DATA COMPRESSION
A COVQ-Based Image Coder for Channels with Bit Errors and Erasures .......................... 161
Tomas Andersson and Mikael Skoglund

FADING/EQUALIZATION
Method-of-Moments Parameter Estimation for Compound Fading Processes .......................... 166
Frantz Bouchereau and David Brady

MODULATION/Detection
Adaptive Ordering for Imperfect Successive Decision Feedback Multiuser Detection .................. 173
Vladimir D. Trajkovic, Predrag B. Rapajic, and Rodney A. Kennedy
Variable Rate and Variable Power MQAM System Based on Bayesian Bit Error Rate and Channel Estimation Techniques .......................................................... 177
Lay Teen Ong, Mohammad Shikh-Bahaei, and Jonathon A. Chambers
Second-Order Statistics for Diversity-Combining of Non-Identical Correlated Hoyt Signals .......... 183
Gustavo Fraidenraich, Michel Daoud Yacoub, José Ricardo Mendes, and J. C. S. Santos Filho
Performance of a CDMA System Employing AMC and Multicodes in the Presence of Channel Estimation Errors ...... 189
Raymond Kwan and Cyril Leung

TRANSMISSION SYSTEMS
Capacity of Large Array Receivers for Deep-Space Communications in the Presence of Interference ............ 194
Victor Vilnrotter and Meera Srinivasan

TRANSACTIONS PAPERS
CODING
Generalized LDPC Codes and Generalized Stopping Sets .................................................. 201
Nenad Miladinovic and Marc P. C. Fossorier

FADING/EQUALIZATION
Soft Handover Overhead Reduction by RAKE Reception with Finger Reassignment ......... 213
Seyeong Choi, Mohamed-Slim Alouini, Khalid A. Qaraqe, and Hong-Chuan Yang

MODULATION/Detection
Novel GLRT Packet-Data Receivers ........................................................................ 222
Haoli Qian, Stella N. Batalama, and Bruce W. Suter
Unitary Peak Power Reduction in Multiple Transmit Antennas ........................................... 234
Heechoon Lee and Michael P. Fitz
Frequency Domain Detectors for Ultra-Wideband Communications in Short-Range Systems .................................................. 245
Simone Morosi and Tiziano Bianchi

(Table of Contents continued on back cover)
MULTIPLE ACCESS
Proportional QoS Adjustment for Achieving Feasible Power Allocation in CDMA Systems ............................................ 254
Rudolf Mathar and Anke Schmeink
The Impact of Space Division Multiplexing on Resource Allocation: A Unified Treatment of TDMA, OFDMA and CDMA. .................................................................................................................. 260
Iordanis Koutsopoulos and Leandros Tassiulas
A Practical Algorithm for Power Minimisation in Wireless Networks by Means of Multi-Hop and Load Partitioning ...... 270
Lorenzo Piazzo

OPTICAL COMMUNICATIONS
Power Reduction Techniques for Multiple-Subcarrier Modulated Diffuse Wireless Optical Channels .......................... 279
Weiwei Kang and Steve Hranilovic

SPREAD SPECTRUM
A Comparison of Frequency-Division Systems to Code-Division Systems in Overloaded Channels. ................................. 289
Joon Ho Cho, Qu Zhang, and Lixin Gao

SYNCHRONIZATION
Serial Search Code Acquisition Using Smart Antennas with Single Correlator or Matched Filter ................................. 299
Henri Puska, Harri Saarnisaari, Jari Iinatti, and Pekka Lilja

TRANSMISSION /RECEPTION
Performance Analysis for MIMO Systems with Lattice-Reduction Aided Linear Equalization ................................. 309
Xiaoli Ma and Wei Zhang
Transactions Letters

A COVQ-Based Image Coder for Channels with Bit Errors and Erasures

Tomas Andersson and Mikael Skoglund

Abstract—We illustrate how channel optimized vector quantization (COVQ) can be used for channels with both bit-errors and bit-erasures. First, a memoryless channel model is presented, and the performance of COVQ’s trained for this channel is evaluated for an i.i.d. Gaussian source. Then, the new method is applied in implementing an error-robust sub-band image coder, and we present image results that illustrate the resulting performance. Our experiments show that the new approach is able to outperform a traditional scheme based on separate source and channel coding.

Index Terms—Binary erasure channels, binary symmetric channels, channel-optimized quantization, image coding, joint source-channel coding.

I. INTRODUCTION

F OR packet data networks where parts of the overall transmission are over wireless links, phenomena occur that are not present in traditional wired networks (over optical fiber and/or cable links). In particular, it is not a reasonable assumption to neglect bit-errors over wireless channels. In addition to bit-errors, source coder robustness relates to its sensitivity to packet-loss, which may occur in the network due to overload (or, sometimes, over wireless paths due to detected bit-errors in packets that are then declared lost).

A famous result from information theory, known as the source–channel separation theorem, states that there is no loss in treating source and channel coding as two separate problems [2]. It is however important to notice that in the general case, the separation can be made without loss only in the limit of infinite delay and coding complexity; a fact that has been frequently pointed out to motivate the use of combined source–channel codes [3]–[7]. Many current source coders employ only error concealment to make the coder robust, while we emphasize the use of combined source–channel coding, with a focus on techniques based on channel optimized vector quantization (COVQ).

The concept of COVQ originates in [3]–[7], and the use of COVQ has since been proposed in many different contexts. However, although much has already been said on the subject, most of the previous works have been focused towards channels with bit-errors, such as the binary symmetric channel. In contrast, the main contributions of the present work are an extension to channels with both bit-errors and bit-erasures, and a demonstration of how the new COVQ technique can be implemented to enhance the performance of a subband image coder. Related work on trellis-based source coding for the binary erasure channel can be found in [8].

II. THE BINARY SYMMETRIC ERASURE CHANNEL

We consider channels where both erasures and bit-errors occur. Such channels can arise, e.g., when information is communicated over several consecutive channels with different properties. Consider for instance the case where a packet-switched network is used for long distance communication, but local access is made through a wireless link. The network may fail to deliver a packet on time, causing packet erasures. The wireless link, on the other hand, is prone to introduce bit-errors.

A wireless link, as defined by its physical properties, modulation, non-perfect channel coding, etc., can be replaced by an equivalent discrete channel. Often this discrete channel can be modeled as a binary symmetric channel (BSC) with a transition probability \( \alpha \), corresponding to the bit-error rate (BER) of the channel. In packet loss channels, erasures come in groups of many bits. We assume however, that bit-level interleaving, using a pseudo-random spreading sequence, is implemented in such a way that all bits in a symbol are transmitted in different packets. This way, from a symbol perspective, independent packet losses are turned into independent bit-erasures. The resulting channel is the binary erasure channel, defined by the erasure probability \( \beta \) which will be referred to as the bit-loss rate (BLR). Note that in our case the BLR is equal to the packet loss rate (PLR).

If the binary symmetric channel and the binary erasure channel are concatenated, the binary symmetric erasure channel (BSEC), depicted in Fig. 1, is obtained. For the BSEC, bits are complemented with probability \( \alpha \) and lost with probability \( \beta \). Thus, the probability of a correctly received bit is \( (1 - \alpha)(1 - \beta) \), the probability of a complemented bit is \( \alpha(1 - \beta) \) and the probability of an erasure is \( \beta \). Note that both the binary symmetric channel and the erasure channel are obtained as special cases of the BSEC.

In practical scenarios the BSEC is naturally a quite crude model. For example the channel parameters may vary during the transmission of a set of interleaved packets, since packets may follow different routes (in this case, \( \alpha \) and \( \beta \) model...
the average behavior). Still, the BSEC serves as a useful approximation in a scenario where both errors and losses are present. The parameters $\alpha$ and $\beta$ can be estimated using similar techniques as those used for related Markov models for Internet transmission [9].

III. CHANNEL OPTIMIZED VQ FOR THE BSEC

While general results for COVQ are well known [4], [6], [7], some basic results are repeated here for convenience. In general, a vector quantizer (VQ) is defined by two basic operations, the encoder and decoder. The encoder, $\varepsilon(\cdot)$, transforms a source vector, $\mathbf{X} \in \mathbb{R}^k$, into a quantization index, $I = \varepsilon(\mathbf{X})$, $I \in \{0, 1, \ldots, M - 1\}$. Assuming $M = 2^L$, this corresponds to sending $L$ bits per source vector, that is, the rate is $R = L/k$. The encoder operation is defined by a partitioning, $\mathcal{P} = \{S_0, S_1, \ldots, S_{M-1}\}$, of $\mathbb{R}^k$ such that $\varepsilon(\mathbf{x}) = i$, iff $\mathbf{x} \in S_i$. The decoder, $\delta(\cdot)$, is a mapping from a finite set of integers to an associated set of vectors, $\mathbf{Y} = \delta(J)$, $J \in \{0, 1, \ldots, N - 1\}$. The set of reconstruction vectors, $\mathcal{C} = \{\mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_{N-1}\}$, $\mathbf{y}_j \in \mathbb{R}^k$, is called the decoder codebook.

Suppose that the index $I = \varepsilon(\mathbf{X})$ is sent over a noisy channel, and that $J$ is observed at the receiver. Assume also that the distortion measure $d(\mathbf{x}, \mathbf{y})$ associated with mapping an input vector, $\mathbf{x}$, into an output vector, $\mathbf{y}$, is given by the squared Euclidean distance, i.e.,

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^2.$$ 

Then necessary conditions for minimizing the expected distortion, $D(\mathcal{P}, \mathcal{C}) = E[d(\mathbf{X}, \mathbf{Y})]$, are:

$$S_i = \left\{ \mathbf{x} : \sum_{j=0}^{N-1} P(j|i) \|\mathbf{x} - \mathbf{y}_j\|^2 \leq \sum_{j=0}^{N-1} P(j|i') \|\mathbf{x} - \mathbf{y}_{j'}\|^2, \forall i' \neq i \right\}$$

(1)

and

$$\mathbf{y}_j = E[\mathbf{X}|J = j] = \frac{\sum_{i=0}^{M-1} \text{Pr}(I = i)P(j|i)\mathbf{c}_i}{\sum_{i=0}^{M-1} \text{Pr}(I = i)P(j|i)}$$

(2)

where in (2) we defined the encoder centroids $\mathbf{c}_i = E[\mathbf{X}|I = i]$, and where $P(j|i) = \text{Pr}(J = j|I = i)$ are the transition probabilities of the channel. Generally [4], [6], [7], COVQ design is based on iterating between (1) and (2) until convergence to a (local) optimum in terms of a stationary point of $D(\mathcal{P}, \mathcal{C})$.

The precise model assumed for the discrete channel influences the design only through the transition probabilities $P(j|i)$. When transmitting $L$ bits, $b_L, \ldots, b_1$, over the BSEC, with $b_1$ the LSB and $b_L$ the MSB of the index $i$, we can define the output index $j$ as $j = \sum_{l=1}^{L} d_l 2^{L-1}$, where: no error $\Rightarrow d_l = b_l$, bit-error $\Rightarrow d_l = 1 - b_l$, and erasure $\Rightarrow d_l = 2$. In this case, the transition probabilities can be computed easily, for example with $L = 2$, we get $P(7|0) = \{b_1 = 0 \rightarrow d_1 = 1 \text{ and } b_2 = 0 \rightarrow d_2 = 2\} = \alpha(1 - \beta)$, and $P(8|0) = \{b_1 = 0 \rightarrow d_1 = 2 \text{ and } b_2 = 0 \rightarrow d_2 = 2\} = \beta^2$, and so on.

Note that when transmitting over the BSEC, the decoder codebook is larger than the number of encoder indexes. More specifically, $N = 3^L > M = 2^L$. This, effectively, results in soft decision source decoding at the receiver, c.f. [13], [14]. The soft information used by the decoder stems from the additional erasure output symbol of the BSEC. We stress that under the assumptions made and since the decoder codebook is optimal subject to these assumptions, the decoder utilizes the available soft information in an optimal manner. An alternative interpretation of COVQ’s trained for the BSEC, is in terms of multiple description coding. Each received bit can be used to decrease the distortion of the reproduced value, independently of which other bits are received. Especially in the case $\alpha = 0$, $\beta > 0$ (no bit-errors, only erasures) we see that, as in multiple description quantization, each bit is carried on a “channel” that either works without error or is turned off, corresponding to descriptions (bits) which are either received or lost. Hence, training an $L$-bit COVQ for the BSEC is similar to designing a multiple description VQ for $L$ channels. Related recent work uses channel optimized scalar quantizers to design multiple description coders for the two channel case [15], see also [16] for an extension to VQ and multiple channels.

The basic principle of optimizing a COVQ jointly for bit-errors and erasures, that is for $P(j|i)$’s corresponding to the channel in Fig. 1, is illustrated in Fig. 2. The figure shows the performance of COVQ’s trained for uncorrelated Gaussian data with unit variance. An ordinary VQ, trained with the splitting algorithm [17] followed by the binary switching

![Fig. 1. Transition probabilities of the BSEC.](image1)

![Fig. 2. Performance of COVQ over different channel parameters.](image2)
algorithm [18], is used to initialize the COVQ training. In this example, the rate is 2 bits/dimension for 3-dimensional data, and the channel is perfectly matched to the training parameters. The $x$- and $y$-axes correspond to different bit error and bit loss probabilities $\alpha$ and $\beta$ respectively, and the $z$-axis shows performance in terms of reproduced signal-to-noise ratio $E\|X\|^2/E[d(X,Y)]$.

IV. A VQ-BASED SUBBAND IMAGE CODER

We have constructed a simple subband image coder using vector quantization of the image subband samples. VQ-based subband image coding can be done in a variety of ways; a good overview can be found in [19]. Our coder uses a four-level pyramid subband decomposition [20]–[22], giving a total of 13 different subbands. Vectors are formed within subbands (no crossband vectors) and each subband is assigned a certain bit rate according to the “equal-slope” method (c.f. [19]). The output indices from the VQ’s are put directly in the bit stream, without any additional entropy- or channel coding (except for bit-level interleaving), resulting in a fixed bit rate. Although better compression is possible by using variable-rate entropy coding, we motivate our fixed bit-rate structure by its avoidance of error propagation.

Designing COVQ’s for the described image coder requires the use of training data. The approach we use to generate training data is to fit a Gaussian mixture model to the empirical probability density function of vectors from each subband [23]. The model parameters are estimated from a set of images not including the test images. The estimated models can then easily be applied in generating an arbitrary amount of training data. The main reason for using a model of the source distribution instead of training on image data directly, is to avoid over-fitting the COVQ’s to a small set of image data.

V. IMAGE RESULTS

Fig. 4 shows the performance of a subband coder based on ordinary VQ compared with two COVQ-based subband coders. In all three cases the simulated channel is the BSEC with 10% erasures and 5% bit errors. The COVQ’s used in Fig. 4(b) are trained for a BSC with 5% bit errors, without knowledge about the erasures. The COVQ’s of Fig. 4(c) are trained according to our proposed scheme with full knowledge of the BSEC compared with assuming bit-errors only is clearly visible, but not as striking. Still this gain comes at almost no cost in increased complexity (the only essential difference is in storing larger decoder codebooks).

VI. COMPARISON WITH FORWARD ERROR CORRECTION

To compare the performance of the COVQ approach with the performance of a more traditional one, the same image coder structure is used, but with VQ’s optimized for the source statistics only (as given by the Gaussian mixture models). The resulting bit stream from the image coder is then further encoded using BCH codes for protection against errors and erasures. Using forward error correction (FEC) increases the total bit rate, so to get a fair comparison, the number of bits allocated to the image coder has to be reduced until the total transmission rate is the same for COVQ and VQ+FEC.

Table I

<table>
<thead>
<tr>
<th>Subband number</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vector size</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>VQ rate</td>
<td>8</td>
<td>10</td>
<td>10</td>
<td>9</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>COVQ rate</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>9</td>
<td>10</td>
<td>10</td>
<td>9</td>
<td>7</td>
<td>7</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Fig. 3. Performance of COVQ compared with VQ+FEC.

![Fig. 3. Performance of COVQ compared with VQ+FEC.](image)
optimized for one specific point on the curve. The flat part of the curves for FEC corresponds to the case when the BCH code can correct all errors and erasures. When the code breaks down, the performance drops rapidly. COVQ shows a smoother degradation, observed by others as being typical of COVQ.

Note that it should always be possible to find a COVQ that performs equally well or better than VQ+FEC for any given channel (and a certain channel block length). Such a COVQ can be found trivially, by using VQ+FEC as the initial encoder partitioning in the COVQ training. Since training only can improve performance, one iteration is sufficient to guarantee equally good or better performance (at the design point). Furthermore, the scheme based on COVQ will in general be more robust toward mismatch in assumptions about the channel parameters.

VII. SUMMARY

We have presented a simple but important extension of previous work on COVQ for binary memoryless channels with bit-errors to channels with both errors and erasures. Such channels are motivated by a scenario where parts of a network connection are over wireless links. Bit-errors stem from transmission errors and bit-erasures stem from packet losses. An
implicit assumption is that no retransmissions are utilized in the network, e.g., due to strong real-time requirements. Hence, packets potentially recognized to contain bit-errors are not declared lost. Instead, packet losses are assumed to be due to other phenomena, such as network overload.

The results show that COVQ is a useful tool for designing joint source-channel codes for the binary symmetric erasure channel (BSEC). In particular, advantages of joint source-channel coding over separate source and channel coding were demonstrated.

REFERENCES
Method-of-Moments Parameter Estimation for Compound Fading Processes

Frantz Bouchereau, Member, IEEE, and David Brady, Member, IEEE

Abstract—A novel and general parameterized fading model for the instantaneous received path power is presented, which accounts for both wide-sense stationary shadowing and small-scale fading. A method-of-moments parameter estimator is derived and implemented using a non-linear least-squares algorithm. Performance is explored by analysis and simulation under different noisy channel scenarios, and compared to exact Cramer-Rao bounds in the case of uncorrelated shadowing, and to idealized CR bounds, where shadowing and small-scale fading processes are assumed to be observed separately, in the case when shadowing correlation was not negligible.

Index Terms—Compound fading, gamma-lognormal processes, and method of moments.

I. INTRODUCTION

Accurate modeling of the fading process and estimation of the corresponding fading parameters are fundamental problems in radio system engineering. Their solutions enable the determination of outage statistics [1], appropriate diversity techniques, and adaptation rates for receiver structures. Some modeling methods involve estimation of parameters for a family of distributions rich enough to include multiple fading models. Compound fading models were first proposed to explain the transition between small-scale (local) fading and shadowing [2]. A byproduct of this work was a class of distributions which included Rayleigh, Lognormal, and mixed fading models. Recently, this family of distributions was further enlarged to include Nakagami-m small-scale fading (see [3] and references [1-10] in [4]). Maximum likelihood (ML) based estimation of the parameters for this family of distributions, hereafter referred to as the Gamma-Lognormal family (GL), has been addressed under the assumptions of block-constant shadowing with known block coherence length and constant mean [4]. An extension to this work, presented in [5], lifted the block-constant shadowing assumption by considering a first-order autoregressive model (AR(1)) for the shadowing process in decibels (dB). This model limits the shadowing process to have a known zero dB mean and a strictly exponential correlation function. Two estimation procedures for the unknown AR(1) parameters were presented in [5]. The first one, denoted as FB/AML, consists in iterating between estimating a shadowing time series using forward-backward (FB) prediction and then using asymptotic maximum likelihood (AML) estimates for the AR(1) correlation coefficient and the innovation variance. The second one, denoted as EL, consists of shadowing time series prediction, substitution of the predicted time series into the log-likelihood function of the AR(1) process (see reference 24 in [5]), and maximization of this approximate log-likelihood function with respect to the AR(1) parameters. Both FB/AML, and EL estimators require knowledge of the small-scale fading parameter $m$ which is found using the ML procedure based on the assumption of block-constant shadowing with known block coherence length described in [4].

