Towards a formal theory of wireless networking

PI: Alejandro Ribeiro (email: aribeiro@seas.upenn.edu)

Dept. of Electrical and Systems Engineering
University of Pennsylvania
Philadelphia, PA 19096

December 13, 2009

Agency: National Science Foundation (NSF)
Program: Faculty Early Career Development (CAREER)
Solicitation: 08-557
Directorate: Computer and Information Science and Engineering (CISE)
Computing and Communication Foundations (CCF)
Communications and Information Foundations (CIF)
Deadline: July 21, 2009
A Project Summary

The last decades of the 20th century witnessed the emergence of networks that have become ever more pervasive and important. At the start of the 21st century developing a scientific understanding of networks remains an unresolved intellectual endeavor. This proposal is framed into this effort by trying to understand wireless communication networks on a fundamental level.

Research Objective. The proposer’s long term career goal is to develop a formal theory of wireless networks. The goal of this project is to lay the foundations for this longer term endeavor. The research agenda is to formulate generic optimization problems that model wireless networks in a variety of settings. Being a defining characteristic, fading runs through all the proposed research threads. Properties and solution methods of these optimization problems are explored and translated into architectural properties of wireless networks as well as algorithms to find optimal operating points. Problems are specialized to study different types of physical layers. Strategies to learn fading distributions are pursued. Problems associated with acquiring channel state information are addressed. Cutting across these research topics is the issue of developing mechanisms to translate algorithms into protocols, i.e., actual message exchanges letting nodes determine optimal network variables.

Intellectual merit. Progress in optimal design of wireless networks is hampered by the complexity of the associated optimization problems. They are highly dimensional and non convex, therefore falling on the unlucky side of the watershed division between easy and difficult. The novel technical approach is to use statistical signal processing tools to exploit structure introduced by fading. It is well known that when correctly taken into account fading can be beneficial for wireless communications. The related fact that resource allocation problems are simpler to solve in the presence of fading, however, is sometimes overlooked. Even though it might seem counterintuitive that seemingly more complicated resource allocation problems in the presence of fading are easier to solve that their deterministic – i.e., without fading – counterparts, science is full of examples where randomness yields a simpler problem. In fact, it has been argued that this is in fact the core epistemological value of the theory of probability.

In an interesting example of structure introduced by randomness, it has been recently proved that while it is widely believed that the conventional protocol stack results in poorly performing wireless networks, separating the design of wireless networks in layers is in fact optimal in the presence of fading. This result is a consequence of the fact that even though the associated optimization problem is generally non-convex the gap with its Lagrange dual problem is zero. The separation principles provide a motivation for the proposed research and exemplify the feasibility of the proposed technical approach.

Education plan. The proposed education plan revolves around the excitement, challenge and discipline gaps. The excitement gap is about the excitement people feels about science and technology versus the lack of interest to pursue careers in science and technology. The challenge gap refers to the ongoing trend to reduce the complexity of the material we teach in our classes. The discipline gap alludes to the compartmental experience offered to our students and the reality of an increasingly hazy separation between disciplines. The goal of the education plan is to contribute to the closing of these gaps through the development of new courses.

The proposer believes that scientific progress depends on the continuous interaction of teaching, research and development. Research results are taught to students that incorporate them in new technology whose development uncovers novel research challenges. The research agenda outlined in this proposal lends itself naturally to the creation of a class on optimal design of wireless networks that is being developed to realize this vision. A significant part of the current version of the class exposes students to open research problems and provides a formal framework to face their analysis. As the research agenda advances the class will morph into a more conventional graduate-level offering.

Broader impact. Success in the research effort proposed here has the potential to change teaching and design of wireless networks. As has happened with digital communications in the last fifty years, a shift from current heuristic based approaches to rigorous formulations is foreseen. The importance of this shift should not be underestimated. Existing wireless networks, for example, utilize less than one tenth of the spectrum allocated to their operation. Ultimately, the goal is to improve this meager use of resources.
# Table of Contents

## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Project Summary</td>
<td>A-1</td>
</tr>
<tr>
<td>B Table of Contents</td>
<td>B-1</td>
</tr>
<tr>
<td>C Project Description</td>
<td>C-1</td>
</tr>
<tr>
<td>C.1 Research and education plan preview</td>
<td>C-2</td>
</tr>
<tr>
<td>C.2 Related work</td>
<td>C-2</td>
</tr>
<tr>
<td>C.3 Technical approach</td>
<td>C-4</td>
</tr>
<tr>
<td>C.3.1 Link capacities and transmitted power</td>
<td>C-4</td>
</tr>
<tr>
<td>C.3.2 Optimal wireless networking</td>
<td>C-5</td>
</tr>
<tr>
<td>C.3.3 Lagrangian dual problem</td>
<td>C-6</td>
</tr>
<tr>
<td>C.4 Optimality of dual relaxation for ergodic networks</td>
<td>C-6</td>
</tr>
<tr>
<td>C.4.1 Separation principles revisited</td>
<td>C-7</td>
</tr>
<tr>
<td>C.5 Proposed research</td>
<td>C-8</td>
</tr>
<tr>
<td>C.5.1 Wireless networks architectures</td>
<td>C-8</td>
</tr>
<tr>
<td>C.5.2 Optimal physical layer design</td>
<td>C-9</td>
</tr>
<tr>
<td>C.5.3 Adaptive algorithms</td>
<td>C-11</td>
</tr>
<tr>
<td>C.5.4 Channel state information</td>
<td>C-13</td>
</tr>
<tr>
<td>C.5.5 Optimal wireless networking protocols</td>
<td>C-13</td>
</tr>
<tr>
<td>C.6 Education</td>
<td>C-14</td>
</tr>
<tr>
<td>C.6.1 Excitement, challenge and discipline</td>
<td>C-14</td>
</tr>
<tr>
<td>C.6.2 Optimal design of wireless networks</td>
<td>C-15</td>
</tr>
<tr>
<td>C.7 Perspective</td>
<td>C-15</td>
</tr>
</tbody>
</table>

## References

D-1
C  Project Description

My long term career goal is to develop a formal theory of wireless networks providing a scientific basis to understand their fundamental properties and guide their design. The goal of this CAREER proposal is to develop core results conducive towards that objective.

In their classical work on limit distributions Gnedenko and Kolmogorov wrote that the “epistemological value of the theory of probability is based on this: that large-scale random phenomena in their collective action create strict, nonrandom regularity” [11]. In simpler words, randomness generates structure. It is often possible to infer properties of large-scale stochastic systems even if analogous deterministic counterparts are intractable. Randomness, in the form of fading – random variations in propagation coefficients between network nodes –, is inherent in wireless networks. The technical approach is to explore structure introduced by fading to understand fundamental properties of wireless networks. An example of the desired outcomes are recent results concerning the optimality of separating wireless networking problems in layers and per-fading state subproblems [32].

The goal of a communication network is to administer given resources to support information flows with some required level of service. In conventional wired networks the resource given is a set of physical connections between nodes. Supporting information flows requires finding routes between source and destination, determining link sharing strategies, and controlling the amount of traffic injected into the network. It was an early design specification to separate these problems in layers – routing, link and transport for the problems in the previous sentence – that operate more or less independently, and interact through standardized interfaces. While this was mostly a matter of ensuring inter-operability it is remarkable that this separation can be optimal. Specifically, it is possible to define separate per-layer optimization problems whose outcome coincides with the solution of a joint non-layered optimization. Mathematically, separability comes from the fact that the wired networking problem is convex – a linear program in fact. The Lagrangian dual problem can thus be solved instead. As it often happens, the Lagrangian exhibits a separable structure, which, as it turns out coincides with the conventional layers.

In a wireless network, the given resources are not connections but bandwidth and power. Therefore, on top of routes, link shares and rate control, a wireless networking problem entails determining which connections, among those possible for the given bandwidth and power, should be established to support the required level of service. Earlier approaches to wireless networking migrated the conventional layers and defined the power and frequency assignment as physical layer subproblems. This yields poor results though, and over time lead to the surge of cross-layer design as synonym of joint optimization across layers. Ultimately, the poor performance of layered wireless networks stems from the non-convexity of the joint cross-layer networking problem. As in wired networks, the Lagrangian exhibits a separable structure that can be mapped to layers. But non-convex problems have positive duality gap explaining the poor results of layered wireless networks.

