})], \quad \forall i, f \in \mathcal{F}(i) \tag{4b}$$

$$\sum_{f \in \mathcal{F}(i)} r_{i,j}^f(\mathbf{h}) \leq \sum_k w_{i,j}^k(\mathbf{h}) C_{i,j}^k(\mathbf{h}, p_{i,j}^k(\mathbf{h})), \quad \forall \mathbf{h}, i, j \in \mathcal{N}(i) \tag{4c}$$

$$\mathbb{E}\left[\sum_k \sum_{j \in \mathcal{N}(i)} w_{i,j}^k(\mathbf{h}) p_{i,j}^k(\mathbf{h})\right] \leq \bar{p}_i, \quad \forall i \tag{4d}$$

$$w_{i,j}^k(\mathbf{h}) \leq \sum_{s \in \mathcal{S}(i,j)} w_s^k(\mathbf{h}), \qquad \forall \mathbf{h}, k, i, j \in \mathcal{N}(i) \tag{4e}$$

$$\sum_{s \in \mathcal{S}} w_s^k(\mathbf{h}) \leq 1, \quad \forall \mathbf{h}, k \tag{4f}$$

$$r_{i,j}^f(\mathbf{h}) \geq 0, \quad \forall \mathbf{h}, i, j \in \mathcal{N}(i), f \in \mathcal{F}(i) \tag{4g}$$

$$w_{i,j}^k(\mathbf{h}) \geq 0, \quad w_s^k(\mathbf{h}) \geq 0, \quad \forall \mathbf{h}, k, i, j \in \mathcal{N}(i), s \in \mathcal{S} \tag{4h}$$

$$\bar{a}_i^f \geq 0, \qquad \forall i, f \in \mathcal{F}(i) \tag{4i}$$

$$0 \leq p_{i,j}^k(\mathbf{h}) \leq \check{p}_{i,j}^k, \ \forall \mathbf{h}, k, i, j \in \mathcal{N}(i); \quad 0 \leq \bar{p}_i \leq \check{p}_i, \ \forall i. \tag{4j}$$

The cross-layer nature of the resource allocation problem is apparent because variables of different layers are jointly optimized. The channel-adaptive attribute is also apparent since the optimization variables $r_{i,j}^f(\mathbf{h})$, $w_{i,j}^k(\mathbf{h})$, $w_s^k(\mathbf{h})$, and $p_{i,j}^k(\mathbf{h})$ are all functions of $\mathbf{h}$.

The objective in (4a) is constrained to several conditions. Constraints (4g)–(4j) effect lower and upper bounds on different variables, and are known as box constraints, which are easy to tackle. The remaining constraints enforce relationships among variables described in Section 2. Constraint (4b) corresponds to (1) and guarantees the flow conservation. Constraints (4e) and (4f) encapsulate the scheduling decisions at the link layer. Different from (4b) which needs only to hold on average, constraints (4e) and (4f) need to hold for *every* channel realization $\mathbf{h}$. The interaction among layers is manifested in (4c), which ensures that the number of packets routed during channel realization $\mathbf{h}$ never exceeds the instantaneous capacity of the wireless channel. Finally, (4d) represents the average power constraint introduced in Section 2.3. In fact, it can be easily shown that a problem with the same solution as (4) could be formulated by replacing $\bar{p}_i$ with $\mathbb{E}[\sum_k \sum_{j\in\mathcal{N}(i)} w_{i,j}^k(\mathbf{h}) p_{i,j}^k(\mathbf{h})]$ and eliminating (4d). The reason behind keeping $\bar{p}_i$ as a variable in (4d) is that it will be helpful for decoupling the optimality conditions of (4).

Problem (4) is nearly convex. In fact, the only source of non-convexity are the monomials $w_{i,j}^k(\mathbf{h})p_{i,j}^k(\mathbf{h})$ and $w_{i,j}^k(\mathbf{h})C_{i,j}^k(\mathbf{h}, p_{i,j}^k(\mathbf{h}))$. This source of non-convexity can be eliminated by introducing the auxiliary variables $u_{i,j}^k(\mathbf{h}) := w_{i,j}^k(\mathbf{h})p_{i,j}^k(\mathbf{h})$. It can be shown that if $p_{i,j}^k(\mathbf{h})$ in (4) is replaced by $u_{i,j}^k(\mathbf{h})/w_{i,j}^k(\mathbf{h})$, the problem becomes convex, and the reformulated problem yields the same Lagrangian as well as the same optimality conditions as those of (4); see e.g., [3, 37, 62]. Since both problems yield the same optimality conditions, the original formulation in (4) will be retained for brevity, without explicitly introducing the auxiliary variables $u_{i,j}^k(\mathbf{h})$.

4 Optimal Resource Allocation

In this section, the optimal solution of (4) is characterized as a function of the optimal multipliers associated with the constraints in (4), and the instantaneous CSI $\mathbf{h}$. This is accomplished using duality theory in order to find necessary and sufficient conditions for optimality (Section 4.1). The optimal resource allocation schemes at the physical, link, and network layers are developed based on these conditions (Section 4.2). Finally, an alternative solution for asymptotically optimal scheduling and routing schemes is presented and shown to offer distinct advantages relative to the optimal ones (Section 4.3).

4.1 Lagrangian and Optimality Conditions

Using the Lagrangian dual approach, necessary and sufficient conditions that the optimal solution of (4) needs to satisfy are identified in this section. Let ρ_i^f and π_i denote Lagrange multipliers associated with the average constraints in (4b) and (4d), respectively. Similarly, let $\nu_{i,j}(\mathbf{h})$, $\vartheta_{i,j}^k(\mathbf{h})$, $\xi^k(\mathbf{h})$, $\eta_{i,j}^{R,f}(\mathbf{h})$, $\eta_{i,j}^{W,k}(\mathbf{h})$, and $\eta_s^{W,k}(\mathbf{h})$ denote the Lagrange multipliers associated with instantaneous constraints (4c) and (4e)–(4h), respectively.[3] Multipliers associated with (4i) and (4j) are not introduced because they are not needed in the subsequent analysis.

[3]The dependence of the multipliers associated with instantaneous constraints on $\mathbf{h}$ is written explicitly throughout.

Furthermore, let $\mathbf{y}$ be a vector containing all the average primal variables, $\mathbf{x}(\mathbf{h})$ be a vector containing all the instantaneous primal variables, $\boldsymbol{\lambda}$ a vector containing all the Lagrange multipliers (dual variables) associated with average constraints, and $\boldsymbol{\chi}(\mathbf{h})$ a vector containing all the dual variables associated with instantaneous constraints.

The full Lagrangian of (4) is

$$\begin{aligned}\mathcal{L}(\mathbf{y},\mathbf{x}(\mathbf{h}),\boldsymbol{\lambda},\boldsymbol{\chi}(\mathbf{h})) := &-\sum_{i,f\in\mathcal{F}(i)} U_i^f(\bar{a}_i^f)+\sum_{i,f\in\mathcal{F}(i)}\rho_i^f\left(\bar{a}_i^f+\sum_{j\in\mathcal{N}(i)}\mathbb{E}[r_{j,i}^f(\mathbf{h})]-\sum_{j\in\mathcal{N}(i)}\mathbb{E}[r_{i,j}^f(\mathbf{h})]\right)\\ &+\sum_i J_i\left(\bar{p}_i\right)+\sum_i \pi_i\left(\mathbb{E}\left[\sum_k\sum_{j\in\mathcal{N}(i)} w_{i,j}^k(\mathbf{h})p_{i,j}^k(\mathbf{h})\right]-\bar{p}_i\right)\\ &+\sum_{i,j\in\mathcal{N}(i)}\nu_{i,j}(\mathbf{h})\left(\sum_{f\in\mathcal{F}(i)} r_{i,j}^f(\mathbf{h})-\sum_k w_{i,j}^k(\mathbf{h})C_{i,j}^k(\mathbf{h},p_{i,j}^k(\mathbf{h}))\right)\\ &+\sum_{k,i,j\in\mathcal{N}(i)}\vartheta_{i,j}^k(\mathbf{h})\left(w_{i,j}^k(\mathbf{h})-\sum_{s\in\mathcal{S}(i,j)} w_s^k(\mathbf{h})\right)+\sum_k\xi^k(\mathbf{h})\left(\sum_{s\in\mathcal{S}} w_s^k(\mathbf{h})-1\right)\\ &-\sum_{i,j\in\mathcal{N}(i),f\in\mathcal{F}(i)}\eta_{i,j}^{R,f}(\mathbf{h})r_{i,j}^f(\mathbf{h})-\sum_{k,i,j\in\mathcal{N}(i)}\eta_{i,j}^{W,k}(\mathbf{h})w_{i,j}^k(\mathbf{h})-\sum_{k,s\in\mathcal{S}}\eta_s^{W,k}(\mathbf{h})w_s^k(\mathbf{h}). \quad (5)\end{aligned}$$

Let $\mathbf{y}^*$, $\mathbf{x}^*(\mathbf{h})$, $\boldsymbol{\lambda}^*$, and $\boldsymbol{\chi}^*(\mathbf{h})$ denote the optimal solution and the associated Lagrange multipliers of (4). Due to the convexity of (4), the Karush-Kuhn-Tucker (KKT) conditions yield the following conditions for optimality [8, Sec. 5.5] ($\dot{g}$ denotes the derivative of g, and $f_{\mathbf{h}}(\mathbf{h})$ the probability density function of random vector $\mathbf{h}$):

$$\dot{J}_i(\bar{p}_i^*)-\pi_i^* = 0 \quad (6a)$$

$$-\dot{U}_i^f(\bar{a}_i^{f*})+\rho_i^{f*} = 0 \quad (6b)$$

$$-\nu_{i,j}^*(\mathbf{h})w_{i,j}^k(\mathbf{h})\dot{C}_{i,j}^k(\mathbf{h},p_{i,j}^{k*}(\mathbf{h}))+\pi_i w_{i,j}^k(\mathbf{h})f_{\mathbf{h}}(\mathbf{h})d\mathbf{h} = 0 \quad (6c)$$

$$\left(\rho_j^{f*}-\rho_i^{f*}\right)f_{\mathbf{h}}(\mathbf{h})d\mathbf{h}+\nu_{i,j}^*(\mathbf{h})-\eta_{i,j}^{R,f*}(\mathbf{h}) = 0 \quad (6d)$$

$$-\nu_{i,j}^*(\mathbf{h})C_{i,j}^k(\mathbf{h},p_{i,j}^{k*}(\mathbf{h}))+\pi_i^* p_{i,j}^{k*}(\mathbf{h})f_{\mathbf{h}}(\mathbf{h})d\mathbf{h}+\vartheta_{i,j}^{k*}(\mathbf{h})-\eta_{i,j}^{W,k*}(\mathbf{h}) = 0 \quad (6e)$$

$$-\sum_{(i,j)\in s}\vartheta_{i,j}^{k*}(\mathbf{h})+\xi^{k*}(\mathbf{h})-\eta_s^{W,k*}(\mathbf{h}) = 0 \quad (6f)$$

Conditions (6) must be supplemented with the complementary slackness conditions [8, Chapter 5] and the box constraints in (4i) and (4j). The complementary slackness conditions dictate that if at the optimal point a constraint is satisfied with strict inequality, then the corresponding optimal Lagrange multiplier is zero.

4.2 Characterizing the Optimal Solution

Conditions in Section 4.1 are used to characterize the optimal policies in the present section. To this end, define $\rho_{i,j}^* := \max_f\{\rho_i^{f*}-\rho_j^{f*}\}$, which will play an instrumental role in describing the optimal policies. (The convention $\rho_i^{f*}=0$ whenever $i=d(f)$

is adopted.) Let also $\dot{J}_i^{-1}(\cdot)$, $\dot{U}_i^{f\ -1}(\cdot)$ and $\dot{C}_{i,j}^{k\ -1}(\mathbf{h},\cdot)$ denote, respectively, the inverse functions of the derivatives of $J_i(\cdot)$, $U_i^f(\cdot)$, and $C_{i,j}^k(\mathbf{h},\cdot)$. The optimal allocation of average power, average arrival rate, and instantaneous power is described by the following two propositions.[4] (Hereafter, $[.]_a^b$ with a, b real denotes projection onto $[a,b]$, and $[.]_0^\infty$ denotes componentwise projection onto the nonnegative reals.)

Proposition 1 The optimal average power allocation and average arrival rate allocation are, respectively,

$$\bar{p}_i^* = \left[\dot{J}_i^{-1}(\pi_i^*)\right]_0^{\check{p}_i} \tag{7}$$

$$\bar{a}_i^{f*} = \left[\dot{U}_i^{f\ -1}(\rho_i^{f*})\right]_0^{\infty}. \tag{8}$$

The result in (7) follows after solving (6a) with respect to $\bar{p}_i^*$ and projecting such solution onto the feasible set $[0,\check{p}_i]$ defined by the average constraints in (4j). Similarly, (8) follows after solving (6b) with respect to $\bar{a}_i^{f*}$, and projecting the result onto the feasible set $[0,\infty)$. Note that (8) dictates the flow control implemented at the transport layer. To gain intuition, (8) can be rewritten as $\dot{U}_i^f(\bar{a}_i^{f*}) = \rho_i^{f*}$. The latter reveals that the optimal flow control policy consists of increasing $\bar{a}_i^{f*}$ until the marginal utility reaches the cost of injecting more exogenous traffic (measured by ρ_i^{f*}) into the network. To ensure stability in practice, one can conservatively select the long-term average rate of each node's exogenous traffic to stay slightly smaller than $\bar{a}_i^{f*}$.

Proposition 2 The optimal instantaneous power allocation is given by

$$p_{i,j}^{k*}(\mathbf{h}) = \left[\dot{C}_{i,j}^{k\ -1}\left(\mathbf{h}, \frac{\pi_i^*}{\rho_{i,j}^*}\right)\right]_0^{\check{p}_{i,j}^k}. \tag{9}$$

Interestingly, when $C_{i,j}^k(\mathbf{h},x) = \log(1+h_{i,j}^k x)$, (9) reduces to the well-known waterfilling formula $p_{i,j}^{k*}(\mathbf{h}) = [\rho_{i,j}^*/\pi_i^{k*} - 1/h_{i,j}^k]_0^{\check{p}_{i,j}^k}$ [20]. For the waterfilling case, higher values of π_i^* entail lower power and rate loadings (average power is a limiting factor), while higher values of $\rho_{i,j}^*$ entail higher power and rate loadings (satisfaction of average flow conservation constraint is critical, thus high rates are required). In fact, it is easy to see that the previous observations hold for any $C_{i,j}^k(\mathbf{h},\cdot)$ increasing and concave. The inverse of the derivative of the rate-power function naturally arises in different power control and resource allocation problems; see, e.g., [6, 29, 37].

While the optimal values $\bar{p}_i^*$, $\bar{a}_i^{f*}$, and $p_{ij}^{k*}(\mathbf{h})$ can be found in closed form, obtaining the optimal expressions for $r_{i,j}^{f*}(\mathbf{h})$, $w_{i,j}^{k*}(\mathbf{h})$, and $w_s^{k*}(\mathbf{h})$ is more intricate. The reason for this is that the Lagrangian in (5) is linear with respect to those variables, and dual Lagrangian methods are known to be challenged by linear constraints.

For this reason, before characterizing the optimal $w_{i,j}^{k*}(\mathbf{h})$ and $w_s^{k*}(\mathbf{h})$, some definitions are needed. First, for each link consider the functional

$$\varphi_W(\mathbf{h},k,i,j) := -\rho_{i,j}^* C_{i,j}^k(\mathbf{h}, p_{i,j}^{k*}(\mathbf{h})) + \pi_i^* p_{i,j}^{k*}(\mathbf{h}) \tag{10}$$

[4] Proofs of propositions are not presented due to space limitations. In some cases, a few lines discussing the main idea of the proof are provided.

which represents the instantaneous cost of scheduling channel k to link (i,j); that is, the cost of selecting $w_{i,j}^{k*}(\mathbf{h}) = 1$. Secondly, for each maximal independent set consider the aggregate functional ($\mathbb{1}_{\{X\}}$ is the indicator function taking the value 1 if expression X is true, and 0 otherwise)

$$\varphi_S(\mathbf{h}, k, s) := \sum_{(i,j)\in s} \left(\varphi_W(\mathbf{h}, k, i, j)\mathbb{1}_{\{\varphi_W(\mathbf{h},k,i,j)<0\}}\right) \tag{11}$$

which represents the cost of scheduling channel k to the maximal independent set s when the CSI is $\mathbf{h}$; i.e., the cost of selecting $w_s^{k*}(\mathbf{h}) = 1$. Define also the collection of independent sets attaining the minimum cost ($\wedge$ denotes the logical operator "and")

$$\mathcal{S}_S(\mathbf{h}, k) := \{s : s = \arg\min_{s'} \varphi_S(\mathbf{h}, k, s') \quad \wedge \quad \varphi_S(\mathbf{h}, k, s) < 0\}. \tag{12}$$

The following result holds. (For a set $\mathcal{X}$, $|\mathcal{X}|$ denotes cardinality.)

Proposition 3 The optimal instantaneous scheduling $w_s^{k*}(\mathbf{h})$ and $w_{i,j}^{k*}(\mathbf{h})$ satisfy:
(i) If $s \notin \mathcal{S}_S(\mathbf{h}, k)$, then $w_s^{k*}(\mathbf{h}) = 0$;
(ii) If $|\mathcal{S}_S(\mathbf{h}, k)| > 0$, then $\sum_{s\in\mathcal{S}_S(\mathbf{h},k)} w_s^{k*}(\mathbf{h}) = 1$; and
(iii) $w_{i,j}^{k*}(\mathbf{h}) = \mathbb{1}_{\{\varphi_W(\mathbf{h},k,i,j)<0\}} \sum_{s\in\mathcal{S}(i,j)} w_s^{k*}(\mathbf{h})$.