In this work we extend the GL family to include shadowing time series whose logarithm is a realization from a wide-sense stationary (WSS) Gaussian random process with an arbitrary unknown mean and an arbitrary finite-parameter autocovariance function, and derive a simple method-of-moments (MOM) estimator for the model parameters. This extension enables accurate modeling of non independent and identically distributed (i.i.d.) observations of received path powers and WSS shadowing. Our model is equivalent to that in [5] in the special case when the shadowing Gaussian process has zero dB (and known) mean and an exponential autocovariance function. In this special case we will compare the performance of the parameter estimation algorithms presented in [5] to the performance of the MOM estimator. The proposed GL family of distributions contains the path power distributions corresponding to Lognormal, Nakagami, Rayleigh, half-Gaussian, and Suzuki fading, as well as a rich set of compound fading processes.

This paper is organized as follows: Section II presents the GL model and moment equations. Section III presents the MOM estimator, and analytical expressions and convergence analysis for the mean-squared error of the sample moments. Section IV presents numerical analysis of bias and MSE performance of the MOM estimates.

II. GAMMA-LOGNORMAL FADING PROCESSES

In this section we present a brief overview of the proposed GL family of distributions. In this setting we envision a uniformly spaced set of observations of instantaneous received powers from a flat fading environment. For each $t = 1, \ldots, N$, we denote the observation of the instantaneous path power as $y_t$. The entire observation will be denoted by the $N$-dimensional column vector $y$. The distribution of $y$ may be described by first considering the continuous time, WSS Gaussian random process $z_t$ with mean $\mu$ and autocovariance function $E[z_{t} z_{t+\tau}] = \sigma^2 R_\alpha(\tau)$. It is assumed that $R_\alpha(\tau)$ has a maximum of unity and that parameter vector
TABLE 1
REGIONS OF \{ 2 \, 1 \, 1 \} FOR CLASSICAL FADING MODELS

<table>
<thead>
<tr>
<th>Fading Model</th>
<th>2</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>GL</td>
<td>R</td>
<td>R+</td>
</tr>
<tr>
<td>Rayleigh</td>
<td>R</td>
<td>0</td>
</tr>
<tr>
<td>Lognormal</td>
<td>R+</td>
<td>0</td>
</tr>
<tr>
<td>Nakagami</td>
<td>R+</td>
<td>0</td>
</tr>
<tr>
<td>Suzuki</td>
<td>R+</td>
<td>0</td>
</tr>
<tr>
<td>Half-Gaussian</td>
<td>R+</td>
<td>0</td>
</tr>
</tbody>
</table>

a has range \( A \). The distribution of the random process \( \{ z_t \} \) is completely characterized by \( \text{dim}(a) + 2 \) real parameters. In physical terms, \( \{ e^{x_t}, \; t = 1, \ldots, N \} \) denotes samples of the Lognormal shadowing process. We also define a vector \( x \) of i.i.d. Gamma random variables which is independent of the vector \( z \). Here, the \( l \)-th component of the vector \( x \) has unit mean and variance \( 1/m \). The instantaneous path power \( y_t \) is statistically equivalent to the product \( e^{x_t} z_t \), which exhibits a multiplicative effect of small-scale fading and shadowing. Conditioned on \( z \), the components of \( y \) are mutually independent. Letting \( f_a(z) \) denote the \( N \)-dimensional multivariate normal density function with common mean \( \mu \) and covariance matrix \( \Lambda_{i,j} = \sigma^2 R_{a}(t_i - t_j) \), the likelihood function of the observation takes the form

\[
L(\mu, \sigma^2, m, a) = \left[ \frac{m^m}{\Gamma(m)} \right]^N \int_{\mathbb{R}^N} f_a(z) \times \exp \left\{ \sum_{l=1}^N (m - 1) \log y_t - mz_l - m \psi(t) \right\} dz.
\]

(1)

To simplify notation, we have not defined the Lognormal distribution using the classical dB \((10 \log_{10}(\cdot))\) definition. Instead, we have chosen to use the natural logarithm without any scaling factor. To convert the shadowing standard deviation values used in this work to corresponding values in dB we apply the simple conversion \((\sigma \text{ in dB}) = 4.3 \sigma \).

Three real parameters \( \{ \mu, \sigma^2, 1/m \} \) determine the marginal distribution of the instantaneous power \( y_t \). Table I illustrates the regions for this parameter set which corresponds to classical fading models. Note that classical fading models usually refer to the distribution of the envelope, not that of the instantaneous power.

Defining \( \tilde{y}_l = \frac{y_l}{E[y_l]} \), it is not difficult to find the following moments:

\[
E[\tilde{y}_l] = Q_r(m) \exp \left[ \frac{\sigma^2 (r^2 - r)}{2} \right],
\]

(2)

\[
E[\tilde{y}_l \tilde{y}_{l+r}] = \exp \left[ \sigma^2 R_{a}(r) \right] ; \; \tau \neq 0,
\]

(3)

\[
E[\log(y_l)] = \psi(m) - \log(m) - \frac{\sigma^2} {2},
\]

(4)

\[
E[\log(y_l)] = \mu + \psi(m) - \log(m),
\]

(5)

where \( Q_r(m) = \Gamma(m + r)/\Gamma(m)m^r \), and \( \psi(m) \) denotes the Digamma function defined as \( \psi(m) = \frac{1}{\Gamma(m)} \left[ \frac{d}{dm} \frac{\Gamma(m)}{m} \right] \).

1\text{log denotes the natural logarithm in this work.}

III. METHOD-OF-MOMENTS PARAMETER ESTIMATION

In this section we propose a simple method to estimate the parameters of the GL model. We use the moment equations (2), (3) and (4) to construct the non-linear least-squares (NLLS) problem

\[
\hat{\theta} = \arg \min_{\theta \in \mathbb{R}^2 \times A} \left\{ \| f(\hat{\theta}) - g(\hat{\theta})^T h(\theta) \| \right\}^2,
\]

(6)

where \( \theta = [1/m \; \sigma^2/2] \) and \( \hat{\theta} \) is the vector of unknown parameters and \( \hat{\theta} \), its estimate counterpart. For any collection of \( R \) powers \( \{ r_1, \ldots, r_R \} \) and \( K \) correlation lags \( \{ \tau_1, \ldots, \tau_K \} \), the differences between model vectors \( f(\hat{\theta}), g(\hat{\theta}), h(\hat{\theta}) \) and measurement vectors \( \hat{f}, \hat{g}, \hat{h} \) are given as

\[
\hat{f}_i - f(\hat{\theta}) = \frac{\omega_i}{N} \sum_{t=1}^N \tilde{y}_t - \omega_i, Q_r(m) \exp \left[ \frac{\sigma^2}{2} (t_i^2 - r) \right],
\]

(7)

\[
\hat{g}_j - g(\hat{\theta}) = \frac{\omega_j}{N} \sum_{t=1}^N \tilde{y}_t \tilde{y}_{t+r} - \omega_j \exp \sigma^2 R_{a}(r),
\]

(8)

\[
\hat{h}_i - h(\hat{\theta}) = \frac{\omega_i}{N} \sum_{t=1}^N \tilde{y}_t \tilde{y}_w - \omega_i \psi(m) - \log(m) - \sigma^2/2.
\]

For \( i = 1, \ldots, R, \; j = 1, \ldots, K, \) and where \( \hat{f}_i, f(\hat{\theta}), \tilde{g}_j, \; \tilde{g}(\hat{\theta}) \) are the \( i \)-th and \( j \)-th components of vectors \( \hat{f}, f(\hat{\theta}), \hat{g}, \; \tilde{g}(\hat{\theta}) \) respectively. \( \tilde{y}_w = y_i / (\sum y_i) \), and \( \{ \omega_1, \omega_2, \omega_3 \} \) are weights used to emphasize each equation. From (5), the shadowing mean parameter \( \mu \) can be estimated as \( \hat{\mu} = \frac{1}{N} \sum_{t=1}^N \log y_t - \omega_i \tilde{y}_w + \log(\tilde{y}_w) \).

In this work, the NLLS problem presented in equations (6), and (7) has been solved using a subspace trust region method based on the interior-reflective Newton algorithm described in [7].

The performance of the proposed MOM estimation depends on the bias and variances of the sample moment estimators. It is straightforward to show that the moment estimators are unbiased.\(^2\) However due to the non i.i.d. nature of \( Y \), the rate at which the estimates approach their limits is less obvious.

For the power moments it is possible to show that

\[
\text{Var} \left[ \frac{\sum_{t=1}^N \tilde{y}_t}{N} \right] = \frac{Q_r(m)^2 \exp(2\mu R + \sigma^2 r^2)}{N} \times \left[ 1 + \frac{1}{N} \sum_{t=1}^N \sum_{r=1}^N \exp \left[ \sigma^2 r^2 \right] \right],
\]

(8)

For large \( N \), and under the weak assumption that \( \lim_{t \to \infty} R_{a}(s) = 0 \), the variance of the sample \( r \)-th power moments behaves as

\[
\lim_{t \to \infty} \text{Var} \left[ \frac{1}{N} \sum_{t=1}^N \tilde{y}_t \right] \approx Q_r(m)^2 \exp(2\mu R + \sigma^2 r^2) \times \left[ 1 + \lim_{t \to \infty} \int_1^w \frac{1}{1 + \delta(x)} \left[ \frac{\exp(\sigma^2 x^2) - 1}{\epsilon N^2} \right] dx \right],
\]

(9)

where \( \epsilon \) is a constant that depends on \( a \). Then, it follows from (9) that the power moment estimation variance vanishes as \( O(\theta) \).

Variances for \( \log y \) and correlation (with \( N \geq 2\tau \)) sample moments were found to be respectively given by equations 10, and 11.

\[
\text{Var} \left[ \frac{1}{N} \sum_{t=1}^N \log y_t \right] = \frac{1}{N^2} \sum_{t=1}^N \sum_{r=1}^N \sigma^2 R_{a}(t - r) + \frac{1}{N} \sum_{k=0}^{\infty} \frac{1}{(m+k)^2}.
\]

(10)

\(^2\text{Notice that the use of unbiased sample-moment estimators does not guarantee that the solution to the NLLS problem will yield unbiased parameter estimators.} \)
Following an asymptotic analysis similar to the one used for the sample power moments, it can be shown that (10), and (11) also vanish as $O(1/N)$.

Note that these variance calculations are indirectly related to the MOM estimation accuracy. While the moment estimation accuracy could be related to the MOM performance by the Jacobian, this would only yield an asymptotic result. Instead, we present bias and mean-squared error (MSE) performance curves for the MOM estimator via numerical calculations in the following section.

**IV. Numerical Results**

In this section we set $a$ to be a scalar, and let $R_a(\tau) = e^{-a|\tau|}$, and $A = \mathbb{R}_+$. The constraints for the optimization procedure were set to: $m \in [0.5, 15]$, $\sigma^2 \in [0.001, 5]$, and $a \in [0.001, 30]$, and the initial search points were set to the minimum values in these intervals. A set of $R = 10$ power moments was chosen as $r_i \in \{-0.25, -0.2, -0.1, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.5\}$. A set of $K = 10$ correlation lags was chosen as $\tau_i \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ samples. Finally, to obtain equally weighted equations, all weights were set to unity except $\omega$, which was set to ten. The power moments, correlation lags, and weights were chosen based on the best bias and MSE performance observed over several combinations of these tuning parameters. Note however that this set of values may not correspond to the optimal choice.

Ten thousand realizations of the MOM algorithm were obtained to analyze bias and MSE performance of the GL parameter estimates. True parameter values were set to $\mu = -3$, $\sigma^2 = 0.005$, $a = 0.1$, and $1/m = 1$ to obtain Rayleigh observations, to $\mu = -3$, $\sigma^2 = 1$, $a = 0.1$, and $1/m = 0.1$ to obtain Lognormal observations, and to $\mu = -3$, $\sigma^2 = 1$, $a = 0.1$, and $1/m = 1$ to obtain Suzuki observations. As stated in Section II, in this work $\sigma^2 = 1$ is equivalent to a shadowing variance of 18.5 in dB.

**A. Performance in Different Channel Scenarios**

In this section we analyze the bias and MSE performance of the MOM estimator in the previously described Rayleigh, Lognormal and Suzuki channel scenarios. Further, to analyze robustness of the MOM to noise, bias and MSE results were also obtained in the case where complex additive white Gaussian noise (AWGN) was added to the complex envelope observations. The complex AWGN variance $\sigma^2_{AWGN}$ value was adjusted to achieve, in all channel scenarios, an average signal to AWGN power ratio $E[y_k]/\sigma^2_{AWGN}$ of 7 dB. Table II summarizes the bias findings while Figure 1 presents MSE curves versus number of observations.

From Table II we observe that estimates of $\mu$ have a fairly small bias in all three channel scenarios. Estimates of $\sigma^2$, and $\sigma$ have a moderate bias in the Lognormal and Suzuki cases and fairly large bias in Rayleigh scenarios. Bias for estimates of $1/m$ is large for the Lognormal scenario and small for the Rayleigh and Suzuki cases. Finally, as expected, bias increases in the presence of noise but the effects are moderate.

Figure 1 shows that the MSE performance for estimates of $\mu$ is fairly insensitive to noisy observations except for the Rayleigh case. Notice that the MOM has comparable MSE performance in the Lognormal and Suzuki scenarios. The best MSE performance for estimates of $\mu$, $\sigma^2$, and $1/m$ is achieved under the Rayleigh scenario.

**B. Effects of Shadowing Coherence Time**

The same type of numerical calculations described in Section IV-A were now performed keeping the number of observations constant at $N = 1000$ and letting the shadowing coherence parameter take values from a grid in $a = [0.01, 1]$. The results of these calculations are presented in Figure 2.

Observe that, in the Rayleigh scenario, MSE performance for estimates of $\mu$, $\sigma^2$, $a$, and $1/m$ is invariant to changes in the shadowing covariance parameter $a$ as expected since Rayleigh has no Lognormal components.

**C. Lower Bounds on Performance**

In this section we numerically calculate Cramér-Rao (CR) bounds for unbiased estimates of GL parameters. Recall that the MOM estimator is biased. Then, comparison of MSE curves with respect to CR bound curves will simply yield lower bounds on the relative performance of the proposed MOM estimator to that of efficient, unbiased estimators.

1) Uncorrelated Shadowing: When $a >> 1$ the shadowing process becomes uncorrelated. Then our GL model is equivalent to the GL model presented in [4] when the constant shadowing block coherence length is equal to one sample. For this scenario, the gradient of the log-likelihood function presented in [4] is equivalent to the gradient of the likelihood function for our GL model. The gradient equations involve several sums and integrals that can be solved numerically assuming that $a$ is known and using Hermite polynomials [6]. Empirical CR bounds were found by plugging a large number of GL realizations into the gradient equations, solving the integrals numerically (Hermite polynomials of order 20 were used for this purpose), obtaining gradient outer products, and averaging [8]. Neglecting the Hermite integration approximation errors, the empirical CR bounds would become exact if an infinite number of averaging points were used in the calculations (in this work ten thousand GL observations were used for averaging).

MSE curves for estimates of $\mu$, $\sigma^2$, and $1/m$ were obtained for Rayleigh, Lognormal and Suzuki scenarios with $a = 30,$
TABLE II
BIAS FOR MOM parameter estimates

<table>
<thead>
<tr>
<th>Fading model</th>
<th>Bias ((N = 200, 2500))</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rayleigh noiseless</td>
<td>-0.024 -0.004</td>
<td>1.815 1.104</td>
</tr>
<tr>
<td>Rayleigh noisy</td>
<td>0.074 0.061</td>
<td>1.853 1.170</td>
</tr>
<tr>
<td>Lognormal noiseless</td>
<td>0.089 0.045</td>
<td>-0.303 -0.119</td>
</tr>
<tr>
<td>Lognormal noisy</td>
<td>0.159 0.123</td>
<td>-0.340 -0.147</td>
</tr>
<tr>
<td>Suzuki noiseless</td>
<td>0.055 0.02</td>
<td>-0.265 -0.056</td>
</tr>
<tr>
<td>Suzuki noisy</td>
<td>0.153 0.117</td>
<td>-0.325 -0.129</td>
</tr>
</tbody>
</table>

Fig. 1. MSE curves for estimates of \(\mu, \sigma^2, a\), and \(1/m\) as a function of the number of observations \(N\) for Rayleigh, Lognormal and Suzuki scenarios. In all cases \(a = 0.1\). The dashed line curves correspond to GL observations with AWGN where \(10\log_{10}(\frac{E[y]}{\sigma^2_{AWGN}}) = 7\) dB. The solid line curves correspond to noiseless GL observations.

and these were compared to empirical CR bounds. Figure 3 presents these results. The validity of the CR bound calculations was examined by averaging the gradient realizations and verifying that the result was close to zero.

Recall that the CR bound calculations assume known \(a\) whereas the MOM estimator needs to estimate this unknown shadowing correlation parameter. Nevertheless the MSE and CR bound curves for the estimates of \(\mu, \sigma^2,\) and \(1/m\) are fairly close to each other in the Rayleigh and Lognormal scenarios. These curves are less similar in the Suzuki case. Notice that the MSE curve for estimates of \(1/m\) lies below the CR bound in the Rayleigh scenario. This behavior is possible due to the fact that the MOM estimator is a biased estimator.

2) Correlated Shadowing: When \(a\) is not large enough (i.e., correlated shadowing) empirical calculations of CR bounds become very difficult since one can not apply Hermite integration to calculate the gradient of the log-likelihood function in (1). In this section we consider an alternative method to lower bound the performance of unbiased estimators in the correlated shadowing.

Recall that the observation \(y\) is the result of a component-wise product of small-scale vector \(x\) and the shadowing vector \(e^z\). It is clear that parameter estimation could only be improved by the observation of \(x\) and \(z\) separately. In considering this idealized observation, easy bounds on performance may be obtained for the small-scale and shadowing processes separately. Notice that some of the bounds will be trivial and equal to zero when exactly one of the processes is deterministic (i.e., Nakagami fading or Lognormal fading), as that would amount to estimating a constant using non-random observations. These trivial cases will not be taken into account in our analysis. In any case, separability-based bounds provide a simple and convenient way to measure performance of the proposed MOM estimator. The ratio of these bounds to the
mean-squared error of unbiased estimators would yield a lower bound to Pittman efficiency. Since the MOM estimators are biased in general, we refer to this ratio as the lower bound to relative performance.

The CR bound for the separately observed shadowing parameters follows from the well-known result for the multivariate normal model [8] when \( z \) is observed, and will not be repeated here. The CR bound for the Nakagami fading factor \( 1/m \) when the vector \( x \) is observed separately is also readily determined from the CR bound for the \( m \)-parameter [9] as

\[
\text{CRB}_{1/m} = \frac{1}{N m^4} \left( \psi'(m) - \frac{1}{m} \right)^{-1},
\]

where \( \psi'(m) \) is the first-order derivative of the Digamma function [6].

Figure 4 presents the ratio of the idealized CR bound curves to MSE curves obtained in Section IV-A for Rayleigh, Lognormal and Suzuki scenarios (with \( a = 0.1 \)) and for the cases where the proposed bounding technique yields non-trivial results.

The lower bound on relative performance for the \( 1/m \) estimates in a Rayleigh scenario increases with the number of observations and is close to unity for all the values of \( N \). In all other cases, the lower bounds on relative performance decrease with the observation size. This is due, mainly, to the looseness of the bounding technique. For small and moderate sample sizes (\( N \in [100, 500] \)), the relative performance of the proposed MOM estimator is above 70\% for estimates of \( \mu \) in the Lognormal and Suzuki scenarios, above 50\% for the estimates of \( \sigma^2 \) in the Lognormal scenario, and above 30\% for the estimates of \( \sigma^2 \) in the Suzuki scenario. Relative performance for estimates of \( 1/m \) in the Suzuki scenario is only above 10\%. Finally, relative performance for estimates of \( a \) is above 20\% in both Lognormal and Suzuki scenarios. Recall that the relative performance is being measured with respect to idealized CR bounds so small relative performance percentages should not discourage the reader.