In what constitutes an interesting example of structure introduced by randomness, it has been proved that general wireless networking problems in the presence of fading, while non-convex, have zero Lagrangian duality gap [32]. Exploiting the separability of the Lagrangian, this result yields the following principles.

First separation principle of wireless networking. This principle pertains to the separability of wireless networking problems into layers. It states that it is possible to define separate optimization problems to obtain optimal routes, optimal link capacity allocations, and optimal power/frequency assignments.

Second separation principle. Another difficulty in optimal wireless networking is the need to optimize jointly for all fading states. Given that fading coefficients take on a continuum of values, this is a variational problem that requires finding optimal functions of the fading coefficients. This principle states that network optimization is further separable in per-fading-state subproblems. The practical importance of this result is that it is not necessary to find optimal functions but only the values of the functions for those channels actually observed.

The separation principles hold under specific assumptions, e.g., networks operating in an ergodic setting and availability of perfect channel state information (CSI) that restrict applicability of the separation principles to particular settings. Part of the proposed research is to study the extent to which these assumptions can be lifted and what implications follow when this is not possible. Nonetheless, it has to be recognized that they do establish a fundamental property of wireless networks in the presence of fading that is not true for networks in a deterministic setting. In the context of this proposal the separation principles are provided as motivation for part of the proposed research and to exemplify the feasibility of the proposed technical approach.
C.1 Research and education plan preview

The research agenda is to formulate generic optimization problems that model wireless networks in a variety of settings. Questions about such optimization problems are then posed and corresponding answers translated to fundamental properties of wireless communication networks. The analysis relies heavily on statistical signal processing tools to exploit structure generated by randomness. The proposed research comprises the following topics.

Architectural principles. The emphasis in this thrust is not on solving optimization problems associated with wireless networks but on studying their structural properties, e.g., lack of duality gap and computational complexity. Such structural properties of optimization problems will be translated into architectural principles for wireless networks, e.g., separation principles – see Section C.5.1.

Resource allocation at the physical layer. A consequence of the separation principles is that optimal wireless network design can be achieved if optimal resource allocation problems at the physical layer can be solved; see Section C.4 and Fig. 2. Algorithms to find exact and approximate solutions for different physical layer models will be pursued – see Section C.5.2.

Adaptive algorithms. Optimal networks depend on the probability distributions of fading. But these distributions vary from link to link and are not known beforehand. Adaptive algorithms and protocols to learn fading’s probability distributions while searching for optimal operating points will be developed. – see Section C.5.3.

Channel state information. Adaptation to varying fading coefficients requires channel estimation and percolation of channel state information (CSI) through the network. Two issues arise in this context, incorporation of the cost of sensing into optimality criteria and operation with imperfect and outdated CSI – see Section C.5.4. Note that imperfect CSI arises due to estimation errors and outdated CSI appears because of communication delays.

Protocols. Protocols are defined as mechanisms to exchange variables between neighboring terminals to determine optimal operating points. They are a particular type of algorithm in which only certain operations, those between nearby nodes, are allowed. They will be developed drawing from tools used in parallel and distributed optimization – see Section C.5.5.

This agenda is related to works on network utility maximization (NUM), e.g., [15, 21], and layer decompositions, e.g., [6, 7, 49] as detailed in upcoming Section C.2. Note though, that much of existing work assumes convexity of the associated optimization problems motivating the use of convex approximations and relaxations to deal with wireless networks. The proposed approach is different in that we will work with the original non-convex formulations and study the extent to which randomness simplifies problem analysis and solution.

I believe that engineering is an exciting field and I strive to communicate this feeling to an audience as broad as possible. I also believe that scientific progress depends on the continuous interaction of teaching, research and development. Research results are taught to students that incorporate them in new technology whose development uncovers novel research challenges. These two beliefs underly the following proposed education plan.

Excitement, challenge and discipline gaps. The excitement gap is about the excitement people feels about science and technology versus the lack of interest to pursue careers in science and technology. The challenge gap refers to the ongoing trend to reduce the complexity of the material we teach in our classes. The discipline gap alludes to the compartmental experience offered to our students and the reality of an increasingly hazy separation between engineering disciplines. My goal as an educator is to contribute to the closing of these gaps through the development of new courses – see Sections C.6.1.

Optimal design of wireless networks. The research agenda outlined above lends itself naturally to the development of a class on optimal design of wireless networks. A first version of this class has been created and a second iteration will be offered in Spring 2010 [43]. A significant part of the current version of the class exposes students to open research problems and provides a formal framework to face their analysis. As the research agenda advances the class will morph into a more conventional graduate-level offering – see Section C.6.2.

C.2 Related work

Optimization as a mathematical tool to analyze network protocols appeared first in the NUM framework independently proposed by [15] and [21]. The gist of these works is that congestion control protocols can be viewed
as distributed implementations of algorithms that solve utility maximization problems. Source rates are regarded as primal variables and congestion parameters constitute variables of the corresponding Lagrange dual problems. Recursive schemes updating these variables boil down to subgradient descent iterations on the dual function – the kind of optimization algorithm also known as dual decomposition. The connection between congestion control and NUM has been fruitful in understanding which NUM problems are solvable by heuristic congestion control schemes and also for introducing protocols as solvers of suitably formulated NUM problems; see e.g., [22, 20] and [38] for recent accounts on NUM.

Extending the NUM framework to wireless networking problems is not a simple pursuit. Different from wireline networks where pairs of nodes are individually connected at fixed capacities, node links in wireless networks are not predetermined and their capacities are not fixed. Rather, connections and link capacities are variables of the optimization problem itself. Nonetheless, wireless physical layer models have been incorporated into NUM formulations, and arguably represent one of the most promising research directions on cross-layer network design; see e.g., [5, 9, 19, 47, 50] and references therein.

Although some recent works advocate alternative decomposition methods [14, 28, 29], most of the NUM literature relies on dual decomposition. This is because the associated Lagrangian function exhibits a separable structure reminiscent of layered network designs. This has been pointed out first in e.g., [6, 49] and recently popularized by [7]. Indeed, it is possible using the NUM formulation to understand specific layered architectures as decompositions of optimization problems [7]. On this issue, it is important to stress that since wireless networking problems are non-convex, the duality gap is generally nonzero. As a result, the dual optimum is generally different from the primal optimum and for this reason layering is believed to come at the price of optimality loss.

A different approach to the wireless networking problem is the work on stochastic network optimization, originally reported by [27] and subsequently extended in [25, 26]; see also [10] for a comprehensive treatment and [18] for a related approach. These works build on the back-pressure algorithm [40, 41, 42], and extend it to deal with a finite number of random network states as those generated by fading. While the original back-pressure algorithm was introduced to stabilize all queues in the network without further optimality considerations, [25, 26, 27] developed modified versions that can provably approximate solutions of various wireless networking problems. These modified versions are obtained after observing a similarity between Lagrange multiplier updates in dual decomposition and evolution of (virtual) queue lengths in a communication network.

All of the approaches outlined so far pertain to either wireless networks with deterministic links, or, if random fading effects are accounted for, the channel links are confined to take on a finite number of values. This is simply not the case in wireless fading propagation and, given subtleties involved in limits of stochastic processes, cannot be regarded as practically irrelevant. Further motivation to consider fading channels with infinite number of states is provided by the fact that when this has been incorporated in related settings it has actually turned out to yield a simpler problem.

Rate utility maximization of time or frequency division multiple access (TDMA/FDMA) with fading coefficients taking on a continuum of values was considered in [46]. To deal with this non-convex problem, time sharing of carriers is introduced, a strategy that renders optimization tractable but complicates implementation. However, it is observed that with probability 1 time sharing is not needed. This result does not hold when the number of fading states is finite [45]. In this sense, TDMA and FDMA problems become simpler when fading takes on an infinite number of states. Another related problem is that of optimal subcarrier allocation at the physical layer of digital subscriber lines (DSL). Here too the problem is nonconvex, but when formulated in a continuous (as opposed to discrete) frequency-domain the duality gap becomes zero. This was first observed for sum-rate maximization in [51] and extended to general utilities by [23] as an application of Lyapunov’s convexity theorem [24].

