Measurement-Based Opportunistic Scheduling for Heterogenous Wireless Systems

Shailesh Patil and Gustavo de Veciana
{patil, gustavo}@ece.utexas.edu

Abstract— **We study the performance of measurement-based opportunistic scheduling strategies for wireless systems in practical scenarios where user's heterogenous capacity distributions are unknown. We make the case for using *maximum quantile scheduling*, i.e., scheduling a user whose current rate is in the highest quantile relative to its current *empirical distribution.* Under the assumption of fast fading, we prove a bound on the relative penalty associated with such estimates, showing that number of independent samples need only grow linearly with the number of active users. This is a fairly limited cost, suggesting one could track distributional changes in users' channels. By contrast other opportunistic scheduling schemes require estimating or setting weights/thresholds that implicitly depend on the number of users, their channel distributions, and possibly their traffic characteristics and/or are queue dependent. Our results show that it is easier to estimate users' distributions than to infer good weights, and that maximum quantile scheduling is more robust to changes in the activity levels of users and/or changes in the number of users. This allows it to maintain opportunism without loss in performance in dynamic and/or unsaturated regimes. In addition, for a saturated regime, we show that maximum quantile scheduling not only maximizes 'opportunism,' but if rates are bounded and number of users is high, it is sum average throughput maximizing subject temporal fairness. Furthermore we show that the distributions for the vector of rates allocated to various users on a typical slot by maximum quantile cannot be stochastically dominated by any other non idling scheduler. As such our analysis and simulations suggest that maximum quantile scheduling might provide the best features both in terms of performance and robustness for practical scenarios.**

I. Introduction

Motivation. The scheduling of users' data transmissions at a wireless access point has recently attracted a substantial amount of attention, see e.g., [6][19][4]. A key feature of wireless systems relative to the traditional wireline systems is that, the channel capacity, or service rate, may exhibit temporal variations. This allows one to consider scheduling policies that choose to send to, or receive from, a user (or a subset of users) which at a given point in time has (have) the 'best', e.g., highest, capacity. Such 'opportunistic scheduling' can lead to good increases in the aggregate capacity of a wireless system, and has thus been adopted in various wireless standards such as CDMA-HDR, HSDPA [2][1], and will almost certainly play a role in future wireless systems.

In practice users' channel capacity variations are unknown and heterogenous, e.g., users close to an access point see significantly different channel capacity than those further off. Thus it is important to devise opportunistic schedulers that do not starve some users, e.g., those with poor channels, to achieve some degree of fairness among users sharing an access point. To this end many opportunistic scheduling schemes have been devised that make decisions by selecting the user that currently has the highest weighted channel capacity. In practice the weights may be hard to determine, because they depend in a complex way on the users' channel capacity distributions, the number of users, and the characteristics of their traffic. Thus they either need to be estimated or tuned based on the service users have received or their queue lengths.

Unfortunately, the complex dependence of weights may make them very sensitive to changes in the system, i.e., if a user's traffic characteristics changes, or a user leaves or enters the system (e.g., a mobile user comes out of the shadow of a building), or the channel characteristics of a user change, then the weights associated with *all* users may need to change. Therefore, it is likely that a significant fraction of time will be spent in estimating/tuning weights to their 'ideal' values. In fact, if the system is dynamic enough and/or the tuning algorithm is not sensitive enough, one may never converge, possibly compromising fairness but also, and more importantly leading to poor throughput performance. Consider a simple example. Due to the stochastic or time varying nature of channel capacity and user's traffic a measurement-based opportunistic scheduler may be biased in favor of a user who has not received service in the recent past or one that currently has a high queue. While, this myopic approach is good for short term fairness, the scheduler may end up serving a user even though it is not currently experiencing a high channel rate. This in turn decreases the achieved opportunism and long term throughput the system can sustain. In heavily loaded systems, at a given moment of time, it is very likely that there exists a group of users which are starved. If those users are served, others may become starved, leading to a cycle, in which the level of opportunism and throughput are low. In this paper we will see that indeed the performance of many proposed opportunistic scheduling schemes in such regimes are subject to such performance penalties.

Recently, distribution based opportunistic schedulers have been proposed by several researchers under different guises [9][10][3][14]. In this paper, we shall refer to this family of schemes as *maximum quantile schedulers*. The idea is to schedule a user whose current rate is highest relative to his *own* distribution, i.e., in the highest quantile. As will be explained in the sequel because the quantile of each users' rate is uniformly distributed, maximum quantile scheduling is automatically temporally fair – i.e., no weights required to achieve fairness. However, in practice maximum quantile scheduling would involve estimating each user's channel capacity distribution. In this paper we will show that the throughput penalty incurred from estimating user's distributions can be limited. Furthermore, unlike other schemes, maximum quantile does not require estimation/tuning of weights which depend on users' joint channel capacity distributions, and so it is robust to fast changes in the number of users or their activity levels. In other words the performance penalties associated with estimation/tuning are substantially reduced.

Contributions. The following is the list of the key contributions of this paper:

- We investigate the throughput performance of maximum quantile scheduling and show that if the achievable instantaneous rate of users' is bounded, then among the class of scheduling policies that serve each user an equal fraction of time, maximum quantile scheduling maximizes the long term system throughput when there is a large number of users. Furthermore, we show that the marginal distribution for the rate when users are selected for service under maximum quantile scheduling can not be stochastically dominated by any other non-idling scheduler.
- Under the assumption of fast fading, we prove a bound on each user's relative throughput penalty when maximum quantile scheduling is based on empirical distributions for users' channel capacity. This is significant because it shows that such penalties can be controlled if the number of independent samples used to estimate the empirical distribution is roughly proportional to the *number of users in the system*. Thus maximum quantile scheduling can be used even when users' channel distributions are not known or slowly changing.
- We conjecture that the best way to serve a user is to serve it when its rate is high compared to its distribution, rather than favoring a user that has not been served for some period of time. This conjecture is supported by simulating the performance of various measurement based opportunistic scheduling schemes for various network and traffic scenarios. We find that maximum quantile scheduling can have significantly better performance in terms of both packet delay and file transfer delay, e.g., up to 40% improvement. We stress that this observation has not been made by previous works.

Paper Organization. This paper is organized as follows. In Section II we revisit and critique representative prior work in the area of opportunistic scheduling and introduce some known features of maximum quantile scheduling. Throughput performance and optimality of maximum quantile scheduling is studied in Section III. We prove our bound on the relative throughput penalty associated with measuring distributions in Section IV. Simulation results comparing the performance of maximum quantile scheduling to other schemes are presented in Section V and VI. Section VII concludes the paper.

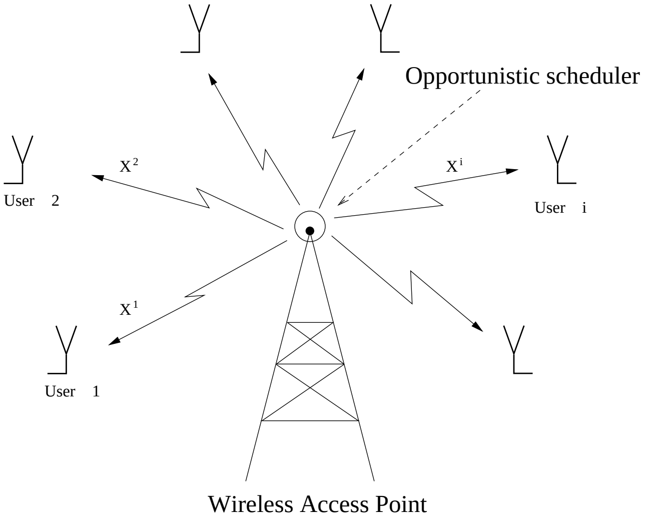

Fig. 1. Downlink scheduling to users from a wireless access point.

II. Revisiting Opportunistic Scheduling

A. System Model and Notation

We begin by introducing our system model and some notation. For simplicity, we focus on downlink scheduling from an access point to multiple users (see Figure 1). Suppose time is divided into equal sized slots and at most one user gets served per slot, e.g., for the CDMA-HDR systems defined in the CDMA2000 IS-856 standard, the slot time has a duration of 1.67 ms [2]. During each slot, each user feeds back the data rate it can support and the access point makes a decision on which user should get served. In the sequel we use the terms 'channel capacity' and 'rate' interchangeably and make the following assumption on user's channel characteristics over time slots.

For analysis purposes, we make the following assumptions on users' channel capacity distribution(s) across slots for this section of the paper.

Assumption 2.1: We assume the channel capacity (rate) for each user is a stationary ergodic process and these processes are independent across users. Further we assume that the marginal distribution for each user is continuous and is either known a priori, or estimated by the access point.

