Measurement-Based Opportunistic Feedback and Scheduling for Wireless Systems

Gustavo de Veciana and Shailesh Patil
Wireless Networking and Communications Group
Dept. of Electrical and Computer Engineering
The University of Texas at Austin
Austin, TX 78712
{gustavo,patil}@ece.utexas.edu

Abstract

We study opportunistic feedback of channel state and scheduling of users transmissions at a wireless access point. Our focus is on a regime where users have unknown possibly slowly-varying channel characteristics and it is desirable to limit the resources expended to exchanging channel state opportunistic. The paper argues that opportunistically scheduling the user whose current rate is in the highest quantile with respect to an empirical distribution has substantial advantages. Indeed by comparing to other measurement-based opportunistic scheduling schemes it systematically achieves high 'opportunism' and thus throughput in fair and robust manner. Further the throughput penalty resulting from estimating users distributions is limited as long as the number of independent samples grows linearly with the number of active users. Second, a maximum quantile based scheduling scheme is consistent the use of a simple opportunistic feedback strategy that can be used to substantially reduce the overheads associated with deciding which user should in fact be scheduled at any point in time. Using both analytical and simulation results we will show that the benefits can be substantial for a variety of practical wireless system scenarios.

1 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][15][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.

Whenever an access point makes an opportunistic decision on the user(s) to serve, it needs to know the current channel capacity (or a function of it) for all (or a subset) of the users. Therefore, before *each* decision *all* users need to transmit their current channel condition to the access point. This can be a huge overhead, compared to the gain in throughput due to opportunistic scheduling. For example, consider a system where all users see i.i.d. (independent, identically distributed) Rayleigh channel capacity fading. Here, the gain due to opportunistic scheduling grows at most logarithmically with the number of users in the system, while the amount of feedback, i.e., the number of transmissions increases linearly with the number of users. This clearly underlines the need for reducing the amount of feedback in opportunistic scheduling.

Related Work. Several researchers have studied the problem of reducing feedback in opportunistic scheduling. A simple threshold based scheme was proposed in [5]. There, the idea was to allow only users with current channel capacity above a threshold to feedback their current state. This, it was shown, reduced the amount of feedback significantly for a small reduction in the overall throughput achieved. However, there it was assumed that the resources used to feedback channel capacity were not shared among users. Whereas, the issue of reducing feedback becomes more relevant when the resources used for feedback are shared among users, for example in a TDMA system where the time used in collecting the feedback grows linearly with the number of users. In this paper we will focus on such a setup.

In [13], an ‘opportunistic splitting’ algorithm is proposed in a TDMA set up for uplink scheduling (although it can be easily modified to downlink scheduling). There, each data transmission was preceded by mini slots, which are used to learn the current channel capacity of users via feedback. Once the user with highest channel capacity is identified, data transmission ensues.

In opportunistic splitting, initially a pair of thresholds depending on the number of users is set. At the start of the first mini slot, every user with current channel capacity between the pair of thresholds contend, i.e., transmit to the access point. The access point then broadcasts to all the users whether no user contended, exactly one user contended, or a collision occurred, i.e., more than one user contended and the access point was unable to decode any information. Depending on the broadcast message received, each user modifies his threshold according to a binary search like algorithm and users’ whose channel capacity is between the new thresholds contend in the next mini slot. This process continues until exactly one user contends, therefore the number of mini slots before a transmission vary. This user is guaranteed to be the user with the highest current channel capacity. It was shown that an average of only 2.5 mini slots are required for the algorithm to find the user with the highest current channel capacity. This is significant as compared to linear number of slots required for a naive feedback scheme.

Opportunistic splitting requires two way communication and coordination between the access point and users in every mini slot. This may be a high overhead since the time scales involved are quite small. (In practical systems, a slot is of the order of milliseconds, therefore each mini slot has to be smaller than a millisecond.) To overcome this coordination problem, a random access based feedback protocol was proposed in [16], where only users transmit. Here, each data transmission is preceded by a fixed number of mini slots (the smaller the number of mini slots, the lesser the time used in feedback). In each mini slot, users with current channel capacity above a threshold contended with some probability. If in a mini slot exactly one user contended, then that user’s identity is stored at the access point, and the access point randomly serves one of the identified users. However, if no user is identified a user is selected at random. The threshold and probability were optimized to maximize the overall sum capacity. However, simulation results presented in the paper show that opportunistic splitting performs better than the proposed scheme.

