# Power Spectrum Optimization of Multi-Carrier MIMO Ad Hoc Networks: A Dual Stochastic Approximation Approach

Jia Liu  Y. Thomas Hou*

The Bradley Department of Electrical and Computer Engineering

Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

**Abstract**

Multi-carrier MIMO (MC-MIMO) has emerged as a promising technology to provide significant capacity gain for broadband wireless ad hoc networks. However, due to the non-convex and large-scale problem structure, little work has been done on power spectrum optimization to harvest MC-MIMO's capacity gain for broadband ad hoc networks. Our principal goal in this work is to fill this challenging gap. First, we make advance in general non-convex optimization theory by showing a surprising result that the duality gap of a certain class of non-convex optimization problems is *zero* if they satisfy the "*concave perturbation*" condition. Further, we show that the power spectrum optimization problem of MC-MIMO ad hoc networks satisfies the concave perturbation condition. This important result allows us to tackle the problem with a much lower complexity in the dual domain, where we design efficient centralized and distributed algorithms based on complete knowledge of channel distribution information (CDI).

To further eliminate the requirement of complete CDI knowledge, we devise an online adaptive algorithm by stochastic approximation. We show that, when the feedback of channel state information (CSI) is error free, our proposed subgradient stochastic approximation (SSA) algorithm converges with probability one to the same optimal solution obtained by the off-line algorithm. In the case where there is error in CSI, we show that SSA can still converge with probability one to the same optimal solution if the resultant subgradients are asymptotically unbiased, or is recurrent to some neighborhood of the optimal solution when the resultant subgradients are biased.

# 1 Introduction

## 1.1 Motivation

Multiple-input multiple-output (MIMO) and multi-carrier systems are two new technologies that have received the most attention in recent years. By employing multiple antennas on both sides

*Please direct all correspondence to Prof. Thomas Hou, The Bradley Department of Electrical and Computer Engineering, 302 Whittemore Hall (0111), Virginia Tech, Blacksburg, VA, 24061, USA. Phone: +1-540-231-2950; Fax: +1-540-231-8292; Email: thou@vt.edu, URL: http://www.ece.vt.edu/thou/index.html.

of a wireless channel, MIMO can improve the spectral efficiency of *narrowband* communications by multiple-fold. On the other hand, multi-carrier techniques (e.g., orthogonal frequency-division multiple access (OFDMA)) offer an effective mechanism for communications over *broadband* wireless channels. Thus, it is not surprising to see that the combination of MIMO and multi-carrier systems is becoming an enabling technology for next generation broadband wireless networks [7].

Recently, there has been a surge of research interest in multi-carrier MIMO (MC-MIMO) technology at the physical layer (see, e.g., [7, 10, 14–16, 18, 22, 26]). Roughly speaking, the basic idea of MC-MIMO originates from the following holistic view: by partitioning the entire broadband spectrum that experiences frequency selective fading into multiple narrowband subcarriers, the fading on each subcarrier can be considered frequency-flat, which is much easier to mitigate. Further, for each narrowband subcarrier, MIMO can be employed to significantly increase the spectral efficiency.

Unfortunately, for wireless ad hoc networks, little work has been done so far on power spectrum optimization, which is the key to harvest MC-MIMO's significant capacity gain. This is perhaps due to the fact that power spectrum optimization for MC-MIMO ad hoc networks is indeed challenging. First, co-channel interferences among links (due to the lack of centralized infrastructure) in an ad hoc network renders a non-convex mathematical structure for the power spectrum optimization problem. To warrant optimality, state-of-the-art global optimization techniques (e.g., BB/RLT [17]) still have exponential complexity in $KM$ matrix variables, where $K$ and $M$ denotes the number of transmissions in the network and the number of subcarriers, respectively. In multi-carrier systems, $M$ is usually large (at least 128 and could be as large as 4096). Thus, an attempt to solve the power spectrum optimization problem by global optimization is not feasible. Second, compared to simple scalar channels in single antenna systems, power allocations are now performed over complex matrix channels in time, frequency, and space domains. Optimal solutions need to be searched in a much higher dimensional space. Finally, in MC-MIMO ad hoc networks, wireless channel responses are not only frequency varying but also time varying. How to judiciously exploit channel state information (CSI) to improve network performance is also very complicated. In light of the increasingly compelling need of extending MC-MIMO to broadband ad hoc networks, one of the major goals in this work is to fill this challenging gap and develop efficient and near-optimal

power spectrum optimization algorithms for MC-MIMO ad hoc networks.

In this paper, we consider optimizing power spectrum for a MC-MIMO ad hoc networks so that the maximum weighted sum rate (MWSR) of the network is maximized. The choice of MWSR objective is motivated by recent results in [1,4,13,25], where it has been shown that adaptive policies based on solving a MWSR problem in each time slot can be used to stabilize the transmission buffers for any set of arrival rates in the ergodic network capacity region. Further, it has been shown in [23] that by appropriately updating the weights across time slots, a proportionally fair scheduler can be designed based on MWSR.

In this paper, we tackle the MWSR problem in two steps: 1) off-line centralized and distributed algorithms based on complete knowledge of the underlying channel distribution information (CDI); and 2) an online adaptive algorithm that does not require full CDI knowledge. Accordingly, the results are organized into two thrusts.

## 1.2 Summary of Main Results

In our first thrust, we make advance in general non-convex optimization theory by showing that the duality gap of a certain class of non-convex optimization problems is *zero* if they satisfy the "*concave perturbation*" condition. This result is surprising because it is well-known that non-convex optimization problems have non-zero duality gap in general. Further, we show that the power spectrum optimization problem in MC-MIMO ad hoc networks satisfies the concave perturbation condition. This results legitimates the approach to tackle the MWSR problem in its dual domain, which has a *linear* complexity in $M$. This improvement is significant because, as we indicated earlier, $M$ is large in multi-carrier systems. This complexity reduction is made possible by decomposing the dual problem into $M$ subproblems, each of which corresponds to a single subcarrier. However, it turns out that each subproblem remains a difficult non-convex problem that has an exponential complexity in $K$ $n_t$-dimensional matrix variables, where $n_t$ is the number of transmit antennas at each node. To circumvent this difficulty, we develop efficient centralized and distributed algorithms by appealing to matrix differential calculus theory [12]. We show that the proposed algorithms can quickly identify solutions that are very close to the global optimum obtained by exhaustive search.

In the first thrust, our proposed algorithms are based on full CDI knowledge. We should point out that in practice, CDI knowledge is acquired by an off-line estimation stage. Thus, the algorithms developed in the first thrust are off-line algorithms. However, we show that they can serve as the baseline that provides important insights and frameworks for our online adaptive algorithm.

In our second thrust, we focus on developing an online adaptive algorithm that does not assume full CDI knowledge. We devise a subgradient stochastic approximation (SSA) algorithm. SSA can adapt to any *unknown* underlying fading distribution on the fly, lending itself an online algorithm desirable for practical implementation. We further establish its convergence theorem that SSA converges with probability one to the same optimal solution obtained by the off-line algorithms. We note that the proof of the convergence theorem is non-trivial since standard techniques (e.g., closedness of algorithmic maps) are not applicable for our problem. Instead, we use a novel approach by appealing to recent results on Martingale theory [8]. Finally, we investigate the impact of CSI error on SSA. We consider two cases, depending on whether the resultant stochastic subgradients are asymptotically biased or unbiased. We prove that, if the stochastic subgradients are unbiased, SSA still converges with probability one to the same optimal solution obtained by the off-line methods. On the other hand, in the biased case, we show that the iterates generated by SSA is recurrent to some neighborhood of the optimal solution.

## 1.3 Related Work and Paper Organization

As indicated earlier, there has been tremendous interest on MC-MIMO techniques at the physical layer in recent years, e.g., joint transceiver design [14,15], channel estimation [18,26], multiple access strategies [22], limited feedback and antenna selection techniques [16], etc. An excellent overview of MC-MIMO physical layer techniques can be found in [7]. Another closely-related line of research is on joint subcarrier scheduling and power spectrum optimization for multi-carrier communications systems [1, 11, 19, 20, 24]. However, these studies are limited to single-antenna systems. A more relevant work to this paper can be found in [10], where Liu and Hou investigated MC-MIMO power spectrum optimization for the downlinks of WiMAX access networks. The major difference between our work and [10] is that our focus is on ad hoc communications (which is prone to co-channel

interference), while the network setting in [10] is infrastructure-based (which relies on centralized scheduling to eliminate interference). So far, the research on MC-MIMO ad hoc networks is still in its infancy and results in this area remain limited. It is worth pointing out that our work is different to another line of research termed "multi-channel multi-radio" (MC-MR) wireless networks (see, e.g., [9] and references therein). In MC-MR wireless networks, the number of channels is usually small (e.g., IEEE 802.11 networks [9]).

The remainder of this paper is organized as follows. Section 2 introduces the network model and problem formulation. Section 3 presents the zero duality gap theorem. In Section 4, we design off-line algorithms for solving the MWSR problem in the dual domain. In Section 5, we present an online stochastic approximate algorithm and establish its convergence theorem. In Section 6, we investigate the impact of CSI error on our proposed stochastic approximation algorithm. Section 7 concludes this paper.

# 2 Network Model and Problem Formulation

In this paper, we use boldface to denote matrices and vectors. For a matrix $\mathbf{A}$, $\mathbf{A}^\dagger$ denotes the conjugate transpose, $\mathrm{Tr}\{\mathbf{A}\}$ denotes the trace of $\mathbf{A}$, $|\mathbf{A}|$ denotes the determinant of $\mathbf{A}$, and $\|\mathbf{A}\|$ denotes the Frobenius norm of $\mathbf{A}$. $\mathrm{Diag}\{\mathbf{A}_1, \ldots, \mathbf{A}_n\}$ represents the block diagonal matrix with matrices $\mathbf{A}_1, \ldots, \mathbf{A}_n$ on its main diagonal. We denote $\mathbf{I}$ the identity matrix with dimension determined from the context. $\mathbf{A} \succeq 0$ represents that $\mathbf{A}$ is Hermitian and positive semidefinite (PSD). $\mathbf{1}$ and $\mathbf{0}$ denotes a vector whose elements are all ones and zeros, respectively, and their dimensions are determined from the context. For a real vector $\mathbf{v}$ and a real matrix $\mathbf{A}$, $\mathbf{v} \geq \mathbf{0}$ and $\mathbf{A} \geq \mathbf{0}$ mean that all entries in $\mathbf{v}$ and $\mathbf{A}$ are nonnegative, respectively. The operator "$\langle \cdot, \cdot \rangle$" represents the inner product operation for vectors or a matrices.

We consider a MC-MIMO ad hoc network consisting of $K$ concurrent transmission pairs and $M$ subcarriers, as shown in Figure 1. For convenience, each transmission pair is referred to as a link, although such a link does not physically exist. Suppose that the transmitting node of each link has

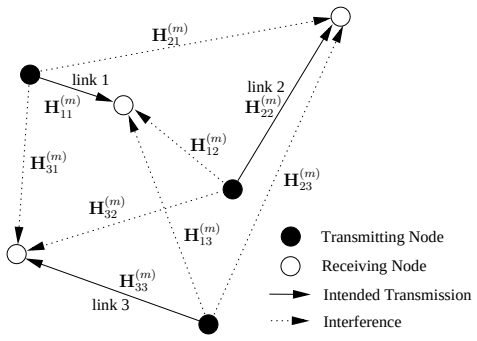


Figure 1: Network model of a MC-MIMO ad hoc network.

