Opportunistic Communication: From Theory to Practice

Raul Etkin and David Tse
Dept. of Electrical Engineering and Computer Science
University of California, Berkeley
e-mail: {retkin,dtse}@eecs.berkeley.edu

Abstract

The variation of the channel gains in a fading channel can be exploited by opportunistic communication. In a multiuser system throughput performance can be improved considerably by transmitting to the users only when their channels are strong. For the downlink channel, the proportional fair scheduling algorithm that we propose tries to achieve multiuser diversity gains while maintaining resource fairness among users. Essential for the performance of this algorithm is the accurate measurement and prediction of the channel fading states. It is shown that mobile speed has a strong impact on channel uncertainty, which can be controlled if multiuser diversity is considered in the early stages of system design. Some techniques to reduce channel state uncertainty are proposed, and the corresponding performance gains are evaluated. In multi-cell systems it is possible to obtain important multiuser diversity gains if the system is designed to operate in an interference limited regime, and the fading states can be accurately tracked. We analyze how much performance can be gained by operating in an interference limited scenario, as opposed to a noise limited one, and how sensitive is this performance gain to channel uncertainty. Finally we consider the particular issues involved in exploiting multiuser diversity in the uplink channel.

1 Introduction

The past decade has seen a resurgence of research activities in wireless communication theory. Many of the recent conceptual advances in wireless communication theory are centered around the question of how to communicate effectively and efficiently over *fading channels*. Indeed, perhaps the most fundamental and unique characteristic of wireless channels is the time-variation of the channel strengths due to multipath fading, shadowing and path loss effects. The traditional view of fading is that it is a source of *unreliability* that has to be *compensated* for by various techniques. Central among them are *power control* and *diversity combining*. The basic goal of all these techniques is to convert the time-varying fading channel into a constant non-faded one.

In contrast, the modern view of fading is that it is a source of *randomization* that can be *exploited* to get a significant capacity boost, even beyond that of a non-faded channel. Fading can be exploited by *opportunistic communication*. By tracking the channel at the transmitter, dynamic rate and power allocation can be performed over the dimensions of *time, frequency, antennas* and *users* in a wireless system. In contrast to traditional approaches attempting to average out or invert the fading, transmission is instead done opportunistically, only *where* and *when* the channel is very good. Compared to a non-faded channel with the same average signal-to-noise ratio (SNR), fading

is *beneficial* to these schemes. This is because the randomness ensures that the channel is very strong sometime and somewhere, far above the average. For these opportunistic schemes, it is the peak channel conditions that govern performance, not the average channel condition.

The idea of opportunistic communication is first proposed in the flat fading point-to-point context [1]. Assuming perfect channel state information (CSI) at the transmitter, it was shown that for a given average power constraint, the optimal (capacity achieving) strategy is to perform waterfilling over the fading states, transmitting more information when the channel is good and less (or not at all) when the channel is bad. However, for typical fading channel models and SNR levels, it was found that there is limited performance gain compared to the strategy of transmitting at constant power and rate (where feedback is not required.) In particular, at high SNR, it can be shown that the capacity difference between the two strategies approaches 0. In this regime, most channel states are very good and the dynamic power allocation makes little difference.

The situation gets more interesting in multiuser systems. Here, in addition to the time dimension, there is a user dimension: resource can be adaptively allocated to different users depending on their time-varying channel conditions. It was shown that in the uplink (multiple users to the base-station), the optimal strategy to achieve the total information capacity (sum capacity) of the system is to allow only the user with the strongest channel strength to transmit at any one time [2]. For the downlink from the base station to the users, the same strategy is also optimal [3, 5]

To illustrate the performance gain from opportunistic communication, let us focus on the downlink channel. Under the optimal strategy, the spectral efficiency (total bits/s/Hz, summed over all users) of downlink fading channels can be evaluated. Figure 1 tells the story. We plot as a function of the number of users the spectral efficiencies of two downlink channels: a non-faded channel where each user has the same constant SNR, and a fading channel where each user undergoes independent Rayleigh fading with the same average SNR.

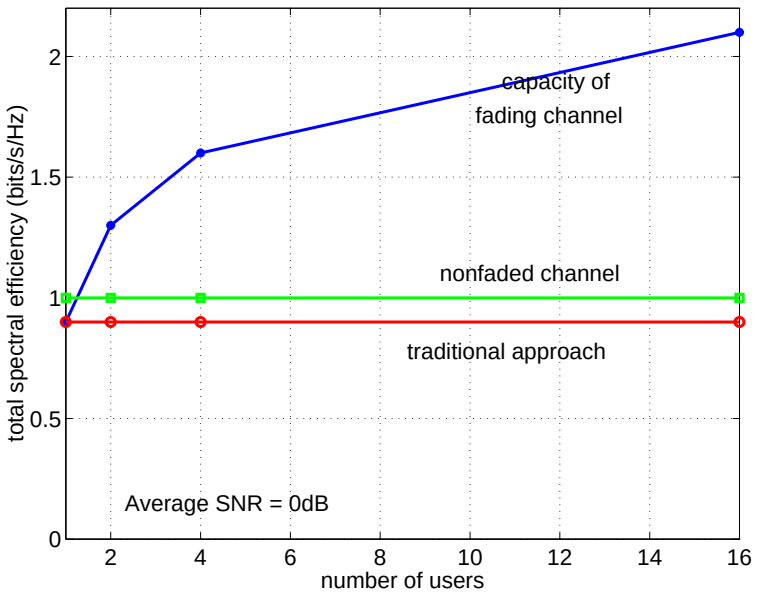

Figure 1: To fade or not to fade?

When there is one user in the system, the capacity of the fading channel is less than the

non-faded one. However, the *reverse* is true when there are multiple users in the system. While the spectral efficiency of the non-faded channel does not depend on the number of users in the system, that of the fading channel *increases* with the number of users; in a system with 16 users, the spectral efficiency of the fading channel is more than twice that of the non-faded channel. This is a *multiuser diversity* effect: when there are a large number of users in the system, with high probability there is always a user whose channel strength is near its peak. By transmitting to that user, the channel is used most efficiently. Observe that users are kept orthogonal under the optimal strategy. In second-generation systems like GSM or CDMA, users are orthogonalized as well (by time or by codes) but in a channel *independent* way. These traditional approaches cannot exploit multiuser diversity; the third line in the plot is indicative of their performance.

Based on this idea of multiuser diversity, a scheduling algorithm, the Proportional Fair Algorithm, has been designed and is now implemented in Qualcomm's HDR system [4, 5]. HDR is a wide-area wireless data system operating on the 1.25 MHz IS-95 band [6]. In the downlink, each user measures its own channel strength based on a common pilot signal, predicts the channel quality in the next time slot and feeds back the information to the base station in terms of a requested data rate that the channel is predicted to support. The downlink operates on a TDMA basis, and in each time-slot, the scheduler decides which user to serve based on the information fed back. Once it decides on a user, it will serve it at the requested data rate.

Direct implementation of the strategy of always serving the strongest user is not feasible due to issues of fairness (asymmetric users' channel statistics) and delay (a user cannot wait for arbitrarily long time for its channel condition to improve.) The PF scheduling algorithm exploits the inherent multiuser diversity gain but at the same time shares the benefit fairly among individual users, providing good throughput over a pre-specified latency time-scale. In brief, rather than serving the user with the strongest channel in *absolute* terms, the scheduler instead serves the user whose channel strength *relative* to its own average channel condition over the latency time-scale is the strongest. The goal is to serve each user when its channel condition is near its peak within the latency time scale.

Random channel variations provide *opportunity* for performance gain, but if the variations are fast, there is *uncertainty* about where and when the good channel states are. Moreover, if the channel state is tracked by the receiver feeding back the information to the transmitter, the delay in the feedback loop incurs additional uncertainty. The issues of channel measurement, prediction and feedback are essential for opportunistic communication.

Figure 2 gives some insights on the issues involved in realizing multiuser diversity benefits in practice. The plot shows the total throughput of the HDR downlink in three simulated environments:

- fixed: users are fixed but there are movements of objects around them (2 Hz Rician, K-factor = 5);
- low mobility: users move at walking speeds (3 km/hr, Rayleigh) ;
- high mobility: users move at 30 km/hr (Rayleigh).

While these simulation results pertain to a single cell with users having symmetric statistics, a full-scale simulation with multiple cells and users spread across the cells shows similar trends [7].

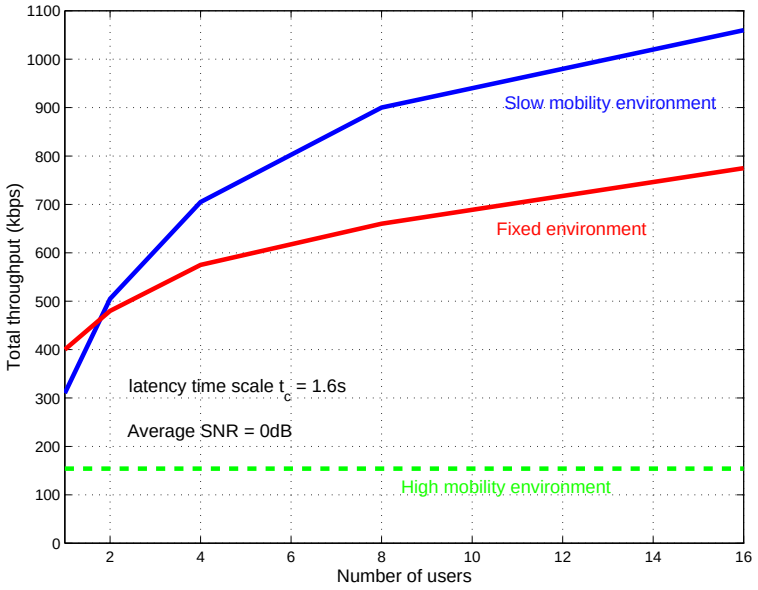

Figure 2: Performance of scheduling algorithm in different fading environments.

The total throughput increases with the number of users in both the fixed and low mobility environments, but the increase is more dramatic in the low mobility case. While the channel fades in both cases, the dynamic range and the rate of the variations is larger in the mobile environment than in the fixed one. This means that over the latency time-scale (1.67s in these examples) the peaks of the channel fluctuations are likely to be higher in the mobile environment, and the peaks are what determine the performance of the scheduling algorithm. Thus, the inherent multiuser diversity is higher in the low mobility environment than in the fixed environment.

Should one then expect an even higher throughput gain in the high mobility environment? In fact quite the opposite is true. The total throughput hardly increases with the number of users! It turns out that at this speed, the receiver has trouble tracking and predicting the channel variations, so that the predicted channel is a low-pass smoothed version of the actual fading process. Thus, even though the actual channel fluctuates, opportunistic communication is impossible without knowing when the channel is actually good.

Summarizing: there is inherently very significant multiuser diversity gain to exploit in mobile environments, but in practice the challenge is to be able to exploit it despite the channel uncertainty, particularly in high mobility environments.

Experience in implementing opportunistic communication in the face of channel uncertainty in Qualcomm's HDR system suggests that the performance gain is limited in highly mobile environments if uncertainty is not directly taken into account in the design of the dynamic resource allocation scheme. Our goal is to address the issues that the system designer must face in order to exploit multiuser diversity in conditions of high mobility. Some of these issues are choice of scheduling algorithm, feedback delay and overheads, rate selection, and interference nulling.

We will use $\ln(\cdot)$ for natural logarithm, $\log_2(\cdot)$ for base-2 logarithm, $\lfloor \cdot \rfloor$ for the floor function, $\mathcal{N}(\mu, \sigma^2)$ for the distribution of a Gaussian random variable with mean μ and variance σ^2, and $\mathcal{CN}(0, \sigma^2)$ for the distribution of a circularly symmetric complex Gaussian random variable with variance σ^2.

2 Scheduling

2.1 Throughput and fairness

There are two main objectives that a forward-link scheduling algorithm should fulfill:

1. It should allocate resources fairly to the different users,
2. It should try to maximize the total throughput of the system while being fair.

The design of the scheduling algorithm that we will describe stems from the observation that under dynamic channel conditions when the requested rates vary with time, imposing a fairness constraint at each system state may be too stringent. Rather, it may be sufficient to maintain fairness in the throughputs to the users, not at every instant of time but averaged over a certain time-scale. This fairness time-scale is naturally tied to the worst-case latency requirement of the traffic. For data traffic with a somewhat lax delay requirement (in contrast to voice), this time-scale can be of the order of seconds. If the channels vary significantly over this time-scale, the claim is that increased throughput can be obtained by relaxing the instantaneous fairness constraint to the average fairness constraint.

To get a feel of how this performance gain can come by, consider an idealized symmetric situation when the channel statistics of all the users are independent and identical over the fairness time-scale. Then a scheduling algorithm which, at every time-slot, always transmits to the user with the highest DRC[1] clearly maximizes the total throughput, but at the same time it is perfectly fair in the sense that the throughputs of the users, averaged over the fairness time-scale, are the same. This is because over the fairness time-scale, each user will have, on the average, equal amount of time to be selected for transmission. On the other hand, this algorithm is certainly not fair at each time instant: the user with the best channel at a given time seizes all resources.

The performance gain comes from the fact that the scheduler is always using the transmission resource (i.e. time) in the most efficient way: it tries to hit the users while they are at their peaks in the requested rates. If we were to try to maintain fairness at each channel state, then efficiency is compromised because some time-slots will have to be given to users with lower DRCs at the moment. Assuming that the channel fading is independent from user to user, then at any one time there is likely to be a user with a strong channel. Thus, this performance gain can also be thought of as a form of selection diversity. Whereas normally selection diversity is derived from multiple receive antennas (antenna diversity) or signals from multiple base stations (soft handoff), this diversity is obtained by exploiting the fact that multiple users have statistically independent channels. This is the multiuser diversity gain.