The above described work assume that all users in the system see i.i.d. channel capacity distribution. Furthermore, to set the thresholds correctly, it is assumed that the channel capacity

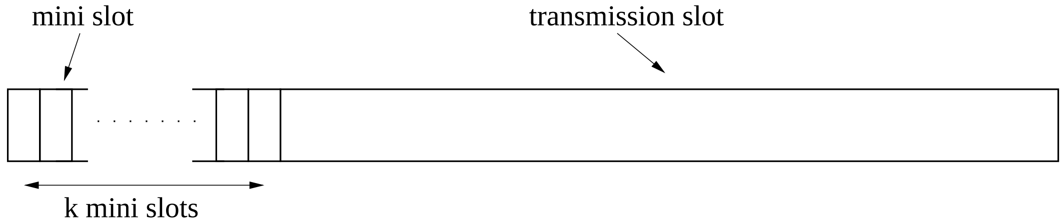

Figure 1: Structure of a time unit.

distributions are either known at the access point, or by the users. In practice both of the assumptions are unlikely to be true. (Note that an extension of opportunistic splitting to the case where users can experience one of two possible channel capacity distributions is presented in [14], however this is also not a reasonable assumption.)

In practice users channel capacity variations are heterogenous, e.g., users close to an access point see significantly different channel capacity than those further off. Also, the channel capacity distributions are usually unknown (both at the access point and by users) and need to be estimated via measurement. Therefore, one needs to evaluate the penalty incurred due to estimation.

Contributions. In this paper we propose a threshold based scheme known as *static splitting* to reduce the amount of feedback. We will consider a TDMA setup similar to that in [13][16]. However unlike opportunistic splitting, static splitting requires only one way communication. We will combine static splitting with a distribution based scheduler that we call *maximum quantile scheduling* to handle heterogeneity in users' channel capacity variations. Maximum quantile scheduling has been proposed by several researchers under different guises [8][9][3][14], the idea is to schedule the user whose current rate is highest relative to his *own* distribution, i.e., in the highest quantile. Maximum quantile scheduling allows the access point to calculate a common threshold independent of users' distribution for our scheme. In this paper we also develop insight into throughput penalty due to estimation errors using a distribution free bound developed in [11]. Finally, simulation results indicate that static splitting can perform better than a truncated form of opportunistic splitting.

Paper Organization. This paper is organized as follows. In Section 2 we give a brief introduction to maximum quantile scheduling and describe the proposed scheme. We discuss the throughput penalty incurred due to estimation errors in Section 3. Simulation results are presented in Section 4. Section 5 concludes the paper.

2 Static Splitting

2.1 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. Suppose time is divided equal sized 'time units'. Each time unit consists of k equal sized mini slots (for collecting feedback) followed by a transmission slot during which at most one user can be served (see Figure 1). In the sequel we use the terms 'channel capacity' and 'rate' interchangeably and make the following assumption on user's channel characteristics over data transmission time slots.

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 known a priori at the user.*

Some comments on this assumption. First the channel capacities seen by users might indeed be roughly stationary over a reasonable period of time particularly if users are at fixed locations. 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 a user knows, and in particular can estimate, the marginal distributions of the channel capacity processes may not be completely reasonable, but simple book keeping of the currently achievable rate can be used to estimate distributions. We will discuss estimation of such distributions in Section 3. 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.

In the sequel we will let $x^i(t)$ denote the realization of the downlink 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 $F_{X^i}(\cdot)$. Note that by assumption $F_{X^i}(\cdot)$ is known at the user.

For analysis purposes, in this paper we will only consider a 'fixed saturated' regime where the number of users in the system stay fixed with time, and each user in the system is infinitely backlogged. The number of users in the system are denoted by n.

Before explaining the proposed scheme, we give a brief introduction to maximum quantile scheduling.