In words, the optimal solution schedules only maximal independent sets with minimum negative cost, which can be viewed as a greedy policy because for a given channel k not all the links are scheduled. This policy is oftentimes referred to as opportunistic allocation or *winner-takes-all* allocation, and it is known to be optimal for different problems; see e.g., [29], [60], [36] for user selection in cellular systems operating over fading channels, or, [27], [57] for max-weight scheduling in packet networks.

To find the optimal scheduling percentages among independent sets that attain the minimum cost, two different cases must be considered. If the minimum cost in channel k is attained by a single independent set—denote that set as $s^*(\mathbf{h}, k)$—then the second part of Proposition 3 allows writing the optimal instantaneous link scheduling in closed form as

$$w_s^{k*}(\mathbf{h}) = \mathbb{1}_{\{s=s^*(\mathbf{h},k)\}} \tag{13}$$

and

$$w_{i,j}^{k*}(\mathbf{h}) = \mathbb{1}_{\{(i,j)\in s^*(\mathbf{h},k) \ \wedge \ \varphi_W(\mathbf{h},k,i,j)<0\}}. \tag{14}$$

If several independent sets attain the minimum cost, i.e., if $|\mathcal{S}_S(\mathbf{h}, k)| > 1$, then the calculation of the percentage for each of the independent sets is more complicated. Section 4.3 discusses this issue. Last but not least, although for a given channel realization $\mathbf{h}$ the scheduling is opportunistic, long-term fairness is ensured in the sense that the set of links that wins access is different per channel realization.

A similar approach is followed to find the optimal $r_{i,j}^{f*}(\mathbf{h})$. The first step is to define the flow cost functional

$$\varphi_F(i, j, f) := \rho_j^{f*} - \rho_i^{f*} \tag{15}$$

which represents the cost of routing flow f through link (i,j). The second step is to define the set of optimal flows

$$\mathcal{S}_F(i,j) := \{f : f = \arg\min_{f'} \varphi_F(i,j,f') \quad \wedge \quad \varphi_F(i,j,f) < 0\}. \tag{16}$$

The optimal instantaneous rate of link (i,j) is $C^*_{i,j}(\mathbf{h}) := \sum_k w^{k*}_{i,j}(\mathbf{h}) C^k_{i,j}(\mathbf{h}, p^{k*}_{i,j}(\mathbf{h}))$.

Using these notational conventions, the following result holds.

Proposition 4 The optimal instantaneous routing $r^{f*}_{i,j}(\mathbf{h})$ satisfies:
(i) If $f \notin \mathcal{S}_F(i,j)$, then $r^{f*}_{i,j}(\mathbf{h}) = 0$; and
(ii) If $|\mathcal{S}_F(i,j)| > 0$, then $\sum_{f \in \mathcal{S}_F(i,j)} r^{f*}_{i,j}(\mathbf{h}) = C^*_{i,j}(\mathbf{h})$.

As before, the optimal solution is greedy, meaning that only flows with minimum negative cost are routed. In fact, flows only are allowed to use routes (hops) that decrease the value of the price ρ^{f*}_i. It can also be verified that the more negative the flow cost in (15) is, the more negative the link channel indicator in (10) becomes. This means that links that could give rise to a significant reduction of the flow cost are more likely to be scheduled. Furthermore, the more negative the flow cost is, the higher the value of the routing variable becomes if the link is scheduled. This is because the power in (9) is higher, thus the capacity is higher. Finally, it is stressed that the gain of a given link does not affect how different flows share the link, but only $C^*_{i,j}(\mathbf{h})$, which represents the total number of packets actually routed through that link.

If the minimum cost is attained by a single flow, we have that

$$r^{f*}_{i,j}(\mathbf{h}) = \mathbb{1}_{\{f \in \mathcal{S}_F(i,j)\}} C^*_{i,j}(\mathbf{h}). \tag{17}$$

If several flows attain the minimum cost, the specific value for each of the flows can be found using the results of Section 4.3.

4.3 Tie Resolution: Winner-Takes-Almost-All

The event of having different flows (or maximal independent sets) attaining the minimum cost will be henceforth referred to as a *tie*. The main difficulty in dealing with a tie is that Proposition 3-(ii) and Proposition 4-(ii) do not specify how resources have to be split among winners. The underlying reason is that although any arbitrary splitting minimizes the Lagrangian in (5), only a subset of those splittings (in many cases a single one) is the actual solution to the original constrained problem. One way to find the optimal primal solution when a tie occurs consists of selecting, among all possible tied schedulings (flows), the one that satisfies the *average* constraints with equality [37]. Although this approach is optimal, it does not lead to a closed-form solution. Furthermore, it requires knowing the exact Lagrange multiplier values; hence it is very sensitive to small inaccuracies.

To bypass such problems, we advocate a smooth *suboptimal* scheme to resolve ties that: (a) can be implemented for any number of elements in $\mathcal{S}_F(i,j)$ and $\mathcal{S}_S(\mathbf{h},k)$; (b) is available in closed form (thus incurs reduced computational burden); and (c) is continuous with respect to the Lagrange multipliers. Equally important, the next section establishes analytically that the proposed scheme is asymptotically optimal (meaning that the loss of optimality can be made arbitrarily small). To distinguish the smooth near-optimal

schemes developed next from their optimal counterparts, the notation $\tilde{x}^*$ will be used henceforth to denote the near-optimal version of the optimal x^*.

The first step to derive the smooth schemes is to realize that the number of elements in $\mathcal{S}_S(\mathbf{h}, k)$ and $\mathcal{S}_F(i, j)$ is critical to describe the optimal allocation. Since the condition for being a member of those sets is very restrictive (the cost has to be exactly equal to the minimum cost), we relax the definition of sets $\mathcal{S}_S$ and $\mathcal{S}_F$ so that more elements belong to them. Specifically, for the case of instantaneous link scheduling, define the minimum independent set cost

$$\varphi_S^*(\mathbf{h}, k) := \min_{s'} \varphi_S^k(\mathbf{h}, k, s'). \tag{18}$$

Then, we relax the definition of a tie so that now the (suboptimal) collection of independent sets which tie is [cf. (12)]

$$\tilde{\mathcal{S}}_S(\mathbf{h}, k) := \{s : \varphi_S(\mathbf{h}, k, s) - \varphi_S^*(\mathbf{h}, k) < \varepsilon_S \quad \wedge \quad \varphi_S(\mathbf{h}, k, s) < 0\} \tag{19}$$

where ε_S is a small positive number. Based on $\tilde{\mathcal{S}}_S(\mathbf{h}, k)$, the following suboptimal instantaneous link scheduling is proposed:[5]

$$\tilde{w}_s^{k*}(\mathbf{h}) := \mathbb{1}_{\{s \in \tilde{\mathcal{S}}_S(\mathbf{h},k)\}} \frac{\left(1 - \frac{\varphi_S(\mathbf{h},k,s) - \varphi_S^*(\mathbf{h},k)}{\varepsilon_S}\right)^2}{\sum_{s' \in \tilde{\mathcal{S}}_S(\mathbf{h},k)} \left(1 - \frac{\varphi_S(\mathbf{h},k,s') - \varphi_S^*(\mathbf{h},k)}{\varepsilon_S}\right)^2} \tag{20}$$

$$\tilde{w}_{i,j}^{k*}(\mathbf{h}) := \mathbb{1}_{\{\varphi_W(\mathbf{h},k,i,j) < 0\}} \sum_{s \in \mathcal{S}(i,j)} \tilde{w}_s^{k*}(\mathbf{h}). \tag{21}$$

This new allocation allows maximal independent sets whose cost is not minimum but ε_S-close to the minimum to be scheduled for transmission too, but in a proportional way; that is, sets with lower cost will access the channel during more time. Next, we consider several examples of two independent sets that tie at channel k: if $\varphi_S(\mathbf{h}, k, s_1) = \varphi_S(\mathbf{h}, k, s_2)$, then $\tilde{w}_{s_1}^{k*}(\mathbf{h}) = \tilde{w}_{s_2}^{k*}(\mathbf{h}) = 1/2$; if $\varphi_S(\mathbf{h}, k, s_1) + \varepsilon_S/2 = \varphi_S(\mathbf{h}, k, s_2)$, then $\tilde{w}_{s_1}^{k*}(\mathbf{h}) = 4\tilde{w}_{s_2}^{k*}(\mathbf{h}) = 4/5$; and if $\varphi_S(\mathbf{h}, k, s_1) + \varepsilon_S \leq \varphi_S(\mathbf{h}, k, s_2)$, then $\tilde{w}_{s_1}^{k*}(\mathbf{h}) = 1$ and $\tilde{w}_{s_2}^{k*}(\mathbf{h}) = 0$.

Proposition 5 The smooth suboptimal instantaneous schedulings $\tilde{w}_s^{k*}(\mathbf{h})$ and $\tilde{w}_{i,j}^{k*}(\mathbf{h})$ satisfy:
(i) If $s \notin \tilde{\mathcal{S}}_S(\mathbf{h}, k)$, then $\tilde{w}_s^{k*}(\mathbf{h}) = 0$;
(ii) If $|\tilde{\mathcal{S}}_S(\mathbf{h}, k)| > 0$, then $\sum_{s \in \mathcal{S}_S(\mathbf{h},k)} \tilde{w}_s^{k*}(\mathbf{h}) = 1$; and
(iii) $\tilde{w}_s^{k*}(\mathbf{h})$ and $\tilde{w}_{i,j}^{k*}(\mathbf{h})$ are continuous functions of $\boldsymbol{\lambda}^*$.

Note that Proposition 5 is true due to definition (20). Properties (i)–(ii) are similar to those in Proposition 3, while (iii) ensures continuity with respect to $\boldsymbol{\lambda}^*$. The sharing coefficient $\tilde{w}_s^{k*}(\mathbf{h})$ is continuous with respect to $\varphi_S(\mathbf{h}, k, s)$, and the latter is continuous with respect to $\boldsymbol{\lambda}^*$.

Proceeding in a similar manner for the instantaneous routing, consider the minimum flow cost

$$\varphi_F^*(i, j) := \min_{f'} \varphi_F(i, j, f'). \tag{22}$$

[5] From an optimization perspective, (20) is a smooth version of the optimal discontinuous scheduling; see e.g., [69] and [45]. Alternative smooth schedulings, for example based on sigmoidal functions, are also possible.

Note that the latter is very similar to the definition of $\rho^*_{i,j}$ given in Section 4.2; specifically, it holds that $\rho^*_{i,j} = -\varphi^*_F(i,j)$. Moreover, the set of suboptimal flows is [cf. (16)]

$$\tilde{\mathcal{S}}_F(i,j) := \{f : (\varphi_F(i,j,f) - \varphi^*_F(i,j)) < \varepsilon_F \quad \wedge \quad \varphi_F(i,j,f) < 0\} \tag{23}$$

where ε_F is a small positive number. Based on $\tilde{\mathcal{S}}_F(i,j)$, we propose the suboptimal instantaneous routing

$$\tilde{r}^{f*}_{i,j}(\mathbf{h}) := \tilde{C}^*_{i,j}(\mathbf{h}) \mathbb{1}_{\{f \in \tilde{\mathcal{S}}_F(i,j)\}} \frac{\left(1 - \frac{\varphi_F(i,j,f) - \varphi^*_F(i,j)}{\varepsilon_F}\right)^2}{\sum_{f \in \tilde{\mathcal{S}}_F(i,j)} \left(1 - \frac{\varphi_F(i,j,f) - \varphi^*_F(i,j)}{\varepsilon_F}\right)^2} \tag{24}$$

where $\tilde{C}^*_{i,j}(\mathbf{h}) := \sum_k \tilde{w}^{k*}_{i,j}(\mathbf{h}) C^k_{i,j}(\mathbf{h}, p^{k*}_{i,j}(\mathbf{h}))$. The proposed instantaneous routing satisfies the properties described in the next proposition.

Proposition 6 The smooth suboptimal instantaneous routing $\tilde{r}^{f*}_{i,j}(\mathbf{h})$ satisfies:
(i) If $f \notin \tilde{\mathcal{S}}_F(i,j)$, then $\tilde{r}^{k*}_{i,j}(\mathbf{h}) = 0$;
(ii) If $|\tilde{\mathcal{S}}_F(i,j)| > 0$, then $\sum_{f \in \tilde{\mathcal{S}}_F(i,j)} \tilde{r}^{f*}_{i,j}(\mathbf{h}) = \tilde{C}^*_{i,j}(\mathbf{h})$; and
(iii) $\tilde{r}^{f*}_{i,j}(\mathbf{h})$ is a continuous function of $\boldsymbol{\lambda}^*$.

As before, Proposition 6 is due to (24), and properties (i) and (ii) are similar to those in Proposition 4.

4.4 Layered Resource Allocation

Although the formulation in (4) allows for arbitrary dependence among variables of different layers, it turns out that the optimal schemes presented so far exhibit a layered structure. Indeed, the power and rate loadings at the physical layer depend only on the channel gain and the Lagrange multipliers [cf. (9)]. Once the physical layer is fixed, the links to be activated can be found via (20) and (21); hence, link scheduling does depend on variables of other layers, but in a simple way. With the optimal link and the physical allocation available, $C^*_{i,j}(\mathbf{h})$ can be readily obtained. Then, the network layer can find which flow(s) to route using (24), which requires only knowledge of the Lagrange multipliers. In other words, the way flows share a specific link does not depend on the lower layers; only the total (aggregate) number of packets is given by $C^*_{i,j}(\mathbf{h})$. Finally, the flow control implemented at the transport layer is based on $\bar{a}^{f*}_i$, which according to (8), only depends on the Lagrange multipliers.

The intuition behind this solution is that the Lagrange multipliers act as layer interfaces encapsulating all the cross-layer information which is relevant from a resource allocation point of view. These findings are consistent with those in [12] (non-fading case), and in [52] (fading case).

5 Ergodic Resource Allocation

5.1 Finding the Optimal Lagrange Multipliers

To implement the optimal resource allocation schemes of the previous section, the optimal multiplier vector $\boldsymbol{\lambda}^*$ must be known. However, $\boldsymbol{\lambda}^*$ cannot be obtained analytically from

the optimality conditions in Section 4.1, and numerical search is needed. This is possible using dual methods. Toward this objective, consider first the partial Lagrangian of (4) where only the contribution of the average constraints is included

$$\mathcal{L}(\mathbf{y},\mathbf{x}(\mathbf{h}),\boldsymbol{\lambda}) := -\sum_{i,f\in\mathcal{F}(i)} U_i^f\left(\bar{a}_i^f\right) + \sum_{i,f\in\mathcal{F}(i)} \rho_i^f \left(\bar{a}_i^f + \sum_{j\in\mathcal{N}(i)} \mathbb{E}[r_{j,i}^f(\mathbf{h})] - \sum_{j\in\mathcal{N}(i)} \mathbb{E}[r_{i,j}^f(\mathbf{h})]\right) + \sum_i J_i\left(\bar{p}_i\right) + \sum_i \pi_i \left(\mathbb{E}\left[\sum_k \sum_{j\in\mathcal{N}(i)} w_{i,j}^k(\mathbf{h}) p_{i,j}^k(\mathbf{h})\right] - \bar{p}_i\right). \quad (25)$$

Recall that all the instantaneous constraints (link scheduling and instantaneous routing), as well as nonnegativity constraints were already satisfied by the solution of the previous section. Thus, the focus here is to find the Lagrange multipliers associated with average constraints, namely, ρ_i^f and π_i for all i, f.

With $\mathcal{A}$ denoting the feasible set for the primal variables, $\mathcal{A} := \{\mathbf{y}, \mathbf{x}(\mathbf{h}) : (4c), (4e), (4f), (4g), (4h),$ and (4j) are satisfied$\}$, the dual function is defined as

$$D(\boldsymbol{\lambda}) := \inf_{(\mathbf{y},\mathbf{x}(\mathbf{h}))\in\mathcal{A}} \mathcal{L}(\mathbf{y},\mathbf{x}(\mathbf{h}),\boldsymbol{\lambda}) = \mathcal{L}(\mathbf{y}^*(\boldsymbol{\lambda}),\mathbf{x}^*(\mathbf{h},\boldsymbol{\lambda}),\boldsymbol{\lambda}), \quad (26)$$

which is always concave with respect to $\boldsymbol{\lambda}$ [4, Sec. 6.2]. Note that $\mathbf{y}^*(\boldsymbol{\lambda})$ and $\mathbf{x}^*(\mathbf{h},\boldsymbol{\lambda})$ in (26) can be obtained by substituting $\boldsymbol{\lambda}^* = \boldsymbol{\lambda}$ into the expressions of the optimal primal solutions presented in Section 4.2. Based on (26), the dual problem of (4) is

$$\mathsf{D} := \max_{\boldsymbol{\lambda}\geq\mathbf{0}} D(\boldsymbol{\lambda}). \quad (27)$$

Since problem (4) is convex, as long as it is strictly feasible, the duality gap between the primal and dual problems (4) and (27) is zero, i.e., $\mathsf{P} = \mathsf{D}$ [8, Sec. 5.2]. As a result, the value of $\boldsymbol{\lambda}$ optimizing (27) can be used to find the optimal primal solution. A standard approach to obtain $\boldsymbol{\lambda}^*$ is through a gradient iteration. However, this is impossible here because the linear constraints in (4) render $D(\boldsymbol{\lambda})$ non-differentiable with respect to some of the entries of $\boldsymbol{\lambda}$. In this case, one can resort to subgradient iterations. For the dual function (26) at a given point $\boldsymbol{\lambda}$, it is known that the constraint violation evaluated at the primal solution $\mathbf{x}^*(\mathbf{h},\boldsymbol{\lambda})$ and $\mathbf{y}^*(\boldsymbol{\lambda})$ is a subgradient [4, Sec. 8.1].

