

Channel State Quantization in MIMO Broadcast Systems: Architectures and Codes

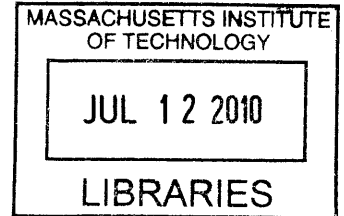
by

Charles Swannack

B.S. Computer Engineering
Clemson University (2003)

S.M. Electrical Engineering and Computer Science
Massachusetts Institute of Technology (2005)

ARCHIVES



Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

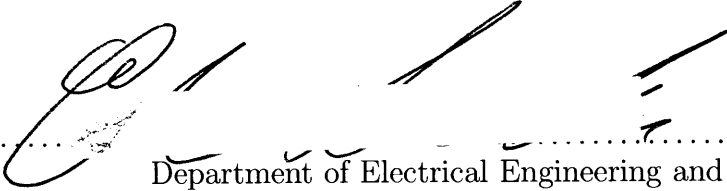
Doctor of Philosophy in Electrical Engineering and Computer Science

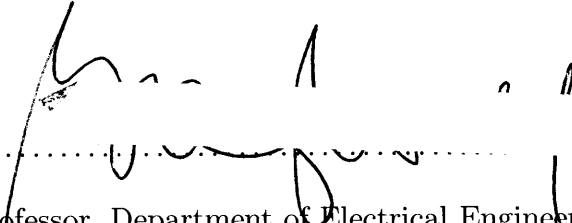
at the


MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2010

© Massachusetts Institute of Technology 2010. All rights reserved.

Author 
Department of Electrical Engineering and Computer Science
March 19, 2010

Certified by 
Professor, Department of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by 
Terry P. Orlando
Professor, Department of Electrical Engineering and Computer Science
Chairman, Department Committee on Graduate Students

Channel State Quantization in MIMO Broadcast Systems: Architectures and Codes

by
Charles Swannack

Submitted to the Department of Electrical Engineering and Computer Science
on March 19, 2010, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

It is now well understood that the use of a multiple-element antenna array at the transmitter can, in principle, greatly increase the capacity of wireless systems. However, little is known about the performance characteristics of such wireless systems in a network setting, or about how to optimize the design of such systems, especially when complexity is taken into account as a practical constraint. This thesis studies the problem of multi-user multiple-antenna broadcast system design with an emphasis on the role that channel feedback plays in a network setting. We develop new design principles for channel feedback design in such systems and show that the system designer is afforded extra degrees of freedom in the choice of the channel quantizer due to the multi-user diversity of the system. As such, the system designer may use the extra degrees of freedom to design structured quantizers that aid in user selection and allow the system to adapt to heterogeneous user populations with different fading characteristics. We construct an adaptive quantization framework which, when paired with low-complexity graph algorithms, enables efficient and robust user scheduling for multi-user multiple-antenna broadcast systems.

Thesis Supervisor: Gregory W. Wornell

Title: Professor, Department of Electrical Engineering and Computer Science

Acknowledgments

This work could not have been completed if not for the wonderful discussions I have had while at MIT. I am particularly grateful for the supervision, support and mentorship of my thesis advisor Greg Wornell. Greg was a great influence in my technical development and provided me with great insights to relevant engineering problems. I am additionally thankful for what have now become quite worn copies of [145] and [146].

I would also like to thank my thesis committee, Professor Vivek Goyal and Dr. Uri Erez for their time and valuable comments. Both have played an important role in my development ever since my arrival at MIT and have always been available for discussion whenever I needed it.

An important element of my grad school experience was my interactions with group members and other students in the signal processing and information theory community at MIT. It is a pleasure to thank them as I complete this thesis. In particular, I feel deeply privileged to have known Anthony Accardi, Petros Boufounos, Albert Chan, Venkat Chandar, Sourav Dey, Vijay Divi, Qing He, Ying-Zong Huang, Everest Huang, Ashish Khisti, Yuval Kochman, James Krieger, Julius Kusuma, Emin Martinian, Urs Niesen, Maryam Shanechi, Aslan Tchamkerten, Elif Uysal, Lav Varshaney, Da Wang and Chen-Pang Yeang. All have played a large part in the development of this thesis.

Most importantly I would like to thank my family, my wife, children, and parents, and dedicate this thesis to them. Without their love, support, and patience this work would not have been possible.

This material is based upon work supported in part by a National Science Foundation Graduate Research Fellowship, the National Science Foundation under Grant No. CNS-0434974, the MITRE Corporation, and by HP through the MIT/HP Alliance.

Contents

1	Introduction	13
2	Multi-User MIMO System Models and Metrics	19
2.1	Single-Antenna Systems	20
2.2	Multiple-Antenna Systems	24
2.3	Figures of Merit for MIMO Channels and Beamforming	32
2.4	Bounds on MIMO System Performance with Finite Rate Feedback	38
3	Systematic Design of MIMO Channel Quantizers	43
3.1	Structured Quantization for MIMO Systems	47
3.2	Systematic Construction of Channel Quantizers	53
3.3	Systematic Construction of Component Codes	71
3.4	Component Codes with Varying Degrees of Orthogonality	89
3.5	Component Codes at Intermediate Rates	104
3.6	Low Complexity Rate Doubling Operations	115
4	Multi-User MIMO System Design with Finite Rate Feedback	125
4.1	A System Architecture to Optimize System Tradeoffs	126
4.2	An Introduction to Channel-Aware Scheduling	129
4.3	Optimization of the Input Occupancy Distribution	143
4.4	Analysis of the Output Occupancy Distribution	149
4.5	Asymptotic Decoupling with the Rayleigh Assumption	159
4.6	Quantizer Performance with Many Users	166
4.7	Practical System Design for Developed Quantizers	172
5	Multi-User MIMO Systems Design with Non-Rayleigh Fading	181
5.1	Modeling the User Assignment Distribution	186
5.2	The EM Algorithm and Homogeneous Class Modeling	199
5.3	Robustness of the Systematic Construction for Multi-User Systems	203
6	Algorithms for Scheduling in Multi-User MIMO Systems	215
6.1	Fast Maximal Clique Algorithms	216
6.2	Complexity of Systematic Quantization Framework	223
7	Conclusions and Future Work	233
7.1	MIMO System Design	235
7.2	Coding and Approximation Theory	236

A	Linear Codes over Rings	239
A.1	Systematic Unitary Space-Time Constructions	242
A.2	Generalized Reed-Muller Construction	243
A.3	Affine-Invariant Constructions	244
B	Bounds on SINR_{sat}	247
B.1	Bounds on SINR_{sat} without Order Statistics	247
B.2	Bounds on SINR_{sat} with Order Statistics	248
C	Proofs	251
C.1	Proofs for Chapter 2	251
C.2	Proofs for Chapter 3	253
C.3	Proofs for Chapter 4	258
	List of Symbols	263
	Bibliography	271

List of Figures

1-1	The MIMO downlink system with an m -antenna transmitter and n uncoordinated receivers.	14
2-1	An illustration of how the shape of the Voronoi cell effects the mean square error for users with isotropic fading.	28
2-2	A plot of the spectral efficiency of each user in a MIMO system with 4 transmit antennas and a given quantization error.	38
3-1	An example of the trade-off between mean squared quantization error and the number of orthogonal bases contained in the code.	44
3-2	The performance of a few channel quantizers for a 4 transmit antenna system which we construct relative to the best known upper bound on SINR_{sat} . . .	46
3-3	The difference in SINR_{sat} between random vector quantization the upper bound (2.44) and various existing constructions for a 4 antenna system. . . .	50
3-4	The difference in SINR_{sat} between random vector quantization the upper bound (2.44) and various constructions for 4 antennas.	52
3-5	A depiction of the general quantization framework for component codes. . .	54
3-6	The cross correlation spectrum of the codewords from Example 3.2.2. . . .	57
3-7	The cross correlation spectrum for the quantizer from Example 3.2.4. . . .	60
3-8	A depiction of the orthogonality relations between the codevectors of Example 3.2.4 as a graph.	60
3-9	Two additional orthogonal bases for the codevectors of Example 3.2.4 as a graph. Here two vectors from basis \mathcal{B}_1 have been swapped with two vectors from \mathcal{B}_2 so that the resulting sets remain orthogonal.	61
3-10	The performance of random vector quantization and the sequence of codes $\mathcal{C}_{\text{sparse}}^{(2,4)}(k)$	62
3-11	An illustration of the poor performance of the sequence of sparse codes. . .	63
3-12	The performance of random vector quantization and a sequence of systematic constructions of codes constructed by first taking the union of sparse and dense codes then increasing the cardinality of the integer ring underlying the construction of each of the component codes in the union.	65
3-13	A depiction of the code in \mathbb{R}^3 that corresponds to the vertices of the icosahedron and an associated universal code	67
3-14	The performance of random vector quantization and our complete systematic constructions of codes.	68

3-15	A depiction of the systematic construction of the 5-bit quantizer $\mathcal{C}_{\text{ASC}}^*(2, 2)$ and the 10-bit quantizer $\mathcal{C}_{\mathbf{F}}(\alpha, \gamma, \mathcal{C}_{\text{ASC}}^*(2, 2))$	70
3-16	The relation of the parameters of our general construction to our geometric interpretation.	73
3-17	A depiction of the actions of $\mathbf{T}(\boldsymbol{\lambda})$ and $\mathbf{S}(\tilde{\boldsymbol{\beta}})$ on the codebook $\mathcal{C}(\Upsilon_1, \Upsilon_2; L)$	75
3-18	A depiction of the actions of $\mathcal{H}_{L,a}$ and $\mathcal{H}_{L^c,a}$ on two complimentary codes $\mathcal{C}(\Upsilon_1, \Upsilon_2; L)$ and $\mathcal{C}(\tilde{\Upsilon}_1, \tilde{\Upsilon}_2; L^c)$	78
3-19	A depiction of the performance of two 6 bit quantizers in \mathbb{C}^4	79
3-20	A depiction of the relationships between the four orthogonal bases of Example 3.2.4.	86
3-21	An illustration of the orthogonal sets of the code from Example 3.3.3.	88
3-22	An illustration of the orthogonal sets of the code from Example 3.3.4. Note that this shows only 4 non-intersecting orthogonal bases while the code of Example 3.3.3 had 12 orthogonal bases.	90
3-23	The cross correlation spectrum of the quantizers from Example 3.3.3 and Example 3.3.4.	91
3-24	An example of the orthogonality relations between codewords of the quantizer developed using the lift $\vartheta_{\mathcal{I}}(x)$ in 8 complex dimensions.	105
3-25	An example of the cross correlation spectrum of the quantizer developed using the lift $\vartheta_{\mathcal{I}}(x)$ in 8 complex dimensions.	106
4-1	The MIMO system architecture of interest.	128
4-2	Two 8 input and 4 output input-queued cross-bar switches.	132
4-3	A depiction of the input-queued cross bar switch in which users are randomly assigned to switch inputs at each scheduling interval.	134
4-4	A single matching representing a given processing mode $k \in K(\mathbf{m})$	137
4-5	A depiction of the static generalized switch of Example 4.2.1 as a graph	139
4-6	An alternate view of a generalized switch of Stolyar in the case of finite rate feedback.	140
4-7	An alternate view of Stolyar's generalized switch for channel aware scheduling with finite rate feedback as a best random server process.	142
4-8	The quantization order, $n_{\delta}(\alpha)$, as a function of δ and α for a few distributions of interest.	147
4-9	A view of the statistical dependencies of switch outputs in the BRS model as a three level urn process.	152
4-10	The trade-off between $p_{\sigma,\rho}$ and $p_{\mathcal{G}}$ predicted by Theorem 4.4.3 for $n = 8, 12, 16, 24$ with 4 transmit antennas.	157
4-11	The trade-off between $p_{\sigma,\rho}$ and $p_{\mathcal{G}}$ predicted by Theorem 4.4.3 for $n = 16, 24, 32, 48$ with 8 transmit antennas. The smallest number of users is at top and the largest at bottom. Note, even when using the large deviation bound of Theorem 4.4.2 the plots show a rapid transition from 0 to 1 so long as $p_{\sigma,\rho} > 0.4$	158
4-12	Two possible arrangement of 12 lines in \mathbb{R}^3 which in the absence of order statistics have differing mean squared quantization error.	167
4-13	The two arrangement of 12 lines in \mathbb{R}^3 from Figure 3-1 where spherical caps of equal half angles are depicted around the codewords.	169

4-14	The upper bound $\text{SINR}_{\text{sat}}^{(\text{UB})}(n, \ell)$ in a 32 user system for various values of ℓ as well as the upper bound on $\text{SINR}_{\text{sat}}^{\text{UB}}(32, \ell)$, (B.6a). Note that the for a large number of bits there is an approximately equal slope for each curve with a fixed offset due to the number of users selected as predicted by (B.6a).	170
4-15	The upper bound $\text{SINR}_{\text{sat}}^{(\text{UB})}(n, 4)$ in a n user system for various values of r as well as the upper bound on $\text{SINR}_{\text{sat}}^{\text{UB}}(n, 4)$, (B.6a). Note that the growth in the SNR is linear in $\log_2 m$ with slope $3/(m - 1) = 1$ as predicted by (B.6a). The linear growth in r predicted by (B.6a) may also be observed through the difference of every pair of curves (lines).	171
4-16	The performance of existing and developed quantizers in \mathbb{C}^4 relative to the upper bound $\text{SINR}_{\text{sat}}^{\text{UB}}(n, \ell)$ in a 32 user system. We note that as all users are considered the achieved performance is independent of the number of users in the system.	174
4-17	The performance of existing and developed quantizers in \mathbb{C}^4 relative to the upper bound $\text{SINR}_{\text{sat}}^{\text{UB}}(n, \ell)$ in a 32 user system where only the 16 users which achieve the highest value of SINR_{sat} at each scheduling interval are considered.	175
4-18	The performance of existing and developed quantizers in \mathbb{C}^4 relative to the upper bound $\text{SINR}_{\text{sat}}^{\text{UB}}(n, \ell)$ in a 32 user system where only the 8 users which achieve the highest value of SINR_{sat} at each scheduling interval are considered.	176
4-19	The performance of existing and developed quantizers in \mathbb{C}^4 relative to the upper bound $\text{SINR}_{\text{sat}}^{\text{UB}}(n, \ell)$ in a 32 user system where only the 4 users which achieve the highest value of SINR_{sat} at each scheduling interval are considered.	177
4-20	The performance of existing and developed quantizers in \mathbb{C}^4 relative to the upper bound $\text{SINR}_{\text{sat}}^{\text{UB}}(n, \ell)$ in a 16 user system where only the 8 users which achieve the highest value of SINR_{sat} at each scheduling interval are considered.	178
4-21	The performance of existing and developed quantizers in \mathbb{C}^4 relative to the upper bound $\text{SINR}_{\text{sat}}^{\text{UB}}(n, \ell)$ in a 16 user system where only the 4 users which achieve the highest value of SINR_{sat} at each scheduling interval are considered.	179
4-22	The performance of existing and developed quantizers in \mathbb{C}^4 relative to the upper bound $\text{SINR}_{\text{sat}}^{\text{UB}}(n, \ell)$ in a 8 user system where only the 4 users which achieve the highest value of SINR_{sat} at each scheduling interval are considered.	180
6-1	An illustration of the importance of the input vertex order for the algorithm of Carraghan and Pardalos.	221
6-2	An illustration of the results of a vertex ordering which excludes every maximally sized clique after 4 iterations. The ordering is taken from a coloring of the graph \mathcal{G}	222

Introduction

Wireless communication systems have seen remarkable growth in the past century with an impressive rate of expansion in the past few decades. In 1895, Guglielmo Marconi succeeded in establishing the first documented wireless communication link via radio signals sending Morse code (i.e. a sequence of dots and dashes) over a wireless channel to a receiver 18 miles away. While Marconi's message was accurately received over this long distance, at the time there was little understanding on the fundamental limits of the wireless signaling and the rate at which one could transmit over such a wireless channel reliably. It was not until Shannon's pioneering work in 1948 on the capacity of the additive white Gaussian noise (AWGN) channel that communication engineers understood the fundamental limits on the communication rate for reliable transmission [108]. Presently, third and fourth generation communications technologies are being designed to push the limits of the wireless channel aiming to deliver data rates of up to 100 Mbit/s. More ambitiously, system designers are developing wireless system to replace the standard wired last mile of service providing a wireless alternative to cable modems and digital subscriber lines, a wireless backbone for Wi-Fi (IEEE 802.11) hotspots as well as providing general telecommunications and data services. The current IEEE 802.16 standard (WiMAX) aims to deliver local as well as metropolitan network service where the base stations are mounted on homes or buildings rather than towers. Current development of such next-generation wireless networks call for the support of a wide variety of data services. To provide this functionality, the current IEEE 802.16 Standard [1] provides a high-rate framework aimed to replace conventional last mile of networking with a wireless link which provides five quality-of-service (QOS) classes, three for real-time data connections and an additional two classes for delay tolerant [1]. Thus, state of the art wireless design is a problem of cross-level design where both the aspects of the network as well as the physical channel must be considered as part of the design.

It is now well understood that the use of a multiple-element antenna array at the transmitter can, in principle, greatly increase the capacity of such systems. However, little is known about the performance characteristics of such wireless systems in a network setting, or about how to optimize the design of such systems, especially when complexity is taken into account as a practical constraint. The richness of this system design problem stems from the fact that it is one of spatio-temporal scheduling, *i.e.*, both *temporal scheduling* and *spatial multiplexing* aspects of the design must be considered. This thesis investigates key aspects of joint scheduler-multiplexer design problem for multi-input multi-output (MIMO) systems, focusing on the problem of delivering high throughput as well as broader quality of service (QOS) guarantees while being subject to complexity and limited feedback constraints. In particular, we consider such a system when the number of users who must be served is greater than the number of elements in the antenna array. A general depic-

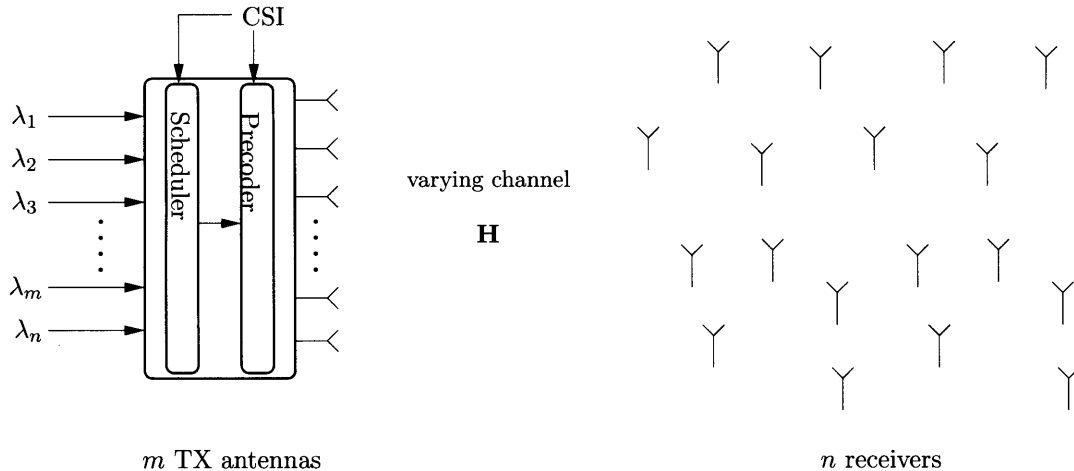


Figure 1-1. The MIMO downlink system with an m -antenna transmitter and n uncoordinated receivers each having a single receive antenna. Arrivals occur at the beginning of every scheduling interval and each arriving packet is destined for a single-user.

tion of this network scenario may be seen in Figure 1-1. An m -antenna transmitter and n uncoordinated receivers each having a single receive antenna are distributed throughout some geographic area. At the beginning of each scheduling arrivals occur at the transmitter destined for a single-user. These messages are placed in a queue for the appropriate user. Then, based on the state of both the queue as well as the channel, the transmitter precodes and transmits messages for a subset of users.

Current MIMO systems must be developed in a way as to be robust to a variety of radio environments to be easily (and quickly) deployed on a large scale. To do such a system designer may design a system under some minimum number of assumptions (for example number of users, user mobility etc.) while leaving free a few degrees of freedom in the design which may be set independently at each deployment site. An even more desirable approach is to design a system that may infer these parameters through some set of minimal training data as this removes much of the complexity of system deployment as well as provides the system with the ability to adapt to possible future changes in the radio environment. A simple approach to provided this functionality is to design a feedback link for users of the system to report the current state of their radio channel to the transmitter. It is well known that this approach (knowledge of the channel state at the transmitter) can yield considerable increase in the system throughput and hence should be incorporated in system design. However, little is known about how this feedback effects QOS in this broader network context or how to optimally design this feedback link in a variety of radio environments.

In this thesis we consider how to design efficient MIMO systems with a particular emphasis on the role of channel feedback plays in this broader network context. In particular, we examine present feedback design rules used currently in the IEEE 802.16 standard [1] when the broader network problem as well as the overall system complexity is considered. As the design of the feedback link is intimately tied to the radio environment we additionally present methods which allow a system to adapt an existing channel feedback design to more closely match the characteristics of the fading process at a given deployment site in a way that boosts transmission rates while keeping the overall system complexity low.

In a MIMO system where the number of users exceeds the number of transmit elements one expects to improve a particular performance criterion as the number of users in the system grows. In particular, by taking a heuristic approach [6, 7, 49, 107] or restricting the scheduler to a pure time-division strategy, whereby at most one user is selected to transmit to at any time, it has been shown that the system throughput can be improved by selecting the user with the highest signal to noise ratio (SNR). Under infinite backlogs such a schedule can maximize throughput for a *single-antenna* broadcast channel [80], provided the channel is allocated to the strongest user at any time. However, for the multiple-antenna broadcast channel, such approaches, while low in complexity, fail to exploit the large available throughput gains from spatial multiplexing. In particular, it is clear that in a system where the transmit array has multiple elements one may more generally consider selecting a subset of users for which the corresponding sub-channel achieves the highest rate.

With perfect channel state information (CSI) at the transmitter and infinite backlogs, a throughput maximizing scheduling/multiplexing scheme is to successively encode (i.e. employ dirty paper coding) the set of users which at that time can achieve the highest sum rate. Such an encoding strategy has a quite high complexity as it is sensitive to the order in which users are encoded and lower complexity solutions are of interest in practical systems. When many users are present in the system it is reasonable to expect that there is a subset of users that negligible interfere with one another. As such, it is reasonable to expect that lower complexity multiplexing schemes will achieve a similar rate to that of successive encoding for such a set. However, there is no guarantee that a subset of users that negligible interfere with one another will in general be the subset of users that achieve the highest rate and in general all subsets of users will have to be considered. This search has high algorithmic complexity in the number of users n and transmit dimension m , and for purposes of implementability it is of interest to find lower-complexity solutions. The complexity of such an optimization is dominated by the underlying search for the best user subset to multiplex across the transmitter array, which must be performed each time an arriving packet or channel variation changes the system state. To reduce this complexity, one may limit the search to a smaller pool of users while ensuring that a set of users can be found in this restricted pool that obtains a sum rate which is close to optimal with high probability. When the number of users in the system approaches infinity it has been shown that both successive encoding and a much lower complexity multiplexing strategy achieve the optimal scaling in rate [110, 111]. More precisely, the ratio of the rates achieved by successive encoding and a random beamforming strategy tend to one as the number of users tends to infinity. However, this says little about the actual system performance for a user population of fixed size which employs a sub-optimal multiplexing or the complexity of the search for a subset of users that are nearly orthogonal that can be multiplexed with a chosen sub-optimal multiplexing scheme with negligible penalty in rate.

In a practical system finding the set of users that achieve the maximum sum rate may not be feasible due to complexity constraints. Thus, it is of interest to develop scheduling algorithms that choose a set of users who achieve a rate close to that of the optimal set with as few operations as possible. It has been recognized that using sets for which there are guarantees on the channel norms and the magnitudes of pairwise inner products can provide close to optimal performance [111, 120–124, 131, 140–142]. Such an approach aims to find a set of users that are nearly orthogonal so that the penalty in rate incurred using a sub-optimal multiplexing scheme will be negligible for the *selected set*. However, in general there is no guarantee that a nearly orthogonal set will in fact be the optimal set and one

in general expects to pay some price in SNR for using such multiplexing and scheduling. In particular, it is not clear that one may simultaneously find a subset of users with good channel gains that simultaneously are nearly orthogonal. More precisely, with this greedy approach to reduce the interference between users there are two competing forms of multi-user diversity:

1. the order statistic gain, the multi-user diversity stemming from one's ability to schedule the users that are individually at high SNR
2. the multi-node matching gain, the multi-user diversity stemming from one's ability to schedule users that negligibly interfere with one another

For example, if one attempts to select only the users whose channels are individually at high SNR it may not be possible to find a subset of users that are nearly orthogonal. Alternatively, if one first searches for sets that are nearly orthogonal it may not be possible to select users that are individually at high SNR. Hence, for this greedy approach and more generally in the interest of system complexity, it is of interest to understand when these two problems decouple. In the sequel we say that the order statistic gain decouples from the multi-node matching gain if, with high probability, one is able to find a subset of users that are nearly orthogonal from the restricted pool of users that are individually at high SNR. We have shown that in the large user limit and a fixed number of transmit antennas the order statistic gain decouples from the multi-node matching gain [123] (i.e. [123] shows that asymptotically one can first select users based off their individual statistics then find an *orthogonal set* from the resulting population). However, it is not clear how large n must be for these asymptotic insights to be relevant for system design. In particular, for a user pool of fixed size it is not clear how to design systems to jointly optimize the order statistic gain and multi-node matching gain. Moreover, as obtaining exact knowledge of the channel is unrealistic in many MIMO channels due to bandwidth limitations on the feedback link, in practice each user terminal must quantize the observation of its channel and feed back this representation to the transmit base. Thus, in practice one must additionally optimize the trade-off between the rate and structure of the quantizer and the order statistic gain and multi-node matching gain.

In a multi-user MIMO system knowledge of the channel state at the transmitter is necessary to realize the multiplexing gain. However, when finite rate feedback is used to convey a users channel state to the transmitter there is some uncertainty at the transmitter of each users channel state. Hence, the transmitter can not employ an intelligent multiplexing method to fully eliminate the co-channel interference. As this interference scales linearly with the SNR one must decrease the co-channel interference proportionally with the SNR, leading to the need to linearly increase the feedback rate. Thus, high rate systems with few users and finite rate feedback must use large codebooks to ensure that the system performance is not limited [68]. In such cases it is of interest to develop *structured* codebooks that enable user terminals to efficiently quantize their channel vectors. One of the most crucial insights to our development in the sequel is that when the *number of users is greater than the number of transmit elements the diversity of the system decreases the uncertainty at the transmitter of each users channel state for a subset of the user pool largely independent of the particular feedback design allowing the system designer to rather focus the feedback design on increasing the transmitter's knowledge of the co-channel interference offsetting the rate of scaling of the feedback per user*. This results stems from our asymptotic development that when the number of users tends to infinity a quantizer which consist of a single

orthonormal basis is sufficient to achieve the optimal rate. Thus, in a multi-user system there is a natural desire to develop quantization schemes that contain orthogonal bases as well as have good mean square error characteristics.

In Chapter 5, we show that in a multi-user MIMO system with finite rate feedback the effect on the expected quantization error from adding an additional user to the system roughly equals the effects of adding an additional codeword to an optimally designed codebook. Thus, multi-user diversity in a system makes the constraint that a quantizer be good in terms of the quantization error largely irrelevant and a system designer may choose a quantization scheme that helps identify users with low co-channel interference to boost the achieved signal-to-interference-to-noise ratio (SINR). However, to make this precise we must first overview our system model. This is done in Chapter 2 after a brief summary of this thesis.

Thesis Outline and Contributions

We have identified the problem of feedback design as a central issue in reducing the complexity of both multiplexing as well as user selection in a multi-user MIMO channel. Hence, in the subsequent sections we develop a systematic framework to treat both of these issues so that the trade-off between the complexity reduction of multiplexing as well as user selection can be optimized by the system designer. We note that these trade-offs are highly dependent on the particular model assumed for the MIMO channel. Thus, this thesis will develop the relevant insights needed for system design in two parts. First, which includes the majority of this thesis, we assume that the MIMO channel is isotropic and develop a framework for feedback design and scheduling under this assumption. Then, we consider a class of channel models for which a system developed assuming an isotropic channel model will degrade and develop a simple method with which a system may adapt the feedback framework to compensate.

We begin this development by first making the model for our system precise in Chapter 2. Then, we proceed to develop a systematic structured finite rate feedback framework in Chapter 3 which can be used to balance the trade-off between the mean squared quantization error and the number of orthogonal bases contained in the quantizer. Then, in Chapter 4 we present a simple model and associated base station architecture in which the system designer may study the trade-off between the order statistic gain and the multi-node matching gain and how this trade-off is affected by the variations in the structure of the feedback design. Further, in Chapter 4, we present efficient algorithms for user selection that exploit the structure of our systematic feedback. A benefit of the models and algorithms of Chapter 4 is that they additionally allow one to examine the effects that variations in the channel model have on the system performance of such a system. As such, we proceed to identify the relevant statistical models for the fading process in multi-user MIMO systems as well as present a discrete model for user feedback in Chapter 5. Further, in Chapter 5, we show that our systematic feedback framework of Chapter 3 may also be viewed as a method to adapt the channel feedback to better match the covariance structure of the channels which significantly degrades system performance. This has practical relevance as the feedback framework of Chapter 3 provides a common framework in which one may simultaneously develop good structured high rate quantizers as well quantizers that may adapt to unknown channel covariances.

To summarize, the major contributions of this thesis are:

1. Identifying the problem of feedback design as an integral part of the joint design of efficient channel aware schedulers as well as robust low complexity multiplexing schemes (Chapter 1)
2. Providing a systematic feedback framework in which the system designer may trade-off between the order statistic gain and the multi-node matching gain to meet certain system objectives (Chapter 3)
3. Providing a simple base station architecture to understand to trade-off between the order statistic gain, the multi-node matching gain and system complexity (Chapter 4)
4. Identifying an appropriate discrete model for user feedback and identifying an associated expectation-maximization algorithm to estimate this distribution under unknown channel conditions and identify clusters of users with similar channel correlation (Chapter 5)
5. Providing a systematic method to adapt our feedback framework so that the resulting design remains stable as the statistics of the underlying channel change (Chapter 5)
6. Providing a new class of algorithms for user selection that exploit the structure of our feedback framework to solve the user scheduling problem (Chapter 6)

We now proceed to provide our model for the multi-user MIMO system of interest before moving to our new design for the feedback problem.

Multi-User MIMO System Models and Metrics

Wireless communications systems continue to develop providing increased data throughput and enhanced quality of service. However, wireless transmission is hampered by the time varying nature of the channel of which the transmitter only has causal knowledge. Such channel variations are caused by the underlying physical structure of the environment for which an electromagnetic information-bearing signal propagates between a transmit and receive pair. In the most simple scenario where the electromagnetic signal propagates through free space in the absence of physical structure one minimally expects the signal to be attenuated proportionally to the inverse square of the distance between the transmit and receive pair. However, current high rate wireless systems are being deployed in urban environments that are wrought with obstacles and as such one would expect that the channel variations to be much different than that experienced in free space. In particular, wireless signal propagation may be affected by [100]

1. reflections which occurs when a electromagnetic wave impinges upon a smooth surface of much larger size than the signal wavelength
2. diffractions which occur when dense bodies of size greater than the signal wavelength are present in the propagation path between the transmit and receive pair
3. scattering which occurs when a electromagnetic wave impinges upon either a rough surface of size greater than the signal wavelength or any surface whose size is on the order of the wavelength

In the sequel we do not need to distinguish between these effects but rather their bulk effect and refer to reflection, diffraction and scattering simply as scattering and the objects causing these effects as scatterers.

If there are scatters in the propagation path between the transmitter and receiver one expects the received waveform to be attenuated. However, the particular scale of this attenuation is dictated by the particular number, position and physical properties of these scatters. Given the location and relevant material properties of the scatters along the propagation path one could solve¹ the relevant wave equations to find the signal attenuation between the transmitter and receiver. However, any change in this geometry, whether due to the mobility of the receiver, transmitter or other scatter can dramatically alter the signal attenuation over a short period of time. The length of time one may assume this

¹Minimally, one can find a close numerical approximation via finite difference methods or the method of moments [8]

fading is constant we call the coherence² time of the channel. In a channel with such fluxuations, one may not be able to ensure the channel is in a sufficiently good state for reliable transmission due to the signal attenuation. Moreover, as in practice one does not know the precise characteristics of every scatter for every deployment site one generally forms parametrized statistical models that can characterize a variety of possible propagation environments. That is, in practice one more generally seeks to form statistical models for the channel fluxuations based on some basic assumptions on the dynamics of the transmitter and receiver as well as the distribution and dynamics of the scatterers. To this end one needs to understand how the modeling assumptions and prior information given about the channel effect the figure of merits used to measure the performance of the system and avoid any assumptions that are not supported by prior information that unduly influences these figures of merit. In the sequel, we first consider the relevant figures of merit and models for a single-antenna system before returning to the more general question of multiple antenna systems.

■ 2.1 Single-Antenna Systems

Wireless communication systems have seen remarkable growth in the past 100 years, in large part, due the ability for one to accurately model and predict the relevant aspects of the wireless communication channel. In 1895, Guglielmo Marconi succeeded in establishing the first documented wireless communication link via radio signals using a very fundamental understanding of the electromagnetic associated to radio wave propagation which enabled him to send Morse code (i.e. a sequence of dots and dashes) over a wireless channel. However, at the time there was little understanding on the fundamental limits of the wireless signaling and the rate at which one could transmit over such a wireless channel reliably. It was not until Shannon's pioneering work in 1948 on the capacity of the additive white Gaussian noise (AWGN) channel that communication engineers understood the fundamental limits on the communication rate for reliable transmission [108]. In particular, Shannon considered the discrete time power-constrained AWGN channel given by

$$y[k] = x[k] + z[k] \quad (2.1)$$

where the power constraint is $\frac{1}{n_b} \|\mathbf{x}\|^2 \leq P$ (n_b being the block length) and where the noise $z[k]$ is a zero mean Gaussian random variable with variance σ^2 . Shannon showed that for sufficiently long transmissions one may signal at a rate that scaled linearly in the spatial degrees of freedom reliably with a nominal spectral efficiency of

$$\log_2 \left(1 + \frac{P}{\sigma^2} \right) \text{ bits/complex dimension.}$$

However, the capacity of the AWGN channel is not sufficient to fully characterize a wireless channel as in general the channel fluxuations led to a time varying signal quality.

A wireless communication system is subject to not only the thermal additive noise effecting wireline channels, but also from the structure of the propagation environment of the signal. In particular, in urban environments the location and geometry of near by buildings and other scatterers may introduce self interference due to copies of the same signal arriving

²The channel coherence is normally only used to describe a narrow band block fading channel which we introduce in the sequel.

at the receiver delayed in time due to the increased path length caused by the scattering. Hence, in the absence of a relevant model of the propagation of the electromagnetic signal it is unclear how to address the limits of such a wireless communication system. Moreover, dynamics of the system may make the overall channel input response time varying. That is, the input-output relationship for the wireless channel in general must be described as [128]

$$y(t) = \int_{-\infty}^{\infty} h(\tau, t) d\tau.$$

where $h(\tau, t)$ is the time varying channel impulse response. Assuming a multipath channel with finitely many scatterers one may write the channel impulse response, $h(\tau, t)$, as

$$h(\tau, t) = \sum_i a_i(t) \delta(\tau - \tau_i(t))$$

Sampling the channel outputs at multiple of $1/W$, where W is the system bandwidth, the resulting baseband discrete time model for the channel becomes [128]

$$y[k] = \sum_n x[n] \sum_i a_i^b(m/W) \text{sinc}[m - n - \tau_i(m/W)W] \quad (2.2)$$

where

$$a_i^b(t) = a_i(t) \exp(-2\pi\sqrt{-1}f_c\tau_i(t))$$

and in turn where f_c is the carrier frequency of the signal.

In a wireless system with sufficiently high bandwidth, the scattered signals, which arrive at the receiver delayed in time, may be resolved and coherently combined for a gain in overall received signal power. However, in more narrow band systems the delayed signals can not be resolved and combined and either constructively or destructively attenuates the received signal. Such attenuation of the received signal we refer to as fading. It is important to note that by (2.2) the position, number and dynamics of the scatterers completely determine the signal propagation at a given frequency. However, in general the transmission frequency influences the signal propagation. It is reasonable to suspect that, at least in the cases of interest, frequency response of the channel at near by frequencies will be quite similar and hence attenuate a signal equally over narrow frequency band. If the transmitted signal is attenuated approximately equally over the frequency band used for transmission we say that the system experiences flat fading. The largest possible bandwidth that can be used while ensuring a flat fading behavior is called the coherence bandwidth. In this thesis we assume a narrowband system for which the transmission bandwidth is less than the coherence bandwidth so that the resulting system experiences flat fading.

In a narrowband flat fading channel with a single transmit element the complex discrete time baseband model for each user in the system is:

$$y_i[k] = h_i^*[k] \cdot x[k] + z_i[k] \quad (2.3)$$

where $y_i[k]$ is the received signal, $x[k]$ is the transmitted signal, $h_i[k]$ is the channel fading coefficients and $z_i[k]$ is independent identically distributed (i.i.d.) $\mathcal{CN}(0, 1)$ noise, and where the channel gain $h_i[k] \in \mathbb{C}$. The noises are independent from receiver to receiver, from block to block and further are independent of the channel gains. The transmitter is subject to an average total power constraint P . In a single-antenna system the instantaneous signal to

noise ratio (SNR) is of interest as it describes the instantaneous capacity of a user's channel. For the k -th signaling interval the instantaneous signal to noise ratio (SNR) is

$$\text{SNR}_i[k] = \frac{P \cdot |h_i[k]|^2}{\sigma^2}$$

which leads to a corresponding time varying spectral efficiency for the i -th user of

$$\log(1 + \text{SNR}_i[k]) \text{ bits/complex dimension} \quad (2.4)$$

If the channel varies rapidly the coherence time may not be long enough to enable reliable transmission over a single interval. Hence, one can transmit over multiple fades to achieve some overall performance which, if the transmission occurs over sufficiently many realization of the fading coefficient, typically becomes deterministic. For such a transmission approach, the expected rate

$$C_{\text{ergodic}} = \mathbb{E}_{\mathbf{h}} [\log(1 + \text{SNR}[k])] \quad (2.5)$$

is the relevant figure of merit which we call the ergodic capacity. The ergodic capacity alone does little to guarantee the channel quality at a particular instance in time will be good but rather measures the quality of the signal over several fades. We are interested in the role feedback plays in the broader design and in the sequel we assume a flat fading model with a sufficiently long coherence time to allow for reliable transmission with in each fading block.

If the fading process varies slowly, i.e. if the coherence time of the channel is significantly long, one may transmit a signal over a single fade reliably at a rate determined by the fading process. In practice one uses one or more fixed rate coding schemes to ensure reliability. When the channel quality drops below the SNR threshold for which the fixed rate coding scheme can be used reliably one will have a high probability of bit errors. Thus, in practice there is some SNR threshold τ_0 for which communication can not be performed reliably and the probability that the SNR is not sufficiently high as to not support reliable transmission is an important figure of merit. We call this the outage probability. More precisely the outage probability is

$$P_{\text{outage}}(\tau_0) = \Pr [\text{SNR}[k] \leq \tau_0]. \quad (2.6)$$

While the ergodic capacity does not directly relate to the problem of interest it is important to note that the outage probability has a very useful interpretation in the problem of channel aware scheduling. In particular, a single-user system for which the outage probability for every selected user is low implies that the service rate for the wireless link becomes approximately constant. Thus, the outage probability can alternatively be viewed as a coarse measure of how strongly coupled the particular channel realization is to the scheduling decision. That is, if a system employs a given fixed rate coding scheme and the probability of outage is low then one only needs to first determine the subset of users that are over this threshold then do a simple weight matching. As noted in Chapter 1, under infinite backlogs opportunistic scheduling of the system (allocating the channel to the strongest user at any time) can maximize throughput for this *single-antenna* broadcast channel [80]. Thus, in such a system one is more generally interested in a generalized notion of outage

$$P_{\text{fail}}(\tau_0) = \Pr \left[\left(\max_{i=0,1,\dots,n-1} \text{SNR}_i[k] \right) \leq \tau_0 \right]. \quad (2.7)$$

It is clear that in such a multi-user single-antenna broadcast channel one may have a dramatically lower probability of outage for the selected user assuming that $P_{\text{fail}}(\tau_0)$ is low as

$$P_{\text{outage}}(\tau_0) > P_{\text{fail}}(\tau_0).$$

While this identification is not needed to reduce the complexity in a single-user system it will help us substantially reduce the system complexity for a multiple-antenna system in the sequel. This is exactly the perspective that led to our definitions of order statistic gain and multi-node matching gain in Chapter 1 and we wish to develop a similar definition to $P_{\text{fail}}(\tau_0)$ for the multi-user MIMO channel.

For the transmitter to make an informed scheduling decision, as in (2.7), the transmitter must have some knowledge of each users signal strength. If the channel is not reciprocal (i.e. the propagation characteristics from the transmitter to receiver is not identical to that from the receiver to the transmitter) then the transmitter must receive some sort of feedback from the users to indicate their signal strength for inference of the channel state to be possible. Moreover, each user must be able to measure their signal strength for such feedback to be possible. Throughout this thesis we assume that each user has perfect knowledge of their channel state and that some imperfect representation of this channel state is known by the transmitter. In particular, we assume that each user has fed back some quantized representation of the fading state through a finite bandwidth communication link. In this thesis we do not consider the design of this link nor do we consider how much bandwidth is needed by such a link. Rather, we assume that this feedback link has been sufficiently designed so that every transmission occurs without error and examine how the rate of the associated quantization scheme affects the system throughput.

As seen in the single-antenna broadcast system the figures of merit (for both the outage probability as well as the ergodic capacity) rely heavily on the distribution of the fading process and hence one must accurately model the fading process for the results to be meaningful. In a single-antenna system the effects of user dynamics and the geometry of the propagation environment are well understood [100, 128]. However, in the MIMO channel there are far more effects that must be modeled which not only effects the system throughput but also the feedback design. In particular, one must model the effects of the array geometry, electromagnetic coupling of the transmit elements as well as the co-channel interference between the different users.

In order to model the co-channel interference in a multi-user MIMO system one in general must understand effects the propagation environment has on the users in the system. In particular, one must model the effects the propagation environment has on the co-channel interference. As the multiple transmit elements led to more propagation paths the problem of modeling the multi-user MIMO channel is far more complex than a system with a single element. This modeling problem is compounded by the many different propagation environments for which current multi-user MIMO devices and standards are being designed. In particular, as the current IEEE 802.16 standard has modes of operation for urban, suburban, and rural radio transmission and it is not clear what assumptions can be made about the multi-user MIMO channel, or more generally, the number of degrees of freedom available in the multi-user MIMO channel. In the absence of strong modeling it seems that a system designer must make either too strong or too weak assumptions on the channel model which may be overly optimistic or pessimistic causing poor performance at one or more deployment sites. However, in the sequel we show that one may design robust multi-user MIMO systems by constructing quantizers for an isotropic channel which have a large

degree of symmetry. Thus, in the sequel we provide a brief introduction to multiple-antenna systems and proceed to design quantizers that perform well assuming isotropic fading. Then, in Chapter 5 return to the question of modeling multi-user MIMO channel more generally.

■ 2.2 Multiple-Antenna Systems

Current MIMO wireless systems have shown the potential for increasing wireless system capacity without the price of power or bandwidth [126]. These results stem from the fact that a MIMO channel allows the construction of parallel communication channels that are separated in space affording path diversity for the transmitted signal. As the transmitted signal follows multiple spatial paths it is likely (under reasonable assumptions) that each path does not simultaneously undergo poor fading and hence common figures of merit used to measure the performance of single-antenna systems (i.e. the ergodic capacity as well as the outage probability) are likely to be improved. In particular, [126] has shown that in a rich scattering environments the resulting ergodic capacity scales approximately linearly in the minimum of the number of transmit and receive antennas.

In a MIMO system with m transmit antennas and n receive antenna one must in general consider all transmit and receive pairs to accurately model the channel. More precisely, in order to derive the input-output relationship for a MIMO system one must generally derive the input-output relationship for each transmit and receive pair. That is, the general input-output relationship for the MIMO channel is

$$\mathbf{y}(t) = \int_{-\infty}^{\infty} \mathbf{H}(t, \tau) \mathbf{x}(t - \tau) d\tau + \mathbf{z}(t) \quad (2.8)$$

where $\mathbf{y}(t)$ is the vector of received signals for the users, $\mathbf{x}(t)$ it the signal transmitted from the array, $\mathbf{z}(t)$ is the time varying noise process and in turn where

$$\mathbf{H}(t, \tau) = \begin{bmatrix} h_{0,0}(t, \tau) & h_{0,1}(t, \tau) & \cdots & h_{0,m-1}(t, \tau) \\ h_{1,0}(t, \tau) & h_{1,1}(t, \tau) & \cdots & h_{1,m-1}(t, \tau) \\ \vdots & \vdots & \ddots & \vdots \\ h_{n-1,0}(t, \tau) & h_{n-1,1}(t, \tau) & \cdots & h_{n-1,m-1}(t, \tau) \end{bmatrix} \quad (2.9)$$

is the time varying impulse response of the channel. In the sequel, we assume that each one of these mn links are narrowband flat fading with a sufficiently long coherence time to allow for reliable transmission with in each fading block. More precisely, we assume a narrowband discrete-time channel model that is block fading where, in any particular block, the signal $y_j[k]$ received by user j at time k in response to a signal $\mathbf{x}[k]$ transmitted from the array is of the form

$$y_j[k] = \mathbf{h}_j^\dagger[k] \mathbf{x}[k] + z_j[k] \quad (2.10)$$

where $z_j[k]$ is independent identically distributed (i.i.d.) $\mathcal{CN}(0, 1)$ noise, and where the (normalized) channel gain vectors $\mathbf{h}_j[k] \in \mathbb{C}^m$ are of length m . The noises are independent from receiver to receiver, from block to block and further are independent of the channel gains. The transmitter is subject to an average total power constraint P , i.e.

$$\mathbb{E} \left[\text{Tr} \left(\mathbf{x}[k] \mathbf{x}[k]^\dagger \right) \right] \leq P, \quad (2.11)$$

within each signaling interval which is equivalent to power constraint imposed on the single-

antenna system. As in the single-antenna system we assume that channel gains in each signaling interval are known perfectly (i.e., measured to arbitrary accuracy) at the respective receivers at the beginning of each such interval. Moreover, a feedback link exists by which individual users can inform the transmitter of their channel gains (or more generally quantized versions thereof), also at the beginning of each associated signaling interval. Further, we assume the users do not know each others channel gains, nor are they able to more generally share information between each other. As results on modeling and measurement for the multi-user MIMO channel have only recently begun to be reported we from time to time appeal to cooperative results. If this is the case we refer to the system as the cooperative MIMO system or as a MIMO system with cooperative receivers. Note, however, that unless otherwise identified we assume that the users may not cooperate.

Any message scheduled for delivery is transmitted within one block and the blocks are long so that the messages can be reliably received. Thus each block corresponds to a new signaling (and hence scheduling) interval. Within each signaling interval, the transmitter sends from its array a group of messages, one for each of a subset of the user pool. We denote the set of n users as $\mathcal{U} = \{0, 1, 2, \dots, n-1\}$ and the set of user selected to receive a message we call the *active* set of users which is denoted by \mathcal{A} . We further refer to \mathcal{A} as the activation set.

In Chapter 5 we examine appropriate models for the joint distribution of each users channel gains and hence collect every users channel gain vector in a matrix $\mathbf{H}[k]$ where

$$\mathbf{H}^\dagger[k] = \begin{bmatrix} \mathbf{h}_0^\dagger[k] \\ \vdots \\ \mathbf{h}_{n-1}^\dagger[k] \end{bmatrix}.$$

However, as previously noted, a main contribution of the thesis is that one may design channel feedback for many multi-user MIMO systems with general fading distributions given that one has a class of “good” quantizers for a system with an isotropic fading distribution. Hence, in the sequel we assume that $\mathbf{H}[k]$ is modeled as a random matrix where by each element of $\mathbf{H}[k]$ are *i.i.d* complex Gaussian $\mathcal{CN}(0, 1/2m)$ random variables. In particular, let

$$\mathbf{H}[k] = \mathbf{G}[k] \tag{2.12}$$

where $\mathbf{G}[k]$ is a $m \times n$ random matrix with *i.i.d* $\mathcal{CN}(0, 1/2m)$ elements. We refer to this model for the MIMO channel as the Rayleigh model. In order to extend our results to more general model we, from time to time, also assume that each user’s channel vector is spatially correlated to examine how non-isotropic channel distribution effect our results. More precisely, from time to time, we assume each user channel is distributed as

$$\mathbf{h}_i = \Sigma^{1/2} \cdot \mathbf{h}_i^{(0)}$$

where the elements of $\mathbf{h}_i^{(0)}$ are *i.i.d* $\mathcal{CN}(0, 1/2m)$ and make clear when this assumption is made. Such an approach leads to developing a quantization framework which is described in terms of the relevant model and geometric parameters thus leading to a quantization framework that may be adapted to match channel conditions for general fading distributions. In order to make this precise we now state our general quantization model and examine the effects a system with finite rate feedback has on multi-user system performance. Then, in Chapter 3 we present our systematic quantization framework.

■ 2.2.1 Channel Quantization

In a multi-user MIMO system the quantizer design not only effects the order statistic gain through the mean square error, but also the multi-node matching gain through the transmitters ability to infer channel interference. This relationship can be quite difficult to model exactly and as such, we outline the effects a correlated Gaussian random vector has on a *general quantization scheme* and latter address how this general picture relates to the relevant channel model of the multi-user MIMO channel identified in Chapter 5. Such an approach has practical relevance. Indeed, part of our motivation for the feedback design problem is to develop a system which is robust to a variety of fading conditions. However, such a modeling approach leads to developing a quantization framework which is described in terms of the relevant model and geometric parameters leading to a quantization framework that may be chosen to match channel conditions. This allows us to later develop a framework that may be dynamically modified to adapt to changes in statistics of the channel. All of the relevant insights and geometric motivation may be gained by considering a Gaussian fading model. However, we first require a few more details concerning channel quantization before proceeding to develop this framework.

We assume that the quantization codebook \mathcal{C} is such that the codewords $\mathbf{c} \in \mathcal{C}$ all lie on the unit sphere in m (complex) dimensions. We let r denote the number of bits to which a channel direction is quantized, so the codebook is of size 2^r . We label the codewords in the codebook $\mathcal{C} = \mathcal{C}_r$ as $\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_{2^r}$. An important property of a code is the *sparsity* of the code. We say that a code is k -sparse if every vector of a code has at most k non-zero entries. We note that every quantizer in \mathbb{C}^m may be viewed as the union of a 0-sparse, 1-sparse, \dots , $m - 1$ -sparse and m -sparse codes.

The quantization codebook \mathcal{C} is fixed and the same for all users and the corresponding quantization rule corresponds to

$$\mathcal{Q}(\mathbf{h}_j) = \arg \max_{\mathbf{c} \in \mathcal{C}} d(\mathbf{c}, \mathbf{h}_j) \quad \text{where} \quad d(\mathbf{c}, \mathbf{h}_j) = \left| \mathbf{c}^\dagger \mathbf{h}_j \right|. \quad (2.13)$$

We denote the quantization of \mathbf{h}_j as

$$\hat{\mathbf{h}}_j \triangleq \mathcal{Q}(\mathbf{h}_j)$$

and for any subset of channel vectors $\{\mathbf{h}_{a_1}, \dots, \mathbf{h}_{a_\ell}\}$ we denote by

$$\hat{\Phi}_{\mathcal{A}} = \mathcal{Q}(\mathbf{H}_{\mathcal{A}}) \triangleq \begin{bmatrix} \mathcal{Q}(\mathbf{h}_{a_1})^\dagger \\ \vdots \\ \mathcal{Q}(\mathbf{h}_{a_\ell})^\dagger \end{bmatrix} \quad (2.14)$$

the set of quantized channel vectors for the set of users $\mathcal{A} = \{a_1, \dots, a_\ell\}$.

The quantization rule (2.13) leads to a quantizer design which may be thought as of a *system of lines through the origin* rather than discrete points on the unit sphere. Thus, the current system only quantizes the channels direction and not the gain. We note one may more generally quantize the gain. However, the corresponding results do not dramatically effect our results and thus we only consider feedback schemes which quantize the direction of each users channel.

In a system where the channel state is quantized the set of rates that may be achieved

by the system may be considered to be discrete³. Moreover, in such a system, the number and distribution of these discrete operating points is directly tied to the structure of the associated feedback scheme as the channel feedback is the only knowledge the transmitter has of the channel state. Thus, the transmitter may only infer each users channel and the co-channel interference from the descriptions of users channels given by the feedback scheme. Hence, the transmitter may only schedule users based on the discrete set of channel vectors used by a feedback scheme. In the sequel we leave many parameters of the quantizer and channel model free and as such it is unrealistic to precisely compute the joint distribution of the quantization error and the transmitters estimate of the co-channel interference and achievable rates for the system in general. To circumvent this issue we present a parametric model for the feedback process in Section 5.1 that may be used to estimate the fading distribution of the channel. More precisely, for any given fading distribution, in Section 5.1 we present a systematic method to estimate the probability that any user is quantized to a given codeword. With this in hand one may then in turn approximate the distribution of the joint fading statistics. More precisely, in Section 5.1 we present a systematic method to estimate the probability vector

$$\mathbf{p}_i(\mathcal{C}_r) = (p_{i,0}(\mathcal{C}_r), p_{i,1}(\mathcal{C}_r), \dots, p_{i,2^r-1}(\mathcal{C}_r))$$

where

$$p_{i,j}(\mathcal{C}_r) = \Pr[\text{user } i \text{ is quantized to codeword } j \mid \mathcal{C}_r].$$

In the sequel we present how one may compute this distribution exactly for a user which has a spatially correlated channel vector. Although, we do not use the following methods to compute the exact user assignment distribution directly from the channel model we examine how one may compute the marginal distribution for the feedback from user i as this development provides useful insights we use in the sequel.

Assume for the present that the channel vector of each user in the system is marginally distributed as a jointly Gaussian random complex vector of length m and covariance Σ . In particular, in the sequel we assume

$$\mathbf{h}_i = \Sigma^{1/2} \cdot \mathbf{h}_i^{(0)}$$

where the elements of $\mathbf{h}_i^{(0)}$ are *i.i.d* $\mathcal{CN}(0, 1/2m)$. With this assumption each user's channel vector has a norm that has a Chi-squared distribution (for some suitable parameters) and a direction that is distributed non-uniformly over the complex unit m -sphere. As we are interested in quantizing the direction of each users channel gain vector the quantization rule (2.13) determines a set of 2^r regions on the complex unit m -sphere which determine which points of the sphere are quantized to each codeword. That is, (2.13) determines the collection of Voronoi regions for any code \mathcal{C}_r . We let \mathcal{V}_i be the Voronoi region for \mathbf{c}_i , i.e. \mathcal{V}_i is the set of all points on the complex unit m -sphere that are closer to \mathbf{c}_i than any other codeword in \mathcal{C}_r (where ties are broken arbitrarily). More precisely,

$$\mathcal{V}_i = \{\mathbf{x} \in \mathbb{C}^m : \|\mathbf{x}\| = 1 \text{ and } d(\mathbf{c}_i, \mathbf{x}) \leq d(\mathbf{c}_j, \mathbf{x}) \quad \forall \mathbf{c}_j \in \mathcal{C}_r \setminus \{\mathbf{c}_i\}\}. \quad (2.15)$$

and the probability that user i is quantized to any codeword of \mathcal{C}_r , say \mathbf{c}_j , is equal to the

³This is true, for example, in a system which omits power control and time-division schemes

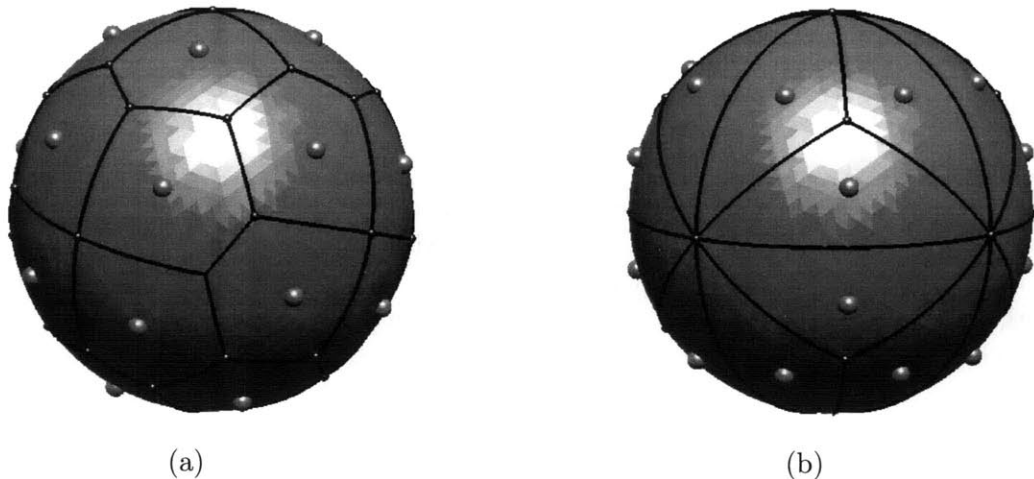


Figure 2-1. An illustration of how the shape of the Voronoi cell effects the mean square error for users with isotropic fading. Two possible arrangement of 12 lines in \mathbb{R}^3 . (a), a uniform collection of lines that has a low mean square error. (b), a structured collection of 12 lines with higher mean square error. Note that by assuming Voronoi regions are isomorphic a high coherence implies the Voronoi region has points that lay far from center increasing the inertia of the region and hence the mean square error.

weighted volume of \mathcal{V}_j . That is,

$$\Pr[\mathcal{Q}(\mathbf{h}_i/\|\mathbf{h}_i\|) = \mathbf{c}_j] = \int_{\mathbf{x} \in \mathcal{V}_j} d\mu_m(\mathbf{x}; \Sigma) \quad (2.16)$$

where $d\mu_m(\mathbf{x}; \Sigma)$ is a continuous measure on the unit m -sphere induced by the covariance matrix Σ . A similar argument holds if one is interested in computing the mean squared quantization error.

In a system with correlated fading the expected MSE error is directly related to the size and shape of the Voronoi cells. In particular, the MSE of any cell is the (weighted) second moment of the cell,

$$\int_{\mathbf{x} \in \mathcal{V}_j} \|\mathbf{c}_j - \mathbf{x}\|^2 d\mu_m(\mathbf{x}; \Sigma).$$

Thus, a code book with a smaller (weighted) second moment has a smaller MSE and hence achieves a higher expected rate. To see how the shape of the Voronoi cell effects the mean square error for users with isotropic fading consider the two codebooks in \mathbb{R}^3 in Figure 2-1. Note, that the quantizer on the left has a much smaller second moment than the one on the right as the mass of Voronoi cells for the quantizer on the left is more evenly distributed about its center. However, channel correlation may significantly change this picture and a significant mismatch may led to a high mean square quantization error regardless of the codebook coherence.

If a MIMO system has isotropic fading and one employs a quantizer which has isomorphic Voronoi regions both the MSE and the cell probability, p_j , are the same for every region. However, if the fading process is correlated or the Voronoi cells have irregular shape then one must compute the probability of every cell directly using (2.16). That is, repeating

(2.16) for every Voronoi cell of the code yields a discrete distribution

$$\mathbf{p}_i(\mathcal{C}_r) = (p_{i,0}(\mathcal{C}_r), p_{i,1}(\mathcal{C}_r), \dots, p_{i,2^r-1}(\mathcal{C}_r))$$

which describes the probability that user i is quantized to a codeword in \mathcal{C}_r .

In practice one does not have knowledge of the particular covariance matrix of each user and hence can not in general compute $\mathbf{p}_i(\mathcal{C}_r)$. However, observations of the feedback process from every user does allow one to make reasonable inference of \mathbf{p}_i and may be further used to estimate Σ . In a multiple-antenna system one may, through observation of a users feedback, estimate the covariance of the i th users channel

$$\mathbf{K}_{\mathbf{h}_i} \triangleq \mathbb{E} [\mathbf{h}_i \mathbf{h}_i^\dagger]$$

by first forming an estimate of $\mathbf{p}_i(\mathcal{C}_r)$, say $\hat{\mathbf{p}}_i(\mathcal{C}_r)$, and then estimate the covariance of the i th user's channel through the empirical covariance

$$\hat{\mathbf{K}}_{\mathbf{h}_i} \triangleq \sum_{j=0}^{2^r-1} \hat{p}_j \mathbf{c}_j \mathbf{c}_j^\dagger. \quad (2.17)$$

With this approach it is additionally possible to estimate the principle eigenmode of the channel covariance. Indeed, given the empirical covariance the principle eigenmode of $\hat{\mathbf{K}}_{\mathbf{h}_i}$ is the ML estimate of the principal invariant subspace of the covariance [115]. Hence one can identify the dominate mode of the correlation to aid in adapting the quantization codebook. This is an important observation as the ability to infer characteristics of the propagation environment coupled to a quantization scheme which has the ability to adapt to match the dominate features of the propagation environment allows a system to be stable under a wide range of channel conditions. We exploit this observation in Chapter 5. In particular, in Section 5.1 we develop a discrete framework to model the feedback process directly which allows one to make reasonable inference of the propagation environment and adapt the feedback framework to better match the channel. However, this requires a base design, which in the absence of a prior on the channel covariance, must perform well for the Rayleigh model. In this direction we turn to the relevant figures of merit for channel quantization assuming an isotropic channel distribution.

■ 2.2.2 Quantization Figures of Merit and MSE vs. Orthogonality Trade-off

The figures of merit chosen to evaluate a quantization codebook must be chosen to adequately reflect the problem of interest. In our development, we have advocated a quantization design which balances one's ability to estimate the co-channel interference with the incurred mean squared quantization error. However, most feedback designs for the MIMO channel at present choose a figure of merit that characterizes the mean square quantization error characteristics of the quantizer. It is natural to consider how these two approaches differ. Thus, in this section we develop the relevant figures of merit for MSE centered designs as well as the design we advocate. In particular, we show how the problem of designing a quantizer for which the mean square quantization error is low is often at odds with a design which increases one's ability to estimate co-channel interference by enforcing that every codeword is pairwise orthogonal with a specified number of other codewords.

With the quantization rule (2.13) a key figure of merit for the codebook is its *coherence*

$$\mu_0(\mathcal{C}) = \max_{i \neq j} |\mathbf{c}_i^\dagger \mathbf{c}_j|. \quad (2.18)$$

In general, $0 \leq \mu_0 \leq 1$, and, for a given r and many codes of interest, smaller values of μ_0 correspond to quantizers in which the lines are more equally spaced relative to the quantization rule (2.13). We note that there is not a one to one correspondence between the coherence of a quantizer and the mean squared quantization error as the coherence of a quantizer only describes the distance of the closest codeword and not the second moment. However, as seen in the sequel, codes in which $\mu_0(\mathcal{C})$ is small, often have symmetric Voronoi cells and hence low mean squared quantization error.

Previous work on MIMO feedback design [82, 90, 105, 137, 144] has taken the coherence as the sole figure of merit. Indeed, for a given code rate lowering the coherence by making the Voronoi cell more symmetric reduces the mean squared quantization error in isotropic fading, which increases the user's SNR on average. Thus, with the implicit assumption that small $\mu_0(\mathcal{C})$ implies a more symmetric Voronoi cell, minimizing $\mu_0(\mathcal{C})$ is a relevant design rule for minimum mean squared error quantizer design. In this thesis we show that in a multi-user system one often should consider other figures of merit for the system as well. In this direction consider a second, weaker, figure of merit of a code book, the k -norm of the cross correlation

$$\mu_k(\mathcal{C}) = \sqrt[k]{\sum_{\mathbf{c}_i, \mathbf{c}_j \in \mathcal{C}} |\mathbf{c}_i^\dagger \mathbf{c}_j|^{2k}}. \quad (2.19)$$

The corresponding lower bound on $\mu_k(\mathcal{C})$, for a codebook with 2^r codewords, is [135]

$$\mu_k(\mathcal{C}) \geq \bar{\mu}_k(2^r, m) = \sqrt[k]{\frac{2^{2r}}{\binom{m+k-1}{k}}}. \quad (2.20)$$

While the coherence roughly describes the minimal angle between codewords the k -norm of the cross correlation relates to the average angle between codewords. As the maximum of a sum with 2^r terms must be greater than $1/2^r$ times the sum, we can use $\bar{\mu}_k(2^r, m)$ to arrive at a lower bound on $\mu(\mathcal{C})$. That is,

$$\sum_{\mathbf{c}_i, \mathbf{c}_j \in \mathcal{C}} |\mathbf{c}_i^\dagger \mathbf{c}_j|^{2k} = |\mathcal{C}| + \sum_{\mathbf{c}_i, \mathbf{c}_j \in \mathcal{C}} |\mathbf{c}_i^\dagger \mathbf{c}_j|^{2k} \quad (2.21a)$$

$$\leq |\mathcal{C}| + (|\mathcal{C}| - 1)|\mathcal{C}| \max_{\mathbf{c}_i, \mathbf{c}_j \in \mathcal{C}} |\mathbf{c}_i^\dagger \mathbf{c}_j|^{2k}. \quad (2.21b)$$

Hence,

$$\mu_0(\mathcal{C}) \geq \sqrt[k]{\frac{\bar{\mu}_k(2^r, m) - 2^r}{2^r(2^r - 1)}}.$$

This yields the best known bound on $\mu_0(\mathcal{C})$ [70, 135], which, for any positive integer k , is

$$\mu_0(\mathcal{C}) \geq \sqrt[k]{\frac{1}{2^r - 1} \left(\frac{2^r}{\binom{m+k-1}{k}} - 1 \right)}. \quad (2.22)$$

With this derivation one can see that a code meeting (2.22) has a uniform minimum distance and hence symmetric Voronoi cells. While finding codes with optimal coherence is in general

an open problem finding codes meeting (2.20) has been largely solved [106]. In fact, a large number of codebooks are known to meet (2.20).

In this thesis we seek to understand the trade-off between one's ability to represent any user's channel well (with respect to (2.13)) and one's ability to infer the co-channel interference between user groups. In the preceding discussion we have provided a bound on how well one may hope to do in terms of coherence which roughly corresponds to the achieved mean squared quantization error. However, it is of interest to understand how this bound and the weaker bound on the k -norm on the cross correlation are influenced by placing some constraint on the codebook to help the transmitter infer the co-channel interference between users. A particularly natural constraint to place on the codebook to help the transmitter infer the co-channel interference between users is a requirement that each codeword in the quantization codebook should have many orthogonal vectors from which many orthogonal sets may be selected. Such an approach allows a user to indicate a plurality of subspaces for which it is near. As such, we let

$$\eta(\mathcal{C}) = \min_{\mathbf{c}_i \in \mathcal{C}} \left| \left\{ \mathbf{c}_j : \mathbf{c}_i^\dagger \mathbf{c}_j = 0 \right\} \right|.$$

To see how constraining a code to have a given number of orthogonal vectors has on the coherence we begin by noting that any feedback scheme should minimally meet the k -norm of the cross correlation (2.19). Thus, repeating (2.21), this time adding in prior knowledge of $\eta(\mathcal{C})$, yields⁴

$$\sum_{\mathbf{c}_i, \mathbf{c}_j \in \mathcal{C}} |\mathbf{c}_i^\dagger \mathbf{c}_j|^{2k} = |\mathcal{C}| + \sum_{\substack{\mathbf{c}_i, \mathbf{c}_j \in \mathcal{C} \\ |\mathbf{c}_i^\dagger \mathbf{c}_j| > 0}} |\mathbf{c}_i^\dagger \mathbf{c}_j|^{2k} \quad (2.23a)$$

$$\leq |\mathcal{C}| + (|\mathcal{C}| - \eta(\mathcal{C}) - 1)|\mathcal{C}| \max_{\mathbf{c}_i, \mathbf{c}_j \in \mathcal{C}} |\mathbf{c}_i^\dagger \mathbf{c}_j|^{2k} \quad (2.23b)$$

We note that (2.23b), while simple to derive, illustrates the necessary trade-off between the order statistic gain and multi-node matching gain in terms of the feedback design. That is, as we have previously shown, the multi-node matching gain is enhanced when the quantizer has many orthogonal sets while the order statistic gain is improved when the mean squared quantization error is decreased. Equation (2.23b) shows exactly how these two design objectives are at odds. To see this suppose, in order to increase the multi-node matching gain, one designs a quantizer such that every codevector is orthogonal with η other codevectors. Then, by inserting (2.19) in (2.23b) for a fixed k , the bound on the maximum cross correlation for the resulting code is

$$\mu_0(\mathcal{C}) \geq \bar{\mu}_k(2^r, m; \eta) = \sqrt[2k]{\frac{\bar{\mu}_k(2^r, m)^{2k} - 2^r}{(2^r - \eta - 1)2^r}}. \quad (2.24)$$

If η is chosen to be a constant fraction of the codebook size then, for large 2^r , (2.24)

⁴This bound is loose as in general as one could re-derive the result for the k -norm on the cross correlation for non-orthogonal codewords or given the number of distinct cross correlation values employ the results of [44, 57].

can be substantially larger than (2.22) as

$$\begin{aligned}\bar{\mu}_k(2^r, m; \alpha \cdot 2^r) &= \bar{\mu}_k(2^r, m; 0) \cdot \sqrt[2k]{\frac{(2^r - 1)}{(2^r - \eta - 1)}} \\ &\approx \bar{\mu}_k(2^r, m; 0) \cdot (1 - \alpha)^{-1/(2k)}\end{aligned}$$

Thus, if one imposes a strong orthogonality constraint on our codebook, i.e. $\eta \propto 2^r$, then (2.24) predicts a non-negligible increase in the maximum cross correlation. Alternatively, any attempt to maximize the multi-node matching gain by increasing the number of orthogonal sets will, by (2.24), likely increase the codebook coherence. In Chapter 3 we describe a quantization framework in which the system designer can balance these design objectives. However, we first must identify the effects that quantization has of system performance in a multi-user MIMO system.

■ 2.3 Figures of Merit for MIMO Channels and Beamforming

The figures of merit we consider in the sequel are identical to those provided for the single-antenna channel. That is, we again consider the appropriate generalizations of the ergodic capacity, outage probability as well as the scheduling failure probability. To begin, we consider the ergodic capacity of the MIMO channel. If the fading between transmit pairs vary rapidly one may have to again transmit over multiple realizations of the fading process to achieve reliable transmission. To be concrete we at present assume the Rayleigh model. For such a model the channel matrix $\mathbf{H}[k]$ is modeled as a random matrix where by each elements of $\mathbf{H}[k]$ are *i.i.d* complex Gaussian $\mathcal{CN}(0, 1/2m)$ random variables. Assuming the Rayleigh model one can show that with high probability every realization of the channel matrix \mathbf{H} provides approximately $\min\{n, m\}$ parallel paths from the transmitter to each receiver. With such path diversity it is likely that if one path undergoes a deep fade the remaining paths will be better provided that these paths are not highly correlated. This is the basic intuition behind the ergodic capacity scaling results of [126]. That is, if one assumes the channel follows the Rayleigh model (2.12) then the ergodic capacity of a cooperative MIMO channel is

$$C_{\text{ergodic}} = \mathbb{E}_{\mathbf{H}} \left[\log \left(\left| \mathbf{I} + \frac{P}{\sigma^2 m} \mathbf{H} \mathbf{H}^\dagger \right| \right) \right] \approx \min\{M, N\} \cdot \log \left(1 + \frac{P}{\sigma^2} \right)$$

If the MIMO channel is sufficiently slow fading so that one may reliably transmit over a single fade one is again interested in the instantaneous SNR of the channel. However, in multi-user MIMO there are two ways in which a transmitter may exploit the extra spatial degrees of freedom afforded by the MIMO channel. The transmitter may transmit to only a single-user, thereby providing that single-user with full path diversity or the transmitter may more generally multiplex signals for multiple users together using the spatial degrees of freedom to transmit multiple streams of data simultaneously. If the transmitter only exploits the spatial diversity of the array by transmitting to a single-user then the instantaneous SNR of user i is

$$\text{SNR}_i[k] = \frac{|\mathbf{h}_i[k]^\dagger \mathbf{x}[k]|}{\sigma^2}$$

resulting in a spectral efficiency of

$$\log(1 + \text{SNR}_i[k]) \text{ bits/complex dimension.}$$

If the user's channel state is known to the transmitter the transmitter may, in order to provide a desired signal-to-interference-to-noise ratio (SNR) for a user, use transmit and receive *beamforming*. That is, one may select

$$\mathbf{x}[k] = u_i \mathbf{w}_i$$

where u_i is the message symbol for users i and where in turn \mathbf{w}_i is the *beamforming vector* for user i . We assume throughout that $|u_i|^2 = P_i$, where P_i is the power allocated to user i . Using transmit and receive beamforming, the instantaneous SNR of user i becomes

$$\text{SNR}_i[k] = \frac{P \cdot |\mathbf{h}_i[k]^\dagger \mathbf{w}_i|}{\sigma^2}. \quad (2.25)$$

In a system with perfect channel state information at the transmitter one may optimize the SNR in (2.25) by choosing $\mathbf{w}_i = \mathbf{h}_i[k]$ and hence (2.25) becomes

$$\text{SNR}_i[k] = \frac{P \cdot \|\mathbf{h}_i[k]\|^2}{\sigma^2}.$$

A beamforming system with channel state information thus can significantly increase the performance of a system. However, as in a single-antenna system, the channel fading may still cause a user to have a significantly poor fading state and hence the channel quality may be below the SNR threshold for which a chosen fixed rate coding scheme can be used reliably. Thus, one is again interested in the outage probability,

$$P_{\text{outage}}(\tau_0) = \Pr [\text{SNR}_i[k] \leq \tau_0]. \quad (2.26)$$

It is important to note that due to the spatial diversity of MIMO the outage probability of a multiple-antenna system may be much lower than that of a single-antenna system for a given SNR threshold. If one additionally has multiple users in the system and a scheduler allocates the channel to the strongest user at any time one may see an additional increase in the SNR of the channel and even further reduce outage probability. Thus, in a multi-user MIMO system it is of further interest to know when a scheduler which chooses the single best user is in outage, i.e.

$$P_{\text{fail}}^{(S)}(\tau_0) = \Pr \left[\left(\max_{i=0,1,\dots,n-1} \text{SNR}_i[k] \right) \leq \tau_0 \right]. \quad (2.27)$$

As previously noted, in a multi-user MIMO system there are additional ways one may choose to exploit the degrees of freedom. In particular, in a multi-user MIMO system one may alternately use the spatial degrees of freedom of the MIMO channel to *multiplex* many users across the array simultaneously by reducing the diversity of each user. This may, however, introduce interference if the users channels are not orthogonal and one must balance the system gains one receives by increasing the number of users multiplexed across the array with the decrease in each user's rate.

In a MIMO system in which multiple users are multiplexed across the array it is the job of the *multiplexer* to construct a beamforming matrix \mathbf{W}_A which balances each users

SNR and level of interference for a given set of users \mathcal{A} . The achieved signal to interference-plus-noise ratio (SINR) is a function of every users channel state and when the multiplexer is informed of this channel state the multiplexer may intelligently choose the beamforming matrix $\mathbf{W}_{\mathcal{A}}$. In the sequel we consider *linear* multiplexers as they are an attractive choice when overall system complexity is of interest.

We focus on the case where the instantaneous signal \mathbf{x} can be represented as the linear combination

$$\mathbf{x} = \sum_{i \in \mathcal{A}} u_i \mathbf{w}_i = \mathbf{W}_{\mathcal{A}} \mathbf{u} \quad (2.28)$$

where again u_i is the message symbol for users i , \mathbf{w}_i is the *beamforming vector* for user i and $\mathbf{W}_{\mathcal{A}}$ is the *beamforming matrix* for the set of user \mathcal{A} . The vectors \mathbf{w}_i in general may be optimized for each transmission but may also come from some finite codebook. With this definition, assuming flat power allocation as we do throughout, the power allocated to each user is,

$$P_i = \frac{P}{\text{Tr}(\mathbf{W}_{\mathcal{A}} \mathbf{W}_{\mathcal{A}}^\dagger)}. \quad (2.29)$$

Thus, the baseband model for the system becomes

$$y_i = \mathbf{h}_i^\dagger \mathbf{w}_i \cdot u_i + \sum_{\substack{j \in \mathcal{A} \\ j \neq i}} \mathbf{h}_i^\dagger \mathbf{w}_j \cdot u_j + n_i.$$

We now examine the achievable signal to interference-plus-noise ratio (SINR) using common beamforming techniques.

Let σ_i be the correlation between the normalized channel vector

$$\tilde{\mathbf{h}}_i = \frac{\mathbf{h}}{\|\mathbf{h}\|}$$

and \mathbf{w}_i ,

$$\sigma_i = \tilde{\mathbf{h}}_i^\dagger \mathbf{w}_i.$$

Further, let $\boldsymbol{\sigma}_{i,\mathcal{A}}$ be the vector of correlations between the i th channel vector and the beamforming vectors of the other users in the set \mathcal{A} ,

$$\boldsymbol{\sigma}_{i,\mathcal{A}} = \mathbf{W}_{\mathcal{A} \setminus i} \tilde{\mathbf{h}}.$$

If the receiver employs an MMSE receiver to maximize the receive SINR the resulting SINR for the i th user is

$$\text{SINR}_i(\mathbf{W}_{\mathcal{A}}, \mathbf{H}_{\mathcal{A}}, P) = \frac{P \|\mathbf{h}_i\|^2 \sigma_i^2}{\text{Tr}(\mathbf{W}_{\mathcal{A}} \mathbf{W}_{\mathcal{A}}^\dagger) + P \|\mathbf{h}_i\|^2 \|\boldsymbol{\sigma}_{i,\mathcal{A}}\|^2}. \quad (2.30)$$

Note that (2.30) illustrates the trade-off between the order statistic gain and multi-node matching gain in a beamforming system. Indeed, if the channel state is perfectly known at the transmitter, one may, by ignoring the interference from the other users (letting $\boldsymbol{\sigma}_{i,\mathcal{A}}$ be arbitrary) greedily take $\mathbf{W}_{\mathcal{A}} = \tilde{\mathbf{H}}_{\mathcal{A}}$ in an attempt to increase the channel SNR (by ensuring $\sigma_i = 1$). Alternatively, one may attempt to precancel the interference from the other users (by ensuring $\boldsymbol{\sigma}_{i,\mathcal{A}} = 0$) using some of the possible transmit power to null the co-channel interference.

In a MIMO system the multiplexer that ignores the co-channel interference by taking $\mathbf{W}_{\mathcal{A}} = \tilde{\mathbf{H}}_{\mathcal{A}}$ we call the *interference ignoring* multiplexer and write

$$\mathbf{W}^{\text{II}}(\mathbf{H}_{\mathcal{A}}) = \tilde{\mathbf{H}}_{\mathcal{A}}. \quad (2.31)$$

The interference ignoring multiplexer transmits a signal of power

$$P_i^{\text{II}} = \frac{P}{|\mathcal{A}|}$$

to every user $i \in \mathcal{A}$ which yields a corresponding SINR equal to

$$\text{SINR}_i^{\text{II}}(\mathbf{H}_{\mathcal{A}}, P) = \frac{P \|\mathbf{h}_i\|^2}{|\mathcal{A}| + P \|\mathbf{h}_i\|^2 \sum_{j \in \mathcal{A}, j \neq i} |\mathbf{h}_j^\dagger \mathbf{h}_i|^2}. \quad (2.32)$$

At the other end of the spectrum in the zero-forcing multiplexer which uses some of the available transmit power to precode the signal so there is no co-channel interference. We call this multiplexer the *interference-cancelling* multiplexer. More precisely, the interference-cancelling multiplexer chooses the pseudo-inverse of the channel as the beamforming matrix, i.e.

$$\mathbf{W}^{\text{IC}}(\mathbf{H}_{\mathcal{A}}) = \mathbf{H}_{\mathcal{A}} \cdot (\mathbf{H}_{\mathcal{A}}^\dagger \mathbf{H}_{\mathcal{A}})^{-1}.$$

If one multiplexes users with the interference-cancelling multiplexer the power allocated to every user $i \in \mathcal{A}$ is

$$P_i^{\text{IC}} = \frac{P}{\text{Tr} \left((\mathbf{H}_{\mathcal{A}}^\dagger \mathbf{H}_{\mathcal{A}})^{-1} \right)}$$

with corresponding SNR

$$\text{SINR}_i^{\text{IC}}(\mathbf{H}_{\mathcal{A}}, P) = \frac{P}{\text{Tr}((\mathbf{H}_{\mathcal{A}} \mathbf{H}_{\mathcal{A}}^\dagger)^{-1})} \quad (2.33)$$

Examining (2.33) one may see that if the channel matrix $\mathbf{H}_{\mathcal{A}}$ is ill conditioned each user receives only a small fraction of the peak transmit power P . To combat this power loss one may more generally consider a regularized inverse of the channel [98],

$$\mathbf{W}^{\text{MMSE}}(\mathbf{H}_{\mathcal{A}}; \rho_{\text{MMSE}}) = \mathbf{H}_{\mathcal{A}} \cdot (\rho_{\text{MMSE}} \cdot \mathbf{I}_{|\mathcal{A}|} + \mathbf{H}_{\mathcal{A}}^\dagger \mathbf{H}_{\mathcal{A}})^{-1} \quad (2.34)$$

where $\rho_{\text{MMSE}} \geq 0$ and $\mathbf{I}_{|\mathcal{A}|}$ is an $|\mathcal{A}| \times |\mathcal{A}|$ unitary matrix. We call this multiplexer the MMSE beamforming multiplexer. Such a multiplexer trades off the received signal power with the co-channel interference which may be seen through examining the power allocated to every user [98]

$$P_i^{\text{MMSE}} = \frac{P}{\sum_{i=0}^{|\mathcal{A}|-1} \frac{\lambda_i(\mathbf{H}_{\mathcal{A}}^\dagger \mathbf{H}_{\mathcal{A}})}{(\lambda_i(\mathbf{H}_{\mathcal{A}}^\dagger \mathbf{H}_{\mathcal{A}}) + \rho_{\text{MMSE}})^2}}$$

where $\{\lambda_i(\mathbf{H}_{\mathcal{A}}^\dagger \mathbf{H}_{\mathcal{A}})\}_{i=0}^{|\mathcal{A}|-1}$ are the eigenvalues of $\mathbf{H}_{\mathcal{A}}^\dagger \mathbf{H}_{\mathcal{A}}$. Note that if $\rho_{\text{MMSE}} = 0$ then the MMSE beamforming multiplexer is simply the zero forcing multiplexer. We do not provide the explicit SINR expression for the MMSE beamforming multiplexer as it is generally quite complex.

In a multi-user MIMO system which employs multiplexing one may opportunistically allocate the channel to the subset of users with the highest SINR at any time. In such a multi-user MIMO system it is of interest to know when a scheduler which opportunistically searches for a subset of user meeting a prescribed SINR target fails to meet its objective. That is, one may further generalize the notion of outage to include

$$P_{\text{fail}}^{(M)}(\text{SINR}_0) = 1 - \Pr[\mathcal{A} \subset \mathcal{U} \text{ such that } \text{SINR}_i(\mathcal{A}) \geq \text{SINR}_0 \forall i \in \mathcal{A}]. \quad (2.35)$$

Of particular interest is whether one may meet specified SINR targets in a system with finite rate feedback. We address this question in detail in Chapter 4. At present we examine the effects of finite rate feedback on the achievable SINR in a multi-user MIMO system.

In a system with finite rate feedback one may employ the same beamforming techniques used as when the transmitter had perfect feedback. In particular, we focus our attention on the interference-cancelling multiplexer as the insights for the interference-ignoring multiplexer do not differ greatly from that when the transmitter has perfect channel state information. In order to derive the relevant expression for the SINR we, for simplicity, fix

$$\mathbf{W}_{\mathcal{A}} = \mathbf{W}^{\text{IC}}(\mathcal{Q}(\mathbf{H}_{\mathcal{A}}))$$

and

$$\mathbf{R}_{\mathcal{A}} = \hat{\Phi}_{\mathcal{A}}^{\dagger} \hat{\Phi}_{\mathcal{A}}.$$

Let $\mu_{i,\mathcal{A}}$ be the vector of correlations between the i th beamforming vector and the remaining beamforming vectors in the set \mathcal{A} , i.e.

$$\mu_{i,\mathcal{A}} \triangleq \hat{\Phi}_{\mathcal{A} \setminus i}^{\dagger} \mathbf{w}_i.$$

Then, we show in Appendix C.1.2 the received SNR for user i is

$$\frac{P}{\text{Tr}(\mathbf{R}_{\mathcal{A}}^{-1})} \|\mathbf{h}_i\|^2 c_i^2(\mathcal{A})$$

where

$$c_i(\mathcal{A}) \triangleq \frac{|\sigma_i - \sigma_{i,\mathcal{A}}^{\dagger} \mathbf{R}_{\mathcal{A} \setminus i}^{-1} \mu_{i,\mathcal{A}}|}{1 - \mu_{i,\mathcal{A}}^{\dagger} \mathbf{R}_{\mathcal{A} \setminus i}^{-1} \mu_{i,\mathcal{A}}} \quad (2.36)$$

and the corresponding co-channel interference caused by choosing $\mathbf{W}_{\mathcal{A}}$ is

$$\frac{P \|\mathbf{h}_i\|^2 c_{i,\perp}^2(\mathcal{A})}{\text{Tr}(\mathbf{R}_{\mathcal{A}}^{-1})}$$

where

$$c_{i,\perp}(\mathcal{A}) \triangleq \|(\sigma_{i,\mathcal{A}}^{\dagger} - \sigma_i \mu_{i,\mathcal{A}}^{\dagger})(\mathbf{R}_{\mathcal{A} \setminus i} - \mu_{i,\mathcal{A}} \mu_{i,\mathcal{A}}^{\dagger})^{-1}\|. \quad (2.37)$$

Thus, the SINR for the quantized interference canceling multiplexer may be written as

$$\text{SINR}_i^{\text{QIC}}(\mathcal{A}, P) = \frac{P \|\mathbf{h}_i\|^2 c_i^2(\mathcal{A})}{\text{Tr}(\mathbf{R}_{\mathcal{A}}^{-1}) + P \|\mathbf{h}_i\|^2 c_{i,\perp}^2(\mathcal{A})}. \quad (2.38)$$

We note that (2.38) makes explicit the need to reduce the uncertainty of the co-channel

interference at the transmitter as the SNR of the channel scales. That is, examining (2.37) one may see that using a quantized zero-forcing multiplexer the co-channel interference is a weighted function of the difference in the estimated co-channel interference $\mu_{i,\mathcal{A}}$ and the realized co-channel interference $\sigma_{i,\mathcal{A}}$. Additionally, this error is scaled based on the metric properties of the beamforming matrix. In particular,

$$c_{i,\perp}(\mathcal{A}) = \|(\sigma_{i,\mathcal{A}} - \sigma_i \mu_{i,\mathcal{A}})(\mathbf{R}_{\mathcal{A}\setminus i} - \mu_{i,\mathcal{A}}^\dagger \mu_{i,\mathcal{A}})^{-1}\| \quad (2.39a)$$

$$\leq \frac{\|\sigma_{i,\mathcal{A}} - \sigma_i \mu_{i,\mathcal{A}}\|}{\lambda_{\min}(\mathbf{R}_{\mathcal{A}\setminus i} - \mu_{i,\mathcal{A}}^\dagger \mu_{i,\mathcal{A}})} \quad (2.39b)$$

$$\leq \frac{\|\sigma_{i,\mathcal{A}} - \sigma_i \mu_{i,\mathcal{A}}\|}{\lambda_{\min}(\mathbf{R}_{\mathcal{A}\setminus i}) - \|\mu_{i,\mathcal{A}}\|^2} \quad (2.39c)$$

Hence, a feedback scheme which better estimates the co-channel interference and leads to beamforming matrices with better singular values likely leads to high rates when paired with a quantized zero-forcing scheme.

If the beamforming matrix is unitary, which corresponds to an orthogonal set of code-words, one has

$$c_{i,\perp}(\mathcal{A}) = \|\sigma_{i,\mathcal{A}}\|^2.$$

Then, from (2.38), one has the received SINR of user i is⁵,

$$\text{SINR}_i(\mathcal{A}) = \frac{P/|\mathcal{A}| \cdot \|\mathbf{h}_i\|^2 |\tilde{\mathbf{h}}_i^\dagger \mathbf{c}_i|^2}{\sigma_n^2 + P/|\mathcal{A}| \cdot \|\mathbf{h}_i\|^2 \sum_{j \notin \mathcal{A}} |\tilde{\mathbf{h}}_i^\dagger \mathbf{c}_j|^2} \quad (2.40a)$$

$$= \frac{|\tilde{\mathbf{h}}_i^\dagger \mathbf{c}_i|^2}{\frac{|\mathcal{A}| \sigma_n^2}{P \|\mathbf{h}_i\|^2} + \sum_{j \notin \mathcal{A}} |\tilde{\mathbf{h}}_i^\dagger \mathbf{c}_j|^2}. \quad (2.40b)$$

Examining (2.40b) it is easy to see that as the SNR of the system grows it is not necessary that the SINR does if there is finite rate feedback. In particular, in the limit of infinite SNR, assuming that the channel feedback vectors from each user are pairwise orthogonal, one has

$$\text{SINR}_i(\mathcal{A}) = \frac{|\tilde{\mathbf{h}}_i^\dagger \mathbf{c}_i|^2}{\sum_{j \notin \mathcal{A}} |\tilde{\mathbf{h}}_i^\dagger \mathbf{c}_j|^2} \quad (2.41a)$$

$$\approx \frac{|\tilde{\mathbf{h}}_i^\dagger \mathbf{c}_i|^2}{1 - |\tilde{\mathbf{h}}_i^\dagger \mathbf{c}_i|^2} \cdot \frac{1}{|\mathcal{A}| - 1}. \quad (2.41b)$$

If $|\tilde{\mathbf{h}}_i^\dagger \mathbf{c}_j|^2$ does not tend to 1 as the SNR scales it is clear from (2.41b) that the SINR of the system saturates as the co-channel interference scales proportionally with the SNR. This phenomenon may be seen in Figure 2-2. Thus, in order for the spectral efficiency of the system to scale as the SNR grows one must ensure that $|\tilde{\mathbf{h}}_i^\dagger \mathbf{c}_j|^2$ tends to 1. In a system with finitely many users this implies that systems with higher transmit powers must have higher feedback rates to fully realize the gains one expects with an increase in power [68]. Thus, in the sequel we use a normalized version of the expected value of the high SNR approximation of the SINR to characterize the performance of a beamforming scheme with

⁵We note that in this special case the SINR of the quantized zero-forcing multiplexer coincides with the interference ignoring multiplexer.

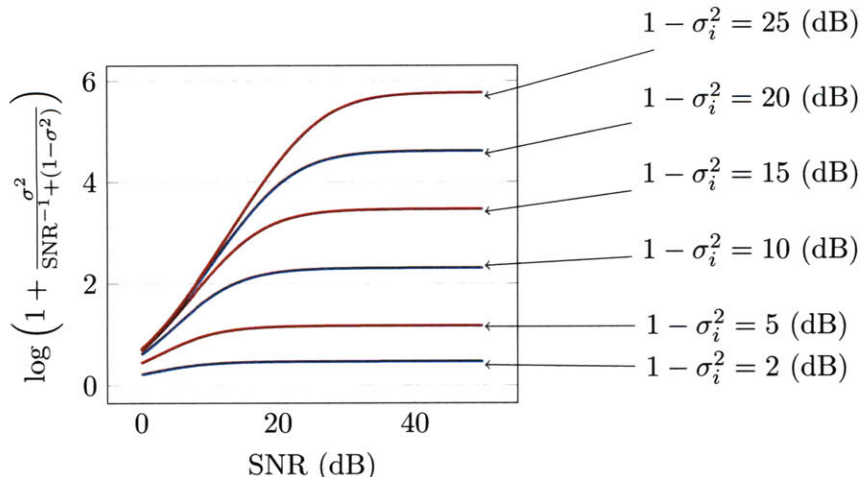


Figure 2-2. A plot of the spectral efficiency of each user in a MIMO system with 4 transmit antennas and a given quantization error. Note that as the SNR scales the spectral efficiency of a user is limited by the quantization error. Thus, a MIMO system which operates in the high SNR regime must use high rate feedback codebooks to ensure the system achieves high throughput.

finite rate feedback. That is, from (2.41b) one has

$$\text{SINR}_i(\mathcal{A}) \approx \frac{|\tilde{\mathbf{h}}_i^\dagger \mathbf{c}_i|^2}{1 - |\tilde{\mathbf{h}}_i^\dagger \mathbf{c}_i|^2} \cdot \frac{1}{|\mathcal{A}| - 1}.$$

Hence, for any code \mathcal{C}_r , we let

$$\text{SINR}_{\text{sat}}(\mathcal{C}_r) \triangleq \mathbb{E}_{\mathbf{h}_i} \left[\max_{\mathbf{c} \in \mathcal{C}_r} \frac{|\tilde{\mathbf{h}}_i^\dagger \mathbf{c}|^2}{1 - |\tilde{\mathbf{h}}_i^\dagger \mathbf{c}|^2} \right] \quad (2.42)$$

be the relevant metric for a MIMO beamforming system with finite rate feedback.

It is natural to consider how well one may do with regards to this metric. In the following section we consider an achievable lower bound on SINR_{sat} based on random vector quantization (RVQ) and provide a simple argument due to Shannon to provide an upper bound on SINR_{sat} .

■ 2.4 Bounds on MIMO System Performance with Finite Rate Feedback

In this section we provide an upper bound on SINR_{sat} and derive the performance of random vector quantization. To begin, we recall some basic facts about the distribution of the inner product between an isotropic vector distributed on the complex unit m -sphere and a fixed vector. In particular, let $\tilde{\mathbf{h}}_i$ be the direction of any user's channel vector. Then, from [16] one has that $\tilde{\mathbf{h}}_i$ is isotropic and

$$\Pr \left[|\tilde{\mathbf{h}}_i^\dagger \mathbf{c}|^2 < x \right] = 1 - (1 - x)^{m-1} \quad (2.43)$$

for any unit norm vector \mathbf{c} . Random vector quantization is a simple technique to analyze the achievable performance of quantization schemes which exploits the simple form of

(2.43). Random vector quantization simply generates 2^r quantization vectors independently at random with a uniform distribution over the complex unit m -sphere. We denote the ensemble of every such code as \mathcal{W}_r . Using this code ensemble one may analyze the system performance by averaging over the codebook ensemble \mathcal{W}_r as well as the channel fading distribution. Such an approach can be shown to yield an achievable lower bound that is quite close to the best known upper bound on the mean squared quantization error. In particular, it has been shown that the expected mean square quantization error for random vector quantization is

$$\mathbb{E}_{\mathbf{h}_i, \mathcal{W}_r} [\|\tilde{\mathbf{h}}_i - \mathcal{Q}(\tilde{\mathbf{h}}_i)\|^2] = 2 \cdot \mathbb{E}_{\mathbf{h}_i, \mathcal{W}_r} [1 - \tilde{\mathbf{h}}_i^\dagger \mathcal{Q}(\tilde{\mathbf{h}}_i)] = 2 \cdot \left(1 - 2^r \cdot B\left(2^r, \frac{m}{m-1}\right)\right)$$

where $B(\cdot, \cdot)$ is the beta function

$$B(a, b) = \int_0^1 x^{a-1} (1-x)^{b-1} dx.$$

In a multi-user MIMO system one is interested in not only the expected value of the quantization error, but also the expectation of the ratio of the channel correlation to the mean square quantization error, SINR_{sat} . In this direction, let

$$\text{SINR}_{\text{sat}}^{\text{RVQ}}(r, m) = \mathbb{E}_{\mathbf{h}_i, \mathcal{W}_r} \left[\max_{\mathbf{c} \in \mathcal{C}} \frac{|\mathbf{h}_i^\dagger \mathbf{c}|^2}{1 - |\mathbf{h}_i^\dagger \mathbf{c}|^2} \right].$$

Then, we have the following lemma as a direct extension of the results of [16].

Lemma 2.4.1. *Consider the ensemble of rate r random vector quantizers \mathcal{W}_r . Then,*

$$\text{SINR}_{\text{sat}}^{\text{RVQ}}(r, m) = -1 + 2^r B\left(\frac{m-2}{m-1}, 2^r\right).$$

Further, for large r

$$\text{SINR}_{\text{sat}}^{\text{RVQ}}(r, m) \sim -1 + 2^{r/(m-1)} \Gamma\left(\frac{m-2}{m-1}\right).$$

Proof. The achievable performance of random vector quantization may be derived through direct computation. The asymptotic expression follows directly from the asymptotic expression for the beta function with one fixed parameter [10]. ■

Lemma 2.4.1 provides important insights into the performance of random vector quantization. In particular, for high rate codebooks one gains approximately 3 (dB) in $\text{SINR}_{\text{sat}}^{\text{RVQ}}$ for each additional $m-1$ bits of feedback. Thus, without multi-user diversity and user selection one must increase the feedback rate linearly with SNR for the system not to saturate [67]. It is natural to consider whether one may do better in general. In the sequel we present a few quantization schemes that outperform random vector quantization for a fixed (small) number of feedback bits.

It is of additional interest to determine an upper bound on $\text{SINR}_{\text{sat}}(r, m)$ for arbitrary quantization schemes to see if the scaling predicted by Lemma 2.4.1 may be improved as

well as determine how far our constructed quantizers are from the optimal scheme. We let,

$$\text{SINR}_{\text{sat}}^{\text{UB}}(r, m) = -1 + 2^{r/(m-1)} \frac{m-1}{m-2}. \quad (2.44)$$

Then we have the following lemma providing an upper bound on SINR_{sat} .

Lemma 2.4.2. *Let \mathcal{C}_r be any rate r quantizer in \mathbb{C}^m . Then,*

$$\mathbb{E}_{\mathbf{h}_i} \left[\frac{|\mathbf{h}_i^\dagger \mathbf{c}_i|^2}{1 - |\mathbf{h}_i^\dagger \mathbf{c}_i|^2} \right] \leq \text{SINR}_{\text{sat}}^{\text{UB}}(r, m). \quad (2.45)$$

Proof. See Appendix B.1. ■

Note that the achievable values for SINR_{sat} provided using random vector quantization and the upper bound in (2.45) are quite similar. The main difference being the presence of the beta function for random vector quantization. In general, one can show that these two expressions are quite close. Examining the extremes one may see that they are equal for $r = 0$ as $B(1/x, 1) = x$ and similarly using the asymptotic expression for $\text{SINR}_{\text{sat}}^{\text{RVQ}}(r, m)$ one may see that the asymptotic gap⁶ in dB is not too large. In particular,

$$10 \log_{10} \text{SINR}_{\text{sat}}^{\text{UB}}(r, m) - 10 \log_{10} \text{SINR}_{\text{sat}}^{\text{RVQ}}(r, m) \sim 10 \log_{10} \frac{m-1}{m-2} - 10 \log_{10} \Gamma\left(\frac{m-2}{m-1}\right).$$

Hence, for large m and high quantization rates the gap between the random vector quantization and the upper bound vanishes. This is to be expected due to the asymptotic optimality of RVQ in large dimensions [16]. However, this asymptotic gap is, for $m > 2$, a decreasing function of m and hence as $r \rightarrow \infty$ is never larger than

$$10 \log_{10} \frac{2}{1} - 10 \log_{10} \Gamma\left(\frac{1}{2}\right) = 0.5246 \text{ dB}$$

which corresponds to the asymptotic gap for $m = 3$.

It is important to note that SINR_{sat} is a high SNR approximation of the achieved SINR of a system that uses a particular quantization scheme and not a measure of the achieved SINR for a given SNR. For a multi-user system to approach the limit predicted by SINR_{sat} one needs a subset of users which simultaneously have large channel norms, small quantization error as well as have nearly orthogonal quantized channel vectors. Thus, for a system to achieve an SINR close to the limit predicted by SINR_{sat} one needs a quantizer with orthogonal codewords as well as an algorithm to select users that are nearly orthogonal. We turn to the problem of user selection in Chapter 4 after first developing our quantization framework. However, before proceeding we note that there are two system regimes of interest; one regime where the number of users is fixed and the SNR growth is a function of power and a second regime where the number of users in the system grows and the SNR growth is caused by the order statistic gain.

In a MIMO system with a fixed number of users which operates in the high SNR regime one must scale the feedback rate per user linearly with the signal-to-noise ratio (SNR) of

⁶We caution the reader attempting to compute $\text{SINR}_{\text{sat}}^{\text{RVQ}}(r, m)$ for large r that care needs to be taken to ensure the numerical accuracy of $\text{SINR}_{\text{sat}}^{\text{RVQ}}(r, m)$ as the direct expression is often numerically unstable.

the channel for SINR_{sat} to grow unbounded [68]. Hence, one needs to develop high rate quantizers for high data rate systems. For a fixed value of SINR_{sat} a system will have improved performance if the codebook has multiple orthogonal bases as one can better estimate the co-channel interference and perform a more accurate interference cancelling scheme. Alternatively, when number of users in the system increases and the SNR scaling is caused by the order statistic gain a scaling in the feedback rate is not needed. In particular, we show in Chapter 4 that the order statistic gain for the quantization error tends to zero faster than the growth of the order statistic for the channel norm. Hence as the SNR approaches infinity the throughput scales unbounded with only $\log m$ bits of feedback per user. That is, as the number of users in the system increases the optimal quantization scheme in isotropic fading tends to any arbitrary basis of \mathbb{C}^m . This observation is the underpinning of our order statistic gain and multi-node matching gain trade-off and suggests a general design rule for feedback design in a multi-user MIMO system:

If one is interested in optimizing the SINR as the SNR scales in a multi-user system one should jointly design the feedback link to balance the trade-off between the quantization error and the number of orthogonal bases contained in the quantization codebook.

It is this perspective we take in our development in the sequel. In particular, in Chapter 3 we develop a systematic quantization framework to balance the number of orthogonal bases contained in a code with the quantization error.

Systematic Design of MIMO Channel Quantizers

Obtaining exact knowledge of the channel is unrealistic in many MIMO channels. One in practice must often quantize a channel realization and feed this finite rate representation back to the transmit base. In a single-user system the relevant aspects of feedback design have been well studied for an isotropic fading channel. Much of this work originated from the work of Narula et. al. [91] that studied the relevant aspects of system design when minimizing the mean square error (MSE) or maximizing the mutual information is of interest. Subsequent work has shown that both of the problems of minimizing the mean square error (MSE) as well as the problem of maximizing the mutual information may be treated in a common framework by considering the problem of minimizing the weighted mean square quantization error [105]. The authors of [91, 105] have proposed the use of a numerical algorithm to design a quantization codebook with near minimum (weighted) mean square error as well as maximum mutual information for a specified quantizer rate [91, 105].

In a multi-user MIMO system knowledge of the channel state at the transmitter is necessary to realize the multiplexing gain. Specifically, it has been shown that a MIMO system with a fixed number of users must scale the feedback rate per user linearly with the signal-to-noise ratio (SNR) of the channel for the spectral efficiency of the system to scale unbounded [68]. The nexus of this result is that when finite rate feedback is used to convey a users channel state to the transmitter there is some uncertainty at the transmitter of each users channel state. Hence, the transmitter can not employ an intelligent multiplexing method to fully eliminate the co-channel interference. As this interference scales linearly with the SNR one must decrease the co-channel interference proportionally with the SNR, leading to the need to linearly increase the feedback rate. In Section 2.3, we encapsulated this observation in our definition of $\text{SINR}_{\text{sat}}(\mathcal{C})$,

$$\text{SINR}_{\text{sat}}(\mathcal{C}) \triangleq \mathbb{E}_{\mathbf{h}_i} \left[\max_{\mathbf{c} \in \mathcal{C}} \frac{|\mathbf{h}_i^\dagger \mathbf{c}|^2}{1 - |\mathbf{h}_i^\dagger \mathbf{c}|^2} \right]$$

which we let be the relevant metric for MIMO beamforming systems with finite rate feedback. Thus, MIMO systems which operate in the high SINR regime with a fixed number of users and finite rate feedback must use large codebooks to ensure that the system performance is not limited. In such cases it is of interest to develop *structured* codebooks that enable user terminals to efficiently quantize their channel vectors as often the user terminals are power and complexity constrained.

In a multi-user MIMO system with finite rate feedback one may show that the feedback design directly effects the statistics of both order statistic gain and multi-node matching

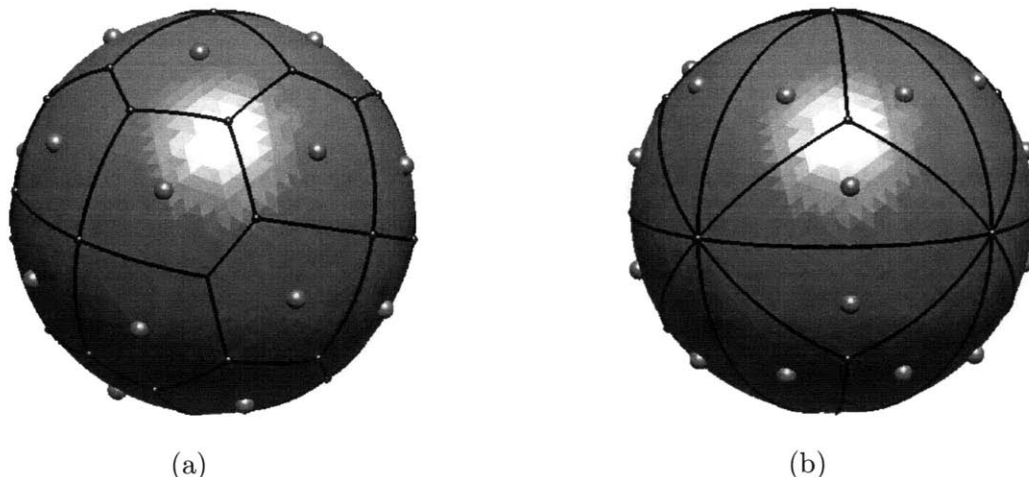


Figure 3-1. An example of the trade-off between mean squared quantization error and the number of orthogonal bases contained in the code. Two possible arrangement of 12 lines in \mathbb{R}^3 . (a), a uniform collection of lines that has a low mean square error. (b), a structured collection of 12 lines containing more orthogonal bases at the cost of higher mean square error.

gain. In particular, as the feedback received from the user terminals is the only knowledge that the transmitter has of each users channel the choice of the representation of the channel vectors in the feedback design affect both the order statistic gain and multi-node matching gain. More precisely, by one's choice in the feedback design one may reduce the mean squared quantization error (as in the currently proposed schemes) or increase the number of orthogonal bases so the transmitter may better identify users with low interference. An example of this trade-off in \mathbb{R}^3 can be seen in Figure 3-1. Thus, in a multi-user system there is a natural desire to develop quantization schemes that contain orthogonal bases as well as have good mean square error characteristics. Such an approach uses some of the feedback rate to identify when the interference between users is low and uses the remaining code rate to decrease the quantization error. Thus, in this chapter we provide a systematic way to design codebooks that have many orthogonal bases as well as regular structure to ensure the mean squared quantization error is low. More precisely, in this chapter we develop a systematic construction of channel quantizers which consists of three main structural components; a family of low-rate codes which contain many orthogonal bases, a systematic method to construct intermediate rate codes through unions of low-rate codes and a rate doubling operation which may be used to construct high rate codes with low complexity quantization algorithms.

To construct low-rate channel quantizers, we construct a family of structured codes in which one may trade-off the mean squared quantization error and the number of orthogonal bases contained in the code. We call these low-rate codes *component codes*. As we have previously stated, a MIMO system which operates in the high SNR regime with a fixed number of users needs high rate feedback from each user so that the co-channel interference does not limit the system performance. In order to increase the rate of a code one may form a union of low-rate component codes. However, in order to ensure that the resulting quantizer has low mean squared error one must ensure that the chosen component codes pair together well. In particular, from rate distortion theory one would like, in the limit of high quantization rates, the distribution of the codewords of the quantizer to approximately

match that of the channel vectors. In the particular case of isotropic fading one would like a quantizer to be distributed as uniformly over the surface of the complex unit m -sphere as possible. That is, in the high quantization rate limit the distribution of a codeword selected uniformly at random should be isotropic and hence the quantization codebook should be invariant to every unitary transformation. Thus, we develop a systematic framework in which one may form a union of component codes which ensures the resulting code has a large group of symmetries and hence good mean square quantization error.

Our development of component codes and methods to construct unions of component codes only produce quantizers that perform well up to intermediate rates. Specifically, our constructions of component codes of length 4 produce codes that outperform random vector quantization up to 8 bits. However, as we have previously noted, multi-user MIMO systems which operate in the high SNR regime must have good high rate channel quantizers. As the channel quantization occurs at the user terminals, we would like to find a way to extend our constructions to higher rates which allow for low complexity quantization.

As a final component of our systematic construction, we develop a methodology to double the rate of *any* existing channel quantizer. Such a method may be used in conjunction with random vector quantization as well as with our construction of component codes. Specifically, we develop a method to construct quantizers with 2^{2r} codewords from an existing rate r channel quantizer, say \mathcal{C}_r , which allows the application of multi-stage quantization algorithms. This is achieved by taking the union of the image of the code \mathcal{C}_r under a set of 2^r linear transformations¹ producing a rate $2r$ code, \mathcal{C}_{2r} . An important characteristic of these linear transformations is that they may be chosen to ensure that the resulting quantization complexity is only two times the complexity of quantization associated with the code \mathcal{C}_r . More precisely, the quantization of any channel vector, say \mathbf{h} , to one of the codewords of \mathcal{C}_{2r} according to (2.13) amounts to first quantizing \mathbf{h} to a codeword of \mathcal{C}_r , multiplying \mathbf{h} by the inverse of one of the 2^r linear transformations used in the rate doubling operation (which is determined by the first stage of quantization) and then performing a second quantization of the transformed channel vector to a codeword of \mathcal{C}_r . Thus, one may systematically construct rate $2r$ channel quantizers which have exponentially lower quantization complexity than a general rate $2r$ channel quantizer. We note, however, that codes produced with this component of our systematic construction often suffer slightly in performance when compared to other approaches which have no complexity restrictions. However, our high rate constructions have greater practical applicability than general rate $2r$ channel quantizers as one may have an intolerable quantization complexity with a general scheme leading to a quantizer which is unimplementable in practice.

We plot the performance of our length 4 codes constructed using our systematic framework in Figure 3-2. Component codes are plotted with a circle, the systematic union of component codes with a square and our high rate, low complexity codes with pentagons. These constructions are plotted relative to the best known upper bound on SINR_{sat} , $\text{SINR}_{\text{sat}}^{\text{UB}}$, the values of which are labeled at 0. We also plot the performance of random vector quantization, which provides an achievable bound for quantization with no complexity restrictions. One can see that at low to intermediate rates (3 – 7 bits) our construction of component codes and the associated unions perform quite well and are within 0.5 dB of the upper bound. Additionally, these codes outperform the achievable lower bound provided by random vector quantization. At high quantization rates the performance of our systematic

¹We note that this process may be repeated ad infinitum to produce higher and higher rate codes with low quantization complexity.

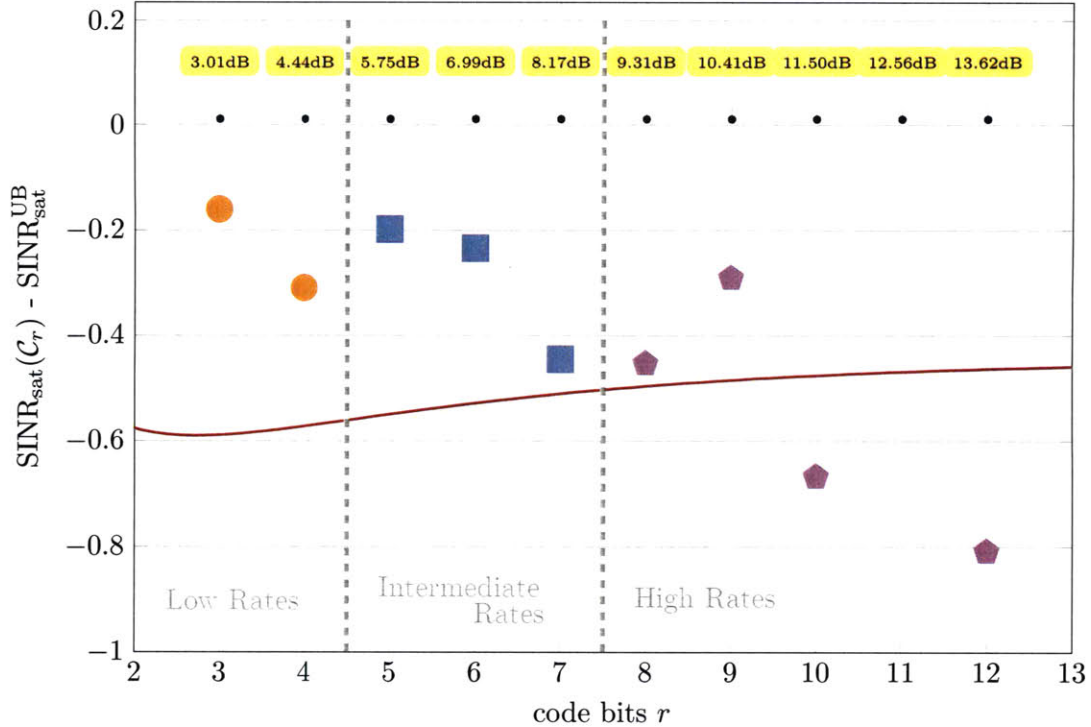


Figure 3-2. The performance of a few channel quantizers for a 4 transmit antenna system which we construct in the sequel. The performance of these quantizers are plotted relative to the best known upper bound on SINR_{sat} . The values taken by the upper bound are labeled at 0 and the performance of random vector quantization is plotted as a solid curve. Component codes are plotted with a circle, constructions consisting of unions of component codes are plotted with squares and codes constructed through the rate doubling framework are plotted with pentagons. For low-rates, specifically 3 and 4 bits, our construction of length 4 component codes perform well as do our constructions consisting of unions of component codes. Note that as the rate of the code increases from 3 bits to 9 bits the achieved performance is within 0.5 dB of the upper bound. Our low complexity codes perform worse than random vector quantization from 10 to 12 bits. However, the performance is within 0.81 dB of the upper bound and 0.35 dB of RVQ.

construction falls slightly. However, these constructions remain within 0.81 dB of the upper bound and within 0.35 dB of random vector quantization.

We develop the basics of our channel quantizer constructions in full in Section 3.2 along with our insights on why different constructions work well. Further, in Section 3.2 we provide a concrete example of how one may develop good channel quantizers for a 4 transmit antenna system. As a channel quantizer for a m transmit antenna system is a set of lines in \mathbb{C}^m in the sequel we use the statements “a channel quantizer in \mathbb{C}^m ” and “a channel quantizers for a m transmit antenna system” interchangeably. We use these constructions in \mathbb{C}^4 throughout this chapter to illustrate important concepts. However, our constructions are applicable to dimensions other than 4 and can be used to develop quantizers of arbitrary length². In Sections 3.3 – 3.6 we proceed to develop each one of the components of our systematic construction in depth. In particular, we present our basic construction for component codes with a fixed sparsity in Section 3.3 and develop how one

²Our discussion will be limited to the when the number of transmit antennas is equal to some prime power. However, we note that the constructions may be extended to arbitrary integers, however that development is overly cumbersome and does not yield any new insights and hence is neglected from the development.

may form low-rate codes with many orthogonal bases and low mean square quantization error in Section 3.4. In Section 3.5 we develop how one may systematically construct channel quantizers at intermediate rates by taking unions of component codes with varying sparsity. Then, in Section 3.6 we present a framework to extend code rates by a factor of two by using a family of linear operators, yielding a method to extend existing codes to higher rate codes for which channel quantization may be performed with multi-stage quantization algorithms. However, before proceeding to these construction we begin by examining the performance of a few known quantization schemes.

■ 3.1 Structured Quantization for MIMO Systems

The study of quantizer design to maximize the achieved rate in a single-user system leads to design criterion which minimizes the mean squared quantization error and often, for high quantization rates, is unstructured as some of the best known codes are designed through a Lloyd algorithm or RVQ. When only approaching system design from the standpoint of optimizing SINR_{sat} , random vector quantization is an appealing option as there is little room for improvement as we have shown that asymptotically there is at most a 0.5246 dB gap between RVQ and the optimal scheme. However, one drawback of RVQ is that it is *unstructured*. Thus, one must do an exhaustive search over a list to perform quantization which becomes prohibitive in terms of complexity and power use at the user terminals for high feedback rates. Further, in a multi-user system RVQ does little in terms of helping the transmitter identify the users that are nearly orthogonal which leads to an SINR penalty due to the inversion of non-orthogonal users. In particular, if a code contains no orthogonal bases then any set of users with small quantization error are not orthogonal and hence will suffer a SINR penalty caused by channel inversion with the interference canceling multiplexer or higher co-channel interference using the interference ignoring multiplexer. Thus, as previously noted, it is natural to consider embedding as many orthogonal bases in a code as possible, while not substantially degrading SINR_{sat} , to enable a transmitter to select users that are orthogonal, boosting the overall SINR. Thus, while RVQ has good performance in terms of SINR_{sat} there are other practical system objectives which make the development of *structured* quantizers that have performance close to that of RVQ of interest.