Discussion on the assumption. First the channel capacity distributions seen by users might indeed be roughly stationary over a reasonable period of time particularly if users are at fixed locations. As will be discussed later, we conjecture that the channel should remain stationary for roughly $O(n^2)$ (here n is the number of users in the system) samples for the result on measurement based performance proved in Section IV to be practically viable. The assumption that users' rates are independent is also likely to be true, though a notable exception is the case where mobile users move in a correlated manner, e.g., along a highway. The assumption that the access point knows, and in particular can estimate, the marginal distributions of the channel capacity processes may seem unreasonable, but simple book keeping on the users' feedback of the currently

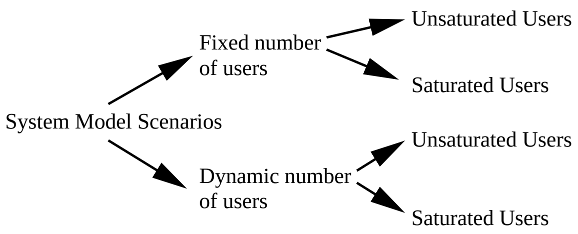

Fig. 2. Different scenarios for system model.

achievable rate can be used to estimate distributions. We will discuss estimation of such distributions in Section IV. Note that channel capacities are not restricted to any specific distribution, or class of distributions, i.e., users can undergo any fading process. This makes the analysis presented later applicable to real world scenarios. Note that the we require the marginal distributions of rates to be continuous only for simplicity sake, the results presented here can be extended to the discrete case.

Notation. In the sequel we will let $x^i(t)$ denote the realization of the channel capacity of user i at time slot t, and let X^i be a random variable whose distribution is that of the channel capacity of user i on a *typical* slot. Recall that we will be assuming X^i to be independent across users but need not be identically distributed. We denote the distribution function of X^i by $F_{X^i}(\cdot)$. For simplicity, we will assume that $F_{X^i}(\cdot)$ is a strictly increasing function, so that its inverse denoted by $F^{-1}_{X^i}(\cdot)$ is defined.

System Scenario. There are several system scenarios (Figure 2) one can consider. In a real world scenario, the number of users in the system may be changing, and users may not be infinitely backlogged, i.e., unsaturated dynamic. However, such a scenario is analytically intractable, therefore we usually study different idealizations. The first idealization is the 'fixed saturated' case, where the number of users in the system does not change with time and each user is infinitely backlogged. Such a scenario is an approximation where the number of users in the system changes slowly and packet queues for each user are always non empty at the access point. This idealization is often studied in literature, and we will largely focus on this case. We will also perform some simulations in the 'fixed unsaturated' and 'dynamic saturated' case, the former referring to the scenario where even though the number of users remain static, they are not necessarily backlogged, and the later refer to the scenario where the number of users changes with time, but whenever a user is present, it is infinitely backlogged.

We denote the number of users present in the system on slot t by $n(t)$. We simplify this to n in a fixed system (saturated or unsaturated) since the number of users is constant. The set of active, i.e., backlogged users on slot t is denoted by $A(t)$. In other words, $A(t)$ is the set of users that wish to be served on slot t. Note that in a dynamic saturated system $|A(t)| = n(t)$, while in fixed saturated system $|A(t)| = n$.

B. Weight based Opportunistic Schemes

Opportunistic scheduling was first proposed in [6]. They proposed *maximum rate scheduling*, where the user with maximum channel capacity at that point of time is served, i.e., user $k(t)$ is selected for service on time slot t if

$$k(t) \in \arg\max_{i \in A(t)} x^i(t).$$

This maximizes system throughput in a fixed saturated system, but in a system where users have heterogenous rate distributions, may neglect those with poor channels.

Subsequently a myriad of approaches have been proposed to address both unfairness and/or performance issues. One of the more cited schemes is *proportional fair scheduling* [5][20][4] which serves the user whose current rate normalized by a moving average of his allocated rate is the highest, i.e., user $k(t)$ is selected for service at time slot t if

$$k(t) \in \arg\max_{i \in A(t)} \frac{x^i(t)}{\mu^i(t)}, \tag{1}$$

where

$$\mu^i(t) = (1 - \frac{1}{t_c})\mu^i(t-1) + \frac{1}{t_c} x^i(t) \mathbf{1}_{S^i_{pf}(t)}$$

and t_c is the moving average parameter, $S^i_{pf}(t)$ is the event that user i gets served on slot t by the scheme, and $\mathbf{1}_{S^i_{pf}(t)}$ is the indicator function of $S^i_{pf}(t)$.

As a simple experiment we compare the throughput achieved by proportionally fair to that achieved by maximum quantile scheduling (described in the next subsection) in a fixed saturated system. Our setup consists of two classes of users having a mean signal to noise ratio (SNR) of 2 and 0.1, with both classes experiencing Rayleigh fading and containing an equal number of users. The channel capacity for all users is fast fading, i.e., rate supported by users is independent across slots, and the slot size is set to 1.67 msec. The bandwidth associated with each user is 500 KHz and we assume that coding achieves the Shannon rate. This setup will be used throughout the paper for simulations, and unless specified otherwise, both classes will contain an equal number of users.The parameter t_c is set equal to 1000 slot length [5].

Figure 3 exhibits the ratio of the per class throughput achieved by maximum quantile scheduling versus that achieved by proportional fair based on allocated rate for an increasing number of users. As can be observed, maximum quantile achieves a 50% gain in throughput with less than 8 users in the system for both the classes, and the gain exceeds 100% for a larger number of users. *This clearly illustrates that scheduling based on the recent service given to a user can lead to large loss in opportunism.*

Queue based opportunistic scheduling schemes that factor the magnitude of ongoing users' queue lengths (as a measure of recent service given to a user) and their channel capacity in deciding which to serve have also been proposed in the literature. For example, the *exponential rule* [19][18][17], chooses to serve user $k(t)$ on slot t if

$$k(t) \in \arg\max_{i \in A(t)} [\gamma^i x^i(t) \exp(\frac{a^i q^i(t)}{1 + \sqrt{\bar{q}(t)}})],$$

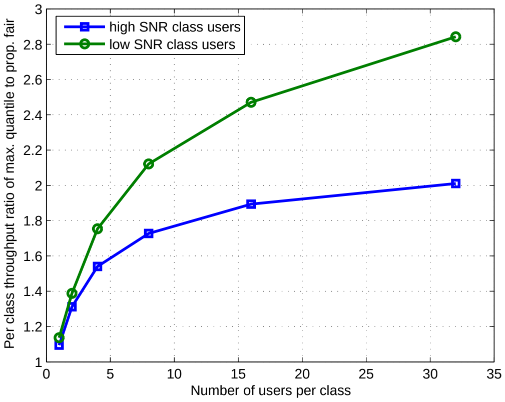

Fig. 3. Ratio of per class throughput achieved by maximum quantile scheduling to that achieved by proportional fair.

where $q^i(t)$ is the queue length of user i at time t, a^i is the weight associated with user i's queue, $\bar{q}(t)$ is the average weighted queue length across users at time t, and γ^i is the weight associated with user i's channel rate $x^i(t)$. Factoring users' queue length has the potential advantage of reducing packet delays. Indeed, it has been shown in [17] that under heavy traffic scenario, the exponential rule will minimize the maximum of weighted queue length, i.e., $a^i q^i$. Good packet delay performance of the rule has been supported by simulations shown in [18]. We will revisit this point in Section V, and show that in practice, not unlike proportional fair, such queue based schemes introduce biases that may compromise opportunism thus compromising packet delay performance.

Finally, [7] proposed strategies that maximize system throughput under fairness constraints. For example, they show that a scheduling policy of the form

$$k(t) \in \arg\max_{i \in A(t)} [x^i(t) + \nu^i], \tag{2}$$

maximizes the overall sum/system throughput subject to constraints on the fraction of time each user i is served in a fixed saturated regime. Here ν^i is a weight associated with user i that ensures that users get served the desired fraction of time. Similar optimal schemes were proposed for rate and utility based fairness.

While the fairness and optimality characteristics of these schemes are desirable, in practice they would require estimating thresholds ν^i which are complicated functions of users' rate distributions, number of users and temporal constraints. In the sequel (Section IV and VI), we show that such estimates may converge slowly and are not robust to changes in unsaturated and/or dynamic regime.

C. Maximum Quantile Scheduling

Maximum quantile scheduling has been proposed independently by several researchers. Specifically [9][10] proposed a 'CDF based scheme'. While [3] proposed a so called 'score based scheduler' and [14] proposed a 'distribution fairness' based scheduler. We have studied the properties of maximum quantile scheduling under greedy user behavior in [13], and in [12], we evaluate its use to achieve quality of service guarantees for real-time traffic.