$n_t$ antennas and the receiving node has $n_r$ antennas. Let $\mathbf{H}_{ij}^{(m)} \in \mathbb{C}^{n_r \times n_t}$ represent the channel gain matrix from the transmitting node of link $j$ to the receiving node of link $i$ over subcarrier $m$. We let $\mathbf{H} = \{\mathbf{H}_{ij}^{(m)} : i, j = 1, \ldots, K,\ m = 1, \ldots, M\}$ denote the collection of all channel gain matrices. The entries in each channel gain matrix are assumed to be i.i.d. complex Gaussian distributed. The total bandwidth of the network is $B$. As a result, the bandwidth of each subcarrier is $\frac{B}{M}$. We assume that $M$ is sufficiently large such that the fading on each subcarrier can be considered frequency-flat.

The normalized received complex base-band signal vector at the receiving node of link $k$ over subcarrier $m$ can be computed as

$$\mathbf{r}_k^{(m)} = \sqrt{\rho_{kk}}\mathbf{H}_{kk}^{(m)}\mathbf{t}_k^{(m)} + \sum_{i=1, i\neq k}^{K} \sqrt{\rho_{ki}}\mathbf{H}_{ki}^{(m)}\mathbf{t}_i^{(m)} + \mathbf{n}, \tag{1}$$

where $\mathbf{t}_k^{(m)} \in \mathbb{C}^{n_t}$ and $\mathbf{r}_k^{(m)} \in \mathbb{C}^{n_r}$ represent the transmitted signal vector (with unit power) and the received signal vector of link $k$ over subcarrier $m$, respectively, $\mathbf{n}$ represents the normalized complex additive white Gaussian noise vector with zero mean and unit variance. In (1), $\rho_{ki}$ denotes the signal-to-noise ratio (SNR) of link $k$ if $i = k$, or the interference-to-noise ratio (INR) from link $i$ to link $k$ if $i \neq k$.

Let matrix $\mathbf{Q}_k^{(m)}$ denote the covariance matrix of input symbol vector $\mathbf{t}_k^{(m)}$, i.e., $\mathbf{Q}_k^{(m)} = \mathbb{E}\{\mathbf{t}_k^{(m)} \cdot \mathbf{t}_k^{(m)\dagger}\}$. It is evident from this definition that $\mathbf{Q}_k^{(m)} \succeq 0$ and $\mathrm{Tr}(\mathbf{Q}_k^{(m)}) \leq 1$. Physically, $\mathbf{Q}_k^{(m)}$ represents the power allocation among the $n_t$ antennas of the transmitting node of link $k$ when transmitting

data over subcarrier $m$. We let $\mathbf{Q} = \{\mathbf{Q}_k^{(m)} : k = 1, \ldots, K,\ m = 1, \ldots, M\}$ denote the collection of all input covariance matrices. The ergodic capacity of link $k$ over subcarrier $m$ can be computed as [5]

$$C_k^{(m)} = \frac{B}{M} \mathbb{E}_{\mathbf{H}} \left[ \log_2 \left| \mathbf{I} + \rho_{kk} \mathbf{H}_{kk}^{(m)} \mathbf{Q}_k^{(m)} \mathbf{H}_{kk}^{(m)\dagger} (\mathbf{R}_k^{(m)})^{-1} \right| \right], \tag{2}$$

where $\mathbb{E}_{\mathbf{H}}[\cdot]$ represents the expectation taken over the distribution of $\mathbf{H}$. In (2), $\mathbf{R}_k^{(m)}$ represents the aggregated interferences at the receiving node of link $k$ over subcarrier $m$, and can be computed as

$$\mathbf{R}_k^{(m)} = \mathbf{I} + \textstyle\sum_{i=1, i\neq k}^{K} \rho_{ki} \mathbf{H}_{ki}^{(m)} \mathbf{Q}_i^{(m)} \mathbf{H}_{ki}^{(m)\dagger}. \tag{3}$$

Since that the average consumed power cannot be larger than the maximum transmit power, we have the following constraint for each transmitting node:

$$\mathbb{E}_{\mathbf{H}} \left[ \textstyle\sum_{m=1}^{M} \mathrm{Tr}(\mathbf{Q}_k^{(m)}) \right] \leq 1, \quad \forall k. \tag{4}$$

Putting together these expressions, constraints, and definitions, the MWSR problem of a MC-MIMO ad hoc network can be formulated as follows:

$$\begin{array}{rl} \text{Maximize} & \sum_{m=1}^{M} \sum_{k=1}^{K} w_k C_k^{(m)} \\ \text{subject to} & \mathbb{E}_{\mathbf{H}} \left[ \sum_{m=1}^{M} \mathrm{Tr}(\mathbf{Q}_k^{(m)}) \right] - 1 \leq 0, \forall k \\ & \mathbf{Q}_k^{(m)} \succeq 0, \forall k, m, \end{array} \tag{5}$$

where $w_i$, $i = 1, 2, \ldots, K$, are given weights. In (5), $\mathbf{Q}_k^{(m)}$'s are optimization variables.

Problem (5) is a complex non-convex optimization problem because the objective function of (5) is not concave in $\mathbf{Q}_k^{(m)}$. As a result, numerical optimization of (5) is difficult. To warrant optimality, state-of-the-art global optimization techniques (e.g., BB/RLT [17]) still have exponential complexity in the total number of variables $KM$. In addition, to ensure flat fading on each subcarrier, the value of $M$ is large. Thus, solving (5) by global optimization technique is not feasible.

In the next section, we establish a general theory to show that a class of non-convex optimization problems having the same structure of (5) have *zero* duality gap if they satisfy certain perturbation condition. Further, we show that (5) indeed satisfies such condition. Thus, we can equivalently solve the dual problem of (5), which, fortunately, has linear complexity in $M$.

# 3 Non-convex Optimization Problems with Zero Duality Gap

Let us consider the general form of (5), which can be written as

$$\begin{array}{ll} \text{Maximize} & \sum_{m=1}^{M} f_m(\mathbf{x}_m) \\ \text{subject to} & \sum_{m=1}^{M} \mathbf{g}_m(\mathbf{x}_m) \leq \mathbf{0}, \end{array} \tag{6}$$

where $\mathbf{x}_m \in \mathbb{S} \subseteq \mathbb{C}^K$ are vectors of the optimization variables, $f_m(\cdot) : \mathbb{C}^K \rightarrow \mathbb{R}$ are functions that may or may not be concave, and $\mathbf{g}_m(\cdot) : \mathbb{C}^K \rightarrow \mathbb{C}^L$ are matrix-valued functions that may or may not be convex.

By associating a dual vector variable, denoted by $\mathbf{u} \in \mathbb{R}^L$, with the constraint, the Lagrangian of (6) can be written as

$$L(\mathbf{x}_m, \mathbf{u}) = \sum_{m=1}^{M} f_m(\mathbf{x}_m) + \mathbf{u}^T \left( \mathbf{0} - \sum_{m=1}^{M} \mathbf{g}_m(\mathbf{x}_m) \right). \tag{7}$$

The Lagrangian dual function is an unconstrained optimization with respect to $\mathbf{x}_m$, i.e.,

$$\Theta(\mathbf{u}) \triangleq \max_{\mathbf{x}_m} L(\mathbf{x}_m, \mathbf{u}). \tag{8}$$

The Lagrangian dual optimization problem is

$$\begin{array}{ll} \text{Minimize} & \Theta(\mathbf{u}) \\ \text{subject to} & \mathbf{u} \geq \mathbf{0}. \end{array} \tag{9}$$

If $f_m(\mathbf{x}_m)$'s are concave and $\mathbf{g}_m(\mathbf{x}_m)$'s are convex, (6) becomes a convex optimization problem. If Slater's condition[1] holds (i.e., strictly feasible region is non-empty), then it is known that the primal problem in (6) and the dual problem in (9) have the same solution. On the other hand, when convexity does not hold, weak duality theorem says that the solution of the dual problem is an upper bound of the primal problem, and the difference between the upper bound and the

[1]Slater condition holds for almost all practical optimization problems [2].

true optimum is referred to as "duality gap." However, in what follows, we will show that when a non-convex optimization problem with the structure of (6) satisfies certain conditions, the duality gap is zero. Toward this end, we first introduce the concept of *perturbation function*.

**Definition 1.** *The perturbation function of (6), denoted as* $\theta(\mathbf{y}) : \mathbb{C}^K \rightarrow \mathbb{C}^L$*, is defined as the optimal value function of the following problem,*

$$\theta(\mathbf{y}) = \max \left\{ \sum_{m=1}^{M} f_m(\mathbf{x}_m) : \sum_{m=1}^{M} \mathbf{g}(\mathbf{x}_m) \leq \mathbf{y} \right\}, \ \mathbf{y} \in \mathbb{C}^K.$$

From the definition, it is clear that $\theta(\mathbf{0})$ yields the same optimal objective value as that of the primal problem in (6). Now, we introduce the first main result of this paper.

**Theorem 1** (Concave Perturbation Condition)**.** *If the perturbation function* $\theta(\mathbf{y})$ *is concave, then the non-convex optimization problem in (6) and its Lagrangian dual problem in (9) has the same optimal objective value, i.e., the duality gap is zero.*

*Proof.* The basic idea of the proof is to show the existence of a *saddle point*, which implies absence of duality gap.

For convenience, we let matrix $\mathbf{X} = [\mathbf{x}_m : m = 1, \ldots, M] \in \mathbb{S}^M \subseteq \mathbb{C}^{K \times M}$ denote collection of all $\mathbf{x}_m$ variables. Let $\bar{\mathbf{X}} = [\bar{\mathbf{x}}_m : m = 1, \ldots, M]$ be the solution to the primal problem in (6). Since $\theta(\mathbf{y})$ is concave, there exists a hyperplane that supports the hypograph of $\theta(\mathbf{y})$ for any $\mathbf{y} \in \mathbb{C}^L$ [2]. Thus, for $\theta(\mathbf{0})$, there exists some $\bar{\mathbf{u}} \in \mathbb{C}^L$ such that, for all $\mathbf{y} \in \mathbb{C}^K$,

$$\theta(\mathbf{y}) \leq \theta(\mathbf{0}) + \bar{\mathbf{u}}^T(\mathbf{y} - \mathbf{0}). \tag{10}$$

Our approach is to show that $(\bar{\mathbf{X}}, \bar{\mathbf{u}})$ is indeed a saddle point of the Lagrangian function in (7), i.e.,

$$L(\mathbf{X}, \bar{\mathbf{u}}) \leq L(\bar{\mathbf{X}}, \bar{\mathbf{u}}) \leq L(\bar{\mathbf{X}}, \mathbf{u}), \tag{11}$$

for all $\mathbf{X} \in \mathbb{S}^M$ and all $\mathbf{u} \geq 0$. The complete proof of this theorem is provided in Appendix A. □

**Remark 1.** *When all* $f_m(\mathbf{x}_m)$*'s are concave and all* $\mathbf{g}_m(\mathbf{x}_m)$*'s are convex, the concave perturbation condition is automatically satisfied. However, the converse is not true.*

Next, we show that the non-convex MWSR problem in (5) satisfies the concave perturbation condition. To see this, let us first consider the continuous version of the MWSR problem in (5). That is, $M = \infty$ and $\mathbf{H}_{ki}(f)$ is the matrix-valued channel response function of $f$ from the transmitting node of link $i$ to the receiving node of link $k$. The problem formulation is re-written as follows:

$$\begin{aligned} \text{Maximize} \quad & \sum_{k=1}^{K} w_k \int C_k(f) df \\ \text{subject to} \quad & \int \mathbb{E}_{\mathbf{H}} \left[ \text{Tr}(\mathbf{Q}_k(f)) \right] df - 1 \leq 0, \forall k \\ & \mathbf{Q}_k(f) \succeq 0, \forall k, f. \end{aligned} \tag{12}$$

It is not difficult to see that the optimal solution to (12) must possess a frequency division multiplex (FDM) structure, where each band corresponds to a transmission strategy for which a subset of $K$ links transmit. Thus, there are at most $\sum_{i=0}^{K} \binom{K}{i} = 2^K$ different transmission strategies. The FDM structure is shown in Figure 2. For example, the solid block corresponding to link 1 and band 1 in Figure 2 represents that link 1 transmits in band 1.