An alternative line of work for single-user systems has considered the design of quantization codebooks with near minimum mean square error that have added structure [82,90,105,137,144]. In particular, if one is interested in the probability of outage, i.e. the probability that the channel realization cannot support a desired rate, the authors of [82,90,137,144] suggest the use of structured, so called *Grassmannian line packings*, as efficient quantization alternatives to the less structured quantization codebooks proposed by [91]. The term “Grassmannian line packings” is a misnomer when used in the context of MIMO beamforming. We note that as the SINR and hence rate and outage probability are a function of the quantization error. As such one is more interested in a “Grassmannian line covering” rather than a packing. Indeed, one may have quite good mean square error performance without having a large minimum distance, i.e. with out having a large packing radius. However, as noted in Section 2.2.2, with the implicit assumption that a large minimum distance implies a uniform distribution in the distance between codewords, implying a small covering radius, optimization of a codebook with regard to this metric should perform well.

One of the simplest approaches to low complexity structured quantization is scalar quantization [91,104]. Scalar quantization is a simple scheme where by each coordinate of

a channel vector is quantized independently with a fixed number of bits. Although this scheme is extremely simple, it has been shown to perform reasonably well when compared to RVQ [66] as it has been shown numerically to have a constant rate gap relative to RVQ. Thus, scalar quantization appears to provide a throughput scaling with the same slope as random vector quantization and can be shown numerically in \mathbb{C}^6 to have a 2.7 dB loss relative to RVQ [66]. Thus, from a pure complexity standpoint, scalar quantization is a natural choice. However, this scheme in general has no orthogonal bases and hence will suffer the same SINR penalty due to the inversion of non-orthogonal users as RVQ. Thus, while we have gained in complexity we now suffer in performance and still have a code which lacks orthogonal bases. As practical system design requires choosing a balance between the complexity of quantization at the user terminals as well as the achieved throughput of the system it is of interest to develop structured quantization methods that balance the quantization complexity, number of orthogonal bases as well as achieved high values for SINR_{sat} .

A class of structured quantizers of great interest in the sequel are the quantization schemes developed by Hochwald [56] and the subsequent modifications which have been incorporated in to the 802.16e standard [1, 143]. The quantization scheme of Hochwald [56] forms a rate r codebook in \mathbb{C}^m by choosing m columns of the scaled $2^r \times 2^r$ DFT matrix

$$\text{DFT}(2^r, m) \triangleq \frac{1}{\sqrt{m}} \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & e^{\sqrt{-1} \frac{2\pi}{2^r} 1} & e^{\sqrt{-1} \frac{2\pi}{2^r} 2} & \cdots & e^{\sqrt{-1} \frac{2\pi}{2^r} (2^r-1)} \\ 1 & e^{\sqrt{-1} \frac{2\pi}{2^r} 1} & e^{\sqrt{-1} \frac{2\pi}{2^r} 2} & \cdots & e^{\sqrt{-1} \frac{2\pi}{2^r} (2^r-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & e^{\sqrt{-1} \frac{2\pi(2^r-1)}{2^r} 1} & e^{\sqrt{-1} \frac{2\pi(2^r-1)}{2^r} 2} & \cdots & e^{\sqrt{-1} \frac{2\pi(2^r-1)}{2^r} (2^r-1)} \end{bmatrix}.$$

More precisely, one may systematically construct rate r quantizers by choosing some subset of columns \mathbf{u} and letting

$$\mathcal{C}_{\text{DFT}}(r, \mathbf{u}) = \{\text{DFT}(2^r, m)[i, \mathbf{u}]\}_{i=0}^{2^r-1}. \quad (3.1)$$

One may then systematically design quantizers with high SINR_{sat} by solving the discrete optimization problem

$$\mathbf{u}^* = \underset{\substack{\mathbf{u} \in \mathbb{Z}_{2^r}^m \\ 0 \leq u_0 < \cdots < u_{m-1} \leq 2^r-1}}{\arg \max} \text{SINR}_{\text{sat}}(\mathcal{C}_{\text{DFT}}(r, \mathbf{u})).$$

Such a design may be shown to have good performance for a small number of bits, but performs worse than RVQ at higher rates. Thus, [143] proposed removing the constraint that the codewords are columns of the DFT matrix by performing a rotation to all but one vector of $\mathcal{C}_{\text{DFT}}(r, \mathbf{u})$. More precisely, let \mathbf{a} be any complex vector and let \mathbf{e}_0 be the vector where a 1 stands in the first coordinate and is 0 elsewhere. Then, [143] proposed a systematic codebook construction in \mathbb{C}^m by fixing the first codeword of every codebook to be

$$\mathbf{c}_0 = \text{DFT}(m, m)[:, 1]$$

and using a sequence of transformations to the codeword to form the remaining codewords.

Index	Construction	Reference
(3,1)	WiMax 3-bit	[1, 143]
(3,2)	$\mathcal{C}_{\text{DFT}}(3, [1, 2, 7, 6])$	[56]
(4,1)	MUB(4)	[61, 76]
(6,6)	WiMax 6-bit	[1, 143]
(6,4)	$\mathcal{C}_{\text{DFT}}(6, [1, 45, 22, 49])$	[56]

Table 3.1. A table of existing channel quantization constructions from literature and existing standards for 3, 4 and 6 bits. The performance of these channel quantizers may be seen in Figure 3-3.

In particular, let

$$\mathbf{P}(\mathbf{a}) = \mathbf{I} - \frac{2}{\|\mathbf{w}^\dagger \mathbf{w}\|} \mathbf{w} \mathbf{w}^\dagger$$

where in turn $\mathbf{w} = \mathbf{c}_0 - \mathbf{a}$ and let

$$\mathbf{Q}(\mathbf{u}) = \text{diag}(\text{DFT}(2^r, m)[1, \mathbf{u}]).$$

Then, one may form a codebook with 2^r codewords by letting

$$\mathbf{c}_i = e^{\sqrt{-1}\phi_i} \cdot \mathbf{P}(\mathbf{a})\mathbf{Q}(\mathbf{u})^i \mathbf{P}(\mathbf{a})^\dagger \mathbf{c}_0$$

for some chosen phase ϕ_i which makes the first coordinate have zero phase. We denote the resulting quantizer as $\mathcal{C}_{\text{WiMax}}(r, \mathbf{u}, \mathbf{a})$. Inside this framework one may systematically design a rate r quantizer by solving the mixed optimization problem

$$(\mathbf{a}^*, \mathbf{u}^*) = \underset{\substack{(\mathbf{a}, \mathbf{u}) \in \mathbb{C}^m \times \mathbb{Z}_2^r \\ 0 \leq u_0 < \dots < u_{m-1} \leq 2^r - 1}}{\arg \max} \text{SINR}_{\text{sat}}(\mathcal{C}_{\text{WiMax}}(r, \mathbf{u}, \mathbf{a})).$$

Due to fewer constraints this scheme in general does better than the construction (3.1). A depiction of the performance of these schemes as well as other well known constructions, which we list in Table 3.1, maybe seen, relative to the performance of RVQ and the upper bound (2.44), in Figure 3-3. Note that the WiMax construction does quite well relative to RVQ for 3 bits, but is much closer to RVQ at 6 bits. Additionally, the WiMax design outperforms Hochwald's constructions at both 3 and 6 bits. However, in general these constructions contain no orthogonal bases and have no guarantee that at higher rates there exist efficient quantization schemes with complexity comparable to multi-stage quantization.

We note that the design of structured quantizers with many orthogonal bases has been considered previously by Ashikhmin et. al. in [13]. In [13] a quantization framework was developed which produces at most one channel quantizer per dimension. Each quantizer performs quite well in terms of SINR_{sat} for the given rate, relative to the upper bound, but yields no systematic construction for various rates in a given dimension. We seek a more systematic approach to the design of MIMO feedback codebooks that allows a system designer to trade-off the quantization error for more orthogonal bases if, for instance, one knows apriori there are a large number of users in the system. At present we do not describe the quantization scheme of [13] as it follows from our general quantization scheme, which we develop in full in Section 3.2. At present, we only plot a few of our best constructions, which are listed in Table 3.2 alongside the existing results in Figure 3-4.

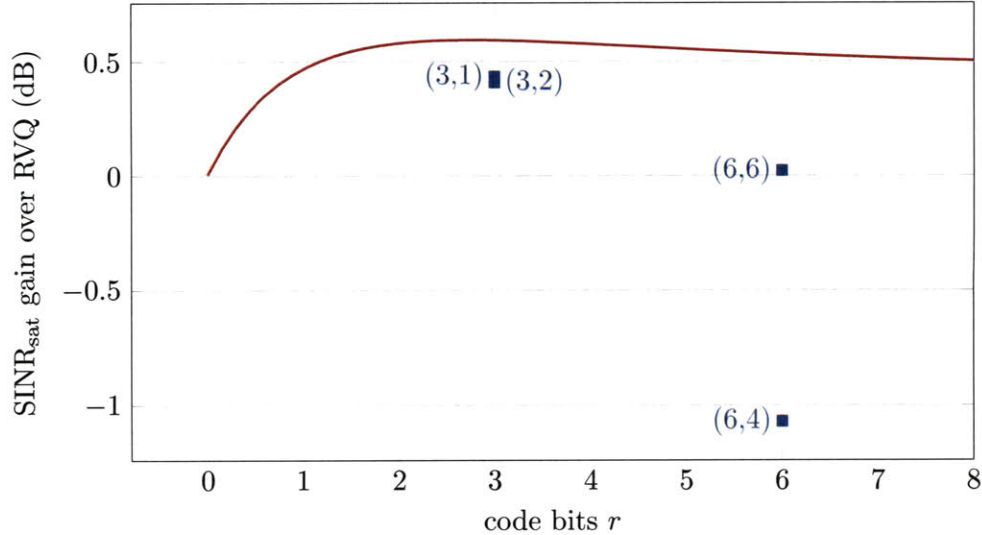


Figure 3-3. The difference in SINR_{sat} between random vector quantization the upper bound (2.44) and various existing constructions for a 4 antenna system. The upper bound is plotted as a solid curve and the performance of random vector quantization is the reference and corresponds to a value of 0 at each rate. Note that both the construction of Hochwald and that used in the WiMax standard perform similarly for 3 bits. However, the gap is much larger at 6 bits. However the quantizer from the WiMax standard performs similar to that of RVQ at 6 bits.

Examining Figure 3-4 one may see that the constructions presented do very well in terms of SINR_{sat} relative to the performance of RVQ as well as contain many orthogonal bases for a system in which there are very few users. However, in general, codes that contain many orthogonal bases perform worse than those which contain fewer orthogonal bases in terms of the quantization error and SINR_{sat} . However, this does not mean in general that a system which employs a channel quantizer with many orthogonal bases will in fact suffer an average loss in SINR as great as depicted in Figure 3-4. It is important to recall that SINR_{sat} is a high SNR approximation of the achieved SINR of a system that uses a particular quantization scheme and not a measure of the achieved SINR for a given SNR. Further, SINR_{sat} by definition assumes that there is a set of nearly orthogonal users and hence SINR_{sat} by definition does not favor codebooks with many orthogonal bases. In particular at moderate SNR there may be a considerably smaller gap between the expected SINR achieved by one of our constructions and RVQ as in general there will be a SINR penalty due to channel inversion with RVQ. Further, the definition of SINR_{sat} is only in terms of the quantization error of a single-user. As previously noted, in MIMO systems with many users the order statistic for the quantization error leads to similar performance. In the sequel, we show that the same is true for systems in which the number of users is only a small multiple of the size of the transmit array. Hence, in such systems one expects, by choosing the users that have the best quantization error, the gap between the achieved average SINR of a system which uses a channel quantizer with many orthogonal bases and one without many orthogonal bases to be smaller. As our general constructions perform well independent of this effect we postpone this discussion until Section 4.6.

Index (r, \perp -Bases)	\perp -Bases	Construction	Reference
(3,4)	4	$\mathcal{C}_{\mathbb{Z}}^{(2,4)}(3; [[0, 1]])$	(3.6)
(3,Z1)	0	Hochwald 3-bit	[56]
(3,Z2)	0	WiMax 3-bit	[1, 143]
(4,8)	8	$\mathcal{C}_{\mathbb{Z}}^{(2,4)}(3; [[0, 1]]) \cup \mathcal{C}_{\mathbb{Z}}^{(2,4)}(3; [[1, 0]])$	(3.6)
(4,4)	4	MUB(4)/ $\mathcal{C}_{\mathcal{T}}(2, [0, 0], 0)$	[61, 76]/(3.54)
(4,12)	12	$\mathcal{C}_{\mathcal{T}}(2, [1, 0], 0)$	(3.54)
(5,26)	26	$\mathcal{C}_{\text{ASC}}^*(2, 2)$	Example 3.2.6
(5,36)	36	$\mathcal{C}_{\mathbb{Z}}^{(2,4)}(3; [[0, 0], [0, 1]]) \cup \mathcal{C}_{\mathbb{Z}}^{(2,4)}(3; [[1, 0]]) \cup \mathcal{C}_{\mathcal{T}}(2, [0, 0], 0)$	(3.6), (3.54)
(5,32)	32	$\mathcal{C}_{\mathcal{T}}(2, [0, 0], 0) \cup \mathcal{C}_{\mathcal{T}}(2, [0, 0], 2)$	(3.54)
(5,12)	12	$\mathcal{C}_{\text{sparse}}^{(2,4)}(2)$	(3.7)
(6,105)	105	$\mathcal{C}_{\text{ASC}}(2, 0)$	[13]/Example 3.2.6
(6,16)	16	$\mathcal{C}_{\mathcal{T}}(3, [1, 0], 0)$	(3.54)
(6,4)	4	$\mathcal{C}_{\mathbb{F}}(0.6777, 0.5305 + 0.7425 \cdot i, \mathcal{C}_{\mathbb{Z}}^{(2,4)}(3; [[0, 1]]))$	(3.11)
(6,Z3)	0	Hochwald 6-bit	[56]
(6,48)	48	$\mathcal{C}_{\text{sparse}}^{(2,4)}(3)$	(3.7)
(6,Z5)	0	WiMax 6-bit	[1, 143]
(7,233)	233	$\mathcal{C}_{\text{ASC}}(3, 2)$	Example 3.2.6
(7,112)	112	$\mathcal{C}_{\mathbb{Z}}^{(2,4)}(4; [[0, 1]]) \cup \mathcal{C}_{\mathbb{Z}}^{(2,4)}(4; [[1, 0]]) \cup \mathcal{C}_{\mathcal{T}}(3, [0, 0], 0)$	(3.6), (3.54)
(7,128)	128	$\mathcal{C}_{\mathcal{T}}(3, [0, 0], 0) \cup \mathcal{C}_{\mathcal{T}}(3, [0, 0], 2)$	(3.54)
(7,192)	192	$\mathcal{C}_{\text{sparse}}^{(2,4)}(4)$	(3.7)
(8,393)	393	$\mathcal{C}_{\text{ASC}}(3, 1)$	Example 3.2.6
(8,4)	4	$\mathcal{C}_{\mathbb{F}}(0.2303, 0.6817 + 1.9577 \cdot i, \mathcal{C}_{\mathcal{T}}(2, [0, 0], 0))$	(3.11)
(8,768)	768	$\mathcal{C}_{\text{sparse}}^{(2,4)}(5)$	(3.7)
(9,1097)	1097	$\mathcal{C}_{\text{ASC}}(3, 0)$	Example 3.2.6
(9,26)	26	$\mathcal{C}_{\mathbb{F}}(0.0100, 0, \mathcal{C}_{\text{ASC}}(2, 2))$	(3.11)
(10,2289)	2289	$\mathcal{C}_{\text{ASC}}(4, 1)$	Example 3.2.6
(10,1521)	1521	$\mathcal{C}_{\text{ASC}}(4, 2)$	Example 3.2.6
(10,26)	26	$\mathcal{C}_{\mathbb{F}}(0.5872, 0.4628 + 0.6790 \cdot i, \mathcal{C}_{\text{ASC}}(2, 2))$	(3.11)
(11,14577)	14577	$\mathcal{C}_{\text{ASC}}(4, 0)$	Example 3.2.6
(12,105)	105	$\mathcal{C}_{\mathbb{F}}(0.3639, 1.9529, \mathcal{C}_{\text{ASC}}(2, 1))$	(3.11)

Table 3.2. A list of good quantizers in \mathbb{C}^4 we develop in the sequel. Pre-existing constructions are highlighted. The first column is used to index the simulated performance of each code in Figure 3-4 and Figures 3-10 – 3-14. The second column contains the number of orthonormal bases for \mathbb{C}^4 contained in the code and the last column contains a reference (possibly forward in the thesis) to the construction. The performance of these constructions may be seen in Figure 3-4.

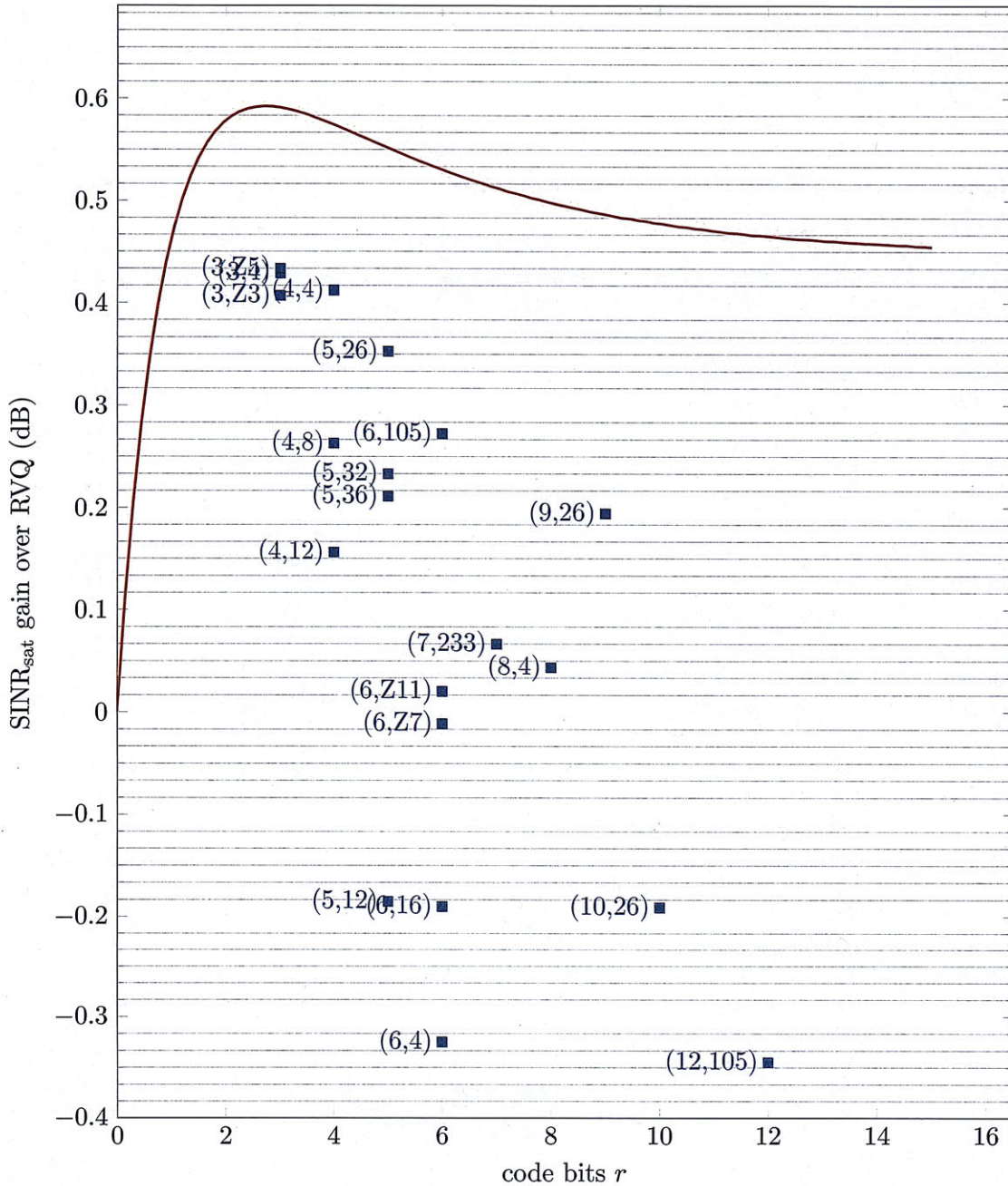


Figure 3-4. The difference in SINR_{sat} between random vector quantization the upper bound (2.44) and various constructions for 4 antennas as labeled in Table 3.2. The upper bound is plotted as a solid curve and the performance of random vector quantization is the reference and corresponds to a value of 0 at each rate. We note that the quantizers from Table 3.2 which achieve a value for SINR_{sat} that is 1.3 dB or more dB below RVQ are not depicted. The constructions presented do very well in terms of SINR_{sat} relative to the performance of RVQ as well as contain many orthogonal bases for a system in which there are very few users. However, in general, codes that contain many orthogonal bases perform worse than those which contain fewer orthogonal bases in terms of the quantization error and SINR_{sat} .

■ 3.2 Systematic Construction of Channel Quantizers

In a multi-user MIMO system it is of interest to have a quantization codebook for which the average quantization error is small and the codebook contains many orthogonal bases. In this section we examine a framework to construct channel quantizers that can balance these properties. To achieve this flexibility we begin by constructing codebooks of fixed sparsity, i.e. codebooks in which every codeword has a support of fixed size. We then overview the constraints that one must place on codes of varying sparsity used in a union to form higher rate codes with good mean squared quantization error. Then, we proceed to our geometric motivation for the linear operations used to form high rate codes with low complexity quantization algorithms.

■ 3.2.1 Introduction to Component Code Constructions

In order to derive our systematic construction of component codes we start with a simple construction that leads to our more general construction to follow. We are interested in forming a code with fixed sparsity. That is, a code in which every codeword has a support of fixed size. A natural way to form such a channel quantizer is to embed a lower dimensional channel quantizer in a higher dimensional space. Suppose one is given a dense matrix, say $\mathbf{C}_B \in \mathbb{C}^{m_0 \times J}$, where $m_0 < m$ and suppose that the columns of \mathbf{C}_B form a “good” channel quantizer in \mathbb{C}^{m_0} . At present we let \mathbf{C}_B be an arbitrary complex $m_0 \times J$ matrix and note we develop a family of good dense matrices which we use in our construction in Section 3.5. Now, the most natural way to construct a quantizer in \mathbb{C}^m from \mathbf{C}_B is to view the columns of \mathbf{C}_B as the non-zero components of a set of sparse vectors in \mathbb{C}^m by choosing a constant support for each vector. In particular, let

$$\mathcal{I}_0 = \{i_0, i_1, \dots, i_{m_0}\} \subset \{0, 1, 2, \dots, m-1\}$$

be the support chosen for the code in \mathbb{C}^m and let

$$\mathbf{s} = [i_0, i_1, \dots, i_{m_0}]$$

be the vector³ which indexes the non-zero coordinates of the constructed code in \mathbb{C}^m . Then, one may construct a quantizer in \mathbb{C}^m associated to the columns of the matrix

$$\mathbf{C}_0[\mathbf{s}, :] = \mathbf{C}_B.$$

More precisely, one may construct a code $\mathcal{C}_0 \subset \mathbb{C}^m$, where

$$\mathcal{C}_0 = \{\mathbf{C}_0[:, i]\}_{i=0}^J.$$

The code \mathcal{C}_0 will leave portions of the complex unit m -sphere poorly covered leading to a small value of SINR_{sat} at high rate as any channel vector which has a dominate component off the support of \mathcal{C}_0 will have a large quantization error. Thus, in our construction we permute the code \mathcal{C}_0 several times to form a higher rate code which better covers the complex unit m -sphere. In particular, let $\{\Pi_{\tau_1}, \Pi_{\tau_2}, \dots, \Pi_{\tau_t}\}$ be a set of t matrices describing permutations to the *rows* of \mathbf{C}_0 . Then, one may consider a channel quantizer which consists

³We note that the ordering may be taken arbitrarily with out effecting the results.

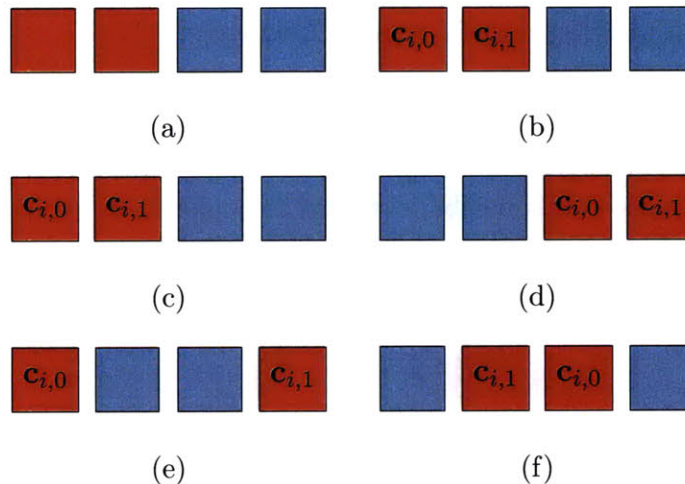


Figure 3-5. A depiction of the general quantization scheme for component codes. First a subset of coordinates are selected for the base code as depicted in (a) where the first two coordinates have been selected. Then a code is formed over this subset of indices as depicted in (b). Last, a larger code is formed by permuting the coordinates of the base code as seen in (c)–(f).

of the columns of the $m \times (J \cdot t)$ complex matrix

$$\mathbf{C} = [\mathbf{C}_0, \Pi_{\tau_1} \cdot \mathbf{C}_0, \Pi_{\tau_2} \cdot \mathbf{C}_0, \dots, \Pi_{\tau_t} \cdot \mathbf{C}_0].$$

That is, one can construct a channel quantizer

$$\mathcal{C} = \{\mathbf{C}[:, i]\}_{i=0}^{J \cdot t}$$

with $J \cdot t$ codewords. Thus, in this framework every quantizer of interest is specified by

1. A dense $m_0 \times J$ matrix \mathbf{C}_B
2. A support for \mathbf{C}_B in $\mathbb{C}^m, \mathcal{I}_0$
3. A collection of permutations $\{\Pi_{\tau_1}, \Pi_{\tau_2}, \dots, \Pi_{\tau_t}\}$

A depiction of this construction may be seen in Figure 3-5.

One may systematically construct component codes of varying rates through one's choice of $\mathbf{C}_B, \mathcal{I}_0$ and $\{\Pi_{\tau_1}, \Pi_{\tau_2}, \dots, \Pi_{\tau_t}\}$. In particular, one may systematically construct a rate r code by solving

$$\max_{0 < m_0 < m} \max_{J > 0} \max_{\mathbf{C}_B \in \mathbb{C}^{m_0 \times J}} \max_{\mathcal{I}_0 \subset \{0, 1, 2, \dots, m-1\}} \max_{\Pi_{\tau_1}, \Pi_{\tau_2}, \dots, \Pi_{\tau_{\lfloor 2^r/J \rfloor}}} \text{SINR}_{\text{sat}}(\mathcal{C})$$

However, as one may expect, optimization of codes from this construction is quite hard in general as there are many free parameters. Thus, we take a more formal position to specify our component codes in the sequel which allows us to identify good systems of dense matrices as well as structured sets of permutations.

To begin we note that our general construction may introduce non-distinct codewords. That is, as we have placed no restriction on the relationship of the support \mathcal{I}_0 of \mathbf{C}_B in \mathbb{C}^m , the structure of \mathbf{C}_B and the set of permutations $\{\Pi_{\tau_1}, \Pi_{\tau_2}, \dots, \Pi_{\tau_t}\}$ we have no guarantee that a chosen “rate r ” construction contains 2^r distinct codewords. Thus, for more efficient optimization of codes one may develop a systematic method to develop channel quantizers

which intelligently excludes combinations of \mathcal{I}_0 , \mathbf{C}_B and $\{\Pi_{\tau_1}, \Pi_{\tau_2}, \dots, \Pi_{\tau_t}\}$ that yield non-unique codewords. In this direction, we note that the elements of \mathbf{C}_0 determine a bi-variate function. That is, for any matrix \mathbf{C}_0 we may associate a bi-variate function

$$c(i, j) = \mathbf{C}_0[i, j].$$

More precisely, $c(i, j)$ is a function from $\{0, 1, 2, \dots, m-1\} \times \{0, 1, 2, \dots, J-1\}$ to \mathbb{C} where $c(i, j) = \mathbf{C}_0[i, j]$.

A crucial observation we exploit in the sequel is that one has the freedom to choose both the domain and the range of the function $c(i, j)$. That is, our present choice of labels for the rows and columns of \mathbf{C}_0 and \mathbf{C}_B are irrelevant. We may rather choose two abstract sets \mathcal{D}_1 and \mathcal{D}_2 such that $|\mathcal{D}_1| = m$ and $|\mathcal{D}_2| = J$ as labels for the rows and columns of \mathbf{C}_0 and \mathbf{C}_B . Then, by determining a function $\tilde{c}(i, j)$ from $\mathcal{D}_1 \times \mathcal{D}_2$ to \mathbb{C} equivalent to $c(i, j)$ one may obtain an equivalent definition for any quantizer in our previous framework. This is an important observation as one's choice for \mathcal{D}_1 , \mathcal{D}_2 and function $c(i, j)$ effect the mean square error performance as well as one's ability to provide a succinct representation of codewords which makes identifying orthogonal bases simple.

In the sequel we label the rows of \mathbf{C}_0 by the set \mathcal{I} and label the support of a code by \mathcal{I}_0 . Alternately, one may view \mathcal{I}_0 as the row labels of \mathbf{C}_B . We label the columns of \mathbf{C}_0 (or alternatively \mathbf{C}_B) by Υ_1 and denote the set of permutations $\{\Pi_{\tau_1}, \Pi_{\tau_2}, \dots, \Pi_{\tau_t}\}$ now defined on an abstract set as Υ_2 . Thus, every quantizer in this framework may be alternatively given by

1. \mathcal{I} , row labels for \mathbf{C}_0
2. \mathcal{I}_0 , the support of the rows of \mathbf{C}_0
3. Υ_1 , an index set for the columns of \mathbf{C}_0 (or \mathbf{C}_B)
4. Υ_2 , a set of permutations of \mathcal{I}
5. $c(i, j)$, a map from $\mathcal{I}_0 \times \Upsilon_1$ to \mathbb{C} which describes the entries of \mathbf{C}_B

This new characterization of our component codes gives rise to a new representation in which the relationships between $c(i, j)$, Υ_1 and Υ_2 may be better understood. In particular, this yields a framework in which we can identify orthogonal codewords and orthogonal bases. We present the framework we use to identify orthogonal bases as well as constructions that produce co-linear codewords in Section 3.3. There we also develop a usefully choice for the function $c(i, j)$. Then, in Section 3.5 we present a family of good choices for the matrix \mathbf{C}_B which yield codes of varying rates.

To begin, let \mathbf{e}_i be the element of the standard basis such that a one stands in the i th position and is otherwise zero. More precisely,

$$\mathbf{e}_i = (0, 0, \dots, 0, 1, 0 \dots, 0).$$

Now, for any given codebook $\mathcal{C}_0 \subset \mathbb{C}^m$ of cardinality J let \mathcal{I}_0 be the support of the code \mathcal{C}_0 , i.e. the subset of $\{0, 1, \dots, m-1\}$ for which there is an element $\mathbf{c}_0 \in \mathcal{C}_0$ such that $\mathbf{c}_0^\dagger \mathbf{e}_i \neq 0$ if $i \in \mathcal{I}_0$. In the sequel we index the codewords in \mathcal{C}_0 via a set $\Upsilon_1 = \{j_1, j_2, \dots, j_J\}$. Thus, the code \mathcal{C}_0 consist of the vectors

$$\mathbf{c}(j) = \sum_{i \in \mathcal{I}_0} c(i, j) \cdot \mathbf{e}_i \tag{3.2}$$

for some set of complex numbers $\{c(i, j)\}_{i \in \mathcal{I}_0, j \in \Upsilon_1}$. We note that the set of coefficients

$\{c(i, j)\}_{i \in \mathcal{I}_0, j \in \Upsilon_1}$ are simply the elements of \mathbf{C}_B in our previous construction. Given a set of permutations $\Upsilon_2 = \{\hat{\tau}_1, \hat{\tau}_2, \dots, \hat{\tau}_t\}$ we extend the codebook \mathcal{C}_0 by including the complex vectors that are permutations of vectors in \mathcal{C}_0 . In particular, we extend \mathcal{C}_0 to a higher rate code, say \mathcal{C} , by including the complex vectors

$$\mathbf{c}(j, \hat{\tau}_k) = \sum_{i \in \mathcal{I}_0} c(i, j) \cdot \mathbf{e}_{\hat{\tau}_k(i)}. \quad (3.3)$$

We note that the above description of our quantization framework is still quite general. In particular, this description may be used to characterize any quantizer by taking $\mathcal{I}_0 = \{0, 1, \dots, m-1\}$ and $\{c(i, j)\}$ to be an arbitrary set of complex numbers. The case when the numbers $c(i, j)$ come from a structured set is of particular interest. In fact in the sequel we present a systematic construction of sets of coefficients $c(i, j)$ that is general enough to describe a large set of constructions of quantizers found in literature [13, 48, 55, 56, 113, 138]. Before proceeding we briefly consider some examples of quantizers of the form (3.3).

Example 3.2.1 A Unit Cube in \mathbb{C}^4

We now consider a construction of a code that is the standard basis in \mathbb{C}^4 using the framework of (3.3). To begin we first construct the standard basis using our original notation. In this direction, let $\mathbf{C}_B = [1]$ and $\mathbf{s} = [0]$. Then, $\mathbf{C}_0 = [1, 0, 0, 0]^\dagger$. As this is obviously a poor choice for a channel quantizer we permute this code using the “right circular shift” permutation 3 times.

$$\Pi_{\text{rshift}} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

That is, we permute the code using the permutations Π_{rshift} , Π_{rshift}^2 and Π_{rshift}^3 . In our more formal framework this may alternatively be constructed as

1. $\mathcal{I}_0 = \{0\}$
2. $\Upsilon_1 = \{0\}$,
3. $\Upsilon_2 = \{(0, 0), (0, 1), (0, 2), \dots, (0, m-1)\}$
4. $c(i, j) = \delta(i - j)$

where $\delta(x) = 1$ if $x = 0$ and is 0 otherwise and in turn where (i, j) is the permutation that takes $i \rightarrow j$ and $j \rightarrow i$ and leaves all other elements fixed. Then, it is easy to see by direct computation that

$$\begin{aligned} \mathbf{c}(0, (0, 0)) &= [1, 0, 0, 0]^\dagger, & \mathbf{c}(0, (0, 1)) &= [0, 1, 0, 0]^\dagger, \\ \mathbf{c}(0, (0, 2)) &= [0, 0, 1, 0]^\dagger, & \mathbf{c}(0, (0, 3)) &= [0, 0, 0, 1]^\dagger \end{aligned}$$

We note that one may also construct the standard basis in \mathbb{C}^4 using $\mathcal{I}_0 = \{0, 1, 2, 3\}$, $\Upsilon_2 = \{(0, 0)\}$ and $c(i, j) = \delta(i - j)$ in this framework.

A second simple example considers the selection of columns of a discrete Fourier transform matrix (DFT). This serves as the core construction for the WiMax (802.16e) standard.

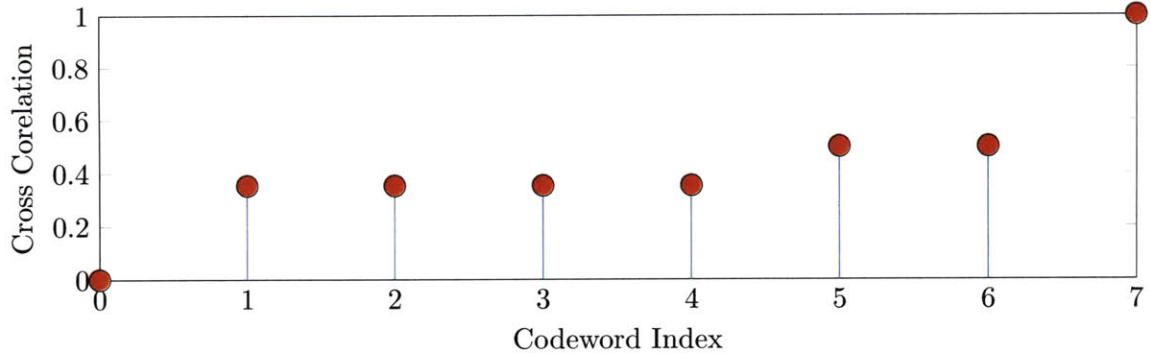


Figure 3-6. The cross correlation spectrum of the codewords from Example 3.2.2. Each stem represents an inner product between a fixed (arbitrary) codeword selected for the code of Example 3.2.2 and another codeword from this code. Note that there is only one vector orthogonal to any given vector of the code while the remaining correlation values are approximately constant. Thus, the associated Voronoi has a low second moment and small mean squared quantization error. The inner product which obtains a value 1 is the inner product of the fixed codeword with itself.

Example 3.2.2 Hochwald 3-Bit DFT Code in \mathbb{C}^4

We now examine the systematic construction of beamforming vectors of Hochwald et. al. [56]. Recall that the $N \times N$ DFT matrix is the matrix for which the entry at position (i, j) is the complex exponential ζ^{ij} where $\zeta = \exp(2\pi\sqrt{-1}/N)$ is an N -th root of unity. One can construct a 4 dimensional codebook by selecting 4 out of the N columns of the $N \times N$ DFT matrix to use as the components of the codeword. That is, let

1. $\mathcal{I}_0 = \{1, 2, 3, 4\}$
2. $\Upsilon_1 = \{0, 1, 2, \dots, 7\}$
3. $\Upsilon_2 = \{(0, 0)\}$
4. $c(i, j) = \zeta^{iu_j}$

Hence,

$$\mathbf{c}(i, (0, 0)) = \sum_{j=0}^3 \zeta^{iu_j} \mathbf{e}_j. \quad (3.4)$$

We note that proposals for the WiMax standard [143] the 3 bit quantizer considered uses a vector $\mathbf{w} = [1, 2, 7, 6]$. We note that with this choice of vector there are is a very diverse range in the magnitude of the cross correlation between codewords. This can be seen in Figure 3-6.

Note that (3.3) is general enough to describe every set of lines in \mathbb{C}^m and hence we will look a subclass of this framework in order to identify quantizers with many subsets of orthogonal bases. In particular, we will show that by considering quantizers for which only a subset of permutations are allowed in the choice of Υ_2 will provide a useful mechanism in understanding the configurations of lines in the corresponding quantizer. To begin, we first slightly extend (3.3) by allowing a general indexing of the standard basis. In particular, we let \mathcal{I} be an arbitrary indexing of the standard basis such that there is a one-to-one correspondence between \mathcal{I} and $\{0, 1, 2, \dots, m-1\}$. We assume that this indexing has been

chosen so that the permutations from Υ_2 act as linear translations on \mathcal{I} , i.e.

$$\hat{\tau}(i) = i + \tau,$$

for some $\tau \in \mathcal{I}$ and all $i \in \mathcal{I}$. Hence, in the sequel we require that \mathcal{I} is closed under addition and restrict Υ_2 to only contain permutations that act as translation on \mathcal{I} . Then, (3.3) becomes

$$\mathbf{c}(j, \hat{\tau}_k) = \sum_{i \in \mathcal{I}_0 \subset \mathcal{I}} c(i, j) \cdot \mathbf{e}_{\tau_k+i}. \quad (3.5)$$

In the sequel we let $\Upsilon_2 = \{\tau_1, \tau_2, \dots, \tau_k\} \subset \mathcal{I}$ be the set of linear shifts that describe the coordinate permutations. Thus, every quantizer in this framework is given by

1. \mathcal{I} , an indexing of the standard basis
2. \mathcal{I}_0 , the support of the base code \mathcal{C}_0
3. Υ_1 , an index set for the base code \mathcal{C}_0
4. Υ_2 , a subset of \mathcal{I} describing the “shifts” on the basis
5. $\{c(i, j)\}_{i \in \mathcal{I}_0, j \in \Upsilon_1}$, a set of complex numbers describing the codewords of \mathcal{C}_0

We now provided a re-derivation of Example 3.2.1 using the framework in (3.5).

Example 3.2.3 A Second Construction of a Unit Cube in \mathbb{C}^4 .

We now consider a less trivial construction of the code of Example 3.2.1 where we index elements of the standard basis by elements of \mathbb{F}_2^2 , i.e. binary vectors of length 2. If $\mathbf{a} = [a_1, a_2] \in \mathbb{F}_2^2$ we let

$$\mathbf{e}_{\mathbf{a}} = \mathbf{e}_{a_1+2a_2}.$$

To construct our quantizer we let

1. $\mathcal{I} = \mathbb{F}_2^2$
2. $\mathcal{I}_0 = \{[0, 0]\}$,
3. $\Upsilon_1 = \{[0, 0]\}$,
4. $\Upsilon_2 = \mathcal{I}$
5. $c(\mathbf{a}, \mathbf{b}) = (\sqrt{-1})^{\mathbf{a}^\dagger \mathbf{b}}$

where $\mathbf{a}^\dagger \mathbf{b}$ is the inner product^a of \mathbf{a} and \mathbf{b} as vectors in \mathbb{C}^2 (not as binary vectors). That is

$$\mathbf{a}^\dagger \mathbf{b} = a_1 b_1 + a_2 b_2.$$

Then it is easy to see by direct computation that

$$\begin{aligned} \mathbf{c}_{[0,0],[0,0]} &= [1, 0, 0, 0]^\dagger & \mathbf{c}_{[0,0],[1,0]} &= [0, 1, 0, 0]^\dagger \\ \mathbf{c}_{[0,0],[0,1]} &= [0, 0, 1, 0]^\dagger \text{ and } & \mathbf{c}_{[0,0],[1,1]} &= [0, 0, 0, 1]^\dagger \end{aligned}$$

^aThis could be defined equivalently to be the inner product modulo 4 as $\sqrt{-1}$ is a fourth root of unity and the elements of \mathbf{a} and \mathbf{b} are integral.

In the sequel we consider a framework for the development of quantizers similar to that of Example 3.2.3. In particular, we will consider quantizers in which the basis is labeled by a finite field and the support is described by a sub-field (sub-space) of the finite field used to label the basis. Further, we use functions $c(i, j)$ which have a range that is a subset of the

unit circle. Hence, every quantizer in this framework contains codewords with coordinates that have a magnitude of zero or one. We show that this framework is general enough to yield a design framework that is flexible enough to meet a variety of design objectives. In this direction we provide the following example of a 3-bit quantizer over \mathbb{C}^4 . We note that this particular example illustrates many of our insights to follow and use it frequently in the sequel.

Example 3.2.4 Four Orthogonal Bases for \mathbb{C}^4 with 3-Bits

We now consider a simple code of the form (3.5) that is the union of two non-standard orthogonal bases in \mathbb{C}^4 . As in Example 3.2.3 we index elements of the standard basis by elements of \mathbb{F}_2^2 and let $c(\mathbf{a}, \mathbf{b}) = (\sqrt{-1})^{\mathbf{a}^\dagger \mathbf{b}}$ where $\mathbf{a}^\dagger \mathbf{b}$ is the inner product of \mathbf{a} and \mathbf{b} as vectors over \mathbb{C}^2 . However, here we let $\mathcal{I}_0 = \{[0, 0], [0, 1]\}$ and index codewords by the elements of \mathbb{Z}_4^2 . In particular, we let:

1. $\mathcal{I} = \mathbb{F}_2^2$
2. $\mathcal{I}_0 = \{[0, 0], [0, 1]\}$
3. $\Upsilon_1 = \{[0, 0], [0, 1], [0, 2], [0, 3]\}$
4. $\Upsilon_2 = \{[0, 0], [1, 0]\}$
5. $c(\mathbf{a}, \mathbf{b}) = (\sqrt{-1})^{\mathbf{a}^\dagger \mathbf{b}}$

Then, by direct computation it is easy to see that the resulting code is the union of the two orthogonal bases:

$$\mathcal{B}_1 = \{[1, 1, 0, 0], [0, 0, 1, 1], [1, -1, 0, 0], [0, 0, -1, 1]\}$$

for $\Upsilon_1 = \{[0, 0], [0, 2]\}$ and

$$\mathcal{B}_2 = \{[1, \sqrt{-1}, 0, 0], [1, -\sqrt{-1}, 0, 0], [0, 0, \sqrt{-1}, 1], [0, 0, -\sqrt{-1}, 1]\}$$

for $\Upsilon_1 = \{[0, 1], [0, 3]\}$. Note using this construction the magnitude of any inner product between the two bases is quite regular. In particular, the magnitude of the inner product between any vector from \mathcal{B}_1 with any vector from \mathcal{B}_2 is 0 or $\sqrt{2}$. The orthogonal codewords may be seen in Figures 3-8 and 3-9. The remaining inner product relations may be seen in Figure 3-7.

The quantizer of Example 3.2.4 is the best performing three bit quantizer we develop. It is natural to wonder if such a simple construction will perform as well in general.

■ 3.2.2 Introduction to Systematic Unions of Component Codes

As a first attempt to systematically construct a quantizer of varying rates, one may consider constructing higher and higher rate codes by considering a quantizer for which the codewords are indexed by elements of \mathbb{Z}_{2^k} as opposed to \mathbb{Z}_4 . In the sequel, we consider a class of quantizers similar to that in Example 3.2.4 where \mathcal{I}_0 is chosen to be a subspace of \mathbb{F}_2^2 and Υ_1 is taken over a ring of larger and larger cardinality to increase the code rate. In particular, in the sequel we consider a quantizer with:

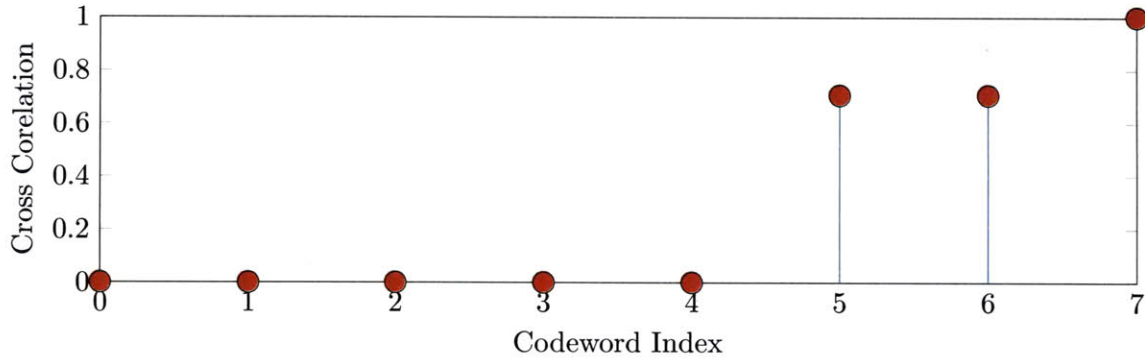


Figure 3-7. The cross correlation spectrum of the codewords from Example 3.2.4. Each stem represents an inner product between a fixed (arbitrary) codeword selected for the code of Example 3.2.4 and another codeword from this code. Note that each codeword is orthogonal to 5 codewords while the code from Example 3.2.2 only has one. However, for this property the code of Example 3.2.4 has considerably higher coherence and an irregularly shaped Voronoi cell leading to higher mean squared quantization error.

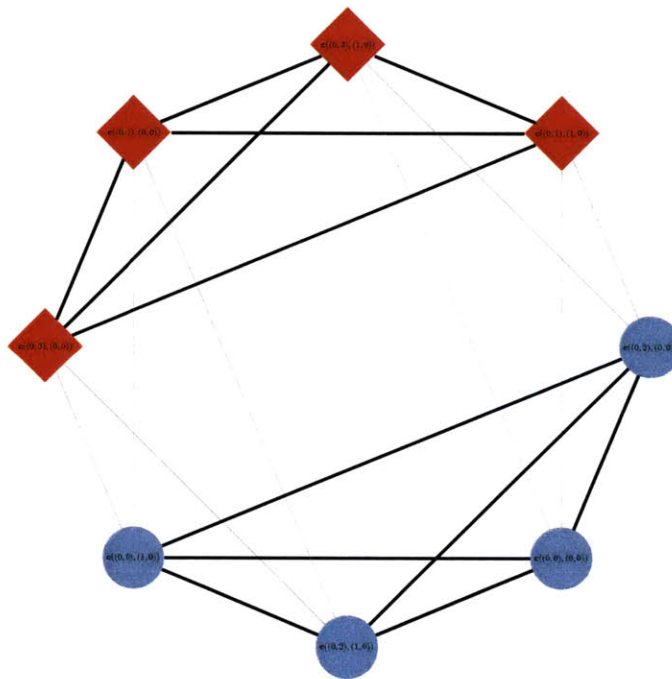


Figure 3-8. A depiction of the orthogonality relations between the codevectors of Example 3.2.4 as a graph. The codevectors of Example 3.2.4 are the vertices and an edge is placed between any two vertices if the corresponding codevectors are orthogonal. The vectors of basis \mathcal{B}_1 are depicted as circles while the vectors of basis \mathcal{B}_2 are depicted with a diamond. Note that this graph has 20 of the possible $\binom{8}{2} = 28$ edges. Moreover, there are four subsets of vectors that form an orthogonal basis. Two such subsets of nodes are depicted that correspond to the *orthogonal* bases \mathcal{B}_1 (filled red nodes) and \mathcal{B}_2 (filled blue nodes). The remaining two orthogonal bases can be seen in Figure 3-9.

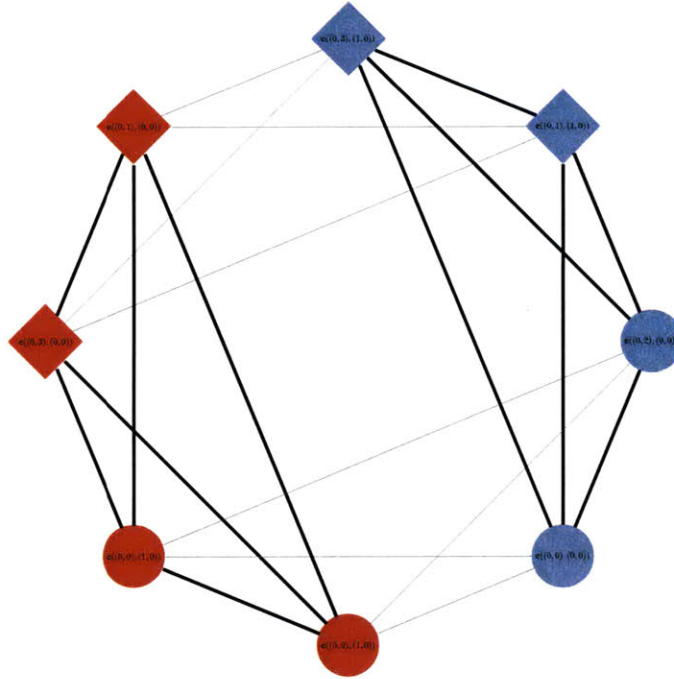


Figure 3-9. Two additional orthogonal bases for the codevectors of Example 3.2.4 as a graph. Here two vectors from basis \mathcal{B}_1 have been swapped with two vectors from \mathcal{B}_2 so that the resulting sets remain orthogonal.

1. $\mathcal{I} = \mathbb{F}_2^2$
2. $|\mathcal{I}_0| = 2$ and a subspace of \mathbb{F}_2^2
3. Υ_1 is the additive subset $\mathbb{Z}_{2^{k-1}}^2$,

$$\Upsilon_1 = \langle \mathbf{v} \mid \mathbf{v} \in \mathcal{I}_0 \rangle_{\mathbb{Z}_{2^{k-1}}} = \left\{ \sum_{\mathbf{v}_i \in \mathcal{I}_0} a_i \mathbf{v}_i \mid a_i \in \mathbb{Z}_{2^{k-1}} \right\}$$

4. $\Upsilon_2 = \mathbb{F}_2^2 / \mathcal{I}_0$
5. $c(\mathbf{a}, \mathbf{b}) = \exp\left(\frac{2\pi \cdot \sqrt{-1}}{2^{k-1}} \cdot \mathbf{a}^\dagger \mathbf{b}\right)$

which we denote as

$$\mathcal{C}_{\mathbb{Z}}^{(2,4)}(k; \mathcal{I}_0 \setminus \{[0, 0]\}). \quad (3.6)$$

With this definition the quantizer of Example 3.2.4 is simply $\mathcal{C}_{\mathbb{Z}}^{(2,4)}(3; [0, 1])$. Thus, as $\mathcal{C}_{\mathbb{Z}}^{(2,4)}(3; [0, 1])$ performs well, it is natural to consider the sequence of codes $\mathcal{C}_{\mathbb{Z}}^{(2,4)}(k; [0, 1])$ as this sequence inherits the same structure as $\mathcal{C}_{\mathbb{Z}}^{(2,4)}(3; [0, 1])$. However, this will do quite poorly for isotropic channel vectors as $\mathcal{C}_{\mathbb{Z}}^{(2,4)}(k; \cdot)$ only quantizes a few subspaces of dimension two.

In an attempt to more uniformly cover the complex unit m -sphere one may increase the quantization rate by forming codes over different supports by choosing different subspaces of \mathbb{F}_2^2 to index the support of the code. For example, in \mathbb{C}^4 , one may construct a new code

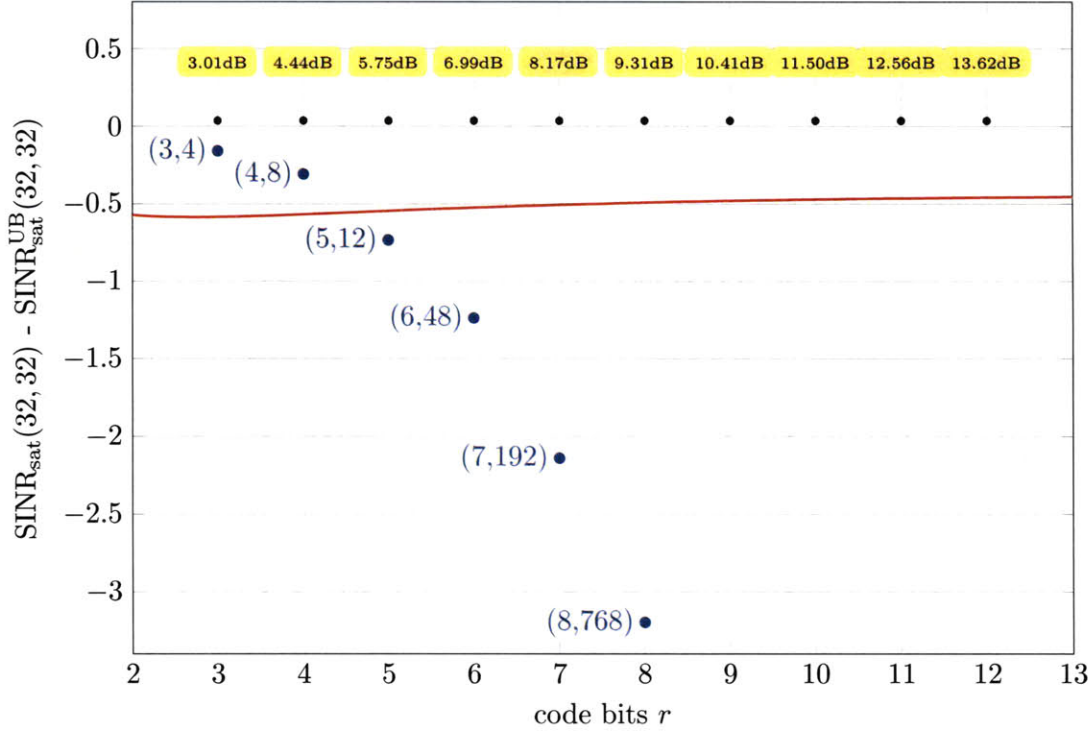


Figure 3-10. The performance of RVQ (solid curve) and the sequence of codes $\mathcal{C}_{\text{sparse}}^{(2,4)}(k)$ relative to the upper bound (2.44). Note that as the rate of the code increases from 3 bits to 8 bits the achieved performance rapidly deteriorates as the only channel vectors that have two dominate components will have low quantization error. That values taken by the upper bound are labeled at 0.

by taking the union of the codes which use the 3 subspaces of \mathbb{F}_2^2 ,

$$\{[0, 0], [0, 1]\}, \{[0, 0], [1, 0]\}, \{[0, 0], [1, 1]\},$$

to index the support of the codes. This, yields the code of size $3 \cdot 2^{k+1}$,

$$\mathcal{C}_{\text{sparse}}^{(2,4)}(k) \triangleq \mathcal{C}_{\mathbb{Z}}^{(2,4)}(k; [0, 1]) \cup \mathcal{C}_{\mathbb{Z}}^{(2,4)}(k; [1, 1]) \cup \mathcal{C}_{\mathbb{Z}}^{(2,4)}(k; [1, 0]). \quad (3.7)$$

As $\mathcal{C}_{\text{sparse}}^{(2,4)}(k)$ covers the sphere more uniformly than $\mathcal{C}_{\mathbb{Z}}^{(2,4)}(k+2; \mathcal{I}_0 \setminus \{[0, 0]\})$ for any choice of \mathcal{I}_0 , one should expect $\mathcal{C}_{\text{sparse}}^{(2,4)}(k)$ to perform better. However, it is still unclear how close to the upper bound (2.44) this code will be. We plot the performance of $\mathcal{C}_{\text{sparse}}^{(2,4)}(k)$ for $k = 3, 4, \dots, 8$ in Figure 3-10.

Note that as the rate of the code $\mathcal{C}_{\text{sparse}}^{(2,4)}(k)$ increases from 3 bits to 8 bits the achieved performance rapidly deteriorates compared to the upper bound. As the code $\mathcal{C}_{\mathbb{Z}}^{(2,4)}(3; [0, 1])$ performs quite well one may be curious to understand why this sequence does so poorly. The answer to this question may be seen naturally in \mathbb{R}^3 as depicted in Figure 3-11. As one increase the cardinality of the underlying ring only a few subspaces are more accurately quantized and regions of the sphere will be poorly covered as depicted in Figure 3-11 (a). However, by adding vectors from the standard basis, as seen in Figure 3-11 (b), one may obtain a more uniform covering of the sphere.

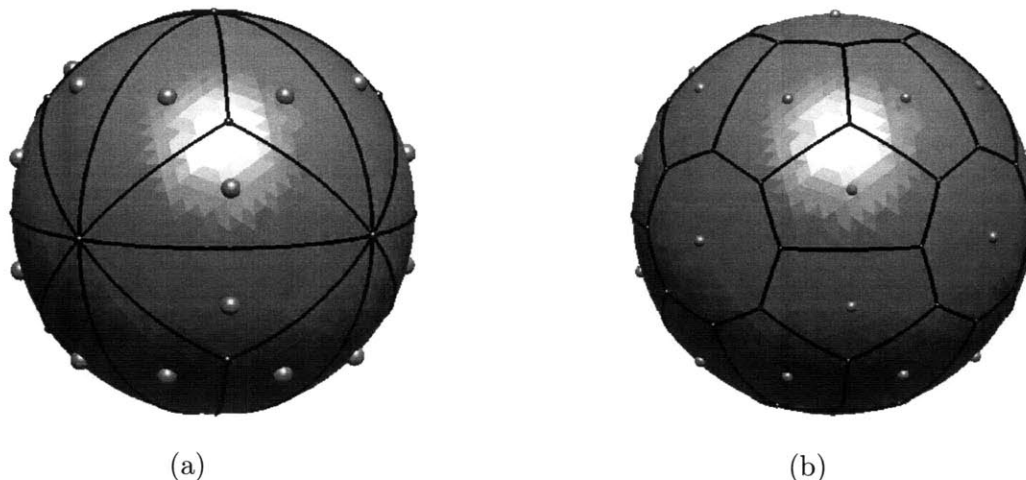


Figure 3-11. An illustration of the poor performance of the sequence of sparse codes. In (a) a code which is the union of vectors from a system of two dimensional subspaces. If one only increases the quantization rate in these subspaces regions of the sphere will be poorly covered as depicted in (a). However, by adding vectors from the standard basis for \mathbb{R}^3 to the code yields a more uniform covering of the sphere.

Only using codewords with a fixed sparsity to quantize channel vectors led to portions of the unit sphere being poorly covered. In particular, vectors that are more or less sparse will fall in regions that are distant from codewords. Thus, as the distribution of the user's channel vectors are assumed to be isotropic, channel vectors which have a single dominate coordinate will fall in one of the “wholes” of the quantizer. Similarly vectors with no dominate coordinate will also fall into these “wholes”. More precisely, if the magnitude of each coordinate of a channel vector is approximately constant will incur a quantization error that is approximately one half. Hence, for isotropic channel distributions it is of interest to develop codes which have both sparse and dense subcodes for accurate quantization of channel vectors with variations in the number of dominate components.

To design quantizers with low mean squared quantization error one in general should consider both dense and sparse codes. However, if one is interested in forming a union of such codes, one should not design the sparse and dense codes independently. One should rather ensure that they pair well together. In particular, from rate distortion theory one would like, in the limit of high rates, that the distribution of the codewords of the quantizer approximately match that of the channel vectors. In the particular case of isotropic fading one would like a quantizer to be distributed as uniformly on the sphere as possible. In the high rate limit this would imply that the resulting code is invariant to every unitary transformation to the code book. That is, in the high rate limit the distribution of a codeword selected uniformly at random should be isotropic and hence invariant to every unitary transformation. As such, an important metric for channel quantizers is the number of unitary transformations that fix the codebook. In this direction, we say that a unitary matrix, U , acts transitively⁴ on a codebook \mathcal{C} if every element of \mathcal{C} can be represented as

⁴We note that this definitions varies slightly from what is common in literature. However, in the cases we will study in the sequel a unitary matrix U acting on \mathcal{C} will be an element of a doubly transitive matrix group acting on the codebook by left multiplication.

the multiplication of U and an element of \mathcal{C} . More precisely, U acts transitively on \mathcal{C} if

$$U \cdot \mathcal{C} = \{U \cdot \mathbf{c} \mid \forall \mathbf{c} \in \mathcal{C}\} = \mathcal{C} \quad (3.8)$$

and we let

$$\text{Sym}(\mathcal{C}) = \{U \mid U \cdot \mathcal{C} = \mathcal{C}\} \quad (3.9)$$

be the set of all unitary matrices that act transitively on \mathcal{C} . Thus, $|\text{Sym}(\mathcal{C})|$ is a measure of how isotropic the quantizer is. We note that as RVQ uses an isotropic distribution to generate the codebook, one expects that code books from this ensemble to not behave poorly with regard to this metric, especially at high rates. However, for finite rates $|\text{Sym}(\mathcal{C})| = 0$ with probability one for any randomly generated vector quantizer and hence it is not unreasonable to expect that explicit constructions perform well at low-rates relative to RVQ if one ensures that $|\text{Sym}(\mathcal{C})|$ is large. In our present development this means that one must find sparse and dense codes that have similar symmetries. Using an equivalent constructions for every code on support of fixed sized one can ensure that this collection of codes has similar symmetries. However, this approach will led to the collection of codes with small supports naturally having more symmetries than the union of denser codes. This results from the symmetries arising from the “shifts” in the support. More precisely, to construct a plurality of sparse codes our quantization framework took translations of the linear space that indexed the coordinates. As this describes a coordinate permutation, which is a unitary transformation, sparse codes will in general have larger groups of symmetries than denser codes. Thus, to ensure the union of a dense code and sparse code have a large symmetry group, it is natural to impose this same structure on the dense quantizers to ensure the symmetries of the sparse code may be extended to the entire code increasing $|\text{Sym}(\mathcal{C})|$. This is an important subtlety of our construction that will take a bit of care and exposition to develop and make precise. However, we note that this is developed fully in Section 3.5 where we we define a family of “good” component codes with varying degrees of sparsity and rate which are all invariant to “shifts” in the support of the code. Thus, using the identified family of good sparse and dense codes⁵, say C_{good} , one may systematically construct quantizers by solving the design problem

$$\max_{T \subset C_{\text{good}}} \text{SINR}_{\text{sat}} \left(\bigcup_{\mathcal{C} \in T} \mathcal{C} \right). \quad (3.10)$$

This is developed fully in Section 3.5 and at present overview how one may use this forthcoming result to systematically design quantizers for multi-user MIMO systems.

■ 3.2.3 Introduction to Constructions of Low Complexity, High Rate Quantizers

In the preceding discussion we have described the key ingredients to our quantizer construction as a union of codes with differing supports which are all invariant to a set of shifts to the coordinate set. To increase the rate of the quantizer one may take one or many possible unions of codes and increase the cardinality of the integer ring underlying the construction of each of the constituent codes in the union. As each component code only contains code-words with coordinates that have a magnitude of zero or one, increase the cardinality of

⁵In the sequel we show that this is equivalent to optimizing over subsets of $\{0, 1, 2, \dots, m\}$ which satisfies a system of constraints provided in Theorem 3.5.5 which in general is much easier than the construction (3.1) at high rates.

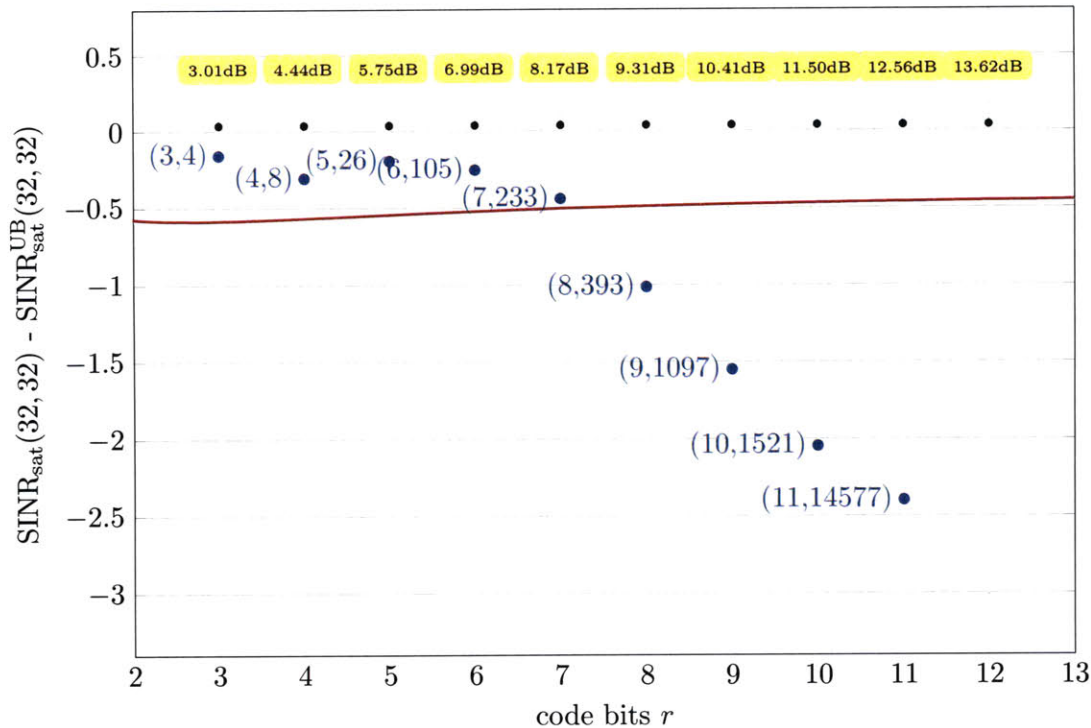


Figure 3-12. The performance of RVQ (solid curve) and a sequence of systematic constructions of codes constructed by first taking the union of sparse and dense codes then increasing the cardinality of the integer ring underlying the construction of each of the constituent codes in the union. The performance is plotted relative to the upper bound (2.44), the value of which is labeled at 0. Note that as the rate of the code increases from 3 bits to 7 bits the achieved performance is within 0.5 dB of the upper bound and performs better than RVQ. However, from 8 bits to 11 bits this approach once again rapidly deteriorates as only the phase of each coordinate is known more precisely.

the integer ring underlying the construction of each of the constituent codes in the union constructs codes of higher and higher rates by increasing the precision of the phase of each coordinate. Thus, in the high rate limit this scheme will only produce a code in which the *phase* of each coordinate is known precisely while the magnitude of each coordinate is only known only to finite precision. The performance of a code which takes unions of sparse and dense codes may be seen in Figure 3-12. Note the construction does quite well, outperforming RVQ from 3 bits to 7 bits and is within 0.5 dB of the upper bound. However, the performance begins to degrade at higher rates as only the phase of each coordinate is known more precisely.

For a truly systematic structured construction of channel quantizers one must find a systematic way to increase precision of the magnitude of every coordinate and not just the phase. To do this, one may consider taking unions of codes that are simple linear transformations of a “good” base code, say \mathcal{C}_r , in order to construct higher rate codebooks which uses some of the rate to increase the precision of the magnitude of each coordinate. In Section 3.6 we introduce a simple parametric family of operators that serves this purpose. In particular, we introduce a “localization” operation, $\mathbf{F}(\mathbf{c}_0, \alpha, \gamma)$, which takes any point on the complex sphere to a neighborhood of \mathbf{c}_0 described by α and γ . The freedom of α and γ allows one to tune this operation to optimize the performance of the resulting code. In

this direction, let

$$\mathcal{C}_{\mathbf{F}}(\alpha, \gamma, \mathcal{C}) = \bigcup_{\mathbf{c}_i \in \mathcal{C}} \mathcal{C}(\mathbf{c}_i; \alpha, \gamma, \mathcal{C}) \quad \text{where} \quad \mathcal{C}(\mathbf{c}_i; \alpha, \gamma, \mathcal{C}) = \mathbf{F}(\mathbf{c}_i; \alpha, \gamma) \cdot \mathcal{C} \quad (3.11)$$

One of the greatest benefits to this approach is it allows one to form *multi-resolution* codebooks which greatly simplifies the problem of quantization in high rate codes. In particular, by appropriately choosing the parameters α and γ one may ensure that each element of $\mathcal{C}(\mathbf{c}_i; \alpha, \gamma, \mathcal{C})$ is inside the Voronoi cell of the codeword \mathbf{c}_i in the original code. To illustrate this concept we now present an example of a universal codebook associated to a codebook in \mathbb{R}^3 .

Example 3.2.5 An Interpolated Icosahedron

In the following we successively refine the icosahedron to obtain a finer and finer quantization of the unit sphere in \mathbb{R}^3 . We do this by using the interpolation in (3.68). To begin, let $t = (1 + \sqrt{5})/2$, $\tau = t/\sqrt{1+t^2}$ and $\omega = 1/\sqrt{1+t^2}$. Then, the set of unit norm vectors that form the vertices of the icosahedron are the rows of

$$\mathbf{Q}_{\text{icos}} = \begin{bmatrix} \tau & \omega & 0 \\ -\tau & \omega & 0 \\ -\tau & -\omega & 0 \\ \tau & -\omega & 0 \\ \omega & 0 & \tau \\ \omega & 0 & -\tau \\ -\omega & 0 & -\tau \\ -\omega & 0 & \tau \\ 0 & \tau & \omega \\ 0 & -\tau & \omega \\ 0 & -\tau & -\omega \\ 0 & \tau & -\omega \end{bmatrix}$$

The rows of \mathbf{Q}_{icos} are a set of 12 points forming 6 lines in \mathbb{R}^3 . We note that each row of the Gram matrix $\mathbf{Q}_{\text{icos}}^T \mathbf{Q}_{\text{icos}}$ takes on the values 1 once, -1 once, $\pm(\tau^2 - \omega^2)$ twice, and $\pm\tau\omega = \pm(\tau^2 - \omega^2)$. Hence, the rows of \mathbf{Q}_{icos} are a set of 6 equiangular lines in \mathbb{R}^3 . We form a refinement of \mathbf{Q}_{icos} by adding the 30 lines (60 points) corresponding to the (unnormalized) set of points $a_1 \mathbf{c}_i + a_2 \mathbf{c}_j$ for $0 \leq \mathbf{c}_i^T \mathbf{c}_j < 1$ to \mathbf{Q}_{icos} and a fixed choice for a_1 and a_2 . The points of the icosahedron and the points of the universal code can be seen in Figure 3-13.

Thus, in this special case one may quantize any channel vector by first performing quantization using \mathcal{C} then, using the same quantization algorithm, perform quantization inside the local code $\mathcal{C}(\mathbf{c}_i; \alpha, \gamma, \mathcal{C})$ where \mathbf{c}_i is the result of the first stage of decoding. A multi-resolution codebook is a quite important property for a quantizer to have in a MIMO system as the quantization is performed at the user terminals. In many cases the user terminals are power and complexity limited and hence may not have the resources to perform high complexity quantization needed to obtain high rates. However, employing a well chosen base code \mathcal{C}_r and parameters α and γ one has the complexity of quantization at the user terminals using $\mathcal{C}_{\mathbf{F}}(\alpha, \gamma, \mathcal{C}_r)$ is two times that of the complexity of quantization using \mathcal{C}_r . Hence, irregardless of the performance of the codes $\mathcal{C}_{\mathbf{F}}(\alpha, \gamma, \mathcal{C}_r)$ relative to RVQ there is great practical relevance in a high rate system to employ the codes $\mathcal{C}_{\mathbf{F}}(\alpha, \gamma, \mathcal{C}_r)$.

The performance of the resulting codes may be seen in Figure 3-14. One can see that

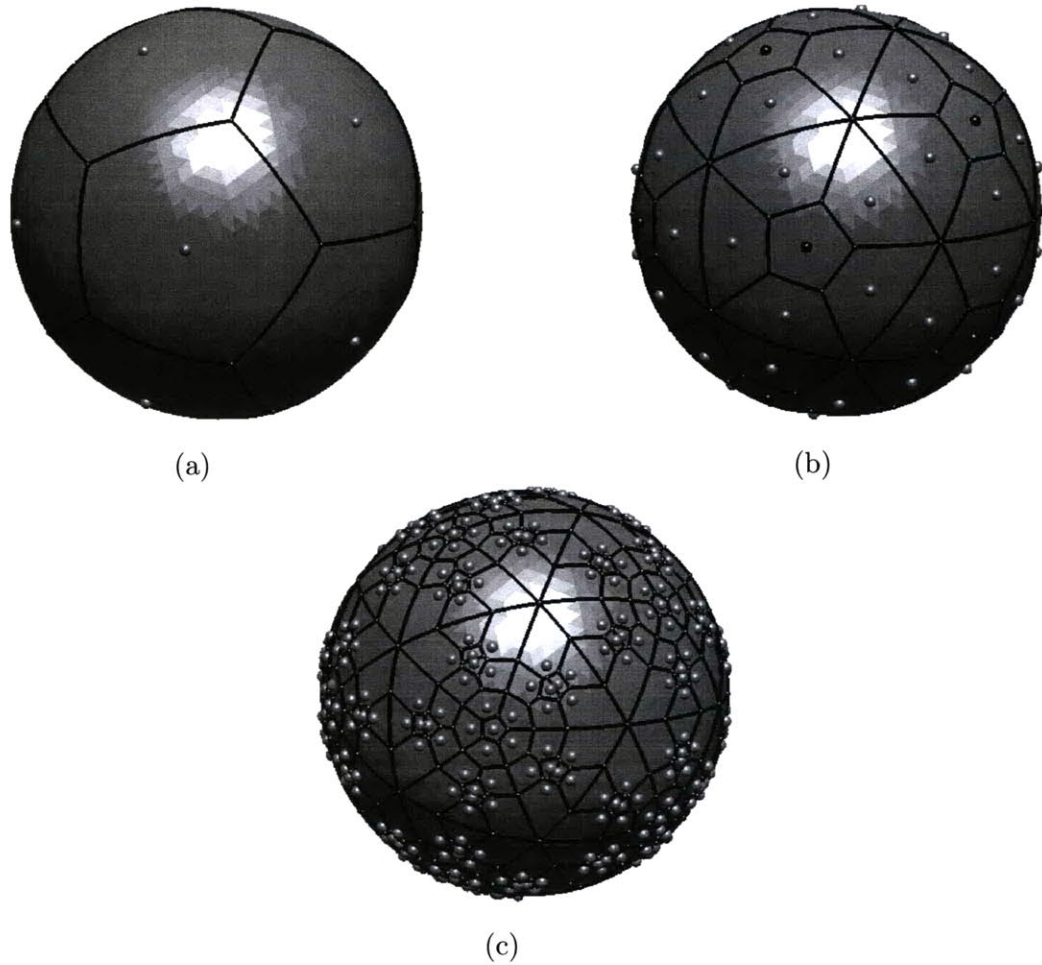


Figure 3-13. A depiction of the code in \mathbb{R}^3 that corresponds to the vertices of the icosahedron (a) and an associated universal code constructed by interpolating between the lines defined by the code in (a). The lines corresponding to the codewords in (a) are colored black. Note that each interpolation adds lines locally around each codeword from (a). The code in (c) is an additional interpolation of the lines defined by the code in (b).

these codes do perform quite well in \mathbb{C}^4 and up to 12 feedback bits are no more than 0.81 dB away from the optimal quantization scheme and no more than 0.35 dB away from random vector quantization.

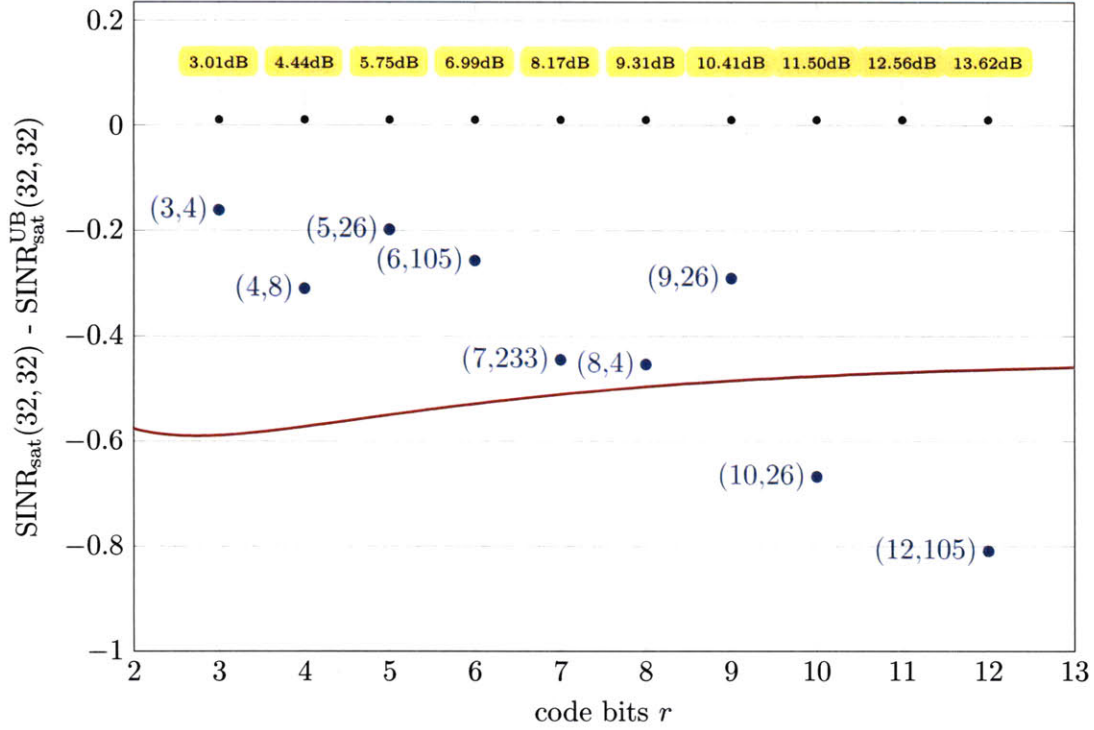


Figure 3-14. The performance of RVQ (solid curve) and our complete systematic constructions of codes. The performance is plotted relative to the upper bound (2.44), the value of which is labeled at 0. One firsts takes the union of sparse and dense codes then uses a family of 2^r linear transformation to double the rate of the code. Note that as the rate of the code increases from 3 bits to 9 bits the achieved performance is within 0.5 dB of the upper bound, while from 10 to 12 bits the performance is with in 0.81 dB of the upper bound and 0.35 dB of RVQ.

■ 3.2.4 Systematic Code Construction Summary

Our approach to channel quantization has quite a few components in the design. One must first find a good base code then solve for the parameters α and γ which maximize SINR_{sat} . A particularly useful method to construct quantizers of varying rate is to find a family of “good” component codes of varying degrees of sparsity and rate, all of which are invariant to shifts in the coordinate set. Then one may pair these codes together to form larger and larger rate codes by increasing the cardinality of the integer ring underlying the construction of each of the constituent codes. Lastly, one may increase the code rate by taking the union of codes resulting from applying a system of linear transforms as in (3.11). This allows one to systematically construct good low-rate quantizers then, using these good low-rate quantizers as building blocks, construct higher and higher rate codes that have associated low complexity quantization algorithms. Thus, our systematic construction first finds a family of good sparse and dense codes, say C_{good} , then solves the design problem

$$\max_{T \subset C_{\text{good}}} \max_{\substack{(\alpha, \gamma) \\ 0 < \alpha < 1, \gamma \in \mathbb{C}}} \text{SINR}_{\text{sat}} \left(C_{\mathbf{F}}(\alpha, \gamma, \bigcup_{\mathcal{C} \in T} \mathcal{C}) \right). \quad (3.12)$$

An example construction following this principal may be seen in Example 3.2.6.

Example 3.2.6

In this example we provide a systematic construction for a family of good quantizers in \mathbb{C}^4 with an arbitrary number of bits. We construct several good dense codes that may be paired with sparser codes to yield a systematic construction of quantizers with good performance in terms of SINR_{sat} . In this direction we denote $\mathcal{C}_{\text{Identity}}$ as the set

$$\mathcal{C}_{\text{Identity}} = \left\{ [1, 0, 0, 0]^\dagger, [0, 1, 0, 0]^\dagger, [0, 0, 1, 0]^\dagger, [0, 0, 0, 1]^\dagger \right\}$$

and let $\mathcal{C}_{\text{Sparse}}(k)$ be the union of the codes with supports indexed by the 3 subspaces of \mathbb{F}_2^2 ,

$$\{[0, 0], [0, 1]\}, \{[0, 0], [1, 0]\}, \{[0, 0], [1, 1]\},$$

with index sets, Υ_1 ,

$$\{[0, 0], [0, 1], [0, 2], \dots, [0, 2^{k_1} - 1]\},$$

$$\{[0, 0], [1, 0], [2, 0], \dots, [2^{k_1} - 1, 0]\}$$

and

$$\{[0, 0], [1, 1], [2, 2], \dots, [2^{k_1} - 1, 2^{k_1} - 1]\},$$

respectively. To specify the dense codes we use a slightly different notation as the map in to our general framework will require more formalities. However, at present we may specify our dense codes by letting the coordinate set be indexed by the integers $\{0, 1, 2, 3\}$ and codewords be indexed by integer vectors of length 4 over $\mathbb{Z}_{2^{k_2}}$. We employ 3 dense codes each of which may be described by a simple generator matrix. In this direction, let

$$\mathbf{G}_0(k) = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 2 \end{bmatrix}$$

$$\mathbf{G}_1(k) = \begin{bmatrix} 2 & 3 & 3 & 0 \\ 3 & 2 & 3 & 0 \\ 0 & 0 & 0 & 2^{k-1} \end{bmatrix}$$

$$\mathbf{G}_2(2) = \begin{bmatrix} 0 & 2 & 2 & 0 \\ 2 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix} \quad \text{and for } k > 2 \quad \mathbf{G}_2(k) = \begin{bmatrix} 4 & 6 & 6 & 0 \\ 6 & 4 & 6 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}$$

Then, we let the codes $\mathcal{C}_{\text{Dense}}(j, k_2)$ be indexed by the unique elements of the set

$$\mathcal{I}_{\text{Dense}}(j, k_2) = \{\mathbf{G}_j(k_2) \cdot \mathbf{v}\}$$

where $\mathbf{v} \in \mathbb{Z}_{2^{k_2}}^3$ and all operation are performed modulo 2^{k_2} . To form codewords we let

$$c(\mathbf{a}, j) = \exp\left(\frac{2\pi\sqrt{-1} \cdot \mathbf{a}_j}{2^{k_1}}\right).$$

With this formality we let

$$\mathcal{C}_{\text{Dense}}(k_2, j) = \{[c(\mathbf{a}, k)]_{k=0}^3 \mid \mathbf{a} \in \mathcal{I}_{\text{Dense}}(j, k_2)\}.$$

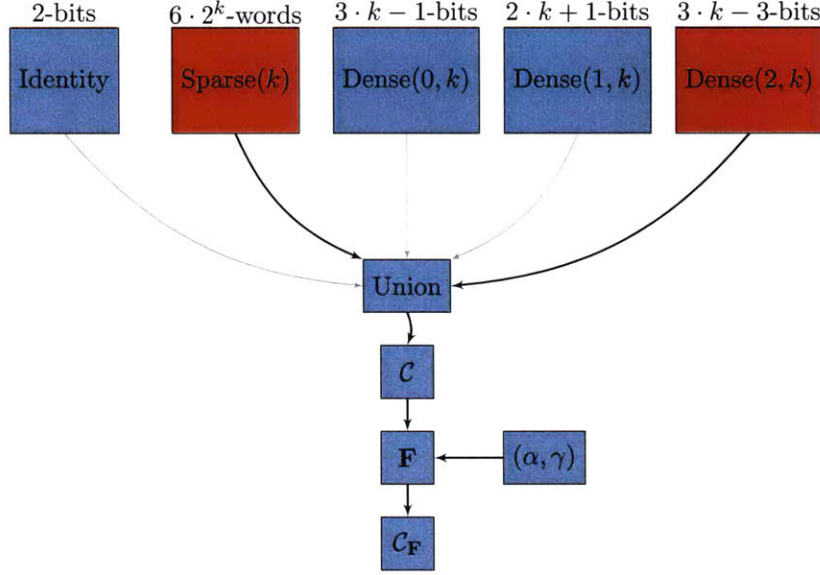


Figure 3-15. A depiction of the systematic construction of the 5-bit quantizer $\mathcal{C}_{\text{ASC}}^*(2, 2)$ and the 10-bit quantizer $\mathcal{C}_{\mathbf{F}}(\alpha, \gamma, \mathcal{C}_{\text{ASC}}^*(2, 2))$. Our systematic construction first chooses several good dense and sparse codes of varying rates which may be paired together to yield a higher rate code with low mean squared error. In this particular example both a sparse and a dense code are selected and a union of these two codes is formed to yield, for $k = 2$, a 5 bit code which is indicated by the dark arrows and shaded boxes. Then, to construct a 10 bit code one may optimize over the choice of α and γ in the construction of the universal code $\mathcal{C}_{\mathbf{F}}(\alpha, \gamma, \mathcal{C}_{\text{ASC}}^*(2, 2))$.

Example Continued

It should be clear from the definition of $\mathbf{G}_j(k_2)$ that

$$|\mathcal{C}_{\text{Dense}(0, k_2)}| = 2^{2 \cdot k + k - 1},$$

$$|\mathcal{C}_{\text{Dense}(1, k_2)}| = 2^{2 \cdot k + 1},$$

and

$$|\mathcal{C}_{\text{Dense}(2, k_2)}| = 2^{3 \cdot (k - 1)}.$$

Thus, the code

$$\mathcal{C}_{\text{ASC}}(k, j) = \mathcal{C}_{\text{Dense}(j, k)} \cup \mathcal{C}_{\text{Sparse}(k)} \cup \mathcal{C}_{\text{Identity}}$$

has $4 + 3 \cdot 2^{k-1} + 2^{2 \cdot k + k - 1}$, $4 + 3 \cdot 2^{k-1} + 2^{2 \cdot k + 1}$ or $4 + 3 \cdot 2^{k-1} + 2^{3 \cdot (k-1)}$ codewords for $j = 0, 1$ and 2 respectively. We similarly let

$$\mathcal{C}_{\text{ASC}}^*(k, j) = \mathcal{C}_{\text{Dense}(j, k_2)} \cup \mathcal{C}_{\text{Sparse}(k_1)}.$$

which yields codes with $3 \cdot 2^{k-1} + 2^{2 \cdot k + k - 1}$, $3 \cdot 2^{k-1} + 2^{2 \cdot k + 1}$ or $3 \cdot 2^{k-1} + 2^{3 \cdot (k-1)}$ codewords for $j = 0, 1$ and 2 respectively. Then, using these quantizers, one can form higher and higher rate codes by forming universal codes $\mathcal{C}_{\mathbf{F}}$, improving performance by optimizing one's choice for α and γ . A depiction of the construction may be seen in Figure 3-15.

■ 3.3 Systematic Construction of Component Codes

In Section 3.2 we have argued that the performance of a quantizer is enhanced if the quantization codebook has a large set of unitary transformation that act transitively on the codebook. Hence, in this section we consider the design of quantizers that have this property. This construction relies heavily on the theory of linear codes over rings. We present a full discussion of this in Appendix A and provide overview here that is less reliant on that theory. For the uninterested reader, or one who wishes to experiment with these codes before proceeding, the developed codes may be found at [119].

Recall from Section 3.2, that in order for a quantizer to uniformly cover the sphere⁶ the quantizer in the architecture of interest is described by the union of several permutations of a fixed base code \mathcal{C}_0 . As seen in Example 3.2.4 such constructions can yield regular structures that aid in algorithm development for user selection. To further simplify algorithms for user selection as well as quantization⁷ we consider codes that are images of linear codes over rings. As we show in the sequel such constructions yield large groups of unitary transformations that act transitively on the codebook. In order to derive the group of transitive actions we use a generalization of a method of Sidelnikov [113] which in turn can be viewed as an extension of the quantum coding frame work of Calderbank, Shor and Stean [28, 116, 117]. We present our first and most simple generalization in the following and present our most general quantization framework in Section 3.4.

■ 3.3.1 A Generalization of Sidelnikov's Codes

In the sequel, we consider the case when the number of transmit antennas is equal to some prime power⁸, say $m = p^{m'}$, and we index the standard basis with the elements of the vector space $(\mathbb{Z}_p)^{m'}$. That is, in the sequel we let

$$\mathcal{I} = (\mathbb{Z}_p)^{m'}$$

and, for any $\boldsymbol{\lambda} = [\lambda_0, \dots, \lambda_{m-1}] \in \mathcal{I}$,

$$\mathbf{e}_\lambda = \mathbf{e}_a \text{ where } a = \sum_{i=0}^{a-1} \lambda_i p^i.$$

Recall from our previous discussion we require that the support for the base quantizer, \mathcal{I}_0 , as well as \mathcal{I} to be closed under addition. Hence, \mathcal{I}_0 is a sub-space of $(\mathbb{Z}_p)^{m'}$. We denote the sub-space \mathcal{I}_0 as L to stress the fact that this set is linear, i.e. closed under addition.

Every code supported on a subset, L , of $(\mathbb{Z}_p)^{m'}$ is indexed by a subset Υ_1 of the \mathbb{Z}_{p^a} -module $(\mathbb{Z}_{p^a})^{m'}$. We note that an element of $(\mathbb{Z}_{p^a})^{m'}$ may be viewed as a “vector” of length m' over \mathbb{Z}_{p^a} . Moreover, in the current context the term module and vector space and sub-module and sub-space may be used interchangeably. We will make clear when the distinction is needed in later sections. The set of coefficients $\{c(\boldsymbol{\lambda}, \tilde{\gamma})\}_{\boldsymbol{\lambda} \in \Upsilon_1, \tilde{\gamma} \in L}$ will be a function of the inner product between $\boldsymbol{\lambda}$ and $\tilde{\gamma}$ where $\tilde{\gamma}$ is viewed as an element of $(\mathbb{Z}_{p^a})^{m'}$

⁶ We note that this construction has a greater applicability than described at present. In particular, this construction can be used to design interesting space-time codes [2] quantum stabilizer codes [113], nested diversity codes [47] as well as geometrically uniform frames [48].

⁷ Efficient quantization algorithms can be derived by direct extension of [13].

⁸ We note that the present discussion may be extended to arbitrary integers, however that development is overly cumbersome and does not yield any new insights and hence is neglected from the present development

in a natural way. In particular, for $\boldsymbol{\lambda} = [\lambda_0, \dots, \lambda_{m-1}] \in \Upsilon_1$ and $\bar{\boldsymbol{\beta}} = [\beta_0, \dots, \beta_{m-1}] \in L$, we let

$$c(\boldsymbol{\lambda}, \bar{\boldsymbol{\beta}}) = \zeta_{p^a}^{\langle \boldsymbol{\lambda}, \bar{\boldsymbol{\beta}} \rangle}$$

where $\zeta_{p^a} = \exp\left(\frac{2\pi\sqrt{-1}}{p^a}\right)$ is a p^a -th root of unity and

$$\langle \boldsymbol{\lambda}, \bar{\boldsymbol{\beta}} \rangle = \langle \boldsymbol{\lambda}, \boldsymbol{\beta} \rangle_{\mathbb{Z}_{p^a}} = \sum_{i=0}^{m'-1} \lambda_i \beta_i \quad (3.13)$$

where in turn β_i is a natural lifting of $\bar{\beta}_i$ to \mathbb{Z}_{p^a} , i.e. where $\bar{\beta}_i$ is regarded as an element of \mathbb{Z}_{p^a} . In particular, β_i is the element of \mathbb{Z}_{p^a} for which

$$\bar{\beta}_i \equiv \beta_i \pmod{p} \text{ and } \beta_i - ((\beta_i \pmod{p}) \cap \mathbb{Z}_{p^a}) = 0. \quad (3.14)$$

As we will see in the sequel *the choice of this lift dramatically alters the structure of the associated quantizer*. In particular, we will show that by altering how this lift is defined (or alternatively how we define the inner product in (3.13)) one can trade off between the coherence of a quantizer and the number of orthogonal bases contained in the quantizer. However, the current definition of lift illuminates the trade-off while not obfuscating the results with the precision we require to fully describe lifts in the sequel. Hence, at present, every quantizer in the architecture of interest is described by:

1. \mathcal{I} , the vector space $(\mathbb{Z}_p)^{m'}$
2. L , a sub-space of $(\mathbb{Z}_p)^{m'}$
3. Υ_1 , a subset, $(\mathbb{Z}_{p^a})^{m'}$ which describes the base quantizer \mathcal{C}_0
4. Υ_2 , a subset of $(\mathbb{Z}_p)^{m'}$ which describes the “shifts” of L
(i.e. the coordinate permutations to be applied to \mathcal{C}_0)
5. the function $c(\boldsymbol{\lambda}, \bar{\boldsymbol{\beta}}) = \zeta_{p^a}^{\langle \boldsymbol{\lambda}, \bar{\boldsymbol{\beta}} \rangle}$

where at present we have left the degree of freedom for the choice of $\langle \boldsymbol{\lambda}, \bar{\boldsymbol{\beta}} \rangle$ implicit. We let

$$\mathcal{C}(\Upsilon_1, \Upsilon_2; L) = \bigcup_{\bar{\boldsymbol{\beta}} \in \Upsilon_2} \bigcup_{\boldsymbol{\lambda} \in \Upsilon_1} \{c(\boldsymbol{\lambda}, \bar{\boldsymbol{\beta}}; L, p^a)\} \quad (3.15)$$

where, for $\boldsymbol{\lambda} \in (\mathbb{Z}_{p^a})^{m'}$ and $\bar{\boldsymbol{\beta}} \in (\mathbb{Z}_{p^e})$

$$c(\boldsymbol{\lambda}, \bar{\boldsymbol{\beta}}; L, p^a) = \sum_{\bar{\boldsymbol{\gamma}} \in L} \zeta^{\langle \boldsymbol{\lambda}, \bar{\boldsymbol{\gamma}} \rangle} \mathbf{e}_{\bar{\boldsymbol{\gamma}} + \bar{\boldsymbol{\beta}}}. \quad (3.16)$$

Recall that it is our ultimate goal to determine a group of transitive unitary actions on the codebook. Hence, in the sequel we will characterize the effects one’s choice of a , L , Υ_1 and Υ_2 has on the associated group of transitive unitary actions for our present choice of lift. At present the geometric interpretation of these parameters may seem a bit abstract. Closely examining these parameters one can see that these parameters do in fact relate closely to our physical description of our quantizer thus far. In particular, the parameters a , L , Υ_1 specify the precision of the quantizer in the subspace of \mathbb{C}^m described by L while the choice of Υ_2 specifies additional subspaces of \mathbb{C}^m in which the quantizer has this specified precision. More precisely, a describes that rate one allocates to quantize the phase of each coordinate of the channel vector, $\dim L$ is equal to the dimension of the subspace the base

Quantizer Parameter	Geometric Interpretation
a	rate allocated to phase of each coordinate
L	subspace describing support of base codebook
$ \Upsilon_1 $	rate allocated to each subspace
$ \Upsilon_2 $	number of subspaces

Figure 3-16. The relation of the parameters of our general construction to our geometric interpretation. The parameters a , L , Υ_1 specify the precision of the quantizer in the subspace of \mathbb{C}^m described by L . The choice of Υ_2 specifies additional subspaces of \mathbb{C}^m in which the quantizer has this specified precision.

code quantizes, $|\Upsilon_1|$ describes the rate allocated to each one of the dominate subspaces and $|\Upsilon_2|$ describes the number of subspaces of \mathbb{C}^m in which the quantizer has the this specified precision. Hence, the choice of Υ_1 and Υ_2 allow one to balance how bits are allocated on the feedback link. That is $|\Upsilon_1|$ approximately describes the coherence in the subspaces described by Υ_2 and L while $|\Upsilon_2|$ approximately describes the number of subspaces in which the quantizer measures. We summarize these points in Figure 3-16. Hence, it is of interest to determine the effects a , L , Υ_1 and Υ_2 has on the associated group of transitive unitary actions as this will allow the system designer to balance the algorithmic complexity of user selection with the precision and robustness of the associated quantizer. In order to proceed in this direction we require the following lemma.

Lemma 3.3.1. *For any $\boldsymbol{\lambda} \in (\mathbb{Z}_{p^a})^{m'}$ and $\bar{\boldsymbol{\beta}} \in (\mathbb{Z}_p)^{m'}$ the map $\langle \boldsymbol{\lambda}, \bar{\boldsymbol{\beta}} \rangle$ is linear in both of its arguments.*

Proof. See Appendix C.2.1. ■

We now proceed and address how L , Υ_1 and Υ_2 may be chosen so that there exists a large group of unitary transformations that act transitively on $\mathcal{C}_L(\Upsilon_1, \Upsilon_2)$. In this direction, we let $T(\boldsymbol{\lambda})$ be the matrix that acts diagonally on the basis $\{\mathbf{e}_{\bar{\alpha}}\}$ by

$$T(\boldsymbol{\lambda})\mathbf{e}_{\bar{\alpha}} = \zeta^{\langle \boldsymbol{\lambda}, \bar{\alpha} \rangle} \mathbf{e}_{\bar{\alpha}}. \quad (3.17)$$

It is clear that $T(\boldsymbol{\lambda})$ is unitary as

$$\begin{aligned} T(\boldsymbol{\lambda})^\dagger T(\boldsymbol{\lambda})\mathbf{e}_{\bar{\alpha}} &= T(\boldsymbol{\lambda})^\dagger \left(\zeta^{\langle \boldsymbol{\lambda}, \bar{\alpha} \rangle} \mathbf{e}_{\bar{\alpha}} \right) \\ &= \zeta^{-\langle \boldsymbol{\lambda}, \bar{\alpha} \rangle} \zeta^{\langle \boldsymbol{\lambda}, \bar{\alpha} \rangle} \mathbf{e}_{\bar{\alpha}} \\ &= \mathbf{e}_{\bar{\alpha}}. \end{aligned}$$

Moreover,

$$T(\boldsymbol{\lambda}')\mathbf{c}(\boldsymbol{\lambda}, \bar{\boldsymbol{\beta}}; L, p^a) = \sum_{\bar{\gamma} \in L} \zeta^{\langle \boldsymbol{\lambda} + \boldsymbol{\lambda}', \bar{\gamma} \rangle} \mathbf{e}_{\bar{\gamma} + \bar{\boldsymbol{\beta}}} = \mathbf{c}(\boldsymbol{\lambda} + \boldsymbol{\lambda}', \bar{\boldsymbol{\beta}}; L, p^a) \quad (3.18)$$

as the map $\langle \boldsymbol{\lambda}, \bar{\gamma} \rangle$ is linear in both its arguments by Lemma 3.3.1. Thus, if Υ_1 is closed under addition each matrix $T(\boldsymbol{\lambda})$ for $\boldsymbol{\lambda} \in \Upsilon_1$ acts transitively on $\mathcal{C}(\Upsilon_1, \Upsilon_2; L)$ for any Υ_2 and L . We state this in the following lemma.

Lemma 3.3.2. *Let Υ_1 be a sub-module of $(\mathbb{Z}_{p^a})^{m'}$ and let Υ_2 and L be non-empty subsets of $(\mathbb{Z}_p)^{m'}$. Then, $T(\boldsymbol{\lambda})$ acts transitively on $\mathcal{C}(\Upsilon_1, \Upsilon_2; L)$ for all $\boldsymbol{\lambda} \in \Upsilon_1$.*

Proof. This trivially follows from the fact that Υ_1 is a subspace of $(\mathbb{Z}_p)^{m'}$. Hence,

$$\mathbf{T}(\boldsymbol{\lambda}')\mathbf{c}(\boldsymbol{\lambda}, \bar{\boldsymbol{\beta}}; L, p^a) = \mathbf{c}(\boldsymbol{\lambda} + \boldsymbol{\lambda}', \bar{\boldsymbol{\beta}}; L, p^a) \in \mathcal{C}(\Upsilon_1, \{\bar{\boldsymbol{\beta}}\}; L)$$

as $\boldsymbol{\lambda} + \boldsymbol{\lambda}' \in \Upsilon_1$. ■

We seek a results similar to Lemma 3.3.2 for the set Υ_2 . In this directions, let $\mathbf{S}(\bar{\boldsymbol{\beta}})$ be the matrix that permutes the basis $\{\mathbf{e}_{\bar{\boldsymbol{\alpha}}}\}$ by translations. More precisely,

$$\mathbf{S}(\bar{\boldsymbol{\beta}})\mathbf{e}_{\bar{\boldsymbol{\alpha}}} = \mathbf{e}_{\bar{\boldsymbol{\alpha}} + \bar{\boldsymbol{\beta}}}. \quad (3.19)$$

It is clear that $\mathbf{S}(\bar{\boldsymbol{\beta}})$ is unitary as $\mathbf{S}(\bar{\boldsymbol{\beta}})$ is a permutation matrix. Additionally,

$$\mathbf{S}(\bar{\boldsymbol{\beta}}')\mathbf{c}(\boldsymbol{\lambda}, \bar{\boldsymbol{\beta}}; L, p^a) = \sum_{\bar{\boldsymbol{\gamma}} \in L} \zeta^{\langle \boldsymbol{\lambda}, \bar{\boldsymbol{\gamma}} \rangle} \mathbf{e}_{\bar{\boldsymbol{\gamma}} + \bar{\boldsymbol{\beta}} + \bar{\boldsymbol{\beta}}'} = \mathbf{c}(\boldsymbol{\lambda}, \bar{\boldsymbol{\beta}} + \bar{\boldsymbol{\beta}}'; L, p^a) \quad (3.20)$$

and hence if Υ_2 is closed under addition each matrix $\mathbf{S}(\bar{\boldsymbol{\beta}})$ for $\bar{\boldsymbol{\beta}} \in \Upsilon_2$ acts transitively on $\mathcal{C}(\Upsilon_1, \Upsilon_2; L)$ for any Υ_1 and L . We state this in the following lemma.

Lemma 3.3.3. *Let Υ_2 be a sub-space of $(\mathbb{Z}_p)^{m'}$, let Υ_1 be a non-empty subset of $(\mathbb{Z}_{p^a})^{m'}$ and let L be a non-empty subset of $(\mathbb{Z}_p)^{m'}$. Then, $\mathbf{S}(\bar{\boldsymbol{\beta}})$ acts transitively on $\mathcal{C}(\Upsilon_1, \Upsilon_2; L)$ for all $\bar{\boldsymbol{\beta}} \in \Upsilon_2$.*

Proof. This trivially follows from the fact that Υ_2 is a subspace of $(\mathbb{Z}_p)^{m'}$. Hence,

$$\mathbf{S}(\bar{\boldsymbol{\beta}}')\mathbf{c}(\boldsymbol{\lambda}, \bar{\boldsymbol{\beta}}; L, p^a) = \mathbf{c}(\boldsymbol{\lambda}, \bar{\boldsymbol{\beta}} + \bar{\boldsymbol{\beta}}'; L, p^a) \in \mathcal{C}(\{\boldsymbol{\lambda}\}, \Upsilon_2; L)$$

as $\bar{\boldsymbol{\beta}} + \bar{\boldsymbol{\beta}}' \in \Upsilon_2$. ■

We note that the matrices $\mathbf{T}(\boldsymbol{\lambda})$ and $\mathbf{S}(\bar{\boldsymbol{\beta}})$ have very simple interpretations in terms of their actions on subcodes. In particular, by Lemma 3.3.2 the matrix $\mathbf{T}(\boldsymbol{\lambda})$ acts transitively on the subcode $\mathcal{C}(\Upsilon_1, \{\bar{\boldsymbol{\beta}}\}; L)$ for any fixed $\bar{\boldsymbol{\beta}}$ while by Lemma 3.3.3 $\mathbf{S}(\bar{\boldsymbol{\beta}})$ acts transitively on the subcode $\mathcal{C}(\{\boldsymbol{\lambda}\}, \Upsilon_2; L)$ for any fixed $\boldsymbol{\lambda}$. This can be seen in Figure 3-17. Note, if Υ_1 and Υ_2 are *both* closed under addition then $\mathbf{S}(\bar{\boldsymbol{\beta}})\mathbf{T}(\boldsymbol{\lambda})$ and $\mathbf{T}(\boldsymbol{\lambda})\mathbf{S}(\bar{\boldsymbol{\beta}})$ act transitively on $\mathcal{C}(\Upsilon_1, \Upsilon_2; L)$. Hence, one may guess that any choice for Υ_1 and Υ_2 such that both Υ_1 and Υ_2 are linear will produce a quantizer with a large set of transitive unitary transformations. However, note by combining (3.17) – (3.20) one can see that for any $\bar{\boldsymbol{\alpha}} \in (\mathbb{Z}_p)^{m'}$

$$\mathbf{T}(\boldsymbol{\lambda})\mathbf{S}(\bar{\boldsymbol{\beta}})\mathbf{e}_{\bar{\boldsymbol{\alpha}}} = \zeta^{\langle \boldsymbol{\lambda}, \bar{\boldsymbol{\alpha}} + \bar{\boldsymbol{\beta}} \rangle} \mathbf{e}_{\bar{\boldsymbol{\alpha}} + \bar{\boldsymbol{\beta}}} \quad (3.21)$$

while

$$\mathbf{S}(\bar{\boldsymbol{\beta}})\mathbf{T}(\boldsymbol{\lambda})\mathbf{e}_{\bar{\boldsymbol{\alpha}}} = \zeta^{\langle \boldsymbol{\lambda}, \bar{\boldsymbol{\alpha}} \rangle} \mathbf{e}_{\bar{\boldsymbol{\alpha}} + \bar{\boldsymbol{\beta}}}. \quad (3.22)$$

Hence,

$$\mathbf{T}(\boldsymbol{\lambda})\mathbf{S}(\bar{\boldsymbol{\beta}}) = \zeta^{\langle \boldsymbol{\lambda}, \bar{\boldsymbol{\beta}} \rangle} \mathbf{S}(\bar{\boldsymbol{\beta}})\mathbf{T}(\boldsymbol{\lambda}) \quad (3.23)$$

and the actions of $\mathbf{S}(\bar{\boldsymbol{\beta}})\mathbf{T}(\boldsymbol{\lambda})$ and $\mathbf{T}(\boldsymbol{\lambda})\mathbf{S}(\bar{\boldsymbol{\beta}})$ on $\mathcal{C}(\Upsilon_1, \Upsilon_2; L)$ only differ by the phase $\zeta^{\langle \boldsymbol{\lambda}, \bar{\boldsymbol{\beta}} \rangle}$. As in our problem we are only interested in the magnitude of the correlation we are interested in the lines defined by a quantizer, not the points. Thus, as considering $\mathbf{T}(\boldsymbol{\lambda})$ and $\mathbf{S}(\bar{\boldsymbol{\beta}})$ that do not commute will only produce results differing in phase we would like to only consider the matrices for which $\mathbf{S}(\bar{\boldsymbol{\beta}})\mathbf{T}(\boldsymbol{\lambda}) = \mathbf{T}(\boldsymbol{\lambda})\mathbf{S}(\bar{\boldsymbol{\beta}})$. More precisely, we like to identify the sets of $\mathbf{S}(\bar{\boldsymbol{\beta}})$ and $\mathbf{T}(\boldsymbol{\lambda})$ that form a commutative group.

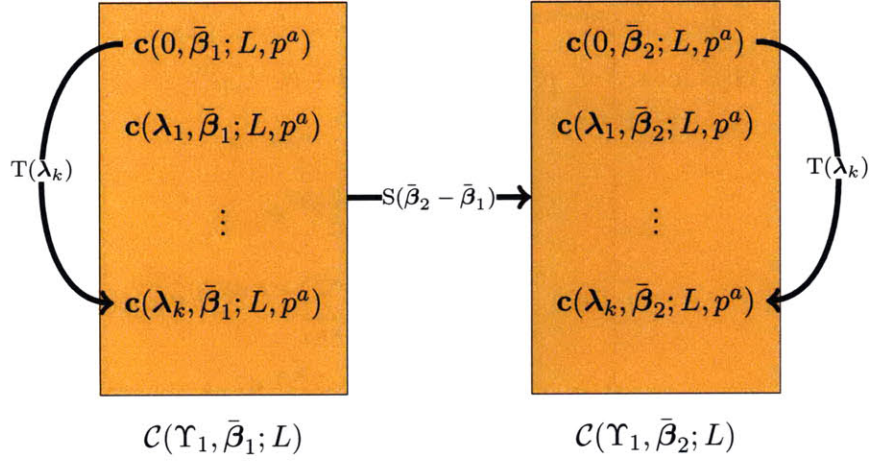


Figure 3-17. A depiction of the actions of $T(\boldsymbol{\lambda})$ and $S(\bar{\boldsymbol{\beta}})$ on the codebook $\mathcal{C}(\Upsilon_1, \Upsilon_2; L)$ where Υ_i is closed under addition. For any $\boldsymbol{\lambda} \in \Upsilon_1$, the matrix $T(\boldsymbol{\lambda}_k)$ permutes the elements of any subcode $\mathcal{C}(\Upsilon_1, \{\bar{\boldsymbol{\beta}}_i\}; L)$. In particular, for $\boldsymbol{\lambda}_k \in \Upsilon_1$, the matrix $T(\boldsymbol{\lambda}_k)$ maps $\mathbf{c}(0, \bar{\boldsymbol{\beta}}_i; L, p^a)$ to $\mathbf{c}(\boldsymbol{\lambda}_k, \bar{\boldsymbol{\beta}}_i; L, p^a)$. Additionally, for $\bar{\boldsymbol{\beta}} \in \Upsilon_2$, $S(\bar{\boldsymbol{\beta}})$ permutes the subcodes $\mathcal{C}(\Upsilon_1, \{\bar{\boldsymbol{\beta}}_i\}; L)$. In particular, for any two elements $\beta_1 \neq \beta_2$ of Υ_2 , the matrix $S(\bar{\boldsymbol{\beta}}_2 - \bar{\boldsymbol{\beta}}_1)$ maps the subcode $\mathcal{C}(\Upsilon_1, \{\bar{\boldsymbol{\beta}}_1\}; L)$ to $\mathcal{C}(\Upsilon_1, \{\bar{\boldsymbol{\beta}}_2\}; L)$.

It is clear from (3.23) that the matrices $S(\bar{\boldsymbol{\beta}})$ and $T(\boldsymbol{\lambda})$ commute if and only if $\langle \boldsymbol{\lambda}, \bar{\boldsymbol{\beta}} \rangle = 0$. Hence, if $S(\bar{\boldsymbol{\beta}})$ and $T(\boldsymbol{\lambda})$ commute for all $\boldsymbol{\lambda} \in \Upsilon_1$ and $\bar{\boldsymbol{\beta}} \in \Upsilon_2$, the set Υ_1 and Υ_2 must lay in “orthogonal” spaces. More precisely, for any subspace L of $(\mathbb{Z}_p)^{m'}$ let L_a^\perp be the set of elements of $(\mathbb{Z}_{p^a})^{m'}$ orthogonal to the lifted elements of L . That is,

$$L_a^\perp = \left\{ \bar{\boldsymbol{\alpha}} \in (\mathbb{Z}_{p^a})^{m'} \mid \langle \bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\gamma}} \rangle = 0 \quad \forall \bar{\boldsymbol{\gamma}} \in L \right\}$$

Then the set of matrices

$$\mathcal{H}_{L,a} = \left\{ T(\boldsymbol{\lambda})S(\bar{\boldsymbol{\beta}}) \mid \boldsymbol{\lambda} \in L_a^\perp, \bar{\boldsymbol{\beta}} \in L \right\}$$

is commutative.

Recall from Example 3.2.4 transitive unitary actions on the codebook that have fixed points were shown to be a valuable tool for searches for orthogonal bases. In fact, such a sequence of transformations allowed us to enumerate all orthogonal bases in the code by swapping in and out pairs of vertices in the graph. We would like to develop this approach more generally. That is we would like to identify unitary transformations that act transitively on the codebook for which a portion of the code words are in the eigenspace of the transformation. This allows one to embed many orthogonal bases into a single quantizer. In this direction, we have the following regarding the eigenspace of $\mathcal{H}_{L,a}$.

Lemma 3.3.4. *Let $(\boldsymbol{\lambda}', \bar{\boldsymbol{\beta}}') \in L_a^\perp \times L$ be given. Then, $T(\boldsymbol{\lambda}')S(\bar{\boldsymbol{\beta}}') \in \mathcal{H}_{L,a}$ and $\mathbf{c}(\boldsymbol{\lambda}, \bar{\boldsymbol{\beta}}; L, p^a)$ is an eigenvector of $T(\boldsymbol{\lambda}')S(\bar{\boldsymbol{\beta}}')$ with eigenvalue $\zeta^{-\langle \boldsymbol{\lambda}, \bar{\boldsymbol{\beta}}' \rangle}$ for all $\boldsymbol{\lambda} \in (\mathbb{Z}_{p^a})^{m'}$ and $\bar{\boldsymbol{\beta}} \in (\mathbb{Z}_p)^{m'}$.*

Proof. For any $(\boldsymbol{\lambda}', \bar{\boldsymbol{\beta}}') \in L_a^\perp \times L$ one has that

$$\mathbf{T}(\boldsymbol{\lambda}')\mathbf{S}(\bar{\boldsymbol{\beta}}')\mathbf{c}(\boldsymbol{\lambda}, \bar{\boldsymbol{\beta}}; L, p^a) = \mathbf{c}(\boldsymbol{\lambda} + \boldsymbol{\lambda}', \bar{\boldsymbol{\beta}} + \bar{\boldsymbol{\beta}}'; L, p^a) \quad (3.24a)$$

$$= \sum_{\bar{\gamma} \in L} \zeta^{\langle \boldsymbol{\lambda} + \boldsymbol{\lambda}', \bar{\gamma} \rangle} \mathbf{e}_{\bar{\gamma} + \bar{\boldsymbol{\beta}} + \bar{\boldsymbol{\beta}}'} \quad (3.24b)$$

$$= \sum_{\bar{\gamma} \in L} \zeta^{\langle \boldsymbol{\lambda} + \boldsymbol{\lambda}', \bar{\gamma} - \bar{\boldsymbol{\beta}}' \rangle} \mathbf{e}_{\bar{\gamma} + \bar{\boldsymbol{\beta}}} \quad (3.24c)$$

$$= \zeta^{-\langle \boldsymbol{\lambda} + \boldsymbol{\lambda}', \bar{\boldsymbol{\beta}}' \rangle} \sum_{\bar{\gamma} \in L} \zeta^{\langle \boldsymbol{\lambda} + \boldsymbol{\lambda}', \bar{\gamma} \rangle} \mathbf{e}_{\bar{\gamma} + \bar{\boldsymbol{\beta}}} \quad (3.24d)$$

$$= \zeta^{-\langle \boldsymbol{\lambda}, \bar{\boldsymbol{\beta}}' \rangle} \sum_{\bar{\gamma} \in L} \zeta^{\langle \boldsymbol{\lambda}, \bar{\gamma} \rangle} \mathbf{e}_{\bar{\gamma} + \bar{\boldsymbol{\beta}}} \quad (3.24e)$$

$$= \zeta^{-\langle \boldsymbol{\lambda}, \bar{\boldsymbol{\beta}}' \rangle} \mathbf{c}(\boldsymbol{\lambda}, \bar{\boldsymbol{\beta}}; L, p^a) \quad (3.24f)$$

where (3.24a) follows from (3.18) and (3.20), (3.24b) follows from the definition of the codeword $\mathbf{c}(\boldsymbol{\lambda} + \boldsymbol{\lambda}', \bar{\boldsymbol{\beta}} + \bar{\boldsymbol{\beta}}'; L, p^a)$ in (3.16), (3.24c) uses the fact that $\bar{\boldsymbol{\beta}}' \in L$ and L is a linear space, (3.24d) uses the fact that the map $\langle \boldsymbol{\lambda}, \bar{\boldsymbol{\beta}} \rangle$ is linear in both of its arguments, (3.24e) use the condition that $(\boldsymbol{\lambda}', \bar{\boldsymbol{\beta}}') \in L_a^\perp \times L$ and (3.24f) follows from the definition of $\mathbf{c}(\boldsymbol{\lambda}, \bar{\boldsymbol{\beta}}; L, p^a)$, (3.16). \blacksquare

Examining Lemma 3.3.4 one can see that every codevector $\mathbf{c}(\boldsymbol{\lambda}, \bar{\boldsymbol{\beta}}; L, p^a)$ for $\boldsymbol{\lambda} \in (\mathbb{Z}_{p^a})^{m'}$ and $\bar{\boldsymbol{\beta}} \in (\mathbb{Z}_p)^{m'}$ is an eigenvector of $\mathcal{H}_{L,a}$. Hence, as this describes $p^{(a+1) \cdot m'}$ codewords a large subset of the vectors must be linearly dependent⁹ and hence correlated. In particular, examining (3.24a) and (3.24f) in the proof of Lemma 3.3.4 it is clear that the codewords $\mathbf{c}(\boldsymbol{\lambda}, \bar{\boldsymbol{\beta}}; L, p^a)$ and $\mathbf{c}(\boldsymbol{\lambda} + \boldsymbol{\lambda}', \bar{\boldsymbol{\beta}} + \bar{\boldsymbol{\beta}}'; L, p^a)$ are colinear if $(\boldsymbol{\lambda}', \bar{\boldsymbol{\beta}}') \in L_a^\perp \times L$. However, if $(\boldsymbol{\lambda}', \bar{\boldsymbol{\beta}}') \notin L_a^\perp \times L$ then it is not clear when the codewords are colinear, correlated or orthogonal.

