An Overview of Limited Feedback in Wireless Communication Systems

David J. Love, Member, IEEE, Robert W. Heath Jr, Senior Member, IEEE, Vincent K. N. Lau, Senior Member, IEEE, David Gesbert, Senior Member, IEEE, Bhaskar D. Rao, Fellow, IEEE, Matthew Andrews, Member, IEEE

(I Invited Tutorial Paper)

Abstract—It is now well known that employing channel adaptive signaling in wireless communication systems can yield large improvements in almost any performance metric. Unfortunately, many kinds of channel adaptive techniques have been deemed impractical in the past because of the problem of obtaining channel knowledge at the transmitter. The transmitter in many systems (such as those using frequency division duplexing) cannot leverage techniques such as training to obtain channel state information. Over the last few years, research has repeatedly shown that allowing the receiver to send a small number of information bits about the channel conditions to the transmitter can allow near optimal channel adaptation. These practical systems, which are commonly referred to as limited or finite-rate feedback systems, supply benefits nearly identical to unrealizable perfect transmitter channel knowledge systems when they are judiciously designed. In this tutorial, we provide a broad look at the field of limited feedback wireless communications. We review work in systems using various combinations of single antenna, multiple antenna, narrowband, broadband, single-user, and multiuser technology. We also provide a synopsis of the role of limited feedback in the standardization of next generation wireless systems.

Index Terms—Wireless communications, Limited feedback, MIMO systems, Quantized precoding, Multiuser MIMO systems techniques. These sorts of signaling approaches allow the transmitter to adapt to the propagation conditions. This implies that the transmitter requires some form of knowledge of the wireless channel conditions, often referred to as channel state information (CSI) at the transmitter (CSIT). Employing most kinds of channel adaptive techniques has been impossible in the past because two-way communication is accomplished using frequency division duplexing (FDD). The forward and reverse links in FDD generally have highly uncorrelated channels because they are separated in frequency.

One way of overcoming this problem is by using other forms of reciprocity (e.g., statistical reciprocity). These sorts of systems use the fact that the forward and reverse links often share the same fading distribution. Statistical approaches can perform very well in situations where the channel exhibits some form of (slowly varying) structure, such as having a large mean component (i.e., a large Rician K-factor) or strong correlation (either in space, time, or frequency). Generally, however, statistical adaptation comes with a non-negligible performance loss compared with adaptation techniques that use the instantaneous channel realization.

The big innovation that has overcome the challenge of making instantaneous channel adaptation practical is the use of feedback. A system employing feedback uses a low rate data stream on the reverse side of the link to provide information to the transmitter of the forward side of the link. This information conveys some notion of the forward link condition (e.g., channel state, received power, interference level, etc.), and the transmitter uses the information to adapt forward link transmission. The value of feedback varies with the system scenario. However, generally speaking, the value is greater when the channel introduces some form of disturbance (such as spatial interference, intersymbol interference, multiuser interference, etc.) that cannot be handled by the receiver alone. The feedback information itself can be digital or analog. In this tutorial, we concentrate on digital feedback, which is commonly referred to as limited feedback or finite-rate feedback.

The history of feedback in communication systems traces back to Shannon [233] and other early work such as [76], [229], [230], [259], [260]. Interest has continued to grow in uses of feedback. Feedback has had broad impact in areas such as control systems, source coding, information theory, and communication theory. We concentrate and summarize the
present state of research into applications of limited feedback in wireless communication systems, where its interest has recently seen much revival, particularly in relation with multiple-input multiple-output (MIMO) systems. Our goal is to examine what has been accomplished and make some comments on the direction of this area of research.

We will divide the work into two main areas: single-user (see Section II) and multiuser communication (see Section III). Because the true measure of the impact of research is into the applications it generates, we look at the role of limited feedback in current and future standardized wireless systems in Section IV. We provide some concluding remarks in Section V.

Throughout the paper we use some common notation. The complex numbers are denoted by $\mathbb{C}$. The transpose of a vector is denoted by a superscript $^T$, and the conjugate transpose by a superscript $^*$. A diagonal matrix is created from a vector with the function $\text{diag}(\cdot)$. The two-norm of a vector (or matrix) is represented by $\| \cdot \|_2$, and the Frobenius norm of a matrix is represented by $\| \cdot \|_F$. The ceiling function is written as $\lceil \cdot \rceil$, and the floor function is similarly written as $\lfloor \cdot \rfloor$. The base two logarithm is written $\log_2(\cdot)$. The determinant of a matrix is evaluated with $\det(\cdot)$.

II. FEEDBACK IN SINGLE-USER WIRELESS SYSTEMS

The design of single-user wireless systems has a long and storied history. We address the role of limited feedback in single and multiple antenna systems.

A. Single Antenna Systems

Single antenna wireless links are the most commonly found wireless links. Single-user wireless systems are often split into the categories of narrow and broadband depending on the relationship between the bandwidth and delay spread of the propagation channel. For this reason, the benefits of channel adaptation using limited feedback will be divided into narrowband and broadband systems.

1) Narrowband Systems: The $k$th channel use of a narrow-band system is mathematically modeled as

$$y[k] = h[k]x[k] + n[k].$$

where $y[k]$ is a complex received symbol, $h[k]$ is the complex channel response, $x[k]$ is the transmitted symbol, and $n[k]$ is noise distributed according to $\mathcal{CN}(0,1)$ (assuming the noise is normalized to unit variance). The transmitted signal $x[k]$ is subject to a long term power constraint where $\mathbb{E}_{h,x}[|x[k]|^2] \leq \rho$. To allow the receiver to perform coherent detection, channel estimation techniques are usually performed. Most of the work on limited feedback assumes that the receiver has perfect knowledge of the $h[k]$ for all $k$. We will note when discussing work that makes other assumptions. Additionally, various ergodicity and stationarity assumptions must hold for the process $\{ h[k] \}$, but these are beyond the scope of this paper.

Because our focus is on adapting the transmitted signal to the channel conditions, modeling how the channel varies across a codeword block is critical. We primarily focus on a block-fading channel model, where the channel is constant for several channel uses before changing independently. Therefore, the $l$th channel block satisfies $h[tK_{ch} + l] = h[tK_{ch} + l + 1] = \cdots = h[(t + 1)K_{ch} - 1] = h(t)$ where $K_{ch}$ is the length of the fading block. The transmitted data will also have a block structure. Let $K_{bl}$ denote the codeword block length. We refer to the vector $[x[0] \ x[1] \ \cdots \ x[K_{bl} - 1]]$ as the transmitted codeword. The relationship between the channel block length $K_{ch}$ and the codeword block length $K_{bl}$ is important. In this tutorial, we will refer to the case when $K_{ch} = K_{bl}$ as the slow-fading scenario and the case when $K_{ch} \to 0$ when $K_{bl} \to \infty$ as the fast-fading scenario. More discussion on the relation between codeword block length and time variation of the fading process is available in [30] and the references therein.

Depending on the time evolution properties of the channel, both power and/or rate control provide benefits. For the $l$th codeword block, denote the average power constraint as $\mathbb{E}_y[|x[k]|^2 | h[k] = h(t)] \leq \rho_l$ where the expectation is over all possible codewords. To satisfy the long-term power constraint, we have to require that $\mathbb{E}_x[|p_k|] \leq \rho$. If the transmitter has knowledge of the channel conditions for each channel block, $\rho_l$ could be adaptively chosen to maximize performance. Variable rate encoding is also very common. In this kind of framework, the rate is varied according to the instantaneous channel conditions.

Assuming perfect knowledge of the magnitude of the channel, the ergodic capacity is [33], [69]

$$R = \mathbb{E}_h \left[ \log_2 \left( 1 + \rho(h) |h|^2 \right) \right]$$

where $\rho(h)$ is a function that allocates power subject to water-filling. Interestingly, this rate can be achieved asymptotically with fixed rate codeword sets. For the fast-fading case, we can construct the codewords as

$$x[k] = \sqrt{\rho(h[k])} s[k]$$

where $\{s[k]\}_{k=0}^{K_{bl}-1}$ is a codeword designed independently of the channel conditions (but whose rate is determined using distribution information) such that $\mathbb{E}[|s[k]|^2] \leq 1$ and $\rho(h[k])$ is chosen according to the water-filling algorithm.

The problem with capacity achieving power allocation frameworks is that they require the transmitter to perfectly know $h[k]$ (or at least its magnitude). As mentioned earlier, in systems such as those using FDD, this knowledge is not available. For this reason, the solution is for the receiver to utilize the reverse link as a feedback channel, send channel state information on this channel, and give the transmitter some kind of side information $u[k]$ about the current channel realization $h[k]$. This is generally shown in Figure 1. The receiver can obtain some level of channel information using techniques such as training. Using this knowledge, the receiver can design feedback to be sent as overhead on the reverse link.

The problem of codeword design with side information was brought up in [31]. This paper considers more general channel models than just (1), without restriction to block fading. In addition, [31] does not require the receiver to perfectly know $h[k]$ but instead assumes the receiver has access to some side information $w[k]$. Thus, the problem becomes one of encoding and decoding using this side information along with
knowledge of the joint probability density function $p(h, u, w)$. The interesting innovation in this paper is the observation that capacity of these systems with side information can be achieved with multiple codebooks properly multiplexed together.

This work was later extended to the fast-fading case (through a block-fading construction) in [138] adding the additional requirement of a cardinality constraint on the side information $u[k]$. The problem of properly designing the side information $u[k]$ is shown to be one of scalar quantization that can be solved using the Lloyd algorithm. The fast-fading assumption employed in this paper allows the codeword rate to be fixed because a codeword block spans a large number of channel realizations.

For the case of a fast-fading block channel model and perfect receiver channel knowledge, the multiplexed coding approach has later been extended and enhanced in [122] when the transmitter is provided with a quantized version of the magnitude of $h[k]$. This quantized version is taken by dividing up the non-negative part of the real line into quantization regions. This quantization approach is similar to techniques used in the temporal waterfilling proof in [69], which took the limit as the quantization noise goes to zero. In [122], the power allocation strategy then uses the quantized channel realization subject to either a short-term power constraint (where $p_t \leq \rho$ for any channel block $t$) or a long-term power constraint (where the power allocated to the $t$th channel block $\rho_t$ is restricted in expected value to be bounded by $\rho$). An overview of the possible power constraints is available in [30].

A model other than block fading was discussed in [212]. This work assumed periodic feedback, where feedback is sent every fixed number of channel uses. The channel model considered was a finite-state first-order Markov model.