D. Comparison of MOM, AML/FB, and EL Parameter Estimators

As described in the introductory paragraphs of this work, our proposed GL model is equivalent to the GL model presented in [5] in the special case when the shadowing Gaussian process has zero mean and an exponential autocovariance function. Recall that in this section we have assumed that \( a \) is a scalar, and that \( R_a(\tau) \) is indeed exponential. Then the two models will be readily comparable if we set \( \mu = 0 \).

The GL model in [5] is given by the multiplicative process \( y_t = x_t 10^{\beta_t} \). The shadowing process \( \beta_t \) in dB is described by the AR(1) recursion \( \beta_t = \alpha \beta_{t-1} + \omega_t \), where \( \omega_t \) are independent, zero dB mean innovations with variance \( \sigma^2_{\omega} \). The small-scale fading process \( x_t \) is modeled as Nakagami-\( m \). It is straightforward to translate parameters \( \alpha \) and \( \sigma^2_{\omega} \) in [5] to our set of shadowing parameters \( a \), and \( \sigma^2 \). Numerical calculations were obtained in [5] to analyze the MSE performance of the FB/AML and EL estimators for the AR(1) parameters and of
Fig. 3. CR bound and MSE curves for the estimates of $\mu$, $\sigma^2$, and $1/m$ for GL observations with uncorrelated shadowing.

**TABLE III**

**COMPARISON OF MOM, FB/AML, EL AND ML ESTIMATORS**

<table>
<thead>
<tr>
<th>$m$</th>
<th>MOM</th>
<th>FB/AML</th>
<th>EL</th>
</tr>
</thead>
<tbody>
<tr>
<td>$= 1$</td>
<td>2.970</td>
<td>2.006</td>
<td>0.38</td>
</tr>
<tr>
<td>$= 3$</td>
<td>0.767</td>
<td>0.587</td>
<td>0.35</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$m$</th>
<th>MOM</th>
<th>FB/AML</th>
<th>EL</th>
</tr>
</thead>
<tbody>
<tr>
<td>$= 1$</td>
<td>0.011</td>
<td>0.006</td>
<td>0.01</td>
</tr>
<tr>
<td>$= 3$</td>
<td>0.005</td>
<td>0.003</td>
<td>0.003</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$m$</th>
<th>MOM</th>
<th>Block constant shadowing ML</th>
</tr>
</thead>
<tbody>
<tr>
<td>$= 1$</td>
<td>0.054</td>
<td>0.032</td>
</tr>
<tr>
<td>$= 3$</td>
<td>0.877</td>
<td>0.847</td>
</tr>
</tbody>
</table>

Fig. 4. Ratio of the idealized CR bound to MSE for estimates of $\mu$, $\sigma^2$, and $1/m$ for GL observations with correlated shadowing ($\alpha=0.1$). Only non-trivial relative performance curves are presented in this figure.

The performance of FB/AML, and EL is far superior to the block-constant shadowing ML estimator for $m$. In these calculations true parameter values were set to $\alpha = 0.9704$, $\sigma^2 = 0.9318$ dB, and $m = [1, 3]$. Translation to our GL model parameters yields $\sigma^2 = 0.845$, and $\alpha = 0.03$ (values for $m$ remain unchanged). Ten thousand realizations of the MOM estimator were obtained using this set of parameters, and, by simple transformations, MSE performance curves were obtained in terms of parameters $\alpha$ and $\sigma^2$. These MSE performance results are compared to the ones obtained in [5] for $N = 200$ and $N = 250$ in Table III. Recall that the FB/AML, and EL algorithms assume that the shadowing mean $\mu$ is known whereas this parameter is unknown and needs to be estimated in the MOM approach.

The performance of FB/AML, and EL is far superior to the MOM when estimating $\sigma^2$ in the $m = 1$ scenario. However, this performance becomes comparable in the $m = 3$ case. On the other hand, the MOM performance is superior (in some cases by an order of magnitude) to the performance of estimates of the shadowing correlation parameter $\alpha$ obtained with FB/AML, and EL for the $m = 1$ scenario, and this performance becomes comparable in the $m = 3$ case. Finally, the block-constant shadowing ML algorithm performance is an order of magnitude superior to that of the MOM for the case when $m = 1$, and this performance becomes comparable...
in the \( m = 3 \) case. Notice however that prior knowledge of the shadowing coherence time was used in [5] to choose an appropriate constant-shadowing block length for the ML estimator.

V. CONCLUSIONS

We have extended the GL family of distributions to include a correlated shadowing process with an arbitrary mean and an arbitrary finite-parameter autocovariance function. This extension enables accurate modeling of non i.i.d. observations of received path powers with WSS shadowing.

A MOM parameter estimator for the proposed GL distribution has been derived and implemented using a NLLS algorithm. Performance of the algorithm was compared to exact CR bound curves in the case of uncorrelated shadowing and to simplified idealized CR bounds in the case when correlation was not negligible. The estimation algorithm was shown to have good performance under different scenarios such as Rayleigh, Lognormal and Suzuki channels, and under the presence of AWGN. When the large-scale fading process follows a first order Gauss-Markov behavior, we were able to show that the performance of the MOM estimator is, in certain cases, comparable to other GL parameter estimators existing in literature for this same fading scenario such as the FB/AML, EL, and block-constant shadowing ML schemes.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the General Electric Global Research Center, Niskayuna, NY, for their support of this work. In particular, we would like to thank Mr. Daniel Sexton for defining the relevance of this work.

REFERENCES

Adaptive Ordering for
Imperfect Successive Decision Feedback Multiuser Detection

Vladimir D. Trajković, Member, IEEE, Predrag B. Rapajic, Senior Member, IEEE, and Rodney A. Kennedy, Fellow, IEEE

Abstract—A method for selecting the order in which the users are detected in communication systems employing adaptive successive decision feedback multiuser detection is proposed. Systems employing channel coding without the assumption of perfect decision feedback are analyzed. The method is based on the Mean Squared Error (MSE) measurements during a training period for each user. The analysis shows that the method delivers BER performance improvement relative to other previously proposed ordering methods.

Index Terms—Adaptive ordering, multiuser detection, decision feedback.

I. INTRODUCTION

In SUCCESSIVE Decision Feedback multi-user Detection (S-DFD) [1], [2], users are detected one by one in a successive manner. Consequently, in a multiuser system serving K users, there are K! different ordering possibilities implying K! different (although structurally equivalent) detectors [3]. In previous work on conventional S-DFD [4] users are ordered according to non-increasing powers after matched filtering. An extensive analysis of users’ detection order as a design parameter of the DFD with no assumption about perfect feedback has been provided in [3]. A rule for ordering the users according to non-increasing Asymptotic Effective Energies (AEE) has shown that any DFD always outperforms its linear counterpart in AEE for every user, e.g., D-DFD always outperforms the linear decorrelator. Both ordering methods require perfect knowledge of the system parameters such as amplitudes, spreading sequences, and channel.

In previous work on MMSE S-DFD, [5], the users are ordered according to received powers, i.e., the strongest users are detected first. In [6] the problem of the ordering in V-BLAST in a laboratory environment has been analyzed. An ordering method based on the post-detection SNRs has been proposed showing that at each stage during the cancellation process the signal with the largest post-detection SNR among the remaining (non cancelled) signals has to be detected first.

In this paper we present a new adaptive ordering method which employs an adaptive algorithm and linear multiuser detection (MUD). The method is shown to outperform all previously proposed ordering methods for the analyzed scenarios, delivering an SNR gain for moderate loads. For the overloaded case, a system employing the proposed ordering improves the BER performance, providing lower BER floor values, in comparison with previously proposed ordering methods such as power ordering [5], arbitrary ordering, AEE based ordering [3] or SNR based ordering [6].

In addition, we also show that when all users are received with the same powers it is still possible to improve BER performance applying the adaptive ordering relative to the arbitrary ordering (power based ordering is not applicable for that case). This makes the adaptive ordering more robust in respect to the users’ power variations.

II. SYSTEM MODEL

We consider an up-link of symbol-asynchronous CDMA multiuser detection scheme. At the transmitting end all signals are encoded first using Recursive Systematic Convolutional (RSC) Codes [7]. After interleaving and spreading, the users transmit their signals over an unknown multi-path communication channel.

A multi-cell mobile communications system is considered. Two groups of users are recognized in the system: the users within the cell of interest (denoted 1, 2, ..., K) and the users from neighboring cells whose signals are strong enough to cause interference within the cell of interest (inter-cell interferers denoted K+1, ..., K+K_f). The total number of users is K_T = K + K_f. The received signal is the superposition of all signals plus Additive White Gaussian Noise (AWGN) and for the asynchronous case it can be expressed as [1], [4]

\[ r(t) = \sum_{k=1}^{K+K_f} \sum_{L_k}^{L_k-1} s_k(t - iT - \tau_k) x_k(i) + n(t) \]  

(1)

where i indexes time instants, L_k is the number of symbols transmitted by each user, s_k(t) is the received signature of user k, x_k(i) is the kth user’s transmitted symbol, n(t) is AWGN. The kth user received signature waveform is the convolution s_k(t) = p_k(t) * g_k(t), where p_k(t) is the transmitted signature waveform, and g_k(t) = \sum_{n=0}^{L_k-1} g_{kn} \delta(t - \tau_{kn}) is the channel impulse response, where \tau_{kn} are path delays, g_{kn} are path gains and \delta(\cdot) is the Dirac delta function, while L_k is the length of kth user’s channel impulse response. The multi-path taps g_{kn} are assumed to be time-invariant or very slowly time-varying in comparison to the speed of the adaptive algorithm.
The MUD front-end is designed as a bank of linear MMSE filters. Here, we use the Sliding Window MMSE Filtering approach as explained in [1], which says that in practice an MMSE filter need not cover the entire received signal vector, but the length needs to cover the useful part of the effective signature vector (including multi-path).

For S-DFD, the kth user’s output at time instant i can be expressed as

\[ z_k(i) = f_k^T(i)r(i) - b_k^T(i)\tilde{x}_k(i) \]  

(2)

where \( f_k \) and \( b_k \) are the feedforward and feedback coefficients of user k, respectively, and \( \tilde{x}_k \) are the estimated symbols of users 1, 2, …, K-1 obtained during the detection process.

In order to estimate the MMSE S-DFD coefficients, in this paper we use the LMS algorithm. The first stage during the training period is devoted to ordering and will be discussed in more detail in the following section. Once the users’ order is determined, the LMS algorithm is used again to estimate the receiver coefficients, as proposed in [1]. While receiving data (coded signals), the S-DFD still operates as an adaptive S-DFD employing the same adaptive LMS algorithm.

In this work we use Maximum A Posteriori Probability (MAP) algorithm [7]. It was observed in [8] that the interference plus noise at the output of the MMSE filters can be assumed to be Gaussian. Therefore, the signal at the time instant \( i \) corresponding to the kth user used to feed the channel decoder \( k \) can be expressed as \( z_k(i) = A_k x_k(i) + \xi_k(i) \), where \( A_k \) is a constant and \( \xi_k(i) \) is a random variable with variance \( \sigma^2_{\xi_k} \). Quantities \( A_k \) and \( \sigma^2_{\xi_k} \) can be expressed as \( A_k = E[x_k(i)z(i)] \) and \( \sigma^2_{\xi_k} = E[z_k(i) - A_k x_k(i)]^2 \) so that estimates \( A_k \) and \( \sigma^2_{\xi_k} \) can be obtained from the MMSE filter output. The MAP algorithm delivers soft outputs of the th bit of the kth user defined as the Log-Likelihood Ratio [7] (LLR), \( \Lambda(x_k(i)) = \log(p(x_k(i) = +1|z_k(i))/p(x_k(i) = -1|z_k(i))) \). Finally, the expectations (or soft decisions) \( \tilde{x}_k(i) \) used in the feedback of S-DFD are obtained as \( \tilde{x}_k(i) = \tanh(\Lambda(x_k(i))/2) \) [9].

### III. Optimal Ordering and Adaptive Ordering MMSE Analysis

For the S-DFD with the error propagation the estimation of the kth user symbol after MMSE filtering is \( z_k(i) = f_k^T(i)r(i) - b_k^T(i)\tilde{x}_k(i) \), where \( f_k = [f_{k,1} f_{k,2} \ldots f_{k,L}]^T \) and \( b_k = [b_{k,1} b_{k,2} \ldots b_{k,k-1}]^T \) are vectors of kth user feed-forward and feedback coefficients, \( L_f \) is the length of the filter \( f_k \), and \( \tilde{x}_k(i) = [\tilde{x}_1(i) \tilde{x}_2(i) \ldots \tilde{x}_{k-1}(i)]^T \) are soft channel decoders output of the users 1, 2, …, k-1. \( r(i) \) is a part of the entire received sequence \( r \) (eq. (2) in [1]), that covers the th symbol of the kth user, and it is obtained according to the Sliding Window approach in [1]. Furthermore, the soft decoder output contains errors and the output symbols can be expressed as \( \tilde{x}_k(i) = x_k(i) + \xi_k(i) \), where \( x_k(i) \) is the correct symbol value and \( \xi_k(i) \) is the error at the decoder output. For the k-th user, the error can be expressed as the difference \( e_k(i) = f_k^T(i)r(i) - b_k^T(i)\tilde{x}_k(i) - x_k(i) \). The MSE is obtained as

\[ e_k = E[|e_k(i)|^2] = f_k^T R f_k - 2 f_k^T S_{D,k} b_k - 2 f_k^T s_k + b_k^T R_{\xi,k} b_k + \sigma^2_k \]  

(3)

where \( R = E[r(i)r^T(i)] \), \( R_x = \diag(\sigma^2_1, \sigma^2_2, \ldots, \sigma^2_K) \), \( S_{D,k} = E[x_k(i)\tilde{x}_k^T(i)] \), \( \sigma^2_k = E[x_k(i)x_k^T(i)] \), and \( \sigma^2_k = E[|x_k(i)|^2] \) is kth user received symbol energy. \( R_{\xi,k} = E[\tilde{x}_k(i)\tilde{x}_k^T(i)] \) can be expressed as \( \diag(\sigma^2_1, \sigma^2_2, \ldots, \sigma^2_{k-1}) + \tilde{N}_k \) where \( \tilde{N}_k = E[\tilde{x}_k(i)\tilde{x}_k^T(i)] \) is covariance matrix of decision feedback error, \( x_k(i) = [x_1(i) x_2(i) \ldots x_{k-1}(i)]^T \) and \( \tilde{x}_k(i) = [\tilde{e}_1(i) \tilde{e}_2(i) \ldots \tilde{e}_{k-1}(i)]^T \). The MMSE filters \( f_k \) and \( b_k \) can be expressed by

\[ f_k = \sigma^2_k (R - S_{D,k} R_{\xi,k}^{-1} S_{D,k}^T)^{-1} s_k \]  

(4)

\[ b_k = R_{\xi,k}^{-1} S_{D,k}^T f_k \]  

(5)

Combining (4), (5) and (3), we find the following expression for MMSE \( e_k \)

\[ e_k = \sigma^2_k [1 - \sigma^2_k S_{D,k} R_{\xi,k}^{-1} S_{D,k}^T]^{-1} s_k \]  

(6)

Equation (6) takes into account error propagation during the previous detection. At each stage during the detection process, the optimal order is determined according to increasing \( e_k \). This means that, among the undetected users, the user with the smallest \( e_k \) has to be chosen and detected next.

To achieve an optimal ordering we would have to use coded training sequences and to estimate \( e_k \) as described in (6) taking into account error propagation. However, this method is too complicated and it increases the complexity of the ordering process significantly due to use of channel coding. Here, we propose a simple ordering method that determines the users’ order according to increasing users’ error variances during the training period, employing linear MUD and the adaptive LMS algorithm, which estimates the kth user’s linear coefficients according to

\[ f_k(i+1) = f_k(i) + \mu e_k(i) r(i) \]  

(7)

The LMS algorithm calculates the error with each training bit as a difference between correct and the estimated value, i.e. \( e_k(i) = x_k(i) - f_k^T(i)r(i) \). The kth user’s MMSE or error variance is obtained by averaging \( |e_k(i)|^2 \) during the training process, i.e.,

\[ \text{Error Variance } k = \frac{1}{L - l'} \sum_{i=l'+1}^{L} |x_k(i) - f_k^T(i)r(i)|^2 \]  

(8)

where \( l' + 1 \) is the training bit from which we start measuring the error variance. Our experience shows that the initial training period should be avoided when determining the error variance since they tend to be unreliable. Once the convergence is achieved, we start measuring error variances. After measuring each user’s error variance, the users are then reordered according to non-decreasing error variances.

### IV. Simulation Results

In this section the simulation results are provided for different scenarios. In the transmitter, users are encoded using four-state (5,7) in octals) RSC codes. The encoded sequences are interleaved by 2048 bits long random interleavers. The linear filters \( f_k \) are all of the same length \( L_f = 21 \) chips. The channels \( g_k \) are modelled as five-tap chip-spaced multipath channels, i.e., \( L_k = 5 \) with randomly chosen coefficients in each realization. The adaptation constant is chosen to be...
The improvement in BER floor is around 3 times relative to SNR and AEE-based orderings, with respect to the weakest user. The results of Fig. 1 show that the proposed ordering outperforms SNR and AEE-based orderings by shifting the BER floor towards lower values. The simulation set-up, however, differs from the ones used to obtain the results in Fig. 2. Here, for the cases of post-SNRs and AEEs calculations. However, for moderate loads $\rho = 0.75$ the adaptive ordering delivers improvement of about 5 dB at BER of $2 \times 10^{-4}$ relative to the arbitrary ordering. For the overloaded case $\rho = 1.25$ the adaptive ordering provides significant improvement shifting the BER floor for almost two orders of magnitude relative to arbitrary ordering. These observations about BER floor improvement are made for high SNRs ($SNR > 20 \text{ dB}$).

Fig. 3 shows the BER performance of a multi-cell scenario where certain number of interfering users are in the neighboring cells. Number of users within the cell of interest is $K = 16$, while number of inter-cell interferers is $K_I = 4$. The simulation set-up, however, differs from the ones used to obtain the results in Fig. 2. Here, for the cases of post-SNR and AEE-based orderings, it is assumed that the inter-cell interferers are perfectly known. This knowledge is taken into account during post-SNRs and AEEs calculations. However, during the detection of information data, those inter-cell users are assumed to be still unknown, i.e., their signals are not used in the decision feedback part of the receiver. The results confirm that the proposed adaptive ordering still outperforms all other orderings and that the main cause of the better performance of the proposed algorithm is accounting for the
effects of feedback error propagation, caused not only by inter-cell interference, but also by intra-cell interferers.

The results in Fig. 4 are obtained for the single-cell scenario where hard decisions are used in the decision feedback part. Two sets of BER curves are obtained. The first set of results is obtained for $\rho = K/N = 0.75$, i.e., $K = 12$, and the second set of results obtained for $\rho = K/N = 1$, i.e., $K = 16$. Both sets of curves again confirm better BER performance obtained when the users are ordered according to the proposed algorithm, comparing to all other orderings, i.e., post-SNR based, AEE-based, and power-based. However, it can be noted that the BER performance are significantly degraded with all ordering methods comparing to the soft-decision results. For $\rho = 0.75$, however, the performances of all ordering methods are very close to each other, although the adaptive ordering shows still slightly better BER results than other orderings. However, for increased number of users ($\rho = 1$), the gap between the proposed adaptive ordering and other orderings becomes more significant. This is not surprising because the feedback error propagation increases with the increasing number of users within the system.

V. Conclusion

A new method for determining the ordering in an adaptive S-DFD scheme has been presented. It was shown that significant BER improvement can be obtained by the proposed adaptive MSE based ordering relative to the post-detection SNR based [6], AEE based [3], power based [5] or arbitrary ordering, especially for over-saturated CDMA systems where the number of users exceeds the spreading gain.