Lemma 3.3.5. *The codewords are $\mathbf{c}(\boldsymbol{\lambda}, \bar{\boldsymbol{\beta}}; L, p^a)$ and $\mathbf{c}(\boldsymbol{\lambda}', \bar{\boldsymbol{\beta}}'; L, p^a)$ are colinear if and only if $\bar{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}' \in L$ and $\boldsymbol{\lambda} - \boldsymbol{\lambda}' \in L_a^\perp$*

Proof. See Appendix C.2.2 \blacksquare

Examining Lemma 3.3.5 we can see that so long as Υ_1 is chosen such that $\boldsymbol{\lambda}' - \boldsymbol{\lambda} \notin L_a^\perp$ and $\bar{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}' \notin L$ we can guarantee that the constructed quantizer does not contain colinear points. As L , Υ_1 and Υ_2 are all linear this requires us to choose Υ_1 and Υ_2 from a set complimentary to L_a^\perp and L respectively. In this direction let L^c be any sub-space of $(\mathbb{Z}_p)^{m'}$ complimentary to L and let L_a^d be any sub-module of $(\mathbb{Z}_{p^a})^{m'}$ that is complimentary to L_a^\perp . More precisely, L^c is any sub-space of $(\mathbb{Z}_p)^{m'}$ such that

$$(\mathbb{Z}_p)^{m'} = L \oplus L^c$$

and L_a^d is any sub-module of $(\mathbb{Z}_{p^a})^{m'}$ such that

$$(\mathbb{Z}_{p^a})^e = L_a^\perp \oplus L_a^d.$$

Then we have the following theorem.

⁹How does this related to nested diversity space time codes

Theorem 3.3.6. *Let L be a fixed subspace of $(\mathbb{Z}_p)^m$ and suppose Υ_1 is a sub-module of L_a^d and Υ_2 is a subspace of L^c . Then, $\mathcal{C}(\Upsilon_1, \Upsilon_2; L)$ is invariant to multiplication by any element of $\mathcal{H}_{L,a}$. Moreover, any matrix $H' \in \mathcal{H}_{L^c,a}$ such that $H' = T(\lambda')S(\bar{\beta}')$ where $(\lambda', \bar{\beta}') \in \Upsilon_1 \times \Upsilon_2$, acts transitively on the code $\mathcal{C}(\Upsilon_1, \Upsilon_2; L)$ and leaves no codeword fixed if $\bar{\beta}' \neq 0$ or $\lambda' \neq 0$. More precisely, for all $\mathbf{c} \in \mathcal{C}(\Upsilon_1, \Upsilon_2; L)$, if $H' = T(\lambda')S(\bar{\beta}') \in \mathcal{H}_{L^c,a}$ and $(\lambda', \bar{\beta}') \in \Upsilon_1 \times \Upsilon_2$ then*

$$H' \cdot \mathbf{c} \in \mathcal{C}(\Upsilon_1, \Upsilon_2; L) \text{ and if } \bar{\beta}' \neq 0 \text{ or } \lambda' \neq 0 \text{ then } H' \cdot \mathbf{c} \neq \mathbf{c}$$

and for any $H \in \mathcal{H}_{L,a}$,

$$H \cdot \mathbf{c} = \mathbf{c}.$$

Proof. This is a direct consequence of the preceding discussion. That is by Lemma 3.3.4 one can see that for any $H \in \mathcal{H}_{L,a}$, $H \cdot \mathbf{c} = \mathbf{c}$. Moreover, as $(\lambda', \bar{\beta}') \in \Upsilon_1 \times \Upsilon_2$ and

$$T(\lambda')S(\bar{\beta}')\mathbf{c}(\lambda, \bar{\beta}; L, p^a) = \mathbf{c}(\lambda + \lambda', \bar{\beta} + \bar{\beta}'; L, p^a)$$

we have that $\mathbf{c}(\lambda + \lambda', \bar{\beta} + \bar{\beta}'; L, p^a) \in \mathcal{C}(\Upsilon_1, \Upsilon_2; L)$ as Υ_1 and Υ_2 are both linear. Thus, $H' = T(\lambda')S(\bar{\beta}')$ acts transitively on $\mathcal{C}(\Upsilon_1, \Upsilon_2; L)$ for $(\lambda', \bar{\beta}') \in \Upsilon_1 \times \Upsilon_2$. To see that no codeword of $\mathcal{C}(\Upsilon_1, \Upsilon_2; L)$ is fixed if $\bar{\beta}' \neq 0$ or $\lambda' \neq 0$ note that if $\bar{\beta}' \neq 0$ then $\bar{\beta} - (\bar{\beta} + \bar{\beta}') = -\bar{\beta}' \in L^c$ and hence by Lemma 3.3.5, we have that $\mathbf{c}(\lambda, \bar{\beta}; L, p^a)$ and $\mathbf{c}(\lambda + \lambda', \bar{\beta} + \bar{\beta}'; L, p^a)$ are not colinear. Similarly, if $\lambda' \neq 0$ then $\lambda - (\lambda + \lambda') = -\lambda' \in L_a^d$ and by Lemma 3.3.5, we have that $\mathbf{c}(\lambda, \bar{\beta}; L, p^a)$ and $\mathbf{c}(\lambda + \lambda', \bar{\beta} + \bar{\beta}'; L, p^a)$ are not colinear. ■

We note that Theorem 3.3.10 only considered the case when the subspace L was fixed. However, it should be clear that one may want to create quantizers that are indexed over multiple subspaces or for that matter other maps that are linear in both arguments. Hence, in the sequel we consider how one may choose additional subspaces and maps in a “good” way, i.e. in a way as to yield many orthogonal subsets which cover the sphere well. In particular, recall that we previously noted that unitary transformations that fix part of the codebook provided a structure that aided in the design of user selection algorithms. However, to present we have only exhibited unitary transformations that either fix the entire codebook or leaves no codevector fixed (if the transformation is of course not the identity). In particular, as a consequence of Theorem 3.3.6 we saw that the matrix group $\mathcal{H}_{L,a}$ acted invariantly on any code while $\mathcal{H}_{L^c,a}$ acted strictly as translation. However, if we exchange L with L^c we obtain a code for which $\mathcal{H}_{L,a}$ acts transitively while $\mathcal{H}_{L^c,a}$ acts invariantly on the code. Thus, any quantizer that is the union of eigenvectors of $\mathcal{H}_{L,a}$ and $\mathcal{H}_{L^c,a}$ will yield a codebook for which a subsets of $\mathcal{H}_{L,a}$ and $\mathcal{H}_{L^c,a}$ will act invariantly on a faction of the code while strictly transitive on the remaining fraction. We will say that such codes are complimentary. That is, the codes $\mathcal{C}(\Upsilon_1, \Upsilon_2; L)$ and $\mathcal{C}(\tilde{\Upsilon}_1, \tilde{\Upsilon}_2; L^c)$. We make the preceding discussion more precise in the following Theorem. Then, to be concrete, we present a simple example of complimentary codes in Example 3.3.1 and a simple diagram illustrating the effects of the actions of elements of $\mathcal{H}_{L,a}$ and $\mathcal{H}_{L^c,a}$ on cosets in Figure 3-18.

Theorem 3.3.7. *Let L be a fixed subspace of $(\mathbb{Z}_p)^m$ and suppose Υ_1 is a sub-module of L_a^d and Υ_2 is a subspace of L^c . Further, suppose that $\tilde{\Upsilon}_1$ is a sub-module of L_a^d and $\tilde{\Upsilon}_2$ is a subspace of L . Then, every $H' = T(\lambda')S(\bar{\beta}')$ for $(\lambda', \bar{\beta}') \in \Upsilon_1 \times \Upsilon_2$ acts transitively on the code $\mathcal{C}(\Upsilon_1, \Upsilon_2; L)$ and invariantly on the code $\mathcal{C}(\tilde{\Upsilon}_1, \tilde{\Upsilon}_2; L^c)$. Moreover, every $H = T(\lambda)S(\bar{\beta})$*

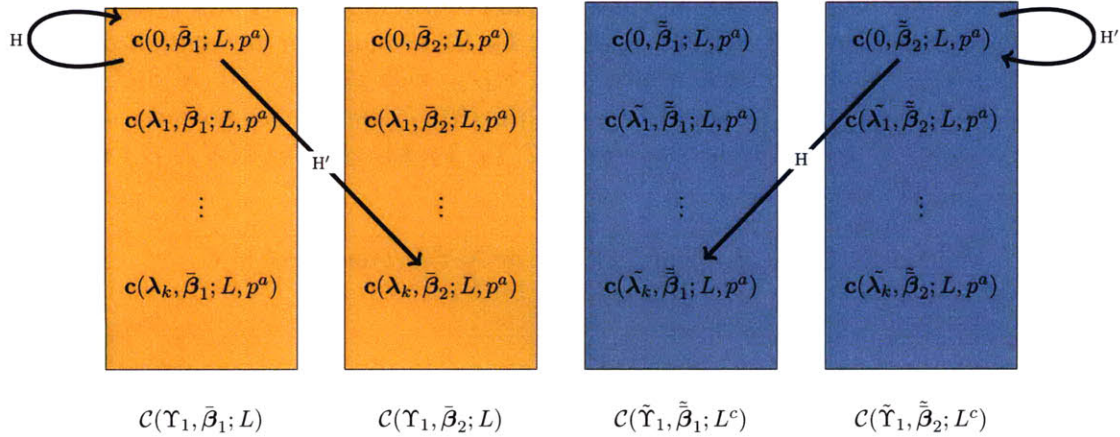


Figure 3-18. A depiction of the actions of $\mathcal{H}_{L,a}$ and $\mathcal{H}_{L^c,a}$ on two complimentary codes $\mathcal{C}(\Upsilon_1, \Upsilon_2; L)$ and $\mathcal{C}(\tilde{\Upsilon}_1, \tilde{\Upsilon}_2; L^c)$. Any $H \in \mathcal{H}_{L,a}$ acts invariantly on $\mathcal{C}(\Upsilon_1, \Upsilon_2; L)$ and hence maps $\mathbf{c}(0, \bar{\beta}_1; L, p^a)$ to itself. However, if $H = T(\tilde{\lambda})S(\tilde{\beta})$ where $(\lambda, \bar{\beta}) \in \tilde{\Upsilon}_1 \times \tilde{\Upsilon}_2$ then H acts transitively on $\mathcal{C}(\tilde{\Upsilon}_1, \tilde{\Upsilon}_2; L^c)$. Further, and $H' \in \mathcal{H}_{L^c,a}$ acts invariantly on $\mathcal{C}(\tilde{\Upsilon}_1, \tilde{\Upsilon}_2; L^c)$ while operating as translation on $\mathcal{C}(\Upsilon_1, \Upsilon_2; L)$ if $H' = T(\lambda')S(\bar{\beta}')$ where $(\lambda', \bar{\beta}') \in \Upsilon_1 \times \Upsilon_2$.

for $(\lambda, \bar{\beta}) \in \tilde{\Upsilon}_1 \times \tilde{\Upsilon}_2$ acts transitively on the code $\mathcal{C}(\tilde{\Upsilon}_1, \tilde{\Upsilon}_2; L^c)$ and invariantly on the code $\mathcal{C}(\Upsilon_1, \Upsilon_2; L)$. Moreover, the magnitude of the inner product between any two elements of $\mathcal{C}(\Upsilon_1, \Upsilon_2; L)$ and $\mathcal{C}(\tilde{\Upsilon}_1, \tilde{\Upsilon}_2; L^c)$ is $1/\sqrt{m}$.

Proof. We note that everything but the last statement follows from the discussion preceding the statement of the theorem. To see the last statement regarding the inner product between any two codewords from $\mathcal{C}(\Upsilon_1, \Upsilon_2; L)$ and $\mathcal{C}(\tilde{\Upsilon}_1, \tilde{\Upsilon}_2; L^c)$ note that by definition $L^c \cap L = \{0\}$. Hence, $\{\beta_c + L\} \cap \{\beta + L\} = \{\beta_c + \beta\}$ for every $\beta_c \in L^c$ and $\beta \in L$. Hence the supports of any two codeword from $\mathcal{C}(\Upsilon_1, \Upsilon_2; L)$ and $\mathcal{C}(\tilde{\Upsilon}_1, \tilde{\Upsilon}_2; L^c)$ intersect in exactly one location. As the component of the codeword from $\mathcal{C}(\Upsilon_1, \Upsilon_2; L)$ has a modulus $1/\sqrt{|L|}$ at this location and the component of the codeword from $\mathcal{C}(\tilde{\Upsilon}_1, \tilde{\Upsilon}_2; L^c)$ has a modulus of $1/\sqrt{|L^c|} = \sqrt{|L|/m}$ at this location the inner product of any two codewords from $\mathcal{C}(\Upsilon_1, \Upsilon_2; L)$ and $\mathcal{C}(\tilde{\Upsilon}_1, \tilde{\Upsilon}_2; L^c)$ is $\sqrt{1/m}$. ■

Example 3.3.1 Two Complimentary Codes

Recall from Example 3.2.4 we constructed a code that was the union of two orthogonal bases by an appropriate choice for $\mathcal{I}_0, \Upsilon_1$ and Υ_2 . In the sequel we consider a yet larger codebook by take the union of two codes with different choices of $\mathcal{I}_0, \Upsilon_1$ and Υ_2 . In particular, here we derive a 6 bit quantizer by letting we letting Υ_1 be a one dimensional subspace of \mathbb{Z}_{16}^2 and hence use 4 bits to index an element of Υ_1 . We then use the remaining 2 bits to index which code is being used and which element of Υ_2 . In particular,

For \mathcal{C}_1 let:

1. $\mathcal{I}_{0,0} = \{[0, 0], [0, 1]\}$,
2. $\Upsilon_{1,0} = \{[0, i] \mid 0 \leq i < 16\}$,
3. $\Upsilon_{2,0} = \{[0, 0], [1, 0]\}$

For \mathcal{C}_2 let:

1. $\mathcal{I}_{0,1} = \{[0, 0], [1, 0]\}$,
2. $\Upsilon_{1,1} = \{[i, 0] \mid 0 \leq i < 16\}$
3. $\Upsilon_{2,1} = \{[0, 0], [0, 1]\}$

We note that with this choice of parameters there is a regular structure to the magnitude of the cross correlation between codewords as. This can be seen in Figure 3-19.

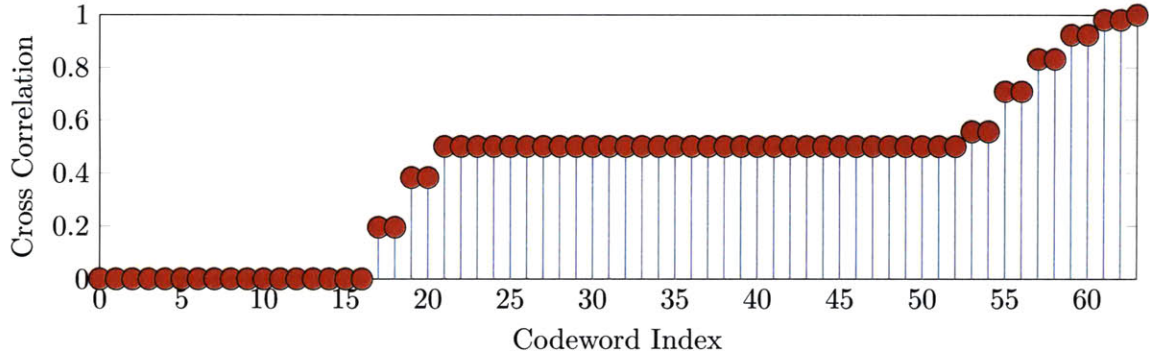


Figure 3-19. FIX ME..A depiction of the performance of two 6 bit quantizers in \mathbb{C}^4 . (a) The complimentary code from Example 3.3.1 and (b) a 6 bit code used in the 802.16 standard (c) The Grassmannian Packing from [13]. We note that while complimentary code has quite regular the code from the 802.16 standard does not.

We note that the code \mathcal{C}_2 paired with \mathcal{C}_1 in Example 3.3.1 is not unique. In fact, one may have alternatively chosen $\Upsilon_2 = \{[0, 0], [1, 1]\}$. This should be reminiscent of our development of our quantization framework in the preceding section. Indeed, we saw that one may form a sparse code of increasing rate by choosing, in the present context, $\Upsilon_2 = \{[0, 0], [0, 1]\}$, $\Upsilon_2 = \{[0, 0], [1, 1]\}$ and $\Upsilon_2 = \{[0, 0], [1, 0]\}$. Thus, from Theorem 3.3.7 one can see that there is a far greater motivation for choosing this system of codes. By choosing codes of this form there is a large group of accompanying unitary transformations which, from our preceding discussion, should make the resulting code appear more isotropic reducing the mean squared quantization error.

We now identify the orthogonal subset of the code $\mathcal{C}(\Upsilon_1, \Upsilon_2; L)$ and their structure so that we may further develop how one may develop isotropic codes with many orthogonal bases. In this direction we call any a set of vectors, say \mathcal{C} , self orthogonal if

$$\mathbf{c}_i^\dagger \mathbf{c}_j = 0 \quad \forall \mathbf{c}_i, \mathbf{c}_j \in \mathcal{C} \text{ and } \mathbf{c}_i \neq \mathbf{c}_j$$

and say that two sets of vectors, \mathcal{C}_1 and \mathcal{C}_2 , are mutually orthogonal if

$$\mathbf{c}_{1,i}^\dagger \mathbf{c}_{2,j} = 0 \text{ for all } \mathbf{c}_{1,i} \in \mathcal{C}_1 \text{ and } \mathbf{c}_{2,j} \in \mathcal{C}_2$$

Lemma 3.3.8. *If $\bar{\beta} - \bar{\beta}' \notin L$ then the codes $\mathcal{C}(\Upsilon_1, \{\bar{\beta}\}; L)$ and $\mathcal{C}(\Upsilon_1, \{\bar{\beta}'\}; L)$ are mutually orthogonal for any choice of Υ_1 .*

Proof. We note that if $\bar{\beta} - \bar{\beta}' \notin L$ then $\bar{\beta} + L$ and $\bar{\beta}' + L$ define different cosets of L . Hence, any two codewords $\mathbf{c}(\boldsymbol{\lambda}, \bar{\beta}; L, p^a)$ and $\mathbf{c}(\boldsymbol{\lambda}', \bar{\beta}'; L, p^a)$ have non-intersecting supports and are hence orthogonal. \blacksquare

We note that Lemma 3.3.8 provides valuable insights into the construction of orthogonal sets. In particular, given that Υ_1 has been found such that $\mathcal{C}(\Upsilon_1, \{\bar{\beta}\}; L)$ is self orthogonal we can form larger orthogonal sets by taking a union over different choices of $\bar{\beta}$. This observation allows us to easily identify all possible orthogonal bases that are contained in $\mathcal{C}(L_a^d, L^c; L)$ in Theorem 3.3.10. However, before proceeding we require the following lemma.

Lemma 3.3.9. *Let $a = 1$. Then, for any $\Upsilon_1 \subset L_1^d$ the code $\mathcal{C}(\Upsilon_1, L^c; L)$ is self orthogonal.*

Moreover, $\mathcal{C}(L_1^d, L^c; L)$ is an orthogonal basis for \mathbb{C}^m .

Proof. See Appendix C.2.3 ■

Note that Lemma 3.3.9 states that the code $\mathcal{C}(L_1^d, L^c; L)$ is a single orthogonal basis for \mathbb{C}^m . As $\mathcal{C}(L_a^d, L^c; L)$ contains many more lines than $\mathcal{C}(L_1^d, L^c; L)$ it is natural to guess that $\mathcal{C}(L_a^d, L^c; L)$ contains more than one orthogonal basis for \mathbb{C}^m . In the sequel we will show that every orthogonal basis for \mathbb{C}^m that is contained in $\mathcal{C}(L_a^d, L^c; L)$ is not too different from that in Lemma 3.3.9. In order to see this note that $\zeta_{p^a}^{p^{a-1}} = \zeta_p$ as

$$\zeta_{p^a}^{p^{a-1}} = \exp\left(\frac{2\pi\sqrt{-1}p^{a-1}}{p^a}\right) = \exp\left(\frac{2\pi\sqrt{-1}}{p}\right) = \zeta_p.$$

Hence, any codevector derived over \mathbb{Z}_p can be lifted to a codevector over \mathbb{Z}_{p^a} using the lift defined in (3.14) and multiplying this lifted element by p^{a-1} . Thus, for any $a > 1$ we can embed the orthogonal basis described by Lemma 3.3.9 (constructed with a $\tilde{\Upsilon}_1 \subset \mathbb{Z}_p$) in a code derived over \mathbb{Z}_{p^a} as

$$\mathcal{C}(p^{a-1} \cdot L_1^d, L^c; L).$$

We note, as $\mathcal{C}(p^{a-1} \cdot L_1^d, L^c; L)$ forms as basis for \mathbb{C}^m , so will

$$\mathbf{T}(\boldsymbol{\lambda}) \cdot \mathcal{C}(p^{a-1} \cdot L_1^d, L^c; L) \tag{3.25}$$

for all $\boldsymbol{\lambda} \in L_a^d$ as $\mathbf{T}(\boldsymbol{\lambda})$ is a unitary transformation and preserves inner product relations. However, there will clearly be an equivalence between some orthogonal bases if one naively tries to enumerate all orthogonal bases using every element of L_a^d and (3.25). In this direction, we let

$$\downarrow L_a^d = (L_a^d \pmod{p^{a-1}}) \cap L_a^d$$

be the set of elements of L_a^d that are complimentary to $p^{a-1} \cdot L_1^d$, i.e. the set of elements in L_a^d such that each element $\boldsymbol{\lambda}$ of L_a^d can be written uniquely as

$$\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}} + \bar{\boldsymbol{\lambda}}$$

where $\hat{\boldsymbol{\lambda}} \in \downarrow L_a^d$ and $\bar{\boldsymbol{\lambda}} \in p^{a-1} \cdot L_1^d$. Intuitively $\downarrow L_a^d$ is the subset of L_a^d for which each coordinate of every element of L_a^d has been reduced modulo p^{a-1} . Thus, as $\mathcal{C}(p^{a-1} \cdot L_1^d, L^c; L)$ is an orthogonal basis for \mathbb{C}^m , so will $\mathcal{C}(\bar{\boldsymbol{\lambda}} + p^{a-1} \cdot L_1^d, L^c; L)$ for all $\bar{\boldsymbol{\lambda}} \in \downarrow L_a^d$. Moreover, each $\bar{\boldsymbol{\lambda}} \in \downarrow L_a^d$ defines a unique basis as the vectors from any two orthogonal bases $\mathcal{C}(\bar{\boldsymbol{\lambda}}_1 + p^{a-1} \cdot L_1^d, L^c; L)$ and $\mathcal{C}(\bar{\boldsymbol{\lambda}}_2 + p^{a-1} \cdot L_1^d, L^c; L)$ have zero intersection for $\bar{\boldsymbol{\lambda}}_1 \neq \bar{\boldsymbol{\lambda}}_2 \in \downarrow L_a^d$. However, it is not at all clear whether two arbitrary codewords $\mathbf{c}_1 = \mathbf{c}(\bar{\boldsymbol{\lambda}}_1 + \hat{\boldsymbol{\lambda}}_1, \boldsymbol{\beta}; L, p^a)$ and $\mathbf{c}_2 = \mathbf{c}(\bar{\boldsymbol{\lambda}}_2 + \hat{\boldsymbol{\lambda}}_2, \boldsymbol{\beta}; L, p^a)$ are orthogonal where $(\bar{\boldsymbol{\lambda}}_i, \hat{\boldsymbol{\lambda}}_i) \in \downarrow L_a^d \times (p^{a-1} \cdot L_1^d)$. It is clear from our previous discussion that if $\bar{\boldsymbol{\lambda}}_1 = \bar{\boldsymbol{\lambda}}_2$ and $\hat{\boldsymbol{\lambda}}_1 \neq \hat{\boldsymbol{\lambda}}_2$ then \mathbf{c}_1 and \mathbf{c}_2 are orthogonal. We note that this observation is a special case of our more general theorem to follow. However, before stating this more general theorem we require a few definitions.

To begin, note that in the current framework the inner product between any two vectors $\mathbf{c}_1 = \mathbf{c}(\boldsymbol{\lambda}_1, \boldsymbol{\beta}; L, p^a)$ and $\mathbf{c}_2 = \mathbf{c}(\boldsymbol{\lambda}_2, \boldsymbol{\beta}; L, p^a)$ from $\mathcal{C}(L_a^d, L^c; L)$ is a function of the difference

of λ_1 and λ_2 . More precisely,

$$\mathbf{c}_2^\dagger \mathbf{c}_1 = \sum_{\bar{\gamma} \in L} \zeta^{\langle \lambda_1, \bar{\gamma} \rangle} \zeta^{-\langle \lambda_2, \bar{\gamma} \rangle} \mathbf{e}_{\bar{\gamma} + \bar{\beta}} \quad (3.26a)$$

$$= \sum_{\bar{\gamma} \in L} \zeta^{\langle \lambda_1 - \lambda_2, \bar{\gamma} \rangle} \mathbf{e}_{\bar{\gamma} + \bar{\beta}}. \quad (3.26b)$$

Hence, we let

$$\Gamma_C(\mathbf{a}; \beta, L) = \sum_{\bar{\gamma} \in L} \zeta^{\langle \mathbf{a}, \bar{\gamma} \rangle} \mathbf{e}_{\bar{\gamma} + \bar{\beta}}. \quad (3.27)$$

With this definition it is easy to see from (3.26b) that

$$\mathbf{c}(\lambda_2, \beta; L, p^a)^\dagger \mathbf{c}(\lambda_1, \beta; L, p^a) = \Gamma_C(\lambda_1 - \lambda_2; \beta, L).$$

Thus, in order to understand the orthogonality properties of the code $\mathcal{C}(L_a^d, L^c; L)$ it is sufficient to understand when the function $\Gamma_C(\mathbf{a}; \beta, L)$ is 0. In this direction, note that as the sum of (3.27) is over the elements of the subspace L of $(\mathbb{Z}_p)^{m'}$, we may alternatively write (3.27) as the sum

$$\sum_{x_{i_0}=0}^{p-1} \sum_{x_{i_1}=0}^{p-1} \cdots \sum_{x_{i_{d-1}}=0}^{p-1} \zeta^{\langle \mathbf{a}, \mathbf{x} \rangle} \mathbf{e}_{\mathbf{x} + \bar{\beta}} \quad (3.28)$$

where $\{i_0, i_1, \dots, i_{d-1}\} = \text{sup } L$, $d = \dim L$ and in turn where

$$\mathbf{x} = [x_0, x_1, \dots, x_{m'-1}]^\dagger.$$

where we let $x_j = 0$ if $j \notin \text{sup } L$. Representing (3.27) as the multivariate sum (3.28) is quite important in understanding when two codewords are orthogonal. In particular, for any vector $\mathbf{x} \in (\mathbb{Z}_p)^{m'}$ let

$$\tilde{\mathbf{x}}_j = \mathbf{x} - x_j \mathbf{e}_j.$$

Then, for any \mathbf{a} and any $0 \leq j < m'$ one may write

$$\langle \mathbf{a}, \mathbf{x} \rangle = \langle \tilde{\mathbf{a}}_j, \tilde{\mathbf{x}}_j \rangle + a_j \cdot x_j = \langle \tilde{\mathbf{a}}_j, \tilde{\mathbf{x}}_j \rangle + (\hat{a}_j + p^{a-1} \cdot \bar{a}_j) x_j \quad (3.29)$$

for all $\mathbf{x} \in L$. Thus, we may rewrite the sum from (3.28) as

$$\sum_{x_{i_j}=0}^{p-1} \zeta_{p^a}^{(\hat{a}_j + p^{a-1} \cdot \bar{a}_j) \cdot x_j} \left(\sum_{x_{i_0}=0}^{p-1} \sum_{x_{i_1}=0}^{p-1} \cdots \sum_{x_{i_{j-1}}=0}^{p-1} \sum_{x_{i_{j+1}}=0}^{p-1} \cdots \sum_{x_{i_{d-1}}=0}^{p-1} \zeta_{p^a}^{\langle \tilde{\mathbf{a}}_j, \tilde{\mathbf{x}} \rangle} \mathbf{e}_{\mathbf{x} + \bar{\beta}} \right). \quad (3.30)$$

That is, if $\mathbf{a} \in (\mathbb{Z}_{p^a})^{m'}$ permits the decomposition (3.29) then we can “marginalize out” the variable x_j in the multivariate sum from (3.28). However, from elementary Fourier analysis on groups [92] (e.g. from our knowledge of the discrete Fourier transform) we know that

$$\left| \sum_{x_{i_j}=0}^{p-1} \zeta_{p^a}^{a_j \cdot x_j} \right| = \begin{cases} 0 & \text{if } a_j \neq 0 \text{ and } p^{a-1} \mid a_j \\ > 0 & \text{if } a_j \neq 0 \text{ and } p^{a-1} \nmid a_j \\ p & \text{if } a_j = 0 \end{cases}$$

Thus, if for a given \mathbf{a} and some $0 \leq j \leq m'$ we can marginalize out a variable in the sum

(3.28) such that

$$a_j = 0 + p^{a-1} \cdot \bar{a}_j \quad (3.31)$$

for some $0 < \bar{a}_j < p$ then $\Gamma_C(\mathbf{a}; \boldsymbol{\beta}, L) = 0$. Hence, to show that $\Gamma_C(\mathbf{a}; \boldsymbol{\beta}, L) = 0$ it is sufficient to show that (3.28) can be marginalized as in (3.30). Thus, we next examine a simple condition to test for this property. In this direction, recall that the *Hamming weight* of an element $\beta \in \mathbb{Z}_{p^a}$ is 1 if $\beta \neq 0$ and is 0 otherwise. We denote the Hamming weight of $\beta \in \mathbb{Z}_p$ as $\text{wt}_H(\beta)$ and the Hamming weight of any vector $\boldsymbol{\beta} \in (\mathbb{Z}_{p^a})^{m'}$ as $\text{wt}_H(\boldsymbol{\beta})$. Thus,

$$\text{wt}_H(\beta) = \begin{cases} 0 & \text{if } \beta = 0 \\ 1 & \text{o.w.} \end{cases}$$

and

$$\text{wt}_H(\boldsymbol{\beta}) = \sum_{i=0}^{m'-1} \text{wt}_H(\beta_i)$$

In order to identify orthogonal bases we will need a slightly modification to the Hamming weight which incorporates our prior observation that $\mathcal{C}(p^{a-1} \cdot L_1^d, L^c; L)$ is an orthogonal basis for \mathbb{C}^m . In particular, for any two codewords $\mathbf{c}_1 = \mathbf{c}(\boldsymbol{\lambda}_1, \boldsymbol{\beta}; L, p^a)$ and $\mathbf{c}_2 = \mathbf{c}(\boldsymbol{\lambda}_2, \boldsymbol{\beta}; L, p^a)$ from $\mathcal{C}(\boldsymbol{\lambda} + p^{a-1} \cdot L_1^d, L^c; L)$ one has

$$\text{wt}_H(\tilde{\boldsymbol{\lambda}}_1 - \tilde{\boldsymbol{\lambda}}_2) = 0 \text{ while } \text{wt}_H(\hat{\boldsymbol{\lambda}}_1 - \hat{\boldsymbol{\lambda}}_2) = m'. \quad (3.32)$$

However, from our preceding discussion it is clear that if $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ satisfy (3.32) then one may marginalize *any coordinate* of the sum (3.28) such that (3.31) holds. However, from our discussion it is clear that in general a far less strict requirement can be placed on the difference to determine orthogonality. In particular, reexamining (3.28) it is clear that so long as there is *some coordinate* for which (3.31) holds then $\Gamma_C(\mathbf{a}; \boldsymbol{\beta}, L) = 0$. In this direction, we let the *twisted Hamming weight* of an element $\boldsymbol{\beta} = (\hat{\boldsymbol{\beta}}, \bar{\boldsymbol{\beta}}) \in \downarrow L_a^d \times (p^{a-1} \cdot L_a^d)$ be the number of coordinates for which $\hat{\boldsymbol{\beta}}$ is zero and for which the corresponding entry of $\bar{\boldsymbol{\beta}}$ is non-zero. We denote this by quantity as $\text{twt}_H(\boldsymbol{\beta})$. More precisely,

$$\begin{aligned} \text{twt}_H(\boldsymbol{\beta}) &= \sum_{i=0}^{m'-1} \left(1 - \text{wt}_H(\hat{\beta}_i)\right) \text{wt}_H(\bar{\beta}_i) \\ &= \left| \left\{ i \mid \hat{\beta}_i = 0 \text{ and } \bar{\beta}_i \neq 0 \right\} \right|. \end{aligned} \quad (3.33)$$

This leads us to our characterization of all of the orthogonal bases for \mathbb{C}^m contained in the code $\mathcal{C}(L_a^d, L^c; L)$.

Theorem 3.3.10. *Let $\mathbf{c}_1 = \mathbf{c}(\boldsymbol{\lambda}_1, \boldsymbol{\beta}; L, p^a)$ and $\mathbf{c}_2 = \mathbf{c}(\boldsymbol{\lambda}_2, \boldsymbol{\beta}'; L, p^a)$ be any two codevectors of $\mathcal{C}(L_a^d, L^c; L)$. Let $\hat{\boldsymbol{\lambda}}_i = (\boldsymbol{\lambda}_i \pmod{p^{a-1}}) \cap L_a^d$ and $\bar{\boldsymbol{\lambda}}_i = \boldsymbol{\lambda}_i - \hat{\boldsymbol{\lambda}}_i$. Then, \mathbf{c}_1 and \mathbf{c}_2 are orthogonal if and only if one of the following hold:*

- (i) $\boldsymbol{\beta}' - \boldsymbol{\beta} \notin L$
- (ii) $\bar{\boldsymbol{\lambda}}_1 \neq \bar{\boldsymbol{\lambda}}_2$ and $\hat{\boldsymbol{\lambda}}_1 = \hat{\boldsymbol{\lambda}}_2$
- (iii) $0 < \text{twt}_H((\hat{\boldsymbol{\lambda}}_1 - \hat{\boldsymbol{\lambda}}_2, \bar{\boldsymbol{\lambda}}_1 - \bar{\boldsymbol{\lambda}}_2))$

Moreover, for any set $\mathcal{S} \subset \downarrow L_a^d \times (p^{a-1} \cdot L_a^d)$ such that (ii) and (iii) holds for every pair of

distinct elements, the set of vectors

$$\bigcup_{\beta \in L^c} \bigcup_{(\hat{\lambda}_s, \bar{\lambda}_s) \in S_\beta} \mathcal{C}(\hat{\lambda}_s + \bar{\lambda}_s, \{\beta\}; L) \quad (3.34)$$

is an orthogonal basis of \mathbb{C}^m contained in $\mathcal{C}(L_a^d, L^c; L)$. Additionally, every basis contained in $\mathcal{C}(L_a^d, L^c; L)$ is of the form (3.34).

Proof. See Appendix C.2.4 ■

Note that Theorem 3.3.10 encapsulates our discussion this far on the conditions needed for two codewords from our codebook to be orthogonal. Moreover, Theorem 3.3.10 shows that these conditions are in fact necessary to be orthogonal. Additionally, we note that condition (iii), by our previous discussion, implies condition (ii). That is, as we have seen any two distinct vectors from $\mathcal{C}(\hat{\lambda} + p^{a-1} \cdot L_a^d, L^c; L)$ have a twisted hamming weight of m' . However, we keep this case separate as it will be useful in the sequel. In particular, examining Example 3.2.4 one can see that vectors from the associated code that are orthogonal meet not only condition (iii) but (ii). Additionally, we note that this condition identifies a special type of orthogonality relations. That is, condition (ii) identifies the orthogonal codewords using a twisted hamming weight of m' , i.e. for which $\hat{\lambda}_i = \hat{\lambda}_j$. We note that this particular case is important as it identifies *disjoint* orthonormal bases contained in a code. In particular, we have the following corollary.

Corollary 3.3.11. *Let $\hat{\Upsilon}_1$ be any arbitrary subset of $\downarrow L_a^d$ and let*

$$\Upsilon_1 = \hat{\Upsilon}_1 + p^{a-1} \cdot L_1^d.$$

Then, $\mathcal{C}(\Upsilon_1, L^c; L)$ is a disjoint union of $|\hat{\Upsilon}_1|$ orthonormal bases forming $p^{m'} \cdot |\hat{\Upsilon}_1|$ distinct lines.

Proof. This is a simple extension of Theorem 3.3.10. As, $\lambda + p^{a-1} \cdot L_1^d$ forms a basis for \mathbb{C}^m and $\lambda_1 + p^{a-1} \cdot L_1^d \cap \lambda_2 + p^{a-1} \cdot L_1^d = \emptyset$ and hence $\hat{\Upsilon}_1 + p^{a-1} \cdot L_1^d$ determine a disjoint union of orthogonal bases. ■

Note that Theorem 3.3.10 greatly simplifies the problem of finding codebooks with many orthogonal bases with a large number of unitary matrices that act transitively on the codebook. In particular, by Theorem 3.3.10 it is sufficient to select a set from $\downarrow L_a^d$ that is closed under addition modulo p^{a-1} , say $\hat{\Upsilon}_1$, and select $\Upsilon_1 = \hat{\Upsilon}_1 + p^{a-1} \cdot L_1^d$. Then, the number of orthogonal bases can be determined by counting the number of subsets of $\hat{\Upsilon}_1 \times (p^{a-1} \cdot L_a^d)$ of cardinality $|p^{a-1} \cdot L_1^d|$ that satisfy Theorem 3.3.10. In this direction, we let $\Omega_{k,m'}(\hat{\Upsilon}_1)$ be the collection of sets of $\hat{\Upsilon}_1 \times (p^{a-1} \cdot L_a^d)$ cardinality k that satisfy Theorem 3.3.10. That is,

$$\Omega_{k,m'}(\hat{\Upsilon}_1) = \left\{ \mathcal{S} \subset \hat{\Upsilon}_1 \times (p^{a-1} \cdot L_a^d) \mid \text{for every } (\hat{\lambda}_i, \bar{\lambda}_i) \neq (\hat{\lambda}_j, \bar{\lambda}_j) \in \mathcal{S} \right. \\ \left. \text{either (ii) or (iii) of Theorem 3.3.10 holds} \right\} \quad (3.35)$$

This leads to the following corollary to Theorem 3.3.10.

Corollary 3.3.12. *Let $\hat{\Upsilon}_1$ a subset of $\downarrow L_a^d$ that is closed under addition modulo p^{a-1} . Then,*

$$\Upsilon_1 = \hat{\Upsilon}_1 + p^{a-1} \cdot L_1^d$$

is a linear subset of $(\mathbb{Z}_{p^a})^{m'}$ and $\mathcal{C}(\Upsilon_1, L^c; L)$ contains $|\hat{\Upsilon}_1| \cdot |L_1^d| \cdot |L^c|$ distinct lines which form $|\Omega_{k,m'}(\hat{\Upsilon}_1)|$ orthogonal bases.

Proof. See Appendix C.2.5. ■

In the sequel we will identify the set of orthogonal bases given in Theorem 3.3.10 as $U_a(L)$. Then, so long as we can identify a group of matrices that act two transitively on the set of orthogonal bases we can naturally find a subset that will act two transitively on any subset. That is, we let

$$U_a(L) = \left\{ \bigcup_{\beta \in L^c} \bigcup_{(\hat{\lambda}_s, \bar{\lambda}_s) \in S_\beta} \mathcal{C}(\hat{\lambda}_s + \bar{\lambda}_s, \{\beta\}; L) \mid S_\beta \in \Omega_{|p^{a-1} \cdot L_a^d|, m'}(\downarrow L_a^d) \right\} \quad (3.36)$$

It is clear that for any $\lambda' \in L_a^d$ and for any $\bar{\beta} \in T(\lambda')$ $T(\lambda')$ acts transitively on the collection of orthogonal sets $U_a(L)$ as the set of differences of cosets are equal, i.e. $\Delta S = \Delta(\lambda' + S)$. In this direction, we let

$$R(\lambda; \bar{\beta}) e_{\bar{\alpha}} = \begin{cases} T(\lambda) e_{\bar{\alpha}} & \text{if } \bar{\alpha} \in \bar{\beta} + L \\ e_{\bar{\alpha}} & \text{otherwise} \end{cases}$$

be the unitary transformation which acts as the identity for $\bar{\alpha} \notin \bar{\beta} + L$ and diagonally for $\bar{\alpha} \in \bar{\beta} + L$. For any subset \mathcal{D} of $\downarrow L_a^d$ we let the set of matrices

$$\mathcal{R}_L(\mathcal{D}) = \left\{ \prod_{\bar{\beta} \in L^c} R(\lambda_{\bar{\beta}}; \bar{\beta}) \mid \lambda_{\bar{\beta}} \in \mathcal{D} \subset \downarrow L_a^d \right\}.$$

Clearly, $\mathcal{R}_L(\downarrow L_a^d)$ acts transitively on $U_a(L)$ and the subgroup $\mathcal{R}_L(\Upsilon_1)$ will act transitively on a code selected according to Corollary 3.3.12 as $\hat{\Upsilon}_1$ is closed under addition. Moreover, $S(\bar{\gamma})$ acts transitively on the set $U_a(L)$ by permuting the terms of (3.34). These observations are the content of the following theorem.

Theorem 3.3.13. *Let $\hat{\Upsilon}_1$ be a subset of $\downarrow L_a^d$ that is closed under addition modulo p^{a-1} and let $\Upsilon_1 = \hat{\Upsilon}_1 + p^{a-1} \cdot L_1^d$. Then, every element of $\mathcal{R}_L(\hat{\Upsilon}_1)$ acts transitively on the orthogonal bases of $\mathcal{C}(\Upsilon_1, L^c; L)$ as well as transitively on the code. Moreover, $S(\bar{\gamma})$ acts transitively on the orthogonal bases of $\mathcal{C}(\Upsilon_1, L^c; L)$ for all $\bar{\gamma} \in L^c$ as well as transitively on the code. Further,*

$$\left\langle S(\bar{\gamma}) \cdot R(\lambda_{\bar{\beta}}; \bar{\beta}) \mid R(\lambda_{\bar{\beta}}; \bar{\beta}) \in \mathcal{R}_L(\hat{\Upsilon}_1) \text{ and } \bar{\gamma} \in L^c \right\rangle$$

acts transitively on the code $\mathcal{C}(\Upsilon_1, L^c; L)$ as well as the collection of orthogonal bases contained in $\mathcal{C}(\Upsilon_1, L^c; L)$.

Proof. See Appendix C.2.6 ■

We now return to examine Example 3.2.4 in light of Theorem 3.3.10 to provide a more concrete illustration of how Theorem 3.3.10 applies to the problem of interest.

Example 3.3.2 Two Orthogonal Bases Continued

Recall from Example 3.2.4 that we considered a set of lines in \mathbb{C}^4 that was the union of two orthogonal bases. Moreover, upon closer examination there were two additional orthogonal bases that came from exchanging two elements from each basis. To be more precise recall from Example 3.2.4 we chose:

- 1) $\mathcal{I} = \mathbb{F}_2^2 = \{[0, 0], [0, 1], [1, 0], [1, 1]\}$
- 2) $\mathcal{I}_0 = L = \{[0, 0], [0, 1]\}$
- 3) $\Upsilon_1 = \{[0, 0], [0, 1], [0, 2], [0, 3]\} \subset \mathbb{Z}_4$
- 4) $\Upsilon_2 = \{[0, 0], [1, 0]\} \subset \mathbb{F}_2^2$

which yielded a codebook that was the union of the two orthogonal bases:

$$\mathcal{B}_1 = \{[1, 1, 0, 0], [1, -1, 0, 0], [0, 0, 1, 1], [0, 0, -1, 1]\}$$

for $\Upsilon_1 = \{[0, 0], [0, 2]\}$ and

$$\mathcal{B}_2 = \{[1, \sqrt{-1}, 0, 0], [1, -\sqrt{-1}, 0, 0], [0, 0, \sqrt{-1}, 1], [0, 0, -\sqrt{-1}, 1]\}$$

for $\Upsilon_1 = \{[0, 1], [0, 3]\}$.

We now examine this codebook along the lines of Theorem 3.3.10. To begin, note that $\Upsilon_2 = L^c$, i.e. it is complimentary to L , and $L_1^d = L$. Thus,

$$2 \cdot L_1^d = \{[0, 0], [0, 2]\},$$

$$\mathcal{B}_1 = \mathcal{C}([0, 0] + \{[0, 0], [0, 2]\}, \Upsilon_2; L)$$

and

$$\mathcal{B}_2 = \mathcal{C}([0, 1] + \{[0, 0], [0, 2]\}, \Upsilon_2; L).$$

Hence, by Theorem 3.3.10 the two orthogonal bases \mathcal{B}_1 and \mathcal{B}_2 are self orthogonal as they are cosets of $2 \cdot L_1^d$. Moreover,

$$\mathcal{B}_2 = \mathbb{R}([0, 1]; [0, 0]) \mathbb{R}([0, 1]; [1, 0]) \cdot \mathcal{B}_1 = \mathbb{T}([0, 1]) \cdot \mathcal{B}_1.$$

However, by Theorem 3.3.10 the orthogonal bases

$$\mathcal{B}'_1 = \mathbb{R}([0, 1]; [1, 0]) \cdot \mathcal{B}_1 = \{[1, 1, 0, 0], [1, -1, 0, 0], [0, 0, \sqrt{-1}, 1], [0, 0, -\sqrt{-1}, 1]\}$$

and

$$\mathcal{B}'_2 = \mathbb{R}([0, 1]; [1, 0]) \cdot \mathcal{B}_2 = \{[1, \sqrt{-1}, 0, 0], [1, -\sqrt{-1}, 0, 0], [0, 0, 1, 1], [0, 0, -1, 1]\}$$

are orthogonal as well and as $[0, 1] \in \hat{\Upsilon}_1$ are contained in the code as. In fact, again by Theorem 3.3.10 these are all of the orthogonal subsets of $\mathcal{C}(L_a^d, L^c; L)$. The relationship between the orthogonal bases $\mathcal{B}_1, \mathcal{B}_2, \mathcal{B}'_1$ and \mathcal{B}'_2 can be seen in Figure 3-20.

Note that we have discussed the relevant aspects of the construction of sparse quantizers, but have yet to discuss how one may develop dense codes that are invariant to the shifts in the coordinate sets which act transitively on the base code. A naive approach to construct

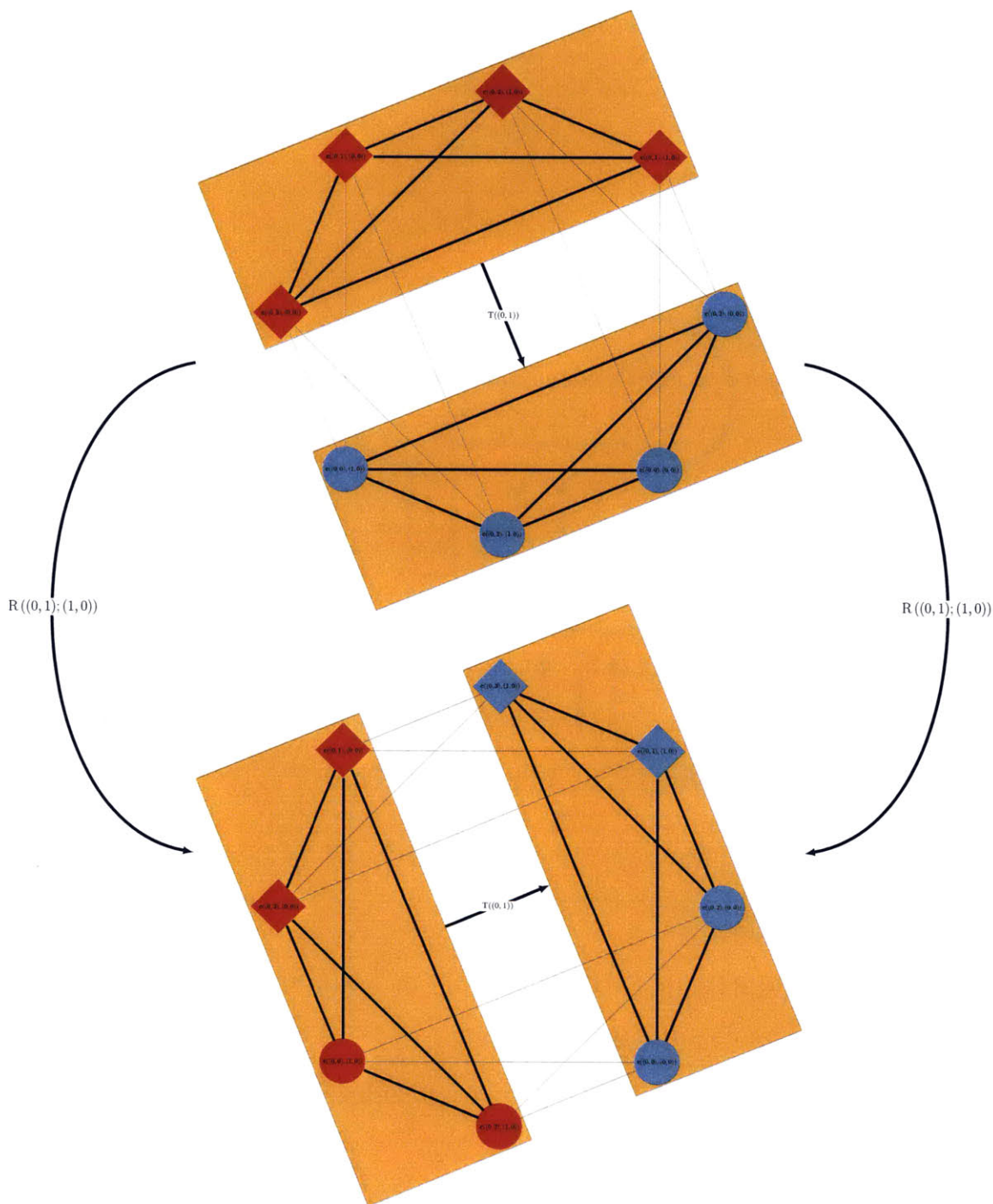


Figure 3-20. A depiction of the relationships between the four orthogonal bases of Example 3.2.4. Two copies of the code from Example 3.2.4 are depicted in the figure; one copy is located at the top of the figure and the other at the bottom. Note that the operator $T(\lambda)$ translates between pairs of orthogonal bases, as seen in both the copy of the code at the top of the figure as well as at the bottom. The operator $R(\lambda; \beta)$ interchanges elements of a basis. The action of the operator $R(\lambda; \beta)$ is depicted by the change in groupings (represented by rectangles) between the two copies of the code.

such dense codes is to consider indexing the dense codes by a weighted sum elements of the linear space indexing the coordinated sets, for example $\mathbb{F}_2 + 2 \cdot \mathbb{F}_2$, in the natural manner over the larger integer ring. We now provide a second illustration of Theorem 3.3.10 using a dense code of this form.

Example 3.3.3 A 4-bit Quantizer with Near Optimum Correlation

We now examine a 4-bit quantizer in the context of Theorem 3.3.10 and Corollary 3.3.12. In particular, we consider the code defined by

1. $\mathcal{I} = \mathbb{F}_2^2 = \{[0, 0], [0, 1], [1, 0], [1, 1]\}$
2. $L = \mathbb{F}_2^2$
3. $\Upsilon_1 = \mathcal{I} + 2 \cdot \mathcal{I}$
4. $\Upsilon_2 = \{[0, 0]\}$

which yields a code containing 16 codewords. We note that this could be obtained directly from Corollary 3.3.12 as $|\mathbb{F}_2^2| \cdot |L_1^d| \cdot |L^c| = 4 \cdot 4 \cdot 1 = 16$. Further, note that in the current example $L_2^d = (\mathbb{Z}_4)^2$ and $2 \cdot L_2^d = 2 \cdot \mathbb{F}_2^2$. Now, in order to identify the orthogonal bases, we explicitly enumerate the elements of $\Omega_{4,2}(\hat{\Upsilon}_1)$ which contain $[0, 0]$. These elements are (with the corresponding elements of $2 \cdot L_a^d$ at the top of each column):

		[0, 0]	[2, 0]	[0, 2]	[2, 2]
$\mathcal{S}_1 = \{$		[0, 0],	[0, 0],	[0, 0],	[0, 0] }
$\mathcal{S}_2 = \{$		[0, 0],	[0, 0],	[1, 0],	[1, 0] }
$\mathcal{S}_3 = \{$		[0, 0],	[0, 1],	[0, 0],	[0, 1] }

Repeating this for each element of $\hat{\Upsilon}_1$ and noting that each step defines 3 unique orthogonal bases one can see that there are $3 \cdot |\hat{\Upsilon}_1| = 12$ orthogonal bases for \mathbb{C}^4 contained in this code. These are depicted in Figure 3-21. By direct computation one can further see that this code is orthogonal to 7 codevectors, has correlation of magnitude of $1/\sqrt{2}$ with 4 codevectors and correlation of magnitude of $1/2$ with the remaining 4 codevectors.

We note that the code of Example 3.3.3 meets the RMS Welch bound for correlation (2.20) while having many orthogonal sets. This may be seen in Figure 3-23 (a). Moreover, if for every codevector the 4 codevectors with correlation of magnitude of $1/\sqrt{2}$ could be moved so that they have correlation of magnitude $1/2$ and 4 vectors that are orthogonal could be moved so that they have correlation of magnitude $1/2$ then this code would meet the Welch bound for coherence and hence would have a uniform cross correlation yielding low mean squared quantization error. This may be seen in Figure 3-23 (b). We note, however, if the vectors that are orthogonal are not moved then the resulting code would violate the average Welch bound and hence such a rearrangement would not be possible. That is, as previously mentioned, there is a trade-off between the number of orthogonal sets and maximum correlation between any two codewords. It is natural to consider whether there is a systematic way to see this trade-off in our current quantization framework. We note that this, in part, has been answered by Theorem 3.3.10. Indeed, if we modify our choice for $\hat{\Upsilon}_1$ Theorem 3.3.10 describes how many orthogonal bases are removed. We now illustrate this observation with a very important example, the Kerdock line set [30].

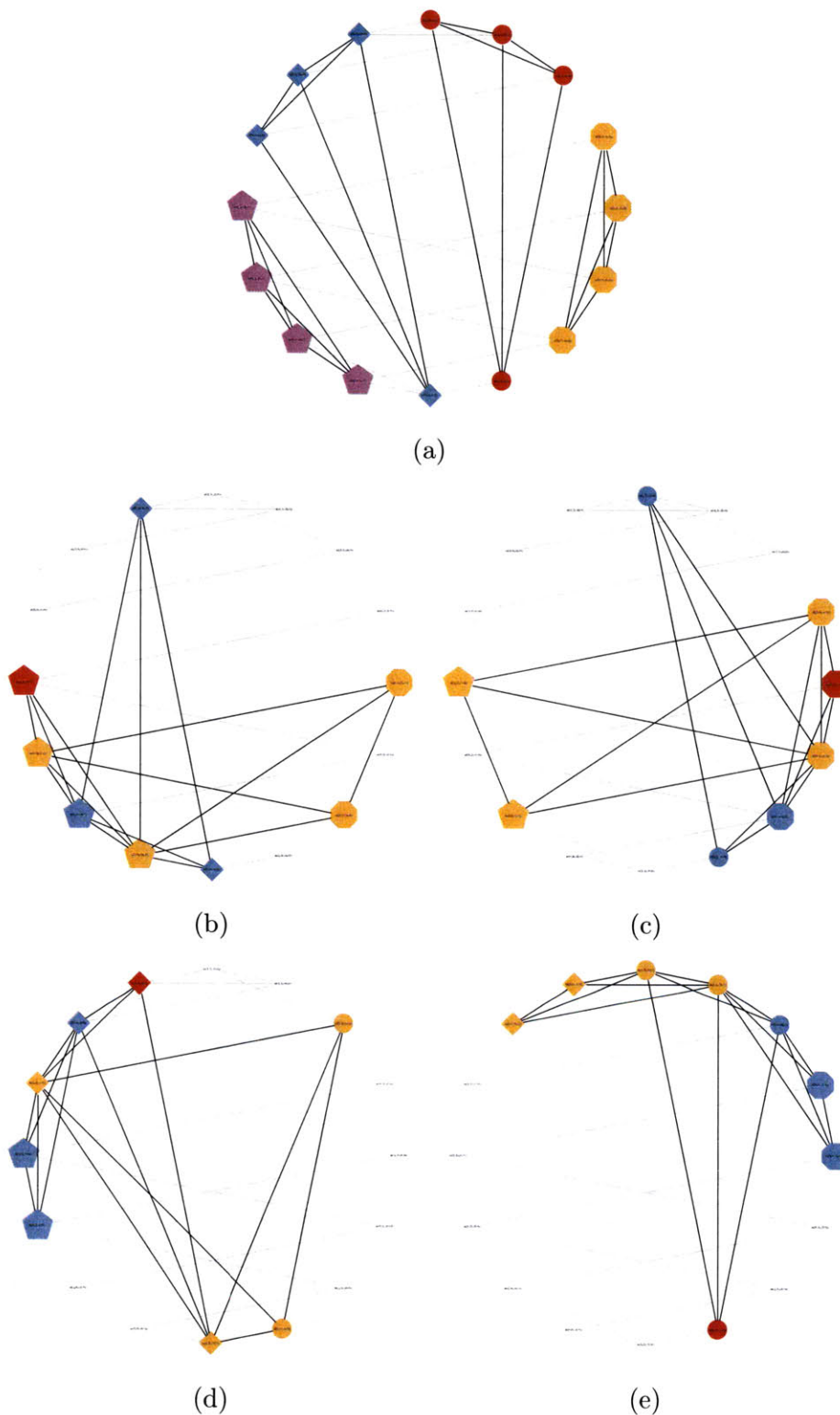


Figure 3-21. An illustration of the orthogonal sets of the code from Example 3.3.3. (a) The orthogonal sets that correspond to condition (ii) of Theorem 3.3.10 and (b-e) the orthogonal sets that correspond to condition (iii) of Theorem 3.3.10 which contain a fixed element of $\Omega_{4,2}(\hat{\Upsilon}_1)$ (b) $[0, 0]$ is fixed, (c) $[0, 1]$ is fixed, (d) $[1, 0]$ is fixed and (e) $[1, 1]$ is fixed.

Example 3.3.4 The Kerdock Line Set in \mathbb{C}^4

We now examine a 4-bit quantizer that trades the number of orthogonal bases for better codebook coherence as compared to the 4-bit quantizer from Example 3.3.3. For this example we do not assume that the basis is labeled by a linear space and hence revert to the notation of (3.5). In particular, we consider the code defined by

1. $\mathcal{I} = \{[0, 0], [0, 1], [1, 0], [3, 3]\}$
2. $\mathcal{I}_0 = \mathcal{I}$
3. $\Upsilon_1 = \mathcal{I} + 2 \cdot \mathcal{I}$
4. $\Upsilon_2 = \{[0, 0]\}$

which yields a code containing 16 codewords. Now, in order to identify the orthogonal bases, we explicitly enumerate the elements of the code that form orthogonal bases (this can be done with a slight modification to Theorem 3.3.10 that is not provided here). These elements are (with the corresponding elements of $2 \cdot \mathcal{I}$ at the top of each column):

	[0, 0]	[2, 0]	[0, 2]	[2, 2]
$\mathcal{S}_1 = \{$	$[0, 0],$	$[0, 0],$	$[0, 0],$	$[0, 0] \}$
$\mathcal{S}_2 = \{$	$[0, 1],$	$[0, 1],$	$[0, 1],$	$[0, 1] \}$
$\mathcal{S}_3 = \{$	$[1, 0],$	$[1, 0],$	$[1, 0],$	$[1, 0] \}$
$\mathcal{S}_4 = \{$	$[3, 3],$	$[3, 3],$	$[3, 3],$	$[3, 3] \}$

Thus, the only orthogonal bases of this code are the ones satisfying condition (ii) of Theorem 3.3.10. These are depicted in Figure 3-22. By direct computation one can further see that every codevector is orthogonal to 3 codevectors and has correlation of magnitude of $1/2$ with 12 codevectors.

It should be clear from Examples 3.3.3 and 3.3.4 that two similarly defined quantizers can result in quite different objects. In particular, upon closer examination the set Υ_1 in Examples 3.3.3 and 3.3.4 are equal. Thus, the only difference was in the set chosen for the basis, or alternatively, the bilinear form used for the inner product. In the sequel we will provide a generalization of our present quantization framework that will make this subtlety more clear. In particular we explicitly give quantizer constructions that interpolate between the competing design objects of orthogonality and coherence.

■ 3.4 Component Codes with Varying Degrees of Orthogonality

In the previous section we developed a framework to construct a family of component codes which contained many orthogonal bases. To do this we fixed a natural “lift” from \mathbb{Z}_p to \mathbb{Z}_{p^a} . However, in Examples 3.3.3 and 3.3.4 we saw that while these two quantizers were almost identically defined the “lift” caused the number of orthogonal bases contained in the code as well as the distribution of the inner product between codewords to vary. In this section we provide an explanation of this phenomenon and show that the generalizations of these two examples contain the fewest and greatest number of orthogonal bases in our framework. Hence, in this section we provide a method in which one may interpolate between these two extremes, providing a family of good low-rate component codes. To do this we generalize our preceding results to include more general bilinear maps. That is, recall that in our

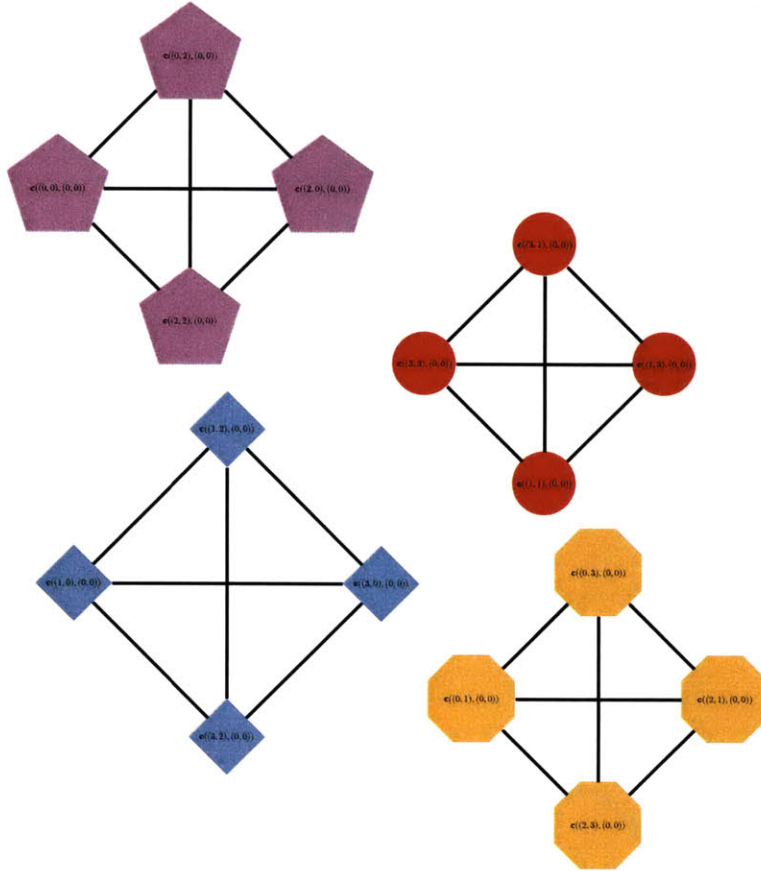


Figure 3-22. An illustration of the orthogonal sets of the code from Example 3.3.4. Note that this shows only 4 non-intersecting orthogonal bases while the code of Example 3.3.3 had 12 orthogonal bases (see Figure 3-21).

derivations of the unitary matrices that acted transitively (or invariantly) on the codebooks $\mathcal{C}(\Upsilon_1, \Upsilon_2; L)$ it was the bilinear nature of the inner product that allowed us to identify how the actions of $T(\lambda)$, $S(\beta)$ and their products behaved on the codebook. In particular, the key equations (3.18), (3.24a)–(3.24f) that led to the insights in to the matrices that act transitively on the codebook relied on the fact that the inner product defined in (3.13) was a bilinear map. In particular, the inner product allowed us to explicitly characterize the orthogonal bases as well as index the codewords that were eigenvectors of the set of matrices $\mathcal{H}_{L,a}$. Thus, it is natural to extend the quantization framework (3.5) of Section 3.2 in terms of a set of bilinear maps. In the sequel we consider a more general class. In particular, we consider \mathbb{Z}_{p^a} -valued bilinear forms on a module that is a finite extension of \mathbb{Z}_{p^a} . To make this more precise we make the following definitions and offer a more complete exposition in Appendix A.

■ 3.4.1 Finite Extensions of \mathbb{Z}_{p^a}

We now make the definitions regarding rings that are a finite extension of \mathbb{Z}_{p^a} that we require in the sequel. For the reader who is unfamiliar with the theory of finite extensions

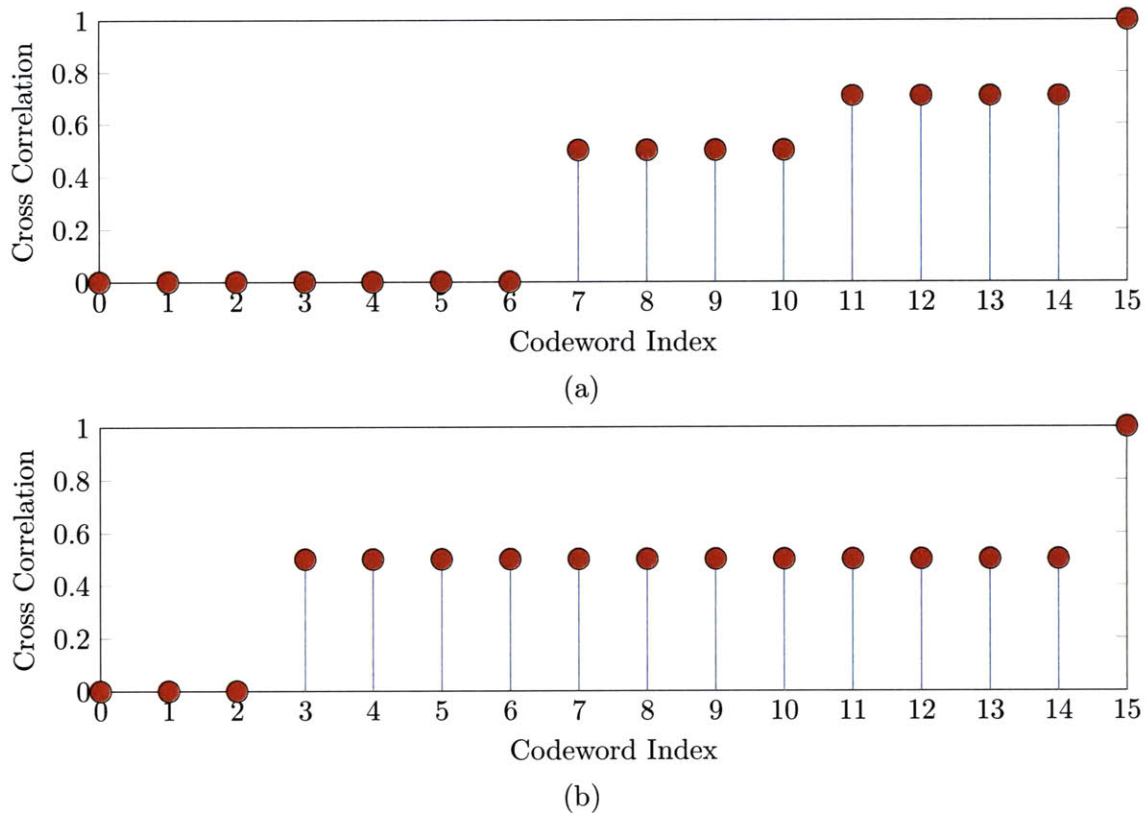


Figure 3-23. The cross correlation spectrum of the quantizers from Example 3.3.3 and Example 3.3.4. (a) The correlation spectrum of the quantizer from Example 3.3.3 which has more orthogonal vectors per code word but higher coherence. (b) The correlation spectrum of the quantizer from Example 3.3.4 which has few orthogonal vectors per code word but lower coherence and uniform cross correlation spectrum yielding a low mean squared error.

of \mathbb{Z}_{p^a} we note that in many ways this theory coincides with the theory of finite fields¹⁰ (i.e. finite extensions of \mathbb{Z}_p). To be more precise, recall that a polynomial $\bar{f}(x) \in (\mathbb{Z}_p)[x]$ of degree m over \mathbb{Z}_p is the a polynomial such that

$$\bar{f}(x) = \sum_{i=0}^m \bar{a}_i x^i$$

where $\bar{a}_i \in \mathbb{Z}_p$. The polynomial $\bar{f}(x)$ is monic if $\bar{a}_m = 1$ and $\bar{f}(x)$ is irreducible over \mathbb{Z}_p if it does not factor over $(\mathbb{Z}_p)[x]$, i.e. if $\bar{f}(x) = \bar{g}(x)\bar{h}(x)$ where $\bar{g}(x), \bar{h}(x) \in (\mathbb{Z}_p)[x]$ then either $\bar{g}(x)$ or $\bar{h}(x)$ is constant. The polynomial $\bar{f}(x)$ is primitive over \mathbb{Z}_p if it is irreducible and the smallest natural number n such that $\bar{f}(x)$ divides $x^n - 1$ is $n = p^m - 1$. It is natural to wonder whether knowledge of a the characteristics of a polynomial over \mathbb{Z}_p in anyway relates to those over \mathbb{Z}_{p^a} . In this direction, let μ be the homomorphism from \mathbb{Z}_{p^a} to \mathbb{Z}_p that reduces any element of \mathbb{Z}_{p^a} modulo p . We then have the following lemma from [29, 85].

Lemma 3.4.1. *Let $\bar{f}(x) \in (\mathbb{Z}_p)[x]$ be a degree m' monic irreducible divisor of $x^{p^{m'}-1} - 1$*

¹⁰However, there is one notable exception: the generator of the Galois group of an extension of \mathbb{Z}_{p^a} is not the power map on every element in an extension of \mathbb{Z}_{p^a} .

over \mathbb{Z}_p . Then there exists a unique irreducible polynomial $f(x) \in (\mathbb{Z}_{p^a})[x]$ which divides $x^{p^{m'}-1} - 1$ over \mathbb{Z}_{p^a} such that $\bar{f}(x) = \mu f(x)$.

In the sequel we will for any monic irreducible divisor of $x^{p^{m'}-1} - 1$ over \mathbb{Z}_p , say $\bar{f}(x)$, denote the unique divisor of $x^{p^{m'}-1} - 1$ over \mathbb{Z}_{p^a} such that $\bar{f}(x) = \mu f(x)$ simply as $f(x)$ and say that $f(x)$ is the lift of $\bar{f}(x)$. We note that Lemma 3.4.1 describes a quite strong correspondence between polynomials over \mathbb{Z}_p and \mathbb{Z}_{p^a} . In particular it provides a correspondence between the roots of $\bar{f}(x)$ and $f(x)$. In particular, if ζ is a primitive root of $f(x)$ then $\bar{\zeta} = \mu\zeta$ is a primitive root of $\bar{f}(x)$. As the finite field $\mathbb{F}_{p^{m'}}$ is by definition $\mathbb{Z}_p[\bar{\zeta}]$ it is likely that many of the properties of an extension of \mathbb{Z}_p will carry over to an extension of \mathbb{Z}_{p^a} . Thus, it is natural to define an analogous object over \mathbb{Z}_{p^a} . In this direction let ζ be a primitive root of a monic irreducible polynomial over \mathbb{Z}_{p^a} of degree m' . Then, we let $\text{GR}(p^a, m') = \mathbb{Z}_{p^a}[\zeta]$ be the Galois ring of degree m' over \mathbb{Z}_{p^a} . The reader should note that if $a = 1$ the Galois ring $\text{GR}(p, m')$ is simply the standard Galois field $\text{GF}(p^{m'}) = \mathbb{F}_{p^{m'}}$ and if $m' = 1$ the Galois ring $\text{GR}(p^a, 1)$ is simply the ring of integers modulo p^a , \mathbb{Z}_{p^a} .

In the sequel we will say that a primitive root $\zeta \in \text{GR}(p^a, m')$ is the “lift” of the primitive element $\bar{\zeta} \in \text{GF}(p^{m'})$ if $\bar{\zeta} = \mu\zeta$. That is, $\zeta \in \text{GR}(p^a, m')$ is the “lift” of $\bar{\zeta}$ to $\text{GR}(p^a, m')$ if $f(x)$ is the unique lift of $\bar{f}(x)$ implied by Lemma 3.4.1, ζ is a root of $f(x)$, $\bar{\zeta}$ is a root of $\bar{f}(x)$ and $\bar{\zeta} = \mu(\zeta)$. Recall that primitive element of $\bar{\zeta} \in \mathbb{F}_{p^{m'}}$ generates the non-zero elements of $\mathbb{F}_{p^{m'}}$. Hence, μ induces an isomorphism from $\mathbb{F}_{p^{m'}}$ to the set

$$\mathcal{T}_{p^a, m'} = \{0, \zeta, \zeta^2, \dots, \zeta^{p^{m'}-1}\}$$

The elements of $\text{GR}(p^a, m')$ have two simple representations. First, for any $r \in \text{GR}(p^a, m')$, we have (analogous to the representation of finite fields)

$$r = \sum_{i=0}^{m'-1} r_i \zeta^{p^i} \quad (3.37)$$

where $r_i \in \mathbb{Z}_{p^a}$ and ζ is a primitive element of $\text{GR}(p^a, m')$. Second, for any $r \in \text{GR}(p^a, m')$, we have

$$r = \sum_{i=0}^{a-1} p^i u_i \quad (3.38)$$

where $u_i \in \mathcal{T}_{p^a, m'}$. In order to use the results of the preceding section we will need a bilinear map from $\text{GR}(p^a, m') \times \text{GR}(p^a, m')$ to \mathbb{Z}_{p^a} . In this direction let for any $r \in \text{GR}(p^a, m')$, using the expansion (3.38)

$$\phi(r) = \sum_{i=0}^{a-1} r_i^p p^i. \quad (3.39)$$

This is the *Frobenius* automorphism of $\text{GR}(p^a, m')$ which acts as a power map on the elements of $\mathcal{T}_{p^a, m'}$ and leaves the elements of \mathbb{Z}_{p^a} fixed. We caution the reader that the *Frobenius* automorphism does not in general act as a power map on every element of $\text{GR}(p^a, m')$ as it does on $\text{GF}(p^{m'})$. That is, examining (3.39) one can see that the *Frobenius* automorphism in general only the power map if $a = 1$. Now, we define the trace map from

$\text{GR}(p^a, m')$ to \mathbb{Z}_{p^a} as

$$\text{Tr}_{\text{GR}(p^a, m')/\mathbb{Z}_{p^a}}(r) = \sum_{i=0}^{m'-1} \phi^i(r) \quad (3.40)$$

We note that both of the representations defined in (3.37) and (3.38) will be useful in the sequel. In fact, carefully examining (3.37) and (3.38) one can notice that both of these representations appeared in the previous section in a veiled form. In particular note by examining (3.37) that $\text{GR}(p^a, m')$ is a \mathbb{Z}_{p^a} -module. More precisely, to any $r \in \text{GR}(p^a, m')$ such that,

$$r = \sum_{i=0}^{m'-1} r_i \zeta^{p^i} \quad (3.41)$$

we can associate a vector

$$\mathbf{r} = [r_0, r_1, \dots, r_{m'-1}] \in (\mathbb{Z}_{p^a})^{m'} \quad (3.42)$$

One may be tempted to use this representation to define the inner product. Although this is not possible in general a simple alternative is. In this direction, let $\{\zeta_{\perp}^p, \zeta_{\perp}^{p^2}, \dots, \zeta_{\perp}^{p^{m'}}\}$ be the trace dual basis for the normal basis $\{\zeta^p, \zeta^{p^2}, \dots, \zeta^{p^{m'}}\}$. More precisely, $\{\zeta_{\perp}^p, \zeta_{\perp}^{p^2}, \dots, \zeta_{\perp}^{p^{m'}}\}$ is the set of elements

$$\text{Tr}(\zeta^{p^i} \cdot \zeta_{\perp}^{p^j}) = \delta(i - j)$$

which always exists [25, 85]. Then, one may write the inner product between the two elements $r, s \in \text{GR}(p^a, m')$ by

$$\text{Tr}_{\text{GR}(p^a, m')/\mathbb{Z}_{p^a}}(r \cdot s) = \sum_{i=0}^{m'-1} r_i s_i^{\perp} \pmod{p^a}. \quad (3.43)$$

where

$$r = \sum_{i=0}^{m'-1} r_i \zeta^{p^i} \quad \text{and} \quad s = \sum_{i=0}^{m'-1} s_i^{\perp} \zeta_{\perp}^{p^i}.$$

Comparing (3.43) to (3.13) it is clear that

$$\langle \mathbf{r}, \mathbf{s}_{\perp} \rangle = \text{Tr}_{\text{GR}(p^a, m')/\mathbb{Z}_{p^a}}(r \cdot s).$$

Moreover, by examining (3.38) one can see that it is quite easy to identify the elements of interest from our previous discussion. In particular,

$$\downarrow L_a^d = \left\{ \mathbf{r} \mid r = \sum_{i=0}^{a-2} p^i u_i \text{ where } u_i \in \mathcal{T}_{p^a, m'} \right\}. \quad (3.44)$$

Hence, many of our results from Section 3.3.1 can be restated by replacing $(\mathbb{Z}_{p^a})^{m'}$ with $\text{GR}(p^a, m')$, $\langle \mathbf{r}, \mathbf{s} \rangle$ with $\text{Tr}_{\text{GR}(p^a, m')/\mathbb{Z}_{p^a}}(r \cdot s)$ and $\downarrow L_a^d$ with the natural embedding of $\text{GR}(p^{a-1}, m')$ in $\text{GR}(p^a, m')$, (3.44). This will be done in our most general results to follow. However, we first need to generalize our existing results to account for the bilinear form $\text{Tr}(\cdot)$

■ 3.4.2 Codes Defined Through The Trace Map

Now with these new definitions at hand we return to the question of explaining why Examples 3.3.3 and 3.3.4 have such different properties. Then, we provide a method in which one may interpolate between these two extremes. This will be done by generalizing our results for our bi-variate map $c(i, j)$ to include more general bilinear maps. In this direction, a \mathbb{Z}_{p^a} -valued bilinear map on $\text{GR}(p^a, m')$ is a map $\beta : \text{GR}(p^a, m') \times \text{GR}(p^a, m') \rightarrow \mathbb{Z}_{p^a}$ such that for all $a, b, a_i, b_i \in \text{GR}(p^a, m')$ and $\alpha \in \mathbb{Z}_{p^a}$ one has

$$\beta(a_1 + a_2, b) = \beta(a_1, b) + \beta(a_2, b) \quad (3.45)$$

$$\beta(a, b_1 + b_2) = \beta(a, b_1) + \beta(a, b_2) \quad (3.46)$$

$$\beta(\alpha a, b) = \alpha \beta(a, b) = \beta(a, \alpha b) \quad (3.47)$$

Further, a \mathbb{Z}_{p^a} -valued quadratic map on $\text{GR}(p^a, m')$ is a map $f_Q : \mathbb{K} \rightarrow \mathbb{F}$ such that

$$\beta(y, z) = f_Q(z + y) - f_Q(z) - f_Q(y) + f_Q(0)$$

is a bilinear map. To make this more concrete, recall that for any two vectors $\alpha, \beta \in (\mathbb{Z}_{p^a})^{m'}$

$$2\langle \alpha, \beta \rangle = \|\alpha + \beta\| - \|\alpha\| - \|\beta\|$$

is a bilinear map where

$$\|\alpha\| = \sum_{i=0}^{n-1} \alpha_i^2$$

and in turn where $\alpha = [\alpha_0, \alpha_1, \dots, \alpha_{m'-1}]$. Hence, $\|\beta\|$ is a quadratic map on $(\mathbb{Z}_{p^a})^{m'}$. From our discussion in the previous section (Section 3.4.1)

$$2\langle \alpha, \beta_{\perp} \rangle = 2 \text{Tr}(\alpha \cdot \beta) \quad (3.48)$$

where α and β_{\perp} were the elements in $\text{GR}(p^a, m')$ corresponding to the vectors α and β in $(\mathbb{Z}_{p^a})^{m'}$ respectively. Hence $\text{Tr}(x^2)$ is a quadratic map on $\text{GR}(p^a, m')$ and $\text{Tr}(x \cdot y)$ is a \mathbb{Z}_{p^a} -valued bilinear map on $\text{GR}(p^a, m')$. Of particular interest in the sequel is the trace map $\text{Tr}(\cdot)$.

For any polynomial with coefficients in $\text{GR}(p^a, m')$, say $f(x) \in \text{GR}(p^a, m')[x]$, any subfield \mathbb{K} of \mathbb{Z}_p and any index set $\Upsilon_1 \subset \text{GR}(p^a, m')$ and shifts $\Upsilon_2 \subset \mathbb{Z}_p$ let

$$\mathcal{C}(\Upsilon_1, \Upsilon_2; \mathbb{K}, f) = \bigcup_{\tau \in \Upsilon_2} \bigcup_{y \in \Upsilon_1} \{\mathbf{c}(y, \tau; \mathbb{K}, f)\} \quad (3.49)$$

where in turn

$$\mathbf{c}(y, \tau; \mathbb{K}, f) = \sum_{z \in \mathbb{K}} \zeta^{\text{Tr}(y \cdot f(z))} \mathbf{e}_{z+\tau}.$$

In the sequel we show that the codes $\mathcal{C}(\Upsilon_1, \Upsilon_2; \mathbb{K}, f)$ has a large group of transitive actions that are easily identifiable for appropriately chosen Υ_1 and Υ_2 analogous to our derivation in Section 3.3.1. In this direction as a natural analogue to (3.17) we let for any polynomial $f \in \text{GR}(p^a, m')[x]$,

$$\mathbf{T}(\ell; f) \mathbf{e}_{\alpha} = \zeta^{\text{Tr}(y \cdot f(z))} \mathbf{e}_{\alpha}$$

be the diagonal transform associated with the polynomial f . Further, we let $S(\tau)$ be the

corresponding permutation matrix corresponding to shifting every basis element by τ , i.e.

$$S(\tau)\mathbf{e}_z = \mathbf{e}_{z+\tau}.$$

In order to relate these current results to those previously we again need to show that $T(\ell; f)$ and $S(\tau)$ act linearly on the index sets for the codewords, i.e.

$$T(\ell; f)\mathbf{c}(y, \tau; \mathbb{K}, f) = \mathbf{c}(y + \ell, \tau; \mathbb{K}, f)$$

and

$$S(\tau')\mathbf{c}(y, \tau; \mathbb{K}, f) = \mathbf{c}(y, \tau + \tau'; \mathbb{K}, f).$$

This is stated in the following lemmas which are direct analogues to Lemma 3.3.2 and Lemma 3.3.3.

Lemma 3.4.2. *Let Υ_1 be a subring of $\text{GR}(p^a, m')$ and let Υ_2 and \mathbb{K} be non-empty subsets of $\mathbb{F}_{p^{m'}}$. If the image of $f(\mathbb{K})$ in $\text{GR}(p^a, m')$ is an additive group then, $T(\ell; f)$ acts transitively on $\mathcal{C}(\Upsilon_1, \Upsilon_2; \mathbb{K}, f)$ for all $\lambda \in \Upsilon_1$.*

Proof. This follows directly from the proof of Lemma 3.3.2 with the assertion that if $f(\mathbb{K})$ forms an additive group then $\zeta^{\text{Tr}(y \cdot f(z))}$ is a linear character of the group and hence $\zeta^{\text{Tr}(y \cdot f(z_1))} \cdot \zeta^{\text{Tr}(y \cdot f(z_2))} = \zeta^{\text{Tr}(y \cdot f(z_3))}$ for some $z_3 \in \mathbb{K}$. ■

Lemma 3.4.3. *Let Υ_2 and \mathbb{K} be subfields of $\mathbb{F}_{p^{m'}}$ and let Υ_1 be a non-empty subset of $\text{GR}(p^a, m')$. Then, $S(\tau)$ acts transitively on $\mathcal{C}(\Upsilon_1, \Upsilon_2; \mathbb{K})$ for all $\tau \in \Upsilon_2$.*

Proof. This follows directly from Lemma 3.3.3. ■

Lemma 3.4.2 and Lemma 3.4.3 again form the base to our results on the unitary matrices that act transitively on the orthogonal bases contained in a code. However, we note one important subtlety that has appeared that was absent in our prior discussion. In Lemma 3.4.2 the introduction of the polynomial introduced an important constraint on the polynomial f , its image must form an additive group for our previous results to push through. This is an important observation exploited in the sequel. However, from the proof of Lemma 3.4.2 we can see that this simple constraint puts us in a quite unnatural position. That is, we no longer have a guarantee that the lift respects the addition in \mathbb{K} . Thus, the definition of a dual space and commutativity of the operators no longer may be clearly interpreted. In this direction, we note that nowhere in our design do we require the set indexing the basis to be a finite field. One may just as easily take it to be any additive group. In this direction, for any polynomial for which the image of $f(\mathbb{K})$ is an additive group, we let

$$\mathcal{R}_f(\mathbb{K}) = \{r \in \text{GR}(p^a, m') : r = f(k) \text{ some } k \in \mathbb{K}\}$$

be the image of $f(\mathbb{K})$. We emphasize¹¹ that *the addition which defines the additive group $\mathcal{R}_f(\mathbb{K})$ need not follow the addition law of the Galois ring $\text{GR}(p^a, m')$* . That is, as this set only describes actions permuting the basis one may choose any additive group inside this framework. Indeed, this was implicit in our discussion in Section 3.3.1 regarding our

¹¹We note that this approach was similarly used in [113] to construct a new class of quantum Hamming codes.

original lift. With this in mind one may naturally extend our preceding definitions for any additive group \mathcal{R} contained in $\text{GR}(p^a, m')$. In this direction, let

$$\mathcal{C}(\Upsilon_1, \Upsilon_2; \mathcal{R}, \text{id}) = \bigcup_{\tau \in \Upsilon_2} \bigcup_{y \in \Upsilon_1} \{\mathbf{c}(y, \tau; \mathcal{R}, \text{id})\} \quad (3.50)$$

where $\Upsilon_1 \subset \text{GR}(p^a, m')$, $\Upsilon_2 \subset \mathcal{R}$ and

$$\mathbf{c}(y, \tau'; \mathcal{R}, \text{id}) = \sum_{z' \in \mathcal{R}} \zeta^{\text{Tr}(y \cdot z')} \mathbf{e}_{z' + \tau'}.$$

where $\tau' \in \mathcal{R}$. With this formality we let, abusing notation as to illuminate an equivalence, let $\mathbf{T}(\ell)$ be the operator which acts diagonally on the basis and $\mathbf{S}(\tau)$ be the corresponding coordinate permutation. With this formalism we note that the results of Lemma 3.4.2 and Lemma 3.4.3 with out modification.

We begin our extension of our previous result assuming that the image of $f(\mathbb{K})$ is a additive group contained in $\text{GR}(p^a, m')$ and repeat the steps in (3.21) – (3.22) from this viewpoint. In this direction note one immediately has

$$\mathbf{T}(\ell)\mathbf{S}(\tau)\mathbf{e}_z = \zeta^{\text{Tr}(\ell \cdot \tau)} \mathbf{S}(\tau)\mathbf{T}(\ell)\mathbf{e}_z. \quad (3.51)$$

Thus, the matrices $\mathbf{T}(\ell)\mathbf{S}(\tau)$ commute if and only if $\zeta^{\text{Tr}(\ell \cdot \tau)}$. As we are only interested in quantizers that form a system of lines, codewords that differ by a simple phase are not of interest. Thus, we again consider the subset of $\text{GR}(p^a, m')$ that is “orthogonal” to $\mathcal{R}_f(\mathbb{K})$. We let, for any additive group $\mathcal{R} \subset \text{GR}(p^a, m')$,

$$\mathcal{R}^\perp = \{z \in \text{GR}(p^a, m') \mid \text{Tr}(z \cdot r) = 0 \quad \forall r \in \mathcal{R}\}$$

be the trace dual subset of \mathcal{R} . Then, again abusing notation,

$$\mathcal{H}_{\mathcal{R}, a} = \left\{ \mathbf{T}(\ell)\mathbf{S}(\tau) \mid \ell \in \mathcal{R}^\perp, \tau \in \mathcal{R} \right\}$$

is a commutative group of matrices. Then analogous to Lemma 3.3.4 and Lemma 3.3.5 we have the following lemmas.

Lemma 3.4.4. *Let $(\lambda', \tau') \in \mathcal{R}^\perp \times \mathcal{R}$ be given. Then, $\mathbf{T}(\lambda')\mathbf{S}(\tau') \in \mathcal{H}_{\mathcal{R}, a}$ and $\mathbf{c}(\lambda, \tau; \mathcal{R}, \text{id})$ is an eigenvector of $\mathbf{T}(\lambda')\mathbf{S}(\tau')$ with eigenvalue $\zeta^{-\text{Tr}(\lambda \cdot \tau')}$ for all $\lambda \in \text{GR}(p^a, m')$ and $\tau \in \mathcal{R}$.*

Lemma 3.4.5. *The codewords $\mathbf{c}(\lambda, \tau; \mathcal{R}, \text{id})$ and $\mathbf{c}(\lambda', \tau'; \mathcal{R}, \text{id})$ are colinear if and only if $\tau - \tau' \in \mathcal{R}$ and $\lambda - \lambda' \in \mathcal{R}^\perp$.*

We wish to identify matrices that acts transitively on the codewords. However, to extend the proceeding results one must find a way to decompose the Galois Ring $\text{GR}(p^a, m')$ to permit a direct summand. A natural way to do this is to extend the addition law of $\mathcal{R}_f(\mathbb{K})$ to $\text{GR}(p^a, m')$. In this direction, let $\oplus|_f$ be the addition law on $\mathcal{R}_f(\mathbb{K})$. We say that $\oplus|_f$ may be extended to $\text{GR}(p^a, m')$ if there exists a \oplus for which $(\text{GR}(p^a, m'), \oplus)$ is an additive group and

$$r \oplus s = r \oplus|_f s \quad \forall r, s \in \mathcal{R}_f(\mathbb{K})$$

If the addition law on $\mathcal{R}_f(\mathbb{K})$ may be extended to $\text{GR}(p^a, m')$ we say that $\mathcal{R}_f(\mathbb{K})$ extends to $\text{GR}(p^a, m')$. If $\mathcal{R}_f(\mathbb{K})$ extends to $\text{GR}(p^a, m')$ then $\mathcal{R}_f(\mathbb{K})$ is a subgroup of $(\text{GR}(p^a, m'), \oplus)$

and such a decomposition is natural. In this direction let, for \mathcal{R} which extends to $\text{GR}(p^a, m')$ \mathcal{R}^c be any subgroup of $(\mathcal{R}_f(\mathbb{F}), \oplus)$ that is complimentary to \mathcal{R} and let \mathcal{R}_a^d be any subgroup of $(\text{GR}(p^a, m'), \oplus)$ that is complimentary to \mathcal{R}^\perp . That is, \mathcal{R}^c is any subgroup

$$\mathcal{R}_f(\mathbb{K}) = \mathcal{R} \otimes \mathcal{R}^c$$

and \mathcal{R}_a^d is any subgroup such that

$$\text{GR}(p^a, m') = \mathcal{R}^\perp \otimes \mathcal{R}_a^d$$

where we have used \otimes to represent the direct sum. Then, as an analogue to Theorem 3.3.6 and Corollary 3.3.7 we have the following.

Theorem 3.4.6. *Let \mathcal{R} be additive group which may be extended to $\text{GR}(p^a, m')$ and suppose Υ_1 is an additive subgroup of $\text{GR}(p^a, m')$ and Υ_2 is an additive subgroup of \mathcal{R}^c . Then, $\mathcal{C}(\Upsilon_1, \Upsilon_2; \mathcal{R})$ is invariant to multiplication by any element of $\mathcal{H}_{\mathcal{R}, a}$. Moreover, any matrix $\mathbf{H}' \in \mathcal{H}_{\mathcal{R}, a}$ such that $\mathbf{H}' = \mathbf{T}(\lambda')\mathbf{S}(\tau')$ where $(\lambda', \tau') \in \Upsilon_1 \times \Upsilon_2$, acts transitively on the code $\mathcal{C}(\Upsilon_1, \Upsilon_2; \mathcal{R})$. More precisely, for all $\mathbf{c} \in \mathcal{C}(\Upsilon_1, \Upsilon_2; \mathcal{R})$, if $\mathbf{H}' = \mathbf{T}(\lambda')\mathbf{S}(\tau')$ for some $(\lambda', \tau') \in \Upsilon_1 \times \Upsilon_2$ then*

$$\mathbf{H}' \cdot \mathbf{c} \in \mathcal{C}(\Upsilon_1, \Upsilon_2; \mathcal{R})$$

and for any $\mathbf{H} \in \mathcal{H}_{\mathcal{R}, a}$,

$$\mathbf{H} \cdot \mathbf{c} = \mathbf{c}.$$

Proof. This follows directly from the preceding discussion and the results from the proof of Theorem 3.3.6. ■

Recall we used the analogue of Theorem 3.4.6 to motivate our notion of complimentary codes. That is, as a large part of our quantizer design has been motivated by developing unitary transformations that fix part of the codebook provided as this provides structure to aided in the design of user selection algorithms. However, Theorem 3.4.6 only exhibits unitary transformations that either fix the entire codebook or leaves no codevector fixed (if the transformation is of course not the identity). Recalling our consequence of Theorem 3.3.6 we saw that the matrix group $\mathcal{H}_{L, a}$ acted invariantly on any code while $\mathcal{H}_{L^c, a}$ acted strictly as translation. However, if we exchange L with L^c we obtain a code for which $\mathcal{H}_{L, a}$ acts transitively while $\mathcal{H}_{L^c, a}$ acts invariantly on the code. As our present framework mimics that of Theorem 3.3.6 this is again the case.

Corollary 3.4.7. *Let \mathcal{R} be additive group which may be extended to $\text{GR}(p^a, m')$ and suppose Υ_1 is an additive subgroup of $\text{GR}(p^a, m')$ and Υ_2 is an additive subgroup of \mathcal{R}^c . Further, suppose that $\tilde{\Upsilon}_1$ is an additive subgroup of \mathcal{R}^\perp and $\tilde{\Upsilon}_2$ is an additive subgroup of \mathcal{R} . Then, every $\mathbf{H}' = \mathbf{T}((; \lambda'))\mathbf{S}(\tau')$ for $(\lambda', \tau') \in \Upsilon_1 \times \Upsilon_2$ acts transitively on the code $\mathcal{C}(\Upsilon_1, \Upsilon_2; \mathcal{R})$ and invariantly on the code $\mathcal{C}(\tilde{\Upsilon}_1, \tilde{\Upsilon}_2; \mathcal{R}^c)$. Moreover, every $\mathbf{H} = \mathbf{T}(\lambda)\mathbf{S}(\tau)$ for $(\lambda, \tau) \in \tilde{\Upsilon}_1 \times \tilde{\Upsilon}_2$ acts transitively on the code $\mathcal{C}(\tilde{\Upsilon}_1, \tilde{\Upsilon}_2; \mathcal{R}^c)$ and invariantly on the code $\mathcal{C}(\Upsilon_1, \Upsilon_2; \mathcal{R})$.*

We are now left to identify the orthogonal subset of the code $\mathcal{C}(\Upsilon_1, \Upsilon_2; \mathcal{R})$ and their structure. We begin with our most elementary result.

Lemma 3.4.8. *If $\tau \ominus \tau' \notin \mathcal{R}$ then the codes $\mathcal{C}(\Upsilon_1, \{\tau\}; \mathcal{R})$ and $\mathcal{C}(\Upsilon_1, \{\tau'\}; \mathcal{R})$ are mutually orthogonal for any choice of Υ_1 .*

Recall that the analogue to Lemma 3.4.8 in the preceding discussion provided valuable insights into how one may form many orthogonal bases. In fact, it led to the observation that altering the dimension of linear space L leads to a rapid growth in the number of orthogonal bases. Additionally, Lemma 3.3.8 provided condition (i) in Theorem 3.3.10 to test if any two vectors are orthogonal. Thus, the influence one's choice of the dimension of \mathbb{K} has on the number of orthogonal bases in $\mathcal{C}(\Upsilon_1, \Upsilon_2; \mathcal{R})$ is identical to that of L . That is, the insights developed for one's choice of L in Section 3.3.1 carry over to \mathbb{K} without modification.

In the sequel we show that condition (ii) in Theorem 3.3.10 carries over as well. That is, a constant shift to the “canonical” basis corresponding to a \mathbb{F}_{p^m} is again a basis. Thus, so long as Υ_1 is an additive subgroup then there exists unitary matrices in $\text{Sym}(\mathcal{C}(\Upsilon_1, \Upsilon_2; \mathcal{R}))$ which act transitively on these bases. However, in the present context, i.e. by looking at codes define through the trace map over the Galois Ring $\text{GR}(p^a, m')$, there is no general analogue to condition (iii) in Theorem 3.3.10. In particular, we show that if one chooses $f = \text{id}$ there is no way to generalize the twisted hamming weight to test for orthogonality along the lines of condition (iii) in Theorem 3.3.10. We note that this observation is quite important in the problem of interest. That is *if one is interested only in minimizing the coherence then sequences define by the trace function is the appropriate choice*. More precisely, reexamining (2.24), one can see that if one wishes to minimize the coherence then one should consider the class of codes defined over a Galois Ring as this framework constrains the number of orthogonal sequences to be small. It is this subtlety which allows us to choose polynomials which provide a good trade off between the coherence properties and the orthogonality properties of a quantizer. In particular, we show that by simply modifying once choice of lift from the finite field indexing the basis one can achieve a desired level orthogonality while keeping the coherence low. Moreover, as the twisted hamming weight was the driving force behind our algorithmic insights into the enumeration of orthogonal bases in Section 3.3.1, in the sequel we focus on how one's choice of lift influences when and how an analogue to the twisted hamming weight may be defined.

To begin, recall that our definition of the twisted hamming weight arose by examining the function $\Gamma_{\mathbb{C}}(\mathbf{a}; \boldsymbol{\beta}, L)$. That is, as the correlation of any two codewords was a function of $\Gamma_{\mathbb{C}}(\mathbf{a}; \boldsymbol{\beta}, L)$ it was sufficient to study the elements of $(\mathbb{Z}_{p^a})^{m'}$ for which $\Gamma_{\mathbb{C}}(\mathbf{a}; \boldsymbol{\beta}, L) = 0$. In the sequel we show how a similar analysis will hold. More precisely, let for any $r \in \text{GR}(p^a, m')$

$$\Gamma_{\mathbb{R}}(r; \tau, \mathbb{K}, f) = \sum_{\bar{z} \in \mathbb{K}} \zeta_{p^a}^{\text{Tr}(r \cdot f(z))}. \quad (3.52)$$

Then, by some simple computation one can see that

$$\mathbf{c}(r, \tau; \mathbb{K}, f)^\dagger \mathbf{c}(s, \tau; \mathbb{K}, f) = \Gamma_{\mathbb{R}}(s - r; \tau, \mathbb{K}, f)$$

so again it is sufficient to understand which elements of $a \in \text{GR}(p^a, m')$ satisfy $\Gamma_{\mathbb{R}}(a; \tau, \mathbb{K}, f) = 0$ to develop a test for orthogonality. In Section 3.3.1 this was achieved by identifying coordinates that were divisible by p^{a-2} , so one could “marginalize” $\Gamma_{\mathbb{C}}(\mathbf{a}; \boldsymbol{\beta}, L)$ and easily see the result was zero. From (3.52) it is clear that we could attempt to do the same in the current context. However, this in general is not fruitful as the lift from \mathbb{K} to $\text{GR}(p^a, m')$ is a cyclic group which does not have an additive structure. Thus, in order to generalize our insights one must find a way to chose \mathbb{K} which allows some similar decomposition. We begin by showing a negative result in this direction by choosing $f = \text{id}$. While this in general will not provided the quantizer of interest it does provide a very fundamental insight in our

development. To begin we provide the following example.

Example 3.4.1 A Second Take at the Kerdock Line Set

We now reexamine a 4-bit quantizer from Example 3.3.4. For this example we now assume that the basis is labeled using $\text{GR}(2^2, 2)$. In particular, we consider the code defined by

1. $\mathcal{I} = \mathcal{T}_{2^2, 2}$
2. $\mathcal{I}_0 = \mathcal{T}_{2^2, 2}$
3. $\Upsilon_1 = \mathcal{T}_{2^2, 2} + 2 \cdot \mathcal{T}_{2^2, 2}$
4. $\Upsilon_2 = \{0\}$

which yields a code containing 16 codewords. To see that this is equivalent to Example 3.3.4 we note

$$\begin{aligned} [\text{Tr}(0 \cdot \zeta^2), \text{Tr}(0 \cdot \zeta^4)] &= [0, 0] \\ [\text{Tr}(\zeta \cdot \zeta^2), \text{Tr}(\zeta \cdot \zeta^4)] &= [1, 0] \\ [\text{Tr}(\zeta^2 \cdot \zeta^2), \text{Tr}(\zeta^2 \cdot \zeta^4)] &= [0, 1] \\ [\text{Tr}(\zeta^3 \cdot \zeta^2), \text{Tr}(\zeta^3 \cdot \zeta^4)] &= [3, 3] \end{aligned}$$

In vector form our definition of the code becomes

1. $\mathcal{I} = \{[0, 0], [0, 1], [1, 0], [3, 3]\}$
2. $\mathcal{I}_0 = \mathcal{I}$
3. $\Upsilon_1 = \mathcal{I} + 2 \cdot \mathcal{I}$
4. $\Upsilon_2 = \{[0, 0]\}$

Thus, once again the only orthogonal bases of this code are the ones satisfying condition (ii) of Theorem 3.3.10.

From Example 3.4.1 it is clear that determining the orthogonality properties of a quantizer, at least in the case where $f = \text{id}$, is more subtle in the case of a trace codes over a Galois Ring than in the case examine in Section 3.3.1. That is while in Section 3.3.1 one could attempt to marginalize $\Gamma_{\mathbb{C}}(\mathbf{a}; \boldsymbol{\beta}, L)$ using *every coordinate* it is not clear that in this case there is any coordinate for which one may marginalize $\Gamma_{\mathbb{R}}(a; \tau, \mathbb{K}, f)$. In this direction we have the following theorem from [76]

Lemma 3.4.9. *Let $p > 1$ be a given prime number and let $a \in \text{GR}(p^2, m')$ for some $m' > 1$. Then,*

$$\Gamma_{\mathbb{R}}(a; \tau, \mathbb{K}, \text{id}) = 0$$

if and only if $a = p \cdot \zeta$ for some $\zeta \in \mathcal{T}_{p^2, m}$.

As the sum $\Gamma_{\mathbb{R}}(a; \tau, \mathbb{K}, \text{id})$ is only zero when $a = p \cdot \zeta$ two vectors $\mathbf{c}(r, \tau; \mathbb{K}, \text{id})$ and $\mathbf{c}(s, \tau; \mathbb{K}, \text{id})$ are only orthogonal on a very limited basis. This is a quite discouraging result as this implies that one has no hope in developing codewords with many orthogonal codewords (with $f = \text{id}$) in this framework for low quantization rate. One may attempt to construct a codebook at higher rate by increasing a in the hopes this provided enough freedom to produce more orthogonal vectors. The following lemma shows that this is not possible in general.

Lemma 3.4.10. *Let p be given and consider a code defined over $\text{GR}(p^2, m')$ for some $m' > 1$. If, for $i \geq 2$,*

$$|\Gamma_{\mathbb{R}}(r'; \tau, \mathbb{K}, \text{id})| > 0$$

for all $r' \in p^{e-i}\text{GR}(p^i, m')$ then

$$|\Gamma_{\mathbb{R}}(r; \tau, \mathbb{K}, \text{id})| > 0$$

for all $r \in p^{e-i-1}\text{GR}(p^{i+1}, m')$

Proof. See Appendix C.2.7 ■

Examining Lemmas 3.4.9 – 3.4.10 one can see that in the case $f = \text{id}$ we can not guarantee that there will be any codewords for which condition (iii) of Theorem 3.3.10 holds. We state this in the following theorem.

Theorem 3.4.11. *Let $\mathbf{c}_1 = \mathbf{c}(\tilde{\lambda}_1, \tau; \mathbb{K}, \text{id})$ and $\mathbf{c}_2 = \mathbf{c}(\tilde{\lambda}_2, \tau'; \mathbb{K}, \text{id})$ be any two codevectors of $\mathcal{C}(\mathbb{K}_a^d, \mathbb{K}^c; \mathbb{K}, \text{id})$. Let $\lambda_i = (\tilde{\lambda}_i \pmod{p^{a-1}}) \cap \mathbb{K}_a^d$ and $\lambda'_i = \tilde{\lambda}_i - \lambda_i$. Then, \mathbf{c}_1 and \mathbf{c}_2 are orthogonal if and only if one of the following holds:*

(i) $\tau' - \tau \notin \mathbb{K}$

(ii) $\lambda'_1 \neq \lambda'_2$ and $\lambda_1 = \lambda_2$

Moreover, every orthogonal basis of \mathcal{C}^m contained in $\mathcal{C}(\mathbb{K}_a^d, \mathbb{K}^c; \mathbb{K}, \text{id})$ has the form

$$\bigcup_{\tau \in \mathbb{K}^c} \mathcal{C}(\lambda_{\tau} + p^{a-1} \cdot \mathbb{K}_1^d, \mathbb{K}^c; \mathbb{K}, \text{id}) \quad (3.53)$$

where λ_{τ} are not necessarily distinct elements of $\downarrow \mathbb{K}_a^d$.

We note that while Theorem 3.4.11 appears quite pessimistic in terms of one's hopes to develop codebooks with a large degree of orthogonality it is in fact far more illuminating than one may expect. Before proceeding in this direction we reiterate a key observation:

If one is interested in only minimize the coherence then sequences define by the trace function is the appropriate choice as it constrains the number of orthogonal sequences to be small¹². If one is interested only in maximizing the number of orthogonal bases contained as subcodes then a quantizer defined over the cross product of integers modulo p^a is the appropriate choice as this provides a large number of orthogonal vectors.

In practical systems one is, more often than not, interested in balancing the objectives of coherence and orthogonality it is natural to consider the question on how one may interpolate between these two extremes. In the sequel we show that Theorem 3.4.11 is far more positive for this broader question than one may expect. In particular, we show in Section 6.2 that the orthogonal bases that satisfied condition (ii) of Theorem 3.3.10 (or in the present context condition (ii) of Theorem 3.4.11) have the most orthogonal bases that satisfy Theorem 3.3.10 (iii) within a given distance (a notion we make more precise in the sequel). That is, the orthogonal bases that satisfy condition (ii) of Theorem 3.3.10 are the easiest to modify to obtain new orthogonal bases. Hence, it is reasonable to expect

¹² Hence by (2.23) the coherence of quantizers from this class should be small

that starting from the present framework over Galois rings will provide a good starting point to understand exactly how one may interpolate between the extreme cases described by Theorem 3.3.10 and Theorem 3.4.11. That is, while codes defined with $f = \text{id}$ have quite good coherence, it appears that one may introduce many orthogonal bases while not dramatically altering the cross correlation spectrum starting from this particular design.

We note that known results on sums of the form $\Gamma_{\mathbb{R}}(a; \tau, \mathbb{K}, f)$ extend far beyond those presented to this point. In fact, they may be extended to some what arbitrary *functions* over $\mathcal{T}_{p^a, m'}$. Indeed, this is the well known extension of the theorem of Weil, Carlitz and Uchiyama. In particular, consider a polynomial over $\text{GR}(p^a, m')$,

$$f(x) = \sum_{i=0}^d a_i x^i$$

of degree where $a_i \in \text{GR}(p^a, m')$. Further, let

$$f(x) = \sum_{i=0}^{a-1} p^i F_i(x)$$

be the corresponding p -adic expansion of $f(x)$ where $F_i(x) \in \mathcal{T}_{p^a, m'}[x]$. We note that such an expansion is always possible by considering the p -adic expansion of the coefficients of $f(x)$, a_i . We say a polynomial is degenerate if the degrees of each polynomial in the p -adic expansion of $f(x)$ is divisible by p . Lastly, let n_j be the degree of the polynomial $F_j(x)$. Then, the weighted degree of the function $f(x)$ is

$$\text{wt}_d(f) \triangleq \max\{n_0 p^{a-1}, n_1 p^{a-2}, \dots, n_{a-1}\}$$

Then, we have the following result from [78].

Proposition 3.4.12. *Let $f(x) \in \text{GR}(p^a, m')[x]$ be a polynomial with weighted degree $\text{wt}_d(f)$ and suppose that the degree of each polynomial in the p -adic expansion is not divisible by p . Then,*

$$\left| \sum_{x \in \mathcal{T}_{p^a, m'}} \zeta_{p^a}^{\text{Tr}(f(x))} \right| \leq (\text{wt}_d(f) - 1) \sqrt{p^{m'}}$$

We note that when $p = 2$ and $a = 2$ one can show that this bound is in fact tight for quadratic functions and hence this provides a tight bound on the Kerdock code of Example 3.4.1. However, we note that this results also has a strong influence our our development of codebooks that have many orthogonal bases. Indeed, as we have seen in Section 2.2.1 if one has a codebook with many orthogonal bases then the bound on the maximal inner product, and hence in the present context $\Gamma_{\mathbb{R}}(a; \tau, \mathbb{K}, \text{id})$, will increase. Thus, while one may not find a code which contains many orthogonal bases with a quadratic function, one may with a function of higher degree. However, our approach taken in Section 3.3.1 employed marginalization of the sum $\Gamma_{\mathbb{C}}(\mathbf{a}; \beta, L)$ to identify orthogonal codewords. Thus, it is of interest to find a polynomial for which we have an identifiable set of coordinates for which to marginalize the sum $\Gamma_{\mathbb{R}}(a; \tau, \mathbb{K}, f)$. In particular, as our previous quantizer from Section 3.3.1 was defined over the cross product of the integers and there was a natural way to break up the sum. At present, we have no such identification as the set $\mathcal{T}_{p^a, m'}$ is

cyclic. However, the set $\mathcal{T}_{p^a, m'}$ is only a small subset of $\text{GR}(p^a, m')$ and one may ask if there is a different subset which will fit our purposes. In particular, for our previous results concerning transitive unitary actions on the code we need a map that will break up the sum $\Gamma_{\mathbb{R}}(a; \tau, \mathbb{K}, f)$ as well as have an image that forms an additive group that may be extended to $\text{GR}(p^a, m')$.

A set that is of particular interest is the unit group of $\text{GR}(p^a, m')$ as it is the direct product of $\mathcal{T}_{p^a, m'}$ as well as additional cyclic groups. In particular, the group of units of $\text{GR}(p^a, m')$ is

$$\text{GR}^*(p^a, m') = \mathbb{Z}_{p^{m'-1}} \times \underbrace{\mathbb{Z}_{p^{a-1}} \times \mathbb{Z}_{p^{a-1}} \times \cdots \times \mathbb{Z}_{p^{a-1}}}_{m' \text{ times}}$$

if $p = 2$ and $a \leq 2$ or $p > 2$ and when $p = 2$ and $a \geq 3$ one has

$$\text{GR}^*(2^a, m') = \mathbb{Z}_{p^{m'-1}} \times \mathbb{Z}_2 \times \underbrace{\mathbb{Z}_{2^{a-1}} \times \mathbb{Z}_{2^{a-1}} \times \cdots \times \mathbb{Z}_{2^{a-1}}}_{m'-1 \text{ times}}$$

Recall, we seek a map for which one may marginalize the sum $\Gamma_{\mathbb{R}}(a; \tau, \mathbb{K}, f)$ as well as forms an additive group that may be extended to $\text{GR}(p^a, m')$. In the context of the unit group it appears, conceptually, that a natural choice for the polynomial f is a function which has its image in $\text{GR}^*(p^a, m')$ a few of the groups isomorphic to $\mathbb{Z}_{p^{a-1}}$ and not exclusively $\mathcal{T}_{p^a, m'}$. However, it is not clear at present how one may do this in a way that the image is an additive group that may be extended to $\text{GR}(p^a, m')$. A natural choice is to chose a polynomial which respects that addition law of the underlying finite field \mathbb{K} . That is, we are interested in a map from $\mathcal{T}_{p^a, m'}$ to $\text{GR}^*(p^a, m')$ which, when reduced modulo p again lays in $\text{GR}^*(p^1, m')$. In particular, we want a sequence of maps

$$\vartheta_A(\zeta^i) : \mathbb{F}_{p^{m'}} \rightarrow \mathcal{T}_{p^a, m'} \rightarrow \text{GR}(p^a, m')$$

where the composite map $\bar{\zeta}^i \rightarrow \vartheta_A(\zeta^i)$ from $\mathbb{F}_{p^{m'}} \rightarrow \text{GR}(p^a, m')$ is injective. Any such map for which $\mu \circ \vartheta_A(\mathcal{T}_{p^a, m'}) = \mathbb{F}_{p^a}$ we say is a *lift* of \mathbb{F}_{p^a} .

The notion of a lift played a key role in our previous development. In particular, it allowed us to use the simple addition of the finite field to describe the permutations to the coordinate set that act as shifts. Hence, we seek “lifts” of \mathbb{F}_{p^a} that again will play this role. A particularly useful map is

$$\vartheta_{\mathcal{I}}(x) = x \prod_{i \in \mathcal{I}} \left(1 + p^{a-1} \zeta^{p^i} \text{Tr} \left(x \zeta^{p^i} \right) \right)$$

It should be clear that such a map satisfies the require criterion. What is less clear is that is also provides our desired interpolation. In particular we have the following theorem.

Theorem 3.4.13. *Let $\mathcal{I} \subset \{0, 1, \dots, m'\}$ for $p = 2$ and $a \leq 2$ or $p > 2$ and $\mathcal{I} \subset \{0, 1, \dots, m' - 1\}$ for $p = 2$ and $a \geq 3$ be given. Then, the map $\vartheta_{\mathcal{I}}(\zeta^j)$ from $\mathcal{T}_{p^a, m'} \setminus \{0\}$ is injective and*

$$\vartheta_{\mathcal{I}}(\zeta^i) \equiv \bar{\zeta}^i \pmod{p}$$

so that $\vartheta_{\mathcal{I}}(\mathcal{T}_{p^a, m'})$ is a lift of $\mathbb{F}_{p^{m'}}$ in $\text{GR}(p^a, m')$. Moreover,

$$\vartheta_{\mathcal{I}}(\zeta^i) : \mathcal{T}_{p^a, m'} \rightarrow \text{GR}^*(p^a, m')$$

and $\vartheta_A(\mathcal{T}_{p^a, m'})$ forms an additive group that may be extended to $\text{GR}(p^a, m')$.

