A Low Feedback Robust Opportunistic Scheduler and Beamformer for MISO Wireless Systems

by

Ali Honarvar

A thesis submitted to the
Department of Electrical and Computer Engineering
in conformity with the requirements for
the degree of Master of Science (Engineering)

Queen's University
Kingston, Ontario, Canada
June 2007

Abstract

Opportunistic Beamforming is a promising scheme with potential usage in a variety of applications ranging from cellular audio/video communications to wireless mesh-networks. However, some aspects in this scheme are open to further improvements, such as effective quality of service (QoS) provisioning, and efficient handling of sparse/realistic networks.

We focus on the downlink channel of a multi-rate time-division multiplexing MISO (multiple input single output) point-to-multi-point wireless communication system, and design a cross-layer scheme. We employ a two-stage opportunistic scheduler. The scheduler receives channel state information in the form of SNR (signal to noise ratio) measurements from each terminal, and schedules one terminal in each time-slot. At the first stage in a time-slot a terminal is chosen, and our gradient descent channel estimator finds an estimation of its channel vector. Then the beamformer produces a beamforming vector in an adjacent pilot channel, or in the first part of the time-slot in the data channel. Our beamforming is based on the estimation of the channel, as opposed to the random beamformer used in the opportunistic beamforming. Each terminal measures its perceived SNR even when the channel is assigned to another terminal, and sends this information back to the base-station subject to a feedback-reduction policy. The scheduler uses the new channel state information for its second stage and assigns a terminal for the succeeding time-slot. The channel estimation procedure uses the reported SNRs for each terminal, which is a significantly lower feedback rate than the amount required for coherent beamforming. Our approach efficiently handles sparse/realistic networks, and also provides a better QoS than the opportunistic beamforming scheme.

Acknowledgements

I would like to express my deep gratitude to my supervisor, Prof Saeed Gazor, for his continued guidance that I received during the course of the degree. In addition to providing a great opportunity, he helped me develop the research skills that will always be invaluable to me as a researcher.

I am also thankful to my colleagues for sharing their knowledge and inspiring me. Especially I would like to thank Alireza Heidar-Barghi, Reza Rashidifar, Renita Sudriga, Hamid Saligheh-rad, Ladan Golipoor, Najah Abu-Ali, Sepideh Zarrin and Saeed Moradi, for their true friendship and support.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Future wireless communication systems should support various service types efficiently. In many applications there is a larger throughput demand in the downlink direction than the uplink. In third generation systems, this requirement has been addressed by including a high-speed downlink channel mode; High Data Rate (HDR) mode in CDMA2000 is an example. As the developments in recent years show, multiple-antenna systems are promising choices for next generation transmission schemes. It is well known that employing multiple antenna systems can increase the channel capacity. Many schemes have been proposed to exploit spatial diversity in open loop (e.g. space-time coding) and closed loop (e.g. coherent beamforming) configurations.

Multiuser diversity is another type of diversity which is available in point-to-multi-point communication systems. This form of diversity is an additional way to improve throughput. Exploiting multiuser diversity requires appropriate scheduling algorithms. These algorithms increase the throughput by deliberately modifying the order of serving packets. Serving users when they experience a better channel strength has the advantage that they can be served at higher rates and more data can be transmitted. The immediate downside is

that maintaining fairness and the required delay performance becomes more challenging in such a scheme than in a channel-independent scheduling scheme. Employing channel state information in the scheduling policy requires a cross-layer design. Cross-layer designs tend to improve the efficiency of the system, but additional complexity is often inevitable. That is because abstractions provide separation between layers and allow one to conveniently design a module with minimal considerations of the rest of the system. A system combining multiuser diversity and spatial diversity does away with the separation between physical and MAC layers and higher layers [1]. Such a scheme often requires a joint design of the scheduling algorithm and beamforming/channel-precoding modules. Nevertheless, the scarcity of the air bandwidth and the growing demand for better services justify the difficulty of a cross-layer paradigm.

As the boundary between computer networks and cellular telephony fades out, it is of practical importance for a successful scheme to serve a wide range of configurations efficiently. Computer networks often have fixed or very slow moving devices with a significant elastic traffic (i.e. traffic that does not have hard delay requirements). On the other hand, cellular audio/video communication systems have stringent QoS and delay profile requirements and serve merely non-elastic traffic. Cellular systems also tend to have larger mobility and more channel variations. The breadth of possible applications and configurations raises a deliberate issue. A scheme that improves the service for mobile users but does not exploit the relatively static channel of the fixed devices cannot be considered as a scalable scheme. In this case one might adopt such a scheme for mobile services while using efficient schemes for fixed devices. Therefore, it is of interest to be efficient in both cases. Also, theoretically, it is more interesting to have a scheme that does not have an explicit mode of operation (mobile, fixed) but rather has a smooth transition between these

states. An effective method to improve the throughput especially for fixed devices is the usage of closed loop systems exploiting spatial diversity. The amount of feedback or the measurement equipment required in a closed-loop scheme makes it more appropriate for fixed terminals, although considerable gains are also possible in mobile scenarios. In order to support both cases efficiently, it has been suggested to use typical SNR measurement feedback and improve the throughput as much as possible. One of the promising schemes, namely the "Opportunistic Beamforming scheme" is not effective in sparse networks and does not have a mechanism to improve the throughput of a terminal when it is possible (e.g. in a fixed terminal). This is a crucial issue in major configurations such as wireless mesh networks.

The above open problems form the basis and motivation of this thesis. Our goal is to employ both multiuser diversity and spatial diversity in an efficient cross-layer design. To this end we develop an appropriate channel-aware beamforming and opportunistic scheduling scheme which is able to harness most of the possible gains in the fixed scenarios, while at the same time works well in fast fading environments. We also employ various scheduling algorithms and show the throughput and QoS benchmarks of the system under these disciplines, although none can be considered as optimal.

1.1 Contributions

The contributions in the present work are divided into the following parts.

- Cross-Layer Design: We have designed a cross-layer scheme to combine spatial and multi-user diversity efficiently. The *high level design* of the system and the overall procedure in each time-slot is one of our contributions.

- Channel Estimator: We use the typical SNR feedback from the terminals to estimate the channel for one of them. We employ a maximum likelihood estimator and formulate the optimization problem. To solve this problem we propose a gradient-based algorithm that stochastically converges to the channel signature of the terminal.

- Feedback Reduction Algorithm: This algorithm reduces the volume of measurement results being fed back to the base-station by allowing only significant results (compared to prior results multiplied by a forgetting rate) to be reported back to the base-station. This policy reduces the load on the feedback channel by imposing a lightweight pre-processing rule on the terminals. The performance loss is small even with as low as $\frac{1}{4}$ of the full SNR feedback.

- Beamformer: We use the estimated channel (obtained in the channel estimator module) to form a beamforming vector for succeeding transmissions. Since the beamforming vector has also the role of a measurement for the other terminals (other than the one being served), we need a regularization mechanism to avoid space dimension shrinkage. We employ a randomization mechanism which introduces measurements in new dimensions and maintains the "variety" of the measurement space. It is the same mechanism that reduces channel acquisition time for recently joined terminals.

- Scheduling Mechanism: The scheduler makes two decisions per time-slot. The first decision is proposed to the channel estimator. The second decision actually assigns the next time-slot to the selected terminal. We employ various scheduling disciplines and compare the performance benchmarks of the whole system under each scheduling policy.

1.2 Organization of Thesis

Chapter 2 provides a review of the literature, including the opportunistic beamforming scheme and relevant opportunistic scheduling schemes. The architecture and model of the proposed system is also presented in Chapter 2. There, we briefly explain system modules and the goals they should fulfill. In Chapter 3 we present our solution to the problem of channel estimation using SNR-feedback. There, we proceed to use a gradient-based maximum likelihood estimator and provide simulation results. Chapter 4 presents our beamforming scheme based on the channel estimator of Chapter 3. There, we justify the choice of a randomizer in the beamformer and provide supporting simulations. Chapter 5 presents the impact of various scheduling disciplines on system performance. We present various performance benchmarks for the whole system. The thesis concludes in Chapter 6 where a summary and future directions are presented. The derivation and details of the maximum likelihood channel estimation algorithm based on SNR feedback is presented in Appendix A.

1.3 Notation and Abbreviations

The notations used in the thesis are as follows: Scalars are represented by lower-case letters. Vectors are written in lower-case bold-face letters, e.g. $\mathbf{b}, \mathbf{h}$. Matrices are represented by upper-case letters, e.g. B, U, V. The j^{th} column of a matrix B is represented by $\mathbf{b}_j$. (Note the usage of bold-face lower-case letter). The i^{th} element of a vector $\mathbf{x}$ is represented by x_i. The element on the j^{th} column and i^{th} row of the matrix B is represented by $b_{i,j}$, it also shows the i^{th} element of the vector $\mathbf{b}_j$. Hermitian of a matrix B is shown as B^{H}. Expectation operator is $\text{E}[\cdot]$, norm of a complex vector is $\|\cdot\| = \sqrt{(\cdot)^{\text{H}}(\cdot)}$, and determinant

AMC	Adaptive Modulation and Coding
BER	Bit Error Rate
BPSK	Binary Phase-Shift Keying
ccdf	complementary cumulative distribution function
CDMA	Code Division Multiple Access
DoA	Direction of Arrival
GPS	Generalized Processor Sharing
HDR	High Data-Rate
IIR	Infinite Impulse Response
M-LWDF	Modified Least Weighted Delay First
MIMO	Multiple Input Multiple Output
MISO	Multiple Input Single Output
nSNR	Normalized Signal to Noise Ratio, § 2.3, (2.16)
OFDM	Orthogonal Frequency-Division Multiplexing
pdf	probability distribution function
PFS	Proportional Fair Sharing (Scheduling)
QoS	Quality of Service
RR	Round Robin
SINR	Signal to Interference plus Noise Ratio
SNR	Signal to Noise Ratio
TDM	Time Division Multiplexing
WFQ	Weighted Fair Queueing

Table 1.1: Table of abbreviations

operator is $|\cdot|$. The constant $\gamma = \lim_{n\to\infty}\left(\sum_{k=1}^{n}\frac{1}{k}\right) - \log n$ is the Euler-Mascheroni constant. We refer to mobile or static user devices as terminals and K often denotes the number of terminals or users, while k is often used as an index denoting terminal id $k = 1, 2, ...K$. Table 1.1 presents the list of abbreviations used throughout the thesis.

Chapter 2

Literature Review and System Model

In this thesis we focus on the downlink channel of a multi-rate multiuser TDM (time-division multiplexing) MISO cellular wireless communication system. We also study the scheduling policies in such a system. Scheduling policies that use the channel state feedback (opportunistic schedulers) are of particular interest. A review of relevant schemes and the related previous works follows. A block diagram of a typical system equipped with multiple antennas is shown in Figure 2.1. The scheduling algorithm in the shown system does not depend on the channel state, so it is a non-opportunistic scheduling policy. Thus, no multiuser diversity gain can be exploited by such a system.

2.1 Combining Multiuser Diversity and Spatial Diversity

Multiuser diversity is an important form of diversity which can be exploited in multiuser wireless systems, such as cellular communication systems. Since in a multiuser system, channels of different terminals often vary independently, it is possible to track their channel variations and select the terminal with the strongest channel. When the number of users

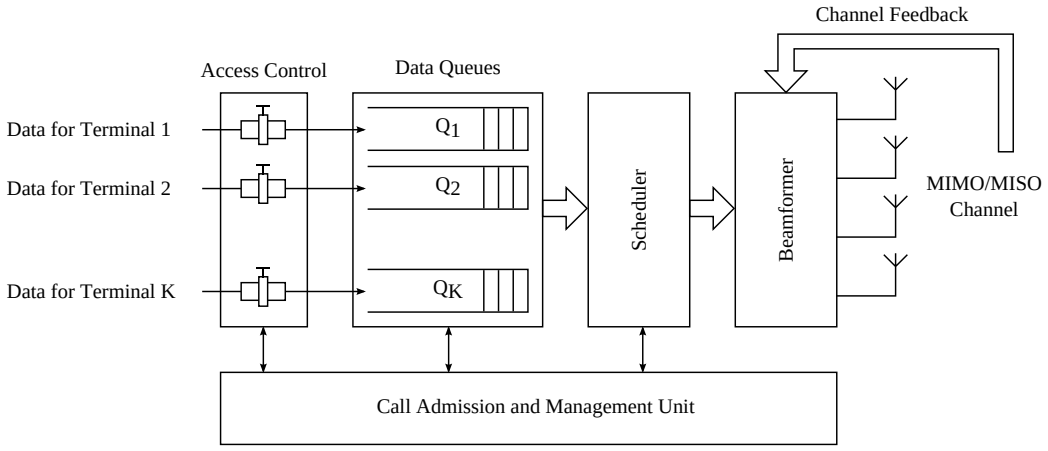


Figure 2.1: Block diagram of a system using a non-opportunistic scheduling policy

increases in such a system, the chance of having a user near its peak channel strength increases. Multiuser diversity is explored by Knopp and Humblet [2]. They have shown that the maximal uplink capacity can be obtained when the common channel is assigned to the user with the best instantaneous channel (typically the user that receives maximum SNR). They present the expression for the average bit error rate (BER) for uncoded BPSK (binary phase shift keying) for a given SNR under Rayleigh fading with unit variance. The average BER, $\bar{P}_e$, for a system with K terminals is given by

$$\bar{P}_e = \frac{1}{2}\sum_{k=1}^{K}(-1)^{k-1}\binom{K}{k}\left(1 - \frac{1}{\sqrt{a + k/SNR}}\right) \tag{2.1}$$

This is shown in Figure 2.2 and Figure 2.3.

However, this scheme can be unfair and its QoS is poor, because some terminals might experience higher channel strengths on average (e.g., users closer to the base-station might always have better channels), so they are more likely to win the competition over the common channel. Although this scheme is inspiring, it needs major modifications and additional mechanisms to provide fairness and QoS guarantees. Often it is important to draw performance comparisons with this scheme.

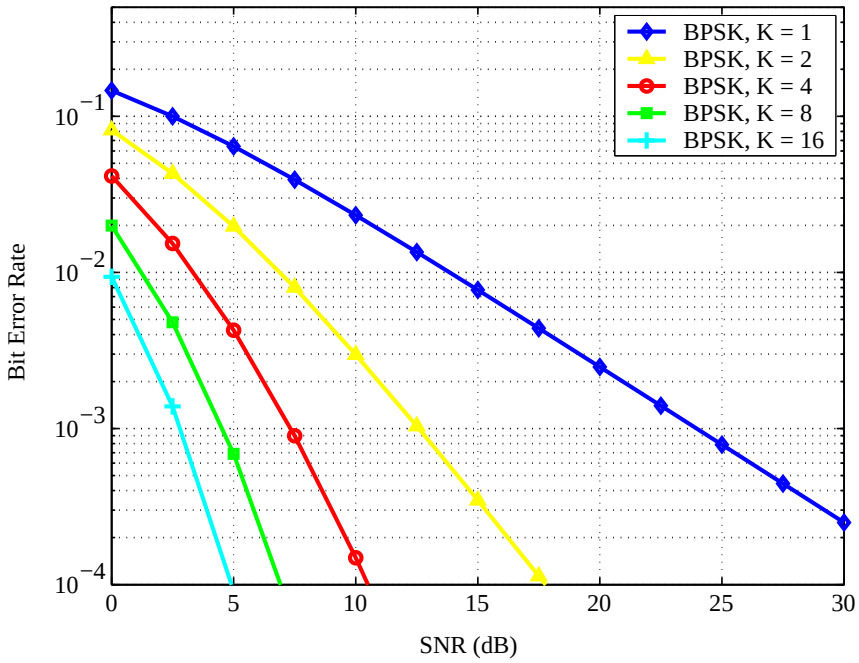


Figure 2.2: Multiuser diversity gain (for uncoded BPSK)

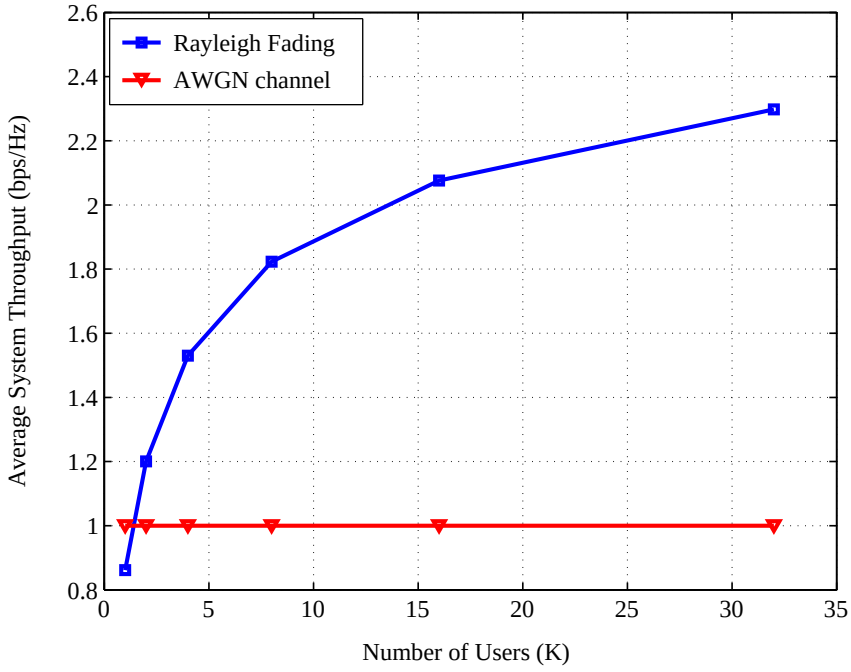


Figure 2.3: Throughput gain in a multiuser system (for AWGN, Rayleigh fading channels $\text{SNR}_{\text{avg}} = 0$) [3]

In order to address the fairness issue of the MaxRate policy, a novel scheduling policy was proposed by Tse [4] and also pursued by Jalali et al [5]. Tse [4] proposes the "proportionally fair" scheduling rule, which employs channel state information in the form of SNR feedback (so it is an opportunistic scheduler) and provides fair bandwidth allocation among terminals. Let R_k denote the supportable rate of kth terminal. As mentioned, the MaxRate scheme picks the terminal with the maximum supportable rate, i.e. $\arg\max_k\{R_k\}$ for the next time slot. The proportional fair scheduling policy, on the other hand, chooses the terminal that maximizes another measure, specifically R_k/T_k, where T_k represents the received throughput of terminal k in a predefined time window. If τ, the length of the time window, is infinite, T_k will represent the total throughput that the kth terminal has received so far. This scheme has a fairness property: asymptotically, the terminals will receive equal service, even if their average SNRs are not the same. The throughput of this scheme is inferior to that of the MaxRate scheme. This throughput loss is the price paid for (asymptotic) fairness. The reason is that the terminal maximizing R_k/T_K does not necessarily have the maximum R_k, so its supportable rate is less than the supportable rate of the terminal that the MaxRate policy would have chosen.