2.2 Maximum Quantile Scheduling

As discussed earlier, maximum quantile scheduling has been proposed independently by several researchers. Specifically [8][9] 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 [12], in [10], we evaluate its use to achieve quality of service guarantees for real-time traffic, and in [11] we evaluate its performance in a measurement based set up.

The main idea of the scheme is to schedule a user who's rate is highest compared to his *own* distribution, i.e., serve user $l(t)$ during slot t if

$$l(t) \in \arg\max_{i=1,\dots,n} F_{X^i}(x^i(t)).$$

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

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

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

$$G^i_{mq}(n) = \frac{E[F^{-1}_{X^i}(U^{(n)})]}{n} = \frac{E[X^{i,(n)}]}{n}. \tag{2}$$

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 $X^i_j \sim X^i,\ \forall j = 1, \ldots, n.$

Compared to other schemes proposed in literature, maximum quantile scheduling is found to be very useful in a realistic scenario, i.e., a scenario where users have heterogeneous channel capacity distributions, and parameters/weights associated with a scheduling scheme have to be estimated via measurement. Maximum quantile not only maximizes the amount of opportunism exploited while being fair, but is also quite robust to measurement errors (this will be discussed later also). In fact maximum quantile scheduling does better than the sum throughput optimal strategy proposed in [7], we refer the reader to [11] for details. Therefore in our proposed scheme we try to serve a user with high quantile, i.e., high $F_{X^i}(x^i(t))$.

2.3 Proposed Scheme

Recall that a slot contained k mini slots used to collect feedback. Using these k mini slots we try to identify a user whose current rate is in a high quantile, i.e., high $F_{X^i}(x^i(t))$, so as to maximize the total quantile served. (Note that we do not necessarily try to identify the user with the highest quantile, we will revisit this point later.) Apart from the several advantages of selecting a user with high quantile discussed above, we will show in sequel, it also simplifies threshold calculation.

For simplicity assume that the number of users n is such that $\frac{n}{k}$ an integral value. In the proposed scheme, a user is associated with exactly one mini slot, with $\frac{n}{k}$ users associated with each mini slot. A user can contend for the slot (i.e., send its feedback) only on the mini slot with which it is associated. In other words, users are split into ‘static’ groups, with users belonging to the same group allowed to contend only in the same mini slot. This reduces the chances of collision in a mini slot. Once a user is associated with a mini slot, it calculates a quantile threshold denoted by q^i, $i = 1, \ldots, n$. We will give the details of calculating q^i later.

Consider slot t, recall that the rate supported by user i during slot t is denoted by $x^i(t)$. At the start of a mini slot, during the first mini slot all users associated with it *and* having current rate above $F^{-1}_{X^i}(q^i)$ contend, i.e., if $x^i(t) > F^{-1}_{X^i}(q^i)$ they transmit the quantile of their current rate $F_{X^i}(x^i(t))$ to the access point.

If exactly one user contends, we assume that the access point is able to identity the user that contended, and the value of its quantile, and store this information. If more than one user contends, i.e., a collision occurs, then the access point stores the fact that there was a collision on the first mini slot. If no user contends, then no action is taken by the access point. This process of contending is repeated in all the subsequent mini slots.

Once the contending for mini slots is finished, if the access point was able to identify at least one user, it serves the user with the highest quantile among the identified users. However, if it fails to identify such a set, it chooses to serve a randomly selected user. This can occur in two ways, if collisions happened in none of the mini slots, then a user is randomly selected from all the users. However, if a collision occurred on at least one of the mini slot, then a user is randomly selected among the groups of users that could have contended on the mini slots on which collisions occurred. This increases the chance of finding a user with a high quantile.

We now discuss calculating the threshold q^i. Let S^i be the event denoting the selection of user i for service, and $\mathbf{1}_{S^i}$ be the indicator function for S^i. Recall that we would like to adjust the thresholds q^i's so as to maximize the overall sum quantile of the users served, i.e., $E[\sum_{i=1}^n F_{X^i}(X^i)\mathbf{1}_{S^i}]$.