From a practical perspective, another approach to the problem of adapting to the channel conditions is to concentrate on selecting from a fixed set of per channel use constellations and varying the density (or equivalently the average energy) of these constellations. On-off rate adaptation was proposed in [22], where the transmission was turned on and off subject to the channel conditions. A more general system where the rate of the transmitter is adjusted based on the channel is addressed in [34]. Here the effect on the probability of error subject to an average rate constraint is analyzed. These ideas were later extended to take into account queue length [35]. Various other work has looked at the application of rate variation [7], [30], [114], [189], [240], [241], [250], some using specific constellation families and some combining the rate variation with adaptive power allocation. Analysis of adaptive modulation with feedback imperfections has been studied in [57], [190]. Discussion can also be found in the overview paper [58]. A diversity-based approach is given in [236].

Work taking practical code designs into account has been relatively limited. Adaptive $M$-ary orthogonal coding for high bandwidth expansion systems (such as CDMA) has been proposed in [137], and adaptive trellis coded modulation for high bandwidth efficiency has been studied in [6], [67], [135], [136], [184]. These works consider joint optimization of the coding rate and modulation level coding based on maintaining a target average error rate or average throughput requirement. Outdated knowledge of channel state information has been considered. In addition to the performance benefit associated with adaptive coded modulation systems, there is another important benefit of channel state knowledge at the transmitter. In [181], the authors studied the concept of incorporating knowledge of channel side information at the transmitter on the LDPC code design. It is shown that substantial reduction of LDPC decoding complexity can be obtained utilizing the side information.

Another approach to feedback is the use of repeat requests when channel conditions cause codeword errors. In fact, regardless of the availability of explicit CSIT, there is always ACK/NAK signaling exchange in the upper layers in most communication systems. Such ACK/NAK exchange is used for automatic repeat request (ARQ) in the upper layers so that an error-free logical channel can be presented to the application layers. In fact, the ACK/NAK signaling exchange can also be utilized at the physical layer of the transmitter to learn about the actual channel conditions. This information is particularly useful when the CSIT (through explicit feedback [FDD] or implicit feedback [TDD]) is not perfect.

Consider the case when the channel state information obtained by limited feedback (or finite-rate feedback) may be outdated or suffering from feedback errors. Because of these errors, the transmitter must adapt the transmit power and/or data rate according to this imperfect CSIT. In order to effectively exploit the imperfect channel information at the transmitter, it is important to take into account the error statistics of the CSIT in the adaptation. However, it is very difficult for the transmitter to obtain and keep track of the error statistics because they usually depend on the channel environment and Doppler spectrum. In such cases, the ACK/NAK signaling from the upper layer ARQ is very useful to provide a truly closed-loop adaptation. For example, if the transmitter is over aggressive in the adaptation (e.g., in adjusting the data rate), the packet will be corrupted at the receiver and a NAK will result. Based on the NAK information, the transmitter can...
reduce the data rate and/or increase the transmit power until an ACK is received. Such an approach is very robust to CSIT errors and does not require explicit knowledge of CSIT error statistics at the transmitter. In fact, this closed-loop adaptation framework has been commercially deployed in IS95 in outer-loop power control.

Selective repeat ARQ is studied in [14]. ARQ schemes with reliable and unreliable feedback are studied in [13]. Power and rate adaptation utilizing ACK/NACK feedback has appeared in [73], [95], [281]. In [108], the authors considered a two level stochastic scheduling based on learning automata. In [266], the authors modeled the power, rate adaptation (as well as user selection) using Markov Decision Process (MDP) and obtained optimal as well as low complexity control policy. From these works, it is found that robust performance can be obtained by jointly considering both limited CSIT feedback as well as ACK/NACK signaling in the design of transmitter adaptation policy.

2) Broadband and Wideband Systems: A single antenna broadband model is complicated by the fact that previously transmitted symbols interfere with the current symbols. A discrete-time model for this kind of set-up is

\[ y[k] = \sum_{\ell=0}^{L} h[k, \ell] x[k - \ell] + n[k]. \]  

(4)

where the channel is now frequency selective and represented by an \((L + 1)\)-tap finite impulse response filter \([h[k, 0] \ldots h[k, L]]\) at the \(k\)th channel use.

The work in [31] derives a capacity formula for the case when the transmitter and receiver have access to some side information under the assumption of perfect receiver channel knowledge and a condition that implies that the transmitter obtains all information about the current channel conditions using only its current feedback (i.e., it can not gain extra knowledge from past feedback information).

Because of the difficulty in dealing with the intersymbol interference resulting from frequency selective channels, especially for recently standardized wideband systems (UMTS-LTE, WiMax, WiFi), industry and academia have turned toward the use of orthogonal frequency division multiplexing (OFDM). In OFDM, the signal \(x[k]\) is jointly designed over \(K_{sc} + L\) channel uses assuming that the channel is constant during a block of \(K_{ch}\) channel uses with \(K_{ch} \geq K_{sc} + L\). The transmitter constructs a \(K_{sc}\) collection of parallel subchannels in the frequency domain. The \(\ell\)th transmission across the parallel subchannels can be written \(\tilde{x} = [x_0[k] \cdots x_{K_{sc}-1}[k]]^T\). This vector is then multiplied by an inverse discrete Fourier transform (DFT) matrix, and the last \(L\) entries of the transformed signal are appended to the beginning of the vector (termed a cyclic prefix).

After reception, the receiver removes this cyclic prefix and multiplies the signal by a DFT matrix. This then gives a post-processing input-output relation in the frequency domain of

\[ \tilde{y}[k] = \text{diag} \left( \tilde{h}[k] \right) \tilde{x}[k] + \tilde{n}[k]. \]  

(5)

at OFDM channel use \(\tilde{k}\). Here vector notation has been used where the \(v\)th entry of each vector corresponds to the input-output relation for the \(v\)th subcarrier.

Adapting the subcarrier powers with limited feedback has been the focus of several works. Using a one bit per subcarrier (or per block of subcarriers) design that simply turns subchannels off and on was proposed by [141]. Later work on quantized feedback in OFDM to activate or deactivate subchannels was the focus of [246], [247]. More general schemes for jointly quantizing the per subcarrier power allocations have been discussed in [161], [164], [209]. Techniques used to address the problem of adaptation with unquantized (but stale or imperfect) CSIT studied in [273] can also be employed. The case of using feedback for bit interleaved coded OFDM was addressed in [249]. An overview of adaptive modulation with OFDM is available in [215].

Besides needing power allocation to achieve optimal performance, a challenge with OFDM is the large number of channel coefficients required when training is done only in the frequency domain. The receiver will require knowledge about the channel conditions for each of possibly thousands of subcarriers. A novel use of limited feedback is for the receiver to feedback previously detected symbols to decrease the amount of training needed in OFDM [51].

With the emergence of systems such as ultra-wideband (UWB) there has been an increased interest in adaptive signaling over very large bandwidths (often on the order of \(10^9\) Hertz). One possible approach to signaling in these systems is to send a narrowband signal over an adaptively chosen frequency band. When a narrowband channel is chosen by probing over a wideband channel, feedback allows the transmitter to choose a frequency band with good performance (generally defined as having a large SINR). The low SNR scaling of the maximum achievable rate is the focus of [26]. Training a wideband channel with feedback to optimize rate is discussed in [1]. Extending feedback analysis to wideband channels that are sparse in the delay and Doppler domains is considered in [74].

B. Multiple Antenna Systems

The application of limited feedback to multiple antenna wireless systems has received much attention in the recent past. The spatial degree-of-freedom and the potentially sizable benefits available by adapting over it make limited feedback a very attractive option.

The degrees of freedom with multiple antenna systems can be exploited to offer rate and diversity benefits as well as beamforming and interference canceling capabilities. While the diversity gain can be typically extracted without the need of CSIT feedback (e.g., space time codes), CSIT plays a crucial role for beamforming and interference mitigation at the transmitter side, as will be clarified below.

1) Narrowband Systems: A single-user narrowband multiple antenna system can be represented by an expression of the form

\[ y[k] = H[k]x[k] + n[k] \]  

(6)

at the \(k\)th channel use. Assuming \(M_t\) transmit antennas and \(M_r\) receive antennas, \(y[k]\) is an \(M_r\)-dimensional receive
vector, $\mathbf{H}[k]$ is an $M_r \times M_t$ channel response matrix, $\mathbf{x}[k]$ is an $M_r$-dimensional transmit vector, and $\mathbf{n}[k]$ is an $M_r$-dimensional noise. We assume the noise to have i.i.d. normalized entries distributed according to $\mathcal{C}\mathcal{N}(0,1)$. The transmitter power constraint requires that $\mathbb{E}_{\mathbf{x}, \mathbf{n}} \left[|\mathbf{x}[k]|^2\right] \leq \rho$. As in the single antenna case, we concentrate on the scenario where the receiver has access to $\mathbf{H}[k]$. Given this, there are a variety of ways to design $\mathbf{x}[k]$ if the transmitter is given access to some quantized information relating to $\mathbf{H}[k]$.

Again, this analysis will depend on the time evolution model of the channel. If we use our previous notation of block-fading, the $t-th$ channel block satisfies $\mathbf{H}[tK_{ch}] = \mathbf{H}[tK_{ch} + 1] = \cdots = \mathbf{H}[(t+1)K_{ch} - 1] = \mathbf{H}(t)$ where $K_{ch}$ is the length of the fading block. For power constraint reasons, $\mathbb{E}_x \left[|\mathbf{x}[k]|^2 | \mathbf{H}[k] = \mathbf{H}(t)\right] \leq \rho_t$ for the $t-th$ block. Varying $\rho_t$ to perform temporal water-filling provides capacity benefits, but unless otherwise noted, our discussion assumes $\rho_t = \rho$ for all channel blocks.