For decreasing and non-absolutely summable stepsizes, subgradient iterations are known to converge in the dual domain [4, Sec. 8.2]. However, for a finite number of iterations finding a (near-)feasible primal solution is not guaranteed. The problem is that $r_{i,j}^{f*}(\mathbf{h},\boldsymbol{\lambda})$ and $w_{i,j}^{k*}(\mathbf{h},\boldsymbol{\lambda})$ are typically discontinuous at $\boldsymbol{\lambda}^*$; therefore, small hovering in the dual domain around $\boldsymbol{\lambda}^*$ can give rise to significant differences in the primal domain; see also Section 7.2 for more details about convergence of subgradient iterations. Fortunately, this is not a problem for the suboptimal scheduling and routing of Section 4.3. The reason is that when viewed as a function of $\boldsymbol{\lambda}$, $\tilde{r}_{i,j}^{f*}(\mathbf{h},\boldsymbol{\lambda})$, $\tilde{w}_s^{k*}(\mathbf{h},\boldsymbol{\lambda})$, and $\tilde{w}_{i,j}^{k*}(\mathbf{h},\boldsymbol{\lambda})$ are Lipschitz continuous. (Recall that while for the optimal scheduling the transition from a tie to a single-winner is abrupt, for the suboptimal scheduling the transition is smooth avoiding any discontinuity.) Lipschitz continuity guarantees that proximity in the dual domain implies proximity in the primal domain. In the context of optimization algorithms, smoothing techniques have been successfully used as a mean to effect continuity or differentiability; see e.g., [69] and [45].

Before presenting the results for the smooth case, define the smooth version of the dual function as

$$\tilde{D}(\boldsymbol{\lambda}) := \mathcal{L}(\mathbf{y}^*(\boldsymbol{\lambda}), \tilde{\mathbf{x}}^*(\mathbf{h}, \boldsymbol{\lambda}), \boldsymbol{\lambda}) \tag{28}$$

where $\mathbf{y}^*(\boldsymbol{\lambda})$ and $\tilde{\mathbf{x}}^*(\mathbf{h}, \boldsymbol{\lambda})$ are obtained by substituting $\boldsymbol{\lambda}^* = \boldsymbol{\lambda}$ into the expressions of the optimal (smooth when needed) primal solutions presented in Sections 4.2 and 4.3. Similarly, the smooth version of the subgradient is defined as the vector $\boldsymbol{\partial}\tilde{D}(\boldsymbol{\lambda})$ with entries

$$\partial\tilde{D}_{\rho_i^f}(\boldsymbol{\lambda}) := \bar{a}_i^{f*}(\boldsymbol{\lambda}) + \sum_{j\in\mathcal{N}(i)} \mathbb{E}[\tilde{r}_{j,i}^{f*}(\mathbf{h}, \boldsymbol{\lambda})] - \sum_{j\in\mathcal{N}(i)} \mathbb{E}[\tilde{r}_{i,j}^{f*}(\mathbf{h}, \boldsymbol{\lambda})], \quad \forall i, f \in \mathcal{F}(i) \tag{29}$$

$$\partial\tilde{D}_{\pi_i}(\boldsymbol{\lambda}) := \mathbb{E}\left[\sum_k \sum_{j\in\mathcal{N}(i)} \tilde{w}_{i,j}^{k*}(\mathbf{h}, \boldsymbol{\lambda}) p_{i,j}^{k*}(\mathbf{h}, \boldsymbol{\lambda})\right] - \bar{p}_i^*(\boldsymbol{\lambda}), \quad \forall i. \tag{30}$$

Based on these definitions, the following convergence and optimality result follows.

Proposition 7 If $\mu > 0$ denotes a small constant stepsize, then for any $\boldsymbol{\lambda}^{(0)}$ there exists μ so that:
(i) the iteration

$$\boldsymbol{\lambda}^{(\ell)} = \left[\boldsymbol{\lambda}^{(\ell-1)} + \mu\boldsymbol{\partial}\tilde{D}(\boldsymbol{\lambda}^{(\ell-1)})\right]_0^\infty \tag{31}$$

converges to some point $\tilde{\boldsymbol{\lambda}}^*$, i.e., $\boldsymbol{\lambda}^{(\ell)} \to \tilde{\boldsymbol{\lambda}}^*$; and
(ii) at the convergence point: $D(\boldsymbol{\lambda}^*) \leq \tilde{D}(\tilde{\boldsymbol{\lambda}}^*) < D(\boldsymbol{\lambda}^*) + f(\varepsilon_S, \varepsilon_F)$, where $f(\cdot,\cdot)$ is a positive increasing function satisfying $f(\varepsilon_S, \varepsilon_F) \to 0$ as $(\varepsilon_S, \varepsilon_F) \to (0, 0)$.

Proposition 7 has various implications. As far as convergence is concerned, it provides a systematic algorithm to compute $\tilde{\boldsymbol{\lambda}}^*$. From a feasibility perspective, it guarantees that if $\tilde{\mathbf{x}}^*(\mathbf{h}, \boldsymbol{\lambda})$ is implemented at $\tilde{\boldsymbol{\lambda}}^*$, the average flow conservation and power constraints are satisfied with equality (recall that $\boldsymbol{\partial}\tilde{D}(\boldsymbol{\lambda}) = \mathbf{0}$ only if this holds). Finally, from an optimality perspective, it guarantees that the overall price paid for implementing the smooth instead of the optimal policy is asymptotically small.[6]. The last assertion is true because the bounds on the dual values given in Proposition 7-(ii), directly translate to bounds on the objective in (4a). A rigorous proof of this proposition for a problem related to the one investigated in this chapter can be found in [37].

5.2 Operational Mode: Offline and Online Phases

The proposed cross-layer channel-adaptive schemes operate in two phases: (a) an offline phase, which takes place before communication starts during the initialization phase; and (b) an online phase, which is executed during the communication process, every time the instantaneous CSI $\mathbf{h}$ is updated.

The main objective of the offline phase is to find the Lagrange multipliers $\tilde{\boldsymbol{\lambda}}^*$. As presented in Section 5.1, $\tilde{\boldsymbol{\lambda}}^*$ is found through (31), which basically describes dual subgradient

[6]In practice, the gap with respect to $D(\boldsymbol{\lambda}^*)$ is almost zero even for finite (small) values of ε_S and ε_F. This is true because the smooth schemes are slightly suboptimal only when ties occur, which are rare events.

iterations. It is known that such methods may have slow convergence in practice. Hence, hundreds of iterations may be needed in order to find Lagrange multipliers reasonably close to the optimal ones. Of course, the specific number of iterations depends on factors such as the initialization point, the stepsize, and the required accuracy.

Moreover, the computational burden for each iteration can be high, especially for large-scale networks. The reason is that the subgradients in (31) involve expectations over the channel distribution [cf. (29) and (30)]. In practice, these expectations are replaced by Monte Carlo estimates. The samples needed for this may be obtained (a) by drawing independent realizations from the distribution of $\mathbf{h}$, if it is known; or (b) by using actual channel measurements, if such are available. Method (b) works even if the measurements are correlated—which may happen in the present context when the fading process is correlated across time. In any case, the higher the size of the network is, the bigger the number of channel realizations needed to obtain a reliable estimate of the actual subgradients becomes. It is also worth mentioning that for each channel realization, the independent set of links that wins access to the channel needs to be found. Enumerating all maximal independent sets has in general exponential complexity; see e.g., [54]. Different from other scheduling schemes, as in e.g., [10, 34], such a computation here needs to be executed only once, before the offline phase starts. All maximal independent sets are identified upon the initialization, and for each channel realization generated, the independent set achieving the highest link quality indicator can be easily found. It should finally be stressed that although the offline phase incurs high computational burden, it only needs to be re-run every time the channel statistics or the system set-up changes.

In contrast, the online phase needs to be executed every time the CSI changes, which depends on the channel coherence interval. Then, based on the current value of the CSI $\mathbf{h}$, and the value of $\tilde{\boldsymbol{\lambda}}^*$ obtained from the offline phase, the resources at the different layers are adapted according to (9)–(24). Note that in the online phase, there is no need for re-computing $\mathcal{S}$, since it is already available from the offline phase. Although the optimal allocation at the physical and network layers only requires local CSI, obtaining the optimal link activations requires knowledge of the full $\mathbf{h}$ vector. Specifically, to find the optimal scheduling, the maximal independent set giving rise to the lowest cost needs to be found. To this end, nodes need to share either the channel gains of their local links, or, the value of their link cost indicators. Exchange of this information can be implemented using different options. In a decentralized approach, control channels may be available, over which nodes exchange their local information—see [3] for a protocol that implements this task. Under a hierarchical approach, the network can contain scheduler-node(s)—regular or dedicated—that gather the information, find the optimal link allocation, and broadcast it to the transmitting nodes.

5.3 Numerical Example

To illustrate how the developed schemes perform, numerical tests were simulated for the very simple case of the network in Fig. 2(a) involving $I = 3$ nodes, all of them connected. Moreover, $F = 3$ different flows—one for each possible destination—and $K = 4$ parallel channels are considered. The channel SNRs are exponentially distributed, their average power gain is 5 dB, and are assumed to be reciprocal (i.e., $h^k_{i,j} = h^k_{j,i}$). The utilities to be maximized are $U^f_i(x) = \log(1+x)$, if $(i, f) = (1, 3)$ or $(i, f) = (3, 2)$; $U^f_i(x) = \log(1/10+x)$, if $(i, f) = (2, 1)$; and zero for all other node-flow combinations. The power cost functions

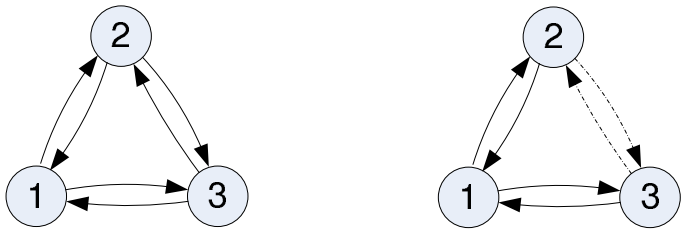


Figure 2: Network with 3 nodes, all connected with each other. For the first test case, all channels have the same SNR (left). For the second test case, the channel between nodes 2 and 3 is very weak (right).

are $J_i(x) = x^2/10$ forall i. Moreover a maximum average transmit power of $\check{p}_i = 5$ per node is considered.

The performance of the network during the online phase when the optimal multipliers are used is the following: $\bar{p}_1 = 1.5$, $\bar{p}_2 = 1.3$, $\bar{p}_3 = 1.5$; $\bar{a}_1^3 = 3.0$, $\bar{a}_2^1 = 2.6$, $\bar{a}_3^2 = 3.0$; and $\bar{r}_{1,3}^3 = 3.0$, $\bar{r}_{2,1}^1 = 2.6$, $\bar{r}_{3,2}^2 = 3.0$; and zero for all other variables. Note first that the network conditions are similar for nodes 1 and 3. This is not true for node 2, because it routes packets from flow 1, which according to the utilities considered, yield lower utility. Taking into account these facts, we observe that the numerical results follow the expected behavior since power and rate performance are similar for nodes 1 and 3; and the optimal exogenous rate injected at node 2 is smaller than that at nodes 1 and 3. To further validate the developed schemes, we slightly modify the set-up and reduce the average channel gain between nodes 3 and 2 by 10 dB. The modified configuration yields the following: $\bar{p}_1 = 1.8$, $\bar{p}_2 = 1.3$, $\bar{p}_3 = 1.2$; $\bar{a}_1^3 = 2.6$, $\bar{a}_2^1 = 2.3$, $\bar{a}_3^2 = 1.5$; and $\bar{r}_{1,3}^3 = 2.6$, $\bar{r}_{2,1}^1 = 2.3$, $\bar{r}_{3,2}^2 = 0.3$, $\bar{r}_{3,1}^2 = 1.2$, $\bar{r}_{1,2}^2 = 1.2$; and zero for all other variables. Since the channel between nodes 3 and 2 is now very poor, most of the packets from 3 destined for 2 are routed through 1. Indeed, this is confirmed by the numerical results. Moreover, we also observe that the power consumed by node 1 increases, the exogenous rate injected at node 3 decreases, and the overall network performance decreases. Clearly, all these changes are caused by the SNR loss between nodes 3 and 2.

6 Stochastic Resource Allocation

The resource allocation algorithms developed in the previous sections are functions of two variables: the current CSI $\mathbf{h}$, and the optimal (smooth) Lagrange multipliers. As mentioned earlier, finding $\tilde{\boldsymbol{\lambda}}^*$ offline requires knowledge of the channel distribution, and incurs considerable computational burden. To bypass these challenges, resources can be allocated using stochastic approximation algorithms. These algorithms learn the unavailable information on-the-fly, exhibit tracking capabilities, and incur moderate computational complexity. Roughly speaking, they could be understood as "intelligent" least mean-square (LMS) type schemes. From an operational perspective, they operate as *fully on-line* solutions because they do not require offline calculations and consume limited computational resources. Different alternatives can be considered to develop such stochastic schemes.

The approach here is to implement the cross-layer resource allocation not based on the optimal Lagrange multipliers, but on a stochastic estimate of them that varies with time n; i.e., $\tilde{\boldsymbol{\lambda}}^*$ is replaced by $\tilde{\boldsymbol{\lambda}}[n]$. It is important to clarify that n indexes blocks whose duration is the channel coherence interval. In other words, the resource allocation and the Lagrange multiplier estimates are updated every time the CSI $\mathbf{h} = \mathbf{h}[n]$ is updated. Recall that the fading process $\{\mathbf{h}[n]\}_{n=1}^{\infty}$ is assumed to be stationary and ergodic.

On top of coping with channel nonstationarities and reducing the computational burden, the stochastic algorithms facilitate distributed implementation, and link Lagrange multipliers with queue lengths analytically. This link is exploited in upcoming sections to: (a) characterize the average queueing delay of the stochastic schemes and, consequently, enable a way to incorporate explicitly the delay into the design; and (b) create links with existing algorithms that allocate resources based on the state of the queues.

6.1 Estimating the Lagrange Multipliers

The first step to obtain the stochastic schemes consists of replacing the original ensemble iterations in (29) and (30) with their instantaneous counterparts. Specifically, a Robbins-Monro approach is used to obtain $\tilde{\boldsymbol{\lambda}}[n]$, whereby all ensemble average terms in (29) and (30) are replaced by unbiased instantaneous (one-shot) estimates, as in e.g., [24]. Specifically, with $a_i^{f*}\big(n, \tilde{\boldsymbol{\lambda}}[n]\big)$ denoting the *instantaneous* arrival of flow f at node i during block n [which is a random variable drawn for a distribution with mean $\bar{a}_m^*(\boldsymbol{\lambda}[n])$], the following iterations are proposed:

$$\partial \tilde{D}_{\rho_i^f}(n, \tilde{\boldsymbol{\lambda}}[n]) := a_i^{f*}\big(n, \tilde{\boldsymbol{\lambda}}[n]\big) + \sum_{j \in \mathcal{N}(i)} [\tilde{r}_{j,i}^{f*}(\mathbf{h}[n], \tilde{\boldsymbol{\lambda}}[n])] - \sum_{j \in \mathcal{N}(i)} [\tilde{r}_{i,j}^{f*}(\mathbf{h}[n], \tilde{\boldsymbol{\lambda}}[n])] \tag{32}$$

$$\partial \tilde{D}_{\pi_i}(n, \tilde{\boldsymbol{\lambda}}[n]) := \left[\sum_k \sum_{j \in \mathcal{N}(i)} \tilde{w}_{i,j}^{k*}(\mathbf{h}[n], \tilde{\boldsymbol{\lambda}}[n]) p_{i,j}^{k*}(\mathbf{h}[n], \tilde{\boldsymbol{\lambda}}[n]) \right] - \bar{p}_i^*(\tilde{\boldsymbol{\lambda}}[n]) \tag{33}$$

where (32) does not apply if $i = d(f)$. Comparing (32) and (33) with (29) and (30), the expectations over $\mathbf{h}$ have indeed been dropped, and the ensemble $\boldsymbol{\lambda}$ has been replaced with its stochastic estimate $\tilde{\boldsymbol{\lambda}}[n]$. Moreover, the optimal average arrival $\bar{a}_i^{f*}(\tilde{\boldsymbol{\lambda}})$ has been replaced by its stochastic counterpart $a_i^{f*}\big(n, \tilde{\boldsymbol{\lambda}}[n]\big)$. Key to convergence analysis is that sequence $\{\partial \tilde{D}_{\lambda_i}(n, \tilde{\boldsymbol{\lambda}}[n])\}$ is bounded, which is true as far as both the instantaneous rates at the physical level and the exogenous instantaneous arrival rates at the network level are bounded. In addition, it holds by construction that $\mathbb{E}[\partial \tilde{D}_{\lambda_i}(n, \tilde{\boldsymbol{\lambda}}[n])] = \partial \tilde{D}_{\lambda_i}(\tilde{\boldsymbol{\lambda}}[n])$, where the expectation is conditioned to the history of the network; i.e., $\{\mathbf{h}[r]\}_{r=0}^{n}$ and $\{\tilde{\boldsymbol{\lambda}}[r]\}_{r=0}^{n}$.

Based on the previous definitions, the original iterations over $\boldsymbol{\lambda}$ in (31) are replaced by their stochastic counterparts

$$\tilde{\boldsymbol{\lambda}}[n+1] = \Big[\tilde{\boldsymbol{\lambda}}[n] + \mu \boldsymbol{\partial} \tilde{D}(n, \tilde{\boldsymbol{\lambda}}[n]) \Big]_0^{\infty} \tag{34}$$

where $\mu > 0$ denotes again a *constant* stepsize. Equation (34) shows clearly that now $\tilde{\boldsymbol{\lambda}}[n]$ depends on the random value of $\mathbf{h}[n]$, thus it is stochastic.