This idealized example is intended to illustrate the nature of the diversity gain. In practice, channel statistics of the different users are not symmetric: users closer to the base station have stronger average channel strength than users farther away; the rates of channel fluctuations may also be different. Moreover, the statistics of channel variations are not known in advance. The challenge is to find scheduling algorithms which can adapt to channel statistics, maintain throughput fairness over a desired time-scale, and exploit the inherent multiuser diversity in the system.

[1] DRC stands for Data Rate Control, the term used in Qualcomm's HDR to denote the requested data rate.

2.2 Proportional fair algorithm

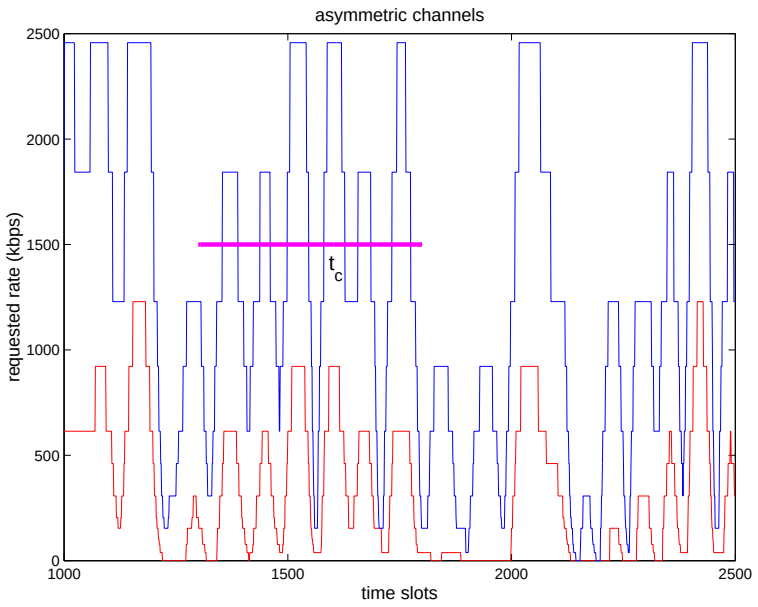

Figure 3: Requested rates of two different users with different channel statistics. t_c denotes the scheduler latency time constant.

Figure 3 shows two DRC trajectories of two different users simultaneously accessing the system. Since the users are located in different places and experience different path losses and fading, the average channel statistics are not the same. This asymmetry between the channel conditions of the users is typical in an HDR scenario, since there is no forward-link power control. Nevertheless, due to fading effects, each of the DRC processes fluctuate around their respective long-term averages, and this suggests that there is still multiuser diversity gain to be exploited as in the symmetrical case.

The algorithm which picks the user with the higher DRC at any time slot maximizes the total throughput, but is highly unfair as the user with the poorer average channel will get almost no throughput. To exploit the multiuser diversity and at the same time maintain fairness among the two users, what we really should strive to do is to schedule transmissions for a user when its current channel condition is good relative to its long term average. This way both users have a fair access to the resource and can use it efficiently. The observation motivates the following algorithm. This simplified version assumes that there is an infinite backlog of data to transmit to all users and that all packets have duration one time-slot. Let K be the number of active users.

Proportional Fair Algorithm (Version 1): The variables $T_1(t), \ldots, T_K(t)$, keep track of the average throughputs of the users in a past window prior to time t. The time constant t_c is pre-specified and dictates the size of this window.

1. Initialization: At time slot $t = 0$, set $T_k(0) = R_{min}/K$ for all k. Here R_{min} is the minimum DRC (9.6 Kbps).

2. Scheduling: At time slot t, given the current $DRC_1(t), \ldots, DRC_K(t)$ from the users, select for transmission the user k^* with the highest ratio $DRC_k(t)/T_k(t)$. Break ties randomly.

3. Updating: For k from 1 to K,

   If $k = k^*$, then

$$T_k(t+1) = \left(1 - \frac{1}{t_c}\right) T_k(t) + \frac{1}{t_c} DRC_k$$

   Else

$$T_k(t+1) = \left(1 - \frac{1}{t_c}\right) T_k(t)$$

In this algorithm, transmission is scheduled to the user with the highest current DRC relative to its own average throughput. A low-pass filter using exponential weighting is used to estimate the average throughput of each user over a past window. There is a built-in feedback mechanism in this algorithm: users which have been given greater access because of the recent strength of their channels will be automatically penalized when competing for resources in the future, since they will have a higher average throughput in the past window. This allows the control of the average throughputs of the users, and the time constant t_c for the filter reflects the fairness time-scale over which such control is exercised.

The choice of the initial conditions for the T_k's are not too important; it can be shown that if the channel statistics are stationary and ergodic, and $t_c = \infty$, the T_k's will converge to a unique equilibrium regardless of the initial conditions.

To get more insights into this algorithm, we look at several special cases:

1. Suppose the DRC's of every user remain constant over time and let the constant requested rate of user k be DRC_k. Then for $t_c = \infty$, it can be seen that at the equilibrium, the average throughput of user k is DRC_k/K and users are given equal amount of time in the scheduling policy. This is because the ratio $DRC_k(t)/T_k(t)$ is the same for each user. Thus, the proportional fair algorithm becomes the equal-time scheduling algorithm in AWGN channel conditions. More generally, if the channel conditions remain more or less constant over the fairness time-scale t_c, then the algorithm operates like the equal-time policy. This occurs when $1/t_c$ is much larger than the Doppler spread of the fading processes. Since the channel varies too slowly compared to the fairness time-scale, there is no opportunity to average over time to exploit the possible diversity gain. The algorithm in this case just maintains fairness among the users for every channel state. In the context of this example, the "fairness" of this algorithm refers to resource-fairness, i.e. users are given the same amount of access time. On the other hand, the throughput that each user gets is proportional to its own DRC; hence the notion of fairness here is also called proportional fairness.

2. If the channel statistics of all users are identical and $t_c = \infty$, then by symmetry, at the equilibrium, the T_k's are all the same and the proportional fair policy reduces to the policy of always scheduling packets to the user with the highest current DRC. In this case, the algorithm maximizes the total throughput as well.

3. Suppose the channel statistics of user k are described by $(C/I)_k = a_k X_k(t)$, where a_k models the large-scale channel propagation effect due to distance and lognormal fading, and is assumed to be constant over time but possibly different from user to user; $X_k(t)$ models the small-scale multipath fading process and is assumed to be independent and identically

distributed for all users k. We also assume that the requested rate of a user is equal to $b(C/I)$ for some constant b (which is approximately true in HDR). Then for $t_c = \infty$, at equilibrium the throughput $T_k(t)$ is proportional to a_k. That this is an equilibrium can be seen from the following argument: when the throughput of user k is proportional to a_k, $DRC_k(t)/T_k(t) = cX_k(t)$ for some constant c which is the same for all users. Thus, the algorithm would just schedule the user with the largest $X_k(t)$. Since the $X_k(t)$'s are identically distributed, users will have equal opportunity to be transmitted data, and the throughput of the kth user is given by $\frac{1}{K}ba_kE[\max(X_1(t),\ldots,X_K(t))]$. Hence the throughput of the kth user remains proportional to a_k, implying that this is an equilibrium.

This example illustrates well the intuitive idea behind the algorithm: by dividing the DRC's of each user by its average throughput, the algorithm normalizes the DRC's of users with different average channel statistics, so that the competition for resources is only dependent on the channel fluctuations $X_k(t)$'s. In this example, the $X_k(t)$'s are identically distributed, so in fact the scheduling algorithm gives on the average equal amount of time for the transmissions of each user. However, users with stronger average channel will naturally get a higher throughput. Again, "fairness" of the algorithm here refers to "resource-fairness".

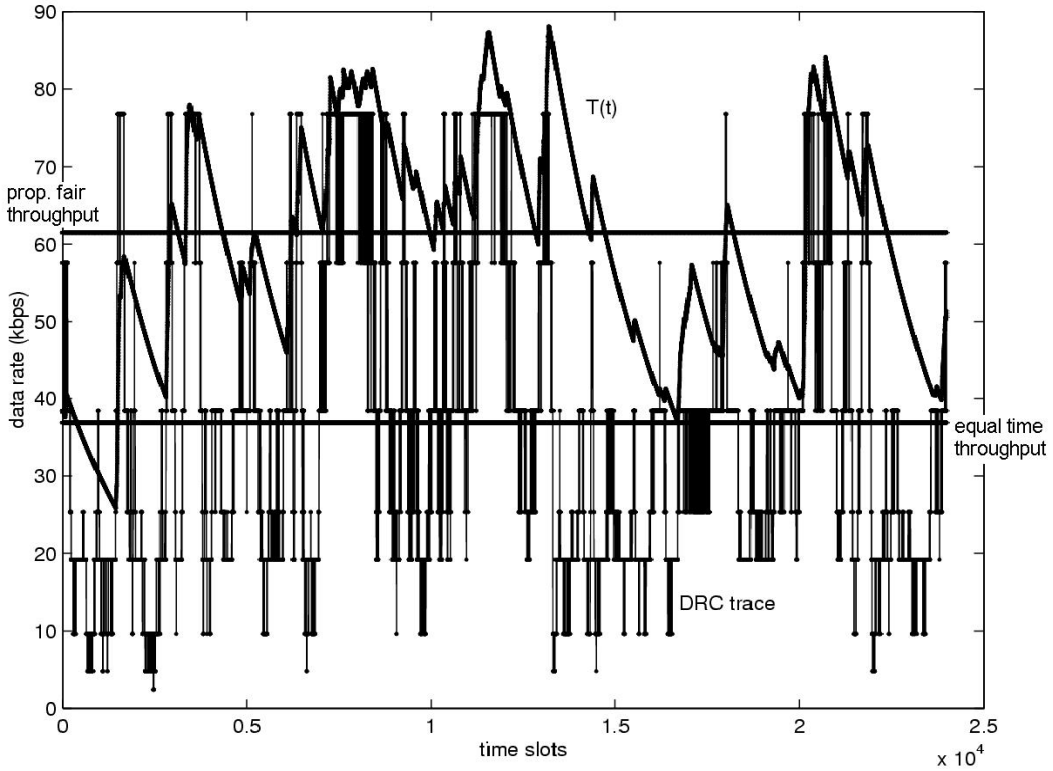

Figure 4: Dynamics of the algorithm for a user.

Figures 4 and 5 give a feel of how the algorithm works. The t_c value is chosen to be 1000 slots = 1.67s. In the first experiment, 8 users with randomly shifted versions of the same DRC trajectory are simultaneously accessing HDR, their transmissions scheduled by the algorithm defined above. Figure 4 shows the DRC trajectory of one such user, normalized by 8; this is the data rate it would get if the equal-time round-robin scheduling were done. The upper curve shows the trajectory of the throughput estimate $T_k(t)$ for this user; this curve increases whenever the user is scheduled transmission, and decays exponentially in between transmissions. Observe that most

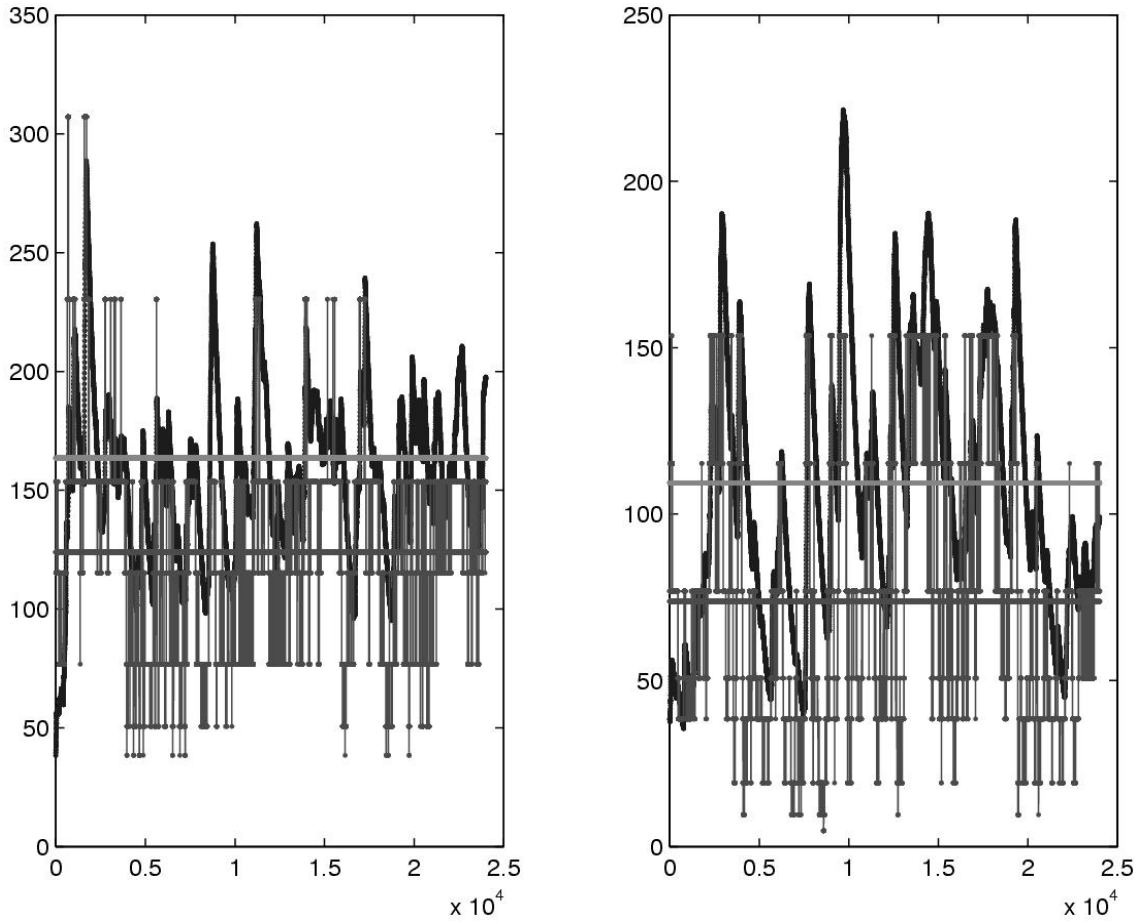

Figure 5: Dynamics of the algorithm for two users with different channel statistics.

of the transmissions are scheduled when the DRC of the user is above the long term average. As a result, the actual throughput of the user is significantly larger than the (normalized) average of the DRC trace, shown by the two horizontal lines in the figure. In the second experiment, we again multiplex 8 users, but now with 4 users having randomly shifted versions of one DRC trace, and the other 4 users having random shifted versions of another DRC trace. Figure 5 shows the behavior of two users, one from each of the two classes. Observe that although the long term statistics of the two users are different, the algorithm manages to schedule transmission to the user whose current DRC is above its own long term average, thus increasing the throughputs of both users, as compared to a channel independent equal-time scheme.