Proof. See Appendix C.2.8. ■

We note that Theorem has a particularly useful consequence in the design of quantizers for the channel-aware scheduling problem. In particular, the map $\vartheta_i(x)$ “unlocks” coordinates which allows us to marginalize the inner product computation and identify orthogonal codewords. This produces a code which has more orthogonality in general than the original trace codes. A code of particular interest chooses $\Upsilon_1 = \text{GR}(p^a, m')$ and $\Upsilon_2 = \{0\}$. In this direction we let

$$\mathcal{C}_{\mathcal{I}}(a, m, \mathcal{I}, h) = \mathcal{C}(\text{GR}(p^a, m'), \{0\}, \mathbb{F}, \vartheta_{\mathcal{I}}(\cdot)) + [\zeta_{p^a}^h, 0, 0, \dots, 0] \quad (3.54)$$

Then in order to identify orthogonal codewords in this new code one may now define a restricted twisted hamming weight for which one may test for orthogonality. In this direction we have the following theorem.

Theorem 3.4.14. *Let p be prime and let $a, m' \in \mathbb{Z}$ be given such that $a > 0$ and $m' > 1$. Further, suppose $\mathcal{I} \subset \{0, 1, \dots, m'\}$ for $p = 2$ and $a \leq 2$ or $p > 2$ and $\mathcal{I} \subset \{0, 1, \dots, m' - 1\}$ for $p = 2$ and $a \geq 3$ is given. Then, for any $y \in \text{GR}(p^a, m')$ if $\mu(y) \in \{0, 1\}$, then*

$$\text{Tr}(y \cdot \vartheta_{\mathcal{I}}(x)) = \sum_{i=0}^{m'-1} x_i \text{Tr}(\hat{y} \cdot \zeta_{\perp}^{p^i}) + p^{a-1} \sum_{i=0}^{m'-1} x_i \left(\text{Tr}(y_{a-1} \cdot \zeta_{\perp}^{p^i}) + \mathbf{1}_{\{i \in \mathcal{I}\}} \cdot x_i \right) \quad (3.55)$$

Proof. See Appendix C.2.9. ■

As a consequence to Theorem 3.4.14 one can see that it is possible to once again marginalize as one may expand any element of $\text{GR}^*(p^a, m')$ into a vector of length m' over \mathbb{Z}_{p^a} by using the representation of the element in terms of the dual basis. That is, one can consider expanding any element of $\text{GR}^*(p^a, m')$, say r , with regard to this basis as

$$\mathbf{r} = [\text{Tr}(r \cdot b_0), \text{Tr}(r \cdot b_1), \dots, \text{Tr}(r \cdot b_{m'-1})].$$

and consider the problem of marginalizing the sum (3.55) as done previously. However, we note that the conditions for the twisted hamming weight have slightly changed. Recall, to define the twisted hamming weight we decomposed any element of $\beta \in (\mathbb{Z}_{p^a})^{m'}$ as $\beta = (\hat{\beta}, \bar{\beta}) \in \downarrow L_a^d \times (p^{a-1} \cdot L_a^d)$ where $\bar{\beta}$ was the component of β divisible by p^{a-1} . Then, in the context of Section 3.3.1, we defined the twisted Hamming weight of an element as

$$\text{twt}_{\text{H}}(\beta) = \left| \left\{ i \mid \hat{\beta}_i = 0 \text{ and } \bar{\beta}_i \neq 0 \right\} \right|.$$

We note that the present context has changed the conditions on $\hat{\mathbf{r}}$ as well as the conditions on $\bar{\mathbf{r}}$ have. Indeed, examining Theorem 3.4.14 one can see that for any index, say j , such that $\hat{\mathbf{r}}_j = 0$ two vectors are orthogonal if *either* $\bar{\mathbf{r}}_j \neq 0$ or $j \in \mathcal{I}$. However, while on one had this have gotten better in terms of the flexibility one has in constructing bases in terms of $\bar{\mathbf{r}}$ things have gotten worse in terms of the constraints of $\hat{\mathbf{r}}$. Indeed, for the conditions of the theorem to hold one must have $\mu(r) \in \{0, 1\}$. Thus, as a natural extension to the twisted weight we let

$$\text{twt}_{\mathcal{I}}(r) = \mathbf{1}_{\{\mu(r) \in \{0, 1\}\}} \cdot \left| \left\{ 0 < i < m' - 1 \mid \hat{\mathbf{r}}_i = 0 \text{ and } (\bar{\mathbf{r}}_i \neq 0 \text{ or } i \in \mathcal{I}) \right\} \right| \quad (3.56)$$

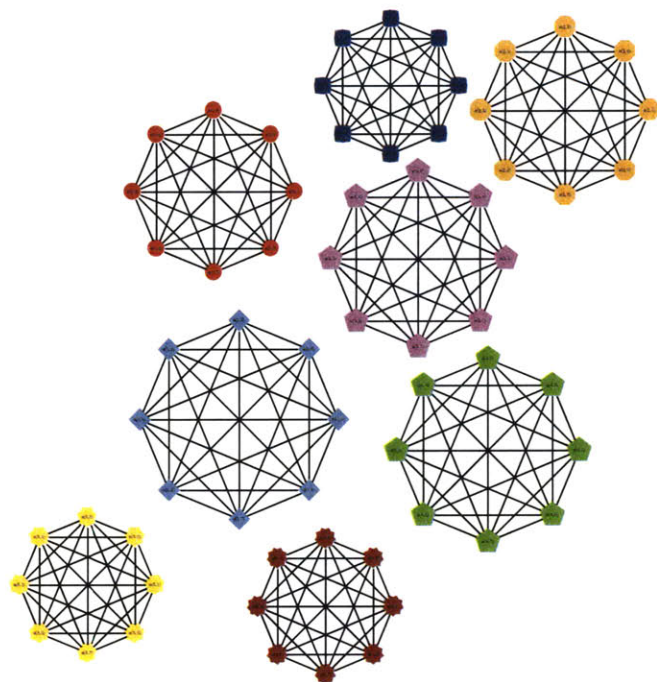
We note that with this approach we have not only our former insights gained through the twisted hamming weight, but also through Proposition 3.4.12 which illustrates that the inner product between two vectors can not grow too fast. As a brief illustration of the generalized switches formed using this approach can be seen in Figure 3-24. The resulting cross correlation spectrum can be seen in Figure 3-25.

We note that with this definition hand one can proceed to extend all the results one had for codes over the cross product of the integers. In particular, one has a natural generalization of Theorem 3.3.10, now using the restricted twisted hamming weight $\text{twt}_{\mathcal{I}}(\mathbf{r})$. Moreover, all of the subsequent discussion and theorems follow directly with the $\text{twt}_{\mathcal{H}}(\mathbf{r})$ replaced by $\text{twt}_{\mathcal{I}}(\mathbf{r})$. In particular, one can again show that $\text{Sym}(\mathcal{C})$ contains unitary matrices that act two transitively on the all orthogonal bases contained in a code using the lift $\vartheta_{\mathcal{I}}(x)$.

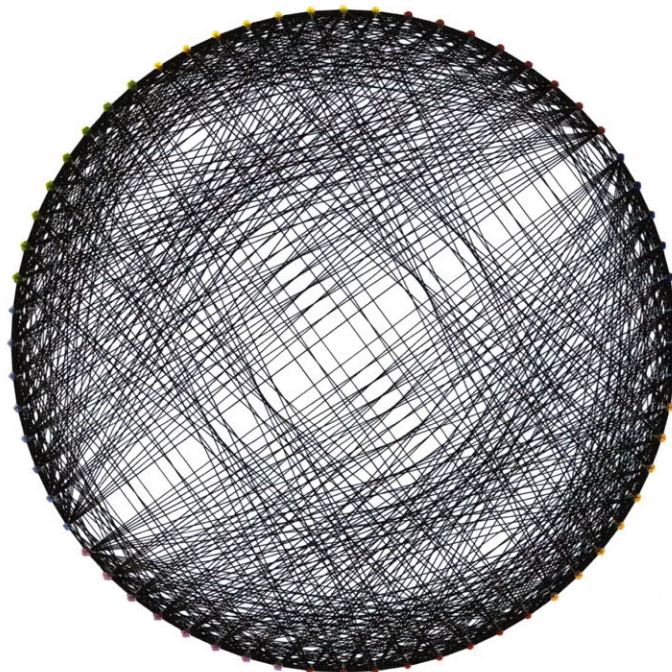
As the map allows one to marginalize over a coordinate set as well forms an additive group which may be extended to $\text{GR}(p^a, m')$ one may for any appropriate choice of \mathcal{I} use this quantizer in conjunction with our previous theorems to see that there exists a large symmetry group, hence leading to low mean squared error. In particular, the present observation to the twisted hamming weight has direct applications to the tradeoff between orthogonality and coherence which has great importance in developing quantizers and associated algorithms which identify users with low co-channel interference. Indeed, in our present framework one may obtain the quantizers with the best mean squared quantization error, i.e. one may choose $\vartheta_{\emptyset}(x)$, if only the mean squared quantization error is of interest. However, in order to maximize this figure of merit one must exclude relations that led to many orthonormal bases and the resulting code only contains a disjoint union of orthogonal bases. Hence, quantizers which use $\vartheta_{\{i\}}(x)$ seem like a natural choice for use in a multi-user MIMO systems as they admit quite a few orthogonal bases with a minimal effect on the mean squared quantization error. The cross correlation spectrum, which relates to the shape of the Voronoi region and the MSE, may be seen in Figure 3-25. Moreover, as seen in Figure 3-24 they yield quite regular structures which reflects the large group of transitive unitary transformations contained in $\text{Sym}(\mathcal{C})$. However, this construction will only yield constructions of size $p^{a \cdot m'}$. As we have seen, in general, codes which increase the quantization rate by increasing the size of the underlying ring do not perform well. Hence, to have a truly systematic approach to channel quantization one must have additional ways to increase the quantization rate. We now turn to this final problem; constructing a quantization framework which allows one to increase the quantization rate with out modifying the cardinality of the underlying ring.

■ 3.5 Component Codes at Intermediate Rates

In Section 3.4 we developed the function $\vartheta_{\mathcal{I}}(x)$ to interpolate between the competing design objectives of orthogonality as well as mean squared error. As previously noted inside this framework the only way one could increase the code rate was to increase the cardinality of the base ring which was shown to yield poor performance. In this section our goal is two fold. First, we develop methods to increase the rate of the code by developing a class of functions which may be paired with $\vartheta_{\mathcal{I}}(x)$ to yield codes of higher rates. Second, we develop how one may choose these functions so that the resulting codes are invariant to shifts in the coordinate set yielding a system of codes that may be paired with sparser codes to construct good high rate quantizers with large symmetry groups. As such, throughout this section we consider the design of dense codes. One may then develop sparse codes using



(a)



(b)

Figure 3-24. An example of the orthogonality relations between codewords of the quantizer developed using the lift $\vartheta_{\mathcal{I}}(x)$ in 8 complex dimensions. The generalized switch for the orthogonal processing modes for (a) $\vartheta_{\emptyset}(x)$ and (b) $\vartheta_{\{1,2\}}(x)$. Note that the code represented in (a) only has non-intersecting orthogonal bases which correspond to condition (ii) in Theorem 3.4.11. However, using the map $\vartheta_{\{1,2\}}(x)$ allows on to marginalize over a set of coordinates leading to more bases which have intersection.

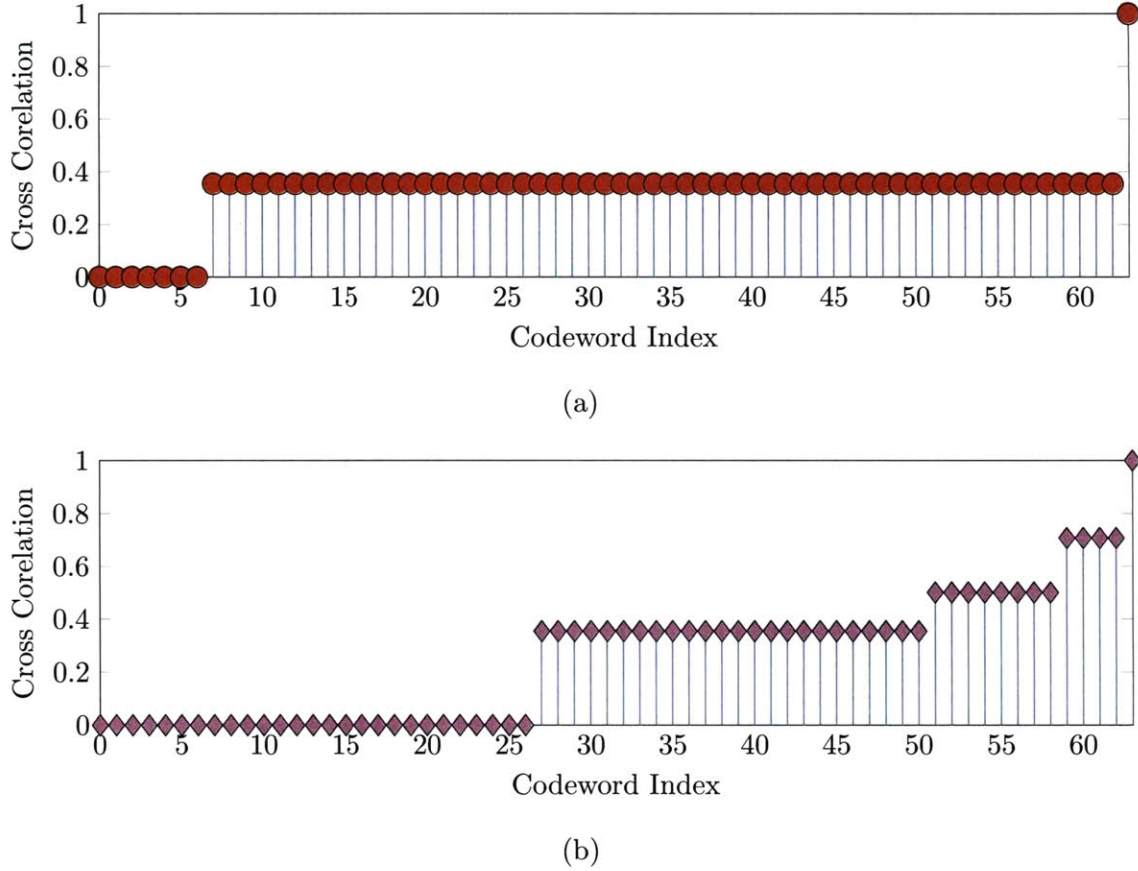


Figure 3-25. An example of the cross correlation spectrum of the quantizer developed using the lift $\vartheta_{\mathcal{I}}(x)$ in 8 complex dimensions. The cross correlation spectrum for any codeword of the quantizer for (a) $\vartheta_{\emptyset}(x)$ and (b) $\vartheta_{\{1,2\}}(x)$. Note that while the code in (b) contains many more orthogonal bases it suffers from a more irregular cross correlation spectrum and hence has a higher mean square error than the code in (a).

these results in a lower dimension using the methods of Section 3.2.

Recall that in our development in Section 3.3.1 we used the structure of a bilinear map (the inner product) to understand the structure of the symmetry group associated with a quantizer in our framework. In order to extend these results we must find other such maps. In this section we begin by using some classically known results from linear codes over finite fields to achieve this goal. Then, using some more contemporary results provide these results in general. To begin, recall that the trace map is a linear map from \mathbb{F}_{p^m} to \mathbb{Z}_p . In fact, every linear map from \mathbb{F}_{p^m} to \mathbb{Z}_p is of the form $\text{Tr}_{\mathbb{F}_{p^m}/\mathbb{Z}_p}(\alpha \cdot x)$ for some $\alpha \in \mathbb{F}_{p^m}$ [106]. Thus, using functions of the form $\text{Tr}_{\mathbb{F}_{p^m}/\mathbb{Z}_p}(y \cdot x)$ one may develop a set of bilinear functions to use in our constructions. However, to develop dense codes which are invariant to shifts we require a larger set of maps.

We note that all of the codes used in the sequel are (affine invariant) extended cyclic codes over integer rings. In particular, we consider extended cyclic codes over the integer ring \mathbb{Z}_{p^a} . We note that these codes have been shown to yield linear representations for some notoriously non-linear codes. In particular, Forney, Sloane and Trott have shown that the Nordstrom-Robinson code is the binary image of the octacode which is a linear code over \mathbb{Z}_4 of length 8 [50]. Later, Hammons et. al. showed that the non-linear binary Kerdock code

may be constructed as the image of a linear code over \mathbb{Z}_4 [55]. Further work has shown that codes over \mathbb{Z}_4 can be used to develop optimal sets of lines in complex spaces [30]. Subsequent generalizations of this work have led to a quite general framework in which one can succinctly describe some of the densest known sphere and lattice packings [93]. We provide a brief introduction to linear codes over integer rings in Appendix A and in the sequel present a representation for these codes that is compatible with our general framework of component codes from Section 3.3.

In order to develop a large family of bilinear maps we begin by identifying the set of quadratic maps from \mathbb{F}_{p^m} to \mathbb{Z}_p , then, using the relation between bilinear and quadratic maps developed in Section 3.4 identify a family of bilinear maps. In this direction, recall from Section 3.4 that

$$\mathrm{Tr}_{\mathbb{F}_{p^m}/\mathbb{Z}_p}(\alpha \cdot x^2)$$

is an example of a quadratic map from \mathbb{F}_{p^m} to \mathbb{Z}_p . As this was the base of our previous construction, it is of interest to identify additional maps with this structure in order to produce higher rate codes. Recall from Section 3.4.1 that the Frobenius automorphism $\phi : x \rightarrow x^p$, as well all of its powers, are linear over \mathbb{F}_{p^m} . Thus for any $\alpha \in \mathbb{F}_{p^m}$, if $d = p^j + p^k$, the function

$$\mathrm{Tr}_{\mathbb{F}_{p^m}/\mathbb{Z}_p}(\alpha \cdot x^d) \tag{3.57}$$

is a quadratic map [86] from \mathbb{F}_{p^m} to \mathbb{Z}_p as $x^{p^j} \cdot x^{p^k}$ is the product of linear functions. It is again natural to ask whether functions of the form (3.57) are the only quadratic functions from \mathbb{F}_{p^m} to \mathbb{Z}_p for some $d = p^j + p^k$. This is indeed true. However, as

$$\mathrm{Tr}_{\mathbb{F}_{p^m}/\mathbb{Z}_p}(x) = \mathrm{Tr}_{\mathbb{F}_{p^m}/\mathbb{Z}_p}(\phi(x)) = \mathrm{Tr}_{\mathbb{F}_{p^m}/\mathbb{Z}_p}(x^p) \tag{3.58}$$

one must take care when forming a system of maps as to not include functions that led to redundant codewords. Thus, one must form an equivalence between the functions $\{x^0, x^1, \dots, x^{p^{m'}-2}\}$ which define the same function under the trace map. Under the correspondence (3.58) there will be a corresponding partition of $\{0, 1, 2, \dots, p^{m'}-2\}$. We call this partition of $\{0, 1, 2, \dots, p^{m'}-2\}$ the p -cyclotomic cosets modulo $p^{m'}-1$. More precisely, the partition of $\{0, 1, 2, \dots, p^{m'}-2\}$, say $\mathcal{P} = \{\mathcal{P}_0, \mathcal{P}_1, \dots, \mathcal{P}_{|\mathcal{P}|-1}\}$, are the p -cyclotomic cosets modulo $p^{m'}-1$ if the following hold:

$$\bigcup_{i=0}^{|\mathcal{P}|-1} \mathcal{P}_i = \{0, 1, 2, \dots, m-2\} \text{ and } \mathcal{P}_i \cap \mathcal{P}_j = \emptyset \text{ for } i \neq j \tag{3.59}$$

$$\mathcal{P}_i = \left\{ s \cdot p^{jm'} \pmod{m-1} \mid 0 \leq j < m_s \right\} \text{ if } s \in \mathcal{P}_i \tag{3.60}$$

where in turn m_s is the smallest positive integer such that

$$s \cdot p^{m'm_s} \equiv s \pmod{m-1}.$$

We identify each coset $\mathcal{P}_i \in \mathcal{P}$ by its smallest element and call this element of \mathcal{P}_i the coset leader. We denote the set of all coset leaders of \mathcal{P} as $\mathcal{I}_{\mathcal{P}}$. This leads to the following theorem [81].

Theorem 3.5.1. *Every function $\tilde{f}(z)$ from \mathbb{F}_{p^m} to \mathbb{Z}_p can be written uniquely as*

$$\tilde{f}(z) = \sum_{i \in \mathcal{I}_P} \text{Tr}_{\mathbb{F}_{p^{m_s}}/\mathbb{Z}_p} (a_i z^i) + a_{m-1} z^{m-1} \quad (3.61)$$

where $a_i \in \mathbb{F}_{p^{m_s}}$, $a_{m-1} \in \mathbb{Z}_p$ and \mathcal{I}_P is the set of coset leaders of the p -cyclotomic partition of $\{0, 1, 2, \dots, m-2\}$.

Examining Theorem 3.5.1 one can see that every function from \mathbb{F}_{p^m} to \mathbb{Z}_p is indexed by elements of the coset leaders of the p -cyclotomic partition modulo m and elements of $\mathbb{F}_{p^{m'}}$. Thus, to systematically design quantizer over finite fields it is sufficient to optimize over this set and the coefficients of the polynomial to design a quantizer. Moreover, by Theorem 3.5.1 and our previous discussion every quadratic map is a linear combination of functions of the form $\text{Tr}_{\mathbb{F}_{p^{m_s}}/\mathbb{Z}_p} (a_i z^s)$ where $s = 1 + p^j$ for some $0 \leq j < m'$ and $a_i \in \mathbb{F}_{p^{m_s}}$. In this direction, let

$$\mathcal{D}_p(m-1) = \left(\bigcup_{j=0}^{m'-1} \{p^j\} \cup \bigcup_{j=0}^{m'-1} \{1+p^j\} \right) \cap \mathcal{I}_P. \quad (3.62)$$

Then, every quadratic function from $\mathbb{F}_{p^{m'}}$ to \mathbb{Z}_p is of the form

$$\tilde{f}_Q(z) = \sum_{i \in \mathcal{D}_p(m-1)} \text{Tr}_{\mathbb{F}_{p^{m_s}}/\mathbb{Z}_p} (a_i z^i) + a_{m-1} z^{m-1}.$$

While the set $\mathcal{D}_p(m-1)$ yields a good set of function to use to develop quantizers, as they yield bilinear functions, there is no guarantee that any arbitrarily chosen set of quadratic functions will yield a set of codewords that are not colinear. Recall in Section 3.3.1 we were able to determine when a code had colinear codewords by examining the actions of the operator $T(\lambda)$ and $S(\beta)$ on codewords which, in turn, relied on the linearity of the inner product. Thus, if we are to consider a system of multiple functions it is natural to expect that we need the system of functions to be closed under addition. Hence, we need a notion of linear independence of a set of functions from \mathbb{F}_{p^m} to \mathbb{Z}_p to achieve this goal.

To begin, recall that we have frequently used the fact that the trace map is a linear function. Thus, the set of functions

$$\{\text{Tr}_{\mathbb{F}_{p^{m_s}}/\mathbb{Z}_p} (a_i z^s)\}_{i=0}^{m_s}$$

are linearly independent if and only if the set $\{a_i\}_{i=0}^{m_s}$ are linearly independent elements in $\mathbb{F}_{p^{m_s}}$ when viewed as a vector space over \mathbb{Z}_p . Alternatively, from Theorem 3.5.1 the functions

$$\{\text{Tr}_{\mathbb{F}_{p^{m_s}}/\mathbb{Z}_p} (a_s z^s)\}_{s \in \mathcal{I}_P}$$

are linearly independent for any choice of $a_s \in \mathbb{F}_{p^{m_s}}$. Thus, one may, for any subset $\mathcal{S} \subset \{0, 1, 2, \dots, m-1\}$ and collection of linearly independent elements $A = \{\{a_{i,s}\}_{i=0}^{m_s-1}\}_{s \in \mathcal{S}}$, form a high rate code by first forming the linear set of functions

$$\mathcal{F}(\mathcal{S}) = \langle \text{Tr}_{\mathbb{F}_{p^{m_s}}/\mathbb{Z}_p} (a_{i,s} z^s) \mid s \in \mathcal{S} \cap \mathcal{I}_P \text{ and } a_{i,s} \in A \rangle.$$

Then, for $f \in \mathcal{F}(\mathcal{S})$, one may construct a code containing the codewords

$$\mathbf{c}(f) = \sum_{\bar{z} \in \mathbb{F}} \zeta^{f(z)} \mathbf{e}_z. \quad (3.63)$$

It is clear from (3.63) that two codewords $\mathbf{c}(f_1)$ and $\mathbf{c}(f_2)$ are colinear if and only if

$$f_1(z) - f_2(z) = a \quad \forall z \in \mathbb{F}$$

for some $a \in \mathbb{Z}_p$. Thus, a set of functions $\mathcal{F}(\mathcal{S})$ defines a code with colinear lines if and only if $0 \in \mathcal{S}$ and it is a simple process to develop high rate codes over finite fields as one may optimize over subsets of $\{1, 2, \dots, m-1\}$. We note, however, this construction is a bit distant from our preceding development. That is, in Section 3.2 our systematic construction of codes consisted of:

1. \mathcal{I} , the vector space $(\mathbb{Z}_p)^{m'}$
2. L , a sub-space of $(\mathbb{Z}_p)^{m'}$
3. Υ_1 , a subset, $(\mathbb{Z}_{p^a})^{m'}$ which describes the base quantizer \mathcal{C}_0
4. Υ_2 , a subset of $(\mathbb{Z}_p)^{m'}$ which describes the “shifts” of L (i.e. the coordinate permutations to be applied to \mathcal{C}_0)
5. the function $c(\lambda, \bar{\beta}) = \zeta_{p^a}^{\langle \lambda, \bar{\beta} \rangle}$

In Section 3.4 we further developed this framework to allow one to choose Υ_1 to be a subring of $\text{GR}(p^a, m')$ and $c(\lambda, \bar{\beta}) = \zeta_{p^a}^{\text{Tr}(\lambda \cdot f(\beta))}$. Thus, it is a far more natural setting to consider extending our codes by considering the addition of polynomials of the form z^s rather than directly applying the definition (3.63) as this construction has no explicit connection to the underlying group of symmetries defined by the operators $T(\lambda)$ and $S(\beta)$. Hence, we rather consider the codes¹³

$$\mathcal{C}(\Upsilon_1, \Upsilon_2; \mathbb{F}, \mathcal{T}) = \bigcup_{\mathcal{S} \subset \mathcal{T}^\perp \cap \mathcal{I}_p} \bigcup_{\lambda, \Upsilon_1} \mathbf{c}(\lambda, 0; \mathbb{F}, \mathcal{S})$$

where

$$\mathbf{c}(\lambda, 0; \mathbb{F}, \mathcal{S}) = \sum_{\bar{z} \in \mathbb{F}} \zeta^{\text{Tr}(\lambda \cdot \sum_{s \in \mathcal{S}} z^s)}.$$

and in turn where

$$\mathcal{T}^\perp = \{s \in [0, m-1] : m-1-s \notin \mathcal{T}\}. \quad (3.64)$$

Then, in the special case over finite fields we have the important theorem as a direct corollary to [22]

Theorem 3.5.2. *Let $\mathcal{T} \subset [0, m-1]$ be given. Then,*

$$S(\beta) \cdot \mathcal{C}(\mathbb{F}, \{0\}; \mathbb{F}, \mathcal{T}) = \mathcal{C}(\mathbb{F}, \{0\}; \mathbb{F}, \mathcal{T})$$

for every $\beta \in \mathbb{F}$ if and only if

$$s = \sum_{i=0}^{m'-1} s_i p^i \in \mathcal{T}$$

¹³We note that the appearance of the set \mathcal{T}^\perp comes from historical developments in cyclic codes and helps identify structure in the sequel.

then

$$s_j > 0 \implies s - p^j \in \mathcal{T}.$$

Note that nowhere in Theorem 3.5.2 is there an explicit requirement that any of the functions associated to $\mathcal{C}(\mathbb{F}, \{0\}; \mathbb{F}, \mathcal{T})$ be quadratic or for that matter bilinear. This is important as it removes the bilinear constraint explicitly from our development of quantizers. That is, by construction the code $\mathcal{C}(\mathbb{F}, \{0\}; \mathbb{F}, \mathcal{T})$ has an associated set of functions which are linear in Υ_1 and hence invariant to multiplication by $T(\lambda)$. More precisely, as the codewords in $\mathcal{C}(\mathbb{F}, \{0\}; \mathbb{F}, \mathcal{T})$ are all defined through a linear map, one has that

$$T(\lambda) \cdot \mathcal{C}(\mathbb{F}, \{0\}; \mathbb{F}, \mathcal{T}) = \mathcal{C}(\mathbb{F}, \{0\}; \mathbb{F}, \mathcal{T}) \quad \forall \lambda \in \mathbb{F}.$$

Further, by appropriately choosing \mathcal{T} one can ensure the resulting code is invariant to multiplication by $S(\beta)$ through Theorem 3.5.2. Thus, in order to systematically design sparse and dense codes over finite fields it is sufficient to choose codes $\mathcal{C}(\mathbb{F}, \{0\}; \mathbb{F}, \mathcal{T})$ which satisfy Theorem 3.5.2.

A particularly well known example of codes which satisfy Theorem 3.5.2 are the Reed Muller codes. In this direction, recall that any integer, say s , has for a prime p , a unique p -adic expansion

$$s = \sum_{j \geq 0} s_j p^j$$

where $0 \leq s_j \leq p - 1$. We let the p -weight of any integer s , denoted $\text{wt}_p(s)$, be the sum of the coefficients in the p -adic expansion of s . That is,

$$\text{wt}_p(s) = \sum_{j \geq 0} s_j \quad \text{where} \quad s = \sum_{j \geq 0} s_j p^j.$$

Then, the set

$$\mathcal{T}_{\text{RM}}(r) = \{s : \text{wt}_p(s) < m'(p - 1) - r\}.$$

defines the Reed Muller codes and we have the important corollary to Theorem 3.5.2 [43].

Corollary 3.5.3. *Let r be given. Then,*

$$S(\beta) \cdot \mathcal{C}(\mathbb{F}, \{0\}; \mathbb{F}, \mathcal{T}_{\text{RM}}(r)) = \mathcal{C}(\mathbb{F}, \{0\}; \mathbb{F}, \mathcal{T}_{\text{RM}}(r))$$

for every $\beta \in \mathbb{F}$.

We note that Corollary 3.5.3 is quite important to our development in the sequel as it provides a specific construction in the case the underlying ring is a finite field. In the sequel we provide a similar result over more general rings. That is, we develop the necessary extensions to Theorem 3.5.1 and Theorem 3.5.2 that allow explicit constructions of codes that may be used in our systematic construction over more general rings. A natural question is whether the code $\mathcal{C}(\mathbb{F}, \{0\}; \mathbb{F}, \mathcal{T}_{\text{RM}}(r))$, extended using the natural lift from $\mathbb{F}_{p^{m'}}$ to $\text{GR}(p^a, m')$, may be used in our systematic construction over $\text{GR}(p^a, m')$. In general this may not be done. However, in larger rings there is a plurality of codes that do have the required invariance properties that may be used in our systematic construction. Moreover, these codes exist at a variety of rates allowing one to design codes which meet specific rate targets rather than being tied to a specific rate. To begin, we first state the generalization of Theorem 3.5.1 from [24, 25].

Theorem 3.5.4. *Let f be a linear function from $\mathcal{T}_{p^a, m}$ to \mathbb{Z}_{p^a} . Then, f can be uniquely written as*

$$f(x) = \sum_{i \in I_{\mathbb{P}}} \text{Tr}_{\text{GR}(p^a, m_s)/\mathbb{Z}_{p^a}}(a_i x^i) + a_{m-1} x^{m-1} \quad (3.65)$$

where $a_i \in \text{GR}(p^a, m_s)$ and $a_{m-1} \in \mathbb{Z}_{p^a}$ and \mathbb{P} is the p -cyclotomic partition of $\{0, 1, 2, \dots, m-2\}$ modulo $m-1$.

Due to the similar structure Theorem 3.5.4 has to Theorem 3.5.1 one may be tempted to apply Theorem 3.5.2 to again characterize when $S(\beta)$ acts invariantly on a code. However, as previously noted, this may not be done in general as while the conditions of Theorem 3.5.2 are necessary to ensure that any code over $\text{GR}(p^a, m')$ is invariant to multiplication by $S(\beta)$ it is far from sufficient. In particular, as every element of $\text{GR}(p^a, m')$ has a p -adic expansion, one may more generally write any function form $\mathcal{T}_{p^a, m}$ to \mathbb{Z}_{p^a} as

$$f(x) = \sum_{i=0}^{a-1} p^i f_i(x).$$

Thus, in this context, every function from $\mathcal{T}_{p^a, m}$ to \mathbb{Z}_{p^a} is rather defined by a *set of functions* $\{f_i(x)\}_{i=0}^{a-1}$ from $\mathcal{T}_{p^a, m}$ to $p^i \cdot \mathbb{Z}_{p^a}$ and one may more generally describe functions from $\mathcal{T}_{p^a, m}$ to \mathbb{Z}_{p^a} using a subsets of $\{0, 1, \dots, m-1\}$. In this direction, we say that the subsets $\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_a$ are the *defining sets* of code over \mathbb{Z}_{p^a} if

$$\{0\} \subseteq \mathcal{T}_a \subseteq \mathcal{T}_{a-1} \subseteq \dots \subseteq \mathcal{T}_1 \subseteq \{0, 1, \dots, m-1\}. \quad (3.66)$$

We note that the nesting of the sets in (3.66) results in our requirement that the associated set of functions be linear. Hence, analogous to (3.64), we let

$$\mathcal{T}_{i-1}^\perp = \{s \in [0, m-1] : m-1-s \notin \mathcal{T}_{a-i+1}\}$$

where $\mathcal{T}_0 = \{0, 1, \dots, m-1\}$. Then, for any $\Upsilon_1 \subset \text{GR}(p^a, m')$ we let

$$\mathcal{C}(\Upsilon_1, \{0\}; \mathbb{F}, \{\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_a\}) = \bigcup_{S_1 \subset \mathcal{T}_1^\perp} \bigcup_{S_2 \subset \mathcal{T}_2^\perp} \dots \bigcup_{S_a \subset \mathcal{T}_a^\perp} \bigcup_{\lambda, \Upsilon_1} \mathbf{c}(\lambda, 0; \mathbb{F}, \{\mathcal{S}\}_{i=1}^a)$$

where

$$\mathbf{c}(\lambda, 0; \mathbb{F}, \{\mathcal{S}\}_{i=1}^a) = \sum_{z \in \mathbb{F}} \zeta^{\text{Tr}(\lambda \cdot f_{\mathcal{T}}(z; \{\mathcal{S}\}_{i=1}^a))}$$

and in turn where

$$f_{\mathcal{T}}(z; \{\mathcal{S}\}_{i=1}^a) = \sum_{i=1}^a p^{i-1} \sum_{s \in \mathcal{S}_{a-i+1}} z^s.$$

Due to the plurality of sets which define the code $\mathcal{C}(\Upsilon_1, \{0\}; \mathbb{F}, \{\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_a\})$ there is a corresponding plurality of codes that are invariant to multiplication by $S(\beta)$. Thus, we would like to understand how to optimally choose the sets $\{\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_a\}$ to ensure that the corresponding code has as large a symmetry group as possible. In this direction, we note that the simple permutation described by the “shifts” are not the largest symmetry group a code may have. In general there may be much larger groups of permutations that act invariantly on the code. In this direction, let, for $m' = kt$, $\text{AGL}_k(p^t)$ be the set of all affine linear transformations on the finite field $\mathbb{F}_{p^{m'}}$ when viewed as a k dimensional vectors

space over \mathbb{F}_{p^t} . More precisely, for any element $\bar{z} \in \mathbb{F}_{p^{m'}}$ an affine transformation of \mathbb{F}_{p^t} , $(A_k, b_k) \in \text{AGL}_k(p^t)$, is such that

$$\begin{aligned} (A_k, b_k) &: \mathbb{F}_{p^{m'}} \rightarrow \mathbb{F}_{p^{m'}} \\ (A_k, b_k) &: z \rightarrow A_k \cdot z + b_k. \end{aligned}$$

Then, we say that the code \mathcal{C} is invariant under the group $\text{AGL}_k(p^t)$ if $\text{AGL}_k(p^t)$, acting on the coordinates of \mathcal{C} , fixes the code \mathcal{C} . Thus, we let

$$S(A_k, b_k) \cdot \mathbf{e}_z = \mathbf{e}_{A_k \cdot z + b_k}$$

and a code \mathcal{C} is invariant under the group $\text{AGL}_k(p^t)$ if

$$S(A_k, b_k) \cdot \mathcal{C} = \mathcal{C}.$$

Clearly, with this notation

$$S(\beta) = S(\mathbf{I}_1, \beta) \in \text{AGL}_k(p^t) \quad \forall k | m'.$$

More generally one has $\text{AGL}_k(p^{tk}) \subseteq \text{AGL}_\ell(p^{t\ell})$ if $k | \ell | m'$. Hence, $S(\beta) \in \text{AGL}_k(p^t)$ for any k and our previous results will hold if we can show that a code is invariant to any affine group of linear transformations $\text{AGL}_k(p^t)$. In this direction, we have the following theorem from [3]

Theorem 3.5.5. *Let $T_a \subset T_{a-1} \subset \dots \subset T_1 \subset [0, m-1]$ be given. Then,*

$$S(A_k, b_k) \cdot \mathcal{C}(\Upsilon_1, \{0\}; \mathbb{F}, \{\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_a\}) = \mathcal{C}(\Upsilon_1, \{0\}; \mathbb{F}, \{\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_a\})$$

for all $(A_k, b_k) \in \text{AGL}_k(p^t)$, for $tk = m'$, if and only if the following four properties hold:

- (i) If, for $d = 1, 2, \dots, a$, $s \in T_d$, $s_j > 0$, then $s - p^j \in T_d$
- (ii) If, for $d = 1, 2, \dots, a$, $s \in T_d$, $s_j > 0$, then $s - p^j + p^{j+tl} \pmod{m-1} \in T_d$ for $l = 0, 1, \dots, m' - 1$
- (iii) If, for $d = 2, \dots, a$, $s_j > 0$, then $s - p^j + p^{j-1} \cdot (p^{tl_1} + p^{tl_2} + \dots + p^{tl_p}) \pmod{m-1} \in T_d$ for any l_1, l_2, \dots, l_p with $0 \leq l_i \leq m' - 1$
- (iv) If, for $d = 1, \dots, a$, $s \in T_d$, $s_j = s_{j+1} = \dots = s_{j+a-1} = 0$, $s_{j+a} > 0$ and $d > a > 0$, then $s - p^j \pmod{m-1} \in T_{d-a}$

where every subscript is taken mod m' .

Examining Theorem 3.5.5 one can see that we now have a system of constraints for our systematic construction of codes. That is, one now has a precise characterization of “good” codes to use in the systematic construction (3.12). To systematically choose both sparse and dense codes one may simply search over nested subsets $\{\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_a\}$ of $\{0, 1, \dots, m-1\}$ and use Theorem 3.5.5 as a certificate as to whether or not the resulting code will lead to large symmetry groups. However, to complete this systematic construction, we must be able to identify the rate of the associated code. In this direction we have the following lemma from [24].

Lemma 3.5.6. *Let $\{\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_a\}$ be the defining sets of a code over \mathbb{Z}_{p^a} of length m . Then,*

$$\log_p |\mathcal{C}(\Upsilon_1, \{0\}; \mathbb{F}, \{\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_a\})| = a \cdot m' - \sum_{i=1}^a |\mathcal{T}_i|.$$

This result is particularly useful as it yields the last constraint needed to systematically design good dense and sparse codes. That is, one may always design a good rate r dense code by solving the discrete optimization problem in \mathbb{C}^m :

$$\begin{aligned} & \underset{a, \{\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_a\}}{\text{maximize}} && \text{SINR}_{\text{sat}}(\mathcal{C}(\Upsilon_1, \{0\}; \mathbb{F}, \{\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_a\})) \\ & \text{subject to} && a \cdot m' - \sum_{i=1}^a |\mathcal{T}_i| = r \end{aligned} \quad (3.67a)$$

$$\mathcal{T}_a \subseteq \mathcal{T}_{a-1} \subseteq \dots \subseteq \mathcal{T}_1 \subseteq \{0, 1, 2, \dots, m-1\} \quad (3.67b)$$

$$\{\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_a\} \text{ satisfy Theorem 3.5.5} \quad (3.67c)$$

To design a system of good sparse and dense codes one may optimize over the rate, value of a and defining sets of each component of the code yielding a much larger optimization problem. In general one would prefer to dispense with the optimization problem (3.67) as much as possible and rather consider a subclass of defining sets that will work well. In this direction, recall that the Reed Muller codes played a key role in the case of finite fields. Hence, in light of Theorem 3.5.5 one may generalize¹⁴ the Reed Muller code by considering defining sets which use differing values of r for each set \mathcal{T}_i [24]. That is, we let

$$T_{\text{GRM}}(r_1, r_2, \dots, r_a) = \{\mathcal{T}_{\text{RM}}(r_1), \mathcal{T}_{\text{RM}}(r_2), \dots, \mathcal{T}_{\text{RM}}(r_a)\}$$

Then, we have the following corollary from [5, 46].

Corollary 3.5.7. *Let $T_{\text{GRM}}(r_1, r_2, \dots, r_a)$ be given. If, for $p = 2$, $i = 2, \dots, a-1$ and $\ell = 1, \dots, i-1$,*

$$r_{i-\ell} \leq 2^{\ell-1}(m' - r_i),$$

then

$$S(A_1, b_1) \cdot \mathcal{C}(\Upsilon_1, \{0\}; \mathbb{F}, T_{\text{GRM}}(r_1, r_2, \dots, r_a)) = \mathcal{C}(\Upsilon_1, \{0\}; \mathbb{F}, T_{\text{GRM}}(r_1, r_2, \dots, r_a))$$

for all $(A_1, b_1) \in \text{AGL}_1(p^{m'})$. Moreover, for any prime p let the following two conditions hold:

(i) If $0 < r_i \leq (m' - 1)(p - 1) - 1$, then $r_{i+1} > r_i + (p - 1)$

(ii) If $r_i = (m' - 1)(p - 1)$ then $r_{i+1} = (m' - 1)(p - 1)$

Then,

$$S(A_{m'}, b_{m'}) \cdot \mathcal{C}(\Upsilon_1, \{0\}; \mathbb{F}, T_{\text{GRM}}(r_1, r_2, \dots, r_a)) = \mathcal{C}(\Upsilon_1, \{0\}; \mathbb{F}, T_{\text{GRM}}(r_1, r_2, \dots, r_a))$$

for all $(A_{m'}, b_{m'}) \in \text{AGL}_{m'}(p)$.

¹⁴We note that this generalization of the Reed Muller codes is over an integer ring and is not the generalized Reed Muller codes of [45].

To fully illuminate the usefulness of Corollary 3.5.7 we show how it may be used to arrive at our systematic construction of quantizers from Example 3.2.6.

Example 3.5.1 Systematic Constructions For \mathbb{C}^4

In this example we show how one may arrive at the systematic construction of Example 3.2.6 through application of Corollary 3.5.7. We begin by noting that each one of the codes used in Example 3.2.6 is a “generalized” Reed Muller code. Thus, it is sufficient to show that there is a choice of r_1, r_2, \dots, r_a for these codes for which Corollary 3.5.7 holds. In this direction we note that

$$\mathcal{T}_0 = T_{\text{GRM}}(1, 2, 2, \dots, 2, 2),$$

$$\mathcal{T}_1 = T_{\text{GRM}}(0, 1, 2, \dots, 2, 2)$$

and

$$\mathcal{T}_2 = T_{\text{GRM}}(0, 2, 2, \dots, 2, 2)$$

satisfy Corollary 3.5.7 as for $i = 1, \dots, a - 1$ and $\ell = 1, \dots, i - 1$

$$r_{i-\ell} \leq 2 \leq 2^\ell \leq 2^{\ell-1}(m' - r_i).$$

Moreover, one may show by using the normal basis $\{\zeta^p, \zeta^{p^2}, \dots, \zeta^{p^{m'}}\}$ the set of functions

$$\bigcup_{k \in \{0, 1, \dots, a-1\}} \bigcup_{s \in \mathcal{T}_{a-k+1}^\perp \cap \mathcal{I}_P} \left\{ \text{Tr} \left(\zeta^{\frac{p^{m'}-1}{p^{ms}-1}} p^i z^s \right) \right\}$$

are linearly independent over \mathbb{Z}_{p^a} . Hence, for any $i \in \{1, 2, 3\}$, $k \in \{0, 1, \dots, a - 1\}$ and $s \in \mathcal{T}_{a-k+1}^\perp \cap \mathcal{I}_P$ the vectors

$$\mathbf{g}_{i,s,k} = p^{k-1} \cdot \left[\text{Tr} \left(\zeta^{\frac{p^{m'}-1}{p^{ms}-1}} p^i z^s \right) \right]_{z \in \mathcal{I}_{p^a, m'}}$$

are linearly independent. Furthermore, any vector associated to a function from $\mathcal{F}(\mathcal{T}_i)$ is a linear combination of $\{\mathbf{g}_{i,s,k}\}$. Alternatively, any vector

$$\mathbf{g}(f) = [f(0), f(1), f(\zeta), \dots, f(\zeta^{m-1})]$$

associated to a function $f \in \mathcal{F}(\mathcal{T}_i)$ is, for some $\mathbf{v} \in \mathbb{Z}_{2^{k_2}}^3$,

$$\mathbf{g}(f) = G_i(k) \cdot \mathbf{v}$$

where, after some suitable change of coordinates,

$$G_0(k) = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 2 \end{bmatrix}, \quad G_1(k) = \begin{bmatrix} 2 & 3 & 3 & 0 \\ 3 & 2 & 3 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix},$$

$$G_2(2) = \begin{bmatrix} 0 & 2 & 2 & 0 \\ 2 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix} \quad \text{and for } k > 2 \quad G_2(k) = \begin{bmatrix} 4 & 6 & 6 & 0 \\ 6 & 4 & 6 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}.$$

This yields the construction from Example 3.2.6.

We have provided a systematic construction for both the dense and sparse codes that may be paired together to form a class of good constituent codes to use in a systematic construction. However, as we have seen in Section 3.2, this construction is not in general robust enough to allow one to increase the rate of a code endlessly as there is no mechanism in place to allow one to more accurately represent the magnitude of each coordinate, but rather only the phase. Thus, in the sequel we introduce a simple system of linear operators which not only allow one to more precisely quantize the magnitude of each coordinate, but also allows one a class of high rate quantizers with low quantization complexity.

■ 3.6 Low Complexity Rate Doubling Operations

In the preceding sections we have described the key ingredients to our quantizer construction. This construction consisted of a union of codes of differing sparsity which are all invariant to a set of shifts to the coordinate set. To increase the rate of the quantizer one may take one or many possible unions of such codes. Additionally, to further increase the rate, one may increase the cardinality of the integer ring underlying the construction of each of the component codes in the union. However, as we have seen such an approach constructs codes of higher and higher rates by increasing the precision of the quantizer in a subspace by increasing the precision of the phase of each coordinate. Thus, in the high rate limit this scheme will only produce a code in which the *phase* of each coordinate is known precisely while the magnitude of each coordinate is known only to finite precision. Thus, one may expect that for high rate quantization our current construction may not outperform simple scalar quantization.

For a truly systematic structured construction of channel quantizers one must find a systematic way to increase precision of the magnitude of every coordinate and not just the phase. To do this, one may consider taking unions of codes that are simple linear transformations of a “good” base code, say \mathcal{C}_r , in order to construct higher rate codebooks which use some of the rate to increase the precision of the magnitude of each coordinate. In this section we introduce a “localization” operation, $\mathbf{F}(\mathbf{c}_0; \alpha, \gamma)$, which takes any point on the complex sphere to a neighborhood of the codeword \mathbf{c}_0 described by α and γ . The freedom of α and γ allows one to tune this operation to optimize the performance of the resulting code. One of the greatest benefits to this approach is it allows one to form *multi-resolution* codebooks which greatly simplify the problem of quantization in high rate codes. In particular, it allows one to use multi-stage quantization algorithms. As we have stated, multi-user MIMO systems which operate in the high SNR regime must use large codebooks to ensure that the system performance is not limited. In such cases it is of interest to develop *structured* codebooks that enable user terminals to efficiently quantize their channel vectors. In particular, by appropriately choosing the parameters α and γ one may ensure that each element of $\mathcal{C}(\mathbf{c}_i; \alpha, \gamma, \mathcal{C})$ is inside the Voronoi cell of the codeword \mathbf{c}_i in the original code. Thus, in this special case one may quantize any channel vector by first performing quantization using \mathcal{C} then, using the same quantization algorithm, perform quantization inside the local code of the codeword which was the result of the first stage of decoding. If this may be done we say the codebook is a multi-resolution codebook.

A multi-resolution codebook is of great interest for MIMO broadcast systems as the quantization is performed at the user terminals. In many cases the user terminals are power and complexity limited and hence may not have the resources to perform high complexity quantization needed to obtain high rates. Indeed, this was part of our motivation to develop structured quantization methods as the complexity of quantization at the user terminals

using a random vector quantizer, in general, requires exponential complexity in the number of feedback bits. However, employing a well chosen base code \mathcal{C}_r and parameters α and γ one has the complexity of quantization at the user terminals using a multi-resolution codebook is two times that of the complexity of quantization using \mathcal{C}_r . Hence, irregardless of the performance of the multi-resolution codes relative to random vector quantization, there is great practical relevance in a high rate system to employ multi-resolution codes.

In this section we construct a framework in which a codebook which has been well designed for the Rayleigh model may be successively refined to higher and higher rate codes which are also good for the Rayleigh model. In this development we call the base design for the Rayleigh model the *root code* and the codebook consisting of the union of transformations of the root code the *universal code*. It is unreasonable to expect that one will preserve the structure of the original design. In particular, the image of any (every) set of orthogonal vectors under a non-unitary transformation will not be orthogonal. However, in the sequel we develop a simple transformations which extends a large part of the structure of the root code.

A Geometric Construction of Rate Doubling Operations

In a multi-user MIMO system there is a need to develop high rate, low complexity quantizers. Recall that our motivation behind a system of linear operators is that there is no mechanism in our systematic construction thus far to more accurately quantize the magnitude of each coordinate of a channel vector. Thus, in the sequel we derive a operator in which the components of the resulting codewords do not have constant modulus. In particular, we consider constructing new codes by interpolating between the lines of an existing constellation using a codebook from our previous framework (3.5). This general approach to construct universal codes is not new. The authors of [34, 102] have considered similar localization methods. However, the authors of [34, 102] did not consider the question of preserving an underlying structure of a code, nor did they address the problem of constructing a universal code which in its own right is a good quantizer for the Rayleigh model which allows for the use of multi-stage quantization algorithms. Hence, in the sequel, we arrive at a quite different form for the interpolation than was used in [34, 102].

In Section 3.1 we presented the 3 bit quantizer of length 4 that is currently an optional part of the 802.16 standard. In order to decrease the mean square quantization error this quantizer used a Householder transform to transform an existing constellation. In the sequel we use a similar approach to develop operators for our universal code. To begin, recall that Householder transform for two points, say \mathbf{x} and \mathbf{y} , in \mathbb{R}^n is the linear transform \mathbf{A} ,

$$\mathbf{A} = \mathbf{I} - 2 \frac{\mathbf{v} \mathbf{v}^T}{\mathbf{v}^T \mathbf{v}}$$

where $\mathbf{v} = \mathbf{x} - \mathbf{y}$. It is easy to see by direct computation that

$$\mathbf{A} \mathbf{x} = \mathbf{y}.$$

The Householder transform is well known for its usefulness in matrix analysis for both its efficiency and numerical stability. In \mathbb{C}^n the Householder transform for two points \mathbf{a} and \mathbf{b} takes a slightly different form and can be shown [36] to be, for $\|\mathbf{a}\| = \|\mathbf{b}\|$,

$$\mathbf{X}(\mathbf{a}, \mathbf{b}) = \mathbf{I} - 2 \frac{\mathbf{z} \mathbf{z}^\dagger}{\mathbf{z}^\dagger \mathbf{a}}$$

were $\mathbf{z} = \mathbf{a} - \mathbf{b}$. Again, some simple computation shows that

$$\mathbf{X}(\mathbf{a}, \mathbf{b})\mathbf{a} = \mathbf{b}.$$

Thus, if one wishes to interpolate between two points \mathbf{c}_1 and \mathbf{c}_2 the transform

$$\mathbf{Y}(\mathbf{c}_1, \mathbf{c}_2; \alpha) = \left(\sqrt{1 - |\alpha|^2} \cdot \mathbf{I} + \alpha \mathbf{X}(\mathbf{c}_1, \mathbf{c}_2) \right) \quad (3.68)$$

is a linear transform from \mathbf{c}_1 to $\mathbf{c}_1 + \alpha \mathbf{c}_2$. This transform depends on both \mathbf{c}_1 and \mathbf{c}_2 and it is not clear how one could simultaneously localize a code while extending a large number of symmetries to the entire code. However, as the Householder transform is a linear transformation and interpolates between two existing codewords it is clear that there is a structure present that preserves some of the existing structure of the root code.

We prefer a representation for the interpolation that is not dependent on the code word \mathbf{c}_2 as the associated transformation should give rise to symmetries for a large subset of the universal code. Hence, in the sequel we describe a different interpolation that can be defined in terms of a basis containing \mathbf{c}_1 and not \mathbf{c}_2 . In this direction, recall that geometrically the Householder transformation performs a rotation in the plane spanned by \mathbf{c}_1 and \mathbf{c}_2 while leaving the rest of the space fixed. Thus, if $\mathcal{B} = \{\mathbf{b}_0, \mathbf{b}_1, \dots, \mathbf{b}_{m-1}\}$ is an ortho-normal basis for \mathbb{C}^m then

$$\mathbf{Y}(\mathbf{b}_1, \mathbf{b}_2; \alpha) \cdot \mathbf{b}_l = \begin{cases} \alpha \cdot \mathbf{b}_2 - \sqrt{1 - |\alpha|^2} \cdot \mathbf{b}_1 & \text{if } \mathbf{b}_l = \mathbf{b}_1 \\ \alpha \cdot \mathbf{b}_1 + \sqrt{1 - |\alpha|^2} \cdot \mathbf{b}_2 & \text{if } \mathbf{b}_l = \mathbf{b}_2 \\ \mathbf{b}_l & \text{otherwise} \end{cases}$$

However, this only defines a single rotation and does not localized codewords as we desire. One could attempt to construct a more general interpolation operator by products and sums of interpolation operators of the form (3.68). However, this leads to complex cross terms that generally destroy any sense of locality of the resulting interpolations which will further inhibit the identification of large symmetries of the code. That is, products and sums of interpolation operators of the form (3.68) do not lead to an easily identifiable root codeword for the interpolation since (3.68) defines a two dimensional rotation. Hence, we rather consider one dimensional rotation operations

$$\tilde{\mathbf{Y}}(\mathbf{b}_1, \mathbf{b}_2; \alpha) = \mathbf{Y}(\mathbf{b}_1, \mathbf{b}_2; \alpha) - \left(\alpha \cdot \mathbf{b}_2 - \sqrt{1 - |\alpha|^2} \cdot \mathbf{b}_1 \right) \mathbf{b}_1^\dagger \quad (3.69a)$$

$$= \mathbf{I} + \left(\alpha \cdot \mathbf{b}_1 + (\sqrt{1 - |\alpha|^2} - 1) \cdot \mathbf{b}_2 \right) \mathbf{b}_2^\dagger \quad (3.69b)$$

so that,

$$\tilde{\mathbf{Y}}(\mathbf{b}_1, \mathbf{b}_2; \alpha) \cdot \mathbf{b}_l = \begin{cases} \alpha \cdot \mathbf{b}_1 + \sqrt{1 - |\alpha|^2} \cdot \mathbf{b}_2 & \text{if } \mathbf{b}_l = \mathbf{b}_2 \\ \mathbf{b}_l & \text{otherwise} \end{cases}$$

Hence, $\tilde{\mathbf{Y}}(\mathbf{b}_1, \mathbf{b}_2; \alpha)$ can be viewed as a rotation of the basis vector \mathbf{b}_2 in the $\mathbf{b}_1 - \mathbf{b}_2$ plane. We note that this interpolation operation has the added benefit that it is quite simple to invert $\tilde{\mathbf{Y}}$. Hence, elements of a local code may be efficiently quantized by first inverting the factor $\tilde{\mathbf{Y}}(\mathbf{b}_1, \mathbf{b}_2; \alpha)$ and using the quantization algorithm of the root quantizer. More precisely, using the inversion formula for a small rank adjustment one has [58],

$$\tilde{\mathbf{Y}}(\mathbf{b}_1, \mathbf{b}_2; \alpha)^{-1} = \mathbf{I} - \frac{1}{\sqrt{1 - |\alpha|^2}} \left(\alpha \cdot \mathbf{b}_1 + (\sqrt{1 - |\alpha|^2} - 1) \cdot \mathbf{b}_2 \right) \mathbf{b}_2^\dagger. \quad (3.70)$$

Comparing (3.70) to (3.69) it is easy to see that inverting $\tilde{\mathbf{Y}}(\mathbf{b}_1, \mathbf{b}_2; \alpha)$ is no more complex than the original interpolation operation. Indeed, this is expected as this operation is again just a one dimensional rotation in the $\mathbf{b}_1 - \mathbf{b}_2$ plane. However, $\tilde{\mathbf{Y}}(\mathbf{b}_0, \mathbf{b}; \alpha)$ as defined will only localize one component of every codeword about \mathbf{b}_0 . For efficient quantization we would like to have the entire root code localized about \mathbf{b}_0 . Hence, we form our interpolation operation as a product of the one dimensional rotations $\tilde{\mathbf{Y}}(\mathbf{b}_0, \mathbf{b}; \alpha)$. In particular, for each codeword $\mathbf{c}_i \in \mathcal{C}$ and an associated basis \mathcal{B}_i such that $\mathbf{c}_i \in \mathcal{B}_i$ we let, for $0 < \alpha < 1$ and $\gamma \in \mathbb{C}$,

$$\mathbf{F}(\mathbf{b}_0; \alpha, \gamma, \mathcal{B}) = \left(\mathbf{I} + (\gamma - 1) \cdot \mathbf{b}_0 \mathbf{b}_0^\dagger \right) \prod_{\mathbf{b} \in \mathcal{B} \setminus \mathbf{b}_0} \tilde{\mathbf{Y}}(\mathbf{b}_0, \mathbf{b}; \alpha) \quad (3.71)$$

be the local interpolation operation for the root codeword \mathbf{c}_i with respect to the basis \mathcal{B}_i and for each local interpolation operation we let

$$\mathcal{C}(\mathbf{c}_i; \alpha, \gamma, \mathcal{B}_i) = \{ \mathbf{F}(\mathbf{c}_i; \alpha, \gamma, \mathcal{B}_i) \cdot \mathbf{c}_j \mid \mathbf{c}_j \in \mathcal{C} \}$$

be the code localized about \mathbf{c}_j . We note that for an arbitrary choice for α and γ we have no general guarantee that the elements of $\mathcal{C}(\mathbf{c}_i; \alpha, \gamma, \mathcal{B}_i)$ are more correlated with \mathbf{c}_i than some other codeword from the root code. However, from (3.71) it is clear that for an appropriate choice of γ and α the elements of the local code $\mathcal{C}(\mathbf{c}_i; \alpha, \gamma, \mathcal{B}_i)$ can be made to be arbitrarily correlated with the \mathbf{c}_i . In particular, for any codeword $\mathbf{c} \in \mathcal{C}(\mathbf{c}_i; \alpha, \gamma, \mathcal{B}_i)$ one can see by inspecting (3.71) that for any $0 < \alpha \leq 1$ as $|\gamma| \rightarrow \infty$

$$\frac{|\mathbf{c}_i^\dagger \mathbf{F}(\mathbf{c}_i; \alpha, \gamma, \mathcal{B}) \mathbf{c}|}{\|\mathbf{F}(\mathbf{c}_i; \alpha, \gamma, \mathcal{B}) \mathbf{c}\|} \rightarrow 1$$

while for any $\gamma \in \mathbb{C}$

$$\frac{|\mathbf{c}_i^\dagger \mathbf{F}(\mathbf{c}_i; \alpha, \gamma, \mathcal{B}) \mathbf{c}|}{\|\mathbf{F}(\mathbf{c}_i; \alpha, \gamma, \mathcal{B}) \mathbf{c}\|} \rightarrow 0$$

as $\alpha \rightarrow 0$. Hence, for some appropriate choice of α and γ we can ensure that every codeword in the local code $\mathcal{C}(\mathbf{c}_i; \alpha, \gamma, \mathcal{B}_i)$ is more correlated with \mathbf{c}_i than any other codeword in the root code and hence the codewords in $\mathcal{C}(\mathbf{c}_i; \alpha, \gamma, \mathcal{B}_i)$ are truly “local” to \mathbf{c}_i . Further, one can always ensure by appropriately choosing α and γ that the resulting union of local codes is a multi-resolution.

Universal Codes From Geometric Operators

As previously noted, our interest in forming the local codes $\mathcal{C}(\mathbf{c}_i; \alpha, \gamma, \mathcal{B}_i)$ is it allows one to form a much larger code from a root code in which each codeword of the root code has an associated local code of equal rate. That is, we can view each local code as a subcode of a “universal” code

$$\mathcal{C}_{\mathbf{F}}(\alpha, \gamma, \mathcal{C}) = \bigcup_{\mathbf{c}_i \in \mathcal{C}} \mathcal{C}(\mathbf{c}_i; \alpha, \gamma, \mathcal{B}_i). \quad (3.72)$$

This large code is of interest when our existing systematic construction fails to yield a desired mean square error performance. Moreover, this is the codebook of interest if one wishes to develop a codebook in which quantization may be performed on the root code then sub-codebooks corresponding to specific local codes¹⁵. Thus, it is of interest to understand

¹⁵ Alternatively the “universal” code is the appropriate setting for slow fading channels where users incrementally feedback a quantized description of their channel.

how one's choice of α and γ affect the properties of this code (i.e. the coherence of the code and when the code is a multi-resolution).

In order to precisely characterize the effects of the parameters α and γ we first examine the eigen and geometric structure of the operators $\mathbf{F}(\mathbf{b}_i; \alpha, \gamma, \mathcal{B})$. For this, it is often more convenient to write $\mathbf{F}(\mathbf{b}_0; \alpha, \gamma, \mathcal{B})$ as a sum rather than a product. This is the content of the following lemma.

Lemma 3.6.1. *For any complex vector \mathbf{b}_0 and basis \mathcal{B} containing \mathbf{b}_0 ,*

$$\mathbf{F}(\mathbf{b}_0; \alpha, \gamma, \mathcal{B}) = \left(\gamma(1 - \alpha) - \sqrt{1 - \alpha^2} \right) \cdot \mathbf{b}_0 \mathbf{b}_0^\dagger + \quad (3.73a)$$

$$\sum_{\mathbf{b} \in \mathcal{B}} \left(\alpha \gamma \mathbf{b}_0 + (\sqrt{1 - \alpha^2}) \mathbf{b} \right) \mathbf{b}^\dagger \quad (3.73b)$$

Proof. See Appendix C.2.10. ■

Now, to characterize the behavior of $\mathbf{F}(\mathbf{b}_0; \alpha, \gamma, \mathcal{B})$ on the code \mathcal{C} , we note that the matrix $\mathbf{F}(\mathbf{b}_0; \alpha, \gamma, \mathcal{B})$ is in general not Hermitian as the term in (3.73b) is not Hermitian if $\alpha \gamma \neq 0$. In fact, it is easy to see that $\mathbf{F}^\dagger \mathbf{F} \neq \mathbf{F} \mathbf{F}^\dagger$ so that \mathbf{F} is not even a normal matrix. Hence, by the spectral theorem for normal matrices [58] the eigenvectors of \mathbf{F} are not orthonormal. Thus, we may only take the weakest form for the eigen-decomposition [58] for the matrix \mathbf{F} . That is, as \mathbf{F} is full rank and not normal, there exists a matrix \mathbf{P} whose columns are the eigenvectors of \mathbf{F} and a diagonal matrix \mathbf{D} , such that

$$\mathbf{F} = \mathbf{P} \mathbf{D} \mathbf{P}^{-1}.$$

Now let,

$$\nu(\alpha, \gamma) = \frac{\alpha \gamma}{\sqrt{1 - \alpha^2} - \gamma}.$$

Then, we have the following description of the eigenvectors of $\mathbf{F}(\mathbf{b}_0; \alpha, \gamma, \mathcal{B})$.

Lemma 3.6.2. *Let \mathcal{B} be an orthonormal basis for \mathbb{C}^m . Then, for any $0 < \alpha < 1$ and $\gamma \in \mathbb{C}$ such that $\gamma \neq \sqrt{1 - \alpha^2}$, \mathbf{b}_0 is an eigenvector for $\mathbf{F}(\mathbf{b}_0; \alpha, \gamma, \mathcal{B})$ with eigenvalue γ and $\{\mathbf{b} + \nu(\alpha, \gamma) \cdot \mathbf{b}_0\}_{\mathbf{b} \in \mathcal{B} \setminus \mathbf{b}_0}$ is a basis for the eigenspace of $\mathbf{F}(\mathbf{b}_0; \alpha, \gamma, \mathcal{B})$ with eigenvalue $\sqrt{1 - \alpha^2}$.*

Proof. See Appendix C.2.11 ■

Examining Lemma 3.6.2 yields some simple intuitions behind the choice of (3.73) as the local interpolation operation. In particular, examining Lemma 3.6.2 one can see that the eigenstructure of $\mathbf{F}(\mathbf{b}_0; \alpha, \gamma, \mathcal{B})$ is aligned with \mathbf{b}_0 as \mathbf{b}_0 is an eigenvector of $\mathbf{F}(\mathbf{b}_0; \alpha, \gamma, \mathcal{B})$ as well as the linear dependence between the eigenvectors implied by non-normality of the matrix $\mathbf{F}(\mathbf{b}_0; \alpha, \gamma, \mathcal{B})$. To more precisely characterize this dependence we now explicitly compute an orthonormal basis for the eigenspace associated with the eigenvalue $\sqrt{1 - \alpha^2}$. In this direction note that any $\mathbf{v} \in \mathbb{C}^m$ can be written as

$$\mathbf{v} = \sum_{i=0}^m a_i \mathbf{b}_i$$

for some $a_i \in \mathbb{C}$ as \mathcal{B} is a orthonormal basis for \mathbb{C}^m . Moreover, examining Theorem 3.6.2 one can see that any vector $\mathbf{v} \in \mathbb{C}^m$ such that $a_0 = 0$ and $\sum_{i=0}^m a_i = 0$ is an eigenvector of $\mathbf{F}(\mathbf{b}_0; \alpha, \gamma, \mathcal{B})$ with eigenvalue $\sqrt{1 - \alpha^2}$ as

$$\begin{aligned} \mathbf{v} &= \sum_{i=1}^m a_i \mathbf{b}_i \\ &= \nu(\alpha, \gamma) \cdot \left(\sum_{i=1}^m a_i \right) \cdot \mathbf{b}_0 + \sum_{i=1}^m a_i \mathbf{b}_i \\ &= \sum_{i=1}^m a_i (\nu(\alpha, \gamma) \cdot \mathbf{b}_0 + \mathbf{b}_i) \end{aligned}$$

if $\sum_{i=1}^m a_i = 0$. Moreover, this set of vectors form as $m - 2$ dimensional subspace of \mathbb{C}^m and every vector from this subspace is trivially orthogonal to any vector

$$\hat{\mathbf{v}} = a_0 \cdot \mathbf{b}_0 + \sum_{i=1}^m \mathbf{b}_i.$$

However, by Lemma 3.6.2, the set of vectors $\{\mathbf{b} + \nu(\alpha, \gamma) \cdot \mathbf{b}_0\}_{\mathbf{b} \in \mathcal{B} \setminus \mathbf{b}_0}$ is a basis for the eigenspace of $\mathbf{F}(\mathbf{b}_0; \alpha, \gamma, \mathcal{B})$ with eigenvalue $\sqrt{1 - \alpha^2}$. Hence,

$$\hat{\mathbf{v}} = (m - 1)\nu(\alpha, \gamma) \cdot \mathbf{b}_0 + \sum_{i=1}^m \mathbf{b}_i$$

is an element of the eigenspace of $\mathbf{F}(\mathbf{b}_0; \alpha, \gamma, \mathcal{B})$ with eigenvalue $\sqrt{1 - \alpha^2}$. Thus, to find an orthonormal basis for the eigenspace we must identify $m - 2$ orthogonal vectors of length $m - 1$ that sum to zero to use in addition to the already identified eigenvector $\hat{\mathbf{v}}$. However, this set of $m - 2$ vectors is quite familiar. It is simply the set of rows (or columns) from that $(m - 1) \times (m - 1)$ discrete Fourier transforms (DFT) matrix which sum to zero. In this direction, let

$$\text{DFT}(m) = \frac{1}{\sqrt{m}} \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & e^{\sqrt{-1} \frac{2\pi}{m} 1} & e^{\sqrt{-1} \frac{2\pi}{m} 2} & \cdots & e^{\sqrt{-1} \frac{2\pi}{m} (m-1)} \\ 1 & e^{\sqrt{-1} \frac{2\pi}{m} 2} & e^{\sqrt{-1} \frac{2\pi}{m} 4} & \cdots & e^{\sqrt{-1} \frac{2\pi}{m} (m-1) 2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & e^{\sqrt{-1} \frac{2\pi}{m} (m-1) 1} & e^{\sqrt{-1} \frac{2\pi}{m} (m-1) 2} & \cdots & e^{\sqrt{-1} \frac{2\pi}{m} (m-1) (m-1)} \end{bmatrix} \quad (3.74)$$

be the $m \times m$ discrete Fourier transforms (DFT) matrix and let

$$\mathbf{B}_j(\mathcal{B}) = \mathbf{B}_j = \left[\begin{array}{c|ccc|c} & \cdots & & \\ \mathbf{b}_{i_0} & \cdots & \mathbf{b}_{i_{m-1}} & \\ & \cdots & & \end{array} \right] \text{ and } \tilde{\mathbf{B}}_j(\mathcal{B}) = \tilde{\mathbf{B}}_j = \left[\begin{array}{c|ccc|c} & \cdots & & \\ \mathbf{b}_{i_1} & \cdots & \mathbf{b}_{i_{m-1}} & \\ & \cdots & & \end{array} \right]$$

where $i_0 = j$ and $\{i_0, i_1, \dots, i_{m-1}\} = \{0, 1, 2, \dots, m-1\}$. Then from the preceding discussion it is clear that the $(m - 1) \times (m - 1)$ submatrix of the DFT for which the rows sum to zero times $\tilde{\mathbf{B}}_j(\mathcal{B})$ forms an orthonormal basis for an $m - 2$ dimensional subspace of the

eigenspace of $\mathbf{F}(\mathbf{b}_0; \alpha, \gamma, \mathcal{B})$ with eigenvalue $\sqrt{1 - \alpha^2}$. Thus, we let

$$\mathbf{U}_F(\alpha, \gamma)^\dagger = \left[\begin{array}{c|cccc} 1 & 0 & \cdots & 0 & 0 \\ \nu(\alpha, \gamma) & & & & \\ 0 & & & & \\ \vdots & & & & \\ 0 & & & & \end{array} \middle| \text{DFT}(m-1) \right], \quad \bar{\mathbf{U}}_F(\alpha, \gamma)^\dagger = \left[\begin{array}{c|cccc} 1 & 0 & \cdots & 0 & 0 \\ \nu(\alpha, \gamma) & & & & \\ \nu(\alpha, \gamma) & & & & \\ \vdots & & & & \\ \nu(\alpha, \gamma) & & & & \end{array} \middle| \text{DFT}(m-1)^\dagger \right]$$

and let

$$\mathbf{\Lambda}_F(\alpha, \gamma) = \left[\begin{array}{ccccc} \gamma & 0 & 0 & \cdots & 0 \\ 0 & & & & \\ 0 & \sqrt{1 - \alpha^2} \cdot \mathbf{I}_{m-2} & & & \\ \vdots & & & & \\ 0 & & & & \end{array} \right] \quad \text{and} \quad \mathbf{D}_F(\alpha, \gamma) = \left[\begin{array}{cccccc} 1 & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{\nu(\alpha, \gamma)^2 + (m-1)}} & 0 & \cdots & 0 \\ 0 & 0 & & & \\ 0 & 0 & & & \mathbf{I}_{m-2} \\ \vdots & & & & \\ 0 & 0 & & & \end{array} \right]$$

The preceding discussion leads to the following theorem.

Theorem 3.6.3. *Let \mathcal{B} be an orthonormal basis for \mathbb{C}^m . Then, for any $\mathbf{b}_0 \in \mathcal{B}$ and any $0 < \alpha < 1$ and $\gamma \in \mathbb{C}$ such that $\gamma \neq \sqrt{1 - \alpha^2}$. Then,*

$$\mathbf{F}(\mathbf{b}_0; \alpha, \gamma, \mathcal{B}) = \mathbf{B}_0 \mathbf{D}_F(\alpha, \gamma) \mathbf{U}_F(\alpha, \gamma) \mathbf{\Lambda}_F(\alpha, \gamma) \cdot (\mathbf{B}_0 \cdot \mathbf{D}_F(\alpha, \gamma) \mathbf{U}_F(\alpha, \gamma))^{-1} \quad (3.75a)$$

$$= \mathbf{B}_0 \mathbf{D}_F(\alpha, \gamma) \mathbf{U}_F(\alpha, \gamma) \mathbf{\Lambda}_F(\alpha, \gamma) \cdot \bar{\mathbf{U}}_F(\alpha, \gamma) \mathbf{D}_F(\alpha, \gamma)^{-1} \mathbf{B}_0^\dagger \quad (3.75b)$$

where $\mathbf{B}_0 = \mathbf{B}_0(\mathcal{B})$.

Proof. This theorem has been proven by the preceding discussion. The only things left to show is the form for the inverse of $\mathbf{B}_0 \cdot \mathbf{D}_F(\alpha, \gamma) \mathbf{U}_F(\alpha, \gamma)$. This is easily seen as $(\mathbf{B}_0)^{-1} = \mathbf{B}_0^\dagger$, $\mathbf{D}_F(\alpha, \gamma)$ is diagonal and the inverse of $\mathbf{U}_F(\alpha, \gamma)$ can be verified by direct multiplication. ■

Before proceeding we more closely examine Theorem 3.6.3. We note that while (3.75) is a quite long chain of matrix multiplications each of the terms requires very little computation. Moreover, due to the specific structure of the DFT matrix it is natural to expect that there is a more efficient way to apply this transform than through the application of the eigenvalue decomposition. Let,

$$\mathbf{F}_I(\alpha, \gamma) = \left[\begin{array}{cccccc} \gamma & 0 & 0 & \cdots & 0 & 0 \\ \alpha\gamma & \sqrt{1 - \alpha^2} & 0 & \cdots & 0 & 0 \\ \alpha\gamma & 0 & \sqrt{1 - \alpha^2} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ \alpha\gamma & 0 & 0 & \cdots & \sqrt{1 - \alpha^2} & 0 \\ \alpha\gamma & 0 & 0 & \cdots & 0 & \sqrt{1 - \alpha^2} \end{array} \right]$$

Then we have the following corollary.

Corollary 3.6.4. *Let \mathbf{c} be any vector in \mathbb{C}^m with unit norm and let $\mathbf{c}_0, \mathcal{B}$ be given such that $\mathbf{c}_0 \in \mathcal{B}$ and \mathcal{B} is an orthonormal basis for \mathbb{C}^m . Then,*

$$\mathbf{F}(\mathbf{c}_0; \alpha, \gamma, \mathcal{B}) \mathbf{c} = \mathbf{B}_0(\mathcal{B}) \cdot \mathbf{F}_I(\alpha, \gamma) \cdot \mathbf{B}_0(\mathcal{B})^\dagger \mathbf{c} \quad (3.76a)$$

Moreover, if $\mathbf{B}_0(\mathcal{B})^\dagger \in \text{Sym}(\mathcal{C})$ then

$$\mathbf{F}(\mathbf{c}_0; \alpha, \gamma, \mathcal{B}) \cdot \mathcal{C} = \mathbf{B}_0(\mathcal{B}) \cdot \mathbf{F}_I(\alpha, \gamma) \cdot \mathcal{C}$$

Proof. Both parts of this corollary are achieved by direct computation. \blacksquare

Note that Corollary 3.76 precisely describes what we geometrically expect. That is, every codeword from the root code gets a gain in the direction of \mathbf{c}_0 and a uniform scaling in the space orthogonal to \mathbf{c}_0 . However, this result is far more applicable in the case of interest when $\mathbf{B}_0(\mathcal{B})^\dagger \in \text{Sym}(\mathcal{C})$. In particular, if $\mathbf{B}_0(\mathcal{B})^\dagger \in \text{Sym}(\mathcal{C})$ for any $\mathbf{c} \in \mathcal{C}$ one has

$$\mathbf{B}_0(\mathcal{B}) \left((\gamma - \sqrt{1 - \alpha^2} + (m - 1)\alpha\gamma)\mathbf{e}_0\mathbf{e}_0^\dagger\mathbf{c} + \sqrt{1 - \alpha^2}\mathbf{c} \right) \in \mathcal{C}_{\mathbf{F}}(\alpha, \gamma, \mathcal{C}).$$

In the sequel we let

$$\mathbf{c}_{\mathbf{F}}^{(0)}(\mathbf{c}; \alpha, \gamma) = \frac{\left((\gamma - \sqrt{1 - \alpha^2} + (m - 1)\alpha\gamma)\mathbf{e}_0\mathbf{e}_0^\dagger\mathbf{c} + \sqrt{1 - \alpha^2}\mathbf{c} \right)}{\left\| \left((\gamma - \sqrt{1 - \alpha^2} + (m - 1)\alpha\gamma)\mathbf{e}_0\mathbf{e}_0^\dagger\mathbf{c} + \sqrt{1 - \alpha^2}\mathbf{c} \right) \right\|} \quad (3.77)$$

By examining (3.77) one may see that if $\mathbf{B}_0(\mathcal{B})^\dagger \in \text{Sym}(\mathcal{C})$, the universal code $\mathcal{C}_{\mathbf{F}}(\alpha, \gamma, \mathcal{C})$ quantizes the magnitude of the first coordinate of a vector in the $\mathbf{B}_0(\mathcal{B})$ coordinate system relative to the other non-zero coordinates. Thus, one now has a method to increase the code rate by varying the magnitude of the coordinates. Moreover, if every basis used in the construction acts two transitively on the code then these results hold for every basis used in the construction and one can infer quite a bit about the structure of the universal code through the structure of the orthogonal bases used in the construction of $\mathcal{C}_{\mathbf{F}}(\alpha, \gamma, \mathcal{C})$. This is the content of the following theorem.

Theorem 3.6.5. *Let $\{\mathcal{B}_i\}_{i=0}^{2^r}$ be the 2^r orthogonal bases used in the construction of the universal code $\mathcal{C}_{\mathbf{F}}(\alpha, \gamma, \mathcal{C})$ where \mathcal{B}_i is the basis used to construct the local code $\mathcal{C}(\mathbf{c}_i; \alpha, \gamma, \mathcal{C})$. If $\mathbf{B}_0(\mathcal{B}_i) \in \text{Sym}(\mathcal{C})$ for $i = 0, 1, \dots, 2^r - 1$, then*

$$\max_{\substack{\mathbf{c}_i, \mathbf{c}_j \in \mathcal{C}_{\mathbf{F}}(\alpha, \gamma, \mathcal{C}) \\ \mathbf{c}_i \neq \mathbf{c}_j}} |\mathbf{c}_i^\dagger \mathbf{c}_j| = \max_{\mathcal{B}_g, \mathcal{B}_h} \max_{\substack{\mathbf{c}_f, \mathbf{c}_j \in \mathcal{C} \\ f \neq i \text{ and } g \neq h}} \left| \mathbf{c}_{\mathbf{F}}^{(0)}(\mathbf{c}_f; \alpha, \gamma)^\dagger \mathbf{B}_0(\mathcal{B}_g)^\dagger \mathbf{B}_0(\mathcal{B}_h) \mathbf{c}_{\mathbf{F}}^{(0)}(\mathbf{c}_i; \alpha, \gamma) \right|$$

We note that Theorem 3.6.5 may not seem directly applicable to the problem of interest at first. However, it greatly simplifies the code optimization process if there is a regular structure to the orthogonal bases used in the construction. In particular, let B be the collection of unique values $\mathbf{B}_0(\mathcal{B}_g)^\dagger \mathbf{B}_0(\mathcal{B}_h)$ assumes for every (non-unique) pair of orthogonal bases used in the construction of a universal code. Then, one has

$$\max_{\substack{\mathbf{c}_i, \mathbf{c}_j \in \mathcal{C}_{\mathbf{F}}(\alpha, \gamma, \mathcal{C}) \\ \mathbf{c}_i \neq \mathbf{c}_j}} |\mathbf{c}_i^\dagger \mathbf{c}_j| = \max_{\mathbf{M} \in B} f_{\text{dist}}(\mathbf{M}; \mathcal{C}, \alpha, \gamma)$$

where

$$f_{\text{dist}}(\mathbf{M}; \mathcal{C}, \alpha, \gamma) = \widetilde{\max}_{\mathbf{c}_f, \mathbf{c}_j \in \mathcal{C}} \left| \mathbf{c}_{\mathbf{F}}^{(0)}(\mathbf{c}_f; \alpha, \gamma)^\dagger \mathbf{M} \mathbf{c}_{\mathbf{F}}^{(0)}(\mathbf{c}_i; \alpha, \gamma) \right|$$

and in turn where $\widetilde{\max}$, for notation convenience, excludes any solution that results in 1. Thus, if $|B|$ is small then computation of the coherence may be greatly simplified. This is

of practical relevance in our design as, by construction, every codeword in our systematic construction is contained in multiple orthogonal bases. Thus, if the set of orthogonal bases is chosen from the code the set B will be quite small and the coherence will be easily computable. However, of greater relevance is that Theorem 3.6.5 may be used to characterize exactly when the resulting universal code is a multi-resolution. That is, when optimal quantization consist of first performing quantization on the root code then on the corresponding local code. Clearly a universal code is a multi-resolution code if and only if for each local code $\mathcal{C}(\mathbf{c}_i; \alpha, \gamma, \mathcal{C})$,

$$\mathbf{c}_i = \arg \max_{\mathbf{c} \in \mathcal{C}} |\mathbf{c}^\dagger \mathbf{c}_\ell|$$

for each $\mathbf{c}_\ell \in \mathcal{C}(\mathbf{c}_i; \alpha, \gamma, \mathcal{C})$. Alternatively, if the set of orthogonal bases used in the construction act transitively on the root code \mathcal{C} one has that a universal code is a multi-resolution if and only if

$$\mathbf{e}_0 = \arg \max_{\mathbf{c} \in \mathcal{C}} \max_{\mathbf{M} \in B} \mathbf{c}^\dagger \mathbf{M} \mathbf{c}_F^{(0)}(\mathbf{c}_i; \alpha, \gamma)$$

for every $\mathbf{c}_i \in \mathcal{C}$. Thus, for a well chosen set of orthogonal bases contained in the root code one may consider optimizing codes by solving the optimization problem:

$$\min_{\mathcal{C}_r \subset \mathbb{C}^m} \min_{\substack{\{\mathcal{B}_i\}_{i=0}^{2^r-1} \\ \mathcal{B}_i \subset \mathcal{C}_r}} \min_{0 < \alpha < 1, \gamma \in \mathbb{C}} \max_{\mathbf{M} \in B} f_{\text{dist}}(\mathbf{M}; \mathcal{C}_r, \alpha, \gamma) \quad (3.78)$$

where one may add the additional constraint

$$(\mathbf{c} - \mathbf{e}_0)^\dagger \mathbf{M} \mathbf{c}_F^{(0)}(\mathbf{c}_i; \alpha, \gamma) < 0$$

for all $\mathbf{c}_i, \mathbf{c} \in \mathcal{C}_r$ if one is interested in a multi-resolution codebook.

When the size of B is small, one may precisely compute the optimal choice of α and γ for a given code \mathcal{C}_r and collection of orthogonal bases by examining the spectrum of the matrices contained in B . Further, one can use these results to ensure that the code is a multi-resolution. However, in general one must use a non-linear optimization routine to solve for good choices of α and γ and when one wants a truly systematic construction the choice for the defining sets and rates of the dense and sparse codes must be optimized according to (3.67). We provide these general methods in [119] along with an archive of our best found codes. The performance of our constructions in \mathbb{C}^4 may be seen in Figure 3-14.

We note while our systematic construction performs well in terms of SINR_{sat} this systematic construction does not guarantee that a set of user selected for transmission will achieve a high rate. That is, as previously noted, quantizers which optimize SINR_{sat} do not necessarily guarantee that the rates achieved in a system will be optimal. This is due to the fact that SINR_{sat} by definition assumes that there is a set of nearly orthogonal users and hence SINR_{sat} by definition does not favor codebooks with many orthogonal bases. At moderate SNR there may be a considerably smaller gap between the expected SINR achieved by one of our constructions and RVQ as in general there will be a SINR penalty due to channel inversion with RVQ. Further, the definition of SINR_{sat} only considers the quantization error of a single-user. As previously noted, in MIMO systems with many users the order statistic for the quantization error will lead to a decrease in the performance gap between a given channel quantizer and the optimal scheme. Moreover, in such systems one expects by choosing the users that have the best quantization error, the gap between the achieved average SINR of a system which uses a channel quantizer with many orthogonal

bases and one without many orthogonal bases to be smaller. In the following section we show that this is true for systems in which the number of users is only a small multiple of the size of the transmit array.

Multi-User MIMO System Design with Finite Rate Feedback

Current standards for multi-user MIMO system [1] require that in addition to high data rates, quality of service (QOS) guarantees must be met as well. These, for example, may be delay and stability guarantees. As previously noted, this problem is well understood for tradition wireline networks and more generally for single-antenna system. However, these methods are not directly applicable in a multi-user MIMO system due to the time varying nature of the fading channel which introduces random co-channel interference between users for each fading state. That is, as one in general only has causal knowledge of the time varying channel due to the CSI feedback at the transmitter, one can not use a simple round-robin or time division scheme and expect to simultaneously provide high throughput while meeting QOS guarantees in multi-user MIMO. Moreover, it is unclear how ones limited knowledge of the channel state influences the broader problem of delivering quality of service guarantees. In particular, it is unclear if the extra degrees of freedom available in our feedback design provides any assistance in delivering quality of service guarantees or more generally how this degree of freedom may be exploited to simplify and/or reduce the complexity of the associated scheduling algorithms.

In this chapter we do not address which particular scheduling mechanism one should use to meet a given quality of service requirement. Rather we more generally consider the effects the order statistic gain, multi-node matching gain and feedback design effect the complexity of user selection when broader quality of service requirements are of interest. To this end, we consider the problem of maximal weight matching, where by a set of users is selected for transmission if the set has the highest weighted achievable rate, is effected by the feedback design. In the sequel we first identify a low complexity system architecture which may be used to perform maximum sum rate scheduling. We then extend these results to a more general quality of service framework which solves the maximal weighted rate scheduling framework and identify its applicability in meeting QOS guarantees. Then, we provide our insights on how the order statistic gain, multi-node matching gain and feedback design effect the complexity of user selection in this framework.

In a multi-user MIMO system the channel aware scheduling problem is of practical interest as current standards demand that quality of service constraints be met in addition to throughput guarantees. Due to the nature of the fading channel it may be impractical and too costly to examine all subsets of users due to computation and power constraints at the transmitter. In Chapter 3 we developed a feedback framework that maximized an upper bound on the achievable SINR, SINR_{sat} , which assumed a high SNR limit as well as users with negligible co-channel interference. However, we left open the question of the ability of one to schedule users with negligible co-channel interference. Central to this question is

whether one may first reduce the size of the user pool by selecting users whose channels are individually at high SNR while still finding a subset of users that are nearly orthogonal. Such a result would imply that there are extra degrees of freedom in the feedback design allowing a system designer to reduce the overall system complexity by developing structured quantizers. In Chapter 1 we encapsulated this question in the trade-off between the order statistic gain and the multi-node matching gain. That is in Chapter 1 we argued the trade-off between the order statistic gain and the multi-node matching gain has algorithmic relevance as it effects the number of subset that must be considered to find a set of users with low co-channel interference and high SNR. In this chapter we provide a system architecture and an associated analysis framework with which one may analyze the trade-off between the channel fading statistics, the order statistic gain, the multi-node matching gain and the structure of the feedback design. We show that in the case of the Rayleigh model this architecture is optimal in a very strong sense as the size of the user pool tends to ∞ as well as provide a simple method of system design when the size of the user pool is small.

To address the effects of the channel fading statistics we begin by weakening our assumptions on our channel model. Henceforth, we assume that the user pool may be partitioned into n_c clusters of users

$$\mathcal{U} = \prod_{\ell=0}^{n_c-1} \mathcal{U}^{(\ell)}.$$

where each user in $i \in \mathcal{U}^{(\ell)}$ has a channel vector with common spatial correlation. More precisely, for $i \in \mathcal{U}^{(\ell)}$,

$$\mathbf{h}_i = \Sigma_\ell^{1/2} \cdot \mathbf{h}_i^{(0)}$$

where where the elements of $\mathbf{h}_i^{(0)}$ are *i.i.d* $\mathcal{CN}(0, 1/2m)$. Thus, as discussed in Chapter 2.2, each user has, in general, a non-uniform probability of being quantizers to a codeword. That is, one has

$$\mathbf{p}_i = \mathbf{p}^{(\ell)} = [p_0^{(\ell)}, p_1^{(\ell)}, \dots, p_{2^r-1}^{(\ell)}]$$

where $p_i^{(\ell)}$ is the probability that a user from the ℓ th cluster is quantized to the i th codeword. For each cluster the associated the channel correlation will effect the ability of user of each cluster to meet specified quantization error and channel norm constraints. Thus, for each cluster, we would like a method to optimize the order statistic gain and the multi-node matching gain. Thus, in the sequel we present a simple system architecture that aids in this optimization and relates directly to ones ability to efficiently select users.

■ 4.1 A System Architecture to Optimize System Tradeoffs

In this section, we present a simple system architecture for subset selection for use in multi-user MIMO systems that directly relates to the trade-off between the order statistic gain and the multi-node matching gain. In particular, we propose a system architecture whereby nodes first perform a decentralized and distributed subset-selection based on each users measurement of their own channel. Then, in this system architecture, the users selected by this decentralized subset-selection feed back a quantized version of their channel to the transmit base. The transmitter then selects from those users reporting the best subset to be used in the reconstruction.

In order to evaluate the complexity of user selection one may examine the effects the distributed subset-selection has on ones ability to optimally select users. For this, we propose

a simple two stage process for user selection at the transmit base. First, a greedy search is used to produce a small collection of candidate subsets. Then, an exhaustive search over this smaller collection is used to determine the final subset to be selected as the active set. We now describe the system architecture of interest.

In the architecture of interest it is the job of the *scheduler* to select the set \mathcal{A} for transmission. At each interval the scheduler selects the activation set \mathcal{A} and message symbols \mathbf{u} and forwards this set and vector to the multiplexer which forms the signal to be multiplexed across the array. It is the job of the *multiplexer*, which was described in Section 2.3 to select the signal vector \mathbf{x} for transmission. In each scheduling interval, a subset \mathcal{R} of users from the full population \mathcal{U} send a quantized representation of their respective channel gain vectors to the transmitter over the feedback link. The subset \mathcal{R} is determined in a decentralized manner, i.e., based on an individual evaluation of each channel gain vector. Specifically, each user j computes the squared norm $\|\mathbf{h}_j\|^2$ of its channel gain vector, and the correlation $|\mathbf{h}_j^\dagger \hat{\mathbf{h}}_j|$ between the channel gain vector and its quantization $\hat{\mathbf{h}}_j$. If these factors fall within certain prescribed ranges, a user will convey its channel gain to the transmitter. As we assume that each user cluster $\mathcal{U}^{(\ell)}$ has a distinct spatial correlation structure it is likely that a single threshold for the user population will have varying effects on each of the clusters. Thus, we assume a different criterion for feedback from each of the clusters which corresponds to

$$\mathcal{R}_{\rho,\sigma}^{(\ell)} \triangleq \{j \in \mathcal{U}^{(\ell)} : \rho_-^{(\ell)} \leq \|\mathbf{h}_j\|^2 \leq \rho_+^{(\ell)} \text{ and } |\tilde{\mathbf{h}}_j^\dagger \hat{\mathbf{h}}_j| \geq \sigma^{(\ell)}\}, \quad (4.1)$$

where $\tilde{\mathbf{h}}_j = \mathbf{h}_j / \|\mathbf{h}_j\|$, and where $\rho_+^{(\ell)}$, $\rho_-^{(\ell)}$, and $\sigma^{(\ell)}$ are prescribed parameters of the protocol¹ for the ℓ th cluster. We assume throughout that $\sigma^{(\ell)}$ is chosen such that $\sigma^{(\ell)} \geq \mu_0(\mathcal{C})$

At the transmitter, there are three relevant stages of processing. First, from the set \mathcal{R} of reporting users, a collection \mathcal{T} of candidate subsets is formed; this is the pre-selection phase. The pre-selection phase² is based on simple pairwise evaluation of the vectors in \mathcal{R} . The particular criterion we consider corresponds to

$$\mathcal{T}_\epsilon^{(\ell)} \triangleq \{\mathcal{A} \subset \mathcal{R}_{\rho,\sigma}^{(\ell)} : |\mathcal{A}| = m \text{ and } |\hat{\mathbf{h}}_i^\dagger \hat{\mathbf{h}}_j| \leq \epsilon, \forall i \neq j \in \mathcal{R}_{\rho,\sigma}^{(\ell)}\}, \quad (4.2)$$

where ϵ is another prescribed parameter of the protocol. Next, one of these subsets, denoted \mathcal{A} , is selected from \mathcal{T} by the scheduler, and corresponds to the active user set for the signaling interval. Finally, one message for each of the active users is selected, and the resulting group of messages is multiplexed across the array for transmission. The architecture of interest is illustrated in Fig. 4-1. The protocol is identical in each signaling interval, so we restrict our attention to a single arbitrary one. We now examine how our system architecture relates to the questions of interest.

As mentioned in Chapter 1 the order statistic gain and the multi-node matching gain are not compatible in general. That is, if one attempts to select only the users individually at high SNR it may not be possible to find a subset of users that negligibly interfere with one another. This particular dependence is embodied in the distribution of the users which feed

¹We note in practice $\rho_+^{(\ell)}$ should typically be set to ∞ .

²This is the particular embodiment of the pre-selection phase we fix in the sequel. However, in general the pre-selection phase may be taken to be more general. In particular, it may be taken to be the solution obtained by multiple runs of a greedy algorithm.

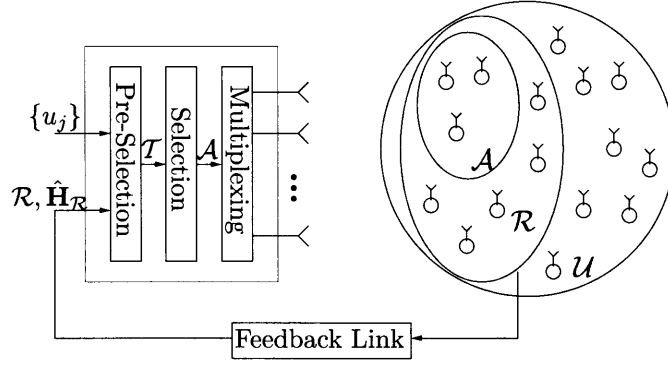


Figure 4-1. The MIMO system architecture of interest. In each scheduling interval a subset \mathcal{S} of the full user pool \mathcal{U} of size n reports quantization $\hat{\mathbf{H}}_{\mathcal{R}}$ of its channel gains to the transmitter via the feedback link using a decentralized (individual) criterion. From the set \mathcal{S} , the transmitter first forms a collection \mathcal{B} of candidate user sets of size m using a pairwise criterion; this is the pre-selection phase. Next, a set $\mathcal{A} \in \mathcal{B}$ is chosen at random as the active set, whose messages $\{u_j, j \in \mathcal{A}\}$ are linearly multiplexed across the array for transmission.

back,

$$\mathcal{R}_{\sigma,\rho} = \prod_{\ell=0}^{n_c-1} \mathcal{R}_{\sigma,\rho}^{(\ell)}$$

and in particular in the cardinality of $|\mathcal{R}_{\sigma,\rho}|$. In the sequel we let,

$$N_{\epsilon,\rho} = |\mathcal{R}_{\sigma,\rho}| \text{ and } N_{\epsilon,\rho}^{(\ell)} = |\mathcal{R}_{\sigma,\rho}^{(\ell)}|$$

be the random variable counting the number of users that feedback from the entire user pool and the ℓ th cluster respectively. If one too ambitiously prescribes a SNR target by ones choice of parameters for $\{\mathcal{R}_{\sigma,\rho}^{(\ell)}\}$ one may have a very high probability that $|\mathcal{R}_{\sigma,\rho}| < m$ or even worse a reasonable probability that $|\mathcal{R}_{\sigma,\rho}| = 0$. Alternatively, by choosing too lax of feedback thresholds for $\{\mathcal{R}_{\sigma,\rho}^{(\ell)}\}$ one may with high probability have every user from the feedback pool reporting their channel to the transmitter and hence in the case of finite rate feedback every quantization index may be reported multiple times. The former of these two scenarios (too ambitiously prescribing a SNR target) is easy for the transmitter to detect and correct as one may slowly decrease the feedback threshold for each cluster until the desired level of feedback is achieved. However, while the latter of these scenarios is easy to detect the appropriate action of the transmitter is less clear as it is unclear which part of the system to adapt. That is, in a system with finite rate feedback increasing the SNR target by a marginal amount may led to a dramatic decrease in the level of feedback spoiling the multi-node matching gain while increasing the rate of the quantizer may strain the feedback link and under exploit the order statistic gain. Hence in the following we consider a model for channel aware scheduling where one can examine the effects of variations in the feedback thresholds in a more general quality of service framework. Then, in the sequel we analyze the effects the variations in the feed back parameters effect the diversity of the user selection problem.

■ 4.2 An Introduction to Channel-Aware Scheduling

In the previous sections we have examined the effects that multi-user diversity has on the problem of feedback design. However, in a system where the number of users outnumbers the available system resources (the number of transmit elements) there is generally competition amongst the users for these resources. As such one must generally provide a scheduling mechanism to ensure some sort of fairness between these demands. In the present context, delay tolerant data arrives at the transmit array destined for some user terminal. As the number of users in the system is assumed greater than the number of transmit elements one must provide a scheduling mechanism which allocates the moments in time data may be transmitted to each user by the array. More generally, in multi-user MIMO system the network design problem concerns how to grant competing users access to the transmit array in order to meet system quality of service objectives.

To provide QOS functionality in a MIMO system, the current IEEE 802.16 Standard [1] provides five quality-of-service classes, three for real-time data connections and an additional two classes for delay tolerant data; one class which must be served with a guaranteed minimum throughput and an additional class with best effort service [1]. Thus, the base station of such a link must be able to provide support for data applications that have fundamentally different traffic and quality of service requirements than the real time data that has strict delay constraints. Hence, in a system where the channel has time varying fading it is attractive to use channel aware scheduling to improve throughput performance for the delay tolerant data. In particular, with delay tolerant data, one can opportunistically use the best channel available to transmit at as high a data rate as possible. As such, one would expect that the fraction of users with favorable channel conditions to have their services demands satisfied sooner. However, using such an approach the delay experienced by the fraction of users with poor channel conditions may be intolerable. Moreover, it may be impossible to meet minimum service levels for this fraction of users with such an opportunistic approach. Alternatively, if one uses a pure time division strategy to schedule users in an attempt to achieve some minimum service guarantee for each user the overall system throughput will be reduced as the proportion of time slots allocated to users with poor channel conditions must be increased to meet the minimum service level. Hence, in a fading channel one must, in general, forgo opportunistic as well as static scheduling if one wishes to balance minimum service level guarantees and the system throughput. A particularly attractive scheduling approach for fading channels to provide such guarantees is the proportionally fair [62, 75] scheduler, or, when only the overall stability, delay and throughput of the system is of interest, the max-weight scheduler [87, 118, 125] which are described in the sequel.

■ 4.2.1 Scheduling Policies for Multi-User MIMO Systems

In a multi-user MIMO system a channel aware scheduler must not only choose the subset of active users for transmission but also a power control policy to control the rate allocation to each of the users from the active subset. For example, in a system for which the power of a transmitted signal must stay below some limit the scheduler must choose how to allocate the power amongst messages for each user so that the resulting signal does not violate the given power constraint. However, optimizing the power control policy will have a negligible impact on our results and as we are primarily interested in the underlying dependence of the scheduler on the feedback design we do not optimize the power control policy. Instead, we assume a naive power control policy which allocates an equal fraction of the available

transmit power to each user. Thus, in the sequel the rate allocated to a user is only a function of each users channel gains and the co-channel interference caused by the other users in the activation set. As such, we denote the rate achieved by user i with an active set of users \mathcal{A} as $R_i(\mathcal{A})$. Further, we assume that the arrival process for each user in the system, $A_i[t]$ for $0 \leq i < n$, is a stationary and ergodic discrete time process describing the arrival of fixed size packets. We let $Q_i[t]$ for $0 \leq i < n$ be the length of the queue for user i and let $W_i[t]$ for $0 \leq i < n$ be the waiting time of the packet at the head of each users queue. With this identification, we now review some common scheduling policies.

In a system with no QOS guarantees it is often of interest to define the total system throughput as the relevant QOS metric. Such a metric yields a scheduler which maximize the system throughput by opportunistically choosing the subset of users that achieve the highest sum rate at each scheduling interval. That is the maximum sum rate (MSR) scheduler selects the set of users

$$\mathcal{A}^* \in \arg \max_{\mathcal{A} \subset \{0,1,2,\dots,n\}} \sum_{i \in \mathcal{A}} R_i(\mathcal{A}). \quad (4.3)$$

as the active set. However, as previously noted to balance minimum service level guarantees and the system throughput one in general needs to forgo such a opportunistic approach. With such a constraint the proportionally fair scheduler is often of interest. This scheduler is currently the default scheduler for the CDMA 1xEV-DO system [32, 62] and is also considered for High-Speed Downlink Packet Access (HSDPA) enhancement to the third generation (3G) mobile telephony protocol [26]. The weighted proportionally fair (WPF) scheduler, chooses the set of users

$$\mathcal{A}^* \in \arg \max_{\mathcal{A} \subset \{0,1,2,\dots,n\}} \sum_{i \in \mathcal{A}} \frac{\gamma_i}{A_i[t]} \cdot R_i(\mathcal{A}) \quad (4.4)$$

as the active set of users where $A_i[t]$ is the exponentially smoothed average service rate of user i ,

$$A_i[t+1] = \begin{cases} (1 - \alpha_{\text{PF}}) \cdot A_i[t] + \alpha_{\text{PF}} \cdot R_i(\mathcal{A}^*) & \text{if } i \in \mathcal{A}^* \\ (1 - \alpha_{\text{PF}}) \cdot A_i[t] & \text{otherwise} \end{cases} \quad (4.5)$$

and in turn where α_{PF} is a given constant such that $0 < \alpha_{\text{PF}} < 1$. While the WPF scheduler has been shown to maximize the sum of the logarithms of the long term average throughput of each user almost surely [132], it also has been shown to be unstable in high data rate systems [11]. Thus using such a scheduler there is no guarantee that all data will be transmitted in bounded time. To circumvent this deficiency one may use a maximum longest delay first (M-MLDF) schedule, which more generally takes the delay and/or queue state of each user into consideration. In particular, the generalized maximum longest delay first (GM-MLDF) scheduler chooses

$$\mathcal{A}^* \in \arg \max_{\mathcal{A} \subset \{0,1,2,\dots,n\}} \sum_{i \in \mathcal{A}} \gamma_i \cdot V_i[t] \cdot R_i(\mathcal{A}) \quad (4.6)$$

as the active set of users where $V_i[t]$ is a function of the queue length and delay for user i at time t . More precisely,

$$V_i[t] = \left(\alpha_{\text{MW}}^{(i)} \cdot Q_i[t] + (1 - \alpha_{\text{MW}}^{(i)}) \cdot W_i[t] \right)^{\beta_{\text{MW}}} \quad (4.7)$$

where in turn $0 \leq \alpha_{\text{MW}}^{(i)} \leq 1$ and $\beta_{\text{MW}} > 0$.

Examining (4.3),(4.4) and (4.6) we can see that the form of the scheduling problem in the WPF and GM-LWDF framework are not too different. In fact, all can be cast as a maximal weight matching problem where the time-varying weights are unity for the MSR policy, the inverse of the smoothed long term average throughput of each user in the WPF framework and a function of the weighted combination of the delay and queue state of each user in the GM-LWDF framework. More precisely, let

$$w_j[t] = \begin{cases} 1 & \text{for the MSR policy} \\ \gamma_j \cdot V_i[t] & \text{for the GM-LWDF policy} \\ \gamma_j/A_i[t] & \text{for the WPF policy} \end{cases} \quad (4.8)$$

Then, the scheduling problem that must be solved using the MSR policy, the GM-LWDF policy or the WPF policy is the determination of any set of users \mathcal{A}^* such that

$$\mathcal{A}^* \in \arg \max_{\mathcal{A} \subset \{0,1,2,\dots,n\}} \sum_{i \in \mathcal{A}} w_i[t] \cdot R_i(\mathcal{A}). \quad (4.9)$$

Hence, in order to understand the complexity of the channel aware scheduling problem in a multi-user MIMO system it is sufficient to understand how the channel variations and rates achievable in the physical layer effect the maximal weight matching problem (4.9).

In a system where the channel state is quantized and a static flat power allocation policy is used the region of achievable rates becomes discrete³. Moreover, in such a system the number and distribution of these discrete operating points is directly tied to the structure of the associated feedback scheme. As previously noted, the feedback scheme is the only knowledge the transmitter has of the channel state. Thus, the transmitter may only infer each users channel and the co-channel interference from the descriptions of users channels given by the feedback scheme. Hence, the transmitter may only allocate rate based on the discrete set of channel vectors used in the feedback scheme. This is a particularly useful observation as this implies that every time the channel changes state the set of possible operating points comes from some finite collection. Thus, one may construct efficient discrete structures and algorithms to aid user selection. In the absence of the co-channel interference this view point is quite familiar. Indeed, if one did not have to worry about the co-channel interference this problem would reduce to the problem of scheduling in a switch with time varying state [12]. In general, the interdependencies caused by the co-channel interference are strong enough that one requires a slightly more general switching framework to fully handle the channel aware scheduling problem from this discrete viewpoint. However, almost all insights needed in the sequel may be gained by considering this less general system. Moreover, the necessary generalization obfuscates these insights and thus before proceeding further to this generalization, we first consider modeling the channel aware scheduling problem by a input-queued cross-bar switch.

■ 4.2.2 A Discrete Model for Channel Aware Scheduling

The problem of complexity, throughput maximization and fairness for an input-queued cross-bar switch has been well studied and a more complete exposition can be found in [39, 87, 88] among others. For our purposes we only recall the basic definitions we require

³This statement uses our assumption that one excludes time sharing as a possibility in the physical layer.

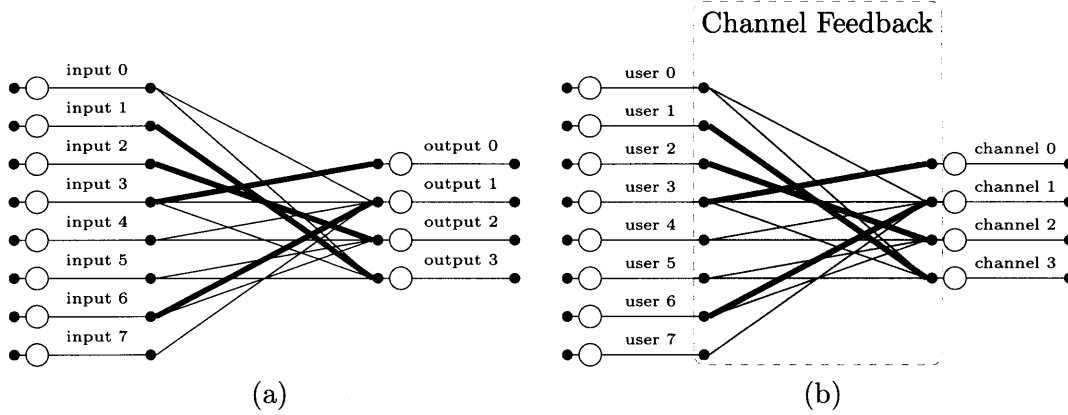


Figure 4-2. Two 8 input and 4 output input-queued cross-bar switches. The open circles at left represent input ports while the open circles at right represent output ports. (a) An edge from an input port to an output port represents a possible allocation and the bold edges represent a matching. (b) The same switch this time with the inputs representing users of a wireless system and the edges representing allowable allocation of physical channels which may or may not interfere with one another.

in the sequel as a more broad framework will be necessary to address quality of service in the problem of interest. An input-queued cross-bar switch with n inputs and m outputs is defined [87] to be an undirected graph $\mathcal{G} = (V, E)$ with vertex set V and edge set E where

1. The vertex set $V = V_i \cup V_o$ consists of the disjoint union of a set of n input vertices V_i and m output vertices V_o
2. The edge set E only consist of edges connecting the vertices in V_i to V_o

A *matching* in \mathcal{G} is a subset $\mathcal{M} \subset E$ of edges of E such that no two edges in \mathcal{M} have vertices in common. This can be seen in Figure 4-2.

To model the channel aware scheduling problem using the input-queued cross-bar switch one must find away to map the users and the time varying channel state to the inputs and edges. The simplest way one may do this is to view each input as a single-user in the system and attempt to represent the scheduling dependencies arising from the co-channel interference through the assignment of edges in the graph. At present we assume that this may be done in such a away that the rate allocated to users in such a switch does not depend on the other users selected in the matching. As such, one may associate a weight $w_{i,j}$ to the edge $(i, j) \in E$ equal to the reward one gets in the linear objective function representing the QOS constraint for assigning user i to slot j at the particular scheduling interval. More precisely, as the rate allocated to user i does not depend on the other users in the matching but rather the particular choice of output, (4.9) becomes

$$M^* \in \arg \max_{M \text{ matching in } \mathcal{G}} \sum_{i \in M} w_{i,j} \quad (4.10)$$

where

$$w_{i,j} = w_i[t] \cdot \mathbf{1}_{\{(i,j) \in E\}} \quad (4.11)$$

and in turn where $\mathbf{1}_{\{(i,j) \in E\}}$ is one if $(i, j) \in E$ and is zero otherwise. We let the weight of a matching in \mathcal{G} be the sum of the weights of the edges in the matching. Thus, the problem of user selection is equivalent to finding a matching of maximal weight in \mathcal{G} .

In the input queued crossbar switch model for channel aware scheduling the interdependencies arising from the channel realization were modeled through the edges in a bipartite graph. However, in a system with multiple transmit elements the problem of rate allocation depends on the state of the underlying channel realization as well as transmission power constraints and thus, at any (every) scheduling interval, the dependencies between users can be arbitrarily complex. As such, the dependencies which may be modeled by an input-queued cross-bar switch are not, at present, sufficient for our purposes to model scheduling in the multi-user MIMO channel. However, one may more adequately model the channel aware scheduling problem by making the role of finite rate feedback more apparent.

Our current model for the channel aware scheduling problem using the input-queued cross-bar switch (see for example Figure 4-2) implicitly incorporates a users channel realization and hence a users feedback through the edge set in the bipartite graph. As we have a fundamental interest in the order statistic gain and the multi-node matching gain tradeoff and the implications this tradeoff has on user selection we need to make the role of feedback more explicit. One relation between feedback and scheduling that is far too implicit in the current model is the constraint that users with common quantized channel vectors may not be scheduled concurrently. That is, in general if two users share a common quantized channel vector then one may ignore the user with the lowest weight when making a scheduling decision. In the input-queued cross-bar switch model this relation may be only modeled by requiring users that have common channel vectors to share common output ports and only be incident with one output port. This requirement over-constrains other relations which may be modeled in the switch and as such it is natural to consider a switch for which only the user of highest weighted is consider for every quantized channel vector as this allows one to model additional dependencies in rate scheduling inherent in the underlying channel. In this model, at each scheduling interval, the subset of users represented by the switch correspond to distinct codewords from the quantization codebook. As the codewords are the sole influence in the rate interdependencies for a set of users it is more natural to model the channel aware scheduling problem with finite rate feedback by assuming the quantized channel vectors are the inputs to the switch rather than individual users. This approach yields a switching model that is independent of the user population and allows one to understand the interactions between feedback design, channel statistics and greedy scheduling approaches.

In a MIMO system with finite rate feedback at each scheduling interval the random feedback from the users determines the associated achievable rates and hence the configurations of the switch. As an alternate model for the channel aware scheduling problem with finite rate one may consider an input-queued cross-bar switch where the input ports of the switch correspond to user feedback rather than the users themselves. Thus, at each scheduling interval the channel fluxuations randomly assign users to an input port based on their particular channel realization. At present we do not assume a particular model for this joint distribution as it is a function of both the feedback design as well as the channel statistics, but rather leave it arbitrary and refer to the joint probability distribution describing user assignment simply as the *user assignment distribution*. We further refer to any input as *occupied* if there is a user which has been assigned to the input and refer the distribution of occupied inputs as the *input occupancy distribution*. As the switch inputs correspond to codewords in the quantization codebook the edge set in the bipartite graph is independent of the channel realization and we refer to this deterministic graph as the *static switch*. Thus, at each scheduling interval, an arbitrary number of inputs may be occupied which in turn select an associated subset of edges from the static switch. We say that an edge $(i, j) \in E$ in

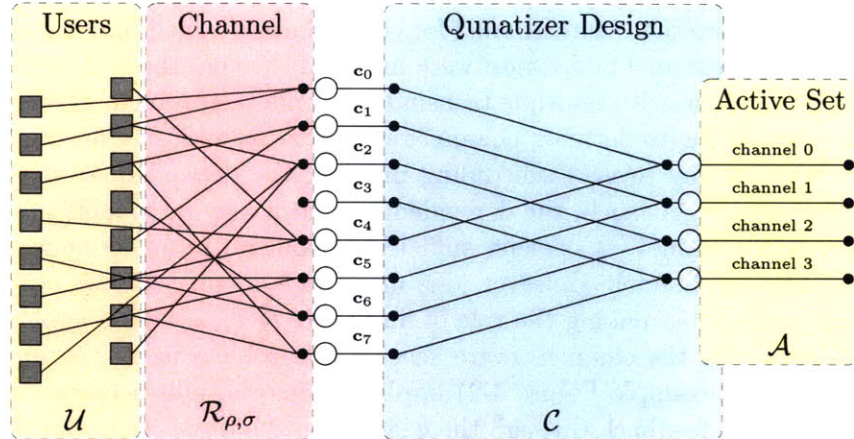


Figure 4-3. A depiction of the input-queued cross bar switch in which users are randomly assigned to switch inputs at each scheduling interval. We do not label the users and simply represent them as filled squares (seen at left). At each scheduling interval edges are randomly drawn from each user to the switch input which represents the users feedback at the scheduling interval.

the static switch is active if input i is occupied and thus the distribution of active edges at each scheduling interval are determined solely by the user assignment distribution and the structure of the static switch. We further say that an output j is occupied if there is an edge $(i, j) \in E$ such that input i is occupied and refer to the distribution of occupied outputs as the *output occupancy distribution*. Hence, there is an intricate connection between the input occupancy distribution and the output occupancy distribution which is described through the structure of the static switch. A depiction of this may be seen in Figure 4-3.

It is important to note that the relation between the occupancy distribution at the input is intimately tied to the occupancy distribution at the output through the structure of the static switch. Thus, a system designer may use the degrees of freedom in the quantizer design to not only develop efficient user selection structures but also to craft a static switch for which the output distribution and hence the scheduled rate is immune to variations of the input occupancy distribution. Clearly if the user assignment distribution causes sufficiently many input ports to be occupied with high probability the probability that a matching of maximal size exists is also trivially high regardless of the quantizer structure. Conversely, in a system where the number of input ports largely out numbers the number of users one in general may only have a small number of active edges at each scheduling interval leading to the possibly of a maximal matching of small size. However, as the system designer is able to design the feedback scheme, the system designer may structure the quantization scheme in an effort to ensure that a maximally sized matching may be found when only a fraction of input ports are occupied. Indeed, one may structure the quantizer in a way to pigeonhole⁴ the output occupancy distribution by imparting a structure on the quantization codebook so that only a subset of inputs must be occupied in order to guarantee that every output port is occupied. For example, examining Figure 4-3, one may see due to the structure of the switch, it is sufficient for any 7 input ports to be occupied to guarantee a maximally sized matching exists. Alternatively the switch in Figure 4-2 (b) needs all 8 inputs to be

⁴ Recall the pigeonhole principle states that if n items are put into m pigeonholes at least one pigeonhole must contain more than one item if $n > m$. More generally, if n items are placed in to m containers, then at least one container must hold $\lceil n/m \rceil$ items.

occupied to guarantee a matching of size 4 to exist as there is a strong dependence on input 3. Further inspecting Figure 4-3 one may see that the switch in Figure 4-3 is guaranteed to have 3 output ports active if any 5 input ports are occupied and 2 output ports active if any 3 input ports are occupied. Thus, this pigeonholing structure not only makes a system more immune to the number of occupied inputs but also variations in the input occupancy distribution itself as there is an inherent ability to exhibit the same output occupancy distribution for a large set of possible user assignment distributions. For the switch of Figure 4-3 one may see that the output occupancy distribution is invariant to any input occupancy distribution which fixes the probability that input i or input $i + 4$ is occupied for $i = 0, 1, 2, 3$. Thus, the static switch plays a strong role in determining how variations in the input occupancy distributions effect the output occupancy distribution.

