Calhoun

Institutional Archive of the Naval Postgraduate School

Calhoun: The NPS Institutional Archive

Faculty and Researcher Publications   Faculty and Researcher Publications

2012

Feedback Allocation For OFDMA Systems With Slow Frequency-domain Scheduling

Ganapathy, Harish

Harish Ganapathy, Siddhartha Banerjee, Nedialko B. Dimitrov and Constantine Caramanis.
Feedback Allocation For OFDMA Systems With Slow Frequency-domain Scheduling. IEEE

Feedback Allocation For OFDMA Systems With Slow Frequency-domain Scheduling

Harish Ganapathy[*], Siddhartha Banerjee, Nedialko B. Dimitrov and

Constantine Caramanis

Abstract

We study the problem of allocating limited feedback resources across multiple users in an orthogonal-frequency-division-multiple-access downlink system with slow frequency-domain scheduling. Many flavors of slow frequency-domain scheduling (e.g., persistent scheduling, semi-persistent scheduling), that adapt user-sub-band assignments on a slower time-scale, are being considered in standards such as 3GPP Long-Term Evolution (LTE). In this paper, we develop a feedback allocation algorithm that operates in conjunction with any arbitrary slow frequency-domain scheduler with the goal of improving the throughput of the system. Given a user-sub-band assignment chosen by the scheduler, the feedback allocation algorithm involves solving a weighted sum-rate maximization at each (slow) scheduling instant. We first develop an optimal dynamic-programming-based algorithm to solve the feedback allocation problem with pseudo-polynomial complexity in the number of users and in the total feedback bit budget. We then propose two approximation algorithms with complexity further reduced, for scenarios where the problem exhibits additional structure.

Index Terms

Limited feedback, multi-user feedback allocation, throughput-optimal, uplink feedback, Random Vector Quantization, sub-modular functions, convex relaxations

*Corresponding author. A preliminary version with a subset of the results was presented at the *Allerton Conference on Communication, Control and Computing*, Monticello, IL, Sep. 2009.

H. Ganapathy, S. Banerjee and C. Caramanis are with the Department of Electrical and Computer Engineering, The University of Texas, Austin, TX 78712. N. B. Dimitrov is with the Operations Research Department, Naval Postgraduate School, Monterey, CA 93943.

I. INTRODUCTION

Orthogonal-frequency-division-multiple-access (OFDMA) is the multiple-access technology of choice for many current and future wireless standards such as 3GPP Long-Term Evolution (LTE), IEEE 802.16e (WiMAX) and Long-Term Evolution Advanced (LTE-A). With the singular goal of achieving higher throughputs in order to keep pace with the every-growing suite of data-hungry applications, OFDMA systems typically operate in conjunction with a fast frequency-domain scheduler that allows for aggressive adaptation to the fading conditions of the channel. Here, user-sub-band assignments typically occur once every 1ms, 2ms or 5ms depending upon the standard under consideration. In the quest for higher data rates, the overhead incurred in *enabling* such fast frequency-domain scheduling is often ignored.

Primarily, there are two types of overhead that facilitate user scheduling in an OFDMA downlink system. These are: the overhead incurred in *(i)* communicating user-sub-band assignments and in *(ii)* collecting channel state information (CSI) from all users commonly referred to as the process of feedback. To address the first issue, recently, there has been an increasing interest in "slow" frequency-domain scheduling [1]–[5] instead of its faster counterpart for applications where the overhead demands of the latter do not justify its use. For example, LTE adopts *(semi-)persistent* scheduling for voice-over-IP applications that typically do not have high throughput demands [1]–[4]. Here, user-sub-band assignments are decided on a slower time-scale while link adaptation (on the fast time-scale) specifically in the semi-persistent approach, is achieved through Hybrid Automatic Repeat Request (HARQ) re-transmissions. Li et al. [5] show that slow OFDMA scheduling can achieve throughputs close to the ideal case in many real-world scenarios.

Moving on to the implications of *(ii)*, we borrow an example of a typical LTE setting recently provided in Ouyang et al. [23]: In LTE, the smallest unit of bandwidth that can be assigned to a user for data transmission is called a resource block, which is essentially a group of OFDM sub-carriers. If we consider a 10MHz LTE system with $L = 50$ resource blocks shared by $K = 50$ users equipped with standard 4-bit codebooks (modulation/coding tables) at the mobiles, we have a total feedback bandwidth of $4KL = 4 \times 50 \times 50 = 10$kb per sub-frame [6]. Given a typical uplink data rate of 48kb per sub-frame, this consumes 20% of the uplink capacity, clearly making feedback bandwidth consumption an important bottleneck. This observation, amongst others, has motivated the development of limited feedback techniques [7]–[15]. In general, adapting the size of

the codebook (e.g., CSI table at the mobiles) [8]–[10] and sub-carrier grouping [11]–[15], subject to a constraint on the total available feedback bandwidth, are two of the most popular multi-user limited feedback approaches in the literature. In the former, the size of the codebook on each OFDMA sub-band, and potentially the codebooks themselves, are periodically chosen based on the "state" of the network. In the latter, feedback reduction is achieved through a grouping technique where one CSI report is generated for a group of OFDMA sub-bands.

In this paper, we propose a feedback allocation policy that operates in conjunction with a slow frequency-domain scheduler assumed given. In particular, given a scheduling assignment on a slower time-scale, i.e., once every T time slots, the feedback allocation policy decides user codebook sizes again on the same time-scale. Thus, in the context of past literature, we focus on the former limited feedback approach of choosing dynamic codebook sizes as a function of the network state (e.g., channels, queues, etc.), a process that we call *feedback allocation* henceforth[1], to address the second type of overhead. The difference between our approach and past work on dynamic codebook selection is that our algorithm adapts to queue sizes and hence user applications, in addition to the channel state thereby generalizing earlier methods.

The main contributions of this paper are the following:

1) We propose a throughput-optimal feedback allocation policy that overlays a given slow scheduler. The proposed policy takes the form of a weighted sum-rate maximization problem that needs to be solved once every T time slots. Throughput-optimality is with respect to the space of all possible feedback allocation policies while *fixing* the particular data scheduler of interest.

Efficient algorithmic implementability of these policies is a critical design requirement, and this is the focus of our remaining contributions. Our focus is in line with several papers over the last decade, which study the algorithmic aspects of queue-based scheduling for specific network structures and resource allocation problems (see, e.g., [24]–[28] and references therein). Needless to say, the difficulty in solving the weighted sum-rate maximization problem in each slot is intricately tied to the resource being optimized. Recently, significant strides were made by Tan et al. [27], [28] in solving the joint queue-based scheduling and power control problem that has attracted much

---

[1]One can in general consider a more comprehensive feedback allocation policy that includes both codebook-size adaptation *and* sub-carrier grouping. However, such policies are beyond the scope of this paper and a subject for future study.

interest over the years (see [26] and references therein). Here, the possible transmission rates in each slot come from a continuous region induced by all possible power allocations. The authors [27], [28] solve non-convex power-control problems (e.g., weighted sum-rate) accurately and efficiently by using solutions to related convex problems (e.g., max-min rate) in an intelligent manner. Optimality is established under many channel settings. While the lack of convexity is due to interfering users in [27], [28], in the model by Huang et al. [26], self-noise due to channel estimation error forms the cause. The authors [26] nevertheless propose both optimal and sub-optimal approaches with varying degrees of complexity. In contrast to power allocation, in our case, the region of possible rates in each slot is discrete and is induced by all possible splits of the total feedback budget. This allows us to leverage many powerful tools from the area of combinatorial optimization. With the exception of the work of Ouyang and Ying [23], the problem of feedback allocations has not been considered in the past, to the best of our knowledge.

2) We develop a dynamic programming algorithm that solves the weighted sum-rate feedback allocation problem with pseudo-polynomial complexity in the number of users and the total feedback bit budget. This approach is exact and requires no assumptions on the structure of the weighted sum-rate function.

3) We show that in many practical wireless systems, the weighted sum-rate is non-decreasing and sub-modular in the feedback budget. Using this observation, we leverage sub-modular optimization results from combinatorial optimization (e.g. [19]–[21]) and propose a reduced-complexity algorithm with an approximation guarantee of $(1 - \frac{1}{e})$.

4) Multiple-input-single-output (MISO) beamforming is being considered as a potential transmission mode in the Long Term Evolution standard [1]. For such systems, we show that when the popular Random Vector Quantization codebook [17], [33], [34] is used, we are able to reduce the complexity even further and provide an approximation guarantee of $\frac{1}{2}$.

The rest of this paper is organized as follows. In Section II, we introduce the system model for feedback allocation and slow data scheduling. In Section III, we discuss the notion of throughput-optimality in queueing networks and introduce a throughput-optimal joint feedback allocation and slow data scheduling policy. In Section IV, we solve the optimal online feedback optimization problem for both objectives while in Section V, we investigate methods of reducing the complexity of the optimal online optimization problem by exploiting more structure in the objective function.

Concluding remarks are made in Section VII.