2.3 Throughput performance of the algorithm

There are several key factors determining the throughput performance gain of the algorithm:

- The dynamic range of the channel fluctuations: the larger the fluctuation, the more potential performance gain there is, since the algorithm thrives by "hitting the peaks" of the users' DRC trajectories. This brings home the point that this scheduling algorithm is really providing another form of diversity to the fading environment. If there is already a lot of diversity for each user so that their effective channel is close to AWGN, there is little gain to be exploited by this algorithm. On the other hand, one would expect this algorithm to be very helpful to

a user with a very unreliable, fluctuating channel.

- The ability of the mobile to predict the channel variations: not only has there to be significant channel variations, they should not be too fast so that the mobile can predict the variation and request the appropriate data rate for the next time slot. For fixed indoor environments where the channel varies relatively slowly (a few Hz's of Doppler) or for walking speeds, this is not expected to be a serious limitation.

- The rate of channel variation compared to the data latency requirement: the channel cannot vary too slowly either, otherwise the delay necessary to wait until the channel of a user gets better becomes intolerable.

- The number of users simultaneously accessing HDR: the larger the number of users, the more the potential gain since at any one time, there is more likely to be a user with a very good channel (good relative to its own average, of course.) Hence, when congestion in the system is caused by having many users accessing HDR at the same time rather than having one or two big users, this algorithm can be expected to be very useful in relieving that congestion. In this sense, the gain provided by this algorithm is not unlike the statistical multiplexing gain from multiplexing many bursty traffic sources together.

2.4 Fairness and optimality properties

One can in fact view the proportional fair algorithm as an adaptive algorithm which solves a certain concave optimization problem. If the DRC processes of the K users are assumed to be stationary and ergodic, then it can be shown that of all scheduling policies, the proportional fair algorithm, with $t_c = \infty$, maximizes the objective function:

$$\sum_{k=1}^{K} \log x_k$$

where x_k is the average throughput for user k. In fact, it can be shown that the present algorithm is of a coordinate steepest ascent type to get to a maxima of the objective function. Moreover, because the objective function is concave, there is a unique global maxima, and thus guarantees the convergence of the algorithm to the optimal solution from any initial conditions. Since in practice channel statistics are rarely stationary, a finite t_c is used so that the algorithm can adapt to time-varying channel statistics, as is customary in adaptive algorithm design. (This is another reason for choosing a small t_c, in addition to keeping the delays within bounds.) The log function is strictly concave, so that increasing the throughput of one user results in diminishing marginal returns, and that prevents the algorithm from giving all the resources to the user with the best channel. Contrast this with maximizing the total throughput $\sum_{k=1}^{K} x_k$ of the system, where the objective function is linear in each of the throughputs, leading to giving all the throughput to the user with the best channel. This interpretation of the proportional fair algorithm as the solution of an optimization problem suggests how it can be generalized to meet specific grade-of-service targets (worst-case ratio of throughputs of users). Instead of maximizing $\sum_{k=1}^{K} \log x_k$, we can consider maximizing $\sum_{k=1}^{K} f(x_k)$, where $f(\cdot)$ is a concave function. By choosing $f(\cdot)$ which is even "more concave" than the log function, we can reduce the disparity between the users' throughputs, at the expense of

losing some efficiency. For example, consider the following class of functions, parameterized by a fairness parameter α:

$$f(x) = -\left[-\log\left(\frac{x}{DRC_{max}}\right)\right]^{\alpha}$$

Here, DRC_{max} is the highest possible data rate, used to normalize the argument of the log function to make it less than 1. This ensures that $f(\cdot)$ as defined above is concave.

When $\alpha = 1$, this corresponds to proportional fairness. For $\alpha > 1$, a tighter form of fairness is enforced. As $\alpha \to \infty$, the throughputs of all users approach the same value, i.e. satisfies grade-of-service $G = 1$; this is also called max-min fairness, in the sense that the minimum throughput among all users is maximized. There is a natural generalization of the proportional fair algorithm to solve the optimization problem for general concave function $f(\cdot)$.

Generalized Proportional Fair Algorithm (Version 2, for arbitrary $f(\cdot)$):

1. Initialization: At time slot $t = 0$, set $T_k(0) = R_{min}/K$ for all k. Here R_{min} is the minimum DRC (9.6Kbps).

2. Scheduling: At time slot t, given the current $DRC_1(t), \ldots, DRC_K(t)$ from the users, select for transmission the user k^* with the highest ratio $DRC_k(t)f'[T_k(t)]$. Break ties randomly.

3. Updating: For k from 1 to K,

   If $k = k^*$, then

   $$T_k(t+1) = \left(1 - \frac{1}{t_c}\right) T_k(t) + \frac{1}{t_c} DRC_k$$

   Else

   $$T_k(t+1) = \left(1 - \frac{1}{t_c}\right) T_k(t)$$

In step 2, f' is the derivative of the function f. In the special case when $f(x) = \log(x)$, $f'(x) = 1/x$, and this algorithm specializes to the proportional fair algorithm. This algorithm solves the optimization problem for general f under stationary and ergodic channel conditions, when t_c is set to be infinity.

3 Channel Measurement and Prediction

As in the previous section we focus on the downlink channel of a multiuser system. As we discussed previously, the throughput performance of the proportional fair algorithm depends on the ability of the mobile to predict channel variations. As mobile speed increases these variations become harder to track, and eventually the mean value of the fading process is the only parameter that can be estimated. There are two sources of channel uncertainty:

- **Measurement error**: this source of error is due to the presence of noise and interference. If we estimate the channel fading coefficient, using training sequences for example, there will be some measurement error due to noise. This source of uncertainty can be reduced by increasing the signal power, or by spending more time in the transmission of the training sequences.

- **Prediction error**: even if we assume a perfect measurement of the channel gain at a given time, due to channel fluctuations there will be some prediction error in the estimate of the future value of the fading. As mobile speed increases, the coherence time of the channel is reduced and prediction error becomes the dominant source of channel uncertainty.

At high SNR measurement error is small and the estimation performance is limited by prediction error. The proportional fair algorithm schedules users when their SNR is high compared to their average, so even if the average SNR of a given user is small, with high probability he will be scheduled when his SNR is large. As a result, we can neglect measurement error in most SNR regimes[2].

In order to exploit multiuser diversity at high mobile speeds it is necessary to reduce prediction error. One possible approach is to increase the frequency of channel measurements and to reduce the frame length, so that the scheduling decision occurs within a short interval of channel measurement. This alternative brings many system design issues that need to be considered:

- **Increased feedback**: a higher frequency of channel measurements results in an increase in the amount of information that must be fed back, and consequently on the bandwidth utilization of the reverse link.
- **Increased overheads**: each frame must include some information to identify the target user. As the frame size is reduced this overhead may become significant.
- **Processing delays**: when the frame size is small, processing delays may limit performance. However, with increasing processor speeds and with reasonable frame lengths, processing delays should not be the bottleneck for exploiting multiuser diversity.

We will address these issues in later sections, where we will show some strategies to overcome these difficulties. In this section we will analyze the effect of mobile speed on channel uncertainty, and evaluate the benefits that can be obtained from the reduction of the frame size.

3.1 Channel model

Consider the downlink channel of a single cell system for the case when the coherence bandwidth is smaller than the reciprocal of the symbol duration. This means that there is a single multipath component arriving at the receiver. We assume that the channel gains remain constant for the frame duration F (measured in number of symbols), and that they evolve in time as a Gauss-Markov process. The frame length in seconds T is given by $F \cdot T_s$ where T_s is the symbol period in seconds. The discrete time received signal at mobile k, $(1 \leq k \leq K)$ is given by:

$$y_k[n] = h_k\,[m]\,x[n] + z_k[n] \tag{1}$$

where $m = \lfloor n/F \rfloor$ is the frame index, K is the number of users, h_k is the channel gain from the base station to the receiver k, x is the transmitted signal constrained to have average power P, and z_k is additive white circularly symmetric complex Gaussian noise of variance σ_k^2 received by user k.

[2] If there is interference, SNR should be interpreted as SINR, signal to interference plus noise ratio.

The channel gains h_k, $(k = 1, \ldots, K)$ are generated using the following autorregressive model

$$h_k[m+1] = \sqrt{1-\epsilon_k} \cdot h_k[m] + \sqrt{\epsilon_k} v_k[m] \tag{2}$$

where $0 \leq \epsilon_k \leq 1$ and $v_k \sim \mathcal{CN}(0, \sigma_{h_k}^2)$ are independent white processes. It follows that $h_k[m] \sim \mathcal{CN}(0, \sigma_{h_k}^2)$. The coherence time of the channel is controlled by the parameter[3] ϵ. As $\epsilon \to 0$ we get the limiting case of a constant channel.

It is useful to compute the typical values of ϵ for different applications. The coherence time of the channel T_c represents the time over which the fading coefficients are highly correlated. If we define T_c as the time over which the autocorrelation function is above 0.5 of its value at 0 then [8]:

$$T_c \approx \frac{9}{16\pi f_m} \tag{3}$$

where f_m is the maximum Doppler shift given by $f_m = v/\lambda$, v is the mobile speed, and λ is the wavelength. We can compute the autocorrelation of the process defined by (2), set it equal to $1/2$ and solve for the corresponding value of ϵ:

$$\epsilon = 1 - 2^{\frac{-2T}{T_c}} \tag{4}$$

For Qualcomm's HDR system the frame length T is 1.67ms. At the 1.9 MHz band, with a mobile speed of 3 km/h, $\epsilon = 6.6 \cdot 10^{-2}$. For a mobile speed of 30 km/h, we have $\epsilon = 0.495$.

As the mobile changes its location or its speed, the channel parameters defined above vary. However these variations occur in a time scale of seconds or minutes, while the fading phenomena that our model represents take place in a time scale of milliseconds. Therefore, we will restrict our attention to a quasi-stationary scenario.

3.2 Channel estimation

Under the assumption of negligible interference we analyze the problem of predicting the channel gain h_0 using a pilot signal for channel measurement. Without loss of generality we normalize the transmitted signal power by setting $P = 1$. At time nT a pilot of duration L is transmitted, resulting in the following received symbols

$$y[nT + j] = h_0[n] \cdot 1 + z[nT + j]$$

where $j = 0, 1, \cdots, L-1$, from which we construct the sufficient statistic

$$\overline{y}[n] = \frac{1}{L} \sum_{j=0}^{L-1} y[nT + j] = h_0[n] + \frac{1}{L} \sum_{j=0}^{L-1} z[nT + j] = h_0[n] + \overline{z}[n]$$

where $\overline{z}[n] \sim \mathcal{CN}(0, \frac{\sigma_z^2}{L})$.

We estimate $h_0[n+N]$ from $(\overline{y}[n-j], j \geq 0)$ using a causal linear least squares predictor. Here N is the time in the future when the prediction will be used. If there is instant feedback the

[3] We will omit the subindex k to simplify notation when the remark applies to all users $1 \leq k \leq K$.

scheduling decision is done for the next frame after the measurement of the pilot, which corresponds to $N = 1$. In practice there will be some delay in the feedback channel so $N \geq 1$. We will use $\hat{h}[n]$ to denote the causal minimum mean square estimate of $h_0[n+N]$.

The mean square prediction error σ^2 is given by:

$$\sigma^2 = \sigma_{h_0}^2 \left(1 - \frac{\sigma_{h_0}^2}{\sigma_z^2} \frac{z_1 \epsilon (1-\epsilon)^{\frac{2N-1}{2}} L}{(1 - z_1\sqrt{1-\epsilon})^2}\right) \tag{5}$$

where

$$z_1 = \frac{1}{2}\left(\frac{\sigma_{h_0}^2}{\sigma_z^2}\frac{\epsilon L}{\sqrt{1-\epsilon}} + \sqrt{1-\epsilon} + \frac{1}{\sqrt{1-\epsilon}}\right) - \sqrt{\left[\frac{1}{2}\left(\frac{\sigma_{h_0}^2}{\sigma_z^2}\frac{\epsilon L}{\sqrt{1-\epsilon}} + \sqrt{1-\epsilon} + \frac{1}{\sqrt{1-\epsilon}}\right)\right]^2 - 1}$$

3.3 Rate selection

After the prediction of the channel each receiver feeds back the estimate $\hat{h}[n]$ to the base station. Following the scheduling algorithm, the base station selects one user for transmission at the next time slot. Based on the estimated SNR $\hat{h}[n]/\sigma_z^2$ the transmitter must select a code and modulation scheme that ultimately determine the rate of data transmission. To make the analysis independent of the coding problem, we assume the use of a large block length and strong codes that approximately achieve capacity with negligible probability of error.