Although this scheme is fair in the limit, the quality of service can be poor when some terminals experience slow fading. Naturally, τ, the length of the time window within which T_k's are calculated, is related to the time window of the service. The relation between τ and the service time window is that choosing a smaller time window forces the scheduler to serve all terminals in a smaller time window. On the other hand, increasing τ moves us further away from the throughput optimality of the MaxRate scheme. Here, a trade-off between average throughput and delay bound is implied, and is controlled by the length of the time window.

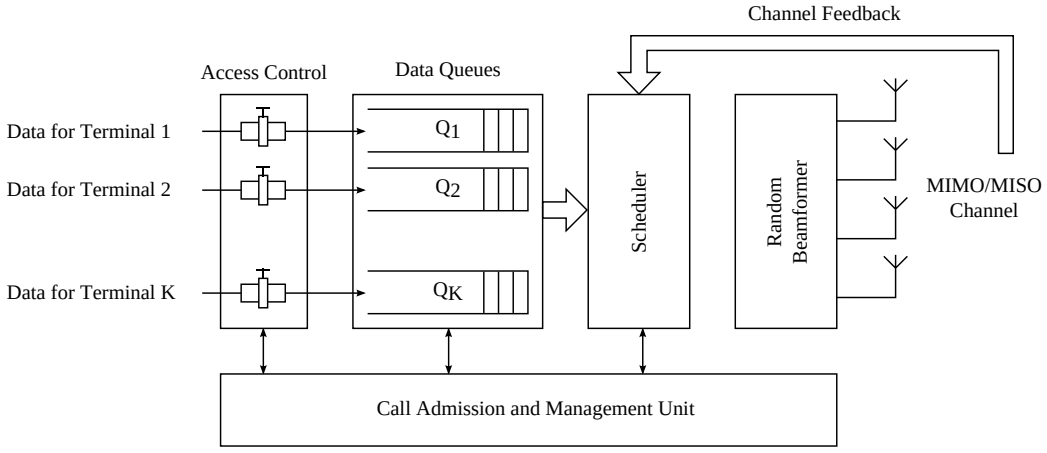


Figure 2.4: Block diagram of a system using opportunistic beamforming [6]

2.1.1 Opportunistic Beamforming

It is not sufficient to improve the fairness and quality of service by modifying *only* the scheduling policy, unless other parts of the system are also subject to modifications. Viswanath, Tse, and Laroia [6] employ multiple antennas and transmit beamforming to introduce additional channel variations, mimicking that of fast fading channels. The result is a significant improvement in the overall delay statistics. This scheme is called "Opportunistic Beamforming" and is a combination of spatial diversity and multiuser diversity. An overall block diagram of a system incorporating opportunistic beamforming is illustrated in Figure 2.4. The beamformer part is also shown in more detail in Figure 2.5.

In this scheme N_t transmit antennas are used in the base station and only one antenna on each terminal. This scheme works as follows:

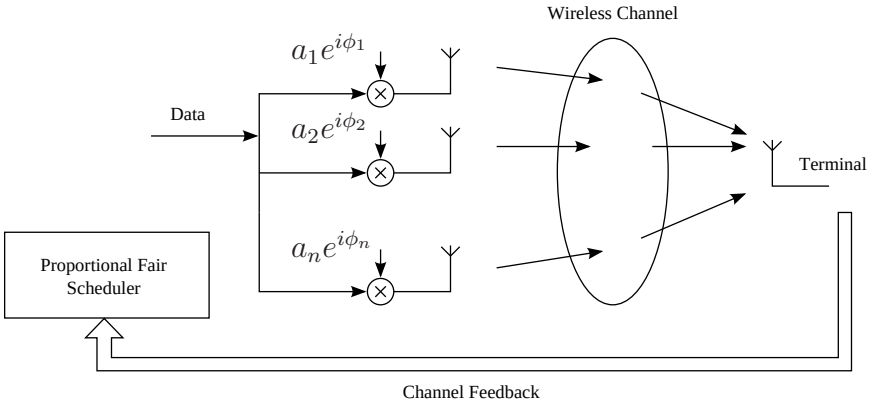


Figure 2.5: Random beamforming in the opportunistic beamforming scheme

1. For time-slot n a random beamforming vector $\mathbf{b}_n$ is drawn from a multivariate circularly normal complex distribution and its magnitude is normalized:

$$\tilde{\mathbf{b}}_n \sim \mathcal{CN}(0, I_{N_t})$$
$$\mathbf{b}_n = \frac{\tilde{\mathbf{b}}_n}{\|\tilde{\mathbf{b}}_n\|}$$

2. Then this beamforming vector is applied to the transmit antennas using known data, as a pilot. In other words, $\mathbf{b}_n$ is multiplied by the known pilot signal, say x_n and this product is scaled subject to the desired transmission power and then applied to the transmit antennas. Each terminal listens and measures its received SNR and sends it back to the base-station. The base-station finds their $R_k(n)$'s (i.e. their supportable rates in this time-slot) according to the modulation scheme.

3. The proportional fair scheduler at the base-station assigns the time-slot to the terminal k^* that maximizes the ratio $R_k(n)/T_k(n)$. Here, T_k is the same as the one mentioned above, that is, the average received service during a time window of length τ. Here, $T_k(n)$ means the received service *right before* time-slot n.

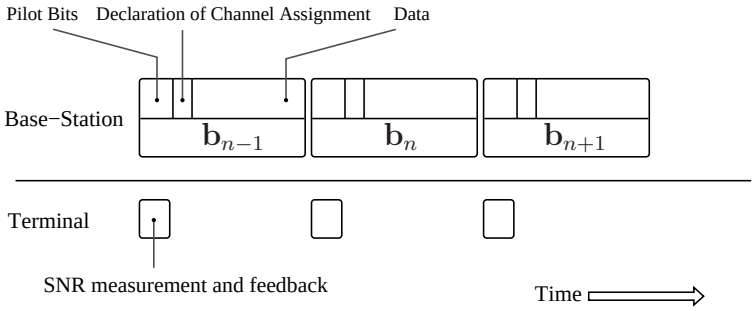
4. Since the terminal k^* has received service with rate $R_{k^*}(n)$, and the other terminals

have received zero bits during this interval, the service measure T_k is updated as follows:

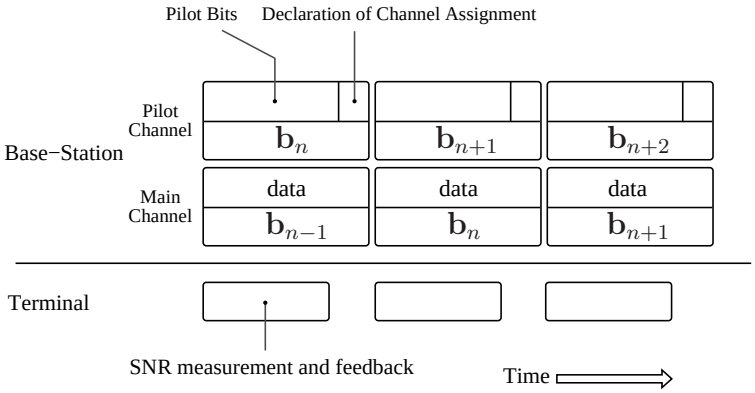
$$T_k(n+1) = \begin{cases} (1-\frac{1}{\tau})T_k(n) & k \neq k^* \\ (1-\frac{1}{\tau})T_k(n) + \frac{1}{\tau}R_k(n) & k = k^* \end{cases} \tag{2.2}$$

The only difference between this method and the proportional fair method (i.e. without spatial diversity) is the usage of the random beamformer, and the fact that it is required to "publish" the beamforming vector prior to the decision of the scheduler. This can be done in the initial proportion of the same time-slot in the same channel as an operation in the header of the time-slot, or in an adjacent pilot channel which has fading characteristics similar to the main channel. In latter case, terminals listen to the pilot channel and report their SNRs one slot earlier. The timing of both modes are illustrated in Figure 2.6. In the figure, $\mathbf{b}(n)$ represents the beamforming vector in time-slot n.

The throughput-optimal scheme combining multiuser diversity and transmit beamforming is, of course, a method based on the MaxRate terminal selection and coherent beamforming. Such a scheme perfectly knows the channels of all terminals, and selects the terminal which has the best MISO channel at the moment and then sets the beamforming vector as the one which optimizes this terminal's received SNR, i.e., the conjugate of the channel vector of the selected terminal. The throughput optimality has the cost of unfairness. The performance gap between the MaxRate + coherent-beamforming scheme and the opportunistic beamforming is large and requires further improvements [7]. Note that increasing the number of users reduces this gap because with large number of users it is more likely that the random beamforming vector lies near the optimal vector for some users. So, for large number of users the performance of this scheme asymptotically converges to that of coherent-beamforming. Unfortunately, the sufficient number of users for this to



(a) employing mini-slot.



(b) employing a separate pilot channel.

Figure 2.6: Timing diagrams for two modes of operation (a) using mini-slots (b) using a separate pilot channel, in the opportunistic beamforming scheme

happen grows exponentially with the number of transmit antennas. This fact is explained by Sharma and Ozarow [8], where the number of users needed in the opportunistic beamforming scheme and a few similar methods are found and compared. The need for a large number of users forms a major burden in sparse/realistic networks, and for structures such as wireless mesh-networks. For some scenarios, having a large performance gap when there are few terminals in the system is a major issue, and throughput improvements in the case of sparse networks are needed to make this scheme desirable in such applications.

Another important advantage of opportunistic beamforming is its lower amount of required feedback. Coherent beamforming needs to know all channel coefficients from any antenna to any terminal. This is a large amount of feedback and would also impose additional measurement equipment on the terminal. Opportunistic beamforming, on the other hand, needs only typical SNR feedback reports and doesn't require any major modification in the terminal device [1] which is an important advantage from a practical perspective.

The simplicity and advantages of this scheme has inspired many researchers to propose new schemes to improve the performance of standard opportunistic beamforming.

One approach to improve the performance of the standard opportunistic beamforming scheme is to use multiple beamforming vectors per time-slot and to choose the one which provides highest SNR. This approach is used, for instance, by Laroia et al [9] and Kim et al [10]. Laroia et al [9] employ opportunistic beamforming scheme in a system with multiple frequency (or code or possibly time) divisions resulting in performance improvement due to frequency (or code or time) diversity in addition to multiuser diversity. Multiple beamforming vectors [10] are also employed, where the maximum throughput and the optimum number of beamforming vectors per time-slot are also given analytically. Q^{opt}, the optimum

[1] Currently many terminal devices are able to report their SNRs. These reports are used for power control, topology control, admission control, etc.

number of beamforming vectors per time-slot is obtained from[2]

$$\hat{Q} = \frac{1}{K} e^{W\left(\frac{LK}{T} exp(1+\gamma)\right) - 1 - \gamma}, \tag{2.3}$$

where $W(.)$ is the Lambert's W-function (the inverse function of xe^x) and γ is the Euler-Mascheroni constant (The approach involves harmonic series).

Both schemes result in short-term performance improvements compared to the standard opportunistic beamforming. The cost paid here is the additional feedback which increases with Q, the number of beamforming vectors per time-slot[3].

A related approach to the previous (i.e. multi-beam per time-slot) approach is to transmit simultaneously to more than one terminal per time-slot (the multi-beam approach) [11]. They considered multiple spatially separated channels per time-slot, and found how many of such parallel channels should co-exist. These parallel channels actually use orthogonal beamforming vectors. They have found that the capacity associated with their scheme scales as $M \log\log KN$, where K is the number of terminals, N is number of antennae per terminal, and M is the number of orthogonal beamforming vectors per time-slot. They further showed that the optimal value of M should be approximately $\log K$. Their scheme offers optimal capacity scaling laws when the number of terminals is large enough. A limitation of this scheme, however, is that it is far from optimal for low to moderate number of users, which is a major drawback from a practical perspective. The performance gap between this scheme and the MaxRate + coherent-beamforming scheme grows even faster than the standard opportunistic beamforming when the number of terminals is decreased. The reason is that it becomes more unlikely that a set of randomly generated orthogonal

[2] As $\hat{Q}$ obtained from (2.3) might be non-integer, Q^{opt} is either $\lfloor \hat{Q} \rfloor$ or $\lceil \hat{Q} \rceil$.

[3] However the amount of feedback for Q divisions is less than Q-times the amount of feedback in standard opportunistic beamforming. That is because the excess feedback can be coded with fewer bits.

beamforming vectors happen to be close to coherent beamforming vectors of any set of terminals, when the number of terminals decreases. Also, unlike Viswanath, Tse and Laroia [6], the scheme of Sharif and Hassibi [11] does not have a fairness guarantee when the system is not interference dominated.

Now, we briefly review the concepts of the scheme by Sharif and Hassibi [11]. In this framework, at each time-slot the transmitter generates N_t random beamforming vectors and serves the users with the highest SINRs. *The random beamforming vectors are generated independently from one time-slot to the other.* Assuming there is only one antenna at each terminal, a $N_t \times N_t$ unitary matrix Q is generated according to an isotropic distribution. At time slot t, the transmitted signal is

$$x(t) = \sum_{m=1}^{N_t} \mathbf{q}_m(t) s_m(t),$$

where $s_m(t)$ is the mth transmit symbol at time slot t and $\mathbf{q}_m$'s, columns of Q, are random orthonormal vectors (beamforming vectors) for $m = 1, ..., N_t$. Therefore, the received signal at the kth terminal is:

$$y_k(t) = \sum_{m=1}^{N_t} \mathbf{h}_k^{\mathrm{H}}(t)\mathbf{q}_m(t)s_m(t) + w_k(t), \quad k = 1, ..., K.$$

Each terminal measures its SINR on each one of the N_t random beamforming vectors according to

$$\mathrm{SINR}_{k,m} = \frac{|\mathbf{h}_k^{\mathrm{H}}(t)\mathbf{q}_m(t)|^2}{\frac{1}{\rho_k} + \sum_{j \neq m} |\mathbf{h}_k^{\mathrm{H}}(t)\mathbf{q}_j(t)|^2},$$

and feeds back its maximum SINR and the index m of the vector for which its SINR is maximized. For each vector $\mathbf{q}_m$, the transmitter assigns the beam to the user with the highest corresponding SINR on that beam, i.e. $\max \mathrm{SINR}_{k,m}$. The sum rate of the above

scheme is given by Sharif and Hassibi [11],

$$C_{\text{sum}} \approx \mathrm{E}\left[\sum_{m=1}^{N_t} \log_2\left(1 + \max_{1\le k\le K}\{\text{SINR}_{k,m}\}\right)\right].$$

It is shown that the sum rate of the opportunistic multiuser beamforming (using this kind of beamformer [4]) scales as $N_t \log\log K$, which is essentially the same as the scaling law of full channel knowledge and using Dirty Paper Coding [11].The drawback is that a large number of users is required for this to happen, which might not be feasible.