1a) Covariance Quantization

When the transmitter and receiver both perfectly know the channel, the ergodic capacity is [68], [256]

$$R = \mathbb{E}_\mathbf{H} \left[ \max_{Q_{tr}(Q) \subseteq 1, Q \geq 0} \log_2 \det \left( \mathbf{I} + \rho \mathbf{H} Q \mathbf{H}^H \right) \right].$$

(7)

Here $Q$ is the covariance of the transmitted signal for each individual instantaneous channel realization. The covariance of the transmitted signal could incorporate both the spatial power allocation as well as unitary precoding. Note that spatial power allocation is important especially for cases when the number of transmit and receive antennas are equal. From an encoding point of view, $\mathbf{x}[k] = \sqrt{\rho} Q[k]^{1/2} \mathbf{s}[k]$, $k = 0, \ldots, K_{ch} - 1$, where $Q[k]$ solves the optimization (based on channel feedback)

$$Q[k] = \arg \max_{Q_{tr}(Q) \subseteq 1, Q \geq 0} \log_2 \det \left( \mathbf{I} + \rho \mathbf{H}[k] Q \mathbf{H}^H [k] \right)$$

and $s[k]$ is the $k-th$ channel use of an open-loop codeword. This codeword set is chosen according to some spatial power constraint criteria such that $\mathbb{E}_Q [s[k] (s[k])^H] = \mathbf{I}$ and such that the encoding rate per channel block approaches the achievable rate of the instantaneous channel. For fast-fading, a fixed rate codeword set can be used satisfying similar conditions to those above but with a fixed encoding rate.

One of the first looks at trying to design the covariance matrix using imperfect channel information was the covariance design for multiple-input single-output (MISO) systems using statistical information published in [262]. For a limited rate feedback approach, the general idea is to use the fact that the receiver knows $\mathbf{H}[k]$ through procedures such as training. Using this channel knowledge, the receiver can quantize some function of $\mathbf{H}[k]$ using vector quantization (VQ) techniques.

Naturally, the aspects of the channel that the transmitter cares about are those that allow the design of the covariance for the $t-th$ channel block [237]. Using this line of reasoning, the receiver can determine a rate maximizing covariance and feed this back to the transmitter. Employing a codebook of possible covariance matrices $Q = \{Q_1, \ldots, Q_{2^B}\}$ that is known to both the transmitter and receiver, the receiver can search for the codebook index that solves

$$n_{\text{opt}}[k] = \arg \max_{1 \leq n \leq 2^B} \log_2 \det \left( \mathbf{I} + \rho \mathbf{H}[k] Q_n \mathbf{H}^H [k] \right)$$

and send the $B$-bit binary label corresponding to covariance $Q_{n_{\text{opt}}[k]}$ to the transmitter. This gives a maximum achievable rate in bits per channel use of

$$R_Q = \mathbb{E}_\mathbf{H} \left[ \max_{Q \in Q} \log_2 \det \left( \mathbf{I} + \rho \mathbf{H} Q \mathbf{H}^H \right) \right]$$

(8)

using a codebook $Q$ known to both the transmitter and receiver.

The covariance codebook can be either fixed or randomly generated (using a seed known to both the transmitter and receiver). Designing a fixed covariance codebook to maximize the average rate is a challenging problem that depends on the stationary distribution of the channel [24], [134]. Vector quantization approaches using the Lloyd algorithm have been shown to efficiently generate codebooks that achieve a large rate [134]. Random approaches for covariance design have also been proposed [45] using ideas pioneered in [222]. In fact, it was shown in [45] that the rate loss with $B$ bits of feedback decreases exponentially with the number of feedback bits.

While the codebook approach is optimal for a block-to-block independently fading channel, temporal correlation between channel realizations can improve quantization. Feedback approaches based on tracking the channel using gradient analysis are studied in [18], [19]. The use of switched codebooks, where the codebook is changed or adapted over time is proposed in [170]. Beamforming codebooks with adaptive localized codebook caps, the orientation and radius of the cap changing over time, was considered in [213]. Markov models to analyze the effects of feedback delay and channel time evolution were proposed in [91]–[93]. These models can be used to implement feedback compression by using Markov chain compression. Statistical characterizations of the feedback side information can be further leveraged [279].

As a final remark, all the above works considered block-fading channels and optimize the ergodic capacity in the covariance optimization problem under limited feedback. However, ergodic capacity may not be an appropriate performance measure in non-ergodic channels (such as the slow fading case). In slow fading channels, there is systematic packet errors due to channel outage despite the use of powerful channel coding because given the limited CSIT, there is still uncertainty about the actual CSI and hence, the transmitted packet will be corrupted whenever the data rate exceeds the instantaneous mutual information. In addition to limited CSIT feedback, there might be feedback error due to noisy feedback links. This will also contribute to packet errors due to channel outage. When there is a noisy feedback link, the index mapping is also an important design parameter that will affect the robustness of the CSIT feedback. As a result, joint adaptation between the data rate, covariance matrix, and feedback index mapping is important to control the packet errors to a reasonable target. In order to account for the potential penalty of packet errors, it is important to consider system goodput (b/s/Hz successfully
delivered to the receiver) instead of ergodic capacity as the system performance measure in the optimization framework. The design of robust limited feedback schemes and the joint rate, covariance, and feedback index mapping optimization for system goodput is a relatively unexplored topic. In [269], the authors extend the VQ optimization framework to consider joint rate and covariance adaptation using Lloyd’s algorithm for slow fading MIMO channels.

1b) Beamforming

While optimal covariance quantization is of interest to analyze how close to perfect transmitter channel knowledge a limited feedback system can perform, limited feedback can have immediate impact enhancing existing closed-loop signaling approaches.

Beamforming is characterized by the use of a rank one covariance matrix. Note that using a rank one Q matrix is optimal whenever the single-user channel is itself rank one. This notably occurs when the user terminal is equipped with a single antenna. In this situation the availability of CSIT is critical.

In beamforming, the single-user MIMO expression in (6) is restricted so that \( x[k] = \sqrt{\rho} \mathbf{f}[k] s[k] \) where \( \mathbf{f}[k] \) is a channel dependent vector referred to as a beamforming vector and \( s[k] \) is a single-dimensional complex symbol chosen independently of the instantaneous channel conditions. For power constraint reasons, \( E_x \left[ |s[k]|^2 \right] \leq 1 \) and \( \mathbf{f}[k] \) is restricted such that \( \| \mathbf{f}[k] \|_2 = 1 \).

Much of the early beamforming work focused on the multiple-input single-output (MISO) case when there is only a single receive antenna. In this case, (6) can be reformulated as

\[
y[k] = \sqrt{\rho} \mathbf{h}^T[k] \mathbf{f}[k] s[k] + n[k]
\]

where a lower case bold symbol has been used to show that \( \mathbf{h}[k] \) is a column vector. With this configuration, the receive SNR at channel use \( k \) (averaged with respect to the transmitted signal and noise) is given by

\[
\text{SNR}[k] = \rho |\mathbf{h}^T[k] \mathbf{f}[k]|^2.
\]

For MIMO beamforming and combining, a receive-side combining vector \( \mathbf{z}[k] \) (typically unit norm) is used so that after processing

\[
y[k] = \sqrt{\rho} \mathbf{z}^T[k] \mathbf{H}[k] \mathbf{f}[k] s[k] + \mathbf{z}^T[k] n[k].
\]

Various forms of combiners exist (e.g., see the discussion in [159], [235] and the references therein).

Allowing the receiver to send some feedback to assist the transmitter’s design was proposed early in [61] and later in works such as [60], [86], [87], [190]–[180], [183]. The simplest form of this feedback is transmit antenna selection [238]. In this scenario, the transmit beamforming vector is restricted such that only one entry is non-zero. With this kind of set-up in a MISO system, the optimal solution is to send data on the antenna that maximizes the receiver SNR meaning all data (and all power) is sent on antenna \( m_{opt}[k] \) where

\[
m_{opt}[k] = \arg \max_{1 \leq m \leq M_t} |h_m[k]|^2
\]

where \( h_m[k] \) denotes the \( m \)th antenna entry of the channel vector \( \mathbf{h}[k] \). Using this approach, the optimal selected antenna can be designed at the receiver and sent back to the transmitter using \( \lfloor \log_2(M_t) \rfloor \) bits. Typically these bits are assumed error free, but work has been done in compensating for errors [142]. Error rates with antenna selection for spatially uncorrelated set-ups have been analyzed in [40], [163], [235].

Clearly antenna selection is limited in terms of its benefits to the overall capacity as it does not allow for the full beamforming gains. If there exists a feedback link, more complicated forms of channel dependent feedback should improve performance. In [182], it was proposed to quantize the channel vector for a MISO system into a set of column vectors \( \mathcal{H} = \{ \mathbf{h}_1, \ldots, \mathbf{h}_{2^B} \} \). Because the system has only a single receive antenna, the channel vector \( \mathbf{h}[k] \) can be quantized over this set by selecting the codebook vector \( \mathbf{h}_{n_{opt}[k]} \) using a phase invariant distortion such that

\[
n_{opt}[k] = \arg \max_{n_{\mathbf{h}} \in \mathcal{H}} |\mathbf{h}_{n_{opt}[k]}^T \mathbf{h}[k]|^2.
\]

The transmitter can then pick a beamforming vector that solves

\[
f[k] = \arg \max_{f \in \mathcal{F}} \left( 1 + \rho |\mathbf{h}_{n_{opt}[k]}^T f|^2 \right) = \frac{\mathbf{h}_{n_{opt}[k]}^T}{\| \mathbf{h}_{n_{opt}[k]} \|_2}.
\]

Later work analyzed the effect of training, feedback, and power quantization on these types of designs [23] and other issues of signal design in [174].

Another early form of limited feedback beamforming was the use of MISO per antenna phase quantization in [79]. Equal gain approaches that attempt to co-phase the signals received from various antennas can give excellent performance. The work in [79] used this concept to quantize the phases of each \( h_m[k], m = 1, \ldots, M_t \), using uniform phase quantization on the unit circle.

These new channel quantization approaches marked a change in thinking. Since the codebooks in [79], [182], [238] fundamentally do nothing more than allow the receiver to directly design the beamforming vector and send this designed vector back to the transmitter, the problem could be approached differently as one of beamforming vector quantization rather than channel quantization. The main idea is to restrict \( f[k] \) to lie in a set or codebook \( \mathcal{F} = \{ f_1, \ldots, f_{2^B} \} \). The receiver can use its channel knowledge to pick the optimal vector from this codebook. This kind of approach is demonstrated in Figure 3 (using the interpretation that beamforming is rank one precoding). The receiver now, in some sense, controls how the signal is adapted to the channel. This makes sense because the receiver will nearly always have higher quality CSI than the transmitter.

This change in thinking lead to significant advancement in feedback techniques. Phase quantization codebooks were created in [159] for MIMO beamforming and combining. This extended some of the concepts in [79] by jointly quantizing the phases across all the transmit antennas and guaranteed full diversity. Quantized equal gain codebooks were later
thoroughly analyzed in [176]. An analysis and summary of designs in quantized equal gain beamforming is available in [287].