If the channel estimate $\hat{h}[n]$ was perfect, the base station could select an appropriate transmission rate for the channel conditions and achieve and arbitrarily small error probability. In practice the channel estimate $\hat{h}[n]$ is not perfect, and for each fixed rate R there is a positive probability (conditioned on the measurement $\hat{h}[n]$) that there is an outage, i.e. the event that the channel $h_0[n+N]$ does not support the selected rate. We emphasize that outages occur due to channel uncertainty.

In the event of an outage there is a high probability of decoding error, and as a result, an ARQ (automatic repeat request) protocol must be used to retransmit corrupted frames. In the selection of $R(\hat{h}[n])$ there is a trade-off between throughput in case of a successful transmission and outage probability. A reasonable design approach is to fix the value of the outage probability γ conditioned on $\hat{h}[n]$ and solve for the resulting rate $R(\hat{h}[n], \gamma)$. In a later subsection we will optimize over the choice of γ to maximize the overall sector throughput. We define a parameter $\beta(\hat{h}[n], \gamma)$ related to $R(\hat{h}[n], \gamma)$ in the following way

$$R(\hat{h}[n], \gamma) = \log_2[1 + \beta^2(\hat{h}[n], \gamma)]$$

The outage probability is determined as follows,

$$\begin{aligned} \gamma &= E[\gamma(R(\hat{h}[n]), \hat{h}[n])] = E\left\{P\left[R(\hat{h}[n]) > \log_2\left(1 + \frac{|h[n+N]|^2}{\sigma_z^2}\right)\middle|\, \hat{h}[n]\right]\right\} \\ &= E\left\{P\left[\beta(\hat{h}[n]) > \frac{|h[n+N]|}{\sigma_z}\middle|\, \hat{h}[n]\right]\right\} \end{aligned} \tag{6}$$

Making $\gamma(R(\hat{h}[n]), \hat{h}[n])$ constant for all values of $\hat{h}[n]$ lets us remove the expectation operator in (6). To calculate the above probability we write $\hat{h} = h + h_\epsilon$, where the time indices have been removed for convenience in notation. By the orthogonality principle the estimation error h_ϵ is uncorrelated with the observations $(\overline{y}[n-j], j \geq 0)$, and hence it is uncorrelated with $\hat{h}$ which is a linear function of the observations. In addition to that, $(\hat{h}, h_\epsilon)$ are jointly Gaussian so the previous remark implies that they are independent. Therefore we can rewrite (6) as

$$\gamma = P\left[\beta(\hat{h}) > \frac{\left|\hat{h} - h_\epsilon\right|}{\sigma_z} \middle| \hat{h}\right] = \int_{\mathcal{D}} dF_{h_\epsilon} \tag{7}$$

where F_{h_ϵ} is the distribution function of a $\mathcal{CN}(0, \sigma^2)$ random variable with σ^2 given by (5) and $\mathcal{D} = \left\{ z \in \mathcal{C} \middle| \beta(\hat{h}) > \frac{|\hat{h}-z|}{\sigma_z} \right\}$. For fixed $\hat{h}$, $\beta(\hat{h})$ remains constant and we can evaluate the integral in (7) numerically. However, we are interested in fixing γ and solving for $\beta(\hat{h}, \gamma, \sigma^2)$. Even though it is possible to do this numerically, this turns out to be computationally intensive and inaccurate for extreme values of the parameters, particularly for values of $\epsilon \approx 0$ (slow mobile speeds). We propose approximating the function $\beta(\hat{h}, \gamma, \sigma^2)$ with first and second order polynomials in the following way:

$$\beta(\hat{h}, \gamma, \sigma^2) = \begin{cases} a\left|\hat{h}\right|^2 + b & 0 \leq \left|\hat{h}\right| \leq e \\ c\left|\hat{h}\right| + d & e < \left|\hat{h}\right| < \infty \end{cases} \tag{8}$$

where the parameters a, b, c, d, e are functions of γ and σ^2.

In Appendix A we show how to obtain these constants as a function of the model parameters. The resulting approximation is given by:

$$\beta(\hat{h}, \gamma, \sigma^2) = \begin{cases} \left[4\sigma\sigma_z\left(\sqrt{\ln\left(\frac{1}{1-\gamma}\right)} - \frac{1}{\sqrt{2}}\Phi^{-1}(\gamma)\right)\right]^{-1}\left|\hat{h}\right|^2 + \frac{\sigma}{\sigma_z}\sqrt{\ln\left(\frac{1}{1-\gamma}\right)} & 0 \leq \left|\hat{h}\right| \leq e \\ \frac{1}{\sigma_z}\left|\hat{h}\right| + \frac{\sigma}{\sqrt{2}\sigma_z}\Phi^{-1}(\gamma) & e < \left|\hat{h}\right| < \infty \end{cases} \tag{9}$$

with $e = 2\sigma\left(\sqrt{\ln\left(\frac{1}{1-\gamma}\right)} - \frac{1}{\sqrt{2}}\Phi^{-1}(\gamma)\right)$, and $\Phi(\cdot)$ being the distribution function of a $\mathcal{N}(0,1)$ random variable.

As we can see in Figure 6, the previous expression gives a good approximation of the exact function $\beta(|\hat{h}|, \gamma, \sigma^2)$.

3.4 System performance

To evaluate the influence of the different parameters and design choices on system performance we calculate an approximate expression for the achievable average data rate. We assume a sufficiently large block length to neglect the probability of decoding error, and consider the use of channel codes with performance close to the capacity of the AWGN channel. If all users have channels with the same statistics, and if we let the scheduler time constant $t_c \to \infty$ the proportional fair algorithm reduces to selecting the user with the largest channel estimate. Therefore at each time slot the

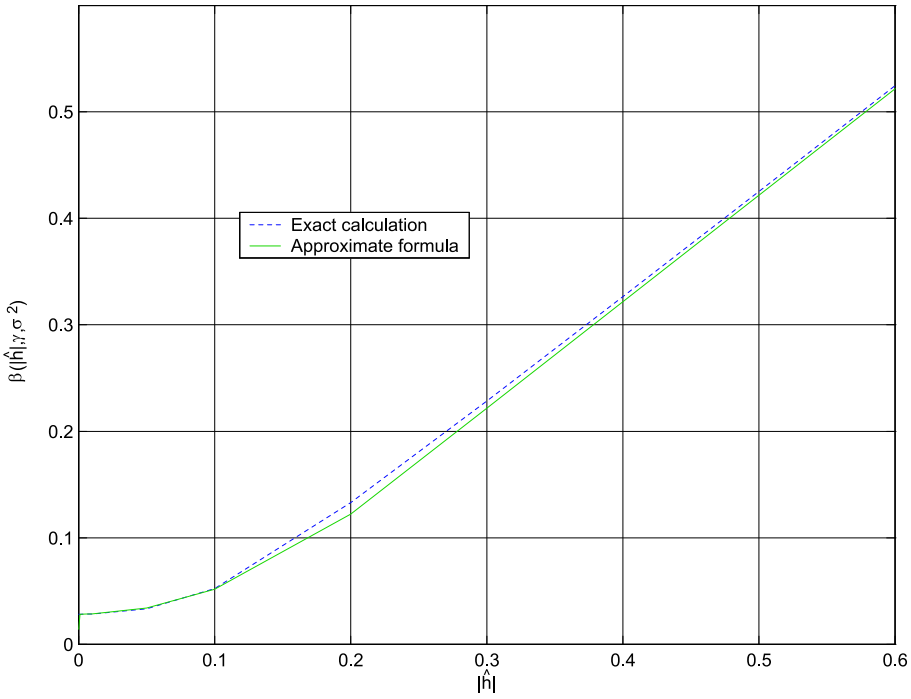

Figure 6: Exact and approximate plots of the function $\beta(|\hat{h}|, \gamma, \sigma^2)$ for $\gamma = 0.1$, $\sigma_z^2 = 1$ and $\sigma^2 = 0.00756$

scheduler selects the user $\bar{k} = \arg\max_{1 \le k \le K} |\hat{h}_k|$. Under these assumptions the spectral efficiency in bits per second per Hertz, for a fixed value of $|\hat{h}_{\bar{k}}|$ for the scheduled user and conditioned on the event that there is no outage, is

$$R(|\hat{h}_{\bar{k}}| \mid \overline{\text{Outage}}) = \log_2[1 + \beta^2(|\hat{h}_{\bar{k}}|, \gamma, \sigma^2)]$$

We now remove the conditioning on the outage event

$$R(|\hat{h}_{\bar{k}}|) = (1-\gamma) R(|\hat{h}_{\bar{k}}| \mid \overline{\text{Outage}}) + \gamma R(|\hat{h}_{\bar{k}}| \mid \text{Outage}) = (1-\gamma) \log_2[1 + \beta^2(|\hat{h}_{\bar{k}}|, \gamma, \sigma^2)]$$

and finally take expectation over $|\hat{h}_{\bar{k}}|$ to get

$$\begin{aligned} R &= E\left\{(1-\gamma)\log_2\left[1+\beta^2(|\hat{h}_{\bar{k}}|,\gamma,\sigma^2)\right]\right\} \\ &= (1-\gamma)\int_0^\infty \log_2\left[1+\beta^2\left(\sqrt{x(\sigma_{h_0}^2-\sigma^2)},\gamma,\sigma^2\right)\right] f_{X_{\max}}(x)dx \end{aligned} \tag{10}$$

where $X_{\max}$ is the maximum of K independent Exp(1) random variables.

We note that the overall sector throughput R depends on the choice of γ, the outage probability. For a given set of model parameters we would like to optimize over the choice of γ to maximize R. In practice it is not possible to find a closed form expression for the optimal value of γ so we optimize over γ numerically. In Figure 7 we plot the sector's spectral efficiency for a 16 user system as a function of the outage probability γ for various mobile speeds and a frame length $T = 0.167$ms. The optimal values of γ for these mobile speeds are indicated. As can be seen from the figure a choice of $\gamma = 0.1$ results in a maximum rate loss of less than 5% for mobile speeds ranging from 1 to 20 km/h.

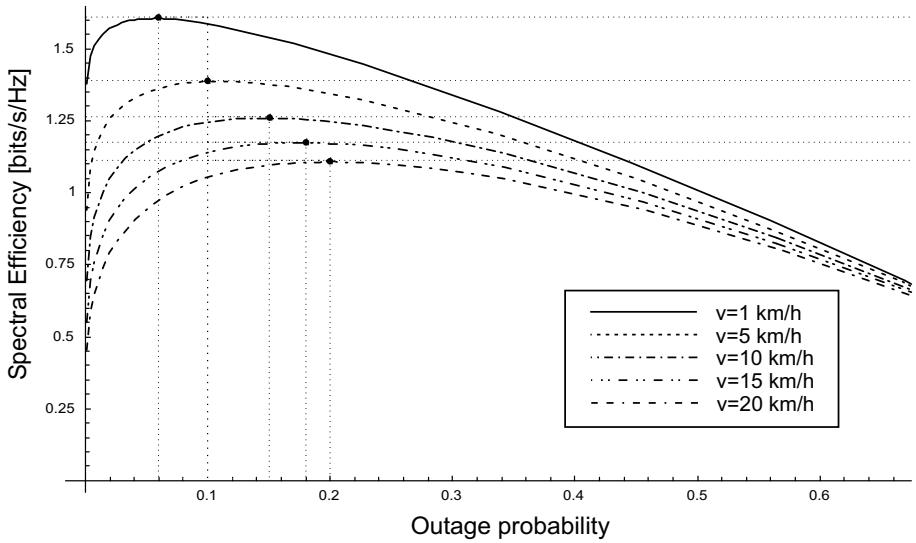

Figure 7: Spectral efficiency as a function of the outage probability γ for various mobile speeds. Model parameters: $K = 16$, $\sigma^2_{h_0} = 1$, $P = 1$, $\sigma^2_z = 1$, $L = 1$, and $T = 0.167$ms.

In Figure 8 we plot a similar set of curves, but computed for a frame length $T = 1.67$ms. In this case choosing $\gamma = 0.2$ results in acceptable performance. Although we omit the curves, a choice of $\gamma = 0.2$ is adequate for both frame lengths when the system has only one user.

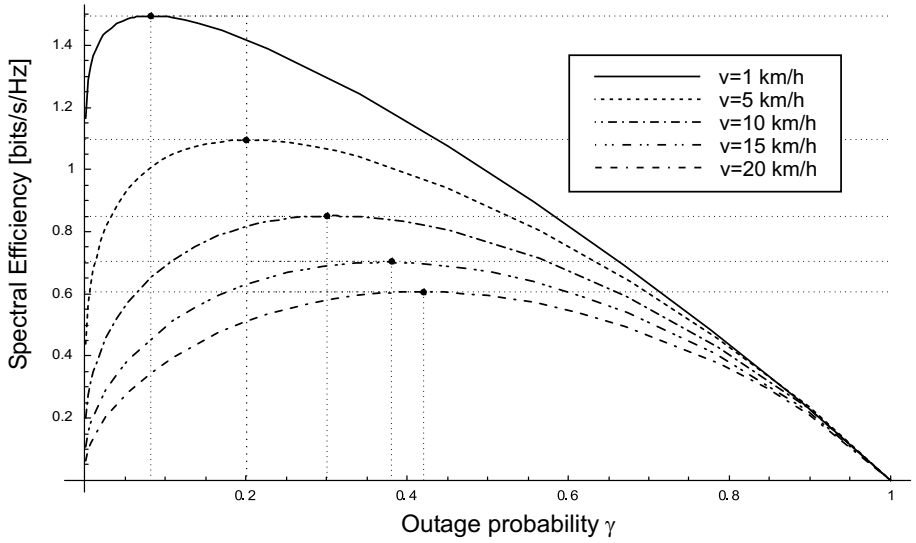

Figure 8: Spectral efficiency as a function of the outage probability γ for various mobile speeds. Model parameters: $K = 16$, $\sigma^2_{h_0} = 1$, $P = 1$, $\sigma^2_z = 1$, $L = 10$, and $T = 1.67$ms.