Recall that $F_{X^i}(X^i)$ are i.i.d. across users, therefore each user is equally likely to achieve a high quantile. Therefore to maximize $E[\sum_{i=1}^n F_{X^i}(X^i)\mathbf{1}_{S^i}]$, the scheduler is equally likely to select a user, i.e., $\forall i$, $\Pr(S^i) = \frac{1}{n}$. Also, the maximum number of users that can contend in any mini slot is $\frac{n}{k}$. These symmetries allows us to have a common threshold q across users, i.e., $q^i = q$, $\forall i = 1, \ldots, n$. Now consider

$$E[\sum_{i=1}^n F_{X^i}(X^i)\mathbf{1}_{S^i}] = \sum_{i=1}^n E[F_{X^i}(X^i)|S^i]\Pr(S^i) = \frac{1}{n}\sum_{i=1}^n E[F_{X^i}(X^i)|S^i].$$

Again by symmetry, $E[F_{X^i}(X^i)|S^i]$ is equal across users, therefore it is sufficient to choose a value of q that maximize $E[F_{X^i}(X^i)|S^i]$ for any user i. For simplicity, from hereon we will drop the super script in notation denoting the user in this subsection.

Note that in our scheme, even if a user is correctly identified by the access point and selected for service, it may not have the highest quantile. Let S_j, $j = \frac{n}{k}, \ldots, n$ denote the correct identification and selection of user when it has the j^{th} highest quantile. (Note that a user has to be at least the highest in its group to to be identified, therefore the minimum value of j is $\frac{n}{k}$.) Let S_r denote random selection of a user (when no user has been identified by the access point). Then

$$E[F_X(X)|S] = \sum_{j=\frac{n}{k}}^n E[F_X(X)|S_j]\Pr(S_j) + E[F_X(X)|S_r]\Pr(S_r). \tag{3}$$

Now

$$\Pr(S_n) = \int_q^1 nu^{n-1}(\frac{q}{u})^{\frac{n}{k}-1}du = \frac{n}{n(1-\frac{1}{k})+1}(q^{\frac{n}{k}-1} - q^n), \tag{4}$$

and

$$E[F_X(X)|S_n] \approx E[U^{(n)}] = \frac{n}{n+1}, \tag{5}$$

where $U^{(n)}$ is as defined in (1). Furthermore since we will be taking maximum among quantiles of successfully identified users with quantiles greater than q, $\forall j = \frac{n}{k}, \ldots, n$,

$$E[F_X(X)|S_j] \geq \frac{1+q}{2}. \tag{6}$$

Note that if the value of q is high, then the above inequality can be treated as an approximate equality.

Then from (5) and (6), one can rewrite (3) as

$$E[F_X(X)|S] \approx \frac{n}{n(1-\frac{1}{k})+1}(q^{\frac{n}{k}-1} - q^n)\Pr(S_n) + \frac{1+q}{2}\sum_{j=\frac{n}{k}}^{n-1}\Pr(S_j) + E[F_X(X)|S_r]\Pr(S_r). \tag{7}$$

Now $\sum_{j=\frac{n}{k}}^{n-1}\Pr(S_j)$ is the probability that the access point is unable to identify the user with the highest quantile, but is able to identify at least one user in the remaining $k-1$ mini slots that do not contain the user with the highest quantile. Let

$$p_s = \frac{n}{k}(1-q)q^{\frac{n}{k}-1}$$

denote the probability of a successful transmission in a typical mini slot. Then one can approximate $\sum_{j=\frac{n}{k}}^{n-1} \Pr(S_j)$ as

$$\sum_{j=\frac{n}{k}}^{n-1} \Pr(S_j) \approx (1 - \Pr(S_n))(1 - (1 - p_s)^{k-1}). \tag{8}$$

Finally

$$\Pr(S_r) = 1 - \sum_{j=\frac{n}{k}}^{n} \Pr(S_j), \tag{9}$$

and we can approximate $E[F_X(X)|S_r]$ as

$$E[F_X(X)|S_r] \approx \frac{1}{2}, \tag{10}$$

i.e., the average quantile of the selected user when it is selected completely randomly (and not conditioned on the access point being unable to identify any user).