Different from NUM, e.g., [5, 6, 9, 19, 47, 49, 50] and stochastic network optimization [10, 25, 26, 27] the goal here is not to propose and analyze a specific algorithm but to look at structural properties of the optimization problems. Different from [7] the research proposed here does not argue that layered architectures can be formally understood as decompositions of optimization problems but tries to exploit the fact that layering is optimal in wireless networks. Instead of a particular case of medium access control problem as in [45, 46] or a DSL problem as in [23, 51], general wireless networking problems are considered.
C.3 Technical approach

Consider an ad-hoc wireless network comprising \( J \) user terminals \( \{T_i\}_{i=1}^J \). Terminal \( T_i \) wishes to deliver packets for different application-level flows generically denoted by \( k \). The flow \( k \) specifies the destination of the flow’s packets, but the same destination may be associated with different flows to, e.g., accommodate different types of traffic (video, voice or data). The destination of flow \( k \) is denoted as \( T^k \) to emphasize that flow indexing is different from terminal indexing. For unicast traffic, each flow is associated with only one source \( T_i(k) \). For multiple access traffic each flow is associated with a set of sources. For every flow \( k \), packet arrivals at \( T_i \) form a stationary stochastic process with mean \( a_i^k \).

Network connectivity is modeled as a graph \( G(v, e) \) with vertices \( v \in V := \{1, \ldots, J\} \) and edges \( e \in E \) connecting pairs of vertices \( (i, j) \) when and only when \( T_i \) and \( T_j \) can communicate with each other; see Fig. 1. The neighborhood of \( i \) is denoted by \( n(i) := \{ j : (i, j) \in E \} \). Each terminal \( \{T_j\}_{j \in n(i)} \) that can communicate with \( T_i \) will be referred to as a neighbor. Given this model, terminals rely on multi-hop transmissions to deliver packets to the intended destination \( T^k \) of the flow \( k \). For that matter, \( T_i \) selects an average rate \( r_{ij}^k \) for transmitting \( k \)-th flow packets to \( T_j \). Assuming that packets are not discarded and queues are stable throughout the network, average rates \( a_i^k \) of exogenous packet arrivals (from the application layer) are related with endogenous (to the network layer) average rates \( r_{ij}^k \) transmitted to and from neighboring nodes. Endogenous and exogenous rates are related through the flow conservation equation per flow \( k \) as (see e.g., [10])

\[
a_i^k = \sum_{j \in n(i)} \left( r_{ij}^k - r_{ji}^k \right). \tag{1}
\]

Consider now the average rates \( r_{ij}^k \) of all flows \( k \) traversing the link \( T_i \rightarrow T_j \). Letting \( c_{ij} \) denote the information capacity of this link, queue stability is ensured by requiring [10]

\[
\sum_k r_{ij}^k \leq c_{ij}. \tag{2}
\]

The constraints in (1) and (2) are basic in describing traffic flow over a wireline network with fixed capacities \( c_{ij} \). In this setting, \( T_i \) needs to determine exogenous arrival rates \( a_i^k \) and transmission rate variables \( r_{ij}^k \) to satisfy certain optimality criteria. In a wireless network however, \( c_{ij} \) is not a fixed resource given to the terminals. In fact, operating conditions are determined by a set of available frequencies (tones) \( \mathcal{F} \) and prescribed powers \( p_{\text{max}} \). Thus, in addition to \( a_i^k \) and \( r_{ij}^k \), terminal \( T_i \) has to decide how to split its power budget \( p_{\text{max}} \) among tones \( f \in \mathcal{F} \) and neighbors \( T_j, j \in n(i) \). Matters are further complicated by fading propagation effects as described in the next section.

C.3.1. Link capacities and transmitted power. For every frequency \( f \in \mathcal{F} \) and \( (i, j) \in E \) denote the channel gain from \( T_i \) to \( T_j \) as \( h_{ij}^f \). As is customary practice in wireless communications \( h_{ij}^f \) is modeled as a random variable. The channel gains of all network links are collected in the vector \( \mathbf{h} \) and all realizations of \( \mathbf{h} \) in the set \( \mathcal{H} \).

Let \( p_{ij}^f(\mathbf{h}) \) denote the power used by \( T_i \) for sending packets to \( T_j \) on the tone \( f \) when the channel vector realization is \( \mathbf{h} \). Consequently, the instantaneous total power \( p_i(\mathbf{h}) \) used by \( T_i \) is the sum of the power used to transmit to all selected neighbors on all selected tones, i.e.,

\[
p_i(\mathbf{h}) := \sum_{j \in n(i)} \sum_{f \in \mathcal{F}} p_{ij}^f(\mathbf{h}). \tag{3}
\]

In an ergodic network the quantity of interest is the amount of power consumed over a period of time. This is obtained by averaging \( p_i(\mathbf{h}) \) in (3) over all possible channel realizations yielding the average power used by \( T_i \)

\[
p_i := \mathbb{E}_\mathbf{h}[p_i(\mathbf{h})] = \mathbb{E}_\mathbf{h} \left[ \sum_{j \in n(i)} \sum_{f \in \mathcal{F}} p_{ij}^f(\mathbf{h}) \right]. \tag{4}
\]
where \( \mathbb{E}_h[\cdot] \) denotes expectation over the channel probability distribution function.

The rate of information transmission over the \( T_i \rightarrow T_j \) link is a function of the power distribution \( p^f_{ij}(h) \) and the channel realization \( h \). To maintain generality of the model, define a function \( C_{ij}(h^f, p^f(h)) \) to map channels and powers to link capacities so that the capacity \( c^f_{ij}(h) \) of the link \( T_i \rightarrow T_j \) on the tone \( f \) is

\[
c^f_{ij}(h) := C_{ij}(h^f, p^f(h)).
\]

Function \( C(\cdot) \) is determined by the capabilities and operating conditions of the terminals. Some properties of the network, e.g., the separation principles, will be true for any physical layer, while others will depend on the chosen mode of operation. The proposed research will consider both, physical layer dependent and physical layer independent properties – see Section C.5.

In any event, the ergodic capacity \( c_{ij} \) of the wireless link \( T_i \rightarrow T_j \) is obtained after averaging over all possible channel realizations to obtain

\[
c_{ij} := \mathbb{E}_h \left[ \sum_{f \in \mathcal{F}} c^f_{ij}(h) \right] = \mathbb{E}_h \left[ \sum_{f \in \mathcal{F}} C_{ij}(h^f, p^f(h)) \right].\tag{6}
\]

The average power and link capacity expressions in (4) and (6) along with the flow and rate constraints in (1) and (2) describe information flow in a generic wireless network. They can be used to define the wireless network optimization problem described in the next section.

**C.3.2. Optimal wireless networking.** Average power \( p_i \) and link capacities \( c_{ij} \) depend on the chosen power profiles \( p^f_{ij}(h) \) as per (4) and (6). The average link rates \( r^k_{ij} \) are then constrained by (2) and the end-to-end flow rates \( a^k_i \) by (1). Problem variables \( p^f_{ij}(h), c_{ij}, p_i, r^k_{ij} \) and \( a^k_i \) that satisfy these equations can be supported by the network. Network designers, want to select out of these set of feasible variables those that are optimal in some sense. To this end, consider concave \( U^k_i(a^k_i) \) and convex \( V_i(p_i) \) utility functions, respectively, representing the reward of rate \( a^k_i \) and the cost of power \( p_i \). Though not required, it is intuitively reasonable to consider \( U^k_i(a^k_i) \) and \( V_i(p_i) \) increasing functions of their arguments. Based on these utility functions, it becomes possible to define the optimal networking problem as [cf. (1), (2), (4) and (6)]

\[
P = \max \sum_i \sum_k U^k_i(a^k_i) - \sum_i V_i(p_i) \tag{7}
\]

\[
c_{ij} \leq \mathbb{E}_h \left[ \sum_{f \in \mathcal{F}} C_{ij}(h^f, p^f(h)) \right], \quad p_i \geq \mathbb{E}_h \left[ \sum_{j \in u(i)} \sum_{f \in \mathcal{F}} p^f_{ij}(h) \right] \tag{8}
\]

\[
a^k_i \leq \sum_{j \in u(i)} \left( r^k_{ij} - r^k_{ji} \right), \quad \sum_k r^k_{ij} \leq c_{ij} \tag{9}
\]