In Figure 9 we plot (10) as a function of the mobile speed for a single user system, i.e. $K = 1$, and a 16 user system, i.e. $K = 16$, for three different frame lengths. In both cases we consider $T = 1.67$ms, Qualcomm's HDR frame length, and also consider frame lengths 2 and 10 times shorter. We choose γ ranging from 0.1 to 0.2 depending on the case to obtain acceptable performance in speeds ranging from 1 to 20 km/h. We also included curves for the capacity of the corresponding systems calculated under the assumption of perfect knowledge of the channel state

information (CSI) at the receiver. This capacity is given by

$$C = E\left\{\log_2\left[1 + X_{max}\sigma_{h_0}^2/\sigma_z^2\right]\right\} = \int_0^\infty \log_2\left[1 + x\sigma_{h_0}^2/\sigma_z^2\right] f_{X_{\max}}(x)dx$$

We can observe that the reduction of the frame size increases the range of speeds over which

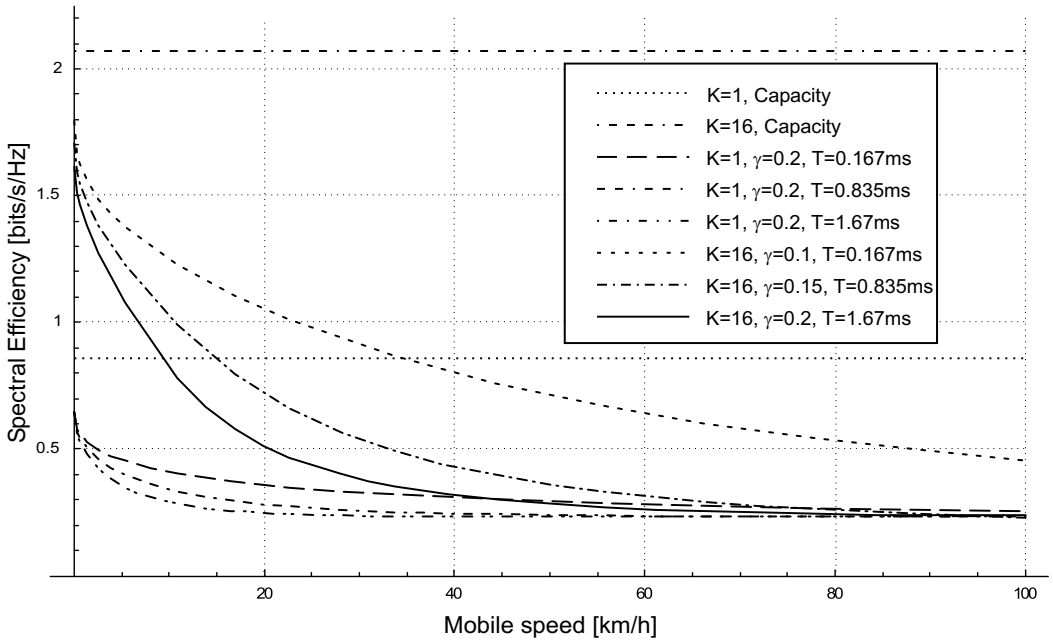

Figure 9: Spectral efficiency as a function of mobile speed for $\sigma_{h_0}^2 = 1$, $P = 1$, $\sigma_z^2 = 1$, $L = 1$ for $T = 0.167$ms, $L = 5$ for $T = 0.835$ms, and $L = 10$ for $T = 1.67$ms. Values of outage probability γ are indicated. Perfect channel knowledge capacities are plotted as a reference for comparison.

multiuser diversity is exploitable. However, an n-times reduction in the frame size does not extend the range of speeds by the same factor. This is because there is a tradeoff between prediction error and measurement error that depends on the choice of frame size.

When the frame size is reduced, the frequency of channel estimation increases and the corresponding prediction error decreases. At the same time, if we assume that the total time and power allocated for channel measurement is fixed, the higher frequency of pilot transmission forces a reduction in the pilot duration L, in order to keep the product $T \cdot L$ constant. This results in an increase in measurement error. It is this increase in measurement error one of the factors that prevents us from using an extremely small frame size.

In addition to that, the reduction of frame size produces an increase in the system overheads that must be taken into consideration when choosing the optimal value for T. In the next section we discuss strategies to keep the overheads approximately constant when reducing the frame length.

4 Reducing Reverse and Forward Link Overheads

4.1 Feedback of channel state information

The scheduling algorithm bases its decisions on the channel states fed back by each of the users. As mobile speed increases, feedback delay results in increased uncertainty about the channel state

and suboptimal scheduling. One way to reduce channel uncertainty is to minimize feedback delay, which requires frequent feedback of channel state information. As the number of users and the frequency of feedback increases, the bandwidth devoted to feedback becomes important. To reduce this overhead we propose to feed back channel state information selectively, only when the SNR measurement of the user is "good" with respect to its own average.

Let K be the number of users in a particular cell and J be the number of users that feed back channel state information at a given time. J is a random variable whose distribution depends on the choice of the selective feedback strategy. The bandwidth available in the reverse link for channel state information feedback is a limited resource, and as a result we impose the following constraint on the selective feedback algorithm: $E[J] = \bar{J}$.

The proposed strategy consists of comparing the estimate $|\hat{h}|^2$ with a threshold θ and transmit the channel state information only when $|\hat{h}|^2 > \theta$. We select θ_k for user k to satisfy

$$P(|\hat{h}_k|^2 > \theta_k) = \frac{\bar{J}}{K} \tag{11}$$

Even though this idea can be applied to any type of channel model or prediction algorithm, for simplicity we will present it for the particular model used in Section 3. From the analysis of the previous section, the random variable $|\hat{h}_k|^2$ has exponential distribution with mean $\sigma^2_{h_{0,k}} - \sigma^2_k$, where we have added subscripts k to show that different users can have different channel statistics and estimation errors. If $X_k \sim \text{Exp}(1)$ then we can rewrite (11) as

$$P\left[X_k > \frac{\theta_k}{(\sigma^2_{h_{0,k}} - \sigma^2_k)}\right] = P(X_k > \bar{\theta}_k) = \frac{\bar{J}}{K}$$

and solving for $\bar{\theta}_k$ we obtain $\bar{\theta}_k = \ln(K/\bar{J})$, or $\theta_k = (\sigma^2_{h_{0,k}} - \sigma^2_k)\ln(K/\bar{J})$. We can write $J = \sum_{k=1}^{K} 1_{(X_k > \theta_k)}$ (where $1_{(\cdot)}$ is the indicator function) and then take expectations to get $E[J] = \sum_{k=1}^{K} E[1_{(X_k > \bar{\theta}_k)}] = \sum_{k=1}^{K} P(X_k > \bar{\theta}_k) = K(\bar{J}/K) = \bar{J}$, verifying the required constraint.

We now analyze how selective feedback impacts system performance. To simplify the analysis we assume that all the users have similar channel statistics, i.e. $\sigma^2_{h_{0,k}} = \sigma^2_{h_0} \forall k$. If we also assume $t_c \to \infty$ in the scheduling algorithm, the scheduler selects the user who has fed back the largest $|\hat{h}|$. Let $\mathcal{A} = \left\{k : |\hat{h}_k|^2 > \theta, 1 \le k \le K\right\}$ be the set of users that feed back their estimates, and $\bar{k} = \arg\max_{1 \le k \le K} |\hat{h}_k|$ be the user chosen by the scheduler when there is complete feedback information. Then if $A \neq \emptyset$ it follows that $\bar{k} \in \mathcal{A}$ and there is no degradation in performance by the use of selective feedback. If on the other hand $\mathcal{A} = \emptyset$ then the transmitter does not have any channel state information. In this case we can obtain a lower bound in performance by assuming that the time slot is wasted and no information is transmitted. This is suboptimal since in general it is possible to randomly select a user and transmit at a low enough rate, or use previous estimates for the scheduling decision.

Let $X_{max} = \max_{1 \le k \le K} X_k$. Then X_{max} is the maximum of K Exp(1) random variables, and has probability density function $f_{X_{max}}(x) = K(1 - e^{-x})^{K-1} e^{-x}$, $x \ge 0$.

Assuming the use of a channel code with performance close to capacity, and neglecting the possibility of an outage, we can approximate the expected spectral efficiency with complete feedback by

$$R_{FF} = E\left[\log_2\left(1 + \frac{(\sigma_{h_0}^2 - \sigma^2)}{\sigma_z^2} X_{\max}\right)\right] = \int_0^\infty \log_2\left(1 + \frac{(\sigma_{h_0}^2 - \sigma^2)}{\sigma_z^2} x\right) f_{X_{\max}}(x)dx$$

When there is selective feedback the throughput is 0 when $X_{max} \leq \bar{\theta}$, and the resulting performance is approximately measured by

$$R_{SF} = E\left[\log_2\left(1 + \frac{(\sigma_{h_0}^2 - \sigma^2)}{\sigma_z^2} X_{\max}\right) 1_{(X_{\max} > \bar{\theta})}\right] = \int_{\bar{\theta}}^\infty \log_2\left(1 + \frac{(\sigma_{h_0}^2 - \sigma^2)}{\sigma_z^2} x\right) f_{X_{\max}}(x)dx$$

Therefore the relative loss in throughput due to the use of selective feedback is

$$\eta = 1 - \frac{R_{SF}}{R_{FF}} = \frac{\int_0^{\bar{\theta}} \log_2\left(1 + \frac{(\sigma_{h_0}^2 - \sigma^2)}{\sigma_z^2} x\right) f_{X_{\max}}(x)dx}{\int_0^\infty \log_2\left(1 + \frac{(\sigma_{h_0}^2 - \sigma^2)}{\sigma_z^2} x\right) f_{X_{\max}}(x)dx} \tag{12}$$

For a large SNR, i.e. $(\sigma_{h_0}^2 - \sigma^2)/\sigma_z^2 \gg 1$, we can approximate $\log_2(1 + \frac{\sigma_{h_0}^2 - \sigma^2)}{\sigma_z^2} x) \approx \log_2(\frac{\sigma_{h_0}^2 - \sigma^2)}{\sigma_z^2}) + \log_2(x)$ and simplify (12):

$$\eta \approx \frac{\log_2\left[\frac{(\sigma_{h_0}^2 - \sigma^2)}{\sigma_z^2}\right] P\left(X_{\max} \leq \bar{\theta}\right) + \int_0^{\bar{\theta}} \log_2(x) f_{X_{\max}}(x)dx}{\log_2\left[\frac{(\sigma_{h_0}^2 - \sigma^2)}{\sigma_z^2}\right] + \int_0^\infty \log_2(x) f_{X_{\max}}(x)dx} \approx P\left(X_{\max} \leq \bar{\theta}\right) = \left(1 - \frac{\bar{J}}{K}\right)^K$$

As we can see in the following table computed for $T = 0.167$ms, $P = 1$, $L = 1$, $\sigma_{h_0}^2 = 1$, $\sigma_z^2 = 1$, and $v = 1$ km/h, the performance loss is relatively small even with a small fraction of users feeding back their measurements.

K	20	20	20	20	50	50	50	50
$\bar{J}$	1	2	4	8	1	2	4	8
η	0.2917	0.0871	$6.714 \cdot 10^{-3}$	$1.442 \cdot 10^{-5}$	0.3141	0.1032	0.0109	$9.540 \cdot 10^{-5}$

It is important to note that even though on average $\bar{J}$ users will feed back their measurements, at any given time J is a random variable that can take any value in $\{0, \ldots, K\}$. Therefore our proposed strategy is better suited for a reverse link multiple access scheme with a soft capacity limit, such as CDMA.

4.2 Identification of target user

The previous subsection dealt with the problem of reducing system overheads in the reverse link when we reduce the frame duration T. The reduction of frame size also produces an increase in the

overheads in the forward link and this increase may become an important performance bottleneck. The main of these overheads is due to target user identification.

Once the scheduler selects a user, data is sent at the rate determined by the channel state information fed back by the user. As a result, the transmission rate is known to the user and need not be encoded with the data. In contrast, all users that fed back channel state information are possible targets so some way of identifying the target user must be included in the forward link transmission.

A valid approach is to fix the maximum allowable number of users, encode the user number with a fixed length binary number, add a channel code to reduce the decoding error probability, and append this information to the header of the frame. If this header is decoded incorrectly and the wrong user decodes the frame, the decoding of the whole packet will fail and many time slots would be wasted. As a result, the channel code used to protect the header information must be very strong. For users with low SINR, this header information may occupy a considerable fraction of the frame. If in addition we reduce the frame length to reduce channel state uncertainty, the overhead due to mobile identification may be excessive.