Once the stochastic multipliers are available, they are substituted into the primal optimal solution presented in Section 4, to yield the stochastic version of the primal

solution. Specifically, the stochastic version of the optimal average power, average arrival rate, and instantaneous power per time slot n are, respectively,

$$\bar{p}_i^*(\tilde{\boldsymbol{\lambda}}[n]) := \left[\dot{J}_i^{-1}(\tilde{\pi}_i[n])\right]_0^{\check{p}_i}, \qquad \bar{a}_i^{f*}(\tilde{\boldsymbol{\lambda}}[n]) := \left[\dot{U}_i^{f\ -1}(\tilde{\rho}_i^f[n])\right]_0^{\infty} \tag{35}$$

$$p_{i,j}^{k*}(\mathbf{h}[n], \tilde{\boldsymbol{\lambda}}[n]) := \left[\dot{C}_{i,j}^{k\ -1}(\mathbf{h}[n], \tilde{\pi}_i[n]/\tilde{\rho}_{i,j}[n])\right]_0^{\check{p}_{i,j}^k} \tag{36}$$

Similarly, the stochastic versions of the suboptimal instantaneous scheduling and routing are given by

$$\tilde{w}_s^{k*}(\mathbf{h}[n], \tilde{\boldsymbol{\lambda}}[n]) := \mathbb{1}_{\{s\in\tilde{\mathcal{S}}_S(\mathbf{h}[n],\tilde{\boldsymbol{\lambda}}[n],k)\}} \left(1 - \frac{\varphi_S(\mathbf{h}[n], \tilde{\boldsymbol{\lambda}}[n], k, s) - \varphi_S^*(\mathbf{h}[n], \tilde{\boldsymbol{\lambda}}[n], k)}{\varepsilon_S}\right)^2 \Bigg/ \sum_{s'\in\tilde{\mathcal{S}}_S(\mathbf{h}[n],\tilde{\boldsymbol{\lambda}}[n],k)} \left(1 - \frac{\varphi_S(\mathbf{h}[n], \tilde{\boldsymbol{\lambda}}[n], k, s') - \varphi_S^*(\mathbf{h}[n], \tilde{\boldsymbol{\lambda}}[n], k)}{\varepsilon_S}\right)^2 \tag{37}$$

$$\tilde{w}_{i,j}^{k*}(\mathbf{h}[n], \tilde{\boldsymbol{\lambda}}[n]) := \mathbb{1}_{\{\varphi_W(\mathbf{h}[n],\tilde{\boldsymbol{\lambda}}[n],k,i,j)<0\}} \sum_{s\in\mathcal{S}(i,j)} \tilde{w}_s^{k*}(\mathbf{h}[n], \tilde{\boldsymbol{\lambda}}[n]) \tag{38}$$

$$\tilde{C}_{i,j}^*(\mathbf{h}[n], \tilde{\boldsymbol{\lambda}}[n]) := \sum_k \tilde{w}_{i,j}^{k*}(\mathbf{h}[n], \tilde{\boldsymbol{\lambda}}[n]) \tilde{C}_{i,j}^{k*}(\mathbf{h}[n], p_{i,j}^{k*}(\mathbf{h}[n], \tilde{\boldsymbol{\lambda}}[n])) \tag{39}$$

$$\tilde{r}_{i,j}^{f*}(\mathbf{h}[n], \tilde{\boldsymbol{\lambda}}[n]) := \tilde{C}_{i,j}^*(\mathbf{h}[n], \tilde{\boldsymbol{\lambda}}[n]) \mathbb{1}_{\{f\in\tilde{\mathcal{S}}_F(\tilde{\boldsymbol{\lambda}}[n],i,j)\}} \left(1 - \frac{\varphi_F(\tilde{\boldsymbol{\lambda}}[n], i, j, f) - \varphi_F^*(\tilde{\boldsymbol{\lambda}}[n], i, j)}{\varepsilon_F}\right)^2 \Bigg/ \sum_{f\in\tilde{\mathcal{S}}_F(\tilde{\boldsymbol{\lambda}}[n],i,j)} \left(1 - \frac{\varphi_F(\tilde{\boldsymbol{\lambda}}[n], i, j, f) - \varphi_F^*(\tilde{\boldsymbol{\lambda}}[n], i, j)}{\varepsilon_F}\right)^2. \tag{40}$$

6.1.1 Convergence Results

This section deals with present convergence results for the dual and primal stochastic iterates. From a practical perspective, only convergence of primal variables is required. Nevertheless, convergence of dual variables offers design insights, and is used in the next section to relate stochastic Lagrange multipliers and queue lengths, where packets are stored before transmission.

First, results that guarantee that the dual stochastic estimates remain within a neighborhood of the optimal solution are presented.[7]

Proposition 8 If the initializations of (31) and (34) are the same, it holds that:
(i) Given $T > 0$, there exist $b_T > 0$ and $\mu_T > 0$ so that

$$\max_{1\le n\le T/\mu} \|\boldsymbol{\lambda}^{(n)} - \tilde{\boldsymbol{\lambda}}[n]\| \le c_T(\mu) b_T, \quad 0 \le \mu \le \mu_T \quad \text{with probability 1} \tag{41}$$

[7]The locking results provided in the proposition can be shown based on the averaging approach in [56, Chapter 9]. The main idea is that the Lipschitz continuity of $\boldsymbol{\partial}\tilde{D}(n, \tilde{\boldsymbol{\lambda}})$ with respect to $\tilde{\boldsymbol{\lambda}}$ can be used to prove that the most challenging conditions required in [56, Theorem 9.1] hold. A similar approach can be used to show the convergence in probability result in (42) when $n \to \infty$ [56, Theorem 9.5].

where $c_T(\mu) \to 0$ as $\mu \to 0$.
(ii) Given $A > 0$, there exists a random variable $W(\mu)$ so that

$$\max_{n\geq 1} \Pr\{\|\boldsymbol{\lambda}^{(n)} - \tilde{\boldsymbol{\lambda}}[n]\| > A\} \leq \Pr\{W(\mu) > A\} \tag{42}$$

where $W(\mu) \to 0$ as $\mu \to 0$ with probability 1.

Proposition 8 states that the dual iterates do not strictly converge to the optimal value but may hover around it. Since the resource allocation is a function of the stochastic multipliers, the stochastic version of the instantaneous primal variables (36)–(40) exhibit the same convergence behavior. This means that the stochastic instantaneous primal variables may hover around their optimal (non-stochastic) counterparts; e.g., $p_{i,j}^{k*}(\mathbf{h}[n], \tilde{\boldsymbol{\lambda}}^*) \neq p_{i,j}^{k*}(\mathbf{h}[n], \tilde{\boldsymbol{\lambda}}[n])$ even when $n \to \infty$.

The convergence of the primal iterates is characterized next, in terms of their sample averages.[8]

Proposition 9 The sample average of the stochastic resource allocation: (i) is feasible and (ii) entails a small loss of performance relative to the average non-stochastic solution of the problem in (4). Specifically, as $n \to \infty$, it holds with probability 1 that

$$\text{(i)} \quad \frac{1}{n}\sum_{r=1}^{n}\left(\sum_{j\in\mathcal{N}(i)}\left[\tilde{r}_{j,i}^{f*}(\mathbf{h}[r], \tilde{\boldsymbol{\lambda}}[r]) - \tilde{r}_{i,j}^{f*}(\mathbf{h}[r], \tilde{\boldsymbol{\lambda}}[r])\right]\right) \;\geq\; \frac{1}{n}\sum_{r=1}^{n}\bar{a}_i^{f*}(\tilde{\boldsymbol{\lambda}}[r]) \tag{43}$$

$$\frac{1}{n}\sum_{r=1}^{n}\left(\sum_{k,j\in\mathcal{N}(i)}\tilde{w}_{i,j}^{k*}(\mathbf{h}[r], \tilde{\boldsymbol{\lambda}}[r])p_{i,j}^{k*}(\mathbf{h}[r], \tilde{\boldsymbol{\lambda}}[r])\right) \;\leq\; \frac{1}{n}\sum_{r=1}^{n}\bar{p}_i^{*}(\tilde{\boldsymbol{\lambda}}[r])) \tag{44}$$

(ii) With $\delta_{\mathsf{P}}(\mu)$ denoting a small number proportional to the stepsize μ

$$-\sum_{i,f\in\mathcal{F}(i)} U_i^f\left(\frac{1}{n}\sum_{r=1}^{n}\bar{a}_i^{f*}(\tilde{\boldsymbol{\lambda}}[r])\right) + \sum_i J_i\left(\frac{1}{n}\sum_{r=1}^{n}\bar{p}_i^{*}(\tilde{\boldsymbol{\lambda}}[r])\right) \leq \tilde{\mathsf{P}}^* + \delta_{\mathsf{P}}(\mu). \tag{45}$$

In other words, both stochastic and non-stochastic allocations are asymptotically ($\delta_{\mathsf{P}}(\mu)$) optimal when $n \to \infty$. Of course, differences between the offline/online (non-stochastic) schemes presented in Section 5 and the fully online (stochastic) schemes presented in Section 6 arise also on issues such as speed of convergence to the optimal average values, or, sensitivity to changes in the channel realization, to name a few.

Equally relevant, convergence results can also be obtained when the averages in (43)-(45) are not defined as sample averages. For example, if they are defined using a finite-size sliding window averaging, or, as an exponentially decaying average, then convergence in distribution to a Gaussian random variable whose mean is the optimal solution of (4) can be shown. A rigorous proof along with characterization of the variance of the Gaussian distribution can be obtained using the results in [24, Chapter 11].

[8] Propositions 9 and 10 can be proved using the results in [7, 24, 49].

Remark 1 Although not presented here, stochastic implementations different that those in (32)–(34) can be proposed. These include variations with decreasing stepsizes, and replacement of the ensemble averages with windowed averages or sample averages; see [39, 61] for examples. Needless to say, the convergence claims in each of those cases are different (stronger) than those presented here. As it is shown in the next section, we focus here on the simpler implementation of (32)–(34), because it allows us to establish connections with queuing theory and existing resource allocation algorithms.

6.2 Queue Stability and Average Delay

Communication systems are typically equipped with queues that store packets from higher layers. Packets leave the queues with a rate that depends on the state of the channel and the resource allocation decisions. So far, such queues have been only indirectly accounted for in the problem formulation via (1), where average flow conservation constraints have been imposed as a necessary condition for stability. In what follows, the queue stability and the delay performance of the stochastic resource allocation algorithms will be characterized.

Characterizing the queuing delay in adaptive systems operating over fading channels is typically complicated because most of the standard assumptions do not hold. Indeed, departure times are typically correlated, and the distributions for the arrival and departure times are difficult to describe because they depend on the resource allocation algorithm and the underlying channel fading distribution. This is the case for the non-stochastic online algorithm developed in Sections 4 and 5. It will be nevertheless seen soon that the analysis for the stochastic algorithms is tractable. Moreover, the full buffer assumption (cf. Section 2.1) was required for non-stochastic algorithms to be stable, but it is not needed for their stochastic counterparts.

Starting with the analysis of queue dynamics, let $q_i^f[n]$ denote the queue size for flow f at node i at the time slot n. Moreover, recall that $a_i^{f*}(n, \tilde{\boldsymbol{\lambda}}[n])$ stands for the random *instantaneous* arrival, which is drawn for a distribution with mean $\bar{a}_m^*(\boldsymbol{\lambda}[n])$ given by (35). Then, the queue obeys for all f and all $i \neq d(f)$ the recursion

$$q_i^f[n+1] = \left[q_i^f[n] \; + \; a_i^{f*}(n, \tilde{\boldsymbol{\lambda}}[n]) + \sum_{j \in \mathcal{N}(i)} [\tilde{r}_{j,i}^{f*}(\mathbf{h}[n], \tilde{\boldsymbol{\lambda}}[n])] - \sum_{j \in \mathcal{N}(i)} [\tilde{r}_{i,j}^{f*}(\mathbf{h}[n], \tilde{\boldsymbol{\lambda}}[n])] \right]_0^{\infty}, \tag{46}$$

In practice, arrivals and departures are magnitudes that vary in a time scale smaller than n. This implies that definitions slightly different from the one in (46) are also possible (e.g., one could alternatively say that packets arriving in time slot n can only be transmitted in time slot $n+1$). Such differences are not relevant for the subsequent analysis and, as it will be apparent next, (46) has been chosen for simplicity.

At this point it is useful to particularize the stochastic iteration in (34) for the specific case of $\rho_i^f[n+1]$ (recall that ρ_i^f is the Lagrange multiplier associated with the flow conservation constraint of flow f at node i). According to (32) and (34), such iteration is

$$\tilde{\rho}_i^f[n+1] = \left[\tilde{\rho}_i^f[n] + \mu \left(a_i^{f*}(n, \tilde{\boldsymbol{\lambda}}[n]) + \sum_{j \in \mathcal{N}(i)} [\tilde{r}_{j,i}^{f*}(\mathbf{h}[n], \tilde{\boldsymbol{\lambda}}[n])] - \sum_{j \in \mathcal{N}(i)} [\tilde{r}_{i,j}^{f*}(\mathbf{h}[n], \tilde{\boldsymbol{\lambda}}[n])] \right) \right]_0^{\infty} \tag{47}$$

for all f and $i \neq d(f)$. Comparing (47) with (46), it is clear that $\rho_i^f[n]$ and $q_i^f[n]$ are related in a way that *the stochastic Lagrange multipliers can be interpreted as scaled values of the queue lengths*. Specifically, if $\rho_i^f[0] = \mu q_i^f[0]$, then it follows that $q_i^f[n] = \rho_i^f[n]/\mu$. Had the definition of the queue update in (46) been different, one could always bound the instantaneous difference between $q_i^f[n]$ and $\rho_i^f[n]$ for all n, and argue that after an initial transient period, the approximation $q_i^f[n] \approx \rho_i^f[n]/\mu$ would be accurate.

The previous finding is meaningful from different points of view, namely, (a) analyzing stability of the resource allocation algorithms; (b) estimating the queueing delay that packets will experience; and (c) establishing connections with other well-known cross-layer resource allocation algorithms. The ensuing subsections briefly elaborate on those issues.

6.2.1 Queue Stability

To analyze stability of the stochastic resource allocation in (34)–(40), the fact that $q_i^f[n] = \rho_i^f[n]/\mu$ implies the following result about the convergence of the sample average of the queue lengths.

Proposition 10 If $\bar{q}_i^f[n] := n^{-1}\sum_{r=1}^{n} q_i^f[r]$ denotes the sample average of the queue size $q_i^f[n]$ and δ_q is a small number proportional to the maximum update in (34), then

$$|\hat{\bar{q}}_i^f[n] - \tilde{\rho}_i^*/\mu| < \delta_q \text{ as } n \to \infty \qquad \text{w.p. 1.} \tag{48}$$

Therefore, it holds that when n grows large, $\bar{q}_i^f[n]$ is finite provided that $\tilde{\rho}_i^{f*}$ is finite, which is guaranteed if the original problem is feasible.

It is worth noting that although Proposition 10 establishes convergence of the sample average of the queue lengths to a finite value, bounds on the instantaneous size of the queues can also be characterized using the Lagrange multipliers bounds given in Proposition 8.

6.2.2 Average Delay

The relationship between queues and Lagrange multipliers can also be used to estimate the average queueing delay that the proposed stochastic resource allocation incurs. To this end, one can invoke Little's result [23] that asserts that with stable queues, the average delay is given by the average aggregate queue length divided by the average aggregate arrival rate. This implies that the average delay of a flow (say f) in the entire network is

$$\bar{d}^f := \frac{\sum_i \bar{q}_i^f}{\sum_i \bar{a}_i^f} \tag{49}$$

where in (49), $\bar{q}_i^f$ generically denotes the expected length of the queue that node i keeps for flow f. Moreover, with $\bar{r}_{j,i}^f$ denoting the average routing variable, the average delay experienced by packets of flow f while waiting in the queue of node i is

$$\bar{d}_i^f := \frac{\bar{q}_i^f}{\bar{a}_i^f + \sum_{j\in\mathcal{N}(i)} \bar{r}_{j,i}^f}. \tag{50}$$

Using the results in Proposition 10, it readily follows that the *average delays* for the stochastic resource algorithms presented in this chapter can be approximated as,

$$\bar{d}^f \approx \frac{1}{\mu}\frac{\sum_i \tilde{\rho}_i^{f*}}{\sum_i \bar{a}_i^{f*}(\tilde{\boldsymbol{\lambda}}^*)} \tag{51}$$

$$\bar{d}_i^f \approx \frac{1}{\mu}\frac{\tilde{\rho}_i^{f*}}{\bar{a}_i^{f*}(\tilde{\boldsymbol{\lambda}}^*) + \sum_{\forall j\in\mathcal{N}(i)} \mathbb{E}[\tilde{r}_{j,i}^{f*}(\mathbf{h}, \tilde{\boldsymbol{\lambda}}^*)]}. \tag{52}$$

In other words, the average delay of the stochastic algorithm can be estimated from the optimal solution of (4) and the stepsize of the proposed iterations.

Changing the Stepsize: Upon examining (51) and (52), it is apparent that changes in the stepsize induce changes in the average delay. Specifically,(51) and (52) reveal that the higher the stepsize, the smaller the average queuing delay. The intuition behind this is that high stepsizes will accelerate convergence and thus improve the ability to react against events that otherwise would increase the queuing delay. However, high stepsize values will also lead to more pronounced hovering in the dual domain, and may endanger convergence and stability if they are set beyond a certain level.

Besides quantifying the average delay, the expressions in (51) and (52) are also useful to effect delay priorities. Key for this purpose is the fact that the iterations in (31) and (34) converge not only if the stepsize is common to all entries of $\tilde{\boldsymbol{\lambda}}$, but also if the stepsize is different for each entry. This way, flows and nodes with stricter delay constraints can use a larger stepsize. In a nutshell, if one allows the stepsize to be dependent on i and f, then different delay performances can be obtained.