While equal gain approaches are of interest, a general design framework is needed. Work in [175] for the MISO case and [162] for the MIMO case showed that for a spatially uncorrelated Rayleigh fading channel, the outage minimizing, SNR maximizing, and rate maximizing design is to i) think of the set $\mathcal{F}$ as a collection of lines in the Euclidean space $\mathbb{C}^{M_t}$ and ii) maximize the angular separation of the two closest lines. This problem is actually well known in applied mathematics as the Grassmannian line packing problem. Mathematically, this means that the set $\mathcal{F}$ is chosen to maximize its minimum distance defined as

$$d(\mathcal{F}) = \sqrt{1 - \max_{1 \leq i < j \leq 2^B} |\mathbf{f}_i^* \mathbf{f}_j|^2} = \min_{1 \leq i < j \leq 2^B} \sin(\theta_{i,j})$$

where $\theta_{i,j}$ is the angle between the lines generated by the column spaces of $\mathbf{f}_i$ and $\mathbf{f}_j$. An example is shown in Figure 2. Each beamforming vector is a point on the sphere because of the transmit power constraint. The column space of each beamforming vector corresponds to a line. Therefore, the codebook minimum distance is a function of the minimum angular separation between codebook lines.

![Fig. 2.](image)

The design of optimal or near-optimal Grassmannian line packings is in general a challenging problem. One approach is construction based on difference sets [244], [271]. Another approach is to use a numerical alternating projection algorithm [258]. Modified line packing codebooks that deal with statistical correlation knowledge via rotations and normalizations are proposed in [157]. The correlation design concept was extended to a systematic codebook design approach in [199].

Several special cases of limited feedback beamforming have been significantly analyzed. Necessary and sufficient conditions that beamforming vector codebooks must satisfy to yield full diversity are derived in [160]. In fact, for any channel, the maximum diversity order is achieved when the rank of the matrix $[\mathbf{f}_1 \cdots \mathbf{f}_{2^B}]$ constructed from the set of beamforming vectors has a rank of $M_t$. Receiver SNR degradation analysis is available in [293]. Insights from the problem of Grassmannian line packing designs can be used to assist analysis [171]. Closed-form integral expressions can be obtained by modeling the feedback problem as one of correlated antenna selection [9]. Other performance analysis results with limited feedback beamforming include [97].

An alternative approach to Grassmannian codebooks is to construct the codebooks using vector quantization (VQ) techniques. A general VQ framework for codebook design was proposed in [207]. The idea is to formulate a distortion function (usually related to rate loss or SNR loss) and then iteratively minimize this distortion to obtain locally optimal solutions. Using multiple iterations with different (possibly randomized) initial settings usually yields an approximately globally optimal codebook. Because of the unit vector constraints on the beamforming vector set, this is actually a problem in spherical vector quantization [270].

VQ designs also have very nice analytical properties when the codebook size (or quantizer resolution) increases. High resolution analysis and codebook design were successfully leveraged in [284] to give new insight into codebook behavior. Spatially correlated VQ designs are presented in [285]. Ref. [99] analyzed the symbol error probability using insights from high resolution quantization. The effect of estimation error and feedback delay is discussed in [98].

Grassmannian and VQ limited feedback designs assume codebooks that are fixed and do not vary as the channel changes. Another approach is to randomly generate the codebook at each block (with the randomly generated codebook perfectly known to both the transmitter and receiver). This sort of codebook design technique is known as random vector quantization (RVQ) and was first proposed in [222], [224]. The idea here is to generate the $2^B$ codebook vectors independently and all identically distributed according to the stationary distribution of the optimal unquantized beamforming vector.

For example, a MISO system with perfect channel information at the transmitter and receiver will use a beamforming vector $\mathbf{f}[k] = (\mathbf{h}[k])^*$ (known as maximum ratio transmission). When the channel distribution is spatially uncorrelated Rayleigh (i.e., each entry of $\mathbf{h}[k]$ is i.i.d. $\mathcal{CN}(0,1)$), this vector follows a uniform distribution on the unit sphere in $\mathbb{C}^{M_t}$. Thus, the RVQ codebook would be constructed by taking $2^B$ independently and uniformly generated points on the unit sphere. These kinds of codebooks have very nice asymptotic properties as the number of antennas scales to infinity [222], [224]. Closed-form analysis is also possible when the channel follows a spatially uncorrelated Rayleigh model [11].

Several other codebook designs have been considered as an alternative to Grassmannian line packings, vector quantization, and RVQ. Equiangular frame based codebooks were suggested in [172] based on the observation that (in the real case) codebooks from equiangular frames maximize the mutual information between the true beamforming vector and the quantized precoding vector. In certain cases Grassmannian line packing also leads to equiangular frames [244].
concepts for codebook design were first introduced in [159]. Using codebooks based on the Fourier concept for limited feedback was later generalized in [162], [175], based on a design in [84]. The key idea (in [162], [175]) is to recognize the noncoherent MIMO space-time code design problem as also the problem of finding packings on the Grassmann manifold [286]. DFT codebooks [159], [162] introduce additional structure in Fourier codebooks, further simplifying their design.

A quantized version of a basis selection algorithm is discussed in [94]. Adaptive modulation has been combined with beamforming codebooks (relating to the mean feedback adaptive modulation work in [290]) as discussed in [272]. Techniques for dealing with time variation of the channel during the feedback phase are studied in [147].

1c) Linear Precoding for Spatial Multiplexing

In beamforming, a data stream is sent spatially by projecting the data symbol onto a beamforming vector. Linear precoding extends beamforming ideas to sending multiple data streams spatially [228]. Mathematically, the input-output model is

\[ y[k] = \sqrt{\rho} H[k] F[k] s[k] + n[k]. \]  

(13)

Here, \( F[k] \) is an \( M_t \times M \) matrix (with \( M \leq M_t \)) known as the precoding matrix. The signal vector \( s[k] \) is chosen independently of the channel conditions (with the exception of the rate of the codeword set in a slow-fading channel) and typically satisfies \( E_s |s[k]|s^*[k] = \frac{1}{\rho} I \). This kind of open-loop modulation is commonly referred to as spatial multiplexing. To satisfy a sum power constraint, the matrix \( F[k] \) should satisfy \( \text{tr} (F[k] F^*[k]) = \| F[k] \|^2_F \leq M \).

The simplest form of linear precoding is antenna subset selection [168], [217]. In this kind of configuration \( F[k] \) consists of \( M \) unique columns chosen from the \( M_t \times M_t \) identity matrix \( I \). Thus, there are \( \binom{M_t}{M} \) possible antenna subset selection matrices. The receiver can use its channel knowledge to select the subset using some criterion and feed back the chosen subset using \( \log_2 \left( \frac{M_t}{M} \right) \) bits of feedback.

Unlike single antenna selection, selection of an antenna subset is challenging because it is not clear what performance metric to use. Early work in [70], [80] considered various designs. The capacity with antenna subset selection was analyzed in [25].

Various papers have analyzed antenna subset selection systems for different models. Transmit and receive antenna selection analysis and the resulting capacity are studied in [72]. The capacity growth and capacity degradation with antenna subset selection is derived in [219]. Antenna subset selection in keyhole channels has been studied in [218]. Antenna subset selection with correlation was analyzed in [191].

Another issue with antenna subset selection is the choice of the dimension \( M \). Typically, this subset value is fixed for all times. Removing this constraint (i.e., allowing \( M \) to vary with the channel conditions) can provide large achievable rate or error rate improvements. Antenna subset selection for spatial multiplexing with a varying number of spatial data streams was proposed [78] under the terminology of multi-mode antenna subset selection (where \( M \) is termed the mode of the system).

Techniques for selecting the mode using limited feedback are given in [125], [202].

Another form of simplified linear precoding is the use of a diagonal \( F[k] \) that only adjusts per antenna powers. This form of adaptation (sometimes referred to as per antenna rate and power control) is a natural addition to spatial multiplexing. Varying the power of open-loop spatial multiplexing using finite-rate feedback with an outage design criteria is the focus of [54], [55].

The best form of performance generally comes when \( F[k] \) is designed to direct data over “good” channel subspace directions. Like [182], the first ideas behind limited feedback linear precoding focused on quantizing \( H[k] \) [106]. The idea being that the receiver would quantize the channel, feed back the quantized value to the transmitter, and then have the transmitter pick the precoder assuming that this quantized side information is the perfect channel realization. The work in [200] further elaborates on the notion of “good” and “bad” channels for a precoding scheme in a correlated channel setting.

Like beamforming, performance improves by using a codebook approach to choose the linear precoder directly. The receiver can use its knowledge of \( H[k] \) to pick the optimal linear transmitter-side linear precoder from a codebook known to the transmitter and receiver. For a codebook of size \( 2^B \), the \( B \) bit binary label of the chosen precoder is sent over the feedback channel. Note that the rate and/or SNR must also be known as side information to facilitate communication and is often fed back.

Fig. 3. A block diagram of a limited feedback linear precoded MIMO system is shown above. The receiver uses its channel estimate to pick the optimal linear transmitter-side linear precoder from a codebook known to the transmitter and receiver. For a codebook of size \( 2^B \), the \( B \) bit binary label of the chosen precoder is sent over the feedback channel. Note that the rate and/or SNR must also be known as side information to facilitate communication and is often fed back.

A common precoding framework is the use of multidimensional eigenbeamforming. The idea is to use \( M \) orthonormal unit vectors to spatially signal \( M \) data streams. This means \( F^*[k] F[k] = I \) where \( I \) is the \( M \times M \) dimensional unit vector. Generally most performance metrics depend only on the product \( F[k] F^*[k] \). For this reason, right multiplication of \( F[k] \) by any unitary matrix does not change the performance metric. Using this invariance, the performance is dependent only on the subspace spanned by the columns of \( F[k] \) not the exact formulation of \( F[k] \). The concepts employed in Grassmannian line packing beamforming codebooks in [162],
important information about antenna correlation, mean and standard deviation of angles of departure/arrival which can be exploited in the design of a precoding matrix so as to minimize BER, maximize mutual information, etc. Such work was done in, e.g., [146], [214], [264], [288], [289].

In most work, a simplified spatial correlation model known as the Kronecker model is assumed, which assumes decoupling between transmit and receive antenna correlation structure and leads to closed-form precoders. To deal with arbitrary propagation scenarios (non-Kronecker) alternative algorithms minimizing the BER were proposed in [83].

Beyond exploiting channel statistical information, instantaneous channel adaptation can yield further improvements. Initial work proposed a general model covering both the cases of statistical and instantaneous feedback in [104], [105]. Instantaneous channel feedback appears in the form of channel matrix estimate whose distance (error) to the true channel is arbitrary, but with known statistics.