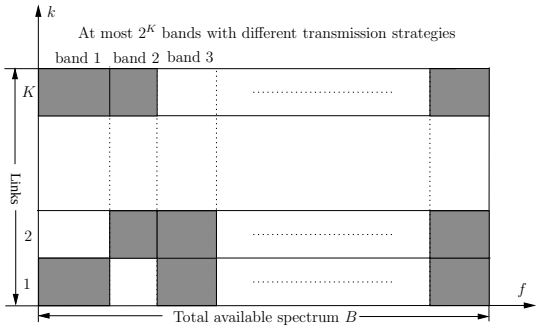


Figure 2: The FDM structure of the solution to (12).

Toward this end, we first consider the case where the channel response functions are constant, for which we have the following theorem.

**Theorem 2.** *Problem (12) satisfies the concave perturbation condition if* $\mathbf{H}_{ki}(f)$ *are constants for all* $k$ *and* $i$*, i.e.,* $\mathbf{H}_{ki}(f) = \mathbf{H}_{ki}$*,* $\forall k, i$*.*

*Proof.* Let $\theta(\mathbf{y})$ be the perturbation function of (12), where $\mathbf{y} \in \mathbb{C}^K$. Suppose that $\mathbf{Q}_1^*(f)$ and $\mathbf{Q}_2^*(f)$ solve the primal problem in $\theta(\cdot)$ when $\mathbf{y} = \mathbf{y}_1$ and $\mathbf{y} = \mathbf{y}_2$, respectively. Let $0 \leq \mu \leq 1$. To prove Theorem 2, it suffices to construct a solution $\mathbf{Q}(f)$ that is feasible to the power constraints in $\theta(\mu\mathbf{y}_1 + (1-\mu)\mathbf{y}_2)$ and $\theta(\mu\mathbf{y}_1 + (1-\mu)\mathbf{y}_2)$ satisfies

$$\theta(\mu\mathbf{y}_1 + (1-\mu)\mathbf{y}_2) \geq \mu\theta(\mathbf{y}_1) + (1-\mu)\theta(\mathbf{y}_2). \tag{13}$$

Without loss of generality, consider $\mathbf{Q}_1^*(f)$, which solves the perturbation function $\theta(\mathbf{y}_1)$. Recall that the solution to (12) has a FDM structure with at most $2^K$ different transmission strategies. Since all channel response functions $\mathbf{H}_{ki}(f)$ are constant functions of $f$ (i.e., a flat function of $f$), the same Karush-Kuhn-Tucker (KKT) condition must hold at optimality for all frequency $f$. As a result, $\mathbf{Q}_1^*(f)$ must be flat within each band. Likewise, we can also conclude that $\mathbf{Q}_2^*(f)$ is flat within each band in another FDM scheme.

Now, suppose that $C_1^*$ and $C_2^*$ are the optimal objective value of $\theta(\mathbf{y}_1)$ and $\theta(\mathbf{y}_2)$, respectively. The way to construct a $\mathbf{Q}(f)$, which is feasible to $\theta(\mu\mathbf{y}_1 + (1-\mu)\mathbf{y}_2)$ and achieves *at least* $\mu C_1^* + (1-\mu)C_2^*$ for all $0 \leq \mu \leq 1$, is illustrated in Figure 3.

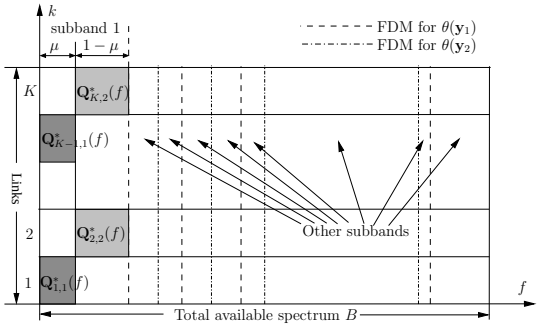


Figure 3: Construct a feasible solution $\mathbf{Q}(f)$ to $\theta(\mu\mathbf{y}_1 + (1-\mu)\mathbf{y}_2)$.

It is clear that each frequency $f$ falls in some intersection of the FDM schemes of $\theta(\mathbf{y}_1)$ and $\theta(\mathbf{y}_2)$. We call each distinct intersection a subband. For each subband, we assign $\mu$ portion of it as $\mathbf{Q}_1^*(f)$ and $(1-\mu)$ portion of it as $\mathbf{Q}_2^*(f)$. Apparently, $\mathbf{Q}(f)$ is feasible because $\int \mathbf{Q}(f) = \int \mu\mathbf{Q}_1^*(f) + (1-\mu)\mathbf{Q}_2^*(f) \leq \mu\mathbf{y}_1 + (1-\mu)\mathbf{y}_2$. Also, $\mathbf{Q}(f)$ achieves an objective value of $\mu C_1^* + (1-\mu)C_2^*$. Hence, the concave perturbation condition holds for the case where channel responses are constant

functions. □

Next, we show that for general cases where the channel response functions are not constant, (12) still satisfies the concave perturbation condition. This is stated formally as follows.

**Theorem 3.** *If $\mathbf{H}_{ki}(f)$ is a continuous function of $f$ for all $k$ and $i$, then (12) satisfies the concave perturbation condition.*

*Proof.* The proof of the general case where the channel response functions are not constant is based on the following asymptotic argument. First, the total available spectrum can be divided into a set of infinitesimal frequency subbands. Since the channel response functions are continuous, $\mathbf{H}_{ki}(f)$ in each subband approaches a constant for all $k$ and $i$ as the number of divisions approaches infinity. Then, from Theorem 2, we know that the concave perturbation condition is satisfied in each infinitesimal band. Therefore, the concave perturbation condition holds for the entire available spectrum. □

Finally, for the original problem (5), which is a discrete version of (12), we have the following proposition.

**Proposition 4.** *Problem (5) satisfies the concave perturbation condition when the number of carriers $M$ is large.*

*Proof.* Since $M$ is large, the channel response functions of adjacent carriers are approximately the same. Therefore, the combined channel response function of all subcarriers can be viewed as approximately continuous. Thus, from Theorem 3, we can conclude that the concave perturbation condition is satisfied for Problem (5) when $M$ is large. □

**Remark 2.** Since the proof of Proposition 4 hinges upon an approximating argument, the duality gap of (5) is not strictly zero. However, the duality gap vanishes as $M$ increases asymptotically. This result allows us to solve (5) in its dual domain.

# 4 Power Spectrum Optimization in the Dual Domain

In the last section, we have shown that (5) can be equivalently solved in its dual domain. Associate a dual variable $u_i$ with each constraint in (5) and denote $\mathbf{u} \triangleq [u_1\, u_2\, \ldots\, u_K]^T$ the collection of such dual variables. Then, the Lagrangian dual function can be written as

$$\Theta(\mathbf{u}) \triangleq \max_{\mathbf{Q}} \left\{ L(\mathbf{Q}, \mathbf{u}) : \mathbf{Q}_k^{(m)} \geq 0,\ \forall k,\ \forall m \right\}, \tag{14}$$

where $\mathbf{Q}$ denotes the collection of all power allocation variables. Then, the master dual problem of MWSR can be written as:

$$\begin{aligned} \mathbf{D}^{\mathbf{MWSR}} : \text{Minimize} \quad & \Theta(\mathbf{u}) \\ \text{subject to} \quad & \mathbf{u} \geq \mathbf{0}. \end{aligned} \tag{15}$$

In (14), the Lagrangian $L(\mathbf{Q}, \mathbf{u})$ is computed as

$$\begin{aligned} L(\mathbf{Q}, \mathbf{u}) = \sum_{k=1}^{K} w_k \mathbb{E}_{\mathbf{H}} \Bigg[ \sum_{m=1}^{M} \log_2 \Big| \mathbf{I} + \rho_{kk} \mathbf{H}_{kk}^{(m)} \mathbf{Q}_k^{(m)} \mathbf{H}_{kk}^{(m)\dagger} \cdot \\ (\mathbf{R}_k^{(m)})^{-1} \Big| \Bigg] + \sum_{k=1}^{K} u_k \Bigg[ 1 - E_{\mathbf{H}} \Big[ \sum_{m=1}^{M} \mathrm{Tr}(\mathbf{Q}_k^{(m)}) \Big] \Bigg]. \end{aligned}$$

After rearranging terms and interchanging expectation and summations, the Lagrangian can be re-written as

$$L(\mathbf{Q}, \mathbf{u}) = \sum_{m=1}^{M} \sum_{k=1}^{K} \mathbb{E}_{\mathbf{H}} \left[ F(\mathbf{Q}_k^{(m)}) \right] + \sum_{k=1}^{K} u_k, \tag{16}$$

where $F(\mathbf{Q}_k^{(m)})$ is defined as

$$\begin{aligned} F(\mathbf{Q}_k^{(m)}) \triangleq w_k \log_2 \left| \mathbf{I} + \rho_{kk} \mathbf{H}_{kk}^{(m)} \mathbf{Q}_k^{(m)} \mathbf{H}_{kk}^{(m)\dagger} (\mathbf{R}_k^{(m)})^{-1} \right| \\ - u_k \mathrm{Tr}(\mathbf{Q}_k^{(m)}). \end{aligned} \tag{17}$$

It is not difficult to see that we can decompose (14) as follows:

$$\Theta(\mathbf{u}) = \sum_{m=1}^{M} \mathbb{E}_{\mathbf{H}} \left[ \max_{\substack{\mathbf{Q}_k^{(m)} \succeq 0 \\ \forall k}} \left\{ \sum_{k=1}^{K} F(\mathbf{Q}_k^{(m)}) \right\} \right] + \langle \mathbf{u}, \mathbf{1} \rangle, \tag{18}$$

i.e., $\Theta(\mathbf{u})$ is decomposed into $M$ subproblems. Thus, $\Theta(\mathbf{u})$ has a linear complexity in $M$, as opposed to the exponential complexity in $M$ in the primal problem.

It can be seen from (18) that each subproblem corresponds to one subcarrier and is a maximization problem of $K$ matrix variables in the following form:

$$\begin{aligned} \text{Maximize} \quad & \textstyle\sum_{k=1}^{K} \Big[ w_k \log_2 \big| \mathbf{I} + \rho_{kk} \mathbf{H}_{kk}^{(m)} \mathbf{Q}_k^{(m)} \mathbf{H}_{kk}^{(m)\dagger} \cdot \\ & \qquad (\mathbf{R}_k^{(m)})^{-1} \big| - u_k \mathrm{Tr}(\mathbf{Q}_k^{(m)}) \Big] \\ \text{subject to} \quad & \mathbf{Q}_k^{(m)} \succeq 0, \forall k. \end{aligned} \tag{19}$$

To solve the MWSR problem in the dual domain, we need to evaluate the optimal Lagrangian dual in (14) with respect to $\mathbf{Q}$ first. Then, we need to solve the master dual problem in (15) with respect to $\mathbf{u}$. In the following subsections, we will discuss the solution procedure for these two problems in more detail.