Incorporation of other forms of diversity such as frequency division has also been a subject of research. For example, the extension of the opportunistic beamforming scheme to OFDM systems has been studied by Svedman et al [12], [13].

2.1.2 Employing Excessive Feedback

The amount of (SNR) feedback required in a proportional fair scheduling policy is studied by Gesbert and Alouini [14]. There, it is shown that the full feedback rate (i.e. when all the terminals report their SNRs for every time-slot) is too large and the system can perform well even with as much as 10 percent of the full feedback rate.

The scheme proposed by Sanayei and Nosratinia [15] further demonstrates this fact. They have shown that even when the feedback rate is reduced to only one bit per terminal per time-slot, the capacity grows as $\log\log n$ which is the sum-rate capacity of greedy (optimal) scheduling. Each terminal in their scheme compares its received SNR with a certain threshold and reports whether it receives better or worse SNR than the threshold. Then the scheduler randomly chooses one of the terminals from those who receive better SNR than the threshold. Recently, the authors have further pursued this direction [16] and have suggested *the usage of excess feedback information in forming beamforming vectors*.

[4] That is when prior SNR measurements are not employed in the forming of beamforming vectors.

The idea here is to incorporate SNR feedback information to reduce the randomness of the beamformer and to make it as accurate as possible in a way to increase the performance. The proposed scheme by Sanayei and Nosratinia [16] uses the limited feedback approach of Love et al [17]. This approach basically obtains channel coefficients and uses quantization methods (e.g. Vector Quantization) to send the required information back to the base-station. Then, the base-station forms the best beamforming vector using this information and the performance gap is significantly reduced. However, this approach imposes the measurement of full channel matrices rather than the typical SNR measurement, so, some practical issues are raised.

A scheme proposed by Avidor, Ling and Papadis [18] also employs the SNR feedback information in the formation of beamforming vectors. We briefly explain this scheme here: The base-station typically receives SNR measurement feedback from each terminal and stores this information. For each time-slot, a terminal who has received the lowest service is chosen. The beamforming module intends to pick a beamforming vector *from the set of recently used beamforming vectors* such that the received SNR of this terminal is increased. This beamforming vector is chosen as the one maximizing the following criterion:

$$j = \arg \max_{n=1,2,\ldots,N} \left\{ \frac{\sum_{\mathbf{g}_{t_i}=\mathbf{b}_n} R_{t_i} a^{t-t_i}}{\sum_{\mathbf{g}_{t_i}=\mathbf{b}_n} a^{t-t_i}} \right\}, \tag{2.4}$$

where t is the current time, t_i is the time-stamp of the ith time-slot during which this terminal was chosen, $R(t_i)$ is the transmission rate at that time-slot, $\mathbf{g}_{t_i}$ is the beamforming vector used in time t_i and $\mathbf{b}_n$ is the nth beamforming vector from the pool of N beamforming vectors. The forgetting rate $0 < a < 1$ undermines the older measurements in favour of newer ones. Basically this rule considers the set of beamforming vectors, and finds a weighted average of the rate this terminal has received when it was being served under that

beamforming vector. The beamforming vector that has previously provided the highest average rate for this terminal is chosen. The rest of the system is the same as the standard opportunistic beamforming. As we see, the contribution of this work is the usage of feedback information in the form of SNR feedback to improve the random beamformer module of the standard opportunistic beamforming. The nature of the approach we use is very similar to the direction pursued here which is different from the previous scheme in that they use typical SNR feedback rather than applying compression and quantization techniques on the feedback link to transfer measured channel coefficients. One subtle difference between this scheme and our approach is that we actually estimate the channel coefficients using the SNR feedback, rather than using this information to choose from a pool of previous ones. The performance of the scheme by Avidor, Ling and Papadis [18] is inferior to the one by Sanayei and Nosratinia [16]. That is because, the former does not form the optimal beamforming vector whereas the latter can provide optimal or sub-optimal beamforming vectors because the latter measures and knows channel matrix and just needs to convey this information from the terminal to the base-station. As we will show later on, it is partially possible to use the same information (and terminal equipment) as Avidor, Ling and Papadis [18] and achieve a performance close to Sanayei and Nosratinia [16] and coherent beamforming, when the channel variations are not too fast. We do so by using that information to *combine* beamforming vectors rather than *choosing* one from them.

A similar method is also proposed by Kountouris and Gesbert [7]. There, a beamformer that uses previous SNR information is called a memory-based beamformer. This scheme applies the concepts of the scheme by Avidor, Ling and Papadis [18] in the general multi-beam case in Sharif and Hassibi's scheme [11]. Since they do not actually combine the beamforming vectors, the performance is essentially similar to Avidor, Ling and Papadis

[18].

There are a few schemes that have considered almost similar approaches as in Avidor, Ling and Papadis [18], while using beamforming combination rather than selection. An example is a recent work by Kim et al [19]. In this work, the channel is assumed to be Rician, and the Direction of Arrival (DoA) is a parameter of interest. They find the DoA using only SNR feedback and use the DoA in the formation of beamforming vectors. Their scheme significantly improves the performance gap of the standard opportunistic beamforming in Rician channels. While this improvement is notable, reducing the performance gap in Rayleigh fading models remains an open problem.

2.2 Scheduling Disciplines

As mentioned before, the MaxRate discipline, which assigns the common channel to the terminal with the maximum supportable rate, is throughput-optimal. The reason MaxRate is not practically acceptable for most applications is its unfairness and extremely poor QoS. In contrast with wire-line packet-switched networks, providing QoS guarantees in packet-switched wireless networks is not an easy problem. The challenge here is to design a policy which sacrifices as little as possible of total system throughput while providing the required QoS and fairness among sessions or terminals. To this end, various scheduling algorithms have been proposed. Although the problem of “scheduling” has been studied

for long and has a rich literature, the cross-layer combination of scheduling and channel-precoding/beamforming has been proposed only recently[5]. Therefore choosing or designing a scheduling algorithm for such a scheme requires revisiting already established algorithms as well as inventing new ones and *investigating their performance in the cross-layer framework*. A brief review of relevant algorithms follows.

2.2.1 Round-Robin

In round-robin (RR) scheduling, arriving packets are queued for each terminal, while the scheduler polls queues for service in a cyclic order irrespective of the wireless link conditions of the terminals to send/receive the packets. As greedy terminals flood their own queues, RR is able to provide terminal isolation. In the scheduling period starting at time t RR picks user k_t^* such that

$$k_t^* = (t \mod K) + 1. \tag{2.5}$$

By scheduling terminals/queues in a cyclic order, RR is able to provide fairness. Hence, it is used as fairness benchmark with respect to other schedulers.

2.2.2 FIFO

The First in, First out (FIFO) scheduler completely ignores wireless channel conditions, as its policy is completely based on the waiting times of packets/terminals. FIFO is also

[5]Here we mean designing scheduling and beamforming in a combined cross-layer scheme in which the traditional separation between physical layer and MAC layer is removed or blurred. Such a cross-layer design has been suggested [1] and shown to be able to provide new benefits that are impossible in the traditional layered system model.

referred to as First Come, First Served (FCFS). In the scheduling period starting at time t the FIFO scheduler picks the terminal k^* according to:

$$k_t^* = \arg\max_k \{D_k(t)\}, \tag{2.6}$$

where $D_k(t)$ is the number of time slots that terminal k has been waiting for service turn at time t, i.e. its starvation period, or service delay. $D_k(t)$ is a congestion indication parameter at terminal (queue) k at time t. Besides poor spectral efficiency, the FIFO scheduling does not provide fairness nor provide protection of well-behaving queues against greedy terminals/queues. However, it can be used as latency benchmark with respect to other scheduling policies.

2.2.3 MaxSNR

Also known as MaxRate, the maximum SNR (MaxSNR) scheduling picks a terminal k^* among all active terminals in the system at time t which has the best SNR, or equivalently, the best feasible instantaneous data rate $R_k(t)$. In other words, the terminal k^* is selected at time t such that

$$k_t^* = \arg\max_k \{R_k(t)\}. \tag{2.7}$$

Note that SNR and data rate have a one-to-one monotonically-increasing mapping for a given modulation scheme and channel bandwidth, so it is also called MaxRate. The MaxSNR rule can easily starve terminals that undergo an extended period of poor link state, but provides very attractive system throughput. Hence, it can be used as throughput benchmark with respect to other scheduling policies.

2.2.4 Proportional Fair Sharing

Proportional Fair Scheduling or Proportional Fair Sharing (PFS) algorithm [4], [6], [5] is an opportunistic MAC scheme which provides fairness across terminals but not for traffic classes. Hence, it cannot quantify service level of individual flows backlogged for a given terminal. At time t PFS picks terminal k^* among all backlogged terminals in the system which has the best supportable data rate normalized by the average short-time throughput it has received so far, i.e.,

$$k_t^* = \arg\max_k \left\{ \frac{R_k(t)}{T_k(t)} \right\}, \tag{2.8}$$

where $T_k(t)$, the short-term average data rate of terminal k until time t, is updated using the exponentially weighted low-pass filter according to (2.2). The PFS is the default scheduler for the downlink of CDMA/HDR (also known as 1xEV-DO, IS-856) which is an evolution of 3GPP2's CDMA2000 sharing a frequency space with IS-95 [5], [20]. However, proportional fair scheduling has been shown to be unstable[6] in HDR according to Andrews [21]. Andrews [21] uses a sample traffic to show the insatiability of this scheme.

2.2.5 M-LWDF

Modified Largest Weighted Delay First (M-LWDF) introduced by Andrews and Kumaran [22] is a scheme based on a combination of FIFO and proportional fair scheduling. The priority value assigned to a terminal by the M-LWDF scheduler is equal to that of the PFS policy scaled by the weighted delay the terminal has endured. In the scheduling interval

[6] A scheduling policy is stable when it can support any traffic that is supportable under any other scheme. In other words, a scheduling scheme is unstable when there exists a traffic that it is not supportable under this scheme, but there exists another scheduling scheme that can support said traffic.

starting at time t, the M-LWDF policy picks terminal k_t^* among all competing terminal satisfying

$$k_t^* = \arg\max_k \left\{ a_k D_k(t) \frac{R_k(t)}{T_k(t)} \right\}, \tag{2.9}$$

where $T_k(t)$ is defined as in (2.2) and $a_k > 0$ is a parameter indicating the QoS level desired by terminal k. $T_k(t)$ can also be the median of $R_k(t)$.

2.2.6 Exponential Scheduler

The exponential scheduler (EXP) [23] attempts to equalize the weighted delays (starvation times) of all queues when their differences become large. It is designed to support real-time services in an AMC/TDM wireless system. In the scheduling period starting at time t, the EXP scheduler picks terminal k_t^* such that

$$k_t^* = \arg\max_k \left\{ \frac{R_k(t)}{T_k(t)} \exp\left(\frac{a_k D_k(t) - \frac{1}{K}\sum_k a_k D_k(t)}{1 + \sqrt{\frac{1}{K}\sum_k a_k D_k(t)}} \right) \right\}, \tag{2.10}$$

$D_k(t)$ can also be the number of packets in terminal k's buffer at time t. As it can be observed in (2.10), a large weighted delay $a_k D_k(t)$ overrides $R_k(t)$, hence restricting packet delays to a certain level. The EXP rule, however, reduces virtually to the PFS policy at low weighted delays. $T_k(t)$ is defined as in (2.2).

2.2.7 OCASD, I-OCASD, CASTI

OCASD (Optimum Channel-Aware Scheduling with Service Differentiation) and I-OCASD (Inverted-Optimum Channel-Aware Scheduling with Service Differentiation) are connection-based scheduling schemes with interesting short-time fairness and throughput properties.

In a multi-class system, a connection $x = (c, k)$ is a pair of class number (or "port" number) c and terminal number k. OCASD, as introduced by Gyasi-Agyei [24], is a traffic-aided opportunistic MAC, which optimizes the trade-off between short-term fairness and throughput maximization, but without any guarantees on the delay constraints of the packets. In the scheduling period starting at time t, OCASD picks the connection x_t^* such that:

$$x_t^* = \arg\max_x \left\{ [|w_c - \frac{B_c(t)}{B(t)}|B(t)]^{\text{sgn}(w_c - \frac{B_c(t)}{B(t)})} R_m(t)/d_c \right\}, \tag{2.11}$$

where w_c is the weight assigned to traffic class $c \in \{\text{real-time}, \text{non-real-time}\}$, $B(T) = \sum_{c \in \mathcal{B}} B_c(t)$, $\mathcal{B}$ is the set of backlogged queues at time t and the amount traffic scheduled from/to traffic class c so far is updated using the first-order IIR (Infinite Impulse Response) filter structure

$$B_c(t+1) = (1 - \frac{1}{\tau})B_c(t) + \frac{1}{\tau} L_c(t)\delta(c_t^*, c), \tag{2.12}$$

where $\delta(\cdot, \cdot)$ is the Kronecker delta function, c_t^* is the class number (c) of x_t^*, the selected connection at time t. The weight w_c depends on the characteristics; hence, the QoS requirements of traffic class c. d_c is the delay deadline (or bound) of the head-of-line packet of size $L_c(t)$ in queue c. Note that queue c is penalized at time t if $w_c < \frac{B_c(t)}{B(t)}$ since in that case $\text{sgn}(w_c - \frac{B_c(t)}{B(t)}) = -1$, and thus its priority scaling factor $()^{\text{sgn}(w_c - \frac{B_c(t)}{B(t)})}$ is less than unity. τ is the sliding time window in which a given set of K terminals share a finite wireless resource. The sliding time window (also known as latency time scale or scheduling window) τ, is an important design parameter as it indicates the trade-off between throughput maximization, fairness and packet delay profile. The delay profile is crucial for real-time traffic. τ determines how long a packet can be delayed before being scheduled. A longer

τ is good for throughput maximization but it is limited by the maximum delay tolerance of network applications. $\tau = 100\text{ms}$ is a typical value proposed for most data applications [25].

I-OCASD [26] is the generalization of proportional fair scheduling to the connection-based schemes. I-OCASD picks the connection x_t^* at time t such that

$$x_t^* = \arg\max_x \left\{ \frac{R_k(t)}{B_c(t)} \right\}, \tag{2.13}$$

where $B_c(t)$ is defined as in (2.12).

2.2.8 GPS-based Policies

An important class of schedulers are schedulers based on "Generalized Processor Sharing". GPS was introduced by Parekh and Gallager [27]. GPS is an ideal scheme in which two non-practical assumptions are made: (a) The common resource (channel/processor/...) can serve multiple clients simultaneously, and (b) Traffic is fluid (i.e. packets have infinitesimal sizes). Based on these assumptions the system works as follows: when a packet arrives it is promptly accepted into the server and its serving time starts immediately. The common server is shared equally between all the current packets being served. The start-time and finish-time of each packet under such policy are ideally fair[7]. As mentioned, this scheme is fair but non-practical. However it is possible to approximate the start-times and/or finish-times (departure times) of this scheme to develop fair scheduling policies. There are many approximations of GPS, varying in complexity, deviation from GPS, and worst case fairness. Weighted Fair Queueing (WFQ) is the most famous one of such approximations. This

[7]This is fairness among sessions(packets). If we care about fairness among terminals/users we can allow only one packet per terminal in the system and keep the rest of the packets of a terminal in its dedicated queue. Upon the departure of each packet of a terminal its next packet in the queue begins to be served.

scheme computes the finish-times of the packets if they were going to be served under GPS. This computation can be done by directly simulating GPS, although more efficient computational methods are often employed. Then, when a scheduling choice is being made, the packet with the minimum GPS-finish-time is chosen and served. In other words, packets get scheduled in the order of their GPS-finish-times.

Although GPS-based schemes have wide applications in many scheduling problems, they are not effective for variable-service-rate systems. In practical wireless communication systems, variable rate transmission cannot be overlooked. So, as shown by Joshi et al [28], processor scheduling is not effective for practical wireless systems where channel feedback is quantized.