Notation: x_{ij} denotes element (i, j) of matrix $\mathbf{X}$ while x_i denotes element i of vector $\mathbf{x}$. Given matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{p \times q}$, $X \leq Y$ means $x_{ij} \leq y_{ij}, \forall i = 1, \ldots, p, \; j = 1, \ldots, q$. $\mathbb{R}_+$, $\mathbb{N}_0$ and $\mathbb{N}$ represent the non-negative real numbers, non-negative integers and positive integers respectively.

II. SYSTEM MODEL

We consider the downlink of a frequency-division-duplex OFDMA system with L sub-carriers/sub-bands and K users that operates in slotted-time. The network model is described below:

Channel State: The true supportable rate for user i on sub-band j at time t is given by $\mu_{ij}[t]$. We assume that $\mu_{ij}[t]$ is ergodic and comes from a continuous set $\mathcal{M}$. We assume that the mobile has perfect knowledge of the channel state $\{\mu_{ij}[t]\}_{j=1}^{L}$ in every time slot. The probability density function for rate $\mu_{ij}[t]$ is given by $\Pr(\mu_{ij}[t] = m) = \pi_{mi}(\alpha_i[t]), \; m \in \mathcal{M}$, where $\alpha_i[t]$ denotes a large-scale fading gain that is dependent on user position and comes from a finite set $\alpha_i[t] \in \Omega$. Users change positions once every T slots where $T \in \{1, 2, 3, \ldots\}$ denotes the large-scale fading coherence time. For ease of notation, we introduce a counter $\bar{t} = \lfloor \frac{t}{T} \rfloor T$ to keep track of the slower large-scale fading time-scale, i.e., $\pi_{mi}(\alpha_i[t]) = \pi_{mi}(\alpha_i[\bar{t}]), \; \forall t$. For convenience, we also set $\pi_{mi}[\bar{t}] = \pi_{mi}(\alpha_i[\bar{t}])$ making implicit the dependence on t and T. Note that the large-scale coefficient is typically only distance-dependent and independent of frequency allowing us to omit the index j when representing it. We assume that the base station has perfect knowledge of $\{\alpha_i[\bar{t}]\}$ and all distribution information $\{\pi_{mi}[\bar{t}]\}$. *Most importantly, $\bar{t}$ represents the time-scale at which feedback optimizations and scheduling assignments are decided.*

Traffic model: Each user $k \in \{1, 2, \ldots, K\}$, has a queue of untransmitted packets with queue-length $q_k[\bar{t}]$ that is maintained at the base station with associated arrival rate λ_k.

Feedback model: The base station allocates a feedback budget of $b_k[\bar{t}]$ bits to user k such that $\sum_{k=1}^{K} b_k[\bar{t}] \leq B$ where B represents the total limited feedback budget of the system. Let the sub-carriers in our OFDMA system be indexed by $\mathcal{S} = \{1, 2, \ldots, L\}$. Assume that user k is allocated sub-bands $\mathcal{N}_k[\bar{t}] \subseteq \mathcal{S}$ by the slow scheduling algorithm. Given a budget of $b_k[\bar{t}]$ bits by the base station and an assignment $\mathcal{N}_k[\bar{t}]$ of size $|\mathcal{N}_k[\bar{t}]| = n_k$, the user employs a codebook of size $b_{kj}[\bar{t}], \; j \in \mathcal{N}_k[\bar{t}]$ bits for sub-band j where $\sum_{j \in \mathcal{N}_k} b_{kj}[\bar{t}] = b_k[\bar{t}]$; $\{b_{kj}[\bar{t}]\}_{j \in \mathcal{N}_k}$ represents the per-sub-band budgets for user k that are chosen by the user to maximize rate.

Quantized channel state: Given a budget of $b_k[\bar{t}]$ and a sub-band assignment $\mathcal{N}_k[\bar{t}]$, the actual post-quantization rate achieved by user k at time t where $\bar{t} \leq t \leq \bar{t}+1$, on sub-band $j \in \mathcal{N}_k[\bar{t}]$ is denoted by $\mu_{kj}^q[b_{kj}[\bar{t}], \mu_{ij}[t]]$. Note that the actual achievable rate is determined by the quantization budgets (along with the codebook of course), that are decided on the slower time-scale indexed by $\bar{t}$, as well as the true state of the channel at current time t.

Network state: The network state at time $\bar{t}$ is given by $M[\bar{t}] = \left(\{\pi_{mi}[\bar{t}]\}_{i=1}^K, \{q_k[\bar{t}]\}_{i=1}^K\right)$, which is a collection of channel distributions and queue lengths on the slower time-scale. In general, the feedback allocation and slow scheduling policies make allocation and assignment decisions, respectively, for T time slots $\bar{t} < t < \bar{t}+1$ based on state $M[\bar{t}]$.

Expected rates and *virtual* users: Let $r_{kj}[\alpha_k[\bar{t}], b_{kj}[\bar{t}]] = \mathbb{E}_{\pi_{mi}[\bar{t}]}\left[\mu_{kj}^q[b_{kj}[\bar{t}], \mu_{ij}[t]]\right]$ denote the expected rate (through the course of T time slots)) for user k on sub-band j. The total expected rate that is achieved by user k given a sub-band assignment $\mathcal{N}_k[\bar{t}]$ and allocation $b_k[\bar{t}]$ is then given by $\sum_{j \in \mathcal{N}_k[\bar{t}]} r_{kj}[\alpha_k[\bar{t}], b_{kj}[\bar{t}]]$. We make an observation at this point that helps us simplify the presentation of the results. Since the rate is additive across sub-bands, and is a function of a band-independent channel gain, one may consider and analyze an equivalent *virtual* system where the number of users is *equal* to the number of sub-bands. This removes the dependence of the feedback allocation policy on the assignments $\mathcal{N}_k[\bar{t}]$. In other words, the equivalent system would consist of L users assigned to L sub-bands with feedback allocations $\{b_{i'}[\bar{t}]\}_{i'=1}^L$ and rates $r_{i'}[\alpha_{i'}[\bar{t}], b_{i'}[\bar{t}]]$. As for the queue lengths, one can simply "replicate" the same queue length $q_k[\bar{t}]$ for all virtual users $k' \in \mathcal{N}_k[\bar{t}]$, i.e., $q_{k'}[\bar{t}] = q_k[\bar{t}], \; \forall k' \in \mathcal{N}_k[\bar{t}]$. Once the optimal feedback allocation $\{b_{i'}^*[\bar{t}]\}_{i'=1}^L$ and virtual rates $r_{i'}[b_{i'}^*[\bar{t}]]$ are computed, we can map back to the original system by calculating the true rate for user k as $\sum_{i' \in \mathcal{N}_k[\bar{t}]} r_{i'}[\alpha_{i'}[\bar{t}], b_{i'}^*[\bar{t}]]$.

Through the remainder of this paper barring the final simulations section, we study the equivalent system mentioned above where we have L users assigned to L sub-bands. Having defined all the ingredients of our OFDMA downlink network, we move on to the next section where we develop the feedback allocation policy that periodically makes decisions based on the network state.

III. THROUGHPUT-OPTIMAL FEEDBACK ALLOCATION WITH SLOW SCHEDULING

In this section, we develop a feedback allocation (codebook size adaptation) protocol that when operated in conjunction with a given slow data scheduling policy, guarantees *throughput-optimality*. This means that given an arrival rate vector $\vec{\lambda}$, if there exists any scheduling policy that can guarantee

bounded expected queue sizes, then so can the proposed policy, which falls under the MaxWeight family of policies that was pioneered by Tassiulus and Ephremedis [29].

As mentioned towards the end of the last section, we now have a virtual system with L users assigned to L sub-bands with feedback allocations $\{b_{k'}[\bar{t}]\}_{k'=1}^L$, rates $\{r_{k'}[\alpha'[\bar{t}], b_{k'}[\bar{t}]]\}_{k'=1}^L$ and queues $\{q_{k'}[\bar{t}]\}_{k'=1}^L$. Through the remainder of this paper, until Section VI, we replace the index k' by k for convenience, with the implicit understanding that we are dealing with virtual users. The feedback allocation policy is presented below.

Algorithm 1 MaxWeight feedback allocation with slow data scheduling

---

1: **while** $t \geq 0$ **do**

2:     **if** $t \ (\mathrm{mod}\ T) = 0$ **then**

3:         Set $\bar{t} = t$

4:         Solve

$$\begin{aligned} \{b_k^*[\bar{t}]\} = \quad & \arg\max \quad \textstyle\sum_{k=1}^L q_k[\bar{t}] r_k[\alpha_k[\bar{t}], b_k] \\ & \text{w.r.t.} \quad b_k \in \{0, 1, \ldots, B\}, \ \forall k \\ & \text{s.t.} \quad \textstyle\sum_{k=1}^L b_k \leq B. \end{aligned} \quad (1)$$

5:     **end if**

6: **end while**

A few remarks about the above algorithm are in order:

(i) Throughput-optimality: The algorithm is throughput-optimal with respect to the space of policies that make feedback allocation and assignment decisions once every T slots. This means that if any policy that makes feedback allocation and assignment decisions once every T slots can stabilize a set of arrival rates $\{\lambda_k\}$, then so can the proposed policy in (1). Let the region of rates $\{\lambda_k\}$ that can be stabilized by the policy in (1) be denoted by $\mathcal{V}$. The above notion of throughput optimality for queueing systems has been used extensively in the literature (see [24], [25], [29], [30] and references therein). We do not prove throughput-optimality as it follows from standard Lyapunov techniques that are well-established in the queueing literature [30].