References

Variable Rate and Variable Power MQAM System Based on Bayesian Bit Error Rate and Channel Estimation Techniques

Lay Teen Ong, Mohammad Shikh-Bahaei, and Jonathon A. Chambers

Abstract—The impact of inaccurate channel state information at the transmitter for a variable rate variable power multilevel quadrature amplitude modulation (VRVP-MQAM) system over a Rayleigh flat-fading channel is investigated. A system model is proposed with rate and power adaptation based on the estimates of instantaneous signal-to-noise ratio (SNR) and bit error rate (BER). A pilot symbol assisted modulation scheme is used for SNR estimation. The BER estimator is derived using a maximum a posteriori approach and a simplified closed-form solution is obtained as a function of only the second order statistical characterization of the channel state imperfection. Based on the proposed system model, rate and power adaptation is derived for the optimization of spectral efficiency subject to an average power constraint and an instantaneous BER requirement. The performance of the VRVP-MQAM system under imperfect channel state information (CSI) is evaluated. We show that the proposed VRVP-MQAM system that employs optimal solutions based on the statistical characterization of CSI imperfection achieves a higher spectral efficiency as compared to an ideal CSI assumption based method.

Index Terms—Adaptive modulation, adaptive power, BER, channel estimation, flat-fading channel, MQAM.

I. INTRODUCTION

INK adaptation is one of the promising approaches to increase spectral efficiency of wireless channels [1]–[9]. In [2] ergodic capacity of a single-user flat-fading channel is obtained when optimal and suboptimal rate and power adaptation are used. The work in [4] and [5] examined a variable rate and variable power MQAM (VRVP-MQAM) system with optimal solutions derived assuming ideal channel state information (CSI) is available at the transmitter. Adaptive transmission under imperfect CSI has been studied in the literature, for instance in [4], [7], [8], [10]–[14]. In [4], the effect of channel estimation error and delay on the bit error rate (BER) performance of the VRVP-MQAM system is also analyzed. The effect of imperfect channel estimation on a VRVP-MQAM system is studied in [7] wherein an instantaneous signal-to-noise ratio (SNR) estimate based on the minimum mean-square error criterion was considered.

Transmitter adaptation using discrete power and/or rate levels has also been considered in a number of works, and more recently, such schemes have been analysed for systems with coding [15]–[17]. Most of these works have used SNR as a system performance indicator. However SNR measurements may not be a direct measure of quality for a mobile system with time-varying channel characteristics due to fading. Therefore power adaptation based on a BER parameter, which is considered to be a more direct representation of the measure of quality has been proposed in a number of published works, e.g. in [18] and [19]. Outer-loop power control based upon BER or FER (frame erasure rate) in W-CDMA systems is a practical example of this kind of power adaptation [20]. Joint optimization of BER-based adaptive modulation and outer loop SNR target was considered in [21].

In this work, we provide a VRVP-MQAM scheme using a BER estimate based on the maximum a posteriori (MAP) criterion, which does not assume exact knowledge of SNR. Instead, we assume the correlation between the true SNR and its estimate is known. Based on the proposed MAP-based BER estimate, we derive analytical expressions for optimal rate and power adaptation to maximize spectral efficiency while satisfying an average power constraint and an instantaneous BER target. We consider a pilot symbol assisted modulation (PSAM) method for channel estimation [22] at the receiver and consider the second order statistical characteristics of the channel imperfection (i.e. correlation coefficient) for transmitter adaptation.

The remainder of this paper is organized as follows. In Section II we describe the proposed system model. We derive the BER estimator in Section III. Then we obtain the analytical expressions for optimum rate and power adaptation in Section IV. In Section V, we evaluate the performance of the proposed VRVP-MQAM system under two scenarios: an ideal CSI scenario and a partial CSI scenario. Finally, conclusions are drawn in Section VI.

II. SYSTEM MODEL

We adopt the system model shown in Fig. 1, and consider a single user flat-fading channel. The channel is modelled in discrete-time, denoted by index $i$, with statistically stationary and ergodic time-varying gain $\alpha(i)$, and zero mean additive white Gaussian noise $n(i)$. We consider a PSAM [22] method for the estimation of the channel gain $\alpha(i)$. At time $i$, based on the channel gain and its respective SNR, the channel estimator has knowledge of the instantaneous received SNR estimate, the average SNR estimate and the correlation coefficient between the true channel SNR and its estimate. We also assume that this information is fed back to the transmitter error-free and without delay. By incorporating the autocorrelation
The transmitter adapts its rate by adjusting the constellation size $M$ of the MQAM scheme based on $\hat{\gamma}$ and a required instantaneous BER target. With data sent at $\gamma(i) = \log_2(M(\hat{\gamma}(i)))$ bits/symbol, the instantaneous data rate is $k(\hat{\gamma})/T_s$ bits/sec (bps), where $T_s$ denotes the symbol period duration. The spectral efficiency of an MQAM scheme can be expressed as its average data rate per unit bandwidth $R/B$. Assuming Nyquist data pulses of duration $T_s = 1/B$, the spectral efficiency (in bps/Hz) can be written as

$$R = \frac{B}{T_s} \int_{\hat{\gamma}} k(\hat{\gamma}) f_{\hat{\gamma}}(\hat{\gamma}) d\hat{\gamma},$$

where $f_{\hat{\gamma}}(\hat{\gamma})$ denotes the probability density function (PDF) of $\hat{\gamma}$. We aim to obtain the optimal rate and power adaptation in order to maximize spectral efficiency (1) subject to an average power constraint

$$E[S(\hat{\gamma})] = \int_{\hat{\gamma}} S(\hat{\gamma}) f_{\hat{\gamma}}(\hat{\gamma}) d\hat{\gamma} \leq \bar{S}$$

and an instantaneous BER requirement. However our proposed algorithm will be based on the estimates of SNR and BER values.

III. BER ESTIMATE

Our objective is to obtain an expression for the BER estimate $\hat{p}_B$ in terms of the observation of $\hat{\gamma}$. The instantaneous BER ($p_B$) will be estimated by adopting the MAP-optimal approach. Hence, the $p_B$ that maximizes the conditional PDF of $p_B$ given $\hat{\gamma}$, $f_{p_B|\hat{\gamma}}(p_B|\hat{\gamma})$, will be derived using Bayes' theorem [24]. Therefore we maximize the function

$$l_{MAP}(p_B) = \int f_{p_B|\hat{\gamma}}(p_B|\hat{\gamma}),$$

where $f_{p_B|\hat{\gamma}}(p_B|\hat{\gamma})$ is the joint PDF of $p_B$ and $\hat{\gamma}$. The estimate of $p_B$ can thus be obtained from the solution to

$$\frac{\partial l_{MAP}}{\partial p_B} = 0.$$ 

We will next derive the closed form expressions for $f_{p_B|\hat{\gamma}}(p_B|\hat{\gamma})$ and $\hat{p}_B$.

Assume the value of the SNR estimate $\hat{\gamma}$ is known through a particular channel gain estimator (where $\hat{\gamma} \neq \gamma$). The transmitter will adapt its power $S(\hat{\gamma})$ and rate $k(\hat{\gamma})$ relative to a BER target ($BERT$) based on the SNR estimate $\hat{\gamma}$ instead of the actual SNR, $\gamma$. Using the generic approximation of the BER expression for an MQAM scheme [4, eqn (42)], the instantaneous BER at the receiver can be expressed as

$$p_B(\gamma, \hat{\gamma}) = c_1 \exp\left(-\frac{c_2 \gamma}{M(\hat{\gamma}) - 1} \frac{S(\hat{\gamma})}{S}\right),$$

where $c_1$ and $c_2$ are positive real numbers [4], [5]. Let $\sigma$ be a particular value of $\sigma(\hat{\gamma})$ at the receiver, and $M$ be the corresponding value of $M(\hat{\gamma})$ determined at the transmitter. The corresponding instantaneous BER at the receiver will be

$$p_B = c_1 \exp\left(-\frac{c_2 \sigma}{M - 1} \frac{\gamma}{\hat{\gamma}}\right).$$

Here we consider a Rayleigh flat-fading channel and incorporate the PSAM method for channel estimation where $\hat{\alpha}$ is derived as the magnitude of a weighted sum of zero
mean complex random variables. Hence $\alpha$ and $\hat{\alpha}$ have a bivariate Rayleigh distribution [25]. Since the SNR ($\gamma$ and $\hat{\gamma}$) is expressed as a function of the channel gain ($\alpha$ and $\hat{\alpha}$ respectively), the joint PDF of $\gamma$ and $\hat{\gamma}$ can be derived (using the transformation of random variables [24, chpt. 6]) as

$$f_{\gamma,\hat{\gamma}}(\gamma, \hat{\gamma}) = \frac{1}{(1 - \rho)\Gamma\Gamma} I_0 \left( \frac{2\sqrt{\rho}}{(1 - \rho)\Gamma} \right) \cdot \exp \left( - \frac{1}{(1 - \rho)\Gamma} \left( \frac{\gamma}{\Gamma} + \frac{\hat{\gamma}}{\Gamma} \right) \right) u(\gamma) u(\hat{\gamma}),$$

which is a bivariate gamma distribution [11]. $I_0(\cdot)$ is the zero order modified Bessel function and $u(\cdot)$ is the unit step function. The parameter $\rho$ denotes the correlation coefficient between $\gamma$ and $\hat{\gamma}$ which we assume to be available at the transmitter. We note that under the above conditions, $\gamma$ and $\hat{\gamma}$ are exponentially-distributed with average values $\Gamma$ and $\Gamma$ respectively. That is, we have $f_{\gamma}(\gamma) = \frac{1}{\Gamma} \exp \left( - \frac{\gamma}{\Gamma} \right)$ and $f_{\hat{\gamma}}(\hat{\gamma}) = \frac{1}{\Gamma} \exp \left( - \frac{\hat{\gamma}}{\Gamma} \right)$. In practical situations, the information about $\rho$ at the transmitter is available through its estimate in the final form of the MAP function as

$$p_B(\gamma) = 1 - \frac{1}{2\pi} \exp \left( - \frac{\gamma}{\Gamma} \right) \cdot \exp \left( - \frac{1}{2\pi} \frac{\gamma^2}{\Gamma} \right) u(\gamma) u(\hat{\gamma}),$$

where $a(\hat{\gamma}) = \frac{1}{1 - \rho} \left( \frac{\gamma}{\Gamma} + \frac{\hat{\gamma}}{\Gamma} \right)$, $b(\hat{\gamma}) = \frac{2\sqrt{\rho}}{(1 - \rho)\Gamma} \frac{1}{\Gamma}$, and $c(\hat{\gamma}) = \frac{(M - 1)\hat{\gamma}}{p_B c_2 \sigma_2 (1 - \rho)\Gamma}$.

Based on (6), and omitting terms that are not functions of $p_B$, the final form of the MAP function can be expressed as

$$IMAP = \frac{1}{p_B} I_0 \left( b(p_B) \hat{\gamma} \right) \exp \left( -a(p_B) \hat{\gamma} \right).$$

To perform the operation $\frac{\partial IMAP}{\partial p_B}$, (10) will be approximated [26] as

$$IMAP_{approx} = \frac{1}{p_B} \frac{\exp \left( b(p_B) \hat{\gamma} \right) \exp \left( -a(p_B) \hat{\gamma} \right)}{\sqrt{2\pi b(p_B) \hat{\gamma}}},$$

where $IMAP_{approx}$ denotes the approximate form of the MAP function for large values of SNR estimate $\hat{\gamma}$. Finally, an approximate closed form expression for $p_B$ is obtained as

$$\hat{p}_B \approx c_1 \exp \left( \frac{\rho \sigma_2 \Gamma}{(1 - \rho) (\sigma_2^2 + \Gamma - 1)^2} \right), \quad \hat{\gamma} > \gamma_{th},$$

where $\gamma_{th}$ can be obtained from $\gamma_{th} = (1 - \rho) \left( -\ln \left( \frac{BERT}{c_1} \right) \right) \Gamma$. Exploiting numerical analysis to obtain the maximum value of $I_{MAP}$ and the corresponding BER estimate using the exact expression (10) leads to the result that realistic $\hat{p}_B$ values greater than zero exist only for the range of $\hat{\gamma} > \gamma_{th}$. Otherwise the optimum value for $\hat{p}_B$ will be zero. Therefore transmission will be stopped for $\hat{\gamma} \leq \gamma_{th}$, as there is no practical and reliable BER estimate (and hence no reliable channel) available in that case.

IV. OPTIMAL RATE AND POWER ADAPTATION

For a required $BERT$, and with only the knowledge of channel estimate available, $M(\hat{\gamma})$ is derived according to [4]

$$M(\hat{\gamma}) = 1 + \frac{c_{2} \hat{\gamma}}{\ln(BERT/c_1)} S(\hat{\gamma}).$$

We note from (13) that for a given $BERT$ and $\hat{\gamma}$, $\frac{c_{2} \hat{\gamma}}{\ln(BERT/c_1)} = \ln(BERT/c_1)$. Using (12) and (13), we can obtain an estimate of BER:

$$\hat{p}_B(\gamma) = c_1 \exp \left( \frac{\ln(BERT/c_1) \frac{\rho}{\Gamma} \hat{\gamma} U}{\ln(BERT/c_1) \frac{\rho}{\Gamma} (1 - \rho) - 1} \right)^2.$$  

Considering that the estimates $\hat{\gamma}$ and $\hat{p}_B(\gamma)$ are available at the receiver and transmitter, the corresponding MQAM constellation size $M_{pB}(\hat{\gamma})$ will be

$$M_{pB}(\hat{\gamma}) = 1 + \frac{c_{2} \hat{\gamma}}{\ln(p_B(\hat{\gamma})/c_1)} S_{pB}(\hat{\gamma}),$$

where $S_{pB}(\hat{\gamma})$ is the power variation at the transmitter. For $\hat{\gamma} > \gamma_{th}$, we substitute (14) into (15) and obtain the adjusted $M$ as

$$M_{pB}(\hat{\gamma}) = 1 + \frac{K T}{\rho U} \left\{ \frac{c_{2} \Gamma}{K^{\gamma}} (1 - \rho) - 1 \right\}^2 S_{pB}(\hat{\gamma}) \hat{\gamma},$$

where $K = \frac{c_{2} \hat{\gamma}}{\ln(BERT/c_1)}$. The corresponding transmission rate is $k_{pB}(\hat{\gamma}) = \log_2 \left( M_{pB}(\hat{\gamma}) \right)$.

For maximizing spectral efficiency subject to the average power constraint $S$, the corresponding power control $S_{pB}(\hat{\gamma})$ obtained using a Lagrangian method is

$$S_{pB}(\hat{\gamma}) = \left\{ U - K T \left[ \frac{p}{\rho} \frac{1}{\ln(BERT/c_1)} \right] \right\}, \quad S_{pB}(\hat{\gamma}) \geq 0, k_{pB}(\hat{\gamma}) \geq 1,$$

where $U$ is a constant value found through numerical search such that the average power constraint (2) is satisfied. Subsequently, the optimal rate adaptation is

$$k_{pB}(\hat{\gamma}) = \left\{ \log_2 \left[ \frac{K T}{\rho U} \left( \frac{c_{2} \Gamma}{K^{\gamma}} (1 - \rho) - 1 \right) \right]^2 \right\}, \quad k_{pB}(\hat{\gamma}) \geq 1,$$

otherwise.

Note that $k_{pB}(\hat{\gamma}) \geq 1$ corresponds to a realistic MQAM constellation size $M \geq 2$. For $S_{pB}(\hat{\gamma}) \geq 0$ and $k_{pB}(\hat{\gamma}) \geq 1$, the SNR cutoff threshold can be shown to be

$$\hat{\gamma}_0 = \gamma_{th} + \chi,$$
where $\chi \triangleq \Gamma / 2KU$ and $\hat{\gamma} = \frac{2 \rho + 2 \sqrt{\rho (2 U c_2 \hat{\Gamma} - 2 U c_2 \rho + \rho)}}{2 \rho}$, which is $\geq 0$. Hence, $S_{PB}(\hat{\gamma}) \geq 0$ and $k_{PB}(\hat{\gamma}) \geq 1$ imply $\hat{\gamma} \geq \gamma_0$ and transmission is allowed. These conditions also verify the appropriateness of using the derived BER estimator (12) since $\gamma_0$ is always $> \gamma_{th}$.

V. NUMERICAL RESULTS

We evaluate the performance of the system over channel variations modelled by Rayleigh fading and we assume that the PSAM technique [25] is used for deriving the channel estimate. For a system employing PSAM under fairly common channel conditions and SNR estimation criterion, $\rho$ has been shown [11] to be obtained from $\rho = \Gamma / \Gamma$. Therefore it is reasonable to assume that $\Gamma = \rho \Gamma$ for our numerical results. We assume $M \geq 2$ can be a non-integer value and consider a BER target of $10^{-3}$, and set $c_1 = 0.2$ and $c_2 = 1.5$.

Based on the analytical expressions derived in Section IV, we evaluate the performance of our proposed VRVP-MQAM system that employs adaptations based on CSI imperfection and a MAP-optimal BER estimate. We refer to this system as ‘VRVP-MQAM-CSI’. We further compare the performance of this system with that of two other MQAM systems: 1) a VRVP-MQAM system that employs adaptations based on an ideal CSI assumption which we refer to as ‘VRVP-MQAM’, and 2) a nonadaptive transmission system that employs a constant-rate constant-power MQAM (CRCP-MQAM) scheme [4].

A. Instantaneous Rate and Power in VRVP-MQAM-CSI

We represent the perfect-CSI scenario by $\rho = 1$ whereas the scenario with CSI imperfection will be represented by $\rho < 1$. For $\Gamma = 25$ dB, the sets of solutions $S_{PB}(\hat{\gamma})$ and $k_{PB}(\hat{\gamma})$ for $\rho = 0.8, 0.9, 1.0$ are obtained using (17) and (18) respectively. The corresponding values for $U$ are obtained numerically. The computed results are illustrated in Fig. 2(a) and Fig. 2(b) respectively. At $\rho = 1$, the transmit power variation $\frac{S_{PB}(\hat{\gamma})}{S}$ follows a smooth water-filling profile for $\hat{\gamma}$ beyond a cutoff value $\gamma_0$, and the rate $k_{PB}(\hat{\gamma})$ increases linearly as $\hat{\gamma}$ increases beyond $\gamma_0$. No data transmission is allowed for $\hat{\gamma}$ below $\gamma_0$. Indeed, when $\rho = 1$, the adaptations turn into the rate and power adaptations observed in [4] with adaptation based on assumption of ideal CSI. There is, however, no transmission in our plot when $k(.) < 1$ but the corresponding result in [4] uses the transmission rate limit of $k(.) = 0$. For $\rho < 1$, higher cutoff SNR values are derived in the rate and power adaptation plots.

As depicted in Fig. 2(a) and 2(b), the power and rate adaptation schemes adapt to the CSI-imperfection by transmitting at a higher power level and a higher transmission rates as $\rho < 1$ for larger SNR values. It is noted that the total average transmitted power is still maintained at $S$.

B. Instantaneous Rate and Power in VRVP-MQAM

In this section, we show the effect of channel imperfections on the performance of a VRVP-MQAM system based on an ideal CSI assumption. An ideal CSI is assumed by considering $\rho = 1$, $\gamma = \hat{\gamma}$, and $\hat{\Gamma} = \hat{\Gamma}$. We note that by substituting $\rho = 1$, $\gamma = \hat{\gamma}$, and $\hat{\Gamma} = \hat{\Gamma}$ into (17) and (18), the respective transmit power and rate adaptations are

\begin{align}
S(\hat{\gamma}) & = U - \frac{1}{K \gamma}, \\
k(\hat{\gamma}) & = \log_2 [K \gamma U],
\end{align}

where $S(\hat{\gamma}) \geq 0$ and $k(\hat{\gamma}) \geq 1$. In fact, optimal solutions (20) and (21) are those of [4], corresponding to a VRVP-MQAM scheme based on actual SNR ($\gamma$, $\Gamma$) knowledge.