2.3 System Model and the Proposed Approach

We consider the downlink of a multiuser TDM wireless MISO (N_t) communication system. The traffic is assumed to be packet-based. The system has a base-station and K terminals. At any given time the common channel is assigned to a single terminal, i.e., it has access to channel bandwidth resources for a time-slot. The channel is modelled by a slow or fast Rayleigh fading process. Let $\mathbf{h}_k(t) \in \mathbb{C}^{N_t \times 1}$ be the channel vector of terminal k at time-slot t. We imply that the channel remains static during each time-slot. We assume that every terminal has an SNR measurement device that listens to the common channel all the time (even when the common channel is assigned to another terminal) and accurately measures the SNR, but due to limitations in the feedback channel the measured quantity will be quantized. Each terminal sends a measurement report when it exceeds a certain fixed or dynamic threshold. The base station receives SNR-measurement feedbacks from terminals and maintains a list of most recent high SNRs for every terminal. Based on

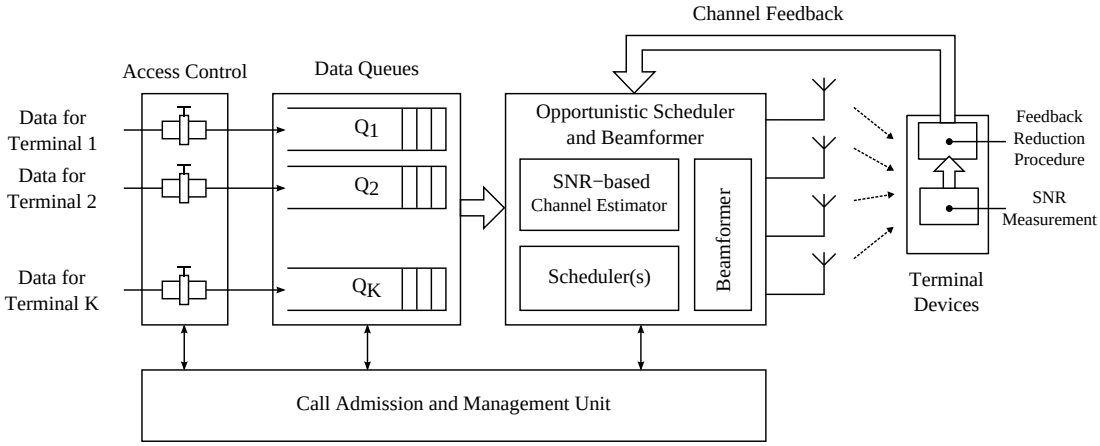


Figure 2.7: An overview of the proposed system

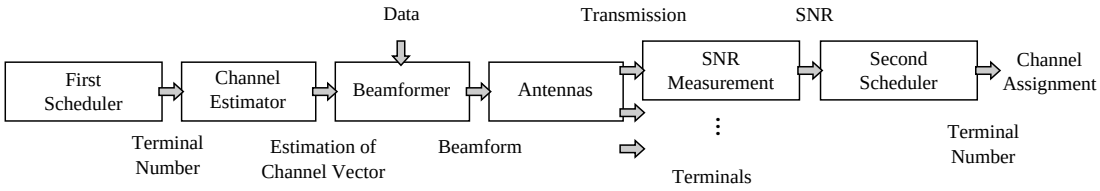


Figure 2.8: The block-diagram of the proposed procedure in one time-slot

SNRs and the status of the queues corresponding to each terminal, the base-station schedules the common channel to one of the terminals in an opportunistic manner. In order to achieve spatial diversity the base station uses the MISO channel by deliberately forming beamforming vectors (rather than random/blind beamforming). The beamformer uses the estimation of the channel. The estimation of the channel is computed using only the SNR feedback reports. We assume that the base-station and the terminals use adaptive modulation and coding (AMC), and we assume that when a terminal is scheduled for transmission, it receives data with rate $\log(1 + \text{SNR})$. We can make this assumption because SNR is known to the base-station and the terminal prior to this transmission.

The overall procedure for every time-slot is briefly as follows:

1. The "first stage of the scheduler" selects a user based on its queue status and possibly[8] previous SNRs.

2. The "channel estimator", which is a maximum-likelihood estimator, finds a best estimation of the selected user's channel vector. The estimator uses only the channel SNR-measurement reports.

3. The "beamformer" develops a beamforming vector based on the estimation obtained in the previous step, and an additional random term whose variance is also determined from the estimation step, or by a higher level management entity. The addition of a random term is a mechanism to prevent the space of "recent beamforming vectors"[9] from shrinking gradually.

4. The beamforming vector is then applied to an adjacent pilot channel. Each terminal has to listen carefully and report its SNR if it is high enough. A "feedback reduction policy" states how high should the received SNR of a terminal be, to allow it to report its SNR and use the feedback channel.

5. The "second stage of the scheduler" receives these reports, and based on the reports and the status of the queues, selects a terminal to which the next time slot will be assigned.

Note that the first and second scheduler are not essentially different. They only differ in that the first scheduler uses *estimations* of channels of terminals, whereas the second scheduler employs the *actual* channel SNR measurements fed back from terminals. Also,

[8]We study both cases for the First Scheduler: Channel state independent, and channel state dependent. In the channel state dependent case, the first and second schedulers are almost identical.

[9]The base-station keeps a number of beamforming vectors. For every time-slot a new beamforming vector is formed and the oldest one is dropped from the list.

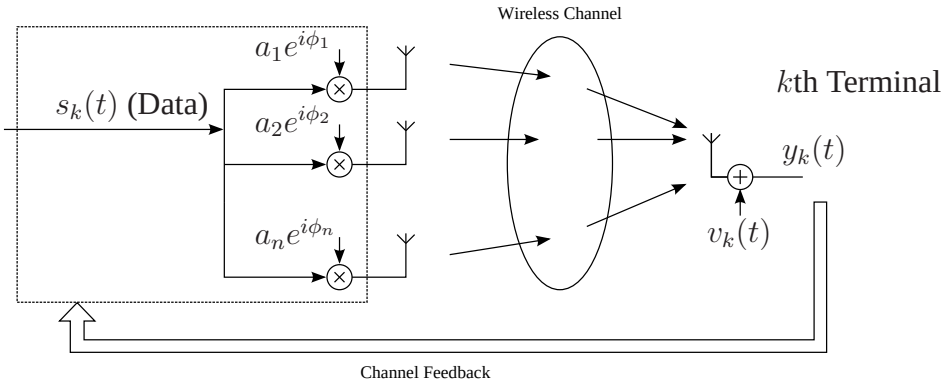


Figure 2.9: Beamformer and channel model

the first scheduler controls the insertion rate of random beamforming vectors. The insertion of random beamforming vectors orthogonal to the space of recent beamforming vectors is a mechanism that prevents the rank of this space from shrinking. The probability p with which these insertions occur, controls the expected time of channel acquisition for new terminals.

The normalized received signal of terminal k, when it is scheduled for transmission is assumed to be as follows

$$y_k(t) = \sqrt{\frac{P_0}{\sigma^2_{v_k}}}\mathbf{b}^{\mathrm{H}}(t)\mathbf{h}_k(t)s_k(t) + v_k, \tag{2.14}$$

where P_0 is the transmission power which is limited by the maximum allowed or desired transmit power from the base-station, and is controlled by the power-control unit. $\sigma^2_{v_k}$ is the power of the noise for terminal k. $\mathbf{b}(t) \in \mathbb{C}^{N_t}$ is the beamforming vector at time t. We assume that $\|\mathbf{b}(t)\| = 1$. Also, $a_j e^{i\phi_j}$'s (in Figure 2.9) are the elements of $\mathbf{b}(t)$. $\mathbf{h}_k(t) \in \mathbb{C}^{N_t}$ is the channel vector (MISO) of terminal k at time t. $s_k(t)$ is the unit-variance data signal for terminal k. v_k is the normalized (unit-power) noise for terminal k. y_k is the normalized received signal of terminal k.

The received SNR of terminal k at time t is given by

$$\text{SNR}_k(t) = \frac{P_0}{\sigma^2_{v_k}} |\mathbf{b}^{\text{H}}(t)\mathbf{h}_k(t)|^2, \tag{2.15}$$

which is measured by each terminal. The second part of the SNR is especially important in our model, so we call it normalized SNR, (nSNR), $r_K(t)$, according to

$$r_k(t) = |\mathbf{b}^{\text{H}}(t)\mathbf{h}_k(t)|^2. \tag{2.16}$$

As we assume the availability of pilot signals, terminals are able to measure their received noise level, so $\sigma^2_{v_k}$ is known at each terminal k. We assume that the base-station keeps the terminals informed of the value of P_0, so terminals can separate $\frac{P_0}{\sigma^2_{v_k}}$, from $r_k(t)$. They apply the feedback reduction policy[10] on the value of $r_k(t)$, and if it is significant according to the policy, they report both $\sigma^2_{v_k}$ (or alternatively $\frac{P_0}{\sigma^2_{v_k}}$), and $\text{SNR}_k(t)$ to the base-station. The base-station uses these reports at time t and some other times after that. By the time t' when the base-station employs the reports, $\mathbf{h}_k(t')$ has changed from the channel state at the time of the report, $\mathbf{h}_k(t)$. Due to this change, and also measurement and quantization errors, the base-station actually has the reports $\rho_k(t)$'s instead of $r_k(t)$'s according to

$$\rho_k(t') = |\mathbf{b}^{\text{H}}(t)\mathbf{h}_k(t')|^2 \tag{2.17}$$

$$= |\mathbf{b}^{\text{H}}(t)\mathbf{h}_k(t)|^2 + \eta_k(t) \tag{2.18}$$

$$= r_k(t) + \eta_k(t), \tag{2.19}$$

where $\eta_k(t) \sim \mathcal{N}(0, \sigma_0^2)$ represents the aforementioned change and measurement and quantization errors. The base-station stores $\rho_k(t)$'s and employs them in channel estimation and scheduling.

[10] The feedback reduction policy in described in § 3.3.

Chapter 3

Channel Estimator

In this chapter we describe a channel estimation procedure based on the limited channel state information fed back from the terminals. The feedback is in the form of typical SNR measurement results. First we describe the goal of this module in § 3.1, where the proposed algorithm is also presented. Then we measure the performance of the proposed method and present the results in § 3.2. Afterwards, we describe our feedback reduction policy in § 3.3. A short summary is also provided at the end, in § 3.4.

3.1 Structure and the Proposed Algorithm

As mentioned before, the base station receives SNR measurement feedback from each terminal. The choice of SNR feedback as the form of channel state information has been made due to the ease of its measurement equipment and its ubiquity in current terminal devices. We assume that each terminal listens to the common channel and measures its perceived SNR, even when another terminal is scheduled for transmission. We assume the received signal of the kth terminal follows (2.14) and its SNR measurement is obtained

according to (2.17). Also, we make an additional assumption in this chapter: We assume that the channel vector of every terminal has not changed much since n time-slots ago[1]. The difference between the perceived normalized SNR in a previous time-slot (not older than n time-slots), and the normalized SNR, that this terminal could receive, if it had its current channel vector at that time-slot, is captured in the error variable $\eta_k(t_i)$. With a proper upper limit for $\eta_k(t_i)$, this assumption is a way of considering a fading channel with channel coherence-time of about n time-slots. The base-station keeps the n most recent beamforming vectors. Since there is no ambiguity, we simply refer to the n recent beamforming vectors as the set of beamforming vectors, forgetting the older beamforming vectors. We employ the measurement results obtained in this time window (i.e. the most recent n time-slots) for a *chosen terminal*[2] and estimate its channel. Note that as SNR is not a linear function of the channel vector, the estimation problem is non-linear and non-trivial, so a linear estimator such as MUSIC [29] is not effective here.

We use a maximum-likelihood estimator based on the assumption about the SNR measurement (2.17). The maximum likelihood estimation leads to minimization of the following error term

$$\epsilon = \sum_{j=1}^{n} \left(\left| \mathbf{b}_j^{\mathrm{H}} \mathbf{g} \right|^2 - \rho_j \right)^2 , \tag{3.1}$$

where $\mathbf{g}$ is the estimation of $\mathbf{h}_k$. For the derivation of (3.1) and additional notes about the estimation procedure see Appendix A. The estimation procedure is explained as follows.

1. Choose a heuristic initial vector $\mathbf{g}_0 \in \mathbb{C}^{N_t}$. We have used $\mathbf{g}_0 = \mathbf{b}(t_{m^*})$ where t_{m^*} is

[1] We can equivalently assume a quasi-static fading channel model, where the channel is constant over a period of time called a frame, and changes in the next frame.

[2] The first stage of the scheduler decides the terminal of choice.

the time stamp of the beamforming vector corresponding to the best SNR that user k has received.

2. Repeat: Find the gradient of the error term ϵ for $\mathbf{g}$, according to

$$\nabla\epsilon = \sum_{j=1}^{n} \mathbf{b}_j \mathbf{b}_j{}^{\mathrm{H}} \mathbf{g}_i \left(\left| \mathbf{b}_j{}^{\mathrm{H}} \mathbf{g} \right|^2 - \rho_j \right), \tag{3.2}$$

and subtract $\mu\nabla\epsilon$ from $\mathbf{g}_i$, where μ is a small coefficient which controls the step-size. We use a fixed step-size for simplicity [3], $\mu = 0.02$. So we have

$$\begin{aligned} \mathbf{g}_{i+1} &= \mathbf{g}_i - \mu\nabla\epsilon\left(\mathbf{g}_i\right) \\ &= \mathbf{g}_i - \mu \sum_{j=1}^{n} \mathbf{b}_j \mathbf{b}_j{}^{\mathrm{H}} \mathbf{g}_i \left(\left| \mathbf{b}_j{}^{\mathrm{H}} \mathbf{g}_i \right|^2 - \rho_j \right). \end{aligned} \tag{3.3}$$

3. Find ϵ_i according to

$$\epsilon_i = \sum_{j=1}^{n} \left(\left| \mathbf{b}_j^{\mathrm{H}} \mathbf{g}_i \right|^2 - \rho_j \right)^2.$$

If ϵ_i is larger than a threshold then repeat step 2, unless the number of repetitions exceeds MAX-Epochs (a predetermined number).

We refer to this procedure as the gradient descent procedure. We also refer to a variation of this procedure as stochastic gradient descent, in which the initial point is chosen randomly and the whole procedure is run several times and the best result is declared as the

[3] There are many ways to improve the gradient descent algorithm, for example, by optimizing the step size. We will add a random term to the estimation result $\mathbf{g}_i$ in the beamformer module. The estimation algorithm does not need to be more accurate than the level of the random term that must be added in the beamforming module. Therefore, it is not necessary to use a more complex estimation algorithm in this application.

final estimation. The initial point is chosen in the subspace spanned by $\mathbf{b}_1, \mathbf{b}_2, ..., \mathbf{b}_n$, since the measurements have no information about the other directions. It is important to note that the estimation ($\mathbf{g}$) always lies within the subspace spanned by $\mathbf{b}_1, \mathbf{b}_2, ..., \mathbf{b}_n$. This fact will create an issue to which we will refer in § 4.1.

3.2 Performance

We compare the performance of alternative algorithms and the proposed algorithm. Since in an adaptive modulation and coding (AMC) scheme such as ours, the supportable rate of each terminal can be computed given its SNR, we compare the SNRs of different methods. Also, we drop the common part of the SNR which is $\frac{P_0}{\sigma^2_{v_k}}$ in (2.15). We compare what remains, i.e.,

$$r = \left|\mathbf{b}^{\mathrm{H}}(t)\mathbf{h}(t)\right|^2, \tag{3.4}$$

where r is the normalized SNR (nSNR). The SNR for coherent-beamforming has the gamma distribution, if we assume that $\mathbf{h}$ is drawn from a circularly normal distribution:

$$f_r(r) = \frac{r^{N_t-1}e^{-r}}{(N_t-1)!} \quad \text{for} r \geq 0. \tag{3.5}$$

In the case of coherent beamforming, the diversity gain and coding gain can be obtained analytically. Since the other schemes also have nearly gamma-distributed SNRs, we approximate them by a gamma distribution and find the diversity and coding gain [4] accordingly. To fit a gamma distribution we use a maximum likelihood estimator [30]. We also find the slope of the BER vs SNR diagram in the high SNR region to find the diversity gain more accurately. The results are similar to the ones obtained through fitting a gamma,

[4] Diversity gain and coding gain are defined, for example, in Tse and Viswanath's book [3].

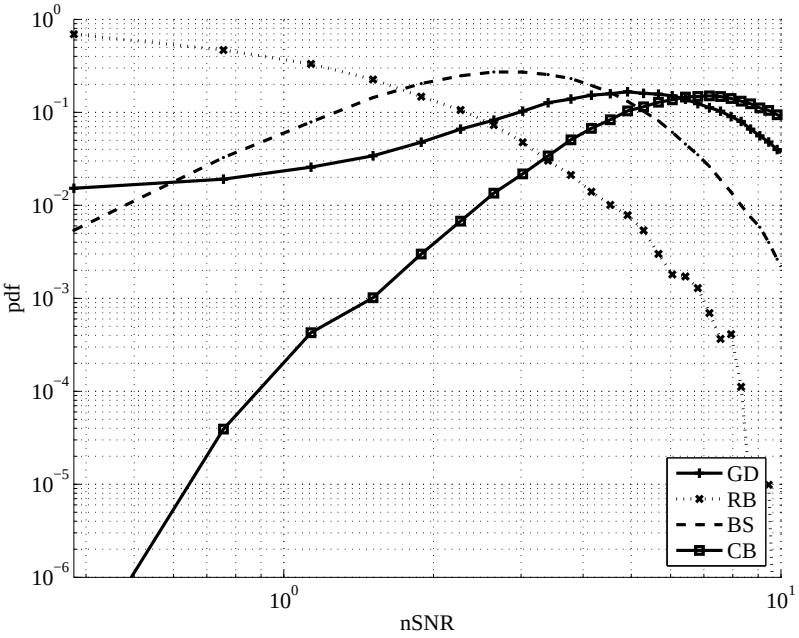


Figure 3.1: PDF of nSNR ($|\mathbf{b}^{\mathrm{H}}(t)\mathbf{h}_k(t)|^2$).

because the SNR distributions are close to the gamma distribution. These parameters are presented in Table 3.1.