Then (7) can be further rewritten as

$$E[F_X(X)|S] \approx \Pr(S_n)\frac{n}{n(1-\frac{1}{k})+1}(q^{\frac{n}{k}-1} - q^n) + \frac{1+q}{2}\sum_{j=\frac{n}{k}}^{n-1} \Pr(S_j) + \frac{1}{2}\Pr(S_r), \tag{11}$$

where the probabilities are defined in (4), (8) and (9).

Recall that we wanted to find the value of threshold q that maximized the value of sum quantile of users served. This can be obtained numerically by searching over $[0, 1]$ and finding the value of q that maximized (11). In Figure 2 we plot the variation of optimum threshold q for an increasing number of users for $k = 5, 7, 9$. The threshold increases with n for a given k, and with k for a given $\frac{n}{k}$, this is expected.

Note that (11) is independent of users' channel capacity distributions. The expression derived in (11) can also be used to find an approximate value of q even when $\frac{n}{k}$ is not an integral value.

Some final comments, one observes that opportunistic splitting can also be combined with maximum quantile scheduling to handle heterogeneity in users' channel capacity. However, opportunistic splitting was designed to only find the user with the highest quantile/rate. In a practical system the number of mini slots may be limited, and if the scheme is unable to find the user with the highest quantile in those many mini slots, a user has to be chosen at random. This is not desirable. Whereas if static splitting is unsuccessful in finding the user with the highest quantile, it tries to serve a user with high quantile. The possibility of serving a high quantile user (and not the highest) is captured in (11), in fact the expression also captures the performance of the scheme even when a user is selected at random. Therefore maximizing (11), can lead to better performance as compared to opportunistic splitting. We will verify this in Section 4.

3 Penalty due to Estimation Errors

As discussed earlier, opportunistic splitting and the scheme proposed in [16] assume that the channel capacity distributions of users are either known to the users or are known at the access point. This assumption is also required for static splitting (rate distribution has to be known by