One possible approach to reduce this overhead is to exploit the memory of the channel and use a variable length code to identify the user. In order to exploit multiuser diversity by hitting the peaks of the fading processes the SINR should remain approximately constant during the frame duration. As a result, one should expect a small variation of the SINR between adjacent frames and, if t_c is large enough, with high probability the scheduler will choose the same user in consecutive frames. This effect can be exploited by encoding the selected user number with only one bit, zero say, if the scheduled user is the same as the one scheduled in the previous slot, and otherwise using the binary representation of the user number with a one appended in the beginning. Finally in both cases a channel code is used to reduce the error probability.

The main goal of reducing frame length is to reduce channel uncertainty for fast moving users. In a practical system users have different channel statistics, and presumably only a small fraction will be highly mobile. We could reduce the user identification overhead by using different frame lengths for slow and fast moving users. Since user mobility changes in a time scale of seconds while channel fluctuations occur in a time scale of milliseconds, we would add a negligible overhead for keeping track of mobility statistics and frame length selection.

Yet another alternative is to use larger frame lengths for low data rates, since low rate frames are the ones with the largest header overhead.

5 Opportunistic Communication in Multi-cell Systems

Whether the channel of a mobile user is noise or interference limited depends on a number of factors such as cell radius, transmission power, shadowing, background noise power, receiver noise figure, etc. Some of these parameters can be controlled by the system designer, so a natural question to ask is which is the most desirable regime of operation, and how much performance improvement can be gained by operating in that regime.

If there is a single interferer, and additive noise is negligible when compared to the interference, for sufficiently slow mobile speeds system capacity grows logarithmically with the number of users. In contrast, when the system is noise limited capacity grows doubly logarithmically with the

number of users. Therefore, if we can design a system to operate in an interference limited regime with negligible channel estimation error, we can expect large performance gains due to multiuser diversity.

In the rest of this section we will provide quantitative results to support the previous statement. We will also find that channel uncertainty plays a fundamental role in the validity of these results and can become a performance bottleneck. We will use K to denote the number of users in a given cell or sector, and $C(K)$ to denote the total cell/sector capacity as a function of K.

5.1 Channel Model

We consider a multi-cell system where cell number 0 is surrounded by M interfering cells, and serves K users. The discrete time received signal at mobile k, $(1 \leq k \leq K)$ is given by[4]:

$$y_k[n] = h_{0,k}[m]\, x_0[n] + \sum_{i=1}^{M} h_{i,k}[m]\, x_i[n] + z_k[n]$$

where $h_{0,k}$ is the channel gain from the base station to the receiver k, $h_{i,k}(i = 1, \ldots, M)$ are the channel gains from the interfering base stations to the receiver, x_0 is the transmitted signal, $x_i(i = 1, \ldots, M)$ are the interfering signals, and z_k is additive white circularly symmetric complex Gaussian noise received by user k, with variance $\sigma^2_{z,k}$. The channel gains $h_{i,k}(i = 0, 1, \ldots, M; k =$

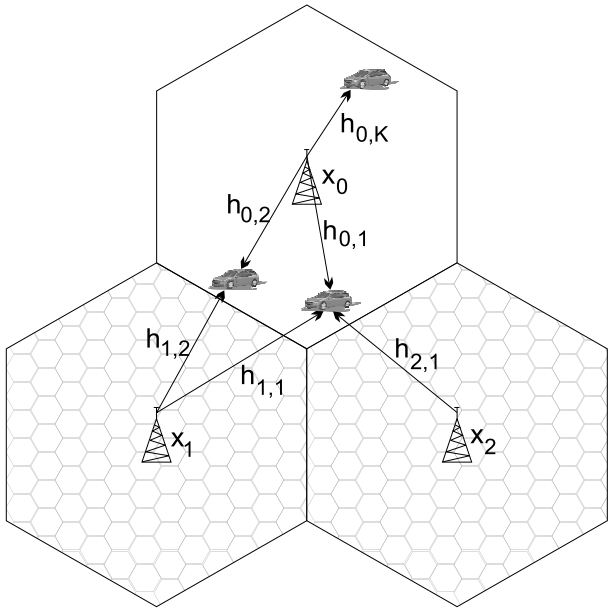

Figure 10: Diagram of a system with full frequency reuse. Hatched cells interfere with the transmissions of the cell of interest.

$1, \ldots, K)$ are generated as in Section 3.1:

$$h_{i,k}[m+1] = \sqrt{1-\epsilon_k} \cdot h_{i,k}[m] + \sqrt{\epsilon_k} v_{i,k}[m]$$

[4]Refer to Section 3.1 for definitions of T, ϵ, m, etc.

where $0 \leq \epsilon_k \leq 1$ and $v_{i,k} \sim \mathcal{CN}(0, \sigma^2_{h_{i,k}})$ are independent white processes. The information bearing signals $x_i[n]$ have an average power constraint 1.

5.2 Single interferer and negligible additive noise

We assume that there is a single interferer, i.e. $M = 1$, and consider the case where the interference dominates the noise term, neglecting σ_z^2. This models the situation of a user in the cell boundary, in a system operating with large transmission power.

5.2.1 Perfect knowledge upper bound

We can obtain an upper bound on the system capacity by assuming that the transmitter has perfect knowledge of the channel gains $h_i, i = 0, 1$. This assumption corresponds to letting $v \to 0$.

The SIR of user k at the time of transmission of the nth frame is

$$\mathsf{SIR}_k[n] = \frac{|h_{0,k}[n]|^2}{|h_{1,k}[n]|^2}$$

Since $h_{i,k}[n] \sim \mathcal{CN}(0, \sigma^2_{h_{i,k}}), (i = 0, 1)$ and hence $|h_{i,k}[n]|^2 \sim \text{Exp}(\text{mean} = 1/\sigma^2_{h_{i,k}}), (i = 0, 1)$ and are statistically independent, we can make a change of variables to obtain the pdf of SIR_k:

$$f_{\mathsf{SIR}_k}(x) = \frac{(\sigma^2_{h_{0,k}} \sigma^2_{h_{1,k}})^{-1}}{(x/\sigma^2_{h_{0,k}} + 1/\sigma^2_{h_{1,k}})^2}$$

valid for $x \geq 0$. If all the users have similar statistics, the scheduler decides to transmit to user $\bar{k} = \arg\max_{1 \leq k \leq K} \mathsf{SIR}_k$. Let $\mathsf{SIR}_{max} = \max_{1 \leq k \leq K} \mathsf{SIR}_k$. Then the capacity of the cell/sector with K users is given by

$$C(K) = E\left[\log_2(1 + \mathsf{SIR}_{max})\right]$$

To analyze the asymptotic behavior of $C(K)$ for large K we use the following lemma. (p.206 of [10])

Lemma 1 *Let $z_1, \ldots, z_K$ be i.i.d. random variables with a common cdf $F(\cdot)$ satisfying*

$$\lim_{z \to \infty} \left[\frac{1 - F(z)}{1 - F(r \cdot z)}\right] = r^\alpha$$

for every $r > 0$ and some $\alpha > 0$. Then $\max_{1 \leq k \leq K} \{z_k / a_K\}$ converges in distribution to a limiting random variable with cdf: $\exp(-x^{-\alpha})$ *, $x > 0$.*

In our case $(\mathsf{SIR}_k)_{k=1}^K$ satisfies the hypotheses of the lemma with $\alpha = 1$, and $a_K = K \cdot \sigma^2_{h_1}/\sigma^2_{h_0}$. Then we can approximate the distribution of SIR_{max} by

$$F_{\mathsf{SIR}_{max}}(x) \approx \exp\left(-\frac{\sigma^2_{h_1} K}{\sigma^2_{h_0} x}\right)$$

$$f_{\mathsf{SIR}_{max}}(x) \approx \frac{\sigma_{h_1}^2 K}{\sigma_{h_0}^2 x^2} \exp\left(-\frac{\sigma_{h_1}^2 K}{\sigma_{h_0}^2 x}\right)$$

valid for $x > 0$. Using this result we can approximately calculate $C(K)$ for large K.

$$\begin{aligned} C(K) &\approx \int_0^\infty \log_2(1+x) \cdot \frac{\sigma_{h_1}^2 K}{\sigma_{h_0}^2 x^2} \cdot \exp\left(-\frac{\sigma_{h_1}^2 K}{\sigma_{h_0}^2 x}\right) dx \\ &= \log_2(e) \left[\exp\left(\frac{K\sigma_{h_1}^2}{\sigma_{h_0}^2}\right) E_1\left(\frac{K\sigma_{h_1}^2}{\sigma_{h_0}^2}\right) + \ln\left(\frac{K\sigma_{h_1}^2}{\sigma_{h_0}^2}\right) + \gamma\right] \end{aligned}$$

where $E_1(x) = \int_x^\infty \frac{e^{-t}}{t} dt$ is the exponential integral and γ is Euler's constant. From the known bounds for $E_1(x)$ [11], the following are tight for large values of x:

$$\frac{e^{-x}}{2} \ln\left(1+\frac{2}{x}\right) \leq E_1(x) \leq e^{-x} \ln\left(1+\frac{1}{x}\right) \tag{13}$$

valid for all $x > 0$. Thus, letting $\alpha \in \{1, 2\}$ we can lower and upper bound $C(K)$:

$$\begin{aligned} C(K) &\lessgtr \frac{1}{\alpha} \log_2\left(1+\frac{\alpha\sigma_{h_0}^2}{K\sigma_{h_1}^2}\right) + \log_2\left(\frac{K\sigma_{h_1}^2}{\sigma_{h_0}^2}\right) + \log_2(e)\gamma \\ &\sim \log_2(K) \end{aligned}$$

where $\lessgtr$ means that the expression on the right is either a lower or upper bound depending on the choice of α, and the approximation is made for $K \to \infty$. We see that in this case the capacity grows logarithmically with the number of users, provided the approximations made are valid.

5.2.2 Influence of prediction error

To analyze the effect of mobile speed in the prediction error, and its influence on the system throughput we relax the assumption of perfect knowledge of the channel gains $h_i, i = 0, 1$ at the time of transmission $n + N$. However, to make the argument independent of the measurement problem, we still assume that the channel gains are perfectly measured by the receiver at time n. This assumption is accurate for large SNR and high mobile speeds, when prediction error is much larger than measurement error.

The scheduler must make a prediction of the maximum achievable rate at time $n+N$ based on the measurements $h_0[n]$ and $h_1[n]$, and then select a user for transmission. If all users have similar statistics and $t_c \to \infty$ in the scheduling algorithm, the scheduler chooses the user for which the predicted transmission rate is maximum. The best predictor $\hat{R}_k$ in mean square error sense is given by the conditional expectation:

$$\hat{R}_k[n+N] = E\left[\log_2(1+\mathsf{SIR}_k[n+N]) | h_0[n], h_1[n]\right]$$

where we assumed the use of channel codes with performance close to the capacity of the AWGN channel. Then the scheduler chooses the user $\bar{k}$ for transmission at time $n + N$ according to:

$$\bar{k} = \arg\max_k \hat{R}_k[n+N]$$

Since there is uncertainty in the estimate $\hat{R}_{\bar{k}}$, for any given transmission rate there exists the probability of an outage, the event that the channel cannot support the selected rate. As was done for the case of noise limited environments in Section 3.3 we fix the value of the outage probability γ and choose a transmission rate $R(h_0[n], h_1[n], \gamma)$ such that:

$$\gamma = E\left\{P\left[R(h_0[n], h_1[n], \gamma) > \log_2(1 + \mathsf{SIR}_{\bar{k}}[n+N])|h_0[n], h_1[n]\right]\right\}$$

Finally the total sector throughput is given by:

$$R = (1-\gamma)E\left[R(h_0[n], h_1[n], \gamma)\right]$$

where the expectation is taken over $h_0[n]$ and $h_1[n]$.

Since it is not possible to compute R in closed form, and even an approximate analysis becomes complicated we performed computer simulations to characterize R as a function of the number of users under different mobility conditions.

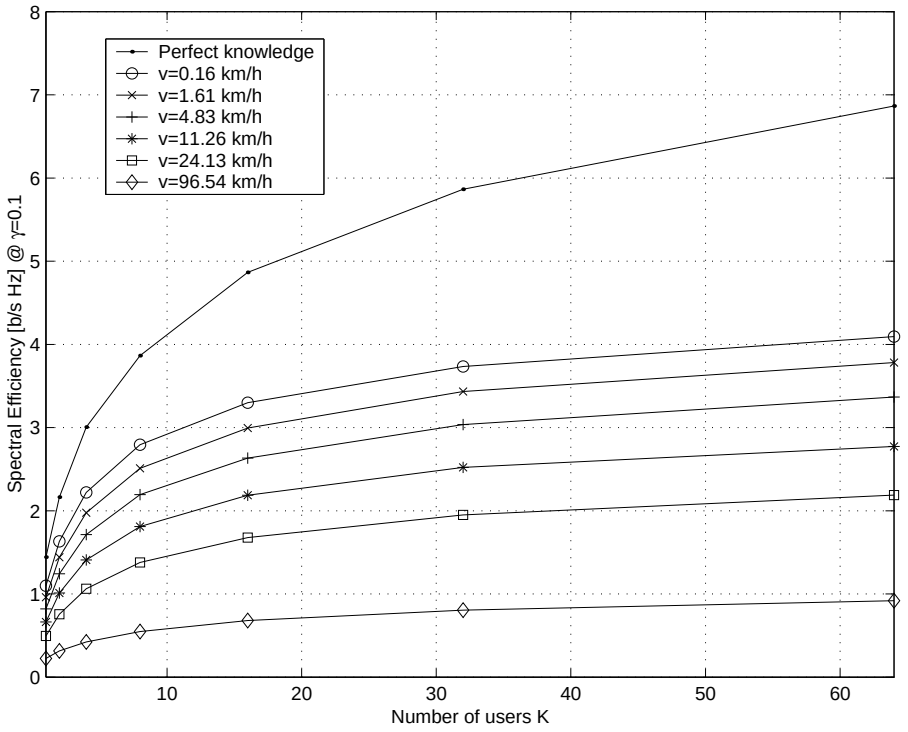

Figure 11: Spectral efficiency as a function of the number of users for $\gamma = 0.1$, SIR=0dB, $T = 0.167$ms, $N = 2$.