Opportunistic beamforming uses a random beamformer. The SNR of a random beamformer has an exponential pdf, i.e., $f_r(r) = e^{-r}$. We also compare our scheme with a Best-Seen Beamforming scheme. Best-seen (BS) beamformer picks the beamforming vector (among the ones already used) that had the best SNR. The best-seen scheme is chosen to draw a comparison with Avidor, Ling and Papadis [18]. We assume that the previous beamforming vectors were drawn from a circularly normal complex distribution.

As a performance measure we compare the pdf of normalized received SNR for coherent-beamforming (CB), random-beamforming (RB), best-seen (BS), and our gradient-based beam-combiner (GD). This performance comparison is shown in Figure 3.1. In this example the number of transmit antennas is $N_t = 8$. Simulation setup is the same as the setup

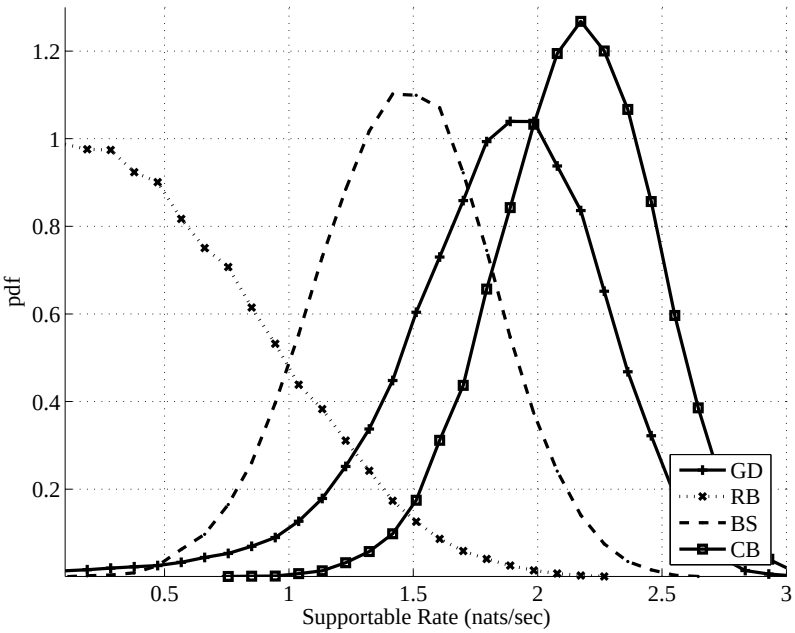


Figure 3.2: PDF of supportable rate ($R_k(t)$)

in § 3.3. Our scheme (GD) outperforms random beamforming (the standard opportunistic beamforming), and has a performance close to the coherent beamforming scheme. Its performance is better than the best-seen scheme except for nSNRs below 0.6. The reason for performance degradation in this particular area is that when nSNR is low, measurement results do not have enough information for an accurate channel estimation. In this case, the combination of a few inaccurate data points produces an inaccurate result which is less accurate than the best single measurement result.

Note that, the average feedback rate used in our scheme (GD) is $\frac{1}{8}$, i.e., on average, terminals report their received SNR once in eight time-slots. Therefore, the performance of GD shown in Figure 3.1 can be improved, should the terminals report their SNRs every time-slot. The effect of feedback rate on the performance will be studied in the next section.

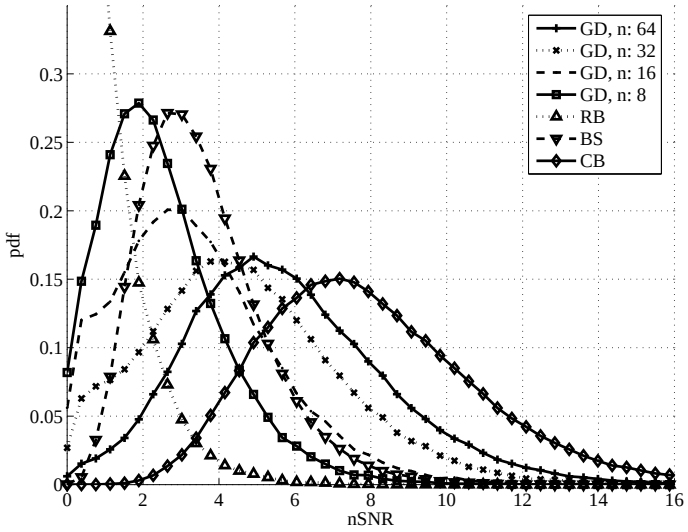
We also compare the pdf of supportable rates of these schemes. This is shown in Figure 3.2. We have assumed that the common part of SNR is 0dB and that the rate can be

found according to

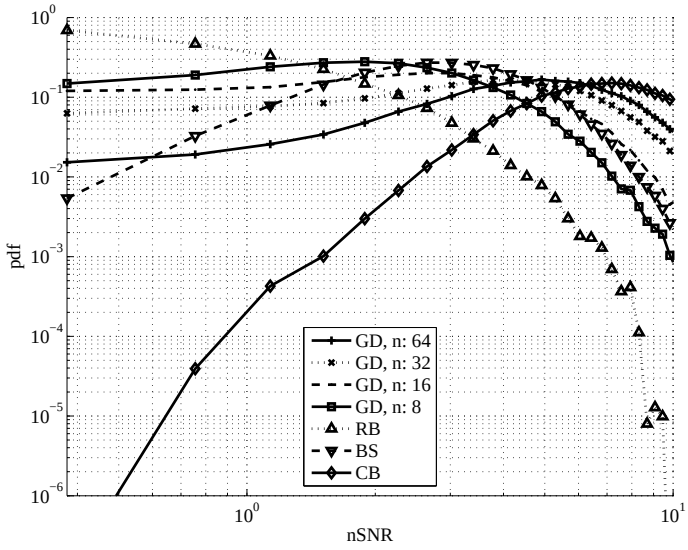
$$R_k(t) = \log\left(1 + \left|\mathbf{b}^{\mathrm{H}}(t)\mathbf{h}_k(t)\right|^2\right) \tag{3.6}$$

The SNR performance of any channel estimation algorithm should be better than the random beamformer, while it cannot exceed the performance of coherent beamforming. Here we can see that the performance of the GD scheme is close to that of the CB scheme. We also show the performance of our scheme for various n's, the number of measurement points used in the estimation, in Figure 3.3. As shown, the performance of our channel estimation algorithm approaches that of coherent beamforming with increasing n, the number of samples. Larger n's require longer channel coherent-times. Table 3.1 presents diversity gain and coding gain for various methods. Figure 3.4 and Figure 3.5 show the supportable-rate performance of various methods in a full system simulation which includes beamforming and scheduling modules explained in next sections. The simulation setup is the same as the one in § 3.3. Figure 3.4 employs MaxSNR scheduling method while Figure 3.5 uses the PFS scheduling policy. It is shown that GD outperforms RB and BS. It is interesting that the performance of BS under PFS scheduling rule with our simulation parameters degrades to the extend that RB outperforms BS when the number of terminals increases. Changing the feedback forgetting rate[5] can change this trend, and let the performance of GD and BS to increase with increasing the number of terminals, but this causes performance loss at low K's.

[5]Forgetting rate is explained in § 3.3



(a) Linear



(b) Logarithmic, low SNR region

Figure 3.3: PDF of received nSNR ($|\mathbf{b}^{\mathrm{H}}(t)\mathbf{h}_k(t)|^2$) for various coherence times.

Table 3.1: Performance parameters for various beamforming methods

Beamforming Method	N_t	Beamforming Diversity Gain (β)	Coding Gain (θ)
Coherent Beamforming	4	4.0	1.0
Gradient Based	4	2.9	1.0
Best-Seen	4	2.7	0.8
Random Beamforming	4	1.0	1.0
Coherent Beamforming	8	8.0	1.0
Gradient Based	8	6.2	1.0
Best-Seen	8	5.6	0.7
Random Beamforming	8	1.0	1.0
Coherent Beamforming	16	16.0	1.0
Gradient Based	16	9.8	1.0
Best-Seen	16	7.5	0.6
Random Beamforming	16	1.0	1.0

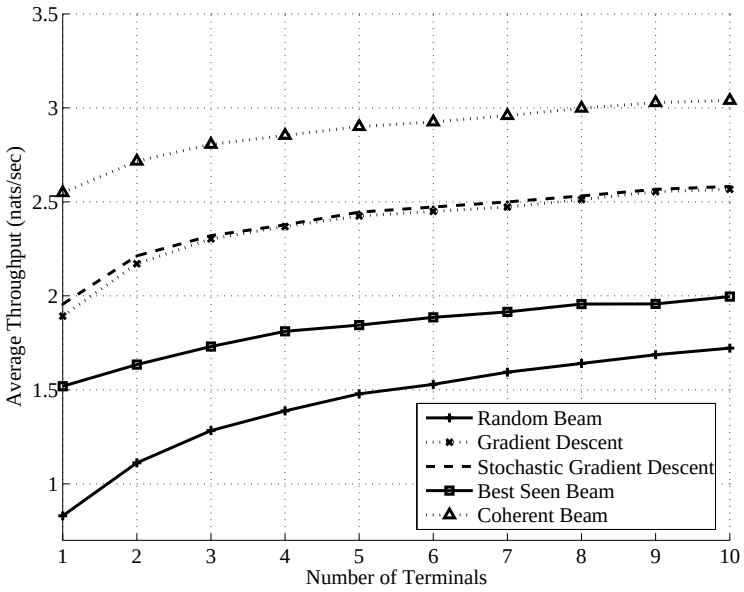


Figure 3.4: Performance of various estimation schemes in complete system simulation. Scheduling method: MaxSNR

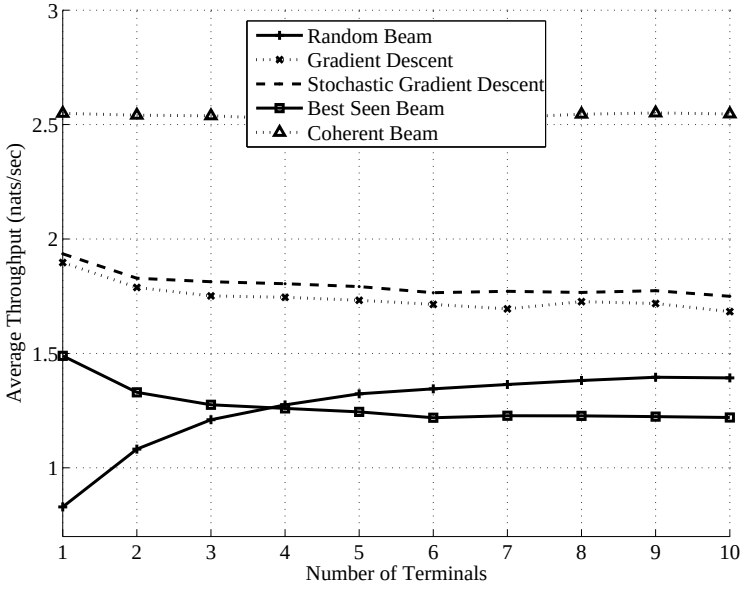


Figure 3.5: Performance of various estimation schemes in complete system simulation. Scheduling method: PFS

3.3 Feedback Reduction Policy

Rather than having all the terminals report their SNRs every time slot, we propose a policy to reduce the amount of feedback. Every terminal should compare its received SNR with its previous reported SNRs within a time-window, say n time-slots. SNRs are multiplied by a forgetting rate of the form α^{t-t_i} to undermine the older measurements, where $t - t_i$ is the age of the measurement, and $\alpha \in \mathbb{R}$, $0 < \alpha < 1$. When a new SNR measurement result is higher than at least one of previous results (multiplied by their forgetting rates), The terminal is allowed to report its SNR. The forgetting rate relates to the coherence time of the channel, when the coherence time is larger the forgetting rate should be closer to 1. The optimal forgetting rate can be computed by incorporating an estimation of the channel, but we assume its value is available. The feedback reduction procedure running at each terminal is described as follows.

foreach *time-slot* i **do**
    Measure SNR. (Even when the channel is assigned to another terminal);
    Compare SNR to the list of previous SNRs;
    if *SNR exceeds the smallest value in the list* **then**
        Report SNR to the base-station;
        Put the SNR in the list;
        if *list is larger than "SNR buffer length"* **then**
            Drop the smallest value of the list;
        end
    end
    Multiply each SNR in the list by $\alpha < 1$, the forgetting rate;
end

We use simulations to explore the effect of various parameters. The following simulation parameters are used throughout the thesis. The number of antennas in the simulations are $N_t = 8$. We assume a Rayleigh-fading channel model, and employ the auto-regressive model for the fading channel as presented by Baddour [31]. Maximum doppler frequency is 100Hz, and the order of AR model is 100. We assumed the symbol frequency is 10K (symb/sec). We have simulated for 2^{14} time-slots for each measurement. $\alpha = \frac{7}{8}$, unless we are testing various α's. BLL $= 16$, unless we are testing the effect of BLLs. We use MaxSNR scheduling method unless otherwise stated. The first randomization method [6] is used with $p = 0.1$. Maximum number of iterations [7]for a gradient descent procedure is set to 512, and the step size coefficient is $\mu = 0.02$. When we use PFS, M-LWDF, or EXP, we set the service window size to $\tau = 10$ time-slots for all terminals.

We show the effect of feedback reduction policy on the channel estimation performance. Figure 3.6 shows the average feedback rate for various values of the forgetting factor. Figure 3.7 shows the average total throughput for the same scenario.

Another parameter that affects the throughput performance of our system is the number of previous beamforming vectors considered in the feedback reduction policy. We refer to this buffer-size as BLL (Beamforming-vector List Length). We show how various BLLs affect the performance and feedback rate. Figure 3.8 shows the feedback rate, and Figure 3.9 shows the average total throughput for various BLLs. In these figures, a feedback rate of 1 means that every terminal reports its received nSNR in every time-slot. This rate is much lower than the amount of feedback needed for coherent beamforming, yet we can provide a significant improvement over best-seen and random beamforming schemes. As expected, increasing α reduces the feedback rate. Also it is interesting to note that increasing α from

[6] See § 4.1.
[7] See § A.

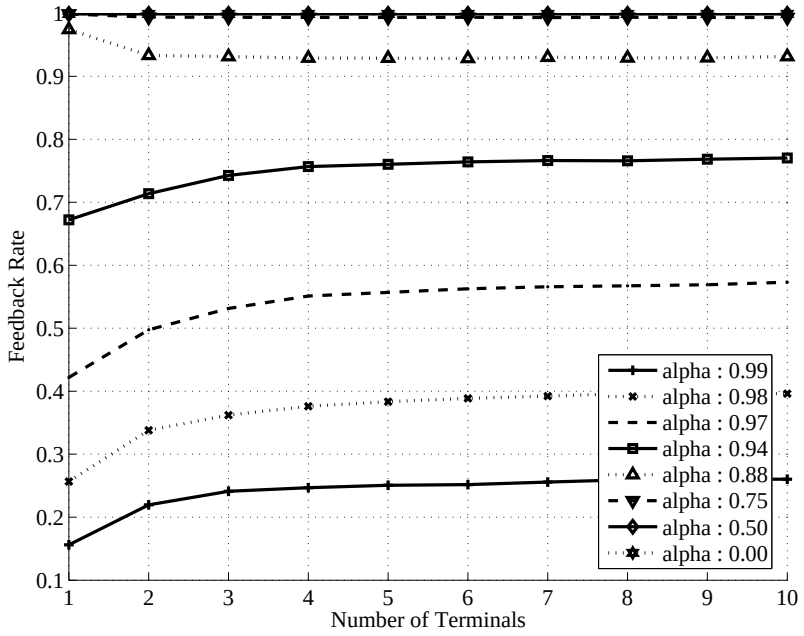


Figure 3.6: Reduction of feedback for various forgetting rates (α's)

0 first increases the throughput, reaches a peak and then further increasing α reduces the throughput. This optimal value of α depends on the channel coherence-time. The optimal value for our simulation setup is $\alpha \approx 0.88$. In next figures, we can see that increasing BLL increases the feedback rate, and for a fixed α, increasing BLL beyond some value does not have a significant effect on throughput. For a fixed α, older values are effectively forgotten, so increasing BLL only adds a few very small terms to the estimation procedure. These terms have very limited effect on the output of the estimation when they are small.