where the constraints (1), (4) and (6) have been relaxed – something that can be done without loss of optimality. All problem variables have to be non-negative, but this is left implicit in (7). Also left implicit are power constraints \( p_i \leq p_{\text{max}} \) and \( p^f_{ij}(h) \leq p_{\text{max}} \), arrival rate requirements \( a^k_{\text{min},i} \leq a^k_i \leq a^k_{\text{max},i} \) as well as upper bound constraints \( c_{ij} \leq c_{\text{max}} \) and \( r^k_{ij} \leq r_{\text{max}} \) on link capacities and link flow rates. For future reference, define the vector-valued power function \( \mathbf{p}(h) \) with entries \( p^f_{ij}(h) \) and \( \mathbf{X} \) the set of primal variables \( c_{ij}, p_i, r^k_{ij} \) and \( a^k_i \) for all possible subindices, i.e., all \( i \) and all \( j \in u(i) \) for \( c_{ij} \), all \( i \) for \( p_i \) and so on. The aforementioned implicit constraints specify a set of feasible variables

\[
\mathcal{B} := \{ (\mathbf{X}, \mathbf{p}(h)) : 0 \leq p^f_{ij}(h) \leq p_{\text{max}}, 0 \leq p_i \leq p_{\text{max},i}, a^k_{\text{min},i} \leq a^k_i \leq a^k_{\text{max},i}, 0 \leq c_{ij} \leq c_{\text{max}}, 0 \leq r^k_{ij} \leq r_{\text{max}} \} \tag{10}
\]

These constraints are henceforth referred to as box constraints. They will be kept implicit for the most part but when required by clarity they will be stated explicitly.

C-5
As function $C(\cdot)$ is not concave in general, problem (7) is a difficult optimization problem. In fact, if channels are deterministic (i.e., per realization $h$), it has been proved that problem (7) is NP-hard [12]. This difficulty notwithstanding, there are fundamental principles of wireless networking problems to be derived from properties of (7). Revealing these principles can be facilitated by looking at the Lagrange dual problem.

**C.3.3. Lagrangian dual problem.** To define the dual problem, associate Lagrange multipliers $\lambda_{ij}$ and $\mu_i$ with the capacity and power constraints in (8) and $\nu_k^i$ and $\xi_{ij}$ with the flow and rate constraints in (9). For notational brevity, call $\Lambda$ the set of all dual variables, i.e., $\lambda_{ij}, \mu_i, \nu_k^i, \xi_{ij}$, and write the Lagrangian as

$$
\mathcal{L}[X, p(h), \Lambda] = \sum_{i,k} U_i^k(a_i^k) - \sum_i V_i(p_i)
$$

$$
+ \sum_{i,j} \lambda_{ij} \left[ \mathbb{E}_h \left[ \sum_{f \in F} C_{ij}(h^f, p^f(h)) \right] - c_{ij} \right] + \sum_i \mu_i \left[ p_i - \mathbb{E}_h \left[ \sum_{j \in n(i)} \sum_{f \in F} p_{ij}^f(h) \right] \right]
$$

$$
+ \sum_{i,k} \nu_k^i \left[ \sum_{j \in n(i)} (r_{ij}^k - r_{ji}^k) - a_i^k \right] + \sum_{i,j} \xi_{ij} \left[ c_{ij} - \sum_k r_{ij}^k \right].
$$

(11)

The dual function $g[\Lambda]$ is then obtained by maximizing the Lagrangian over the primal variables satisfying the box constraints in (10) and the dual problem is defined as the minimization of the dual function; i.e.,

$$
D = \min_{\Lambda \geq 0} g[\Lambda] = \min_{\Lambda \geq 0} \max_{(X, p(h)) \in B} \mathcal{L}[X, p(h), \Lambda].
$$

(12)

Since (7) is non-convex the duality gap is, in principle, nonzero which implies that $D > P$. Solving (12) is thus a relaxation in the sense that it yields an upper bound $D$ on the maximum achievable utility $P$. Consequently, the usefulness of (12) depends on the proximity of $D$ to $P$. For ergodic networks it turns out that $D = P$ implying the optimality of dual relaxations and leading to some interesting conclusions as discussed in Section C.4.

From a mathematical standpoint, the optimization problem in (7)-(9) and its dual form in (12) are the starting points for the research proposed here. Properties and solution methods will be explored. Particular emphasis will be put in architectural principles (Section C.5.1). Problems will be specialized to study different types of physical layers (Section C.5.2). Strategies to learn channel distributions will be pursued (Section C.5.3) and problems associated with acquiring the CSI vector $h$ will be addressed (Section C.5.4). Cutting across these research topics is the issue of developing mechanisms to translate algorithms into protocols, i.e., actual message exchanges letting nodes determine optimal network variables (Section C.5.5). Before describing these research topics we discuss the duality gap in wireless networks and revisit the separation principles.

**C.4 Optimality of dual relaxation for ergodic networks**

The challenges in solving (7) are worth recalling. For deterministic channels, the problem is known to be NP-hard. The Lagrange dual problem in (12) is certainly useful in establishing upper bounds on the achievable utility, but may or may not be close to the actual utility $P$. One expects that introducing fading will complicate matters further. Remarkably, it has been recently established that in the presence of fading the duality gap vanishes, i.e., $P = D$ [32]. This result is stated in the following theorem.

**Theorem 1** Let $P$ denote the optimum value of the primal problem (7) and $D$ that of its dual in (12). If the channel cumulative distribution function (cdf) is continuous, then

$$
P = D.
$$

(13)

To appreciate this result recall that the link capacity function $C(\cdot)$ is not necessarily concave in Theorem 1; hence, the optimization problem is generally nonconvex. The duality gap, however, is null. Continuity of the channel cdf ensures that no channel realization has strictly positive probability. This is satisfied by practical fading channel models including those adhering to Rayleigh, Rice and Nakagami distributions.
Given that the duality gap of the generic optimal wireless networking problem (7) is zero, the dual problem (12) can be solved instead. Because the dual function $g(\Lambda)$ in (12) is convex, descent methods can, in principle, be used to find the optimal multipliers $\Lambda^*$. It is also worth remarking that the primal problem (7) is a variational problem requiring determination of the function $p(h)$ that maps channel realizations to transmission powers. In that sense, it is an infinite-dimensional optimization problem. The dual problem however, involves a finite number of variables.

An important caveat is that zero duality gap does not necessarily mean it is easy to find the minimum of $g(\Lambda)$ in (12). Evaluating $g(\Lambda)$ as per (12) requires maximizing the Lagrangian $\mathcal{L}(x, p(h), \Lambda)$ in (11). This maximization may be difficult to perform depending on the link capacity function $C_{ij}(h^f, p^f(h))$. An important part of the research effort will be dedicated to finding solutions to these problems – see Sections C.5.2 and C.5.5. Nevertheless, Theorem 1 is determinant in establishing the separation principles that follow as corollaries of Theorem 1 and are revisited next.

C.4.1. Separation principles revisited. An implication of Theorem 1 is the optimality of conventional layering in wireless networking problems. As is usually the case, the Lagrangian exhibits a separable structure in the sense that it can be written as a sum of terms that depend on a few primal variables. Rearranging terms in (11) and assuming that the optimal dual argument $\Lambda^*$ is available, yields

$$
\mathcal{L}[X, p(h), \Lambda^*] = \sum_{i,j,k} \left( U^k_i(a^k_i) - \nu^k_j a^k_j \right) + \sum_i \left( \mu^*_i p_i - V_i(p_i) \right) + \sum_{i,j} \left( \xi^*_ij - \lambda^*_ij \right) c_{ij} \\
\quad + \sum_{i,j,k} \left( \nu^k_i - \nu^k_j - \xi^*_ij \right) r^k_{ij} + \mathbb{E}_h \left[ \sum_{i,j,f} \lambda^*_ij \left( h^f, p^f(h) \right) - \mu^*_ip^f_{ij}(h) \right].
$$

The zero duality gap implies that if $\Lambda^*$ is known we can, instead of solving (7), solve the (separable) problem

$$
P = D = g[\Lambda^*] = \max_{X, p(h)} \mathcal{L}[X, p(h), \Lambda^*]
$$

where the maximization is constrained to the $X$ and $p(h)$ that satisfy the box constraints.