## 4.1 Solving the Master Dual Problem

It can be shown that the master dual problem in (15) is a non-differentiable convex problem, which can be solved by subgradient algorithm. For the MWSR problem, the subgradients during the $n^{th}$ iteration can be computed as

$$d_{u_k}(n) = 1 - \mathbb{E}_{\mathbf{H}} \Big[ \sum_{m=1}^{M} \mathrm{Tr}(\mathbf{Q}_k^{(m)*}(n)) \Big], \quad \forall k. \tag{20}$$

where $\mathbf{Q}_k^{(m)*}(n)$ represent the optimal solutions in (18) with $\mathbf{u}$ being replaced by $\mathbf{u}(n)$. The updates of dual variables can be computed as

$$u_k(n+1) = [u_k(n) - s_n d_{u_k}(n)]_+^U, \forall k, \tag{21}$$

where $[\cdot]_+ = \max(\cdot, 0)$, $s_n$ is the step size during the $n^{th}$ iteration, and $U$ is some upper bound for $u_k$. It is shown in [2] that if the step size selection satisfies the following three conditions:

$$s_n \to 0, \quad \sum_{n=1}^{\infty} s_n \to \infty, \text{ and } \sum_{n=1}^{\infty} s_n^2 < \infty, \tag{22}$$

then the iterates generated by the subgradient method converge to the optimal solution $\mathbf{u}^*$. An easy step size selection strategy is the divergent harmonic series: $s_n = \frac{\beta}{n}$, $n = 1, 2, \ldots$, where $\beta$ is some positive constant. We summarize the subgradient algorithm in Algorithm 1.

**Algorithm 1** Subgradient Algorithm for Solving MWSR

1. Choose the initial starting points $\mathbf{u}(0)$. Let $n = 0$.
2. Compute $\mathbf{Q}^*(n)$ by solving $M$ subproblems in the form of (19), each of which corresponds a subcarrier.
3. In the $n^{th}$ iteration, choose an appropriate step size $s_n$. Update subgradient $d_{\mathbf{u}_k}(n)$ using (20) with $\mathbf{Q}^*(n)$.
4. Update dual variables $\mathbf{u}(n)$ using (21) with $d_{u_k}(n)$.
5. If $\|\mathbf{u}_k(n+1) - \mathbf{u}_k(n)\| < \epsilon$, then return $\mathbf{Q}^*(n)$ as the final optimal solution and stop. Otherwise, let $n \leftarrow n + 1$ and goto Step 2.

The most attractive feature of the subgradient algorithm is that it can be implemented in a distributed fashion. This is because the subgradient updates in (20) only involve local power variables at each node. Also, in the updates of dual variables in (21), the step size selection $s_n = \beta/n$ depends only on the iteration index (can be defined as some function of elapsed time) and does not require any other global information.

## 4.2 Solving Single Subcarrier Subproblem

Due to co-channel interferences and MIMO nature of the problem, each subproblem (19) is a non-convex optimization problem of $K$ $n_t$-dimensional matrix variables, which has exponential complexity in $Kn_t^2$. As a result, solving (19) remains very difficult. In this paper, we propose two gradient projection (GP) methods based on matrix differential calculus [12]. GP can efficiently find a local

optimal solution. We will show later that the gap between the solution obtained by GP and the true optimal solution is small (or negligible).

**Gradient Computation** For convenience, we denote the objective function in (19) as $I(\mathbf{Q})$. By using the identity $\frac{\partial}{\partial \mathbf{X}} \ln\det(\mathbf{A}+\mathbf{BXC}) = [\mathbf{C}(\mathbf{A}+\mathbf{BXC})^{-1}\mathbf{B}]^T$ from matrix differential calculus [12], we are able to obtain the gradients $\mathbf{G}_k^{(m)} \triangleq \nabla_{\mathbf{Q}_k^{(m)}} I(\mathbf{Q})$ as:

$$\mathbf{G}_k^{(m)} = \frac{2w_k\rho_{kk}}{\ln 2}\mathbf{H}_{kk}^{(m)\dagger}\left(\mathbf{R}_k + \rho_k \mathbf{H}_{kk}^{(m)}\mathbf{Q}_k\mathbf{H}_{kk}^{(m)\dagger}\right)^{-1}\mathbf{H}_{kk}^{(m)} + \sum_{j=1,j\neq k}^{K} \frac{2w_j\rho_{jk}}{\ln 2}\mathbf{H}_{jk}^{(m)\dagger}\Big[(\mathbf{R}_j^{(m)} + \rho_{jj}\mathbf{H}_{jj}^{(m)}\mathbf{Q}_j^{(m)}\mathbf{H}_{jj}^{(m)\dagger})^{-1} - (\mathbf{R}_j^{(m)})^{-1}\Big]\mathbf{H}_{jk}^{(m)} - u_k\mathbf{I}. \tag{23}$$

**Positive Semidefinite Cone Projection** The way to project a matrix, say $\mathbf{A}$, onto the positive semidefinite cone is as follows. First, we perform eigenvalue decomposition on $\mathbf{A}$ yielding $\mathbf{A} = \mathbf{U}_\mathbf{A}\text{Diag}\{\lambda_i : i = 1,\ldots,p\}\mathbf{U}_\mathbf{A}^\dagger$, where $\mathbf{U}_\mathbf{A}$ is the unitary matrix formed by the eigenvectors. Then, from Moreau decomposition [6], we have the PSD projection of $\mathbf{A}$ as follows:

$$\mathbf{A}_+ = \mathbf{U}_\mathbf{A}\text{Diag}\{\max\{\lambda_1, 0\},\ldots,\max\{\lambda_p, 0\}\}\mathbf{U}_\mathbf{A}^\dagger. \tag{24}$$

**Centralized and Distributed GP** We first propose a centralized gradient projection (CGP) algorithm to solve the power allocation problem in (19). The framework of CGP is illustrated in Algorithm 2.

CGP is a centralized algorithm since global information exchange is needed for gradient computation (23) in every iteration. In light of this, we devise a distributed gradient projection (DGP) algorithm where links update their input covariance matrices in an *iterative* manner: In each iteration, one link uses GP to update its input covariance matrix to maximize $I(\mathbf{Q})$ while all other links' input covariance matrices are held fixed. Thus, except at the initiation, only the exchange of information between link $n$ mod $K$ and link $n+1$ mod $K$ is needed in the $n^{th}$ iteration. The DGP method is illustrated in Algorithm 3.

**Algorithm 2** Centralized Gradient Projection

1. Choose the initial conditions $\mathbf{Q}_1^{(m)}(0)$, $\mathbf{Q}_2^{(m)}(0)$,..., $\mathbf{Q}_K^{(m)}(0)$.

   Let $n = 0$.

2. Use (23) to compute the gradients $\mathbf{G}_k^{(m)}(n)$ for $k = 1, 2, \ldots, K$.

3. Choose an appropriate step size $z_n$. Let $\hat{\mathbf{Q}}_k^{(m)}(n+1) = \mathbf{Q}_k^{(m)}(n)$

   $+z_n \mathbf{G}_k^{(m)}(n)$, for $k = 1, 2, \ldots, K$.

4. Let $\bar{\mathbf{Q}}_k^{(m)}(n+1)$ be the projection of $\hat{\mathbf{Q}}_k^{(m)}(n+1)$ onto the

   positive semidefinite cone.

5. Choose an appropriate step size $\alpha_n$. Let $\mathbf{Q}_k^{(m)}(n+1) =$

   $\bar{\mathbf{Q}}_k^{(m)}(n+1) + \alpha_n(\bar{\mathbf{Q}}_k^{(m)}(n+1) - \mathbf{Q}_k^{(m)}(n))$, $k = 1, 2, \ldots, K$.

6. If $\|\mathbf{Q}_k^{(m)}(n+1) - \mathbf{Q}_k^{(m)}(n)\| < \epsilon$, for $k = 1, 2, \ldots, K$, then stop;

   else let $n \leftarrow n + 1$ and go to Step 2.

**Algorithm 3** Distributed Gradient Projection

1. Choose the initial conditions $\mathbf{Q}_1^{(m)}(0)$, $\mathbf{Q}_2^{(m)}(0)$,...,$\mathbf{Q}_K^{(m)}(0)$.

   Let $n = 0$, $k = 1$.

2. Use (23) to compute the gradients $\mathbf{G}_k^{(m)}(n)$.

3. Choose an appropriate step size $z_n$. Let $\hat{\mathbf{Q}}_k^{(m)}(n+1) = \mathbf{Q}_k^{(m)}(n)$

   $+z_n \mathbf{G}_k^{(m)}(n)$.

4. Let $\bar{\mathbf{Q}}_k^{(m)}(n+1)$ be the projection of $\hat{\mathbf{Q}}_k^{(m)}(n+1)$ onto the

   positive semidefinite cone.

5. Choose an appropriate step size $\alpha_n$. Let $\mathbf{Q}_k^{(m)}(n+1) =$

   $\bar{\mathbf{Q}}_k^{(m)}(n+1) + \alpha_n(\bar{\mathbf{Q}}_k^{(m)}(n+1) - \mathbf{Q}_k^{(m)}(n))$.

6. If $\|\mathbf{Q}_k^{(m)}(n+1) - \mathbf{Q}_k^{(m)}(n)\| \leq \epsilon$, for $k = 1, 2, \ldots, K$, then stop;

   else let $n \leftarrow n + 1$, $k \leftarrow (k+1) \mathrm{mod} K + 1$ and go to Step 2.

## 4.3 Numerical Examples

We provide some numerical examples to illustrate the performance of GP and the subgradient algorithm. First, we examine how close the gap between the solution obtained by GP and the true optimal is. Since that the objective function of (19) is closer to concave for smaller INR's values (i.e., weaker interferences), it is expected that GP performs better under low INR regime.

We first consider the ergodic rate region in the dual domain, i.e., the $K$-dimensional subspace for the maximum achievable link rates of problem (19). Since we can only visualize 2-D graphs, we plot the ergodic rate region for a 2-link 2-antenna example with $\mathbf{u} = [0.8\ 0.8]^T$. By varying the weights $w_1$ and $w_2$ subject to $w_1 + w_2 = 1$, the whole rate region can be achieved [21]. The SNR's in this example are $\rho_{11} = \rho_{22} = 20$ dB. The INR's in this example are $\rho_{12} = 17$ dB and $\rho_{21} = 16$ dB. The ergodic rate region of this example is plotted in Figure 4. In our numerical study, we found

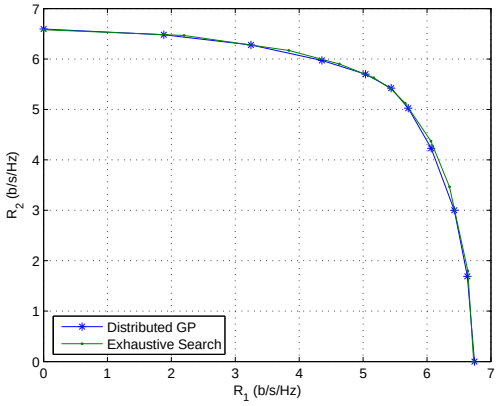


Figure 4: Ergodic rate region in the dual domain of a 2-link example with $\mathbf{u} = [0.8\ 0.8]^T$.

that CGP and DGP have essentially the same objective values for all weights. Thus, we only plot the curve of DGP in Figure 4. It is seen that even for this strong interference example, the gap between DGP and exhaustive search is still very close.

Next, we consider the optimal dual objective values of a 5-link 2-antenna equal-weight example with $\mathbf{u} = 2 \cdot \mathbf{1}$ under various SNR and INR values. The optimal dual objective values computed by DGP and exhaustive search are plotted in Figure 5. Again, we can observe that the gaps between DGP and exhaustive search is very close. For example, when SNR for each link is 10 dB and INR

for each link is 15 dB, i.e., the interference is much stronger than the intended signal, the dual objective values computed by DGP and exhaustive search are 11.17 and 12.13, respectively. In this extreme case, the loss is only 7.9%.