To investigate the impact of CSI imperfection (i.e. $\rho < 1$, $\hat{\Gamma} = \rho \hat{\Gamma}$, $\hat{\gamma} \neq \gamma$) on the ‘ideal-assumed’ system, we first perform numerical search for $U$ at $\rho = 1$ by using (20) in the power constraint formula (2), and determine $U$ numerically. Next, for $\rho < 1$, we use (17) and (18) respectively for power and rate adaptations for $\rho < 1$, but always adopt the value of $U$ that was obtained at $\rho = 1$. We compute numerical results for a set of $\rho$ values, and at $\hat{\Gamma} = 25$ dB. The numerical results are illustrated in Fig. 3(a) and Fig. 3(b). For $\rho = 1$,...
rate and power have a water-filling nature. For $\rho < 1$, higher $\hat{\gamma}_0$ values are derived in the rate and power adaptation plots. Finally, we note that unlike in the VRVP-MQAM-CSI scheme, the power and rate adaptation curves converge to the same value at higher SNRs since rate and power adaptations are not adapting to CSI imperfection. It is therefore clear that the resulting average transmitted power will vary across channel imperfections, which will be verified in section V-D.

C. Instantaneous Rate and Power in CRCP-MQAM

In the CRCP-MQAM system, constant rate and power are transmitted at all times. For comparison with the other two systems, the transmit power is set to $\bar{S}$. The number of constellation points $M$, restricted to be $\geq 2$, is obtained for a given $\Gamma$ value such that the average BER is equal to $BERT$. Subsequently, the corresponding spectral efficiency is $\log_2(M)$.

D. Average Power and Spectral Efficiency

The normalized average power, which is expressed as $\frac{\int_0^\infty \left[ \hat{S}_{\hat{\gamma}}(\hat{\gamma}) / S \right] f_{\hat{\gamma}}(\hat{\gamma}) d\hat{\gamma}}$, is shown in Fig. 4(a) for the VRVP-MQAM-CSI and the VRVP-MQAM systems. In the perfect-CSI assumed VRVP-MQAM system, constant average power is maintained at $\rho = 1$, but variation in the average power occurs for $\rho < 1$. On the other hand, the VRVP-MQAM-CSI system maintains constant average power for all $\rho$ values since the system is adapting to the $\rho$ variations. Though not shown in the figure, the average power for the CRCP-MQAM scheme is clearly constant from $\Gamma \gtrsim 21$ dB, which corresponds to $M \geq 2$. These results confirm that the VRVP-MQAM-CSI system has appropriately exploited the power resource variation in an imperfect CSI scenario, and results in a higher spectral efficiency performance over a wide range of SNRs and for practical ranges of $\rho$, as we will show in Fig. 4(b).

Fig. 4(b) shows the spectral efficiency performance for VRVP-MQAM-CSI, VRVP-MQAM and CRCP-MQAM sys-
tems. The figure confirms that the spectral efficiency for the VRVP-MQAM-CSI and the VRVP-MQAM systems declines as $\rho$ decrease, and both schemes converge to the same performance in the perfect CSI scenario ($\rho = 1.0$). However, it is noted that the VRVP-MQAM-CSI system outperforms the VRVP-MQAM system for all channel imperfection scenarios ($\rho < 1$). In comparison to the CRCP-MQAM system, the results illustrate that the VRVP-MQAM-CSI system provides a better performance for larger values of $\rho$, and for small $\rho$ the CRCP-MQAM system would be a better alternative.

VI. CONCLUSION

We investigated the impact of imperfect channel estimation for a variable rate variable power MQAM system over a Rayleigh flat-fading channel. We have proposed a variable rate and power system model based on the estimates of the instantaneous SNR and BER using, respectively, a PSAM technique and a MAP-optimal BER estimator. Based on our proposed model, we derived optimal solutions for the rate and power algorithms, and the solutions were compared to an ideal CSI assumption based VRVP-MQAM system and a nonadaptive MQAM system. Our proposed system achieves a higher spectral efficiency as compared to the ideal CSI assumption based VRVP-MQAM system in all cases, and performs better than the nonadaptive MQAM system under reasonable channel imperfection scenarios.

ACKNOWLEDGMENT

The authors would like to thank Dr. Sangarapillai Lambetharan and Dr. Ali Olfat for their suggestions in this research work.

REFERENCES


Second-Order Statistics for Diversity-Combining of Non-Identical Correlated Hoyt Signals

Gustavo Fraidenraich, Student Member, IEEE, Michel Daoud Yacoub, José Ricardo Mendes, and J. C. S. Santos Filho

Abstract—In this paper, exact expressions for the level crossing rate (LCR) and average fade duration (AFD) for two-branch selection, equal-gain and maximal-ratio combining systems in a Hoyt fading environment are presented. The expressions apply to unbalanced, non-identically, correlated diversity channels and have been validated by specializing the general results to some particular cases whose solutions are known. In passing, the joint bidimensional envelope-phase Hoyt distribution with arbitrary fading parameters is obtained.

Index Terms—Average fade duration, equal-gain combining, Hoyt fading channels, level crossing rate, maximal-ratio combining, selection combining.

I. INTRODUCTION

LEVEL crossing rate (LCR) and average fade duration (AFD) are widely-used performance measures of wireless diversity systems. However, although the branch signals may be correlated and non-identically distributed in practical systems [1]–[4], the literature on LCR and AFD of diversity techniques over non-identical correlated fading is rather scarce. Pioneering work on this issue was carried out by Adachi et al. [1] for dual branch selection (SC), equal-gain (EGC), and maximal ratio combining (MRC) over balanced correlated Rayleigh channels. The unbalanced correlated Rayleigh case was addressed in [2] for two-, three-, and four-branch MRC. In [5], apart from the MRC, the SC and EGC were also studied for dual branch in a correlated Rayleigh channels. More recently, [3] presented a unified treatment for the LCR and the AFD of M-branch SC over unbalanced correlated Rayleigh, Ricean, and Nakagami-m channels. In [4], the LCR and AFD for the MRC were derived for a correlated, unbalanced Nakagami environment.

The Hoyt distribution has received some attention for its flexibility and simplicity in the analysis of error rate performance, outage analysis, and also for LCR problems in a single branch system [6]. This paper fully generalizes the approach used in [1], and provides expressions for the LCR and AFD of dual-branch SC, EGC, and MRC operating over non-identical, correlated Hoyt (Nakagami-q) channels.

This work is organized as follows: Section II derives the Joint bidimensional envelope-phase Hoyt distribution; Section III presents general expressions for LCR and AFD of the combining output; Section IV-A derives the matrices for the conditional joint distribution; Section IV-B computes the mean and variance for each diversity system; Section V shows some numerical plots, and finally Section VI draws some conclusions.

II. THE JOINT BIDIMENSIONAL ENVELOPE-PHASE HOYT DISTRIBUTION

In a Hoyt fading environment, the received signal at the i-th antenna (i = 1, 2), can be represented in a complex form as

\[ X_i(t) + jY_i(t) = R_i(t) \exp(j \Theta_i(t)) \]  \hspace{1cm} (1)

where, \( X_i(t) \) and \( Y_i(t) \) are zero mean independent Gaussian processes with variance \( \sigma_{X_i}^2 \) and \( \sigma_{Y_i}^2 \), respectively. The variates \( R_i(t) \) and \( \Theta_i(t) \) follow the envelope and phase of the Hoyt distribution [7], respectively.

We now proceed to determine the joint distribution of \( X_i \equiv X_i(t) \), \( Y_i \equiv Y_i(t) \), \( (i = 1, 2) \) Defining the vector \( \mathbf{Z} = [X_1 Y_1 Y_2 Y_2] = [R_1 \cos(\Theta_1) R_1 \sin(\Theta_1) R_2 \cos(\Theta_2) R_2 \sin(\Theta_2)] \), the joint Gaussian distribution \( p_{\mathbf{Z}}(\mathbf{z}) \) can be written as [8]

\[
p_{\mathbf{Z}}(\mathbf{z}) = \frac{1}{(2\pi)^2 |\mathbf{b}|^{1/2}} \exp \left( -\frac{1}{2} \mathbf{z}^T \mathbf{b}^{-1} \mathbf{z} \right) \]  \hspace{1cm} (2)

where:

(1) \( \mathbf{z}^T \) denotes the transpose matrix, \( \mathbf{b} \) is the covariance matrix given by

\[
\mathbf{b} =
\begin{bmatrix}
\sigma_{X_1}^2 & 0 & \sigma_{X_1} \sigma_{X_2} & \sigma_{X_1} \sigma_{Y_2} \\
0 & \sigma_{Y_1}^2 & \sigma_{Y_1} \sigma_{X_2} & \sigma_{Y_1} \sigma_{Y_2} \\
\sigma_{X_1} \sigma_{X_2} & \sigma_{Y_1} \sigma_{X_2} & \sigma_{X_2}^2 & \sigma_{X_2} \sigma_{Y_2} \\
\sigma_{X_1} \sigma_{Y_2} & \sigma_{Y_1} \sigma_{Y_2} & \sigma_{X_2} \sigma_{Y_2} & \sigma_{Y_2}^2
\end{bmatrix}
\] (3)

(2) \( \mu_{ij} = \mu_{ij}(0) \) and \( \eta_{ij} = \eta_{ij}(0) \) are the correlation coefficients, defined as

\[
\mu_{ij}(\tau) = \frac{\text{Cov}(X_i(t), X_j(t+\tau))}{\sqrt{\text{Var}(X_i(t)) \text{Var}(X_j(t+\tau))}} \hspace{1cm} (i \leq j)
\] (4)

and

\[
\eta_{ij}(\tau) = -\frac{\text{Cov}(X_i(t), Y_j(t+\tau))}{\sqrt{\text{Var}(X_i(t)) \text{Var}(Y_j(t+\tau))}} \hspace{1cm} (i \neq j)
\] (5)

and

\[
\eta_{ii}(\tau) = \frac{\text{Cov}(X_i(t), Y_i(t+\tau))}{\sqrt{\text{Var}(X_i(t)) \text{Var}(Y_i(t+\tau))}} \hspace{1cm} (i = 1, 2)
\] (6)


G. Fraidenraich is with the Stanford University Department of Electrical Engineering, David Packard #353, Stanford, CA 94305 USA (e-mail: gfraiden@stanford.edu).

M. D. Yacoub, J. R. Mendes, and J. C. S. Santos Filho are with the Department of Communications, State University of Campinas, 13083-970 Campinas, SP, Brazil (e-mail: [michel, candido]@decom.ime.unicamp.br).

Digital Object Identifier 10.1109/TCOMM.2008.050024.
\[ Var(\cdot) \text{ and } Cov(\cdot) \text{ are the variance and covariance operators, respectively. For Gaussian processes, the following relations are valid: } \mu_{22} = \mu_{11}, \mu_{21} = \mu_{12}, \eta_{21} = -\eta_{12}, \text{ and } \eta_{22} = \eta_{11} [9]. \] 

The joint density \( p_{R_1, R_2, \theta_1, \theta_2} (r_1, r_2, \theta_1, \theta_2) \) can be written as \( p_{R_1, R_2, \theta_1, \theta_2} (r_1, r_2, \theta_1, \theta_2) = |J| p_{\zeta} (\zeta) \), where \( |J| \) is the Jacobian of the transformation. Accordingly, the joint bidimensional envelope-phase Hoyt distribution, as derived here, is given by 

\[ p_{R_1, R_2, \theta_1, \theta_2} (r_1, r_2, \theta_1, \theta_2) = \frac{1}{2 \pi \sigma_x \sigma_y \sqrt{1 - \rho^2}} \exp \left( -\frac{1}{2 (1 - \rho^2)} \left( \frac{r_1^2}{\sigma_x^2} + \frac{r_2^2}{\sigma_y^2} - 2 \rho \frac{r_1 r_2}{\sigma_x \sigma_y} \right) \right) \]

where \( \sigma_x^2 = \mu_1^2 + \eta_1^2 \).

### III. LCR AND AFD

The LCR \( n_R(r) \) and AFD \( T_R(r) \) of a random signal are defined, respectively, as the average number of upward (or downward) crossings per second at a given level and as the mean time the signal remains below this level after crossing it in the downward direction. The LCR and AFD of the combiner output \( R = R(t) \) at level \( r \) are, respectively, given by [10]

\[ n_R(r) = \int_0^\infty \hat{m}_R(r, \hat{r}) d\hat{r} \quad (8) \]

\[ T_R(r) = \frac{P_R(r)}{n_R(r)} \quad (9) \]

where \( P_{R, \hat{R}}(\cdot, \cdot) \) is the joint probability density function (JPDF) of \( R \) and its time derivative \( \hat{R} \), and \( P_R(\cdot) \) is the cumulative distribution function (CDF) of \( R \). In the following, (8) and (9) shall be calculated for the dual-branch, correlated, non-identical Hoyt fading environment using the SC, EGC and MRC techniques.

#### A. Diversity Systems

The output envelope and its time derivative for the SC, EGC and MRC combing systems are given, respectively, by

\[ R = \begin{cases} \max \{ R_1, R_2 \} & \text{SC} \\ \sqrt{R_1^2 + R_2^2} & \text{EGC} \end{cases} \]

\[ \hat{R} = \begin{cases} R_1 & \text{SC} \\ R_2 & \text{SC} \end{cases} \]

For the uncorrelated case, given \( \Theta_i \), it is known that \( \hat{R}_i \) is a zero-mean Gaussian variate [11], i.e., \( E[\hat{R}_i | \Theta_i] = 0 \). Unfortunately, for the correlated case, this assumption is not valid any longer.

Therefore it is clear from (10) that the conditional density \( p_{R | R_1, R_2, \theta_1, \theta_2} (\hat{r} | r_1, r_2, \theta_1, \theta_2) \) is also Gaussian with non zero mean \( m_{R_i}(r_1, r_2, \theta_1, \theta_2) \) and variance \( \sigma_{R_i}^2(r_1, r_2, \theta_1, \theta_2) \). Of course, these quantities depend on the combining scheme and shall be determined later. Now

\[ p_{R_i | R_1, R_2, \theta_1, \theta_2} (\hat{r} | r_1, r_2, \theta_1, \theta_2) = m_{R_i}(r_1, r_2, \theta_1, \theta_2) \]

\[ p_{R_i | R_1, R_2, \theta_1, \theta_2} (\hat{r} | r_1, r_2, \theta_1, \theta_2) p_{R_i, R_1, R_2, \theta_1, \theta_2} (r_1, r_2, \theta_1, \theta_2) \quad (11) \]

Knowing \( p_{R_i | R_1, R_2, \theta_1, \theta_2} (\hat{r} | r_1, r_2, \theta_1, \theta_2) \), as in (11), and the relations of \( R_1, R_2, \eta_1, \eta_2 \), as in (10), the joint density \( p_{R, \hat{R}}(r, \hat{r}) \) can be obtained to be used in (8).

The kernel of the problem now turns out to be the estimation of

\[ p_{R_i | R, \hat{R}} (\hat{r} | r_1, r_2, \theta_1, \theta_2) = \frac{1}{\sqrt{2 \pi \sigma_R(r_1, r_2, \theta_1, \theta_2)}} \exp \left( -\frac{(\hat{r} - m_{R_i}(r_1, r_2, \theta_1, \theta_2))^2}{2 \sigma_{R_i}^2(r_1, r_2, \theta_1, \theta_2)} \right) \quad (12) \]

More specifically, the tricky part of the problem is the determination of \( m_{R}(r_1, r_2, \theta_1, \theta_2) \) and \( \sigma_{R}^2(r_1, r_2, \theta_1, \theta_2) \) for each combining scheme. For the moment, assume that these quantities are known. Then, by means of [1, Eq. 8] for SC and of [11, Eqs. (12) and (17)] for EGC and MRC, respectively, the LCR can be found as in (13), where

\[ \vartheta(r_1, r_2) = \int_0^\infty r p_{R_i | R, \hat{R}} (\hat{r} | r_1, r_2, \theta_1, \theta_2) d\hat{r} \]

Now, with (12) in (14)

\[ \vartheta(r_1, r_2) = \frac{\sigma_R(r_1, r_2, \theta_1, \theta_2)}{\sqrt{2 \pi}} \exp \left( -\frac{m_{R_i}(r_1, r_2, \theta_1, \theta_2)}{2 \sigma_{R_i}^2(r_1, r_2, \theta_1, \theta_2)} \right) \]

\[ + \frac{m_{R_i}(r_1, r_2, \theta_1, \theta_2)}{2} \left( 1 + \text{erf} \left( \frac{m_{R_i}(r_1, r_2, \theta_1, \theta_2)}{\sqrt{2 \sigma_{R_i}(r_1, r_2, \theta_1, \theta_2)}} \right) \right) \quad (15) \]

where \( \text{erf}(\cdot) \) is the error function. The CDF \( P_R(r) \) can be obtained as [12]

\[ P_R(r) = \int_0^\gamma \int_0^2 \int_0^{2 \pi} \int_0^{2 \pi} \frac{p_{R_i | R, \hat{R}} (\hat{r} | r_1, r_2, \theta_1, \theta_2) d\hat{r} d\theta_2 d\phi_2 dr_1}{2 \pi} \quad (16) \]

with

\[ \gamma_1 = \gamma_2 = r \text{ for SC} \]

\[ \gamma_1 = \sqrt{2 \tau} \quad \gamma_2 = \sqrt{2 \tau} - r_1 \text{ for EGC} \]

\[ \gamma_1 = r \quad \gamma_2 = \sqrt{\tau^2 - r_1^2} \text{ for MRC} \]

The AFD follows directly from (9), (13), and (16).

### IV. CONDITIONAL STATISTICS OF \( \hat{R} \)

The aim of this section is to find the mean \( m_{\hat{R}}(r_1, r_2, \theta_1, \theta_2) \) and the variance \( \sigma_{\hat{R}}^2(r_1, r_2, \theta_1, \theta_2) \) of the conditional Gaussian distribution \( p_{\hat{R} | R_1, R_2, \theta_1, \theta_2}(\hat{r} | r_1, r_2, \theta_1, \theta_2) \) for each combining technique.
\[
\begin{align*}
\mathbf{A} &= E \left[ \{ \mathbf{Z} \mathbf{Z} \}^T \{ \mathbf{Z} \mathbf{Z} \} \right] - E \left[ \{ \mathbf{Z} \mathbf{Z} \}^T \right] E \left[ \{ \mathbf{Z} \mathbf{Z} \} \right] \\
&= \begin{bmatrix} a & c \\ c^T & b \end{bmatrix}
\end{align*}
\]

From [13, Eq. 9.106], the following relations are valid
\[
\begin{align*}
E[P(t)P(t+\tau)] &= \frac{dE[P(t)P(t+\tau)]}{d\tau} \\
E[\dot{P}(t)P(t+\tau)] &= -\frac{dE[P(t)P(t+\tau)]}{d\tau} \\
E[\ddot{P}(t)P(t+\tau)] &= -\frac{d^2E[P(t)P(t+\tau)]}{d\tau^2}
\end{align*}
\]

Now we define
\[
\begin{align*}
\dot{\mu}_{ij} &= \left. \frac{d\mu_{ij}(\tau)}{d\tau} \right|_{\tau=0} \\
\ddot{\mu}_{ij} &= \left. \frac{d^2\mu_{ij}(\tau)}{d\tau^2} \right|_{\tau=0}
\end{align*}
\]

Finally, using (18), (19), and (20) the mean and the variance of \( \tilde{R} \) given \( \mathbf{Z} \) will be found.