Because primal variables are decoupled in the Lagrangian $\mathcal{L}[X, p(h), \Lambda^*]$ [cf. (14)], the maximization required in (15) can be split into smaller maximization problems involving less variables.

Let $\lambda^*_ij$, $\mu^*_i$, $\nu^k_j$, and $\xi^*_ij$ denote the optimal dual variables that solve (12). Consider the sub-problems

$$
P(a^k_i) = \max_{a^k_{i_{min}} \leq a^k_i \leq a^k_{i_{max}}} \left[ U^k_i(a^k_i) - \nu^k_j a^k_j \right]
$$

$$
P(r^k_{ij}) = \max_{0 \leq r^k_{ij} \leq r_{max}} \left[ \left( \nu^k_i - \nu^k_j - \xi^*_ij \right) r^k_{ij} \right]
$$

$$
P(c_{ij}) = \max_{0 \leq c_{ij} \leq c_{max}} \left[ \left( \xi^*_ij - \lambda^*_ij \right) c_{ij} \right]
$$

$$
P(p_i) = \max_{0 \leq p_i \leq p_{max}} \left[ \mu^*_i p_i - V_i(p_i) \right].
$$

<table>
<thead>
<tr>
<th>$a^k_i$ for all $k$</th>
<th>$r^k_{ij}$ for all $k$</th>
<th>$c_{ij}$ for all $k$</th>
<th>$p_i$ for all $k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\max_{a^k_{i_{min}} \leq a^k_i \leq a^k_{i_{max}}} \left[ U^k_i(a^k_i) - \nu^k_j a^k_j \right]$</td>
<td>$\max_{0 \leq r^k_{ij} \leq r_{max}} \left[ \left( \nu^k_i - \nu^k_j - \xi^*<em>ij \right) r^k</em>{ij} \right]$</td>
<td>$\max_{0 \leq c_{ij} \leq c_{max}} \left[ \left( \xi^<em>_ij - \lambda^</em><em>ij \right) c</em>{ij} \right]$</td>
<td>$\max_{0 \leq p_i \leq p_{max}} \left[ \mu^*_i p_i - V_i(p_i) \right]$</td>
</tr>
</tbody>
</table>

Figure 2: Having zero duality gap wireless networking problems can be separated in layers without loss of optimality. Thus, we can consider separate optimization problems to determine arrival rates $a^*_i$ [cf. (16)], link rates $r^*_ij$ [cf. (17)] link capacities $c^*_ij$ [cf. (18)] and average transmitted power $p^*_i$ [cf. (19)]. The physical layer problem can be further separated in per-fading-state subproblems [cf. (22)] but in general cannot be separated in per-terminal problems. We may say that the challenge in wireless networking is not as much in cross-layer optimization as in cross-terminal optimization of the physical layer.
Define further the optimal power allocation problem

\[ P[p(h)] = \max_{0 \leq p_{ij}(h) \leq p_{\text{max}}} \mathbb{E}_h \left[ \sum_{i,j,f} \lambda_{ij}^* C_{ij} \left( h^f, p^f(h) \right) - \mu_i^* p_{ij}^f(h) \right] \]  

(20)

Then, the optimal utility yield \( P \) in (7) is given by

\[ P = \sum_{i,k} P(a_k^i) + \sum_{i,j,k} P(r_{ij}^k) + \sum_{i,j} P(c_{ij}) + \sum_i P(p_i) + P[p(h)] \]  

(21)

i.e., the primal problem (7) can be separated into the (sub-) problems (16)-(20) without loss of optimality. The rate problem in (16) dictates the amount of traffic allowed into the network. It therefore solves the flow control problem at the transport layer; see Fig. 2. Likewise, (17) represents the network layer routing problem, (18) determines link-level capacities at the data link layer and (19) is the (link layer) average power control problem. Eq. (20) represents resource allocation at the physical layer. Therefore, it is a consequence of Theorem 1 that layering, in the sense of problem separability as per (21) is optimal in faded wireless networks.

In addition to layer separability, it is not difficult to realize that expectation and maximization can be interchanged in the power allocation subproblem (20). This establishes separability across fading states. Indeed, let \( \lambda_{ij}^* \), denote the optimal dual argument of (12) and define the per-fading-state power allocation problems

\[ P(h) = \max_{0 \leq p_{ij}(h) \leq p_{\text{max}}} \left[ \sum_{i,j,f} \lambda_{ij}^* C_{ij} \left( h^f, p^f(h) \right) - \mu_i^* p_{ij}^f(h) \right] \]  

(22)

Then, the optimal power allocation utility \( P[p(h)] \) in (20) is given by

\[ P[p(h)] = \mathbb{E}_h [P(h)] \]  

(23)

With \( \Lambda^* \) known, it has just been established that the variational problem of finding optimal power distributions in (20) reduces to finding optimal power values for given fading realizations as in (23).

Layer and per-fading state separability assume availability of the optimal Lagrange multipliers \( \lambda_{ij}^* \), \( \mu_i^* \), \( \nu_k^* \), and \( \xi_{ij}^* \). Finding them, while possible, is a non-trivial problem that forms part of the research proposed here (Section C.5.1). However, it has to be appreciated that they establish two fundamental properties of wireless networks in the presence of fading: i) the decomposition of the problem into the traditional networking layers can be optimal; and ii) the separability of the resource allocation problem into per-fading-state subproblems is possible. None of these properties holds true in wireless networks with deterministic channels. The separation principles exemplify the feasibility of the proposed technical approach and provide motivation for the research proposed next.

## C.5 Proposed research

The research agenda proposed here aims at building a theory of wireless networks leveraging the separation principles and related structural properties introduced by fading. Success in this effort would change how we teach and design these networks. As has happened with digital communications in the last fifty years, the aim is to shift from (i) descriptive accounts of network behavior to scientific explanatory understanding of wireless networks; and (ii) heuristics based design to rigorous formulations. Specific lines of research are detailed next.

### C.5.1. Wireless networks architectures

The optimality of dual relaxation for ergodic networks discussed in Section C.4 leads to the separation principles of Section C.4.1. Among other things, these separation principles establish that organizing ergodic wireless networks in layers is optimal. Such architectural principles are important beyond the subject of optimal design yielding light on, e.g., the interaction of routing and admission control protocols. An important proposed research thrust is the investigation of wireless network architectures in different settings.
The layered architecture summarized in Fig. 2 assumes availability of optimal lagrange multipliers $\lambda_{ij}^*, \mu_i^*, \nu_j^*$, and $\xi_{ij}^*$. It is possible to envision scenarios where these dual variables are computed offline but most likely they need to be dynamically updated by network nodes. Algorithms to solve this problem are well known, the most popular being subgradient descent on the dual function, sometimes also called dual decomposition [37]. Preliminary research suggests that subgradient descent algorithms lead to a decomposition in layers and layer interfaces. Layers maintain primal network variables and interact with adjacent layer interfaces only. Likewise, layer interfaces maintain dual variables and interact only with adjacent layers. Primal and dual variables are updated over time, with dual variables eventually converging to the optimal operating point.

However, existing results in subgradient descent algorithms cannot be applied verbatim to the problems considered here. Convergence results, e.g., are applicable to finite-dimensional problems whereas the variational problem in (7)-(9) is infinite-dimensional. The second challenge is the non-convexity of (7)-(9). Because the dual problem in (12) is always convex, lack of convexity of the primal problem is not an issue in the convergence of the dual variables. But it is an important consideration in determining whether the primal variables converge to the optimal operating point. It is worth recalling that dual variables are of interest only inasmuch as they help determining the optimal primal variables. Convergence properties of subgradient descent algorithms to solve (7)-(9) will be investigated. Particular emphasis will be placed on primal variables’ convergence. Notice that to prove convergence of primal variables it is necessary to establish feasibility, i.e., that primal iterates satisfy the constraints (8)-(9), and optimality, i.e., that their utility is close to optimal.

Non-ergodic networks. Theorem 1 and ensuing separation principles hold for ergodic networks. As follows from the presentation in Section C.3 this is so that we can: (i) completely characterize queues in term of average arrival and departure rates; and (ii) replace the time average that determines link capacities by an expected value over channel realizations. Lifting either assumption generates interesting research problems. Assumption (i) is not valid for time sensitive traffic, because delay depends on higher order statistics of departure and arrival processes. End-to-end delay characterizations are well known to be difficult but bounds can be obtained considering delays experienced in individual queues. Do similar separation principles hold for these non-ergodic networks? The answer is probably not, because delay characterizations are not linear and therefore, Lagrangians are probably not separable. Architectural results in this setting will be pursued.