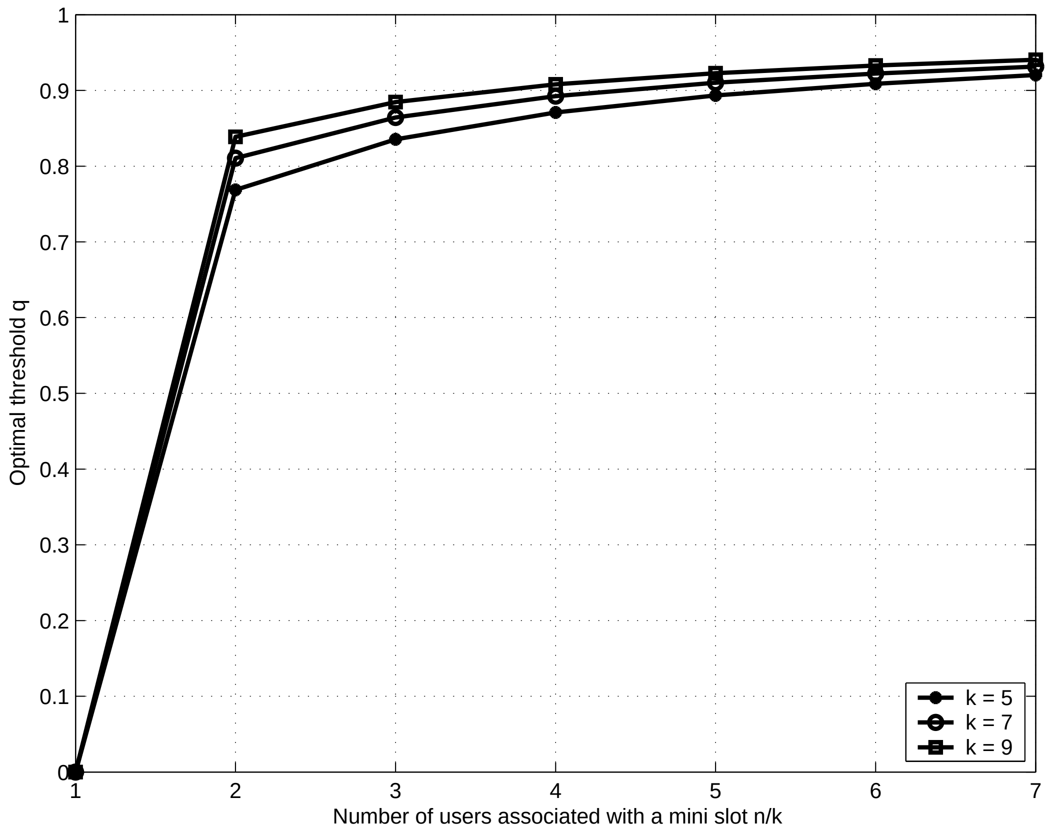

Figure 2: Variation of optimal threshold q with n and k.

the user). However this is unlikely, and one has to estimate the distribution via measurement. Therefore, one needs to evaluate the penalty incurred due to errors in estimation.

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 $\tilde{F}_{X^i}^{m,t}(\cdot)$ and is given by

$$\tilde{F}_{X^i}^{m,t}(x) = \frac{1}{m} \sum_{j=1}^{m} 1\{X^i(t-j) \leq x\}. \tag{12}$$

Note that the above way of estimating is similar to the score function described in [3]. Thus maximum quantile scheduling of users based on estimated distributions, would choose user $l(t)$ for service during slot t if

$$l(t) \in \arg \max_{i=1,\ldots,n} \tilde{F}_{X^i}^{m,t}(x^i(t)), \tag{13}$$

with ties being broken arbitrarily. It can be shown that even with estimated distributions as given in (12), maximum quantile scheduling will serve each user an equal fraction of time.

Recall that the $G_{mq}^i(n)$ (2) is the throughput experienced by user i under maximum quantile scheduling with when channel capacity distribution is perfectly known. Let $\tilde{G}_{mq}^i(n,m)$ be the throughput experienced by user i under maximum quantile scheduling when users' rate distribution is estimated according to (12). We now state a Theorem from [11] that gives a distribution independent bound on the relative error between $G_{mq}^i(n)$ and $\tilde{G}_{mq}^i(n,m)$.

Theorem 3.1 *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 (12) base on m* independent *samples of a user's channel and users are served using maximum quantile scheduling, then*

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

and the relative throughput penalty is bounded by

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

Opportunistic splitting attempts to identify the user with the highest quantile. If the distributions are estimated, the scheme will make the same mistakes as made by the access point under (13). Therefore Theorem 3.1 also applies to opportunistic slitting.

One can show that for reasonably large n, the above stated bound increases with n. Now in static splitting, we may not select the highest user, but we will be selecting the highest among $\frac{n}{k}, \frac{2n}{k}, \ldots, n$ users (i.e., the highest in a mini slot, or the highest among two mini slots and so on). Also, because there is no relative penalty incurred when users are selected at random, the overall relative penalty incurred by static splitting is likely to be lower than that indicated by the bound, or lower than that incurred by opportunistic splitting. (Note that static splitting selects a user only if its current quantile is above a threshold, therefore Theorem 3.1 does not directly apply to it.) We verify this conjecture in Section 4.

4 Simulation Results

We simulated the proposed scheme to observe its performance. We first describe the setup used for simulations. For simplicity, we assume that all users undergo i.i.d. Rayleigh fading with a mean SNR of 2. The bandwidth associated with each user is 500 KHz and we assume that coding achieves the Shannon rate. We set $k = 5$, and increase the number of users from 5 to 35 in steps of 5, i.e., the maximum number of users that can contend in a mini slot are equal across slots and increase from 1 to 7. The threshold q is set as discussed in Section 2.

We also compared our scheme's performance with that of opportunistic splitting. Note that a mini slot in opportunistic splitting consists of two transmissions, whereas a mini slot in our scheme consists of only one transmission. Also note that in our scheme, at the end of mini slots an extra transmission is required from the access point informing the user it has selected for service, i.e., in total $k + 1$ transmissions are required. Therefore to be fair, we compare our scheme with a truncated form of opportunistic splitting where at most $\frac{(k+1)}{2} = 3$ mini slots are used. At the end of 3 mini slots if the algorithm is unable to find the user with the highest quantile, then it selects a user at random. Note that 3 is greater than the average of 2.5 slots needed for the scheme to converge.