Let us briefly introduce this scheme in the fixed saturated regime. The main idea is to schedule a user whose rate is highest compared to his *own* distribution, i.e., serve user $k(t)$ during slot t if

$$k(t) \in \arg\max_{i=1,\ldots,n} F_{X^i}(x^i(t)). \tag{3}$$

It is well known that $F_{X^i}(X^i)$ is uniformly distributed on $[0,1]$. Let $U^i = F_{X^i}(X^i)$, then U^i is also uniformly distributed on $[0,1]$. Maximum quantile can be thought of as picking the maximum among independent realizations of users' (i.i.d.) U^i's on every slot. Thus, it is clear that maximum quantile is equally likely to serve any user on a typical slot, and as a result all users get served an equal fraction of time, i.e., $\frac{1}{n}^{th}$ of time.

Let $U^{(n)} = \max[U_1, \ldots, U_n]$, where U_j is uniformly distributed on $[0,1]$ $\forall j = 1, \ldots, n$, then

$$\Pr(U^{(n)} \le u) = u^n, \ \forall u \in [0,1]. \tag{4}$$

Let S^i_{mq} be the event that user i is the selected for service on a typical slot under maximum quantile scheduling. Then the rate distribution seen by user i on a slot that it gets served is the same as $F^{-1}_{X^i}(U^{(n)})$. Therefore, the average throughput seen by user i is given by $G^i_{mq}(n)$[9][10],

$$G^i_{mq}(n) = \frac{E[F^{-1}_{X^i}(U^{(n)})]}{n} = \frac{E[X^{i,(n)}]}{n}.$$

where $X^{i,(n)}$ is maximum of n i.i.d. copies of X^i, i.e., $X^{i,(n)} := \max[X^i_1, \ldots, X^i_n]$, where $X^i_j \sim X^i$, $\forall j = 1, \ldots, n$. Note that by contrast, with the schemes discussed in the previous subsection, if the users' rate distributions were known, it is fairly easy to evaluate the individual and system throughput for maximum quantile scheduling.

Maximum quantile scheduling can be modified to serve users different fractions of time using easily tuned (distribution independent) weights, see [9][10] for details.

It is clear that maximum quantile scheduling has some very desirable properties: e.g., it is temporally fair, it is amenable to performance prediction in the fixed saturated case, and Figure 3 indicates that it has good throughput performance. However, as discussed earlier, it is unlikely that rate distributions will be known. It is unclear how maximum quantile's performance compares to that of other schemes when distributions are estimated, especially in scenarios other than fixed saturated. In the sequel we will address these issues.

III. Performance of Maximum Quantile Scheduling in Fixed Saturated System

In this section, we look at two metrics to study the performance of maximum quantile scheduling : (1) the amount of opportunism exploited by the scheme, (2) the throughput achieved by the scheme.

'Opportunistically' Optimal. Suppose we consider as measure of opportunism achieved by user i as the quantile of the rate achieved by the user, i.e., $F_{X^i}(x^i(t))$ whenever it is served. A high quantile means a high degree of opportunism and $E[\sum_{i=1}^n F_{X^i}(X^i)\mathbf{1}_{S^i_\beta}]$ denotes the overall expected opportunism realized by a scheduling scheme β. (Here S^i_β is the event that user i is selected for service on typical slot by β.) It should be clear that maximum quantile scheduling maximizes the system opportunism, and does so while serving all users an equal fraction of time.

Not Stochastically Dominated. Maximum quantile scheduling has an optimality in terms of the rates seen by users in the typical slots in which they are served. Let us first introduce the concept of stochastic dominance, before presenting the bound. We say that a random variable Y stochastically dominates random variable V, if $\forall v$, $\Pr(Y > v) \geq \Pr(V > v)$. This is denoted as $Y \geq^{st} V$ and it follows that for any increasing function $g(\cdot)$, we have that $g(Y) \geq^{st} g(V)$.

Let R^i_{mq} represent the rate distribution seen by user i when selected for service on a typical slot by maximum quantile scheduling, and let $\overrightarrow{R}_{mq} = (R^1_{mq}, \ldots, R^n_{mq})$, i.e., the vector of random variables representing the rate distributions. Let $\overrightarrow{R}_\beta = (R^1_\beta, \ldots, R^n_\beta)$ be the same quantity for another distinct non idling scheduling scheme β that may *not* serve all users an equal fraction of time. By distinct we mean that the scheme does not always pick the user with the maximum quantile, and by non idling, we mean that the scheme will never be idle as long as there is at least one backlogged user. Then our claim is that $\overrightarrow{R}_\beta \not\geq^{st} \overrightarrow{R}_{mq}$, i.e., $\exists j$, such that $R^j_\beta \not\geq^{st} R^j_{mq}$. This is formally stated in the following theorem with the proof given in Appendix I.

Theorem 3.1: Consider a fixed saturated system with n users, whose channel capacity variations satisfy Assumption 2.1. Then for any distinct non idling scheme β,

$$\overrightarrow{R}_\beta \not\geq^{st} \overrightarrow{R}_{mq}.$$

Note that a scheduling scheme γ is known to be Pareto optimal if there exists no other scheduling scheme that is able to give an equal or higher average throughput to *all* the users than that received by users under γ. Theorem 3.1 can be thought to be a weak form of Pareto optimality in terms of rate seen in a typical slot, not average throughput. We will next show that maximum quantile is not Pareto optimal in terms of average throughput.

Not Pareto Optimal. We illustrate this with a simple pathological example where users' support only discrete rates. (The example can be extended to the continuous case.) Consider a two user system with ON-OFF channels. The ON and OFF channel states correspond to rates 1 and 0 respectively. User 1 and 2 have an ON probability of 0.6 and 0.4 respectively. Here maximum quantile will serve User 1 a rate of 0.42, and User 2 a rate of 0.32. However, it can be shown that maximum quantile may sometimes serve User 2 in OFF state, even though User 1's channel is ON. Therefore, it is possible to improve performance while still serving each user an equal fraction of time. Consider a scheme that always serves the user with the highest instantaneous rate and breaks ties $\frac{7}{24}^{th}$ of times in favor of User 1. Such a scheme will give User 1 a rate of 0.43, and User 2 will get a rate of 0.33. Hence one can give better performance to both the users, while maintaining temporal fairness.

Throughput Optimal for Large Number of Users. Even though the maximum quantile is not Pareto optimal in general, it does achieve good system throughput performance. If the rates achievable by users in a system are bounded, then maximum quantile scheduling is sum throughput optimal among policies that serve all users an equal fraction of time as the number of users increases. Following lemma is useful to prove this claim.

Lemma 3.2: Consider a fixed saturated system with n users, whose channel capacity variations satisfy Assumption 2.1 and served based on maximum quantile scheduling. Let $\epsilon, \delta \in (0,1)$, then there exists $n_{\epsilon,\delta}$ such that if $n > n_{\epsilon,\delta}$ at any slot where user k gets scheduled for service, the user sees a rate exceeding $F^{-1}_{X^k}(1-\delta)$ with probability greater than $1-\epsilon$.

Proof: As discussed in the previous section, whenever user k gets served under maximum quantile scheduling, it sees a rate $F^{-1}_{X^k}(U^{(n)})$. In order to ensure the desired condition is satisfied we require that

$$\Pr(F^{-1}_{X^k}(U^{(n)}) > F^{-1}_{X^k}(1-\delta)) > 1-\epsilon.$$

Since $F^{-1}_{X^k}(\cdot)$ is an increasing function, the above inequality can be rewritten as

$$\Pr(U^{(n)} > (1-\delta)) > 1-\epsilon.$$

From (4), we get

$$1-(1-\delta)^n > 1-\epsilon.$$

Simplifying and taking log, we get

$$n > \frac{\ln \epsilon}{\ln(1-\delta)}.$$

Defining

$$n_{\epsilon,\delta} = \lceil \frac{\ln \epsilon}{\ln(1-\delta)} \rceil,$$

we have that for any $n \geq n_{\epsilon,\delta}$, whenever user k is served, it will experience a rate greater than $F^{-1}_{X^k}(1-\delta)$ with probability greater than $1-\epsilon$. ■

The following theorem follows from Lemma 3.2 and formally states our claim.

Theorem 3.3: Consider a fixed saturated system with n users, whose channel capacity variations satisfy Assumption 2.1 and are served using maximum quantile scheduling. Suppose each user i has a maximum instantaneous rate of $r^i_{max} < \infty$. Then as $n \to \infty$, each user is likely to be served at his maximum rate, so maximum quantile scheduling is sum throughput optimal.

Summarizing, we observe that even though maximum quantile scheduling is not Pareto optimal, it is likely to give a good throughput performance.

IV. Penalty due to Measurement

We now focus on the measurement aspects of opportunistic scheduling. We will first consider the throughput penalty incurred by maximum quantile scheduling due to estimation of rate distributions of users under fast fading, and present simulation results for the slow fading case. Following this, we will compare the penalty incurred by maximum quantile to that incurred by sum throughput optimal scheme in (2), via simulations.