3.4 Summary

The performance of the channel estimator is especially important in sparse networks where random beamforming is highly unlikely to "illuminate" any terminal. Also in a mixture of mobile and fixed terminals, it is feasible to improve the channel especially for fixed

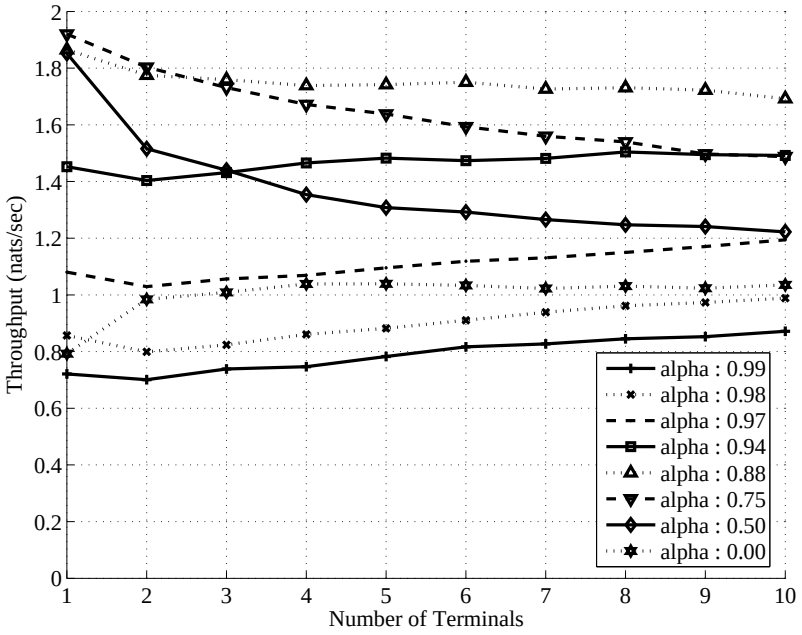


Figure 3.7: Total throughput for various forgetting rates (α's)

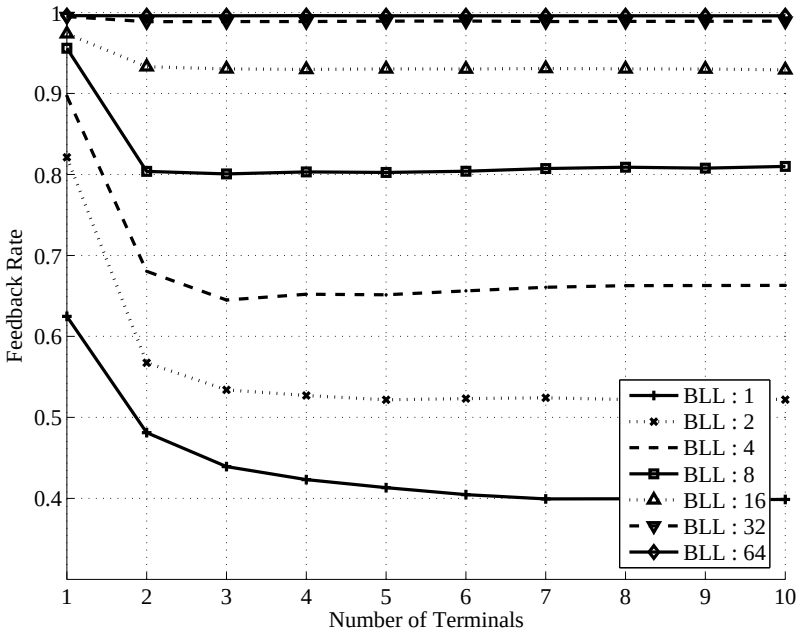


Figure 3.8: Reduction of feedback for various beam-list-lengths

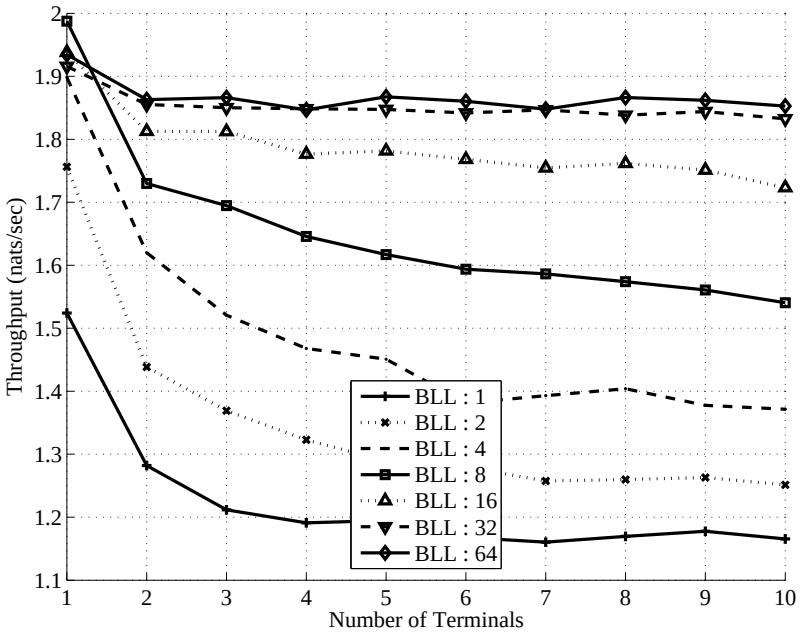


Figure 3.9: Total throughput for various beam-list-length

terminals. Our channel estimator gradually acquires the channel of fixed terminals (since the coherence time of their channels is expected to be higher than the channel coherence time of mobile terminals) while serving mobile terminals with better than a random beamformer. This is in contrast with schemes similar to the standard opportunistic beamforming, where the performance significantly degrades in sparse networks and there is no mechanism to improve the channel of fixed terminals. We have shown that the performance of our scheme is close to that of the coherent-beamforming scheme. When the coherence-time of the channel is sufficiently large, for instance for a fixed device, SNR feedback rate can be reduced accordingly, while the throughput performance remains better than other non-coherent schemes. Also since this scheme acquires the channel vectors, it does not waste resources on blind beamforming, resulting in a larger probability of producing high SNRs.

Chapter 4

Beamformer

In this chapter we use the estimation of the channel of the preselected terminal as the base to form the next beamforming vector. First, in § 4.1, we explain why we should not directly use the estimated channel as the beamforming vector. We also discuss why we used a randomizer to overcome measurement space shrinkage. Then, in § 4.2 we proceed and provide simulation results showing that the randomization mechanism solves the problem.

4.1 Maintaining Measurement Variety

The beamforming vector should provide high SNR at the selected terminal, but, this is not its only purpose. The beamforming vector also acts as a measurement for the other terminals. Therefore, the second objective of the beamformer is to maintain the "variety" of the set of recent beamforming vectors, i.e. the measurement space. The trivial mechanism which is to set the beamforming vector to be the conjugate of the estimated channel of the selected terminal will not achieve the second objective. The intuitive reason is that the estimation of the channel does not (and cannot) leave the space of recent beamforming vectors,

so the trivial mechanism does not allow the space of beamforming vectors to grow. On the other hand, the beamforming vectors will become closer to the actual channel vectors of the present terminals. If the channel vectors of the terminals do not cover the whole possible space sufficiently, the space of beamforming vectors will gradually shrink. Now, if a new terminal joins or the channels of the current terminals change, the estimator will no longer have measurements in the new directions. So it will find the best projection of the channel vectors of the terminals on the space of *recent previous beamforming vectors*. Therefore, the accuracy of the estimation reduces and the resulting SNR will gradually fade out. For the same reason, the expected channel acquisition time for a new terminal might become very long.

This phenomenon can also be explained in terms of $\frac{\lambda_{min}}{\lambda_{max}}$, the ratio of minimum and maximum eigenvalues of the matrix $B^{\mathrm{H}}B$, where

$$B = \Big[\mathbf{b}(t) \mid \mathbf{b}(t-1) \mid \cdots \mid \mathbf{b}(t-BLL+1)\Big]. \tag{4.1}$$

If we directly set the beamforming vector to the estimated channel from the previous block, and transmit using this beamforming vector, and receive nSNRs, and do the estimation for the next time-slot, we have formed a loop that reduces $\frac{\lambda_{min}}{\lambda_{max}}$ with every time-slot. When this ratio is small, there will be directions in the space where there will be practically no measurements. This means that the loop is blinded toward some directions, and can create poor MISO channel conditions for terminals whose channel vectors have large elements in the blinded directions.

In our proposed beamforming algorithm, We use the estimation of the channel, but we also employ the following two mechanisms to achieve the second objective, and prevent measurement space shrinkage.

1. **Mechanism I**: A small random term, orthogonal to the set of recent beamforming

vectors, with magnitude $\sqrt{p}$ is added to the estimated channel vector. So the beamforming vector is formed as

$$\mathbf{b}_{\text{next}} = \sqrt{1-p}\mathbf{g}_k + \sqrt{p}\mathbf{b}' \tag{4.2}$$

where $\mathbf{g}_k$ is the estimated channel vector obtained in the previous module, $\mathbf{b}'$ is a magnitude-normalized $\mathcal{CN}(0, I_{N_t})$ random vector[1] orthogonal to the set $\{\mathbf{b_1}, \mathbf{b_2}, ..., \mathbf{b_n}\}$. This mechanism is shown in Figure 4.1.

2. **Mechanism II**: Sometimes, with probability p, the estimation procedure is skipped, and instead, a random beamforming vector orthogonal to the set of recent beamforming vectors is employed as the current beamforming vector. This is illustrated in Figure 4.2.

In both mechanisms, p is a control parameter which reduces the expected nSNR of the current terminal compared to the nSNR we can get in this time-slot if we do not use randomization. However, this performance reduction is necessary in order to keep the expected channel acquisition time reasonably low, and keep the beamforming space sufficiently large. This parameter depends on the number of users in the system: if we increase the number of terminals, it becomes more unlikely that a beamforming vector "hits" none of the terminals, so it becomes less likely that this beamforming vector fades away.

In the case of $p = 1$ this scheme becomes standard opportunistic beamforming, since in that case we entirely bypass channel estimation and a random beamformer will be used independent of channel estimations.

Simulations agree with the theory and show that the performance will degrade if we do not use any randomization. The results are shown in next section, § 4.2.

[1]That is a random circularly normal complex vector divided by its magnitude.

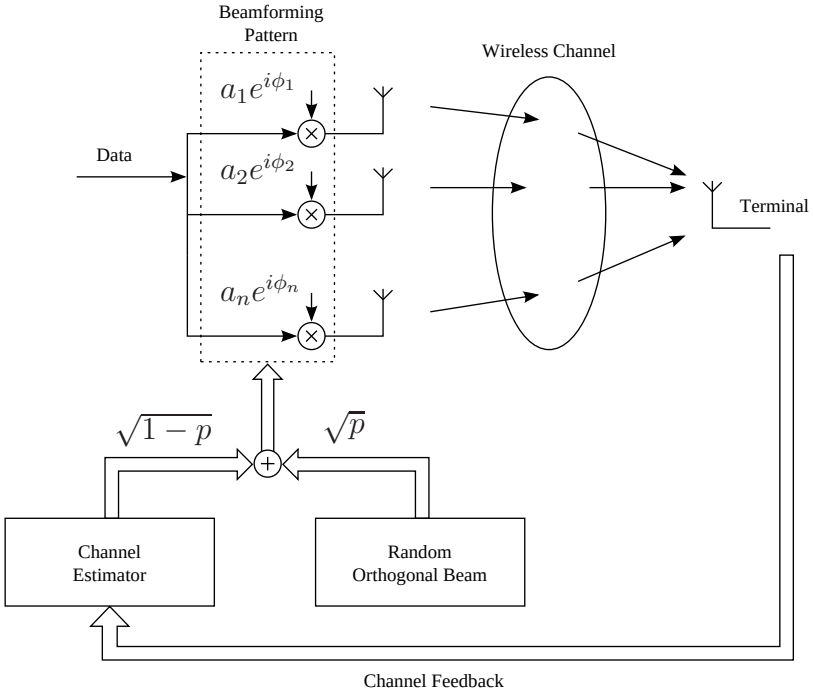


Figure 4.1: Block-diagram of a system with randomization mechanism I

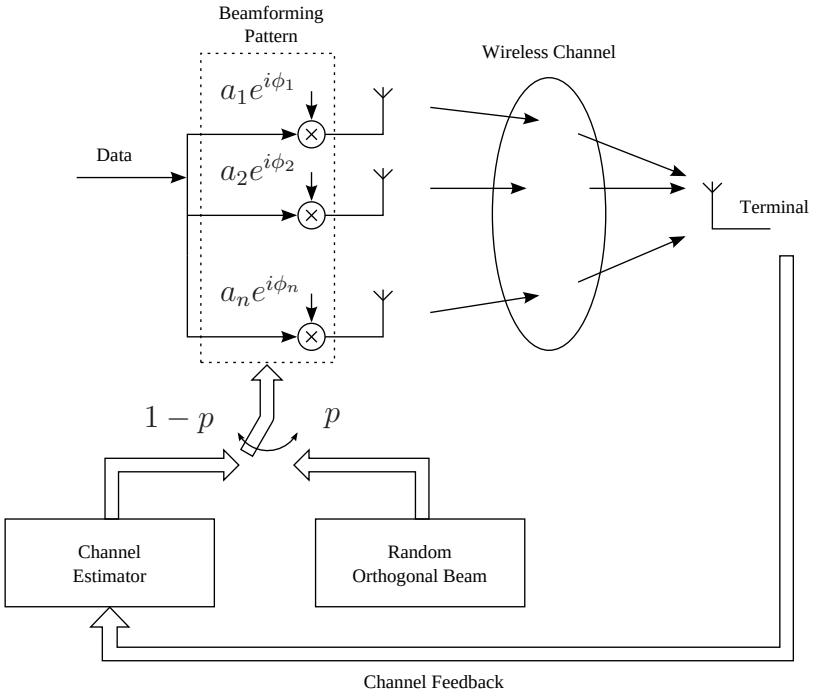


Figure 4.2: Block-diagram of a system with randomization mechanism II

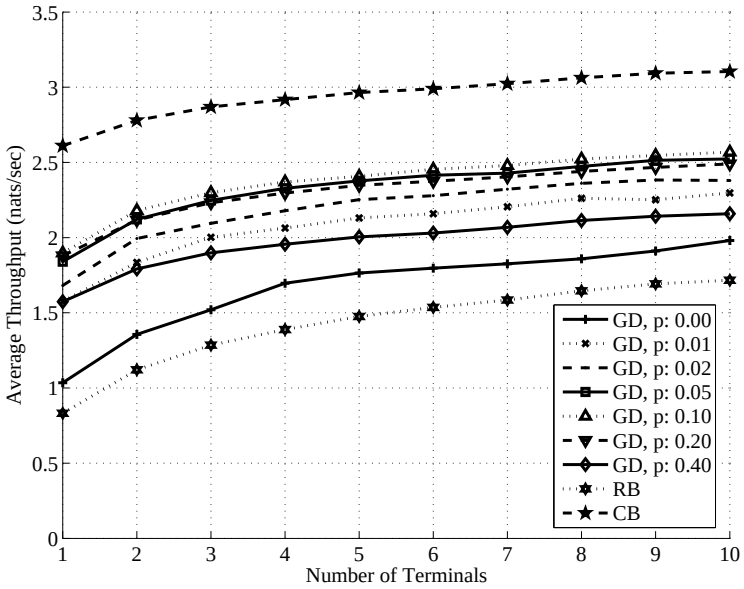


Figure 4.3: Beamforming with randomization mechanism I, for various p's

4.2 Simulation

For our simulations we use the same simulation setup as the one used in § 3.3. The performance of mechanism I is shown in Figure 4.3. The performance of mechanism II is shown in Figure 4.4. Here we use average throughput as the performance measure. As we can see, there is an optimal p for each case. For mechanism I, $p \approx 0.1$ has the best throughput performance. For mechanism II the optimal value is $p \approx 0.2$. The optimal value exists because increasing p will ultimately turn this scheme to the random beamforming scheme (standard opportunistic beamforming) so the performance will degrade. Also having $p = 0$ will have no randomization, thus, the performance of the system will degrade gradually in time. So there is an optimal value between 0 and 1. As shown, the first mechanism has a slightly better performance, so we have used the first mechanism for other simulations in the thesis.