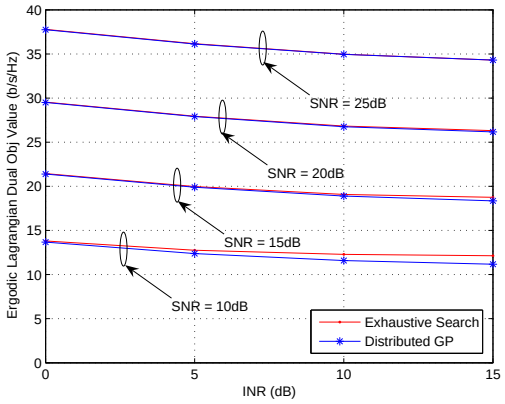


Figure 5: Lagrangian dual objective value of a 5-link example with $\mathbf{u} = 2 \cdot \mathbf{1}$.

We use a 5-link network example to see the convergence behavior of the subgradient algorithm. As shown in Figure 6, 5 links are uniformly distributed in a 2000m×2000m square region. Each node in the network is equipped with 4 antennas. The path-loss index is 4. The solid lines represent intended transmissions and the dotted lines represent interferences. The total available spectrum is divided into 128 subcarriers. The SNR and INR values of this example are given in Table 1. Figure 7

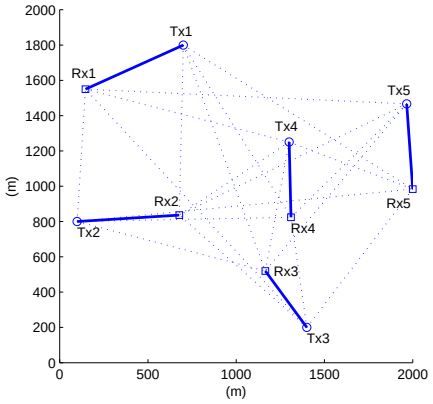


Figure 6: A 5-link MC-MIMO-based ad hoc network example.

shows the convergence behavior of the subgradient algorithm for minimizing the dual problem. The

Table 1: SNR- and INR-values for a 5-link MC-MIMO ad hoc network.

| | SNR (in dB) | INR (in dB) | | | | | |
|---|---|---|---|---|---|---|---|
| | | | $L_{1,r}$ | $L_{2,r}$ | $L_{3,r}$ | $L_{4,r}$ | $L_{5,r}$ |
| $L_1$ | 9.50 | $L_{1,t}$ | – | 1.49 | -4.51 | -1.58 | -6.58 |
| $L_2$ | 10.32 | $L_{2,t}$ | 5.84 | – | -0.84 | -2.48 | -10.38 |
| $L_3$ | 16.93 | $L_{3,t}$ | -9.75 | 1.52 | – | 8.87 | 1.09 |
| $L_4$ | 15.67 | $L_{4,t}$ | -2.20 | 5.90 | 6.01 | – | 5.88 |
| $L_5$ | 13.45 | $L_{5,t}$ | -9.56 | -5.41 | -2.88 | 2.34 | – |

step size selection is $s_n = 10/n$, $n = 1, 2, \ldots$. It can be seen that after approximately 600 iterations, the subgradient algorithm converges and yields a dual objective value of 5.68 b/s/Hz. We recover the corresponding primal feasible solution (by the LP approach [2]) and find that the primal objective value is 5.65 b/s/Hz, i.e., the duality gap is less than 1%. Comparing with the starting point before optimization (equal power allocation to all subcarriers and antennas at each node), which yields an objective value of 3.83 b/s/Hz, we can see the performance gain after optimization is significant (47.5%).

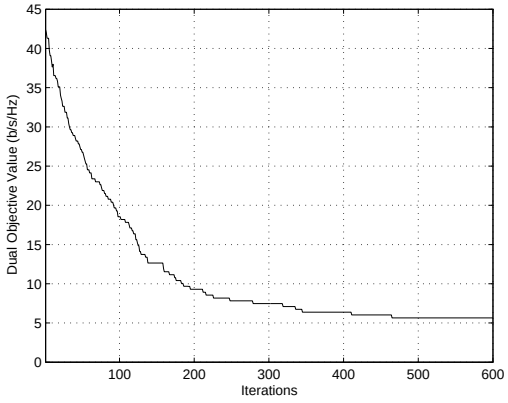


Figure 7: Convergence behavior of the subgradient algorithm.

# 5 Online Adaptive Algorithm

We should point out that the solution procedure proposed in the last section is an off-line algorithm because it requires full CDI knowledge to compute the subgradients in (20). In this section, we will use the off-line algorithm as a baseline and devise an *online adaptive* algorithm by subgradient stochastic approximation (SSA). The basic idea of SSA is that, instead of computing the exact

subgradient based on full CDI, we use the instantaneous CSI feedback to compute an approximation of the true subgradient, thus eliminating the requirement of full CDI knowledge. By appealing to recent results in Martingale theory [8], we show that the stochastic iterates can adapt to *unknown* underlying fading distributions on the fly and lend itself to an online algorithm. In this section, we assume that the feedback of CSI is error free. The impact of CSI error will be discussed in the next section.

## 5.1 Subgradient Stochastic Approximation

The way to eliminate the requirement of full CDI knowledge in (20) is to use the current channel realization in each iteration to compute an approximation of the true subgradient. The stochastic subgradient during the $n^{th}$ iteration is computed as follows:

$$\hat{d}_{u_k}(n) = 1 - \sum_{m=1}^{M} \mathrm{Tr}(\mathbf{Q}_k^{(m)*}), \quad \forall k, \tag{25}$$

and the updates on the Lagrangian dual variables can be computed as

$$\hat{u}_k(n+1) = \left[\hat{u}_k(n) - s_n \hat{d}_{u_k}(n)\right]_+^U, \ \forall k. \tag{26}$$

The complete algorithm is summarized in Algorithm 4. Comparing SSA with Algorithm 1, we can see that the only difference is that expectation computation is not needed and each subgradient is only based on current channel realization. The basic structure of SSA remains the same as that of the off-line subgradient algorithm, which means that SSA can also be implemented in a distributed fashion.

## 5.2 Convergence Theorem

In this section, we prove that SSA converges with probability one to the same optimal solution[2] obtained by the off-line algorithms. Part of the proof needs a recent result on Martingale theory [8], which is stated as follows.

[2]For brevity, we call the near-optimal solution obtained by the off-line algorithms "optimal solution" from now on.

**Algorithm 4** Subgradient Stochastic Approximation

1. Choose the initial starting points $\hat{\mathbf{u}}(0)$. Let $n = 0$.
2. Compute $\mathbf{Q}^*(n)$ using Algorithm 2 or Algorithm 3.
3. In the $n^{th}$ iteration, choose an appropriate step size $s_n$. Update subgradients $\hat{d}_{\mathbf{u}_k}(n)$ using (25) with $\mathbf{Q}^*(n)$.
4. Update dual variables $\hat{\mathbf{u}}(n)$ using (26) with $\hat{d}_{u_k}(n)$.
5. If $\|\hat{\mathbf{u}}_k(n+1) - \hat{\mathbf{u}}_k(n)\| < \epsilon$, then return $\mathbf{Q}^*(n)$ as the final optimal solution and stop. Otherwise, let $n \leftarrow n+1$ and go to Step 2.

**Lemma 1.** *Let $\{X_n\}$ be an $\mathbb{R}^r$-valued stochastic process, and $V(\cdot)$ be a real-valued non-negative function in $\mathbb{R}^r$. Suppose that $\{Y_n\}$ is a sequence of random variables satisfying that $\sum_n |Y_n| < \infty$ with probability one. Let $\{\mathcal{F}_n\}$ be a sequence of $\sigma$-algebra generated by $\{X_i, Y_i, i \leq n\}$. Suppose that there exists a compact set $B \subset \mathbb{R}^r$ such that for all $n$,*

$$\mathbb{E}_n[V(X_{n+1})] - V(X_n) \leq -s_n\delta + Y_n, \text{ for } X_n \notin B,$$

*where $s_n$ satisfies $s_n \to 0$, $\sum_{n=1}^{\infty} s_n \to \infty$, $\sum_{n=1}^{\infty} s_n^2 < \infty$, and $\delta$ is a positive constant. Then, set $B$ is recurrent for $\{X_n\}$, i.e., $X_n \in B$ for infinitely large $n$ with probability one.*

Now we state the convergence theorem as follows.

**Theorem 5.** *If the step size selection of $s_n$ satisfies (22), the stochastic iterations in (26) converge to the optimal solution $\mathbf{u}^*$ of (15) with probability one.*

*Proof.* To prove the convergence theorem, we first examine the structure of the stochastic subgradients during the $n^{th}$ iteration. Let $\{\mathcal{F}_n\}$ be the sequence of $\sigma-$algebra generated by $\{\hat{\mathbf{u}}(i) : i = 1, \ldots, n\}$. It can be readily verified that the stochastic subgradient $\hat{d}_{u_k}(n)$ can be decomposed as $\hat{d}_{u_k}(n) = d_{u_k}(n) + \xi_{u_k}(n)$, where $\xi_{u_k}(n)$ is a martingale noise term and defined as

$$\begin{aligned}\xi_{u_k}(n) &\triangleq \hat{d}_{u_k}(n) - \mathbb{E}_{\mathbf{H}}^{(n)}\left[\hat{d}_{u_k}(n)\right] \\ &= \mathbb{E}_{\mathbf{H}}^{(n)}\left[\sum_{m=1}^{M} \mathrm{Tr}(\mathbf{Q}_k^{(m)*}(n))\right] - \sum_{m=1}^{M} \mathrm{Tr}(\mathbf{Q}_k^{(m)*}(n)). \end{aligned} \tag{27}$$

In (27), $\mathbb{E}_{\mathbf{H}}^{(n)}[\cdot]$ denotes the conditional expectation $\mathbb{E}_{\mathbf{H}}[\cdot|\mathcal{F}_n]$. Then, the stochastic dual updates can be written as

$$\hat{u}_k(n+1) = \hat{u}_k(n) - s_n\left[d_{u_k}(n) + \xi_{u_k}(n)\right] + z_{u_k}(n), \quad \forall k,$$

where $z_{u_k}(n)$ is the correction term that projects the stochastic subgradient back to the non-negative orthant. Then, for $\hat{u}_k(n)$, we have

$$|\hat{u}_k(n+1) - u_k^*|^2 \leq |\hat{u}_k(n) - u_k^*|^2 - 2s_n(\hat{u}_k(n) - u_k^*) \times$$
$$\left[d_{u_k}(n) + \xi_{u_k}(n)\right] + s_n^2\left[d_{u_k}(n) + \xi_{u_k}(n)\right]^2,$$

where the inequality holds because the correction term $\xi_{u_k}(n)$ is non-expansive [3]. Since that $\Theta(\mathbf{u})$ is twice-differentiable with respect to $\mathbf{u}$, $I(\mathbf{Q}^{(n)})$ is bounded. Also, it is evident that $|\hat{u}_k(n) - u_k^*|$ is bounded and $\mathbb{E}_{\mathbf{H}}^{(n)}[\xi_{u_k}] = 0$. From the iteration update in (26), it can be concluded that $|\xi_{u_k}(n)|$ is bounded. From the iteration update in (21), it can be concluded that $|d_{u_k}(n)|$ is bounded. Thus, in vector form, we have

$$\mathbb{E}_{\mathbf{H}}^{(n)}\left[\|\hat{\mathbf{u}}(n+1) - \mathbf{u}^*\|^2\right] \leq \|\hat{\mathbf{u}}^{(n)} - \mathbf{u}^*\|^2 -$$
$$2s_n\langle\hat{\mathbf{u}}(n) - \mathbf{u}^*, \mathbf{d}_{\mathbf{u}}(n)\rangle - 2s_n\langle\hat{\mathbf{u}}(n) - \mathbf{u}^*, \mathbb{E}_{\mathbf{H}}^{(n)}[\xi_{\mathbf{u}}(n)]\rangle +$$
$$O(s_n^2) = \|\hat{\mathbf{u}}(n) - \mathbf{u}^*\|^2 - 2s_n\langle\hat{\mathbf{u}}(n) - \mathbf{u}^*, \mathbf{d}_{\mathbf{u}}(n)\rangle + O(s_n^2).$$