For assumption (ii) to hold in practice the time scale of fading variations has to be much smaller than the time scale of network operation. This is true in most cases, particularly at the high frequencies used in modern wireless networks, but not always. When this is not true, the model at the physical layer changes. Instead of expressing link capacities through expected values, outage probabilities $P_{\text{out}}$ are prescribed and link capacities are defined as the transmission rates that guarantee successful packet delivery with probability $1 - P_{\text{out}}$. Do separation principles hold in this setting? The answer in this case is probably yes because outage probabilities can be written as expected values of indicator functions. Thus, optimization problems to model non-ergodic operation are likely to be have a structure similar to that of (7)-(9). This question will be investigated.

C.5.2. Optimal physical layer design. Of the per-layer problems (16)-(20), physical layer optimization as per (20) has the largest computational cost indicating that the challenge in wireless networks is not as much in cross-layer optimization as in cross-terminal optimization of the physical layer. While some degree of simplification is offered by the second separation principle in (22), joint optimization across terminals is still needed. In that sense the separation principles focus wireless networking problems on the physical layer. We will develop algorithms to solve (20) for interference limited physical layers and orthogonal multiplexing with spatial reuse. These two physical layers are important because they are commonly used in practice. Emergent physical layer technologies, e.g., cooperative communications [17, 30, 35], will also be incorporated into the framework.

Orthogonal multiplexing with spatial reuse. In this setting only terminals sufficiently far apart can share a given tone $f$ when the fading realization is $h$. To model this mode of operation introduce indicator functions $\alpha_{ij}^f(h) \in \{0, 1\}$ to reflect use of tone $f$ to communicate packets from $T_i$ to $T_j$ when the fading vector is $h$. Indicator functions $\alpha_{ij}^f(h) \in \{0, 1\}$ cannot be selected independently. If, e.g., $\alpha_{ij}^f(h) = 1$ for some $j \in n(i)$ then we need at least $\alpha_{ik}^f(h) = 0$ for all other $k \in n(i)$ and $\alpha_{lj}^f(h) = 0$ for all other $l \in n(j)$. Depending on the specific interference model some other exclusion constraints are incorporated. In any event, defining a vector $\alpha^f(h)$ with all indicator functions for given tone and fading channel, it is possible to write any exclusion constraints as $a^T \alpha^f(h) \leq 1$ for a
certain vector \( \mathbf{a} \) with elements in \{0, 1\} – recall that the elements of \( \alpha^f \) are either 0 or 1 too. An arbitrary number of exclusion constraints can be incorporated defining a matrix \( \mathbf{A} \) and requiring \( \mathbf{A} \alpha^f(h) \leq 1 \).

With orthogonal access ensured by the constraint \( \mathbf{A} \alpha^f(h) \leq 1 \), link capacity is determined by the signal to noise ratio (SNR) \( h_{ij}^f p_{ij}^f(h)/N_j^f \), where \( N_j^f \) denotes noise power. Encoding packet transmissions with, e.g., a capacity achieving code the link capacity per tone and fading realization can be written as

\[
C_{ij} \left(h^f, p^f(h)\right) = \alpha_{ij}^f(h) \log \left[1 + h_{ij}^f p_{ij}^f(h)/N_j^f \right].
\]  

(24)

A more practical alternative entails the use of a finite number of adaptive modulation and coding (AMC) modes. In this case, \( C(\cdot) \) is a staircase function defined by the rate of the AMC modes considered – see Fig. 3. Using mode \( m \) when \( p_m \leq h_{ij}^f p_{ij}^f(h)/N_j^f < p_{m+1} \) to achieve a rate \( c_m \), the function \( C_{ij} \left(h^f, p^f(h)\right) \) can then be written as

\[
C_{ij} \left(h^f, p^f(h)\right) = \sum_{m=1}^{M} c_m \left\{ p_m \leq h_{ij}^f p_{ij}^f(h)/N_j^f \left(h^f, p^f(h)\right) < p_{m+1} \right\}
\]  

(25)

because only one term is non-null in the former sum. We emphasize that (24) and (25) are given as examples, but different \( C_{ij} \left(h^f, p^f(h)\right) \) functions can be used.

The capacity constraint in (8) is written using the expressions in, e.g., (24) or (25) with the added constraints \( \mathbf{A} \alpha^f(h) \leq 1 \) and \( \alpha_{ij}^f(h) \in \{0, 1\} \). Note that the resulting optimization problem is not convex because of the integer constraints \( \alpha_{ij}^f(h) \in \{0, 1\} \). Likewise, the resource allocation problem in (22) is written replacing \( C_{ij} \left(h^f, p^f(h)\right) \) by their explicit expressions in (24) or (25) and adding the constraints \( \mathbf{A} \alpha^f(h) \leq 1 \) and \( \alpha_{ij}^f(h) \in \{0, 1\} \). The resulting optimal resource problem is not convex either. However preliminary research suggests that this problem is computationally tractable for some cases of the \( \mathbf{A} \) matrix [31]. This is a promising result as it implies that problems that cannot be solved as stated in primal form become tractable in their dual formulation. Conditions for the validity of this preliminary result will be fully explored. When these conditions do not hold, we will assess the value of approximating their optimal operating point with that of similar networks that do satisfy such conditions.

Interference limited physical layers. If terminals can interfere with each other and perform single user detection, link capacity is determined by the signal-to-interference-plus-noise ratio (SINR). Interference \( I_{ij}^f(h) \) to the \( T_i \rightarrow T_j \) link comes from: (i) terminals \( T_k \) in \( T_j \)'s neighborhood \( n(j) \) transmitting to any of its neighbors \( T_l \in n(k); \) and (ii) transmissions of \( T_l \) itself to terminals other than \( T_j \). Therefore

\[
I_{ij}^f(h) := \sum_{k \in n(j), k \neq i} \sum_{l \in n(k), l \neq j} h_{kl}^f p_{kl}^f(h) + \sum_{l \in n(i), l \neq j} h_{ij}^f p_{ij}^f(h) = \sum_{(k,l) \neq (i,j)} h_{kl}^f p_{kl}^f(h)
\]  

(26)

where \( (k, l) \neq (i, j) \) in the last sum signifies all pairs \( (k, l) \) different from \( (i, j) \) with proper neighborhood restrictions. The first sum includes terms of the form \( h_{ij}^f p_{ij}^f(h) \) to account for the interference of \( T_j \)'s transmissions to packets received at \( T_j \). Typically \( h_{ij}^f \) is very large discouraging transmission and reception of packets over the same tone \( f \). This is not prevented a fortiori but will likely emerge from the solution of the network optimization problem (7)–(9). With interference \( I_{ij}^f(h) \) as in (26), the SINR of the \( T_i \rightarrow T_j \) link on tone \( f \) is

\[
\gamma_{ij}^f(h) = \frac{h_{ij}^f p_{ij}^f(h)}{N_j^f + S^{-1} \sum_{(k,l) \neq (i,j)} h_{kl}^f p_{kl}^f(h)}.
\]  

(27)
where $S$ is the system’s spreading gain and, we recall, $N_j^f$ denotes noise power at $T_j$’s receiver end. The SINR in (27) can be substituted into, e.g., (24) or (25) to yield expressions for the capacity functions $C_{ij}(h^f, p^f(h))$. The resulting expressions are, in turn, substituted in (8) to obtain the optimal wireless networking problem and (22) to obtain the optimal resource allocation problem.

For interference-limited networks and capacity achieving codes the resulting physical layer optimization problem is known to be computationally intractable [23]. We will thus develop methods to find approximate solutions drawing from the literature on digital subscriber lines were similar problems have been studied; see, e.g., [51].