In order to determine \( p_1(\mathbf{Z} \mid \mathbf{Z}) \), the joint multivariate Gaussian distribution given in (2) will be used with the covariance matrix \( \mathbf{A} \) given by

\[
\mathbf{A} = E \left[ \{ \mathbf{Z} \mathbf{Z} \}^T \{ \mathbf{Z} \mathbf{Z} \} \right] - E \left[ \{ \mathbf{Z} \mathbf{Z} \}^T \right] E \left[ \{ \mathbf{Z} \mathbf{Z} \} \right]
\]

where \( \mu_{ij}(\tau) \) and \( \eta_{ij}(\tau) \) are given by (4), (5), and (6). Then

\[
\begin{bmatrix} -\dot{\mu}_{11} \sigma_{x_1} & 0 & -\dot{\mu}_{12} \sigma_{x_1} \sigma_{x_2} & \dot{\eta}_{12} \sigma_{x_1} \sigma_{y_2} \\ -\dot{\mu}_{12} \sigma_{x_1} \sigma_{x_2} & \dot{\mu}_{12} \sigma_{x_1} \sigma_{x_2} & -\dot{\eta}_{12} \sigma_{x_1} \sigma_{y_2} & -\dot{\mu}_{12} \sigma_{y_1} \sigma_{y_2} \\ -\dot{\mu}_{12} \sigma_{x_1} \sigma_{x_2} & -\dot{\eta}_{12} \sigma_{x_1} \sigma_{y_2} & -\dot{\mu}_{12} \sigma_{y_1} \sigma_{y_2} & 0 \\ -\dot{\mu}_{12} \sigma_{x_1} \sigma_{x_2} & 0 & 0 & -\dot{\mu}_{12} \sigma_{y_1} \sigma_{y_2} \end{bmatrix}
\]

and the matrix \( \mathbf{b} \) is given in (3). Note that the diagonal elements in the matrix \( \mathbf{c} \) are all null, because for a stationary process the correlation coefficient between the process and its time derivative is always null at \( \tau = 0 \) (\( \mu_{11} = 0 \)) [13].

Using the results from [8, pp. 495-496], the conditional distribution of \( \tilde{\mathbf{Z}} \) given \( \mathbf{Z} \), \( p_1(\tilde{\mathbf{Z}} \mid \mathbf{Z}) \), is Gaussian distributed with the mean matrix \( \mathbf{M} \) given by

\[
\mathbf{M} = \begin{bmatrix} E[\tilde{\mathbf{Z}}_1\mid\mathbf{Z}] \\ E[\tilde{\mathbf{Y}}_1\mid\mathbf{Z}] \\ E[\tilde{\mathbf{Z}}_2\mid\mathbf{Z}] \\ E[\tilde{\mathbf{Y}}_2\mid\mathbf{Z}] \end{bmatrix} = (\mathbf{c}\mathbf{b}^{-1}) \mathbf{Z}
\]

and the covariance matrix \( \mathbf{A} \) given by (22). Using this and after a tedious procedure, the matrices obtained are given by

\[
\begin{bmatrix} m_{1x_1} + m_{2x_2} \sigma_{x_1} \sigma_{y_1} + m_{3} \sigma_{x_1} x_2 m_{4} \sigma_{x_2} y_2 \\ 0 & -m_{2} \sigma_{x_1} x_1 + m_{1} y_1 - m_{4} \sigma_{x_2} y_2 & -m_{3} \sigma_{x_1} x_2 + m_{4} \sigma_{x_2} y_1 - m_{1} x_2 \sigma_{x_2} y_2 \\ -m_{3} \sigma_{x_1} x_1 - m_{1} x_2 \sigma_{x_1} y_1 + m_{4} \sigma_{x_2} y_1 - m_{2} \sigma_{x_2} y_2 & 0 \\ -m_{4} \sigma_{x_1} x_2 - m_{3} \sigma_{x_1} y_1 - m_{2} \sigma_{x_2} y_2 + -m_{4} \sigma_{x_2} y_2 \end{bmatrix}
\]
\[
\Delta = \begin{bmatrix}
\text{Var}(X_1|Z) & \text{Cov}(X_1, Y_1|Z) & \text{Cov}(X_1, X_2|Z) & \text{Cov}(X_1, Y_2|Z) \\
\text{Cov}(X_1, Y_1|Z) & \text{Var}(Y_1|Z) & \text{Cov}(Y_1, X_2|Z) & \text{Cov}(Y_1, Y_2|Z) \\
\text{Cov}(X_1, X_2|Z) & \text{Cov}(Y_1, X_2|Z) & \text{Var}(X_2|Z) & \text{Cov}(X_2, Y_2|Z) \\
\text{Cov}(X_1, Y_2|Z) & \text{Cov}(Y_1, Y_2|Z) & \text{Cov}(X_2, Y_2|Z) & \text{Var}(Y_2|Z)
\end{bmatrix} = a - cb^{-1}c^T
\] (22)

\[
\Delta = -
\begin{bmatrix}
\sigma_1^2 \Delta_1 & 0 & \sigma_{X_1} \sigma_{X_2} \Delta_2 & -\sigma_{X_1} \sigma_{Y_2} \Delta_3 \\
0 & \sigma_1^2 \Delta_1 & \sigma_{X_1} \sigma_{Y_1} \Delta_3 & \sigma_{Y_1} \sigma_{Y_2} \Delta_2 \\
\sigma_{X_1} \sigma_{X_2} \Delta_2 & \sigma_{X_2} \sigma_{Y_1} \Delta_3 & \sigma_{X_2} \Delta_1 & \sigma_{Y_1} \sigma_{Y_2} \Delta_2 \\
-\sigma_{X_1} \sigma_{Y_2} \Delta_3 & \sigma_{Y_1} \sigma_{Y_2} \Delta_2 & 0 & \sigma_{Y_2}^2 \Delta_1
\end{bmatrix}
\]

where

\[
m_1 = \frac{\mu_{12} \dot{\eta}_{12} + \eta_{12} \dot{\eta}_{12}}{1 - \rho^2}
\]
\[
m_2 = \frac{\eta_{12} \dot{\mu}_{12} - \mu_{12} \dot{\eta}_{12} - \dot{\eta}_{11}}{1 - \rho^2}
\]
\[
m_3 = \frac{\dot{\eta}_{11} \eta_{12} - \mu_{12}}{1 - \rho^2}
\]
\[
m_4 = \frac{\dot{\eta}_{12} + \eta_{11} \mu_{12}}{1 - \rho^2}
\]

\[
\Delta_1 = \frac{\dot{\mu}_{12} + \nu_{12}^2 + \eta_{12}^2 + \dot{\eta}_{12}^2 + 2 \dot{\eta}_{11} (\mu_{12} \dot{\eta}_{12} - \mu_{12} \eta_{12})}{1 - \rho^2}
\]
\[
\Delta_2 = \frac{\eta_{12} (\dot{\mu}_{12} - \dot{\eta}_{11}) - \dot{\mu}_{12} (\nu_{12}^2 - \dot{\eta}_{12}^2 - \dot{\eta}_{12}^2)}{1 - \rho^2}
\]
\[
\Delta_3 = \frac{\dot{\eta}_{12} + \eta_{11} \frac{\dot{\mu}_{12} + \nu_{12}^2 + \eta_{12}^2 + \dot{\eta}_{12}^2 + 2 \dot{\eta}_{11} (\mu_{12} \dot{\eta}_{12} - \mu_{12} \eta_{12})}{1 - \rho^2}}{1 - \rho^2}
\]

3) Maximal Ratio Combining:

\[
m_R (r_1, r_2, \theta_1, \theta_2) = \frac{r_1 m_{R_1} (r_1, r_2, \theta_1, \theta_2) + r_2 m_{R_2} (r_1, r_2, \theta_1, \theta_2)}{\sqrt{r_1^2 + r_2^2}}
\] (29)

\[
\sigma_R^2 (r_1, r_2, \theta_1, \theta_2) = \frac{1}{r_1^2 + r_2^2} \left\{ r_1^2 \sigma_{R_1}^2 (r_1, r_2, \theta_1, \theta_2) + r_2^2 \sigma_{R_2}^2 (r_1, r_2, \theta_1, \theta_2) + 2 r_1 r_2 \sigma_{R_1, R_2} (r_1, r_2, \theta_1, \theta_2) \right\}
\] (30)

C. Special Cases

For the Rayleigh case, \( \sigma_{X_i} = \sigma_{Y_i} = \sigma \) (i = 1, 2), then equations from (23) to (30) reduce in a exact manner to those of [1, Eqs. 26 and 27]. In the case of branch independence (e.g. large separation between the antennas) the mean and variance are not functions of \( R_1, R_2, \Theta_1, \) and \( \Theta_2 \) because \( \mu_{12} = \hat{\mu}_{12} = \mu_{12} = \eta_{12} = \hat{\eta}_{12} = \eta_{12} = 0 \). Then the results coincide with those of [11] for EGC and MRC with the number of branches \( M = 2 \).

V. NUMERICAL RESULTS

The expressions obtained for the LCR and AFD are general and can be applied to any type of diversity (space, frequency or time). In this section, we assume space diversity at the mobile station as [1]. For incoming multipath waves having equal amplitude and independent phases, the crosscorrelation functions are given by [9, [14]

\[
\mu_{11} (\tau) = \frac{J_0 (2\pi f_m \tau)}{1 + (\Delta \omega T)^2}
\]

\[
\mu_{12} (\tau) = \frac{J_0 \left( 2 \pi \sqrt{(f_m \tau)^2 + (d/\lambda)^2 - 2 (f_m \tau)(d/\lambda) \cos (\alpha)} \right)}{1 + (\Delta \omega T)^2}
\]

\[
\eta_{11} (\tau) = \Delta \omega T \mu_{11} (\tau)
\]

\[
\eta_{12} (\tau) = \Delta \omega T \mu_{12} (\tau)
\]

where \( J_0 (\cdot) \) is the zero-order Bessel function, \( \lambda \) is the carrier wavelength, \( f_m \) is the maximum Doppler shift in Hz, \( d \) is the antenna spacing, \( \Delta \omega \) is the angular frequency separation, \( T \) is the time delay spread, and \( \alpha \in [0, 2\pi] \) is the angle between the antenna axis and the direction of the vehicle motion in radians.

For a nil frequency separation, then \( \eta_{11} (\tau) = 0 \) and \( \eta_{12} (\tau) = 0 \). The corresponding correlation coefficients can
Hoyt fading parameter can be calculated as $d/\lambda$

Fig. 2. Normalized LCR and AFD for $d/\lambda$

Fig. 3. Normalized LCR and AFD for $d/\lambda$

Fig. 4. Normalized LCR and AFD for $d/\lambda$

be calculated as

\[
\begin{align*}
\mu_{11} &= 1 \\
\mu_{12} &= J_0 (2\pi d/\lambda) \\
\dot{\mu}_{11} &= 0 \\
\dot{\mu}_{12} &= 2\pi f_m \cos (\alpha) J_1 (2\pi d/\lambda) \\
\ddot{\mu}_{11} &= -2 (\pi f_m)^2 \\
\ddot{\mu}_{12} &= (2\pi f_m)^2 \\
\left\{ \frac{J_1 (2\pi d/\lambda)}{2\pi d/\lambda} \cos (2\alpha) - \cos^2 (\alpha) J_0 (2\pi d/\lambda) \right\}
\end{align*}
\]

where $J_1 (\cdot)$ is the first-order Bessel function. And this is the case explored here (as well as in [1]).

In the illustrations that follow we use the Hoyt parameter [7] $b_i \equiv \frac{\sigma_{X_i}^2}{\sigma_{Y_i}^2}$ and the individual power branches $\Omega_i = \sigma_{X_i}^2 + \sigma_{Y_i}^2$.

Figs. 1 and 2 show the normalized LCR (left vertical axis), $N_{LR}/f_m$, and AFD (right vertical axis), $T_{LR}/f_m$, for $\alpha = 0^\circ$ and $\alpha = 90^\circ$, respectively, as a function of the envelope level, for SC, EGC, and MRC. The following arbitrary parameters have been used: $d/\lambda = 0.06$, $b_i = 0.5$.

Figs. 3 and 4 show the normalized LCR and AFD for two different antenna angles $\alpha = 0^\circ$ and $\alpha = 90^\circ$, respectively, as a function of the parameter $d/\lambda$, for the SC, EGC, and MRC. An envelope level at $r/\sqrt{\Omega_1 + \Omega_2} = -20 \text{ dB}$, identical fading parameters $b_i = 0.5$, and balanced channels have been used. It can be seen that as the antenna spacing becomes larger, the LCR decreases, becoming oscillatory and convergent. Fig. 3 also shows that the MRC has the smaller LCR in both cases of antenna angles. It can be seen in Fig. 4 that the shape of the AFD curves for the SC, EGC, and MRC are loosely dependent on the antenna spacing when $\alpha = \pi/2$.

Figs. 5 and 6 show the effect of power imbalance and the use of non-identical parameters in the normalized LCR as a function of the envelope level and as a function of the parameter $d/\lambda$, respectively, for SC, EGC, and MRC. Fig. 5 shows the balanced ($\Omega_1 = 1$, $\Omega_2 = 1$) and unbalanced cases ($\Omega_1 = 1.8$, $\Omega_2 = 0.2$) for antennas arranged parallel to the direction of the vehicle motion ($\alpha = 0^\circ$), using Hoyt parameters ($b_1 = 0.5$, $b_2 = 0.2$), and $d/\lambda = 0.4$. It can be seen that, for this specific correlation value between the branches, the effect of the power imbalance deteriorates the performance of the LCR for SC, EGC, and MRC cases. Fig. 6 shows the balanced ($\Omega_1 = 1$, $\Omega_2 = 1$) and unbalanced cases ($\Omega_1 = 1.8$, $\Omega_2 = 0.2$) for ($\alpha = 0^\circ$), using Hoyt parameters ($b_1 = 0.5$, $b_2 = 0.2$), and $r/\sqrt{\Omega_1 + \Omega_2} = -20 \text{ dB}$. It also can
be seen that the effect of the power imbalance deteriorates the performance of the LCR for SC, EGC, and MRC cases.

Fig. 7 shows the effect of power imbalance and the use of non-identical parameters in the normalized AFD as a function of the parameter \( d/\lambda \) for SC, EGC, and MRC. This figure shows the unbalanced case \((\Omega_1 = 1.8, \Omega_2 = 0.2)\) for antennas arranged parallel to the direction of the vehicle motion \((\alpha = 0^\circ)\), using Hoyt parameters \((b_1 = 0.5, b_2 = 0.2)\), and \( r/\sqrt{\Omega_1 + \Omega_2} = -20 \text{dB} \). It can be seen in Fig. 7 that the shape of the AFD curves for the SC, EGC, and MRC are extremely dependent on the antenna spacing when \( \alpha = 0^\circ \).

VI. CONCLUSIONS

Exact formulas for level crossing rate and average fade duration of the dual branch SC, EGC and MRC techniques in a unbalanced, non-identical, and correlated Hoyt fading environment have been presented. The results show that the effect of correlation and unbalanced branches have a considerable impact on the LCR and AFD. In passing, this paper derives the joint Hoyt bidimensional envelope-phase distribution. These formulas have been validated by specializing the general results to some particular cases whose solutions are known. Sample numerical results were presented by specializing the general expressions to a space-diversity system with horizontally spaced omnidirectional antennas at the mobile station.

REFERENCES

Performance of a CDMA System Employing AMC and Multicodes in the Presence of Channel Estimation Errors

Raymond Kwan and Cyril Leung

Abstract—The impact of channel state estimation errors in a CDMA system employing adaptive modulation and coding in conjunction with multicodes is studied. The channel is modelled as a finite-state Markov chain and the performances using (1) a simple moving average (SMA) filter (2) a hidden Markov model (HMM) filter to estimate the channel state are compared. The results show that the HMM filter is more robust and provides a significant throughput improvement over the SMA filter, especially when the channel estimate is quite noisy or the normalized Doppler rate is small.

Index Terms—CDMA, adaptive modulation and coding, multicode, imperfect channel estimation, Rayleigh fading, hidden Markov model.

I. INTRODUCTION

ADAPTIVE Modulation and Coding (AMC) has been adopted in the 3GPP standard in order to improve spectral efficiency [1]. In order to increase the granularity of the adaptation and to provide higher bit rates, multicode transmission [2] is employed. Multicode transmission increases the bit rate by dividing a high rate data stream into a number of lower rate sub-streams. These sub-streams are transmitted in parallel synchronous multicode channels so that inter-stream interference is avoided in the absence of multipath.

On the downlink transmission to a target mobile station MS A, the base station (BS) typically acquires the channel state information (CSI) from MS A via an uplink feedback channel. Based on the newly acquired CSI, the BS assigns an appropriate modulation and coding scheme (MCS) and number of multicodes to MS A for use in the next scheduling period. Since the use of AMC requires knowledge of the channel state, it is important to assess the performance degradation which would result from channel state estimation errors.

In the following, we consider the allocation of MCS and multicodes for a single MS. The problem of joint optimal allocation for multiple users is more complex [3]; for example, an important issue is the order in which MS’s are selected for allocation. The near-optimal single user allocation scheme described in Section II can be used as a component in a multi-user scheme. It is assumed that orthogonal multicodes are employed and that signals using different multicodes do not interfere with each other.

This paper examines the performance which results from using (1) a simple averaging filter (2) a hidden Markov model (HMM) based filter to estimate the channel state in a CDMA system employing AMC and multicodes. An HMM formulation of the system is presented in Section II. The estimation of state transition probabilities is described in Section III. The performance evaluation measure is discussed in Section IV, followed by the presentation of some numerical results in Section V.

II. HMM FORMULATION

The assigned bit rate for a given scheduling period can take on one of a finite set of values, depending on the MCS and the number of multicodes which are selected. We choose to associate a channel state with each possible value of the assigned bit rate. The channel is then modelled as a finite state Markov Chain (FSMC) [4]. Since the true channel state is not available at the BS, a HMM formulation is appropriate [5].

Let the number of available MCS’s be denoted by J. The basic bit rate, \( \rho_j \), for MCS \( j \) is given by

\[
\rho_j = \frac{W}{g} R_c^{(j)} \log_2 M_j, \quad 1 \leq j \leq J, \tag{1}
\]

where \( W \) is the chip rate, \( g \) is the spreading factor, \( R_c^{(j)} \) is the code rate for MCS \( j \) and \( M_j \) is the number of points in the modulation constellation for MCS \( j \). Without loss of generality, the MCS’s are assumed to be labelled so that \( \rho_j \) is a monotonically increasing function of \( j \).

It is required that the frame error rate (FER) not exceed a certain maximum tolerable value denoted by \( \varepsilon_0 \). Suppose that \( n \) multicodes are used with equal power for transmission to the MS. Assuming that orthogonal multicodes are employed and that signals using different multicodes do not interfere with each other, for a given (total) received SNR value, \( \gamma \), the SNR per multicode is \( \gamma_1 = \gamma/n \). Let \( \lambda_j \) be the minimum value of \( \gamma_1 \) required for MCS \( j \) to achieve an FER of \( \varepsilon_0 \). Since \( \rho_1 < \rho_2 < \cdots < \rho_J \), it is reasonable to assume that \( \lambda_1 < \lambda_2 < \cdots < \lambda_J \). The maximum number of multicodes that can be assigned to the MS is denoted by \( N_{\text{max}} \).