In Figure 11 we plot the sector's spectral efficiency as a function of the number of users K for different mobile speeds for a frame length $T = 0.167$ms. In all cases we fixed $\gamma = 0.1$. As a reference we included the curve corresponding to the case of perfect prediction, which corresponds to the analysis performed in the previous subsection.

As can be seen from the figure the perfect knowledge upper bound is too optimistic. Even for a mobile speed of 0.16 km/h the perfect knowledge bound overestimates the throughput by about 50%. This shows that the achievable throughput is very sensitive to channel uncertainty. This can be explained by noting that with high probability the scheduled user will be one for which

the channel gain $|h_0[n]|$ is large and the interference gain $|h_1[n]|$ is small. When $|h_0[n]|$ is large $\mathsf{SIR}[n+N]$ does not depend very much on the variations in h_0. In contrast, when $|h_1[n]|$ is small $\mathsf{SIR}[n+N]$ is very sensitive to the variations in h_1, which are significant even at very low speeds. Therefore the SIR of the scheduled user will be varying very fast and becomes hard to predict even at moderate speeds. In Figure 12 we can see how mobile speed and channel uncertainty affect the total sector throughput for systems with different number of users.

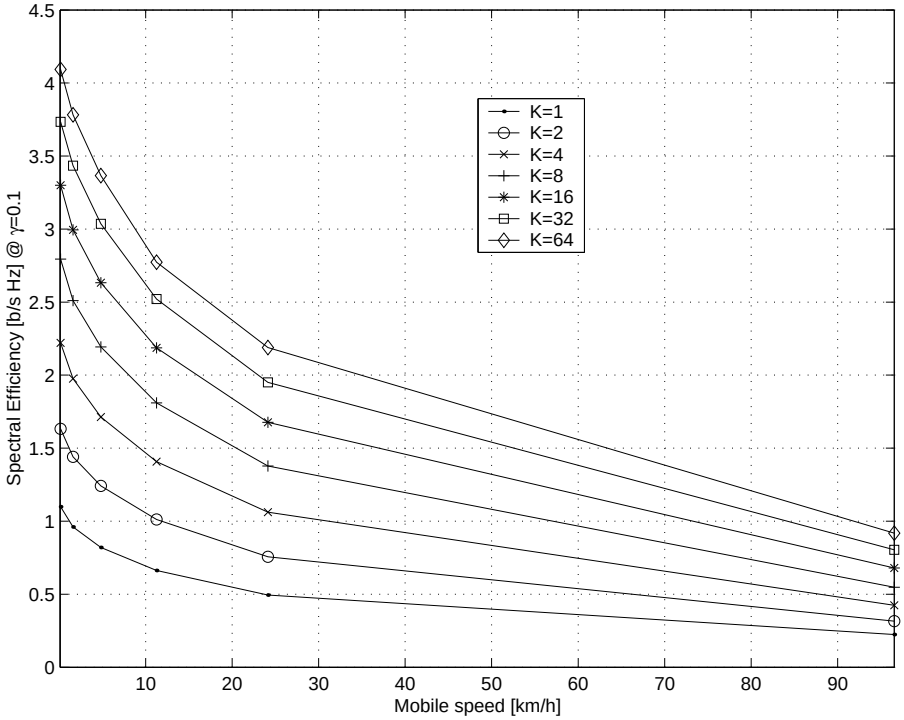

Figure 12: Spectral efficiency as a function of mobile speed for $\gamma = 0.1$, SIR=0dB, $T = 0.167$ms, $N = 2$.

5.3 Single interferer and non-negligible additive noise

We again consider the case of $M = 1$ but relax the assumption of negligible additive noise. In practice, as the number of users in the system increases, the scheduled user will be one for which the interference is small, in which case the additive noise may become a performance limiting factor.

5.3.1 Perfect knowledge upper bound

To get an upper bound on the system capacity we will initially assume that the transmitter has perfect knowledge of the channel gains $h_i, i = 0, 1$, which corresponds to letting $v \to 0$.

The SINR of user k at the time of transmission of the nth frame is

$$\mathsf{SINR}_k[n] = \frac{|h_{0,k}[n]|^2}{|h_{1,k}[n]|^2 + \sigma_{z,k}^2}$$

Since $h_{i,k}[n] \sim CN(0, \sigma_{h_{i,k}}^2), (i = 0, 1)$ it follows after a change of variables that $|h_{i,k}[n]|^2 \sim$

Exp(mean $= 1/\sigma^2_{h_{i,k}}$), $(i = 0, 1)$. Using the statistical independence of $h_{0,k}$ and $h_{1,k}$ we can make a new change of variables to obtain the pdf of SINR_k:

$$f_{\mathsf{SINR}_k}(x) = \frac{\left(\sigma^2_{z,k}/\sigma^2_{h_{0,k}}\right) x + \sigma^2_{z,k}/\sigma^2_{h_{1,k}} + 1}{\sigma^2_{h_{0,k}}\sigma^2_{h_{1,k}}\left(x/\sigma^2_{h_{0,k}} + 1/\sigma^2_{h_{1,k}}\right)^2} \exp\left(-\frac{\sigma^2_{z,k}}{\sigma^2_{h_{0,k}}}x\right) \tag{14}$$

valid for $x \geq 0$. If all the users have similar statistics and $t_c \to \infty$, the scheduler decides to transmit to user $\bar{k} = \arg\max_{1\leq k\leq K} \mathsf{SINR}_k$. Let $\mathsf{SINR}_{max} = \max_{1\leq k\leq K} \mathsf{SINR}_k$. Then the capacity of the cell/sector with K users is given by

$$C(K) = E\left[\log_2(1 + \mathsf{SINR}_{max})\right]$$

We will analyze the asymptotic behavior of $C(K)$ as $K \to \infty$ using the following result for the limiting distribution of the maximum of K i.i.d. random variables with exponential tails. (p.207 of [10])

Lemma 2 *Let $z_1, \ldots, z_K$ be i.i.d. random variables with a common cdf $F(\cdot)$ and pdf $f(\cdot)$ satisfying $F(z)$ is less than 1 for all z and is twice differentiable for all z, and is such that*

$$\lim_{z\to\infty}\left[\frac{1 - F(z)}{f(z)}\right] = c > 0$$

for some constant c. Then $\max_{1\leq k\leq K}\{z_k - l_K\}$ converges in distribution to a limiting random variable with cdf

$$\exp(-e^{-x/c}) \tag{15}$$

In the above, l_K is given by $F(l_K) = 1 - 1/K$.

This result states that the maximum of K such i.i.d. random variables grows like l_K.

Since we assumed that the K users have similar statistics $(\mathsf{SINR}_k)_{k=1}^K$ are i.i.d. random variables with pdf given by (14), whose tail decays exponentially and satisfies the hypotheses of Lemma 2 with $c = \sigma^2_{h_0}/\sigma^2_z > 0$. To find l_K it is necessary to compute the cdf of SINR_k. However, we run into the difficulty that there is no closed form expression for this cdf, but we can get an approximation for large values of its argument.

$$\begin{aligned} 1 - F_{\mathsf{SINR}_k}(x) &= P(\mathsf{SINR}_k > x) = \int_x^\infty f_{\mathsf{SINR}_k}(t)dt \\ &\sim \int_x^\infty \frac{\sigma^2_z}{\sigma^2_{h_1}}\frac{1}{t}\exp\left(-\frac{\sigma^2_z}{\sigma^2_{h_0}}t\right)dt = \frac{\sigma^2_z}{\sigma^2_{h_1}}E_1(x\sigma^2_z/\sigma^2_{h_0}) \end{aligned} \tag{16}$$

where $\sim$ means that the ratio of both sides converges to 1 as $x \to \infty$, and $E_1(\cdot)$ is the exponential integral. From (13) we have that for large x, $E_1(x) \sim e^{-x}/x$ and we can approximate (16) by

$$1 - F_{\mathsf{SINR}_k}(x) \sim \frac{\sigma^2_{h_0}}{\sigma^2_{h_1}}\frac{\exp(-x\sigma^2_z/\sigma^2_{h_0})}{x}$$

Therefore for large values of K we can write

$$l_K = \frac{\sigma_{h_0}^2}{\sigma_z^2} \ln(K) + O[\ln(\ln(K)]$$

The mean of a random variable with cdf given by (15) is $c\gamma$, where $\gamma = 0.577216\ldots$ is Euler's constant. Therefore, $E(\mathsf{SINR}_{max})$ grows as l_K and we can use Jensen's inequality and the concavity of the $\log_2(\cdot)$ function to upper bound $C(K)$ for large K,

$$C(K) \leq \log_2 [1 + E(\mathsf{SINR}_{max})] \sim \log_2 [\ln(K)]$$

We conclude that even with perfect channel state information, capacity grows at most doubly logarithmically with the number of users when there is background noise and one interferer. Also note that for large K, $C(K)$ does not depend on the interference variance $\sigma_{h_1}^2$, and in particular this result also holds for $\sigma_{h_1}^2 \to 0$, i.e. noise limited environment. This is because when K is large with high probability the scheduled user is one for which the interference is small, in which case the noise dominates the interference plus noise term.

5.3.2 Influence of prediction error

As was done in Section 5.2.2 we consider the influence of prediction error on the achievable throughput, assuming that the channel gains are perfectly measured. The SINR of user k for $N = 1$ is given by

$$\mathsf{SINR}_k[n+1] = \frac{\left|\sqrt{1-\epsilon} \cdot h_{0,k}[n] + \sqrt{\epsilon} v_{0,k}[n]\right|^2}{\left|\sqrt{1-\epsilon} \cdot h_{1,k}[n] + \sqrt{\epsilon} v_{1,k}[n]\right|^2 + \sigma_{z,k}^2}$$

The scheduler must make a prediction of the maximum supported rate based on the measurements, and then select a user for transmission at the next time slot. If all users have similar statistics and $t_c \to \infty$ in the scheduling algorithm, the scheduler chooses the user for which the predicted supported rate is maximum. The best predictor in mean square error sense is given by the conditional expectation:

$$\begin{aligned}
\hat{R}_k[n+1] &= E\left\{\log_2\left(1 + \mathsf{SINR}_k[n+1]\right) \middle| h_{0,k}[n], h_{1,k}[n]\right\} \\
&\approx E\left\{\log_2\left(\left|\sqrt{1-\epsilon} \cdot h_{0,k}[n] + \sqrt{\epsilon} \cdot v_{0,k}[n]\right|^2\right) \middle| h_{0,k}[n]\right\} - \\
&\quad E\left\{\log_2\left(\left|\sqrt{1-\epsilon} \cdot h_{1,k}[n] + \sqrt{\epsilon} \cdot v_{1,k}[n]\right|^2 + \sigma_z^2\right) \middle| h_{1,k}[n]\right\} \qquad (17)
\end{aligned}$$

Since there is no closed form expression for these expectations, we make some approximations for the 'large K-slow speed' regime. As the number of users in the system K grows, the scheduled user will be one with a strong channel and weak interference. To make this statement more precise, we will assume that the number of users in the system is large enough so that with

high probability there is a user with $|\sqrt{1-\epsilon}\cdot h_{0,k}|^2 > \epsilon\cdot\sigma_{h_0}^2$ and $|\sqrt{1-\epsilon}\cdot h_{1,k}|^2 < \epsilon\cdot\sigma_{h_1}^2$ so that we can approximate (17) by

$$\begin{aligned}\hat{R}_k[n+1] &\approx E\left\{\log_2\left(|\sqrt{1-\epsilon}\cdot h_{0,k}[n]|^2\right)\Big| h_{0,k}[n]\right\} - E\left\{\log_2\left(\epsilon\cdot|v_{1,k}[n]|^2+\sigma_z^2\right)\Big| h_{1,k}[n]\right\}\\ &= \log_2(1-\epsilon)+\log_2\left(|h_{0,k}[n]|^2\right) - E\left\{\log_2\left(\epsilon\cdot|v_{1,k}[n]|^2+\sigma_z^2\right)\right\} \qquad (18)\end{aligned}$$

The last expectation can be computed as follows:

$$\begin{aligned}E\left\{\log_2\left(\epsilon\cdot|v_{1,k}[n]|^2+\sigma_z^2\right)\right\} &= \int_0^\infty \log_2\left(\epsilon x+\sigma_z^2\right)\cdot e^{-x/\sigma_{h_1}^2}dx\\ &= \sigma_{h_1}^2\log_2(\sigma_z^2)+\sigma_{h_1}^2\log_2(e)e^{\sigma_z^2/(\epsilon\sigma_{h_1}^2)}E_1\left(\frac{\sigma_z^2}{\epsilon\cdot\sigma_{h_1}^2}\right)\\ &\gtrless \sigma_{h_1}^2\log_2(\sigma_z^2)+\frac{\sigma_{h_1}^2\log_2(e)}{\alpha}\ln\left[1+\alpha\frac{\epsilon\cdot\sigma_{h_1}^2}{\sigma_z^2}\right] \qquad (19)\end{aligned}$$

where $E_1(\cdot)$ is the exponential integral and for $\alpha = 1, 2$ we get an upper and lower bound respectively.