Sensitivity Analysis: Another issue of interest is to know how delay varies when any of the variables present in (4) is modified. This is non-trivial because in most cases $\tilde{\rho}_i^{f*}$ is not available in closed form, and even if it is, its dependence on other parameters is difficult to characterize. Convex optimization tools can be used to decipher properties of $\tilde{\rho}_i^{f*}$. Specifically, to rigorously study the effect of modifying variables present in (4) on the average delay of our algorithms, sensitivity analysis has to be used. This analysis is typically complex, although there are a few cases where it can be tractable. For the problem at hand, this is the case for the arrival exogenous rate $\bar{a}_i^{f*}$. In fact, if a node i accepts exogenous packets of a given flow f, the KKT conditions can be used to show that $\tilde{\rho}_i^{f*} = \dot{U}_i^f(\bar{a}_i^{f*})$. The latter implies that $\bar{d}_i^f \approx \mu^{-1}\dot{U}_i^f(\bar{a}_i^{f*})/\bar{a}_i^{f*}$. Upon differentiating, it follows that $\partial\bar{d}_i^f/\partial\bar{a}_i^{f*} \approx \ddot{U}_i^f(\bar{a}_i^{f*})/\bar{a}_i^{f*} - \dot{U}_i^f(\bar{a}_i^{f*})/(\bar{a}_i^{f*})^2$. This implies that $\partial\bar{d}_i^f/\partial\bar{a}_i^{f*} < 0$, because utilities are concave and increasing. As a result, we deduce that problems giving rise to higher $\bar{a}_i^{f*}$, exhibit lower delay. Space limitation prevents further elaboration on this subject.

6.3 Cross-Layer Design and Dynamic Backpressure

Section 4.4 demonstrated that the developed cross-layer schemes exhibit a layered structure, and also that interaction among layers is mainly encapsulated by the Lagrange multipliers ρ_i^{f*} and π_i^*. For the schemes in Section 6, one can go one step further in the interpretation of these multipliers. In fact, the results of Subsection 6.2.1 revealed that a scaled version of the queue length provides an unbiased estimate of ρ_i^{f*}. Changes in the queue lengths (thus changes in $\rho_i^f[n]$) induce changes in the allocation of resources

at all layers. Specifically, node-flow pairs with large queues give rise to: (a) reduced exogenous and high outgoing endogenous rates (transport and network layers), (b) high link quality indicator (link layer), and (c) high power and rate loadings (physical layer). Similarly, changes at the physical layer (e.g., changes in the channel gains) induce changes of resource allocation at physical, link, and network layers.

A major implication of the relationship between the stochastic Lagrange multipliers and queue sizes revealed in Subsection 6.2.1 is that the developed schemes are stable. This was not guaranteed at the outset because queue stability was never explicitly imposed in the present formulation. (Recall that the average flow conservation condition imposed via (1) is necessary but not sufficient.) An intuitive explanation for this behavior is that the developed schemes adapt resources at different layers to react against short-term changes in the state of channels and queues. This way, if the instantaneous arrival rates of a given flow happen to be high or the channel gains happen to be low during several block indices, then the adaptive schemes react to increase the transmit-power, select that flow to be routed, and reduce the average exogenous rate. All those decisions clearly contribute to network stabilization.

Equally important, the relationship between the stochastic Lagrange multipliers and queue sizes is also useful to draw connections between the schemes presented in this chapter and the celebrated dynamic backpressure algorithm. This algorithm was proposed in [58], and later extended to fading networks with cross-layer adaptation [18]. Instead of using dual optimization theory, the schemes in [18] are derived using a Lyapunov stability approach (actual and virtual queues are inputs of the Lyapunov function) under which the resource allocation objective is to stabilize the network. Comparing the results here with those in [18], it easy to infer that: (a) the optimal routing schemes are the same in both cases; and (b) the optimal resource allocation schemes at the link and physical layers are related, with the main difference being that the role which is played here by the Lagrange multipliers, is played by the virtual queues and tuning parameters in [18].

7 Non-Orthogonal Access

In the present section, the focus shifts to multi-hop wireless networks with non-orthogonal access. The main difference here is that all nodes are allowed to access the available tones simultaneously, treating the interfering transmissions as noise. An advantage of using a non-orthogonal approach is that scheduling becomes implicit in the power allocation. Therefore, there is no need for introducing the scheduling variables $w_{i,j}^k(\mathbf{h})$ and $w_s^k(\mathbf{h})$—recall that in the orthogonal case it was necessary to find the collection of all maximal independent sets, which has exponential complexity. A disadvantage is that the optimization problem for non-orthogonal access is generally non-convex, and may again incur exponential complexity. Non-convexity emerges because now the instantaneous capacity of each link depends on the SINR of that link, which couples the power allocation decisions. Non-convexity typically brings two undesirable effects: (a) there may not be efficient algorithms to find the optimal solution; and (b) using a dual approach to solve the problem is not optimal, because zero duality gap is not guaranteed.

Section 7.1 gives the optimal networking formulation with the SINR-limited physical layer, and corresponds to material of Sections 3 and 4. Then, an ergodic resource allocation algorithm is presented in Section 7.2, which is the counterpart of the one given

in Section 5 for the orthogonal case. Stochastic resource allocation algorithms are not presented here; the interested reader is referred to [49]. An alternative online algorithm is presented in [48].

Many of the model parameters and network design variables, such as the exogenous arrival rates $\bar{a}_i^f$, coincide with those for orthogonal access. These are only cursorily mentioned here; see Section 2 for more elaboration. On the other hand, the differences between the two models are stressed throughout this section.

The content of this section draws mainly from material reported in [17, 52], and the proofs for all results in the present section can be found in these works.

7.1 Problem Statement and Duality Properties

The cross-layer resource allocation problem for multi-hop networks with non-orthogonal access is formulated in Subsection 7.1.1, while a fundamental property of the optimization problem is stated in Subsection 7.1.2 using Lagrangian duality.

7.1.1 Problem Statement

The formulation developed here considers the long-term average rate of flow f from node i to node j, $\bar{r}_{i,j}^f$, instead of the instantaneous one; see also [18]. Note that this approach is not essential in the case of non-orthogonal access treated here; a formulation using $r_{i,j}^f(\mathbf{h})$ would also be possible. The flow conservation constraint corresponding to (1) or (4b) is

$$\bar{a}_i^f \leq \sum_{j\in\mathcal{N}(i)} \bar{r}_{i,j}^f - \sum_{j\in\mathcal{N}(i)} \bar{r}_{j,i}^f, \qquad \forall i, f \in \mathcal{F}(i). \tag{53}$$

The total average rate carried by any link cannot exceed the average capacity of the link, $\bar{c}_{i,j}$; this leads to the link capacity constraint

$$\sum_{f\in\mathcal{F}(i)} \bar{r}_{i,j}^f \leq \bar{c}_{i,j}, \qquad \forall i, j \in \mathcal{N}(i) \tag{54}$$

which can be viewed as a long-term average version of (4c). Note that there is no quantity in (4) corresponding to $\bar{c}_{i,j}$. Variables $\bar{c}_{i,j}$ represent the average data rates at the link layer, and they are related to the instantaneous capacities, as described next.

Link capacities are dictated by the SINR at the physical layer, whereby different nodes are allowed to use the same frequency to transmit and treat other nodes' transmissions as noise. This is the case when receiving nodes implement single-user decoding. The instantaneous SINR of link (i, j) over tone k is

$$\gamma_{i,j}^k(\mathbf{h}, \mathbf{p}(\mathbf{h})) := \frac{h_{i,j}^k p_{i,j}^k(\mathbf{h})}{\sigma_j^k + \sum_{(\kappa,l)\in\mathcal{I}_{i,j}} h_{\kappa,j}^k p_{\kappa,l}^k(\mathbf{h})} \tag{55}$$

where σ_j^k is the noise variance at node j over tone k, and $\mathcal{I}_{i,j}$ denotes the set of links causing interference to (i, j). This set consists of the links carrying: (a) incoming transmissions for j over k from nodes other than i, (b) outgoing transmissions from j, and (c) transmissions originating from nodes in $\mathcal{N}(j)$ intended for nodes other than j. Hence, this set takes the form

$$\mathcal{I}_{i,j} := \big\{(\kappa, l)\colon \kappa \in \mathcal{N}(j)\backslash\{i\}, l \in \mathcal{N}(\kappa);\ (i, l)\colon l \in \mathcal{N}(i)\backslash\{j\};\ (j, l)\colon l \in \mathcal{N}(j)\big\}. \tag{56}$$

Based on (56), the term $h^k_{j,j}\sum_{l\in\mathcal{N}(j)}p^k_{j,l}$ in the denominator in (55) represents self-interference to receiving node j from transmissions originating from j. In order to discourage this self-interference, and thus ensure half-duplex operation of the nodes, $h^k_{j,j}$ is set to a high (deterministic) value. Furthermore, interference from "far-away" links, corresponding to (κ,l) with $\kappa\in\mathcal{N}(j)$, $\kappa\neq j$ and $l\in\mathcal{N}(\kappa)$, $l\neq j$, is neglected. Note that all links causing interference to link (i,j) are explicitly accounted for in the orthogonal transmissions constraints considered in Section 2.2. Moreover, the SINR γ^f_{ij} depends on power allocation $\mathbf{p}(\mathbf{h})$ as well as on $\mathbf{h}$. Whenever needed, the latter is emphasized by writing $\gamma^f_{i,j}(\mathbf{h},\mathbf{p}(\mathbf{h}))$.

It is important to remark at this point that P-CSI is assumed throughout Section 7. The reason is that the subsequently developed results rely on the fact that the fading is continuous. This is of course true for P-CSI, but not when Q-CSI is available because $\mathbf{h}$ is then a discrete random vector.

The transmission rate (instantaneous capacity) of link (i,j) over tone k is described via a generic function of the SINR over that link and tone, $C^k_{i,j}(\gamma^k_{i,j})$, which is increasing and concave in the SINR. The instantaneous capacity depends on power allocation $\mathbf{p}(\mathbf{h})$ as well as on $\mathbf{h}$, because the SINR depends on those [cf. (55)]. This function is the counterpart of $C^k_{i,j}(\mathbf{h},p^k_{i,j}(\mathbf{h}))$ in Section 2.3. Note though that the rate of link (i,j) here depends on the power allocations over all neighboring links—via the SINR—and not only on the power allocation over link (i,j), as was the case with the orthogonal access. A few particular examples of $C^k_{i,j}(\gamma^k_{i,j})$ are now in order.

Supposing Gaussian codebooks with sufficiently large blocklengths, $C^k_{i,j}(\gamma^k_{i,j})$ takes the form

$$C^k_{i,j}(\gamma^k_{i,j})=\log\bigl(1+\gamma^k_{i,j}\bigr). \tag{57}$$

It is also possible to include a penalty term in (57), called SINR gap, Γ, in order to account for practical codes and adaptive modulation schemes [20]: $C^k_{i,j}(\gamma^k_{i,j})=\log\bigl(1+\gamma^k_{i,j}/\Gamma\bigr)$.

Furthermore, a high-SINR approximation of (57) can also be used for cross-layer optimization

$$C^k_{i,j}(\gamma^k_{i,j})=\ln\bigl(K^k_{i,j}\gamma^k_{i,j}\bigr) \tag{58}$$

where $K^k_{i,j}$ is a constant. It should be stressed that (58) is meaningful only when $\gamma^k_{i,j}>1/K^k_{i,j}$; otherwise the logarithm is a negative number. A more general model which includes (58) as a special case considers a strictly increasing $C^k_{i,j}(\gamma^k_{i,j})$ satisfying the condition

$$\ddot{C}^k_{i,j}(\gamma^k_{i,j})\cdot\gamma^k_{i,j}+\dot{C}^k_{i,j}(\gamma^k_{i,j})\leq 0,\quad\forall\,\gamma^k_{i,j}>0. \tag{59}$$

A linear function of the SINR (low-SINR approximation) is another model:

$$C^k_{i,j}(\gamma^k_{i,j})=K^k_{i,j}\gamma^k_{i,j} \tag{60}$$

where $K^k_{i,j}$ is a constant. The previously mentioned forms of the capacity function (57)–(60) have been used for cross-layer optimization in [11, 13, 17, 19, 33, 34, 43, 47, 64, 65].

The ergodic capacity $\bar{c}_{i,j}$ of link (i,j) is

$$\bar{c}_{i,j}:=\mathbb{E}\left[\sum_k C^k_{i,j}(\gamma^k_{i,j}(\mathbf{h},\mathbf{p}(\mathbf{h})))\right]. \tag{61}$$

As in Section 3, two kinds of power constraints are considered. The first are instantaneous spectral mask constraints, expressed as $0 \leq p_{i,j}^k(\mathbf{h}) \leq \check{p}_{i,j}^k$, as in (4j). Second, the average power consumed by a node in the network [cf. (4d)] is

$$\bar{p}_i := \mathbb{E}\left[\sum_{j\in\mathcal{N}(i)}\sum_k p_{i,j}^k(\mathbf{h})\right] \tag{62}$$

is constrained by a power budget $\check{p}_i$, i.e., $0 \leq \bar{p}_i \leq \check{p}_i$ for all i, as in (4j).

Average exogenous rates $\bar{a}_i^f$ in practical networks lie within application-specific bounds expressed as $a_{i\,\min}^f \leq \bar{a}_i^f \leq a_{i\,\max}^f$. Furthermore, the network designer may wish to impose upper bounds on link capacities $\bar{c}_{i,j}$ and multicommodity flows $\bar{r}_{i,j}^f$, that is $0 \leq \bar{c}_{i,j} \leq \check{c}_{i,j}$ and $0 \leq \bar{r}_{i,j}^f \leq \check{r}_i$. Note that such upper bounds are not used in the formulation of Section 3 [cf. (4g)], but they could have been included, leading also to appropriate modifications of Propositions 4 and 6. As in Section 4, the notation $\mathbf{y}$ is used to denote the collection of all average variables, namely, $\bar{a}_i^f$, $\bar{r}_{i,j}^f$, $\bar{c}_{i,j}$, and $\bar{p}_i$ for all i, $j \in \mathcal{N}(i)$, $f \in \mathcal{F}(i)$. Then, the previously mentioned box constraints are summarized by the polyhedral set

$$\begin{aligned}\mathcal{B}:=\{&\mathbf{y}, \mathbf{p}(\mathbf{h}) : 0 \leq p_{i,j}^k(\mathbf{h}) \leq \check{p}_{i,j}^k, 0 \leq \bar{p}_i \leq \check{p}_i,\\ &a_{i\,\min}^f \leq \bar{a}_i^f \leq a_{i\,\max}^f, 0 \leq \bar{c}_{i,j} \leq \check{c}_{i,j}, 0 \leq \bar{r}_{i,j}^f \leq \check{r}_i\}.\end{aligned} \tag{63}$$

Increasing and strictly concave utility functions $U_i^f(\bar{a}_i^f)$ for the exogenous arrival rates, and increasing and strictly convex cost functions $J_i(\bar{p}_i)$ for the average powers will be used. The optimal networking problem is [cf. (4)]

$$\mathsf{P} = \max_{(\mathbf{y},\mathbf{p}(\mathbf{h}))\in\mathcal{B}} \sum_{i,f\in\mathcal{F}(i)} U_i^f(\bar{a}_i^f) - \sum_i J_i(\bar{p}_i) \tag{64a}$$

$$\text{subj. to } \bar{a}_i^f \leq \sum_{j\in\mathcal{N}(i)} \bar{r}_{i,j}^f - \sum_{j\in\mathcal{N}(i)} \bar{r}_{j,i}^f, \qquad \forall i, f \in \mathcal{F}(i) \tag{64b}$$

$$\sum_{f\in\mathcal{F}(i)} \bar{r}_{i,j}^f \leq \bar{c}_{i,j}, \qquad \forall i, j \in \mathcal{N}(i) \tag{64c}$$

$$\bar{c}_{i,j} \leq \mathbb{E}\left[\sum_k C_{i,j}^k(\gamma_{i,j}^k(\mathbf{h}, \mathbf{p}(\mathbf{h})))\right] \quad \forall i, j \in \mathcal{N}(i) \tag{64d}$$

$$\mathbb{E}\left[\sum_{j\in\mathcal{N}(i)}\sum_k p_{i,j}^k(\mathbf{h})\right] \leq \bar{p}_i. \qquad \forall i. \tag{64e}$$

A comparison between (4) and (64) is due now. The objective (64a) corresponds to (4a). Constraints (64b) and (64e) parallel (4b) and (4d), respectively. Furthermore, there are no constraints corresponding to (4e), (4f), and (4h), as there are no scheduling variables here. The counterparts of constraints (4g)–(4j) are included in $\mathcal{B}$. Moreover, constraints (64c) and (64d) combined correspond to (4c)—note that there is also an additional variable here, namely $\bar{c}_{i,j}$. Introduction of $\bar{c}_{i,j}$ as optimization variable is not essential, but offers the advantage of providing an interface between the network layer average flows $\bar{r}_{i,j}^f$ and the instantaneous physical layer capacities $C_{i,j}^k(\gamma_{i,j}^k)$.

In practice, the traffic flowing in the network is stored in queues. The constraints in (64) are necessary conditions for stability [18]. The queueing operations in the network

and how the solution of (64) can be used for network control is described in detail in Subsection 7.2.3.

Problem (64) is non-convex in general, due to the coupling of the powers induced by interference, and the form of the instantaneous capacity function. For example, the form (57) appears in power control problems for digital subscriber lines (DSL), and is known to give rise to non-convex problems [22]. On the other hand, the functions (58) and (59) give rise to a convex problem under an appropriate change of variables [11], [64]. More details on the convexity of the problem and algorithmic solutions are discussed in Subsection 7.2.1. Next, a fundamental property of (64) is given, which is valid regardless of the convexity of (64).

7.1.2 Characterizing of the Optimal Solution via Duality

The KKT conditions for problem (64) are not sufficient for optimality, because the problem is non-convex in general (contrast with Section 4). Here, an alternative approach is followed, based on the Lagrangian dual of (64). Although impossible to derive the optimal solution of (64) as a function of the optimal Lagrange multipliers, it turns out that structure similar to convex problems is present in (64), which in turn can be used for efficient algorithmic solutions.