Define a Lyapunov function $V(\hat{\mathbf{u}}) \triangleq \|\hat{\mathbf{u}} - \mathbf{u}^*\|^2$ and define a ball centered at $\mathbf{u}^*$ as $\mathbb{B}_\epsilon = \{\hat{\mathbf{u}} : V(\hat{\mathbf{u}}) \leq \epsilon\}$ for some given $\epsilon > 0$. From Lemma 1, we can see that the proof of the recurrence of $\{\hat{\mathbf{u}}(n) : n = 1, 2, \ldots\}$ boils down to showing that

$$\langle\hat{\mathbf{u}}(n) - \mathbf{u}^*, \mathbf{d}_{\mathbf{u}}(n)\rangle \geq 0. \tag{28}$$

It is not difficult to see (28) holds because for $\hat{\mathbf{u}}(n)$, we have

$$\langle\hat{\mathbf{u}}(n) - \mathbf{u}^*, \mathbf{d}_{\mathbf{u}}(n)\rangle \geq \Theta(\hat{\mathbf{u}}(n)) - \Theta(\mathbf{u}^*) \geq 0,$$

where the first inequality follows from the convexity of $\Theta(\mathbf{u})$ and the second inequality follows from the fact that $\mathbf{u}^*$ is the global minimizer of $\Theta(\mathbf{u})$. This shows that there exists $\delta_\epsilon > 0$ such that

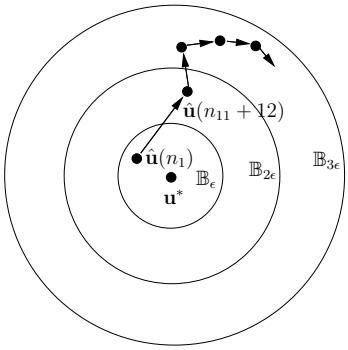


Figure 8: The sketch of the basic idea in the proof of Theorem 5.

$\langle \hat{\mathbf{u}}(n) - \mathbf{u}^*, \mathbf{d}_{\mathbf{u}}(n) \rangle > \delta_\epsilon$ when $\hat{\mathbf{u}} \notin \mathbb{B}_\epsilon$. It then follows from Lemma 1 that $\{\hat{\mathbf{u}}(n) : n = 1, 2, \ldots\}$ returns to $\mathbb{B}_\epsilon$ infinitely often with probability one.

Having shown the recurrence of $\{\hat{\mathbf{u}}(n) : n = 1, 2, \ldots\}$, our next step of proving convergence is to show that for any small $\epsilon > 0$, there exists $n_1$ such that for $n \geq n_1$, even if the martingale noise terms $\xi_{\mathbf{u}}(n)$ drive the $\hat{\mathbf{u}}(n)$ away from $\mathbb{B}_\epsilon$, the trajectory of $\{\hat{\mathbf{u}}(n) : n > n_1\}$ is still bounded by the contraction region $\mathbb{B}_{3\epsilon}$ almost surely. We sketch the basic idea of the proof in Figure 8.

Toward this end, we first show that if $n_1$ is sufficiently large and $\hat{\mathbf{u}}(n_1 + 1)$ moves out of $\mathbb{B}_\epsilon$, then $\hat{\mathbf{u}}(n_1 + 1)$ resides in $\mathbb{B}_{2\epsilon}$ with probability one. This fact holds because by Chebyshev's inequality, we have

$$\Pr\left(s_{n_1} \|\xi_{\mathbf{u}}(n_1)\| > \epsilon\right) \leq \frac{s_{n_1}^2 \mathbb{E}_{\mathbf{H}}^{(n)}[\|\xi_{\mathbf{u}}(n_1)\|^2]}{\epsilon^2}. \tag{29}$$

Since $s_{n_1}^2 \to 0$ when $n_1$ is large, the change in one step is almost surely no greater than $\epsilon$. Thus, $\hat{\mathbf{u}}(n_1 + 1)$ resides in $\mathbb{B}_{2\epsilon}$ with probability one.

Next, we show that if $\hat{\mathbf{u}}(n_1 + 1) \in \mathbb{B}_{2\epsilon} \cap \bar{\mathbb{B}}_\epsilon$, where $\bar{\mathbb{B}}_\epsilon$ denotes the complement of $\mathbb{B}_\epsilon$, then for all $n > n_1 + 1$, $\hat{\mathbf{u}}(n) \in \mathbb{B}_{3\epsilon}$ with probability one. That is, the trajectory of the series starting from $\hat{\mathbf{u}}(n_1 + 1)$ resides in $\mathbb{B}_{3\epsilon}$ almost surely. By using the martingale inequality [8, Eq. (1.4)], we have that, for $m \geq n_1 + 1$,

$$\Pr\left\{\sup\left|\sum_{n=n_1+1}^{m} s_n \|\xi_{\mathbf{u}}^{(n)}\|\right| \geq \epsilon\right\} \leq \frac{\bar{\mathbb{E}}_{\mathbf{H}}^{(n)}[\|\xi_{\mathbf{u}}^{(n)}\|^2]}{\epsilon^2} \sum_{n=n_1+1}^{\infty} s_n^2,$$

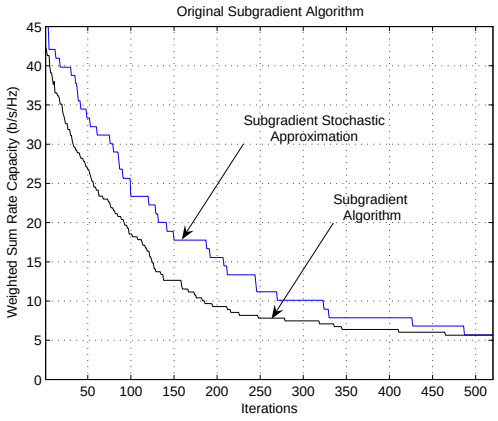


Figure 9: Convergence behavior of SSA.

where $\bar{\mathbb{E}}_{\mathbf{H}}^{(n)}[\|\xi_{\mathbf{u}}(n)\|^2] \triangleq \limsup_n \mathbb{E}_{\mathbf{H}}^{(n)}[\|\xi_{\mathbf{u}}(n)\|^2]$. It then follows from the conditions of $s_n$ and the boundedness of $\|\xi_{\mathbf{u}}(n)\|$ that

$$\lim_{m\to\infty} \Pr\left\{\sup\left|\sum_{n=n_1+1}^{m} s_n\|\xi_{\mathbf{u}}(n)\|\right| \geq \epsilon\right\} = 0, \ \forall \epsilon > 0. \tag{30}$$

That is to say, the accumulated distance deviating from $\hat{\mathbf{u}}(n_1+1)$ is almost surely less than $\epsilon$. As a result, the trajectory resides in $\mathbb{B}_{3\epsilon}$ with probability one. Finally, since $\epsilon$ can be arbitrary small, it then follows that $\{\hat{\mathbf{u}}(n) : n = 1, 2, \ldots\}$ converge to $\mathbf{u}^*$ with probability one. □

## 5.3 Convergence Rate Analysis

The convergence rate of SSA is influenced by the condition number of the negative Hessian matrix of the objective function (i.e., the ratio of the largest to smallest eigenvalues). A smaller condition number implies faster convergence rate. It can be further shown that the limiting process of the convergence is a Gaussian diffusion process if the optimal solution is an interior point of the feasible region; or a stationary reflected linear diffusion process if the optimal solution is achieved on the boundary of the feasible region. Due to space limitation, we omit the proof in this paper.

### 5.4 Numerical Example

We use the same 5-link MC-MIMO ad hoc network example in the last section to illustrate the convergence behavior of the SSA algorithm. For comparison, we plot the convergence processes of SSA and the subgradient algorithm (using the same step size selection) in Figure 9. It can be seen that after approximately 750 iterations, SSA converge to the same solution obtained by the subgradient algorithm. Although the number of iterations is slightly larger than that of the subgradient algorithm, the running time of SSA is much shorter because SSA does not require computing expectations over the fading distribution.

## 6 Subgradient Stochastic Approximation with CSI Error

In the last section, we assume that CSI, i.e., $\mathbf{H}$, is error free. In practice, the knowledge of $\mathbf{H}$ is acquired by sending a number of training symbols from the transmitting nodes. Then, the receiving nodes estimate their channel gains and send back their estimations through the reverse links. As a result, the error free assumption may not hold if the training and estimation phase is not long enough or the reverse links are noisy. In this section, we will study the impact of CSI error on the convergence of SSA.

Denote the channel gain matrices with estimation error as $\tilde{\mathbf{H}} = [\tilde{\mathbf{H}}_k^{(m)} : \forall k, \forall m]$. Let $\tilde{\mathbf{Q}}_k^{(m)*}$ be the resultant power allocation decisions when $\mathbf{H}_k^{(m)}$ is replaced by $\tilde{\mathbf{H}}_k^{(m)}$. The Lagrangian dual variables are updated as follows:

$$\tilde{u}_k(n+1) = \left[\tilde{u}_k(n) - s_n \tilde{d}_{u_k}(n)\right]_+, \tag{31}$$

where the stochastic subgradients with CSI error can be written as:

$$\tilde{d}_{u_k}(n) = 1 - \sum_{m=1}^{M} \mathrm{Tr}(\tilde{\mathbf{Q}}_k^{(m)*}(n)), \quad \forall k, \tag{32}$$

It can be readily verified that these stochastic subgradients can be decomposed as follows:

$$\tilde{d}_{u_k}(n) = d_{u_k}(n) + \mu_{u_k}(n) + \zeta_{u_k}(n), \tag{33}$$

where $\mu_{u_k}(n)$ is the biased estimation error of the stochastic subgradients, and is defined as

$$\mu_{u_k}(n) \triangleq \mathbb{E}_{\mathbf{H}}^{(n)}\left[\tilde{d}_{u_k}(n)\right] - d_{u_k}(n)$$

$$= \sum_{m=1}^{M} \mathbb{E}_{\mathbf{H}}^{(n)}\left[\mathrm{Tr}(\tilde{\mathbf{Q}}_k^{(m)*}(n))\right] - \mathrm{Tr}(\mathbf{Q}_k^{(m)*}(n)), \tag{34}$$

and $\zeta_{u_k}(n)$ is a martingale noise term, which is defined as

$$\zeta_{u_k}(n) \triangleq \tilde{d}_{u_k}(n) - \mathbb{E}_{\mathbf{H}}^{(n)}\left[\tilde{d}_{u_k}(n)\right]$$

$$= \mathbb{E}_{\mathbf{H}}^{(n)}\left[\mathrm{Tr}(\tilde{\mathbf{Q}}_k^{(m)*}(n))\right] - \mathrm{Tr}(\tilde{\mathbf{Q}}_k^{(m)*}(n)), \tag{35}$$

This decomposition leads naturally to the following two cases: $\tilde{d}_{u_k}(n)$ is asymptotically biased or unbiased.

## 6.1 Unbiased Case

Under the unbiased case, $\sum_{n\to\infty} s_n|\mathbb{E}_{\mathbf{H}}^{(n)}\left[\mu_{u_k}(n)\right]| < \infty$ with probability one for all $k$, so that the biased error terms can be asymptotically driven to zero by the diminishing step sizes. For this case, we have the following theorem.