A. Maximum Quantile Scheduling based on Empirical Distributions

Assumption 2.1 required that the channel capacity distribution, i.e., $F_{X^i}(\cdot)$ of each user be known at the access point. This is unlikely, and in this subsection we consider the penalty in throughput seen by users in a n user fixed saturated system due to such mistakes by the scheduler.

Suppose the quantile of the current rate of a user is estimated using the previous m samples of the user's rate. The empirical distribution of user i during slot t based on m previous samples is denoted by $\tilde{F}^{m,t}_{X^i}(\cdot)$ and is given by

$$\tilde{F}^{m,t}_{X^i}(x) = \frac{1}{m}\sum_{j=1}^{m} 1\{X^i(t-j) \leq x\}. \tag{5}$$

Note that the above way of estimating is similar to the score function described in [3], however no attempt was made there to evaluate the penalty due to incorrect distribution estimation as function of n and m.

Thus maximum quantile scheduling of users based on estimated distributions, would choose user $k(t)$ for service during slot t if

$$k(t) \in \arg\max_{i=1,\ldots,n} \tilde{F}^{m,t}_{X^i}(x^i(t)),$$

with ties being broken arbitrarily.

Let us examine the properties of the above scheme. It can be shown that for any user on any slot t, $\tilde{F}^{m,t}_{X^i}(X^i(t))$ is uniformly distributed on $\{0, \frac{1}{m} \ldots, 1\}$. Therefore, it is easy to see that even with estimated distributions, maximum quantile scheduling will still serve each user an equal fraction of time.

Calculating the penalty due to estimation seems to be intractable under slow fading, *therefore we add an additional assumption of fast fading, i.e., channel capacity realization of a user in a slot is independent across slots*. Even though fast fading users' channel capacity is not usually true, independence of samples can be ensured by taking samples that are sufficiently apart in time or for some physical layer follows from system design, see e.g. 'opportunistic beamforming' [20]. The assumption is also likely to be true in OFDM based systems where slot times are relatively long.

We now calculate the long term throughput achieved by users under maximum quantile scheduling based on estimated distributions. Here, since we are interested in the stationary behavior, we simplify notation for the estimated distribution to $\tilde{F}^{m}_{X^i}(\cdot)$. Following theorem characterizes the performance of this scheme, a proof is given in Appendix II.

Theorem 4.1: Consider a fixed saturated system with n users whose channel capacity variations satisfy Assumption 2.1. Suppose the channel capacity distributions in such a system are estimated via (5) base on m *independent* samples of a user's channel and users are served using maximum quantile scheduling, then the long term throughput achieved by user k is given by

$$\tilde{G}^k_{mq}(n,m) = \frac{E[F^{-1}_{X^k}(\tilde{U}_{n,m})]}{n},$$

where $\tilde{U}_{n,m}$ is a continuous r.v. on $[0,1]$ having a probability density function

$$f_{\tilde{U}_{n,m}}(u) = \sum_{j=0}^{m} \left(\begin{array}{c} m \\ j \end{array}\right) u^j(1-u)^{m-j}\frac{((j+1)^n - j^n)}{(m+1)^{n-1}}. \tag{6}$$

Recall that R^i_{mq} represent the rate distribution seen by user i when selected for service on a typical slot by maximum quantile scheduling (with perfect distribution knowledge). Let $\tilde{R}^{i,m}_{mq}$ denote the same quantity for maximum quantile scheduling when distributions are estimated using m samples.

We show that $\tilde{R}^{i,m}_{mq}$ and R^i_{mq} are 'closely related' random variables, i.e., the rate seen by a user when served under empirical distributions case is similar to that seen when distributions are perfectly known. This is used to show that the average throughput achieved by a user when empirical distributions are used is less than or equal to that achieved when distributions are perfectly known, i.e., $\tilde{G}^k_{mq}(n,m) \leq G^k_{mq}(n)$ and bound the relative throughput penalty due to estimation. Our result is formally stated below, the proof given in Appendix III.

Theorem 4.2: Consider a fixed saturated system with n users whose channel capacity variations satisfy Assumption 2.1. Then under fast fading $\forall n, m$,

$$(\frac{m+1}{n}(1-(\frac{m}{m+1})^n)) \leq \frac{\Pr(\tilde{R}^{i,m}_{mq} \leq r)}{\Pr(R^i_{mq} \leq r)} \leq 1, \ \forall r,$$

and

$$G^k_{mq}(n) \geq \tilde{G}^k_{mq}(n,m), \ \forall m,$$

and the relative throughput penalty is bounded by

$$\frac{|G^k_{mq}(n) - \tilde{G}^k_{mq}(n,m)|}{G^k_{mq}(n)} \leq 1 - \frac{m+1}{n}(1-(\frac{m}{m+1})^n).$$

To understand the scaling of the number of independent samples m required to limit the throughput penalty, note that for a reasonably large n, if m scales linearly with n, then

$$(\frac{m}{m+1})^n = (1+\frac{1}{m})^{-n} \approx e^{-\frac{n}{m}}.$$

Expanding $e^{-\frac{n}{m}}$ and simplifying, we get that the penalty is equal to

$$1 - \frac{m+1}{m} + \frac{m+1}{n}(\frac{1}{2}(\frac{n}{m})^2 - \ldots),$$

which is upper bounded by $\frac{n}{2m}$. Therefore to achieve a relative error less than ϵ, approximately $\frac{n}{2\epsilon}$ samples are needed. For

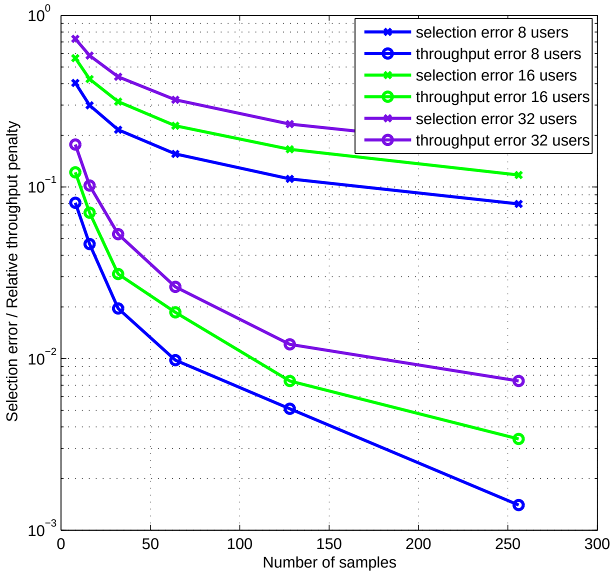

Fig. 4. The top three curves plot the selection error probability for maximum quantile scheduling, due to estimated distributions with increasing number of users. The bottom three curves plot the relative throughput penalty for the same.

example to achieve an error less than 5%, approximately $10n$ samples are needed. Therefore for a given error bound, the number of samples required will at worst grow roughly linearly with the number of users contending.

To validate these results, we ran some simulations. The set up is same as discussed in Section II. We observed the throughput penalty for different values of n and m. The value of n is varied from 8 to 16 to 32, while m is varied by a factor of 2 from 8 to 256 for a given value of n. As shown in Figure 4, the bound is clearly met, in fact the results indicate that our bound is quite conservative (which is not surprising, since the bound is distribution free). For example, a penalty of around 1% is achieved with only 64 samples for 8 users, whereas the bound suggests 5%.

We also plot the selection error probability in the figure, i.e., the fraction of slots where the user selected with maximum quantile is *not* chosen due to error in estimation of distribution. As the plot indicates, this can be quite high. Our analysis (not included in this paper) shows that the number of samples required to achieve a given error probability grows roughly as $O(n^2)$. Therefore, even though mistakes may be made in selecting the user with the highest quantile, the throughput penalty in making an error is not large.

Let us consider the relevance of the bound under slow fading. The need for m independent samples immediately suggests the need for sampling m coherence time intervals to achieve the required penalty. We ran simulations to confirm this conjecture. The simulation consisted of two (earlier described) classes of slow Rayleigh fading users with 5 users each, we aimed for a throughput penalty of 5%. The Doppler spread for the channels was varied from 10 Hz to 50 Hz in steps of 10 Hz. Let f_D denote the Doppler spread, then the coherence time can be estimated using the formula $\frac{9}{16\pi f_D}$ [15]. Given the coherence time, the total number of users and the required penalty, the number of slots needed to estimate the rate distributions can be ascertained. The simulation results are plotted in Figure 5, as can be observed, the required penalty is easily met in all cases.