Recall that the tradeoff between the order statistic gain and multi-node matching gain may be interpreted through a greedy rate scheduler, whereby users meeting an individual SNR target are first selected then the subset of users with the best co-channel interference were selected. Thus, in the present context the SNR target may yield a user assignment distribution that causes the distribution of the number of occupied inputs to be sufficiently small limiting the scheduler's ability to find matching of large weight and/or size. An output centered analysis has the added benefit of describing how variations in the input assignment distribution (the order statistic gain) effect the probability of a matching of maximal size (the multi-node matching gain). Thus, viewing the channel aware scheduling problem as a switch provides a framework in which one may understand the interplay between the channel fading statistics, the order statistic gain, the multi-node matching gain as well as the complexity of user scheduling. From this viewpoint there are two questions of interest. The first question concerns how variations in user assignment probabilities effect the occupancy distribution at the input, the second question concerns the relation between variations in the user assignment probabilities to the occupancy distribution at the output. We consider these questions further in Sections 4.3 and 4.4. However, the single input-queue cross bar switch described still does not model enough of the physical dependencies of the channel. In an attempt to generalize this model one could consider multiple separate switches, each of which describes a subset of achievable rates, and choose the best matching from among the results as the scheduling decision. However, the interdependencies that may be represented through a single input-queue cross bar switch are few leading to a need to consider a large number of switches to make the optimal scheduling decision in general. Thus, in the sequel we consider a slight generalization to this model which captures sufficiently many interdependencies of the channel and leads to efficient scheduling framework that allows one to similarly analyze the tradeoff in the order statistic gain and multi-node matching gain as well as the complexity of user scheduling.

■ 4.2.3 Channel Aware Scheduling as a Generalized Switch

To generalize the input-queued cross-bar switch model to the multi-user MIMO downlink one must find a way to relate the "switch state" in this model to the random and asynchronously varying state of the channel [12, 118]. In particular, one must introduce the interdependencies that arise from interference that is introduced by non-orthogonal channels and additional rate interdependencies that arise from transmission power constraints. We follow the direction of Stolyar and Andrews et. al. [12, 118] and view the problem as a generalized switch which we describe in the sequel.

In the sequel we refer to any (discrete) time varying collection of service vectors as a

generalized switch. More precisely, we let

$$\boldsymbol{\mu} = [R_0, R_1, \dots, R_{n-1}]$$

be a service rate vectors if the system can simultaneously allocate a rate R_0 to user 0, R_1 to user 1 and so on. Then, a generalized switch is simply the time varying collection

$$K[t] = \{\boldsymbol{\mu}_0[t], \boldsymbol{\mu}_1[t], \dots\}.$$

Of particular interest is a generalized switch which only assumes one of a finite set of states $\overline{\mathbf{M}}$ which form a Markov chain. Such a model fits in to the general framework of Stolyar [118] in which strong statements may be made about throughput and stability optimality. For each $\mathbf{m} \in \overline{\mathbf{M}}$ we associate a set of processing modes $K(\mathbf{m}) = \{k_{\mathcal{A}}\}$ which describe a service rate vector for the n users,

$$\boldsymbol{\mu}(k_{\mathcal{A}}; \mathbf{m}) = [R_0(\mathcal{A}), R_1(\mathcal{A}), \dots, R_{n-1}(\mathcal{A})]$$

where $R_i(\mathcal{A})$ is the rate allocated to user i in processing mode $k_{\mathcal{A}}$. In this context, the maxweight scheduling problem (4.9) is equivalent to determining a processing mode, $k_{\mathcal{A}}$, such that

$$k_{\mathcal{A}} \in \arg \max_{i \in K(\mathbf{m})} \sum_{j=1}^n \gamma_j \cdot w_j[t] \cdot \boldsymbol{\mu}_j(i; \mathbf{m}).$$

In order to reuse some of the results one has from scheduling in an input-queued cross-bar switch, it is natural seek an identification between the generalized switch and the less general input-queued cross-bar switch previously described. In this direction, note that every processing mode of a generalized switch corresponds to a matching in a trivial graph. Thus every processing mode is a trivial input-queued cross-bar switch. More precisely, consider an edge-less input-queued cross-bar switch with n inputs and m outputs. Then, by arbitrarily assigning every user that receives non-zero rate in a processing mode $k_{\mathcal{A}} \in K(\mathbf{m})$ to an arbitrary output port in this graph yields a trivial input-queued cross-bar switch that consist of a single matching. Clearly with this identification the associated edge weights for the input-queued cross-bar switch are, analogous to (4.11),

$$\omega_{i,j} = w_\ell[t] \cdot \boldsymbol{\mu}_\ell(k; \mathbf{m}) \cdot \mathbf{1}_{\{(i,j) \in E\}}$$

where we note that the edge set may be taken arbitrarily so long as the edge set is a matching and user ℓ has the largest weight of those users assigned to input i . An example of this may be seen in Figure 4-4. However, at present this particular identification of a processing mode with a trivial input-queued cross-bar switch does nothing to simplify the overall problem of user selection nor does it yield any insights to the tradeoffs between the order statistic gain and the multi-node matching gain. Examining Figure 4-4 it is natural to consider possible ways to add additional matching to this trivial input-queued cross-bar switch to describe other processing modes by cleverly assigning users to output ports. Indeed, if one may find subsets of processing modes which form an input-queued cross-bar switch one may employ standard matching algorithms used in an input-queued cross-bar switch on this subset of processing modes. However, due to the spatial structure of the channel feedback this may not be done in general. To illustrate this concept we provide the following example.

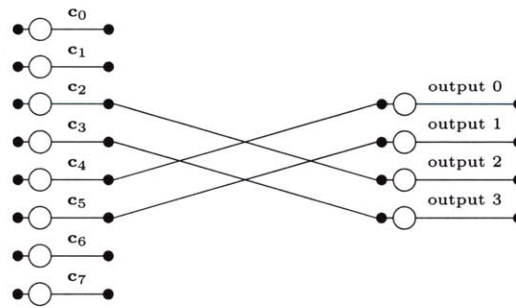


Figure 4-4. A single matching representing a given processing mode $k \in K(\mathbf{m})$. Note that the addition of any additional edge yields a matching which does not correspond to an orthogonal basis.

Example 4.2.1

In this example we show how one may not in general form a single input-queued cross-bar switch to represent the processing modes which correspond to users with orthogonal (quantized) channel vectors using the quantizer from Example 3.2.4. In this direction, we let $\mathcal{C}_0 = \{\mathbf{c}_0, \mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3, \mathbf{c}_4, \mathbf{c}_5, \mathbf{c}_6, \mathbf{c}_7\}$ be the quantization codebook where

$$\begin{aligned} \mathbf{c}_0 &= [1, 1, 0, 0], & \mathbf{c}_4 &= [1, \sqrt{-1}, 0, 0], \\ \mathbf{c}_1 &= [1, -1, 0, 0], & \mathbf{c}_5 &= [1, -\sqrt{-1}, 0, 0], \\ \mathbf{c}_2 &= [0, 0, 1, 1], & \mathbf{c}_6 &= [0, 0, \sqrt{-1}, 1], \\ \mathbf{c}_3 &= [0, 0, -1, 1], & \mathbf{c}_7 &= [0, 0, -\sqrt{-1}, 1] \end{aligned}$$

By some simple computation it is easy to see that the 8 codewords above form four orthonormal bases for \mathbb{C}^4 which are

$$\begin{aligned} \mathcal{B}_0 &= \{\mathbf{c}_0, \mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3\} & \mathcal{B}_1 &= \{\mathbf{c}_4, \mathbf{c}_5, \mathbf{c}_6, \mathbf{c}_7\} \\ \mathcal{B}_2 &= \{\mathbf{c}_0, \mathbf{c}_1, \mathbf{c}_6, \mathbf{c}_7\} & \mathcal{B}_3 &= \{\mathbf{c}_2, \mathbf{c}_3, \mathbf{c}_4, \mathbf{c}_5\} \end{aligned}$$

In an attempt to form an input-queued cross bar switch to represent the processing modes one may begin by mapping the basis \mathcal{B}_3 as follows:

1. \mathbf{c}_4 to output port 0,
2. \mathbf{c}_5 to output port 1,
3. \mathbf{c}_2 to output port 2,
4. \mathbf{c}_3 to output port 3

Now, as one may replace \mathbf{c}_4 and \mathbf{c}_5 in \mathcal{B}_3 with \mathbf{c}_0 and \mathbf{c}_1 and similarly as one may replace \mathbf{c}_2 and \mathbf{c}_3 in \mathcal{B}_3 with \mathbf{c}_6 and \mathbf{c}_7 one may attempt to form an input-queued cross bar switch to simultaneously describe these processing modes by adding these edges to the single matching in Figure 4-4. The resulting switch may be seen depicted in Figure 4-3. However, examining Figure 4-3 it is clear that there are matching which do not represent orthogonal bases. In particular, the matching corresponding to the inputs $\{\mathbf{c}_0, \mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_7\}$ does not define an orthogonal processing mode and thus there is not a consistent way to label edges to represent the orthonormal bases simultaneously in a input-queued cross bar switch.

For optimal scheduling in a multi-user MIMO system the relations between inputs and

outputs in an input-queued crossbar switch are too weak to capture the complex geometric structure required for channel aware scheduling with multiple-antennas. In order to identify the tradeoffs between the order statistic gain and multi-node matching gain and in order to identify how one may simplify the channel aware scheduling problem we wish to find a suitable structure in which one may simultaneously consider subsets of processing modes in an efficient manner. While the interdependencies between rate allocations that may be represented by a bi-partite graph are insufficient to represent the interdependencies required for channel aware scheduling with multiple-antennas a general undirected graph, in large part, is. That is, if one does not include a set of nodes distinguished as outputs one may describe many of the dependencies arising from co-channel interference through the assignments of edges in a general graph $\mathcal{G} = (V, E)$. In particular, we let an edge in \mathcal{G} represent a permissible pairing of codewords. In this setting a set of codewords may be scheduled simultaneously if and only if there is an edge between each codeword in \mathcal{G} . Any set subsets of vertices of \mathcal{G} such that every two vertices in the subset are connected by an edge is called a *clique*. Thus, to each vertex $i \in V$ one may associate a weight w_i representing the reward one gets in the linear objective function representing the QOS constraint by including the user with feedback associated to vertex i . We further let the weight of a clique be the sum of the weights of the vertices in the clique. Thus, the solution to the scheduling problem when restricted to the rate allocations represented by \mathcal{G} is equivalent to finding a maximally weighted clique in \mathcal{G} .

It may not be possible for a single graph to describe every possible processing mode for a given switch state. Indeed, analogous to what we have seen for the input-queued crossbar switch in Example 4.2.1 it may not be possible to consistently include cliques in a single graph that reflect valid processing modes. In the present scenario we require that every vertex in \mathcal{G} has some fixed weight that is independent of the choice of the clique containing it. Thus, one may only include cliques in a graph \mathcal{G} for which every vertex may be assigned a fixed weight. In this direction, we say that a set of processing modes $k_0, k_1, \dots, k_\ell \subset K(\mathbf{m})$ form a generalized switch if there exists a graph $\mathcal{G} = (V, E)$ such that every clique in \mathcal{G} corresponds to a subset of users receiving non-zero rate in one of the processing modes $k_0, k_1, \dots, k_\ell \subset K(\mathbf{m})$. For any set of processing modes $k_0, k_1, \dots, k_\ell \subset K(\mathbf{m})$ that form a switch, we denote the associated graph as $\mathcal{G}(\{k_i\}_{i=1}^\ell)$ and for a given switch state \mathbf{m} we denote the associated set of switches as

$$\mathcal{S}(K(\mathbf{m})) = \left\{ \{k_i\}_{i=1}^\ell \mid \{k_i\}_{i=1}^\ell \subset K(\mathbf{m}) \text{ form a generalized switch} \right\}$$

We note that for any generalized switch in $\mathcal{G} \in \mathcal{S}(K(\mathbf{m}))$ each vertex must have a fixed weight and hence a generalized switch \mathcal{G} may only contain processing modes for which every user assigned to given input has a fixed rate. If the rate allocations of a processing mode vary then one likely needs many generalized switches to represent every possible achievable rate, thus increasing the complexity of user selection. Thus, it is of interest to develop efficient multiplexing techniques which enable many processing modes of the system to be represented through a single generalized switch. For example, examining Example 4.2.1 one may see that every processing mode corresponding to users with orthogonal quantized channel vectors may be represented in a single graph using a flat power allocation. This is depicted in Figure 4-5.

The multi-user MIMO channel aware scheduling problem with finite rate feedback is equivalent to finding a maximally weighted clique from amongst the collection of graphs in

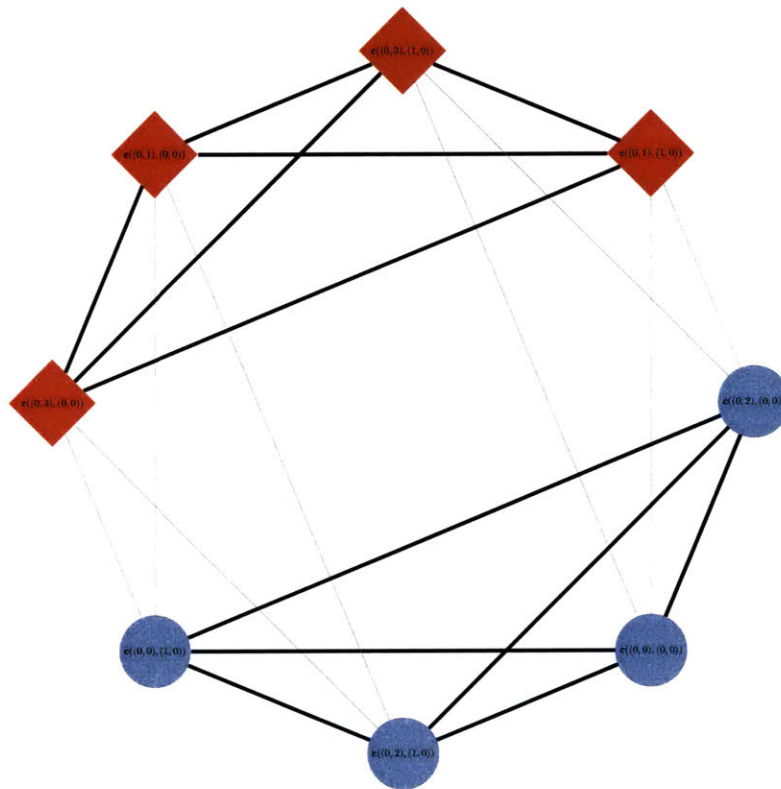


Figure 4-5. A depiction of the static generalized switch of Example 4.2.1 as a graph. The codevectors of Example 4.2.1 are the vertices and an edge is placed between any two vertices if the corresponding codevectors are orthogonal. The vectors of basis \mathcal{B}_1 are depicted as circles while the vectors of basis \mathcal{B}_2 are depicted with a diamond. Any clique in this graph corresponds to a processing mode employing a set of users with orthogonal quantized channel vectors.

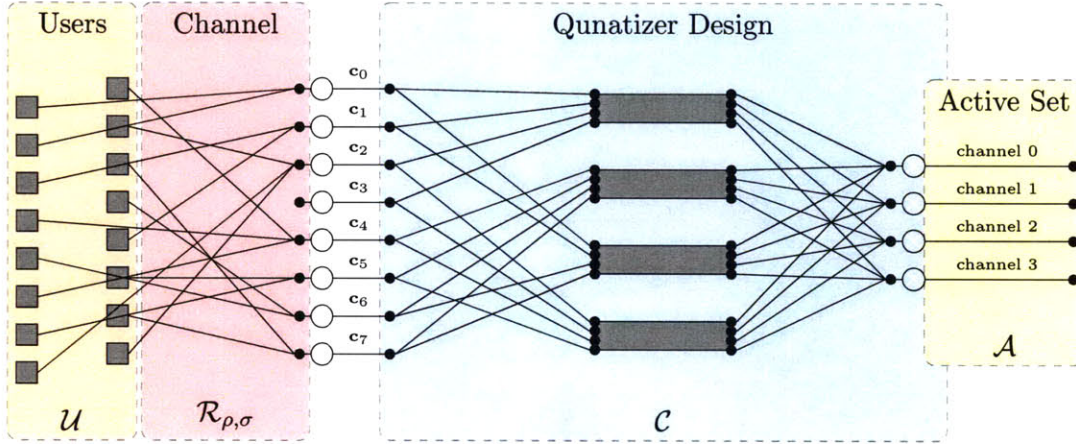


Figure 4-6. An alternate view of a generalized switch of Stolyar [118] in the case of finite rate feedback. At each scheduling interval users are randomly assigned to the inputs. Then, a maximal clique finding algorithm is run on to find the best clique.

$\mathcal{S}(K(\mathbf{m}))$. More precisely, the channel aware scheduling problem is equivalent to

$$S^* \in \underset{\substack{\mathcal{G} \in \mathcal{S}(K(\mathbf{m})) \\ S \text{ clique in } \mathcal{G}}}{\arg \max}}{\sum_{i \in S} w_i} \quad (4.12)$$

We note that (4.12) parallels (4.10). However, as there is no notion of an output present in the generalized switch we do not have an apparent way to relate the input occupancy distribution to the output occupancy distribution which was key in our development of the tradeoffs between the structure of the quantizer used for feedback, the order statistic gain and the multi-node matching gain in the input-queued crossbar switch.

Using a general graph to describe the interdependencies in rate allocations leaves the notion of an output absent. More precisely, by our definition of a generalized switch as a graph we have left implicit the fact that each clique describes a set of “output ports” in this generalized switch. To fully connect to our previous development for the input-queued cross-bar switch we must add an extra layer to the generalized switch in order to describe the multi-node matching gain. To do this, one may think of a generalized switch as a three tiered structure where the first tier describes the channel feedback, the second represents the cliques and the third and final tier represents the cardinality of scheduling decision of the switch and hence the multi-node matching gain. This may be seen in Figure 4-6. However, due to the complex structure of this switch it is unclear how one may analyze the order statistic gain and multi-node matching gain tradeoff due to the large dependency that arises from intersecting cliques. We do not cover this here and postpone the discussion to Section 4.4. Moreover, it is unclear the effects this more complex structure has on the overall scheduling complexity especially as one may have to consider more than one switch. We briefly address the problem of scheduling complexity in the sequel and provide a particular algorithm for scheduling in Chapter 6.

Our motivation for considering the generalized switch was to develop a scheduling framework that mitigates the $\binom{n}{m}$ complexity of examining the rates achieved by every user subset. Thus, it seems a bit unfortunate that one may have to consider more than one generalized switch for channel aware scheduling. However, for efficient channel aware scheduling we

note that one need not consider every graph in $\mathcal{S}(K(\mathbf{m}))$ but rather one only needs to consider the smallest subsets of $\mathcal{S}(K(\mathbf{m}))$ which contains every processing mode of $K(\mathbf{m})$ is as this set is sufficient to make the optimal scheduling decision. More precisely, we say that a collection of graphs $\mathcal{C}(\mathbf{m}) \subset \mathcal{S}(K(\mathbf{m}))$ covers $K(\mathbf{m})$ if

$$K(\mathbf{m}) \subset \bigcup_{\mathcal{G} \in \mathcal{C}(\mathbf{m})} \mathcal{G}.$$

In order to efficiently solve the channel aware scheduling problem it is sufficient to find a small cover of $\mathcal{S}(K(\mathbf{m}))$ for every switch state (or equivalently channel realization). More precisely, one may rewrite (4.12) as

$$S^* \in \underset{\substack{\mathcal{G} \in \mathcal{C}(\mathbf{m}) \\ S \text{ clique in } \mathcal{G}}}{\arg \max} \sum w_i \quad (4.13)$$

However, one typically does not wish to compute this cover for each channel realization as this process is as difficult as optimal user selection in general. Rather, one would like to find a minimal cover for a “global” set of processing modes

$$K_{\text{global}} = \bigcup_{\mathbf{m} \in \bar{\mathbf{M}}} K(\mathbf{m})$$

and use what ever (random) subset of this cover needed to solve the problem. In particular, as the service rates, and hence the switch state, is governed by the descriptions of the channel vectors that were fed back in a multi-user MIMO system with finite rate feedback, one may first minimally decompose a “global” switch state that contains every rate allocation for the feedback scheme. Then, for each channel realization one may use a subset of this minimal decomposition of the global switch state to cover the processing modes for the particular channel realization. More precisely, for any feedback scheme we can consider decomposing the processing modes into a minimal set of generalized switches offline. Then, for every channel realization the user assignment process randomly chooses the switches that must be considered to make the optimal rate allocation. In summary, the channel-aware scheduling problem in a multi-user MIMO system with finite rate feedback can be considered as follows:

1. A minimal cover of the global switch state is computed off line
2. Each time the channel changes state users are randomly connected to input ports in the minimal cover which determine the possible rate allocations for every subset of users
3. For each switch the maximal (weighted) clique is determined
4. The maximal-maximal (weighted) configuration is chosen from amongst the switches which identifies the active set of users

We will refer to above as the best random server (BRS) process. A depiction of the BRS process can be seen in Figure 4-7.

Reexamining the BRS process it is easy to see how one may provide efficient algorithms for the channel-aware scheduling problem. Indeed, it is easy to see that the channel aware

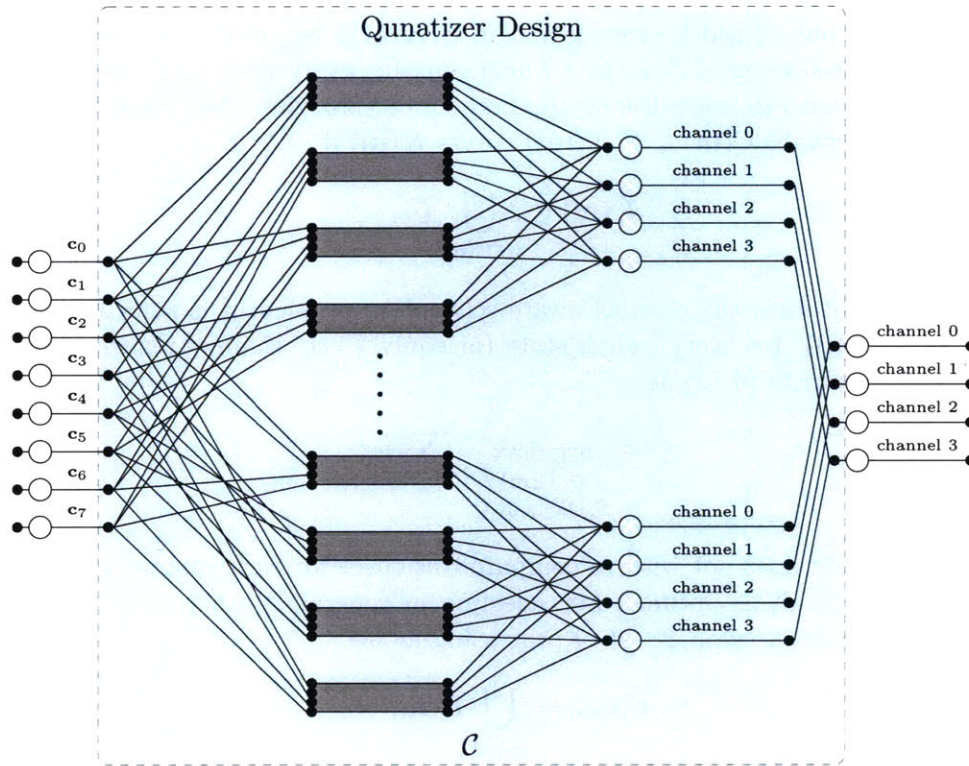


Figure 4-7. An alternate view of Stolyar’s generalized switch [118] for channel aware scheduling with finite rate feedback as a best random server process. At each scheduling interval users are randomly assigned to inputs of the switch. A maximum weighted clique algorithm is run on each switch independently to find the best clique. Then, the best switch and it maximal clique are selected.

scheduling problem is no more complex than multiple runs of existing clique finding algorithms on the random number of switches determined by the channel realization. However, it is unclear how many switches this naive approach will have to consider as the feedback from the channel realization drives the process determining the state of the generalized switch. In the worst case the statics of the fading process, and hence the user assignment distribution, may necessitate examining every switch. However, as the structure of the feedback scheme governs the possible rate allocations, a system designer may design the feedback scheme as to minimize the number of switches needed in the BRS model.

The overall complexity of scheduling using the naive approach of the BRS process is a function of both the fading process as well as the feedback design. Thus, one may by adding extra structure to the quantization codebook reduce the cardinality of the minimal cover of the global set of processing modes. However, if a system designer has any knowledge of the larger *structure* of the cover of the global switch state then it may be possible to employ a more intelligent algorithm to further reduce the system complexity. In particular, there is no guarantee that two switches in a minimal cover have zero intersection. For any switches that share processing modes or contain multiple similar processing modes it may be possible to use the intermediate results of a previous clique algorithm on a different switch or fully exclude a set of switches. For example, given a maximal clique from one switch it may be possible to use this state as a starting point for a maximal weighted clique algorithm on a different switch or used to fully exclude all cliques of a switch without search.

In particular, if the quantizer contains multiple orthogonal bases with common codewords then one may use a maximal clique on one switch as the starting point of another. Hence, in a MIMO channel with finite rate feedback the *structure* of the feedback is intimately tied to the complexity of optimal user selection and should be considered as a factor in feedback design. This observation is somewhat orthogonal to the motivation of feedback schemes that are presently being designed for the MIMO downlink [82, 90, 91, 105, 137, 144] which, as previously noted, advocate a feedback design which involves minimizing some increasing function of the mean square quantization error alone as the relevant design metric. In particular, the current feedback schemes adopted in the IEEE 802.16 standard [1, 143] employ such an MSE centric design.

We have identified a discrete model for the problem of channel aware scheduling. In particular, we have identified the input and output occupancy distributions as the relevant metric for the performance and complexity in a multi-user MIMO system. Moreover, we have provided a system architecture that relates directly to this model. As such, it is of interest to understand the influence ones choice of feedback parameters has on these distributions as it will influence the overall performance of the system.

■ 4.3 Optimization of the Input Occupancy Distribution

In a MIMO system with finite rate feedback it is important that the thresholds on the SNR target are set appropriately as otherwise one may not fully exploit the multiuser diversity of the system. That is, the SNR target may cause either too few users to feed back, limiting the scheduler's ability to find high weight matching, or too many users to feed back thus not sufficiently exploiting the order statistic gain. In the sequel we do not consider the choice of particular feedback parameters to select for each cluster but rather consider the question of determining exactly how the number of users that feedback from each cluster influences the system's ability to exploit the multiuser diversity. As the choice of feedback parameters differs between clusters in general in the sequel we only consider the effects of the SNR threshold on a single cluster.

We consider a single cluster of k users and denote this cluster as ℓ . We let $\mathbf{N}_i^{(\ell)}(k)$ be the random number of users from the cluster ℓ that feedback the codeword \mathbf{c}_i at the scheduling interval of interest and let $\mathbf{X}_i^{(\ell)}(k)$ be the random variable that is 1 if $\mathbf{N}_i^{(\ell)}(k) \geq 1$ and zero otherwise. More precisely,

$$\mathbf{X}_i^{(\ell)}(k) = \begin{cases} 1 & \text{if } i \in \mathcal{R}_{\sigma, \rho}^{o, (\ell)} \\ 0 & \text{otherwise} \end{cases}$$

It is clear that in order to fully exploit the multi-node matching gain one would like the random variable,

$$\mathbf{Y}^{(\ell)}(k) = \sum_{i=0}^{2^r-1} \mathbf{X}_i^{(\ell)}(k),$$

which counts the number of distinct quantization indices that are fed back to be large. However, $\mathbf{Y}^{(\ell)}(k)$ is implicitly a function of the underlying choice of the SNR target and thus is strongly influenced by one's choice of feedback thresholds. For the scheduler to exploit the multi-node matching gain to the fullest extent one would like the user assignment distribution to, with high probability, assign the users to a multitude of inputs. In particular, one would like $\mathbf{Y}^{(\ell)}(k)$ to be a modest fraction of the nodes that feed back to ensure that

there is a reasonable probability of a matching of large weight and size. However, as stated, there is a trade off between the order statistic gain and the multi-node matching gain and one must balance the effects an increase or decrease of the SNR threshold has on the distribution of $\mathbf{Y}^{(\ell)}(k)$. Clearly, when the users of a cluster have channel vectors that are isotropically distributed the expected value of $\mathbf{Y}^{(\ell)}(k)$ should see a modest increase as the SNR threshold is decreased so long as

$$\mathbb{E} \left[\mathbf{Y}^{(\ell)}(k) \right] \ll \min\{2^r, k\}.$$

However, in a correlated channel $\mathbf{N}_i^{(n)}(k)$ may become concentrated on a subset of codewords and hence regardless of the variation in the SNR threshold there may be little variation in $\mathbb{E} \left[\mathbf{Y}^{(\ell)}(k) \right]$ even when $\mathbb{E} \left[\mathbf{Y}^{(\ell)}(k) \right] \ll \min\{2^r, k\}$. That is, in a highly correlated channel any increase or decrease in an SNR threshold may not be able to compensate for the underlying correlation in the channel and one needs to adapt the feedback framework. Thus, it is of interest to understand when the underlying spatial correlation of the users in the cluster causes the expected value of $\mathbf{Y}^{(n)}$ to halt after a very small number of users from the cluster feedback.

In order to characterize when an increase in the SNR target has diminishing returns in the cumulative distribution of the random variable $\mathbf{Y}^{(\ell)}(k)$ we consider a sequential occupancy problem where by users are continually added to a cluster until the distribution of $\mathbf{Y}^{(\ell)}(k)$ becomes roughly constant. We assume in this scenario that no SNR target has been set to study the effects variations in the number users that feedback has on the random variable $\mathbf{Y}^{(\ell)}(k)$. In particular, we study the evolution of the density of $\mathbf{Y}^{(\ell)}(k)$ as a function of k . Such an approach lets one understand how changes in the SNR threshold effects the number of occupied inputs in the generalized switch. We would like to know the smallest number of users, say k_0 , such that the addition of more users in the cluster does not dramatically alter the probability that more than some fixed number, say n_0 , of inputs are occupied. That is, one would like to know for what value of k_0 is

$$\Pr[\mathbf{Y}^{(\ell)}(k_0) \geq n_0] \approx \Pr[\mathbf{Y}^{(\ell)}(k_0 + \Delta) \geq n_0]$$

for small values of Δ .

To make this precise, let \mathbf{V}_r be the random variable that counts the number of nodes required to be added to the system until r nodes are quantized to previously used quantization indices. That is, \mathbf{V}_r is a stopping rule with respect to the decision rules

$$\mathbf{I}_k^{(r)} = \mathbf{1}_{\left\{ \sum_{i=0}^{2^r-1} \mathbf{N}_i^{(\ell)}(k-1) - \mathbf{Y}^{(\ell)}(k-1) < r \right\}}$$

where $\mathbf{1}_A$ is the indicator function of the event A . Intuitively speaking, \mathbf{V}_r stops when a fraction of $r/(k-1)$ of the users in the cluster have been assigned to previously occupied inputs. In order to optimize the SNR target we would like to know the largest k can be such that there is a large probability that a small fraction of user are redundant. That is, the smallest k such that for a given α

$$\Pr[\mathbf{V}_{\alpha k} \leq (1 + \alpha)k] \approx 1.$$

We note that this definition for the trade-off is particularly useful as it accounts for spatial correlation in the fading process of the clusters. That is, this definition allows for $\mathbf{N}_i^{(\ell)}(k)$ to

become concentrated on a subset of inputs. In such a case $\mathbf{V}_{\alpha k}$ stops with high probability for very small values of k . Alternatively, $\mathbf{V}_{\alpha k}$ stops for large k with high probability in an isotropically distributed channel.

Analysis of $\mathbf{V}_{\alpha k}$ for a given user assignment distribution of a cluster indicates when the SNR threshold should be increased or decreased to allow more or less user to feedback and when the current quantization scheme needs to be adapted to more fully exploit the multiuser diversity. In this direction we say that a cluster has a quantization order of $n_\delta(\alpha)$ if $n_\delta(\alpha)$ is the smallest positive integer such that

$$\Pr[\mathbf{V}_{\alpha n_\delta} \geq (1 + \alpha)n_\delta(\alpha)] \leq \delta. \quad (4.14)$$

Note, that if a cluster has a quantization order of $n_\delta(\alpha)$ then with high probability there are no more than $n_\delta(\alpha)$ occupied inputs in the generalized switch when $(1 + 2\alpha)n_\delta(\alpha)$ users feedback. Thus, as there is a negligible probability that feedback from more than $(1 + 2\alpha)n_\delta(\alpha)$ users will yield more than $n_\delta(\alpha)$ occupied inputs one should design the SNR threshold no make sure that the expected number of user that feedback is not too much greater than $(1 + 2\alpha)n_\delta(\alpha)$.

■ 4.3.1 The Quantization Order and Input Occupancy Distribution

The quantization order $n_\delta(\alpha)$ may be used to determine how well a system is exploiting the order statistic gain and the multi-node matching gain. When the number of users who feed back their channel measurement becomes too low (relative to $n_\delta(\alpha)$) then a system is too aggressively setting the SNR target for the order statistic gain. When the number of users feeding back their channel measurement is too great (relative to $n_\delta(\alpha)$) indicates that the system has not exploited the order statistic gain to the fullest. We say that the *multi-node matching gain is saturated* if

$$\mathbb{E}[|\mathcal{R}_{\sigma,\rho}|] \gg (1 + 2\alpha)n_\delta(\alpha)$$

and, if

$$\mathbb{E}[|\mathcal{R}_{\sigma,\rho}|] \ll (1 + 2\alpha)n_\delta(\alpha)$$

we say that a cluster *is order statistic gain centered*. If a system is neither multi-node matching gain saturated nor order statistic gain centered we say the system is *balanced*. It should be clear that one prefers a system to be balanced if one hopes to fully exploit the order statistic gain and multi-node matching gain. However, the quantization order $n_\delta(\alpha)$ is a function of the parameters δ and α which should be set by the system designer to reflect a particular systems bias toward a high order statistic gain or high multi-node matching gain. In particular, for large values of α the definition of $n_\delta(\alpha)$ becomes biased toward an multi-node matching gain saturated system while small values of δ correspond to a order statistic gain centered design. Thus, the quantization order may be used to reflect a system designers preference of system balance.

Understanding when a system is order statistic gain centered, multi-node matching gain saturated or balanced has dramatic effects on the overall system design. In particular, given that a system designer has targeted a design to have, say n_{fb} , users feedback on average the quantization order can be used to determine the minimal quantization rate needed to ensure that the system is balanced. Alternatively, given a particular feedback bandwidth constraint and a fixed quantizer resolution, i.e. for a fixed r , the quantization

order can be used to determine an appropriate choice of the feedback parameters $\rho_-^{(\ell)}, \rho_+^{(\ell)}$ and $\sigma^{(\ell)}$ so that a reasonable fraction of the feedback set $\mathcal{R}_{\sigma,\rho}^{(\ell)}$ are useful in the process of user selection. That is, if a cluster has a quantization order of $n_\delta(\alpha)$ for an appropriately chosen δ and α then with high probability no more than $(1 + 2\alpha)n_\delta(\alpha)$ users are useful at the transmitter. Thus, if

$$\Pr(|\mathcal{R}_{\sigma,\rho}^{(\ell)}| > (1 + 2\alpha)n_\delta(\alpha)) \gg 0,$$

the feedback parameters $\rho_-^{(\ell)}, \rho_+^{(\ell)}$ and/or σ can be decreased without effecting the multi-node matching gain. A particularly attractive solution is the choice for ρ and σ such that

$$\Pr((1 + 2\alpha)n_\delta(\alpha) - \Delta \leq |\mathcal{R}_{\sigma,\rho}^{(\ell)}| \leq (1 + 2\alpha)n_\delta(\alpha) + \Delta) \approx 1$$

for some small positive value of Δ as this ensures that the resulting system is balanced. Note that if the number of users in a cluster is large, say k , $\mathbb{E} [|\mathcal{R}_{\sigma,\rho}^{(\ell)}|] = kp_{\sigma,\rho}$ and there is an exponentially small probability that $|\mathcal{R}_{\sigma,\rho}^{(\ell)}|$ deviates greatly from $\mathbb{E} [|\mathcal{R}_{\sigma,\rho}^{(\ell)}|]$. Thus, when the number of users in a system scales a system is balanced when

$$|\mathcal{R}_{\sigma,\rho}^{(\ell)}| = kp_{\sigma,\rho} \approx (1 + 2\alpha)n_\delta(\alpha)$$

and one must choose

$$p_{\sigma,\rho}(k) \propto \frac{(1 + 2\alpha)n_\delta(\alpha)}{k}.$$

Hence, for a fixed quantization scheme if $p_{\sigma,\rho}(k) = o(1/k)$ the system is asymptotically order statistic gain centered and if $1/k = o(p_{\sigma,\rho}(k))$ the system is asymptotically multi-node matching gain saturated. This distinction is important as in a multi-node matching gain saturated system the order statistic gain decouples from the multi-node matching gain trivially as one has extra degrees of freedom in the choice of the quantization scheme. This is an important observation as in a system with a fixed feedback bandwidth constraint and multiple users the order statistic gain decouples from the multi-node matching gain trivially and the system designer is afforded extra degrees of freedom in the feedback design.

It is of practical relevance to characterize the quantization order as a function of the quantization rate as well as the user assignment distribution of a cluster as it identifies several relevant system regimes, some of which require the system to adapt the feedback scheme to fully exploit the multi-user diversity. However, before proceeding we note that the random variable \mathbf{V}_r is by definition the *complimentary waiting time distribution* of the occupancy distribution [33]. That is, as \mathbf{V}_r stops when a fraction of $r/(k - 1)$ of the users in the cluster have been assigned to previously occupied inputs one has

$$\mathbf{V}_r \leq k \text{ if and only if } k - \mathbf{Y}(k) \geq r. \quad (4.15)$$

This identification is important as waiting time distributions of combinatorial processes are known to exhibit rather sharp phase transitions [65]. That is, if one examines the evolution of the probability of an event as a function of the number trials it is often the case that the probability distribution rapidly transitions from 0 to 1 [65]. The most common example of this phenomenon is the binomial random variable.

In the context of a phase transitions the definition of the quantization becomes a bit more clear. The quantization order simply defines, for a distribution that transitions continuously

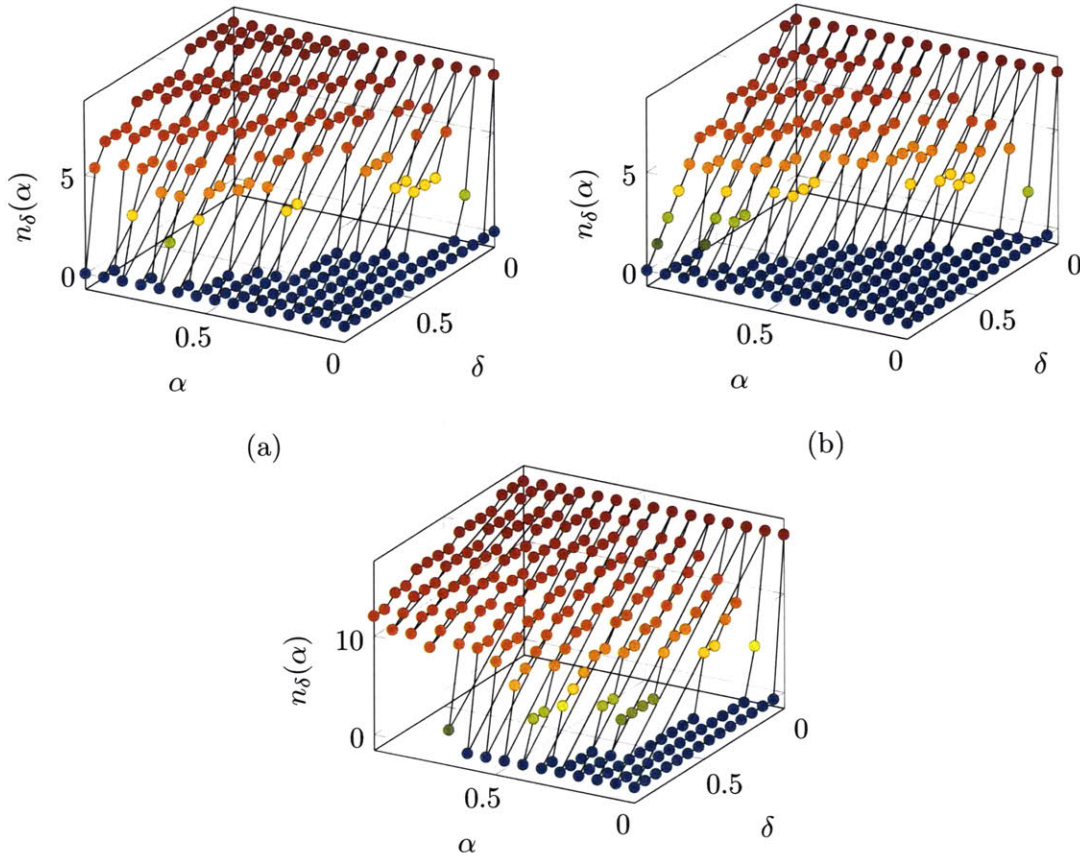


Figure 4-8. The quantization order, $n_\delta(\alpha)$, as a function of δ and α for (a) $\mathbf{p} = \mathbf{p}_{\text{unif}}$ of length 8, (b) $\mathbf{p}_i \propto 1/(i+1)$ of length 8, (c) $\mathbf{p} = \mathbf{p}_{\text{unif}}$ of length 16. The quantization order simply defines, for a distribution that transitions continuously from 0 to 1, by way of δ , the point on the “step” from 0 to 1 one wishes to operate while, by way of α how sharp this step should be. One may see that for a code of length 8 the quantization order rapidly jumps from 0 to 5 when $\alpha + \delta$ is approximately greater than $1/2$ for a uniform distribution as seen in (a). Moreover, (a) is strictly larger than a non-uniform distribution as seen in (b). Further, the same trend occurs for a length 16 code where the quantization order jumps up rapidly for $\alpha \gg 0$ and $\beta \gg 0$.

from 0 to 1, by way of δ , the point on the “step” from 0 to 1 one wishes to operate while, by way of α how sharp this step should be. Thus, it is reasonable to expect that for any choice of δ and α the quantization order may be quite small for modest sized feedback schemes and one will find themselves in the regime where the multi-node matching gain is saturated and the order statistic gain decouples from the multi-node matching gain. This can be seen in Figure 4-8.

However, to make this precise we must know the distribution of \mathbf{V}_τ . In this direction, by way of (4.15), we have the following lemma.

Lemma 4.3.1. *Let $\mathbf{p}^{(\ell)}$ be the user assignment distribution for an r -bit quantizer. Then,*

$$\Pr[\mathbf{V}_{n_1} \leq n_2 \mid \mathbf{p}^{(\ell)}] = \Pr[\mathbf{Y}^{(\ell)}(n_1 + n_2) \leq n_2 - n_1 \mid \mathbf{p}^{(\ell)}]$$

By Lemma 4.3.1 it is sufficient to study the behavior of the distribution of $\mathbf{Y}^{(\ell)}(k)$ in order to characterize the quantization order. In this direction, recall the distribution of the

random variable $\mathbf{Y}^{(\ell)}(k)$ is [33]

$$\Pr(\mathbf{Y}^{(\ell)}(k) \leq y) = \sum_{j=0}^y (-1)^{y-j} \binom{2^r - j - 1}{y - j} S_{2^r, 2^r - j}(\mathbf{p}; k) \quad (4.16)$$

where

$$S_{m_1, m_2}(\mathbf{p}; k) = \sum_{0 \leq j_1 < j_2 < \dots < j_{m_1 - m_2} \leq m_1 - 1} (p_{j_1} + p_{j_2} + \dots + p_{j_{m_1 - m_2}})^k.$$

Note (4.16) is only a function of the distribution $\mathbf{p}^{(\ell)}$ and implicitly the cardinality of \mathcal{C} . Hence, in order to examine the effects the quantizer has on the multi-node matching gain it is sufficient to examine the effects that the distribution $\mathbf{p}^{(\ell)}$ has on (4.16). In this direction, recall that a vector \mathbf{p} majorizes the vector $\mathbf{q} = (q_1, \dots, q_k)$ if, after possible reordering,

$$\sum_{i=1}^r p_i \geq \sum_{i=1}^r q_i \quad \forall r = 1, \dots, k,$$

and $\sum_{i=1}^k p_i = \sum_{i=1}^k q_i$. If \mathbf{p} majorizes the vector \mathbf{q} we write $\mathbf{p} \succeq \mathbf{q}$. Further, recall that a function, say $f(\mathbf{p})$, is Schur convex if,

$$f(\mathbf{p}) \geq f(\mathbf{q}) \quad \forall \mathbf{p} \succeq \mathbf{q}.$$

We now have the following lemma from [94].

Lemma 4.3.2. *The distribution $\Pr[\mathbf{Y}^{(\ell)} \leq k; \mathbf{p}]$ is Schur convex in \mathbf{p} for any $k \geq 0$.*

By Lemma 4.3.2 one can derive upper (resp. lower) bounds on the distribution of $\mathbf{Y}^{(\ell)}(k)$ so long as one can find distributions \mathbf{p}_u (resp. \mathbf{p}_l) that majorizes \mathbf{p} (resp. that is majorized by \mathbf{p}). In this direction, let, for any probability vector \mathbf{p} of length 2^r ,

$$\mathbf{p}_{\text{unif}} = \underbrace{\left(\frac{1}{2^r}, \frac{1}{2^r}, \dots, \frac{1}{2^r} \right)}_{2^r \text{ times}}$$

and

$$\mathbf{p}_{\text{min}} = \underbrace{(p_{\text{min}}, p_{\text{min}}, \dots, p_{\text{min}})}_{2^r - 1 \text{ times}}, 1 - (2^r - 1)p_{\text{min}}$$

where in turn $p_{\text{min}} = \min_{0 \leq i \leq 2^r - 1} p_i$. Clearly, $\mathbf{p}_{\text{unif}} \preceq \mathbf{p} \preceq \mathbf{p}_{\text{min}}$. Thus a uniform user assignment distribution always provides an lower bound on $\Pr[\mathbf{Y}^{(\ell)} \leq k]$ and thus an upper bound on (4.14). In particular, by Lemma 4.3.1 and Lemma 4.3.2, one has

$$\Pr[\mathbf{V}_{\alpha n_\delta} > (1 + \alpha)n_\delta(\alpha); \mathbf{p}] = 1 - \Pr[\mathbf{Y}^{(\ell)}((1 + 2\alpha)n_\delta(\alpha)) \leq n_\delta(\alpha); \mathbf{p}] \quad (4.17a)$$

$$\leq 1 - \Pr[\mathbf{Y}^{(\ell)}((1 + 2\alpha)n_\delta(\alpha)) \leq n_\delta(\alpha); \mathbf{p}_{\text{unif}}] \quad (4.17b)$$

$$= \Pr[\mathbf{V}_{\alpha n_\delta} > (1 + \alpha)n_\delta(\alpha); \mathbf{p}_{\text{unif}}] \quad (4.17c)$$

This yields the more general theorem.

Theorem 4.3.3. *The quantization order $n_\delta(\alpha; \mathbf{p})$ is a Schur concave function of \mathbf{p} .*

Theorem 4.3.3 is a particularly useful theorem as one may study the problem that the order statistic gain decouples from the multi-node matching gain by considering a uniform

distribution for \mathbf{p} which greatly simplifies the analysis. In particular, in Section 4.5 we show that assuming the Rayleigh model for the MIMO channel and hence a uniform distribution for the user assignment distribution if one uses a quantizer with isometric Voronoi cells, that the order statistic gain decouples from the multi-node matching gain in the large user limit. Moreover, this theorem states that when the channel is correlated it is even more likely that the order statistic gain decouples from the multi-node matching gain provided that the covariance structure is not sufficiently mismatched causing few users to feed back. If this is the case one further expects that $n_\delta(\alpha)$ is sufficiently small so that the system would benefit greatly from *adapting* the quantization scheme to more adequately match the covariance of the channel. In Chapter 5 we develop a systematic framework which provides methods to match the feedback codebook to the covariance of each cluster of users, approximately whitening the sampling probabilities. Hence, it will be sufficient to use \mathbf{p}_{\min} and \mathbf{p}_{unif} to bound the quantization order.

We have exhibited how the channel statistics and the user assignment distribution effect the input occupancy distribution of the generalized switch, thus answering our first question of interest. While these insights are sufficient to optimize the order statistic gain and the multi-node matching gain trade off we still have a question on how the output occupancy is effected by this statistic. Thus, we next consider this question.

■ 4.4 Analysis of the Output Occupancy Distribution

In this section we examine the influence of the input occupancy distribution on the output occupancy distribution. To ease the exposition in the sequel we assume that scheduler does not use knowledge of the classification of the users. That we assume that the scheduler does not explicitly use the classification of the users, but rather only forms the weighted average over all clusters⁵, as

$$\bar{\mathbf{p}}_i = \sum_{\ell=0}^{n_c-1} \frac{|\mathcal{U}^{(\ell)}|}{n} \hat{\mathbf{p}}^{(\ell)}.$$

In the sequel we simply denote this probability as \mathbf{p} .

In the preceding section we showed that with very mild assumptions on the size of the user pool one may, through examining the quantization order, determine when and if the order statistic gain decouples from the multi-node matching gain in a given system. A particularly useful results was that the quantization order is a Scours concave function of the user assignment distribution and hence correlation in the fading process only decreases the quantization order. This implies a correlated channel reduces the number of users that need to feedback for the order statistic gain to decouple from the multi-node matching gain. However, the knowledge that the order statistic gain decouples from the multi-node matching gain does not imply that the system achieves a high rate. Rather, it indicates that the system designer has added degrees of freedom in the quantization design. Indeed, if the user assignment distribution leads to only a few inputs of the generalized switch to be occupied at each scheduling interval, while the multi-node matching gain is saturated, there is a high probability that the scheduler has few candidate sets of users, leading to a multi-node matching gain and may result in poor system performance. That is, it is possible that while the multi-node matching gain is saturated the underlying channel correlation has thrust the system in to a quite unfavorable position. Indeed, as we saw in Section 4.2.3 if

⁵We note that this distribution has a statistical relevance which we discuss in Section 5.1.

there is not sufficient structure in the quantizer the system may have an underlying bias to a particular input that makes the system more susceptible to correlation. However, by “pigeonholing” the output distribution we showed that one may develop a system that is invariant to a large number of fading distributions. These observations led to our second and last question regarding the order statistic gain and the multi-node matching gain trade-off in the generalized switch. In particular, we are interested in the relation between variations in the user assignment distribution to the occupancy distribution at the output as a function of the structure of the switch.

In this section we provide methods for this analysis and identify the relevant aspects of feedback design that are needed to make the output distribution immune to a wide range of spatial correlation structures. In particular, we provide a direct relationship between the output distribution of a single generalized switch and the structure of the switch itself. Recall from Section 4.2 that, as the service rates, and hence the switch state, is governed by the descriptions of the channel vectors that were fed back. One may first minimally decompose a “global” switch state that contains every rate allocation for the feedback and multiplexing scheme. This decomposition yielded a collection of generalized switches that may be used to find the subset of users that maximize the scheduling utility function at each scheduling interval. That is, for every channel realization the user assignment process randomly chooses the switches that must be considered to make the optimal rate allocation. Thus, the channel-aware scheduling problem in a multi-user MIMO system with finite rate feedback can be considered as follows:

1. A minimal cover of the global switch state is computed off line
2. Each time the channel changes state users are randomly connected to input ports in the minimal cover which determine the possible rate allocations for every subset of users
3. For each switch the maximal (weighted) clique is determined
4. The maximal-maximal (weighted) configuration is chosen from amongst the switches which identifies the active set of users

More precisely, from (4.13) one has

$$S^* \in \underset{\substack{\mathcal{G} \in \mathcal{C}(K_{\text{global}}) \\ S \text{ clique in } \mathcal{G}}}{\arg \max}} \sum w_i. \quad (4.18)$$

Thus, in order to efficiently search for the optimal subset of users and more generally to understand how the variations in the channel effects the distribution of the rate of the active set of users one must understand how our model for the input occupancy distribution is effected by the structure of the generalized switch.

To begin, recall that in Section 4.2.3, we defined a generalized switch to be a graph with a vertex set that represent the codewords of the quantization scheme and let edges in the graph represent possible pairings for the scheduling decision. Any clique in the graph represents a given processing mode and hence a possible rate allocation of the system. Thus, the processing modes of the systems may be identified with *subsets* of the codewords of the quantization scheme and the BRS model provides a map from these subsets to cliques in one of the graphs used in the cover of the global set of processing modes. This is an important formality as it is useful in understanding the relationship between the structure of the generalized switch is effected by variations in the input occupancy distribution. More

precisely, each level of a generalized switch has an important influence on the stability of the scheduled rate. In particular, for a given input occupancy distribution (the first level) the generalized switch first “disperses” the input occupancy distribution by forming a distribution that describes the occupancy of $\binom{2^r}{m}$ subsets of inputs (the second level). Then, the map from these subsets to cliques defined by the BRS model “collects” the associated probabilities in to a third occupancy model by taking the union of each subsets of inputs described by the switch. Thus, if the map from the BRS model takes many disjoint subsets to a given switch then it is likely that variations in the input occupancy distribution will have little effect on the output occupancy distribution for this switch. Conversely, if the map from the BRS model takes few disjoint subsets to a given switch it is likely that variations in the input occupancy distribution will have a dramatic effect on the output occupancy distribution for this switch. To see this more concretely we provided the following probabilistic model for the BRS model.

In order to understand how the user assignment distribution effects the output distribution of each generalized switch in the BRS model we must understand how the user assignment distribution effects the occupancy distribution of the subsets of codewords and in turn how this occupancy distribution effects the output of each generalized switch via the structure of the switch. We illustrate this general relationship in Figure 4-9 as a three level urn model. In this model the first level contains urns representing the 2^r codewords of the quantization codebook, the second level contains $\binom{2^r}{m}$ contiguous urns which represent the possible subsets of the codewords, each with m distinguished cells labeled by the codewords of the quantization codebook. The final level contains urns representing the generalized switches in the BRS model. Thus, using this model one may view the channel aware scheduling problem as:

1. At each scheduling interval every user places a ball in the urn in the first level that is label by that users quantized channel vector
2. Then, each occupied urn places additional balls in every cell of every urn on the second level which has it as a label
3. In turn each urn on the second level which has every cell occupied places a ball in the urn of the third level corresponding to the switch which contains it.

Thus, in Figure 4-9 the top set of arrows represent the aforementioned “dispersion” of the input occupancy distribution while the bottom set of arrows represent the aforementioned “collection” of the input occupancy distribution.

It is important to note that if one is only able to observe the occurrence that a urn on the second level is fully occupied does not enable one to infer the statistics of the contributing codewords. More precisely, let $\mathbf{1}_{\{\mathcal{S}\}}$ be the indicator random variable which indicates when the urn corresponding to \mathcal{S} is full, i.e.

$$\mathbf{1}_{\{\mathcal{S}\}} = \prod_{i \in \mathcal{S}} \mathbf{1}_{\{i \in \mathcal{R}_{\sigma, \rho}\}}.$$

Then the observation of the frequency that $\mathbf{1}_{\{\mathcal{S}\}} = 1$ does not enable one to make a reliable inference on any of the individual probabilities $\Pr[\mathbf{1}_{\{i \in \mathcal{R}_{\sigma, \rho}\}} = 1]$ for $i \in \mathcal{R}_{\sigma, \rho}$. In particular, the distribution of $\mathbf{1}_{\{\mathcal{S}\}}$ is a symmetric function of the individual probabilities $\Pr[\mathbf{1}_{\{i \in \mathcal{R}_{\sigma, \rho}\}} = 1]$. Thus, the distribution of the indicator $\mathbf{1}_{\{\mathcal{S}\}}$ and hence the marginal distribution of any

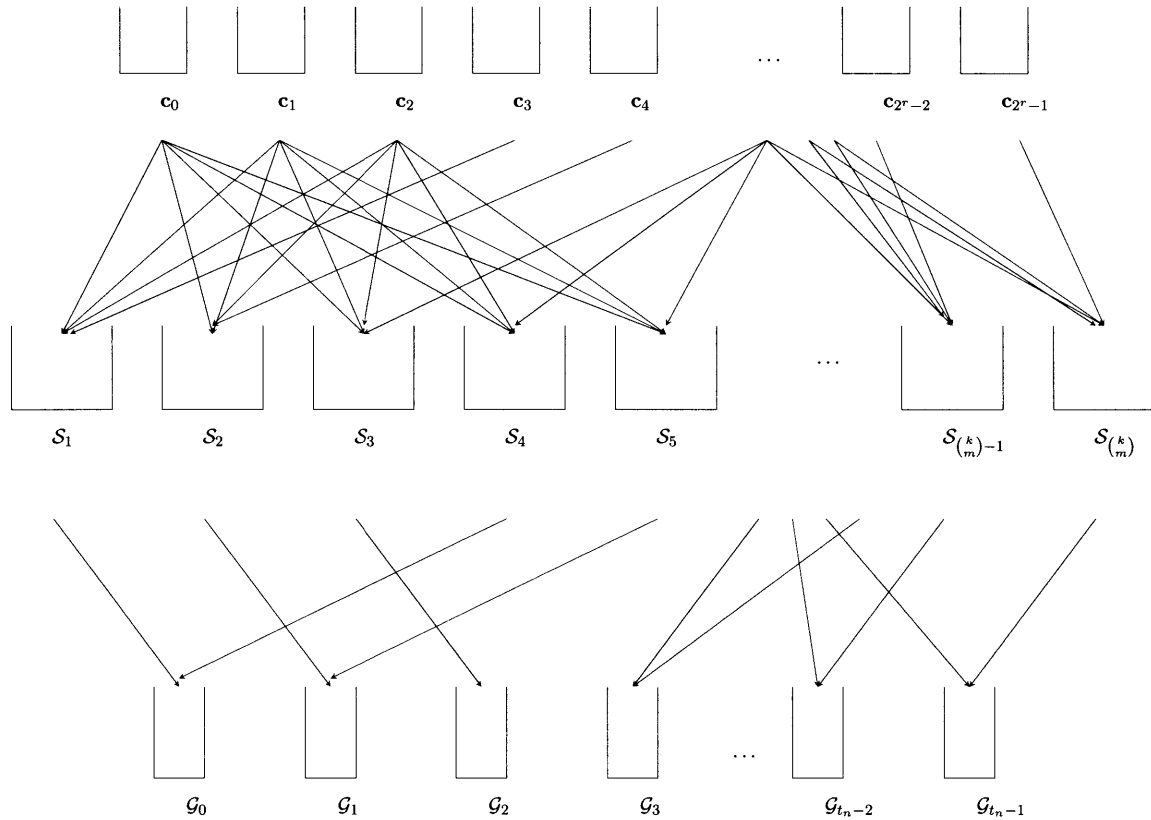


Figure 4-9. A view of the statistical dependencies of switch outputs in the BRS model as a three level urn process. At each scheduling interval every user places a ball in the urn in the first level that is label by that users quantized channel vector. Each occupied urn in turn places additional balls in every cell of every urn on the second level which has it as a label. Then, each urn on the second level, which has every cell occupied, places a ball in the urn of the third level corresponding to the switch which contains it.

subset of codewords is minimally invariant to permutations in the individual probabilities for that subset. For example, suppose that $|\mathcal{S}| = k$ and consider a k users system. Then,

$$p_{\mathcal{S}} \triangleq \Pr[\mathbf{1}_{\mathcal{S}} = 1 \mid |\mathcal{R}_{\sigma,\rho}| = k] = k! \prod_{i \in \mathcal{S}} p_i \quad (4.19)$$

and one may only infer the product of cell probabilities through a history of observations of $\mathbf{1}_{\{\mathcal{S}\}}$. This is an important observation as we are interested in the distribution of the output occupancy of generalized switch. Thus, if one examines any one clique in a switch the frequency it is occupied is invariant to permutations of the marginal probabilities of the events determining it. Thus, one may expect that further combining multiple such sets together to form a generalized switch is likely to further make the output distribution invariant to a large number of input occupancy distributions.

If a switch in the BRS model contains a large number cliques then there are a multitude of ways that one may arrive at a lower bound on the probability that any one of these cliques are occupied. In the sequel, we examine the effects that the input occupancy distribution has on the output distribution of a switch. Since, the input occupancy distribution is a function of the user assignment distribution as well as the number of users that feedback in the sequel we develop all our bounds conditional on the cardinality $|\mathcal{R}_{\sigma,\rho}|$. This allows us to analyze the system performance, by the total law of probability, as well as yields a distribution that is useful to the scheduler which has knowledge of the realization of $|\mathcal{R}_{\sigma,\rho}|$. In this direction, it is useful to know the probability that any set of users of size m channel vectors will yield a maximally sized clique in a switch of interest. That is, the probability

$$p_{\mathcal{G}} = \Pr[\mathcal{A}^{\circ} \text{ maximal clique in } \mathcal{G} \mid |\mathcal{A}| = m] \Pr[|\mathcal{A}^{\circ}| = m]$$

is of interest, where we let \mathcal{A}° denote the unique quantization indices feedback by the set \mathcal{A} . Assuming that there is a large number of cliques in \mathcal{G} , there are a multitude of ways that one may arrive at a lower bound on the probability that a set of users of size m feedback yields a clique. In particular, one may trivially lower bound this probability by examining the probability that the most probable clique is occupied using (4.19) as

$$p_{\mathcal{G}} \geq \max_{\mathcal{S} \text{ maximal clique in } \mathcal{G}} k! \prod_{i \in \mathcal{S}} p_i.$$

Slightly more generally, one may consider forming a lower bound by considering the probability that a disjoint set of cliques are occupied via

$$p_{\mathcal{G}} \geq \Pr \left[\prod_{i=0}^c \mathcal{S} \mid \mathcal{S} \text{ maximal clique in } \mathcal{G} \right] = \sum_{i=0}^c \prod_{i \in \mathcal{S}} p_i.$$

In the most general setting one may arrive at a lower bound by considering the probability that an arbitrary union of cliques are occupied by a subset of users as

$$p_{\mathcal{G}} \geq \Pr \left[\bigcup_{i=0}^c \mathcal{S} \mid \mathcal{S} \text{ clique in } \mathcal{G} \right]. \quad (4.20)$$

In order to to compute (4.20) one may use the principle of inclusion and exclusion. Let $\{\mathcal{S}_i\}_{j \in J}$ be collections of cliques of a given graph \mathcal{G} . Then, by the principle of inclusion and

exclusion one has

$$\Pr \left[\bigcup_{j \in J} \{\mathbf{1}_{\{\mathcal{S}_j\}} = 1\} \right] = \sum_{j=1}^{|J|} (-1)^{j+1} S_j(J) \quad (4.21)$$

where

$$S_j(I) = \sum_{\substack{1 \leq k_1 < \dots < k_j \leq |I| \\ \{k_1 \dots k_j\} \subset I}} \Pr \left[\bigcap_{l=1}^j \{\mathbf{1}_{\{\mathcal{S}_l\}} = 1\} \right]$$

is the j th binomial moment of the number of the events occurring in I . However, this sum is quite complex and one often bounds the union by using the Kwerel lower bounds [79] which yields the following proposition.

Proposition 4.4.1. *Let $\{\mathcal{S}_i\}_{i \in J}$ be collections of cliques of a graph and consider the events $\{\mathbf{1}_{\{\mathcal{S}_i\}} = 1\}_{i \in J}$. Then,*

$$\Pr \left[\bigcup_{j \in J} \{\mathbf{1}_{\{\mathcal{S}_j\}} = 1\} \right] \geq b_l(S_{i,1}(J), S_{i,2}(J), S_{i,3}(J), |J|)$$

where

$$b_l(s_1, s_2, s_3, n) = \max \{l_1 s_1 - l_2 s_2 + l_3 s_3, s_1 - s_2\}$$

and in turn where $l_1 = \frac{h_l + 2n - 2}{nh_l}$, $l_2 = \frac{2(2h_l + n - 4)}{h_l(h_l - 1)n}$, $l_3 = \frac{6}{h_l(h_l - 1)n}$ and $h_l = 2 + \left\lfloor \frac{-6s_3 + 2(n-2)s_2}{-2s_2 + (n-1)s_1} \right\rfloor$.

In the case that the probabilities are uniform one need not apply Proposition 4.4.1 as one may arrive at the exact probability of the union. In particular, conditioned on all users having distinct feedback the probability that any set is occupied is uniform. Thus,

$$p_{\mathcal{G} | \text{distinct}} = \frac{cl(\mathcal{G})}{\binom{2^r}{m}}$$

where

$$cl(\mathcal{G}) = |\{\mathcal{S} : \mathcal{S} \text{ clique in } \mathcal{G}\}|$$

Hence,

$$p_{\mathcal{G}} = \frac{cl(\mathcal{G})}{\binom{2^r}{m}} \prod_{i=2}^m \left(1 - \frac{i-1}{2^r}\right). \quad (4.22)$$

We now examine how this analysis effects the output occupancy distribution.

■ 4.4.1 The Order Statistic Gain/Multi-Node Matching Gain Trade-Off

Due to the large amount of mixing one may conjecture that the output occupancy distribution again behaves like a multinomial distribution. One may then attempt to model the output occupancy distribution as such and use observations of the output occupancy distribution to infer the relevant model parameters ⁶. However, for efficient user selection we prefer to consider only the marginal distribution of the output occupancy distribution for each subset of users as it provides a quite useful form to determine switch occupancy. Also, this provide a direct explanation of the relationship between the input and output

⁶This is doe to model the input occupancy distribution in Section 5.1.

occupancy distributions. In this direction, note that $p_{\mathcal{G}}$ is the probability that any subset of *users* who feed back may be scheduled together via the switch. Thus, in order to determine the probability a maximally sized clique exists in a switch one may consider the sum of indicator functions

$$\mathbf{N}_{\mathcal{G}} = \sum_{\mathcal{A} \subset \mathcal{R}_{\sigma, \rho}} \mathbf{1}_{\{\mathcal{A}^{\circ} \text{ maximal clique in } \mathcal{G}\}} \quad (4.23)$$

Then, one may bound the distribution of $\mathbf{N}_{\mathcal{G}}$ as

$$\Pr(\mathbf{N}_{\mathcal{G}} > 0) = \sum_{j=l}^n \Pr[|\mathcal{R}_{\sigma, \rho}| = j] \Pr[\mathbf{N}_{\mathcal{G}} > 0 \mid |\mathcal{R}_{\sigma, \rho}| = j] \quad (4.24)$$

By ignoring the obvious plurality of subsets of (4.23) one may arrive at a simple lower bound on the probability that $\mathbf{N}_{\mathcal{G}}$ is non-zero by considering an arbitrary partition of $\mathcal{R}_{\sigma, \rho}$ into a disjoint union yielding

$$\Pr[\mathbf{N}_{\mathcal{G}} > 0 \mid |\mathcal{R}_{\sigma, \rho}|] > \Pr[\tilde{\mathbf{N}}_{\mathcal{G}} > 0] \quad (4.25)$$

where $\tilde{\mathbf{N}}_{\mathcal{G}} \sim \text{Binomial}(\lfloor |\mathcal{R}_{\sigma, \rho}|/m \rfloor, p_{\mathcal{G}})$. Using (4.25) one may arrive at a simple lower bound by using known methods to compute the cumulative distribution function of a binomial random variable. However, (4.23) contains exponentially more summands than (4.25) and hence in certain cases one would expect (4.25) to be quite a poor estimate. To remedy this we have the following proposition.

Proposition 4.4.2. [64, Thrm. 2.1] *Let $\mathcal{P}_l(\mathcal{U})$ be the collection of all unordered sets of size l on n items and let*

$$X = \sum_{\mathcal{A} \in \mathcal{P}_l(\mathcal{U})} 1_{\mathcal{A}} \quad (4.26)$$

where $\{1_{\mathcal{A}}\}$ is a family of Bernoulli random variables with $\Pr[1_{\mathcal{A}} = 1] = p$, which are independent if $\mathcal{A} \cap \mathcal{B} = \emptyset$. Then,

$$\Pr[X = 0] \leq \exp\left(-\max\left\{2p^2 \lfloor \frac{n}{l} \rfloor, \frac{8p}{25} \frac{\binom{n}{l}}{\binom{n}{l-1}}\right\}\right) \quad (4.27)$$

Using Proposition 4.4.2 one can easily bound the output occupancy distribution. Indeed, if we are given that $|\mathcal{R}_{\sigma, \rho}| = k$ one may use Proposition 4.4.2 to bound the conditional probability $\Pr[\mathbf{N}_{\mathcal{G}} > 0 \mid |\mathcal{R}_{\sigma, \rho}|]$ and then use the total law of probability to bound the unconditional distribution. However, we first take a slightly different exponent than the one in Theorem 4.4.2 to simplify the resulting expressions. Let,

$$E(p, l) \triangleq \max\left\{\frac{2p^2}{l}, \frac{8p}{25l}\right\}. \quad (4.28)$$

Then as a simple consequence of (4.24) we have the following theorem.

Theorem 4.4.3. *Let \mathcal{G} be a given graph with maximal cliques of size m and let $p_{\mathcal{G}}$ be the probability that an independent and identically distributed selection of m vertices in \mathcal{G} yields*

a clique. Then, if n vertices in \mathcal{G} are selected independently and identically distributed,

$$\Pr[\mathbf{N}_{\mathcal{G}} > 0] \geq \Pr[|\mathcal{R}_{\sigma,\rho}| \geq m] - \min\{c_1 \left(1 - p_{\sigma,\rho} \left(1 - e^{-E(p_{\mathcal{G}},m)}\right)\right)^n, 1\} \quad (4.29)$$

where $c_1 = c_1(p_{\mathcal{G}}, m) = \exp(2p_{\mathcal{G}}^2)$ if $E(p_{\mathcal{G}}, l) = \frac{2p_{\mathcal{G}}^2}{m}$ and $c_1 = c_1(p_{\mathcal{G}}, m) = \exp\frac{8p_{\mathcal{G}}(m-1)}{25m}$ otherwise.

Proof. See Appendix C.3.1 ■

We note that Theorem 4.4.3 exactly characterizes the order statistic gain and multi-node matching gain trade-off for sufficiently large n . Indeed, examining the right hand side of (4.29) one may see that the probability that a switch has m outputs occupied is simply the difference between the cumulative distribution function of a binomial random variable and a function that tends to 0 as n increases. Thus, as the size of the user pool increases the function

$$\min\{c_1 \left(1 - p_{\sigma,\rho} \left(1 - e^{-E(p_{\mathcal{G}},m)}\right)\right)^n, 1\} \rightarrow 0$$

tends to zero so long as

$$n \cdot p_{\sigma,\rho} \left(1 - e^{-E(p_{\mathcal{G}},m)}\right) > 0. \quad (4.30)$$

We note that (4.30) is implicitly a function of the quantizer rate as well as the channel fading distribution. That is, the dependence of the channel statistics and quantizer have been completely characterized through the parameter $p_{\mathcal{G}}$. Thus, particular choices for the quantizer rate and quantizer structure as well as the fading statistics will lead to different tradeoffs between the convergence rate and feedback requirements. However, these parameters do not fundamentally limit the system in terms of the achievable rate asymptotically.

We present a depiction of this behavior in Figures 4-10 – 4-11. In Figure 4-10 we plot the trade-off between $p_{\sigma,\rho}$ and $p_{\mathcal{G}}$ as predicted by the lower bound of Theorem 4.4.3 for a $n = 8, 12, 16, 24$ user system. Note that corresponding bound on the probability of pre-selection success jumps up quite quickly for $p_{\mathcal{G}} > 0.4$ and $p_{\sigma,\rho} > 0.6$ in a 8 user system while in a 32 user system the jump occurs for $p_{\mathcal{G}} > 0.2$ and $p_{\sigma,\rho} > 0.4$. Thus, as the number of users in the system grow there is a smaller requirement that the system contains a large number of orthogonal bases. This may be seen similarly in a 8 transmit antenna system. In Figure 4-11 we plot the trade-off between $p_{\sigma,\rho}$ and $p_{\mathcal{G}}$ as predicted by the lower bound of Theorem 4.4.3 for a $n = 16, 24, 32, 48$ user system. The behavior there is similar, however, relative to the size of the transmit array, the transition from 0 to 1 happens more quickly in a 8 transmit antenna system.

In the following we show the order statistic gain decouples from the multi-node matching gain asymptotically in the case of the Rayleigh model and an almost arbitrarily chosen channel quantization scheme. Thus, as the number of users in a system grows the system designer has a great degree of freedom in the feedback design. However, this question is more subtle for small to moderately sized user pools. As we have seen in Section 3.2 codes which contain many orthogonal bases, in general, have a larger mean squared quantization error. Hence, by choosing a channel quantizer for which $p_{\mathcal{G}}$ is large, and hence contains many orthogonal bases, to ensure successful pre-selection one may increase the mean squared quantization error to an intolerable level. Thus, for practical system design one must balance this trade-off. However, we have yet to thoroughly examine the effects that multi-user diversity has on SINR_{sat} . In particular, one would like to know how well a quantizer which contains many orthogonal bases performs in a system with multi-user diversity. We

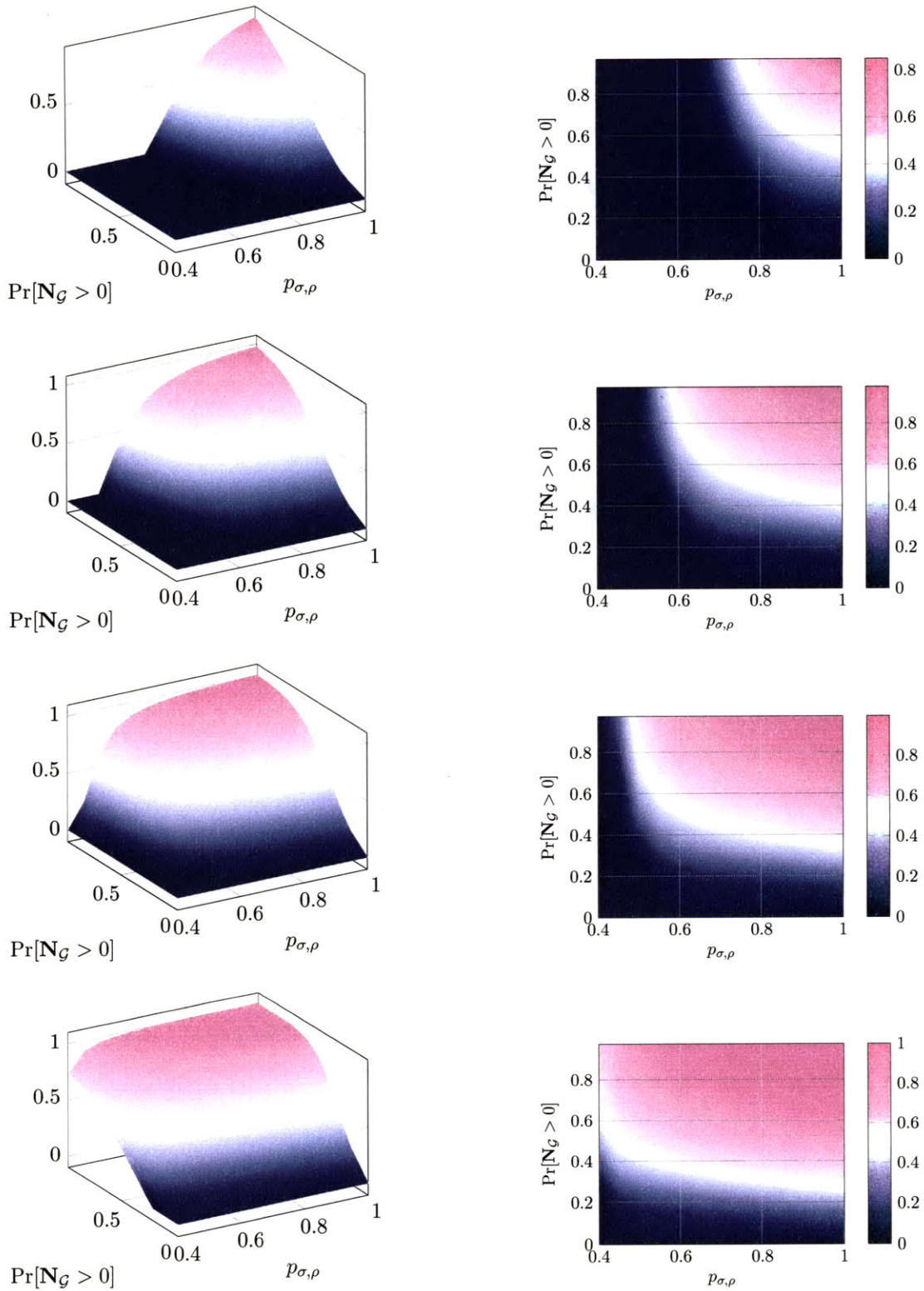


Figure 4-10. The trade-off between $p_{\sigma,\rho}$ and p_G predicted by Theorem 4.4.3 for $n = 8, 12, 16, 24$ with 4 transmit antennas. The smallest number of users is at top and the largest at bottom. Note, even when using the large deviation bound of Theorem 4.4.2 the plots show a rapid transition from 0 to 1 so long as $p_{\sigma,\rho} > 0.4$.

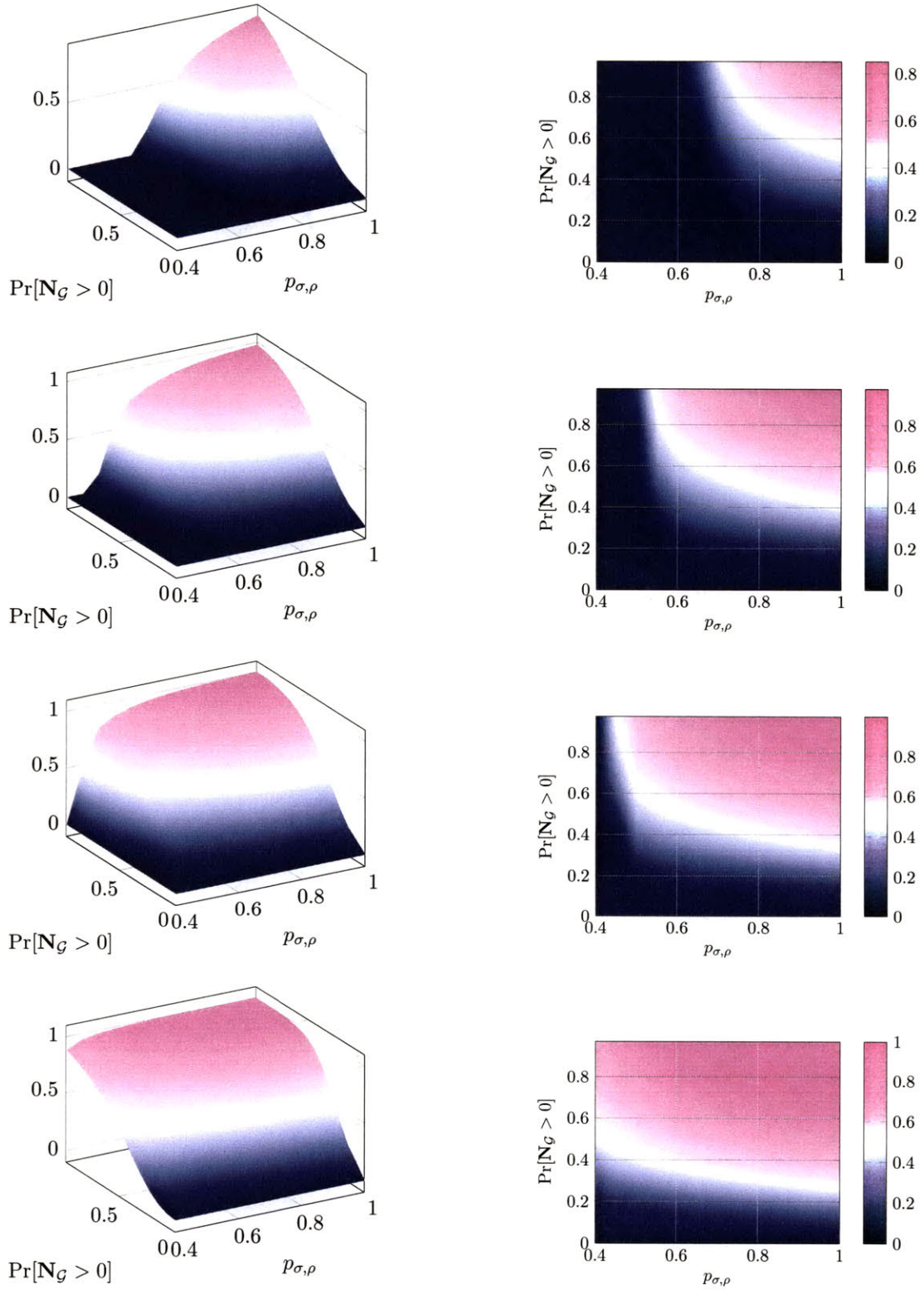


Figure 4-11. The trade-off between $p_{\sigma,\rho}$ and p_G predicted by Theorem 4.4.3 for $n = 16, 24, 32, 48$ with 8 transmit antennas. The smallest number of users is at top and the largest at bottom. Note, even when using the large deviation bound of Theorem 4.4.2 the plots show a rapid transition from 0 to 1 so long as $p_{\sigma,\rho} > 0.4$.

consider this question in Section 4.6 and then in Section 4.7 examine how the quantizers we developed in Chapter 3 perform relative to a derived upper bound. We then, through an explicit example, illustrate how one may use these results in conjunction with Theorem 4.4.3 to practically design a 4 transmit antenna system. We now address the asymptotic decoupling of the order statistic gain and multi-node matching gain for the Rayleigh model.

■ 4.5 Asymptotic Decoupling with the Rayleigh Assumption

In this section we show that the order statistic gain decouples from the multi-node matching gain asymptotically in the case of the Rayleigh model and an almost arbitrarily chosen channel quantization scheme while simultaneously obtaining the maximal achievable throughput. In this direction we let $R^*(n)$ be the maximum rate achieved by any protocol for the Rayleigh model and system model of interest with no constraint on complexity and processing capabilities. In order to examine when order statistic gain decouples from the multi-node matching gain we begin by considering a fairly strong notion of optimality of an architecture.

Definition 1. *An architecture $\mathcal{S}(P, m)$ is said to be strongly asymptotically optimal (with respect to average throughput) if there exists a sequence of protocols*

$$\mathcal{P}(1), \mathcal{P}(2), \dots \in \mathcal{S}(P, m)$$

such that the corresponding average throughputs $R(1), R(2), \dots$ of these protocols satisfies

$$\lim_{n \rightarrow \infty} [R^*(n) - R(n)] = 0, \quad (4.31)$$

Note that replacing (4.31) with the condition

$$\lim_{n \rightarrow \infty} [\log R^*(n) - \log R(n)] = 0 \quad (4.32)$$

corresponds to a much weaker notion of optimality. Preliminary work on asymptotic optimality has focused on this weaker rate-ratio convergence⁷, limiting the practical value of the associated results. To see this weakness, let us define the signal-to-interference plus noise ratio $\text{SINR}(n)$ of the protocol via

$$\text{SINR}(n) \triangleq 2^{R(n)/m} - 1. \quad (4.33)$$

Then weak convergence of rates in the sense of (4.32) can be obtained even when the SINR gap in dB is asymptotically infinite, i.e.,

$$\text{SINR}^*(n)/\text{SINR}(n) \rightarrow \infty.$$

By contrast, strong convergence of rates in the sense of (4.31) ensures that the SINR gap in dB is asymptotically zero. In order to be precise, we begin by defining the asymptotic notation we use in the sequel.

⁷We note that strong convergence of random beamforming has recently been shown in [129].

■ 4.5.1 Asymptotic Notation

We will use the standard asymptotic notation which may be found in [65, 77]. We will say, for two sequences of real numbers $f(n)$ and $g(n)$, that

$$f(n) = O(g(n))$$

if there exists some positive constant C and positive integer n_0 such that for all $n \geq n_0$ $|f(n)| \leq Cg(n)$. Similarly we will say that

$$f(n) = \Omega(g(n))$$

if there exists some positive constant C and positive integer n_0 such that for all $n \geq n_0$ $|f(n)| \geq Cg(n)$. We will say that

$$f(n) = \Theta(g(n))$$

if $f(n) = O(g(n))$ and $f(n) = \Omega(g(n))$. Also, we will say that

$$f(n) = o(g(n))$$

if for any $\epsilon > 0$ there exists some positive integer $n(\epsilon)$ such that for all $n \geq n(\epsilon)$ $|f(n)| \leq \epsilon g(n)$. That is, $\lim_{n \rightarrow \infty} f(n)/g(n) = 0$. We now define the metric we will use to examine how various multiplexing and scheduling algorithms perform in the MIMO downlink when many users are present.

■ 4.5.2 Preliminaries

In the sequel, we show that the simple, low complexity, decentralized protocol architecture of Section 4.1 is strongly asymptotically optimal in the sense of Definition 1 for the Rayleigh model. More specifically, we show that the average throughput achievable by this architecture converges in the sense of (4.31) to

$$R_+^*(n) = m \log(1 + \text{SINR}^*(n)) + o(1) \quad (4.34)$$

with

$$\text{SINR}^*(n) = \frac{P \log n}{m^2}, \quad (4.35)$$

which, as shown in [111], is an asymptotic upper bound on $R^*(n)$, i.e.,

$$\lim_{n \rightarrow \infty} [R_+^*(n) - R^*(n)] \geq 0.$$

The average throughput achievable for a given sequence of protocols in our architecture can be expressed in the form

$$R(n) = \mathbb{E} [R_{\mathbf{H}_{\mathcal{A}}}], \quad (4.36)$$

where the expectation is taken over both the channel realizations and the randomization in the selection of the set $\mathcal{A} \in \mathcal{T}$, and where $R_{\mathbf{H}_{\mathcal{A}}}$ denotes the rate achieved for a particular active set \mathcal{A} .

A bound on the rate gap associated with (4.36) can be readily obtained when there exists, as will be the case in our development, a rate bound $R_-(n)$ such that $R_{\mathbf{H}_{\mathcal{A}}}(n) \geq R_-(n)$ for

all $\mathcal{A} \in \mathcal{T}$. In particular, in this case, we may write

$$R(n) \geq (1 - p_\emptyset(n))R_-(n), \text{ with } p_\emptyset(n) \triangleq \Pr_{\text{fail}}^{(M)} \left(2^{\frac{R_-(n)}{m}} - 1 \right)$$

whence

$$R^*(n) - R(n) \leq [R^*(n) - R_-(n)] + [p_\emptyset(n)R_-(n)]. \quad (4.37)$$

Thus to show strong asymptotic optimality, it suffices to show that each of the two terms in brackets in (4.37) approach zero as $n \rightarrow \infty$. We note that proving the asymptotic optimality has a quite high impact on the broader system design. That is, to show that each of the two terms in brackets in (4.37) approach zero we must show that $p_\emptyset(n) \rightarrow 0$. Thus, proving strong asymptotic optimality by this method implies that the throughput lower bound $R_-(n)$ may be met with probability one. As previously mentioned this is of interest for the broader network design as this implies that as the user population grows a small *subset* of switches in the BRS process model may be considered in order to arrive at an optimal scheduling decision and the overall system behaves as a conventional (wireline) switch.

We now describe suitable choices for $R_-(n)$ for the particular multiplexers of interest. In the sequel, when there is risk of confusion, we use superscripts ^{II} and ^{IC} to distinguish $R(n)$, $R_-(n)$, $\text{SINR}(n)$, and other quantities for the interference ignoring and cancelling

Consider first the case of interference-ignoring multiplexers. In this case, for a given active set \mathcal{A} and channel realization $\mathbf{H}_{\mathcal{A}}$, it is straightforward to verify that the achievable sum rate satisfies

$$R_{\mathbf{H}_{\mathcal{A}}}^{\text{II}}(n) = \sum_{j \in \mathcal{A}} \log(1 + \text{SINR}_j^{\text{II}}) \quad (4.38)$$

where

$$\text{SINR}_j^{\text{II}} = \frac{P \|\mathbf{h}_j\|^2 \sigma_j^2}{m + P \|\mathbf{h}_j\|^2 \|\boldsymbol{\sigma}_j^c\|^2} \quad (4.39)$$

with

$$\sigma_j = \hat{\mathbf{h}}_j^\dagger \tilde{\mathbf{h}}_j, \quad \text{and} \quad \boldsymbol{\sigma}_j^c = \hat{\mathbf{H}}_{\mathcal{A} \setminus j}^\dagger \tilde{\mathbf{h}}_j. \quad (4.40)$$

The case for which there is no quantization corresponds to setting $\hat{\mathbf{h}}_j = \tilde{\mathbf{h}}_j$ in (4.39) and (4.40), so that $\sigma_j = 1$ and $\boldsymbol{\sigma}_j^c = \tilde{\mathbf{H}}_{\mathcal{A} \setminus j}^\dagger \tilde{\mathbf{h}}_j$.

To obtain a lower bound on $R^{\text{II}}(n)$, we define the following (deterministic) lower bound on $\text{SINR}_j^{\text{II}}$:

$$\text{SINR}_-^{\text{II}}(n) \triangleq \min_{\mathcal{A}, j, \mathbf{H} : |\mathcal{T}| \neq 0, \mathcal{A} \in \mathcal{T}, j \in \mathcal{A}} \text{SINR}_j^{\text{II}}, \quad (4.41)$$

from which we obtain, via (4.38) and (4.36),

$$\frac{R^{\text{II}}(n)}{1 - p_\emptyset(n)} \geq \frac{\mathbb{E} \left[R_{\mathbf{H}_{\mathcal{A}}}^{\text{II}} \right]}{1 - p_\emptyset(n)} \geq m \log(1 + \text{SINR}_-^{\text{II}}(n)) \quad (4.42)$$

for any $\mathcal{A} \in \mathcal{T}$. In turn, via (4.33), we obtain

$$\text{SINR}^{\text{II}}(n) \geq (1 + \text{SINR}_-^{\text{II}}(n))^{1 - p_\emptyset(n)} - 1. \quad (4.43)$$

In the absence of quantization there is a corresponding specialization of $\text{SINR}_-^{\text{II}}(n)$. While

for the case without quantization a natural bound analogous to (4.41) is immediate, for the case with quantization it is more convenient to develop an alternative. To this end, we obtain⁸ in Appendix C.1.2

$$\text{SINR}_j^{\text{IC}} \geq \gamma_j \triangleq \frac{P\|\mathbf{h}_i\|^2 \left[|\sigma_j|\tau_j - \sqrt{1 - |\sigma_j|^2} \lambda_{\min} \right]_+^2}{\text{Tr}(\hat{\Phi}_{\mathcal{A}}^{-1})\tau_j^2 + P\|\mathbf{h}_j\|^2(1 - |\sigma_j|^2)\lambda_{\max}} \quad (4.44)$$

where $[x]_+ = \max\{0, x\}$ and where λ_{\min} and λ_{\max} are, respectively, the minimum and maximum eigenvalues of $\hat{\Phi}_{\mathcal{A} \setminus j}$, and where

$$\tau_j = \lambda_{\min} - \|\hat{\sigma}_j^c\|^2. \quad (4.45)$$

Hence, defining

$$\text{SINR}_-^{\text{IC}}(n) \triangleq \min_{\mathcal{A}, j, \mathbf{H} : |\mathcal{T}| \neq 0, \mathcal{A} \in \mathcal{T}, j \in \mathcal{A}} \gamma_j, \quad (4.46)$$

which is deterministic, we obtain

$$R^{\text{IC}}(n) \geq (1 - p_\emptyset(n)) m \log(1 + \text{SINR}_-^{\text{IC}}(n)) \quad (4.47)$$

whence, via (4.33),

$$\text{SINR}^{\text{IC}}(n) \geq (1 + \text{SINR}_-^{\text{IC}}(n))^{1 - p_\emptyset(n)} - 1. \quad (4.48)$$

■ 4.5.3 Asymptotic Optimality with Perfect Feedback

We now develop the key characteristics of our architecture in the absence of quantization effects. We first characterize the amount of feedback required by the protocol as a function of the parameter settings. For this case, we view $N_\rho = |\mathcal{R}_\rho|$ as a measure of the feedback link capacity requirement. Observe that N_ρ is a binomial random variable with mean $\mathbb{E}[N_\rho] = np_\rho$. Since p_ρ is the probability that a user feeds back its channel gain vector, we have, from (4.1), that

$$p_\rho = \Gamma(2m, m\rho_-) - \Gamma(2m, m\rho_+) \quad (4.49)$$

with $\Gamma(\cdot, \cdot)$ denoting the incomplete gamma function. We have the following theorem.

Theorem 4.5.1. *Let, for any $\delta > 0$, $\rho_+(n) = (1 + \delta)(\log n)/m$ and $\rho_-(n) = (\log n)/m - (\log \alpha(n))/m$ where*

$$m \log \log n \leq \log \alpha(n) = o(\log n).$$

Then

$$\mathbb{E}[N_\rho] = 2m\alpha(n)(1 - o(1)) + \mathcal{O}(1/n) \quad (4.50)$$

Proof. See Appendix C.3.2. ■

From this theorem we see that the choice of $\alpha(n) = e^{m(\rho_+(n) - \rho_-(n))}$ effectively controls the amount of feedback required by the system. We next characterize the probability p_\emptyset that the pre-selection phase of the protocol yields no candidate sets.

⁸As will become apparent, the appeal of γ_j as a bound is its simple form as $\sigma_j \rightarrow 1$.

Theorem 4.5.2. *Let $\rho_+(n)$ and $\rho_-(n)$ be as in Theorem 4.5.1. Then provided $0 \leq \varepsilon(n) \leq 1$ we have*

$$p_{\emptyset}(n) \leq e^{-\mathbb{E}[N_{\rho}]\beta(n)/m}, \quad (4.51)$$

where

$$\log \beta(n) = 2(m-1)^2 \log(\varepsilon(n)) \quad (4.52)$$

Proof. See Appendix C.3.3. ■

This theorem characterizes the manner in which successful pre-selection depends on the interference control parameter ε and the feedback parameter ρ . Finally, we establish that our architecture is strongly asymptotically throughput optimal.

Theorem 4.5.3. *Let $\rho_+(n) = (\log n)/m$. For both the interference-ignoring multiplexer and interference-cancelling multiplexer, let*

$$\rho_-(n) = (\log n)/m - \log \log n \text{ and } \varepsilon(n) = 1/(\log n)^{1/(2(m-1))}.$$

Then in both cases the protocol sequence $\mathcal{P}_{\varepsilon,\rho}(n)$ with average throughputs $R_{\varepsilon,\rho}(n)$ and $\text{SINR}_{\varepsilon,\rho}(n)$ satisfies

$$R^*(n) - R_{\varepsilon,\rho}(n) = \mathcal{O}\left(\frac{1}{\log n}\right), \quad (4.53)$$

$$\frac{\text{SINR}^*(n)}{\text{SINR}_{\varepsilon,\rho}(n)} - 1 = o(1). \quad (4.54)$$

Moreover, with this protocol sequence, the feedback link must support, on average,

$$\mathbb{E}[N_{\rho}] = 2m(\log n)^m(1 + o(1)) + \mathcal{O}(1/n) \quad (4.55)$$

users.

Proof. See Appendix C.3.4. ■

We note that this theorem is quite illuminating in terms of the decoupling of the order statistic gain and the multi-node matching gain. In particular, Theorem 4.5.1, in the presence of perfect feedback at the transmitter, shows the number of users which feedback (the order statistic gain) has an exponential effect on the probability that the rate target can be met while the requirement for inner products between channel vectors has little effect on this decay so long as it is bounded away from zero. That is, the multi-node matching gain target for the scheduled rate simply interpolates $\beta(n)$ between 0 and $\exp(2(m-1)^2)$. Moreover, this result clearly implies that the order statistic gain decouples from the multi-node matching gain for large n .

Corollary 4.5.4. *Assuming the Rayleigh model from user user's channel fading the order statistic gain decouples from the multi-node matching gain as $n \rightarrow \infty$ when the transmitter has perfect knowledge of each user's channel realization.*

From Theorem 4.5.3 we see that the use of a cruder multiplexer does not incur a penalty in strong throughput optimality. In particular, in both cases the number of users who must report their channel gains in any scheduling interval is the same sub-linear function. Thus, one may enforce a strong SNR target in a system that employs an interference ignoring

multiplexer. We note that this subtlety is further illuminated in the proof in Appendix C.3.4. There we show the one may lower bound the probability that a set of users meets a prescribed interference control parameter, $\epsilon(n)$, by fixing an arbitrary basis at the transmitter which is unknown to the receivers. Then, the transmitter may select a set of users which meet the constraint $\epsilon(n)$ by ensuring each user is highly correlated with the chosen basis. We note that this is a *centralized* approach to user selection. In the following section we show that a *distributed* approach to this problem is equivalent to the problem of user selection with finite rate feedback. That is, one may rather distribute the basis chosen for selection at the transmitter to the users. Then, one may have only the users that meet the correlation constraint the chosen basis used for selection feedback. Thus, with this approach some of the pre-selection phase is computed in a distributed manner at the user terminals. We now generalize our optimality results to the case in which the feedback is quantized.

■ 4.5.4 Asymptotic Optimality with Finite Rate Feedback

Our previous result for the asymptotic decoupling of the order statistic gain and multi-node matching gain with perfect channel state information generalizes rather naturally, especially in light of Theorem 4.4.3 and (4.22). Since the protocol uses r -bit quantization for each channel gain to be fed back, the total feedback per scheduling interval is $rN_{\rho,\sigma}$ bits, where $N_{\rho,\sigma} = |\mathcal{R}_{\rho,\sigma}|$.

Now $N_{\rho,\sigma}$ is similarly a binomial random variable with mean $\mathbb{E}[N_{\rho,\sigma}] = np_{\rho,\sigma}$. Since $p_{\rho,\sigma}$ is the probability that a user feeds back its channel gain vector, and the channel is assumed to be isotropic, we have that

$$p_{\rho,\sigma} = p_{\rho}p_{\sigma} \quad (4.56)$$

where p_{ρ} is as defined in (4.49) and

$$p_{\sigma} = \Pr\{|\tilde{\mathbf{h}}_j \hat{\mathbf{h}}_j^{\dagger}| \geq \sigma\} = 2^r(1 - \sigma^2)^{m-1}, \quad (4.57)$$

with the right-hand equality following from the protocol constraint that $\sigma \geq \mu_0(\mathcal{C})$, with, as in (2.18), $\mu_0(\mathcal{C})$ denoting the coherence of the code. Hence, (4.56) and (4.57) imply that the expected aggregate feedback per scheduling interval is proportional to

$$\mathbb{E}[N_{\rho,\sigma}] = \mathbb{E}[N_{\rho}] 2^r(1 - \sigma^2)^{m-1}. \quad (4.58)$$

We next characterize the probability p_{\emptyset} that the pre-selection phase of the protocol yields no candidate sets, generalizing our result of Theorem 4.5.3 to the case where there is quantization. In particular, consider a generalized switch in which edges are drawn between any codevectors for which the magnitude of the inner product is less than ϵ . One has, by specializing (4.22),

$$p_{\mathcal{G}} = \frac{k_{\epsilon}(\mathcal{C}_r)}{\binom{2^r}{m}} \prod_{i=2}^m \left(1 - \frac{i-1}{2^r}\right)$$

with $k_{\epsilon}(\mathcal{C}_r)$ denoting the number of codes of size m with coherence at most ϵ that can be constructed from expurgations of \mathcal{C}_r , i.e.,

$$k_{\epsilon}(\mathcal{C}_r) = |\{\mathcal{C}_{\log m} \in \mathcal{C}_r : \mu_0(\mathcal{C}_{\log m}) \leq \epsilon\}|. \quad (4.59)$$

Then we have the following theorem.

Theorem 4.5.5. *Let $\rho_+(n)$ and $\rho_-(n)$ be as in Theorem 4.5.1 and let $0 < \sigma(n) < 1$. Then for any fixed $\varepsilon \geq 0$ we have*

$$p_{\emptyset}(n) \leq e^{-\mathbb{E}[N_{\rho,\sigma}]p_G/m} \quad (4.60)$$

where $\mathbb{E}[N_{\rho,\sigma}] = 2^r \mathbb{E}[N_{\rho}] \cdot (1 - \sigma(n)^2)^{m-1}$ where $\mathbb{E}[N_{\rho}]$ is as in (4.50).

Proof. See Appendix C.3.3. ■

This theorem characterizes the manner in which successful pre-selection depends not only on the feedback parameters (ρ, σ) and the interference control parameter ε , but also on the properties of the quantization codebook \mathcal{C}_r . Finally, we have that our architecture is also strongly asymptotically throughput optimal when the feedback is quantized.

Theorem 4.5.6. *Let $\varepsilon(n) \equiv 0$, let $\rho_+(n) = (\log n)/m$,*

$$\rho_-(n) = (\log n)/m - (2m - 1)/m \cdot \log \log n,$$

and let $\sigma^2(n) = 1 - 1/\log^2 n$. Furthermore, choose a quantization codebook \mathcal{C}_r such that it contains at least one orthonormal basis, i.e., $k_0(\mathcal{C}_r) \geq 1$. Finally, select the interference-cancelling multiplexer. Then the protocol sequence $\mathcal{P}_{\varepsilon,\rho,\sigma}(n)$ with average throughputs $R_{\varepsilon,\rho,\sigma}(n)$ and $\text{SINR}_{\varepsilon,\rho,\sigma}(n)$ satisfies

$$R^*(n) - R_{\varepsilon,\rho,\sigma} = \mathcal{O}\left(\frac{1}{\log n}\right), \quad (4.61)$$

$$\frac{\text{SINR}^*(n)}{\text{SINR}_{\varepsilon,\rho,\sigma}(n)} - 1 = o(1) \quad (4.62)$$

Moreover, with this protocol sequence, the aggregate rate the feedback link must support, on average, is

$$\mathbb{E}[N_{\rho,\sigma}] = 2^{r+1} m \log n (1 + o(1)) + \mathcal{O}(1/n). \quad (4.63)$$

Proof. See Appendix C.3.4. ■

That one can also get such throughput optimality for the case of interference-ignoring multiplexers follows immediately from the fact that when $\varepsilon(n) \equiv 0$ the interference-cancelling and interference-ignoring multiplexers are identical. Additionally, in both cases the result of Theorem 4.63 clearly implies that the order statistic gain decouples from the multi-node matching gain for large n in a system with quantization.

Corollary 4.5.7. *Assuming the Rayleigh model from user user's channel fading the order statistic gain decouples from the multi-node matching gain as $n \rightarrow \infty$ when the transmitter uses a single bases as the quantization codebook.*

For any particular choice of multiplexer, we can also compare the feedback requirement scaling with and without quantization — e.g., (4.55) and (4.63) in the case of an interference-cancelling multiplexer. As this case reveals, and as is true more generally, we see that the number of users reporting back their channel gains scales much more slowly when quantization is used. This is because the common quantization is effectively providing sufficient coordination to enable some pre-selection to happen at the receiver. Hence, with finite rate feedback the multi-node matching gain is enhanced in part by the order statistic gain through this decentralized pre-selection achieved by multiple bases contained in the quantizer.

We also emphasize that the parameter choices in Theorem 4.5.6 (and Theorem 4.5.3 earlier) are sufficient but not necessary for throughput optimality. And in particular different parameter choices will lead to different tradeoffs between the convergence rate and feedback requirement. However, in the case of quantization, it is worth noting that $\varepsilon(n) \rightarrow 0$ is necessary.

Finally, it is also worth remarking that an implication of the theorem is that a large codebook (fine quantization) is not required for strong asymptotic throughput optimality — indeed an orthonormal codebook of size m is sufficient which further implies that as the number of user in the system grows the system designer is afforded extra degrees of freedom in the feedback design. However, these results relied heavily on the isotropic distribution of the users channels. If the users channel correlation a the single basis selected to achieve Theorem 4.5.6 is mismatched to this correlation then a subset of users may have a reduced the ability for users to meet an SNR target and thus slow the rates of convergence. If one has the ability to infer the underlying channel correlation however, this scenario may be easily remedied by adapting the quantization scheme to match a users fading statics. However, adapting each users quantization codebook may yield a complex set of codevectors which make scheduling in the associated generalized switch quite difficult. Hence, in Chapter 5 we present a simple adaptive feedback framework that enables a system to adapt a given quantization scheme while preserving the underlying structure of the base code by using the methods of Section 3.6.

However, the interplay between the achieved mean square error and the pre-selection failure probability is more subtle for small to moderately sized user pools. As we have seen in Section 3.2 codes which contain many orthogonal bases, in general, have a larger mean squared quantization error. Hence, by choosing a channel quantizer for which p_G is large, and hence contains many orthogonal bases, to ensure successful pre-selection one may increase the mean squared quantization error to an intolerable level. Thus, for practical system design one must balance this trade-off. We consider this final question before examining the effects channel correlation has on our system architecture.

■ 4.6 Quantizer Performance with Many Users

In Section 2.2.1 we showed that, under mild constraints, a feedback scheme which better represents orthogonal vectors in general has a larger mean square quantization error. Hence, in a practical system one must balance these two properties in order to meet system design constraints. Thus, it is of interest to know in what regimes it is better to design a feedback scheme which better represents orthogonal vectors or has a lower mean square quantization error. In this direction we note that in a multi-user MIMO system with finite-rate feedback it is reasonable to expect that either:

1. multiple users will be quantized to every cell and the users with the smallest MSE can be selected
2. multiple users will be quantized to distinct quantization indices and the user's whose MSE is below a given threshold may be used

As such, it is reasonable to expect that the transmit base can choose from among the users with the *best* mean squared error. In order to simplify scheduling one can again attempt to choose the users with the best individual SNR, where here the SNR is dependent on the channel fading as well as the quantization error, then attempt to find a set of users

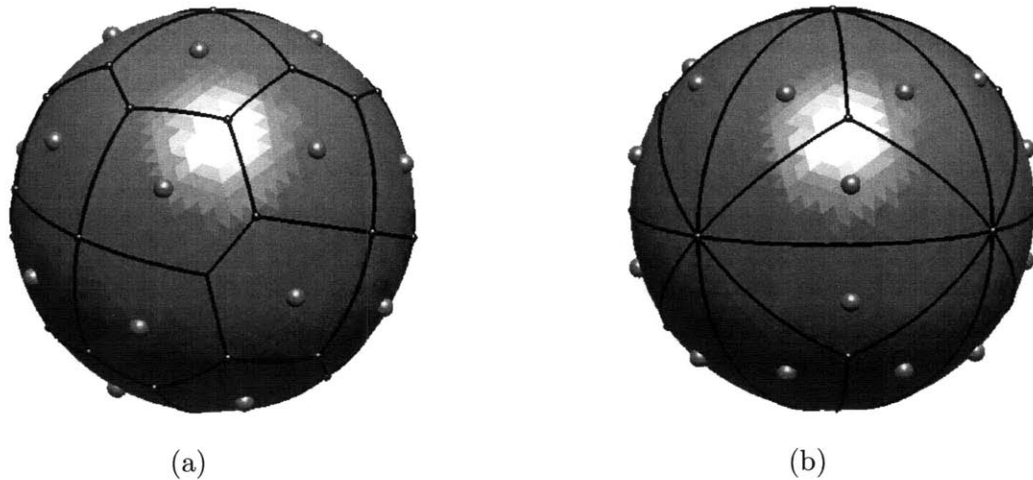


Figure 4-12. Two possible arrangement of 12 lines in \mathbb{R}^3 . (a), a uniform collection of lines that has a low mean square error. (b), a structured collection of 12 lines with more bases. In the absence of order statistics the quantizer in (b) has a higher mean square error.

that negligibly interfere with one another from this reduced set. If such a greedy approach to selecting users is successful with high probability the feedback design problem is less influenced by the mean squared quantization error and the system designer has the freedom to depart from a mean squared error centered feedback design and use this extra degree of freedom to choose a feedback scheme that is that is more convenient for the broader system design. In particular, the system designer may choose to use a channel quantizer from our framework from Section 3.2 to balance the mean squared error of quantizer with the number of orthogonal bases contained in the code.

To begin to develop the relevant insights needed for a system designer to choose the appropriate channel quantizer design in a multi-user MIMO system we begin by considering how the single-user and multi-user quantization problems differ in a time-division system where by one user is selected for transmission in any scheduling interval under the assumption of the Rayleigh model. In rich scattering environments, where the channel gains between each transmit element are modeled as *i.i.d* complex Gaussian, the channel is isotropic so that the direction of any channel vector is uniformly distributed on the complex unit m -sphere. Hence, the codebook design problem may be viewed as a sphere vector quantization problem [137] and one may use a Lloyd like numerical algorithm (see [52] for details) to construct a codebook that minimizes the mean square error by attempting to uniformly space the lines⁹.

In a single-user system the expected MSE error is directly related to the size and shape of the Voronoi cells. In particular, the MSE of any cell is the second moment of the cell. Thus, a Voronoi cell with a smaller second moment has a smaller MSE and hence achieves a higher expected rate. A numerical algorithm that attempts to improve system performance by uniformly spacing lines, as the Lloyd algorithm does, is thus likely to improve system performance in a single-user system. To see how the shape of the Voronoi cell effects the

⁹This algorithm starts by initializing with a random (or deterministic) placement of the codevectors. Then, the Voronoi cell for each codevector is determined creating a partition of Ω_m . Next, a new partition is determined by computing the center of each partition and moving each codevector to the center of its partition. This process is repeated until the process converges to some local optimum.

mean square error consider the two codebooks in \mathbb{R}^3 in Figure 4-12. Note, that the quantizer on the left has a much smaller second moment than the one on the right as the mass of Voronoi cells for the quantizer on the left is more evenly distributed about its center.

In a multi-user system much of the gain in MSE that is achieved by a Lloyd like numerical optimization is achieved by the order statistic gain. That is, when multiple users are quantized to the same cell it is highly likely that one user's channel vector is close to the codeword of the cell (or alternatively far from the boundary of the cell). Indeed, in the multi-user MIMO channel with finite-rate feedback one can select users with the *best* channel correlation in each Voronoi cell. Hence, the MSE performance of a user selected for transmission may be given by an order statistic of the MSE over all users quantized to a cell and not the second moment of that cell. We have shown that in the large user limit the mean squared quantization error tends to 0 with only $\log m$ bits of feedback per user [124]. Thus, in the multi-user scenario the MSE error is less closely tied to the particular size and shape of the Voronoi cells of the quantizer and the overall performance of the system is less closely related to the second moment of the Voronoi cells of the quantizer. This observation is important as when the number of users in the system grows the system designer is afforded an additional degrees of freedom in the feedback design. That is, as the users selected from each cell are likely to have channel vectors that lie in a spherical cap which is strictly contained inside the Voronoi region of each codeword the system designer has the freedom to perturb any arrangement of lines to one that is more convenient for the broader system design. In particular, any irregular quantizer that is designed by a Lloyd like numerical algorithm to optimize the MSE (or the expected rate) of the system can be rearranged to meet broader system design objectives with negligible effect on the MSE (or the expected rate) of the system.

To see that any set of lines designed by a Lloyd like numerical algorithm can be moved to a more regular structure we provide a simple example which is depicted in Figure 4-13. Note that the collection of lines on the left hand side of Figure 4-13 has a smaller second moment relative to that of the right hand side of Figure 4-13 as Voronoi regions that are more symmetric about center have lower second moments. However, examining Figure 4-13 (b) one may see that the quantizer depicted still has a significant mass around the center. In a system with many users it is likely that every user selected for transmission has a channel vector which lies strictly inside a spherical cap contained in the Voronoi region with high probability. Thus, the quantizers in in Figure 4-13 (a) and (b) should achieve approximately the same expected quantization error on average. However, the achieved rates in a multi-user system are not only a function of the quantization error, but also the interference between users. Thus, the rates achieved by the two quantization schemes may continue to differ if one can not ensure that co-channel interference is not approximately equal in the two systems as the structure of the two system of lines may led to a difference in the interference a user sees on average.

In a system with multiple users a natural metric on the performance of a quantizer is the order statistic on SINR_{sat} . More precisely, for an n user system we let

$$\sigma_{(0)} \leq \sigma_{(1)} \leq \cdots \leq \sigma_{(n-1)}$$

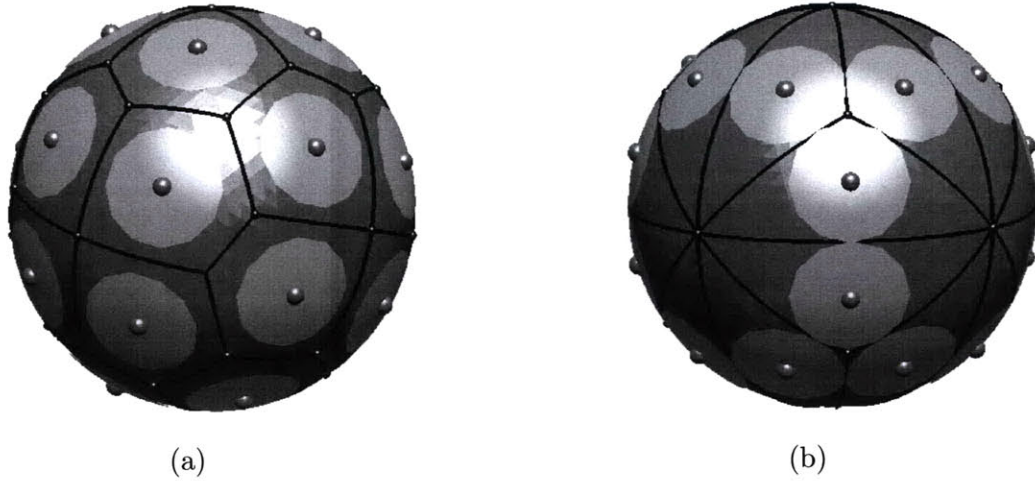


Figure 4-13. The two arrangement of 12 lines in \mathbb{R}^3 from Figure 3-1 where spherical caps of equal half angles are depicted around the codewords. In a system with many users the probability that the quantization error of the user with the *smallest* quantization in each cell falls in the spherical cap is approximately equal.

be the ordered magnitude of the correlation of each user's channel vector with its quantized channel vector. Then, we let for any code \mathcal{C}_r ,

$$\text{SINR}_{\text{sat}}(\mathcal{C}_r; n, \ell) = \mathbb{E}_{\mathbf{H}} \left[\frac{1}{\ell} \sum_{i=n-\ell}^{n-1} \frac{\sigma_{(i)}^2}{1 - \sigma_{(i)}^2} \right]$$

be the expected value of SINR_{sat} for the best ℓ users in a pool of size n .