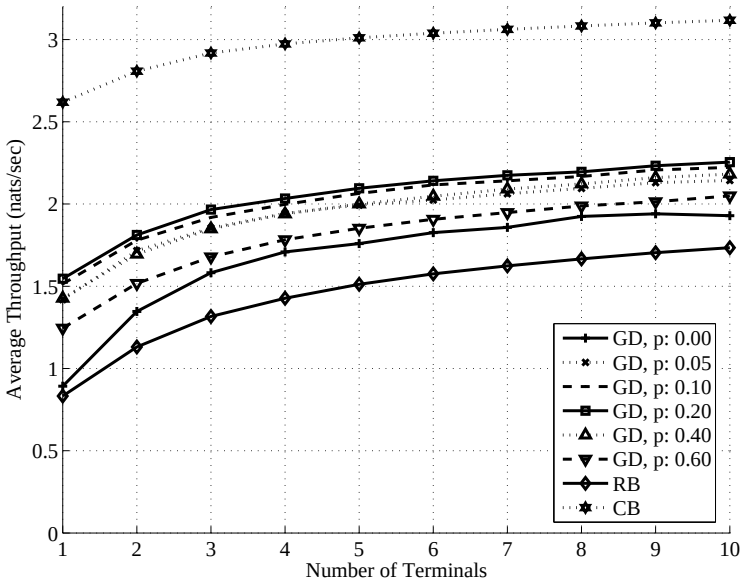


Figure 4.4: Beamforming with randomization mechanism II, for various p's

Chapter 5

Scheduler

In this chapter we employ various scheduling schemes in our system and compare the performance benchmarks. First we describe the scheduling schemes in § 5.1. Then, in § 5.2, we present the benchmarks and compare the schemes.

5.1 Scheduling Schemes

The scheduler in our scheme makes two decisions per time-slot. First it decides a terminal for which the optimal beamforming should be calculated based on the estimation of its channel. Then, after the terminals report their nSNR through the feedback channel (also known as DRC, the Data Request Channel), the scheduler decides the terminal who will be assigned for the next time slot. When the estimation of the channel is accurate, the information available at the scheduler after the SNR feedback reports will not differ much from the information before the reports. In this case, the first scheduling decision and the second scheduling decision will be the same. So we can run the same scheduling algorithm twice in a time-slot (once prior to channel feedback, and, the other one using

new updated nSNR reports). However, when a new terminal joins the channel, or when the channel of a terminal varies rapidly, the estimation procedure cannot be accurate. In this case the scheduling algorithm should prevent the starvation of this terminal. For example, the process of estimation may be replaced with a random orthogonal beamforming vector generation. The mechanism II in the beamformer controls the rate of introducing new random beamforming vectors rather than using the estimation of the channel. When a new terminal is accepted into the system, increasing p, the probability of introducing random beamforming vectors, improves the channel acquisition time for the new terminal at the expense of throughput loss in the other terminals. Controlling and balancing this parameter remains a higher level design parameter. In this thesis we use the same algorithm for the first and second scheduling decisions in a time-slot, and use a fixed p, assuming this parameter is controlled by a higher layer management entity. We assume that there is always backlogged data for each active terminal, and that there are K active terminals.

In this chapter we use various algorithms, namely MaxSNR, Round-Robin (RR), Proportional Fair Sharing (PFS), M-LWDF and Exponential Law (EXP) as our scheduling policies, and compare various benchmarks.

5.2 Benchmarks

In order to compare the performance of the scheduling policies, we employ various benchmarks. These benchmarks include “Throughput”, “Fairness Index” and “Worst Case Service Curve”.

5.2.1 Throughput Benchmark

For this benchmark we measure the average total throughput of the system. We also measure the throughput of terminals separately. This is shown in Figure 5.1. Average total throughput represents the bandwidth utilization by the whole system. It is desirable to increase this value, unless it drastically affects QoS for some terminals. MaxSNR scheduling policy is known to maximize the total throughput, while effectively being unable to support real-time traffic due to its unreliable QoS.

5.2.2 Fairness Benchmark

We use a Fairness Index according to

$$\text{FI} = \frac{(\sum_{k=1}^{K} U_k)^2}{K \sum_{k=1}^{K} U_k^2}, \tag{5.1}$$

where U_k is the average throughput of terminal k. Note that $0 < \text{FI} \leq 1$, and 1 happens when all terminals receive the same total amount of service. This fairness index is widely used in the literature, e.g., in a book by Jain [32]. We measure the fairness index in both homogeneous channel type where all terminals have the same average received SNR, and heterogeneous channel type where the terminals have different average received SNRs. For the heterogeneous case, we have set the terminal with best average SNR to have 10 times the SNR of the terminal with worst average SNR. This ratio is controlled by the "power control" module. Table 5.1 presents the fairness index for various scheduling policies.

Figure 5.1 shows how various scheduling policies distribute throughput among the terminals. We have used the same simulation setup as the ones used in § 3.3, with various scheduling methods. As we see, other than MaxSNR, the other policies have similar

Table 5.1: Fairness indexes for various scheduling policies

Scheduling Policy	Channel Type	Fairness Index
MaxSNR	Heterogeneous	0.1986
Round Robin	Heterogeneous	0.9921
Proportional Fair Scheduling	Heterogeneous	0.7445
Modified-LWDF	Heterogeneous	0.9572
EXP	Heterogeneous	0.9830
MaxSNR	Homogeneous	0.8044
Round Robin	Homogeneous	0.9996
Proportional Fair Scheduling	Homogeneous	0.9576
Modified-LWDF	Homogeneous	0.9995
EXP	Homogeneous	0.9996

throughput fairness indexes.

As shown, the MaxSNR scheme provides a higher total throughput at the expense of fairness. The exponential law behaves very similar to the round-robin scheme, and is almost as fair as the round-robin scheme.

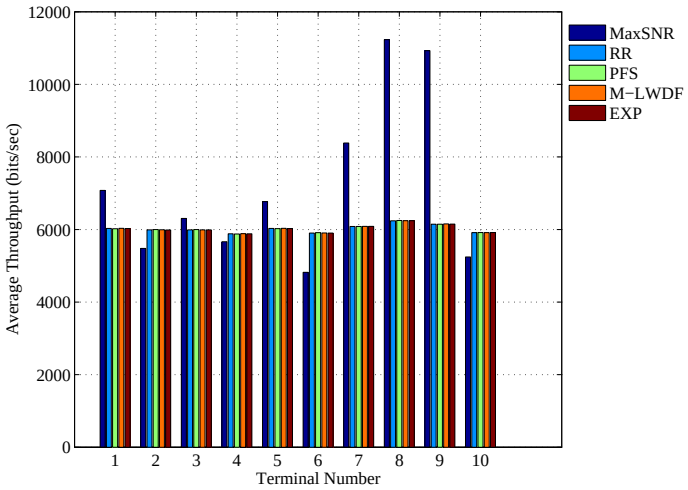
5.2.3 Worst Case Service Curve

Worst case service curve provides a pessimistic insight for QoS. Departure curve, $D(t)$, for a terminal shows the cumulative number of bits sent to this terminal, since some initial point in time. The initial point is often considered $t = 0$. Worst case service curve shows the least number of bits sent for a terminal over any interval with a given length. In other words, worst case service curve ($S(\tau)$) is

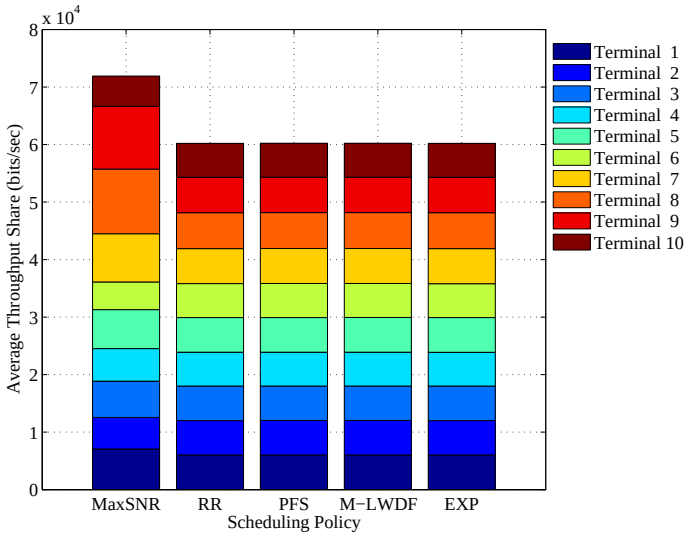
$$S(\tau) = \min_{t \geq 0} \{D(t + \tau) - D(t)\}. \tag{5.2}$$

Worst case service curve (also known as deterministic service curve) provides a service guarantee. We show this measure for our simulation setup[1] for various scheduling methods

[1]The same as the simulation setup in § 3.3.



(a) Grouped



(b) Stacked

Figure 5.1: Throughput distribution under various scheduling schemes

in Figure 5.2. As it is shown, round-robin scheme provides the best guarantee, and the performance of EXP scheme follows that of round-robin, closely. MaxSNR has a drastic guarantee which is zero in the depicted interval length (40 time-slots). Its service curve raises from zero at about 1000 time-slots. M-LWDF is slightly better than PFS here. The reason is that EXP exponentially raises the priority of a terminal after it has waited for some time. M-LWDF also has a similar mechanism (though, the priority increment is not exponential, so it might let the terminal wait longer than EXP). PFS increases the priority when a terminal has not been served in a time window. This increment is not as much as the one in EXP or M-LWDF schemes. Round-robin can provide the best *worst case* guarantee since it is turn-based and channel-independent.

Note that this guarantee is pessimistic and it is not a realistic measure of what happens most of the time. This can be useful when there are very rigid service requirements for an application. Most multimedia applications have a tolerance for service violation. For such applications less pessimistic measures are needed. For instance, a 95-percent service curve might be a better choice for voice streaming applications. A p-service curve ($S_p(\tau)$) means a guarantee that with probability p there will be at least $S_p(\tau)$ bits sent to the terminal in any interval of length τ. A 95-percent service curve is shown in Figure 5.3. Here, MaxSNR's service curve is still zero in the 40 time-slots interval, while other methods have a better performance. EXP scheme outperforms round-robin in some points. The performance of M-LWDF is closer to that of round-robin and EXP schemes in this measure.

5.3 Summary

As it is shown, the choice of scheduling scheme greatly affects the QoS measures. It is not possible to find an "optimal" scheduling policy, unless QoS requirements are accurately

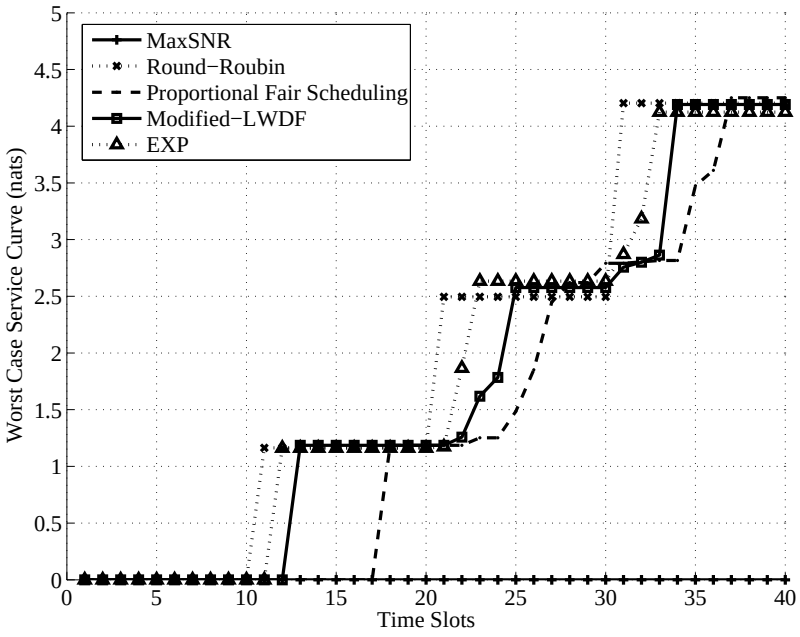


Figure 5.2: Worst case service curve for various scheduling methods, for $K = 10$ terminals.

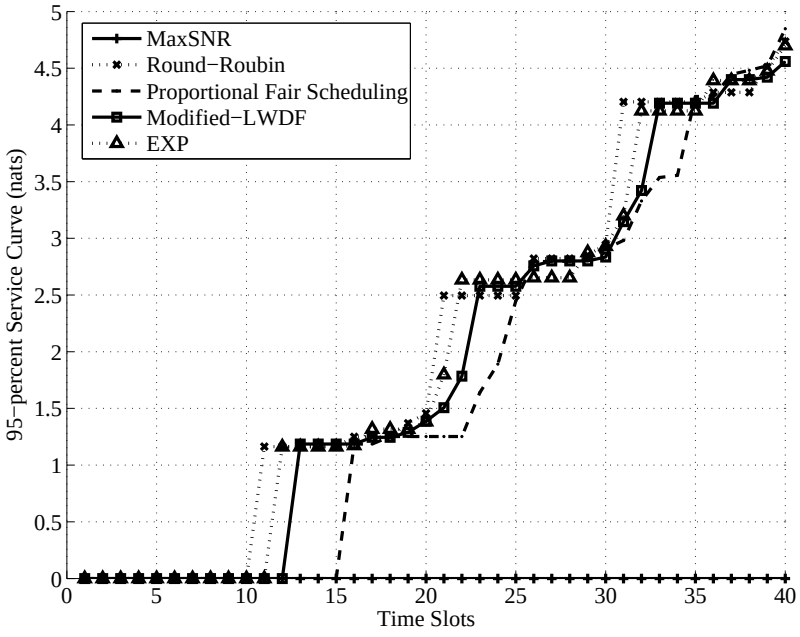


Figure 5.3: 95-precent service curve for various scheduling methods, for $K = 10$ terminals.

declared. However, it is often impractical to foresee the QoS requirements of an application even when the application might seem as trivial as downloading a specific video file. Therefore, practical parameters such as the simplicity of the required Admission Control (AC), the simplicity of the scheduling policy, and the comprehensibility of its analysis are more important in the choice of a scheduling discipline. Among the scheduling schemes for a "good" QoS requirement, e.g. for real-time traffic, we choose PFS with a proper time-window τ, or EXP depending on how hard the deadlines are, and how complex the admission control should be.

Chapter 6

Conclusion

6.1 Final Notes

In the thesis we have presented a cross-layer design based on the opportunistic beamforming scheme. We addressed the non-linear problem of combining beamforming vectors using only SNR measurement feedback. The usage of an opportunistic scheduler, however, allows us to still provide substantial benefits even though the algorithm that combines the beamforming vectors might not reach the global optimal point. We have also presented a two-stage scheduling policy in which the first stage suggests a terminal; the estimator and the beamformer form a beamforming vector for the selected terminal and apply it to an adjacent pilot channel. Then, all the terminals listen to this channel and when they perceive a high SNR they report to the base-station. The second stage scheduler assigns the next time slot to one of the terminals, based on this feedback and queue status. We have also devised a feedback reduction policy which can be incorporated to minimize the required feedback from terminals when their channel coherence time is not very short. Setting a proper forgetting rate for the channel coherence time for relatively fixed terminals allows us to use

the feedback channel efficiently. The proposed scheme fills the performance gap between standard opportunistic beamforming and coherent beamforming in sparse networks when the coherence time of the channel is long enough. Due to this property, this scheme can be a reasonable choice in realistic scenarios. A possible application for the proposed scheme is its usage in wireless mesh-networks. For fast fading channels this scheme performs slightly better than the standard opportunistic beamforming scheme, because the channel estimation procedure can still provide more channel gain than a fully random beamformer.

6.2 Future Work

- Channel Estimator: The proposed gradient-based channel estimation procedure might be improved. For example, turning the problem into a convex optimization problem might be possible. Although an analytical solution for this problem seems unlikely, it is still interesting to search for it. As it is shown in Appendix A, the case $n_0 = 2$ has an analytical solution, but it seems difficult to extend it to higher values of n_0.

- Beamforming: We employed a forgetting-rate to undermine the older measurements. The calculation of an optimal forgetting-rate for each terminal remains as a future work. This rate is related to the coherence-time of the channel for a terminal.

Bibliography

[1] S. Shakkottai, T.S. Rappaport, and P.C. Karlsson. Cross-layer design for wireless networks. *Communications Magazine, IEEE*, 41(10):74– 80, 2003.

[2] R. Knopp and P.A. Humblet. Information capacity and power control in single-cell multiuser communications. *Communications, 1995. ICC 95 Seattle, Gateway to Globalization, 1995 IEEE International Conference on*, 1:331–335 vol.1, 1995.

[3] David Tse and Pramod Viswanath. *Fundamentals of Wireless Communication*. Cambridge University Press, 1 edition, 2004.

[4] D. Tse. Forward link multiuser diversity through proportional fair scheduling, Aug 1999.

[5] A. Jalali, R. Padovani, and R. Pankaj. Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system. *Vehicular Technology Conference Proceedings, 2000. VTC 2000-Spring Tokyo. 2000 IEEE 51st*, 3:1854–1858 vol.3, 2000.