Let ρ_i^f, $\xi_{i,j}$, $\nu_{i,j}$, π_i be Lagrange multipliers corresponding to constraints (64b), (64c), (64d) and (64e), respectively. Note that there are two more Lagrange multipliers here than the ones of problem (4), namely $\xi_{i,j}$ and $\nu_{i,j}$, due to the two additional average constraints, (64c) and (64d). The box constraints (63) are kept implicit. Also let $\boldsymbol{\lambda}$ collectively denote all Lagrange multipliers. The Lagrangian function of (64) reduces after straightforward re-arrangements to

$$\begin{aligned}\mathcal{L}(\mathbf{y},\mathbf{p}(\mathbf{h}),\boldsymbol{\lambda}) = & \sum_{i,f\in\mathcal{F}(i)} \Big(U_i^f(\bar{a}_i^f) - \rho_i^f \bar{a}_i^f \Big) + \sum_i \left(\pi_i \bar{p}_i - J_i(\bar{p}_i)\right) \\ & + \sum_{i,j\in\mathcal{N}(i),k} \mathbb{E}\big[\nu_{i,j} C_{i,j}^k\big(\gamma_{i,j}^k(\mathbf{h},\mathbf{p}(\mathbf{h}))\big) - \pi_i p_{i,j}^k(\mathbf{h})\big] + \sum_{i,j\in\mathcal{N}(i)} \left(\xi_{i,j} - \nu_{i,j}\right) \bar{c}_{i,j} \\ & + \sum_{\substack{i,f\in\mathcal{F}(i),j\in\mathcal{N}(i)\\ j\neq d(f)}} \big(\rho_i^f - \rho_j^f - \xi_{i,j}\big)\bar{r}_{i,j}^f + \sum_{\substack{i,f\in\mathcal{F}(i),j\in\mathcal{N}(i)\\ j=d(f)}} \big(\rho_i^f - \xi_{i,j}\big)\bar{r}_{i,j}^f. \quad (65)\end{aligned}$$

It should be noted that the last sum in (65) might not be present.

In order to facilitate further development, (64) is rewritten in a more compact and generic form, which nevertheless captures all its essential features, as will be explained shortly. Upon rewriting all constraints in (64) with 0 on the right-hand side of the inequalities, (64) can be expressed as

$$\mathsf{P} = \max_{(\mathbf{y},\,\mathbf{p}(\mathbf{h}))\in\mathcal{B}} \quad f(\mathbf{y}) \quad (66a)$$

$$\text{subj. to} \quad \mathbf{g}_1(\mathbf{y}) + \mathbb{E}[\mathbf{g}_2(\mathbf{p}(\mathbf{h}),\mathbf{h})] \leq \mathbf{0} \quad (66b)$$

where the association of functions $f(\mathbf{y})$, $\mathbf{g}_1(\mathbf{y})$ and $\mathbf{g}_2(\mathbf{p}(\mathbf{h}),\mathbf{h})$ in (66) with the objective and constraints in (64) is straightforward. Regarding problem (66) [and hence (64) as a particular case], the following assumption is made.

Assumption 1 Function f is concave, $\mathbf{g}_1$ is convex (entrywise), and $\mathbf{g}_2$ is continuous. Constraint set $\mathcal{B}$ is convex, closed, bounded, and can be written in a decoupled form as $\mathcal{B} = \mathcal{B}_{\mathbf{y}} \times \mathcal{B}_{\mathbf{p}}$, where $\mathcal{B}_{\mathbf{p}}$ is independent of $\mathbf{h}$. Random vector $\mathbf{h}$ is continuous, i.e., has a probability density function without Dirac deltas. Finally, problem (66) is strictly feasible (Slater constraint qualification).

This assumption is satisfied by the setup of problem (64). Note that problem (66) is not convex in general. Moreover, although problem (4) is not a special case of (66), a comparison between Assumption 1 and the setup of problem (4) could be made. The setup of problem (4) does not require continuity of the fading, but instead assumes that the function playing the role of $\mathbf{g}_2(\mathbf{p}(\mathbf{h}), \mathbf{h})$—whether it appears inside or outside the expectation operator $\mathbb{E}[.]$—is convex in $\mathbf{p}(\mathbf{h})$.

Keeping the box constraints implicit, the Lagrangian function of (66) is

$$L(\mathbf{y}, \mathbf{p}(\mathbf{h}), \boldsymbol{\lambda}) = f(\mathbf{y}) - \boldsymbol{\lambda}^T \big(\mathbf{g}_1(\mathbf{y}) + \mathbb{E}[\mathbf{g}_2(\mathbf{p}(\mathbf{h}), \mathbf{h})]\big). \tag{67}$$

When (66) takes the particular form (64), vector $\boldsymbol{\lambda}$ contains the multipliers ρ_i^f, $\xi_{i,j}$, $\nu_{i,j}$, and π_i, and the Lagrangian function (67) clearly reduces to (65).

The dual function and the dual problem are, respectively,

$$D(\boldsymbol{\lambda}) := \max_{(\mathbf{y}, \mathbf{p}(\mathbf{h})) \in \mathcal{B}} L(\boldsymbol{\lambda}, \mathbf{y}, \mathbf{p}(\mathbf{h})) \tag{68}$$

$$\mathsf{D} = \min_{\boldsymbol{\lambda} \geq \mathbf{0}} D(\boldsymbol{\lambda}). \tag{69}$$

The main result is the following [52].

Proposition 11 Under Assumption 1, problem (66) [and hence (64)] has zero duality gap, i.e.,

$$\mathsf{P} = \mathsf{D}. \tag{70}$$

Recall that zero duality gap holds also for problem (4), which is convex. The result holds for (64), despite the possible non-convexity of the capacity function in the power allocation $\mathbf{p}(\mathbf{h})$. Furthermore, the traditional approach for cross-layer design and control has been to solve the dual problem of optimal networking formulations [12]. The result of Proposition 11 is relevant because it renders such approach optimal for wireless networks with continuous fading channels. In the next subsection, the subgradient method is used to solve (69).

7.2 Ergodic Resource Allocation

In this subsection, a solution of the constrained optimization task in (64) is sought via its Lagrangian dual. A subgradient algorithm for the solution of the dual problem (69) is presented in Subsection 7.2.1. As the subgradient algorithm returns Lagrange multipliers, an issue of interest is how to recover near-optimal network variables $\bar{a}_i^f$, $\bar{r}_{i,j}^f$, $\bar{c}_{i,j}$, $\bar{p}_i$, $p_{i,j}^k(\mathbf{h})$ from Lagrange multipliers; this is addressed in Subsection 7.2.2. In Subsection 7.2.3, the resulting (near-)optimal network variables are utilized to obtain a simple strategy for network control.

It should be stressed at this point that there are two complications in problem (64) that create the need for extra effort in order to obtain primal solutions: (a) the non-convexity in $\mathbf{p}(\mathbf{h})$, and (b) the linearity of the Lagrangian in $\bar{r}_{i,j}^f$ and $\bar{c}_{i,j}$. The second issue is also present in problem (4) for the variables $\bar{r}_{i,j}^f(\mathbf{h})$, $w_{i,j}^k(\mathbf{h})$, and $w_s^k(\mathbf{h})$. A smooth subgradient approach is pursued in Section 5 in order to overcome this issue. An alternative approach used in the context of convex optimization is followed here. This is forming the running average of the sequence of primal iterates obtained as byproduct of the subgradient method [26], [40]. In fact, it is argued in Subsection 7.2.2 that this method also works for the variables $\bar{a}_i^f$, $\bar{r}_{i,j}^f$, $\bar{c}_{i,j}$, and $\bar{p}_i$, despite the non-convexity in $\mathbf{p}(\mathbf{h})$.

7.2.1 Offline Phase

The dual problem (69) is solved via subgradient iterations [4, Sec. 8.2]. Different from Section 5, where a smooth subgradient method was used for the solution of (4), standard subgradients are employed here. Both approaches could be applied in both cases. The former approach has the difficulty to actually find a smooth subgradient for the particular problem, while the latter needs additional steps in order to yield the primal solutions, presented in Subsection 7.2.2.

With ℓ denoting the iteration index, the sequence $\boldsymbol{\lambda}^{(\ell)}$ obtained from the subgradient method, with initial $\boldsymbol{\lambda}^{(0)} \geq \mathbf{0}$, is

$$(\mathbf{y}^{(\ell)}, \mathbf{p}^{(\ell)}(\mathbf{h})) \in \underset{(\mathbf{y},\mathbf{p}(\mathbf{h}))\in\mathcal{B}}{\arg\max}\ \mathcal{L}(\mathbf{y}, \mathbf{p}(\mathbf{h}), \boldsymbol{\lambda}^{(\ell)}) \tag{71a}$$

$$\boldsymbol{\lambda}^{(\ell+1)} = \Big[\boldsymbol{\lambda}^{(\ell)} + \mu_\ell\big(\mathbf{g}_1(\mathbf{y}^{(\ell)}) + \mathbb{E}[\mathbf{g}_2(\mathbf{p}^{(\ell)}(\mathbf{h}), \mathbf{h})]\big)\Big]_0^\infty \tag{71b}$$

where the inclusion symbol ($\in$) in (71a) covers the possibility of multiple maximizers, and μ_ℓ is a positive stepsize which is allowed to vary with ℓ.

Using (65), (71a) becomes

$$\bar{a}_i^{f(\ell+1)} \in \underset{a_{i\,\min}^f \leq \bar{a}_i^f \leq a_{i\,\max}^f}{\arg\max}\ \big[U_i^f(\bar{a}_i^f) - \rho_i^{f(\ell)}\bar{a}_i^f\big] \tag{72a}$$

$$\bar{r}_{i,j}^{f(\ell+1)} \in \underset{0\leq\bar{r}_{i,j}^f\leq\check{r}_i}{\arg\max} \begin{cases} (\rho_i^{f(\ell)} - \rho_j^{f(\ell)} - \xi_{i,j}^{(\ell)})\bar{r}_{i,j}^f & \text{if } j \neq d(f) \\ (\rho_i^{f(\ell)} - \xi_{i,j}^{(\ell)})\bar{r}_{i,j}^f & \text{if } j = d(f) \end{cases} \tag{72b}$$

$$\bar{c}_{i,j}^{(\ell+1)} \in \underset{0\leq\bar{c}_{i,j}\leq\check{c}_{i,j}}{\arg\max}\ \big[(\xi_{i,j}^{(\ell)} - \nu_{i,j}^{(\ell)})\bar{c}_{i,j}\big] \tag{72c}$$

$$\bar{p}_i^{(\ell+1)} \in \underset{0\leq\bar{p}_i\leq\check{p}_i}{\arg\max}\big[\pi_i^{(\ell)}\bar{p}_i - J_i(\bar{p}_i)\big] \tag{72d}$$

$$\mathbf{p}^{(\ell)}(\mathbf{h}) \in \underset{\substack{0\leq p_{i,j}^k(\mathbf{h})\leq\check{p}_{i,j}^k \\ \forall k,i,j\in\mathcal{N}(i)}}{\arg\max} \sum_{k,i,j\in\mathcal{N}(i)} \big[\nu_{i,j}^{(\ell)} C_{i,j}^k\big(\gamma_{i,j}^k(\mathbf{h}, \mathbf{p}(\mathbf{h}))\big) - \pi_i^{(\ell)} p_{i,j}^k(\mathbf{h})\big]. \tag{72e}$$

Equation (72e) is obtained by noting that the part of (65) which involves the $\mathbb{E}[.]$ operator is maximized if the term inside the expectation is maximized for each fading state $\mathbf{h}$.

The subgradient updates (71b) take the explicit form

$$\rho_i^{f\,(\ell+1)} = \left[\rho_i^{f\,(\ell)} + \mu_\ell \left(\bar{a}_i^{f(\ell)} - \sum_{j\in\mathcal{N}(i)} \bar{r}_{i,j}^{f(\ell)} + \sum_{j\in\mathcal{N}(i)} \bar{r}_{j,i}^{f(\ell)}\right)\right]_0^\infty \tag{73a}$$

$$\xi_{i,j}^{(\ell+1)} = \left[\xi_{i,j}^{(\ell)} + \mu_\ell \left(\sum_{f\in\mathcal{F}(i)} \bar{r}_{i,j}^{f(\ell)} - \bar{c}_{i,j}^{(\ell)}\right)\right]_0^\infty \tag{73b}$$

$$\nu_{i,j}^{(\ell+1)} = \left[\nu_{i,j}^{(\ell)} + \mu_\ell \left(\bar{c}_{i,j}^{(\ell)} - \mathbb{E}\left[\sum_k C_{i,j}^k\big(\gamma_{i,j}^k(\mathbf{h}, \mathbf{p}^{(\ell)}(\mathbf{h}))\big)\right]\right)\right]_0^\infty \tag{73c}$$

$$\pi_i^{(\ell+1)} = \left[\pi_i^{(\ell)} + \mu_\ell \left(\mathbb{E}\left[\sum_{k,j\in\mathcal{N}(i)} p_{i,j}^{k\,(\ell)}(\mathbf{h})\right] - \bar{p}_i^{(\ell)}\right)\right]_0^\infty . \tag{73d}$$

In order to perform iterations (73), the solution of (72) is required. Each of the problems (72a)–(72d) involves a single variable, concave objective, and box constraints; thus, their solution as a function of the Lagrange multipliers is straightforward. For example, the optimal solutions of (72a) and (72d) can be obtained via (8)—with projection onto $[a_{i\,\min}^f, a_{i\,\max}^f]$ instead of $[0,\infty)$—and (7), respectively. However, the solution of (72e) may pose major challenges, depending on the form of $C_{i,j}^k(\gamma_{i,j}^k)$. Next, the implications of particular forms of the capacity function on the solution of (72e) are reviewed.

- When $C_{i,j}^k(\gamma_{i,j}^k)$ takes the form of Shannon capacity (57), (72e) carries similarities to the spectrum management problem in DSL. This problem may have exponential complexity [22]. Algorithmic solutions to (72e) can be developed by adaptation of successive approximation methods in the DSL literature [46], [59] or the condensation method [2], [1, pp. 151–152]. These algorithms are out of the scope of the present chapter; see [17] for examples. Moreover, approximate methods to deal with the particular non-convex capacity function have appeared in the cross-layer optimization literature. A randomized algorithm where links transmit either at full power or not at all is developed in [43]. Alternative randomized algorithms are also possible [28]. A heuristic algorithm is based on a trick whereby the total instantaneous transmission power at each node is held constant [64].
- In the case of the high-SINR approximation (58) or (59), the change of variables $v_{i,j}^k = \ln p_{i,j}^k$ makes (72e) convex in those variables. Then, any method for convex programs can be used.
- The low-SINR approximation (60), when substituted in (72e), leads after straightforward manipulations to a signomial program, also known as generalized geometric program; see [8, Exercise 4.35], and [1] for definitions. Such optimization problems are in general very hard to solve, and approximate solutions can be found using the methods in [1].

The subgradients in (73a) and (73b) are easily determined, once the solution of (72) is found. On the other hand, (73c) and (73d) involve the expectation $\mathbb{E}[.]$. This can be evaluated efficiently through Monte Carlo methods, as described in Section 5.2.

The following remark is due on the Lagrangian maximization in (72).

Remark 2 The decomposition of the Lagrangian maximization [cf. (71a)] into (72a)–(72e) can be understood as separation of the solution of the wireless networking problem

into conventional layers. This was also the case for the formulation under orthogonal access (cf. Section 4.4). More generally, it is a classical result that layered architectures emerge from the solution of the dual problem of cross-layer formulations [12]. In the present case, the decomposition into layers is optimal due to Proposition 11; the details of the induced layered architecture are described in [52]. In particular, (72a) solves the flow control problem at the transport layer; (72b) performs routing at the network layer; (72c) and (72d) address the link rate control and average power control at the data link layer; and (72e) solves the power allocation at the physical layer.

Similar to the approach in Section 6, it is possible to drop the expectations in (73c) and (73d), while using the maximizers in (72e) for only the current channel realization $\mathbf{h}[n]$. An online stochastic resource allocation algorithm is obtained in this way; the interested reader is referred to [49] for details.

7.2.2 Convergence Results

This subsection develops dual and primal convergence results for the algorithm in the previous subsection.

First, note that the norm of the subgradient, $||\mathbf{g}_1(\mathbf{y}^{(\ell)})+\mathbb{E}[\mathbf{g}_2(\mathbf{p}^{(\ell)}(\mathbf{h}),\mathbf{h})]||$, is bounded, because function $\mathbf{g}_1$ is convex, and hence continuous; function $\mathbf{g}_2$ is continuous; and the set $\mathcal{B}$ is closed and bounded. Thus, there exists a constant G such that $||\mathbf{g}_1(\mathbf{y}^{(\ell)}) + \mathbb{E}[\mathbf{g}_2(\mathbf{p}^{(\ell)}(\mathbf{h}),\mathbf{h})]|| \leq G$ for all $t \geq 0$. This constant will be used in the subsequent analysis.

Two stepsize rules are of interest: (a) constant stepsize $\mu_\ell = \mu > 0$; and (b) non-summable but square-summable stepsize: $\mu_\ell > 0$, $\sum_{\ell=0}^{\infty} \mu_\ell = \infty$, and $\sum_{\ell=0}^{\infty} \mu_\ell^2 < \infty$ with $\lim_{\ell\to\infty} \mu_\ell = 0$. Let

$$D_{\text{best}}^{(\ell)} := \min_{0\leq s\leq \ell} D(\boldsymbol{\lambda}^{(s)}) \tag{74}$$

denote the best dual value up to iteration ℓ; and let

$$\hat{\boldsymbol{\lambda}}^{(\ell)} := \frac{1}{\ell}\sum_{s=0}^{\ell-1} \boldsymbol{\lambda}^{(s)} \tag{75}$$

be the running average of the dual iterates with constant stepsize. The convergence of the dual iterates under the considered stepsize rules is given in the next proposition.