We performed two experiments, in the first experiment we observed the throughput performance of the schemes when the channel capacity distributions are perfectly known. We plot the loss in throughput (due to lower feedback) relative to the case when the user with the highest quantile is known in Figure 3. It is clear from the figure that static splitting outperforms opportunistic splitting. When there are 10 users in the system, static splitting has a 40% lower loss compared to opportunistic splitting.

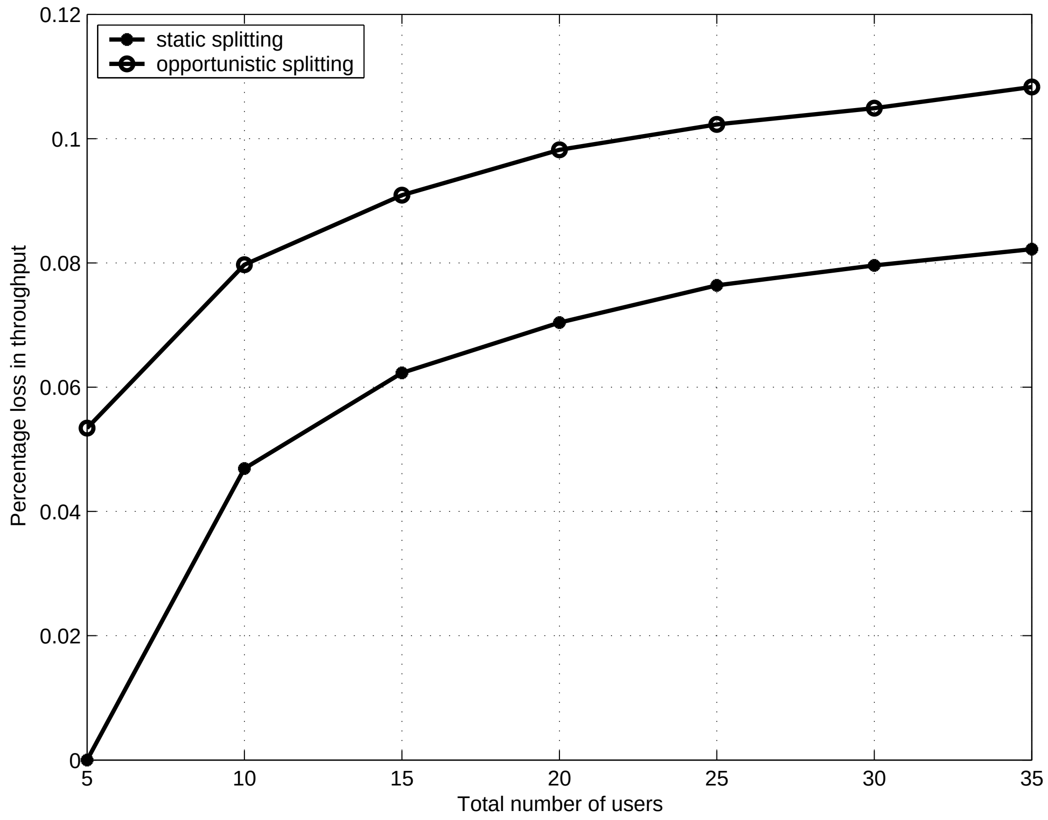

Figure 3: Relative throughput loss due to low feedback under static splitting and opportunistic splitting with increasing number of users.

In the second experiment we study the penalty suffered due to estimating channel capacity distributions. The distributions were estimated using $m = 100$ independent previous samples. We plot the throughput penalty compared to perfectly known distributions case for both static splitting and the truncated opportunistic splitting in Figure 4. As expected static splitting performs better.

5 Conclusion

In this paper we presented a simple scheme for reducing feedback in opportunistic scheduling networks. The scheme is novel in the sense that one can calculate the thresholds independent of users' distribution making it very much applicable to real world scenarios. We also develop insights into the loss incurred when rate distributions are estimated.