**Theorem 6.** *If $\sum_{n\to\infty} s_n|\mathbb{E}_{\mathbf{H}}^{(n)}\left[\mu_{u_k}(n)\right]| < \infty$ with probability one for all $k$, and if the step size selection of $s_n$ satisfies (22), the stochastic iterations in (31) converge to the optimal solution $\mathbf{u}^*$ of (15) with probability one.*

*Proof.* Substitute (33) into (31), the stochastic dual updates can be written as $\tilde{u}_k(n+1) = \tilde{u}_k(n) - s_n[d_{u_k}(n) + \mu_{u_k}(n) + \zeta_{u_k}(n)] + z_{u_k}(n)$, where $z_{u_k}(n)$ is the correction term that projects the stochastic subgradients back to the non-negative orthant. For $\tilde{u_k}(n)$, we have that $|\tilde{u}_k(n+1) - u^*|^2 \leq |\tilde{u}_k(n) - u_k^*|^2 - 2s_n(\tilde{u}_k(n) - u_k^*)[d_{u_k}(n) + \mu_{u_k}(n) + \zeta_{u_k}(n)] + s_n^2[d_{u_k}(n) + \mu_{u_k}(n) + \zeta_{u_k}(n)]^2$,

Applying the similar boundedness arguments in the recurrence proof of Theorem 5 and noting

that $\mathbb{E}_{\mathbf{H}}^{(n)}[\zeta_{u_k}(n)] = 0$ for all $k$, we have

$$\mathbb{E}_{\mathbf{H}}^{(n)}\left[|\tilde{u}_k(n+1) - u_k^*|^2\right] \leq |\tilde{u}_k(n) - u_k^*|^2 - 2s_n(\tilde{u}_k(n) - u_k^*)d_{u_k}(n) + O(s_n|\mu_{u_k}(n)|) + O(s_n^2). \tag{36}$$

Next, define a Lyapunov function $V(\tilde{\mathbf{u}}) \triangleq \|\tilde{\mathbf{u}} - \mathbf{u}^*\|^2$. The remainder of the recurrence proof and contraction proof follows exactly the same line as that in the proof of Theorem 5 and is omitted here for brevity. □

## 6.2 Biased Case

In the biased case, $s_n|\mathbb{E}_{\mathbf{H}}^{(n)}[\mu_{u_k}(n)]|$ is not summable. Thus, we cannot expect SSA still converges to the same solution almost surely. Let $\bar{\mu}_{u_k} \triangleq \limsup_n \mathbb{E}_{\mathbf{H}}^{(n)}[\mu_{u_k}(n)]$. Also, define a neighborhood $\mathbb{N}_\phi$ around the optimal solution $\mathbf{u}^*$ as follows:

$$\mathbb{N}_\phi \triangleq \left\{ \mathbf{u} \;\middle|\; \begin{array}{l} \bar{\mu}_{u_k} \geq \phi |d_{u_k}(\mathbf{u})|, \text{ for some } k \\ 0 \leq \phi < 1. \end{array} \right\}, \tag{37}$$

where $d_{u_k}(\mathbf{u})$ represents the subgradient with respect to $u_k$ at point $\mathbf{u}$. In essence, $\mathbb{N}_\phi$ defines a neighborhood around the optimal point $\mathbf{u}^*$: since $d_{u_k}(\mathbf{u})$ are continuous at $\mathbf{u}^*$ and by KKT conditions we have $d_{u_k}(\mathbf{u}^*) = 0$ for all $k$; also, (37) requires that point in $\mathbb{N}_\phi$ satisfy $|d_{u_k}(u)| \leq \frac{\bar{\mu}_{u_k}}{\phi}$, which means that $\mathbb{N}_\phi$ is a neighborhood around the optimal point. With these definitions, we now present the recurrence theorem for the biased case.

**Theorem 7.** *If $\mu_{u_k}(n)$ is asymptotically upper bounded for all $k$ and if the step size selection satisfies (22), then the iterates $\{\tilde{\mathbf{u}}(n) : n = 1, 2, \ldots\}$, generated by (31) returns to $\mathbb{N}_\phi$ with probability one.*

*Proof.* Define a Lyapunov function as $V(\tilde{\mathbf{u}}) \triangleq \|\tilde{\mathbf{u}} - \mathbf{u}^*\|^2$. Consider a neighborhood around the optimal solution $\mathbf{u}^*$, denoted by $\mathbb{N}_\phi \cup \mathbb{B}_\epsilon$, where $\mathbb{B}_\epsilon = \{\tilde{\mathbf{u}} : V(\tilde{\mathbf{u}}) \leq \epsilon\}$. Recall that for $\tilde{\mathbf{u}}(n)$, we have that $|\tilde{u_k}(n+1) - u_k^*|^2 \leq |\tilde{u}_k(n) - u_k^*|^2 - 2s_n(\tilde{u}_k(n) - u_k^*)[d_{u_k}(n) + \mu_{u_k}(n) + \zeta_{u_k}(n)] + s_n^2[d_{u_k}(n) + \mu_{u_k}(n) + \zeta_{u_k}(n)]^2$.

Note that for $\mathbf{u} \notin \mathbb{N}_\phi \cup \mathbb{B}_\epsilon$, $\limsup_{n\to\infty} \mathbb{E}_{\mathbf{H}}^{(n)}[\mu_{u_k}(n)] < \phi |d_{u_k}(\mathbf{u})|$. Applying the similar boundedness arguments in the recurrence proof of Theorem 5 and based on the fact that $\mathbb{E}_{\mathbf{H}}^{(n)}[\zeta_{u_k}(n)] = 0$, we have

$$\mathbb{E}_{\mathbf{H}}^{(n)}\left[|\tilde{u}_k(n+1) - u_k^*|^2\right] \leq |\tilde{u}_k(n) - u^*|^2 - 2s_n(1-\phi)(\tilde{u}_k(n) - u_k^*)d_{u_k}(n) + O(s_n^2). \tag{38}$$

With $1 - \phi > 0$, the recurrence of $\{\tilde{\mathbf{u}}(n) : n = 1, 2 \ldots\}$ to $\mathbb{N}_\phi \cup \mathbb{B}_\epsilon$ follows exactly the same line as that in the proof of Theorem 5. Next, by letting $\epsilon \to 0$, we have that $\{\tilde{\mathbf{u}}(n) : n = 1, 2 \ldots\}$ is recurrent to $\mathbb{N}_\phi$. □

It is also interesting to study the size of the neighborhood under the biased case. The following proposition characterizes the relationship between the size of the neighborhood and the channel estimation error.

**Theorem 8.** *The neighborhood around the optimal solution* $\mathbf{u}^*$ *in the biased case is outter bounded by points* $\mathbf{u}$ *that satisfy* $|d_{u_k}(\mathbf{u})| = \bar{\mu}_{u_k}$*, for all* $k$*.*

*Proof.* Recall that any point $\tilde{\mathbf{u}} \in \mathbb{N}_\phi$ satisfies $\phi |d_{u_k}(\mathbf{u})| \leq \bar{\mu}_{u_k}$ for some $k$ and $\phi \in [0, 1)$. It then follows that $\phi < 1 \leq \frac{\bar{\mu}_{u_k}}{|d_{u_k}(\mathbf{u})|}$. On the other hand, as $n \to \infty$, $|d_{u_k}(\mathbf{u})|$ decreases because the iterates driven by $d_{u_k}(\mathbf{u})$ move toward the optimal point $\mathbf{u}^*$. Thus, as the value of $|d_{u_k}(\mathbf{u})|$ decreases, $\phi$ asymptotically approaches 1. As a result, the iterates generated by SSA are outter bounded by points $\mathbf{u}$ that satisfy $|d_{u_k}(\mathbf{u})| = \bar{\mu}_{u_k}$. □

## 6.3 Numerical Examples

We use the same 5-link network example to illustrate the impact of CSI error on SSA. For the unbiased case, we use $\mu_{u_k}(n) \sim \frac{1}{n}N(0, 0.01)$ and $\mu_{u_k}(n) \sim \frac{1}{n}N(0, 0.1)$ to approximate the biased error terms, where $N(\nu, \sigma^2)$ represents the normal distribution with mean $\nu$ and a variance of $\sigma^2$. In the biased case, we let $\mu_{u_k}(n)$ be uniformly distributed in [0 0.01] and [0 0.1], respectively. The convergence processes of both cases are plotted in Figure 10.

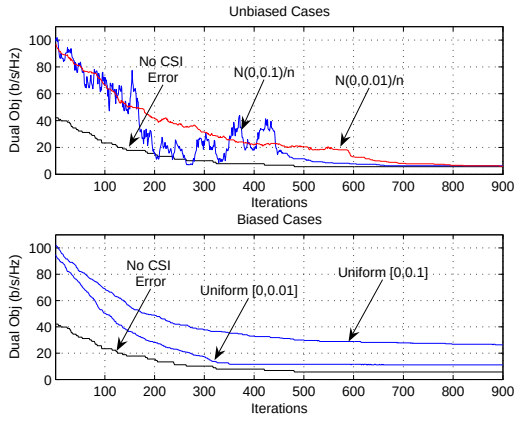


Figure 10: The convergence behavior of SSA with CSI error.

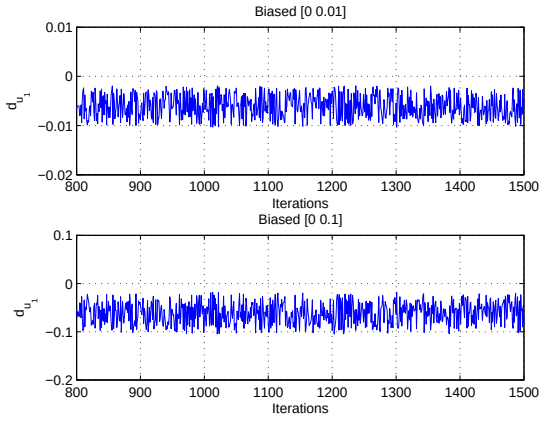


Figure 11: Evolution of $d_{u_1}(\mathbf{u})$ under the biased case.

It can be seen in the unbiased case that, since the biased error are asymptotically diminishing, SSA still converges to the same optimal solution, affirming our analysis in Theorem 6. For the biased case, SSA returns to a neighborhood of the optimal solution, which corroborates our analysis in Theorem 7. The convergence process of the unbiased cases exhibit a lot of fluctuations. This is because $\mu_{\mathbf{u}_k}(n)$ are normally distributed so that they could be either positive or negative, which increases or reduces $|d_{u_k}(n)|$. In contrast, the convergence process of the biased case exhibits little fluctuations because $\mu_{\mathbf{u}_k}(n)$ only increases $|d_{u_k}(n)|$. Comparing with the error-free case, we can see that CSI error slow down the convergence for both unbiased and biased cases.

For both biased examples, the evolution of $d_{u_1}(\mathbf{u})$ is plotted in Figure 11. It can be observed

that the iterates stay in a neighborhood of 0 infinitely often, and are lower bounded by $-0.01$ and $-0.1$, respectively. This corroborates our conclusion in Theorem 8.

# 7 Conclusion

In this paper, we investigated power spectrum optimization for multi-carrier MIMO (MC-MIMO) ad hoc networks. We made advance in general non-convex optimization theory by showing that a certain class of non-convex optimization problems have zero duality gap if they satisfy the "concave perturbation" condition. Further, we showed that the concave perturbation condition is satisfied for the power spectrum optimization problem of MC-MIMO ad hoc networks. This result led to efficient centralized and distributed (off-line) algorithms in the dual domain. Next, by using stochastic approximation technique, we devised an online adaptive algorithm that does not require full knowledge of channel distribution information (CDI). We showed that our subgradient stochastic approximation (SSA) algorithm converges with probability one to the same solution obtained by the off-line algorithm. Further, we explored the impact of CSI error on convergence performance of SSA. We showed that SSA still converges with probability one to the same solution if stochastic subgradients are asymptotically unbiased, or is recurrent to some neighborhood of the solution obtained by the off-line algorithm if the stochastic subgradients are biased.