(ii) Computational complexity: While the optimization problem characterizes optimal performance, solving it exactly may be computationally prohibitive. In fact, a brute-force approach to solving (1) would incur a complexity of $\mathcal{O}\left(\binom{B+L-1}{L-1}\right)$.

The final remark forms the basis for the remainder of this paper. The brute-force approach to solving (1) is clearly infeasible from an implementation perspective. We take up the issue of complexity starting in Section IV and propose a host of computationally-efficient algorithms to solve the feedback allocation problem in (1). We wish to highlight that all algorithmic developments can be applied to full-buffer (saturated) systems where scheduling schemes such as *proportional fairness* become applicable. This is because most schedulers of interest solve a weighted sum-rate maximization problem at each instant [31].

IV. OPTIMAL FEEDBACK ALLOCATION THROUGH DYNAMIC PROGRAMMING

In Section III, we have established that for queue stability, we are interested in solving the following online weighted sum-rate maximization problem

$$\max_{\{b_k\} \in \mathcal{B}} \quad \sum_{k=1}^{L} q_k[\bar{t}] r_k[\alpha_k[\bar{t}], b_k]. \tag{2}$$

The form of the functions $\{r_k[\alpha_k[\bar{t}], b_k]\}$ would of course depend on the underlying physical system and is intimately connected to the computational complexity of the problem. In fact, for complex modulation/coding schemes the function might only be available as a look-up table. While the optimization problem characterizes optimal performance, solving it exactly may be computationally prohibitive. Thus, the focus of this paper becomes algorithmic. We propose novel solutions to (2) through Sections IV and V that explore the natural tradeoffs between accuracy, complexity and the structure of the weighted sum-rate function. We start by showing that by using Dynamic Programming, the exact solution can be obtained in pseudo-polynomial time.

Theorem 1. *The online resource allocation problem (2) can be solved exactly in time $\mathcal{O}\left(LB^2\right)$.*

Proof: Order the users arbitrarily. We choose to work with the existing order w.l.o.g. For any given arbitrary weights $\{w_i\}$, $w_i > 0$, define $A(i,j) \triangleq w_i r_k[\alpha_i, j]$ to be the weighted rate for user i given we allocate j bits to this user and define $R(k,b) \triangleq \max_{\sum_{i=1}^{k} b_i \leq b, \ b_i \in \mathbb{N}_0} \sum_{i=1}^{k} w_i r_i[\alpha_i, b_i]$ to be the maximum weighted sum-rate if we have b bits to allocate amongst the first k users with $R(0,b) = 0$. It follows that $R(1,b) = A(1,b), b = 0, \ldots, B$. We can write a recursion $R(k,b) = \max_{j=0,\ldots,b} \{R(k-1, b-j) + A(k,j)\}$. The optimality of this recursion can be established using standard induction arguments similar to the two-dimensional knapsack problem [16]. This rule gives rise to a table with a total of $L(b+1)$ elements. In order to compute element (k,b) in the table, using

our recursion, we incur a complexity of $\mathcal{O}(b+1)$. Hence, the total complexity can be calculated as $\sum_{k=1}^{L} \sum_{b=0}^{B}(b+1) = L \sum_{b=0}^{B}(b+1) = L\frac{(B+1)(B+2)}{2} = \mathcal{O}(LB^2)$. ∎

Thus, we have proposed an exact solution using dynamic programming, which has pseudo-polynomial[2] complexity $\mathcal{O}\left(LB^2\right)$ and which is applicable to *any* type of weighted sum-rate function. Therefore, in contrast to the joint power-control and scheduling problems in [26]–[28] and owing to the discrete nature of the feedback allocation problem we consider in (2), we do not require any special channel-induced properties of the objective function such as those imposed on its partial derivatives in Lemma 2 of [27], in order to find an optimal solution.

Note that $R(K, B)$ in Theorem 1 with $w_i = q_i$, $\forall i$, is equal to (2) and dynamic programming essentially gives us a technique to compute $R(K, B)$ by solving smaller sub-problems. The following toy example with $K = 2$ users and a total bit budget of $B = 2$ bits illustrates a typical series of computations en route to calculating $R(2, 2)$.

Example (Dynamic programming): Order the two users arbitrarily, say user 1 first followed by user 2. Then, initialize values the following weighted rates appropriately for $b = 0, 1, 2$,

$$R(1, b) = A(1, b) \quad : \quad \text{when user 1 is allocated } b \text{ bits}$$
$$R(2, b) = A(2, b) \quad : \quad \text{when user 2 is allocated } b \text{ bits}.$$

Once initialized, we then compute value $R(2, 1) = \max\{R(1, 1) + A(2, 0), R(1, 0) + A(2, 1)\} = \max\{R(1, 1), A(2, 1)\}$. Finally, we calculate

$$
\begin{aligned}
R(2, 2) \quad = \quad & \max\{R(1, 2) + A(2, 0), R(1, 1) + A(2, 1), \\
& R(1, 0) + A(2, 2)\} \\
= \quad & \max\{R(1, 2), R(1, 1) + A(2, 1), A(2, 2)\},
\end{aligned}
$$

the desired optimal weighted sum-rate with two users and two bits.

To understand the computational requirements in the context of a real-world system, consider the LTE example that was presented in the introduction to this paper. Here, we had the following parameters: $L = 50$, $K = 50$ with 4-bit modulation/coding tables at the mobiles. To model the limited feedback constraint, let $B = 4cL$, $c \in \{1, 2, ..., K - 1\}$, which has an intuitive interpretation of being able to provide full feedback to at most c users; $c = K$ represents no constraint on feedback resources for this setting. Then, a feedback bandwidth of $c = \frac{K}{4}$ corresponds to a complexity of

---

[2] An algorithm has pseudo-polynomial complexity if its running time is a polynomial in the size of the input in unary. The size of the input to (2) in unary at most $LBA_{max} + B = \mathcal{O}(LB)$ where $A_{\max} = \max_{(i,j)} A(i, j)$.

roughly 7×10^{11} operations, which is clearly quite daunting. Thus, while the dynamic programming approach is indeed viable for sufficiently small systems, we require algorithms with faster running times that might be less accurate. This forms the focus of the remainder of this paper.

V. REDUCED-COMPLEXITY RESOURCE ALLOCATION

In this section, we show that if the weighted sum-rate functions have additional structure, we can develop faster algorithms. As is often done for computationally hard problems, one seeks efficient but potentially suboptimal algorithms, but then proves lower bounds on the performance. In this vein, we develop more computationally efficient algorithms that approximately solve (2), and provide theoretical lower bounds on their performance. The long-term performance of these approximate algorithms in achieving queue stability is characterized by Theorem 2 below. The proof is omitted as these are well-known results in queuing systems.

We say that an algorithm is a γ-approximation, $\gamma \in (0, 1]$, to (1) if it provides a solution $\{b_k^{alg}\}$ such that $\sum_k q_k[\bar{t}] r_k[\alpha_k[\bar{t}], b_k^{alg}] \geq \gamma \max_{\{b_k\} \in \mathcal{B}} \sum_k q_k[\bar{t}] r_k[\alpha_k[\bar{t}], b_k]$. The following theorem is a generalization of the original result by Tassiulus and Ephremedis [29]. It essentially states that local approximation is consistent with the long-term objectives we consider.

Theorem 2. *If $\lambda_k \in \{\gamma\nu : \nu \in \mathcal{V}\}, \gamma \in (0, 1]$, then a γ-approximation to the per-instant scheduling rule $\{b^*[t]\} = \arg\max_{\{b_k\} \in \mathcal{B}} \sum_k q_k[\bar{t}] r_k[\alpha_k[\bar{t}], b_k]$ stabilizes the system.*

Recall from remark *(i)* in the Section III that $\mathcal{V}$ represents the region of rates that are stabilizable by Algorithm 1. The theorem essentially states that for queueing systems: If we calculate a γ-approximate solution to (2) in every time slot, one can achieve a γ-fraction of the throughput region. This result paves the way for the design of computationally-efficient algorithms, by constructing approximations to (2).

In Section A, we consider weighted sum-rate functions that are non-decreasing and sub-modular in the feedback bit allocation. In short, sub-modularity refers to *diminishing returns* with respect to the allocation of resources. This is a property that is exhibited quite frequently by wireless systems in general, e.g., point-to-point capacity scales logarithmically in transmit power, achievable rates in multiple antenna precoding systems exhibit diminishing returns in the size of the codebook [17], [18], etc. In the developments that follow, we exploit this property in order to propose a *greedy*

feedback allocation algorithm that has complexity $\mathcal{O}((B+L)\log_2 L)$ with approximation factor of $\left(1 - \frac{1}{e}\right)$. Our main contributions are contained in Lemma 2 and Theorem 3.