Unlike previous work our approach is focussed on finding a user with high quantile, not necessarily the user with the highest quantile. The advantage of such an approach is that it can lead to better overall performance. This is verified by simulation results.

References

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

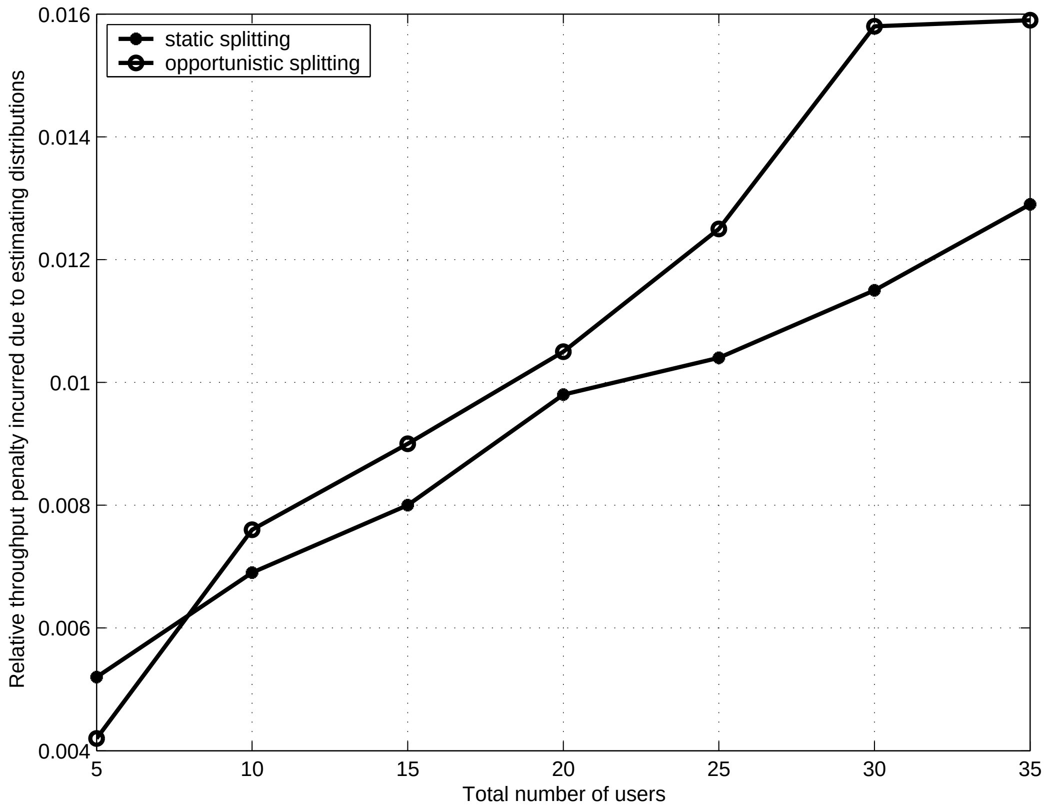

Figure 4: Relative throughput penalty due to estimation error under static splitting and opportunistic splitting.

[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] D. Gesbert and M. Slim-Alouini. How much feedback is multi-user diversity really worth? In *Proc. Int. Conf. on Commun.*, pages 234–238, June 2004.

[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] 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.

[9] 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.

[10] S. Patil and G. de Veciana. Managing resources and quality of service in wireless systems exploiting opportunism, submitted to Infocom 2006.

[11] S. Patil and G. de Veciana. Measurement-based opportunistic scheduling for heterogenous wireless systems, submitted to Infocom 2006.

[12] 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.

[13] X. Qin and R. Berry. Opportunistic splitting algorithms for wireless networks. In *INFOCOM 2004. Twenty-Third Annual Joint Conference of the IEEE Computer and Communications Societies*, March 2004.

[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] 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.

[16] T. Tang and R. W. Heath. Opportunistic feedback for downlink multiuser diversity. *To appear in IEEE Communication Letters*, 2005.