# References


[1] M. Andrews and L. Zhang. Scheduling algorithms for multi-carrier wireless data systems. In *Proc. ACM Mobicom*, pages 3–14, Montréal, Québec, Canada, Sept.9-14, 2007.

[2] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty. *Nonlinear Programming: Theory and Algorithms*. John Wiley & Sons Inc., New York, NY, 3 edition, 2006.

[3] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmon, MA, 1995.

[4] H. Boche and M. Wiczanowski. Stability optimal transmission policy for the multiple antenna multiple access channel in the geometric view. *EURASIP Signal Process. J. (Special Issue on Advances in Signal Processing-Assisted Cross-Layer Designs)*, 86(8):1815–1833, Aug. 2006.

[5] A. Goldsmith, S. A. Jafar, N. Jindal, and S. Vishwanath. Capacity limits of MIMO channels. *IEEE J. Sel. Areas Commun.*, 21(1):684–702, June 2003.

[6] J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis*. Springer-Verlag, Berlin, 2001.

[7] M. Jiang and L. Hanzo. Multiuser MIMO-OFDM for next generation wireless systems. *Proc. IEEE*, 95(7):1430–1469, July 2007.

[8] H. J. Kushner and G. G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, Net York, 2003.

[9] P. Kyasanur and N. H. Vaidya. Capacity of multi-channel wireless networks: Impact of number of channels and interfaces. In *Proc. ACM Mobicom*, pages 43–57, Cologne, Germany, Aug.28–Sept.2, 2005.

[10] J. Liu and Y. T. Hou. Optimal downlink power allocation and scheduling for MIMO-based WiMAX access networks. *IEEE J. Sel. Areas Commun.*, submitted for publication, [Online]. Available: http://filebox.vt.edu/users/kevinlau/publications.

[11] Q. Liu, X. Wang, and G. B. Giannakis. A cross-layer scheduling algorithm with QoS support in wireless networks. *IEEE Trans. Veh. Technol.*, 55(3):839–847, May 2006.

[12] J. R. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Economics*. Wiley, New York, 1999.

[13] M. J. Neely, E. Modiano, and C. E. Rohrs. Power allocation and routing in multibeam satellites with time-varying channels. *IEEE/ACM Trans. Netw.*, 11(2):138–152, Feb. 2003.

[14] D. P. Paloma. Convex primal decomposition for multicarrier linear MIMO transceivers. *IEEE Trans. Signal Process.*, 53(12):4661–4674, Dec. 2005.

[15] D. P. Paloma, J. M. Cioffi, and M. A. Lagunas. Joint Tx-Rx beamforming design for multicarrier MIMO channels: A unified framework for convex optimization. *IEEE Trans. Signal Process.*, 51(9):2381–2401, Sept. 2003.

[16] T. Pande, D. J. Love, and J. V. Krogmeier. Reduced feedback MIMO-OFDM precoding and antenna selection. *IEEE Trans. Signal Process.*, 55(5):2284–2293, May 2007.

[17] H. D. Sherali and W. P. Adams. *A Reformulation-Linearization-Technique for Solving Discrete and Continuous Nonconvex Problems.* Kluwer Academic Publishing, Boston, MA, 1999.

[18] C. Shin, R. W. Heath, and E. J. Powers. Blind channel estimation for MIMO-OFDM systems. *IEEE Trans. Veh. Technol.*, 56(2):670–685, Mar. 2007.

[19] G. Song and Y. Li. Cross-layer optimization for OFDM wireless networks – Part I: Theoretical framework. *IEEE Trans. Wireless Commun.*, 4(2):614–624, Mar. 2005.

[20] G. Song and Y. Li. Cross-layer optimization for OFDM wireless networks – Part II: Algorithm development. *IEEE Trans. Wireless Commun.*, 4(2):625–634, Mar. 2005.

[21] D. Tse and S. Hanly. Multi-access fading channels: Part I: Polymatroid structure, optimal resource allocation and throughput capacities. *IEEE Trans. Inf. Theory*, 44(7):2796–2815, Nov. 1998.

[22] S. Visuri and H. Bölcskei. Multiple-access strategies for frequency-selective MIMO channels. *IEEE Trans. Inf. Theory*, 52(9):3980–3993, Sept. 2006.

[23] P. Viswanath, D. N. C. Tse, and R. Laroia. Opportunistic beamforming using dumb antennas. *IEEE Trans. Inf. Theory*, 48(6):1277–1294, June 2002.

[24] C. Y. Wong, R. S. Cheng, K. B. Letaief, and R. D. Murch. Multiuser OFDM with adaptive subcarrier, bit, and power allocation. *IEEE J. Sel. Areas Commun.*, 17(10):1747–1758, Oct. 1999.

[25] E. M. Yeh and A. S. Cohen. Information theory, queueing, and resource allocation in multi-user fading communications. In *Proc. Conf. Inf. Sci. Syst.*, Princton, NJ, Mar. 2004.

[26] Y. Zeng, A. R. Leyman, and T.-S. Ng. Joint semiblind frequency offset and channel estimation for multiuser MIMO-OFDM uplink. *IEEE Trans. Commun.*, 55(12):2270–2278, Dec. 2006.

# A Proof of Theorem 1

The first step of our proof is to show that $\bar{\mathbf{u}} \geq \mathbf{0}$ must hold. By contradiction, suppose there exists a component of $\bar{\mathbf{u}}$, denoted by $\bar{u}_p$, such that $\bar{u}_p < 0$, then we can choose $\tilde{\mathbf{y}}$ where $\tilde{y}_i = 0$ for $i \neq p$, and $\tilde{y}_p > 0$. Since $\tilde{\mathbf{y}} \geq \mathbf{0}$, $\theta(\tilde{\mathbf{y}})$ is apparently a *relaxation* of $\theta(\mathbf{0})$. It then follows that $\theta(\mathbf{0}) \leq \theta(\tilde{\mathbf{y}})$. On the other hand, from (10), we have $\theta(\tilde{\mathbf{y}}) \leq \theta(\mathbf{0}) + \bar{u}_p\tilde{y}_p$. Thus, we have $\theta(\mathbf{0}) \leq \theta(\mathbf{0}) + \bar{u}_p\tilde{y}_p$, which implies that $\bar{u}_p \geq 0$ since $\tilde{y}_p > 0$. A contradiction.

The second step is to show that $\bar{\mathbf{u}}^T(\sum_{m=1}^{M} \mathbf{g}_m(\bar{\mathbf{x}}_m)) = 0$ must hold. By fixing $\mathbf{y} = \bar{\mathbf{y}} = [\sum_{m=1}^{M} \mathbf{g}_m(\bar{\mathbf{x}}_m)]^T$ in (1), we obtain a *restriction* of the primal problem in (6) since $\sum_{m=1}^{M} \mathbf{g}_m(\bar{\mathbf{x}}_m) \leq \mathbf{0}$. But on the other hand, since $\bar{\mathbf{X}}$ is feasible to the perturbation function with $\bar{\mathbf{y}}$ fixed as such, and $\bar{\mathbf{X}}$ solves the primal problem, we have $\theta(\bar{\mathbf{y}}) = \theta(\mathbf{0})$. By (10), this in turn means that $\theta(\mathbf{0}) \leq \theta(\mathbf{0}) + \bar{\mathbf{u}}^T(\bar{\mathbf{y}} - \mathbf{0}) = \theta(\mathbf{0}) + \bar{\mathbf{u}}^T(\sum_{m=1}^{M} \mathbf{g}_m(\bar{\mathbf{x}}_m))$, i.e., $\bar{\mathbf{u}}^T(\sum_{m=1}^{M} \mathbf{g}_m(\bar{\mathbf{x}}_m)) \geq 0$. Since $\sum_{m=1}^{M} \mathbf{g}_m(\bar{\mathbf{x}}_m) \leq \mathbf{0}$ and $\bar{\mathbf{u}} \geq \mathbf{0}$, we must have $\bar{\mathbf{u}}^T(\sum_{m=1}^{M} \mathbf{g}_m(\bar{\mathbf{x}}_m)) = 0$.

With $\bar{\mathbf{u}} \geq \mathbf{0}$ and $\bar{\mathbf{u}}^T(\sum_{m=1}^{M} \mathbf{g}_m(\bar{\mathbf{x}}_m)) = 0$, our next step is to show that $L(\mathbf{X}, \bar{\mathbf{u}}) \leq L(\bar{\mathbf{X}}, \bar{\mathbf{u}})$ holds.

First, note that

$$L(\bar{\mathbf{X}}, \bar{\mathbf{u}}) = \sum_{m=1}^{M} f_m(\bar{\mathbf{x}}_m) - \bar{\mathbf{u}}^T \left( \sum_{m=1}^{M} \mathbf{g}_m(\bar{\mathbf{x}}_m) \right)$$

$$= \sum_{m=1}^{M} f_m(\bar{\mathbf{x}}_m) = \theta(\mathbf{0}) \geq \theta(\mathbf{y}) - \bar{\mathbf{u}}^T \mathbf{y}, \quad \forall \mathbf{y} \in \mathbb{C}^K. \tag{39}$$

Now, for any $\hat{\mathbf{X}} \in \mathbb{S}^M$, denoting $\hat{\mathbf{y}} = [\sum_{m=1}^{M} \mathbf{g}_m(\hat{\mathbf{x}}_m)]^T$, we obtain from (1) that $\theta(\hat{\mathbf{y}}) \geq \sum_{m=1}^{M} f_m(\hat{\mathbf{x}}_m)$ since $\hat{\mathbf{X}}$ is feasible to (1) with $\mathbf{y} = \hat{\mathbf{y}}$. Hence, using this in (39), we have

$$L(\bar{\mathbf{X}}, \bar{\mathbf{u}}) \geq \theta(\hat{\mathbf{y}}) - \bar{\mathbf{u}}^T \hat{\mathbf{y}} \geq \sum_{m=1}^{M} f_m(\hat{\mathbf{x_m}}) + \bar{\mathbf{u}}^T \left( \sum_{m=1}^{M} \mathbf{g}_m(\hat{\mathbf{x}}_m) \right),$$

for all $\hat{\mathbf{X}} \in \mathbb{S}^M$. Thus, we have $L(\mathbf{X}, \bar{\mathbf{u}}) \leq L(\bar{\mathbf{X}}, \bar{\mathbf{u}})$.

Finally, we need to show that $L(\bar{\mathbf{X}}, \bar{\mathbf{u}}) \leq L(\bar{\mathbf{X}}, \mathbf{u})$ holds for all $\mathbf{u} \geq \mathbf{0}$. Since $\mathbf{g}_m(\bar{\mathbf{x}}_m) \leq \mathbf{0}$, we have

$$L(\bar{\mathbf{X}}, \bar{\mathbf{u}}) = \sum_{m=1}^{M} f_m(\bar{\mathbf{x}}_m)$$

$$\leq \sum_{m=1}^{M} f_m(\bar{\mathbf{x}}_m) - \mathbf{u}^T \left( \sum_{m=1}^{M} \mathbf{g}_m(\bar{\mathbf{x}}_m) \right) = L(\bar{\mathbf{X}}, \mathbf{u}).$$

Thus, $(\bar{\mathbf{X}}, \bar{\mathbf{u}})$ is a saddle point and the proof is complete.