It is shown in [6] that to achieve a near-optimal average throughput for a CDMA system with AMC and multicodes, the assigned MCS, \( j \), and the number, \( k_j \), of multicodes should be chosen as follows. Referring to Fig. 1, the terms \( z_{i,j}^{(\text{min})} \) and \( \tilde{\gamma}_j \) correspond to a set of decision thresholds which yield a near-optimal choice of the MCS and number of multicodes for a given value of \( \gamma \). The decision thresholds \( \tilde{\gamma}_j \) and \( z_{i,j}^{(\text{min})} \)
are given by [6]
\[
\hat{\gamma}_j = \lambda_j N_{\text{max}}, \quad 1 \leq j \leq J
\]
(2)
\[
\hat{\gamma}_j^{(\text{min})} = \begin{cases} 
\lambda_j, & j = 1 \\
\beta_j, & 1 < j \leq J \\
\infty, & j = J + 1
\end{cases}
\]
(3)
where \(\beta_j \in (0, \infty)\) is the smallest value such that
\[
\rho_j \left[ \frac{\beta_j}{\lambda_j} \right] (1 - \varepsilon_0) \geq \rho_{j-1} N_{\text{max}}, \quad 2 \leq j \leq J.
\]
(4)

From Fig. 1, it can be seen that the assigned bit rate, \(R\), for a given value of \(\gamma\) is
\[
R = \begin{cases} 
0, & \text{if } 0 \leq \gamma < \lambda_1 \\
I_{k_j} \rho_{j}, & \text{if } k_j \lambda_j \leq \gamma < (k_j + 1) \lambda_j \\
N_{\text{max}} \rho_j, & \text{if } \hat{\gamma}_j \leq \gamma < \hat{\gamma}_j^{(\text{min})}
\end{cases}
\]
(5)
where \([\hat{\gamma}_j^{(\text{min})} / \lambda_j] \leq k_j \leq N_{\text{max}} - 1\).

Note that (5) establishes a mapping between \(\gamma\) and a finite set of all possible assigned bit rate values. For notational convenience, let this set be denoted by
\[
\mathcal{R} = \{R_1, R_2, \ldots, R_N\},
\]
(6)
where \(R_1 < R_2 < \cdots < R_N\). It is evident from Fig. 1 that \(N \leq J N_{\text{max}}\). For any value of \(\gamma\) which maps to an assigned bit rate \(R_i\), we say that the channel is in state \(i\).

The throughput \(\overline{S}\) can then be written as
\[
\overline{S} \approx (1 - \varepsilon_0) \sum_{i=1}^{N} R_i p(i)
\]
(7)
where \(p(i)\) is the probability that the channel is in state \(i\).

Let \(X_k \in \{1, 2, \ldots, N\}\) denote the channel state at time \(k\). Also, let \(\pi_k(i) = P(X_k = i)\) be the probability that the channel is in state \(i\) at time \(k\), and \(\pi_k = [\pi_k(1), \ldots, \pi_k(N)]\). For small values of the normalized Doppler frequency (defined as the product of the Doppler frequency and the scheduling period), the channel can be assumed to be constant over a scheduling period and the sequence of channel states can be modelled as a first order Markov process; in the High Speed Downlink Packet Access (HSDPA) standard, the scheduling period is 2 ms [1], and such a model would be reasonably accurate for Doppler frequencies up to about 50 Hz. It should be noted that any finite order Markov process can be represented as a first order Markov process over a larger state space. Then, the state probability vector at time \(k+1\) is given by
\[
\pi_{k+1} = \pi_k A
\]
(8)
where \(A = \{a_{i,j}\}, i, j \in \{1, \ldots, N\}\) is the state transition probability matrix, and \(a_{i,j}\) is the probability that \(X_{k+1} = j\) given that \(X_k = i\).

We model the observed channel state as
\[
y_k = x_k + v_k
\]
(9)
where \(v_k\) represents the observation noise and is the outcome of an independent discrete-valued or continuous-valued random variable (rv) \(V\) with pdf \(f_V(v; \phi)\), and \(\phi\) represents a set of parameters associated with the distribution of \(V\). For convenience, we denote the conditional pdf of \(Y_k\) given \(X_k\) by
\[
b_i(y_k; \phi) = f_{Y_k}(y_k | X_k = i, \phi)
\]
(10)
\[
h_i(y_k - i; \phi) = f_{V}(y_k - i; \phi).
\]
(11)

We are interested in minimizing the variance of the state estimate, i.e. \(E[(\hat{X}_k - X_k)^2]\), where \(\hat{X}_k\) is the estimated state and \(X_k\) is the actual state.
Given a sequence of observations \( \mathbf{y}^{(k)} = [y_1, y_2, \ldots, y_k] \), from time 1 to time \( k \), the minimum variance state estimate (MVSE) is given by [7]

\[
\hat{x}_k = E[X_k|\mathbf{y}^{(k)}].
\]

(12)

Let the forward probabilities be \( \alpha_k(j) = P(X_k = j, \mathbf{y}^{(k)}) \). From (12), the MVSE is given by

\[
\hat{x}_k = \frac{\sum_{j=1}^N j \alpha_k(j)}{\sum_{j=1}^N \alpha_k(j)},
\]

(13)

where \( \alpha_k(j) \) can be computed recursively as [7]

\[
\alpha_k(j) = b_j(y_k; \phi) \sum_{i=1}^N a_{i,j} \alpha_{k-1}(i),
\]

(14)

with \( \alpha_1(m) = b_m(y_1; \phi) P(X_1 = m), m = 1, \ldots, N \) as the initial distribution.

In general, the estimated state \( \hat{x}_k \) in (13) is a real number. Since \( X_k \) only assumes values in \( \{1, 2, \ldots, N\} \), for illustration purposes we choose the final estimated state \( \hat{x}_k \) as the integer in \( \{1, 2, \ldots, N\} \) closest to \( \hat{x}_k \). From (13) and (14), we see that the calculation of the MVSE requires approximately \( N^2 + 2N \) operations, where an operation is defined as a combination of one multiplication and one addition.

### III. Estimation of Unknown Parameter Values

In section II, the transition probabilities \( \{a_{i,j}\} \) and the set \( \phi \) of parameter values for the pdf of \( V \) are assumed to be known. In practice, \( \{a_{i,j}\} \) and \( \phi \) may have to be estimated based on the observed sequence \( y_1, y_2, \ldots, y_K \). Here, we consider the use of the expectation-maximization (EM) algorithm [7] for this purpose.

Let \( X^{(K)} = [X_1, X_2, \ldots, X_K] \), where \( K \) is the number of observation samples. Also, let \( \theta \) and \( \theta^{(0)} \) be the true value and the initial estimate of the parameter, respectively. Using \( \theta^{(0)} \), the EM algorithm iteratively generates a sequence of estimates \( \{\theta^{(l)}, l = 1, 2, \ldots, L\} \) based on the following steps:

- Obtain the auxiliary likelihood function as shown in (15), (16), and (17), where \( S = \{1, 2, \ldots, N\}^K \) denotes the set of all possible state sequences of length \( K \). In (15), \( \theta = \{a_{i,j}, \phi\} \) and \( \theta^{(l)} = \{\hat{a}_{i,j}^{(l)}, \hat{\phi}^{(l)}\} \) denote the true and the estimated parameter values after \( l \) iterations respectively. The terms \( \Psi_k^{(l)}(i,j) \) and \( \Gamma_k^{(l)}(i) \) in (17) are given by

\[
\Psi_k^{(l)}(i,j) = \frac{\alpha_k^{(l)}(i) \hat{a}_{i,j}^{(l)} b_j(y_{k+1}; \hat{\phi}^{(l)}) \beta_{k+1}^{(l)}(j)}{\sum_{i=1}^N \sum_{j=1}^N \alpha_k^{(l)}(i) \hat{a}_{i,j}^{(l)} b_j(y_{k+1}; \hat{\phi}^{(l)}) \beta_{k+1}^{(l)}(j)},
\]

(18)

\[
\Gamma_k^{(l)}(i) = \sum_j \Psi_k^{(l)}(i,j).
\]

(19)

Furthermore, let \( \mathbf{y}^{(k+1:K)} = [y_{k+1}, y_{k+2}, \ldots, y_K] \), and the backward probabilities be \( \beta_k(j) = p(\mathbf{y}^{(k+1:K)}|X_k = j) \). The terms \( \alpha_k^{(l)}(j) \) and \( \beta_k^{(l)}(j) \) in (18) are given by

\[
\alpha_k^{(l)}(j) = b_j(y_k; \hat{\phi}^{(l)}) \sum_{i=1}^N \hat{a}_{i,j}^{(l)} \alpha_{k-1}^{(l)}(i),
\]

(20)

\[
\beta_k^{(l)}(j) = \sum_{j=1}^N \hat{a}_{i,j}^{(l)} b_j(y_{k+1}; \hat{\phi}^{(l)}) \beta_{k+1}^{(l)}(j),
\]

(21)

with \( \beta_{k+1}^{(l)}(j) = 1 \). More details regarding the \( Q(\theta|\theta^{(l)}) \) function can be found in [7].

- Maximize the auxiliary likelihood function to obtain the new estimate

\[
\theta^{(l+1)} = \arg \max Q(\theta|\theta^{(l)}).
\]

(22)

The estimated transition probabilities \( \{\hat{a}_{i,j}^{(l+1)}\} \) can be obtained from (22) and are given by [7]

\[
\hat{a}_{i,j}^{(l+1)} = \frac{\sum_{k=2}^{K} \hat{a}_{k-1}^{(l)}(i) \hat{a}_{i,j}^{(l)} b_j(y_k; \hat{\phi}^{(l)}) \beta_{k+1}^{(l)}(j)}{\sum_{k=1}^{K} \hat{a}_{k-1}^{(l)}(i) \beta_{k+1}^{(l)}(j)}.
\]

(23)

For a given analytic form for the pdf of \( V \), the parameter estimates \( \hat{\phi}^{(l)} \) can be obtained using a procedure similar to that described above.

### IV. Performance Measures

As discussed in [8], the effect of estimation error is asymmetrical. Over-estimation typically gives rise to a worse throughput degradation than under-estimation. Thus, the instantaneous throughput \( S_k(\hat{x}_k) \) can be approximated as

\[
S_k(\hat{x}_k) \approx \begin{cases} (1 - \epsilon_0)R_k, & \text{if } x_k \geq \hat{x}_k = i \\ 0, & \text{if } x_k < \hat{x}_k \end{cases}
\]

(24)

The average throughput is then given by

\[
\mathbb{S}_{HMM} = \lim_{M \to \infty} \frac{1}{M} \sum_{k=1}^{M} S_k(\hat{x}_k)
\]

(25)

where \( M \) is the total number of observation instances.

The average throughput, \( \mathbb{S}_{nf} \), for the case when no filtering is performed is given by (25) with \( \hat{x}_k \) replaced by \( \hat{y}_k \), where \( \hat{y}_k \) is the integer in \( \{1, 2, \ldots, N\} \) closest to \( y_k \). For the simple moving average (SMA) filter with a window size, \( Z \), the average throughput, \( \mathbb{S}_{SMA} \), is given by (25) with \( \hat{x}_k \) replaced by \( \hat{z}_k \), where \( \hat{z}_k \) is the integer in \( \{1, 2, \ldots, N\} \) closest to \( \hat{z}_k \) and

\[
\hat{z}_k = \frac{y_k - z + 1 + y_k - z + 2 + \cdots + y_k}{Z}.
\]

(26)

The average improvement factor, \( \eta \), where \( x \) represents either HMM or SMA, is defined as

\[
\eta = \left( \frac{\mathbb{S}_{HMM} - \mathbb{S}_{nf}}{\mathbb{S}_{nf}} \right) \times 100.
\]

(27)

### V. Numerical Results

Simulation results to compare the performance improvement obtainable with the HMM filter and a simple moving average filter are presented in this section. For illustrative purposes, we assume that the MCS set consists of turbo coded QPSK and 16-QAM with code rates of 1/2 or 3/4; a maximum of \( N_{max} = 2 \) multicodes can be assigned to each user; the spreading factor is 16 as specified in [9] and a target frame error rate of 1% is used. The values of \( \gamma_j^{(min)} \) and \( \gamma_j \) are obtained using the first transmission error rates of the selected MCS’s in [10]. A spreading factor of 16, a chip rate of \( 3.84 \times 10^9 \) chips/s, and a frame duration of 2 ms are assumed; these correspond to a baud rate of \( 240 \times 10^3 \) symbols per second. The MCS and number of multicodes assignment is performed at the BS once every
RV’s are independent. As a simple illustration, the observation or discrete observation noise distributions as long as the noise channel states is not exactly first order Markov.

\[ Q(\theta|\theta^{(l)}) = E \left[ \ln \left( p(X^{(K)}; y^{(K)}| \theta) \right) \right] = \sum_{X^{(K)} \in S} p(X^{(K)}| y^{(K)}, \theta^{(l)}) \ln \left[ p(X^{(K)}| y^{(K)}, \theta^{(l)}) \right] \]

(15)

\[ = \sum_{k=1}^{K} \sum_{j=1}^{N} \Gamma_k^{(j)}(j) \ln (b_j(y_k; \phi)) \]

(16)

\[ = \sum_{k=1}^{K} \sum_{j=1}^{N} \Gamma_k^{(j)}(j) \ln (a_{i,j}). \]

(17)

The HMM approach is applicable for arbitrary continuous or discrete observation noise distributions as long as the noise RV’s are independent. As a simple illustration, the observation noise \( V \) is assumed to be Gaussian distributed with a mean of 0 and a variance of \( \sigma^2_V \). Both the noise variance \( \sigma^2_V \) and the state transition probabilities \( \{a_{i,j}\} \) are estimated using the EM algorithm as described earlier. The estimate of the noise variance, \( \hat{\sigma}^2_V \), can be computed as [7]

\[
(\hat{\sigma}^2_V)^{l+1}) = \frac{\sum_{k=1}^{K} \sum_{j=1}^{N} \alpha_k^{(l)}(j) b_k^{(l)}(j)(y_k - j)^2}{\sum_{k=1}^{K} \sum_{j=1}^{N} \alpha_k^{(l)}(j)b_k^{(l)}(j)}.
\]

(28)

Note that although the estimates in (20), (23) and (28) are obtained based on a first order Markov channel model, in the simulation the estimates are applied to the simulated channel which is not exactly first order Markov. The number, \( L \), of iterations in the EM algorithm is set to 5. It was found that using a higher value for \( L \) resulted in little change in the throughput. For the results presented, the 95% confidence intervals are within ±3% of the values shown.

Fig. 2 shows the average % improvement factor, \( \eta \), for the HMM filter as a function of the observation noise variance with normalized Doppler frequencies \( f_d T \) of 0.005 and 0.025; the actual Doppler frequencies are 2.5 Hz and 12.5 Hz respectively. The performance for a system employing a simple (constant coefficients) moving average (SMA) filter with a window size of \( Z \) samples is also shown. The results show that the improvements of both filters relative to the unfiltered case increase with the observation noise variance. In all cases, the performance of the HMM filter is superior to that of the SMA filter. The results also suggest that the performance improvements for both filters decrease with Doppler rate. In particular, the performance of the SMA filter is very sensitive to the choice of \( Z \), i.e. a reasonably good value of \( Z \) for a given Doppler rate can actually result in a performance degradation at another Doppler rate. From a computational standpoint, the SMA filter requires \( Z \) operations per scheduling period which is generally fewer than the \( N^2 + 2N \) operations needed by the HMM filter. For our example, \( J = 4 \) and \( N_{\text{max}} = 2 \) so that \( N \) is at most 8.

Fig. 3 shows a plot of \( \eta \) as a function of the window size, \( Z \), for the SMA filter with an observation noise variance, \( \sigma^2_V \), of 0.36 and normalized Doppler rates \( f_d T \) of 0.005 and 0.025. For comparison, the HMM filter \( \eta \) value (which is independent of \( Z \)) is also shown. It can be seen that the performance improvement achievable with the SMA filter depends quite strongly on the window size \( Z \).
VI. CONCLUSIONS

The impact of channel state estimation errors in a system employing AMC and multicodes over a Rayleigh fading channel was studied. The use of a simple moving average (SMA) filter and an HMM filter to reduce the performance degradation was examined. It is found that the SMA filter can provide a good performance gain if the window size is chosen appropriately based on the Doppler rate and the observation noise power. The HMM filter automatically adjusts to parameter changes and provides a robust, uniformly higher performance gain.

ACKNOWLEDGMENT

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada under Grant OGP0001731 and by the UBC PMC-Sierra Professorship in Networking and Communications.

REFERENCES

Capacity of Large Array Receivers for Deep-Space Communications in the Presence of Interference

Victor Vilnrotter, Senior Member, IEEE, and Meera Srinivasan Member, IEEE

Abstract—We develop a model for a large array ground receiver system for use in deep space communications, and analyze the resulting array channel capacity. The model includes effects of array geometry, time-dependent spacecraft orbital trajectory, point and extended interference sources, and elevation-dependent noise and atmospheric channel variations. Channel capacity is expressed as a simple quadratic form dependent upon covariance matrices characterizing the source, interference, and additive noise. This formulation facilitates inclusion of array and channel characteristics into the model, as well as comparison of optimal, suboptimal and equivalent single-antenna configurations on achievable throughput. Realistic examples of ground array channel capacity calculations are presented, demonstrating the impact of array geometry, planetary interference sources, and array combining algorithm design upon the achievable data throughput.

Index Terms—Array channel capacity, interference cancellation, phased array receivers.

I. INTRODUCTION

ONE of the primary technologies for increasing downlink data rates and enabling more missions for NASA’s Deep Space Network (DSN) is the use of antenna arrays. Large arrays consisting of small to moderately sized antennas provide increased gain as well as robustness and flexibility in receiving greater numbers of spacecraft signals at higher data rates. Long antenna baselines result in greater science capabilities as well as the ability to reject interference due to the resulting narrow beamwidths, especially when coupled with advanced signal processing techniques to further cancel interfering signals. In light of these benefits, it is desirable to formulate a framework for quantifying the potential gains and limitations of array reception, incorporating specific design parameters and signaling conditions. In this paper, we develop an analytical framework in terms of the array channel capacity.

Channel capacity measures the maximum data throughput achievable on a particular channel, and is defined as the maximum mutual information between the input and output of the channel [1], [2]. Array capacity expressions for the AWGN channel have been derived in [2], and in [3] for DSN applications. The topic of multiple-input multiple-output (MIMO) system capacity has been treated extensively in the literature for AWGN channels [4]–[6] as well as for more complicated channels including fading and interference [7]–[10]. Array combining in the presence of interference has been explored specifically in [11], [12]. Here we consider a single-input, multiple-output array system with very low signal-to-interference-noise ratio (SINR) in the presence of spatially extended interference from planetary noise, as well as geographically dispersed antenna locations leading to diverse atmospheric effects. The intent is to provide an analytical basis for the ready evaluation of the extent to which changes in array channel parameters impact achievable data throughput for the specific application of array reception of deep space signals.

II. MODELING OF ARRAY CHANNEL AND SIGNALS

A. Array Geometry and Modulated Signal

The antenna element positions are referenced to the coordinate system origin; hence the position of the $i$th antenna element, $1 \leq i \leq N$, is given by the vector $\mathbf{a}_i = a_{x,i} \hat{x} + a_{y,i} \hat{y} + a_{z,i} \hat{z}$, where $\hat{x}$, $\hat{y}$, and $\hat{z}$ are unit vectors along the coordinate axes, and $\theta_{x,i}$, $\theta_{y,i}$, and $\alpha_{z,i}$ are the projections of the vector corresponding to the $i$th antenna location onto the coordinate axes. The signal direction is always changing due to earth-rotation as well as source trajectory. It can be described by the unit vector $\mathbf{s} = s_x(t) \hat{x} + s_y(t) \hat{y} + s_z(t) \hat{z}$, where the coordinate components are expressed in terms of the source elevation and azimuth angles $(\theta_s(t), \phi_s(t))$ as $s_x(t) = \cos \theta_s(t) \sin \phi_s(t)$, $s_y(t) = \cos \theta_s(t) \cos \phi_s(t)$, and $s_z(t) = \sin \theta_s(t)$. The instantaneous time delay between the arrival of the signal plane wave at the $i$th and $j$th antennas is proportional to the difference of the inner products formed by projecting the instantaneous antenna location vectors onto the source direction vector, e.g., $\mathbf{a}_i \cdot \mathbf{s} = a_{x,i} \cos \theta_s \sin \phi_s + a_{y,i} \cos \theta_s \cos \phi_s + a_{z,i} \sin \theta_s$, where the time dependence has been suppressed for notational convenience. Time dependence will in general be suppressed in the vector and matrix notation henceforth.