In Section B, we focus on a class of weighted sum-rate functions that arise in downlink scenarios where the base station is equipped with multiple antennas and performs transmit beamforming with quantized beamformer feedback. This is a popular transmission strategy that been extensively researched [17], [33]–[35] and adopted into standards such as W-CDMA [32] and LTE [1]. We show that for this choice of physical layer signalling protocol, the weighted sum-rate maximization problem in (2) is sub-modular for certain types of beamformer quantizers. We illustrate the improvement in computational performance by using the LTE example from the Introduction.

A. Reduced-complexity resource allocation through sub-modularity

We begin this section with a quick primer on sub-modular optimization (summarized from [19]–[21]) that will be useful for our purposes. In keeping with the literature, the approach pursued in this section will be graph theoretic in contrast to the rest of this paper. A sub-modular function is defined as follows: Let E be a finite set and 2^E represent all its subsets. Then, $F : 2^E \rightarrow \mathbb{R}_+$ is a *non-decreasing*, *normalized*, *sub-modular* function if $F(\emptyset) = 0$ (normalized), $F(A) \leq F(B)$ when $A \subseteq B \subseteq E$ (non-decreasing) and if $F(A \cup \{e\}) - F(A) \geq F(B \cup \{e\}) - F(B)$, $\forall A \subseteq B \subseteq E$ and $e \in E \setminus B$ (sub-modular).

The following property of sub-modular functions is useful for our analysis.

Lemma 1. *If F_n, $n = 1, \ldots, N$, are sub-modular on set E, then $\sum_{n=1}^{N} w_n F_n(A)$, $A \subseteq E$ is a sub-modular function for $w_n \geq 0, \forall n$.*

Having provided the definition of sub-modularity along with a useful property, we now introduce the kinds of constraint sets that are typically considered in the context of sub-modular optimization: *(i)* A set system $(E, \mathcal{I})$ where E is a finite set and $\mathcal{I}$ is a collection of subsets of E is called an *independence system* if $\emptyset \in \mathcal{I}$ and satisfies if $A \subseteq B$ for $B \in \mathcal{I}$, then $A \in \mathcal{I}$. *(ii)* An independence system is called a *matriod* if it satisfies the following additional property; if $A, B \in \mathcal{I}$ and $|A| < |B|$, then there exists $e \in B \setminus A$ such that $A \cup \{e\} \in \mathcal{I}$. (iii) Set $\mathcal{I}$ is a *uniform matroid* if $\mathcal{I} = \{F \subseteq E : |F| \leq k\}$ for $k \in \mathbb{N}$.

The optimization problem that has been considered in the context of sub-modular functions and

independence systems is

$$F^* = \max_{A \in \mathcal{I}, A \subseteq E} F(A). \tag{3}$$

Since many NP-hard problems can be reduced to a sub-modular function maximization over an independence system, significant research has focused on developing efficient approximation algorithms. In particular, the performance of the greedy algorithm in solving special cases of (3) has been extensively studied. Nemhauser et al. [22] considered problem (3) over uniform matroids and showed that the greedy algorithm provides a $(1 - \frac{1}{e})$ approximation factor for this special case. Please refer to Goundan et al. [19], Calinescu et al. [20] and Vondrak [21] for a summary of related results on sub-modular function optimization over other families of constraint sets. The greedy algorithm is presented later in the section in the context of our specific feedback allocation problem.

Sub-modularity in feedback allocation: We now show that the optimal bit allocation problem in (2) may by posed as a sub-modular maximization over a uniform matroid when the rates exhibit sub-modularity. Let $\mathcal{G} = (U, V, E)$ be a bipartite graph where U contains L *user nodes* and V contains B *bit nodes*, both ordered arbitrarily, i.e. $|U| = L$ and $|V| = B$. Let E contain the set of all edges $E = \{e_{kb} : i = 1, \ldots, L \text{ and } j = 1, \ldots, B\}$. Given $A \subseteq E$, we define $|A|_i \triangleq |\{e_{kb} \in A : k = i\}|$ to represent the number of bits allocated to user i, i.e., $|A|_i = b_i$. The independence we are interested in is $\mathcal{I} = \{A \subseteq E : |A| \leq B\}$ where B is the total bit budget. By definition, $\mathcal{I}$ is a uniform matroid and furthermore, $\mathcal{I}$ is the set of *all* valid allocations since if $A \in \mathcal{I}$, then $\sum_{k=1}^{L} b_k = \sum_{k=1}^{L} |A|_k \leq B$ and if $A \notin \mathcal{I}$, then $\sum_{k=1}^{L} b_k = \sum_{k=1}^{L} |A|_k = |A| > B$. Now the weighted sum-rate maximization problem in (2) in time slot $\bar{t}$ may be re-written as

$$
\begin{aligned}
&\max_{\{b_k\} \in \mathcal{B}} && \sum_{k=1}^{L} q_k r_k[\alpha_k[\bar{t}], b_k] \\
\equiv \quad &\max && \sum_{k=1}^{L} q_k r_k[\alpha_k[\bar{t}], b_k] - r_k[\alpha_k[\bar{t}], 0] \\
&\text{s.t.} && b_k = |A|_k, \ \sum_k |A|_k \leq B, \ A \subseteq E \\
= \quad &\max_{A \in \mathcal{I}} && \sum_{k=1}^{L} q_k r_k[\alpha_k[\bar{t}], |A|_k] - r_k[\alpha_k[\bar{t}], 0].
\end{aligned}
$$

The following result becomes immediate.

Lemma 2. *If the function $r_k[\alpha_k, b_k]$ is non-decreasing and sub-modular in the bit allocation $b_k = |A|_k, \ A \subseteq E$ for all users $k = 1, \ldots, L$, and channel states $\alpha_k \in \Omega$, then $\sum_{k=1}^{L} q_k r_k[\alpha_k, |A|_k] - r_k[\alpha_k, 0]$ is a normalized, non-decreasing, sub-modular function on set E for all channel states* $\mathbf{m} \in \mathcal{M}$.

Proof: The result follows from Lemma 1. ■

Hence, the greedy algorithm can be used to solve the optimal bit allocation problem in (2) with approximation factor $\left(1 - \frac{1}{e}\right)$. The greedy algorithm for the specific case of our bit allocation problem in time slot t can be written as follows where $u_k(b_k) \overset{\triangle}{=} q_k\left(r_k[\alpha_k, b_k + 1] - r_k[\alpha_k, b_k]\right)$ denote the increase in rate or marginal utility if user k is given one extra bit.

Algorithm 2 Greedy feedback allocation

1: Set $b = 1$ and $b_k = 0, \forall k$, which is essentially a bit counter for each user

2: Compute marginal utilities $\{u_k(b_k)\}$.

3: **while** $\sum_k b_k \leq B$ **do**

4:     Sort this list of marginal utilities.

5:     Assign a bit to user k^* who is on top of this list.

6:     Update $b_{k^*} = b_{k^*} + 1$ and re-compute $u_{k^*}(b_{k^*})$

7: **end while**

We end this section by investigating the complexity of the above algorithm in the following theorem.

Theorem 3. *The greedy algorithm approximates the optimal bit allocation problem in (2) to within a factor of $\left(1 - \frac{1}{e}\right)$ while incurring complexity $\mathcal{O}((B + L)log_2 L)$.*

Proof: Step 2 of this algorithm incurs complexity $\mathcal{O}(L log_2 L)$ for the first iteration $b = 1$. Subsequently, every re-sort in Step 3 costs $\mathcal{O}(log_2 L)$ with a maximum of B such re-sorts. Thus, the total complexity is $\mathcal{O}((B + L)log_2 L)$. For the proof of the approximation factor, please refer to Nemhauser et al. [22]. ■

In the context of the LTE example introduced earlier, this means that by exploiting the sub-modular structure in the rates, we reduce the complexity from 7×10^{11} to 15×10^3 operations. Before we move on to the next section, we provide an example of a common wireless setting where sub-modularity is exhibited. Consider a traditional point-to-point single antenna wireless link with a b-bit modulation-coding table at the receiver. The modulation-coding table is constructed as follows. Given a non-negative real number in the interval $[0, \sigma]$, $\sigma >> 0$, we uniformly partition the interval into 2^b sub-intervals and implement the quantization scheme $\lfloor x \rfloor_Q = \frac{i\sigma}{2^b}, i\frac{\sigma}{2^b} \leq x < (i + 1)\frac{\sigma}{2^b}$, $i =$

$0, 1, \ldots, 2^b - 1$. Then, for any fixed position-dependent gain of α, the achievable rate of the system in a fading environment can be written as

$$r[\alpha, b] = \mathbb{E}_h[\log_2(1 + \lfloor \sqrt{\alpha}|h|^2 \rfloor_Q)], \tag{4}$$

where $|h|^2$ is a truncated $\exp(1)$ random variable that has a maximum value of $\sigma >> 0$. The probability density function for such a random variable is given by $f_{|h|^2}(x) = C(\sigma)\exp(-x)$ where $C(\cdot)$ is a normalization factor. Note that this example considers a traditional continuous fading model. One may obtain its discrete version thereby conforming with our system model, by sampling the support $[0, \sigma]$. Thus, the rate expression in (4) may be treated as an approximation that becomes increasingly accurate as we discretize the support more finely. For the case $\alpha = 1$, the rate (4) can be explicitly computed as

$$
\begin{aligned}
r[1, b] &= C(\sigma) \sum_{i=0}^{2^b - 1} \log_2\left(1 + \frac{i\sigma}{2^b}\right) \int_{i\frac{\sigma}{2^b}}^{(i+1)\frac{\sigma}{2^b}} \exp(-x)dx \\
&= \left[1 - \exp\left(-\frac{\sigma}{2^b}\right)\right] \sum_{i=0}^{2^b - 1} \log_2\left(1 + \frac{i\sigma}{2^b}\right)\exp\left(-i\frac{\sigma}{2^b}\right).
\end{aligned}
$$