Under the present assumptions (18) is independent of $|h_{1,k}[n]|$ and is maximized when $|h_{0,k}[n]|$ is maximum. Therefore the scheduler chooses the user $\bar{k} = \underset{1\le k\le K}{\arg\max}\,|h_{0,k}[n]|$ and the corresponding expected achievable rate is

$$\begin{aligned}R &\approx \log_2(1-\epsilon)+E\left[\log_2\left(|h_{0,\bar{k}}[n]|^2\right)\right]-\sigma_{h_1}^2\log_2(\sigma_z^2)-\frac{\sigma_{h_1}^2\log_2(e)}{\alpha}\ln\left[1+\alpha\frac{\epsilon\cdot\sigma_{h_1}^2}{\sigma_z^2}\right]\\ &\le \log_2(1-\epsilon)+\log_2\left[E\left(|h_{0,\bar{k}}[n]|^2\right)\right]-\sigma_{h_1}^2\log_2(\sigma_z^2)-\frac{\sigma_{h_1}^2\log_2(e)}{2}\ln\left[1+2\frac{\epsilon\cdot\sigma_{h_1}^2}{\sigma_z^2}\right]\end{aligned}$$

where the inequality follows from Jensen's inequality and taking $\alpha = 2$ in (19). Here we assumed a transmission rate equal to the expected achievable rate and neglected the possibility of an outage, obtaining an upper bound on the achievable rate.

From Lemma 2 it follows that $E\left[|h_{0,\bar{k}}[n]|^2\right]$ grows like $\sigma_{h_0}^2\log K$ and hence R grows like $\log_2[\ln(K)]$ for large K as in the perfect knowledge case. The effect of prediction error appears as a constant term in the upper bound for the achievable rate.

5.4 Comparison between interference and noise limited systems

In the previous sections we analyzed the performance of interference and noise limited systems. While interference limited systems have a large capacity under the assumption of perfect knowledge of the channel gains, their capacity is very sensitive to channel uncertainty. In contrast, interference limited systems have a much smaller perfect knowledge capacity, but capacity is less sensitive to channel uncertainty. From the system design perspective it is interesting to determine which is the most desirable regime of operation.

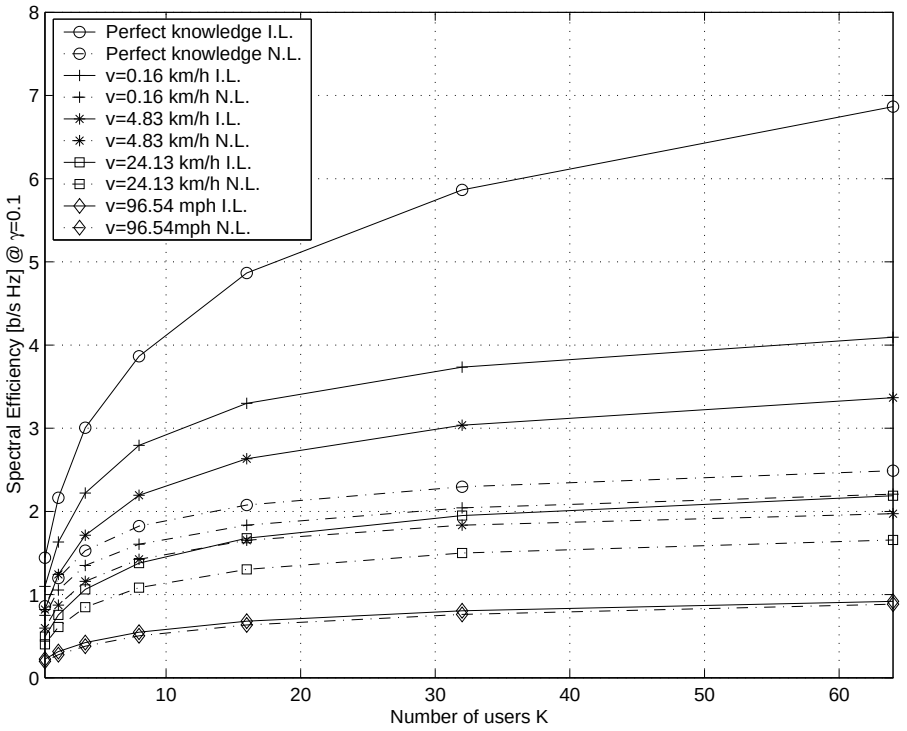

Figure 13: Spectral efficiency as a function of mobile speed for Noise Limited (N.L.) and Interference Limited (I.L) environments, for $\gamma = 0.1$, SNR/SIR=0dB, $T = 0.167$ms, $N = 2$.

Figure 13 shows simulation results for the spectral efficiency of interference and noise limited systems for various mobile speeds. To make the comparison fair the SIR in the interference limited case was set equal to the SNR in the noise limited case. In all cases we assumed perfect measurement of the channel gains, considering channel state uncertainty through prediction error. We see that for low mobile speeds interference limited systems have a considerable performance advantage (twice the throughput for $v = 0.16$km/h). However, as mobile speed increases the performance gap becomes smaller, being insignificant at $v = 96.54$ km/h. For large enough mobile speeds it is impossible to track the interference, so in practice it can be treated as if it was white noise.

We conclude that interference limited systems have a considerable performance advantage over noise limited systems, as long as the interference can be accurately predicted.

5.5 Multiple interfering cells

If $M > 1$ there are multiple interferers that contribute to the noise term. As M increases there is an interference averaging effect that reduces the amplitude of SINR fluctuations. Smaller fluctuations result in easier SINR prediction, but reduced multiuser diversity. Therefore, users located in the edge of a cell receiving interference from 2 or more adjacent cells cannot benefit significantly from interference nulling. If all users in the system receive interference from multiple cells, we expect to see a $\log \log K$ dependence of cell capacity as a function of the number of users in the cell, as was the case for noise limited environments.

6 Opportunistic Communication in the Uplink

In the previous sections we focused on how opportunistic communication can be exploited to improve the performance of the downlink channel of a multiuser system. There are two important features of the downlink channel that allow to exploit multiuser diversity:

- Centralized scheduling: scheduling decisions are made at the base station. The scheduler can use all the measurements fed back by the users simultaneously in a centralized algorithm.
- Efficient channel measurement: the transmission of a single pilot signal can be used simultaneously by all users to estimate the channel. As the number of users increases the channel estimation overhead remains fixed.

In the uplink of an FDD[5] system, channel measurement requires the transmission of pilot signals from each mobile to the base station. The channel measurement overhead grows linearly with the number of users, so it may become an important limiting factor for opportunistic communication.

Scheduling for the uplink can still be done in a centralized way at the base station. When the base station measures the channel gain for each user, it can also make a prediction of the supported rate at the next time slot, and can implement the proportional fair algorithm (or any other centralized algorithm) to select one user. Then the base station only needs to identify the selected user and the supported transmission rate. The corresponding overhead only grows logarithmically with the number of users. Compare this with the overhead of DRC transmission for the downlink channel, which grows linearly with the number of users unless we use a selective feedback strategy.

We see that opportunistic communication for the uplink channel requires significantly more resources for channel measurement, but fewer resources for measurement feedback. Moreover, we can reduce the resources used for channel measurement if the system uses time division duplexing. In this case the transmission of a single pilot from the base station can be used to estimate all the uplink and downlink channels simultaneously at the mobiles. However, to have these measurements available at the base station requires feedback, which grows linearly with the number of users.

We conclude that opportunistic communication can also be exploited in the uplink channel, although there are different challenges to face.

7 Conclusions

Fading provides a source of randomization that can be exploited to obtain significant capacity improvements through opportunistic communication. This capacity boost is particularly important in multiple user systems, and the corresponding performance improvement is called multiuser diversity gain. This gain results from transmitting to the user only when his channel is strong.

A practical system must satisfy some quality of service requirements. The proportional fair scheduling algorithm that we proposed tries to obtain large total throughputs while distributing the system resources with fairness among users. This algorithm schedules users when their fading

[5] FDD stands for frequency division duplexing.

states are good as compared to their own averages, hitting the peaks of the SNR processes. Since the channel states are normalized by the average throughputs, system resources are distributed with fairness even when the users have different fading statistics. A generalized version of this scheduling algorithm was presented to satisfy different grade of service requirements.

Multiuser diversity gains are strongly dependent on the accurate measurement and prediction of the users' fading states. As mobile speed increases channel tracking becomes increasingly challenging, and special considerations are required in the system design to allow for accurate channel state estimation. The reduction of frame length and the increase in the rate of channel measurement can reduce channel prediction uncertainty, resulting in improved performance. However, frame length cannot be reduced arbitrarily due to an increase in measurement error and overheads. In practice, for reasonable frame lengths, we can use some strategies to keep these overheads under control.

Multi-cell systems can be interference or noise limited, depending on various factors that can be controlled by the designer. Interference fluctuations produce variations in the SINR experienced by each user, and provide another source of randomization. If these fluctuations can be accurately tracked, interference nulling can provide a significant capacity boost. Accurately predicting the interference, in particular when it is close to a null, is quite difficult and large prediction errors limit performance even at moderate speeds. However, when the prediction error can be controlled, capacity of interference limited systems is much larger than the one corresponding to noise limited systems. Therefore, from the multiuser diversity perspective, it is desirable to design a system to operate in an interference limited regime.

Most of the ideas presented in this work apply to both uplink and downlink channels, although each type of channel includes some challenges that must be addressed independently.

Appendix A

In this appendix we will derive the approximation of the function $\beta(\hat{h}, \gamma, \sigma^2)$ using first and second order polynomials.

If $\hat{h} = 0$ we can evaluate (7) explicitly:

$$\gamma = \int\limits_{x^2+y^2<\sigma_z^2\beta^2} \int \frac{1}{\pi\sigma^2} \exp\left(-\frac{x^2+y^2}{\sigma^2}\right) dxdy = \int\limits_0^{\sigma_z\beta} \frac{2r}{\sigma^2} \exp\left(-\frac{r^2}{\sigma^2}\right) dr = 1 - \exp\left(-\frac{\sigma_z^2\beta^2}{\sigma^2}\right)$$

and solve for β

$$\beta(\hat{h} = 0, \gamma, \sigma^2) = \frac{\sigma}{\sigma_z}\sqrt{\ln\left(\frac{1}{1-\gamma}\right)} \tag{20}$$

If on the other hand $|\hat{h}|^2 \gg \sigma^2$ we can approximate (7) by

$$\gamma \cong \int\limits_{-\infty}^{\sigma_z\beta} \int\limits_{-\infty}^{\infty} \frac{1}{\pi\sigma^2} \exp\left(-\frac{(x-|\hat{h})^2+y^2}{\sigma^2}\right) dydx$$

$$= \int_{-\infty}^{\sigma_z\beta} \frac{1}{\sqrt{\pi}\sigma} \exp\left(-\frac{(x-|\hat{h}|)^2}{\sigma^2}\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}\sigma} \exp\left(-\frac{y^2}{\sigma^2}\right) dydx$$

$$= \int_{-\infty}^{\sigma_z\beta} \frac{1}{\sqrt{\pi}\sigma} \exp\left(-\frac{(x-|\hat{h}|)^2}{\sigma^2}\right) dx = \Phi\left(\frac{\sigma_z\beta - |\hat{h}|}{\sigma/\sqrt{2}}\right) \quad (21)$$

where $\Phi(\cdot)$ is the distribution function of a $\mathcal{N}(0,1)$ random variable. Solving for β we obtain

$$\beta(\hat{h},\gamma,\sigma^2) = \frac{1}{\sigma_z}|\hat{h}| + \frac{\sigma}{\sqrt{2}\sigma_z}\Phi^{-1}(\gamma) \quad (22)$$

From (20) we obtain $b = \frac{\sigma}{\sigma_z}\sqrt{\ln\left(\frac{1}{1-\gamma}\right)}$ and from (22) we get $c = \frac{1}{\sigma_z}$ and $d = \frac{\sigma}{\sqrt{2}\sigma_z}\Phi^{-1}(\gamma)$. Finally if we require $\beta(\hat{h},\gamma,\sigma^2)$ to be differentiable (with respect to $|\hat{h}|$) we obtain $a = \frac{c^2}{4(b-d)}$ and $e = \frac{2(b-d)}{c}$. Replacing these constants in (8) we obtain (9).

References

[1] A. Goldmisth and P. Varaiya, *Capacity of fading channel with channel side information*, IEEE Trans. on Info Theory, Vol. 43, No. 6, pp. 1986-1992, November, 1997.

[2] R. Knopp and P. Humblet, *Information capacity and power control in single-cell multiuser communications*, Int. Conf. on Communications, Seattle, Washington, June, 1995.

[3] D. Tse, *Optimal power allocation over parallel Gaussian channels*, Proc. of ISIT, 1997.

[4] D. Tse, et al., *Transmitter directed, multiple receiver system using path diversity to equitably maximize throughput*, patent filed, May 24, 1999.

[5] D. Tse, *Downlink multiuser diversity via proportional fair scheduling*, in preparation.

[6] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushayana, A. Viterbi, *CDMA/HDR: A Bandwidth-Efficient High-Speed Wireless Data Service for Nomadic Users*, IEEE Communications Magazine, July 2000.

[7] A. Jalali, R. Padovani, R. Pankaj, *Data throughput of CDMA/HDR*, Vehicular Technology Conference, 2000.

[8] R. Steele, *Mobile Radio Communications*, IEEE Press, 1994.

[9] T. Rappaport, *Wireless Communications: Principles and Practice*, Prentice Hall, 2nd Edition, 2001.

[10] H. A. David, *Order Statistics*, Wiley, 1970 (1st Edition).

[11] H. Alzer, *On Some Inequalities for the Incomplete Gamma Function*, AMS Mathematics of Computation, Vol. 66, No. 218, pp. 771-778, April 1997.