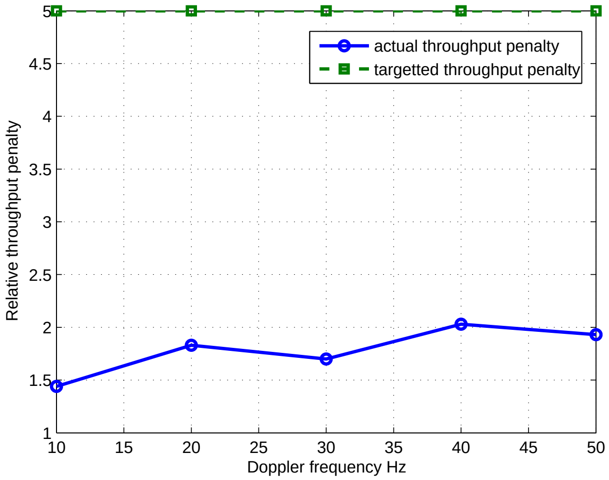

Fig. 5. Relative throughput penalty for 10 users with slow Rayleigh fading channel capacities.

Note that in our simulations we found that for Doppler spread of 10 Hz, 932 slots were needed. (Other Doppler spreads required 466, 311, 233, 187 slots.) This corresponds to 1.55 seconds (slot size is 1.67 msec), it may be reasonable to expect the system to be stationary for such a period because the Doppler spread is quite low, i.e., users/objects are moving quite slowly. In other words, even though very slowly fading systems may require a large number of samples to achieve the desired penalty, it may also be reasonable to expect such channels to be stationary over large periods of time.

Discussion of the bound. Theorem 4.2 has several interesting implications, which we discuss below.

- The bounds shows that the throughput penalty for due to estimation of users' distribution can be bounded for *any* distribution.
- The theorem is strong in the sense that it shows it shows a relationship between distributions of rates seen by the user in both the empirical and perfectly known distribution cases.
- Furthermore, the number of samples needed to achieve small penalty is *only linear in the number of users*. This is fairly limited (at least for the fast fading case) because the slot sizes are usually of the order of milliseconds.
- The dependence of penalty only on the number of users is significant, because this allows the bound to extend to unsaturated and dynamic regime. To achieve a certain penalty, a system designer only needs to estimate the 'average' number of users that will be competing for service at any given time, and not on the users' distri-

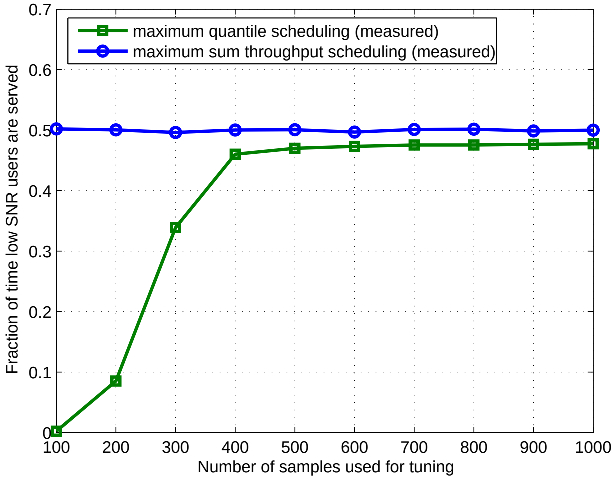

Fig. 6. Fraction of time low SNR users are served by measurement based maximum sum throughput optimal and maximum quantile scheduling schemes, with increasing number of tuning samples.

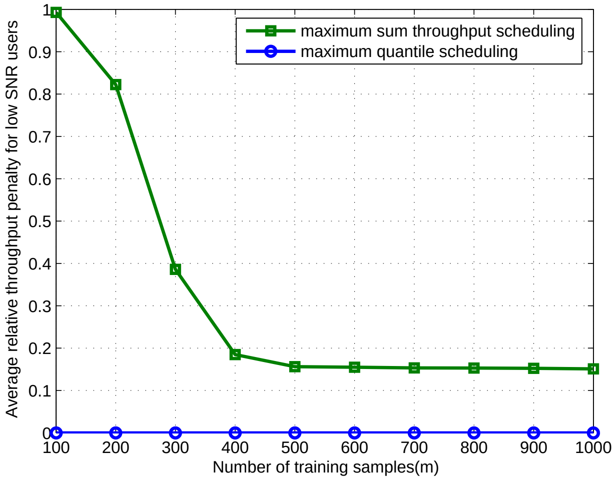

Fig. 7. Average relative throughput penalty incurred by the class of low SNR users for increasing number of tuning samples.

bution or traffic characteristics. We reiterate here that it is difficult to even design heuristics to redefine weights in dynamic and unsaturated scenarios for other weight based schemes.

- The dependence on only the number of users also allows the theorem to extend to quasi stationary rate distributions. We conjecture that if users' channel are stationary for roughly $O(n^2)$ slots (under fast fading), then the desired penalty will be met.

Summarizing, maximum quantile scheduling under estimated distribution case is not only fair, suffers from fairly limited penalty, but is quite easy to design for and to implement.

B. Throughput Penalty Comparisons

Recall that if the users' weight ν^i are properly set in (2), then the scheme maximizes sum throughput under temporal fairness. However in practice the weights for each user needs to be estimated. Let us investigate the sensitivity of system throughput to errors in these weights by performing two controlled experiments.

In the first experiment, there are 5 users in each class (with the previously described setup), and the weights ν^i for all users are initialized to 0. We train the weights for m slots according to the stochastic approximation algorithm suggested in [7], and observe the average penalty in performance due to errors in weights on the $(m+1)^{st}$ slot. We refer the reader to [7] for details on the training algorithm. We evaluate two performance parameters, the fraction of time low SNR users are served, and the relative penalty in throughput achieved by those users as compared to that achieved when weights are perfectly known.

The stochastic approximation algorithm for estimating the ν^i's has several parameters (w, δ, δ_i) that need to be set, we first set these parameters equal to those suggested in [7]. However, the scheduling scheme served the low SNR user less than 0.1% of time even when $m = 2000$ (again demonstrating that measurement based weights may severely affect performance). We changed the parameters to $w = 0.005$, $\delta = 0.2$ and $\delta_i = 0.1$, which exhibited better performance.

Figure 6 shows the fraction of time low SNR users are served as an increasing number of training samples m is used. We also plot the corresponding results for maximum quantile scheduling. Note that maximum quantile scheduling always serves low SNR users close to 0.5 fraction of time. By contrast, maximum sum throughput takes around 400 samples to converge to approximately 0.47 and then shows negligible improvement. This is because the granularity of training is not sufficiently small, however as suggested in the previous paragraph, if one reduces these updates, then the convergence time may be much larger.

Figure 7 shows the throughput penalty for the low SNR users for an increasing number of training samples m. While the throughput penalty is virtually 0 under maximum quantile scheduling, note that there is penalty of 15% even for 1000 training samples. Therefore a 3% loss in temporal fairness can lead to a 15% loss in throughput.

In the second experiment, suppose there are initially 5 users belonging to each class, with estimates for ν^i converged to their true values. Now if a user leaves the system the values of weights would have to change, so if the system does not tune fast enough, then the maximum sum throughput scheme may incur a throughput penalty. We simulated the throughput achieved by the scheme under the previously converged values of weights and compare it to that achieved by maximum quantile scheduling with distribution estimates converged.

Figure 8 shows the throughput achieved by both the schemes when the number of low SNR users is reduced. Note that the throughput difference between the two schemes is small even when both classes have 5 users each. Then if the number of low SNR users goes from 5 to 4, maximum quantile scheduling immediately starts doing better. We observed a similar trend when the high SNR users were reduced.

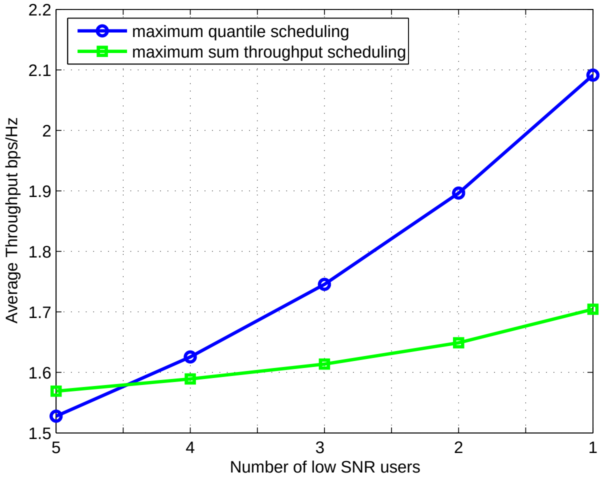

Fig. 8. Throughput achieved by maximum sum throughput under temporal fairness and maximum quantile scheduling with decreasing number of low SNR users.

V. Performance in Fixed Unsaturated Regime

In this section, we consider a fixed unsaturated system. One can show that the throughput achieved by an infinitely backlogged user k in an unsaturated system is lower bounded by $G^k_{mq}(n)$ [11], i.e., the throughput achieved in a fixed saturated system. However, the way in which resources are allocated impacts the delay for e.g. real-time traffic. Therefore, we will evaluate the packet delay in this section.

In our simulations, we compare the performance of maximum quantile scheduling with maximum rate, proportionally fair and the exponential rule. We do not compare the performance with the maximum sum throughput scheme (2), because it is unclear how to set the weights for this scheme in an unsaturated scenario.