The simplest form of closed-loop space-time coding is again antenna subset selection. In this case, an $M$-antenna spacetime code is sent over $M_i$ antennas (with $M_i > M$) antennas. The most popular form of space-time coding, orthogonal space-time block codes (see [5], [133], [194], [255]), is a perfect fit for antenna subset selection particularly because certain values of $M$ lead to “better” code designs. The most popular choice would be $M = 2$ where the simple rate one Alamouti space-time code is available [5]. Antenna subset selection with orthogonal space-time block coding was proposed in [71]. Later, extended orthogonal space-time block codes designed using antenna grouping ideas were proposed in [153], [254].

One of the benefits of orthogonal space-time block codes with antenna subset selection is that they are amenable to closed-form performance analysis. Analysis of the error probability of orthogonal space-time block codes with antenna subset selection has been studied in [109], [150]. The capacity was analyzed in [198].

Like spatial multiplexing, space-time codes also can be linearly precoded for the purpose of achieving both the diversity and beamforming gains. In the case of strong transmit antenna correlation, a precoder based on the knowledge of the correlation matrix alone (like the work mentioned earlier in this section) will achieve some beamforming gain already. However full beamforming capability in general requires instantaneous CSIT.

In this configuration with a space-time code sent over $K_{ST}$ channel uses, the received codeword matrix

\[
Y[k] = \sqrt{\rho}H[k]F[k]S[k] + N[k].
\]

Unlike lines, defining the distance between two subspaces is more complex. There is one principal angle per subspace dimension (for a total of $M$ principal angles). Various distances can be defined using these angles [20]. It was shown in [155] that these various distances (which incidently yield very similar codebooks) can optimize different performance criteria.

The problem of designing $\mathcal{F} = \{F_1, \ldots, F_{2^N}\}$ has continued to be a focus of research. The relation between limited feedback precoding and the Grassmannian manifold was taken into account in [47]. Fourier approaches exploiting the relationship between non-coherent code design in [84] and Grassmannian subspace packing were used in [154], [155]. New precoder designs for spatially correlated channels were the focus of [201]. A bit error rate (BER) minimizing design with linear receivers and quadrature amplitude modulation (QAM) constellations was analyzed in [291]. Unitary rotation precoding with Givens rotations are studied in [211]. An alternative feedback precoding scheme was proposed in [139]. New insights into parameterizing the channel and precoding matrix were given in [208]. Expansions of codebooks using Householder reflections have been used to generate a unitary matrix from a beamforming vector, to enable multimode precoding and certain kinds of multiuser MIMO feedback. Code designs based on mutually unbiased bases or Kerdock codes have been proposed to provide small alphabet near-Grassmannian codebooks that also facilitate multimode precoding (see [173]). Several of these codebooks and other codebooks, often resulting from compromises among different companies, have been included in several wireless standards (see Section IV).

RVQ analysis can also be extended to linear precoding [222], [223]. Again, the codebook $\mathcal{F}$ has i.i.d. entries following the distribution of the capacity maximizing precoder. This precoder is given by the $M$ right singular vectors corresponding to the $M$ largest singular values. In the case of a spatially uncorrelated Rayleigh fading channel, this $M$-dimensional matrix is uniformly distributed on the set of $M$-dimensional matrices with orthonormal columns (a kind of Steifel manifold). Again, this kind of precoding has nice properties that make asymptotic analysis tractable.

Like antenna subset selection, varying the dimensionality (or mode) $M$ of the precoder gives substantial improvement. Ref. [156] discusses precoder design that includes mode feedback as a function of the channel conditions. An analysis of the optimal number of substreams is given in [50].

1d) Improved Space-Time Coding

Typically, space-times codes are designed without channel information, as theory shows that diversity gain can be extracted from the MIMO channel without CSIT (also the case for the single-user spatial multiplexing gain).

Space-time codes, however, can also be enhanced by adapting to channel conditions. If a high rate feedback channel is unavailable, it is possible to limit adaptation to channel statistics which are by nature slow varying, or perhaps even uplink-downlink reciprocal. Channel spatial statistics convey
of designing the codebooks has been shown to relate to the problem of Grassmannian subspace packing [154]. For precoded orthogonal space-time block coding, however, the distance metric used for subspace packing is chordal distance (unlike precoded spatial multiplexing codebooks that use projection two-norm or Fubini-Study distance).

For arbitrary codebooks with precoded orthogonal space-time block codes, necessary and sufficient conditions to guarantee full diversity conditions were derived in [153]. The general idea is that the columns of all codebook precoders should span all of \( \mathbb{C}^{M_t} \). Conditions for optimal precoding were derived in [216]. A partially precoded form of space-time code design is proposed in [56].

An interesting alternative to using precoding is to switch between multiplexing and space-time coding (primarily using orthogonal space-time codes) [81]. The idea is to compare the receive minimum distance of both multiplexing and space-time coding given a fixed rate. To minimize the probability of error, the signaling architecture with the maximum receive minimum distance is chosen. In fact, it is shown in [100] that the optimal signaling approach given quantized feedback in a MISO setting will vary in rank.

In [4], a technique for extending any \( M \) antenna orthogonal space-time block code to any \( M_t \) antenna transmitter (satisfying that \( M_t \) is an integer) is discussed. This approach uses a technique more general than standard linear precoding that requires \( M_t - 1 \) bits of feedback. The idea behind this approach is to use group coherent codes which preserve the low complexity decoding of orthogonal ST codes while enjoying some partial beamforming gain. Additionally, this technique can also be modified to apply to non-orthogonal space-time block codes.

One other enhancement to space-time codes is rate adaptation. The idea being that limited feedback can be sent from the receiver to the transmitter to vary the rate. These techniques were studied in [124] taking delay (i.e., outdated feedback) into account.

There has been some work on codes other than orthogonal space-time block codes. Power allocation using limited feedback for linear dispersion codes is the focus of [145]. Quasi-orthogonal space-time block codes are adapted with feedback in [16], [17], [54], [257]. Space-time bit interleaved coded modulation with finite feedback is addressed in [120], [121] using a precoding-like structure. Space-time trellis coding using quantized phase information (in a way similar to the work in quantized equal gain beamforming) was studied in [148].

1e) Feedback Overhead Loss

In understanding the benefits of finite rate feedback, it is also important to characterize the detriment. Using feedback creates overhead on one side of the link to benefit the achievable data rate on the other side. This overhead can often be non-negligible. The first work on analyzing the overhead of feedback was for the SISO case in [34]. In this paper, rate degradation was taken into account in a symmetric way by adjusting the fraction of channel resources used for feedback. A symmetric MIMO FDD model was used in [151] to analyze the penalty (in bits per second) of feedback.

Training and feedback overhead optimized were optimized [226], [227] assuming a symmetric RVQ system. These papers used an asymptotic (large antenna) analysis to gain insight. A numerical approach to trading off resources for training, feedback, and transmission was given in [10]. The role of overhead is analyzed using an RVQ beamforming assumption was discussed in [152]. Related work also has looked at time division cases [242].

2) Broadband Systems: Fourth generation (4G) and beyond cellular standards are expected to use MIMO-OFDM technology. As mentioned during the single antenna OFDM feedback discussion, fully characterizing the complete channel state information for any OFDM system can be challenging.

Generalizing the input-output relation of (5) to MIMO for the \( v \)th subcarrier yields

\[
\mathbf{\hat{y}}_v[k] = \mathbf{H}_v[k] \mathbf{\hat{x}}_v[k] + \mathbf{n}_v[k]
\]

(15)

for OFDM channel use \( k \). In (15), \( \mathbf{\hat{y}}_v[k] \) is an \( M_r \)-dimensional received signal for subcarrier \( v \), \( \mathbf{H}_v[k] \) is the \( M_r \times M_t \) channel realization (in the frequency domain) for the \( v \)th subcarrier, \( \mathbf{\hat{x}}_v[k] \) is an \( M_t \)-dimensional transmitted signal for subcarrier \( v \), and \( \mathbf{n}_v[k] \) is \( M_r \)-dimensional normalized additive noise with i.i.d. \( \mathcal{CN}(0, 1) \) entries.

MIMO channel adaptation must be done on a per-subcarrier basis. For example, a linear precoded spatial multiplexing system would set

\[
\mathbf{\hat{x}}_v[k] = \sqrt{\rho_v} \mathbf{\hat{F}}_v[k] \mathbf{\hat{s}}_v[k]
\]

where \( \rho_v \) is the SNR on subcarrier \( v \), \( \mathbf{\hat{F}}_v[k] \) is the \( M_r \times M \) precoder on subcarrier \( v \), and \( \mathbf{\hat{s}}_v[k] \) is an \( M \)-dimensional normalized spatial multiplexing vector satisfying \( \mathbb{E}_s[\mathbf{\hat{s}}_v[k]\mathbf{\hat{s}}_v[k]^\ast] = \frac{1}{M} \mathbf{I} \). The precoder \( \mathbf{\hat{F}}_v[k] \) must be adapted directly to \( \mathbf{H}_v[k] \).

Like the single antenna case, MIMO-OFDM feedback systems often send feedback only for pilot subcarriers \( v_0, \ldots, v_{K_{pilots} - 1} \) where \( K_{pilots} \) is the number of pilots. For example, a precoding system using limited feedback with a common codebook for all pilots of \( \mathcal{F} = \{F_1, \ldots, F_{2^p}\} \) would send \( B \) bits for each pilot subcarrier for a total feedback load of \( BK_{pilots} \) bits per channel block. Given this information, the challenge is determining the precoders for non-pilots.

The first paper to address this problem was for the case of beamforming in [41]. Inspired by spherical interpolation, the idea was to weight and sum together the feedback vectors from the two nearest pilots. The weights were designed to maximize the receive SNR of the subcarrier half-way between the two pilots. A transform domain quantization approach was discussed in [169]. In [187], the precoder interpolation problem was formulated as a weighted least squares problem. The weights correspond to the distance (in number of subcarriers) from the different pilot precoders. The technique in [41] was later generalized to larger rank precoding interpolation techniques in [42]. A geodesic approach (i.e., linear interpolation on the Grassmann manifold) was the focus of [188]. Other interpolation ideas are also available in [27], [36]. Instead of trying to interpolate a much simpler approach
is clustering [169], [188], where a common precoder is chosen for several contiguous subcarriers. Interestingly, the clustering approach can yield an antenna subset selection criterion when the cluster is extended to cover all subcarriers (i.e., only one pilot) and the precoding codebook has the $\left( \binom{M}{M_t} \right)$ antenna subset matrices.