Proposition 12 Let Assumption 1 hold. If the stepsize is constant, $\mu_\ell = \mu > 0$, the following hold:

(i) The sequence $\{\boldsymbol{\lambda}^{(\ell)}\}$ is bounded;

(ii) $\lim_{\ell\to\infty} D_{\text{best}}^{(\ell)} \leq \mathsf{D} + \mu G^2/2$;

(iii) $\limsup_{\ell\to\infty} q(\hat{\boldsymbol{\lambda}}^{(\ell)}) \leq \mathsf{D} + \mu G^2/2$; and

(iv) The sequence $\{\hat{\boldsymbol{\lambda}}^{(\ell)}\}$ has at least one limit point $\tilde{\boldsymbol{\lambda}}^*$, and every limit point of this sequence satisfies $\mathsf{D} \leq D(\tilde{\boldsymbol{\lambda}}^*) \leq \mathsf{D} + \mu G^2/2$.

If the stepsize is non-summable but square-summable, that is, $\mu_\ell > 0$, $\sum_{\ell=0}^{\infty} \mu_\ell = \infty$, and $\sum_{\ell=0}^{\infty} \mu_\ell^2 < \infty$, then the sequence $\{\boldsymbol{\lambda}^{(\ell)}\}$ converges to some optimal dual solution.

The result on the diminishing stepsize and part (ii) for the case of constant stepsize are standard [4, Propositions 8.2.6 and 8.2.3]. Part (i) follows the lines [40, Lemma 3]. Part (iii) can be found in [52, Theorem 8]; see also [12] for dual averaging in the context of network utility maximization. Part (iv) follows readily from parts (i) and (iii).

The sequence of primal variables $\{\mathbf{y}^{(\ell)}, \mathbf{p}^{(\ell)}(\mathbf{h})\}$ obtained as a byproduct of the subgradient method [cf. (71a)] does not converge in general under either stepsize rule. Surprisingly, it is possible to recover optimal or approximately optimal primal variables from the sequence $\{\mathbf{y}^{(\ell)}, \mathbf{p}^{(\ell)}(\mathbf{h})\}$ for the generally non-convex primal problem in (66).

Let $K_\ell := \sum_{s=0}^{\ell-1} \mu_s$, and define the sequence of weighted running averages of $\mathbf{y}^{(\ell)}$

$$\hat{\mathbf{y}}^{(\ell)} := \frac{1}{K_\ell} \sum_{s=0}^{\ell-1} \mu_s \mathbf{y}^{(s)}, \quad \ell \geq 1 \tag{76}$$

where $\hat{\mathbf{y}}^{(\ell)}$ lies in $\mathcal{B}_\mathbf{y}$ due to the convexity of $\mathcal{B}_\mathbf{y}$, and the fact that $\sum_{s=0}^{\ell-1} \mu_s / K_\ell = 1$.

The convergence properties of the sequence $\{\hat{\mathbf{y}}^{(\ell)}\}$ and its proximity to the optimal solution of (66) are characterized in the following proposition for the considered stepsize rules [17].

Proposition 13 Let Assumption 1 hold. If the stepsize is constant, $\mu_\ell = \mu > 0$, then there exists a sequence $\{\mathring{\mathbf{p}}^{(\ell)}(\mathbf{h})\}$ in $\mathcal{B}_\mathbf{p}$ so that $\{\hat{\mathbf{y}}^{(\ell)}, \mathring{\mathbf{p}}^{(\ell)}(\mathbf{h})\}$ satisfies

(i) $\lim_{\ell\to\infty} \left\| \left[\mathbf{g}_1(\hat{\mathbf{y}}^{(\ell)}) + \mathbb{E}[\mathbf{g}_2(\mathring{\mathbf{p}}^{(\ell)}(\mathbf{h}), \mathbf{h})]\right]_0^\infty \right\| = 0$; and

(ii) $\liminf_{\ell\to\infty} f(\hat{\mathbf{y}}^{(\ell)}) \geq \mathsf{P} - \mu G^2/2$, and $\limsup_{\ell\to\infty} f(\hat{\mathbf{y}}^{(\ell)}) \leq \mathsf{P}$.

If the stepsize is non-summable but square-summable, that is, $\mu_\ell > 0$, $\sum_{\ell=0}^\infty \mu_\ell = \infty$, and $\sum_{\ell=0}^\infty \mu_\ell^2 < \infty$, then there exists a sequence $\{\mathring{\mathbf{p}}^{(\ell)}(\mathbf{h})\}$ in $\mathcal{B}_\mathbf{p}$ so that $\{\hat{\mathbf{y}}^{(\ell)}, \mathring{\mathbf{p}}^{(\ell)}(\mathbf{h})\}$ satisfies

(i) $\lim_{\ell\to\infty} \left\| \left[\mathbf{g}_1(\hat{\mathbf{y}}^{(\ell)}) + \mathbb{E}[\mathbf{g}_2(\mathring{\mathbf{p}}^{(\ell)}(\mathbf{h}), \mathbf{h})]\right]_0^\infty \right\| = 0$; and

(ii) $\lim_{\ell\to\infty} f(\hat{\mathbf{y}}^{(\ell)}) = \mathsf{P}$.

Part (i) of Proposition 13 under both stepsize rules means that the running average $\hat{\mathbf{y}}^{(\ell)}$ is asymptotically feasible, and there exists some feasible $\mathring{\mathbf{p}}^{(\ell)}(\mathbf{h})$ associated with it. In more precise terms, the constraint violation caused by $\{\hat{\mathbf{y}}^{(\ell)}, \mathring{\mathbf{p}}^{(\ell)}(\mathbf{h})\}$ converges to zero. Moreover, $\hat{\mathbf{y}}^{(\ell)}$ with constant stepsize incurs loss of optimality that is at most $\mu G^2/2$, while $\hat{\mathbf{y}}^{(\ell)}$ with vanishing stepsize is optimal. It should be stressed at this point that Proposition 13 establishes the optimality of the average network variables only, i.e., end-to-end rates, network layer flows, link capacities, average powers. It does not provide a way to obtain the instantaneous $\mathring{\mathbf{p}}^{(\ell)}(\mathbf{h})$ associated with those. A method to obtain an instantaneous power allocation $\mathbf{p}(\mathbf{h})$ regardless of the convexity of the problem is described in the next subsection.

It is further important to remark that results similar to Proposition 13 would hold for the orthogonal access case if in Section 5 standard subgradients were used instead of smooth subgradients. Similarly, primal averaging would not be needed in the present case if a smooth version of the dual function (68) were considered, and then, a result similar to Proposition 7 would be applicable here too.

Iterations (72) and (73) together with (76) solve the dual problem (69), and find near-optimal $\bar{a}_i^f$, $\bar{r}_{i,j}^f$, $\bar{c}_{i,j}$, $\bar{p}_i$ for problem (64). In the next subsection, a simple strategy for network control based on the optimal solution of (64) is described, including a method to obtain an instantaneous power allocation $\mathbf{p}(\mathbf{h})$.

7.2.3 Online Phase

This subsection describes a network control algorithm that utilizes the solution of problem (64)—that is, the optimal $\bar{a}_i^{f*}$, $\bar{r}_{i,j}^{f*}$, $\bar{c}_{i,j}^{*}$, $\bar{p}_i^{*}$, $\mathbf{p}^*(\mathbf{h})$. The offline algorithm of Subsections 7.2.1 and 7.2.2 needs to be run before the communication starts, in order to obtain this solution. This was also the case for the algorithm in Section 5.2. Simulation results from the offline and online phases are presented in [17].

Recall that the network operates in time slots, indexed by $n = 1, 2, 3, \ldots$, whose duration is the coherence time of the channel. Hence, the vector $\mathbf{h}$ will change from slot to slot, but is assumed to remain constant for the duration of a slot. If the channel value at slot n is $\mathbf{h}[n]$, and the power allocation is $\mathbf{p}(\mathbf{h}[n])$, then the transmission rate at link (i, j) over tone k is $C_{i,j}^k\big(\gamma_{i,j}^k\big(\mathbf{h}[n], \mathbf{p}(\mathbf{h}[n])\big)\big)$. The fading process $\{\mathbf{h}[n]\}_{n=1}^{\infty}$ is stationary and ergodic. The algorithm determines how the various flows and powers are allocated per time slot, based on the solution of (64). Next, it is clarified what is meant by the solution $\bar{a}_i^{f*}$, $\bar{r}_{i,j}^{f*}$, $\bar{c}_{i,j}^{*}$, $\bar{p}_i^{*}$, $\mathbf{p}^*(\mathbf{h})$.

In particular, near-optimal $\bar{a}_i^{f}$, $\bar{r}_{i,j}^{f}$, $\bar{c}_{i,j}$, and $\bar{p}_i$ are obtained in the offline algorithm as the value of the running averages sequence $\{\hat{\mathbf{y}}^{(\ell)}\}$ at the last iteration (cf. Proposition 13). Those near-optimal values will be used whenever $\bar{a}_i^{f*}$, $\bar{r}_{i,j}^{f*}$, $\bar{c}_{i,j}^{*}$, and $\bar{p}_i^{*}$ are mentioned here. In order to obtain the power allocation at time slot n, (72e) is solved, where the final values of the Lagrange multipliers (or of their running average under constant stepsize) obtained with the subgradient method are used in place of $\nu_{i,j}^{(\ell)}$ and $\pi_i^{(\ell)}$ (the same for all n), and $\mathbf{h}[n]$ is used for the SINR. The notation $\mathbf{p}^\dagger(\mathbf{h}[n])$ will be used for the aforementioned solution. This procedure for obtaining $\mathbf{p}^\dagger(\mathbf{h}[n])$ is reminiscent of the one in Proposition 2, which gives the optimal power allocation as function of the current fading state and the optimal Lagrange multipliers. Note though that obtaining the power allocation in the present case through maximization of the Lagrangian [cf. (72e)] with the optimal Lagrange multipliers is not necessarily optimal.

Problem (72e) is solved at a central network controller which knows the current fading state $\mathbf{h}[n]$ without delay. The need for a central controller comes from the fact that problem (72e) couples the power allocations over all links via the SINR. Works developing online algorithms also make use of a central controller [18]. Even though distributed solvers are also desirable, as for example in the decentralized approach in Section 5.2 for the case of orthogonal access, they go beyond the scope of the present chapter. It is mentioned nevertheless that the message passing protocols of [3,19,64,65] might be useful to this end. Next, a closer look at the queues maintained at each layer is due, before the network control algorithm is detailed.

Each node i keeps a queue for each commodity $f \in \mathcal{F}(i)$ at the network layer, and a queue for each neighbor $j \in \mathcal{N}(i)$ at the physical layer; see Fig. 3. Every network layer queue accepts exogenous traffic—from the transport layer—with instantaneous rate $a_i^f[n]$. Every physical layer queue sends to the corresponding neighbor bits with instantaneous rate $c_{i,j}[n]$, which depends on the instantaneous power allocation and the fading; this effect will be described in detail later. There is an interface connecting the network layer and physical layer queues at node i. This interface is responsible for routing, because it takes bits from network layer queues, and places them into physical layer queues. The individual rate from the fth network layer queue to the jth physical layer queue is denoted by $r_{i,j}^f[n]$. Then, bits leave the fth network layer queue with instantaneous

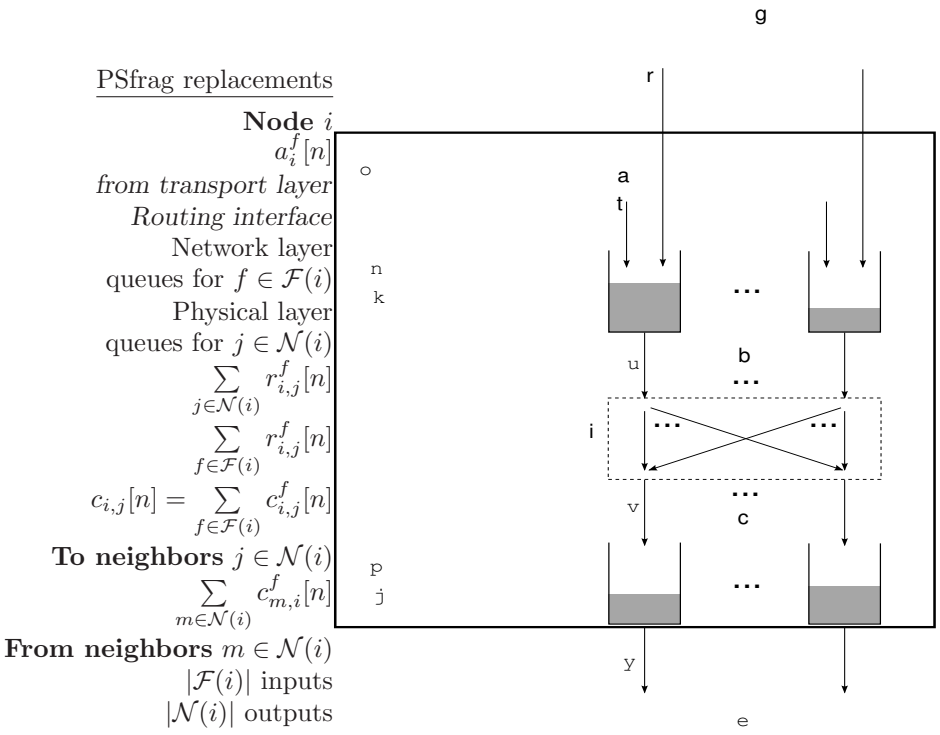


Figure 3: Queues at node i and connections to neighbors.

rate $\sum_{j\in\mathcal{N}(i)} r_{i,j}^f[n]$, and arrive at the j physical layer queue with rate $\sum_{f\in\mathcal{F}(i)} r_{i,j}^f[n]$. Note that the data in the physical layer queues are labeled with the flow index f, which determines their destination. The bits $c_{i,j}[n]$ from node i to node j are placed into the corresponding network layer queues of node j according to their label, except those with $d(f) = j$, which have arrived at their destination, and are not placed in any queue. Now, consider this operation for the bits $c_{m,i}[n]$ arriving to i from its neighbors $m \in \mathcal{N}(i)$; the endogenous arrivals at the fth network layer queue are $\sum_{m\in\mathcal{N}(i)} c_{m,i}^f[n]$, where the $c_{m,i}^f[n]$ are determined by splitting $c_{m,i}[n]$ according to the destination. Each queue operates in a first-in-first-out (FIFO) fashion, and has unlimited storage space. The number of bits entering or leaving the queues is conventionally assumed to take continuous values. This means equivalently that the packet sizes are small as compared to the number of bits that the network control algorithm specifies to be moved at each slot.

A network control algorithm must determine $a_i^f[n]$, $r_{i,j}^f[n]$, and $c_{i,j}[n]$. Variables $c_{i,j}[n]$ also depend on the fading, which of course cannot be controlled. Here, $a_i^f[n]$, $r_{i,j}^f[n]$, and $c_{i,j}[n]$ will be determined by $\bar{a}_i^{f*}$, $\bar{r}_{i,j}^{f*}$, and $\mathbf{p}^\dagger(\mathbf{h}[n])$.

The network operates under the premise that the random arrival process has the optimal long-term average, i.e., $\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} a_i^f[n] = \bar{a}_i^{f*}$. This operating condition for $a_i^f[n]$ is adopted, because $\bar{a}_i^{f*}$ is the optimal operating point of the network, determined by problem (64). As long as there are always enough packets available at the trasport layer—which amounts to the full buffer assumption at the transport layer—this condition

is ensured, provided that the controller admits packets with rate $a_i^f[n]$ drawn from a distribution with mean $\bar{a}_i^{f*}$. In practice, the long-term average arrival rates are chosen to be slightly less than $\bar{a}_i^{f*}$ to conservatively effect stability.

Moreover, the routing variables are set to the optimal endogenous flows $\bar{r}_{i,j}^{f*}$, namely,

$$r_{i,j}^f[n] = \bar{r}_{i,j}^{f*}, \quad n = 1, 2, \dots \tag{77}$$

Note that in the formulation of Section 3, the endogenous flows for *every* fading state $\mathbf{h}$ were variables (i.e., $r_{i,j}^f$ as function of $\mathbf{h}$ was the optimization variable). Then routing at every time slot is straightforward; it is determined by $r_{i,j}^f(\mathbf{h}[n])$. Here, only the long-term average flows are variables, and the optimal ones are used for instantaneous routing decisions, as given by (77).

Power allocation $\mathbf{p}^\dagger(\mathbf{h}[n])$ will be used per time slot n. The instantaneous physical layer rate $c_{i,j}[n]$ at each time slot is

$$c_{i,j}[n] = \sum_k C_{i,j}^k\Big(\gamma_{i,j}^k\big(\mathbf{h}[n], \mathbf{p}^\dagger(\mathbf{h}[n])\big)\Big), \quad n = 1, 2, \dots \tag{78}$$

For the scheme described so far, the full buffer assumption is considered to hold. Furthermore, it is important to stress the role of the physical layer queues in the interaction between the routing decisions $r_{i,j}^f[n]$ and the instantaneous capacity supported by the physical layer $c_{i,j}[n]$. These buffers effectively store the bits that cannot be transmitted if the instantaneous capacity $c_{i,j}[n]$ drops below the rate that the layer above wants to pump out, that is, $\sum_{f\in\mathcal{F}(i)} r_{i,j}^f[n]$. Note that in the formulation of Section 3 there is no issue of the instantaneous capacity not supporting the network layer flows, because constraint (4c) is enforced for every fading realization. In the present formulation, the link capacity constraint is stated in a long-term average (ergodic) form [cf. (64c)], and is satisfied via the physical layer queues. In the formulation of Section 3, it is stated in an instantaneous form [cf. (4c)], and there is no need for physical layer queues.