Setting $l[j, b] = \log_2\left(1 + \frac{j\sigma}{2^b}\right)$, the normalized incremental gain with one extra bit can be calculated as

$$
\begin{aligned}
&C(\sigma)^{-1}(r[1, b+1] - r[1, b]) \\
&= \left[1 - e^{-\frac{1}{2}\frac{\sigma}{2^b}}\right] \sum_{i=0}^{2\times 2^b - 1} l[i, b+1]e^{-i\frac{1}{2}\frac{\sigma}{2^b}} - \left[1 - e^{-\frac{\sigma}{2^b}}\right] \\
&\quad \sum_{i=0}^{2^b - 1} l[i, b]e^{-i\frac{\sigma}{2^b}} \\
&= \left[1 - e^{-\frac{1}{2}\frac{\sigma}{2^b}}\right] \left[\sum_{j=0}^{2^b - 1} l[2j, b+1]e^{-j\frac{\sigma}{2^b}} + \sum_{j=0}^{2^b - 1} l[2j+1, \right. \\
&\quad \left. b+1]e^{-(2j+1)\frac{\sigma}{2^{b+1}}}\right] - \left[1 - e^{-\frac{\sigma}{2^b}}\right]\left[\sum_{i=0}^{2^b - 1} l[i, b]e^{-i\frac{\sigma}{2^b}}\right] \\
&\qquad \text{by splitting odd and even terms} \\
&= \left[e^{-\frac{1}{2}\frac{\sigma}{2^b}} - e^{-\frac{\sigma}{2^b}}\right]\left[\sum_{j=0}^{2^b - 1} e^{-j\frac{\sigma}{2^b}}\log_2\left(1 + \frac{0.5}{\frac{2^b}{\sigma} + j}\right)\right].
\end{aligned} \tag{5}
$$

Through simple numerical enumeration, one may confirm that the 1-bit rate gain given above in (5) decreases over realistic bit sizes of $b \in \{1, 2, \ldots, 25\}$ and hence, $r[1, b]$ is a sub-modular function. With a little more algebra, one may derive a similar result for the general case with any arbitrary, non-negative, position-dependent gain α.

In the next section, we provide another example of a wireless system that exhibits sub-modularity. In particular, we consider a class of multiple antenna wireless links and solve (2) in the context of these systems.

B. Reduced-complexity resource allocation for MISO systems

When the user rates $r_k[\alpha, b]$ are sub-modular in the bit allocation b in every channel state $\alpha \in \Omega$, we use the greedy algorithm in Section A to approximately solve the online feedback allocation problem in (2) with complexity $\mathcal{O}((B + L)\log_2 L)$. In this section, we show that 2×1 MISO quantized transmit diversity systems exhibit sub-modular expected rates bringing into use the results from the previous section. Furthermore, in the context of these specific transmission schemes, we develop an approximation algorithm based on convex relaxations with a further-reduced complexity of $\mathcal{O}(L\log_2 L)$ and an approximation guarantee of $\frac{1}{2}$ for typical operational signal-to-noise ratios (SNR). Thus, aside from the usual impact on precision that is typically omitted from running time calculations, the running time of our algorithm no longer depends on the feedback budget B. In the example above, the running time is reduced even further from 15×10^3 operations to roughly 300 operations.

We begin this section by investigating the effects of limited feedback on the aforementioned class of MISO systems. It is well-known that the instantaneous SNR for a classical 2×1 single-stream beamforming MISO link is given by $\overline{\text{SNR}}(\alpha)||\mathbf{h}||^2$ where $\overline{\text{SNR}}(\alpha) = \frac{P\alpha}{N_o}$, P is the transmit power, N_o is the noise power and $\mathbf{h} = [h_1 \ h_2]^T$, $h_i \in \mathbb{C}$ represents the MISO channel with zero mean, unit variance complex Gaussian entries. As with the example in the previous section, the analytical rate expressions in this section are derived for continuous vector channels, which are increasingly accurate approximations as we sample the support $\mathbb{C}^2$ more finely. Recall from Section II that $\alpha \in \Omega$ models the effects of large-scale fading. To achieve this maximum instantaneous SNR, the user requires perfect feedback of the channel vector $\mathbf{h}$. However, feedback in realistic systems is imperfect due to limited feedback budgets and quantization, the primary motivation for this work. We assume that the channel vector $\mathbf{h}$ is quantized using the popular *Random Vector Quantization* (RVQ) technique [17], [34]. This technique is briefly reviewed in the next section when we present simulation results. Recent results [17], [33], [34] bound (upper and lower) the loss in rate due to quantization of $\mathbf{h}$ when using RVQ codebooks. In particular, both upper and lower bounds on the rate loss due to quantization for user k take the form $c(\alpha_k)2^{-b_k}$ for some $c(\alpha_k) > 0$. Motivated by these results, we assume that the post-quantization rate for user k in the presence of large-scale fading takes the form $r_k[\alpha_k, b_k] = \mathbb{E}\left[\log_2(1 + \overline{\text{SNR}}_k||\mathbf{h}||^2)\right] - \left(\mathbb{E}\left[\log_2(1 + \overline{\text{SNR}}_k||\mathbf{h}||^2)\right] - \mathbb{E}\left[\log_2(1 + \overline{\text{SNR}}_k|h|^2)\right]\right)2^{-b_k}$, where we have omitted the dependence

on $\bar{t}$ for brevity. We validate the use of the above approximation through numerical testing in the next section for many values of α_k from a typical operational range in a wireless system.

Thus, the optimization in (2) for the 2×1 MISO case takes the specific form

$$\max_{\{b_k\} \in \mathcal{B}} \sum_{k=1}^{L} q_k \left[\beta_2(\overline{\text{SNR}}_k) - \left(\beta_2(\overline{\text{SNR}}_k) - \beta_1(\overline{\text{SNR}}_k) \right) 2^{-b_k} \right], \tag{6}$$

where $\text{SNR}_k = \text{SNR}(\alpha_k)$ for short, $\beta_1(\overline{\text{SNR}}) = \mathbb{E} \left[\log_2(1 + \overline{\text{SNR}}|h_1|^2) \right]$ and $\beta_2(\overline{\text{SNR}}) = \mathbb{E} \left[\log_2(1 + \overline{\text{SNR}}||\mathbf{h}||^2) \right]$ denote the one-tap and two-tap expected rates, respectively, for a Rayleigh fading channel with the given SNR.

Relaxation and approximation guarantees: Through the remainder of the section, we develop an approximation algorithm to solve (6) in closed-form while incurring a complexity of $\mathcal{O}(L \log_2 L)^3$. We provide an approximation guarantee of $\frac{1}{2}$.

Theorem 4. *Consider the following continuous relaxation of (6) formed by replacing the discrete set $\mathcal{B}$ with its natural continuous extension and dropping terms that are independent of the variables $\{b_k\}$:*

$$\{b_k^*\} = \arg \min_{\sum_k b_k \leq B, \, b_k \in \mathbb{R}_+} \sum_{k=1}^{L} q_k \beta_1(\overline{SNR}_k) 2^{-b_k}. \tag{7}$$

The solution to this relaxation is $b_k^* = \left[-\log_2 \left(\frac{\eta^*}{q_k(\log 2)} \frac{1}{\beta_1(\overline{SNR}_k)} \right) \right]^+$, where η^* is chosen such that $\sum_k b_k^* = B$ and $[x]^+ = \max\{x, 0\}$.

Proof: See Appendix A. ∎

Next, we comment on the complexity of computing the above fractional solution.

Theorem 5. *Computing the above solution in Theorem 4 incurs a complexity of $\mathcal{O}(L \log_2 L)$.*

Proof: See Appendix A. ∎

The following lemma states that weighted sum-rate function in (6) is non-decreasing and sub-modular on set $E = \{e_{kb} : i = 1, \ldots, L \text{ and } b = 1, \ldots, B\}$, thereby allowing us to connect and compare the results in this section with those in the previous section on sub-modular functions. The proof is standard in the literature on sub-modular functions and follows from the fact that

---

[3]We recognize that there is an additional storage cost of $\mathcal{O}(\log B)$.

the fractional relaxation of the weighted sum-rate function is concave in $\{b_k\}$ over the domain $\{[0, B]^K : \sum_k b_k \leq B\}$. It is hence omitted.

Lemma 3. *The weighted sum-rate function in (6) where $b_k = |A|_k$, $A \subseteq E$, $E = \{e_{kb} : i = 1, \ldots, L$ and $b = 1, \ldots, B\}$ is non-decreasing and sub-modular on this set E.*

Comparing the results in Theorems 3 and 5, we see that by assuming less about the exact form of the communication system, we are incurring an added complexity cost of $\mathcal{O}(B\log_2 L)$, while providing a system-independent approximation guarantee of $(1 - \frac{1}{e})$.