The general idea behind interpolation is demonstrated in Figure 4. Sending precoder feedback only on a subset of the subcarriers, the transmitter must use this information and the channel correlation in the frequency domain to recreate all precoders as shown in Figure 4 (a). One way (though not necessarily general) of thinking of interpolation is shown in Figure 4 (b). In this figure, the beamforming vector for a subcarrier is found from the point on the unit sphere corresponding to a line drawn on the surface of the sphere. The beamforming vectors for subcarriers between the pilots are chosen from this line using the subcarrier index.

Fig. 4. Figure (a) shows an example of an interpolation MIMO-OFDM limited feedback system. Feedback is only sent for a select number of subcarriers (called pilots here). The pilot feedback information is used to design precoders for all subcarriers. In Figure (b), a beamforming example (i.e., rank one precoding) shows one way of choosing the beamforming vector using the pilot information. A line is traced on the surface of the sphere. The beamforming vectors for subcarriers between the pilots are chosen from this line using the subcarrier index.

for precoder interpolation are discussed in [292]. The work in [149] uses a different approach called successive beamforming to take into account correlation in time and frequency. A reduced CSI feedback approach for MIMO-OFDM is proposed in [278] that takes into account the fact that highly correlated channels will have highly correlated feedback values; thus, the number of bits can be effectively reduced by taking the actual correlation between binary sequences into account.

The multi-mode precoding problem is even more difficult to quantify. In this scenario, both the matrix and the rank of the matrix can evolve over the OFDM symbol subcarriers.

An interpolation framework for multi-mode precoding was published in [110].

While OFDM is the most popular MIMO broadband approach, single carrier systems are still of interest. Beamforming concepts (assuming receiver equalization) have been extended to these systems in [143], [144]. Here the beamforming is actually a finite impulse response filter. Feedback techniques similar to the narrowband scenario can be employed for these systems.

III. FEEDBACK IN MULTIUSER WIRELESS SYSTEMS

Adapting the transmitted signal across multiple users is an additional degree-of-freedom that can be leveraged in most communication systems. Clearly, the level of channel knowledge required typically increases proportionally to the number of users. This creates challenging problems in practical system implementations when the transmitter does not have a priori channel information. In this section, we discuss applications of limited feedback in multiuser systems.

A. Single Antenna Scheduling and Spectrum Sharing

In multiuser systems, users compete for resources to ensure larger rates and/or better reliability. In a cellular framework with single antennas, the maximum throughput decision is to transmit to the users with the largest receive SNR at each channel use in order to achieve the so-called multiuser diversity gain [123]. To perform this scheduling, the base station needs knowledge of users’ SNR conditions.

One solution proposed in [62], [63] is for the users to threshold their receive SNRs and notify the base station only if their SNR exceeds some predetermined threshold. This is a very rudimentary approach to SNR (or channel magnitude)
quantization since it is basically a one bit per user feedback set-up. One issue is that there is a small probability that no user will report their SNR thereby leaving the scheduler with no CSIT. A multiple-stage version of this threshold-based technique solves this problem at the expense of some extra latency [75]. These one-bit feedback techniques are very bandwidth efficient. Using more feedback bits, however, may offer diminishing returns. In [221], the benefits of employing only one feedback bit per user and the minor rate enhancements of more feedback bits are analyzed. Other uses of feedback include varying rate and power to minimize outage [203]. An excellent overview of feedback and the compression of feedback in multiuser wireless systems is available in [59].

More discussion on channel aware scheduling and cross-layer design is contained in [267].

Multiple antenna enhancements of these limited feedback scheduling works generally are enhanced versions of opportunistic beamforming [263]. Random opportunistic beamforming offers a way of handling both the beamforming and scheduling problems simultaneously with scalar-only SNR (thus limited) feedback. A limited feedback version of opportunistic beamforming was studied in [185], [220]. A comparison of diversity versus opportunistic beamforming is in [126]. The problem of scheduling in multiuser MIMO systems is visited later in this paper.

In broadband systems with multiple access in frequency, users can be scheduled in various subchannels. Feedback and subcarrier allocation in orthogonal frequency division multiple access (OFDMA) can be done using limited feedback [39]. Additionally, opportunistic scheduling techniques can be combined with OFDM [221]. The problem of using feedback with OFDMA scheduling is also discussed in [140], [165]. A thorough summary of the design issues facing multiuser 3G and beyond OFDMA systems is available in [243].

When multiple access is done in the time domain, limited feedback can allow the system to map users to time slots and adapt coding and modulation [167]. This can lead to transmit power minimizing solutions when the problem is thought of as one of distributed antenna beamforming [166].

Another interesting area is signature optimization using limited feedback [204]. In fact, spreading code design using randomly generated codes formed the basis behind the development of RVQ ideas [225]. Spreading code design has been addressed from the point-of-view of multi-carrier CDMA [195]. Reduced rank signature optimization leads to further designs using subspace concepts [248]. Performance analysis of signature optimization with CDMA can be found in [48].

One issue in multiple access systems where possibly many users are sending feedback to a central controller (e.g., base station) is the issue of limited feedback resources. Clearly multiple access strategies are necessary when users must compete for limited feedback channels. Shared feedback resources were first examined in [251] using a shared random access feedback channel. A code division framework is available in [192].

An exciting area where feedback can have impact is in spectrum access aware cognitive radio systems. One of the big problems among open-access systems is determining how users can adequately share spectrum resources. In [2], a technique allowing users to compete for access to spectrum using limited feedback is discussed.

B. Multiuser Multiple-Input Multiple-Output Systems

While single-user multiple antenna systems provide many benefits, multiuser multiple antennas systems can provide even larger total system rates when the spatial resources are spread among multiple users. Often, the user terminals have limited or no (if single antenna, as considered below) interference canceling capability by themselves. In the downlink, this leaves the base station with the task of precoding the signals in view of suppressing the inter-user interference. CSIT is crucial for this task. Thus, for cellular systems that commonly use FDD, limited feedback is critical to making multiuser MIMO practical.

1) Single Receive Antennas:

A downlink multiuser MIMO system with each user possessing a single receive antenna and normalized noise will give the $i$th user an input-output relation

$$y_i[k] = h_i^T[k] x[k] + n_i[k]$$

where the subscript denotes the user number. The transmitted signal is restricted such that $E_{h_i} \left\| x[k] \right\|_2^2 \leq \rho$. Typically, $\{x[k]\}$ is designed to support at most $M_t$ users per block length.

Various signaling approaches can be employed to divide the spatial resources. Most relate to using at least some form of precoding. In this case, $x[k] = \sqrt{\rho} F[k] s[k]$. Here $F[k]$ is the linear precoding matrix. The signal $\{s[k]\}$ could have independently generated rows that correspond to different users or do more complicated interference mitigation schemes such as dirty paper coding (e.g., see [32], [239], [261], [268], [275]).

Most limited feedback multiuser MIMO schemes fall into two categories. The first category is to let users quantize some function of $h_i[k]$ and send this channel information to the base station. The problem is that the purpose of $F[k]$ is to (in some sense) orthogonalize the various user signals. When the channel is quantized, the user signals can not be perfectly orthogonalized due to inherent quantization error [53], [101]. This leads to a sum rate ceiling as the SNR increases.

Improving this sum rate ceiling is a difficult problem. When the number of users increases, scheduling users with channels that satisfy near orthogonality conditions provides many improvements [274]. Further interesting aspects on multiuser scheduling are discussed later in this paper. The sum rate ceiling can also be raised by leveraging quantization distortion in an MMSE-type of precoder design [46].

Innovative work has also been done characterizing what type of and how much feedback should be used. A comparison between quantized and analog feedback methods with sum rate maximization is available in [28]. The effect of training, feedback quantization, and feedback error on sum rate is analyzed in [29].

The second broad category is a multiuser version of the opportunistic beamforming approach mentioned earlier (as initially discussed in [234]). In the approach of [234], $F[k]$ is chosen randomly according to a known distribution. Pilots
are sent out on all of the spatial beams (i.e., columns of the precoder). Users then measure their receive SINRs on the columns and report this information back to the base station. The base station picks the subset of users that maximizes the sum rate. The best property of this algorithm is that it approaches the optimal sum rate as the number of users increases assuming independently fading spatially uncorrelated Rayleigh channels. This technique can also be extended to probe with multiple precoders [89]. This will allow the sum rate to scale faster as the number of users increases.

Multiuser opportunistic beamforming techniques typically exhibit a degradation of performance when the number of users is low. This is because some of the random beams fail to "hit" a user. To mitigate this problem, several approaches were proposed including enhancing the randomly launched beams with a power control algorithm allowing to reduce the resource allocated to the beams for which no user reported a good enough SINR [129]. Another method improves the beam design with the help of known statistical information. The correlation matrices associated with the transmit channel of some users reveal a great deal about the mean separability of these users. Such information can be exploited both for scheduling and precoder design [64].

This kind of probing idea relates to the popular industrial feedback technique known as Per User Unitary Rate Control (PU2RC) in [115]. Typically PU2RC systems use a deterministic (rather than random codebook) [113]. They can be designed using probing or allowing the users to perform channel estimation and then compute SINRs for all codeword matrices. The throughput scaling of PU2RC was analyzed in [88] for both the noise limited and interference limited regimes.

These similarities and differences in multiuser MIMO approaches are roughly described in Figure 6. The primary difference is in the information that is fed back. Channel quantization does not constrain the form of the multiuser MIMO precoding matrix, but it often leads to inferior performance due to quantization error. Channel sounding allows users to measure (or compute) actual SINR performance. This typically comes at the expense of constraining the precoder to a finite set (or codebook).

2) Multiple Receive Antennas: When users have multiple receive antennas, performance of multiuser MIMO systems can be improved by leveraging the added degrees-of-freedom at the receiver. With enough receive antennas even simple per antenna scheduling without precoding can provide good performance [77]. It was shown in [102] that combining the signals received at the multiple receive antennas provides substantial sum rate benefit because the negative effect of channel quantization error is reduced. Block diagonalized multiuser transmission with limited feedback is discussed in [205], which takes into account that each receive antenna should not be treated as a separate user when the antennas are co-located. A vector quantization framework combined with improved scheduling is discussed in [118]. The amount of feedback necessary to avoid a sum rate ceiling is asymptotically analyzed in [21].