An important consequence of the lack of duality gap is that it becomes possible to determine how close to optimal an approximate candidate solution is. To accomplish this evaluation it suffices to compare primal and dual values at the candidate operating point. If duality gaps can be strictly positive this comparison is of limited value. The candidate point can be a good approximation to the optimal and the optimality gap significant because it is lower bounded by the duality gap. With null duality gap, a candidate point is close to optimal if and only if the gap between primal and dual values is small. Interestingly, preliminary simulations indicate that even simple approximation methods can yield reasonably good approximations to the optimal operating point. We present these simulations after discussing adaptive algorithms in Section C.5.3

**C.5.3. Adaptive algorithms.** Wireless link capacities are expressed as expected values over fading channel realizations. Their evaluation, therefore, requires access to the channel probability distribution function (pdf). Assuming a particular fading model the channel pdf is acquired by estimating channel moments; e.g., for Raleigh fading estimating mean channels is sufficient. While this is a possible approach it does restrict practical applicability as fading models are only rough approximations of distributions observed in field deployments. Algorithms that learn channel distributions simultaneously with the determination of the optimal operating point will be developed.

As stated in (15), solution of wireless optimization problems can be carried out by maximizing the Lagrangian in (14). The Lagrangian maximand involves expected values of stochastic functions suggesting similarities with the least mean-square (LMS) algorithm [13, Chapter 4]. The LMS algorithm solves a linear minimum (LM) mean squared error (MSE) estimation problem, where a stationary signal is estimated from stationary observations. To find the optimal LMMSE estimate a gradient descent algorithm can be used. As it turns out, the gradient depends on the cross-correlation of signal and observations and the observations’ correlations matrix. These matrices need not be known if the signal model is not accessible. In the LMS algorithm the unknown (cross-) covariance matrices are replaced by products of the instantaneous values of the variables, whose expected values are precisely
the (cross-) covariance matrices involved. Because this stochastic gradient is, on average, a descent direction, LMS iterates converge to the (optimal) LMMSE estimate – this is not strictly true, but some form of convergence can be proved for most LMS variations, see e.g., [13, Chapter 4].

To find optimal operating points of wireless networks we will develop descent algorithms in the dual domain to avoid non-convexities of the primal problem. Similar to LMMSE, the subgradients required to implement these algorithms necessitate evaluation of a expected value with respect to the fading’s distribution. Thus, we can propose a stochastic subgradient algorithm by replacing the expected value over channel realizations with the equivalent function for a single measured channel realization. As is the case in LMS, the stochastic subgradient points, on average, towards the optimal argument. Convergence is thus likely. Tools to study convergence properties of the resultant algorithms can be drawn from stochastic gradient descent algorithms of which the LMS is a particular case [16, Chapter 2], [33].

Besides formalizing this argument there are, at least, two more research challenges to be considered here. Even for deterministic convex problems, convergence in the dual domain does not always implies convergence in the primal domain, e.g., [4, Sec. 5.5.5]. Convergence of primal variables for a stochastic algorithm with non-covex constraints is not a simple corollary of dual variable’s convergence. Recall that to establish the latter it is necessary to prove feasibility and optimality, i.e., satisfaction of constraints (8)-(9) and near optimal utility. We will also incorporate in the analysis the use of approximate solutions to physical layer resource allocation problems. Emphasis here will be in determining methods to evaluate optimality gaps for stochastic iterates and studying methods to find close to optimal operating points.

Preliminary simulations. A stochastic subgradient descent algorithm was implemented to find the optimal operating point for the network in Fig. 1 using an interference limited physical layer and adaptive modulation. Nodes operate on 5 frequency tones, using direct sequence spread spectrum in each of them with spreading gain $S$. Three AMC modes corresponding to capacities 1, 2 and 3 bits/s/Hz are used with transitions at SINR 1, 3 and 7. Fading channels are generated as i.i.d. Rayleigh with average powers 1/2 for the links $4 \leftrightarrow 7$, $5 \leftrightarrow 9$, $7 \leftrightarrow 11$, $9 \leftrightarrow 10$, $11 \leftrightarrow 8$, $10 \leftrightarrow 6$, $8 \leftrightarrow 4$ and $6 \leftrightarrow 5$ and 1 for the remaining links. Noise power is $N_i^f = 0.1$ for all terminals and tones and maximum power per terminal is $p_{\text{max}} = 2$ – chosen so that if a terminal with 4 neighbors spreads power uniformly across all neighbors and tones the signal to noise ratio is 0dB. Four flows with destination at terminals 1, 7, 8 and 14 are considered with all other terminals required to deliver at least $a_{\text{min}}^k = 0.5$ bits/s/Hz and at most $a_{\text{max}}^k = 2$ bits/s/Hz to each of these flows. Beyond that, the optimality criteria is sum rate maximization. Maximum link capacity and routing rates are set at $c_{\text{max}} = r_{\text{max}} = 6$ bits/s/Hz. The power spectral mask is fixed at $p_{\text{max}} = 2$. To determine power allocations we choose a link $(i, j)$ and find the optimal power $p_{ij}^f(h)$ assuming all other powers $p_{lk}^f(h)$ with $(l, k) \neq (i, j)$ fixed. We do the same for all links and repeat the process $N = 3$ times. Note that this algorithm finds only a local, not necessarily global, minimum of the resource allocation problem in (22). Simulation results are presented in Figs. 4-7.

Convergence of the algorithm is corroborated by Figs. 4 and 5 respectively showing feasibility and optimality. Feasibility of all constraints takes in the order of 80 to 100 iterations. Fig. 5 shows that to achieve an optimality gap smaller than 10 percent takes 70 iterations. Figs. 4 and 5 show that asymptotic feasibility and optimality of
primal variables can be expected even if: (i) descent is along a stochastic direction; (ii) the primal problem is not convex; and (iii) we use approximate solutions to the power allocation problem (20). This is not likely to be true for all problems, but this preliminary result warrants further investigation.

Power allocation for the link from terminal 1 to terminal 4 is shown in Fig. 6. Without interference, optimal power distribution consists of transitions between AMC modes and within each mode, allocation is proportional to the inverse channel, i.e., $p^f_{1j}(h) \propto 1/h^f_{1j}$. With interference, some spread around this functional form is expected and indeed observed in Fig. 6. Routes to terminal 7 are shown in Fig. 7 to corroborate that packets are indeed being forwarded to the intended destination. Similarly reassuring figures are obtained for link capacities but are omitted due to space considerations.

C.5.4. Channel state information. Adapting network operation to time varying fading channels requires acquisition and dissemination of the CSI vector $h$. Due to channel estimation errors, estimates $\hat{h}$, not actual channels $h$, are available. Furthermore, since CSI is percolated through the network, estimates $\hat{h}$ are more accurately modeled as terminal dependent vectors $h_i$. Recall that the vector $h$ groups all channels in the network. The subindex in estimates $\hat{h}_i$ does not signify the channel being estimated but the location at which such estimate is available. We expect the $i$-th node to have accurate estimates of its own channels, $h^f_{ij}$ and $h^f_{ji}$, reasonable estimates of neighboring channels and otherwise rough estimates, or perhaps no information at all. We will investigate the incorporation of the cost of estimation in optimality criteria and the design of operation modes to deal with inaccurate and dissimilar estimates $\hat{h}_i$.

Power allocation decisions at the $i$-node are based on $\hat{h}_i$ yielding functions of the form $p^f_{ij}(\hat{h}_i)$. Link capacities and power consumptions become double expectations with respect to the distributions of the actual channels and their estimates. It is not difficult to see that as a consequence of Jensen’s inequality, better channel estimates yield larger link capacities $c_{ij}$ for the same average power consumption $p_i$. But acquiring CSI incurs a cost in terms of power and bandwidth used for channel probing and communication of channel estimates. A tradeoff arises between acquiring better estimates $\hat{h}_i$ and the cost of doing so. This tradeoff will be captured by introducing suitable constraints and objectives in the optimization problem. Pertinent questions here range from architectural to algorithmic cutting across the research proposed in Sections C.5.1-C.5.3.

A consequence of having different estimates $\hat{h}_i$ is that nodes cannot predict resource allocation of other network elements. Therefore, terminals’ resource allocation is based on the allocations they expect from the rest of the network. If this is captured by a double expectation as in the previous paragraph, network optimization yields an operating point that maximizes utility on an average sense. Because estimates may have significant variances, actual performance could be much worse than expected. An alternative approach is to formulate robust optimization problems whereby resource allocations $p^f_{ij}(\hat{h}_i)$ are hedged against the worst possible actions of other terminals. This provides an interesting motivation for game theoretic formulations in a cooperative setting. Nodes are indeed cooperating in forwarding each other’s packets, but compete in resource allocation because of inaccurate and dissimilar CSI.