Once we solve for b_k^*, we apply a floor operation in order to enforce the integer constraints, i.e., we set $b_{k,INT}^* = \lfloor b_k^* \rfloor$ if $b_k^* \geq 1$ and $b_{k,INT}^* = 0$ if $b_k^* < 1$. This leads us to the task of quantifying loss due to integrality, which we do next. We consider two cases: For $b_k^* \geq 1$, we have that

$$
\begin{aligned}
&\frac{\beta_2(\overline{\mathrm{SNR}}_k)(1-2^{-b_{k,INT}})+\beta_1(\overline{\mathrm{SNR}}_k)2^{-b_{k,INT}}}{\beta_2(\overline{\mathrm{SNR}}_k)(1-2^{-b_k^*})+\beta_1(\overline{\mathrm{SNR}}_k)2^{-b_k^*}} \\
&\geq \frac{\beta_2(\overline{\mathrm{SNR}}_k)(1-2^{-b_k^*+1})+\beta_1(\overline{\mathrm{SNR}}_k)2^{-b_k^*-1}}{\beta_2(\overline{\mathrm{SNR}}_k)(1-2^{-b_k^*})+\beta_1(\overline{\mathrm{SNR}}_k)} \\
&\qquad \text{since } b_k^* - 1 \leq b_{k,INT}^* \leq b_k^* + 1 \\
&\geq \frac{1}{2}\frac{\beta_1(\overline{\mathrm{SNR}}_k)2^{-b_k^*}}{\beta_1(\overline{\mathrm{SNR}}_k)2^{-b_k^*}} \text{ since } 1 \leq b_k^* < \infty \\
&= \frac{1}{2}.
\end{aligned}
\tag{8}
$$

Similarly for $b_k^* < 1$ and $b_{k,INT} = 0$, we have that

$$
\begin{aligned}
&\frac{\beta_2(\overline{\mathrm{SNR}}_k)(1-2^{-b_{k,INT}})+\beta_1(\overline{\mathrm{SNR}}_k)2^{-b_{k,INT}}}{\beta_2(\overline{\mathrm{SNR}}_k)(1-2^{-b_k^*})+\beta_1(\overline{\mathrm{SNR}}_k)2^{-b_k^*}} \\
&\geq \frac{\beta_1(\overline{\mathrm{SNR}}_k)}{\frac{1}{2}\beta_2(\overline{\mathrm{SNR}}_k)+\beta_1(\overline{\mathrm{SNR}}_k)} \text{ since } b_k^* < 1 \\
&= \frac{1}{\frac{1}{2}\frac{\beta_2(\overline{\mathrm{SNR}}_k)}{\beta_1(\overline{\mathrm{SNR}}_k)}+1}.
\end{aligned}
\tag{9}
$$

From (8) and (9), we can compute the approximation factor as

$$
\min\left\{\frac{1}{2}, \frac{1}{\frac{1}{2}\max_k\left\{\frac{\beta_2(\overline{\mathrm{SNR}}_k)}{\beta_1(\overline{\mathrm{SNR}}_k)}\right\}+1}\right\}.
\tag{10}
$$

Thus, the approximation factor critically depends on the ratio $\frac{\beta_2(\overline{\mathrm{SNR}}_k)}{\beta_1(\overline{\mathrm{SNR}}_k)}$, which essentially represents the rate gain due an extra tap or antenna. In Fig. 1, we numerically compute $\frac{\beta_2(\overline{\mathrm{SNR}}_k)}{\beta_1(\overline{\mathrm{SNR}}_k)}$ for a typical cellular range of -15dB to 15dB and see that $\frac{\beta_2(\overline{\mathrm{SNR}}_k)}{\beta_1(\overline{\mathrm{SNR}}_k)} \leq 2$ over this range. Combining the results in Fig. 1 with (10), we can conclude that the proposed relaxation/rounding algorithm has an approximation factor of $\frac{1}{2}$.
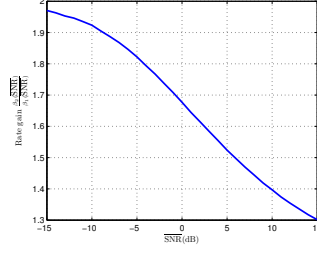
Fig. 1.   Rate gain due to the addition of an extra antenna as a function of $\overline{\mathrm{SNR}}$.

In summary, the two proposed algorithms exploit the structure of the feedback allocation in settings such MISO with quantized beamforming, in delivering lower complexity than the optimal dynamic programming approach accompanied by guarantees on accuracy. Note that the accuracy guarantees, namely, $(1 - \frac{1}{e})$ for the greedy algorithm and $\frac{1}{2}$ for the convex program are independent of any system parameters such as channel statistics, total bit budget B, etc., and are hence, a clear measure of worst-case performance. We now move on to performance evaluation through numerical simulations in the next section, which helps us understand the actual performance against the backdrop of these worst-case guarantees.

VI. NUMERICAL SIMULATIONS

In this section, we evaluate the performance of the greedy feedback allocation algorithm in a MISO downlink network. The simulations serve as a proof of concept for the proposed dynamic feedback allocation approach. As the baseline case, we introduce a static equal allocation algorithm that we describe in detail below along with the rest of the simulation setup. Note that we now revert back (from the virtual system with L users) to the original system with K users, i.e., the indices $k = 1, 2, \ldots, K$, now track actual users.

Number of users, OFDMA bands and data scheduling policy: There are $K = 4$ users in a 10MHz system with a total of $L = 8$ OFDMA sub-bands. Since the focus of these simulations (and this paper) is primarily on quantifying the gains of dynamic feedback allocation, the users are assigned equal amounts of spectrum for data transmission at the beginning of the communication epoch *that do not change with time*, i.e., user i is always assigned to bands $\{2i - 1, 2i\}$.

Small-scale fading, average user SNRs and traffic model: The users are stationary and have fixed average SNRs through the entire epoch of communication. We consider two average SNR

profiles - *(i)* Large asymmetry with average SNRs $10\log_{10}(\overline{\text{SNR}}_1[\bar{t}]) = -10\text{dB}$, $10\log_{10}(\overline{\text{SNR}}_2[\bar{t}]) = -8\text{dB}$, $10\log_{10}(\overline{\text{SNR}}_3[\bar{t}]) = 10\text{dB}$, $10\log_{10}(\overline{\text{SNR}}_4[\bar{t}]) = 10\text{dB}$ and *(ii)* Nearly symmetric with average SNRs $10\log_{10}(\overline{\text{SNR}}_1[\bar{t}]) = -1\text{dB}$, $10\log_{10}(\overline{\text{SNR}}_2[\bar{t}]) = -1\text{dB}$, $10\log_{10}(\overline{\text{SNR}}_3[\bar{t}]) = 1\text{dB}$, $10\log_{10}(\overline{\text{SNR}}_4[\bar{t}]) = 1\text{dB}$. Asymmetric profiles are of interest because this is the regime where dynamic allocation would arguably have most value. The small-scale fading channel realizations $\mathbf{h}$ in each slot are generated according to the standard complex Gaussian distribution. The arrivals are assumed to be deterministic and symmetric with rates $\lambda_k = \lambda$, $\forall k$.

Feedback budget and baseline equal allocation: The feedback budget is set to $B = 12$ bits. The baseline algorithm allocates an equal number of bits to each user, i.e., $b_k = 3$, $\forall k$. Each user in turn distributes these three bits as follows - two bits to the first sub-band it is assigned and one bit to the second. In other words, the per sub-band allocation for user k is $b_{k1} = 2$ and $b_{k2} = 1$. The allocation is changed every $T = 10$ slots.

MISO RVQ codebooks and post-quantization rate: For each bit allocation b, we generate codebook $\mathcal{C}(b)$ by choosing two points uniformly at random from the sphere $\mathbb{C}^2$. For such a codebook $\mathcal{C}(b)$, we compute the ergodic rate over 1000 standard (zero mean, unit variance), complex Gaussian channel realizations. We repeat this experiment over 100 codebooks and choose the codebook $\mathcal{C}^*(b)$ that provides maximum ergodic rate. We repeat this procedure for each $b \in \{0, 1, \ldots, B\}$ and create a *super-codebook* $\{\mathcal{C}^*(0), \ldots, \mathcal{C}^*(B)\}$. Note that the codebook generation procedure is done once at the beginning of the communication epoch. In the previous section, we proposed

$$r_k[\alpha, b] = \beta_2(\overline{\text{SNR}})(1 - 2^{-b}) + \beta_1(\overline{\text{SNR}})2^{-b} \tag{11}$$

as an approximation for the ergodic rate given b bits. In Fig. 2, we compare (11) with the true (numerically computed) ergodic rate given b bits at various $\overline{\text{SNR}}$ values in a typical operational range. We see that (11) is indeed an accurate approximation.