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

[7] M. Kountouris and D. Gesbert. Memory-based opportunistic multi-user beamforming. *Information Theory, 2005. ISIT 2005. Proceedings. International Symposium on*, pages 1426– 1430, 2005.

[8] N. Sharma and L.H. Ozarow. A study of opportunism for multiple-antenna systems. *Information Theory, IEEE Transactions on*, 51(5):1804– 1814, 2005.

[9] R. Laroia, Junyi Li, S. Rangan, and M. Srinivasan. Enhanced opportunistic beamforming. *Vehicular Technology Conference, 2003. VTC 2003-Fall. 2003 IEEE 58th*, 3:1762– 1766 Vol.3, 2003.

[10] Il-Min Kim, Seung-Chul Hong, S.S. Ghassemzadeh, and V. Tarokh. Optimum opportunistic beamforming based on multiple weighting vectors. *Communications, 2005. ICC 2005. 2005 IEEE International Conference on*, 4:2427– 2430 Vol. 4, 2005.

[11] M. Sharif and B. Hassibi. On the capacity of MIMO broadcast channels with partial side information. *Information Theory, IEEE Transactions on*, 51(2):506– 522, 2005.

[12] P. Svedman, S.K. Wilson, Jr. Cimini, L.J., and B. Ottersten. A simplified opportunistic feedback and scheduling scheme for OFDM. *Vehicular Technology Conference, 2004. VTC 2004-Spring. 2004 IEEE 59th*, 4:1878– 1882 Vol.4, 2004.

[13] P. Svedman, S.K. Wilson, and B. Ottersten. A qos-aware proportional fair scheduler for opportunistic OFDM. *Vehicular Technology Conference, 2004. VTC2004-Fall. 2004 IEEE 60th*, 1:558– 562 Vol. 1, 2004.

[14] D. Gesbert and M.-S. Alouini. How much feedback is multi-user diversity really worth? *Communications, 2004 IEEE International Conference on*, 1:234– 238, 2004.

[15] S. Sanayei and A. Nosratinia. Exploiting multiuser diversity with only 1-bit feedback. *Wireless Communications and Networking Conference, 2005 IEEE*, 2:978– 983 Vol. 2, 2005.

[16] S. Sanayei and A. Nosratinia. Opportunistic beamforming with limited feedback. *Signals, Systems and Computers, 2005. Conference Record of the Thirty-Ninth Asilomar Conference on*, pages 648– 652, 2005.

[17] D.J. Love, Jr. Heath, R.W., W. Santipach, and M.L. Honig. What is the value of limited feedback for MIMO channels? *Communications Magazine, IEEE*, 42(10):54–59, 2004.

[18] D. Avidor, J. Ling, and C. Papadias. Jointly opportunistic beamforming and scheduling (JOBS) for downlink packet access. *Communications, 2004 IEEE International Conference on*, 5:2959– 2964 Vol.5, 2004.

[19] Il-Min Kim, Zhihang Yi, Dongwoo Kim, and Wonsuk Chung. Improved opportunistic beamforming in ricean channels. *Communications, IEEE Transactions on*, 54:2199–2211, Dec 2006.

[20] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushyana, and S. Viterbi. CDMA/HDR: a bandwidth efficient high speed wireless data service for nomadic users. *Communications Magazine, IEEE*, 38(7):70–77, 2000.

[21] M. Andrews. Instability of the proportional fair scheduling algorithm for HDR. *Wireless Communications, IEEE Transactions on*, 3(5):1422– 1426, 2004.

[22] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, R. Vijayakumar, and P. Whiting. Cdma data qos scheduling on the forward link with variable channel conditions. *Bell Labs. preprint*, Apr 2000.

[23] S. Shakkottai and A.L. Stolyar. Scheduling algorithms for a mixture of real-time and non-real-time data in hdr. *Proceedings of International Teletraffic Congress (ITC)*, pages 793–804, 2001.

[24] A. Gyasi-Agyei. Multiuser diversity based opportunistic scheduling for wireless data networks. *Communications Letters, IEEE*, 9(7):670– 672, 2005.

[25] T. Bonald. Flow-level performance analysis of some opportunistic scheduling algorithms. *European Transactions on Telecommunications*, 16:65–75, 2005.

[26] A.Gyasi-Agyei. I-ocasd: A multiservice cross-layer scheduler for packetized wireless networks. *Submitted to IEEE Vehicular Technology Conference*, 2006.

[27] A.K. Parekh and R.G. Gallager. A generalized processor sharing approach to flow control in integrated services networks: the single-node case. *Networking, IEEE/ACM Transactions on*, 1(3):344–357, 1993.

[28] Niranjan Joshi, Srinivas R. Kadaba, Sarvar Patel, and Ganapathy S. Sundaram. Downlink scheduling in cdma data networks. In *MobiCom '00: Proceedings of the 6th annual international conference on Mobile computing and networking*, pages 179–190, New York, NY, USA, 2000. ACM Press.

[29] R. Schmidt. Multiple emitter location and signal parameter estimation. *Antennas and Propagation, IEEE Transactions on [legacy, pre - 1988]*, 34(3):276– 280, 1986.

[30] K. Bowman and L. Shenton. *Properties of Estimators for the Gamma Distribution*. Marcel Dekker Inc, 1 edition, 1988.

[31] K.E. Baddour and N.C. Beaulieu. Autoregressive modeling for fading channel simulation. *Wireless Communications, IEEE Transactions on*, 4(4):1650– 1662, 2005.

[32] R. Jain. *The Art of Computer Systems Performance Analysis*. John Wiley and Sons, New York, NY, 1991.

Appendix A

ML Gradient-Descent Algorithm for Adaptive Beamforming

Based on the model of the received signal (2.15) and measurements (2.17), we develop a maximum likelihood estimator to find the channel vector for a given terminal. For terminal k we have

$$\rho_k(t) - |\mathbf{b}^{\mathrm{H}}(t)\mathbf{h}_k(t)|^2 = \eta_k(t) \sim \mathcal{N}(0, \sigma_\eta^2). \tag{A.1}$$

Since there is no ambiguity we can safely drop the index k in this chapter. Assume ρ is the vector of size n (the number of samples) according to

$$\boldsymbol{\rho} = [\rho(1), \rho(2), ..., \rho(n)]^{\mathrm{T}}. \tag{A.2}$$

We choose $\hat{\mathbf{h}}$ as the estimated channel vector such that

$$\log f_{\mathbf{h}|\boldsymbol{\rho}}(\hat{\mathbf{h}}|\boldsymbol{\rho}) \geq \log f_{\mathbf{h}|\boldsymbol{\rho}}(\mathbf{h}'|\boldsymbol{\rho}) \quad \forall \mathbf{h}' \in \mathbb{C}^{N_t}, \tag{A.3}$$

so we have

$$\begin{aligned}&\left|\boldsymbol{\rho}-\left[|\hat{\mathbf{h}}^{\mathrm{H}}\mathbf{b}(1)|^2,|\hat{\mathbf{h}}^{\mathrm{H}}\mathbf{b}(2)|^2,...,|\hat{\mathbf{h}}^{\mathrm{H}}\mathbf{b}(n)|^2\right]^{\mathrm{T}}\right|^2\\&\quad\leq\left|\boldsymbol{\rho}-\left[|\mathbf{h}'^{\mathrm{H}}\mathbf{b}(1)|^2,|\mathbf{h}'^{\mathrm{H}}\mathbf{b}(2)|^2,...,|\mathbf{h}'^{\mathrm{H}}\mathbf{b}(n)|^2\right]^{\mathrm{T}}\right|^2.\end{aligned}\tag{A.4}$$

Thus

$$\begin{aligned}&\left|\rho(1)-|\hat{\mathbf{h}}^{\mathrm{H}}\mathbf{b}(1)|^2\right|^2+\left|\rho(2)-|\hat{\mathbf{h}}^{\mathrm{H}}\mathbf{b}(2)|^2\right|^2+...+\left|\rho(n)-|\hat{\mathbf{h}}^{\mathrm{H}}\mathbf{b}(n)|^2\right|^2\\&\leq\quad\left|\rho(1)-|\mathbf{h}'^{\mathrm{H}}\mathbf{b}(1)|^2\right|^2+\left|\rho(2)-|\mathbf{h}'^{\mathrm{H}}\mathbf{b}(2)|^2\right|^2+...+\left|\rho(n)-|\mathbf{h}'^{\mathrm{H}}\mathbf{b}(n)|^2\right|^2,\end{aligned}\tag{A.5}$$

which means that we need to find $\hat{\mathbf{h}}$ which minimizes ϵ according to

$$\epsilon(\hat{\mathbf{h}})=\sum_{i=1}^{n}\left(|\mathbf{b}^{\mathrm{H}}(i)\hat{\mathbf{h}}|^2-\rho(i)\right)^2.\tag{A.6}$$

For the case of $n=2$ one can proceed as follows. Assume $\hat{\mathbf{h}}$ is expressed as

$$\hat{\mathbf{h}}=x_1\mathbf{b}(1)+x_2\mathbf{b}(2)+\mathbf{h}_\perp,\quad x_1,x_2\in\mathbb{C},\tag{A.7}$$

where $\mathbf{h}_\perp$ is the perpendicular to both $\mathbf{b}(1)$ and $\mathbf{b}(2)$. We have

$$\epsilon=\left\{\left(|\mathbf{b}^{\mathrm{H}}(1)\hat{\mathbf{h}}|^2-\rho(1)\right)^2+\left(|\mathbf{b}^{\mathrm{H}}(2)\hat{\mathbf{h}}|^2-\rho(2)\right)^2\right\}\tag{A.8}$$

$$=\left\{\left(\left|x_1\|\mathbf{b}(1)\|^2+x_2\mathbf{b}(1)^{\mathrm{H}}\mathbf{b}(2)\right|^2-\rho(1)\right)^2+\left(\left|x_1\mathbf{b}^{\mathrm{H}}(2)\mathbf{b}(1)+x_2\|\mathbf{b}(2)\|^2\right|^2-\rho(2)\right)^2\right\}\tag{A.9}$$

So the problem can be expressed as follows:

$$\min \left\| \begin{bmatrix} \|\mathbf{b}(1)\|^4 & |\mathbf{b}(1)^{\mathrm{H}}\mathbf{b}(2)|^2 & 2\|\mathbf{b}(1)\|^2|\mathbf{b}(1)^{\mathrm{H}}\mathbf{b}(2)| \\ |\mathbf{b}(1)^{\mathrm{H}}\mathbf{b}(2)|^2 & \|\mathbf{b}(2)\|^4 & 2\|\mathbf{b}(2)\|^2|\mathbf{b}(1)^{\mathrm{H}}\mathbf{b}(2)| \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ Y_1 \end{bmatrix} - \boldsymbol{\rho} \right\|^2, \tag{A.10}$$

s.t.

$$Y_1^2 \leq X_1 X_2,$$

$$X_1, X_2 \geq 0,$$

where

$$X_1 = |x_1|^2, \tag{A.11}$$

$$X_2 = |x_2|^2, \tag{A.12}$$

$$Y_1 = \Re\left\{ \frac{\mathbf{b}^{\mathrm{H}}(1)\mathbf{b}(2)}{|\mathbf{b}^{\mathrm{H}}(1)\mathbf{b}(2)|} \bar{x_1} x_2 \right\}. \tag{A.13}$$

If $\mathbf{b}(1) = c\mathbf{b}(2)$ for $c \in \mathbb{C}$, then for the minimum of ϵ we have:

$$\min \epsilon = \rho(1)/\|\mathbf{b}(1)\|^4 + \rho(2)/\|\mathbf{b}(2)\|^4. \tag{A.14}$$

Otherwise we can find the inverse of $\begin{bmatrix} \|\mathbf{b}(1)\|^4 & |\mathbf{b}(1)^{\mathrm{H}}\mathbf{b}(2)|^2 \\ |\mathbf{b}(1)^{\mathrm{H}}\mathbf{b}(2)|^2 & \|\mathbf{b}(2)\|^4 \end{bmatrix}$, and a minimum of zero is obtained. In the case $n > 2$ this approach is not effective, so we adopt a gradient-based iterative procedure to find the minimum value for ϵ.

Gradient of ϵ is as follows:

$$\nabla \epsilon_{\hat{\mathbf{h}}} = \sum_{i=1}^{n} \mathbf{b}(i)\mathbf{b}^{\mathrm{H}}(i)\hat{\mathbf{h}} \left(\left| \mathbf{b}^{\mathrm{H}}(i)\hat{\mathbf{h}} \right|^2 - \rho(i) \right), \tag{A.15}$$

And its Hessian is

$$\nabla^2 \epsilon_{\hat{\mathbf{h}}} = \sum_{i=1}^{n} \left(2 \left| \mathbf{b}^{\mathrm{H}}(i) \hat{\mathbf{h}} \right|^2 - \rho(i) \right) \mathbf{b}(i) \mathbf{b}^{\mathrm{H}}(i). \tag{A.16}$$

Consider a vector $\hat{\mathbf{h}}$ with large $\|\hat{\mathbf{h}}\|$. In this case the $\rho(i)$ terms in ϵ become very small, so the negative gradient $-\nabla\epsilon$ points toward the origin. Hence a gradient-based algorithm does not diverge away to infinity, because when $\|\hat{\mathbf{h}}\|$ becomes large, the gradient moves it to the direction in which it gets smaller. The gradient-based approach we incorporate is described as follows:

1. Pick an initial vector $\mathbf{g}_0$.
2. Find the gradient $\nabla\epsilon$ for the current $\mathbf{g}_j$.
3. Subtract $\mu \nabla \epsilon_{\mathbf{g}_j}$ from $\mathbf{g}_j$ to obtain $\mathbf{g}_{j+1}$.
4. Repeat 2 and 3 unless i exceeds a constant MaxIter, or ϵ is below a fixed threshold.

In the above algorithm, μ is a "step-size" parameter and can be fixed to a small number (we used fixed $\mu = 0.02$) or it can be set adaptively. Figure A.1 shows the convergence of the algorithm in the following setting. Here, $\mathbf{b}(t)$'s are drawn randomly (from a magnitude normalized circularly normal set), and we have set the number of previous measurements incorporated in an estimation $n = 32$.

Although this procedure does not diverge, it cannot be guaranteed to converge and find the global minimum point. In fact, it is possible to propose a set of $\mathbf{b}(i)$'s along with corresponding $\rho(i)$'s such that there are many local minimum points. Our procedure does not have a guarantee not to fall into a local minimum in one run. However, if we run it in the whole system, the $\rho(i)$'s will be relatively large, because, according to the feedback reduction policy, the terminals report their SNR when they perceive a significant SNR. If

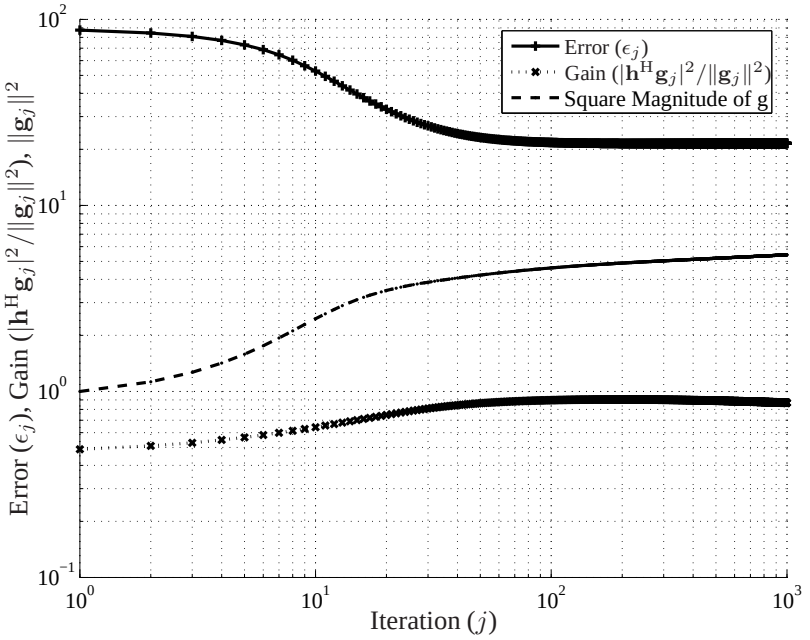


Figure A.1: Convergence and gain of the gradient descent procedure

the $\rho_k(t)$'s are large enough and their error ($\eta_k(t)$) is not large, and we start the procedure from the $\mathbf{b}_k(t)$ corresponding to the largest $\rho_k(t)$, there will be a single minimum. Figure A.1 shows the convergence rate and the gain at each iteration for the gradient descent procedure.

Finally, note that as there are other measurements in other directions (when the channel is assigned to other terminals, and also when a random beamforming vector is drawn), we expect the $\rho_k(t)$'s to become large enough (after a few time-slots), and the procedure *statistically* converges.