The expected order statistics of a general distribution has been well studied. In particular, given a sequence of n identically distributed positive random variables (not necessarily independent), X_0, X_1, \dots, X_{n-1} with common mean μ and variance ς one has [23]

$$\mathbb{E} \left[\frac{1}{\ell} \sum_{i=n-\ell}^{n-1} X_{(i)} \right] \leq \mu + \varsigma \sqrt{\frac{n-\ell}{\ell}}. \quad (4.64)$$

One can show that this bound is in fact tight, i.e. there exists a probability distribution for which the inequality in (4.64) may be replaced with equality, and does not vary greatly with the assumption of independence. This bound on the order statistic is quite useful in understanding the behavior one should expect from the order statistic for SINR_{sat} . Examining (4.64) one may see that by only using a small fraction of the user population for the order statistic, i.e. $\ell = m$ where $m \ll n$, $\text{SINR}_{\text{sat}}(\mathcal{C}_r; n, \ell)$ will grow at a rate no greater than \sqrt{n} times the variance. Thus, one may use (4.64) to arrive at an upper bound on the rate of growth of any quantization scheme using Lemma 2.4.2 and an upper bound on the rate of growth of RVQ using Lemma 2.4.1. However, this bound, while yielding the appropriate behavior of the order statistic, is far too optimistic in the exponent of n in the scaling. That is, in the sequel we show that using the upper bound on the quantizer performance one has

$$10 \log_{10} \text{SINR}_{\text{sat}}^{\text{UB}}(n, \ell) \approx \frac{3}{m-1} \cdot (r + \log_2 n) + C(\ell, m)$$

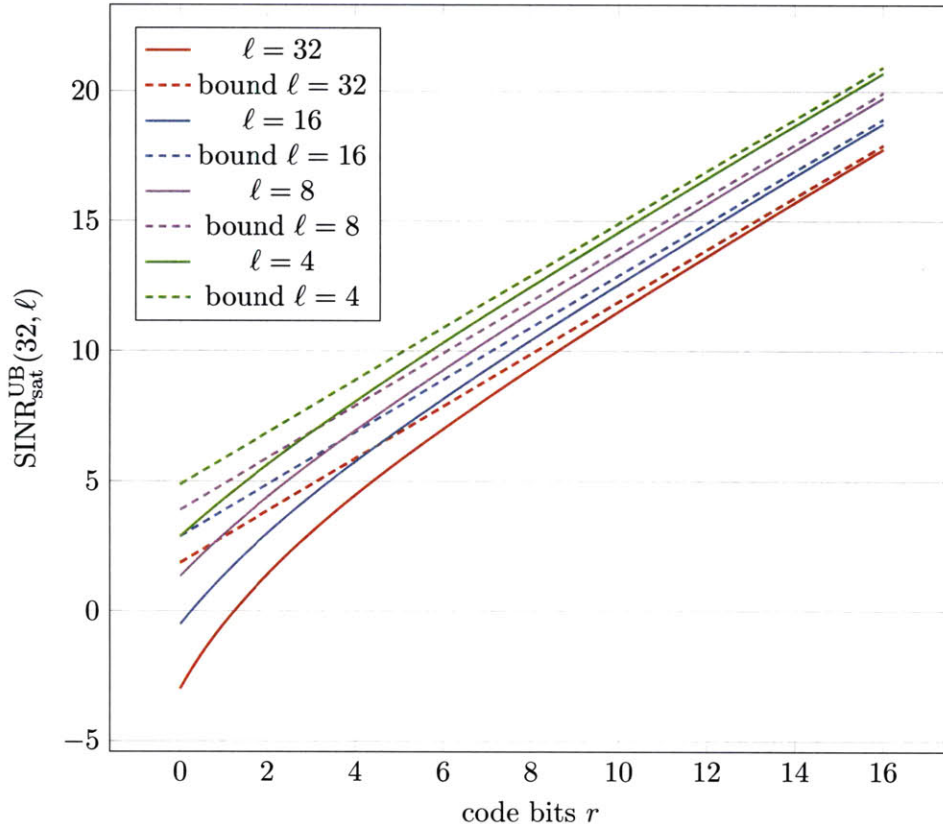


Figure 4-14. The upper bound $\text{SINR}_{\text{sat}}^{\text{UB}}(n, \ell)$ in a 32 user system for various values of ℓ as well as the upper bound on $\text{SINR}_{\text{sat}}^{\text{UB}}(32, \ell)$, (B.6a). Note that the for a large number of bits there is an approximately equal slope for each curve with a fixed offset due to the number of users selected as predicted by (B.6a).

for some constant $C(\ell, m)$ which does not depend on n or r . Thus, in a multi-user system doubling the size of the user pool has roughly the same effect of adding a bit of feedback using the optimal quantization scheme.

We plot the approximation of the upper bound $\text{SINR}_{\text{sat}}^{\text{UB}}(n, \ell)$ in Figures 4-14 and 4-15 along with its exact value. Examining Figures 4-14 and 4-15 one can see that the behavior of the approximation of the upper bound is accurate for high rates. In particular, the curves are approximately linear for rates greater than 10 bits. However, examining Figures 4-14 and 4-15 one can see that the approximation of the upper bound for SINR_{sat} is even more accurate when examining the effects of a growing user population. That is, one may see in Figure 4-14 that while the curves themselves are not linear the gaps between the curves are. This is depicted in Figure 4-15. Note that the approximation parallels the upper bound and there is an approximately constant gap between the groups of curves. This is an important observation for any multi-user MIMO system which aims to operate at or above a fixed SINR in the high SNR regime. In particular, each time the number of users in the system doubles the system designer may decrease the feedback rate per user by a bit and expect to achieved the fixed SINR target. We develop our upper bound on SINR_{sat} with order statistics in Appendix B.2 and we conclude this chapter by presenting the performance of the quantizers we have developed in \mathbb{C}^4 for a 32, 16 and 8 user system and show how this may aid in one's choice of quantizer in a system of interest.

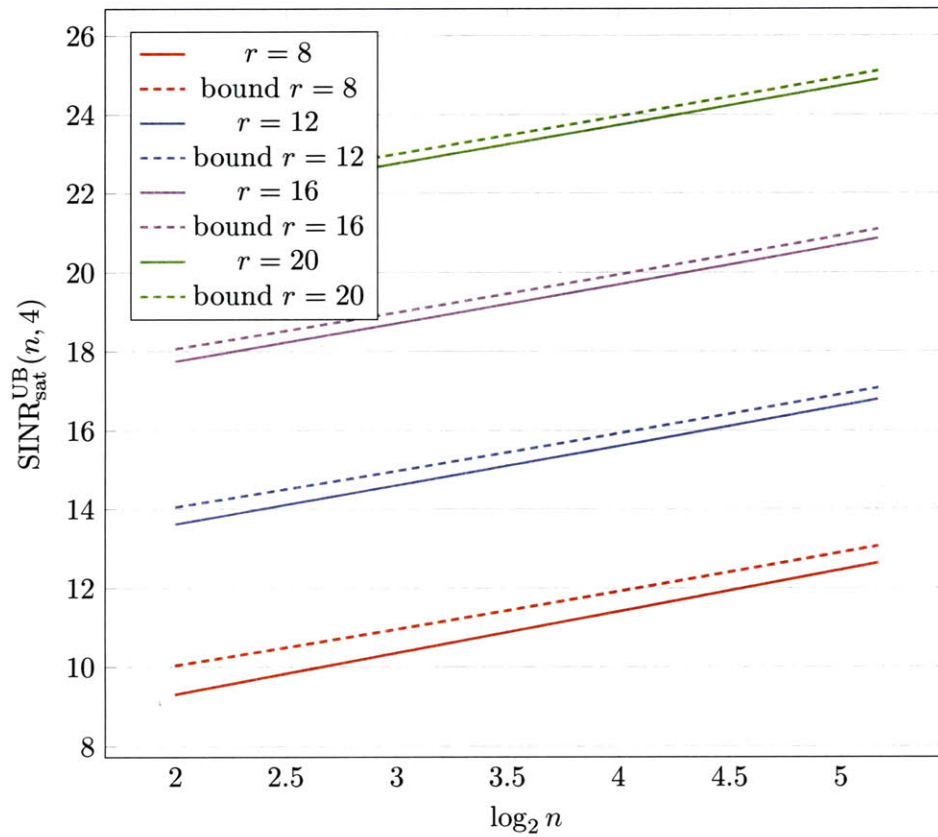


Figure 4-15. The upper bound $\text{SINR}_{\text{sat}}^{\text{UB}}(n, 4)$ in a n user system for various values of r as well as the upper bound on $\text{SINR}_{\text{sat}}^{\text{UB}}(n, 4)$, (B.6a). Note that the growth in the SNR is linear in $\log_2 m$ with slope $3/(m-1) = 1$ as predicted by (B.6a). The linear growth in r predicted by (B.6a) may also be observed through the difference of every pair of curves (lines).

■ 4.7 Practical System Design for Developed Quantizers

In this section we present the performance of the quantizers that we have developed in a system which uses the order statistic on the quantization error of each user to determine a candidate subset of users for scheduling in a system with 4 transmit antennas. We consider a 32, 16 and 8 user system. We label each code using the table in Table 4.1. Then we show how one may use these figures to design practical systems.

Figure 4-16 illustrates the performance in a 32 user system when *no order statistic* is used. This is the performance a system with *any number of users* will obtain. As previously seen our quantizers do quite well for 3 to 12 bits and are at most -0.81 dB from the optimal channel quantizer. We again note that codes with many orthogonal bases perform worse, in terms of the achieved value of SINR_{sat} , than codes with fewer orthogonal bases. However, one may see in Figures 4-16 – 4-19 the gap becomes smaller when the order statistics are considered. A useful example is the code $\mathcal{C}_{\text{ASC}}(3,0)$ labeled as (9,0). Note that this code contains 1097 orthogonal bases. When no order statistics are used this code performs approximately 1.6 dB worse than the optimal scheme. While this performance is still within the range of applicability, one may be compelled to use alternate schemes due to this large gap. However, this gap is cut in half when only the 4 users with the best quantization error are selected as seen in Figure 4-19. The evolution of the performance of this code may be seen in Figures 4-16 – 4-19. As the competing scheme only has 26 orthogonal bases it is wise, if scheduling and multiplexing complexity are of great concern, to use the code $\mathcal{C}_{\text{ASC}}(3,0)$ to increase the probability there is an orthogonal set. This trade-off may also be seen in the 10 bit code $\mathcal{C}_{\text{ASC}}(4,2)$ which contains 2289 orthogonal bases and is labeled by (10,1) as well as by the 11 code $\mathcal{C}_{\text{ASC}}(4,0)$ which contains 14577 orthogonal bases and is labeled by (11,0).

The gains seen in a 32 user system are depicted in Figures 4-16 – 4-19 may also be seen to a lesser extent in 16 and 8 users systems. In particular, the code $\mathcal{C}_{\text{ASC}}(3,0)$ has a gap that is approximately 0.95 dB in a 16 user system in which the 4 users with the best quantization error are selected and a gap that is approximately 1.2 dB in a 8 user system in which the 4 users with the best quantization error are selected. This may be seen in Figure 4-21 and Figure 4-22 respectively.

To intelligently design a 4 transmit antenna system one may use Figure 4-10 in conjunction with Figures 4-16 – 4-22 to determine an appropriate quantizer for a problem of interest. In particular in a 32 user system one may determine a value for $p_{\sigma,\rho}$ such that 8 users on average feedback. Then, given this value of $p_{\sigma,\rho}$ and prescribed probability of pre-selection success one may determine the number of orthogonal bases required to be contained in the quantizer using Figure 4-10. Finally, one may then turn to Figure 4-18 to select a quantizer which contains the required number of orthogonal bases and ensure that it has a tolerable mean squared quantization error. If this is not the case, one can reduce the prescribed probability of pre-selection success, determine the number of orthogonal bases required to be contained in the quantizer and then again turn to Figure 4-18 to select a quantizer which contains the require number of orthogonal bases and ensure that it has a tolerable mean squared quantization error. This may be repeated iteratively until one achieves a desired balance. We note that is may be done similarly for a 8 transmit antenna system using Figure 4-11. However, we do not provide 8 dimensional quantizer performance in this thesis. Plots for the performance of such codes may be found at [119].

Index (r, \perp -Bases)	\perp -Bases	Construction	Reference
(3,4)	4	$\mathcal{C}_Z^{(2,4)}(3; [[0, 1]])$	(3.6)
(3,Z1)	0	Hochwald 3-bit	[56]
(3,Z2)	0	WiMax 3-bit	[1, 143]
(4,8)	8	$\mathcal{C}_Z^{(2,4)}(3; [[0, 1]]) \cup \mathcal{C}_Z^{(2,4)}(3; [[1, 0]])$	(3.6)
(4,4)	4	MUB(4)/ $\mathcal{C}_T(2, [0, 0], 0)$	[61, 76]/(3.54)
(4,12)	12	$\mathcal{C}_T(2, [1, 0], 0)$	(3.54)
(5,26)	26	$\mathcal{C}_{ASC}^*(2, 2)$	Example 3.2.6
(5,36)	36	$\mathcal{C}_Z^{(2,4)}(3; [[0, 0], [0, 1]]) \cup \mathcal{C}_Z^{(2,4)}(3; [[1, 0]]) \cup \mathcal{C}_T(2, [0, 0], 0)$	(3.6), (3.54)
(5,32)	32	$\mathcal{C}_T(2, [0, 0], 0) \cup \mathcal{C}_T(2, [0, 0], 2)$	(3.54)
(5,12)	12	$\mathcal{C}_{sparse}^{(2,4)}(2)$	(3.7)
(6,105)	105	$\mathcal{C}_{ASC}(2, 0)$	[13]/Example 3.2.6
(6,16)	16	$\mathcal{C}_T(3, [1, 0], 0)$	(3.54)
(6,4)	4	$\mathcal{C}_F(0.6777, 0.5305 + 0.7425 \cdot i, \mathcal{C}_Z^{(2,4)}(3; [[0, 1]]))$	(3.11)
(6,Z3)	0	Hochwald 6-bit	[56]
(6,48)	48	$\mathcal{C}_{sparse}^{(2,4)}(3)$	(3.7)
(6,Z5)	0	WiMax 6-bit	[1, 143]
(7,233)	233	$\mathcal{C}_{ASC}(3, 2)$	Example 3.2.6
(7,112)	112	$\mathcal{C}_Z^{(2,4)}(4; [[0, 1]]) \cup \mathcal{C}_Z^{(2,4)}(4; [[1, 0]]) \cup \mathcal{C}_T(3, [0, 0], 0)$	(3.6), (3.54)
(7,128)	128	$\mathcal{C}_T(3, [0, 0], 0) \cup \mathcal{C}_T(3, [0, 0], 2)$	(3.54)
(7,192)	192	$\mathcal{C}_{sparse}^{(2,4)}(4)$	(3.7)
(8,393)	393	$\mathcal{C}_{ASC}(3, 1)$	Example 3.2.6
(8,4)	4	$\mathcal{C}_F(0.2303, 0.6817 + 1.9577 \cdot i, \mathcal{C}_T(2, [0, 0], 0))$	(3.11)
(8,768)	768	$\mathcal{C}_{sparse}^{(2,4)}(5)$	(3.7)
(9,1097)	1097	$\mathcal{C}_{ASC}(3, 0)$	Example 3.2.6
(9,26)	26	$\mathcal{C}_F(0.0100, 0, \mathcal{C}_{ASC}(2, 2))$	(3.11)
(10,2289)	2289	$\mathcal{C}_{ASC}(4, 1)$	Example 3.2.6
(10,1521)	1521	$\mathcal{C}_{ASC}(4, 2)$	Example 3.2.6
(10,26)	26	$\mathcal{C}_F(0.5872, 0.4628 + 0.6790 \cdot i, \mathcal{C}_{ASC}(2, 2))$	(3.11)
(11,14577)	14577	$\mathcal{C}_{ASC}(4, 0)$	Example 3.2.6
(12,105)	105	$\mathcal{C}_F(0.3639, 1.9529, \mathcal{C}_{ASC}(2, 1))$	(3.11)

Table 4.1. A list of quantizers in \mathbb{C}^4 developed with our channel quantization framework. The first column is used to index the simulated performance of each code in Figure Figures 4-16 – 4-22. The second column contains the number of orthonormal bases for \mathbb{C}^4 contained in the code and the last column contains a reference to the construction.

4.7.1 Performance of Developed Quantizers in 32 User Systems

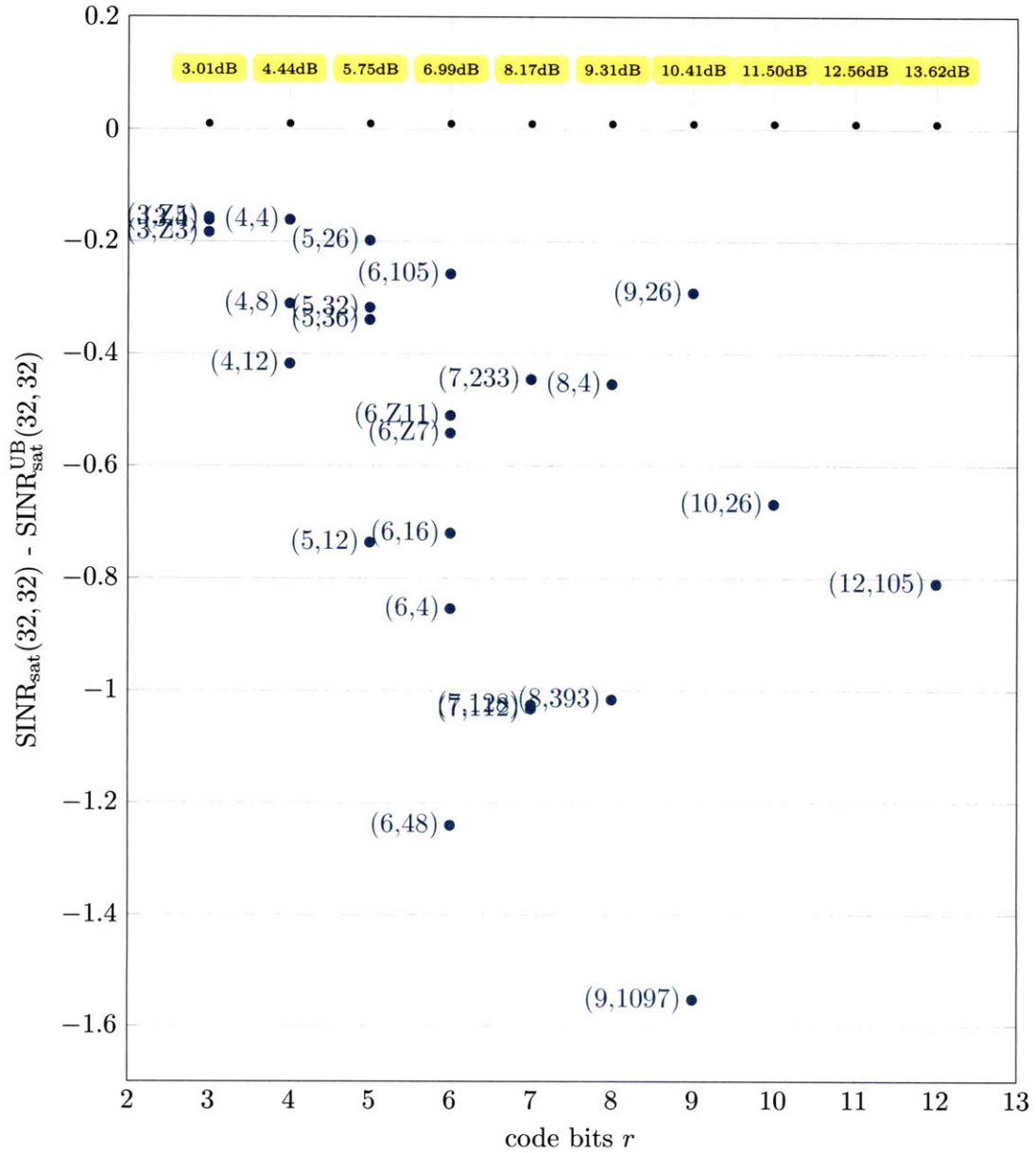


Figure 4-16. The performance of existing and developed quantizers in \mathbb{C}^4 relative to the upper bound $\text{SINR}_{\text{sat}}^{\text{UB}}(n, \ell)$ in a 32 user system. We note that as all users are considered the achieved performance is independent of the number of users in the system. The value taken by the upper bound is labeled at 0 for each rate. Each point corresponds to a specific quantizer as labeled in Table 4.1. For each point the average SINR_{sat} for the best 32 users is computed. Hence, for this example there is no exploitation of the order statistic. Note that the code corresponding to (9, 0) has approximately a 1.6 dB loss compared to the upper bound.

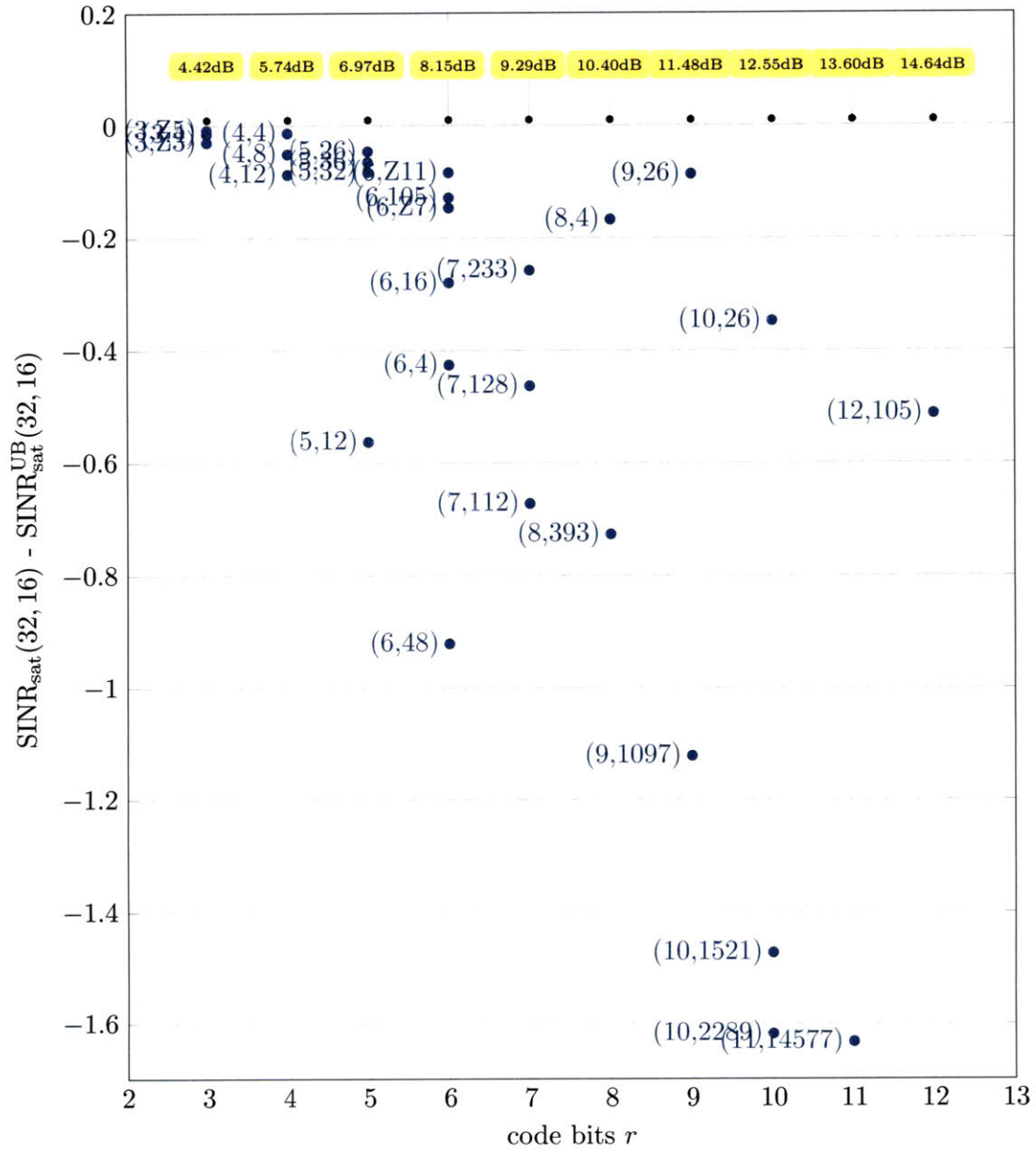


Figure 4-17. The performance of existing and developed quantizers in C^4 relative to the upper bound $SINR_{sat}^{UB}(n, \ell)$ in a 32 user system where only the 16 users which achieve the highest value of $SINR_{sat}$ at each scheduling interval are considered. The value taken by the upper bound is labeled at 0 for each rate. Each point corresponds to a specific quantizer as labeled in Table 4.1. For each point the average $SINR_{sat}$ for the best 16 users is computed. Note that the code corresponding to (9, 0) now has approximately a 1.1 dB loss compared to the upper bound.

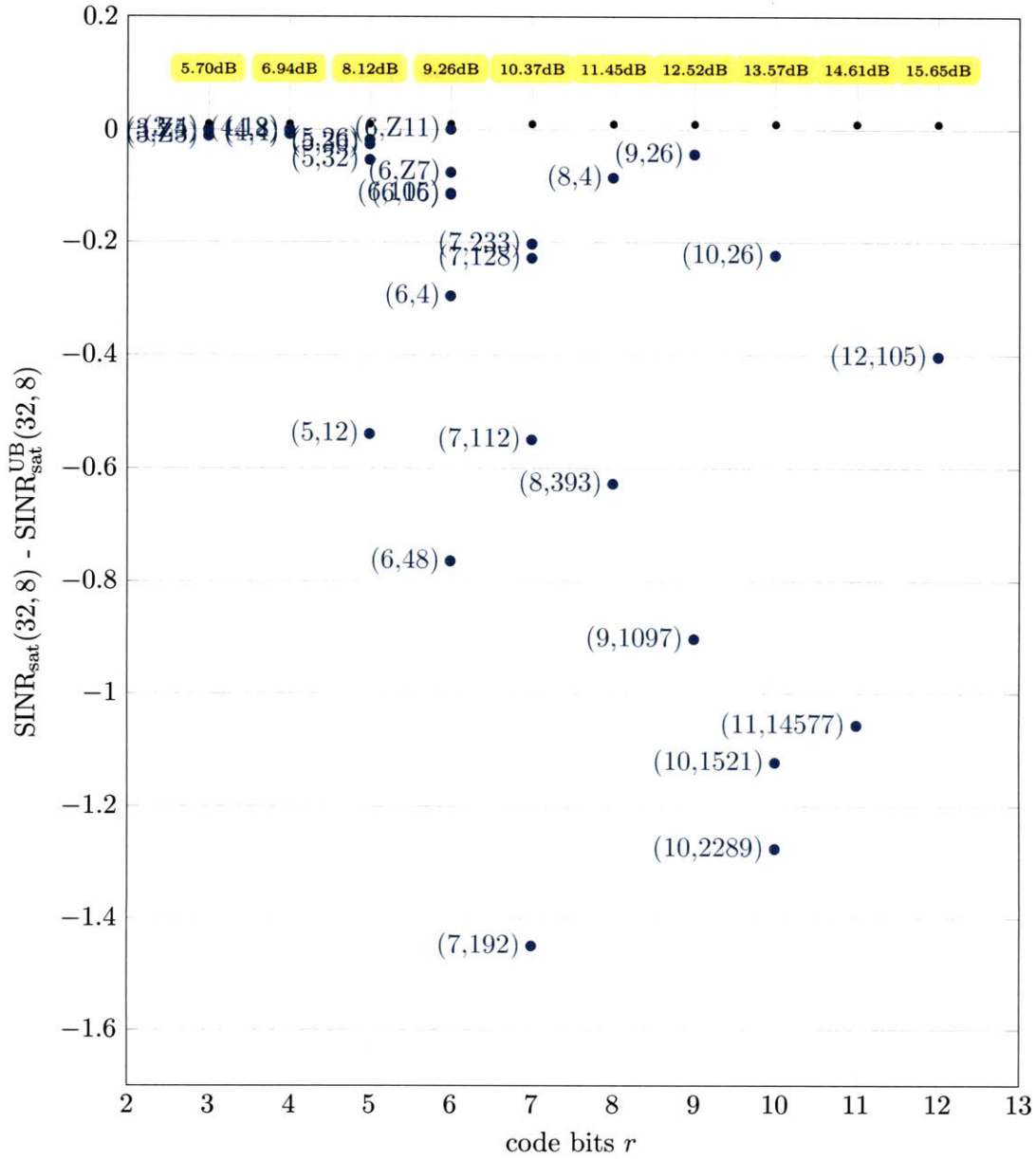


Figure 4-18. The performance of existing and developed quantizers in \mathbb{C}^4 relative to the upper bound $\text{SINR}_{\text{sat}}^{\text{UB}}(n, \ell)$ in a 32 user system where only the 8 users which achieve the highest value of SINR_{sat} at each scheduling interval are considered. The value taken by the upper bound is labeled at 0 for each rate. Each point corresponds to a specific quantizer as labeled in Table 4.1. For each point the average SINR_{sat} for the best 8 users is computed. Note that the code corresponding to (9,0) now has approximately a 0.9 dB loss compared to the upper bound.

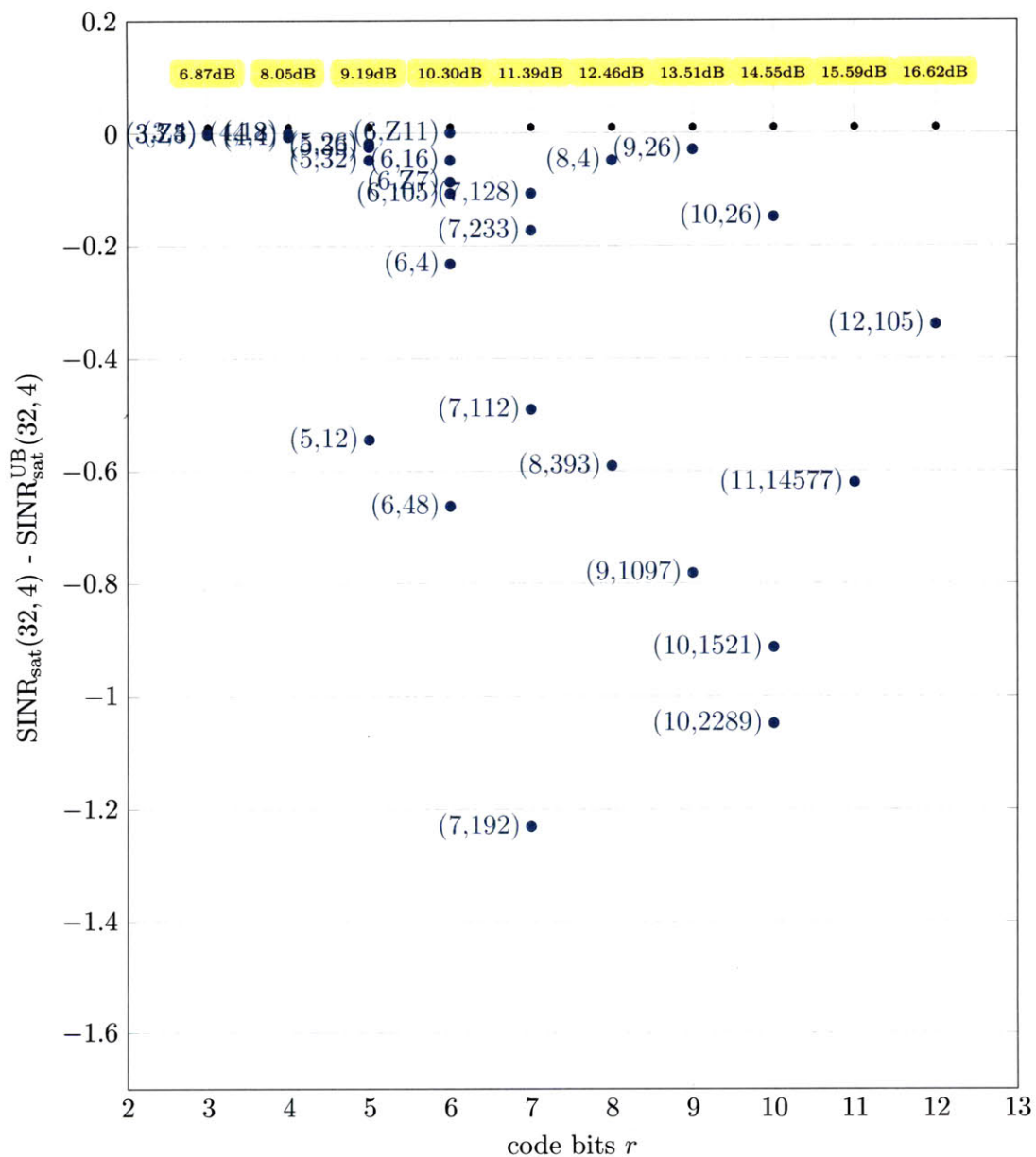


Figure 4-19. The performance of existing and developed quantizers in \mathbb{C}^4 relative to the upper bound $\text{SINR}_{\text{sat}}^{\text{UB}}(n, \ell)$ in a 32 user system where only the 4 users which achieve the highest value of SINR_{sat} at each scheduling interval are considered. The value taken by the upper bound is labeled at 0 for each rate. Each point corresponds to a specific quantizer as labeled in Table 4.1. For each point the average SINR_{sat} for the best 4 users is computed. Note that the code corresponding to (9,0) now has approximately a 0.8 dB loss compared to the upper bound.

4.7.2 Performance of Developed Quantizers in 16 User Systems

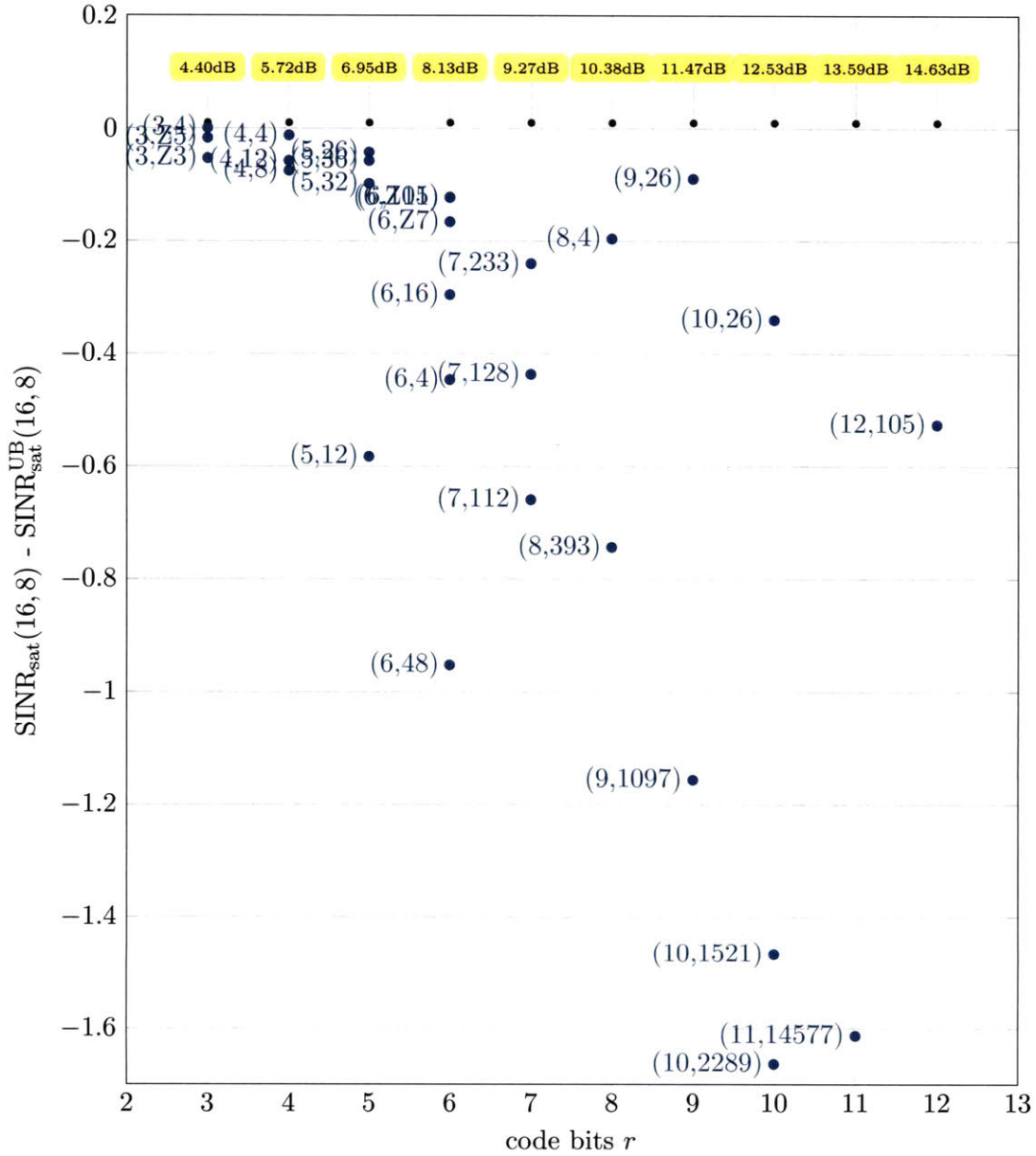


Figure 4-20. The performance of existing and developed quantizers in \mathbb{C}^4 relative to the upper bound $\text{SINR}_{\text{sat}}^{\text{UB}}(n, \ell)$ in a 16 user system where only the 8 users which achieve the highest value of SINR_{sat} at each scheduling interval are considered. The value taken by the upper bound is labeled at 0 for each rate. Each point corresponds to a specific quantizer as labeled in Table 4.1. For each point the average SINR_{sat} for the best 8 users is computed. Note that the code corresponding to $(9, 0)$ now has approximately a 1.15 dB loss compared to the upper bound.

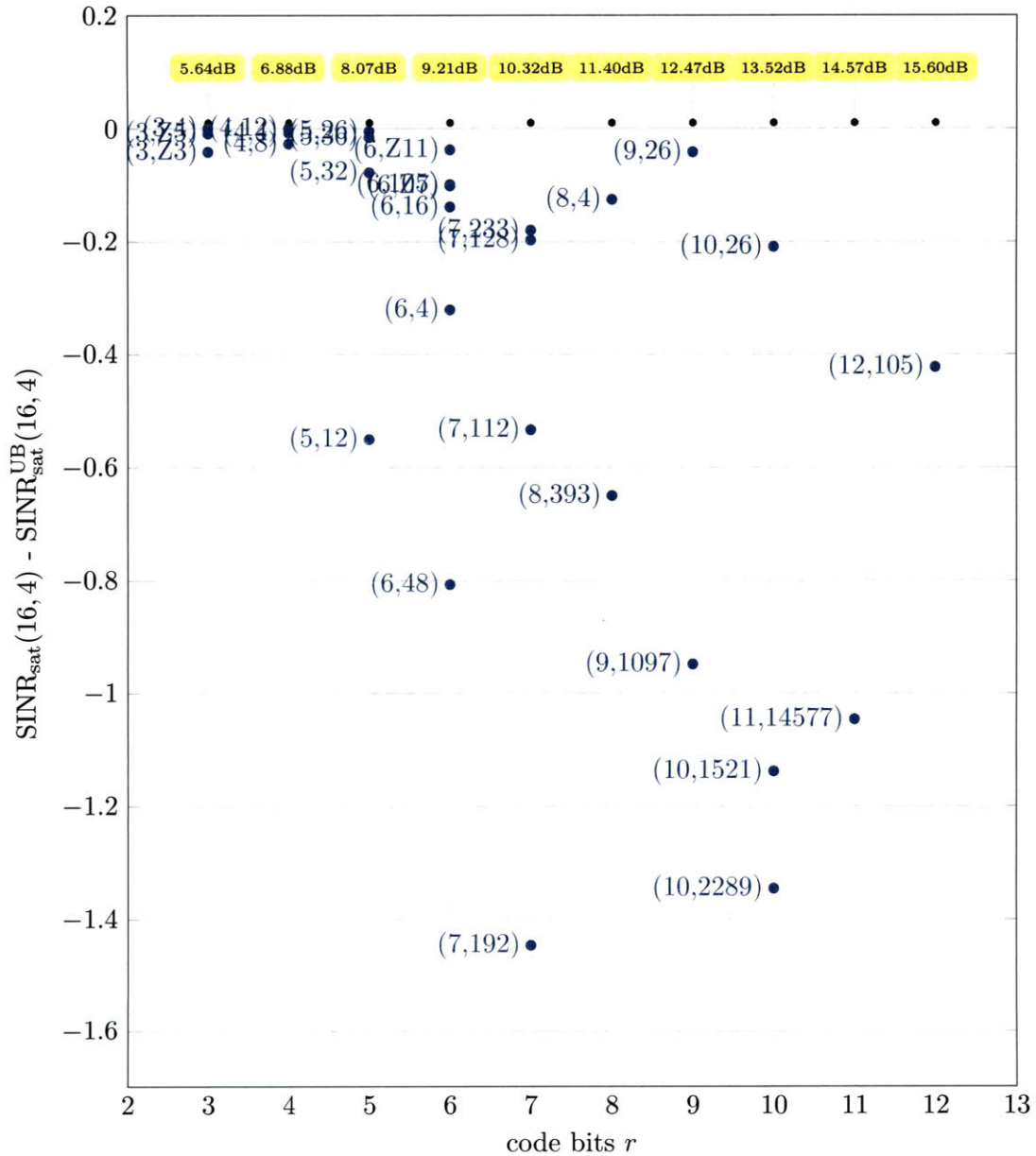


Figure 4-21. The performance of existing and developed quantizers in \mathbb{C}^4 relative to the upper bound $\text{SINR}_{\text{sat}}^{\text{UB}}(n, \ell)$ in a 16 user system where only the 4 users which achieve the highest value of SINR_{sat} at each scheduling interval are considered. The value taken by the upper bound is labeled at 0 for each rate. Each point corresponds to a specific quantizer as labeled in Table 4.1. For each point the average SINR_{sat} for the best 4 users is computed. Note that the code corresponding to (9, 0) now has approximately a 0.95 dB loss compared to the upper bound.

4.7.3 Performance of Developed Quantizers in 8 User Systems

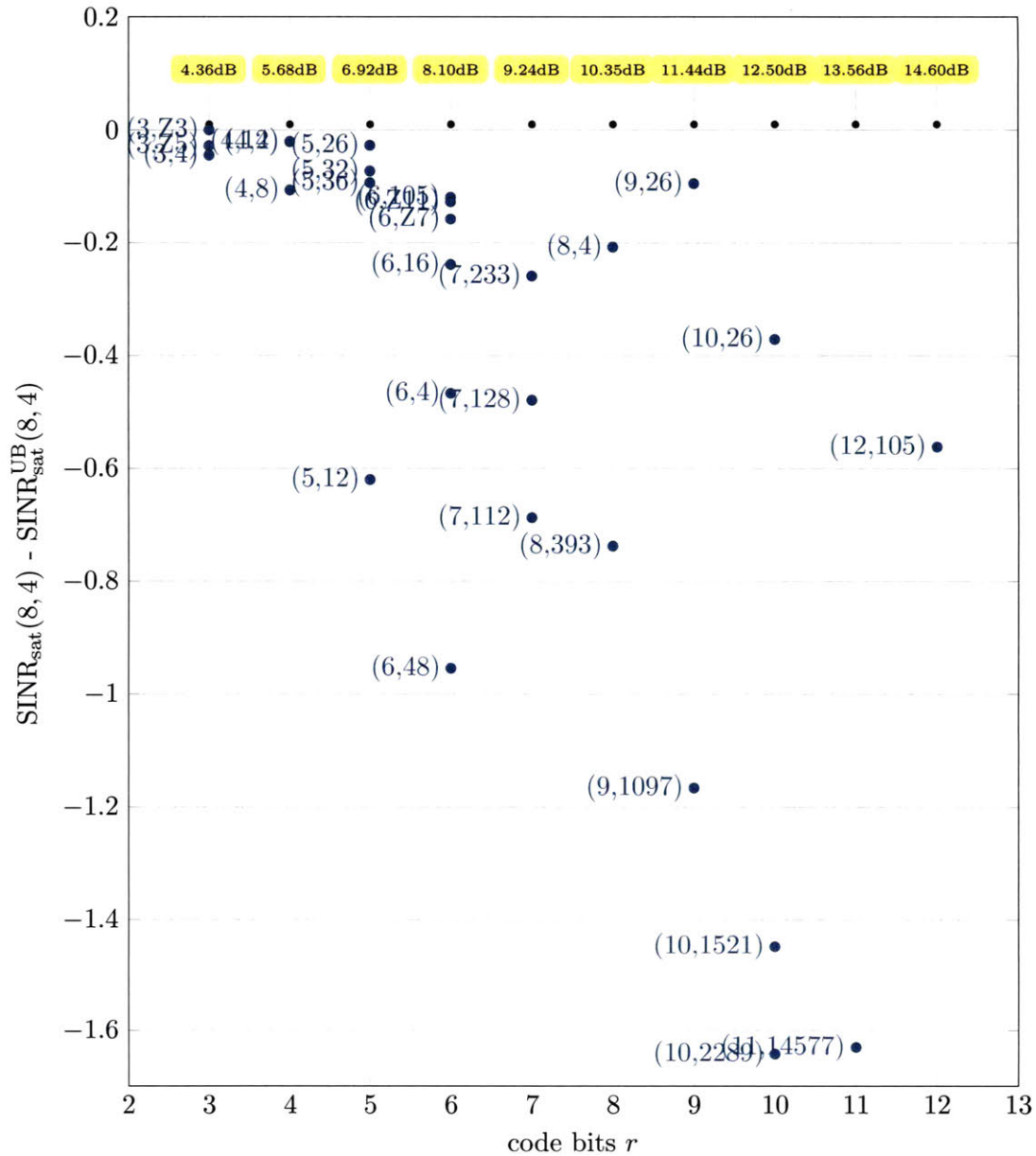


Figure 4-22. The performance of existing and developed quantizers in \mathbb{C}^4 relative to the upper bound $\text{SINR}_{\text{sat}}^{\text{UB}}(n, \ell)$ in a 8 user system where only the 4 users which achieve the highest value of SINR_{sat} at each scheduling interval are considered. The value taken by the upper bound is labeled at 0 for each rate. Each point corresponds to a specific quantizer as labeled in Table 4.1. For each point the average SINR_{sat} for the best 4 users is computed. Note that the code corresponding to (9, 0) now has approximately a 1.15 dB loss compared to the upper bound.

Multi-User MIMO Systems Design with Non-Rayleigh Fading

Original analysis of MIMO wireless systems have shown the potential for increasing wireless system capacity with out the price of power or bandwidth [126] by exploiting the spatial degrees of freedom available multiple transmit and receive elements. It is well understood that the capacity of a wireless communication channel scales linearly with the number of spatial degrees of freedom. These results stem from the path diversity afforded by the MIMO channel previously described under the assumption all transmit and receive pairs are independent and identically distributed. However, the characteristics of MIMO wireless systems rely heavily on the underlying wireless channel and correlations between the transmit and receive elements can be shown to be a limiting factor in MIMO systems [35, 112]. Thus, understanding the correlation and more generally the expected power coupled between the transmit and receive elements should play an import role in modeling the multi-user MIMO channel [133, 134].

In the pioneering work of Teletar [126] and Foschini and Gans [51] it was shown that under the assumptions of the Rayleigh model the capacity of the MIMO channel scales approximately linearly in the minimum of the number of transmit and receive elements. Hence, under the assumptions of the Rayleigh model the capacity of the MIMO channel may grow unbounded if one may simultaneously increase the number of transmit and receive elements. However, in general there are physical limitations on the length and/or area of an antenna array. As such, one would expect that packing more and more antennas into a fixed area will make the fading process between the transmit and receive pairs correlated thus limiting the capacity growth. It has been shown that the physical constraints of antenna arrays and the underlying propagation environment put deterministic limits to the spatial degrees of freedom [99]. In particular, constraints on the areas of the transmit array and receive array led to deterministic limits on the spatial degrees of freedom [99]. As such it is natural to wonder what, if any, limits are put on the multi-user MIMO system as the number of user grows above the number of transmit elements in light of these limitations on the spatial degrees of freedom. Moreover, in an environment with a finite number of scatterers it is likely that as the user population grows there is some subset of users that will be positioned such that the scattering characteristics of the propagation paths for each user's signal are similar. Hence, it is not unreasonable to expect that the propagation environment may have limited degrees of freedom which has a significant influence the structure of the joint fading process in a multi-user MIMO system. Thus, in the sequel we examine explicit ways to characterize these effects.

■ 5.0.4 Physical Modeling and Measurements of MIMO Channels

If one has knowledge of the geometric structure of the propagation environment one may aim to reproduce the actual signal propagation for that environment. That is, one may deterministically compute the multipath components including amplitude and delay for each transmitted signal. More precisely, one may store the geometric and electromagnetic characteristics of the site and simulate the corresponding propagation process. However, the results are only relevant for the specific site measured and in general one must repeat this process many times to get an accurate model for channel. In urban environments one of the most appropriate methods to physically model the channel, taking the actual physical propagation environment in to account, is ray tracing. As one expects many multipath components to dominate the characteristics of the fading process in such an environment, one may consider a set of “rays” emanating from the transmit antennas and arriving at the receiver. Each ray models the radio wave interaction with scatterers in the propagation path. To determine the rays to use in the ray tracing model one constructs the so-called visibility tree to represent the particular propagation environment. This visibility tree is computed by recursively adding nodes to the visibility tree corresponding to line of sight paths between objects. More precisely, the ray tracing algorithm may be described as follows: One begins by adding the transmitter as the root of the tree, every scatterer that has line of sight path with the transmitter is added as a leaf. For each one of these leafs every scatterer with line of sight path to a leaf is added and this process is repeated until a desired number of layers has been reached or the receiver is contained as a leaf. Each branch that contains the receiver as a leaf is then selected as a ray. The ray tracing algorithm has the nice property that once the visibility tree has been built it is a simple process to determine the statistics of the fading process by backtracking from each leaf to the root incorporating an appropriate physical rule at each step to determine the amplitude and delay of the path. We note that repeating this process for each user may produces visibility trees with common branches. If this is the case, and the branches significantly contribute to the fading, then two such users may have a quite similar fading process. Any two users who have similar fading process in the sequel we say have *clustered fading* or form a *cluster*. We note that this definition does not imply a spatial relation between the users nor correlated channels. This model rather simply describes a similarity in signal propagation and hence have similarly spatially correlated fading, i.e. two users i and j form a cluster if and only if

$$\mathbf{K}_{\mathbf{h}_i} \approx \mathbf{K}_{\mathbf{h}_j}.$$

Geometry based models, such as ray-tracing, are determined by the particular scatterer location and hence only succeed at modeling a specific site. In order to form a more applicable model one may rather consider randomly placed scatterers and then model the statistics of the resulting fading. While these physical models provided valuable insights into how the physical environment effects the signal propagation and hence how one should model the channel, they do little to help with our analysis as they do not fully describe the channel impulse response (2.9) or provide an analytic model for the distribution of the fading. Recall from (2.2) the fading coefficient for a single-antenna system in a narrowband flat fading channel could be derived by the system transfer function for the single transmit and receive pair. In a multi-user MIMO channel one must specify nm such transfer functions to characterize the system fully (analytically). However, in the presence of a finite number of scatterers and/or clustered users it is unclear what relationships the physical structure of

the propagation environment has on the nm transfer functions that characterize the system. In order to make realistic assumptions about the choice of model and associated parameters we first review some preliminary results on the role that the structure of the propagation environment has on the multi-user MIMO channel.

The full structure of the fading characteristics of the multi-user MIMO channel are only beginning to emerge through empirical measurements [8, 19, 37, 42, 71, 72] which point to a channel model for which the users channel are spatially correlated. In particular, [71] examined a MIMO system with 4 transmit elements and 4 user where the users are in an indoor environment, outdoor environment near the transmit base or an outdoor environment far from the transmit base. These measurements showed that the receive end covariance

$$R_{RX} = \mathbb{E} \left[\mathbf{H}\mathbf{H}^\dagger \right]$$

is roughly diagonal while the transmit covariance

$$R_{TX} = \mathbb{E} \left[\mathbf{H}^\dagger \mathbf{H} \right]$$

can take on one of many forms depending on the location of users. In particular, [71] showed that the transmit covariance of the users was approximately uniform and dense for users that were near the transmit base, non-uniform and dense for indoor users and sparse for users far from the base. Further, these measurement campaigns as well as the analysis of some simple scattering models for multi-user MIMO channels [139] have shown that the assumption that each transmit and receive pair follow *i.i.d* Rayleigh fading is more often than not an exception rather than the norm. As such, it is unreasonable to assume in general that multi-user MIMO channel follows the Rayleigh model, but rather one should assume a more general model for which the Rayleigh model is a particular case. Thus, in a multi-user MIMO system one expects heterogeneity in user fading not only in magnitude but also direction if a large geographic region is to be served. In particular, in a multi-user MIMO system users may form clusters. Thus, any model for a multi-user MIMO system should have degrees of freedom to model the *structure* as well as rank of the covariance matrix.

From the above discussion it is clear that the realistic models for the multi-user MIMO channel should have some way to account for geometry of the propagation environment. Moreover, the parameters of such a model should have some way to map physical channel measurements and other physical prior information in a way as to accurately predict the relevant figures of merits. In the sequel we show these degrees of freedom play an important role in the performance of a multi-user MIMO system. As such we seek an analytic model that will allow one, with some underlying knowledge of the scattering environment, to accurately (and tractably) model the channel. As direct knowledge of the location of scatters will not be useful in our analysis and will be cumbersome to use at the transmit base we rather select a more analytical approach to multi-antenna channel modeling as to ignore the physical properties of the scattering objects focusing rather on directly modeling the correlation of the fading coefficients between the transmit and receive pairs. Moreover, as we have previously shown the order statistic gain and multi-node matching gain are of fundamental importance. Thus, in the sequel we develop our choice for the channel model of a multi-user MIMO system. Then in the sequel, we provide a discussion which parallels this develop to select a discrete model for the user assignment distribution.

■ 5.0.5 Analytic Models for the MIMO Channel

Physical models and measurements have shown that the multi-user MIMO channel should be assumed to have spatially correlated users. Thus, it is our goal to develop an analytic model for the multi-user MIMO channel. However, the model must be chosen in a way that enables one to model the relevant aspects of the problem of interest in as simple a way as possible to enable simple analysis and estimation of the channel. The original analysis of multiple-antenna systems assumed tractable and practically motivated models in order for system designers to determine the fundamental limits of such systems. The simplest such analytic model for the $m \times n$ MIMO channel is the Rayleigh model which describes the fading process as

$$\mathbf{H}[k] = \mathbf{G}[k]$$

where $\mathbf{G}[k]$ is an $m \times n$ random matrix with elements distributed as *i.i.d* zero-mean complex Gaussian variable with variance $1/2m$. Such a model assumes a rich scattering environment that is uncorrelated. However, as previously noted, there is sufficient evidence that the link pairs in a MIMO channel are spatially correlated [8, 19, 37, 42, 71, 72] and as such one more generally wishes to introduce parameters into the channel model to capture the correlation of the links. In the most general such parametrization one must prescribe the correlation of each of the nm transmit and link pairs resulting in n^2m^2 free parameters. That is, the most general parametrization must describe the $nm \times nm$ full channel covariance matrix,

$$R_{\mathbf{H}} = \mathbb{E}_{\mathbf{H}} \left[\text{vect}(\mathbf{H})\text{vect}(\mathbf{H})^\dagger \right] \quad (5.1)$$

This approach involves many free parameters and several models have been proposed that imposes particular structures on the MIMO covariance matrix to reduce the number of free parameters. As we are interested in the system capacity we seek a model which sufficiently captures enough of the physical structure of the multi-user MIMO channel to predict system performance, while the parametrization is of low enough dimension to allow for simple and direct mappings from channel measurements and other prior information about the channel to be incorporated with low complexity. In the sequel we examine a few simple extensions of the Rayleigh model. As such, throughout this section we let $\mathbf{G}[k]$ be an $m \times n$ random matrix with elements distributed as *i.i.d* zero-mean complex Gaussian variable with variance $1/2m$.

In order to reduce the number of free parameters one has in describing the channel covariance one may rather consider the correlation at both link ends. That is, instead of considering the full channel correlation matrix (5.1) one may rather consider the $m \times m$ and $n \times n$ matrices

$$R_{\text{TX}} = \mathbb{E}_{\mathbf{H}} \left[\mathbf{H}^\dagger \mathbf{H} \right] = \mathbf{U}_{\text{TX}} \mathbf{\Lambda}_{\text{TX}} \mathbf{U}_{\text{TX}}^\dagger$$

and

$$R_{\text{RX}} = \mathbb{E}_{\mathbf{H}} \left[\mathbf{H} \mathbf{H}^\dagger \right] = \mathbf{U}_{\text{RX}} \mathbf{\Lambda}_{\text{RX}} \mathbf{U}_{\text{RX}}^\dagger.$$

Any model that uses these matrices directly has $m^2 + n^2$ degrees of freedom as opposed to the n^2m^2 free parameters of the full covariance matrix. One of the most useful models for our purposes which uses this decomposition is the Weichselberger model. While the Weichselberger model is not specifically developed for the multi-user MIMO channel it is particularly appealing for our use as it succinctly characterizes the spatial degrees of freedom in the MIMO channel. In particular, it captures the expected energy that is

transferred between the modes of the transmitter and receive elements. More precisely, the Weichselberger model for the MIMO channel [134] is

$$\mathbf{H} = \mathbf{U}_{\text{RX}}^\dagger \left(\tilde{\mathbf{\Omega}} \odot \mathbf{G} \right) \mathbf{U}_{\text{TX}}^\dagger$$

where $\tilde{\mathbf{\Omega}}$ is the element wise square root of the $n \times m$ coupling matrix $\mathbf{\Omega}$ which describes the expected energy coupled between the transmit and receive eigenmodes. The matrix $\mathbf{\Omega}$ is easily obtained from measurements of the MIMO channel [8, 134] via the relationship

$$\Omega_{i,j} = E_{\mathbf{H}} \left[\left| \mathbf{u}_{\text{RX},i}^\dagger \mathbf{H} \mathbf{u}_{\text{TX},j}^* \right|^2 \right]$$

where $\mathbf{u}_{\text{RX},i}$ and $\mathbf{u}_{\text{TX},j}$ are the eigenvectors of the receive side and transmit side covariances respectively. As the Weichselberger model uses both the eigenmodes of the transmit and receive covariance as well as the coupling matrix the full model has $n(n-1) + m(m-1) + nm$ real parameters. This general model is of particular interest as there are nm model parameters that coarsely captures the *spatial structure* of the MIMO channel. In particular, the structure of the coupling matrix determines whether the underlying propagation environment has rich scattering as in the Rayleigh model (by taking $\Omega_{i,j} = 1$) or reflects a more sparse environment where $\Omega_{i,j} = 0$ for a large number of i and j pairs. In the sequel we say a channel is a sparse multi-path channel if $\Omega_{i,j} = 0$ for a large number of i and j pairs and say a channel is a dense multipath channel otherwise. This distinction is important as this will determine how one may exploit the gains inherent in MIMO systems. Most importantly the spatial structure of the coupling matrix will largely influence the design of the channel quantization. Indeed, if the propagation environment is a sparse multipath channel then one should restrict the channel quantization to the subspace(s) in which most of the transmit energy propagates. However, there have been many other approaches presented in literature to model correlation in the MIMO channel. We present on such a model in the sequel as it will further motivate our model for the multi-user MIMO channel as well as our choice of model for the user assignment distribution. We have previously mentioned the Rayleigh model has been beneficial in modeling the MIMO channel assuming *i.i.d* fading. However, the Kronecker model has become the one of the most popular and commonly used analytical models for a correlated MIMO channel [8]. We briefly discuss the Kronecker model with a particular emphasis on its deficiencies in modeling the multi-user MIMO channel.

The Kronecker model factorizes the channel correlation matrix into a product of the marginal covariance matrices of the link-ends, R_{TX} and R_{RX} . In particular, the Kronecker model for the MIMO channel selects

$$\mathbf{H} = R_{\text{RX}}^{1/2} \mathbf{G} R_{\text{TX}}^{1/2}. \quad (5.2)$$

It is clear from the definition that this model requires the specification of R_{TX} and R_{RX} directly and as such this model has the aforementioned $n^2 + m^2$ degrees of freedom. However, in an attempt to simply model the correlation the Kronecker makes an implicit assumption that the joint DOA-DOD spectrum is separable which has a large effect on the capacity [134]. Moreover, rewriting (5.2) using the eigenvalue decompositions of R_{TX} and R_{RX} one equivalently has

$$\mathbf{H}[k] = \mathbf{U}_{\text{RX}} \Lambda_{\text{RX}}^{1/2} \mathbf{G}[k] \Lambda_{\text{TX}}^{1/2} \mathbf{U}_{\text{TX}}^\dagger = \mathbf{U}_{\text{RX}} \left(\boldsymbol{\lambda}_{\text{RX}} \boldsymbol{\lambda}_{\text{TX}}^\dagger \odot \mathbf{G} \right) \mathbf{U}_{\text{TX}}^\dagger \quad (5.3)$$

where λ_{RX} and λ_{TX} are the vectors of eigenvalues for the receive and transmit covariances respectively. Thus, the Kronecker model also makes an implicit assumption on the propagation environment. Indeed, from (5.3) the Kronecker model implicitly assumes that the energy coupled between the eigenmodes of the transmitter and the receiver is rank one. In a single-user system this may provide an adequate model for the MIMO system in some cases. However, in a multi-user system an assumption of a rank one coupling matrix is too restrictive as the propagation paths associated to different eigenmodes may fade quite differently for different users. In particular, consider a single coordinated MIMO system where one link end is spatially rich while the other end is rank deficient (i.e. environments where there may be physical obstructions that reduce the rank of the covariance at one link end while the other link end has many local scatterers). In the Kronecker model this scenario is modeled through a rank deficient covariance at one of the link ends and hence the rank one coupling matrix may null the correct modes of the system. However, consider a more general case in which the transmit array is spatially rich while a geographically distributed, uncoordinated receive array is amongst a collection of physical obstructions that null certain *pairings of transmit and receive modes*. Such a scenario will be poorly captured by the Kronecker model as it is rank one. However, there are sufficient degrees of freedom in the Weichselberger model to capture this scenario. Thus, in order to model the MIMO channel accurately in terms of both the correlation as well as the capacity one must use a more general model, such as the Weichselberger model, to accurately model the possible energy couplings between transmit and receive modes of the system.

The particular structure of the coupling matrix in the Weichselberger model determines whether the underlying propagation environment is a sparse or dense multipath channel which strongly influences the quantizer design. This is important as this will determine how one may exploit the gains inherent in MIMO systems. Most importantly the spatial structure of the coupling matrix will largely influence the design of the channel quantization. In Chapter 3 we developed a method to exploit knowledge of the spatial structure of the fading. Indeed, using our systematic construction one may select only sparse codes to quantize sparse multi-path channels or use the construction as described to quantize dense multipath channels. However, as we have stated previously one in general does not have prior knowledge of the spatial correlation of the fading process for each user in the system for every deployment site. In order to effectively design a feedback scheme we must infer this structure and develop a quantization scheme that may adapt to this knowledge.

■ 5.1 Modeling the User Assignment Distribution

Current MIMO systems must be developed in a way as to be robust to a variety of radio environments to be easily (and quickly) deployed on a large scale. To do such a system designer may design a system under some minimum number of assumptions (for example number of users, user mobility etc.) while leaving free a few degrees of freedom in the design which may be set independently at each deployment site. An even more desirable approach is to design a system that may infer these parameters through some set of minimal training data as this removes much of the complexity of system deployment as well as provides the system with the ability to adapt to possible future changes in the radio environment. The simplest approach to providing this functionality is to design a feedback link for users of the system to report the current state of their radio channel to the transmitter. However, if the radio propagation environment is unknown one may not be able to simultaneously design a feedback scheme that has a tolerable quantization error for a variety of fading environments

for a given feedback rate. If this is the case then one may rather design a feedback scheme that is good for a large class of fading distribution and provide a mechanism with which a system may adapt the feedback scheme in environments where this scheme performs poorly. To illustrate how this may be done we first consider a simple example for a single-antenna system then generalize it to a MIMO channel.

In a single-antenna system, one is generally only interested in the received signal power as this describes the instantaneous capacity of the channel (2.4). Hence, in a single-antenna systems one typically only models the magnitude of the fading. Common models for the magnitude of the fading are

- The Rayleigh model, which assumes non-line of sight signal propagation where by the fading coefficients $h[k]$ is modeled as a complex Gaussian random variable
- The Rician model, where the received signal contains a significant line of sight component

We note that both of these models reflect an underlying assumption on the location and distribution of the scatterers. Both assume a large number of scatterers while the Rician model assume a scatterer is not present in the line of sight while the Rayleigh model does. If the transmitter has perfect knowledge of $|h_i[k]|^2$ for each user over a sufficiently long length of time it is a relatively simple process to determine which model to use based on this series of observations. To see this, recall that the Rician distribution is for $x \geq 0$

$$\text{Rician}(x, A, \sigma) = \frac{x^2}{\sigma^2} \exp\left(-\frac{x^2 + A^2}{2\sigma^2}\right) I_0\left(\frac{Ax}{\sigma^2}\right)$$

where $I_0(x)$ is the modified zeroth order Bessel function of the first kind (see [10] for the exact expression for this function). Furthermore, recall that the Rayleigh model is simply a Rician distribution for which $A = 0$. Hence for the transmitter to make an inference on the underlying propagation environment for any user in the system (i.e. whether the user has line of sight or non-line of sight) it is sufficient to test whether $A = 0$ or $A > 0$ based on a series of observations of $h_i[k]$. Hence for each user one may define a hidden random variable

$$Z_i = \begin{cases} 0 & \text{if user } i \text{ channel follows Rayleigh model} \\ 1 & \text{if user } i \text{ channel follows Rician model} \end{cases}$$

and form a maximum likelihood estimate for each Z_i in an attempt to determine whether user i has line of sight. One can show that the ML estimate for Z_i after perfectly observing the k -th fading coefficient is [114]

$$\hat{Z}_i = \begin{cases} 0 & \text{if } \frac{1}{k} \sum_{i=0}^{k-1} |h[k]|^2 \leq 2\sigma^2 \\ 1 & \text{otherwise} \end{cases}$$

Thus, in practice if the transmitter is informed of the fading state of each user it can infer relevant aspects of each users propagation environment (in this case line of sight) to help optimize scheduling decisions.

It is important for any feedback design to account for the underlying fading process and/or be able to adapt to estimates of the fading process. For example, if one uses finite rate feedback to convey the channel state in a single-user system the optimal scalar quantization scheme depends heavily on the value of A in the Rician model. Thus, if

estimates of the fading process are not taken into account a feedback design may have to use a scheme of much higher rate than needed or an intolerable distortion in the channel representation may be incurred. We address how one may design intelligent feedback scheme which may adapt to fluxuations in the Chapter 3 after first considering the relevant figures of merit for single-user systems and sufficiently generalizing the present discussion to the multi-user MIMO system.

In Section 4.2 we asked the question whether a single generalized switch could be constructed which, with high probability, solves global scheduling problem as this indicates when the multi-user MIMO system behaves like a wire-line system. In the sequel we provide this generalized definition. However, as seen in the single-antenna broadcast system the figures of merit (for both the outage probability as well as the ergodic capacity) rely heavily on the distribution of the fading process and hence one must accurately model the fading process for the results to be meaningful. In a single-antenna system the effects of user dynamics and the geometry of the propagation environment are well understood [100,128]. Indeed, we have previously described two models for the single-antenna system that stem from different assumption on the line of sight. In the MIMO channel there are far more effects that must be modeled which not only effects the system throughput but also the feedback design. In particular, one must model the effects of the array geometry, electromagnetic coupling of the transmit elements as well as the co-channel interference between the different users.

In the preceding example for a single-antenna system we showed that one may make a coarse estimate of the underlying propagation environment through observations of each users fading state. In the sequel we describe how one may do similarly in the multi-antenna case. That is, in the sequel we examine how one may, through observations of the feedback process, form an estimate of the coupling matrix $\mathbf{\Omega}$. However, we do not do this directly. As we have shown in Section 2.2.1 one may first estimate the assignment distribution for each user, $\hat{\mathbf{p}}_i$, and then use this distribution to form an estimate of each users channel covariance and hence of $\mathbf{\Omega}$. This approach is of interest as it may be used more broadly throughout the system to influence scheduling decisions and adapt the feedback scheme. In particular, by first estimating the user assignment distribution one can aid in the search for a maximal matching in the BRS model by identifying the switches that are most likely to contain the clique of maximum weight. This approach is a far more attractive approach than estimating $\mathbf{\Omega}$ directly as it is quite cumbersome in general to compute the user assignment distribution from knowledge of $\mathbf{\Omega}$ as exact computation of the user assignment distribution requires the evaluation of multidimensional integrals. Thus, in the sequel we develop an appropriate model for the user assignment distribution. However, to estimate the user assignment distribution one must first find an appropriate class of discrete distributions to model this process.

Any model for the user assignment distribution in the multi-user MIMO channel should be strongly tied to the underlying distribution of the multi-antenna channel. Indeed, from Section 2.2.1 we have shown that the probability that any user is quantized to a specified codeword is simply the integral over the Voronoi cell of the codeword where the measure used in the integration is tied to the channel covariance. If the Weichselberger model has degenerated to the Rayleigh model then, given that the codebook is symmetric, it is easy to see that it is equally likely that a user is quantized to every codeword. This result stems from the fact that the Rayleigh model assumption corresponds to an assumption that each user's channel vector is isotropically distributed and hence uniformly distributed over the unit sphere. If the Voronoi cells for each user are isomorphic then the associated integrals (2.16)

are equivalent yielding a uniform distribution. That is, assuming the Rayleigh model and a symmetric quantizer, the probability that any (every) user is quantized to a given codeword index, say \mathbf{c}_j , is

$$\Pr(\mathcal{Q}(\mathbf{h}_i) = \mathbf{c}_j \mid \text{Rayleigh model}) = \frac{1}{2^r}. \quad (5.4)$$

As we have previously discussed many measurements for MIMO channels [8, 19, 37, 42, 71, 72] have shown that the Rayleigh model is not a good model for the multi-user MIMO channel and hence it is unreasonable to assume the uniform distribution (5.4) is a good model for the user assignment distribution. That is, measurements have shown these channels to have an underlying correlation structure that is not captured by the *i.i.d* white process described by the Rayleigh model. Thus, one minimally expects the user assignment distribution to depend on the codeword index j . Rather, if one assumes every users signal follows a similar propagation path (i.e. assuming an appropriately rotated Kronecker model for the channel) the probability that any (every) user is quantized to a given codeword index, say \mathbf{c}_i , is, for some set of positive real numbers $\{p_{i,0}, p_{i,1}, \dots, p_{i,2^r-1}\}$ such that $\sum_{j=0}^{2^r-1} p_{i,j} = 1$,

$$\Pr(\mathcal{Q}(\mathbf{h}_i) = \mathbf{c}_j \mid \text{Kronecker model}) = p_{i,j}. \quad (5.5)$$

Thus, the underlying assumption on the channel model and in particular the coupling matrix $\mathbf{\Omega}$ strongly influences the associated sampling distribution. As the Weichselberger model generalizes the Rayleigh model and the Kronecker model one must find the appropriate generalization of (5.4) and (5.5) to accurately model the MIMO channel feedback process.

In Section 5.0.5 the Weichselberger model was chosen as the model for the fading process as it has been shown to effectively model the instantaneous mutual information between the transmit array and a cooperative receive array well. The important aspect of this model was that it modeled the MIMO channel using both the eigenmodes of the transmit and receive correlation as well as the expected energy coupling between the transmit and receive modes. It is easy to see that the rows of $\mathbf{\Omega}$ roughly correspond to the associated sampling probabilities in (5.4). Indeed, as we have seen in (5.4) and (5.5) a rank 1 coupling matrix with an appropriately rotated codebook leads to *i.i.d* statistics for each users feedback. As the Weichselberger model generalizes both the Rayleigh model (by taking $\mathbf{\Omega}_{i,j} = 1$) as well as the Kronecker model (by taking a rank 1 coupling matrix $\mathbf{\Omega}$) in the sequel we seek a model for the user assignment distribution that can minimally be reduced to both of the *i.i.d* models in (5.4) and (5.5) while describing the more general covariance structure associated to the Weichselberger model as well. In particular, as the coupling matrix $\mathbf{\Omega}$ may have arbitrary rank and an arbitrary number of distinct rows one must select a discrete distribution which can model clustering effects for the users with similar propagation characteristics. We begin by consider how one may form a discrete model for a single-user then turn to the question of multiple clusters.

■ 5.1.1 Models for The User Assignment Distribution for a Single User

When developing a model of a process one must understand the figure of merits of interest as to not make unnecessary or unneeded assumptions that may influence the results. Our motivation for choosing the Weichselberger model for the fading process stemmed from our desire to model the capacity of the system so that one may analyze the achievable rates in the system, which in turn came from the problem of channel aware scheduling. Our

main interest in developing a model for the user assignment distribution is to enable us to understand the trade-off between the order statistic gain and the multi-node matching gain and the complexity of user scheduling as well as identifying when covariance of the fading process of a cluster of users warrants adapting the feedback scheme. We note that the interdependencies of users channels have been captured through the generalized switch and the BRS model. Thus, in our model of the user assignment distribution for a single-user it is sufficient for us to model only the relations between the a single-user's channel fading and the quantization codebook. In particular, a motivating factor behind our development of a generalized switching framework was that assuming a flat power allocation the set of rates which may be allocated at each scheduling interval is completely determined by the feedback received at each scheduling interval. Thus, we have a particular interest in frequency each user is assigned to a given input as this may be used to derive the full input occupancy distribution for the user pool which in turn directly relates to the output occupancy distribution through the structure of the switch yielding the require insights in to the system tradeoffs. In this direction we let

$$\mathbf{X}_{i,j} = \begin{cases} 1 & \text{if input } j \text{ is occupied by user } i \\ 0 & \text{otherwise} \end{cases}$$

and call the joint distribution of $\{\mathbf{X}_{i,j}\}$ the input occupancy distribution of user i . It is sufficient for one to only model the input occupancy distribution if one only has interest in the system tradeoffs and scheduling complexity. However, we also are interested in the design of a system that has the ability to adapt to a variety of environments as current MIMO system are deployed in a wide range of environments. Thus, in general one does not know or is not able to sufficiently model the distribution of $\mathbf{X}_{i,j}$ in advance for each deployment site. In order to make a strong inference on the input occupancy distribution one generally must keep a count of the number of times each codeword has been reported. That is, to infer the distribution of the input occupancy distribution one has additional interest in the joint distribution of the random variables

$$\mathbf{N}_{i,j}[k_1, n_k] = |\{\mathbf{h}_i[k] : \mathcal{Q}(\mathbf{h}_i[k]) = \mathbf{c}_j \text{ for } k \in [k_1 - n_k, k_1 - 1]\}|,$$

which we call the user assignment distribution of length n_k . We note that the user assignment distribution of length n_k records a history of a user's feedback over a window of length n_k and thus may be used to estimate the distribution of $\{\mathbf{N}_{i,j}[k_1, n_k]\}_{j=0}^{2^r-1}$ accurately given that n_k is sufficiently large and given that we have a sufficiently parametrized prior distribution for $\{\mathbf{N}_{i,j}[k_1, n_k]\}_{j=0}^{2^r-1}$. In the sequel we develop the relevant models for the user assignment distribution $\{\mathbf{N}_{i,j}[k_1, n_k]\}_{j=0}^{2^r-1}$ for the previously considered channel models. To motivate our chosen model for the user assignment distribution we provide a discussion which parallels our development of our channel model by providing a model for the user assignment distribution corresponding to the Rayleigh model, the Kronecker model and the Weichselberger model.

We begin with the simplest discrete model for $\{\mathbf{N}_{i,j}[k_1, n_k]\}_{j=0}^{2^r-1}$ which follows from the assumption of a user channel having a correlated Gaussian channel. We note this assumption is valid for the Rayleigh model, the Kronecker model or the Weichselberger model. That is, assume at present that one may prescribe some set of positive real numbers

$\{p_{i,0}, p_{i,1}, \dots, p_{i,2^r-1}\}$ such that $\sum_{j=0}^{2^r-1} p_{i,j} = 1$ and

$$\Pr[\mathcal{Q}(\mathbf{h}_i[k]) = \mathbf{c}_j] = p_{i,j}. \quad (5.6)$$

Further, assume that distribution of $\mathcal{Q}(\mathbf{h}_i[k])$ and $\mathcal{Q}(\mathbf{h}_i[k'])$ for $k' \neq k$ are independent and identically distributed (as described by (5.6)) from block to block. As each of the quantized channel vectors are independent the joint distribution of the random variables $\{\mathbf{N}_{i,j}[k_1, n_k]\}_{j=0}^{2^r-1}$ have a joint distribution equal to the multinomial distribution of index n_k with cell probabilities $(p_{i,0}, p_{i,1}, \dots, p_{i,2^r-1})$ [69]. More precisely,

$$\{\mathbf{N}_{i,j}[k_1, n_k]\}_{j=0}^{2^r-1} \sim \text{Multinomial}(p_{i,0}, p_{i,1}, \dots, p_{i,2^r-1}; n_k)$$

where

$$\Pr\left[\{\mathbf{N}_{i,j}[k_1, n_k] = n_j\}_{j=0}^{2^r-1}\right] = n! \prod_{i=0}^{2^r-1} \frac{p_i}{n_i!} \quad (5.7)$$

and $n = \sum_{i=0}^{2^r-1} n_i = n$ and $n_i \geq 0$. In practice one is not given the particular covariance of each user's channel a priori and hence to use such a model one must first fit the parameters of the Multinomial distribution to match, as closely as possible, the true distribution of a user's assignment distribution. This may be done by a simple training process at each site the system is deployed and more generally it may be done contiguously to estimate and track the correlation structure of each cluster which may vary due to user or scatterer dynamics.

In practice one must, for each site the system is deployed and subsequently each time the user and scatterer dynamics subsequently change the system state, estimate the cell probabilities $(p_{i,0}, p_{i,1}, \dots, p_{i,2^r-1})$. In the absence of any prior assumptions on the distribution of the cell probabilities the maximum likelihood (ML) estimate is simply the relative frequency of the codeword occurrence [69]. That is, the ML estimate of the probability any (every) user is quantized to the i -th codeword based on the observation of $\{\mathbf{N}_{i,j}[k_1, n_k]\}_{j=0}^{2^r-1}$ is, absent any prior for $p_{i,j}$,

$$\hat{p}_{i,j}^{(\text{NP})}[k_1] = \frac{\mathbf{N}_{i,j}[k_1, n_k]}{n_k}. \quad (5.8)$$

However, if users channels are not *i.i.d* over time (5.8) can be shown to be quite poor. Thus, if our assumption that $\mathcal{Q}(\mathbf{h}_i[k])$ is independent of $\mathcal{Q}(\mathbf{h}_i[k'])$ for $k' \neq k$ is too strong we must generalize the Multinomial distribution to account for these dependencies. However, the simplicity of the Multinomial distribution (5.7) and the related simple form of the ML estimate for its parameters make it desirable to find a simple augmentation to the ML estimate (5.8) to account for the short comings of the Multinomial model rather than generalize the Multinomial distribution itself. An efficient way to do this is to introduce a parametrized family of a prior distributions on the cell probabilities and chose the best prior for the data using some simple training data or observations of the process to bias the estimates of $(p_{i,0}, p_{i,1}, \dots, p_{i,2^r-1})$ to better match the data. That is, one may assume that the cell probabilities themselves are random with some underlying joint distribution and choose the parameters of the prior distribution of the cell probabilities in a way as to capture the dependencies of the process determining $\{\mathbf{N}_{i,j}[k_1, n_k]\}_{j=0}^{2^r-1}$.

The introduction of a prior distribution on the cell probabilities may at first seem a bit abstract. However, there is a large physical motivation behind this choice. Note that the introduction of prior distribution for the cell probabilities $(p_{i,0}, p_{i,1}, \dots, p_{i,2^r-1})$ reflects a

relevant and valid prior assumption on the values the cell probabilities should take. Assuming that the Voronoi diagrams of the quantizer are isomorphic the only source of variability in the cell probabilities is from the channel covariance. Thus, a choice for a prior distribution for the cell probabilities reflects a prior assumption on the channel covariance and hence from our development of the Weichselberger model a prior assumption on the coupling matrix $\mathbf{\Omega}$ and the physical propagation environment. Thus, adding a prior distribution on the cell probabilities with free parameters that must be inferred through observations of the user feedback not only enables a system designer to more fully model the user assignment distribution, but when used in practice allows the system to *learn* some coarse information of the propagation environment at a particular deployment site. Such a model incorporated into a MIMO system when paired with an inference engine to estimate the free parameters allows a system to learn and adapt to a wide range of deployment sites given that the class of priors chosen to provide this flexibility adequately captures the relevant physical aspects of the system.

To accurately model the user assignment distribution one must be sure to choose a prior that encapsulates the relevant aspects of the fading distribution. As we have selected the Weichselberger model as the relevant prior on the fading distribution we must be sure that we choose a prior on the cell probabilities that reflects this assumption. In the sequel we seek to find a valid prior for a single cluster of users, i.e. a set of users whose signal undergoes a similar propagation path and hence have similar spatial correlation. For such a propagation environment one minimally expects the cell probabilities associated to codevectors that represent the dominate scatters at the specific deployment site to be positively correlated. Thus, we seek to find a prior with such positively correlated cell probabilities.

As a first attempt to find a valid prior on the cell probabilities, we examine the Dirichlet distribution which is often chosen as a prior for the Multinomial distribution [27,38] as it is a conjugate prior to the Multinomial distribution [69]. More precisely, the $2^r - 1$ dimensional Dirichlet distribution with parameters $\boldsymbol{\theta} = (\theta_0, \theta_1, \dots, \theta_{2^r-1})$ has the density function, for $p_i \geq 0$ and $\sum_{i=0}^{2^r-1} p_i = 1$,

$$\text{Dirichlet}(p_0, p_1, \dots, p_{2^r-1}; \boldsymbol{\theta}) = \frac{\Gamma(\theta_{\text{sum}})}{\prod_{j=0}^{2^r-1} \Gamma(\theta_j)} \prod_{i=0}^{2^r-1} p_i^{\theta_i-1}$$

where $\theta_{\text{sum}} = \sum_{j=0}^{2^r-1} \theta_j$ and $\theta_i > 0$. With some simple computation it can be shown that [69]

$$\mathbb{E}[p_i] = \frac{\theta_i}{\theta_{\text{sum}}} \quad (5.9)$$

and

$$\text{Var}(p_i) = \frac{\theta_i \cdot (\theta_{\text{sum}} - \theta_i)}{\theta_{\text{sum}}^2 \cdot (\theta_{\text{sum}} + 1)} = \mathbb{E}[p_i] \frac{1 - \mathbb{E}[p_i]}{1 + \theta_{\text{sum}}}. \quad (5.10)$$

Moreover, the posterior distribution for $(p_{i,0}, p_{i,1}, \dots, p_{i,2^r-1})$ given $\mathbf{N}_i[k_1, n_k]$ is

$$f_{\mathbf{p}_i | \mathbf{N}_i}(\mathbf{p}_i | \mathbf{N}_i[k_1, n_k]; \boldsymbol{\theta}_i) = \text{Dirichlet}(p_{i,0}, p_{i,1}, \dots, p_{i,2^r-1}; \boldsymbol{\theta} + \mathbf{N}_i[k_1, n_k]) \quad (5.11)$$

as the Dirichlet distribution is a conjugate prior to the multinomial distribution. Combining (5.9) and (5.11) the corresponding Bayesian estimate of $p_{i,j}$ based on the Dirichlet

distribution as prior is a biased frequency count

$$\hat{p}_{i,j}^{(B)}[k_1] = \frac{\theta_{i,j} + \mathbf{N}_{i,j}[k_1, n_k]}{\theta_{i,\text{sum}} + \sum_{k=0}^{2^r-1} \mathbf{N}_{i,j}[k_1, n_k]}. \quad (5.12)$$

As (5.12) is quite similar to (5.8) it is natural to question how the additional degrees of freedom obtained by adding a Dirichlet prior distribution on the cell probabilities effects ones ability to model the user assignment distribution. Examining (5.9) and (5.10) one can see that all but one degree of freedom one has in the choice of θ_i is used to fix the expected values of the cell probabilities while the remaining single degree of freedom is used to uniformly scale the variance of the cell probabilities. In particular, one may think of the parameter θ_i as a hidden bias one may add to the frequency counts $\mathbf{N}_i[k_1, n_k]$ in order to temper the variability in the estimate given in (5.8). However it should be noted, that for a fixed θ_i , as $n \rightarrow \infty$ the Bayesian estimate of $p_{i,j}$ based on assigning a Dirichlet prior to the cell probabilities converges to (5.8) and hence for large sample sizes the effects of the Dirichlet prior are irrelevant. Moreover, the Dirichlet prior makes a far more restrictive assumption on the covariance of the cell probabilities than one at first realizes and may desire. In fact, it is simple to see that

$$\text{Cov}(p_{i,k}, p_{i,j}) = -\frac{\theta_{i,k} \cdot \theta_{i,j}}{\theta_{i,\text{sum}}^2 \cdot (\theta_{i,\text{sum}} + 1)} \quad (5.13)$$

and hence a Dirichlet prior assumes that the cell probabilities are negatively correlated (one may similarly show that $\text{Corr}(p_{i,k}, p_{i,j}) < 0$). As noted, in a MIMO system codewords which correspond to dominate scatterers should correspond to positively correlated cell probabilities. Thus, one must generalize this prior to remove this deficiency if one in general expects positive correlation as we do for multi-user MIMO.

In order to generalize the Dirichlet prior one must introduce additional degrees of freedom which when appropriately chosen yield the Dirichlet prior while offering significant enough freedom to model a more general covariance structure. This will allow one to better fit the user assignment distribution that arises from the assumption of the Weichselberger model as well as fits the more degenerate case of the Rayleigh model and Kronecker model. As such, we again seek a distribution which is a conjugate prior to the multinomial distribution. A simple way to do this was discussed by Connor and Mosimann in [38]. In particular, Connor and Mosimann noted that

$$S_{i,j} \triangleq \frac{p_{i,j}}{1 - \sum_{k=0}^{j-1} p_{i,k}}$$

for $j = 1, \dots, 2^r - 2$, $S_{i,0} = p_{i,0}$ and $S_{i,2^r-1} = 1$ are independent random variables that are marginally distributed as a univariate beta distribution. More precisely,

$$S_{i,j} \sim \text{Beta}(\theta_{i,j}^{(a)}, \theta_{i,j}^{(b)}) = \frac{\Gamma(\theta_{i,j}^{(a)} + \theta_{i,j}^{(b)})}{\Gamma(\theta_{i,j}^{(a)})\Gamma(\theta_{i,j}^{(b)})} z_{i,j}^{\theta_{i,j}^{(a)}} (1 - z_{i,j})^{\theta_{i,j}^{(b)}}$$

where $\theta_{i,j-1}^{(b)} = \theta_{i,j}^{(a)} + \theta_{i,j}^{(b)}$ for $j = 1, \dots, 2^r - 2$ and $\theta_{i,2^r-1}^{(b)} = \theta_{i,2^r-1}^{(a)}$ and further where $S_{i,j}$ is independent of $S_{i,k}$ for $k \neq j$. In order to develop a more general covariance structure Connor and Mosimann [38] suggested to allow the distribution of $S_{i,j}$ to follow a more

general univariate beta distribution where $\theta_{i,j}^{(a)} > 0$ and $\theta_{i,j}^{(b)} > 0$ and have no predetermined relationship. The resulting generalized prior is, by solving for $p_{i,j}$ in terms of the $S_{i,j}$ [38],

$$\begin{aligned} & \text{GDirichlet}(p_0, p_1, \dots, p_{2^r-1}; \boldsymbol{\theta}^{(a)}, \boldsymbol{\theta}^{(b)}) \\ &= \left(\prod_{k=0}^{2^r-2} \frac{\Gamma(\theta_i^{(a)} + \theta_i^{(b)})}{\Gamma(\theta_i^{(a)})\Gamma(\theta_i^{(b)})} \right) p_{2^r-1}^{\theta_{2^r-2}^{(b)}-1} \cdot \prod_{i=0}^{2^r-2} p_i^{\theta_i^{(a)}-1} \left(\sum_{j=i}^{2^r-1} p_j \right)^{\theta_{i-1}^{(b)} - (\theta_i^{(a)} + \theta_i^{(b)})}. \end{aligned}$$

As this distribution is, in large part, similar to the Dirichlet distribution one can show that it is again conjugate prior to the Multinomial distribution [38]. Thus, along similar lines to our previous development for the Dirichlet distribution, it can be shown that [38]

$$\mathbb{E}[p_{i,j}] = \frac{\theta_{i,j}^{(a)}}{\theta_{i,j}^{(a)} + \theta_{i,j}^{(b)}} \prod_{k=0}^{j-1} \frac{\theta_{i,k}^{(b)}}{\theta_{i,k}^{(a)} + \theta_{i,k}^{(b)}}$$

and

$$\text{Cov}(p_k, p_j) = \mathbb{E}[p_j] \left(\frac{\theta_k^{(a)}}{\theta_k^{(a)} + \theta_k^{(b)} + 1} \prod_{\ell=0}^{k-1} \frac{\theta_\ell^{(b)} + 1}{\theta_\ell^{(a)} + \theta_\ell^{(b)} + 1} - \mathbb{E}[p_k] \right). \quad (5.14)$$

It is important to note that from (5.14) the extra degrees of freedom incorporated into the prior allows us a much more general covariance structure for the cell probabilities. In particular, one now has the freedom to set the covariance of the cell probabilities to be positive. Moreover these new degrees of freedom have been incorporated while the resulting distribution remains conjugate prior to the multinomial distribution allowing efficient estimation of the cell probabilities and hence the channel covariance. As the GDirichlet distribution is a conjugate prior for the multinomial distribution the posterior distribution for the cell probabilities given $\mathbf{N}_i[k_1, n_k]$, is given by [27]

$$f_{\mathbf{p}_i | \mathbf{N}_i}(\mathbf{p}_i | \mathbf{N}_i[k_1, n_k]; \boldsymbol{\theta}_i^{(a)}, \boldsymbol{\theta}_i^{(b)}) = \text{GDirichlet}(p_{i,0}, p_{i,1}, \dots, p_{i,2^r-1}; \boldsymbol{\theta}_i^{(a)} + \mathbf{N}_i[k_1, n_k], \tilde{\boldsymbol{\theta}}_i^{(b)})$$

where

$$\tilde{\boldsymbol{\theta}}_{i,j}^{(b)} = \boldsymbol{\theta}_{i,j}^{(b)} + \boldsymbol{\vartheta}_{i,j}^{(b)}$$

and in turn where

$$\boldsymbol{\vartheta}_{i,j}^{(b)} = \sum_{k=j+1}^{2^r-1} \mathbf{N}_{i,k}[k_1, n_k].$$

It is again quite simple to find the Bayesian estimate for the cell probabilities based on the observation of $\mathbf{N}_i[k_1, n_k]$ as the GDirichlet distribution is conjugate prior to the multinomial distribution. In particular,

$$\hat{p}_{i,j}^{(G)} = \frac{\theta_{i,j}^{(a)} + \mathbf{N}_{i,j}[k_1, n_k]}{\theta_{i,j}^{(a)} + \theta_{i,j}^{(b)} + \boldsymbol{\vartheta}_{i,j}^{(b)} + \mathbf{N}_{i,j}[k_1, n_k]} \prod_{k=0}^{j-1} \frac{\theta_{i,k}^{(b)} + \mathbf{N}_{i,k}[k_1, n_k]}{\theta_{i,k}^{(a)} + \theta_{i,k}^{(b)} + \boldsymbol{\vartheta}_{i,k}^{(b)} + \mathbf{N}_{i,k}[k_1, n_k]} \quad (5.15)$$

$$= \frac{\theta_{i,j-1}^{(b)} + \mathbf{N}_{i,j-1}[k_1, n_k]}{\theta_{i,j-1}^{(a)} + \mathbf{N}_{i,j-1}[k_1, n_k]} \hat{p}_{i,j-1}^{(G)}. \quad (5.16)$$

Note that the additional degree of freedom of the GDirichlet distribution prior yields a

Bayesian estimate for the cell probabilities with much more structure. If again we interpret the parameters $\theta_{i,j}^{(a)}$ and $\theta_{i,j}^{(b)}$ as a statistical bias one may see that there are far more ways that ones may bias ones estimate of the cell probabilities through the choice of $\theta_{i,j}^{(a)}$ and $\theta_{i,j}^{(b)}$. In particular, as one may now model the cell probabilities with positive correlation, one may bias $\hat{p}_{i,j}^{(G)}$ based on the frequency counts of the occurrence of other codewords that are indicative of the dominate scatterers at a particular site, yielding a far more effective way to model the effects of the propagation environment on the feedback. We use the GDirichlet prior to model users in the sequel and hence make frequent use out the estimate (5.15). Thus, we let

$$\hat{\varrho}(\mathbf{n}, j; \boldsymbol{\theta}^{(a)}, \boldsymbol{\theta}^{(b)}) = \frac{\theta_j^{(a)} + \mathbf{n}_j}{\theta_j^{(a)} + \theta_j^{(b)} + \vartheta_j^{(b)} + \mathbf{n}_j} \prod_{k=0}^{j-1} \frac{\theta_k^{(b)} + \mathbf{n}_k}{\theta_k^{(a)} + \theta_k^{(b)} + \vartheta_k^{(b)} + \mathbf{n}_k} \quad (5.17)$$

be the Bayesian estimate for the cell probabilities based on the observation \mathbf{n} assuming a GDirichlet distribution as a prior on the cell probabilities. However, the GDirichlet prior is only sufficient to model users that have similar propagation environments. As current MIMO system aim to cover large geographic regions one should expect subsets of users to have very different propagation environments and hence need to be modeled by different GDirichlet priors. Hence, we now examine how one may infer the number of such clusters as well as the relevant parameters of the associated GDirichlet prior.

■ 5.1.2 The User Assignment Distribution for the Weichselberger model

In the preceding section we have argued that the GDirichlet distribution is the appropriate choice of a prior distribution for the cell probabilities of the multinomial distribution as it allows for accurate modeling of the propagation environment by incorporating the effects of dominate scatterers along different propagation paths, by ones choice of $\theta_i^{(a)}$ and $\theta_i^{(b)}$, while enabling efficient estimation of the cell probabilities. The cell probabilities may then in turn be used to estimate a slow, time varying covariance structure by using the estimate of the cell probabilities (5.15) to estimate the empirical covariance of a user channel. For our assumed model of the MIMO channel, the Weichselberger model, it is possible that users do not have a uniform channel correlation, but rather there may be many clusters of users. Indeed, if there are multiple distinct rows in the coupling matrix $\boldsymbol{\Omega}$ users may undergo dramatically different fading. This may be due, to among other effects, spatial separation of the user leading to the larger scale effects of the propagation environment amongst users to be very different. In particular, due to the large geographic regions current MIMO devices aim to serve the statistics of signals received by users may vary greatly as the may follow quite different propagation paths. Thus, in a multi-user MIMO system it is unlikely that a single GDirichlet distribution will be sufficient to act as a prior to accurately model the feedback from every user and the feedback process for a multi-user MIMO system should be assumed to be over-dispersed.

A frequent method used to model over-dispersion in data is to form a finite mixture of distributions [18, 89]. Before proceeding to describe the general model of interest and to further motivate our final choice we begin by describing a simple probabilistic model for multi-user MIMO systems with user clustering. Suppose, prior to a system deployment, one is able to accurately model and/or measure the characteristics of the propagation environment. Further, suppose that this model is able to identify n_c not necessarily contiguous

geographic regions in the area of coverage for which a number of dominate scatters lead the users in the region to have roughly similar signal propagation paths. For example, there may be users who are indoor or outdoor and near and far from the transmit base. For each of these regions suppose we are able to assign a GDirichlet with an appropriate prior to model the feedback from each of these regions. That is, suppose for each region $i = 0, \dots, n_c - 1$ we model the feedback from, say k , users in the region via a random variable \mathbf{C}_i where is a compound multinomial random variable

$$\mathbf{C}_i \sim \text{Multinomial}(\mathbf{p}_i; k) \bigwedge_{\mathbf{p}_i} \text{GDirichlet}(p_{i,0}, p_{i,1}, \dots, p_{i,2^r-1}; \boldsymbol{\theta}_i^{(a)}, \boldsymbol{\theta}_i^{(b)}), \quad (5.18)$$

That is,

$$\mathbf{C}_i | \mathbf{p}_i \sim \text{Multinomial}(\mathbf{p}_i; k)$$

and \mathbf{p}_i is marginally distributed as $\text{GDirichlet}(p_{i,0}, p_{i,1}, \dots, p_{i,2^r-1}; \boldsymbol{\theta}_i^{(a)}, \boldsymbol{\theta}_i^{(b)})$. Now suppose that the system is deployed and users enter the system randomly amongst the identified n_c geographic regions, each user selecting a region *i.i.d* with some predetermined set of probabilities $\{\pi_j\}$ for $j = 0, \dots, n_c - 1$. Further, suppose the location of each user is not revealed to the transmitter. Then the transmitter may model the state of users i with the “hidden” random variables

$$Z_{i,j} = \begin{cases} 1 & \text{if user } i \text{ in region } j \\ 0 & \text{otherwise} \end{cases}$$

for $i = 0, 1, \dots, n - 1$ and $j = 0, \dots, n_c - 1$. Using this model the transmitter may form an estimate of each of the $Z_{i,j}$ if one wishes to identify the spatial prior of each user. To do this, the transmitter may use the record of the feedback process for each user, $\mathbf{N}_i[k_1, n_k]$, and use this empirical data to infer to which of the n_c regions each user belongs. It is important to recall that $\mathbf{N}_i[k_1, n_k]$ is distributed as a multinomial random variable conditioned on the knowledge of the realization of the cell probabilities \mathbf{p}_i . However, absent knowledge of the realization of the cell probabilities one may only assume that $\mathbf{N}_i[k_1, n_k]$ follows the more general compound multinomial distribution. Given that our prior modeling is correct, user i can be grouped into one of the n_c classes and hence $\mathbf{N}_i[k_1, n_k]$ should be distributed similar to \mathbf{C}_j for some j . That is, one may alternately write the hidden random variables as

$$Z_{i,j} = \begin{cases} 1 & \text{if } \mathbf{N}_i[k_1, n_k] \sim \mathbf{C}_j \\ 0 & \text{otherwise} \end{cases}$$

for $i = 0, 1, \dots, n$ and $j = 0, \dots, n_c - 1$. Given a sufficiently long observation of $\mathbf{N}_i[k_1, n_k]$ it is a simple problem to determine a good estimate for each $Z_{i,j}$ given the distribution of \mathbf{C}_i . That is, given the site modeling is accurate one may form a maximum likelihood estimate for $Z_{i,j}$ by computing

$$j_i^* = \arg \max_{j=0, \dots, n_c-1} f_{\mathbf{C}_j}(\mathbf{N}_i[k_1, n_k]; \boldsymbol{\theta}_j^{(a)}, \boldsymbol{\theta}_j^{(b)})$$

and taking

$$\hat{Z}_{i,j} = \begin{cases} 1 & \text{if } j = j_i^* \\ 0 & \text{otherwise} \end{cases}$$

where $f_{\mathbf{C}_j}(\mathbf{n}; \boldsymbol{\theta}_j^{(a)}, \boldsymbol{\theta}_j^{(b)})$ is the density of the compound multinomial distribution [27]

$$f_{\mathbf{C}_j}(\mathbf{n}; \boldsymbol{\theta}_j^{(a)}, \boldsymbol{\theta}_j^{(b)}) = \frac{\Gamma\left(1 + \sum_{k=0}^{2^r-1} n_i\right)}{\prod_{k=0}^{2^r-1} \Gamma(n_i)} \prod_{k=0}^{2^r-1} \frac{\Gamma\left(\boldsymbol{\theta}_{j,k}^{(a)}\right) \Gamma\left(\boldsymbol{\theta}_{j,k}^{(b)}\right)}{\Gamma\left(\boldsymbol{\theta}_{j,k}^{(a)} + \boldsymbol{\theta}_{j,k}^{(b)}\right)} \\ \times \prod_{k=0}^{2^r-1} \frac{\Gamma\left(\boldsymbol{\theta}_{j,k}^{(a)} + n_k\right) \Gamma\left(\boldsymbol{\theta}_{j,k}^{(b)} + \sum_{\ell=k+1}^{2^r-1} n_\ell\right)}{\Gamma\left(\boldsymbol{\theta}_{j,k}^{(a)} + \boldsymbol{\theta}_{j,k}^{(b)} + \sum_{\ell=k}^{2^r-1} n_\ell\right)}.$$

Thus, with this estimate the transmitter may partition the user pool \mathcal{U} into n_c different classes which we denote by $\mathcal{U}^{(\ell)}$. That is,

$$\mathcal{U} = \prod_{\ell=0}^{n_c-1} \mathcal{U}^{(\ell)}$$

where $i \in \mathcal{U}^{(\ell)}$ if and only if $\hat{Z}_{i,\ell} = 1$. We note that given two users are in the same class does not imply that the two users assignment distribution follow the same multinomial distribution and hence does not imply that the two users have the same channel covariance. That is, if $i_1, i_2 \in \mathcal{U}^{(\ell)}$ then, in general,

$$\mathbf{N}_{i_1}[k_1, n_k] | \mathbf{p}_{i_1} \not\sim \mathbf{N}_{i_2}[k_1, n_k] | \mathbf{p}_{i_2}.$$

Thus, in general one may not assume that all users that have been assigned to a class follow a single multinomial distribution and one must independently model each users assignment distribution with a different multinomial distribution to accurately model the feedback.

It seems a bit unfortunate to have to model each users distribution with a distinct distribution. However, we note that a heterogeneous user population is beneficial to the multi-user scheduling problem. That is, if the statistics of each users channels are sufficiently different then it is less likely that users will be assigned to the same input in the generalized switch and hence increases the transmitters ability to find sets with a small level of co-channel interference. This is important to note as the practical motivation behind modeling the user assignment distribution stems for the desire to detect and correct an underlying channel correlation that is *detrimental to the system performance*. Thus, there is little need to model heterogeneity in each class of users so long as one can find a sufficient homogeneous model for each class that allows one to identify underlying channel correlation that is detrimental to the system performance. If a homogeneous model is sufficient this allows for a dramatic state reduction at the transmitter and simplifies the process of detecting channel correlation that is detrimental to the system performance. However, for a system of have this capability one must first have an accurate model of each deployment site of interest. As one may not have the time and/or resources to form such an accurate model of every deployment site it is of interest to develop a way to infer not only the values of the hidden random variables to classify users, but it is also of interest to develop a method to infer the parameters of the prior distribution of each class as well as the number of classes. While this seems like a tall task there are many ways in which one may solve this problem. We describe one possible solution in the sequel by employing an expectation-maximization EM algorithm after first describing our full model for the user assignment distribution assuming the underlying channel fading follows the Weichselberger model.

The model for user clustering described in the preceding discussion may be described more generally as a *finite mixture model* for the user assignment distribution [89]. In general, a finite mixture model for a random vector decomposes the density of a random vector, say \mathbf{m} , in to a weighted sum of a finite number of component densities. That is, a finite mixture model for \mathbf{m} with g component densities models the density of \mathbf{m} as

$$f(\mathbf{m}) = \sum_{i=0}^{g-1} \pi_i f_i(\mathbf{m}) \quad (5.19)$$

where $f(\mathbf{m})$ is the density of \mathbf{m} , $f_i(\mathbf{m})$ are the component densities of the mixture and $0 \leq \pi_i \leq 1$ are the weights which sum to 1. A simple way to generate a random variable distributed as (5.19), which parallels our preceding discussion, is by considering a categorical random variable Z which takes on values from a finite set of categories $\{0, 1, \dots, g-1\}$ with probability $\pi_0, \pi_1, \dots, \pi_{g-1}$ respectively. One may interpret Z as a random variable which labels the component density the random variable \mathbf{m} follows. More precisely, one may consider the joint distribution of \mathbf{m} and Z as

$$f(\mathbf{m}, Z) = \sum_{i=0}^{g-1} \mathbf{1}_{\{Z=i\}} \cdot f_i(\mathbf{m}).$$

Thus, assuming that the conditional density of \mathbf{m} given Z follows

$$f(\mathbf{m} | Z = i) = f_i(\mathbf{m}).$$

the total law of probability yields that the *unconditional* density of \mathbf{m} is simply the mixture (5.19). It is easy to see that this interpretation of the mixture model is exactly the scenario described for the multi-user MIMO channel. That is, in our preceding discussion each user was selected from one of n_c possible geographic regions and given the random variable $\{Z_{i,j}\}_{j=0}^{n_c-1}$ the assignment distribution of user i was conditional distributed a compound multinomial distribution. More precisely, in the preceding discussion one has $g = n_c$, $\mathbf{m} = \mathbf{N}_{i_2}[k_1, n_k]$, $Z = \sum_{j=0}^{n_c-1} j \cdot Z_{i,j}$ and each one of the component densities is simply the compound multinomial distribution. Hence, in the sequel we assume the user assignment distribution may be modeled as a generalized mixture of compound multinomial distributions. In this direction we let

$$\Theta = \left(\boldsymbol{\theta}_0^{(a)}, \boldsymbol{\theta}_1^{(a)}, \dots, \boldsymbol{\theta}_{n_c-1}^{(a)}, \boldsymbol{\theta}_0^{(b)}, \boldsymbol{\theta}_1^{(b)}, \dots, \boldsymbol{\theta}_{n_c-1}^{(b)} \right)$$

and

$$\boldsymbol{\pi} = (\pi_0, \pi_1, \dots, \pi_{n_c-1}).$$

Then, we model the assignment distribution of each user as

$$\Pr[\mathbf{N}_{i_2}[k_1, n_k] = \mathbf{n}; \Theta, \boldsymbol{\pi}] = \sum_{j=0}^{n_c-1} \pi_j \cdot f_{\mathbf{C}_j}(\mathbf{n}; \boldsymbol{\theta}_j^{(a)}, \boldsymbol{\theta}_j^{(b)}) \quad (5.20)$$

where $f_{\mathbf{C}_j}(\mathbf{n}; \boldsymbol{\theta}_j^{(a)}, \boldsymbol{\theta}_j^{(b)})$ are the compound multinomial component distributions and in turn where $0 < \pi_i < 1$ and satisfy $\sum_{j=0}^{n_c-1} \pi_j = 1$. We refer to (5.20) as a mixed multinomial generalized Dirichlet distribution (MMGDD). Using a finite mixture model for the multi-

user MIMO system one may, given the distribution of the component mixtures, find the ML estimate for the hidden random variables to identify the appropriate prior on each user's assignment distribution. Then one may use the identified prior to estimate the cell probabilities for the multinomial distribution modeling each users assignment distribution. However, as previously noted, it is of interest to model each class with a single multinomial distribution as this allows the system to more easily identify when there is an underlying channel correlation for a class that is detrimental to system performance. That is, if a class of users forms a cluster of users then one may need to adapt the feedback scheme for this cluster (class) of users. However, to have this capability one must first have an accurate estimate of the parameters that describe the fading distribution of each class.

To specify the MMGDD distribution one must specify the number of clusters of users n_c , the mixing proportions π_i as well as the parameters for each of the compound multinomial random variables \mathbf{C}_i . These parameters give system designers many degrees of freedom to model the feedback process and indirectly the coupling matrix $\mathbf{\Omega}$. In particular, π_i may be considered as the proportion of users who on average classified to belong to class i and similarly the parameters $\theta_i^{(a)}$ and $\theta_i^{(b)}$ roughly describe a bias to particular covariance matrices determined by the propagation environment of class i . Thus, so long as the user dynamics do not change rapidly one may use many realization of the feedback process $\mathbf{N}_{i,j}[k, n_1]$ to estimate π , $\theta_i^{(a)}$ and $\theta_i^{(b)}$ and use shorter histories of the feedback process $\mathbf{N}_{i,j}[k, n_2]$ in order to identify the spatial correlation of each user in each class via the Bayesian estimate (5.15). One many attempt to approximate these parameters using the aforementioned relationship to the geometry of the propagation environment, by direct measurements or other physical modeling techniques. However, it is important to note that all of these model parameters including the hidden variables can be determined through an expectation-maximization (EM) algorithm [27] described in the sequel.

■ 5.2 The EM Algorithm and Homogeneous Class Modeling

The EM algorithm is a general method of finding the maximum-likelihood estimate of the parameters of an underlying distribution from a given data set when the data is incomplete or has missing values. In the current context the transmitter must estimate the parameters of the GDirichlet prior for the cell probabilities given its observation of the user feedback process and absent knowledge on which class each user belongs to and more generally absent how many classes are needed to model the system. There are several methods one can use in conjunction with the EM algorithm to estimate the number of classes [27, 89] and in the sequel we assume that the number of classes is known and do not develop the joint estimation of the number of classes and the parameters of the mixture models. Ignoring the problem of estimating the hidden random variables $Z_{i,j}$, one may attempt to compute the ML estimate of Θ using the incomplete set of data $\{\mathbf{N}_{i,j}[k, n_1]\}_{i=0}^{n-1}$,

$$(\hat{\Theta}, \hat{\pi}) = \arg \max_{(\Theta, \pi)} L(\Theta, \pi, \{\mathbf{N}_{i,j}[k, n_1]\}_{i=0}^{n-1})$$

where $L(\Theta, \pi, \mathbf{N}_{i,j}[k, n_1]_{i=0}^{n-1})$ is the log likelihood function,

$$L(\Theta, \pi, \{\mathbf{N}_{i,j}[k, n_1]\}_{i=0}^{n-1}) = \log \left(\prod_{i=0}^{n-1} \Pr[\mathbf{N}_{i,j}[k, n_1]; \Theta, \pi] \right) \quad (5.21)$$

$$= \sum_{i=0}^{n-1} \log \left(\sum_{j=0}^{n_c-1} \pi_j \cdot f_{\mathbf{C}_j}(\mathbf{N}_{i,j}[k, n_1]; \theta_j^{(a)}, \theta_j^{(b)}) \right). \quad (5.22)$$

However, the direct ML solution is in general quite difficult to solve directly and as we have a more general interest in estimating the hidden random variables $Z_{i,j}$ one may consider the likelihood function of the *complete data set* for each user

$$\Psi_i = (\mathbf{N}_{i,j}[k, n_1], Z_{i,0}, Z_{i,1}, \dots, Z_{i,n_c-1})$$

which includes the hidden random variables. More precisely, the log likelihood of the complete data is [89]

$$\log L_c(\Theta, \pi, \{\Psi_i\}_{i=0}^{n-1}) = \sum_{j=0}^{n_c-1} \sum_{i=0}^n Z_{i,j} \cdot \left(\log(\pi_j) + \log \left(f_j(\mathbf{N}_{i,j}[k, n_1]; \theta_j^{(a)}, \theta_j^{(b)}) \right) \right). \quad (5.23)$$

In order to approximate a solution to the ML parameter estimation problem one may make a series of guesses at the values of the hidden random variables $Z_{i,j}$ and the use the likelihood function of the complete data in order to approximate the ML parameter estimate for Θ and π . To do this we first compute the joint distribution of the complete data. As we have shown in the case of the finite mixture model (5.19) one may derive the joint distribution of the complete data by assuming the hidden data labels the distribution which describes the random variable of interest. That is,

$$\Pr[\Psi_i = \psi; \Theta, \pi] = \sum_{j=0}^{n_c-1} Z_{i,j} \cdot f_j(\psi; \theta_j^{(a)}, \theta_j^{(b)}).$$

One may then use the total law of probability to find the marginal distribution of $\mathbf{N}_{i,j}[k, n_1]$. However, as the hidden data are indicator functions is easy to see that the expected distribution is exactly this marginal distribution for $\mathbf{N}_{i,j}[k, n_1]$, (5.20). That is,

$$\Pr[\mathbf{N}_{i,j}[k, n_1] = \mathbf{n}; \Theta, \pi] = \mathbb{E}_{Z_{i,j}} \left[\sum_{k=0}^{n_c-1} Z_{i,j} \cdot f_k(\mathbf{n}; \theta_k^{(a)}, \theta_k^{(b)}) \right]. \quad (5.24)$$

However, this approach has a shortcoming if one wishes to estimate the parameters of Θ and π as well as the hidden random variables $Z_{i,j}$ using the likelihood function of the *complete data set*. That is, if one is interested in the ML estimate of Θ and π given $\{\mathbf{N}_{i,j}[k, n_1]\}_{i=0}^{n-1}$ one may along the lines of (5.24) compute the likelihood function for the complete data, then, as the $Z_{i,j}$ are random and unobserved, consider the *expected* likelihood in order to approximate the ML estimate of Θ and π given $\{\mathbf{N}_{i,j}[k, n_1]\}_{i=0}^{n-1}$. Examining (5.23) one may see that if one, in an attempt to remove the randomness in (5.23), computes the expected

value with respect to $Z_{i,j}$ given $\mathbf{N}_{i,j}[k, n_1]$ then one must specify Θ and π . That is,

$$\mathbb{E}[Z_{i,j} | \mathbf{N}_{i,j}[k, n_1]; \Theta, \pi] = \Pr[Z_{i,j} = 1 | \mathbf{N}_{i,j}[k, n_1]; \Theta, \pi]$$

and from Bayes' law one has that the conditional distribution of $Z_{i,j}$ given $\mathbf{N}_{i,j}[k, n_1]$ is

$$\Pr[Z_{i,j} = 1 | \mathbf{N}_{i,j}[k, n_1]; \Theta, \pi] = \frac{\pi_j \cdot f_j(\mathbf{N}_{i,j}[k, n_1]; \theta_j^{(a)}, \theta_j^{(b)})}{\sum_{k=0}^{n_c-1} \pi_k \cdot f_k(\mathbf{N}_{i,j}[k, n_1]; \theta_k^{(a)}, \theta_k^{(b)})}$$

Thus, to compute the conditional expectation of $Z_{i,j}$ given $\mathbf{N}_{i,j}[k, n_1]$ one must have Θ and π at hand. In turn in order to compute an approximate ML estimate of Θ and π one needs the expected value of the log likelihood of the complete data which is a function of the conditional expectation of $Z_{i,j}$ given $\mathbf{N}_{i,j}[k, n_1]$. Thus, one seems to be in quite a precarious position. However, the EM algorithm exploits this circular structure to iteratively refine the estimates of both the hidden random variables as well as the estimates for Θ and π . In this direction, let

$$Q(\Theta, \pi, \hat{\Theta}[t], \hat{\pi}[t]) = \sum_{i=0}^{n-1} \sum_{j=0}^{n_c-1} \hat{Z}_{i,j}[t] \log \left(\Pr[\mathbf{N}_{i,j}[k, n_1]; \hat{\Theta}[t], \hat{\pi}[t]] \right)$$

be the conditional expectation of the complete data likelihood assuming $\hat{\Theta}[t]$ and $\hat{\pi}[t]$ as the current estimate of parameters Θ and π . The EM algorithm produces a sequence of estimates for the free parameters $\hat{\Theta}[t]$ and $\hat{\pi}[t]$ by alternating between two steps. The first step computes the expected value of the *complete-data* log-likelihood with respect to the hidden random variable $Z_{i,j}$ by way of computing the conditional expectation of $Z_{i,j}$. Then the second step computes an updated set of parameter estimates based on the expected value of the complete-data log-likelihood. That is, the EM algorithm alternates between the following two steps until the estimates converge:

1. **E-step:** Compute estimates of the hidden variables $Z_{i,j}$ as:

$$\hat{Z}_{i,j}[t] = \frac{\hat{\pi}_j[t] \cdot f_j(\mathbf{N}_{i,j}[k, n_1]; \hat{\theta}_j^{(a)}[t], \hat{\theta}_j^{(b)}[t])}{\sum_{k=0}^{n_c-1} \hat{\pi}_k[t] f_k(\mathbf{N}_{i,j}[k, n_1]; \hat{\theta}_k^{(a)}[t], \hat{\theta}_k^{(b)}[t])}$$

2. **M-step** Update the parameter estimates as

$$(\hat{\Theta}[t+1], \hat{\pi}[t+1]) = \arg \max_{(\Theta, \pi)} Q(\Theta, \pi, \hat{\Theta}[t], \hat{\pi}[t])$$

The EM algorithm may be used to find an approximation to the ML parameter estimate for $\hat{\Theta}$ and well as $\hat{\pi}$. However, as the EM algorithm is an iterative algorithm there may be issues with the rate of convergence and other numerical issues. In this thesis we do not consider these issues and rather assume that they may be adequately addressed (using deterministic annealing for example) so that the resulting estimates of Θ and π are accurate.

Using the EM algorithm one may accurately estimate the parameters of the GDirichlet prior distribution on the cell probabilities of each user. Thus, one may incorporate aspects of the propagation environment in ones estimate of the user assignment distribution by inferring the appropriate parameters of the GDirichlet through training and observations

of the users feedback. Most importantly, due to the structure of the GDirichlet and the fact that it is conjugate prior to the Multinomial distribution there are efficient methods to estimate the cell probabilities which in turn may be used to estimate a slow, time varying covariance structure of each user. More precisely, using the EM algorithm one may form an estimate of parameters of the prior distribution that reflect the dominate scatterers of each class using a long training sequence (or after a long observation of the feedback process). Then, one may use this prior to estimate the cell probabilities for each user based on shorter histories of feedback to form an estimate of each user's cell probabilities. More precisely, given the EM algorithm's estimate of $\hat{\Theta}$ one may estimate the cell probabilities of the multinomial distribution modeling the user assignment distribution from (5.15), using any appropriately chosen window of the feedback process, as

$$\hat{p}_{i,j}^{(G)} = \frac{\hat{\theta}_{i,j}^{(a)} + \mathbf{N}_{i,j}[k_1, n_k]}{\hat{\theta}_{i,j}^{(a)} + \hat{\theta}_{i,j}^{(b)} + \vartheta_{i,j}^{(b)} + \mathbf{N}_{i,j}[k_1, n_k]} \prod_{k=0}^{j-1} \frac{\hat{\theta}_{i,k}^{(b)} + \mathbf{N}_{i,k}[k_1, n_k]}{\hat{\theta}_{i,k}^{(a)} + \hat{\theta}_{i,k}^{(b)} + \vartheta_{i,k}^{(b)} + \mathbf{N}_{i,k}[k_1, n_k]} \quad (5.25)$$

where $\hat{\theta}_i^{(a)}$ and $\hat{\theta}_i^{(b)}$ are the parameter estimates for the prior distribution of the cell probabilities for the class for which user i belongs.