A recent technique for multiuser MIMO with multiple receive antenna users is coordinated beamforming with limited feedback. In coordinated beamforming, both the transmit beamformers and receive combining vectors are jointly optimized to maximize sum rate under a zero interference constraint [186]. When implemented with limited feedback [38], the users send limited feedback by sending to the base station the quantized coefficients of a certain normalized Hermitian symmetric matrix that is a function of their channel coefficients. The base station then uses the channel conditions of multiple users to solve for the transmit beamforming and receive combining vectors for each user. Both iterative algorithms [37] and direct solutions can be employed [38]. The receive combining vectors are quantized as part of the joint optimization and subsequently broadcast to the users in what is called limited feedforward [37]. The joint design improves performance versus receive antenna selection [102] while incurring only a small feedback overhead.

An enhanced version of PU2RC using multiple receive antennas was proposed in [116]. WCDMA/HSDPA approaches to include multiuser MIMO with feedback are studied in [117]. Note that practical systems will still suffer from the problem of many users competing for limited feedback resource. Contention feedback approaches for multiuser MIMO, where users compete to send feedback on a shared contention feedback channel may solve this problem [192], [252].

Feedback designs have also been studied for MIMO-CDMA systems, assuming single antenna subscriber units. For example, antenna partitioning based on limited feedback from the subscriber has been proposed to improve the receive SINR of each user by assigning them to the best transmit antenna [43].
The multiuser MIMO discussion above has concentrated on the downlink. In the uplink, user synchronization makes the feedback problem almost identical to those studied for point-to-point MIMO as discussed in [112]. The primary difference is that the codebook precoders are geographically distributed with different precoder columns corresponding to different users. The multiuser MIMO uplink also represents another form of spatial division multiple access [90]. This view provides intuition into system throughput scaling and scheduling.

C. Scheduling in Multiuser Multiple-Input Multiple-Output Systems

When the number of users actively receiving or transmitting packets exceeds the number of antennas at the base station, it becomes necessary to select a subset of terminals which will be allowed to communicate with the base station in any given time or frequency slot. Although this selection could in principle be realized without any prior information on the channel state of the users (as in round-robin scheduling for instance), it has been shown that multiple antenna systems can extract a substantial capacity gain from using carefully designed selection rules. In brief, a rate maximizing scheduler will aim at realizing a good compromise between the multiuser diversity gain, by selecting users with high received power levels, and the multiplexing gain, by choosing the users with sufficiently orthogonal channels, in view of, e.g., linear precoding.

In a precoded SDMA-type systems, when the transmitter lacks CSIT feedback must be used to perform scheduling and to design the proper multiuser MIMO precoding matrix for the selected users. Clearly, the required feedback resolution for each of these two tasks is different. Scheduling typically requires only rough quantization resolution to distinguish between high and low rate users (e.g., see the discussion in [66]). Precoding, however, requires higher resolution accuracy to mitigate the problems caused by interuser interference. This motivates flexible feedback designs which partition optimally the feedback bit budget across those used for scheduling and those for precoding [276].

There is also a non-trivial optimization problem related to how many users should be scheduled given the quantization level. Scheduling more simultaneous users can sometimes create more interuser interference and not actually enhance the achievable sum rate. The tradeoff between multiuser diversity and multiplexing gain is analyzed in [127]. Work in beam selection using SINR feedback has been done in [44]. Scheduling with limited feedback information for multiuser MIMO has been enhanced in [128]. Another way to enhance the performance is by actively varying the rank of the precoder (which corresponds to the number of users selected) using interference prediction approaches [12]. A thresholding technique for the decision of multiuser MIMO feedback is discussed in [265]. The optimal number of users to be supported was characterized in [49].

Another way of exploring the trade-off between multiuser diversity gain and user multiplexing gain was proposed in [206]. In this work, the total feedback bit budget is divided equally across a subset of users, while users outside the subset are dropped from the system. As the considered subset size is reduced, more weight is given on multiplexing gain and less on the multiuser diversity. The optimum trade-off point is a function of SNR, system parameters. This relates to the idea of per-user rate-adaptive feedback which considers an average, rather than fixed, feedback bit budget per user [277].

D. Relaying

In a wireless relay channel, a helper node without data to send acts as a relay between a source and a destination. There are three channels in the basic relay channel: the source-to-relay channel, the relay-to-destination channel, and the source-to-destination channel. More sophisticated relay channels may involve multiple relays, multiple stages of relaying, bidirectional relaying, and MIMO relays. Relays have different levels of assumed sophistication. Amply-and-forward relays (also known by other names) generally apply some kind of analog transformation to the received signal while decode-and-forward relays detect the transmitted bits and re-encode. Relays may operate under full-duplex assumptions (transmitting and receiving at the same time) or more practical half-duplex assumptions (transmitting or receiving). Relaying is now being considered for cellular coverage enhancement through efforts like IEEE 802.16j [196]. It is an extensive area in its own right - this section discusses some applications of limited feedback in relay channels.

Exploiting limited feedback concepts in relay channels is challenging due to the number of different channels. There are multiple channels to be estimated, feedback may be required at multiple nodes, and feedback information may traverse multiple paths. Further, relays are often assumed to be incapable of complex PHY and MAC operation, making the amount and type of possible training and feedback more restricted than in other channel models.

A rough block diagram of a limited feedback relay system is shown in Figure 7. The general idea is to enhance the rate and/or reliability of a transmitter and destination terminal with poor channel conditions. The relay retransmits (and possibly reencodes) the source signal meant for the destination node. Limited feedback links could be used for sink-to-destination feedback and relay-to-source feedback. Because of the poor channel conditions sink-to-source feedback is unusual.

While there remain many open problems in the area, research in limited feedback for relay systems has made significant progress in just the past few years. The seminal paper [131] shows that a single bit of feedback from destination to relay allows decode-and-forward to achieve the full diversity available in the single-antenna relay channel. In [3], the authors derive finite-rate feedback for power control to reduce the probability of outage in the single-antenna amplify-and-forward relay channel, finding that just one bit of feedback can double the diversity order and deriving a strategy that accommodates any size of desired feedback. A similar scenario for the decode-and-forward channel is studied in [119], where it is shown that very low levels of feedback...
Transmitter was not decoded correctly [282]. In another approach, parity information upon receiving feedback that a packet is known as relay selection. Feedback can be used to help communication between source and destination, which is to select the relay that will be used to forward packets to the destination [253]. Alternative to relay selection, multiple relays can be co-phased based on limited feedback when a certain amount of synchronization is available to improve [130].

IV. CODEBOOK BASED FEEDBACK IN STANDARDS

In the past five years, several applications of feedback have appeared in emerging wireless standards including adaptive coding and modulation, power control, hybrid automatic repeat request, and codebook based limited feedback precoding. In this section we review codebook based limited feedback precoding concepts that have been decided or are under consideration in several emerging standards.

A. 3GPP Release - 99

3GPP (third generation partnership program) is the name for the official evolution of the GSM system. It features a wideband code division multiple access (CDMA) physical layer with $5MHz$ channel bandwidths. It uses fast power control feedback on both the uplink and downlink. It is also the first standard to support codebook based beamforming on the downlink with two transmit antennas [52].

Two types of codebook based feedback are supported in the standard [85]. Mode 1 uses a type of quantized equal gain combining [79], [159], [177], where the phase of the second antenna is adjusted based on the average of two one bit feedback commands, effectively implementing a two-bit codebook. Mode 2 uses a type of quantized maximum ratio combining. By averaging over multiple feedback slots, effectively two bits of amplitude information and three bits for phase information are realized.

B. IEEE 802.11n

IEEE 802.11n is a developing standard that is an extension to the IEEE 802.11 wireless local area network standard with a stated goal of achieving 100 $Mbps$ of effective throughput [193]. It uses a MIMO-OFDM physical layer and supports two to four transmit antennas and two to four receive antennas. It supports two flavors of single user feedback based MIMO precoding (generically called beamforming in the standard): implicit and explicit [245].

Implicit feedback precoding uses the channel reciprocity that results from TDD operation. In theory, with reciprocity there is no limited feedback required. In practice though, the RF chains at the transmitter and receiver require calibration to ensure reciprocity across the entire analog signal path. Reciprocity is enabled through the use of feedback. Specifically, two users exchange training information. Then each user sends their quantized channel estimate per subcarrier to the other user. Based on this feedback, each user is able to calibrate their baseband and precoding can be performed using reciprocity. Note that the calibration procedure must be performed during a coherence time but does not have to be performed frequently (perhaps repeated in minutes or hours).

There are three different explicit feedback precoding modes of operation. The first mode is called CSI Matrices feedback...
and consists of sending back for each subcarrier a quantized maximum amplitude (3 bits) and the quantized channel matrix (4−8 bits per scaled real and imaginary entry). The second mode is called non-compressed beamforming matrix feedback. In this case the receiver computes the precoder with orthogonal columns, quantizes each entry, and sends this information back. In the final mode of operation, known as compressed beamforming matrix feedback, the receiver represents the precoder with orthogonal columns using Given’s rotations (inspired by [208]). The parameters (called angles) are then quantized and sent back. The CSI Matrix mode is the most general and allows the transmitter to compute its own precoder. The non-compressed mode reduces the amount of feedback required to just the precoder while the compressed mode further reduces feedback and preserves orthogonality of the columns with quantization.

C. IEEE 802.16e / WiMax

IEEE 802.16e is the mobile extension to IEEE 802.16, for wireless metropolitan area networks [8]. It is often known by the name WiMAX (Wireless Interoperability for Microwave Access), coming from the WiMAX Forum, which is an industry consortium selecting certain subsets of IEEE 802.16e for interoperability and certification. IEEE 802.16e has several different physical layers and MIMO modes of operation. The current WiMAX mobile profile 1.0 supports OFDMA and some basic MIMO features. The next release (1.5) will likely support some of the codebook feedback modes including in the standard.

Several single-user codebook based limited feedback techniques are supported in IEEE 802.16e in the downlink. Codebooks are given in the standard for several configurations. Two, three, and four antenna three bit codebooks are listed; they appear to be Grassmannian codebooks. Six bit beamforming codebooks for three and four antennas are derived from a generating vector multiplied by a Householder reflection matrix, exponentiated diagonal matrix, and another Householder reflection matrix. This approach saves some storage. Precoding codebooks for 3 and 6 bits for up to four antennas are found by taking subsets of the columns of a Householder reflection matrix generated using the beamforming codebook. This result in codebooks with that can be generated from the beamforming codebooks; storing precoding codebooks is not required. Note that Householder reflection matrices are unitary thus the precoders constructed from them have orthogonal columns.