Having described the simulation setup in detail, we now present the results of our experiments. We compare the performance of three algorithms – the greedy dynamic feedback allocation algorithm in Algorithm 2, the equal allocation case, and the case with perfect feedback (i.e., where the bit budget $B = \infty$) – under the two average SNR profiles. The results for SNR profile with large asymmetry are plotted in Fig. 3. In Fig. 3(a), we see that the greedy dynamic algorithm outperforms the static equal allocation approach by almost 13% while consuming only an additional $\frac{\log_2\left(\binom{B+L-1}{L-1}\right)}{T} = 0.88$ bits per second of overhead. Furthermore, greedy dynamic algorithm achieves within 1.5% of the
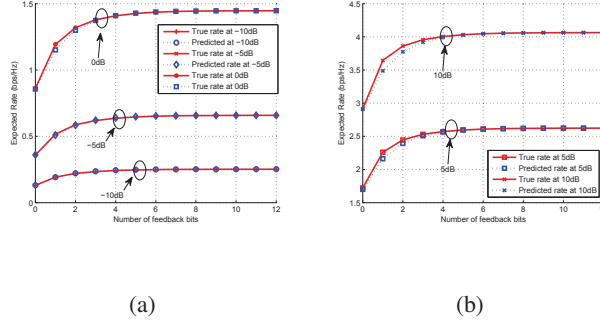
(a)   (b)

Fig. 2.   Comparison of predicted rate (11) with true numerically computed ergodic rate given for codebooks $\{\mathcal{C}^*(b)\}_{b=0}^{B}$ at various values of $\overline{\text{SNR}}$; (a) Low-to-moderate $\overline{\text{SNR}}$ (b) Moderate-to-high $\overline{\text{SNR}}$.

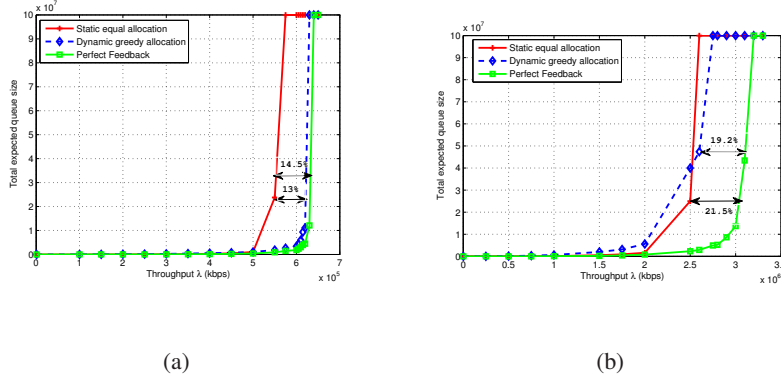(a)   (b)

Fig. 3.   Throughput under the two feedback schemes with different average SNR profiles. The average queue length is measured over 10000 iterations; (a) Large asymmetry with $\overline{\text{SNR}}$ profile $\{-10, -8, 10, 10\}$dB (b) Nearly symmetric case with $\overline{\text{SNR}}$ profile $\{-1, -1, 1, 1\}$dB

optimal dynamic programming allocation[4] thereby rendering the performance guarantee of $\left(1 - \frac{1}{e}\right)$ in Theorem 3 quite conservative.

In the nearly symmetric case however, the gains due to dynamic allocation decrease and almost vanish in the particular case that we consider in Fig.3(b), as would be expected. We see that in this case, the greedy algorithm achieves within 20% of the optimal.

Thus, with minimal expenditure in overhead, the dynamic allocation approach achieves notable gains in throughput for asymmetric settings, thereby showing considerable promise for systems with larger feedback budgets and a greater degree of asymmetry (in traffic loads and channels).

---

[4]The optimal weighted sum-rate is at most as large as the weighted sum-rate with perfect feedback.

VII. Concluding remarks

We summarize the algorithmic contributions presented in Sections IV and V in Table I. We observe from the table that these algorithms explore the tradeoffs between accuracy, computational efficiency and the structure of the weighted sum-rate function. An interesting question and future direction pertaining to the section on MISO systems is whether such an analysis can be extended to cover other commonly-deployed multiple-antenna architectures. Finally, the design of joint data scheduling and feedback allocation policies is another direction for future research.

TABLE I

PROPERTIES OF PROPOSED ONLINE FEEDBACK ALLOCATION ALGORITHMS

Algorithm	Required structure on weighted sum-rate	Complexity	Approx. factor
Dynamic Programming	None	$\mathcal{O}(LB^2)$	1
Greedy	Non-decreasing Sub-modular	$\mathcal{O}((B+L)\log_2 K)$	$\left(1-\frac{1}{e}\right)$
Convex Relaxation	Non-decreasing Sub-modular MISO RVQ Systems	$\mathcal{O}(L\log_2 L)$	$\frac{1}{2}$

In summary, we propose optimal feedback allocation policies for cellular downlink systems where the base station has a limited feedback budget. This problem is solved using dynamic programming incurring pseudo-polynomial complexity in the number of users and the total bit budget. When the weighted sum-rate is a non-decreasing sub-modular function, we leverage the theory of sub-modular function maximization to propose a greedy algorithm with polynomial complexity that has a approximation guarantee of $\left(1-\frac{1}{e}\right)$. For MISO transmit beamforming physical layer communication schemes with quantized beamformer feedback, we recognize that the weighted sum-rate function is non-decreasing and sub-modular for RVQ codebooks. More importantly, it takes a special form that allows us to develop an approximation algorithm based on convex relaxations that can be solved in closed-form, incurring lesser complexity than the greedy algorithm. We connect the performance of the proposed approximate online algorithms to the long-term throughput region of the system.

References

[1] H. Holma and A. Toskala, "LTE for UMTS : OFDMA and SC-FDMA based radio access", *Chichester : John Wiley and Sons, Ltd.*, 2009.

[2] S. Shrivastava and R. Vannithamby, "Performance analysis of persistent scheduling for VoIP in WiMAX networks", *Proc. of IEEE Wireless and Microwave Technology Conference (WAMICON) 2009*, pp. 1 - 5, Apr. 2009, Clearwater, FL.

[3] H. Wang and D. Jiang, "Performance comparison of control-less scheduling policies for VoIP in LTE UL", *Proc. Wireless Communications and Networking Conference (WCNC) 2008*, pp. 2497 - 2501, Las Vegas, NV, Apr. 2008.

[4] H. Jin, C. Cho, N-O. Song and D. K. Sung, "Optimal rate selection for persistent scheduling with HARQ in Time-Correlated Nakagami-m Fading Channels", *IEEE Trans. Wireless Commun.*, pp. 637 - 647, vol. 10, Feb. 2011.

[5] Li, W.W.-L. Ying Jun Zhang So, A.M.-C. Win, M.Z., "Slow adaptive OFDMA systems through chance constrained programming", *IEEE Trans. Sig. Proc.*, vol. 58, pp. 3858 - 3869, July 2010.

[6] 3GPP, "3GPP TS 36.300-870", 3rd Generation Partnership Project.

[7] D. J. Love, R. W. Heath Jr., V. K. N. Lau, D. Gesbert, B. D. Rao and M. Andrews, "An overview of limited feedback in wireless communication systems", *IEEE Journ. Sel. Areas Commun.*, vol. 26, pp. 1341-1365, Oct. 2008.

[8] R. W. Heath, Jr., T. Wu, and A. C. K. Soong, "Progressive refinement for high resolution limited feedback beamforming", *EURASIP Journal on Adv. in Sig. Proc., special issue on Multiuser Limited Feedback*, vol. 2009, Article ID 463823, 2009.

[9] L. Liu and H. Jafarkhani, "Novel transmit beamforming schemes for time-selective fading multiantenna systems", *IEEE Trans. Sig. Proc.*, vol. 54, pp. 4767-4781, 2006.

[10] L. Liu and H. Jafarkhani, "Successive transmit beamforming algorithms for multiple-antenna OFDM systems", *IEEE Trans. Wireless Commun.*, vol. 6, pp. 15121522, 2007.

[11] S. Donthi and N. Mehta, "Joint performance analysis of channel quality indicator feedback schemes and frequency-domain scheduling for LTE", *IEEE Trans. Vehicular Tech.*, vol. PP, pp. 1-13, 2011.

[12] M. -O. Pun, J. K. Kyeong and H. V. Poor, "Opportunistic scheduling and beamforming for MIMO-OFDMA downlink systems with reduced feedback", *IEEE Intern. Conf. on Communications (ICC)*, Bejing, China, May 2008.

[13] D. Gesbert and M.-S. Alouini, "How much feedback is multiuser diversity really worth?", *in Proc. of Intern. Conf. on Communications (ICC)*, pp. 234-238, Jul. 2004.

[14] J. Chen, R. A. Berry, and M. L. Honig, "Limited feedback schemes for downlink OFDMA based on sub-channel groups", *IEEE Journ. Sel. Areas Commun.*, vol. 26, pp. 1451-1461, Oct. 2008.

[15] R. Agarwal, V. Majjigi, Z. Han, R. Vannithamby and J. Cioffi, "Low complexity resource allocation with opportunistic feedback over downlink OFDMA networks", *IEEE Journ. Sel. Areas Commun.*, vol. 26, pp. 1462-1472, Oct. 2008.