Our setup is the same as before with 5 users per class. All users have Poisson packet arrivals with equal average arrival rate. Each packet is 1500 bytes. Packet delay for a packet is measured by finding the difference between packet arrival time and the time when the packet has been *completely* transmitted. We set all the weights equal to 1 in exponential rule (We also experimented by weighting a user's queue inversely proportional to its channel mean, but that increased the average delay for the exponential rule.). We assume that the distribution is perfectly known at the scheduler for maximum quantile scheduling, i.e., the estimates of the distributions have converged (this may be reasonable for fixed systems).

Figure 9 shows the average packet delay across users as the load increases.(Proportional fair has the worst performance among all schemes, for simplicity its plot has not been shown). Maximum quantile always has the lowest packet delay, and achieves more than 35% reduction in packet delay as compared to the exponential rule. This is surprising, since unlike exponential rule, maximum quantile is completely *insensitive* to queue lengths. This underscores the importance of scheduling according to opportunism, rather than simply the rate and/or queue lengths.

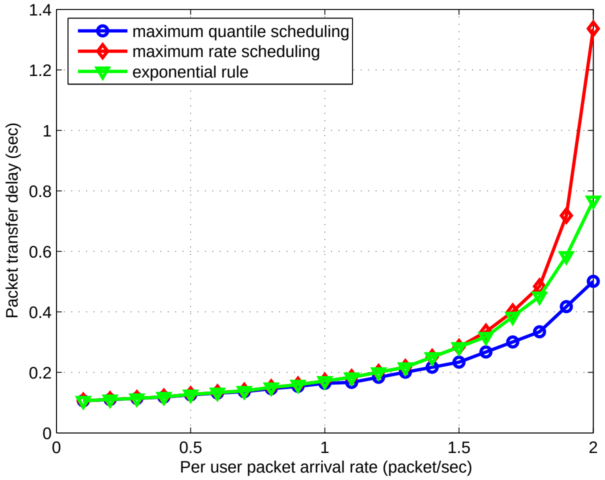

Fig. 9. Packet delay performance of maximum quantile, exponential rule and maximum rate scheduling.

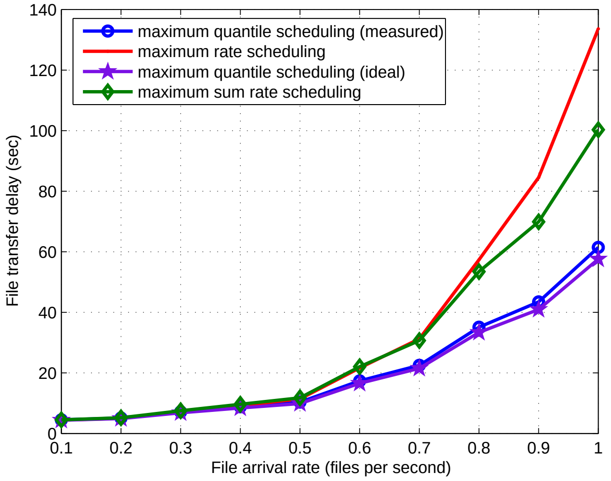

Fig. 10. File transfer delay performance of maximum quantile, maximum rate and maximum sum throughput scheduling.

VI. Performance in Dynamic Saturated Regime

In this section, we compare the performance of maximum quantile scheduling to maximum rate, proportional fair and maximum sum throughput under temporal constraints. Note that the exponential rule does not make sense in a saturated scenario.. Dynamic saturated system is a good model for an access point supporting file transfers, therefore, a good metric for performance here is the average file transfer delay.

Again our setup is the same as before, however since the system is dynamic, the number of users will change with time. Users arrive to the system according to a Poison process, and are equally likely to belong to one of the two classes. Each user has a file associated with it. The file sizes are exponentially distributed with a mean size of 60KB. We keep track of the time taken from a user's arrival to departure. For maximum quantile, estimate for users rate distributions are generated by keeping track of previous samples. While the weights for maximum sum throughput under temporal fairness are trained using the stochastic approximation algorithm referred to in Section IV, with the values of the parameters same as before.

The average file transfer delay experienced by users is plotted with increasing load in Figure 10 (we again do not plot proportional fair because its performance is the worst). The number of samples used for estimating users' distributions is 50 at a load of 0.1. This was increased linearly by 50 samples for every load increase of 0.1. As can be seen in the figure, maximum quantile scheduling outperforms both maximum rate and maximum sum throughput. In fact the reduction in delay is almost 40% (as compared to maximum sum throughput) at a load of 1. This again underscores the importance of scheduling according to the quantile. Also note that due to non convergence of weights, maximum rate and maximum sum throughput up to a load of 0.8 have quite similar performance. Therefore maximum sum throughput can easily degrade to maximum rate in a dynamic scenario.

We also plot the delay experienced by users under maximum quantile with perfect rate distribution knowledge. Observe that the ideal performance is close to measurement based one.

VII. Conclusion

In summary we have evaluated measurement based opportunistic scheduling schemes from various perspectives and under various system regimes, e.g., dynamic/fixed, saturated/unsaturated. The key take away, is that, perhaps surprisingly, maximum quantile scheduling which would require estimation of each users channel rate distribution, realizes excellent performance, relative to proportionally fair, the exponential rule, and schemes that are optimal in terms of sum throughput subject to fairness. The main reason is that maximum quantile places systematic emphasis on scheduling users when they are high relative to their own distribution, while achieving temporal fairness. By contrast other schemes measure the degree to which fairness is achieved and bias scheduling decisions to compensate for biases. This compromises opportunism and also performance. Although the estimation of users distributions seems fairly straightforward and would be necessary to enable resource management and call admission decisions at a wireless point, the question remains as to whether the additional complexity over simple schemes such as proportionally fair is warranted.

Appendix I
Proof of Theorem 3.1

Proof: Define $U_\beta := \sum_{i=1}^n U^i \mathbf{1}_{S^i_\beta}$, i.e., the total opportunism achieved by β. Let $U^i_\beta = U^i|S^i_\beta$, i.e., the quantile of user i conditioned on getting served by β. Then

$$\Pr(U_\beta > u) = \sum_{i=1}^n \Pr(U^i_\beta > u)\Pr(S^i_\beta),\ \ u \in [0,1].$$

Let $j(u) = \arg\min_{i=1,\ldots,n}\Pr(U^i_\beta > u)$. Since β is non idling, $\sum_{i=1}^n \Pr(S^i_\beta) = 1$, so

$$\Pr(U_\beta > u) \geq \Pr(U_\beta^{j(u)} > u).$$

Recall that $U^{(n)}$ is the maximum of n i.i.d. uniformly distributed random variables, then since β is distinct, there must be a u' such that

$$\Pr(U^{(n)} > u') > \Pr(U_\beta > u') \geq \Pr(U_\beta^{j(u')} > u').$$

Let U^i_{mq} be the same quantity as U^i_β for maximum quantile scheduling. Now recall that under maximum quantile scheduling, a user i is selected for service only when its quantile is the highest, i.e., $U^i_{mq} \sim U^{(n)}$. Then

$$\Pr(U_{mq}^{j(u')} > u') > \Pr(U_\beta^{j(u')} > u').$$

So $U_\beta^{j(u')} \ngeq^{st} U_{mq}^{j(u')}$. Note that for any user i, $R^i_\beta = F^{-1}_{X^i}(U^i_\beta)$ and $R^i_{mq} = F^{-1}_{X^i}(U^i_{mq})$. Now since $F^{-1}_{X^{j(u')}}(\cdot)$ is an increasing function, then

$$R_\beta^{j(u')} \ngeq^{st} R_{mq}^{j(u')}.$$

■

Appendix II
Proof of Theorem 4.1

Proof: Recall that S^k is the event denoting the selection of user k for service. Since each user is equally likely to be served, $\Pr(S^k) = \frac{1}{n}$, and

$$\tilde{G}^i_{mq}(n,m) = E[X^k|S^k]\Pr(S^k) = \frac{E[X^k|S^k]}{n}.$$

Let us now evaluate $E[X^k|S^k]$ by conditioning on $\tilde{F}^m_{X^k}(X^k)$, we have that

$$E[X^k|S^k] = \sum_{j=0}^m E[X^k|S^k, \tilde{F}^m_{X^k}(X^k) = \frac{j}{m}]\Pr(\tilde{F}^m_{X^k}(X^k) = \frac{j}{m}|S^k).$$

Note that the selection of a user in a slot is independent of its current rate, given its estimated current quantile, so

$$E[X^k|S^k, \tilde{F}^m_{X^k}(X^k) = \frac{j}{m}] = E[X^k|\tilde{F}^m_{X^k}(X^k) = \frac{j}{m}].$$