As previously noted a single-user being mismatched to a given feedback scheme is not the phenomenon that one wishes to model in the multi-user MIMO system. That is, our figure of merit and our ultimate question of interest is the effects of the input occupancy distribution has on the output occupancy distribution for the entire user pool as this describes the distribution in achievable rates. Thus, if a user in the system has a high degree of spatial correlation the broader system is not effected unless a subset of users in the system exhibit the same spatial correlation. Hence, we are interested rather in constructing a model for user feedback from each class of users that may identify when there is a large subset of users in the user pool which share the same spatial correlation. Note that the EM algorithm has already done much of required work in this direction. That is, the EM algorithm has classified the users the share a similar compound multinomial distribution and hence are more likely to have similar spatial correlation. However, the fact that a users feedback follows a similar compound multinomial distribution neither guarantees nor excludes the possibility that the realized multinomial distributions of each user are the same. That is, there is no guarantee that the feedback from each class is homogeneous.

To detect when a class of users is homogeneous one may compute the likelihood that this is the case. More precisely, let

$$\mathbf{N}^{(\ell)}[k_1, n_k] = \sum_{i \in \mathcal{U}^{(\ell)}} \mathbf{N}_{i,j}[k_1, n_k]$$

be the cumulative history of the feedback for a class of users. Then, the likelihood that this cumulative history for the class follows a single multinomial distribution is

$$\lambda_\ell = f_\ell(\mathbf{N}^{(\ell)}[k, n_1]; \boldsymbol{\theta}_\ell^{(a)}, \boldsymbol{\theta}_\ell^{(b)}).$$

In order to determine if a class of users is homogeneous one may check that λ_ℓ is greater than a prescribed threshold, say h_0 . That is, if $\lambda_\ell \geq h_0$ we say that the ℓ th class of users form a cluster and model the feedback from this class via a common multinomial distribution with cell probabilities

$$\hat{p}_j^{(\ell)} = \hat{q}(\mathbf{N}^{(\ell)}[k, n_1], j; \boldsymbol{\theta}_\ell^{(a)}, \boldsymbol{\theta}_\ell^{(b)}). \quad (5.26)$$

If the users channel are not homogeneous then there is not a single multinomial distribution that simultaneously models each users well. However, as previously noted users in a class that each exhibit significantly different fading is not of significant importance to model as the heterogeneity of the users will not degrade system performance significantly as it affords more selection diversity in the system. More precisely, users with significantly different fading are less likely to be assigned to the same input thus yielding more scheduling options for the transmitter. As our feedback model has been motivated by the problem of detecting when the system performance has been significantly degraded by the channel correlation in the sequel we model any heterogeneous class with a homogeneous model by averaging the cell probabilities for the class. That is, if $\lambda_\ell < h_0$ then we let

$$\hat{\mathbf{P}}_j^{(\ell)} = \frac{1}{|\mathcal{U}^{(\ell)}|} \sum_{i \in \mathcal{U}^{(\ell)}} \hat{p}(\mathbf{N}_i[k, n_1], j; \boldsymbol{\theta}_\ell^{(a)}, \boldsymbol{\theta}_\ell^{(b)}). \quad (5.27)$$

Using the methods of Chapter 4 one can determine exactly how the homogenized assignment distributions are effecting system performance. In particular, by computing the quantization order for the estimated distribution one can determine if the multi-user diversity is being reduced by the channel covariance and identify if the system would benefit from adapting the quantization scheme. An important feature of using the EM algorithm estimate the parameters of the prior distribution used to model the cell probabilities of the Multinomial distribution is it allows one to efficiently estimate the spatial covariance of the fading of each user. More precisely, recall in Section 2.2.1 we showed that given an estimate of the cell probabilities one may estimate the covariance of a user's channel through the empirical covariance. Thus, using (5.25) in conjunction with (2.17) one has

$$\hat{\mathbf{K}}_{\mathbf{h}_i} = \sum_{j=0}^{2^r-1} \hat{p}_{i,j}^{(G)} \mathbf{c}_j \mathbf{c}_j^\dagger. \quad (5.28)$$

Most importantly, using (5.28) one can estimate the principal eigenmode and eigenvalue for the spatial covariance of each user indicating when a user's covariance is "sufficiently mismatched" to the current feedback scheme to warrant adaptation as well as indicating the principal direction the new feedback scheme should be biased toward. We now turn to the problem of designing a framework for feedback design with an emphasis on how one may use the estimates of the covariance to intelligently adapt the feedback design.

■ 5.3 Robustness of the Systematic Construction for Multi-User Systems

It is now well understood that the MIMO channel is more often than not correlated which can have dramatic effects on system performance as the rates achieved by users in the system may sharply decline. Moreover, in a multi-user MIMO system the users in the system may have distinct channel correlation leading to a need to adapt to several distinct correlation matrices. It is well understood principle from vector quantization theory [52] that a quantizer should be designed to match the statistics of the channel that are relevant to the problem of interest as closely as possible [83, 105, 137]. In order to maximize throughput under the assumption of the Rayleigh model this means choosing codewords that are uniformly distributed over the sphere. However, if the underlying propagation environment is correlated, which we have shown in Section 5 is often the case, then the Rayleigh model is not an accurate model of channel. Indeed this was the underlying assumptions that led us

to consider the more general Weichselberger model for the MIMO channel in Section 5.0.5. If the channel correlation $K_{\mathbf{h}_i}$ is not approximately the identity, i.e. the Weichselberger model has not degenerated to the Rayleigh model, then there is reason to suspect that the uniform quantizer should not be close to optimal. We are interested in the expected rate and as such the first and second order statistics of the channel are of interest. We note that the ability for a quantizer to approximate the first order statistics can be addressed by adjusting the rate of a scalar quantizer designed for the Rayleigh model, i.e. if there is no prior on the channel means (or rather that the channel means are isotropic) one may consider this again as a vector quantization problem. However, if the channel correlation matrix $K_{\mathbf{h}_i}$ for each user is not the identity matrix then extra care must be taken to ensure that the (empirical) second order moment of the quantizer

$$K(\mathcal{C}) \triangleq \frac{1}{|\mathcal{C}|} \sum_{\mathbf{c} \in \mathcal{C}} \mathbf{c}_i \mathbf{c}_i^\dagger$$

approximately matches the correlation of the underlying channel, $K_{\mathbf{h}_i}$. More precisely, the relevant design principle is to design the quantizer such that

$$K(\mathcal{C}) \approx \frac{1}{m} K_{\mathbf{h}_i}$$

From a vector quantization perspective such a quantizer first whitens the source and then performs quantization on this whitened source. Thus, if $\mathcal{C} = \{\mathbf{c}_0, \mathbf{c}_1, \dots, \mathbf{c}_{|\mathcal{C}|}\}$ has been designed for the Rayleigh model then

$$A \cdot \mathcal{C} = \left\{ \frac{A\mathbf{c}_0}{\|A\mathbf{c}_0\|}, \frac{A\mathbf{c}_1}{\|A\mathbf{c}_1\|}, \dots, \frac{A\mathbf{c}_{|\mathcal{C}|}}{\|A\mathbf{c}_{|\mathcal{C}|}\|} \right\} \quad (5.29)$$

is well matched to a channel with correlation matrix $K_{\mathbf{h}_i} = AA^\dagger$ as [83, 137]

$$K(A \cdot \mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum_{\mathbf{c} \in \mathcal{C}} A\mathbf{c}_i \mathbf{c}_i^\dagger A^\dagger = A \left(\frac{1}{m} I \right) A^\dagger = \frac{1}{m} K_{\mathbf{h}_i}$$

Conversely, if a codebook $\tilde{\mathcal{C}}$ can be factored as

$$\tilde{\mathcal{C}} = \{\tilde{\mathbf{c}}_0, \tilde{\mathbf{c}}_1, \dots, \tilde{\mathbf{c}}_{|\mathcal{C}|}\} \quad (5.30)$$

$$= \left\{ \frac{A\mathbf{c}_0}{\|A\mathbf{c}_0\|}, \frac{A\mathbf{c}_1}{\|A\mathbf{c}_1\|}, \dots, \frac{A\mathbf{c}_{|\mathcal{C}|}}{\|A\mathbf{c}_{|\mathcal{C}|}\|} \right\} \quad (5.31)$$

where $\mathcal{C} = \{\mathbf{c}_0, \mathbf{c}_1, \dots, \mathbf{c}_{|\mathcal{C}|}\}$ has been designed for the Rayleigh model, i.e. $K(\mathcal{C}) = \frac{1}{m} I$, then we say that A is a factor of the code. Note if A is a factor of the code $\tilde{\mathcal{C}}$ and a matrix B is such that $AA^\dagger = BB^\dagger$ then it is not necessarily true that B is a factor of $\tilde{\mathcal{C}}$. In particular, if $B = A \cdot U$ then B is a factor of $\tilde{\mathcal{C}}$ if and only if

$$\tilde{\mathcal{C}} = \left\{ \frac{A \cdot U\mathbf{c}_0}{\|A \cdot U\mathbf{c}_0\|}, \frac{A \cdot U\mathbf{c}_1}{\|A \cdot U\mathbf{c}_1\|}, \dots, \frac{A \cdot U\mathbf{c}_{|\mathcal{C}|}}{\|A \cdot U\mathbf{c}_{|\mathcal{C}|}\|} \right\}.$$

Thus, B is a factor of $\tilde{\mathcal{C}}$ if U is a factor of \mathcal{C} , i.e. if $U \in \text{Sym}(\mathcal{C})$. We note that this distinction on its own is not necessarily helpful in developing an adaptive feedback framework. That

is, if one is only concerned with adapting the quantization codebook to match a single channel covariance, say $K_{\mathbf{h}}$, then so long as there is some factor of the codebook, say A , such that $AA^\dagger \approx K_{\mathbf{h}}$, then one would not expect the performance of the system to differ compared to another system which has AU as a factor for some unitary transformation U . However, providing robustness to a plurality of covariance is of great concern in a multi-user MIMO system as the users are expected to have heterogeneous fading. Thus, as we would like the overall system design to be robust to a large class of covariance structures as well as unbiased to any particular transmit direction it is natural to require that the system behaves the same for similar covariance matrices. That is, one would like a system that has been designed for a covariance matrix K to have the same performance as one that has been designed for $U^\dagger KU$, i.e. any system achieves approximately the same performance for any correlation matrices with the same singular values. Thus, if one is given a particular correlation spectrum of interest, say $\mathbf{\Lambda}$, one may consider forming a large “universal” codebook consisting of all codewords

$$\bigcup_{U \in \mathcal{W}} U^\dagger \mathbf{\Lambda} U \cdot \mathcal{C} \quad (5.32)$$

for some appropriately chosen set of unitary transforms \mathcal{W} . However, in order for the universal code of (5.32) to have a many of subcodes that are matched to a plurality of similar covariance matrices one in general must take \mathcal{W} to be quite large yielding a codebook which may overly encumber the scheduler by creating too large a search space to examine to find the maximally weighted clique. Thus, it is natural to consider if there is a more effective way to construct such a universal codebook.

In Section 5.1 we argued that in a multi-user MIMO system if each users channel vectors have negligibly correlated fading then the resulting system performance does not sufficiently deteriorate from the rates achieved assuming the Rayleigh model. More precisely, if the users channel vectors have distinct spatial correlation then the heterogeneity likely causes the number of occupied inputs in a generalized switch to increase yielding sufficiently diversity to exploit the multi-node matching gain. However, the system or individual users may see a substantial decrease in performance if the channel correlation of any user(s) is too great. In particular, if the dominate mode of a user’s channel covariance is dramatically larger than the other modes of the channel covariance then one expects that the user is likely to be assigned to a small set of inputs at each scheduling interval or unable to meet the required feedback thresholds. Geometrically, this corresponds to the direction of a user’s channel vector falling on to a small region of the complex unit m -sphere at each scheduling interval. Thus, in an attempt to adapt the feedback scheme to better match the channel covariance one may consider redistributing the points of the original codebook on this small region to reduce the mean squared quantization error as well as increase the diversity of channel vectors that the users feed back. We call this process *localization of the codewords*. Note this is exactly the same perspective we took in Section 3.6 to construct high rate quantizers and hence use the same operators developed there to enable covariance adaptation. In a system which localizes codewords the covariance matrices that are of interest are those with one dominant mode. Thus, we next examine how our systematic quantization framework may be used to adapt the feedback scheme to match users channel covariance which has been inferred through the EM algorithm and the history of past feedback.

■ 5.3.1 Covariance Structure of Local Codes

In order to develop efficient channel quantization methods for increase the quantization rate we developed geometric operations which had one dominate mode. However, this led to a linear transform that was not normal. Thus, we must take some care in developing the associated results for the covariance matrices that are matched with this scheme. Recall that if \mathbf{F} is a factor of a code $\tilde{\mathcal{C}}$ then $\tilde{\mathcal{C}}$ is matched with the channel covariance,

$$\mathbf{K}_{\mathbf{F}} = \mathbf{F}\mathbf{F}^\dagger$$

If \mathbf{F} were Hermitian the eigenvectors of \mathbf{F} are the eigenvectors $\mathbf{K}_{\mathbf{F}}$ and a similar assertion on the eigenvalues would follow. However, as previously noted, in general $\mathbf{F}(\mathbf{b}_0; \alpha, \gamma, \mathcal{B})$ is not normal and hence is neither Hermitian nor unitary. In order to understand to eigenstructure of the covariance matrices that are matched with these factors we must proceed cautiously as the developed eigenvalues and eigenvectors of $\mathbf{F}(\mathbf{b}_0; \alpha, \gamma, \mathcal{B})$ will not generally correspond to the eigenvalues and eigenvectors of $\mathbf{K}_{\mathbf{F}}$.

Recall that in order to find the eigenvalues and eigenvectors of \mathbf{F} we first examined the behavior of \mathbf{F} on a basis. This again proves useful and with some simple, yet tedious arithmetic, one can by applying (3.73) find that

$$\mathbf{K}_{\mathbf{F}}\mathbf{b}_l = \begin{cases} \gamma(1 + \alpha^2 \cdot (m-1))\mathbf{b}_0 + \alpha\sqrt{1 - \alpha^2} \sum_{\mathbf{b} \in \mathcal{B} \setminus \mathbf{b}_0} \mathbf{b} & \text{if } \mathbf{b}_l = \mathbf{b}_0 \\ (1 - \alpha^2)\mathbf{b}_l + \alpha\sqrt{1 - \alpha^2}\mathbf{b}_0 & \text{if } \mathbf{b}_l \neq \mathbf{b}_0 \end{cases} \quad (5.33a)$$

$$(5.33b)$$

As an immediate consequence of (5.33a) one can see that if an eigenvector of $\mathbf{K}_{\mathbf{F}}$ is correlated with \mathbf{b}_0 for $0 < \alpha < 1$ then every element of \mathcal{B} is correlated with this eigenvector. More precisely, consider an arbitrary vector in $\mathbf{v} \in \mathbb{C}^m$ where

$$\mathbf{v} = a_0\mathbf{b}_0 + \sum_{\mathbf{b}_i \in \mathcal{B} \setminus \mathbf{b}_0} a_i\mathbf{b}_i$$

for some a_0, a_1, \dots, a_{m-1} as \mathcal{B} is a basis for \mathbb{C}^m . Then by (5.33a),

$$\mathbf{K}_{\mathbf{F}}\mathbf{v} = \left(a_0 \cdot \gamma(1 + \alpha^2 \cdot (m-1)) + \alpha\sqrt{1 - \alpha^2} \sum_{i=1}^m a_i \right) \mathbf{b}_0 \quad (5.34a)$$

$$+ \sum_{\mathbf{b}_i \in \mathcal{B} \setminus \mathbf{b}_0} \left(a_0 \cdot \alpha\sqrt{1 - \alpha^2} + a_i \cdot (1 - \alpha^2) \right) \mathbf{b}_i \quad (5.34b)$$

Hence, if $\mathbf{v}^\dagger\mathbf{b}_0 \neq 0$ then $\mathbf{b}_i^\dagger\mathbf{K}_{\mathbf{F}}\mathbf{v} \neq 0$ if $a_0/a_i \neq -\alpha^{-1}\sqrt{1 - \alpha^2}$. As $\mathbf{K}_{\mathbf{F}}$ is non-singular there is a least one eigenvector of $\mathbf{K}_{\mathbf{F}}$ that is a linear combination of all of the basis vectors. However, as there is an $m - 1$ dimension eigenspace orthogonal to \mathbf{b}_0 one should expect there is at least an $m - 2$ dimensional invariant subspace of $\mathbf{K}_{\mathbf{F}}$ orthogonal to \mathbf{b}_0 . Examining the case where $a_0 = 0$ and $\sum_{i=1}^m a_i = 0$ then

$$\begin{aligned} \mathbf{K}_{\mathbf{F}}\mathbf{v} &= \sum_{\mathbf{b}_i \in \mathcal{B} \setminus \mathbf{b}_0} a_i \cdot (1 - \alpha^2)\mathbf{b}_i \\ &= (1 - \alpha^2)\mathbf{v}. \end{aligned}$$

That is, if \mathbf{v} is chosen such that $a_0 = 0$ and $\sum_{i=1}^m a_i = 0$ then \mathbf{v} is an eigenvector for \mathbf{K}_F with eigenvalue $1 - \alpha^2$. We note as there is a $m - 2$ dimensional subspace of \mathbb{C}^m with $a_0 = 0$ for which the a_i have zero sum, \mathbf{K}_F must have a $m - 2$ dimensional eigenspace associated with the eigenvalue $1 - \alpha^2$. Thus, in order to understand the eigenvalue decomposition for \mathbf{K}_F we must find an orthonormal basis for this space. More precisely, we must find a set of $m - 2$ orthogonal vectors each of which sums to zero. In this direction we let $\text{DFT}^*(m)$ be the $m \times (m - 1)$ submatrix of the DFT matrix for which the rows sum to zero. More precisely we let,

$$\text{DFT}^*(m) = \frac{1}{\sqrt{m}} \begin{bmatrix} 1 & e^{\sqrt{-1}\frac{2\pi}{m}1} & e^{\sqrt{-1}\frac{2\pi}{m}2} & \dots & e^{\sqrt{-1}\frac{2\pi}{m}(m-1)} \\ 1 & e^{\sqrt{-1}\frac{2\pi}{m}2} & e^{\sqrt{-1}\frac{2\pi}{m}4} & \dots & e^{\sqrt{-1}\frac{2\pi}{m}(m-1)2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & e^{\sqrt{-1}\frac{2\pi}{m}(m-1)1} & e^{\sqrt{-1}\frac{2\pi}{m}(m-1)2} & \dots & e^{\sqrt{-1}\frac{2\pi}{m}(m-1)(m-1)} \end{bmatrix}$$

Then, from the preceding discussion it is clear that

$$\tilde{\mathbf{B}}_0(\mathcal{B})\text{DFT}^*(m)^\dagger$$

is a basis for the eigenspace of \mathbf{K}_F associated with the eigenvalue $1 - \alpha^2$. Thus, we are left to find the eigenvectors for the subspace of \mathbb{C}^m that is complimentary to this eigenspace. In particular we are left to find the eigenvectors for the two dimensional subspace of \mathbb{C}^m for which $a_0 \neq 0$ and $a_i = a_j$ for all $i, j \neq 0$. That is, we must find the two values for ν such that

$$\nu \cdot \mathbf{b}_0 + \sum_{\mathbf{b} \in \mathcal{B} \setminus \mathbf{b}_0} \mathbf{b}$$

is an eigenvector of \mathbf{K}_F . In this direction, let

$$\begin{aligned} \nu_+(\alpha, \gamma) &= \frac{-1 + \alpha^2 + |\gamma|^2 (1 + \alpha^2 \cdot (m - 1))}{2\alpha\gamma\sqrt{1 - \alpha^2}} \\ &\quad + \sqrt{\frac{(-1 + |\gamma|^2 + \alpha^2 \cdot (1 + |\gamma|^2 \cdot (m - 1)))^2 + 4\alpha^2|\gamma|^2 \cdot (1 - \alpha^2)(m - 1)}{4\alpha^2|\gamma|^2 \cdot (1 - \alpha^2)}}, \\ \nu_-(\alpha, \gamma) &= \frac{-1 + \alpha^2 + |\gamma|^2 (1 + \alpha^2 \cdot (m - 1))}{2\alpha\gamma\sqrt{1 - \alpha^2}} \\ &\quad - \sqrt{\frac{(-1 + |\gamma|^2 + \alpha^2 \cdot (1 + |\gamma|^2 \cdot (m - 1)))^2 + 4\alpha^2|\gamma|^2 \cdot (1 - \alpha^2)(m - 1)}{4\alpha^2|\gamma|^2 \cdot (1 - \alpha^2)}} \\ &= \frac{m - 1}{\nu_+(\alpha, \gamma)} \end{aligned}$$

and

$$\sigma_{\pm}(\alpha, \gamma) = \nu_{\pm}(\alpha, \gamma) \cdot \alpha\gamma\sqrt{1 - \alpha^2} + 1 - \alpha^2.$$

Then we have the following theorem.

Theorem 5.3.1. *The eigenvalues of \mathbf{K}_F are $\sigma_+(\alpha, \gamma)$, $\sigma_-(\alpha, \gamma)$ and $(1 - \alpha^2)$ with multiplicity $m - 2$. Further, $\nu_+(\alpha, \gamma) \cdot \mathbf{b}_0 + \sum_{\mathbf{b} \in \mathcal{B} \setminus \mathbf{b}_0} \mathbf{b}$ and $\nu_-(\alpha, \gamma) \cdot \mathbf{b}_0 + \sum_{\mathbf{b} \in \mathcal{B} \setminus \mathbf{b}_0} \mathbf{b}$ are the eigenvectors associated with the eigenvalues $\sigma_+(\alpha, \gamma)$ and $\sigma_-(\alpha, \gamma)$ respectively and $\tilde{\mathbf{B}}_0(\mathcal{B})\text{DFT}^*(m - 1)^\dagger$ is an orthonormal basis for the $m - 2$ dimensional eigenspace as-*

sociated to the eigenvalue $1 - \alpha^2$.

Proof. This proof is a direct consequence of the preceding discussion and simple arithmetic by computing

$$\mathbf{K}_{\mathbf{F}} \left(\nu \cdot \mathbf{b}_0 + \sum_{\mathbf{b} \in \mathcal{B} \setminus \mathbf{b}_0} \mathbf{b} \right) = \nu \cdot \mathbf{b}_0 + \sum_{\mathbf{b} \in \mathcal{B} \setminus \mathbf{b}_0} \mathbf{b}$$

using (5.3.1) and solving the resulting quadratic. \blacksquare

As an immediate corollary to Theorem 5.3.1 we can deduce the eigenvalue decomposition of $\mathbf{K}_{\mathbf{F}}$. As was seen in the eigenvalue decomposition of the quantizer factor \mathbf{F} , the eigenvalue decomposition of $\mathbf{K}_{\mathbf{F}}$ has a conical covariance structure which is rotated by the basis used in the definition. In this direction, let

$$\mathbf{U}_{\mathbf{K}}(\alpha, \gamma)^\dagger = \begin{bmatrix} \frac{\nu_+(\alpha, \gamma)}{\sqrt{\nu_+(\alpha, \gamma)^2 + (m-1)}} & \frac{1}{\sqrt{\nu_+(\alpha, \gamma)^2 + (m-1)}} & \cdots & \frac{1}{\sqrt{\nu_+(\alpha, \gamma)^2 + (m-1)}} \\ \frac{\nu_-(\alpha, \gamma)}{\sqrt{\nu_-(\alpha, \gamma)^2 + (m-1)}} & \frac{1}{\sqrt{\nu_-(\alpha, \gamma)^2 + (m-1)}} & \cdots & \frac{1}{\sqrt{\nu_-(\alpha, \gamma)^2 + (m-1)}} \\ 0 & & \text{DFT}^*(m-1) & \\ \vdots & & & \\ 0 & & & \end{bmatrix}$$

and

$$\mathbf{\Sigma}_{\mathbf{K}}(\alpha, \gamma) = \begin{bmatrix} \sigma_+(\alpha, \gamma) & 0 & 0 & \cdots & 0 \\ 0 & \sigma_-(\alpha, \gamma) & 0 & \cdots & 0 \\ 0 & 0 & & & \\ 0 & 0 & & (1 - \alpha^2)\mathbf{I}_{m-2} & \\ \vdots & & & & \\ 0 & 0 & & & \end{bmatrix}$$

Then we have the following corollary to Theorem 5.3.1 regarding the structure of $\mathbf{K}_{\mathbf{F}}$.

Corollary 5.3.2. *Let \mathbf{c}_j be and arbitrary complex vector and let \mathcal{B} be a orthonormal basis for \mathbb{C}^m containing \mathbf{c}_j . Then, for any $\gamma \in \mathbb{C}$, $0 < \alpha < 1$, such that $\gamma \neq \sqrt{1 - \alpha^2}$,*

$$\mathbf{F}(\mathbf{c}_j; \alpha, \gamma, \mathcal{B})\mathbf{F}(\mathbf{c}_j; \alpha, \gamma, \mathcal{B})^\dagger = \mathbf{B}_j(\mathcal{B})\mathbf{U}_{\mathbf{K}}(\alpha, \gamma)\mathbf{\Sigma}_{\mathbf{K}}(\alpha, \gamma)\mathbf{U}_{\mathbf{K}}(\alpha, \gamma)^\dagger\mathbf{B}_j(\mathcal{B})^\dagger$$

While Corollary 5.3.2 is illuminating in terms of the structure of the covariance of the factor it is still unclear whether Corollary 5.3.2 is in fact the eigenvalue decomposition of $\mathbf{K}_{\mathbf{F}}$. In particular, it is unclear if the matrix $\mathbf{U}_{\mathbf{K}}(\alpha, \gamma)$ is unitary so that the diagonal elements of $\mathbf{\Sigma}_{\mathbf{K}}(\alpha, \gamma)$ are the eigenvalues of $\mathbf{K}_{\mathbf{F}}$. This is in fact so which we state in the following theorem.

Theorem 5.3.3. *Let \mathbf{c}_j be and arbitrary complex vector and let \mathcal{B} be a orthonormal basis for \mathbb{C}^m containing \mathbf{c}_j . Then, for any $\gamma \in \mathbb{C}$, $0 < \alpha < 1$, such that $\gamma \neq \sqrt{1 - \alpha^2}$. Then, the matrix $\mathbf{U}_{\mathbf{K}}(\alpha, \gamma)$ is unitary and the eigenvectors of $\mathbf{F}(\mathbf{c}_j; \alpha, \gamma, \mathcal{B})\mathbf{F}(\mathbf{c}_j; \alpha, \gamma, \mathcal{B})^\dagger$ are the columns of $\mathbf{B}_j(\mathcal{B})\mathbf{U}_{\mathbf{K}}(\alpha, \gamma)$ and the diagonal elements of $\mathbf{\Sigma}_{\mathbf{K}}(\alpha, \gamma)$ are the associated eigenvalues.*

Proof. We note that the rows of $\text{DFT}^*(m-1)$ have zero sum by definition and it is clear that first two columns of $\mathbf{U}_K(\alpha, \gamma)$ are orthogonal to the last $m-2$ columns as these vectors are constant over the last $m-1$ coordinates. As $\text{DFT}^*(m-1)$ is a sub matrix of the $m-1$ dimensional DFT matrix the last $m-2$ columns of $\mathbf{U}_K(\alpha, \gamma)$ are orthogonal. Thus, it is left to show that first two columns of $\mathbf{U}_K(\alpha, \gamma)$. To see this note that,

$$\nu_+(\alpha, \gamma)\nu_-(\alpha, \gamma) = 1 - m$$

and hence the first two columns of $\mathbf{U}_K(\alpha, \gamma)$ are orthogonal. \blacksquare

Examining Theorem 5.3.3 and the preceding discussion one can see that the universal code has an eigen space of dimension $m-2$ with eigenvalue and, in general, two one dimensional eigenspaces of dimension 1 with eigenvalues $\nu_+(\alpha, \gamma)$ and $\nu_-(\alpha, \gamma)$. It is clear from the definition that $\nu_+(\alpha, \gamma) \geq \nu_-(\alpha, \gamma)$ and hence the dominate mode of the covariance matrix is in the direction

$$\nu_+(\alpha, \gamma)\mathbf{b}_0 + \sum_{\mathbf{b} \in \mathcal{B} \setminus \mathbf{b}_0} \mathbf{b}.$$

Thus, given an estimate of a cluster covariance one may choose values for α and γ to construct a factor that is matched to a channel covariance that is estimated through the EM algorithm. However, this result has more practical relevance in a high rate system. That is, to construct a high rate code we used factors to double the code rate. More precisely, given a rate r code \mathcal{C}_r , we formed a rate $2 \cdot r$ code $\mathcal{C}_F(\alpha, \gamma, \mathcal{C}_r)$ by forming unions of local codes and optimizing over the choice of α and γ . Thus, if one use the universal code $\mathcal{C}_F(\alpha, \gamma, \mathcal{C}_r)$, then one will have a rate $2 \cdot r$ code that is matched to a white channel as well as 2^r rate r codes that are matched to covariance matrices that have

$$\nu_+(\alpha, \gamma)\mathbf{c}_i + \sum_{\mathbf{b} \in \mathcal{B}_i \setminus \mathbf{c}_i} \mathbf{b} \quad (5.35)$$

for each $\mathbf{c}_i \in \mathcal{C}_r$ as principal directions where \mathcal{B}_i is the basis used in the construction of the local code $\mathcal{C}(\mathbf{c}_i; \alpha, \gamma, \mathcal{C}_r)$. Moreover, as we have have shown in Section 4.4, only cluster's of users that have a channel correlation that is highly mismatched with code substantially degrades system performance. From this perspective, one can see that a user with a very highly correlated channel, i.e. user with a channel vector with a dominate principal direction, will achieve approximately the same quantization error as as a user with a white channel vector in a code with half the rate. Thus, the systematic construction allows one to not only double the code rate, but also provides a robustness to channel correlations that are detrimental to system performance in the process. However, while this construction will ensure that the quantization error is low and the multi-user diversity is exploited it does not guarantee in any way orthogonality. Indeed, if users have channel vectors that undergo a fading with a common spatial correlation then it is unlikely these users will ever be orthogonal and one will have to select users that have non-zero co-channel interference. Thus, in such cases one would expect to benefit from intelligent multiplexing.

■ 5.3.2 Efficient Multiplexing in the Universal Code

Our preceding discussion has indicated that in the multi-user MIMO downlink it is of interest to design codebooks that contain many orthogonal bases as such an approach helps mitigate interference as well as simplifies the problem of multiplexing. Hence, it is natural

to ask whether extending the root codebook by adding local code is a way to produce new orthogonal sets. In this direction, we will say that two local codes $\mathcal{C}(\mathbf{c}_i; \alpha, \gamma, \mathcal{B}_i)$ and $\mathcal{C}(\mathbf{c}_j; \alpha, \gamma, \mathcal{B}_j)$ are orthogonal if \mathbf{c}_i is orthogonal to \mathbf{c}_j . Then, we have the following import theorem concerning the orthogonality properties of the code $\mathcal{C}_{\mathbf{F}}(\alpha, \gamma, \{\mathcal{B}_i\})$.

Theorem 5.3.4. *Let \mathcal{C}_0 be given and let $\mathcal{C}_{\mathbf{F}}(\alpha, \gamma, \{\mathcal{B}_i\})$ be the universal code associated with \mathcal{C}_0 for some choice of α, γ and collection of bases $\{\mathcal{B}_i\}$. If, $\mathbf{c}_i, \mathbf{c}_j, \mathbf{c}_1, \mathbf{c}_2 \in \mathcal{C}_0$ are such that $\mathbf{c}_i^\dagger \mathbf{c}_j = 0$*

$$\mathbf{c}_1^\dagger \mathbf{c}_j = 0 \text{ and } \mathbf{c}_i^\dagger \mathbf{c}_2 = 0 \text{ and } \mathbf{c}_1^\dagger \mathbf{c}_2 = 0 \quad (5.36)$$

then

$$\mathbf{c}_1^\dagger \mathbf{F}(\mathbf{c}_i; \alpha, \gamma, \mathcal{B}_i)^\dagger \mathbf{F}(\mathbf{c}_j; \alpha, \gamma, \mathcal{B}_j) \mathbf{c}_2 = 0.$$

That is, if $\mathcal{C}(\mathbf{c}_i; \alpha, \gamma, \mathcal{B}_i)$ and $\mathcal{C}(\mathbf{c}_j; \alpha, \gamma, \mathcal{B}_j)$ are orthogonal local codes then $\mathbf{F}(\mathbf{c}_i; \alpha, \gamma, \mathcal{B}_i) \mathbf{c}_1$ and $\mathbf{F}(\mathbf{c}_j; \alpha, \gamma, \mathcal{B}_j) \mathbf{c}_2$ are orthogonal if (5.36) is satisfied.

Proof. This may be proved by direct application of Lemma 3.6.1. In particular, every codeword in a local code is of the form $a_0 \mathbf{c}_i + a_1 \mathbf{c}_1$. As (5.36) describes the 3 inner products arising from inner product of two vectors of this form, the resulting inner product is zero. ■

Closely examining Theorem 5.3.4, one can see that in general there is no guarantee that this method will produce new orthogonal bases and hence likely that one may benefit from more intelligent multiplexing methods. This, however, is not unexpected as the code factors were designed to combat channel correlation and not produce orthogonal sets that are full rank. However, we note that if one is not interested in full rank transmission (i.e. selecting sets of user of size m) or such a transmission is not possible/optimal due to the channel correlation or power constraints, then by examining Theorem 5.3.4 one can see that it is possible that the universal code introduces new orthogonal sets of small size that are not included in the root code. More importantly, from Theorem 5.3.4 it is possible that there are subsets of codevectors of the universal code that are orthonormal bases for subspaces of \mathbb{C}^m for which there is no orthonormal basis in the root code \mathcal{C}_0 . Moreover, the number of such sets in the universal code is governed by the orthogonality relations of the root code. Thus, while not introducing new bases, the universal code does introduce new orthogonal configurations of lower rank which span a subspace which is not spanned by any subset of vectors of the root code.

As our adaptive framework does not introduce new orthogonal bases in a multi-user MIMO system it may not be able to find a size m subset of users that have orthogonal quantized channel vectors. If this is the case one may attempt to find a smaller set of user that do have orthogonal quantized channel vectors. However, if a smaller orthogonal configuration can not be found, or one wishes to use sets of users for transmission, one may need to multiplexing a non-orthogonal configuration for the universal code.

In the sequel we consider how one in the present framework may efficiently multiplex non-orthogonal configurations from the universal code. In particular, we consider multiplexing configurations from the universal code for which

1. all vectors are elements of a single local code
2. all vectors are elements of distinct non-orthogonal local codes
3. all vectors are elements of distinct orthogonal local codes

We note each one of these cases correspond to different system regimes. The first, corresponds to a system in which all users in the system have highly correlated channel vectors and hence are highly correlated with a single root codeword. The second multiplexing regime corresponds to a system in which the users channels are largely independent, however the overall system performance is not dominated by the multi-user diversity gain and hence we can not find nearly orthogonal terminals from subsets of orthogonal local codes. The third multiplexing regime correspond to a system in which the system performance is dominated by the multi-user diversity gain so that the configuration chosen for transmission are nearly orthogonal and lay in orthogonal local codes.

To begin, we consider the case of multiplexing when all vectors are elements of a single local code. In such a case it is desirable to first remove this correlation, then multiplex the resulting configuration from the root code. We note that the product structure of the factor $\mathbf{F}(\mathbf{b}_0; \alpha, \gamma, \mathcal{B})$ makes this approach quite easy to achieve. In fact, from (3.71) it is easy to see that the inverse of $\mathbf{F}(\mathbf{b}_0; \alpha, \gamma, \mathcal{B})$ has a similar form to that in (3.71). That is,

$$\mathbf{F}(\mathbf{b}_0; \alpha, \gamma, \mathcal{B})^{-1} = \left(\prod_{\mathbf{b} \in \mathcal{B} \setminus \mathbf{b}_0} \tilde{\mathbf{Y}}^{-1}(\mathbf{b}_0, \mathbf{b}; \alpha) \right) \left(\mathbf{I} - \frac{(\gamma - 1)}{\gamma} \cdot \mathbf{b}_0 \mathbf{b}_0^\dagger \right) \quad (5.37)$$

where $\tilde{\mathbf{Y}}(\mathbf{b}_1, \mathbf{b}_2; \alpha)^{-1}$ was given in (3.70). Thus, if one wishes to multiplex a set of vectors which are all elements of a single local code one may first invert the factor $\mathbf{F}(\mathbf{b}_0; \alpha, \gamma, \mathcal{B})$ via (5.37), then apply an existing linear multiplexing strategy to the resulting configuration from the root code. More precisely, if

$$\hat{\Phi}_{\mathcal{A}} = \mathbf{C}^\dagger \cdot \mathbf{F}(\mathbf{b}_0; \alpha, \gamma, \mathcal{B})^\dagger$$

for some $\mathbf{b}_0 \in \mathcal{C}$, basis \mathcal{B} and set of codewords from \mathcal{C} , represented in matrix form as \mathbf{C} then

$$\hat{\Phi}_{\mathcal{A}} \cdot \left(\mathbf{F}(\mathbf{b}_0; \alpha, \gamma, \mathcal{B})^\dagger \right)^{-1} \mathbf{W}^{\text{IC}}(\mathbf{C}) = \mathbf{I}$$

Thus, the interference canceling multiplexer is not too much more complex in this regime then it was prior to adaptation as it only require the product of a few simple rotations.

When the codewords lay in distinct local codes it is clear that we do not wish to invert the local factors as we know a priori that these vectors are somewhat (depending on the choice of α and γ) dispersed as they lay in separate local codes. That is, as the codewords lay in different local codes it is more natural to consider a multiplexer which first performs a small perturbation to align all the codewords with their root, then apply an existing linear multiplexing strategy to the resulting configuration from the root code.

Recall from Corollary 3.76 precisely describes what we geometrically expect. That is, every codeword from the root code gets a gain in the direction of \mathbf{c}_0 and a uniform scaling in the space orthogonal to \mathbf{c}_0 . More precisely, from Corollary 3.76 we can see that the inner product of every vector of the local code $\mathcal{C}(\mathbf{F}(\mathbf{c}_0; \alpha, \gamma, \mathcal{B}), \mathbf{c}_0)$ with \mathbf{c}_0 can be written, by examining (3.76), as

$$\mathbf{c}_0^\dagger \mathbf{F}(\mathbf{c}_0; \alpha, \gamma, \mathcal{B}) \mathbf{c} = \gamma(1 - \alpha) \cdot \mathbf{c}_0^\dagger \mathbf{c} + \alpha\gamma \cdot \sum_{\mathbf{b} \in \mathcal{B}} \mathbf{b}^\dagger \mathbf{c}. \quad (5.38)$$

In the sequel we let

$$\varpi(\mathbf{c}, \mathbf{c}_0; \alpha, \gamma, \mathcal{B}) = \mathbf{c}_0^\dagger \mathbf{F} \mathbf{c} - \sqrt{1 - \alpha^2} \cdot \mathbf{c}_0^\dagger \mathbf{c}.$$

Thus, every element of a local code is of the form

$$\frac{\sqrt{1-\alpha^2}\mathbf{c} + \varpi(\mathbf{c}, \mathbf{c}_0; \alpha, \gamma, \mathcal{B})\mathbf{c}_0}{\|\mathbf{F}\mathbf{c}\|}$$

This is a particularly useful form as it allows for quite simple multiplexing of elements from arbitrary local codes. In the sequel we let $(\mathbf{c}_i, \mathbf{c}_j, \mathcal{B}_k)$ denote the element of the universal code

$$\mathbf{F}(\mathbf{c}_i; \alpha, \gamma, \mathcal{B}_k)\mathbf{c}_j.$$

Then, given a set of say m codewords from the universal code $\{(\mathbf{c}_{i_\ell}, \mathbf{c}_{j_\ell}, \mathcal{B}_{k_\ell})\}_{\ell=0}^{m-1}$ which are the quantized channel vectors for some set of users \mathcal{A} we have

$$\hat{\Phi}_{\mathcal{A}} = \mathbf{D}_1 \cdot \mathbf{C}_1 + \mathbf{D}_2 \cdot \mathbf{C}_2 \quad (5.39)$$

where

$$\mathbf{D}_1 = \text{diag} \left[\left\{ \frac{\varpi(\mathbf{c}_{i_\ell}, \mathbf{c}_{j_\ell}; \mathcal{B}_{k_\ell})}{\|\mathbf{F}(\mathbf{c}_{i_\ell}; \alpha, \gamma, \mathcal{B}_{k_\ell})\mathbf{c}_{j_\ell}\|} \right\} \right]$$

and

$$\mathbf{D}_2 = \text{diag} \left[\left\{ \frac{\sqrt{1-\alpha^2}}{\|\mathbf{F}(\mathbf{c}_{i_\ell}; \alpha, \gamma, \mathcal{B}_{k_\ell})\mathbf{c}_{j_\ell}\|} \right\} \right]$$

where in turn

$$\mathbf{C}_1 = \begin{bmatrix} - & \mathbf{c}_{i_0}^\dagger & - \\ \vdots & \vdots & \vdots \\ - & \mathbf{c}_{i_{m-1}}^\dagger & - \end{bmatrix} \quad \text{and} \quad \mathbf{C}_2 = \begin{bmatrix} - & \mathbf{c}_{j_0}^\dagger & - \\ \vdots & \vdots & \vdots \\ - & \mathbf{c}_{j_{m-1}}^\dagger & - \end{bmatrix} \quad (5.40)$$

It is simple to see that if either \mathbf{C}_1 is unitary or \mathbf{C}_2 is unitary then

$$\hat{\Phi}_{\mathcal{A}} = \mathbf{D}_j \left(\mathbf{D}_j^{-1} \mathbf{D}_i + \mathbf{C}_j \cdot \mathbf{C}_i^\dagger \right) \mathbf{C}_i.$$

where $i \in \{1, 2\}$ is the index of the unitary matrix and $j \in \{1, 2\} \setminus \{i\}$ is the remaining index. Hence, in the case either \mathbf{C}_1 is unitary or \mathbf{C}_2 is unitary then

$$\mathbf{W}^{\text{IC}}(\hat{\Phi}_{\mathcal{A}}) = \mathbf{C}_i^\dagger \left(\mathbf{D}_j^{-1} \mathbf{D}_i + \mathbf{C}_j \cdot \mathbf{C}_i^\dagger \right)^{-1} \mathbf{D}_j^{-1} \quad (5.41)$$

where the same convention with i and j is used. We note that (5.41) is far more efficient to compute in practice than it may first appear. In particular, let

$$\mathbf{\Pi} = \text{diag} \left[\left\{ \frac{\varpi(\mathbf{c}_{i_\ell}, \mathbf{c}_{j_\ell}; \mathcal{B}_{k_\ell})}{\sqrt{1-\alpha^2}} \right\} \right] \quad (5.42)$$

Then, (5.41) becomes

$$\mathbf{W}^{\text{IC}}(\hat{\Phi}_{\mathcal{A}}) = \mathbf{C}_i^\dagger \left(\mathbf{\Pi}^{2(j-\frac{3}{2})} + \mathbf{C}_j \cdot \mathbf{C}_i^\dagger \right)^{-1} \mathbf{D}_j^{-1}.$$

Thus, the inverse in (5.41) may be indexed by $(j, \mathbf{\Pi}, \mathbf{C}_j \cdot \mathbf{C}_i^\dagger)$. In the root codes developed in the sequel there will be very few distinct Gram matrices $\mathbf{C}_j \cdot \mathbf{C}_i^\dagger$. Hence, so long as the

distinct number of matrices $\mathbf{\Pi}$ may assume is not too large then (5.41) may be computed by table look-up. However, if this is not the case then inverting

$$\left(\mathbf{\Pi}^{2(j-\frac{3}{2})} + \mathbf{C}_j \cdot \mathbf{C}_i^\dagger\right)^{-1}$$

is not much harder than the MMSE beamformer presented in Section 2.3 which has been implemented in many wireless systems. We do not develop the particular set of matrices $\mathbf{\Pi}$ may assume.

As we are interested in developing root quantizers it is likely that a set selected from the universal code will have either \mathbf{C}_i or \mathbf{C}_j unitary. If this is not the case then one may always compute the standard for for the pseudo inverse,

$$\mathbf{W}^{\text{IC}}(\hat{\Phi}_{\mathcal{A}}) = \mathbf{W}^{\text{IC}}(\mathbf{D}_1\mathbf{C}_1 + \mathbf{D}_2\mathbf{C}_2)$$

We do not develop specific insights for this inversion.

In the preceding section we have shown our geometric quantizer factors have quite nice multiplexing properties due to the eigenstructure of the factors. Thus, in correlated channels one has efficient methods to precancel known interference. However, as we have shown, in a multi-user system it is of equal importance to be able to efficiently identify and select users that have low co-channel interference. However, in a complexity constrained system one may not have the time and/or resources to do optimal selection. Thus, it is of great importance to develop efficient scheduling algorithms that, with high probability, choose the optimal set.

Algorithms for Scheduling in Multi-User MIMO Systems

In the preceding Chapters we have developed an adaptive quantization scheme as well as a framework to develop quantizers with low mean squared quantization error and many orthogonal bases. In particular, in Chapter 3 we found a simple, geometrically motivated, linear transform that may be used to construct high rate quantizers as well as quantizers matched to many users channel covariance. We showed that this quantizer preserves much of the underlying structure to aid in search. Further, in Chapter 3.3 we developed the symmetry group of a quantizer and studied an abstract notion of complexity and flexibility of a basis. However, at present we have not addressed exactly how one may solve the maximal weighted clique problem for a chosen quantizer or more generally how one may address the broader question of how to schedule users from the universal codebook. In this chapter we consider exactly how this may be done.

Recall that our underlying motivation for examining the order statistic gain and multi-node matching gain trade off was that it allowed us the ability to realize the multi-user diversity gains inherent in a multi-user system. This was done by employing simple thresholds on each users individual SNR to limit the search to a smaller pool. We showed that so long as the SNR threshold was not set too aggressively one may ensure that a set of users can be found in this restricted pool that obtains a sum rate which is close to optimal with high probability. However, in a practical system finding the set of users from this reduce pool that achieve the maximum sum rate may still not be feasible due to complexity constraints. Thus, it is of interest to develop scheduling algorithms that choose a set of users who achieve a rate close to that of the optimal set with as few operations as possible. It has been recognized that further restricting ones search to sets of users for which there are guarantees on the channel norms and the magnitudes of pairwise inner products can provide close to optimal performance [111, 120–124, 131, 140–142]. Such an approach aims to find a set of users that are nearly orthogonal so that the penalty in rate incurred using a sub-optimal multiplexing scheme will be negligible for the selected set. In Chapter 4 we showed that this approach is optimal in the large user limit, but also that this may be done successfully even when the user population is a small multiple of the number of the number of antennas. This was done by examining when the order statistic gain decouples from the multi-node matching gain as this implies that greedily selecting the users with best SNR targets first then the subset of users are chosen with the best co-channel interference does not incur a penalty in throughput asymptotically. Depending on the SNR threshold one may still have too many subsets of users to consider for full search to be feasible. To combat this complexity one may employ a code book that contains many orthogonal bases. Then, using the feedback threshold σ one has the added benefit that some of the search complex-

ity may be offset through a decentralized self selection where by users only report back if there channel vectors are near one of a plurality of subspace described by the quantization scheme. If one uses this approach it is reasonable to suspect that more often than not there is a subset of users which have orthogonal channel vectors and once again a search using pairwise inner products can provide close to optimal performance.

In Section 4.2 we presented a model for channel aware scheduling which modeled the dependencies of users feedback through a general graph. A large motivation for this architecture, and the subsequent analysis and quantization design, is if one can develop large switches with many possible processing modes, modeled by cliques, one could use efficient existing algorithms on the plurality of switches to arrive at the optimum scheduling decision for a given channel and queue state. Here we develop the necessary tools for efficient user selection to find the optimal set using our model of channel aware scheduling through a generalized switch. While finding the optimal solution to the channel aware scheduling problem is theoretically simple, as one may simply enumerate all subsets of users and evaluate each one with respect to the quality-of-service (QOS) objective function, from a practical perspective such a search may not be possible as complete enumeration of all subsets grows exponentially with the size of the user pool. In particular, in a MIMO system with 4 transmit elements and 8 users there are 70 subsets of size 4 and 162 subsets of users in all. Thus, as the time available to make a scheduling decision in a communication system can be quite small, one must find efficient ways to search among the subsets of users to find the optimal or approximate solution to the channel aware scheduling problem which is not enumerative. Moreover, entirely greedy algorithms may arrive at a local optimum which has a much lower weighted rate compared to that of the global optimum. Hence, in practice, a natural choice is a hybrid of these two methods. That is, a greedy search which has some knowledge of the combinatorial structure of the problem that allows the search algorithm to backtrack or restart is of interest.

■ 6.1 Fast Maximal Clique Algorithms

For optimal scheduling in a multi-user MIMO system we have chosen cliques in a general graph to represent the inputs and outputs of a generalized switch. This model was chosen as it sufficiently captures the complex geometric structure required for channel aware scheduling with multiple-antennas. We showed the interdependencies between rate allocations that may be represented by a bi-partite graph are insufficient to represent the interdependencies required for channel aware scheduling with multiple-antennas. However, we have shown that a plurality of general undirected graphs are. As a general graph does not include a set of distinguished outputs one may model the dependencies arising from co-channel interference through the assignments of edges in a general graph and use a clique in a graph to model a possible processing mode. Recall, we let an edge in \mathcal{G} represent a permissible pairing of codewords. In this setting a set of codewords may be scheduled simultaneously if and only if there is an edge between each codeword in \mathcal{G} . To each vertex $i \in V$ we associate a weight w_i representing the reward one gets in the linear objective function representing the QOS constraint by including the user with feedback associated to vertex i . We further let the weight of a clique be the sum of the weights of the vertices in the clique. Thus, the solution to the scheduling problem when restricted to the rate allocations represented by \mathcal{G} is equivalent to finding a maximally weighted clique in \mathcal{G} .

It is well understood that finding a maximally weighted clique in a general graph is NP-complete and this problem is counted among Karp's 21 NP-complete problems [73].

That is the problem of finding the maximum clique is intractable and hard to approximate as listing all maximal cliques of a given graph may require exponential time as graphs may contain exponentially many maximal cliques. As we have developed our generalized switch to contain many orthogonal bases it is unclear if this approach has driven us in to a problem that requires exponential time to solve through enumerative algorithms. Unfortunately this is the case in general. While the graphs associated to our quantization scheme may have exponentially many maximal cliques the graph it self is quite structured which allows one to determine an approximation, and often the exact, maximally weighted clique rapidly.

Before proceeding we first recall some basic definitions from graph theory. We have an interest in finding the maximally weighted clique in a graph and the size of the largest clique in \mathcal{G} is of interest. We denote this quantity as

$$\omega(\mathcal{G}) = \max_{\mathcal{S} \text{ clique in } \mathcal{G}} |\mathcal{S}|$$

and say that $\omega(\mathcal{G})$ is the *clique number* of \mathcal{G} . A related figure of merit of a graph is it's (vertex) colorability. We say that a vertex coloring of a graph $\mathcal{G} = (V, E)$ is a labeling of the vertex set V with "colors" such that such that no two adjacent vertices share the same color. More formally, a graph is k -colorable if there exists a map, say f_C , from the vertex set V to the color set $\{0, 1, \dots, k-1\} \times V$ where $f_C(v_i) = (f_c(v_i), v_i)$ and $f_c(v_i) \neq f_c(v_j)$ if $(i, j) \in E$. The *chromatic number* of a graph \mathcal{G} is the smallest coloring which we denote as $\chi(\mathcal{G})$. It is clear that the chromatic number is always greater than the clique number as one needs at least $\omega(\mathcal{G})$ many colors to color the maximal clique. As the chromatic number is an upper bound on the clique number, it is natural to suspect that the chromatic number plays a large role in methods to bound the size of the largest clique. In particular, if the clique number of every induced subgraph of a graph equals the chromatic number of the induced subgraph we say the graph is perfect.

The general problem of finding a maximum weighted clique in a graph is NP-complete. However, there is a large classes of graphs for which the maximum weighted clique may be solved exactly in polynomial time. A well known class of graphs for which the maximum weighted clique may be solved exactly in polynomial time is, not surprisingly, the class of *perfect* graphs. In the sequel we develop some of this theory surrounding perfect graphs in order to motivate a heuristic approach to efficiently solving the channel aware scheduling problem. In particular, *when the clique number equals that chromatic number there are very efficient algorithms to solve the maximum weighted clique problem* [17, 20, 53].

Grötschel, Lovász and Schrijver have shown that if a graph is perfect then the maximum weighted clique problem may be solved in polynomial time [53]. We note that perfect graphs require a very special structure that may not be met in general. However, if the clique number equals the chromatic number one may expect that there are efficient algorithms to solve the maximum clique problem which more often then not only require polynomial time to find the maximal clique due to the similarity with perfect graphs. That is, if a graph is not perfect then one may attempt to find alternate solutions to the maximal clique problem using either an approximation or exact algorithm using the insights one has from a perfect graph.

In general one may formulate the maximal clique problem as an integer programming problem¹. In particular, for a given graph with k vertices, $\mathcal{G} = (V, E)$, and a given set

¹This, should not be surprising as our formulation of the channel aware scheduling problem is an integer programming problem.

of vertex weights $\{w_i\}_{i \in E}$ one may consider the integer linear program based on the edge constraints [97]

$$\begin{aligned} & \underset{\mathbf{x}}{\text{maximize}} && \sum_{i=0}^{n-1} w_i \cdot x_i \\ & \text{subject to} && x_i + x_j \leq 1 \quad \forall (i, j) \in E && (6.1a) \\ & && x_i \in \{0, 1\}, \quad i = 0, 1, \dots, k-1 && (6.1b) \end{aligned}$$

With this formulation one may write down a simple linear relation of (6.1) by replacing the 0, 1 constraint in (6.1b) with a positivity constraint $x_i \geq 0$ for $i = 0, 1, \dots, k-1$. More precisely, one may consider the relaxed program

$$\begin{aligned} & \underset{\mathbf{x}}{\text{maximize}} && \sum_{i=0}^{n-1} w_i \cdot x_i \\ & \text{subject to} && x_i + x_j \leq 1 \quad \forall (i, j) \in E && (6.2a) \\ & && x_i \geq 0, \quad i = 0, 1, \dots, k-1 && (6.2b) \end{aligned}$$

The relaxation (6.2), in most cases, results in few variables having the true optimum values leading to a large gap between the optimal values of (6.1) and the solution to the relaxed problem [97]. In fact, it has been shown that the inequalities (6.2a) and (6.2b) are only sufficient to solve (6.1) if the graph \mathcal{G} is bipartite. Thus, to find more exact solutions one must consider a way to better formulate the problem so that a linear relaxation is successful. To do this one may first find a collection of subsets of vertices for which a constraint stronger than (6.1a) may be written so that the relaxation to the resulting program does not deviate from the optimal solution to the integer program. In this direction we say a set of vertices \mathcal{V} is an independent (stable) set if no two vertices of \mathcal{V} are adjacent. More precisely, \mathcal{V} is an independent set if

$$(i, j) \notin E \quad \forall i, j \in \mathcal{V}.$$

It is clear that a clique does not contain a pair of vertices from an independent set. Thus, suppose that an oracle has given us a list of every maximal clique in a graph, say S . Then, one may alternatively write (6.1) via the independent set formulation [97]

$$\begin{aligned} & \underset{\mathbf{x}}{\text{maximize}} && \sum_{i=0}^{n-1} w_i \cdot x_i \\ & \text{subject to} && \sum_{i \in \mathcal{V}} x_i \leq 1 \quad \forall \mathcal{V} \in S && (6.3a) \\ & && x_i \in \{0, 1\}, \quad i = 0, 1, \dots, k-1 && (6.3b) \end{aligned}$$

The integer program (6.3) clearly reflects the hardness of solving this problem. That is, the list of every maximal clique in a graph S may be exponential leading to an exponential set of constraints in (6.3a). One may again consider a relaxation of (6.3a) by changing (6.3b)

to a weaker positivity constraint. More precisely, one may consider the relaxed program

$$\begin{aligned} & \underset{\mathbf{x}}{\text{maximize}} && \sum_{i=0}^{n-1} w_i \cdot x_i \\ & \text{subject to} && \sum_{i \in \mathcal{V}} x_i \leq 1 \quad \forall \mathcal{V} \in S && (6.4a) \\ & && x_i \geq 0, \quad i = 0, 1, \dots, k-1 && (6.4b) \end{aligned}$$

However, the relaxation (6.4) is, once again, only exact on a specific class of graphs. This, is one of the main results of [53, 54]

Proposition 6.1.1. *A graph \mathcal{G} is perfect if and only if the solution to (6.4) has an integral solution for any set of weights $\{w_i\} \in \mathbb{R}^n$. Moreover, if \mathcal{G} is perfect then (6.3) can be solved in polynomial time.*

As we noted previously, perfect graphs have very efficient methods to solve the maximally weighted clique problem. However, in general the static set of switches developed in the BRS model for our architecture will not be perfect, especially in the case of the universal code. Thus, one must develop enumerative methods if one wishes to solve the maximally weighted clique problem, and hence the channel aware scheduling problem, exactly.

The need to enumerate a large set of cliques in a graph is at the core of difficulty of solving the maximally weighted clique problem. Indeed, as we have seen in (6.3) one may have to enumerate an exponential number of subgraphs, as in (6.3a), to solve the problem in general. While it may take exponential time to definitively *solve* (6.3) the optimal solution may, in some cases, be found much faster by excluding large subset of cliques. That is, to find the maximally weighted clique one may “intelligently” enumerate the cliques of a graph, by not exploring portions of the graph that can be shown to not include that maximally weighted clique. The most well known and common approach to this is the use of a branch and bound algorithm that finds good lower and upper bounds on portions of the graph and breaks the solution of the exact problem into smaller subproblems [97].

There has been considerable historical development of solutions to the max clique problem. We do not overview all of these results here but rather refer the reader to [97]. In the sequel, we develop the most efficient exact algorithms for solving the maximum-weighted clique problem using branch and bound algorithms. We note that the weighted and unweighted cases do not differ greatly and hence in the sequel only develop the unweighted algorithm leaving the extension for the weighted case for the final algorithm.

To date the most efficient clique finding algorithms are extensions of the branch and bound algorithm of Carraghan and Pardalos [31]. An important feature of the algorithm of Carraghan and Pardalos, and a key to its efficiency, is it requires one to specify an order of the vertices of the graph and considers searching for the maximal clique by enumerating the cliques containing a given vertex with respect to this order. This is useful if one can ascertain some properties of the graph to assist in how to find cliques rapidly. For any sequence of vertices of a graph \mathcal{G} with k vertices let τ be a given permutation of $\{0, 1, 2, \dots, k-1\}$. Then, the algorithm of Carraghan and Pardalos produces a sequence of cliques by finding the largest clique in \mathcal{G} which contains $v_{\tau(0)}$, then the largest clique in $\mathcal{G} \setminus \{v_{\tau(0)}\}$ containing $v_{\tau(1)}$ and so on. The crucial observation of Carraghan and Pardalos, which leads to the efficiency of their algorithm, was that one may stop the i th iteration early if the current largest clique found for \mathcal{G} is bigger than one that may be formed on $\mathcal{G} \setminus \{v_{\tau(0)}, v_{\tau(1)}, \dots, v_{\tau(i-1)}\}$.

In particular, Carraghan and Pardalos noted that one may apply this observation to not only the number of iterations but also to the search amongst the cliques containing a given vertex. An important notion in this development is the *distance* between two vertices in a graph. The distance between two vertices in a graph is the number of edges in a shortest path connecting them. Carraghan and Pardalos’s algorithm recursively searches for cliques by searching for cliques which only includes vertices up to distance d . That is, at depth d on the i th iteration Carraghan and Pardalos’s algorithm has found a clique of size d from the vertices $V \setminus \{v_{\tau(0)}, v_{\tau(1)}, \dots, v_{\tau(i-1)}\}$. Hence, if the size of the maximal clique is larger than $d + (k - i)$ one need not recurse further. It should be clear that the vertex order is key to this approach. In particular, if one lists vertices that are only contained in small cliques first one may have to enumerate every clique in the graph. Alternatively, if one provides a vertex which lists the maximal clique first the algorithm will halt far sooner.

Branch and bound algorithms are highly sensitive to the order specified for the vertices. To see this we examine the graph associated to the orthogonal processing modes of the quantizer from Example 3.3.3 which we illustrate in Figure 6-1. The quantizer from Example 3.3.3 may be described by the disjoint union of 4 bases, say $\mathcal{B}_0, \mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_3$. To see how sensitive the algorithm is to the vertex order suppose that one considers a vertex order which takes elements from each basis in order assuming this will locate one of the “most flexible” solutions sooner. More precisely, suppose

$$v_{\tau(i)} \in \mathcal{B}_{\lfloor i/4 \rfloor}. \quad (6.5)$$

Then, the algorithm of Carraghan and Pardalos will enumerate one size 4 clique almost immediately. Then this algorithm will proceed to enumerate many of the size 3 cliques and almost all of the size 2 cliques. This is due to the fact that the algorithm does not use any of the past search history to infer that no larger cliques exist larger than size 4. Clearly this particular ordering is not the optimal ordering as one spends a good bit of each iteration of the algorithm searching on a small region of the graph. An example of the first four iterations of this algorithm may be seen in Figure 6-1. With a vertex ordering (6.5) the algorithm of Carraghan and Pardalos first examines the cliques seen in Figure 6-1 (b). Once the search over this first basis completes the elements of this basis are deleted and the search continues over elements of another basis. However, as seen in Figure 6-1 (c) after four iterations of the search none of the bases which intersect \mathcal{B}_3 have had any vertex pruned.

In order to explore more regions of a graph early in the search the vertex order (6.5) is a poor choice. Indeed, the algorithm spends most of its time searching a local part of the graph. However, if one properly chooses a vertex sequence which takes an element of each basis in turn will do much better. That is, by appropriately choosing

$$v_{\tau(i)} \in \mathcal{B}_{i \pmod{4}}. \quad (6.6)$$

may allow one to exclude more large cliques earlier in the algorithm, leading to earlier termination. In particular, one may first find a coloring of the graph of interest and then take each color class in order as the elements of a color class form an independent set and hence should disperse the search through the graph. A coloring of the graph from Example 3.3.3 may be seen in Figure 6-2 (a). The result of pruning the depicted color class may be seen in Figure 6-2 (b). Note that with this ordering every clique of size 4 has been removed and the resulting graph is much more sparse. However, we note that the results from Figure

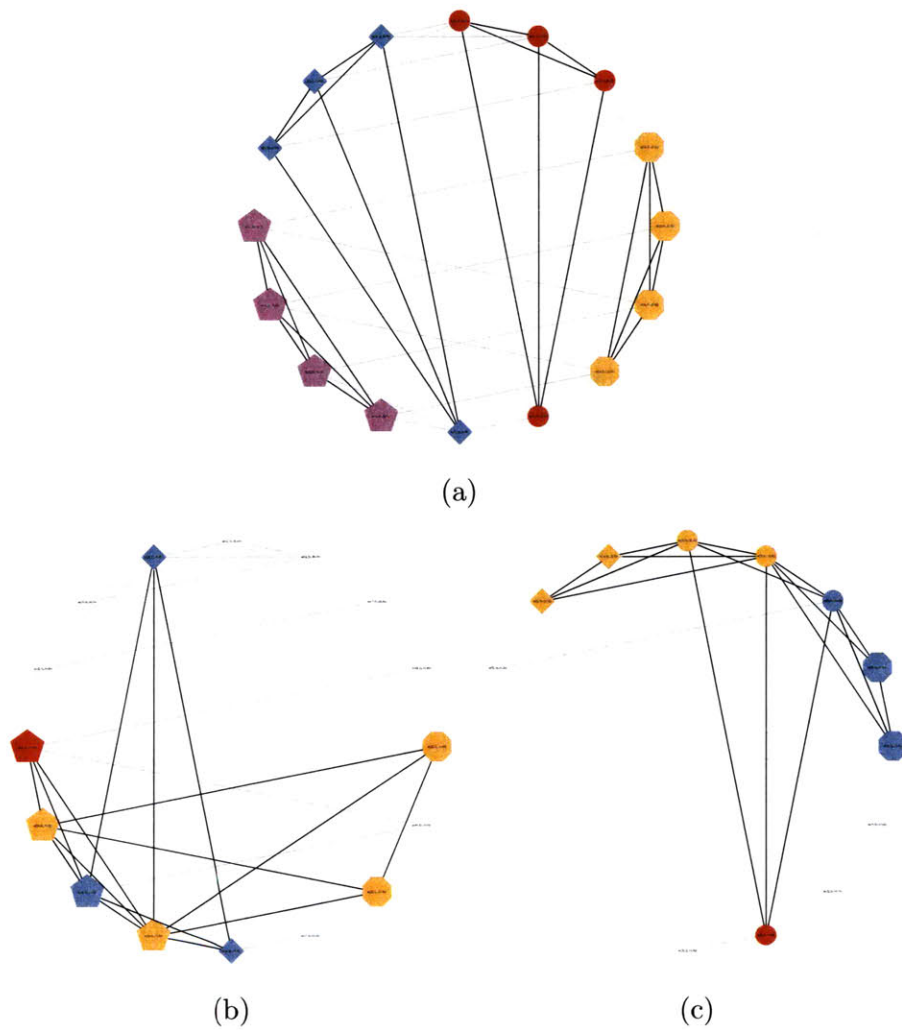


Figure 6-1. An illustration of the importance of the input vertex order for the algorithm of Carraghan and Pardalos. We assume that the bases are ordered from left to right as depicted in (a). That is, at far left is \mathcal{B}_0 , then \mathcal{B}_1 is middle left and so on. (b) the maximal cliques found in the first iteration (c) the cliques unaffected by the pruning of the first 4 iterations.

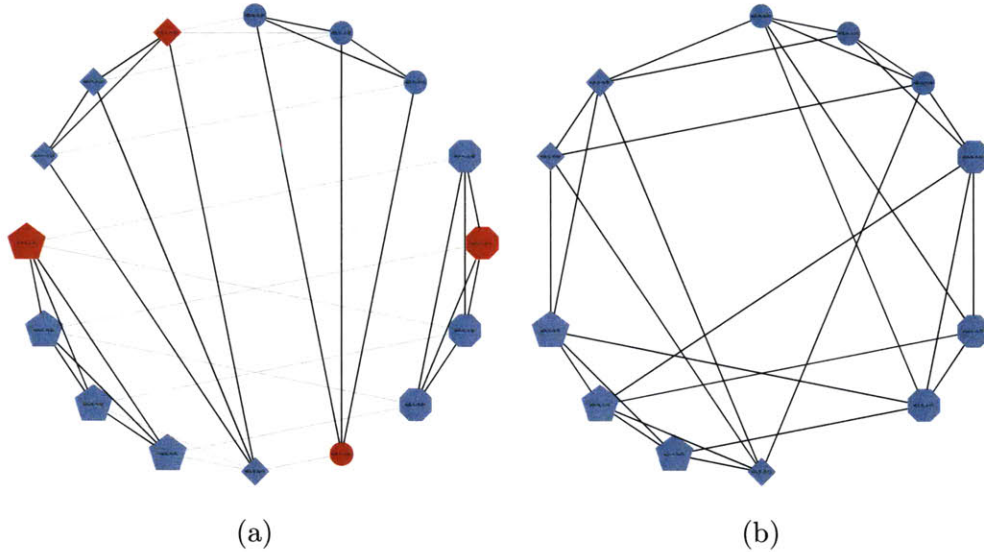


Figure 6-2. An illustration of the results of a vertex ordering which excludes every maximally sized clique after 4 iterations. The ordering is taken from a coloring of the graph \mathcal{G} . (a) A depiction of the first 4 vertices used in to exclude every clique of size 4 after 4 iterations of the algorithm of Carraghan and Pardalos. (b) The maximal cliques effected by removing the elements of the first color class.

6-2 are not true in general. That is, given a coloring a graph, even if it is minimal, removing a color class does not guarantee that size of the maximal clique decreases. Indeed, if the graph is not perfect then there is some sequence of color classes that when deleted does not reduce the size of the maximal clique in the graph. Examining the ordering used in Figure 6-2 one may see that there exists a 4 coloring of the graph. As this graph is 4 colorable and the largest clique is of size 4 the graph has chromatic number 4. As the chromatic number is an upper bound on the clique number one may, in this special case, guarantee a reduction in the cardinality of the largest clique by removing this color class. This has practical relevance in our system as we know the size of the maximal clique in the graph which represents orthogonal processing modes. This graph will always have clique number at most m due to the underlying geometry of the problem. Thus, if one can find an m -coloring of the graph it is easy to determine the existence of a clique of size m by exploiting the structure of the color classes. As we show in the sequel one may find an m -coloring of the graph for the quantizers of interest and hence (6.6) is a natural ordering on the set of vertices. However, as in general our graphs are not perfect, nor will there always be a maximal clique in the graph, there is no guarantee that removing a color class necessarily decreases the size of the maximal clique in the resulting graph if this clique is not of size 4. Hence, in general the algorithm of Carraghan and Pardalos would continue to search amongst the graph until a single independent set of size 4 remains, the last color used in the ordering. It is natural to ask if there is a modification to this algorithm that will detect the absence of a clique larger than size 4.

In order to determine a different heuristic to use find maximal cliques which halts before the algorithm of Carraghan and Pardalos one may consider running this algorithm backwards in an attempt to bound the size of cliques that would have been discovered and/or pruned by that algorithm had it been run forward. This approach has the added benefit that it starts from the smallest possible graph, a single vertex, and builds up to the full

graph, likely stopping before ever reaching this full graph and allows one to *record history of the size of cliques in a subgraph*. To begin in this direction, consider a graph \mathcal{G} with k vertices let τ be a given permutation of $\{0, 1, 2, \dots, k-1\}$. Then, Östergård [95, 96] has proposed keeping a table, say $c(i)$, which keeps track of the size of the largest clique in the subgraph $\mathcal{G} \setminus \{v_{\tau(0)}, v_{\tau(1)}, \dots, v_{\tau(i-1)}\}$. With this approach, $c(i) = c(i+1) + 1$ if and only if there is a clique of size $c(i+1) + 1$ containing v_{i-1} . We note that this approach enables a new pruning strategy based on the prediction of the results one would have had running the algorithm forward. In particular, one may again recurse to find cliques of maximal size up to a given distance, as was done in the algorithm of Carraghan and Pardalos. However, in Östergård's algorithm one may now use the history $c(i)$ to prune the search. In fact, it is easy to see that if vertex $v_{\tau(i)}$ is at a distance d one need not progress if $d + c(i)$ is less than the largest clique found. As the algorithm of Östergård is similar to the algorithm of Carraghan and Pardalos, with the ability to better prune the search, one should similarly expect using a coloring to order the vertices would again be fruitful. However, examining the graph from Example 3.3.3 in Figure 6-1 one can see that in graphs for which the chromatic index equals the clique number the algorithm of Östergård must wait until the addition of the last color class before beginning able to find a maximally sized clique. Thus, we must find a way to better exploit the structure of our graph if one hopes to efficiently solve the maximum weighted clique problem as it relates to channel aware scheduling. To do this we return the structure of the orthogonal bases that we found in Section 3.3.1.

■ 6.2 Complexity of Systematic Quantization Framework

Recall in Chapter 3.3 we identified the form that any orthogonal basis must take in our architecture through our definition of the twisted hamming weight. Further, in Section 3.4 showed that more generally one can define quantizers that have fewer orthogonality relationships by defining a restricted twisted hamming weight. In the sequel we identify how one may search for orthogonal sets in a code designed using Corollary 3.3.12 using the insights from Theorem 3.3.10. This view point will be particularly useful in the development of flexible search algorithms that can adapt to enable the system designer to meet quality of service constraints while simultaneously opportunistically use the maximum rate afforded by the time varying channel. In this direction we note that every codeword in the system of interest, say, $\mathbf{c}(\boldsymbol{\lambda}_j, \boldsymbol{\beta}; L, p^a)$ can be described by the vector $(\hat{\boldsymbol{\lambda}}_j, \bar{\boldsymbol{\lambda}}_j, \boldsymbol{\beta}) \in \hat{\Upsilon}_1 \times (p^{a-1} \cdot L_a^d) \times L^c$. In the sequel, we will ignore the parameter $\boldsymbol{\beta}$ as the effects this parameter has on orthogonality is trivial by condition (i) of Theorem 3.3.10. Thus, in the sequel we will instead study how one's choice of Υ_1 effects the resulting orthogonality properties. It should be clear from Theorem 3.3.10 one may determine if the set of vectors

$$\{\mathbf{c}(\boldsymbol{\lambda}_i, \boldsymbol{\beta}; L, p^a)\}_{i=0}^{\ell} \quad (6.7)$$

is self orthogonal by examining set of pairs of vectors

$$\{(\hat{\boldsymbol{\lambda}}_i, \bar{\boldsymbol{\lambda}}_i)\}_{i=0}^{\ell} \quad (6.8)$$

where $\lambda_i = \hat{\lambda}_i + \bar{\lambda}_i$. Thus, for every such set we may associate two arrays

$$\hat{\Lambda} \left(\{\hat{\lambda}_i\}_{i=0}^{\ell} \right) = \begin{bmatrix} \hat{\lambda}_{0,0} & \hat{\lambda}_{0,1} & \cdots & \hat{\lambda}_{0,m'-2} & \hat{\lambda}_{0,m'-1} \\ \hat{\lambda}_{1,0} & \hat{\lambda}_{1,1} & \cdots & \hat{\lambda}_{1,m'-2} & \hat{\lambda}_{1,m'-1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \hat{\lambda}_{\ell,0} & \hat{\lambda}_{\ell,1} & \cdots & \hat{\lambda}_{\ell,m'-2} & \hat{\lambda}_{\ell,m'-1} \end{bmatrix}$$

and

$$\bar{\Lambda} \left(\{\hat{\lambda}_i\}_{i=0}^{\ell} \right) = \begin{bmatrix} \bar{\lambda}_{0,0} & \bar{\lambda}_{0,1} & \cdots & \bar{\lambda}_{0,m'-2} & \bar{\lambda}_{0,m'-1} \\ \bar{\lambda}_{1,0} & \bar{\lambda}_{1,1} & \cdots & \bar{\lambda}_{1,m'-2} & \bar{\lambda}_{1,m'-1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \bar{\lambda}_{\ell,0} & \bar{\lambda}_{\ell,1} & \cdots & \bar{\lambda}_{\ell,m'-2} & \bar{\lambda}_{\ell,m'-1} \end{bmatrix}$$

From Theorem 3.3.10 it is clear that if a set of vectors is self orthogonal it is necessary that we can find a column for which every pair of rows has a common element in $\hat{\Lambda}$ element while the corresponding entries in $\bar{\Lambda}$ differ. It should be clear that in general one must examine each pair of entries in $\hat{\Lambda}$ and $\bar{\Lambda}$ in order to check if the associated set of vectors is self orthogonal. In particular, a self orthogonal set may have an associated $\hat{\Lambda}$ with $\lceil \ell/p \rceil$ distinct elements in each column or a self orthogonal set may have an associated $\bar{\Lambda}$ for which each column only has one distinct entry. These two cases play an important role in the discussion. As such, we denote the self orthogonal set which has an associated $\hat{\Lambda}$ with $\lceil \ell/p \rceil$ distinct elements in each column as \mathcal{O}_c the set which only has one distinct row as \mathcal{O}_f . More concretely, if $\hat{\Lambda}$ is such that $\hat{\lambda}_{i,j} = \hat{\lambda}_{i+1,j}$ then for an appropriate choice of $\bar{\Lambda}$ the pair $\hat{\Lambda}$ and $\bar{\Lambda}$ correspond to an orthogonal set. Moreover, any orthogonal set in which each pair of codewords satisfy condition (ii) in Theorem 3.3.10 have a $\hat{\Lambda}$ for which there is only one distinct element in each column. Thus, \mathcal{O}_f is self orthogonal set that meets condition (ii) in Theorem 3.3.10 while \mathcal{O}_c is a set with distinct rows meeting condition (iii) in Theorem 3.3.10. Thus, the frequency and number of distinct elements in the columns of $\hat{\Lambda}$ which define a self orthogonal set can be quite diverse.

Note that the preceding examples all define orthogonal sets for appropriate choices of $\bar{\Lambda}$ while having quite dissimilar structure. Thus, it is natural to wonder how these sets differ. In the sequel we will show that while both of the sets \mathcal{O}_f and \mathcal{O}_c are self orthogonal, the set \mathcal{O}_f is more flexible (by a measure we define in the sequel) to modification to a different orthogonal set than \mathcal{O}_c . In order to make this concept more precise we must first identify a notion of a type for $\hat{\Lambda}$. For any vector $\mathbf{v} \in (\mathbb{Z}_p)^{m'}$ we will let the type of the vector \mathbf{v} be the partition of the coordinates of \mathbf{v} for which \mathbf{v} has a constant value and denote this as $\text{type}_{\text{tw}}(\mathbf{v})$. That is, the type of $\mathbf{v} \in (\mathbb{Z}_p)^{m'}$, is the partition of $\{0, 1, \dots, m'-2, m'-1\}$, say $\text{type}_{\text{tw}}(\mathbf{v}) = \{\mathcal{P}_0, \mathcal{P}_1, \dots, \mathcal{P}_r\}$, such that

$$\{0, 1, \dots, m'-2, m'-1\} = \prod_{i=0}^r \mathcal{P}_i$$

and

$$\mathbf{v}_i = \mathbf{v}_j \quad \forall i, j \in \mathcal{P}_k \quad \text{and} \quad k = 0, 1, \dots, r.$$

For any matrix $\hat{\Lambda} \in (\mathbb{Z}_p)^{\ell \times m'}$ we let the type of the matrix be the vector of column types

and denote this as $\text{type}_{\text{twt}}(\widehat{\Lambda})$. That is,

$$\text{type}_{\text{twt}}(\widehat{\Lambda}) = \left[\text{type}_{\text{twt}}(\widehat{\Lambda}[:, 0]), \text{type}_{\text{twt}}(\widehat{\Lambda}[:, 1]), \dots, \text{type}_{\text{twt}}(\widehat{\Lambda}[:, m' - 1]) \right].$$

It should be clear that the $\text{type}_{\text{twt}}(\widehat{\Lambda})$ encapsulates the combinatorial structure of $\widehat{\Lambda}$ needed to test for orthogonality. More precisely, for $\widehat{\Lambda}$ to correspond to an orthogonal set it is necessary for there exists a subset of columns, say \mathcal{J}' , such that

$$\forall i \neq j \in \{0, 1, \dots, \ell\}, \quad \{i, j\} \subset \mathcal{P}_k \quad (6.9a)$$

$$\text{where } \mathcal{P}_k \in \text{type}_{\text{twt}}(\widehat{\Lambda}[:, c]) \quad (6.9b)$$

$$\text{and } c \in \mathcal{J}'. \quad (6.9c)$$

That is in general to check if the matrix $\widehat{\Lambda}$ corresponds to an orthogonal set we must at least check that the union of the column types contains all pairs of row indices. However, we note that the constraint imposed on the relationship between the $\bar{\lambda}_i$ make this far from sufficient. In particular, as it is required for $\bar{\lambda}_i$ to differ pairwise in the coordinates which satisfy (6.9) and the $\bar{\lambda}_i$ are isomorphic to vectors over $(\mathbb{Z}_p)^{m'}$ it is clear that the partitions with parts bigger than p lead to an overly opportunistic constraint. Hence, to identify orthogonal sets we need to check for the existence of a subset of columns, say \mathcal{J} , such that

$$\forall i \neq j \in \{0, 1, \dots, \ell\}, \quad \{i, j\} \subset \mathcal{N}_{k,c} \subset \mathcal{P}_k \quad (6.10a)$$

$$\text{where } \mathcal{P}_k \in \text{type}_{\text{twt}}(\widehat{\Lambda}[:, c]) \quad (6.10b)$$

$$\text{and } |\mathcal{N}_{k,c}| < p \quad (6.10c)$$

$$\text{and } c \in \mathcal{J}. \quad (6.10d)$$

In order to test for orthogonality one would like to dispense with the complexity of the search over rows as much as possible. That is, we would like to identify the types for which verifying (6.10) is as trivial as possible. In this direction we let $\text{comp}_{\text{twt}}(\widehat{\Lambda})$ denote the smallest number of columns which need to be examined to verify (6.10). More precisely, if

$$\mathcal{J}(\widehat{\Lambda}) = \{ \mathcal{J} : (6.10) \text{ is true for } \widehat{\Lambda} \}$$

then

$$\text{comp}_{\text{twt}}(\widehat{\Lambda}) = \min_{\mathcal{J} \in \mathcal{J}(\widehat{\Lambda})} |\mathcal{J}|. \quad (6.11)$$

The quantity $\text{comp}_{\text{twt}}(\widehat{\Lambda})$ is a very coarse measure of the difficulty one has testing if $\widehat{\Lambda}$ corresponds to an orthogonal set. In particular, if some oracle has given us \mathcal{J} then one would only have to examine the submatrix $\widehat{\Lambda}[:, \mathcal{J}]$ in order to verify (6.10) to check to see if the set was orthogonal. However, in practice one is not given this set so the true number of columns that must be search may greatly exceed this number. However, in the sequel we show $\text{comp}_{\text{twt}}(\widehat{\Lambda})$ plays a very important role in determining the number of orthogonal configurations as well as the *flexibility* of a matrix $\widehat{\Lambda}$ to either be extended to a larger orthogonal configuration or modified to a new orthogonal configuration of the same size by replacing a row. More precisely, one may view $\text{comp}_{\text{twt}}(\widehat{\Lambda})$ as a coarse measure of the size and number of the sets $\mathcal{N}_{k,c}$ in (6.10a). That is, if $\text{comp}_{\text{twt}}(\widehat{\Lambda})$ is small then by examining (6.10) it is likely that the $\mathcal{N}_{k,c}$ found to satisfy (6.10) are large and few in number. To illustrate this relation, note that in the case that every row of $\text{comp}_{\text{twt}}(\widehat{\Lambda})$ is

identical (i.e. the set of associated vectors satisfy condition (i) of Theorem 3.3.10) then one may, in order to verify (6.10), greedily take the sets $\mathcal{N}_{k,c}$ based on the p -adic representation of the row index. More precisely, let

$$i = i_0 + p \cdot i_1 + \cdots + p^{m'-1} \cdot i_{m'-1}$$

be the p -adic representation of i and let

$$\mathfrak{N}_{k,c}(\ell) = \{i : 0 \leq i < \ell \text{ and } i - p^c \cdot i_c = k\} \quad (6.12)$$

Then, if

$$\mathcal{N}_{k,c} = \mathfrak{N}_{k,c}(\ell) \quad (6.13)$$

it is simple to see that $\{i, j\} \in \mathcal{N}_{k,c}$ for $i \neq j$ if and only if $i_c = j_c$. Hence, the conditions of (6.10) are satisfied and one has

$$\text{comp}_{\text{twt}}(\widehat{\Lambda}) = \lceil \lg_p(\ell) \rceil.$$

We note, however, that if the original set of vectors had distinct rows while having common indices on a set of size $\lceil \lg_p(\ell) \rceil$ we could use the same set of $\mathcal{N}_{k,c}$ to arrive at the result. That is, if $\text{comp}_{\text{twt}}(\widehat{\Lambda}) < m'$ then there may be additional $\widehat{\Lambda}' \in (\mathbb{Z}_p)^{\ell \times m'}$ such that

$$\widehat{\Lambda}'[:, \mathcal{J}] = \widehat{\Lambda}[:, \mathcal{J}]$$

where $|\mathcal{J}| = \text{comp}_{\text{twt}}(\widehat{\Lambda})$ and is a valid subset of column indices for (6.10). In particular, it is clear that there are $p^{a \cdot \ell \cdot (m' - \text{comp}_{\text{twt}}(\widehat{\Lambda}))}$ such $\widehat{\Lambda}' \in (\mathbb{Z}_p)^{\ell \times m'}$. Thus, if $\text{comp}_{\text{twt}}(\widehat{\Lambda}) < m'$ there are many possible ways to naively adapt an orthogonal set by replacing a row with another one that is constant on \mathcal{J} if $\text{comp}_{\text{twt}}(\widehat{\Lambda}) < m'$. However, from the preceding example it is clear that this is not the only way one may adapt $\widehat{\Lambda}$. In particular, so long as we can find a subset of column indices such that under a suitable permutation of columns and rows the set of $\mathcal{N}_{k,c}$ from (6.13) are valid in (6.10) the resulting set is orthogonal. Thus, while it may not be possible to replace a row by only examining a specific subset of columns it may be possible to replace a row using a different subset of indices. In particular, in the current example where every row of $\widehat{\Lambda}$ is equal, one may search over every subset of columns of size ℓ to find a subset of indices to use to verify (6.10). Thus, we let

$$\text{flex}_{\text{twt}}(\widehat{\Lambda}; t) = \frac{|\{(\tau, \sigma) : \text{ such that (6.10) is true for } \{\mathcal{N}_{\tau(k), \sigma(c)} = \mathfrak{N}_{k,c}(t)\}\}|}{\min(\ell!, t!)} \quad (6.14)$$

be the number of row and column permutations for which the set of standard configurations $\mathfrak{N}_{k,c}(t)$ can be used to verify (6.10). We note that the term $\min(\ell!, t!)$ in the denominator of (6.14) comes from the structure of $\mathfrak{N}_{k,c}(t)$. That is, for every τ , $\mathfrak{N}_{k, \sigma(c)}(t) = \mathfrak{N}_{\tau\sigma(k), c}(t)$ where

$$\tau_\sigma(i) = p^{\sigma(0)} \cdot i_0 + p^{\sigma(1)} \cdot i_1 + \cdots + p^{\sigma(m'-1)} \cdot i_{m'-1}$$

is the equivalent permutation on the row indices. Hence the numerator of (6.14) over counts $\min(\ell!, t!)$ times too many permutations. We note our definition of flexibility excludes configurations that requires more than t columns to be examined to verify (6.10). That is, if $\text{flex}_{\text{twt}}(\widehat{\Lambda}; t) = 0$ it does not necessarily imply that one can not find a row in $\widehat{\Lambda}$ that can be replaced by another vector in $(\mathbb{Z}_p)^{m'}$ to yield an new orthogonal set. It simply implies

that it can not be done using fewer than $t + 1$ columns. This illustrates that there is a fundamental relationship between $\text{flex}_{\text{tw}}(\widehat{\Lambda}; t)$ and $\text{comp}_{\text{tw}}(\widehat{\Lambda})$. To make this more precise we have the following theorem.

Theorem 6.2.1. *For any matrix $\widehat{\Lambda} \in (\mathbb{Z}_p)^{\ell \times m'}$*

$$\text{comp}_{\text{tw}}(\widehat{\Lambda}) = \min\{t : \text{flex}_{\text{tw}}(\widehat{\Lambda}; t) > 0\}$$

Proof. This proof is clear from the definitions. In particular, from (6.11) one has that the complexity of $\widehat{\Lambda}$ is the smallest subset of columns of $\widehat{\Lambda}$ need to verify that $\widehat{\Lambda}$ defines a self orthogonal set for a chosen $\overline{\Lambda}$. Further, by (6.14) one has $\text{flex}_{\text{tw}}(\widehat{\Lambda}; t) = 0$ for any non-orthogonal set. Thus, $\text{flex}_{\text{tw}}(\widehat{\Lambda}; t) > 0$ on the for every t such that a subset of columns of $\widehat{\Lambda}$ of cardinality t may be used to verify that $\widehat{\Lambda}$ defines a self orthogonal set. This yields the result. ■

We note that this observation has great consequence on the development of algorithms that we develop in the sequel. That is, the most flexible configurations are lowest complexity. Thus, if one greedily tries to find a basis of the form \mathcal{O}_f then one will not likely end up in a position that can not be adapted if a set of the form \mathcal{O}_f can not be found. More precisely, to proceed in a manner that is the most flexible as possible (i.e. to keep $\text{flex}_{\text{tw}}(\widehat{\Lambda}; m')$ as large as possible at each stage of the search) one would like to keep $\widehat{\Lambda}$ constant on as large a set of column indices as possible as it trivially admits the largest number of column and row permutations that can be used to satisfy (6.10). That is, a basis in which $\widehat{\Lambda}$ is constant on every row is the most flexible basis, i.e. such a basis has $\text{flex}_{\text{tw}}(\widehat{\Lambda}; m) = m'!$. However, we note that it is not necessary for a basis to have such large flexibility. In fact it is easy to see that \mathcal{O}_c has $\text{flex}_{\text{tw}}(\widehat{\Lambda}; m) = 1$. Thus, it is natural to wonder how two such bases differ. In the sequel we will show that while \mathcal{O}_f has the lowest complexity the number of bases with this form are fewer in number than those with lower complexity. This observation yields additional insights in to how one might develop algorithms to find a basis. In particular, if one greedily tries to find a basis of the form \mathcal{O}_f and one is not successful then there are many other bases in the neighborhood of all basis of the form \mathcal{O}_f for which one may turn the search algorithm to.

In the preceding discussion we neglected mention of $\overline{\Lambda}$ by inserting the constraint we have on its choice in (6.10). That is, given any $\widehat{\Lambda}$ such that $\text{flex}_{\text{tw}}(\widehat{\Lambda}; m) > 0$ we have shown that there is some choice for $\overline{\Lambda}$ such that the set corresponding to the pair is orthogonal. However, something that is far less clear is that the number of possible $\overline{\Lambda}$ that may be paired with a given $\widehat{\Lambda}$ that yield distinct configurations varies inversely to $\text{flex}_{\text{tw}}(\widehat{\Lambda}; m)$. In this direction let

$$\overline{\Lambda}_0 = [i_j]_{i,j=0}^{m'-1}$$

where again we let i_j be the coefficient in the p -adic expansion of i , i.e. $i = i_0 + p \cdot i_1 + \dots + p^{m'-1} \cdot i_{m'-1}$. Then we have the following lemma.

Lemma 6.2.2. *If $(\widehat{\Lambda}, \overline{\Lambda})$ determine an orthonormal basis in \mathbb{C}^m then $\overline{\Lambda}$ is a row permutation of $\overline{\Lambda}_0$.*

Proof. This follows directly from the definition of the twisted Hamming weight. That is, as there are only $m = p^{m'}$ distinct elements $\overline{\Lambda}$ may assume one must use each such element to define $\overline{\Lambda}$. ■

We let

$$\bar{\Lambda}_0(\tau) = [\tau(i)_j]_{i,j=0}^{m'-1} \quad (6.15)$$

denote the row permutation of $\bar{\Lambda}_0$. Now, if (τ, σ) is a pair of row and column permutations such that (6.10) can be verified for

$$\{\mathcal{N}_{\tau(k),\sigma(c)} = \mathfrak{N}_{k,c}(t)\}$$

then the vectors corresponding to $\hat{\Lambda}$ and $\bar{\Lambda}_0(\tau_\sigma)$ yield a basis where τ_σ corresponds to the effective row permutation caused by the pair of row and column permutations (τ, σ) being applied to $\hat{\Lambda}$. Thus, if $\text{flex}_{\text{tw}}(\hat{\Lambda}; m) = m!$ then any (every) choice of permutation for $\bar{\Lambda}_0$ will yield a basis while any $\hat{\Lambda}$ such that $\text{flex}_{\text{tw}}(\hat{\Lambda}; m) = 1$ only one permutation will yield a basis. Thus, we must find an alternate way to understand the number of $\bar{\Lambda}$ that may be paired with $\hat{\Lambda}$. Note that while row and column permutations of $\bar{\Lambda}$ in general may not be used to find new pairings for $\hat{\Lambda}$ permutations to the values of the matrix will. That is, let σ_p be any permutation of $\{0, 1, 2, \dots, p-1\}$ and let

$$\tau_{\sigma_p}(i) = \sigma_p(i_0) + p \cdot \sigma_p(i_1) + p^{m'-1} \cdot \sigma_p(i_{m-1})$$

be the corresponding row permutations. Then if, $\hat{\Lambda}$ and $\bar{\Lambda}_0(\tau)$ correspond to a basis for \mathbb{C}^m then so will $\bar{\Lambda}_0(\tau \circ \tau_{\sigma_p})$. However, it is clear that this will only yield a unique basis if there are enough distinct rows in $\hat{\Lambda}$. That is, if $\hat{\Lambda}$ has any non-distinct rows then there are some row permutations counted by $\text{flex}_{\text{tw}}(\hat{\Lambda}; m')$ for which the corresponding $\bar{\Lambda}_0(\tau)$ do not define unique bases. Then we have the following lemma.

Theorem 6.2.3. *Let $\hat{\Lambda} \in (\mathbb{Z}_p)^{p^{m'} \times m'}$ and suppose there exists some permutation τ such that $\hat{\Lambda} + \bar{\Lambda}_0(\tau)$ forms a basis. Let $\{\tau_{i,p}\}$ be the set of permutations of $\{0, 1, 2, \dots, p-1\}$ which acts on the elements of $\bar{\Lambda}_0(\tau_j)$ which yield a distinct ordering of the rows of $\hat{\Lambda}$. Then,*

$$\Lambda_j = \hat{\Lambda} + \bar{\Lambda}_0(\tau_{j,p})$$

are distinct up to row permutations. Moreover, each set of complex vectors

$$\{\mathbf{c}(\Lambda_j[i, :], \boldsymbol{\beta}; L, p^a)\}_{i=0}^m$$

defines a unique basis for \mathbb{C}^m .

Proof. This theorem is a direct consequence of the discussion preceding it. ■

In the preceding discussion we have mainly focused on the existence of orthogonal sets as well as how one may enumerate them with an emphasis of search and scheduling. In particular we have shown that any $\hat{\Lambda}$ which only has one distinct row was shown to be the most flexible. Thus, from an algorithmic perspective it is natural to consider the bases that may be derived from such a basis by interchanging only p rows. Then, in turn, the bases which may be derived from this derived basis in a similar manner and so forth, constructing a tree where at the root one has the configuration that is the easiest to adapt and all configurations that may be derived are children. More precisely for each $\hat{\lambda}_0 \in (\mathbb{Z}_{p^{a-1}})^{m'}$ we

may consider a tree in which

$$\widehat{\Lambda} = \begin{bmatrix} \widehat{\lambda}_{0,0} & \widehat{\lambda}_{0,1} & \cdots & \widehat{\lambda}_{0,m'-2} & \widehat{\lambda}_{0,m'-1} \\ \widehat{\lambda}_{0,0} & \widehat{\lambda}_{0,1} & \cdots & \widehat{\lambda}_{0,m'-2} & \widehat{\lambda}_{0,m'-1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \widehat{\lambda}_{0,0} & \widehat{\lambda}_{0,1} & \cdots & \widehat{\lambda}_{0,m'-2} & \widehat{\lambda}_{0,m'-1} \end{bmatrix}$$

labels the root and each node on the i th level is labeled by a matrix $\widehat{\Lambda}$ containing $p^{m'-i}$ copies of $\widehat{\lambda}_0$ which satisfies (6.10). Further, we say that a node at level $i+1$, labeled by $\widehat{\Lambda}_{i+1}$, is a child of a node at i , labeled by $\widehat{\Lambda}_i$, if the set of distinct rows of $\widehat{\Lambda}_i$ are contained in the set of distinct rows of $\widehat{\Lambda}_{i+1}$. While this yields an efficient method and structure to enumerate every basis, taking a slightly more intuitive approach yields a more effective way to search. Consider building up a basis for a set of given vectors in a search to find the maximally weighted basis. In this direction, recall that every basis formed by a code derived over a cross product of the integers is of the form

$$\widehat{\Lambda} + \overline{\Lambda}_0(\tau_j)$$

where $\overline{\Lambda}_0(\tau_j)$ was defined in (6.15). If one attempts to construct a basis one may consider a process whereby one first selects a codeword $\mathbf{c}_0 = (\widehat{\lambda}, \bar{\lambda})$ and one temporarily forms a basis by choosing $\widehat{\Lambda}$ to have each row equal to $\widehat{\lambda}$. Then, in order to keep track of the selected codeword one may label one position in $\overline{\Lambda}_0(\tau_j)$ corresponding to \mathbf{c}_0 and mark the remaining $m-1$ positions with don't cares. Then, one may sequentially add in additional codevectors making sure that at each stage the constraints (6.10) are met by ensuring there is a vacant row in $\overline{\Lambda}_0(\tau_j)$ for which one may meet an appropriate constraint on $\bar{\lambda}$. More importantly, an entry on $\overline{\Lambda}_0(\tau_j)$ that is labeled with don't cares tells one exactly which constraints must be examined to ensure that the twisted hamming weight is positive. This development may sound quite familiar. Indeed, as we have seen every maximal clique (and hence in the present context basis) in a graph must have a unique color. As each basis must have a distinct $\bar{\lambda}$ to form a basis it should be clear that the set of distinct $\bar{\lambda}$ color the graph that describes orthogonality relationships. This is the content of the following theorem.

Theorem 6.2.4. *Let p be a prime and suppose that $m', a \in \mathbb{Z}$, $m' > 0$ and $a > 0$. Then, consider a graph \mathcal{G} with vertex set $\mathcal{C}(\Upsilon_1, \{0\}; \mathbb{F}_{p^{m'}})$ and edge set*

$$(\mathbf{c}(\boldsymbol{\lambda}, 0; \mathbb{F}_{p^{m'}}, p^a), \mathbf{c}(\boldsymbol{\lambda}', 0; \mathbb{F}_{p^{m'}}, p^a)) \in E$$

if and only if $\mathbf{c}(\boldsymbol{\lambda}, 0; \mathbb{F}_{p^{m'}}, p^a)^\dagger \mathbf{c}(\boldsymbol{\lambda}', 0; \mathbb{F}_{p^{m'}}, p^a) = 0$. Then, assigning $\mathbf{c}(\boldsymbol{\lambda}, 0; \mathbb{F}_{p^{m'}}, p^a)$ the color $\bar{\lambda}$ is a $p^{m'}$ -coloring of \mathcal{G} .

We note that Theorem 6.2.4 is quite important in terms of user selection algorithms. In fact, finding a minimal coloring in general is an NP-complete problem and often secondary heuristics must be employed to find approximate coloring to use to find a maximal clique. Thus, one obtains a reduction in the complexity of finding a maximal clique by using the deterministic coloring of Theorem 6.2.4. However, Theorem 6.2.4, as stated, only describes a coloring for the root code. As we have argued a multi-user MIMO system in general is correlated. Hence, one more often than not must consider scheduling users from local codes. In Chapter 3 we argued that this problem is not much more complex than scheduling the root code. However, as the graph for such a code is much more complex it is natural to

suspect that the coloring of Theorem 6.2.4 is not directly applicable to coloring the universal code. However, in Section 3.6 we showed through Theorem 5.3.4 that the universal code did not introduce any new maximal cliques. In fact, we saw that in general one must show three inner products are 0 to determine if a pair of codewords from the universal code were orthogonal. As one of these inner products were between the associated root code one may extend a coloring of the root code to the universal code by coloring each element of a local code with the color of its root. This is the content of the following theorem.

Theorem 6.2.5. *Let p be a prime and suppose that $m', a \in \mathbb{Z}$, $m' > 0$ and $a > 0$. Consider a root code $\mathcal{C}(\Upsilon_1, \{0\}; \mathbb{F}_{p^{m'}})$ and let $\mathcal{C}_{\mathbf{F}}(\alpha, \gamma, \{\mathcal{B}_i\})$ be the universal code associated with $\mathcal{C}(\Upsilon_1, \{0\}; \mathbb{F}_{p^{m'}})$ for some chosen design basis. Now, consider a graph \mathcal{G} with vertex set $\mathcal{C}(\Upsilon_1, \{0\}; \mathbb{F}_{p^{m'}})$ with an edge between any two orthogonal vectors in $\mathcal{C}_{\mathbf{F}}(\alpha, \gamma, \{\mathcal{B}_i\})$. Then, assign to each member of a local code in the universal code the color of its root. This yields a $p^{m'}$ -coloring of \mathcal{G} .*

Proof. This follows simply from the fact that for two elements of the $\mathcal{C}_{\mathbf{F}}(\alpha, \gamma, \{\mathcal{B}_i\})$ to be orthogonal the corresponding roots must be orthogonal by Theorem 5.3.4. Hence, two elements of the $\mathcal{C}_{\mathbf{F}}(\alpha, \gamma, \{\mathcal{B}_i\})$ are adjacent if the corresponding roots are adjacent which implies each root has a distinct color. ■

From Theorem 6.2.4 and Theorem 6.2.5 one may easily obtain colorings for the graphs associated with orthogonal processing modes and hence improve the performance of any branch and bound algorithm we have considered. However, we still have not exploited the fact that we have a graph in which the chromatic number equals the clique number. As we are in a quite special case one may suspect that there is a way to exploit the situation and indeed there is. Note that in Östergård's algorithm one provided bounds on the size of any graph contained in a subgraph by using cardinality of the underlying set. However, as we know that the colorings of the graph relates directly to the size of the maximal clique one may consider using the number of colors in a subgraph as a better indication of the size of a possible clique. That is, even if a subgraph has a large number of vertices if there are few colors in the subgraph then proceeding to search on such a subgraph will not dramatically increase the size of the clique. Moreover, for the problem of channel aware scheduling, one is interested in finding the maximally *weighted* clique. As the weight of a clique is the sum of the included vertices, one may easily extend the described search algorithm by using the sum of the largest weighted vertices from each color class on the subgraph as an upper bound on reward one receives by considering a subgraph. More precisely, let \mathcal{G}_d be any subgraph of a given graph \mathcal{G} . Then, we let $\deg(\mathcal{G}_d)$ be the number of color class that exists on \mathcal{G}_d and let $\text{Deg}(\mathcal{G}_d)$ be the sum of the maximum weight for each color class that is contained in \mathcal{G}_d . Then as done previously one may keep a table of degrees of every subgraph consider by a branch a bound algorithm.

It may seem that computing the number of color class that exist on a graph and the corresponding maximal weight for each color class that is contained in \mathcal{G}_d increases the complexity of the search. However, we note that we already have an efficient structure in place to compute just these quantities. In particular, our codebook and our "trees" play just this role. More precisely, employing a code from Section 3.3.1 that is a disjoint union, one may use each one of these bases for table lookup of degree or weighted degree. One just needs to form a tree with each one of these bases as root and join every child that represents the same basis. Using $\bar{\Lambda}$ to index the colors one only needs to inspect which entries in the root are active and their associated weights to determine $\deg(\mathcal{G}_d)$ or $\text{Deg}(\mathcal{G}_d)$.

More precisely, to determine $\deg(\mathcal{G}_d)$ or $\text{Deg}(\mathcal{G}_d)$ one may just examine the elements of the disjoint union which exist on the subgraph.

The maximum weighted clique problem may be solved quite efficiently on the graphs associated to the quantizers of interest. In particular, for the graphs associated to orthogonal processing modes of the quantizers in Section 3.3.1 have chromatic indices that equal the clique number. Thus, one may use efficient algorithms, such as Östergård's algorithm, with an appropriate coloring to solve the maximally weighted clique problem and the channel aware scheduling problem. Most importantly this may be done with or without the assumption of the Rayleigh model or heterogeneous fading amongst the users in the system. More precisely one may use the geometric factors we developed in Section 5.3 and still preserve chromatic number of the much larger graph. Thus, the codes developed in Chapter 3.3 to contain many orthogonal bases may be paired with our adaptive covariance methods of Chapter 5 to yield a framework for robust and efficient scheduling in the multi-user MIMO channel.

Conclusions and Future Work

In this thesis we have identified the problem of feedback design as a central issue in both increasing throughput and reducing the complexity in a multi-user MIMO system. To show this we developed a systematic channel quantization framework which treats the issues of mean squared quantization error and scheduling complexity in a common framework. This allows a system designer to optimize the trade-off between throughput and the complexity of user selection. An added benefit of this framework is that it enabled us to analyze the stability of a system to variety of channel models.

In Section 2.3 we examined the results of [67] which showed that high rate systems with few users and finite rate feedback must use large codebooks to ensure that the system performance is not limited. In such cases it is of interest to develop *structured* codebooks that enable user terminals to efficiently quantize their channel vectors. A central contribution of this thesis was the development of a systematic construction of channel quantizers in Chapter 3. This construction allowed one to trade-off the achieved mean squared quantization error and the number of orthogonal bases contained in the quantizer. As a particular figure of merit we chose a high SNR approximation to the SINR of set of users, SINR_{sat} . To yield codes with large values of SINR_{sat} our systematic construction of channel quantizers consisted of three main structural components; a family of low-rate codes which contain many orthogonal bases, a systematic method to construct intermediate rate codes through unions of low-rate codes and a rate doubling operation which may be used to construct high rate codes with low complexity quantization algorithms. With an appropriate choice of parameters one may use our framework to construct a high rate channel quantizer for which multi-stage quantization is optimal. This may be done by first quantizing a channel vector to a base code of half the rate. Then, using the same quantization algorithm, by performing second quantization on a transformation of the channel vector where the transformation is determined by the first stage of quantization. Such a codebook is of great interest for MIMO broadcast systems as the quantization is performed at the user terminals. In many cases the user terminals are power and complexity limited and hence may not have the resources to perform high complexity quantization needed to obtain high rates.

However, SINR_{sat} is a high SNR approximation of the achieved SINR of a system that uses a particular quantization scheme and not a measure of the achieved SINR for a given SNR. A multi-user system may not be optimized using this criterion alone. In a multi-user system one must develop intelligent scheduling algorithms to exploit the multi-user diversity by selecting users with low co-channel interference. Thus, in Chapter 4 we presented a simple model and associated base station architecture in which the system designer may study the trade-off between the order statistic gain and the multi-node matching gain.

With the model and system architecture of Chapter 4 one may further analyze how the order statistic gain and the multi-node matching gain trade-off is affected by variations

in the structure of the feedback design. A benefit of this approach is that it allows one to examine the effects that variations in the channel model have on the performance of a system using our quantization framework and system architecture. As such, we identified the relevant statistical models for the fading process in multi-user MIMO systems as well as presented a discrete model for user feedback in Chapter 5. This model allowed a base station to estimate the covariance matrix of each user and identify the users with poor fading conditions. For users that have been estimated to have poor channel conditions we showed that one may use the systematic feedback framework from Chapter 3 to adapt the channel feedback to better match the covariance structure of these users channels. Thus, our systematic feedback framework has broad practical relevance as it provides a common framework in which one may simultaneously develop good structured high rate quantizers, develop low complexity scheduling frameworks as well as provides a systematic framework in which a system may adapt to unknown channel correlation.

An additional benefit of the model and system architecture of Chapter 4 is it allows one to examine the complexity of user selection. That is, the model and system architecture of Chapter 4 allowed us the ability to realize the multi-user diversity gains inherent in a multi-user system by employing simple thresholds on each user's individual SNR to limit the search to a smaller pool. Depending on the SNR threshold, however, one may still have too many subsets of users to consider for full search to be feasible. Thus, in Chapter 6, we presented efficient algorithms for user selection that exploit the structure of our systematic feedback. This allows one to greedily search for users with low co-channel interference. Thus, the codes developed in Chapter 3.3 may be paired with the adaptive methods of Chapter 5 and used in conjunction with our system architecture of Chapter 4 to yield a framework for robust and efficient scheduling in the multi-user MIMO channel. To summarize, the major contributions of this thesis are:

1. Identifying the problem of feedback design as an integral part of the joint design of efficient channel aware schedulers as well as robust low complexity multiplexing schemes (Chapter 1)
2. Providing a systematic feedback framework in which the system designer may trade-off between the order statistic gain and the multi-node matching gain to meet certain system objectives (Chapter 3)
3. Providing a simple base station architecture to understand to trade-off between the order statistic gain, the multi-node matching gain and system complexity (Chapter 4)
4. Identifying an appropriate discrete model for user feedback and identifying an associated expectation-maximization algorithm to estimate this distribution under unknown channel conditions and identify clusters of users with similar channel correlation (Chapter 5)
5. Providing a systematic method to adapt our feedback framework so that the resulting design remains stable as the statistics of the underlying channel change (Chapter 5)
6. Providing a new class of algorithms for user selection that exploit the structure of our feedback framework to solve the user scheduling problem (Chapter 6)

Future Work

The results contained in this thesis have a broad scope, much of which was kept implicit in the discussion. As the quantizers developed in this thesis have good mean squared error performance they are of interest in their own right in broader contexts of coding and approximation theory. These applications are discussed in Section 7.2. Additionally, the framework to construct and analyze the sparse and dense codes in our systematic construction use results from quantum coding theory and may be of additional interest in that context. However, there are many additional questions for practical system design which we have left open. We next overview extensions to our work for MIMO system design. Then we provide possible extensions in these broader areas.

■ 7.1 MIMO System Design

We consider two areas of practical MIMO system design for which our results are of use.

■ 7.1.1 Effects on Service Rate Variance

In Chapter 3 and Chapter 6 we argued that the symmetry group of the quantizer reduces the mean squared quantization error as well as reduces the complexity of user selection. These arguments were based on the fact that a large symmetry group implies that there is a large number of unitary matrices which fix the code. Thus, from a quantization perspective, the code is well matched to an isotropic source and the resulting mean squared quantization error is low. However, as the rates achieved by a group of users is also invariant to unitary transformations this implies that there are a small number of large generalized switches that may be formed to represent the set of achievable rates. Moreover, in a system with many users (a small multiple of the size of the transmit array) we showed that the probability that a maximally size clique may be found in any one of these switches is quite high. Thus, from a quality of service standpoint one may, with high probability, guarantee that there is a set that achieves a desired level of service. It is of broader interest to understand how the reduction in the variance in the services rates provided by a quantizer with large symmetry groups has on the ability for a system to provide quality of service. In particular, it is unclear whether the proportionally fair algorithm can or should be augmented to meet additional quality of service constraints. More broadly, it is of interest to consider how the reduction in the variance of the service rates enables one to provide a secondary quality of service guarantee for a system while simultaneously achieving a delay guarantee. Many of these answers appear to be able to be addressed inside the framework of Stolyar [118]. However, a result that is much easier to address is how quality of service is effected by variations in the channel model.

■ 7.1.2 Channel Modeling and Stability

In this thesis we developed several models for the MIMO channel. In particular, we began by assuming the Rayleigh model and proceeded to develop a systematic quantization framework with this assumption. We then analyzed the system performance and showed that the resulting system performance is not greatly effected by mild spatial correlation assumptions and proceed to develop a model in which one could estimate the underlying channel covariance. We further exhibited how high rate codes have a natural immunity to correlation and how one may adapt low rate quantizers using our high rate framework to

improve system performance. However, we did not make any mention of temporal correlation of channel vectors. In particular, in many practical systems there may be a slow (or fast) varying channel mean or some Markov structure to the underlying fading process. If one is only interested in the quality of service we note that the framework of Stolyar [118] is sufficient to address a generalized switch for which the switch state follows a finite irreducible Markov chain. As the switch state is completely determined by the realization of the channel state in a system with finite rate feedback, the question of stability and throughput for a MIMO channel with a Markov structure may be addressed without modification. However, these results will be improved if one can estimate and predict the state of the fading process and use our adaptive framework to match the feedback scheme to the fading state of the channel.

As we have illustrated in Chapter 4 in a multi-user MIMO system one is not interested in tracking and estimating when a user experiences a minimal degree of channel correlation, but rather when the spatial covariance of a user's channel vectors has a principal component which is much larger than the remaining modes. If this is the case the adaptive method we have presented may be applied to combat the possible degradation to the system performance. Thus, it is of interest to understand the ability of one to track this phenomenon, especially with user mobility, and whether, for practical channels, it is reasonable to assume that one may form accurate estimates of the channel. What should be noted is there is already a natural robustness to temporal correlation embedded in our existing channel estimation framework. Indeed, as we modeled the prior distribution on the channel using a generalized Dirichlet distribution there are free parameters for which one may make some inference on the underlying propagation environment. Implicitly this prior was chosen as it has been shown empirically [27] to model "temporal" correlation in 2D-images and more general non-independent samples in time. Thus, it is of interest to classify the fading environments for which the present architecture fails to produce valid estimates.

■ 7.2 Coding and Approximation Theory

Our construction of channel quantizers as well as our system framework also make progress in other directions. We briefly describe these areas and other problems that may be addressed with our channel quantization and MIMO system framework.

■ 7.2.1 Code Analysis

We note that our systematic construction of quantizer produces a large family of codes, some of which outperform existing constructions. A natural question is: How good are the codes that have been developed in our framework in terms of the quantization error? In this thesis we resorted to simulation to answer this question. However, as our constructed codes are quite structured and it is a natural question to ask how one may analyze these codes. An exact method exists to analyze random vector quantization as well as the upper bound. One would like a similar expression, minimally an approximation, for a general quantizers in our framework in small dimensions. In particular, one would like to be able to derive the performance of the order statistics for the quantization error for channel quantizers constructed with our systematic framework.

Given the ability to analyze the performance of a quantizer in our framework with order statistics it is natural to consider a further upper bound on the performance when one places a constraint on the number of bases contained in the code. That is, our current

upper bound does not have a constraint on the number of orthogonal bases contained in the code. We briefly discussed the effects of this constraint in the absence of order statistics, but the result is quite loose. We note that one may use the results of [44] and [57] to arrive at a bound on the number of vectors orthogonal with any codeword when the number of distinct inner products between every pair of codewords in a code is small. With these results one can provide a simple upper bound on the distribution of the inner products of codewords for a code with a fixed number of orthogonal vectors. However, bootstrapping this result to a result on the number of orthogonal bases contained in the code using general results from graph theory produce results that are overly optimistic [136]. Thus, it is of interest to develop an upper bound on the quantization error given a constraint that there are a fixed number of orthogonal bases contained in the code.

Linear Codes over Rings

Recall that a ring $\mathcal{R} = (R, \oplus, \otimes)$ is a non-empty set R together with two binary operations \oplus and \otimes such that (R, \oplus) is a commutative group and multiplication is both associative and right and left distributive. For example, the set of integers, \mathbb{Z} , is a ring as well as \mathbb{Z}_ℓ (the integers modulo some composite number ℓ). As is standard in algebraic coding theory, one can view codewords of length m with symbols taken from the ring \mathcal{R} as polynomials of degree m with coefficients from \mathcal{R} . In this direction, let $\mathcal{R}[X]$ be the polynomial ring over the ring \mathcal{R} . That is, $\mathcal{R}[X]$ is the set of all finite sums of the form $a_0 + a_1X + a_2X^2 + \dots + a_kX^k$ where $a_i \in \mathcal{R}$. Analogous to the case of polynomial rings over finite fields we will say that a function $f \in \mathcal{R}[X]$ is monic if $f = a_0 + a_1X + a_2X^2 + \dots + 1 \cdots X^k$. Moreover, we will say that the polynomial $f \in \mathcal{R}[X]$ is:

- (a) a *unit* if there exists an element $h \in \mathcal{R}[X]$ such that $f \cdot h = 1$,
- (b) *regular* if f is not a zero divisor and
- (c) *irreducible* if f is not a unit and when ever $f = g \cdot h$ then either g or h is a unit.

It is natural to wonder whether knowledge of the characteristics of a polynomial over a ring in anyway correspond to a equivalent polynomial over a finite field. In this direction, let $\mathcal{R} = \mathbb{Z}_{p^\ell}$ and let μ be the homomorphism from \mathbb{Z}_{p^ℓ} to \mathbb{Z}_p that reduces any element of \mathcal{R} modulo p . We now recall the following lemma from [85].

Lemma A.0.1. *Let $f \in \mathcal{R}[X]$ be given. Then,*

- (a) *if f is irreducible, then μf is irreducible*
- (b) *if μf is irreducible, then f is irreducible*
- (c) *if f is a zero divisor, the $\mu f = 0$.*

This lemma is particularly useful in the context of cyclic codes. Recall that a cyclic code defined over a finite field is isomorphic to an ideal in $\mathbb{F}_q[X]/(X^n - 1)$ that is generated by a single polynomial $\bar{g}(x)$. That is, a cyclic code $C = \langle \bar{g}(x) \rangle$ for some *generator polynomial* $\bar{g}(x)$. It is natural to ask if a cyclic code defined over a finite fields of characteristic p , for some prime p , has a corresponding code over the ring \mathbb{Z}_{p^ℓ} . That is, given a generator polynomial for a code over \mathbb{Z}_p is it possible to “lift” the generator polynomial up to a generator polynomial over \mathbb{Z}_{p^ℓ} . This question was answered in [29] and is described by the following lemma [85].

Lemma A.0.2. *Let $f \in \mathcal{R}[X]$ and suppose*

$$\mu f = \bar{g}_1 \cdot \bar{g}_2 \cdots \bar{g}_k.$$

where the \bar{g}_i are pairwise co-prime. Then, there exists $g_1, g_2 \dots g_k$ that are pairwise co-prime such that

$$f = g_1 \cdot g_2 \cdots g_k$$

and $\bar{g}_i = \mu g_i$.