[16] J. Kleinberg and E. Tardos, "Algorithm Design", *Addison Wesley*, 2006.

[17] A. D. Dabbagh and D. J. Love, "Feedback rate-capacity loss tradeoff for limited feedback MIMO Systems", *IEEE Trans. Inform. Theory*, vol. 52, pp. 2190-2202, May 2006.

[18] D. J. Love, R. W. Heath Jr. and T. Strohmer, "Grassmannian beamforming for multiple-input-multiple-output wireless systems", *IEEE Trans. Info. Theory*, vol. 49, pp. 2735-2747, Oct. 2003.

[19] P. R. Goundan and A. S. Schulz, "Revisiting the greedy approach to submodular set function maximization", Jan. 2009.

[20] G. Calinescu, C. Chekuri, M. Pal, J. Vondrak, "Maximizing a submodular set function subject to a matroid constraint (Extended Abstract)", *Lecture Notes In Computer Science, Proc. 12th Intern. Conf. Integer Prog. and Comb. Optimization*, vol. 4513, pp. 182 - 196, Ithaca, NY, 2007

[21] J. Vondràk, "Submodularity in combinatorial optimization", *Ph.D. thesis*, Charles University, Prague, 2007.

[22] G. L. Nemhauser and L. A. Wolsey, "Best algorithms for approximating the maximum of a submodular set function", *INFORMS*, vol. 3, pp. 177-188, Aug. 1978.

[23] M. Ouyang and L. Ying, "On scheduling in multi-channel wireless downlink networks with limited feedback", *Proceedings of the Allerton Conference on Communication, Control, and Computing*, Monticello, IN, Oct. 2009.

[24] C. Joo, X. Lin, and N. B. Shroff, "Greedy maximal matching: Performance limits for arbitrary network graphs under the node-exclusive interference model", *IEEE Trans. Auto. Control*, vol. 54, no. 12, pp. 2734–2744, Dec. 2009.

[25] A. Gupta, X. Lin and R. Srikant, "Low-complexity distributed scheduling algorithms for wireless networks", *IEEE/ACM Trans. Networking*, vol. 17, pp. 1846-1859, Dec. 2009.

[26] J. Huang, V. G. Subramanian, R. Agarwal, and R. A. Berry, "Downlink scheduling and resource allocation for OFDM systems", *IEEE Trans. Wireless Commun.*, vol. 8, pp. 288-296, Jan. 2009.

[27] C. W. Tan, M. Chiang and R. Srikant, "Fast algorithms and performance bounds for sum rate maximization in wireless networks", *IEEE Infocom 2009*, Rio De Janeiro, Brazil, Apr. 2009.

[28] C. W. Tan, M. Chiang and R. Srikant, "Maximizing sum rate and minimizing MSE on multiuser downlink: Optimality, fast algorithms, and equivalence via max-min SINR", *IEEE Trans. Sig. Proc.*, vol. 59, No. 12, pp. 6127-6143, Dec. 2011.

[29] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks", *IEEE Trans. Auto. Control*, vol. 37, pp. 1936-1949, Dec. 1992.

[30] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, R. Vijayakumar and P. Whiting, "Scheduling in a queuing system with asynchronously varying service rates", *Probability in the Engineering and Informational Sciences*, vol. 18, pp. 191-217, Apr. 2004.

[31] A. L. Stolyar, "On the asymptotic optimality of the gradient scheduling algorithm for multiuser throughput allocation", *INFORMS*, vol. 53 , Issue 1, pp. 12-25, Jan. 2005.

[32] H. Holma and A. Toskala, "WCDMA for UMTS: Radio access for third generation mobile communications", *Revised Edition. New York: John Wiley & Sons*, 2001.

[33] V. Raghavan, M. L. Honig, V. V. Veeravalli, "Performance analysis of RVQ-based limited feedback beamforming codebooks", *Proc. IEEE ISIT 2009*, pp. 2437-2441, Seoul, Korea, June 2009.

[34] C. K. Au-Yeung and D. J. Love, "On the performance of random vector quantization limited feedback beamforming in a MISO System", *IEEE Trans. Wireless Commun.*, vol. 6, pp. 458-462, Feb. 2007.

[35] D. Gesbert, M. Shafi, D.-S. Shiu, P. J. Smith, and A. Naguib, "From theory to practice: an overview of MIMO space-time coded wireless systems," *IEEE Journ. Sel. Areas Commun.*, vol. 21, no. 3, pp. 281-302, 2003.

APPENDIX A

PROOF OF THEOREMS 4-5

Proof of Theorem 4: The objective function is clearly convex since 2^{-b_k} is convex. By studying (7) closely, we can also say that b_k^* is such that $\sum_{k=1}^{L} b_k^* = B$ since if this not true, we can increase the bit allocation for at least one user thereby decreasing the objective function. Since $B > 0$, $b_k = 0, \forall k$ is in the interior of our constraint set $\mathcal{B}$, which implies that Slater's constraint for strong duality is satisfied and that the Karush-Kuhn-Tucker (KKT) conditions are sufficient in nature. The Lagrangian cost function can be written as $\mathcal{L}(b_k, \lambda_k, \eta) = -\sum_{k=1}^{L} q_k r_1(\text{SNR}_k) 2^{-b_k} - \lambda_k b_k + \eta \left(\sum_k b_k - B \right)$ for which the KKT conditions are $b_k^* \geq 0$, $\lambda_k^* \geq 0$, $b_k^* \lambda_k = 0$, and $\eta^* = q_k r_1(\text{SNR}_k)(\log 2) 2^{-b_k} + \lambda_k^*$. Since 2^{-b} is a decreasing function in b, it follows that if $\eta^* \leq q_k r_1(\text{SNR}_k)(\log 2)$, then $\lambda_k^* = 0$ and $b_k^* = -\log_2 \left(\frac{\eta^*}{q_k(\log 2)} \frac{1}{r_1(\text{SNR}_k)} \right)$ is a valid solution to (7). If $\eta^* > q_k r_1(\text{SNR}_k)(\log 2)$, $\lambda_k^* = \eta^* - q_k r_1(\text{SNR}_k)(\log 2)$ and $b_k^* = 0$. Hence, we can write the solution as $b_k^* = \left[-\log_2 \left(\frac{\eta^*}{q_k(\log 2)} \frac{1}{r_1(\text{SNR}_k)} \right) \right]^+$ where η^* is chosen such that $\sum_k b_k^* = B$.

Proof of Theorem 5: In order to compute the solution in Theorem 4, we first need to sort $\{\theta_k\}_{k=1}^{L}$ in ascending order where $\theta_k = q_k r_1(\text{SNR}_k)(\log 2)$. This has complexity $\mathcal{O}(L\log_2 L)$. Call this sorted set $\{\theta_m\}$. Once sorted, we need to set $\eta^* = \theta_m$ for each m and test feasibility. Testing feasibility incurs $\mathcal{O}(L)$, as it is a L-term addition and scanning through each θ_m incurs $\mathcal{O}(\log_2 L)$ through the use of binary search. As we increase η^*, more b_m^* terms are set to zero. Once we locate m_1 and m_2 such that $\eta^* = \theta_{m_1}$ is infeasible while $\eta^* = \theta_{m_2}$ is feasible, we can compute η^* in closed-form since it satisfies $\sum_{m \geq m_2} b_m^* = B$. Hence, the total complexity is $\mathcal{O}(L\log_2 L) + \mathcal{O}(L\log_2 L) = \mathcal{O}(L\log_2 L)$.

Harish Ganapathy (M'12) received his B.S. and M.S. degrees in electrical engineering from the State University of New York at Buffalo in May 2003 and May 2005, respectively. He received his Ph.D. in electrical engineering from The University of Texas, Austin in August 2011. His research interests lie broadly in optimization as applied to wireless networks, including both physical layer and networking aspects.

Siddhartha Banerjee (S'09) received a B.Tech in electrical engineering from the Indian Institute of Technology, Madras in May 2007 and an M.S. in electrical and computer engineering from The University of Texas, Austin in December 2009. He is currently pursuing his Ph.D. at the latter. He briefly worked as a visiting researcher at Bell Labs, Alcatel-Lucent in 2009, and at the Technicolor Paris Research Lab in 2011. His research focuses on the learning and control of information flows in networks.

Nedialko B. Dimitrov is an Assistant Professor in the Operations Research Department of the Naval Postgraduate School. He received a Ph.D. in theoretical computer science from the University of Texas at Austin (2008) and B.S. degrees in mathematics and computer science from the University of Michigan, Ann Arbor (2002). Prior to joining the Naval Postgraduate School, he held a two year postdoctoral fellowship in the Operations Research Program of the University of Texas at Austin. His research interests are in network modeling and stochastic combinatorial optimization, with applications in infectious disease control and national security.

Constantine Caramanis (M'06) received his Ph.D. in electrical engineering and computer science from the Massachusetts Institute of Technology in 2006. Since then, he has been on the faculty in the Department of Electrical and Computer Engineering at The University of Texas at Austin. He received the NSF CAREER Award in 2011. His current research interests include robust and adaptable optimization, machine learning and high-dimensional statistics, with applications to large scale networks.