Since $\tilde{F}^m_{X^k}(X^k)$ are uniformly distributed on $\{0, \frac{1}{m}, \ldots, 1\}$ and ties are broken randomly,

$$\Pr(\tilde{F}^m_{X^k}(X^k) = \frac{j}{m}|S^k) = \frac{(j+1)^n - j^n}{(m+1)^n}.$$

Now consider $E[X^k|\tilde{F}^m_{X^k}(X^k) = \frac{j}{m}]$, by using Bayes' formula and the fact that $\left(\begin{array}{c} m \\ j \end{array}\right)\int_0^1 y^j(1-y)^{m-j}dy = \frac{1}{m+1}$, one can show that

$$E[X^k|\tilde{F}^m_{X^k}(X^k) = \frac{j}{m}] = (m+1)\left(\begin{array}{c} m \\ j \end{array}\right)\int_0^\infty x(F_{X^k}(x))^j(1-F_{X^k}(x))^{m-j}f_{X^k}(x)dx,$$

where $f_{X^k}(\cdot)$ is the probability density function of the SNR associated with user k. Now using a change of variables this can be rewritten as

$$E[X^k|\tilde{F}^m_{X^k}(X^k)=\frac{j}{m}]=$$
$$(m+1)\left(\begin{array}{c}m\\ j\end{array}\right)\int_0^1 F^{-1}_{X^k}(u)u^j(1-u)^{m-j}du.$$

So it follows that $\tilde{G}^k_{mq}(n,m)$ is given by

$$\frac{1}{n}\int_0^1 F^{-1}_{X^k}(u)(\sum_{j=0}^{m}\left(\begin{array}{c}m\\ j\end{array}\right)u^j(1-u)^{m-j}\frac{((j+1)^n-j^n)}{(m+1)^{n-1}})du.$$

This completes the proof. ■

Appendix III
Proof of Theorem 4.2

We present a few useful lemmas before proving Theorem 4.2.

Lemma 3.1: Let H be a binomial r.v. with parameters (m,u). Consider the moment generating function of H, $M(s):=(1-u+ue^s)^m$. Its l^{th} derivative is given by

$$\frac{d^l M(s)}{ds^l}=\sum_{j=1}^{l} b_{j,l}\frac{m!}{(m-j)!}(1-u+ue^s)^{m-j}(ue^s)^j. \quad (7)$$

Here $b_{j,l}$'s are constants with the following properties:

- $b_{1,1}=1$
- $b_{j,l}=jb_{j,l-1}+b_{j-1,l-1}$, $\forall j=1,\dots,l$, $\forall l$
- $b_{0,l}=b_{l+1,l}=0$, $\forall l$.

Note that since $b_{1,1}=1$ and $b_{l+1,l}=0$, $\forall l$, from the second property one can show that $b_{l,l}=b_{l-1,l-1}=1$, $\forall l$.

Proof: The lemma clearly holds for $l=1$. We give a proof by induction on l. Assume the lemma holds for l, i.e., (7) is true. Then, to prove the lemma for $l+1$, we differentiate (7) and after some rearrangement get

$$\frac{d^{l+1}M(s)}{ds^{l+1}}=\sum_{j=1}^{l+1}[(jb_{j,l}+b_{j-1,l})\frac{m!}{(m-j)!}$$
$$(1-u+ue^s)^{m-j}(ue^s)^j].$$

This completes the proof. ■

From Lemma 3.1 it follows that the l^{th} order moment of H is given by

$$E[H^l]=\sum_{j=1}^{l} b_{j,l}\frac{m!}{(m-j)!}u^j. \quad (8)$$

The following lemma exhibits an inequality between the moments of H.

Lemma 3.2: Let H be a binomial r.v. with parameters (m,u). Then for all l such that $l\leq m$,

$$E[H^{l+1}]\leq(mu+l(1-u))E[H^l]. \quad (9)$$

Proof: The right side of (9) can be expressed as

$$((m-l)u+l)E[H^l].$$

Using (8), the above equation can be rewritten as

$$\frac{m!}{(m-l-1)!}u^{l+1}+ \quad (10)$$
$$\sum_{j=1}^{l}[lb_{j,l}\frac{m!}{(m-j)!}+(m-l)b_{j-1,l}\frac{m!}{(m-j+1)!}]u^j. \quad (11)$$

If one splits $lb_{j,l}\frac{m!}{(m-j)!}$ in the following way

$$lb_{j,l}\frac{m!}{(m-j)!}=jb_{j,l}\frac{m!}{(m-j)!}+(l-j)b_{j,l}\frac{m!}{(m-j)!},$$

then (10) in turn can be expressed as

$$\frac{m!}{(m-l-1)!}u^{l+1}+\sum_{j=1}^{l}[(jb_{j,l}\frac{m!}{(m-j)!}+(m-l)b_{j-1,l}$$
$$\frac{m!}{(m-j+1)!})u^j+(l-j+1)b_{j-1,l}\frac{m!}{(m-j+1)!}u^{j-1}].$$

Now since $0\leq u\leq 1$, then $\forall j$, $u^{j-1}\leq u^j$. So, from the above equation we get

$$(mu+l(1-u))E[H^l]\geq\frac{m!}{(m-l-1)!}u^{l+1}+$$
$$\sum_{j=1}^{l}[(jb_{j,l}\frac{m!}{(m-j)!}+(m-l)b_{j-1,l}\frac{m!}{(m-j+1)!})+$$
$$(l-j+1)b_{j-1,l}\frac{m!}{(m-j+1)!}]u^j.$$

Combining the last two terms in the summation of the above inequality, we get

$$(mu+l(1-u))E[H^l]\geq\frac{m!}{(m-l-1)!}u^{l+1}+$$
$$\sum_{j=1}^{l}(jb_{j,l}+b_{j-1,l})\frac{m!}{(m-j)!}u^j$$

This proves (9). ■

Next we show that $U^{(n)}$ dominates $\tilde{U}_{n,m}$ in a likelihood ratio ordering sense, i.e., $U^{(n)}\geq^{lr}\tilde{U}_{n,m}$ [8][16]. This is a strong form of dominance which means that $f_{U^{(n)}}(u)/f_{\tilde{U}_{n,m}}(u)$ is non decreasing in u, or $f_{\tilde{U}_{n,m}}(u)/f_{U^{(n)}}(u)$ is non increasing in u (here $f_{U^{(n)}}(u)$ is the probability density function of $U^{(n)}$). If $U^{(n)}\geq^{lr}\tilde{U}_{n,m}$, it follows that $U^{(n)}\geq^{st}\tilde{U}_{n,m}$.

Lemma 3.3: For the random variables $U^{(n)}$ and $\tilde{U}_{n,m}$ given by (4) and (6) respectively, then $\forall n,m$ $U^{(n)}\geq^{lr}\tilde{U}_{n,m}$.

Proof: To prove the lemma, we need to show

$$\frac{d}{du}\left[\frac{f_{\tilde{U}_{n,m}}(u)}{f_{U^{(n)}}(u)}\right]\leq 0,$$

$\forall u \in (0,1]$. To prove this, it is sufficient to show

$$f_{U^{(n)}}(u)\left[\frac{df_{\tilde{U}_{n,m}}(u)}{du}\right] - f_{\tilde{U}_{n,m}}(u)\left[\frac{df_{U^{(n)}}(u)}{du}\right] \leq 0.$$

Note that $f_{U^{(n)}}(u) = nu^{n-1}$. Then expanding, we get

$$\frac{1}{(m+1)^{n-1}}[nu^{n-1}(-m(1-u)^{m-1} + \sum_{j=1}^{m-1}\binom{m}{j}u^{j-1}(1-u)^{m-j-1}(j-mu)((j+1)^n-j^n) + mu^{m-1}((m+1)^n-m^n)) - n(n-1)u^{n-2}(\sum_{j=0}^{m}\binom{m}{j}u^j(1-u)^{m-j}((j+1)^n-j^n))] \leq 0.$$

Simplifying and multiplying both sides by $(1-u)$, we get

$$(-mu(1-u)^m + \sum_{j=1}^{m-1}\binom{m}{j}(j-mu)u^j(1-u)^{m-j}((j+1)^n-j^n) + (m-mu)u^m((m+1)^n-m^n)) - (n-1)(1-u)(\sum_{j=0}^{m}\binom{m}{j}u^j(1-u)^{m-j}((j+1)^n-j^n)) \leq 0.$$

The above inequality can be rewritten as

$$\sum_{j=0}^{m}\binom{m}{j}(j-mu-(n-1)(1-u))u^j(1-u)^{m-j}((j+1)^n-j^n) \leq 0.$$

Then the inequality clearly holds for $m < n$. However the more interesting case is when $m \geq n$, and this requires a few more steps. Note that $\binom{m}{j}u^j(1-u)^{m-j}$ is the probability that a binomial r.v. with parameter (m,u) has a value j, i.e., the same as that of H. Then the inequality can be rewritten in terms of expectations as

$$E[(H-mu)((H+1)^n-H^n)] - (n-1)(1-u)E[(H+1)^n-H^n] \leq 0.$$

This can be further rewritten as

$$E[H((H+1)^n-H^n)] \leq (mu+(n-1)(1-u))E[(H+1)^n-H^n]. \tag{12}$$

Expanding $(H+1)^n$ and simplifying, one can show that (12) will hold if

$$E[H^{l+1}] \leq (mu+l(1-u))E[H^l],$$

$\forall l < n \leq m$. This follows from Lemma 3.2. This completes the proof. ■

We now prove Theorem 4.2.