To build intuition about the algorithm, note that the long-term average flow rates $\frac{1}{N}\sum_{n=1}^N a_i^f[n]$ and $\frac{1}{N}\sum_{n=1}^N r_{i,j}^f[n]$ converge to the optimal $\bar{a}_i^{f*}$ and $\bar{r}_{i,j}^{f*}$ by construction. Moreover, the endogenous rates $c_{j,i}^f[k]$ will satisfy $\sum_{n=1}^N c_{j,i}^f[n] \le \sum_{n=1}^N r_{j,i}^f[n]$, because all packets placed endogenously into the network layer queues of node i must have been routed to i from its neighbors. Hence, (64b) will be satisfied by the long-term averages of the respective processes; that is, $\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^N a_i^f[n] + \lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^N c_{j,i}^f[n] \le \lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^N r_{i,j}^f[n]$. A similar conclusion is true for (64c) with $\bar{c}_{i,j} = \bar{c}_{i,j}^*$. Also, if $\mathbf{p}^\dagger(\mathbf{h}[n])$ is a good approximation of $\mathbf{p}^*(\mathbf{h}[n])$, then the long-term average link capacities $\frac{1}{N}\sum_{n=1}^N c_{i,j}[n]$ and long-term average power consumptions converge, respectively, to their expected values $\mathbb{E}\big[\sum_k \ln\big(1+\gamma_{ij}^f(\mathbf{h}, \mathbf{p}^*(\mathbf{h}))\big)\big]$ and $\mathbb{E}\big[\sum_k \sum_{j\in\mathcal{N}(i)} p_{ij}^{f*}(\mathbf{h})\big]$, because the fading process is stationary and ergodic. Then, (64d) and (64e) are satisfied with $\bar{c}_{i,j} = \bar{c}_{i,j}^*$ and $\bar{p}_i = \bar{p}_i^*$.

8 Concluding Summary

This chapter outlined a framework for cross-layer resource allocation in wireless fading multi-hop networks with orthogonal and non-orthogonal access. In the orthogonal case,

the goal is to optimize average end-to-end rates, instantaneous network layer flows, link schedules, average power consumptions, and instantaneous power allocation across tones. Based on the KKT conditions, the optimal solution is derived in terms of the fading realization and the optimal Lagrange multipliers. Although most allocation variables can be found in closed form, there are cases where analytical expressions for optimal link schedules and routing variables are not available. A low-complexity scheme that is asymptotically optimal is designed to address this issue.

Different alternatives to obtain the Lagrange multipliers required for the channel-adaptive policies are presented. First, an algorithm is proposed based on smooth subgradients that is run offline during an initialization phase. The optimal Lagrange multipliers obtained can then be used for online network control; that is, to allocate the resources every time the instantaneous CSI is updated. Second, stochastic iterations that are run online and do not require knowledge of the channel distribution are proposed. In this case, a fully online algorithm is developed, whereby the resource allocation decisions are determined by the current CSI and the current Lagrange multipliers. Optimality and convergence of such algorithms is analyzed. Then, by drawing connections between the Lagrange multipliers and the queue lengths, it is established that when the online algorithm is used for network control, all queues in the network are guaranteed to be stable, in the sense that the sample averages of the queue lengths converge with probability 1 (to some finite number). The expected delay is explicitly given as function of the optimal Lagrange multipliers.

Finally, attention is turned to the non-orthogonal access case, whereby link capacities become functions of the SINR, and thus couple the power allocation decisions. The resulting cross-layer problem is in general non-convex, but has zero duality gap. Capitalizing on this, a subgradient descent algorithm along with weighted running averages of the primal iterates is developed. This scheme yields near-optimal primal and dual variables, which can then be used for online network control.

References

[1] M. Avriel, R. Dembo, and U. Passy, "Solution of generalized geometric programms," *Int. J. Numer. Methods Eng.*, vol. 9, no. 1, pp. 149–168, 1975.

[2] M. Avriel and A. C. Williams, "Complementary geometric programming," *SIAM J. Appl. Math.*, vol. 19, no. 1, pp. 125–141, July 1970.

[3] J.-A. Bazerque and G. B. Giannakis, "Distributed scheduling and resource allocation for cognitive OFDMA radios," *Mobile Networks and Applications*, vol. 13, no. 5, pp. 452–462, Oct. 2008.

[4] D. P. Bertsekas, A. Nedić, and A. Ozdaglar, *Convex Analysis and Optimization.* Belmont, MA: Athena Scientific, 2003.

[5] R. Bhatia and M. Kodialam, "On power efficient communication over multi-hop wireless networks: Joint routing, scheduling and power control," in *Proc. IEEE INFOCOM*, Hong Kong, China, Mar. 2004, pp. 1457–1466.

[6] H. Boche and E. Jorswieck, "Universal Approach for Performance Optimization in Multiuser MIMO Systems," in *European Trans. on Telecommunications*, vol. 18, no. 3, pp. 217–233, Apr. 2007.

[7] S. Boyd and A. Mutapcic, "Stochastic Subgradient Methods," in *Notes for EE364b, Stanford University*, Jan. 2007. [Online.] Available: `http://www.stanford.edu/class/ee364b/notes/stoch_subgrad_notes.pdf`

[8] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge University Press, 2004.

[9] L. Bui, A. Eryilmaz, R. Srikant, and X. Wu, "Asynchronous congestion control in multi-hop wireless networks with maximal-based scheduling," *IEEE/ACM Trans. Netw.*, vol. 16, no. 4, pp. 826–839, Aug. 2008.

[10] L. Chen, S. H. Low, M. Chiang, and J. C. Doyle, "Cross-layer congestion control, routing and scheduling design in ad hoc wireless networks," in *Proc. IEEE INFOCOM*, Barcelona, Spain, Apr. 2006.

[11] M. Chiang, "Balancing transport and physical layers in wireless multi-hop networks: Jointly optimal congestion control and power control," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 1, pp. 104–116, Jan. 2005.

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

[13] R. L. Cruz and A. V. Santhanam, "Optimal routing, link scheduling and power control in multi-hop wireless networks," in *Proc. IEEE INFOCOM*, San Francisco, CA, Mar.–Apr. 2003, pp. 702–711.

[14] C. Curescu and S. Nadjm-Tehrani, "A bidding algorithm for optimized utility-based resource allocation in ad hoc networks," *IEEE Trans. Mobile Comput.*, vol. 7, no. 12, pp. 1397–1414, Dec. 2008.

[15] A. Eryilmaz, A. Ozdaglar, D. Shah, and E. Modiano, "Distributed cross-layer algorithms for the optimal control of multi-hop wireless networks," *IEEE/ACM Trans. Netw.*, to be published. [Online.] Available: `http://www.mit.edu/~modiano/papers/J62.pdf`

[16] A. Eryilmaz and R. Srikant, "Joint congestion control, routing, and MAC for stability and fairness in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1514–1524, Aug. 2006.

[17] N. Gatsis, A. Ribeiro, and G. B. Giannakis, "A class of convergent algorithms for resource allocation in wireless fading networks," *IEEE Trans. Wireless Commun.*, vol. 9, no. 5, pp. 1808–1823, May 2010.

[18] L. Georgiadis, M. J. Neely, and L. Tassiulas, "Resource allocation and cross-layer control in wireless networks," *Foundations and Trends in Networking*, vol. 1, no. 1, pp. 1–144, 2006.

[19] A. Giannoulis, K. P. Tsoukatos, and L. Tassiulas, "Lightweight cross-layer control algorithms for fairness and energy efficiency in CDMA ad-hoc networks," in *Proc. 4th Int. Symp. Modeling and Optimization in Mobile, Ad hoc, and Wireless Networks*, Boston, MA, Apr. 2006.

[20] A. Goldsmith, *Wireless Communications*. New York, NY: Cambridge University Press, 2005.

[21] M. Johansson and L. Xiao, "Cross-layer optimization of wireless networks using nonlinear column generation," *IEEE Trans. Wireless Commun.*, vol. 5, no. 2, pp. 435–445, Feb. 2006.

[22] S. Hayashi and Z.-Q. Luo, "Spectrum management for interference-limited multiuser communication systems," *IEEE Trans. Inf. Theory*, vol. 55, no. 3, pp. 1153–1175, Mar. 2009.

[23] L. Kleinrock, *Queueing Systems, Vol. I: Theory*, New York: Willey, 1975.

[24] H. J. Kushner and G. G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, 2nd ed. New York, NY: Springer, 2003.

[25] A. Lapidoth and S. Shamai, "Fading channels: How perfect need 'perfect side information' be?," *IEEE Trans. Inf. Theory*, vol. 48, no. 5, pp. 1118–1134, May 2002.

[26] T. Larrson, M. Patriksson, and A.-B. Strömberg, "Ergodic, primal convergence in dual subgradient schemes for convex programming," *Math. Programming*, vol. 86, no. 2, pp. 283–312, 1999.

[27] J.-W. Lee, R. R. Mazumdar, and N. B. Shroff, "Opportunistic power scheduling for dynamic multi-server wireless systems," *IEEE Trans. Wireless Commun.*, vol. 5., no. 6, pp. 1506-1515, June 2004.

[28] H.-W. Lee, E. Modiano, and L. B. Le, "Distributed throughput maximization in wireless networks via random power allocation," in *Proc. 7th Int. Symp. Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks*, Seoul, Korea, June 2009.

[29] L. Li and A. J. Goldsmith, "Capacity and optimal resource allocation for fading broadcast channels—Part I: Ergodic capacity," *IEEE Trans. Inf. Theory*, vol. 47, no. 3, pp. 1083–1102, Mar. 2001.

[30] Y.-H. Lin and R. L. Cruz, "Opportunistic link scheduling, power control, and routing for multi-hop wireless networks over time-varying channels," in *Proc. 43rd Annu. Allerton Conf. Communication, Control, and Computing*, Monticello, IL, Sep. 2005, pp. 976–985.

[31] Y.-H. Lin, T. Javidi, R. L. Cruz, and L. B. Milstein, "Distributed link scheduling, power control and routing for multi-hop wireless MIMO networks," in *Proc. 40th Asilomar Conf. Signals, Systems and Computers*, Pacific Grove, CA, Oct.–Nov. 2006, pp. 122–126.

[32] L. Lin, X. Lin, and N. B. Shroff, "Low-complexity and distributed energy minimization in multi-hop wireless networks," *IEEE/ACM Trans. Netw.*, 2010, to be published.

[33] X. Lin and N. B. Shroff, "Joint rate control and scheduling in multi-hop wireless networks," in *Proc. 43rd IEEE Conf. Decision and Control*, Atlantis, Bahamas, Dec. 2004, pp. 1484–1489.

[34] X. Lin and N. B. Shroff, "The impact of imperfect scheduling on cross-layer congestion control in wireless networks," *IEEE/ACM Trans. Netw.*, vol. 14, no. 2, pp. 302–315, Apr. 2006.

[35] X. Lin, N. B. Shroff, and R. Srikant, "A tutorial on cross-layer optimization in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1452–1463, Aug. 2006.

[36] A. G. Marques, G. B. Giannakis, F. Digham, and F. J. Ramos, "Power-efficient wireless OFDMA using limited-rate feedback," *IEEE Trans. Wireless Commun.*, vol. 7, no. 2, pp. 685–696, Feb. 2008.

[37] A. G. Marques, G. B. Giannakis, and J. Ramos, "Optimizing orthogonal multiple access based on quantized channel state information," *IEEE Trans. Inf. Theory*, submitted for publication. [Online.] Available: `http://arxiv.org/abs/0909.0760`

[38] A. G. Marques, G. B. Giannakis, and J. Ramos, "Stochastic resource allocation for orthogonal access based on quantized CSI: Optimality, convergence and delay analysis," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Taipei, Taiwan, Apr. 2009, pp. 2357–2360.

[39] A. G. Marques, X. Wang, and G. B. Giannakis, "Dynamic resource management for cognitive radios using limited-rate feedback," *IEEE Trans. Signal Process.*, vol. 57, no. 9, pp. 3651–3666, Sep. 2009.

[40] A. Nedić and A. Ozdaglar, "Approximate Primal Solutions and Rate Analysis for Dual Subgradient Methods," *SIAM J. Optimization*, vol. 19, no. 4, pp. 1757–1780, 2009.

[41] M. J. Neely, "Energy optimal control for time-varying wireless networks," *IEEE Trans. Inf. Theory*, vol. 52, no. 7, pp. 2915–2934, July 2006.

[42] M. J. Neely, E. Modiano, and C.-P. Li, "Fairness and optimal stochastic control for heterogeneous networks," *IEEE/ACM Trans. Netw.*, vol. 16, no. 2, pp. 396–409, Apr. 2008.

[43] M. J. Neely, E. Modiano, and C. E. Rohrs, "Dynamic power allocation and routing for time-varying wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 1, pp. 89–103, Jan. 2005.

[44] M. J. Neely and R. Urgaonkar, "Optimal backpressure routing for wireless networks with multi-receiver diversity," *Ad Hoc Networks*, vol. 7, no. 5, pp. 862–881, July 2009.

[45] Y. Nesterov, "Smooth minimization of non-smooth functions," *Math. Programming*, vol. 103, no. 1, pp. 127-152, May 2005.

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

[47] B. Radunović and J.-Y. Le Boudec, "Power control is not required for wireless networks in the linear regime," in *Proc. 6th IEEE Int. Symp. World of Wireless Mobile and Multimedia Networks*, Taormina, Italy, June 2005, pp. 417–427.

[48] K. Rajawat, N. Gatsis, and G. B. Giannakis, "Cross-layers designs in coded wireless fading networks with multicast," *IEEE/ACM Trans. Netw.*, to be published. [Online]. Available: `http://arxiv.org/abs/1003.5239`

[49] A. Ribeiro, "Ergodic stochastic optimization algorithms for wireless communication and networking," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Dallas, TX, Mar. 2010, pp. 3326–3329.

[50] A. Ribeiro, "Stochastic soft backpressure algorithms for routing and scheduling in wireless ad-hoc networks," in *Proc. 3rd IEEE Int. Workshop Computational Advances in Multi-Sensor Adaptive Processing*, Aruba, Dutch Antilles, Dec. 2009, pp. 137–140.

[51] A. Ribeiro and G. B. Giannakis, "Optimal FDMA over wireless fading ad hoc networks," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Las Vegas, NV, Mar.–Apr. 2008, pp. 2765–2768.

[52] A. Ribeiro and G. B. Giannakis, "Separation principles in wireless networking," *IEEE Trans. Inf. Theory*, vol. 56, no. 9, pp. 4488–4505, Sep. 2010.

[53] S. Shakkottai and R. Srikant, "Network optimization and control," *Foundations and Trends in Networking*, vol. 2, no. 3, pp. 279–379, 2007.

[54] G. Sharma, R. R. Mazumdar, and N. B. Shroff, "On the complexity of scheduling in wireless networks," in *Proc. 12th Annu. Int. Conf. Mobile Computing and Networking*, Los Angeles, CA, Sep. 2006, pp. 227–238.

[55] P. Soldati, B. Johansson, and M. Johansson, "Proportionally fair allocation of end-to-end bandwidth in STDMA wireless networks," in *Proc. 7th ACM Int. Symp. Mobile Ad Hoc Networking and Computing*, Florence, Italy, May 2006, pp. 286–297.

[56] V. Solo and X. Kong, *Adaptive Signal Processing Algorithms: Stability and Performance*. Upper Saddle River, NJ: Prentice Hall, 1995.

[57] A. Stolyar, "Maximizing queueing network utility subject to stability: Greedy primal-dual algorithm," *Queueing Systems*, vol. 50, no. 4, pp. 401–457, Aug. 2005.

[58] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multi-hop radio networks," *IEEE Trans. Autom. Control*, vol. 37, no. 12, pp. 1936–1948, Dec. 1992.

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

[60] D. Tse and S. V. Hanly, "Multiaccess fading channels—Part I: Polymatroid structure, optimal resource allocation and throughput capacities," *IEEE Trans. Inf. Theory*, vol. 44, no.7, pp. 2796–2815, Nov. 1998.

[61] X. Wang, G. B. Giannakis, and A. G. Marques, "A unified approach to QoS-guaranteed scheduling for channel-adaptive wireless networks," *Proc. IEEE*, vol. 95, no. 12, pp. 2410–2431, Dec. 2007.

[62] C. Y. Wong, R. S. Cheng, K. B. Lataief, R. D. Murch, "Multiuser OFDM with Adaptive Subcarrier, Bit, and Power Allocation," *IEEE J. Sel. Areas Commun.*, vol. 17, no. 10, pp. 1747–1758, Oct. 1999.

[63] Y. Xi and E. M. Yeh, "Distributed algorithms for spectrum allocation, power control, routing, and congestion control in wireless networks" in *Proc. 8th ACM Int. Symp. Mobile Ad Hoc Networking and Computing*, Montréal, QC, Sep. 2007, pp. 180–189.

[64] Y. Xi and E. M. Yeh, "Node-based optimal power control, routing, and congestion control in wireless networks," *IEEE Trans. Inf. Theory*, vol. 54, no. 9, pp. 4081–4106, Sep. 2008.

[65] Y. Xi and E. M. Yeh, "Throughput optimal distributed power control of stochastic wireless networks," *IEEE/ACM Trans. Netw.*, 2010, to be published.

[66] L. Xiao, M. Johansson, and S. P. Boyd, "Simultaneous routing and resource allocation via dual decomposition," *IEEE Trans. Commun.*, vol. 52, no. 7, pp. 1136–1144, July 2004.

[67] Y. Xue, B. Li, and K. Nahrstedt, "Optimal resource allocation in wireless ad hoc networks: a price-based approach," *IEEE Trans. Mobile Comput.*, vol. 5, no. 4, pp. 347–364, Apr. 2006.

[68] Y. Yi and S. Shakkottai, "Hop-by-hop congestion control over a wireless multi-hop network," in *Proc. IEEE INFOCOM*, Hong Kong, China, Mar. 2004, pp. 2548–2558.

[69] S. A. Zenios, M. C. Pinar, and R. S. Dembo, "A smooth penalty function algorithm for network-structured problems", *European J. of Operational Research*, vol. 83, no. 1, pp. 220–236, May 1995.