C.5.5. Optimal wireless networking protocols. Research proposed in Sections C.5.1-C.5.4 is concerned with properties of optimal wireless networks and algorithms to find optimal operating points. The goal of this thrust is to translate these properties and algorithms into protocols. Protocols are defined as message exchange mechanisms between neighboring terminals that let the nodes determine suitable network variables. This will be accomplished by the development of algorithms amenable to distributed implementation. Such algorithms are borrowed from the parallel processing literature – see e.g., [2, Chapter 3] – and have been extensively used in NUM – see e.g., [22]. The discussion here is brief since the idea is to revisit problems in Sections C.5.1-C.5.4 to develop distributed implementations. We will derive distributed implementations for the resource allocation problems of Section C.5.2.
and consider their stochastic counterparts using the tools discussed in Section C.5.3. The cost of CSI acquisition will be incorporated in the resulting protocols along the lines discussed in Section C.5.4.

C.6 Education

In our research driven careers we sometimes forget that our future legacy is as much in our contributions to the advancement of science as in the lives we touch as educators. Though sometimes regarded as related but different tasks, research and teaching are in my mind two faces of the same coin. I like to say that we have to excel at research so that we have interesting things to share; and that we have to excel at teaching so that things we share actually sound interesting.

I am passionate about teaching. I think we are given wonderful opportunities to work with people during a stage of their lives that they will always regard fondly. Teaching is clearly about giving students solid knowledge and skills. But I also think that students are entitled to more than that. Engineering connotes curiosity, ingenuity and improvement. We are supposed to be curious to discover new problems, ingenious to solve them and ultimately develop things that improve our society. Even if I accept that this is an idealistic romanticized conception, I became an engineer because of it and I believe that this is true of most students. My education philosophy is to provide students with opportunities to satisfy their curiosity, nurture their ingenuity and understand how their work impacts our society. This philosophy underlies the proposed education activities that are discussed next.

C.6.1. Excitement, challenge and discipline gaps. The proposed education plan at the undergraduate level revolves around the excitement, challenge and discipline gaps. The excitement gap is about the excitement people feel about science and technology versus the lack of interest to pursue careers in science and technology. The challenge gap refers to the ongoing trend to reduce the complexity of the material we teach in our classes. The discipline gap alludes to the compartmental experience offered to our students and the reality of an increasingly hazy separation between disciplines. The goal of the proposed activities is to contribute to the closing of these gaps through the development of new courses.

To some extent the challenge gap is an effort to reduce the excitement gap. Engineering education is perceived as difficult, therefore, we can broaden its appeal by making it more accessible. I think that a rather opposite path is worth trying. I believe the gaps can be reduced by creating courses that introduce challenging applications drawn from a variety of disciplines. I have found that undergraduate students can grapple advanced level material if they can see how a particular piece of theory allows them to understand interesting problems. The idea is then to use interesting applications to lead students into complex material. Drawing applications from different disciplines serves to provide multidisciplinary perspectives and to further motivate students by showing that seemingly unrelated problems are, in fact, similar.

I am currently responsible for the undergraduate class in stochastic systems whose purpose is to introduce fundamental concepts used in discrete event simulations. To test feasibility of the proposed approach I am developing a curriculum that revolves around applications of stochastic processes to communications, biochemistry, social sciences and economics [44]. The class will be offered for the first time in Fall 2009.

As part of the class we study simulation of chemical reactions in biochemistry, where stochastic effects manifest because the number of molecules inside a cell is small. Interesting particular cases can be covered, e.g., the lac operon, a set of adjacent genes that form an auto-regulatory network to control the metabolism of lactose in some bacteria [48, Ch. 7]. Cells can only use glucose to generate energy, but they can reduce lactose to glucose if the latter is unavailable. To do this cells have to produce the enzyme $\beta$-galactosidase, which in itself requires some energy expenditure. Thus, production of $\beta$-galactosidase is only justified when lactose is abundant and glucose scarce. Production of $\beta$-galactosidase is controlled by repressor and promoter proteins. Upstream of the lac operon is a gene coding for a repressor protein that hinders mRNA transcription of $\beta$-galactosidase. When no lactose is present, the repressor binds to the operon thus interfering with mRNA transcription and resulting in low enzyme production. When lactose is present, however, the repressor binds preferentially to lactose therefore not interfering with transcription leading to increased production of $\beta$-galactosidase. The second part of the regulation involves the catabolite activator protein (CAP) that when bound to the operon facilitates mRNA transcription of $\beta$-galactosidase. The amount of CAP present is inversely proportional to the amount of glucose. Hence, enzyme production increases when glucose level decreases. When lactose and glucose are present this control mechanism results in a distinctive diauxie pattern with glucose consumed first and lactose processed after glucose is depleted.
We also study stochastic systems in social sciences covering, e.g., the random algorithm used by Google to rank web pages. Consider a web surfer that visits a page and clicks on any of the page’s links at random. She repeats this process forever. What fraction of his time will be spent on a given page? The answer to this question is the rank assigned to the page. This ranking algorithm can be also used to study reputation in social graphs with persons in place of web pages and social connections in place of links. We use this algorithm in class to find the student at the center of the class’s social graph. Applications from communication networks, e.g., wireless multiple access channels and economics, e.g., Black-Scholes’s model of option pricing roundup the class’s curriculum. As can be appreciated, the examples discussed above are quite complex, but chosen so that students can easily understand their relevance.

In the next five years, the class on stochastic process will be developed to maturity. Analogous ideas will be investigated to develop a new curriculum for the undergraduate class on signal processing.

C.6.2. Optimal design of wireless networks. The proposer believes that scientific progress depends on the continuous interaction of teaching, research and development. Research results are taught to students that incorporate them in new technology whose development uncovers novel research challenges. The natural place for this interaction is the graduate curriculum. To realize this vision a class on Optimal Design of Wireless Networks will be developed in the context of this proposal.

A first version of the class was developed in 2008 and a second iteration will be offered in Spring 2010 [43]. The research agenda described here can be naturally mapped into a reasonably broad coverage of wireless networks. The class starts with a brief introduction to wireless communications and networking explaining different methods used for multiple access, the role of fading and the problem formulations they lead to. We continue with an exposition of the separation principles that establishes an important architectural property of wireless networks while giving rise to five important problems: (i) Architectural principles in different settings. (ii) Exact and approximate solutions to resource allocation problems at the physical layer. (iii) Translation of algorithms that solve optimization problems into wireless networking protocols. (iv) Because the optimal operating point depends on the fading’s probability distribution, online learning of this distribution is required. (v) Acquisition and distribution of channel state information. The class is then devoted to explore these problems.

The contents of the class follow closely the list of research topics proposed in Section C.5. Since most problems are still open, the current version exposes students to problems (i)-(v) and provides a formal framework to face their analysis. As the research agenda advances, actual solutions of problems (i)-(v) will be found and incorporated into the class’s topics. Eventually, a formal theory of wireless networks will emerge from this and other research efforts, simpler ways of explaining networks will be devised and the class will morph into a conventional graduate-level offering. This progression coincides with a broadening of the spectrum of potential students. In its current form the offer is of interest to advanced-level Ph. D. students pursuing research in wireless networking. More mature versions of the class will be open to the general Ph. D. population and be possibly open to Master’s students.

C.7 Perspective

In characterizing the current understanding of networks the National Research Council report on Network Science compares the situation with metallurgy in the Renaissance period [8]. Steel forming technology of the day was sufficiently well developed so as to support the industrial revolution, but the scientific understanding of metallurgy was limited. When a science of metallurgy was finally developed in the late nineteenth century, better alloys and novel processes ignited the industrialization surge of the early twentieth century. In the last fifty years we have been able to engineer networks that have certainly transformed our world. Yet, while we obviously master the pertinent technology, our scientific understanding of fundamental issues arising in networked systems remains limited.

Success in the research effort proposed here has the potential to change how we teach and design wireless networks. As has happened with digital communications in the last fifty years, I believe we will shift from current heuristic based approaches to rigorous formulations. The importance of this shift should not be underestimated. Existing wireless networks, for example, utilize less than one tenth of the spectrum allocated to their operation. Ultimately, the goal is to improve this meager use of resources.
References