Hence, for any classical code defined over a finite field \mathbb{Z}_p , one may use Lemma A.0.2 to construct a similar code over \mathbb{Z}_{p^ℓ} . This yields a simple way to find linear codes over the rings \mathbb{Z}_{p^ℓ} and systematically construct nested codes. However, the analysis of these codes require additional machinery. In particular, the Mattson-Solomon polynomials (i.e. the Discrete Fourier Transforms) [130] for cyclic codes over rings do not reside in a polynomial ring over a finite field as was the case for cyclic codes over finite fields [103]. That is, the Mattson-Solomon polynomial is an element of a polynomial ring defined over an extension of the base ring, which is not in general a finite field (unless of course the base ring is a finite field). In this direction we briefly review the necessary results for extensions of finite rings.

We begin by briefly reviewing the theory of extensions of finite fields and then apply these ideas to extensions of finite rings. Suppose K and F are fields and $K \subset F$. That is, K is a subfield of F . Then, we say that F is an extension of K . The field F can be thought of as a vector space over K and we denote the dimension of this vector space as $[F : K]$. We are particularly concerned with the Galois structure of finite extensions as it will be useful in characterizing the structure of associated quantization codebooks. That is, we are interested in the set of all automorphism of F that leave K fixed. We denote this set of automorphism of F as $\text{Gal}(F, K)$. In particular, we are interested in the Galois structure of finite *separable* extensions of finite rings. That is, if $K_r \subset F_r$ are rings then we say that F_r is an extension of the ring K_r and that F_r is a separable extension if and only if F_r is isomorphic to the quotient $K_r[X]/(f)$ for some monic irreducible polynomial $f \in K_r[X]$. As was in the case of finite fields one can show that there is only one unique (up to isomorphism) separable extension of a finite ring of a given degree. In this direction, let $\text{GR}(p^\ell, r)$ be the degree r Galois extension of the ring \mathbb{Z}_{p^ℓ} . The reader should note that,

$$\text{GR}(p, r) = \text{GF}(p^r) = \mathbb{F}_{p^r}$$

and $\text{GR}(p^\ell, 1) = \mathbb{Z}_{p^\ell}$. Due to Lemma A.0.1 it is natural to wonder if the Galois structure of a finite separable ring extension is at all related to the Galois structure of a finite separable field extension. We now have the following relation from [85].

Lemma A.0.3. *Let $\text{GR}(p^\ell, r)$ be the degree r Galois extension of the ring \mathbb{Z}_{p^ℓ} . Then, $\text{Gal}(\text{GR}(p^\ell, r), \mathbb{Z}_{p^\ell})$ is cyclic, isomorphic to $\text{Gal}(\mathbb{F}_{p^r}, \mathbb{F}_p)$ and generated by a power map on a primitive element of $\text{GR}(p^\ell, r)$. That is, if $\zeta \in \text{GR}(p^\ell, r)$ is primitive and $\sigma(\zeta) = \zeta^p$, then, $\langle \sigma \rangle = \text{Gal}(\text{GR}(p^\ell, r), \mathbb{Z}_{p^\ell})$.*

We caution the reader that the generator of the Galois group in Lemma A.0.3 does not in general act as a power map on every element of $\text{GR}(p^\ell, r)$ as $\text{GR}(p^\ell, r)$ is not cyclically generated. More precisely, for any element $u \in \text{GR}(p^\ell, r)$ let ζ be a primitive element of $\text{GR}(p^\ell, r)$. Then,

$$u = \sum_{i=0}^{\ell-1} u_i p^i$$

where $u_i \in \mathcal{T}_\zeta = \{0, 1, \zeta, \zeta^2, \dots, \zeta^{p^r-2}\}$. We let $\phi(u)$ be the map suggested by Lemma A.0.3. That is, the automorphism that acts on primitive elements as a power map. Thus, since

$\phi(u)$ is homomorphism,

$$\phi(u) = \sum_{i=0}^{\ell-1} u_i^p p^i. \quad (\text{A.1})$$

The automorphism $\phi(u)$ is the *Frobenius* automorphism and is a cyclic generator for the Galois group $\text{Gal}(\text{GR}(p^\ell, r), \mathbb{Z}_{p^\ell})$. Further, for a separable extension $\mathcal{R}_1 \subset \mathcal{R}_2$ one defines the *trace* map as

$$\text{Tr}_{\mathcal{R}_2/\mathcal{R}_1}(\alpha) = \sum_{i=0}^{[\mathcal{R}_2:\mathcal{R}_1]-1} \phi^i(\alpha) \quad (\text{A.2})$$

where $\text{Gal}(\mathcal{R}_2, \mathcal{R}_1) = \langle \phi \rangle$. This is a surjective homomorphism from \mathcal{R}_2 to \mathcal{R}_1 . We now consider how one may generalize the idea of a linear code over a finite field to one over a ring.

Recall that a linear code of length m over a finite field \mathbb{F}_q is a subspace of the vector space \mathbb{F}_q^m . A similar concept will hold in the case of codes over a ring, say \mathcal{R} . In this direction recall [60] that a (left) \mathcal{R} module is an additive abelian group A together with a map from $\mathcal{R} \times A \rightarrow A$ (for which the image of (r, a) is denoted ra) such that for all $r, s \in \mathcal{R}$ and $a, b \in A$ one has [60]:

- (i) $r(a + b) = ra + rb$
- (ii) $(r + s)a = ra + sa$
- (iii) $r(sa) = (rs)a$

If \mathcal{R} has an identity element $1_{\mathcal{R}}$, then

- (iv) $1_{\mathcal{R}}a = a$ for all $a \in A$

Let V be a left¹ \mathcal{R} module. Any \mathcal{R} -submodule is called a *code*. In particular, \mathcal{R}^n is a \mathcal{R} -module and a code of \mathcal{R}^n is any sub-module of \mathcal{R}^n . One should note that in the cases of interest one may think of a code over a ring in exactly the same way as one would a code over a finite field. In particular we have the following theorem from Huffman [29, 59].

Lemma A.0.4. *A non-zero linear code \mathcal{L} over $\text{GR}(p^\ell, r)$, for finite ℓ , has a generator matrix, which after a suitable permutation of the coordinated has the form,*

$$G = \begin{bmatrix} I_{k_0} & A_{0,1} & A_{0,2} & \cdots & A_{0,\ell-1} & A_{0,\ell} \\ 0 & pI_{k_1} & pA_{1,2} & \cdots & pA_{1,\ell-1} & pA_{1,\ell} \\ 0 & 0 & p^2I_{k_2} & \cdots & p^2A_{2,\ell-1} & p^2A_{2,\ell} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & p^{\ell-1}I_{k_{\ell-1}} & p^{\ell-1}A_{\ell-1,\ell} \end{bmatrix} \quad (\text{A.3})$$

where $A_{i,j}$ has elements from $\text{GR}(p^\ell, r)$. That is, \mathcal{L} consists of all codewords of the form

$$[\mathbf{g}_0 \mathbf{g}_1 \dots \mathbf{g}_{\ell-1}] G$$

where each vector \mathbf{g}_i is a vector of length k_i with components in $\text{GR}(p^\ell, r)$.

If a code \mathcal{L} has a generator of the form (A.3) then we say that the code has *type* $(k_0, k_1, \dots, k_{\ell-1})$. Moreover, it is easy to see via a simple counting argument that a code \mathcal{L}

¹We note that in general this definition in terms of non-commutative rings that are either finite or infinite. This is not needed here, but will be necessary in the proofs. For a complete introduction to this theory we refer the reader to [93]

of the form (A.3) has $p^{r\alpha}$ many codewords, where

$$\alpha = \sum_{i=0}^{\ell-1} (\ell - i)k_i.$$

Moreover, if a code \mathcal{L} is of the form (A.3) it is easy to compute the form of the dual of \mathcal{L} . More precisely, let \mathcal{L} be of type $(k_0, k_1, \dots, k_{\ell-1})$. Then we define the *dual* of \mathcal{L} to be

$$\mathcal{L}^\perp = \{\boldsymbol{\alpha} \in \mathcal{R}^m : \langle \boldsymbol{\alpha}, \boldsymbol{\beta} \rangle \quad \forall \boldsymbol{\beta} \in \mathcal{L}\}$$

where $\langle \boldsymbol{\alpha}, \boldsymbol{\beta} \rangle$ is the standard inner product. That is,

$$\langle \boldsymbol{\alpha}, \boldsymbol{\beta} \rangle = \sum_{i=1}^m \alpha_i \beta_i.$$

It is again easy to see via a simple counting argument that \mathcal{L}^\perp is of type $(k_\ell, k_{\ell-1}, \dots, k_1)$ where

$$k_\ell = m - \sum_{i=0}^{\ell-1} k_i.$$

Now, we return to to problem of interest. That is, we now consider developing complex codebooks that are the images of linear codes.

■ A.1 Systematic Unitary Space-Time Constructions

Consider the \mathbb{Z}_ℓ -submodule generated by the element $\mathbf{u} = (u_0, u_1, \dots, u_{m-1})$, which we denote $\mathcal{L}_{\mathbf{u}}$. That is,

$$\mathcal{L}_{\mathbf{u}} = \{k \cdot \mathbf{u} : k \in \mathbb{Z}_\ell\}.$$

This yields an associated code complex codebook $\mathcal{C}_{\mathbf{u}}^{(d)} \subset \mathbb{C}^m$ of cardinality ℓ . This class of codes can be thought of as a subset of m columns of the $\ell \times \ell$ DFT matrix [56, 138]. These code books are known to achieve the Welch bound in very special cases. To be more precise recall that a set of integers \mathbf{u} is a (ℓ, m, λ) perfect difference set in \mathbb{Z}_ℓ if the set $\{u_i - u_j \pmod{\ell} : j \neq i\}$ contains exactly λ copies of every integer in $\{1, 2, \dots, \ell - 1\}$. Clearly the parameters of a perfect difference set are not independent. In this direction, let

$$\lambda_\Delta(m, \ell) = \frac{m(m-1)}{\ell-1}.$$

Then, if \mathbf{u} is a (ℓ, m, λ) perfect difference set $\lambda = \lambda_\Delta(m, \ell)$. We have the following theorem from [138].

Theorem A.1.1. *The codebook $\mathcal{L}_{\mathbf{u}}$ is maximum Welch bound achieving if and only if \mathbf{u} is a $(\ell, m, \lambda_\Delta(\ell, m))$ difference set.*

Thus, the code $\mathcal{C}_{\mathbf{u}}^{(d)}$ achieves the Welch bound and is linear. However, in general the construction of perfect difference sets is very difficult. In particular, if $\lambda \neq \lambda_\Delta(m, \ell)$ then a (ℓ, m, λ) difference set trivially does not exist. Hence, one must select \mathbf{u} such that it is not a perfect difference set. Thus, in order to determine the stability in general, one must characterize the distinct difference sets of size K for the code $\mathcal{L}_{\mathbf{u}}$. We now turn to a less trivial construction that will serve as the base for our most general construction.

■ A.2 Generalized Reed-Muller Construction

Recall a ν -variate Boolean function is a function f from \mathbb{Z}_{2^ν} to \mathbb{Z}_2 . The simplest Boolean functions are the monomial functions

$$x_0^{i_0} x_1^{i_1} \cdots x_{\nu-1}^{i_{\nu-1}} \text{ for } \mathbf{i} \in \mathbb{Z}_{2^\nu}$$

where $i = \sum_{j=0}^{\nu-1} i_j 2^j$ is the 2-adic expansion of i . The degree of a Boolean monomial function is the number of variables that have an exponent of 1. For example, $x_1 x_3 x_5$ is of degree 3.

It is well known that every Boolean function can be written in algebraic normal form as the sum of monomial functions. That is, if f is a ν -variate Boolean function then,

$$f(x_0, x_1, \dots, x_{\nu-1}) = \sum_{\mathbf{i} \in \mathbb{Z}_{2^\nu}} x_0^{i_0} x_1^{i_1} \cdots x_{\nu-1}^{i_{\nu-1}}.$$

Moreover, to each ν -variate Boolean function f we may associate a sequence \mathbf{f} of length 2^ν by listing the values taken by f over \mathbb{Z}_{2^ν} in lexicographic order. This identification yields the r th order binary Reed-Muller Code $\mathcal{RM}(r, \nu)$. That is, the r th order binary Reed-Muller Code $\mathcal{RM}(r, \nu)$ is the subspace of \mathbb{Z}_{2^ν} spanned by the sequences associated to monomial functions of degree at most r . Clearly, this is a length 2^ν code and the dimension of this code is

$$\dim(\mathcal{RM}(r, \nu)) = \sum_{i=0}^r \binom{\nu}{i}.$$

Using the generalization of [41], define a generalized ν -variate Boolean function as a function f from \mathbb{Z}_{2^ν} to \mathbb{Z}_{2^h} for $h \geq 1$. As before, one may associate a sequence \mathbf{f} of length 2^ν by listing the values taken by f over \mathbb{Z}_{2^ν} in lexicographic order. Let the generalized² r th order Reed-Muller Code $\mathcal{RM}_{2^h}(r, \nu)$ be the subspace of \mathbb{Z}_{2^ν} spanned in $\mathbb{Z}_{2^h}^{2^\nu}$ by the sequences associated to monomial functions of degree at most r . Note, that while the sequences from $\mathcal{RM}_{2^h}(r, \nu)$ are sequences in \mathbb{Z}_{2^h} and not \mathbb{Z}_2 the generator matrix for $\mathcal{RM}_{2^h}(r, \nu)$ and $\mathcal{RM}(r, \nu)$ are the same.

Let the r th order code $\mathcal{L}_h^{(r, \nu)}$ be the span of the linear code $\mathcal{RM}(r-1, \nu)$ and twice the monomials of degree r over \mathbb{Z}_{2^h} . Clearly this is a code of length 2^ν and has cardinality

$$|\mathcal{L}_h^{(r, \nu)}| = 2^{h \dim(\mathcal{RM}(r-1, \nu))} \cdot 2^{(h-1) \binom{\nu}{r}}.$$

As before, we will consider the associated complex sequences and let $\mathcal{C}_h^{(r, \nu)} \subset \mathbb{C}^{2^\nu}$ be the associated set of complex sequences. We note that there are intimate connections between Reed-Muller codes (or more generally combinatorial designs) and the construction of perfect difference sets [15]. Hence, it is reasonable to expect that the code $\mathcal{L}_h^{(r, \nu)}$ has many sets with the same differences.

We note that the code $\mathcal{L}_h^{(r, \nu)}$ is related to some of the best known non-linear binary codes. That is, for $h = 2$ (i.e. for codes over \mathbb{Z}_4) these linear codes are in fact related to Kerdock and Preparata Codes via the Gray map [41, 55]. Note, that Kerdock and Preparata codes, while being non-linear over the binary field, have a large amount of structure. That is, the

²That these generalized Reed-Muller codes are *not* those of Delsarte et al. [45] as they come from a sequence in the integer ring \mathbb{Z}_{2^h} and not a field.

codes automorphism group contains a large set of permutations. In particular the Kerdock and Preperata Codes are invariant under affine permutation [3] (a notion we make more precise in the following section). Note that any permutation of the coordinate positions that fixes the code extends trivially to a unitary transformation for the associated complex quantization codebook. That is, the set of differences is not the only equivalence for sets of codewords. Hence, in the following section we consider quantizers that are images of affine invariant codes.

■ A.3 Affine-Invariant Constructions

In “classical” algebraic coding theory it was desirable to develop linear block codes with large automorphism group to aid in decoding. In particular, cyclic codes with large groups of permutations of the coordinate positions that leave the code fixed was of particular interest [84]. It can be shown that such codes reduce the complexity of encoders and decoders [127]. In a seminal work, Kasami, Lin and Peterson [74] characterized the necessary and sufficient conditions for a linear code over a finite field to be invariant under a large group permutations (the group of affine permutations). These results have been extended by Berger and Charpin [21,22] in a quite general way, which aids in the construction of codes with large permutation groups. It is this approach we take in the sequel. In particular, the generality of the results in [21, 22] have made the extensions to more general integer rings [4] and the broader class of Galois Rings [24, 46] amenable. We now develop the necessary results on affine invariant cyclic codes. We assume in the following that the base ring used in the construction of the code is a Galois ring $\text{GR}(p^\ell, r)$ for some finite ℓ , r and prime p .

We begin by reviewing the relevant concepts from cyclic codes that we require in the sequel. We refer the reader to [14, 15, 29] for a complete introduction. Recall that a cyclic code of length m over a ring \mathcal{R} is an ideal in the modular algebra $R_m = \mathcal{R}[X]/(X^m - 1)$. That is, if \mathcal{L} is a cyclic code of length m then for any $\alpha \in \mathcal{L}$ the map

$$(\alpha_0, \alpha_1, \dots, \alpha_{m-1}) \rightarrow \alpha_0 + \alpha_1 X + \dots + \alpha_{m-1} X^{m-1}$$

is an isomorphism between \mathcal{R}^m and R_m which identifies the cyclic code \mathcal{R} with an ideal in R_m . Recall that in the case that \mathcal{R} is a field, every cyclic code could be identified with a generator polynomial $g(x)$ which generates the ideal in R_m corresponding to the cyclic code. In the more general setting, a similar statement can be made [29].

Lemma A.3.1. *Every ideal of R_m is of the form*

$$(f_0, p f_1, p^2 f_2, \dots, p^{\ell-1} f_{\ell-1})$$

where the f_i are monic irreducible divisors of $X^m - 1$ in $\mathcal{R}[X]$ and $f_{\ell-1} \mid f_{\ell-2} \mid \dots \mid f_0$.

Thus, by Lemma A.3.1 every cyclic code can be characterized by the functions $f_0, f_1, \dots, f_{\ell-1}$. Moreover, as the f_i are monic irreducible divisors of $X^m - 1$ every cyclic code can be characterized by the roots of $f_0, f_1, \dots, f_{\ell-1}$. In this direction, let r' be a multiple of r such that the field $\mathbb{F}_{p^{r'}}$ contains a primitive m th root of unity, say ζ . Then, as the f_i are divisors of $X^m - 1$,

$$f_i = \prod_{j \in T_{i-1}} (X - \zeta^j)$$

for some set $T_{i-1} \subset \{0, 1, \dots, m-1\}$. We will call the collection $\{T_1, T_2, \dots, T_\ell\}$ the defining set of the cyclic code corresponding to $(f_0, pf_1, p^2f_2, \dots, p^{\ell-1}f_{\ell-1})$. Note, since $f_{i-1} \mid f_i$, one has $T_{i-1} \subset T_i$. We now review a standard representation of cyclic codes before discussing how defining sets can be used to characterize affine invariant cyclic codes.

Recall codewords of a cyclic code of length m can, via the Mattson-Solomon transform, be thought of as a listing of values taken on by a given function evaluated at every element of a finite field. More precisely, let $c(X)$ be a codeword of \mathcal{L} . Then, the Mattson-Solomon polynomial of $c(X)$ is

$$C(Z) = \sum_{i=0}^{m-1} \hat{c}(m-i)Z^i$$

where $\hat{c}(i) = c(\zeta^i)$ where ζ is a primitive m th root of unity. Then, using Fourier inversion [24, 103] one has

$$c_k = \frac{1}{m} C(\zeta^k).$$

That is, one can think of indexing the coordinate positions of codewords by elements of the finite field \mathbb{F}_{p^m} . Thus, any permutation that fixes the code can be described via permutations of the field elements that index the code. This is exactly the group algebra approach used by [21, 22, 74]. Recall, for $m = kt$, the group $\text{AGL}_k(p^t)$ acts on the field \mathbb{F}_{p^m} via affine linear transformations viewing the field \mathbb{F}_{p^m} as a k dimensional vectors space over \mathbb{F}_{p^t} . We say that a code \mathcal{L} is invariant under the group $\text{AGL}_k(p^t)$ if $\text{AGL}_k(p^t)$, acting on the coordinates of \mathcal{L} , fixes the code \mathcal{L} . More generally, we will say that such a code is affine invariant.

Defining sets are particularly useful in determining when a cyclic code is affine invariant. Indeed, this is exactly the result of [21, 22] which, in its generality, can be extended to codes over integer rings and more generally Galois rings. At present, we do not provide necessary and sufficient conditions for a cyclic code with defining sets $\{T_1, T_2, \dots, T_\ell\}$ to be affine invariant, but rather refer the reader to [3, 21, 22, 24, 46]. However, we note that if a code \mathcal{L} is invariant under the action of $\text{AGL}_k(p^t)$ then so is any set of codewords. More precisely, let, for any element $\sigma \in \text{AGL}_k(p^t)$, P_σ be the matrix representation of the permutation σ . Then, for any $\alpha_1, \alpha_2, \dots, \alpha_k \in \mathcal{L}$,

$$\begin{bmatrix} - & \alpha_1 & - \\ - & \alpha_2 & - \\ \vdots & \ddots & \vdots \\ - & \alpha_k & - \end{bmatrix} \cdot P_\sigma = \begin{bmatrix} - & \alpha'_1 & - \\ - & \alpha'_2 & - \\ \vdots & \ddots & \vdots \\ - & \alpha'_k & - \end{bmatrix}$$

for some $\alpha'_1, \alpha'_2, \dots, \alpha'_k \in \mathcal{L}$. Hence, the sets $\{\alpha_1, \alpha_2, \dots, \alpha_k\}$ and $\{\alpha'_1, \alpha'_2, \dots, \alpha'_k\}$ are equivalent for any regular reconstruction algorithm. We note that for any \mathcal{L} that is invariant under $\text{AGL}_k(p^t)$ there are at most $|\mathcal{L}|/|\text{AGL}_k(p^t)|$ distinct error values. Hence, it reasonable to expect that such a code will be quite stable to unmodded correlation in the measurement model as previously discussed. Moreover, as decoding and encoding of such codes is quite efficient it is reasonable to expect that the subset selection problem can also be solved efficiently. It is this question to which we now turn.

Bounds on SINR_{sat}

■ B.1 Bounds on SINR_{sat} without Order Statistics

It is well known that one may use the arguments of Shannon [109] to provide an upper bound on the mean square error of any quantization scheme. In particular, one can show [67, 144] that the distribution of the quantization error for every quantizer is upper bounded by

$$F_{\text{UB}}(x; r, m) \triangleq \begin{cases} 0 & \text{if } x > 1 - 2^{-r/(m-1)} \\ 2^{-r} \cdot (1 - x)^{m-1} & \text{o.w.} \end{cases}. \quad (\text{B.1})$$

More precisely, let X be a random variable distributed according to $F_{\text{UB}}(x)$. Then, for any rate r quantizer in \mathbb{C}^m , say \mathcal{C}_r , one has

$$\Pr \left[\|\tilde{\mathbf{h}}_i - \mathcal{Q}(\tilde{\mathbf{h}}_i)\| > 2 \cdot x \right] \leq 1 - F_{\text{UB}}(x; r, m).$$

Thus, X *stochastically dominates* the quantization error for any quantization scheme.

The distribution $F_{\text{UB}}(x; r, m)$ has a quite intuitive explanation that can be derived from [67, 109]. In particular, as $|\mathbf{h}_i^\dagger \mathbf{c}_j|$ is an increasing function of the angle between \mathbf{h}_i and \mathbf{c}_j the best shape a Voronoi region of codeword may take for a fixed volume is perfectly symmetric about the codeword. In particular, the mean squared quantization error incurred when a channel vector is quantized to a given codeword may be improved by shaping the Voronoi region to have the smallest second moment as possible by taking portions of the Voronoi region that lay the furthest from center of the Voronoi cell and moving them closer to center. Thus, for a rate r quantizer, the best possible scenario is to have 2^r Voronoi cells that are perfectly symmetric of equal volume which cover the surface of the complex m -sphere. As the channel vectors are assumed to be isotropic, such a Voronoi region contains all the points on the complex unit m -sphere such that

$$(1 - |\mathbf{u}^\dagger \mathbf{c}|)^{m-1} \leq 2^r.$$

Such a rate r code has 2^r congruent Voronoi regions,

$$\mathcal{V}_i^{\text{ub}}(r) = \left\{ \mathbf{u} : |\mathbf{u}^\dagger \mathbf{c}_i| \geq 1 - 2^{-r/(m-1)} \right\}.$$

Using this argument leads to (B.1). This is consistent with our previous example, Example 3-1. In particular, reexamining Figure 3-1 one can see that by reshaping the Voronoi regions of the quantizer depicted in Figure 3-1 (b), and hence necessarily moving the centers, one may arrive at the quantizer depicted in Figure 3-1 (a) which has a smaller second moment. As $|\mathbf{u}^\dagger \mathbf{c}_i| / (1 - |\mathbf{u}^\dagger \mathbf{c}_i|)$ is an increasing function in the inner product $|\mathbf{u}^\dagger \mathbf{c}_i|$ one may use (B.1)

to additionally upper bound SINR_{sat} . The Lemma 2.4.2 follows directly from computation of the integral

$$\int \frac{x}{1-x} dF_{\text{UB}}(x) \text{ and } \int \frac{x^2}{(1-x)^2} dF_{\text{UB}}(x).$$

■ B.2 Bounds on SINR_{sat} with Order Statistics

In the sequel, we let $\text{SINR}_{\text{sat}}^{\text{UB}}(n, \ell)$ be the expected SINR_{sat} for the ℓ best users in a n user system using a code distributed as in (B.1). As (B.1) stochastically dominates the distribution of the quantization error for any quantization scheme, it also stochastically dominates the order statistics [40]. More precisely, for any two random variables X and Y if

$$\Pr[X > x] \geq \Pr[Y > y]$$

then the distribution of the order statistics of any sequence of n *i.i.d* samples satisfies

$$\Pr[X_{(\ell)} > x] \geq \Pr[Y_{(\ell)} > y].$$

In order to derive exact expressions for $\text{SINR}_{\text{sat}}^{\text{UB}}(n, \ell)$ recall that the expected value for the k -th order statistics for a sample of n *i.i.d* random variables with a sample space $(0, 1)$ and distribution function $F(x)$ is [40]

$$\mu_{(k)} = k \binom{n}{k} \int_0^1 x [F(x)]^{k-1} [1 - F(x)]^{n-k} f(x) dx. \quad (\text{B.2})$$

Integrating above for the special case of (B.1) one has the following lemma.

Lemma B.2.1. *Consider a quantizer in which the distribution of the quantization error for each cell follows (B.1). Then,*

$$\mathbb{E} \left[\frac{\sigma_{(k)}}{1 - \sigma_{(k)}} \right] = -1 + 2^{\frac{r}{m-1}} \cdot \frac{\Gamma\left(\frac{m-2}{m-1} + n - k\right)}{\Gamma(1 + n - k)} \frac{\Gamma(1 + n)}{\Gamma\left(\frac{m-2}{m-1} + n\right)}$$

Further, for any rate r code,

$$\text{SINR}_{\text{sat}}(C_r; n, \ell) \leq \frac{1}{\ell} \sum_{i=n-\ell}^{n-1} -1 + 2^{\frac{r}{m-1}} \cdot \frac{\Gamma\left(\frac{m-2}{m-1} + n - i\right)}{\Gamma(1 + n - i)} \frac{\Gamma(1 + n)}{\Gamma\left(\frac{m-2}{m-1} + n\right)} \quad (\text{B.3})$$

We let,

$$\mu_{(k)}^{\text{UB}} = -1 + 2^{\frac{r}{m-1}} \cdot \frac{\Gamma\left(\frac{m-2}{m-1} + n - k\right)}{\Gamma(1 + n - k)} \frac{\Gamma(1 + n)}{\Gamma\left(\frac{m-2}{m-1} + n\right)} \quad (\text{B.4})$$

be the upper bound on the expected value of the order statistic of the $n - k$ -th best user in a n user system and let

$$\text{SINR}_{\text{sat}}^{\text{UB}}(n, \ell) = \frac{1}{\ell} \sum_{i=n-\ell}^{n-1} \mu_{(i)}^{\text{UB}}. \quad (\text{B.5})$$

To provide an upper bound on $\text{SINR}_{\text{sat}}^{\text{UB}}(n, \ell)$ that reveals the effects of increasing the number of users as well as the number of feedback bits we first require the following definition. Recall that the digamma function, $\psi(x)$, is defined to be the rate of the exponential growth of the Gamma function, i.e.

$$\psi(x) = \frac{d}{dx} \ln \Gamma(x).$$

We now have the following theorem.

Theorem B.2.2. *Consider any rate r quantization scheme. Then, for any integers $n > 0$ and $0 < \ell \leq n$,*

$$\text{SINR}_{\text{sat}}(n, \ell) \leq \text{SINR}_{\text{sat}}^{\text{UB}}(n, \ell)$$

Further,

$$\text{SINR}_{\text{sat}}^{\text{UB}}(n, \ell) \leq \frac{2^{\frac{r}{m-1}}}{\ell} \frac{\Gamma(1+n)}{\Gamma\left(n + \frac{m-2}{m-1}\right)} \left(\Gamma\left(\frac{m-2}{m-1}\right) + \sum_{i=n-\ell+1}^{n-1} (n-i)^{\frac{-1}{m-1}} \right) \quad (\text{B.6a})$$

$$\leq 2^{\frac{r}{m-1}} \exp\left(\frac{\psi(1+n)}{m-1}\right) \left(\frac{1}{\ell} \Gamma\left(\frac{m-2}{m-1}\right) + \frac{\ell-1}{\ell} \right) \quad (\text{B.6b})$$

Proof. First follows from stochastic domination order statistics (B.2). The following sequence of bounds follows from applying both Kershaw's upper and lower bounds on the ratio of Gamma functions [101] for $n \in \mathbb{Z}^+$ and $0 < s < 1$,

$$\exp((s-1)\psi(1+n)) \leq \frac{\Gamma(n+1)}{\Gamma(n+s)} \leq n^{s-1}$$

■

Examining Theorem B.2.2 reveals quite a lot about the limits in $\text{SINR}_{\text{sat}}^{\text{UB}}(n, \ell)$ in terms of both the quantizer rate as well as the number of users. In particular, the growth in dB is linear in the quantizer rate with slope independent of the number of users as well as linear in $\psi(1+n)$ where the slope is independent of the quantizer rate r . Using the asymptotic expansion for the digamma function one further has

$$\psi(1+x) \sim \ln x + \frac{1}{2x} + O\left(\frac{1}{2x}\right).$$

Hence, for large user populations

$$\begin{aligned} 10 \log_{10} \text{SINR}_{\text{sat}}(n, \ell) &\leq \frac{10 \cdot \log_{10} 2}{m-1} \cdot r + \frac{10 \cdot \log_{10} e}{m-1} \cdot \ln(n)(1 + o(1)) \\ &\quad + 10 \cdot \log_{10} \left(1 + \frac{1}{\ell} \Gamma\left(\frac{m-2}{m-1}\right) \right) \end{aligned}$$

Thus,

$$10 \log_{10} \text{SINR}_{\text{sat}}^{\text{UB}}(n, \ell) \approx \frac{3}{m-1} \cdot r + \frac{3}{m-1} \cdot \log_2 n + C(\ell, m)$$

for some constant $C(\ell, m)$ which does not depend on n or r . Thus, in a multi-user system doubling the size of the user pool has roughly the same effect of adding a bit of feedback

using the optimal quantization scheme. We note that this still does not address the last question we have concerning the achieved SINR of a system. In particular, the definition of SINR_{sat} assumes that there is a set of nearly orthogonal users.

■ C.1 Proofs for Chapter 2

■ C.1.1 Proof of Equation (2.36)

$$\begin{aligned}
 c_i(\mathcal{A}) &= \frac{|\sigma_i - \sigma_{i,\mathcal{A}} \mathbf{R}_{\mathcal{A} \setminus i}^{-1} \mu_{i,\mathcal{A}}^\dagger|}{1 - \mu_{i,\mathcal{A}} \mathbf{R}_{\mathcal{A} \setminus i}^{-1} \mu_{i,\mathcal{A}}^\dagger} \\
 &= \frac{|\sigma_i - \mathbf{h}_i \mathbf{W}_{i,\mathcal{A}}^\dagger \mathbf{R}_{\mathcal{A} \setminus i}^{-1} \mu_{i,\mathcal{A}}^\dagger|}{1 - \mu_{i,\mathcal{A}} \mathbf{R}_{\mathcal{A} \setminus i}^{-1} \mu_{i,\mathcal{A}}^\dagger} \\
 &= \frac{|\sigma_i - (\sigma_i \mathbf{w}_i + \mathbf{h}_i^\perp) \mathbf{W}_{i,\mathcal{A}}^\dagger \mathbf{R}_{\mathcal{A} \setminus i}^{-1} \mu_{i,\mathcal{A}}^\dagger|}{1 - \mu_{i,\mathcal{A}} \mathbf{R}_{\mathcal{A} \setminus i}^{-1} \mu_{i,\mathcal{A}}^\dagger} \\
 &= \left| \sigma_i - \frac{\mathbf{h}_i^\perp \mathbf{W}_{i,\mathcal{A}}^\dagger \mathbf{R}_{\mathcal{A} \setminus i}^{-1} \mu_{i,\mathcal{A}}^\dagger}{1 - \mu_{i,\mathcal{A}} \mathbf{R}_{\mathcal{A} \setminus i}^{-1} \mu_{i,\mathcal{A}}^\dagger} \right|
 \end{aligned}$$

where \mathbf{h}_i^\perp is the component of \mathbf{h}_i that is orthogonal to \mathbf{w}_i . Continuing, we have

$$\begin{aligned}
 c_i(\mathcal{A}) &\geq |\sigma_i| - \frac{|\mathbf{h}_i^\perp \mathbf{W}_{i,\mathcal{A}}^\dagger \mathbf{R}_{\mathcal{A} \setminus i}^{-1} \mu_{i,\mathcal{A}}^\dagger|}{1 - \mu_{i,\mathcal{A}} \mathbf{R}_{\mathcal{A} \setminus i}^{-1} \mu_{i,\mathcal{A}}^\dagger} \\
 &\geq |\sigma_i| - \frac{\sqrt{\|\mathbf{h}_i\|^2 - |\sigma_i|^2} \|\mathbf{W}_{i,\mathcal{A}}^\dagger \mathbf{R}_{\mathcal{A} \setminus i}^{-1} \mu_{i,\mathcal{A}}^\dagger\|}{1 - \mu_{i,\mathcal{A}} \mathbf{R}_{\mathcal{A} \setminus i}^{-1} \mu_{i,\mathcal{A}}^\dagger} \\
 &\geq |\sigma_i| - \frac{\sqrt{\|\mathbf{h}_i\|^2 - |\sigma_i|^2}}{1 - \mu_{i,\mathcal{A}} \mathbf{R}_{\mathcal{A} \setminus i}^{-1} \mu_{i,\mathcal{A}}^\dagger}
 \end{aligned}$$

Thus, since $c_i(\mathcal{A}) \geq 0$ we have

$$c_i(\mathcal{A}) \geq \left(|\sigma_i| - \frac{\sqrt{\|\mathbf{h}_i\|^2 - |\sigma_i|^2}}{1 - \mu_{i,\mathcal{A}} \mathbf{R}_{\mathcal{A} \setminus i}^{-1} \mu_{i,\mathcal{A}}^\dagger} \right)_+$$

■ C.1.2 Proof of Quantized Channel Rates

Note that under the covariance constraint $\mathbb{E} [\text{Tr}(\mathbf{x}\mathbf{x}^\dagger)] \leq P$ we have

$$\mathbb{E} \left[\mathcal{Q}(\mathbf{H}_A)^+ \mathbf{u}\mathbf{u}^\dagger (\mathcal{Q}(\mathbf{H}_A)^+)^{\dagger} \right] = \mathbb{E} [\text{Tr}(\mathbf{R}_A^{-1})]$$

Thus, taking $P_i = P/\text{Tr}(\mathbf{R}_A^{-1})$ yields a valid power allocation. Now, since the channel is modeled using the standard input/output model (2.8), we have

$$\mathbf{y} = \mathbf{H}_A^\dagger \mathcal{Q}(\mathbf{H}_A) \left(\mathcal{Q}(\mathbf{H}_A)^\dagger \mathcal{Q}(\mathbf{H}_A) \right)^{-1} \mathbf{u}_A + \mathbf{n} \quad (\text{C.1})$$

With out loss of generality consider the signal recieved by user 1. Then, using the inverse of a partitioned matrix [58],

$$\mathbf{h}_1^\dagger \mathbf{x} = \mathbf{h}_1^\dagger \mathcal{Q}(\mathbf{H}_A) \left(\mathcal{Q}(\mathbf{H}_A)^\dagger \mathcal{Q}(\mathbf{H}_A) \right)^{-1} \mathbf{u}_A \quad (\text{C.2})$$

$$= \left[\mathbf{h}_1^\dagger \mathcal{Q}(\mathbf{h}_1) \ \sigma_{1,A} \right] \left[\begin{array}{c|c} 1 & \mu_{1,A} \\ \hline \mu_{1,A}^\dagger & \mathbf{R}_{A \setminus 1} \end{array} \right]^{-1} \mathbf{u}_A \quad (\text{C.3})$$

$$= \left[\mathbf{h}_1^\dagger \mathcal{Q}(\mathbf{h}_1) \ \sigma_{1,A} \right] \left[\begin{array}{c|c} \left(1 - \mu_{1,A} \mathbf{R}_{A \setminus 1}^{-1} \mu_{1,A}^\dagger \right)^{-1} & \mu_{1,A} \left(\mu_{1,A}^\dagger \mu_{1,A} - \mathbf{R}_{A \setminus 1} \right)^{-1} \\ \hline \left(\mu_{1,A}^\dagger \mu_{1,A} - \mathbf{R}_{A \setminus 1} \right)^{-1} & \mu_{1,A}^\dagger \end{array} \right] \left(\mathbf{u}_A \right)$$

Now, using the formula for the inverse of a matrix with a small rank adjustment [58], we have

$$\begin{aligned} \left(\mu_{1,A}^\dagger \mu_{1,A} - \mathbf{R}_{A \setminus 1} \right)^{-1} \mu_{1,A}^\dagger &= -\mathbf{R}_{A \setminus 1}^{-1} \mu_{1,A}^\dagger - \frac{\mathbf{R}_{A \setminus 1}^{-1} \mu_{1,A}^\dagger \mu_{1,A} \mathbf{R}_{A \setminus 1}^{-1} \mu_{1,A}^\dagger}{1 - \mu_{1,A} \mathbf{R}_{A \setminus 1}^{-1} \mu_{1,A}^\dagger} \\ &= -\frac{\mathbf{R}_{A \setminus 1}^{-1} \mu_{1,A}^\dagger}{1 - \mu_{1,A} \mathbf{R}_{A \setminus 1}^{-1} \mu_{1,A}^\dagger} \end{aligned}$$

Thus, we may write

$$\begin{aligned} \mathbf{h}_1^\dagger \mathbf{x} &= \left(\frac{\mathbf{h}_1^\dagger \mathcal{Q}(\mathbf{h}_1)}{1 - \mu_{1,A} \mathbf{R}_{A \setminus 1}^{-1} \mu_{1,A}^\dagger} + \sigma_{1,A} \left(-\frac{\mathbf{R}_{A \setminus 1}^{-1} \mu_{1,A}^\dagger}{1 - \mu_{1,A} \mathbf{R}_{A \setminus 1}^{-1} \mu_{1,A}^\dagger} \right) \right) u_1 \\ &\quad + \left(\sigma_{1,A} - \mathbf{h}_1^\dagger \mathcal{Q}(\mathbf{h}_1) \mu_{1,A} \right) \left(\mathbf{R}_{A \setminus 1} - \mu_{1,A}^\dagger \mu_{1,A} \right)^{-1} \mathbf{u}_{A \setminus 1} \\ &= \left(\frac{\mathbf{h}_1^\dagger \mathcal{Q}(\mathbf{h}_1) - \sigma_{1,A} \mathbf{R}_{A \setminus 1}^{-1} \mu_{1,A}^\dagger}{1 - \mu_{1,A} \mathbf{R}_{A \setminus 1}^{-1} \mu_{1,A}^\dagger} \right) u_1 + \left(\sigma_{1,A} - \mathbf{h}_1^\dagger \mathcal{Q}(\mathbf{h}_1) \mu_{1,A} \right) \left(\mathbf{R}_{A \setminus 1} - \mu_{1,A}^\dagger \mu_{1,A} \right)^{-1} \mathbf{u}_{A \setminus 1} \end{aligned}$$

which yields the result.

■ C.2 Proofs for Chapter 3

■ C.2.1 Proof of Lemma 3.3.1

Note that the sum in the right hand side of (3.13) is clearly a linear function of both λ and β . Thus, it is left to show that the lifting of $\bar{\beta}$ to β is a linear function of $\bar{\beta}$. This result will follow from our more general discussion in Section 3.4.

■ C.2.2 Proof of Lemma 3.3.5

To begin, note that the two codewords $\mathbf{c}(\lambda, \bar{\beta}; L, p^a)$ and $\mathbf{c}(\lambda', \bar{\beta}'; L, p^a)$ only have the same support if $\bar{\beta} + L$ and $\bar{\beta}' + L$ define the same coset of L . Thus, if $\bar{\beta} - \bar{\beta}' \notin L$ the codewords $\mathbf{c}(\lambda, \bar{\beta}; L, p^a)$ and $\mathbf{c}(\lambda', \bar{\beta}'; L, p^a)$ do not have the same support and hence can not be colinear. Thus, we now suppose that $\bar{\beta} - \bar{\beta}' \in L$, i.e. the codewords $\mathbf{c}(\lambda, \bar{\beta}; L, p^a)$ and $\mathbf{c}(\lambda', \bar{\beta}'; L, p^a)$ have the same support. To show the if part of the lemma note that if $\langle \lambda - \lambda', \bar{\gamma} \rangle = k$ and $\bar{\beta} - \bar{\beta}' \in L$

$$\mathbf{c}(\lambda, \bar{\beta}; L, p^a) = \sum_{\gamma \in L} \zeta_p^{\langle \lambda, \bar{\gamma} \rangle} \mathbf{e}_{\bar{\gamma} + \bar{\beta}} \quad (\text{C.5a})$$

$$= \sum_{\gamma \in L} \zeta_p^{\langle \lambda', \bar{\gamma} \rangle} \zeta_p^{\langle \lambda - \lambda', \bar{\gamma} \rangle} \mathbf{e}_{\bar{\gamma} + \bar{\beta}} \quad (\text{C.5b})$$

$$= \zeta_p^k \sum_{\gamma \in L} \zeta_p^{\langle \lambda', \bar{\gamma} \rangle} \mathbf{e}_{\bar{\gamma} + (\bar{\beta} - \bar{\beta}') + \bar{\beta}'} \quad (\text{C.5c})$$

$$= \zeta_p^k \sum_{\tilde{\gamma} \in L} \zeta_p^{\langle \lambda', \tilde{\gamma} - (\bar{\beta} - \bar{\beta}') \rangle} \mathbf{e}_{\tilde{\gamma} + \bar{\beta}'} \quad (\text{C.5d})$$

$$= \zeta_p^{k - \langle \lambda', (\bar{\beta} - \bar{\beta}') \rangle} \sum_{\tilde{\gamma} \in L} \zeta_p^{\langle \lambda', \tilde{\gamma} \rangle} \mathbf{e}_{\tilde{\gamma} + \bar{\beta}'} \quad (\text{C.5e})$$

$$= \zeta_p^{k - \langle \lambda', (\bar{\beta} - \bar{\beta}') \rangle} \mathbf{c}(\lambda', \bar{\beta}'; L, p^a) \quad (\text{C.5f})$$

For the only if part of the lemma note that if $\bar{\beta} - \bar{\beta}' \in L$ then

$$\zeta_p^{-\langle \lambda', (\bar{\beta} - \bar{\beta}') \rangle} \mathbf{c}(\lambda', \bar{\beta}'; L, p^a) = \sum_{\gamma \in L} \zeta_p^{\langle \lambda', \bar{\gamma} \rangle} \mathbf{e}_{\bar{\gamma} + (\bar{\beta} - \bar{\beta}') + \bar{\beta}'} \quad (\text{C.6a})$$

$$= \sum_{\gamma \in L} \zeta_p^{\langle \lambda', \bar{\gamma} \rangle} \mathbf{e}_{\bar{\gamma} + \bar{\beta}} \quad (\text{C.6b})$$

Note that (C.6b) is only a complex multiple of (C.5a) if $\langle \lambda - \lambda', \bar{\gamma} \rangle$ is constant for all $\gamma \in L$. However, as L is a sub-space of $(\mathbb{Z}_p)^{m'}$, we have $0 \in L$ and $k = 0$.

■ C.2.3 Proof of Lemma 3.3.9

This is a direction result of elementary character theory [63] or equivalently Fourier Analysis on groups [92].

■ C.2.4 Proof of Theorem 3.3.10

We note that the sufficiency of the conditions of the theorem follow immediately from the discussions preceding it. That is, if the twisted hamming weight is greater than zero than one may marginalize over a coordinate and produce a zero sum. To see that this is necessary

suppose that there is an element λ such that $\text{twt}_H(\lambda) = 0$ and $\Gamma_C(\lambda; \{0\}, L) = 0$. Then, for some j

$$\sum_{x_{i_j}=0}^{p-1} \zeta_{p^a}^{(\hat{a}_j+p^{a-1}\cdot\bar{a}_j)\cdot x_j} \left(\sum_{x_{i_0}=0}^{p-1} \sum_{x_{i_1}=0}^{p-1} \cdots \sum_{x_{i_{j-1}}=0}^{p-1} \sum_{x_{i_{j+1}}=0}^{p-1} \cdots \sum_{x_{i_{d-1}}=0}^{p-1} \zeta_{p^a}^{(\bar{\mathbf{a}}, \bar{\mathbf{x}})} \mathbf{e}_{\mathbf{x}+\bar{\beta}} \right) = 0$$

where

$$\sum_{x_{i_j}=0}^{p-1} \zeta_{p^a}^{(\hat{a}_j+p^{a-1}\cdot\bar{a}_j)\cdot x_j} \neq 0$$

as $\text{twt}_H(\lambda) = 0$. Hence,

$$\sum_{x_{i_0}=0}^{p-1} \sum_{x_{i_1}=0}^{p-1} \cdots \sum_{x_{i_{j-1}}=0}^{p-1} \sum_{x_{i_{j+1}}=0}^{p-1} \cdots \sum_{x_{i_{d-1}}=0}^{p-1} \zeta_{p^a}^{(\bar{\mathbf{a}}, \bar{\mathbf{x}})} \mathbf{e}_{\mathbf{x}+\bar{\beta}} = 0.$$

Thus, as $\text{twt}_H(\lambda) = 0$, for some j' we can marginalize out one coordinate where the multivariate sum is 0 while the outer sum is non-zero. Thus, proceeding recursively one has

$$\sum_{x_{i_{d-1}}=0}^{p-1} \zeta_{p^a}^{(a_{j_0}, x_{j_0})} \mathbf{e}_{\mathbf{x}+\bar{\beta}} = 0.$$

However, this sum is zero if and only if $x_{j_0} = p^{a-1}x'$ for some $x' \neq 0$ which implies $\text{twt}_H(\lambda) > 0$ which is a contradiction.

■ C.2.5 Proof of Corollary 3.3.12

The fact that Υ_1 is closed under addition modulo p^a follows directly from the fact that $\hat{\Upsilon}_1$ is closed under addition modulo p^{a-1} and $p^{a-1} \cdot L_1^d$ contains every element of the form $p^a \cdot \lambda$. In particular, consider two general elements of $\hat{\Upsilon}_1 + p^{a-1} \cdot L_1^d$, say $\lambda_1 = \hat{\lambda}_1 + p^{a-1} \cdot \bar{\lambda}_1$ and $\lambda_2 = \hat{\lambda}_2 + p^{a-1} \cdot \bar{\lambda}_2$. Then, either $p \cdot (\hat{\lambda}_1 + \hat{\lambda}_2) = 0$ and $\hat{\lambda}_1 + \hat{\lambda}_2 \in p^{a-1} \cdot L_1^d$ thus $\lambda_1 + \lambda_2 \in p^{a-1} \cdot L_1^d$. Otherwise, $\hat{\lambda}_1 + \hat{\lambda}_2 \in \hat{\Upsilon}_1$, as $\hat{\Upsilon}_1$ is closed under addition modulo p^{a-1} and $\lambda_1 + \lambda_2 \in \hat{\Upsilon}_1 + p^{a-1} \cdot L_1^d$.

■ C.2.6 Proof of Theorem 3.3.13

The proof of this statement is a simple consequence of the discussions preceding it. Note, by Theorem 3.3.10 every basis must have a collection of codewords which satisfy conditions (i), (ii) or (iii). We note that as multiplication by $T\lambda$ yields an orthogonal set then so will $R(\lambda, \beta)$ as this operation preserves the twisted hamming weight for the elements with support on $\beta + L$. Thus, the vectors supported on $\beta + L$ will remain orthogonal using condition (iii). Moreover, as $R(\lambda, \beta)$ leaves the elements with support which does not intersect $\beta + L$ fixed this set will remain orthogonal. Thus, we are left to check that the elements with non-intersecting supports are orthogonal. However, this is trivial and the image of any orthogonal basis contained in the code is again orthogonal. To see that this image again is contained in the code we note that $R(\lambda, \beta)$ acts linearly on the set of λ which define the code and hence, by the linearity of Υ_1 is again in Υ_1 and hence an element of the code. We note that the image of multiplication by $S\bar{\gamma}$ is trivially again in the code as

L is linear. Hence, every such product is again in the code and

$$\left\langle S(\bar{\gamma}) \cdot R(\lambda_{\bar{\beta}}; \bar{\beta}) \mid R(\lambda_{\bar{\beta}}; \bar{\beta}) \in \mathcal{R}_L(\hat{\Upsilon}_1) \text{ and } \bar{\gamma} \in L^c \right\rangle$$

acts transitively on the code $\mathcal{C}(\Upsilon_1, L^c; L)$ as well as the collection of orthogonal bases contained in $\mathcal{C}(\Upsilon_1, L^c; L)$.

■ C.2.7 Proof of Lemma 3.4.10

Without loss of generality assume that $r \in p^{e-i-1}\text{GR}(p^{i+1}, m') \setminus p^{e-i}\text{GR}(p^i, m')$ as $p^{e-i}\text{GR}(p^i, m') \subset p^{e-i-1}\text{GR}(p^{i+1}, m')$ and $\Gamma_{\mathbb{C}}(r; p, i) > 0$ for all $p^{e-i}\text{GR}(p^i, m')$ by assumption. Thus,

$$r = p^{e-i-1} \cdot \zeta + p^{e-i} \cdot r_0 \quad (\text{C.7})$$

for some $\zeta \in \mathcal{T}_{p^e, m'}$ and $r_0 \in p^{e-i}\text{GR}(p^i, m')$. Suppose, in order to arrive at a contradiction, there is some $r \in p^{e-i-1}\text{GR}(p^{i+1}, m') \setminus p^{e-i}\text{GR}(p^i, m')$ such that $\Gamma_{\mathbb{C}}(r; p, i) = 0$. Then, there exists some basis for $p^{e-i-1}\text{GR}(p^{i+1}, m')$ over \mathbb{Z}_{p^e} , say \mathcal{B} , such that

$$\{\text{Tr}(r \cdot s_i)\}_{r_i \in \mathcal{B}}$$

is a (coset) of a subgroup of \mathbb{Z}_{p^e} by elementary character theory [63]. That is, the elements of the vector

$$\mathbf{v} = [\text{Tr}(r \cdot s_0), \text{Tr}(r \cdot s_2), \dots, \text{Tr}(r \cdot s_{m-1})]$$

form a (coset of a) subgroup of \mathbb{Z}_{p^e} . However, from (C.7) $r = p^{e-i-1} \cdot \zeta + p^{e-i} \cdot r_0$ and hence

$$\mathbf{v} = p^{e-i-1} \cdot [r_0, r_1, \dots, r_{m-1}] + p^{e-i} \cdot [\bar{r}_0, \bar{r}_1, \dots, \bar{r}_{m-1}]$$

where $r_i \in p^{e-i-1}\mathbb{Z}_{p^e}$ and $\bar{r}_i \in p^{e-i}\mathbb{Z}_{p^e}$. However, if the elements of \mathbf{v} form a (coset of a) subgroup of \mathbb{Z}_{p^e} then so must $p \cdot \mathbf{v}$. Moreover,

$$p \cdot \mathbf{v} = [\text{Tr}(\bar{r} \cdot s_0), \text{Tr}(\bar{r} \cdot s_2), \dots, \text{Tr}(\bar{r} \cdot s_{m-1})]$$

for some $\bar{r} \in p^{e-i}\text{GR}(p^i, m')$. Namely,

$$\bar{r} = \sum_{j=0}^{m-1} (p^{e-i}r_j + p^{e-i+1}\bar{r}_j) t_j$$

where t_0, t_1, \dots, t_{m-1} is a trace dual basis to \mathcal{B} , i.e.

$$\text{Tr}(s_j \cdot t_i) = \delta(i - j)$$

where $\delta(x) = 1$ if and only if $x = 0$. However, $\bar{r} \in p^{e-i}\text{GR}(p^i, m')$ and $\Gamma_{\mathbb{C}}(\bar{r}; p, i) = 0$, a contradiction. Hence,

$$\Gamma_{\mathbb{C}}(r; p, i) > 0$$

for all $r \in p^{e-i-1}\text{GR}(p^{i+1}, m')$.

■ C.2.8 Proof of Theorem 3.4.13

The proof of the first part of the theorem is trivial from the definition of $\vartheta_i(\zeta^j)$. The fact that $\vartheta_i(\zeta^j) \equiv \bar{\zeta}^i \pmod{p}$ follows simply from recalling that reduction modulo p defines

a homomorphism between $\text{GR}(p^a, m')$ and \mathbb{Z}_{p^a} . As $\mu \circ \vartheta_i(\mathcal{T}_{p^a, m'}) = \mathbb{F}_{p^a}$, $\vartheta_i(\zeta^j)$ must be injective. In fact, $\vartheta_i(\zeta^j)$ is an injective map from $\mathcal{T}_{p^a, m'}$ into the unit group of $\text{GR}(p^a, m')$ (see [85] for further details on the unit group of Galois Rings). In particular, from [85] one has for $p = 2$ and $a \leq 2$ or $p > 2$ for any free basis $\{b_i\}$ of $\text{GR}(p^a, m')$,

$$\text{GR}^*(p^a, m') = \left\{ \zeta^\ell \cdot \prod_{i=0}^{m-1} (1 + pb_i)^{n_i} \mid \ell \in \{0, 1, \dots, p^{m'-1}\} n_i \in \{0, 1, \dots, p^{a-1}\} \right\}.$$

Now,

$$\begin{aligned} \vartheta_{\mathcal{I}}(x) &= x \prod_{j=1}^i \left(1 + p^{a-1} \zeta^{p^j} \text{Tr}(x \zeta^{p^j}) \right) \\ &= x \prod_{j=1}^i \left(1 + p \zeta^{p^j} \right)^{p^{a-2} \text{Tr}(x \zeta^{p^j})} \end{aligned}$$

As x runs over $\mathcal{T}_{p^a, m'}$, $p^{a-2} \text{Tr}(x \zeta^{p^j}) \in \{0, p^{a-2}, 2 \cdot p^{a-2}, (p-1) \cdot p^{a-2}\}$ equally many times for each class as $\mathcal{T}_{p^a, m'}$ is congruent to $\mathbb{F}_{p^{m'}}$ modulo p . Now, define addition via $\oplus|_{\vartheta}$, as

$$\vartheta_{\mathcal{I}}(x) + \vartheta_{\mathcal{I}}(y) = \vartheta_{\mathcal{I}}(\mu^{-1}(\mu(x+y))).$$

With this law, $-x = \mu^{-1}(-\mu(x))$ and is unique. To see this defines a group law note

- (i) $\vartheta_{\mathcal{I}}(x) \oplus|_{\vartheta} 0 = \vartheta_{\mathcal{I}}(x)$
- (ii) $\vartheta_{\mathcal{I}}(x) \oplus|_{\vartheta} \mu^{-1}(-x) = 0$
- (iii) $\vartheta_{\mathcal{I}}(x) \oplus|_{\vartheta} \vartheta_{\mathcal{I}}(y) = \vartheta_{\mathcal{I}}(y) \oplus|_{\vartheta} \vartheta_{\mathcal{I}}(x)$
- (iv) $\mu^{-1}(\mu(x+y)) \in \mathcal{T}_{p^a, m'}$ and the image of $\vartheta_{\mathcal{I}}(\cdot)$ is closed

and lastly note

$$\begin{aligned} (\vartheta_{\mathcal{I}}(x) \oplus|_{\vartheta} \vartheta_{\mathcal{I}}(y)) + \vartheta_{\mathcal{I}}(z) &= \vartheta_{\mathcal{I}}(\mu^{-1}(\mu(x+y))) + \vartheta_{\mathcal{I}}(z) \\ &= \vartheta_{\mathcal{I}}(\mu(\mu^{-1}(\mu(x+y)) + z)) \\ &= \vartheta_{\mathcal{I}}(\mu(x+y) + \mu(z)) \\ &= \vartheta_{\mathcal{I}}(\mu(x) + \mu(y+z)) \\ &= \vartheta_{\mathcal{I}}(\mu(\mu^{-1}(\mu(z+y)) + x)) \\ &= \vartheta_{\mathcal{I}}(\mu^{-1}(\mu(z+y))) + \vartheta_{\mathcal{I}}(x) \\ &= (\vartheta_{\mathcal{I}}(z) \oplus|_{\vartheta} \vartheta_{\mathcal{I}}(y)) + \vartheta_{\mathcal{I}}(x) \end{aligned}$$

As this defines a group law on the image of $\mathcal{T}_{p^a, m'}$, we extend this map linearly on $\text{GR}(p^a, m')$ via the p -adic expansion of every element. That is, we let

$$r \oplus s = \sum_{i=0}^{a-1} p^i \cdot (r_i \oplus|_{\vartheta} s_i)$$

where $r_i, s_i \in \mathcal{T}_{p^a, m'}$ and

$$r = \sum_{i=0}^{a-1} p^i r_i \text{ and } s = \sum_{i=0}^{a-1} p^i s_i$$

is the p -adic expansion of r and s .

■ C.2.9 Proof of Theorem 3.4.14

$$\text{Tr}(y \cdot \vartheta_{\mathcal{I}}(x)) = \sum_{i=0}^{m'-1} \text{Tr}(x \zeta^{p^i}) \text{Tr}(y \cdot \vartheta_{\mathcal{I}}(x)/x \cdot \zeta_{\perp}^{p^i}) \quad (\text{C.8a})$$

$$= \sum_{i=0}^{m'-1} x_i \cdot \text{Tr}\left(y \cdot \zeta_{\perp}^{p^i} \cdot \prod_{j \in \mathcal{I}} (1 + p^{a-1} \zeta^{p^j} x_j)\right) \quad (\text{C.8b})$$

$$= \sum_{i=0}^{m'-1} x_i \cdot \text{Tr}\left(y \cdot \zeta_{\perp}^{p^i} \cdot \left(1 + p^{a-1} \sum_{j \in \mathcal{I}} \zeta^{p^j} \cdot x_j\right)\right) \quad (\text{C.8c})$$

$$= \sum_{i=0}^{m'-1} x_i \text{Tr}\left((\hat{y} + \bar{y}) \cdot \zeta_{\perp}^{p^i} \cdot \left(1 + p^{a-1} \sum_{j \in \mathcal{I}} \zeta^{p^j} \cdot x_j\right)\right) \quad (\text{C.8d})$$

$$= \sum_{i=0}^{m'-1} x_i \text{Tr}\left(\hat{y} \cdot \zeta_{\perp}^{p^i}\right) \quad (\text{C.8e})$$

$$+ p^{a-1} \sum_{i=0}^{m'-1} x_i \text{Tr}\left(y_{a-1} \cdot \zeta_{\perp}^{p^i} + y_0 \sum_{j \in \mathcal{I}} \zeta_{\perp}^{p^i} \zeta^{p^j} \cdot x_j\right) \quad (\text{C.8f})$$

where we have let $x_i = \text{Tr}(x \zeta^{p^i})$ be the expansion of x in terms of the *normal basis* while we let y be expanded through the p -adic representation. That is, suppose

$$y = \sum_{i=0}^{a-1} p^i \cdot y_i.$$

Then,

$$\hat{y} = \sum_{i=0}^{a-2} p^i \cdot y_i$$

and $\bar{y} = p^{a-1} y_{a-1}$ so that $y = \hat{y} + \bar{y}$. We note that (C.8d) and (C.8f) is now quite familiar. That is, by expanding y in terms of the dual basis one has in (C.8d) the inner product between a vector determining the coordinate set and $\downarrow L_a^d$. However, we are in an unfortunate position in (C.8f). That is (C.8f) has a large mixture of variables. However, we note that if $y_0 \in 0, 1$ then (C.8f) becomes

$$\text{Tr}\left(y_{a-1} \cdot \zeta_{\perp}^{p^i} + y_0 \sum_{j \in \mathcal{I}} \zeta_{\perp}^{p^i} \zeta^{p^j} \cdot x_j\right) = \text{Tr}\left(y_{a-1} \cdot \zeta_{\perp}^{p^i}\right) + \sum_{j \in \mathcal{I}} \delta(i-j) x_j$$

Hence, if $y_0 \in 0, 1$

$$\text{Tr}(y \cdot \vartheta_{\mathcal{I}}(x)) = \sum_{i=0}^{m'-1} x_i \text{Tr}(\hat{y} \cdot \zeta_{\perp}^{p^i}) + p^{a-1} \sum_{i=0}^{m'-1} x_i \left(\text{Tr}(y_{a-1} \cdot \zeta_{\perp}^{p^i}) + \mathbf{1}_{\{i \in \mathcal{I}\}} \cdot x_i \right) \quad (\text{C.9})$$

Thus, the map $\vartheta_{\mathcal{I}}(x)$ allows us to marginalize once again provided $\mu y \in \{0, 1\}$.

■ C.2.10 Proof of Lemma 3.6.1

We note that this is easily computed by examining the action of $\mathbf{F}(\mathbf{b}_0; \alpha, \gamma, \mathcal{B})$ on the basis \mathcal{B} . First, note that

$$\gamma \cdot \mathbf{b}_0 = \mathbf{F}(\mathbf{b}_0; \alpha, \gamma, \mathcal{B}) \cdot \mathbf{b}_0.$$

Further, for $\mathbf{b} \in \mathcal{B} \setminus \mathbf{b}_0$,

$$\alpha\gamma\mathbf{b}_0 + \sqrt{1 - \alpha^2}\mathbf{b} = \mathbf{F}(\mathbf{b}_0; \alpha, \gamma, \mathcal{B}) \cdot \mathbf{b}.$$

Hence, as \mathcal{B} is an orthonormal basis we have the result.

■ C.2.11 Proof of Lemma 3.6.2

This can be by direct computation. First, note that image of \mathcal{B} has a non-zero inner product with \mathbf{b}_0 . Further,

$$\gamma \cdot \mathbf{b}_0 = \mathbf{F}(\mathbf{b}_0; \alpha, \gamma, \mathcal{B}) \cdot \mathbf{b}_0.$$

so \mathbf{b}_0 is an eigenvector of $\mathbf{F}(\mathbf{b}_0; \alpha, \gamma, \mathcal{B})$. Now, consider the vector $\mathbf{b} + \nu \cdot \mathbf{b}_0$. Then,

$$\nu\gamma \cdot \mathbf{b}_0 + \alpha\gamma\mathbf{b}_0 + \sqrt{1 - \alpha^2}\mathbf{b} = \mathbf{F}(\mathbf{b}_0; \alpha, \gamma, \mathcal{B}) \cdot (\mathbf{b} + \nu \cdot \mathbf{b}_0).$$

Hence, for $\mathbf{b} + \nu \cdot \mathbf{b}_0$ to be an eigenvector one must have

$$\frac{\nu\gamma + \alpha\gamma}{\sqrt{1 - \alpha^2}} = \nu$$

which yields the result.

■ C.3 Proofs for Chapter 4

■ C.3.1 Proof of Theorem 4.4.3

Note by conditioning on the number of users that fall in the spherical shell defined by ρ^- and ρ^+ we have,

$$\begin{aligned}
 \Pr(\mathbf{N}_{\mathcal{G}} \gg 0) &= \sum_{j=l}^n \Pr(N_{\epsilon, \rho} = j) \Pr(X_l > 0 | N_{\epsilon, \rho} = j) \\
 &> \sum_{j=l}^n \binom{n}{j} p_s^j (1 - p_{\sigma, \rho})^{n-j} \left(1 - c_1 e^{-jE(p_{\mathcal{G}}, l)}\right) \\
 &= \Pr(N_{\epsilon, \rho} \geq l) - c_1 \sum_{j=l}^n \binom{n}{j} \left(p_{\sigma, \rho} e^{-E(p_{\mathcal{G}}, l)}\right)^j (1 - p_{\sigma, \rho})^{n-j} \\
 &> \Pr(N_{\rho} \geq l) - c_1 \left(p_{\sigma, \rho} e^{-E(p_{\mathcal{G}}, l)} + (1 - p_{\sigma, \rho})\right)^n
 \end{aligned}$$

■ C.3.2 Proof of Theorem 4.5.1

We now prove the rate at which one can hope to scale channel norms and asymptotically have a non-zero probability. In this direction note that from Alzer's bound [9] we have for $m > 1$

$$(1 - e^{-s_l x})^m \leq \tilde{\gamma}_{\text{sf}}(m, x) \leq (1 - e^{-x})^m$$

where $s_l \triangleq \Gamma(1 + m)^{-1/m}$ and

$$\tilde{\gamma}_{\text{sf}}(m, x) = \frac{1}{\Gamma(1 + m)} \int_0^x t^{m-1} e^{-t} dt$$

So,

$$\begin{aligned}
 p_{\rho} &\geq (1 - e^{-s_l \rho^+})^{2m} - (1 - e^{-\rho^-})^{2m} \\
 &= \sum_{j=0}^{2m} \binom{2m}{j} (-1)^{j+1} (e^{j\rho^-} - e^{js_l \rho^+})
 \end{aligned}$$

Now, we note that in order for the bound to be non-zero we must have $\rho_- < s_l \rho_+$ so that the probability is non-zero. However, implicit in the proof of the bound given in [9] if we replace the constant s_l in the lower bound by any number $s \in (s_l, 1)$ then there exists a x^* such that

$$(1 - e^{-sx})^m \leq \tilde{\gamma}_{\text{sf}}(m, x)$$

for all $x \in [x^*, \infty)$. So, asymptotically we can replace the constant s_l by $1 - \epsilon$ for any ϵ such that $1 > \epsilon > 0$.

Now, taking $s < 1$ and $m\rho_+(n) = c \log n$ and $m\rho_-(n) = \log n - \tilde{\alpha}(n)$ yields

$$\begin{aligned}
p_\rho &\geq \sum_{j=0}^{2m} \binom{2m}{j} (-1)^{j+1} \left(e^{-j \log n + j \cdot \tilde{\alpha}(n)} - e^{-jcs \log n} \right) \\
&\geq \sum_{j=0}^{2m} \binom{2m}{j} (-1)^{j+1} e^{-j \log n} \left(e^{j \cdot \tilde{\alpha}(n)} - e^{-j(cs-1) \log n} \right) \\
&= \sum_{j=0}^{2m} \binom{2m}{j} (-1)^{j+1} n^{-j} \left(e^{j \cdot \tilde{\alpha}(n)} - n^{-j(cs-1)} \right)
\end{aligned} \tag{C.10}$$

Thus, for $cs \geq 1$ as $n \rightarrow \infty$ then

$$2m(e^{\tilde{\alpha}(n)} - 1) \leq np_\rho \leq 2me^{\tilde{\alpha}(n)}$$

where the lower bound corresponds to $cs = 1$ and the upper bound corresponds to $cs = \infty$. Thus, if $\rho_+(n) = (1+\delta)(\log n)/m$ and $\rho_-(n) = (\log n)/m - (\log \alpha(n))/m$ where $m \log \log n \leq \log \alpha(n) = o(\log n)$ then

$$\mathbb{E}[N_{\epsilon, \rho}] = np_\rho = 2m\alpha(n)(1 - o(1)) + \mathcal{O}(1/n)$$

From the above derivation (interchanging the role of s in the upper and lower bound) it should be clear that if $\log(n) = o(\rho_-(n))$, then

$$\lim_{n \rightarrow \infty} np_\rho \rightarrow 0$$

■ C.3.3 Proof of Theorem 4.5.2 and Theorem 4.5.5

Similar to the proof of Theorem 4.5.1 we can use a Chernoff bound to bound the probability that $N_{\epsilon, \rho} > l$. Thus, (4.29) becomes

$$1 - \Pr(\mathbf{N}_G = 0) \geq 1 - \exp\left(-\frac{(np_{\sigma, \rho} - m)^2}{np_{\sigma, \rho}}\right) - \left(1 + -p_{\sigma, \rho} \left(1 - e^{(-E(p_G, m))}\right)\right)^n$$

So, bounding $(1 - x)^n$ by $\exp(-x)$ we have

$$\begin{aligned}
\Pr(\mathbf{N}_G = 0) &\leq \Theta(n^{-2m}) + \exp\left(-\mathbb{E}[N_{\epsilon, \rho}] \left(1 - e^{(-E(p_G, m))}\right)\right) \\
&= O(n^{-2m\gamma})
\end{aligned} \tag{C.11}$$

where $\gamma = 1 - e^{-E(p_G, m)}$.

Thus, we are left to determine p_G for an inner product constraint $\epsilon(n)$. Let, $\delta(\epsilon(n), m)$ be the probability that any two users fail to meet the inner product constraint $\epsilon(n)$. That is,

$$\delta(\epsilon(n), m) = \Pr\left[|\mathbf{h}_i^\dagger \mathbf{h}_j| > \epsilon(n)\right].$$

Then, one may, by using the chain rule, write the probability that a set of m users meets

the inner product constraint $\epsilon(n)$ as

$$\prod_{i=1}^{m-1} (1 - i \cdot \delta(\epsilon(n), m))$$

and hence bound the probability that a set of m users meets the constraint $\epsilon(n)$ as

$$p_G > (1 - (m-1) \cdot \delta(\epsilon(n), m))^{m-1}.$$

However, with this representation one may not take $\epsilon(n) \rightarrow 0$ with p_G bounded away from zero. Hence, alternatively one may fix a basis and ensure that users are sufficiently close to the basis. In particular, for the inner product constraint $\epsilon(n)$ to hold one must have

$$|\mathbf{b}_i^\dagger \mathbf{h}_i|^2 \geq \frac{1 + \sqrt{1 - \epsilon(n)^2}}{2} \geq 1 - \epsilon(n)^2.$$

Now, let user 0 channel direction determine the first element of a basis, \mathbf{b}_0 and then consider any orthonormal basis $\{\mathbf{b}_0, \mathbf{b}_1, \dots, \mathbf{b}_{m-1}\}$. Then,

$$\begin{aligned} p_G &> \prod_{i=1}^{m-1} \Pr \left[|\mathbf{b}_i^\dagger \mathbf{h}_i|^2 > 1 - \epsilon(n)^2 \right] \\ &= \prod_{i=1}^{m-1} \epsilon(n)^{2(m-1)} = \epsilon(n)^{2(m-1)^2} \end{aligned}$$

where the last line uses the distribution on inner products (2.43). This completes the proof.

■ C.3.4 Proof of Theorem 4.5.3 and Theorem 4.5.6

This is a simple consequence of Theorem 4.5.1, Theorem 4.5.2 and Theorem 4.5.5. In particular the expected number of users that feedback can be computed directly from Theorem 4.5.1. In the case there is no quantization from Theorem 4.5.2 one may see that the resulting bound on $p_\emptyset(n) = o(1/\log \log n)$ and hence $p_\emptyset(n)R_{\epsilon,\rho}(n) \rightarrow 0$. Thus, it is left to show the difference $R^*(n) - R_{\epsilon,\rho}(n) = O(1/\log n)$. We leave this until after our proof of the results using quantization.

In the case where there is quantization, from Theorem 4.5.5 one may see that the resulting bound on $p_\emptyset(n) = o(1/\log \log n)$ and hence $p_\emptyset(n)R_{\epsilon,\rho,\sigma}(n) \rightarrow 0$. Thus, it is left to show the difference $R^*(n) - R_{\epsilon,\rho,\sigma}(n) = O(1/\log n)$. To see this we note that the SINR of each user may be bounded as

$$\text{SINR}_j^{\text{IC}} \geq \gamma_j \triangleq \frac{P \|\mathbf{h}_j\|^2 \left[|\sigma_j| \tau_j - \sqrt{1 - |\sigma_j|^2} \lambda_{\min} \right]_+^2}{\text{Tr}(\hat{\Phi}_{\mathcal{A}}^{-1}) \tau_j^2 + P \|\mathbf{h}_j\|^2 (1 - |\sigma_j|^2) \lambda_{\max}}$$

from (4.44). Thus, if

$$\|\mathbf{h}_j\|^2 \cdot (1 - |\sigma_j(n)|^2) = g(n)$$

for some $g(n) \rightarrow 0$ and $\sigma_j(n) \rightarrow 1$ one has for sufficiently large n

$$\text{SINR}_j^{\text{IC}} \geq \frac{P\|\mathbf{h}_i\|^2\tau_j^2(1-o(1))}{\text{Tr}(\hat{\Phi}_{\mathcal{A}}^{-1})\tau_j^2 + g(n)}.$$

which in the special case $\hat{\Phi}_{\mathcal{A}} = \mathbf{I}_m$ one has for sufficiently large n ,

$$\begin{aligned} \text{SINR}_j^{\text{IC}} &\geq \frac{P\|\mathbf{h}_i\|^2/m(1-o(1))}{1+g(n)} \\ &\geq \frac{\text{SINR}^*(n)(1-o(1))}{1+g(n)} \end{aligned}$$

Thus,

$$\frac{\text{SINR}^*(n)}{\text{SINR}_j^{\text{IC}}} \leq \frac{1+g(n)}{(1-o(1))}.$$

Now, as $g(n) \rightarrow 0$ and $\text{SINR}^*(n) \rightarrow \infty$ we have

$$\frac{\text{SINR}^*(n)}{\text{SINR}_j^{\text{IC}}} \leq (1+o(1)) \cdot (1+g(n)).$$

By direct computation it is easy to see that

$$\rho_-(n)(1-\sigma(n)^2) = \frac{1}{\log n} + o(1)$$

and thus

$$\begin{aligned} R^*(n) - R_{\epsilon,\rho,\sigma}(n) &= \log \left(\frac{1 + \text{SINR}^*(n)}{1 + \text{SINR}_j^{\text{IC}}(n)} \right) \\ &\leq \log \left(\frac{1 + \text{SINR}^*(n)}{\text{SINR}_j^{\text{IC}}(n)} \right) \\ &\leq \log \left(\frac{1}{\text{SINR}_j^{\text{IC}}(n)} + \frac{\text{SINR}^*(n)}{\text{SINR}_j^{\text{IC}}(n)} \right) \\ &\leq \frac{1}{\text{SINR}_j^{\text{IC}}(n)} + \frac{\text{SINR}^*(n)}{\text{SINR}_j^{\text{IC}}(n)} - 1 \\ &= 1/\log n + o(1) \end{aligned}$$

Hence, $R^*(n) - R_{\epsilon,\rho,\sigma}(n) = O(1/\log n)$.

We note that the in the case that there is no quantization is equivalent to the case where quantization is used and

$$1 - \sigma(n)^2 = \epsilon^2(n).$$

As $\rho_-(n)(1-\sigma(n)^2) = \frac{1}{\log n} + o(1)$ again in this case of the interference-ignoring multiplexer one has $R^*(n) - R_{\epsilon,\rho,\sigma}(n) = O(1/\log n)$. To see that this is also the case in the interference-

canceling multiplexer note

$$\begin{aligned}\mathrm{Tr}(\hat{\Phi}_{\mathcal{A}}^{-1}) &\leq m \cdot \frac{1}{\lambda_{\min}(\hat{\Phi}_{\mathcal{A}})} \\ &\leq m \cdot \frac{1}{1 - (m-1)\epsilon(n)^2}\end{aligned}$$

Hence,

$$\mathrm{SINR}^{\mathrm{IC}} \geq P\rho_{-}(n) \frac{1 - (m-1)\epsilon(n)^2}{m}$$

and again $R^{*}(n) - R_{\epsilon,\rho,\sigma}(n) = O(1/\log n)$.

List of Symbols

■ Channel Notation and Metrics

$\mathbf{h}_j^\dagger[k]$ 24 The channel gain vector of user j .	24
$\hat{\mathbf{h}}_j$ 26 The quantized representation of the j th user's channel.	26
$\tilde{\mathbf{h}}_i$ 34 The direction of the i th user's channel vector.	34
$\mathbf{H}[k]$ 25 The collection of channel gain vectors of the users written in matrix form.	25
$\hat{\mathbf{K}}_{\mathbf{h}_i}$ 29 The estimate of the spatial covariance of the i th user's channel.	29
$P_{\text{fail}}^{(M)}(\text{SINR}_0)$ 36 The probability that there is not a subset of users that simultaneously meet the SINR target SINR_0 .	36
$\mathcal{Q}(\mathbf{h}_j)$ 26 The quantized representation of the j th user's channel.	26
R_{RX} 183 The receive end covariance.	183
R_{TX} 183 The transmit end covariance.	183
σ_i 34 The correlation between the normalized channel vector of user i , $\tilde{\mathbf{h}}_i$, and the beam-forming vector \mathbf{w}_i .	34
$\text{SNR}_i[k]$ 33 The signal-to-noise ratio of user i at time k in a time division system.	33
Ω 185 The expected energy coupled between the transmit and receive eigenmodes.	185
$\mathbf{x}[k]$ 24 The signal transmitted from the array at time k .	24
$y_j[k]$ 24 The signal received by user j at time k .	24
$z_j[k]$ 24 Independent identically distributed white Gaussian Noise.	24

■ Quantizer Notation and Metrics

\mathcal{C}_r 26	An r -bit channel quantization codebook..
$K(\mathcal{C})$ 204	The empirical second order moment of the quantizer.
$\mathbf{K}_{\mathbf{h}_i}$ 29	The spatial covariance of the i th user's channel.
$\eta(\mathcal{C})$ 31	The minimum number of codewords orthogonal to any codeword in \mathcal{C} .
$\mathbf{p}_i(\mathcal{C}_r)$ 27	The probability that user i is quantized to a codeword in \mathcal{C}_r .
$\hat{\Phi}_{\mathcal{A}}$ 26	The collection of quantized channel gain vectors of the users in \mathcal{A} written in matrix form.
$\text{SINR}_{\text{sat}}^{\text{UB}}(r, m)$ 40	An upper bound on the achievable value of SINR_{sat} for a rate r channel quantizer of length m .
$\text{SINR}_{\text{sat}}^{\text{RVQ}}(r, m)$ 39	The expected value of SINR_{sat} achieved by the ensemble of rate r random vector quantizers of length m .
$\text{SINR}_{\text{sat}}(\mathcal{C}_r)$ 38	The expected value of the high SNR approximation of the SINR achieved by an orthogonal set of users which use the codebook \mathcal{C} for channel quantization.
$\text{SINR}_{\text{sat}}(\mathcal{C}_r; n, \ell)$ 169	The expected value of SINR_{sat} for the best ℓ users in a pool of size n .
$\text{Sym}(\mathcal{C})$ 64	The set of all unitary matrices that act transitively on \mathcal{C} .
$\mu_0(\mathcal{C})$ 30	The coherence of the codebook \mathcal{C} .
$\mu_k(\mathcal{C})$ 30	The k -norm of the cross correlation of the codebook \mathcal{C} .
$\bar{\mu}_k(2^r, m; \eta)$ 31	A lower bound on the coherence then every codeword in \mathcal{C} is orthogonal to at least η codewords in \mathcal{C} .
$\bar{\mu}_k(2^r, m)$ 30	The Welch lower bound on $\mu_k(\mathcal{C})$.

■ Quantizer Constructions

$\mathcal{C}(\Upsilon_1, \{0\}; \mathbb{F}, \{\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_a\})$	111
An intermediate rate code over \mathbb{Z}_{p^a} .	
$\mathcal{C}(\Upsilon_1, \Upsilon_2; \mathbb{K}, f)$	94
A general channel quantizer construction which replaces the inner product with $\text{Tr}(y \cdot f(z))$.	
$\mathcal{C}(\Upsilon_1, \Upsilon_2; L)$	72
A code defined through the standard inner product.	
$\mathcal{C}(\Upsilon_1, \Upsilon_2; \mathbb{F}, \mathcal{T})$	109
An intermediate rate channel quantizer.	
$\mathcal{C}_{\text{sparse}}^{(2,4)}(k)$	62
A channel quantizer for a 4 transmit antenna system of cardinality $3 \cdot 2^{k+1}$.	
$\mathcal{C}_{\text{ASC}}^*(k, j)$	70
A systematic construction of channel quantizers using both sparse and dense component codes which expurgates the codewords which correspond to the standard basis..	
$\mathcal{C}_{\text{ASC}}(k, j)$	70
A systematic construction of channel quantizers using both sparse and dense component codes.	
$\mathcal{C}_{\text{DFT}}(r, \mathbf{u})$	48
A rate r DFT code constructed using the method of Hochwald [56].	
$\mathcal{C}_{\mathcal{T}}(a, m, \mathcal{I}, h)$	103
A dense code with varying degrees of orthogonality.	
$\mathcal{C}_{\text{WiMax}}(r, \mathbf{u}, \mathbf{a})$	49
The rate r code construction included in the IEEE 802.16e standard [1, 143].	
$\mathbf{c}(\lambda, 0; \mathbb{F}, \mathcal{S})$	109
A codeword of the intermediate rate channel quantizer $\mathcal{C}(\Upsilon_1, \Upsilon_2; \mathbb{F}, \mathcal{T})$.	
$\mathcal{C}_{\mathbb{Z}}^{(2,4)}(k; \mathcal{I}_0 \setminus \{[0, 0]\})$	61
A 2-sparse channel quantizer in for a 4 transmit antenna system.	
$\mathbf{c}(\boldsymbol{\lambda}, \bar{\boldsymbol{\beta}}; L, p^a)$	72
A codeword of $\mathcal{C}(\Upsilon_1, \Upsilon_2; L)$.	
$\mathcal{C}_{\mathbf{F}}(\alpha, \gamma, \mathcal{C})$	118
A universal code consisting of the union of local codes.	
$\mathcal{C}(\mathbf{c}_i; \alpha, \gamma, \mathcal{B}_i)$	118
A local code which has been localized about \mathbf{c}_i .	

■ SystemNotation

ϵ	127
	A pre-selection threshold.	
$K(\mathbf{m})$	136
	The collection of service rates of the switch state \mathbf{m} .	
$\overline{\mathbf{M}}$	136
	The set of all the switch states \mathbf{m} .	
\mathbf{m}	136
	A generalized switch state.	
$N_{\epsilon,\rho}$	128
	The number of users that feedback from the user pool.	
$N_{\epsilon,\rho}^{(\ell)}$	128
	The number of users that feedback from the ℓ th cluster.	
$n_{\delta}(\alpha)$	145
	The quantization order of a code for given fading statistics.	
$\mathbf{N}_{\mathcal{G}}$	155
	The number of cliques in a generalized switch.	
$\rho_+^{(\ell)}, \rho_-^{(\ell)}, \sigma^{(\ell)}$	127
	Channel feedback thresholds.	
$p_{\mathcal{G}}$	153
	The probability that any set of users of size m channel vectors will yield a maximally sized clique in a switch.	
$\mathcal{R}_{\sigma,\rho}$	128
	The subset of the user pool that feedback.	
$\mathcal{R}_{\rho,\sigma}^{(\ell)}$	127
	The user from the ℓ th cluster that feedback.	
$\mathcal{T}_{\epsilon}^{(\ell)}$	127
	A collection of candidate sets of users for transmission.	
\mathcal{U}	126
	The set of users in the system.	
$\mathcal{U}^{(\ell)}$	126
	The cluster of users in the system.	

■ Quantizer Construction Notation

$c(i, j)$	55
	a map from $\mathcal{I}_0 \times \Upsilon_1$ to \mathbb{C} which describes the entries of \mathbf{C}_B .	
\mathbf{C}_0	53
	A high dimensional embedding of the “good low dimensional channel quantizer \mathbf{C}_B ..	
\mathbf{C}_B	53
	A “good low dimensional channel quantizer.	
$\mathbf{F}(\mathbf{c}_i; \alpha, \gamma, \mathcal{B}_i)$	118
	The local interpolation operation for \mathbf{c}_i with respect to the basis \mathcal{B}_i .	
$\Gamma_{\mathbf{C}}(\mathbf{a}; \boldsymbol{\beta}, L)$	81
	A function which describes the inner product of between any two vectors in $\mathcal{C}(L_a^d, L^c; L)$.	
$\bar{\gamma}$	71
	An element of Υ_2 when $\Upsilon_2 \subset (\mathbb{Z}_p)^{m'}$.	
$\mathcal{H}_{L,a}$	75
	A (commutative) set of unitary matrices that act invariantly on the code $\mathcal{C}(\Upsilon_1, \Upsilon_2; L)$.	
$\vartheta_{\mathcal{I}}(x)$	102
	A lift of \mathbb{F}_{p^a} which defines varying orthogonality relations.	
\mathcal{I}	55
	The row labels for \mathbf{C}_0 .	
\mathcal{I}_0	55
	The support of the rows of \mathbf{C}_0 .	
\mathcal{I}_P	107
	The coset leaders of the p -cyclotomic cosets modulo $p^{m'} - 1$, P .	
λ	71
	An element of Υ_1 when $\Upsilon_1 \subset (\mathbb{Z}_{p^a})^{m'}$.	
$\downarrow L_a^d$	80
	The set of elements of L_a^d that are complimentary to $p^{a-1} \cdot L_1^d$.	
L	71
	A label for \mathcal{I} when \mathcal{I} is closed under addition.	
L^c	76
	Any sub-space of $(\mathbb{Z}_p)^{m'}$ complimentary to L .	
L_a^\perp	75
	The set of elements of $(\mathbb{Z}_{p^a})^{m'}$ orthogonal to the lifted elements of L .,.....	76
	Any sub-module of $(\mathbb{Z}_{p^a})^{m'}$ that is complimentary to L_a^\perp ..	
L_a^d	76
	Any sub-module of $(\mathbb{Z}_{p^a})^{m'}$ that is complimentary to L_a^\perp .	
$U_a(L)$	84
	The set of all orthogonal bases of the code $\mathcal{C}(L_a^d, L^c; L)$.	
P	107
	The p -cyclotomic cosets modulo $p^{m'} - 1$.	

$R(\boldsymbol{\lambda}; \bar{\boldsymbol{\beta}})$	84
A unitary operator which acts diagonally on a subset of coordinates and as the identity on others.	
$S(\bar{\boldsymbol{\beta}})$	74
The matrix that permutes the basis $\{\mathbf{e}_{\bar{\alpha}}\}$ by translations..	
$T(\boldsymbol{\lambda})$	73
The matrix that acts diagonally on the basis $\{\mathbf{e}_{\bar{\alpha}}\}$.	
$T(\ell; f)$	94
.	
$\text{twt}_H(\boldsymbol{\beta})$	82
The twisted Hamming weight of the vector $\boldsymbol{\beta}$.	
$T_{\text{GRM}}(r_1, r_2, \dots, r_a)$	113
The defining sets of a Reed Muller code generalized over \mathbb{Z}_{p^a} .	
\mathcal{T}_a	111
A defining set of an intermediate rate code.	
$\Omega_{k,m'}(\hat{\Upsilon}_1)$	83
Collection of subsets of $\hat{\Upsilon}_1 \times (p^{a-1} \cdot L_a^d)$ of cardinality k that satisfy Theorem 3.3.10..	
Υ_1	55
An index set for the columns of \mathbf{C}_0 (or \mathbf{C}_B).	
Υ_2	55
A set of permutations of \mathcal{I} .	
$\text{wt}_H(\boldsymbol{\beta})$	82
The hamming weight of the vector $\boldsymbol{\beta}$.	
$\text{wt}_p(s)$	110
The p -weight of the integer s .	
$\tilde{\mathbf{Y}}(\mathbf{b}_1, \mathbf{b}_2; \alpha)$	117
A one dimensional rotation in the $\mathbf{b}_1 - \mathbf{b}_2$ plane.	
ζ_{p^a}, ζ	72
A p^a -th root of unit.	

■ Discrete Channel Modeling Notation

$\text{Beta}(\theta_{i,j}^{(a)}, \theta_{i,j}^{(b)})$	193
The univariate beta distribution.	
\mathbf{C}_i	196
The compound multinomial random variable modeling feedback from cluster i .	
$\text{Dirichlet}(p_0, p_1, \dots, p_{2^r-1}; \boldsymbol{\theta})$	192
The Dirichlet distribution of length 2^r .	
$\text{GDirichlet}(p_0, p_1, \dots, p_{2^r-1}; \boldsymbol{\theta}^{(a)}, \boldsymbol{\theta}^{(b)})$	194
The generalized Dirichlet distribution of length 2^r .	
$\mathbf{N}_{i,j}[k_1, n_k]$	190
The user assignment distribution of length n_k .	
$\hat{\boldsymbol{\theta}}(\mathbf{n}, j; \boldsymbol{\theta}^{(a)}, \boldsymbol{\theta}^{(b)})$	195
The Bayesian estimate for the j th cell probabilities based on the observation \mathbf{n} assuming a GDirichlet distribution as a prior on the cell probabilities.	
$\mathbf{X}_{i,j}$	190
The input occupancy distribution of user i .	
$Z_{i,j}$	196
The “hidden random variable indicating if user i in cluster j .”	

■ General Nomenclature

- multi-node matching gain**16
the multi-user diversity stemming from one's ability to schedule users that negligibly interfere with one another.
- order statistic gain** 16
the multi-user diversity stemming from one's ability to schedule the users that are individually at high SNR.

Bibliography

- [1] IEEE standard for local and metropolitan area networks Part 16: Air interface for broadband wireless access systems. *IEEE Std 802.16-2009 (Revision of IEEE Std 802.16-2004)*, pages C1–2004, 29 2009.
- [2] A R. Calderbank, S. N. Diggavi, and N. Al Dahir. Space-time signaling based on Kerdock and Delsarte-Goethals codes. In *IEEE International Conference on Communications (ICC) Paris*, pages 483–487, 2004.
- [3] Kanat Abdukhalikov. Defining sets of extended cyclic codes invariant under the affine group. *Journal of Pure and Applied Algebra*, 196(1):1–19, March 2005.
- [4] Kanat S. Abdukhalikov. Affine invariant and cyclic codes over p -adic numbers and finite rings. *Des. Codes Cryptography*, 23(3):343–370, 2001.
- [5] Kanat S. Abdukhalikov. Codes over p -adic numbers and finite rings invariant under the full affine group. *Finite Fields and Their Applications*, 7(4):449 – 467, 2001.
- [6] Manish Airy, Sanjay Shakkottai, and Robert W. Heath, Jr. Spatially greedy scheduling in multi-user MIMO wireless systems. In *Proc. of IEEE Asilomar Conf. on Signals, Systems, and Computers*, Pacific Grove, CA, November 2003.
- [7] Defne Aktas and Hesham El Gamal. Multiuser scheduling for multiple antenna systems. In *Proc. IEEE Vehicular Tech. Conf.*, volume 3, pages 1743–1747, Orlando, USA, October 2003.
- [8] P. Almers, E. Bonek, A. Burr, N. Czink, M. Debbah, V. Degli-Esposti, H. Hofstetter, P. Kyösti, D. Laurenson, G. Matz, A. F. Molisch, C. Oestges, and H. Özcelik. Survey of channel and radio propagation models for wireless MIMO systems. *EURASIP J. Wirel. Commun. Netw.*, 2007(1):56–56, 2007.
- [9] Horst Alzer. On some inequalities for the incomplete Gamma function. *Mathematics of Computation*, 66(218):771–778, April 1997.
- [10] G.E. Andrews, R. Askey, and R. Roy. *Special Functions*. Number 71 in Encyclopedia of Mathematics and its Applications. Cambridge University Press, Cambridge, 1999.
- [11] M. Andrews. Instability of the proportional fair scheduling algorithm for HDR. *IEEE Transactions on Wireless Communications*, 3(5):1422–1426, Sept. 2004.
- [12] Matthew Andrews, Krishnan Kumaran, Kavita Ramanan, Alexander Stolyar, Rajiv Vijayakumar, and Phil Whiting. Scheduling in a queuing system with asynchronously varying service rates. *Probab. Eng. Inf. Sci.*, 18(2):191–217, 2004.

-
- [13] Alexei Ashikhmin and RaviKiran Gopalan. Grassmannian packings for efficient quantization in MIMO broadcast systems. In *IEEE International Symposium on Information Theory*, pages 1811–1815, June 2007.
- [14] E. F. Assmus, Jr. and J. D. Key. *Polynomial codes and finite geometries*, chapter 16, pages 1269–1343. Elsevier, 1998.
- [15] E. F. Assmus, Jr. and J.D. Key. *Designs and their codes*. Number 103 in Cambridge Tracts in Mathematics. Cambridge University Press, New York, NY, USA, 1992.
- [16] Chun Kin Au-Yeung and David J. Love. On the performance of random vector quantization limited feedback beamforming in a MISO system. *IEEE Transactions on Wireless Communications*, 6(2):458–462, Feb. 2007.
- [17] L. Babel. Finding maximum cliques in arbitrary and in special graphs. *Computing*, 46(4):321–341, December 1991.
- [18] T. Banerjee and S. R. Paul. An extension of Morel-Nagaraj’s finite mixture distribution for modelling multinomial clustered data. *Biometrika*, 86(3):723–727, 1999.
- [19] G. Bauch, J. Bach Andersen, C. Guthy, M. Herdin, J. Nielsen, J.A. Nossek, P. Tejera, and W. Utschick. Multiuser MIMO channel measurements and performance in a large office environment. In *IEEE Wireless Communications and Networking Conference*, pages 1900–1905, March 2007.
- [20] Claude. Berge and V. Chvatal, editors. *Topics on perfect graphs*. Annals of Discrete Mathematics. North-Holland Pub. Co., New York, 1984.
- [21] Thierry P. Berger. On the automorphism groups of affine-invariant codes. *Des. Codes Cryptography*, 7(3):215–221, 1996.
- [22] T.P. Berger and P. Charpin. The permutation group of affine-invariant extended cyclic codes. *IEEE Transactions on Information Theory*, 42(6):2194–2209, November 1996.
- [23] Dimitris Bertsimas, Karthik Natarajan, and Chung-Piaw Teo. Tight bounds on expected order statistics. *Probab. Eng. Inf. Sci.*, 20(4):667–686, 2006.
- [24] J.T. Blackford and D.K. Ray-Chaudhuri. A transform approach to permutation groups of cyclic codes over Galois rings. *IEEE Transactions on Information Theory*, 46(7):2350–2358, November 2000.
- [25] T. Blackford. *Permutation groups of extended cyclic codes over Galois rings*. PhD thesis, The Ohio State University, Columbus, Ohio, 1999.
- [26] S. Borst and M. Jonckheere. Flow-level stability of channel-aware scheduling algorithms. In *International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks*, pages 1–6, April 2006.
- [27] Nizar Bouguila. Clustering of count data using generalized Dirichlet multinomial distributions. *IEEE Transactions on Knowledge and Data Engineering*, 20:462–474, 2007.

-
- [28] A. R. Calderbank and Peter W. Shor. Good quantum error-correcting codes exist. *Phys. Rev. A*, 54(2):1098–1105, Aug 1996.
- [29] A. R. Calderbank and N. J. A. Sloane. Modular and p -adic cyclic codes. *Des. Codes Cryptography*, 6(1):21–35, 1995.
- [30] A.R. Calderbank, P.J. Cameron, W.M. Kantor, and J.J. Seidel. Z₄-Kerdock Codes, Orthogonal Spreads, and Extremal Euclidean Line-Sets. *Proc. London Math. Soc.*, 75(2):436–480, 1997.
- [31] Randy Carraghan and Panos M. Pardalos. An exact algorithm for the maximum clique problem. *Operations Research Letters*, 9(6):375 – 382, 1990.
- [32] Etienne F. Chaponniere, Peter J. Black, Jack M. Holtzman, and David Ngar Ching Tse. Transmitter directed code division multiple access system using path diversity to equitably maximize throughput. "US Patent 6,449,490", September 2002.
- [33] Charalambos A. Charalambides. *Combinatorial Methods in Discrete Distributions (Wiley Series in Probability and Statistics)*. Wiley-Interscience, 2005.
- [34] Jihoon Choi and Jr. Heath, R.W. Interpolation based transmit beamforming for MIMO-OFDM with limited feedback. *IEEE Transactions on Signal Processing*, 53(11):4125–4135, Nov. 2005.
- [35] Chen-Nee Chuah, D.N.C. Tse, J.M. Kahn, and R.A. Valenzuela. Capacity scaling in MIMO wireless systems under correlated fading. *IEEE Transactions on Information Theory*, 48(3):637–650, Mar 2002.
- [36] Kuo-Liang Chung and Wen-Ming Yan. The complex Householder transform. *IEEE Transactions on Signal Processing*, 45(9):2374–2376, Sep 1997.
- [37] G.W.K. Colman and T.J. Willink. Limited feedback precoding in realistic MIMO channel conditions. In *IEEE International Conference on Communications*, pages 4363–4368, June 2007.
- [38] Robert J. Connor and James E. Mosimann. Concepts of independence for proportions with a generalization of the Dirichlet Distribution. *Journal of the American Statistical Association*, 64(325):194–206, 1969.
- [39] J.G. Dai and B. Prabhakar. The throughput of data switches with and without speedup. In *Proceedings of the Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies*, volume 2, pages 556–564, 2000.
- [40] H. A. David. *Order statistics*. Wiley New York,, 1970.
- [41] J.A. Davis and J. Jedwab. Peak-to-mean power control in OFDM, Golay complementary sequences, and Reed-Muller codes. *IEEE Transactions on Information Theory*, 45(7):2397–2417, Nov 1999.
- [42] R. de Lacerda, L.S. Cardoso, R. Knopp, D. Gesbert, and M. Debbah. EMOS platform: Real-time capacity estimation of MIMO channels in the UMTS-TDD band. In *4th International Symposium on Wireless Communication Systems*, pages 782–786, Oct. 2007.

-
- [43] P. Delsarte. On cyclic codes that are invariant under the general linear group. *IEEE Transactions on Information Theory*, 16(6):760 – 769, nov 1970.
- [44] P. Delsarte, J.M. Goethals, and J.J. Seidel. Bounds for systems of lines and Jacobi polynomials. *Philips Res. Rep.*, 30:91–105, 1975.
- [45] P. Delsarte, J.M. Goethals, and F.J. Mac Williams. On generalized Reed-Muller codes and their relatives. *Information and Control*, 16(5):403–442, July 1970.
- [46] B.K. Dey and B.S. Rajan. Affine invariant extended cyclic codes over Galois rings. *IEEE Transactions on Information Theory*, 50(4):691–698, April 2004.
- [47] S. N. Diggavi, A. R. Calderbank, S. Dusad, and N. Al-Dhahir. Diversity embedded space-time codes. *IEEE Transactions on Information Theory*, 54(1):33–50, 2008.
- [48] Yonina C. Eldar and H. Bolcskei. Geometrically uniform frames. *IEEE Transactions on Information Theory*, 49(4):993, 2003.
- [49] P. Fernandes, L. T. Berger, J. Mrtires, and P. Kyritsi. Effects of multi user MIMO scheduling freedom on cellular downlink system throughput. In *Proc. IEEE 60th Vehicular Technology Conference*, Los Angeles, USA, September 2004.
- [50] G. David Forney, Mitchell D. Trott, N. J. A. Sloane, and N. J. A. Sloane. The Nordstrom-Robinson code is the binary image of the octacode. In *Proceedings DIMACS/IEEE Workshop on Coding and*, pages 19–26, 1992.
- [51] G. J. Foschini and M. J. Gans. On limits of wireless communications in a fading environment when using multiple antennas. *Wirel. Pers. Commun.*, 6(3):311–335, 1998.
- [52] Allen Gersho and Robert M. Gray. *Vector quantization and signal compression*. Kluwer Academic Publishers, Norwell, MA, USA, 1991.
- [53] M. Goötschel, L. Lovász, and Schrijver A. Polynomial algorithms for perfect graphs. In C. Berge and V. Chvatal, editors, *Topics on perfect graphs*, Annals of Discrete Mathematics. North-Holland Pub. Co., New York, 1984.
- [54] M Grötschel, L Lovász, and A Schrijver. Relaxations of vertex packing. *J. Comb. Theory Ser. B*, 40(3):330–343, 1986.
- [55] A.R. Hammons, P.V. Kumar, A.R. Calderbank, N.J.A. Sloane, and P. Sole. The Z₄-linearity of Kerdock, Preparata, Goethals, and related codes. *IEEE Transactions on Information Theory*, 40(2):301–319, March 1994.
- [56] B.M. Hochwald, T.L. Marzetta, T.J. Richardson, W. Sweldens, and R. Urbanke. Systematic design of unitary space-time constellations. *IEEE Transactions on Information Theory*, 46(6):1962–1973, Sep 2000.
- [57] S. G. Hoggar. t-designs with general angle set. *Eur. J. Comb.*, 13(4):257–271, 1992.
- [58] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, second edition, 1999.

-
- [59] W.C. Huffman. Decompositions and extremal type II codes over Z_4 . *IEEE Transactions on Information Theory*, 44(2):800–809, Mar 1998.
- [60] Thomas W Hungerford. *Algebra*. Graduate Texts in Mathematics. Springer Verlag, New York, 1996.
- [61] T. Inoue and R.W. Heath. Kerdock codes for limited feedback precoded MIMO systems. *IEEE Transactions on Signal Processing*, 57(9):3711–3716, Sept. 2009.
- [62] A. Jalali, R. Padovani, and R. Pankaj. Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system. In *51st IEEE Vehicular Technology Conference Proceedings*, volume 3, pages 1854–1858, 2000.
- [63] Gordon James and Martin Liebeck. *Representations And Characters Of Groups*. Cambridge University Press, Cambridge, 2001.
- [64] Svante Janson. Large deviations for sums of partly dependent random variables. *Random Structures Algorithms*, 24(3):234–248, 2004.
- [65] Svante Janson, Tomasz Luczak, and Andrzej Rucinski. *Random Graphs*. John Wiley, New York, 2000.
- [66] N. Jindal. MIMO broadcast channels with digital channel feedback. In *Fortieth Asilomar Conference on Signals, Systems and Computers*, pages 1506–1510, Nov. 2006.
- [67] N. Jindal. MIMO broadcast channels with finite-rate feedback. *IEEE Transactions on Information Theory*, 52(11):5045–5060, Nov. 2006.
- [68] Nihar Jindal. MIMO broadcast channels with finite rate feedback. *IEEE Trans. on Inform. Theory*, 52:5045–5059, 2006.
- [69] N. Johnson, S. Kotz, and N. Balakrishnan. *Discrete Multivariate Distributions*. John Wiley, New York, 1997.
- [70] G. A. Kabatyanskii and V. I. Lenenshtein. Bounds for packings on a sphere and in space. *Problems Inform. Transm.*, 14:1–17, 1978.
- [71] F. Kaltenberger, D. Gesbert, R. Knopp, and M. Kountouris. Correlation and capacity of measured multi-user MIMO channels. In *IEEE 19th International Symposium on Personal, Indoor and Mobile Radio Communications*, pages 1–5, Sept. 2008.
- [72] F. Kaltenberger, M. Kountouris, L. Cardoso, R. Knopp, and D. Gesbert. Capacity of linear multi-user MIMO precoding schemes with measured channel data. In *IEEE 9th Workshop on Signal Processing Advances in Wireless Communications*, pages 580–584, July 2008.
- [73] R. M. Karp. Reducibility among combinatorial problems. In R. E. Miller and J. W. Thatcher, editors, *Complexity of Computer Computations*, pages 85–103. Plenum Press, 1972.
- [74] T. Kasami, S. Lin, and W.W. Peterson. Some results on cyclic codes which are invariant under the affine group and their applications. *Information and Control*, 11(5-6):475–496, November-December 1967.

-
- [75] Frank Kelly. Charging and rate control for elastic traffic. *European Transactions on Telecommunications*, 8:33–37, 1997.
- [76] Andreas Klappenecker and Martin Rötteler. Constructions of mutually unbiased bases. In *Finite Fields and Applications*, volume 2948 of *Lecture Notes in Computer Science*, pages 262–266. Springer Berlin / Heidelberg, 2004.
- [77] Donald E. Knuth. *Fundamental Algorithms*, volume 1 of *The Art of Computer Programming*, section 1.2, pages 107–116. Addison-Wesley, Reading, Massachusetts, third edition, 1997.
- [78] P.V. Kumar, T. Helleseth, and A.R. Calderbank. An upper bound for Weil exponential sums over Galois rings and applications. *IEEE Transactions on Information Theory*, 41(2):456–468, Mar 1995.
- [79] Seymour M. Kwerel. Most stringent bounds on the probability of the union and intersection of m events for systems partially specified by $S_1, S_2, \dots, S_k, 2 \leq k < m$. *Journal of Applied Probability*, 12(3):612–619, 1975.
- [80] L. Li and A. Goldsmith. Optimal resource allocation for fading broadcast channels-Part I: Ergodic capacity. *IEEE Transactions on Information Theory*, 47(3):1083–1102, March 2001.
- [81] Jacobus Hendricus van Lint. *Introduction to Coding Theory*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1982.
- [82] David J. Love and Jr. Robert W. Heath. Grassmannian beamforming for multiple-input multiple-output wireless systems. *IEEE Transactions on Information Theory*, 49(10):2735–2747, October 2003.
- [83] D.J. Love and Jr. Heath, R.W. Limited feedback diversity techniques for correlated channels. *IEEE Transactions on Vehicular Technology*, 55(2):718–722, March 2006.
- [84] F.J. MacWilliams. Permutation decoding of systematic codes. *Bell System Tech. J.*, 43:485–505, 1964.
- [85] B. R. McDonald. *Finite Rings with Identity*. Marcel Dekker, New York, 1974.
- [86] Robert J. McEliece. *Finite fields for scientists and engineers*. Kluwer Academic Publishers, Norwell, MA, USA, 1987.
- [87] N. McKeown, A. Mekkittikul, V. Anantharam, and J. Walrand. Achieving 100% throughput in an input-queued switch. *IEEE Transactions on Communications*, 47(8):1260–1267, Aug 1999.
- [88] Nick McKeown. The iSLIP scheduling algorithm for input-queued switches. *IEEE/ACM Trans. Netw.*, 7(2):188–201, 1999.
- [89] Geoffrey McLachlan and David Peel. *Finite Mixture Models*. Wiley Series in Probability and Statistics. Wiley-Interscience, October 2000.
- [90] K.K. Mukkavilli, A. Sabharwal, E. Erkip, and B. Aazhang. On beamforming with finite rate feedback in multiple-antenna systems. *IEEE Transactions on Information Theory*, 49(10):2562–2579, Oct. 2003.

-
- [91] Aradhana Narula, Michael J. Lopez, Mitchell D. Trott, and Gregory W. Wornell. Efficient use of side information in multiple-antenna data transmission over fading channels. *IEEE Journal on Selected Areas in Communications*, 16(8):1423–1436, October 1998.
- [92] Melvyn B. Nathanson. *Elementary Methods in Number Theory*. Number 195 in Graduate Texts in Mathematics. Springer, 2000.
- [93] Gabriele Nebe, Eric M. Rains, and Neil J. A. Sloane. *Self-Dual Codes and Invariant Theory (Algorithms and Computation in Mathematics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.
- [94] Ingram Olkin. Monotonicity properties of Dirichlet integrals with applications to the multinomial distribution and the analysis of variance. *Biometrika*, 59(2):303–307, August 1972.
- [95] Patric R. J. Östergård. A new algorithm for the maximum-weight clique problem. *Nordic J. of Computing*, 8(4):424–436, 2001.
- [96] Patric R. J. Östergård. A fast algorithm for the maximum clique problem. *Discrete Appl. Math.*, 120(1-3):197–207, 2002.
- [97] Panos M. Pardalos and Jue Xue. The maximum clique problem. *Journal of Global Optimization*, 4(3):301–328, April 1994.
- [98] C.B. Peel, B.M. Hochwald, and A.L. Swindlehurst. A vector-perturbation technique for near-capacity multiantenna multiuser communication-Part I: Channel inversion and regularization. *IEEE Transactions on Communications*, 53(1):195–202, Jan. 2005.
- [99] A.S.Y. Poon, R.W. Brodersen, and D.N.C. Tse. Degrees of freedom in multiple-antenna channels: A signal space approach. *IEEE Transactions on Information Theory*, 51(2):523–536, Feb. 2005.
- [100] John Proakis. *Digital Communications*. McGraw-Hill, New York, 4 edition, 2000.
- [101] Feng Qi. Bounds for the ratio of two gamma functions—from Gautschi’s and Kershaw’s inequalities to completely monotonic functions, 2009.
- [102] V. Raghavan, R.W. Heath, and A.V. Sayeed M. Systematic codebook designs for quantized beamforming in correlated MIMO channels. *IEEE Journal on Selected Areas in Communications*, 25(7):1298–1310, September 2007.
- [103] Sundar Rajan and M. U. Siddiqi. Transform domain characterization of cyclic codes over Z_m . *Appl. Algebra Eng. Commun. Comput.*, 5(5):261–275, 1994.
- [104] N. Ravindran and N. Jindal. Limited feedback-based block diagonalization for the MIMO broadcast channel. *IEEE Journal on Selected Areas in Communications*, 26(8):1473–1482, October 2008.
- [105] J.C. Roh and B.D. Rao. Transmit beamforming in multiple-antenna systems with finite rate feedback: A VQ-based approach. *IEEE Transactions on Information Theory*, 52(3):1101–1112, March 2006.

-
- [106] D. Sarwate. Meeting the Welch bound with equality. In *Sequences and their applications*, Singapore, 1998.
- [107] Semih Serbetli and Aylin Yener. Time-slotted multiuser MIMO systems: Beamforming and scheduling strategies. *EURASIP Journal on Wireless Communications and Networking*, 2004(2):286–296, 2004.
- [108] Claude Shannon. A mathematical theory of communication. *BSTJ*, pages 379–423,623–656, 1948.
- [109] Claude E. Shannon. Probability of error for optimal codes in a Gaussian channel. *Bell System Technical Journal*, 38(3), May 1959.
- [110] Masoud Sharif and Babak Hassibi. Scaling laws of sum rate using time-sharing, DPC, and beamforming for MIMO broadcast channels. In *Proc. IEEE ISIT 2004*, Chicago, IL, July 2004.
- [111] Masoud Sharif and Babak Hassibi. On the capacity of MIMO broadcast channels with partial side information. *IEEE Transactions on Information Theory*, 51(2):506–522, February 2005.
- [112] Da-Shan Shiu, G.J. Foschini, M.J. Gans, and J.M. Kahn. Fading correlation and its effect on the capacity of multielement antenna systems. *IEEE Transactions on Communications*, 48(3):502–513, Mar 2000.
- [113] V. M. Sidelnikov. Quantum codes and abelian subgroups of the extra-special group. *Probl. Inf. Transm.*, 38(3):194–202, 2002.
- [114] J. Sijbers, A.J. den Dekker, P. Scheunders, and D. Van Dyck. Maximum-likelihood estimation of Rician distribution parameters. *IEEE Transactions on Medical Imaging*, 17(3):357–361, June 1998.
- [115] S.T. Smith. Covariance, subspace, and intrinsic Cramér-Rao bounds. *IEEE Transactions on Signal Processing*, 53(5):1610–1630, May 2005.
- [116] A. M. Steane. Error Correcting Codes in Quantum Theory. *Physical Review Letters*, 77:793–797, July 1996.
- [117] A.M. Steane. Enlargement of Calderbank-Shor-Steane quantum codes. *IEEE Transactions on Information Theory*, 45(7):2492–2495, Nov 1999.
- [118] Alexander L. Stolyar. Maxweight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic. *Annals of Applied Probability*, 14(1):1–53, 2004.
- [119] Charles Swannack. Systematic constructions of MIMO channel quantizers, 2010. <http://www.rle.mit.edu/sia/technology.htm>.
- [120] Charles Swannack, Elif Uysal-Biyikoglu, and Gregory W. Wornell. Low complexity multiuser scheduling for maximizing throughput in the MIMO broadcast channel. In *Proc. 42nd Annual Allerton Conf. on Communication, Control, and Computing*, Monticello, Illinois, September 2004.

-
- [121] Charles Swannack, Elif Uysal-Biyikoglu, and Gregory W. Wornell. Finding NEMO: Near mutually orthogonal sets and applications to MIMO broadcast scheduling. In *Proc. IEEE WIRELESSCOM 2005 : International Conference on Wireless Networks, Communications, and Mobile Computing*, Maui, Hawaii, USA, June 2005.
- [122] Charles Swannack, Elif Uysal-Biyikoglu, and Gregory W. Wornell. MIMO broadcast scheduling with limited channel state information. In *Proc. 43rd Annual Allerton Conf. on Communication, Control, and Computing*, Monticello, Illinois, September 2005.
- [123] Charles Swannack, Elif Uysal-Biyikoglu, and Gregory W. Wornell. Efficient quantization for feedback in MIMO broadcasting systems. In *Proc. The Asilomar Conference on Signals, Systems, and Computers*, Asilomar, California, Spetember 2006. (to appear).
- [124] Charles Swannack, Gregory W. Wornell, and Elif Uysal-Biyikoglu. MIMO broadcast scheduling with quantized channel state information. In *Proc. IEEE International Symposium on Information Theory*, Seattle, Washington, July 2006.
- [125] L. Tassiulas and A. Ephremides. Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. *IEEE Transactions on Automatic Control*, 37(12):1936–1948, December 1992.
- [126] I. E. Teletar. Capacity of multi-antenna Gaussian channels. *Eur. Trans. Telecom.*, 10:585–595, November 1999.
- [127] L.M.G.M. Tolhuizen and W.J. Van Gils. A large automorphism group decreases the number of computations in the construction of an optimal encoder/decoder pair for linear block code. *IEEE Transactions on Information Theory*, 34(2):333–338, March 1988.
- [128] David Tse and Pramod Viswanath. *Fundamentals of wireless communication*. Cambridge University Press, New York, NY, USA, 2005.
- [129] A. Vakili, A. F. Dana, M. Sharif, and B. Hassibi. Differentiated rate scheduling for MIMO Gaussian broadcast channels. In *Proc. 43rd Annual Allerton Conf. on Communication, Control, and Computing*, Monticello, Illinois, September 2005.
- [130] J. H. van Lint. *Introduction to coding theory*. Springer-Verlag, Berlin, second edition, 1992.
- [131] H. Vishwanathan and K. Kumaran. Rate scheduling in multiple antenna downlink wireless systems. In *Proc. 39th Annual Allerton Conf. on Communication, Control, and Computing*, Monticello, Illinois, 2001.
- [132] P. Viswanath, D.N.C. Tse, and R. Laroia. Opportunistic beamforming using dumb antennas. *IEEE Transactions on Information Theory*, 48(6):1277–1294, Jun 2002.
- [133] W. Weichselberger. *Spatial structure of multiple antenna radio channels*. PhD thesis, Institut für Nachrichtentechnik und Hochfrequenztechnik, Vienna University of Technology, Vienna, Austria, 2003.

-
- [134] W. Weichselberger, Markus Herdin, Huseyin Ozcelik, and Ernst Bonek. A stochastic MIMO channel model with joint correlation of both link ends. *IEEE Transactions on Wireless Communications*, 5(1):90–100, 2006.
- [135] Lloyd R. Welch. Lower bounds on the maximum cross correlation of signals. *IEEE Transactions on Information Theory*, 20(3):397 – 399, May 1974.
- [136] David R. Wood. On the maximum number of cliques in a graph. *Graphs and Combinatorics*, 23(3):337–352, June 2007.
- [137] Pengfei Xia and G.B. Giannakis. Design and analysis of transmit-beamforming based on limited-rate feedback. *IEEE Transactions on Signal Processing*, 54(5):1853–1863, May 2006.
- [138] Pengfei Xia, Shengli Zhou, and G.B. Giannakis. Achieving the Welch bound with difference sets. *IEEE Transactions on Information Theory*, 51(5):1900– 1907, May 2005.
- [139] Wenjie Xu, Seyed A. Zekavat, and Hui Tong. A novel approach for spatially correlated multi-user MIMO channel modeling: Impact of surface roughness and directional scattering. In *Forty-Fifth Annual Allerton Conference*, 2007.
- [140] T. Yoo and A. J. Goldsmith. Optimality of zero-forcing beamforming with multiuser diversity. In *Proc. IEEE International Conf. on Communications(ICC)*, May 2005.
- [141] T. Yoo and A. J. Goldsmith. Sum-rate optimal multi-antenna downlink beamforming strategy based on clique search. In *Proc. IEEE Globecom 2005*, November 2005.
- [142] Taesang Yoo and A. Goldsmith. On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming. *IEEE Journal on Selected Areas in Communications*, 24(3):528–541, March 2006.
- [143] Jianzhong Zhang and Anthony Reid. Improved Hochwald construction of unitary matrix codebooks via eigen coordinate transformations. US Patent 11/119,513, April 2005.
- [144] Shengli Zhou, Zhengdao Wang, and G.B. Giannakis. Quantifying the power loss when transmit beamforming relies on finite-rate feedback. *IEEE Transactions on Wireless Communications*, 4(4):1948–1957, July 2005.
- [145] Mo Willems. *Knuffle Bunny: A Cautionary Tale*. Hyperion Book, New York, NY 2004.
- [146] Peggy Rathmann. *Good Night, Gorilla*. Putnam Juvenile, New York, NY 2000.