Proof: To prove the first claim, define $u := F_{X^i}(r)$ and consider

$$F_{U^{(n)}}(u) - F_{\tilde{U}_{n,m}}(u),\ \forall u \in (0,1].$$

This is equivalent to

$$\int_0^u f_{U^{(n)}}(u) - f_{\tilde{U}_{n,m}}(u)du.$$

This in turn is equivalent to

$$\int_0^u f_{U^{(n)}}(u)(1-\frac{f_{\tilde{U}_{n,m}}(u)}{f_{U^{(n)}}(u)})du.$$

Then

$$F_{U^{(n)}}(u) - F_{\tilde{U}_{n,m}}(u) \leq \int_0^u f_{U^{(n)}}(u)\max_u(1-\frac{f_{\tilde{U}_{n,m}}(u)}{f_{U^{(n)}}(u)})du.$$

Note from Lemma 3.3,

$$\min_u \frac{f_{\tilde{U}_{n,m}}(u)}{f_{U^{(n)}}(u)} = \frac{f_{\tilde{U}_{n,m}}(1)}{f_{U^{(n)}}(1)} = \frac{m+1}{n}(1-(\frac{m}{m+1})^n).$$

Then

$$F_{U^{(n)}}(u) - F_{\tilde{U}_{n,m}}(u) \leq F_{U^{(n)}}(u)(1-\frac{m+1}{n}(1-(\frac{m}{m+1})^n)).$$

Simplifying, one gets

$$F_{U^{(n)}}(u)(\frac{m+1}{n}(1-(\frac{m}{m+1})^n)) \leq F_{\tilde{U}_{n,m}}(u).$$

Now from Lemma 3.3, it follows that $U^{(n)} \geq^{st} \tilde{U}_{n,m}$, combining this with the above equation we get

$$\frac{m+1}{n}(1-(\frac{m}{m+1})^n) \leq \frac{F_{\tilde{U}_{n,m}}(u)}{F_{U^{(n)}}(u)} \leq 1.$$

Using the definition of u, and the fact that $F_{X^i}(\cdot)$ is an increasing function, the above equation can be rewritten as

$$\frac{m+1}{n}(1-(\frac{m}{m+1})^n) \leq \frac{\Pr(F^{-1}_{X^i}(\tilde{U}_{n,m}) \leq r)}{\Pr(F^{-1}_{X^i}(U^{(n)}) \leq r)} \leq 1.$$

Note that $R^i_{mq} = F^{-1}_{X^i}(U^{(n)})$ and $\tilde{R}^{i,m}_{mq} = F^{-1}_{X^i}((\tilde{U}_{n,m})$, then the above equation can be written as

$$\frac{m+1}{n}(1-(\frac{m}{m+1})^n) \leq \frac{\Pr(\tilde{R}^{i,m}_{mq} \leq r)}{\Pr(R^i_{mq} \leq r)} \leq 1.$$

To prove the second claim, recall that $G^k_{mq}(n) = \frac{E[F^{-1}_{X^k}(U^{(n)})]}{n}$. Note that $F^{-1}_{X^k}(\cdot)$ is an increasing function. Therefore it is sufficient to prove that $U^{(n)} \geq^{st} \tilde{U}_{n,m}$ to prove the theorem, which is shown to be true from Lemma 3.3.

We now prove the third part of the theorem. Note from the second part of the theorem, it is suffices to study

$$\frac{G^k_{mq}(n) - \tilde{G}^k_{mq}(n,m)}{G^k_{mq}(n)}.$$

Consider the difference between the two throughput, i.e., $E[F^{-1}_{X^k}(U^{(n)})] - E[F^{-1}_{X^k}(\tilde{U}_{n,m})]$. The difference can be expressed as

$$\int_0^1 F^{-1}_{X^k}(u)f_{U^{(n)}}(u)du - \int_0^1 F^{-1}_{X^k}(u)f_{\tilde{U}_{n,m}}(u)du.$$

Then following the methodology used in the first part of the proof on can show

$$E[F_{X^k}^{-1}(U^{(n)})] - E[F_{X^k}^{-1}(\tilde{U}_{n,m})] \leq$$
$$\int_0^1 F_{X^k}^{-1}(u) f_{U^{(n)}}(u)(1 - \frac{m+1}{n}(1 - (\frac{m}{m+1})^n))du,$$

or

$$E[F_{X^k}^{-1}(U^{(n)})] - E[F_{X^k}^{-1}(\tilde{U}_{n,m})] \leq$$
$$E[F_{X^k}^{-1}(U^{(n)})](1 - \frac{m+1}{n}(1 - (\frac{m}{m+1})^n)).$$

This completes the proof. ■

References

[1] 3GPP TS 25.308. UTRA high speed downlink packet access (HSDPA): Overall description, Release 5. 2003.

[2] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushayana, and A. Viterbi. CDMA-HDR: A bandwidth-efficient high-speed wireless data service for nomadic users. *IEEE Communication Magazine,*, pages 70–77, July 2000.

[3] T. Bonald. A score-based opportunistic scheduler for fading radio channels. In *Proc. of European Wireless*.

[4] S. Borst. User-level performance of channel-aware scheduling in wireless data networks. In *INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications Societies*, volume 1, pages 321 – 331, March-April 2003.

[5] A. Jalali, R. Padovani, and R. Pankaj. Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system. In *Vehicular Technology Conference Proceedings, 2000. VTC 2000-Spring Tokyo*, volume 3, pages 1854 – 1858, May 2000.

[6] R. Knopp and P. Humblet. Information capacity and power control in single cell multi-user communications. In *Proc. IEEE International Computer Conference*, volume 1, pages 331 – 335, June 1995.

[7] X. Liu, E. K. P. Chong, and N. B. Shroff. A framework for opportunistic scheduling in wireless networks. *Computer Networks*, 41:451–474, March 2003.

[8] A. Marshall and I. Olkin. *Inequalities: Theory of Marjorization and Its Applications*. Academic Press, 1979.

[9] D. Park, H. Seo, H. Kwon, and B. G. Lee. A new wireless packet scheduling algorithm based on the cdf of user transmission rates. In *Proc. IEEE Globecom*, pages 528–532, November 2003.

[10] D. Park, H. Seo, H. Kwon, and B. G. Lee. Wireless packet scheduling based on the cumulative distribution function of user transmission rates. *to appear in IEEE Transactions on Communications*, 2005.

[11] S. Patil. Opportunistic scheduling and resource allocation among heterogeneous users in wireless networks, Ph.D. qualifying proposal, Univeristy of Texas at Austin. 2005.

[12] S. Patil and G. de Veciana. Managing resources and quality of service in wireless systems exploiting opportunism. In *Submitted for journal publication, available at http://www.ece.utexas.edu/˜ patil/noniidQoS.pdf.*

[13] S. Patil and G. de Veciana. Throughput optimality of maximum quantile scheduling under greedy user behaviour. In *Proc. Conference on Information Sciences and Systems (CISS)*, March 2005.

[14] X. Qin and R. Berry. Opportunistic splitting algorithms for wireless networks with heterogeneous users. In *Proc. Conference on Information Sciences and Systems (CISS)*, March 2004.

[15] T. S. Rappaport. *Wireless Communications, Principles and Practice*. Pearson Education, 2002.

[16] S. M. Ross. *Stochastic Processes*. John Wiley, 1983.

[17] S. Shakkottai, R. Srikant, and A. L. Stolyar. Pathwise optimality of the exponential scheduling rule for wireless channels. *Advances in Applied Probability*, 36:1021–1045, December 2004.

[18] S. Shakkottai and A. Stolyar. Scheduling algorithms for a mixture of real-time and non-real-time data in HDR. In *Proc. of the 17th International Teletraffic Congress (ITC-17), Salvador da Bahia, Brazil*, September 2001.

[19] S. Shakkottai and A. Stolyar. Scheduling for multiple flows sharing a time-varying channel: The Exponential rule. *American Mathematical Society Translations, Series 2, A volume in memory of F. Karpelevich, Yu. M. Suhov, Editor*, 207, 2002.

[20] P. Viswanath, D. Tse, and R. Laroia. Opportunistic beamforming using dumb antennas. *IEEE Transactions on Information Theory*, 48:1277 – 1294, June 2002.