D. 3GPP Long Term Evolution

3GPP Long Term Evolution (LTE) is the name for release 8 of the 3GPP standard, the evolution of 3GPP Release 99. 3GPP LTE has a MIMO OFDMA physical layer on the downlink and supports various single user and multiple user MIMO modes of operation [65].

Several different single-user codebook based limited feedback techniques are supported in 3GPP LTE. While multiple user codebook methods were discussed during standard meetings, they did not make it into release 8. 3GPP LTE has support for codebook based precoding on the downlink with two or four transmit antennas. In the case of two antennas, a beamforming codebook with six vectors (including two corresponding to antenna selection) and a precoding codebook with three matrices. For four antennas, a four bit codebook is specified for beamforming and precoding with two, three, and four streams. The precoding codebooks are built by taking specific subsets of Householder reflection matrix generated from each beamforming entries. The subsets are chosen to have a nested structure. For example, for a given generating vector, the two stream codebook will include the original vector and an additional vector. The three stream codebook will add an additional vector and so on. This facilitates multi-mode rank adaptation, where the base station can change the number of active streams, and may offer some computational savings.

A major difference between the 3GPP LTE and IEEE 802.16e codebooks is that the 3GPP codebooks have a finite alphabet structure, which makes them easy to store and simplifies computation. This structure is preserved even in the calculation of the reflection matrices, thus can be exploited in the precoding case as well.

E. 3GPP2 Ultra Mobile Broadband

3GPP2 (third generation partnership program 2) Ultra Mobile Broadband (UMB) is the name for the evolution of the cdma2000 standard. It also uses a MIMO OFDMA physical layer along with some CDMA control channels. It supports codebook based precoding for single-user and multiuser systems on the downlink with two or four antennas. It also has supports rank adaptation and adaptive switching between single-user and multiuser modes.

Two different precoding codebooks are supported: a knockdown codebook and a readymade codebook. A knockdown codebook consists of multiple unitary matrices. The receiver chooses a preferred matrix and columns from that matrix to indicate its preferred precoder. Two default codebooks are a Fourier-based codebook with multiple phase shifted discrete Fourier Transform matrices (inspired from constructions in [154], [155]) and an identity codebook (corresponding to antenna selection). A readymade precoding codebook consists of up to 64 matrices. For a given rank $r$, the receiver sends back the preferred matrix to the transmitter. The transmitter uses the first $r$ columns of the chosen matrix as the precoder. Interestingly, unlike other standards, 3GPP2 UMB has downloadable codebooks (a mandatory feature) so the default codebooks in the standard can be changed. This means that hardware can not exploit special structure in the codebooks since codeword search at the receiver must support the downloaded codebook.

Multiple user MIMO, or SDMA, is also supported using the Knockdown codebook and an appropriate channel quality indicator. In this case users are scheduled onto different beams of a single matrix, inspired by the PU2RC multiuser algorithm [115].

F. IEEE 802.16m / 4G

IEEE 802.16m is tasked with developing an advanced air interface for IEEE 802.16. It is one of what will likely be
several fourth generation cellular standards (versus 3GPP LTE and 3GPP2 UMB that are generally considered to be 3.5 generation standards). While it is still early in the standardization process, it appears that IEEE 802.16m (and other 4G standards) will pick up where IEEE 802.16e left off. It is likely that the physical layer will use MIMO OFDMA but will support more advanced techniques like adaptive feedback, multiuser MIMO, relaying (discussed in Section III-D), and base station cooperation, each with more sophisticated limited feedback requirements.

The aforementioned standards implement limited feedback precoding methods that are essentially one-shot quantization techniques. As discussed in Section II-B1, there are adaptive techniques that can exploit temporal correlation in the channel to reduce feedback requirements [18], [19], [170], [213]. There are several different approaches including quantizing gradients [18], [19], using localized codebooks [213], or adaptive codebook structures [170]. Most of this work is for the case of beamforming; adaptive precoding has received less attention. The area of adaptive feedback is still a fertile ground for research.

Multiuser MIMO communication, also called SDMA, was discussed during development of IEEE 802.16m, 3GPP LTE, and 3GPP2 UMB. Only UMB seems to have included it explicitly. A major challenge with limited feedback based multiuser MIMO is that quantization errors create multiuser interference, which can cause throughput ceilings at high SNRs [53], [101], [205]. Reducing these error effects requires either large codebooks, which scale in size with SNR, a substantial amount of multiuser diversity [274], or a combination of multiuser diversity and structured codebooks like with PU2RC [88], [115]. Practically implementing large codebooks remains a challenge. QAM codebooks using fast vector search algorithms are attractive for realizing codebooks [210] while progressive refinement may enable suitable multiple resolution beamforming codebooks [82]. Adaptive methods may also be useful in implementing feedback reduction. The practical realization of large codebooks for multiuser beamforming and precoding is still under investigation.

Base station coordination [231], [232], [280], also called network MIMO [107], creates base station coordination clusters to treat the system like a “super” MIMO system, leveraging high capacity base station backhauls. Interference is eliminated in coordination clusters since users receive signals from all the base stations and there is effectively no interference. Unfortunately, implementing base station coordination on the downlink requires a substantial amount of CSI in the form of every users channel between to every base station in the coordination cluster. Further this information must be exchanged by all the base stations. Research is only now being conducted on issues like codebook based feedback techniques and the impact of codebook size. Given base station coordination is an extension of multiuser MIMO with more effective transmitters and users, it is likely that large codebooks will also be required. Analysis of base station coordination with limited feedback as well as the development of codebook based limited feedback strategies remains a topic of interest.

V. CONCLUSIONS

In this paper, we presented a look at the state of limited feedback research in wireless communication systems. Interest in applications of limited feedback has exploded over the last few years and is sure to grow with the standardization and deployment of 4G and beyond wireless networks.

Many problems still remain. At present there is no general theory of single or multiuser wireless feedback communication networks. This problem may or may not be tractable. It is complicated by issues such as fundamental problems in source coding, interaction between forward and reverse links, effect of delay and the accompanying necessity for small feedback signal blocklengths, effect of errors in the feedback messages, and uncertainty in the optimal way to jointly encode message information with channel state feedback. Practical issues in the deployment of limited feedback systems often tie directly with the quality of the channel model assumed in the system design. Changes and mismatches in the channel distribution must be anticipated and accounted for in a reliable system.

ACKNOWLEDGEMENTS

The authors would like to thank Nihar Jindal, Kyeongyeon Kim, Vasanthan Raghavan, Shahab Sanayei, and Randa Zakhor for reviewing this document. Their comments improved the paper throughout.

REFERENCES


[294] Robert Heath Jr. (S’96 - M’01 - SM’06) received the B.S. and M.S. degrees from the University of Virginia, Charlottesville, VA, in 1996 and 1997 respectively, and the Ph.D. from Stanford University, Stanford, CA, in 2002, all in electrical engineering. From 1998 to 2001, he was a Senior Member of the Technical Staff then a Senior Consultant at Iospan Wireless Inc, San Jose, CA where he worked on the design and implementation of the physical and link layers of the first commercial MIMO-OFDM communication system. In 2003 he founded MIMO Wireless Inc, a consulting company dedicated to the advancement of MIMO technology. Since January 2002, he has been with the Department of Electrical and Computer Engineering at The University of Texas at Austin where he is currently an Associate Professor and Associate Director of the Wireless Networking and Communications Group. His research interests include several aspects of MIMO communication: limited feedback techniques, multihop networking, multiuser MIMO, antenna design, and scheduling algorithms as well as 60GHz communication techniques and multi-media signal processing.

Dr. Heath has been an Editor for the IEEE Transactions on Communication and an Associate Editor for the IEEE Transactions on Vehicular Technology. He is a member of the Signal Processing for Communications Technical Committee in the IEEE Signal Processing Society. He was a technical co-chair for the 2007 Fall Vehicular Technology Conference, general chair of the 2008 Communication Theory Workshop, and is a general co-chair and co-organizer of the 2009 Signal Processing for Wireless Communications Workshop. He is the recipient of the David and Doris Lybarger Endowed Faculty Fellowship in Engineering and is a registered Professional Engineer in Texas.


David Gesbert is Professeur at Eurecom, France. He obtained the Ph.D degree from Ecole Nationale Superieure des Telecommunications, in 1997. From 1997 to 1999, he has been a research fellow at the Information Systems Laboratory, Stanford University. He was a founding engineer of Iospan Wireless Inc, San Jose, Ca., pioneering MIMO-OFDM. From 2001 to 2003 he has been with the University of Oslo as an adjunct professor. D. Gesbert has published about 140 papers and several patents all in the area of signal processing and communications. D. Gesbert co-edited several special issues for IEEE JSAC (2003, 2008), EURASIP JASP (2004, 2007), IEEE Wireless Communications (2006). He is a member of the IEEE Signal Processing for Communications Technical Committee. He authored or co-authored papers winning the 2004 IEEE Best Tutorial Paper Award (Communications Society), 2005 Best Paper (Young Author) Award for Signal Proc. Society journals, and the Best Paper Award for the 2004 ACM MSWiM Conference.

Bhaskar D. Rao received the B.Tech. degree in electronics and electrical communication engineering from the Indian Institute of Technology, Kharagpur, India, in 1979 and the M.S. and Ph.D. degrees from the University of Southern California, Los Angeles, in 1981 and 1983, respectively. Since 1983, he has been with the University of California at San Diego, La Jolla, where he is currently a Professor with the Electrical and Computer Engineering Department and holder of the Ericsson endowed chair in wireless access networks. His interests are in the areas of digital signal processing, estimation theory, and optimization theory, with applications to digital communications, speech signal processing, and human-computer interactions. His paper received the best paper award at the 2000 speech coding workshop and his students have received student paper awards at both the 2005 and 2006 International conference on Acoustics, Speech and Signal Processing conference as well as the best student paper award at NIPS 2006. A paper he co-authored with B. Song and R. Cruz received the 2008 Stephen O. Rice Prize Paper Award in the Field of Communications Systems. He also received the graduate teaching award from the graduate students in the Electrical Engineering department at UCSD in 1998. He was elected to the fellow grade in 2000 for his contributions in high resolution spectral estimation. Dr. Rao has been a member of the Statistical Signal and Array Processing technical committee, the Signal Processing Theory and Methods technical committee as well as the Communications technical committee of the IEEE Signal Processing Society. He currently serves on the editorial board of the EURASIP Signal Processing Journal.