In the context of array reception, the quantity of interest is the phase accumulated at each antenna due to the geometry of the array and the source, since it is primarily the accumulated phase at each antenna that must be estimated and removed before the signal components are combined. The phase accumulated by the signal field in traversing the distance from the antenna first encountered by the field to the $i$th antenna is proportional to the difference between the first antenna-source inner product and the $i$th antenna-source inner product: $2\pi(a_{i1} - a_{i}) \cdot \mathbf{s}/\lambda$. Assuming that geometry-dependent delays have been removed based on predicts, we can use a narrowband array model in which the residual time delay...
results only in a phase difference between elements, so that the complex baseband signal received at the $i$th antenna without thermal noise or other impairments can be represented as

$$s_i(t) = \sqrt{P_{s,i}}d(t)e^{j(2\pi(n_i - a_i)/\lambda + \varphi_i)}.$$  

(1)

Here $P_{s,i} = \gamma_i(t)P$ is the signal power received at the $i$th antenna, where $P$ is the signal power at the transmitter and $\gamma_i(t)$ is a scale factor that includes space loss, antenna gains, etc. for the noise covariance matrix. The noise components at each antenna are independent; deterministic elevation-dependent effects, and time-varying complex baseband signal received at the $i$th antenna. In addition, $d(t)$ is the data modulation, which is assumed to be a narrowband signal, $\lambda$ is the signal wavelength, and $\varphi_i$ is a fixed phase offset due to effects other than antenna geometry, e.g., instrumentation errors. Note that since only phase differences are of interest, we can simplify (1) by ignoring the phase term common to all antennas, yielding

$$s_i(t) = \sqrt{P_{s,i}}d(t)e^{-j(2\pi n_i/\lambda - \varphi_i)}.$$  

(2)

If we denote the gain and phase terms for the $i$th antenna by $g_i(t) = \sqrt{\gamma_i(t)}e^{-j(2\pi n_i/\lambda - \varphi_i)}$, and define the column vector $G = (g_1(t), \ldots, g_N(t))^T$, then we may write the received signal vector as $s = \sqrt{P}Gd(t)$.

Since both the carrier phasor and the signal modulation are slowly time-varying terms, we continue to denote them as explicit functions of time, keeping in mind that the source vector is a slowly time-varying quantity that changes on a time-scale of tens of seconds to minutes over the narrow field-of-view of the array.

### B. Additive Noise

Even if no interfering sources are present, the receiver at each antenna observes the signal in the presence of additive noise, composed of cosmic background, a frequency and elevation-dependent atmospheric contribution that varies with meteorological conditions, and thermal noise produced within the front-end electronics of the receiver itself. These noise components are assumed to be independent zero-mean white Gaussian processes; hence a second moment characterization provides a complete statistical description.

The variance of the total additive noise at each antenna is the sum of the component variances. Cosmic background and receiver thermal noise can be combined into a single noise term with variance $\sigma^2_{\text{n},i} = \sigma^2_c + \sigma^2_{\text{th},i}$, for the $i$th antenna. The contribution of the atmosphere is denoted as a function of elevation and time as $\sigma^2_{\text{atm},i}(\theta_s,t) = \sigma^2_{\text{atm},i}(t)/\sin(\theta_s(t))$, for $\theta_s > \sim 3^\circ$. This notation allows the modeling of both deterministic elevation-dependent effects, and time-varying atmospheric conditions such as clouds rolling over the array. The noise components at each antenna are independent; hence their noise covariance matrices add, yielding the total noise covariance matrix

$$\Theta_{\text{n,total}}(\theta_s,t) = \Theta_{\text{n}} + \Theta_{\text{atm}}(\theta_s,t) = \sigma^2_cI + \text{diag}[\sigma^2_{\text{th},i}] + \text{diag}[\sigma^2_{\text{atm},i}(t)/\sin(\theta_s(t))].$$  

(3)

where $I$ is the $N \times N$ identity matrix.

### C. Interference

The primary sources of interference in the deep space channel are planetary noise and undesired modulated signals from other spacecraft. Terrestrial interference may also occur, but will likely be more sporadic in nature, unless it is due to array self-interference. Here we consider only planetary noise in the interference model. A distributed interference source such as a planet may be modeled as a spatially sampled set of narrowband (i.e., filtered to the signal bandwidth) Gaussian point interferers [13] followed by a spatial low-pass filter operation. Planetary noise sources are modeled here as a triangular lattice of point sources representing spatial samples whose spacing depends upon the beamwidth of the array. Hence, the total number of point sources $K$ that are used to model the planet depends upon the angular diameter of the planet as well as the array diameter. Each interference point source can be modeled in the same manner as the signal case developed in Section II-A, except the modulation is noise-like. The complex baseband waveform received from the $k$th interfering point source, $1 \leq k \leq K$, at the $i$th antenna, within the bandwidth of the receiver, can therefore be represented as

$$b^{(k)}_i(t) = B(t)e^{-j(2\pi n_i/\lambda + \varphi_i)},$$  

(4)

where $B(t)$ is a circularly symmetric complex Gaussian random process spanning the bandwidth of the receiver's pre-detection filter. As with the signal source, the $k$th interferer has unit direction vector $\hat{b}^{(k)}_i(t) = b^{(k)}_i(t)x + b^{(k)}_i(t)y + b^{(k)}_i(t)z$, with components $b^{(k)}_x(t) = \cos(\theta^{(k)}_b(t))\sin(\phi^{(k)}_b(t))$, $b^{(k)}_y(t) = \cos(\theta^{(k)}_b(t))\cos(\phi^{(k)}_b(t))$, $b^{(k)}_z(t) = \sin(\theta^{(k)}_b(t))$, where $\theta^{(k)}_b(t)$ and $\phi^{(k)}_b(t)$ are the elevation and azimuth angles of the $k$th interfering signal. Note that the temporal variation of the geometric phase term in (4) is slow relative to the broadband noise-like variation in $B(t)$. Assuming that delay differences between antennas have been removed to an accuracy commensurate with the receiver bandwidth, we can model the background interference as a circularly symmetric complex Gaussian process with zero mean and variance $\text{Var}[B(t)] = E[|B(t)|^2] = 2\sigma^2_B$. The covariance matrix of the $k$th interferer, $\Theta^{(k)}_b$, thus has entries of the form

$$E[b^{(k)}_i(t)b^{(k)*}_j(t)] = 2\sigma^2_b e^{j(2\pi(n_i - n_j)/\lambda + \varphi_i - \varphi_j)}.$$  

(5)

Note that the interference covariance matrix is a function of time, although the time dependence is dropped from the right-hand side of (5). For a cluster of interferers, the covariance entries for each point can be determined as in (5). Assuming that each interferer generates independent noise processes, the covariance matrix of the sum of interferers is the sum of the individual covariance matrices, $\Theta^{(k)}_b$.

### D. Atmospheric Effects

As signals pass through the atmosphere, they are affected in two ways. The signal power is attenuated by an amount proportional to the thickness of the atmosphere, and the phase is perturbed by the turbulent medium. The Gaussian noise contributed by the atmosphere was discussed in Section II-B. Assume that each antenna element sees a column of atmosphere independent of the other antennas, which is valid
for antenna diameters on the order of 10 m or larger at microwave frequencies, and spaced sufficiently far apart to avoid shadowing at low elevations. Then we can define the atmospheric attenuation and phase matrices \( \mathbf{H}_{\text{atm}} = \text{diag} [h_i(\theta, t)] \) and \( \mathbf{\Psi}_{\text{atm}} = \text{diag} [\psi_i(\theta, t)] \), where \( h_i(\theta, t) \) and \( \psi_i(\theta, t) \) are the elevation and time dependent atmospheric gain and phase terms at each antenna. The received signal at the \( i \)th antenna becomes
\[
\hat{s}_i(t) = h_i(\theta, t) \sqrt{P_i} d(t) e^{j(-2\pi n_i/\lambda + \varphi_i(\theta, t))}
\]
and the interference received at the \( i \)th antenna from the \( k \)th interferer may be expressed as
\[
\hat{b}_i^{(k)}(t) = h_i(\theta, t) B(t) e^{j(-2\pi n_i/\lambda + \varphi_i(\theta, t))},
\]
where \( \varphi_i \) represents residual instrumental phase effects, as before. Recalling that for diagonal matrices,
\[
e^{j\varphi} = I + j\varphi - \frac{\varphi^2}{2} + \cdots = \text{diag}[e^{j\varphi}]
\]
we can separately incorporate the amplitude and phase effects of the atmosphere on the observed signal and interference column vectors as
\[
\tilde{s} = \mathbf{H}_{\text{atm}} e^{j\mathbf{\Psi}_{\text{atm}}} s = \mathbf{H}_{\text{atm}} e^{j\mathbf{\Psi}_{\text{atm}}} \mathbf{G} \sqrt{P} d(t)
\]
and
\[
\tilde{b}^{(k)} = \mathbf{H}_{\text{atm}} e^{j\mathbf{\Psi}_{\text{atm}}} \tilde{b}^{(k)}, \quad 1 \leq k \leq K.
\]
Note that we have assumed that both the signal and interference sources are subject to the same atmospheric effects. The signal outer product may now be written as
\[
\tilde{s} \tilde{s}^\dagger = \mathbf{H}_{\text{atm}} e^{j\mathbf{\Psi}_{\text{atm}}} ss^\dagger e^{-j\mathbf{\Psi}_{\text{atm}}} \mathbf{H}_{\text{atm}}^\dagger = P|d(t)|^2 \mathbf{H} \mathbf{G} \mathbf{G}^\dagger \mathbf{H}^\dagger,
\]
where \( \dagger \) is used to denote conjugate transpose. The interference covariance matrices become
\[
\tilde{\Theta}^{(k)} = E[\tilde{b}^{(k)} \tilde{b}^{(k)}^\dagger] = E[\mathbf{H}_{\text{atm}} e^{j\mathbf{\Psi}_{\text{atm}}} \tilde{b}^{(k)} \tilde{b}^{(k)} e^{-j\mathbf{\Psi}_{\text{atm}}} \mathbf{H}_{\text{atm}}^\dagger] = \mathbf{H} \tilde{\Theta}^{(k)} \mathbf{H}^\dagger, \quad 1 \leq k \leq K.
\]
where we define the complex channel matrix \( \mathbf{H} = \mathbf{H}_{\text{atm}} e^{j\mathbf{\Psi}_{\text{atm}}} \). Having modeled all significant noise sources, the general form of the noise plus interference covariance matrix may be expressed as
\[
\tilde{\Theta}_{\tilde{Z}} = \tilde{\Theta}_n + \tilde{\Theta}_{\text{atm}}(\theta, t) + \sum_{k=1}^K \tilde{\Theta}^{(k)},
\]
where \((\theta, t)\) denotes variation with elevation and time. Although not stated explicitly, we note that elevation and time dependence in the signal and interference terms are incorporated into the covariance matrix through the atmospheric attenuation and atmospheric phase matrices.

### III. Array Capacity

The capacity of an array of Gaussian channels has been derived in [3], where the problem was formulated as a single real source, multiple input - multiple output channel. We now generalize the array capacity derivation for the complex channel model shown in Fig. 1 and include array geometry, atmospheric effects, and directional Gaussian interference as described in the previous section. We assume that large delay variations have been removed, leaving only small residual delays that can be adequately modeled as a phase shift on the modulated carrier. We also drop the time dependence in the notation of this section, with the implicit understanding that we are considering a given time instant in all expressions. In Fig. 1, the source \( \mathbf{X} \) undergoes different gains and geometric phase offsets (denoted by the vector \( \mathbf{G} \)) and is subjected to the atmospheric effects represented by the matrix \( \mathbf{H} \). The vector of signals arriving at the \( N \) antennas may be denoted by \( \tilde{\mathbf{X}} = \mathbf{H} \tilde{\mathbf{G}} \mathbf{X} \). The additive noise vector \( \tilde{\mathbf{Z}} \) consists of complex circular Gaussian interference signals subject to geometric phase offsets and atmospheric effects as well as Gaussian cosmic and receiver thermal noise, having covariance matrix \( \tilde{\mathbf{\Theta}}_{\tilde{Z}} \) given in eq. (8). Note that \( \tilde{\mathbf{Z}} \) is a circularly symmetric complex Gaussian random process, as it is the sum of linearly transformed independent circularly symmetric complex Gaussian processes [6]. The received array vector is then \( \tilde{\mathbf{Y}} = \tilde{\mathbf{X}} + \tilde{\mathbf{Z}} \).

The array capacity \( C_{\text{array}} \) is defined as the maximum mutual information between the source \( \mathbf{X} \) and the array output observables \( \tilde{\mathbf{Y}} = (\tilde{Y}_1, \tilde{Y}_2, \ldots, \tilde{Y}_N)^T \) over all possible distributions of source symbols, subject to the power constraint \( E(|\mathbf{X}|^2) = P \). The capacity can be expressed as a maximization of the difference between the entropy of the vector observable \( H(\mathbf{Y}) \) and the entropy of the additive noise \( H(\mathbf{Z}) \) as follows:
\[
C_{\text{array}} = \max_p \{ I(\mathbf{Y}; X) \} = \max_p \{ H(\mathbf{Y}) - H(\mathbf{Y}|X) \} = \max_p \{ H(\mathbf{Y}) - H(\mathbf{H} \mathbf{G} \mathbf{X} + \tilde{\mathbf{Z}}|X) \} = \max_p \{ H(\mathbf{Y}) \} - H(\tilde{\mathbf{Z}})
\]
where \( H(\mathbf{Y}|X) \) is the conditional entropy of the vector \( \mathbf{Y} \) averaged over the source alphabet \( X \), and \( p \) is the probability distribution of the source symbols. The last equality in (9) follows because \( \mathbf{G} \) and \( \mathbf{H} \) are known, and \( \tilde{\mathbf{Z}} \) is independent of \( X \). The power constraint on \( X \) along with the known quantities \( \mathbf{G} \) and \( \mathbf{H} \) determine the covariance matrix of \( \mathbf{Y} \):
\[
\mathbf{\Theta}_Y = E(\mathbf{Y} \mathbf{Y}^\dagger) = E \left( (\mathbf{H} \mathbf{G} \mathbf{X} + \tilde{\mathbf{Z}}) (\mathbf{H} \mathbf{G} \mathbf{X} + \tilde{\mathbf{Z}})^\dagger \right) = \mathbf{P} \mathbf{H} \mathbf{G} \mathbf{H}^\dagger + \tilde{\mathbf{\Theta}}_{\tilde{Z}} = \mathbf{\Theta}_X + \tilde{\mathbf{\Theta}}_{\tilde{Z}}
\]
where we have defined \( \tilde{\mathbf{\Theta}}_X = \mathbf{E}(\tilde{\mathbf{X}} \tilde{\mathbf{X}}^\dagger) = \mathbf{P} \mathbf{H} \mathbf{G} \mathbf{H}^\dagger \).

As \( \tilde{\mathbf{Z}} \) is the sum of independent circularly symmetric complex Gaussian processes, it is also a circularly symmetric complex Gaussian process with entropy [6]
\[
H(\tilde{\mathbf{Z}}) = \log_2((\pi e)^N |\tilde{\mathbf{\Theta}}_{\tilde{Z}}|).
\]
Since $Y$ is the sum of the signal vector modified via a linear transformation plus independent circularly symmetric complex Gaussian processes, it follows that $H(Y)$ is maximized when $X$ is a circularly symmetric complex Gaussian process, yielding the maximum value of the entropy as

$$H_{\text{max}}(Y) = \log_2(\pi e)^N |\Theta_Y|| = \log_2(\pi e)^N |\Theta_X + \Theta_Z||.$$  

Substituting eqs. (12) and (11) into (9), we obtain

$$C_{\text{array}} = \log_2 \left( \frac{|PHG(\Theta H)^\dagger + \Theta Z|}{|\Theta Z|} \right).$$  

Note that the covariance matrix of the modified source vector, $\Theta_X$, is equivalent to the signal outer product as given in Eq. (6), assuming that $E[d(t)]^2 = 1$, i.e., $\Theta_X = E[XX^\dagger] = \tilde{ss}^\dagger$. We can thus write

$$C_{\text{array}} = \log_2 \left( \frac{|\tilde{ss}^\dagger + \Theta Z|}{|\Theta Z|} \right) = \log_2 \left( 1 + \tilde{s}^\dagger \Theta Z^{-1} \tilde{s} \right).$$  

The last expression in Eq. (14) containing the quadratic form is the result of applying an identity from linear algebra [14, Corollary A.3.1].

In order to achieve the full performance gain of an antenna array, the antenna output observables $Y$ must be combined appropriately. Let us consider a receiver structure that uses a linear combination of the array observables, i.e., $w^\dagger Y$, where $w$ is a weight vector. The maximum combined SINR is achieved when the weights are given by the Wiener solution $w^* = \Theta Z^{-1} \tilde{s}$, which also yields the quadratic form appearing in (14):

$$\text{SINR} = \frac{w^* \tilde{s} \tilde{s}^\dagger w^*}{w^\dagger \Theta Z w^*} = \frac{\tilde{s}^\dagger \Theta Z^{-1} \tilde{s} \tilde{s}^\dagger \Theta Z^{-1} \tilde{s}}{\tilde{s}^\dagger \Theta Z^{-1} \tilde{s} \Theta Z^{-1} \tilde{s}} = \tilde{s}^\dagger \Theta Z^{-1} \tilde{s}.$$  

This implies that the operation consisting of multiplying each array output with the optimum Wiener weight and summing, thus converting the vector channel into a single channel, yields the same capacity for the single channel as for the array channel. However, this conclusion holds only for optimum weights; any error in estimating the weights necessarily decreases the SINR [15], causing a decrease in array capacity.

In the absence of spatially resolved background interference, i.e., a point source, the quadratic form in (14) may be expressed as

$$s^\dagger \Theta Z^{-1} \tilde{s} = \sum_{i=1}^{N} P_s|h_i|^2/\left(\sigma^2_{\text{atm},i}/\sin(\theta) + \sigma^2_{n,i}\right) = \sum_{i=1}^{N} \text{SNR}_i,$$

where we recognize the terms inside the summation as the effective signal-to-noise ratio observed by the $i$-th antenna, SNR$_i$, incorporating elevation dependent atmospheric noise and signal attenuation as well as independent thermal noise components. This results in the familiar capacity equation

$$C_{\text{array}} = \log_2 \left( 1 + \sum_{i=1}^{N} \text{SNR}_i \right).$$  

Note that in the preceding analysis, we assumed that the phase could be measured with sufficient accuracy to justify treating it as a deterministic quantity. This would typically be the case when high data-rate telemetry is received, ensuring high SNR in a relatively long phase measurement since a minimum symbol SNR must be maintained for the communications channel.
IV. NUMERICAL RESULTS

In this section we present examples of large array capacity calculations using the models developed in the previous sections. In these calculations, the antenna elements are assumed to be isotropic, and the signal sources are Gaussian. Atmospheric attenuation of the signal is not modeled in these numerical results.

The presence of an interfering source decreases channel capacity by reducing the signal-to-interference-noise ratio (SINR). One of the primary benefits of arraying is the ability to beamform and reject interference. As array diameter increases, the array beamwidth becomes narrower, allowing better discrimination between signal and interference. In addition to array diameter, the effect of array element placement is also of interest, as it is possible to optimize array capacity as a function of antenna position [4]. Fig. 2(a) illustrates a regular (uniformly-spaced) 100-element array as well as an irregular array layout that has been designed for the Deep Space Network [16]. A desired signal source is placed at azimuth angle $\phi_s = \pi/3$ and elevation angle $\theta_s = \pi/4$. A Gaussian point interference source equal in power to the desired signal is placed at the same azimuth angle $\phi_b = \phi_s$, while its elevation angle $\theta_b$ is varied in its spacing from the desired signal. Fig. 2(b) shows the capacity at Ka-band (32 GHz) as a function of the signal-interference separation in elevation angle for the regular and irregular array layouts of Fig. 2(a). We assume that the ideal combined symbol SNR (the combined symbol SNR in the absence of interference) is -5 dB, which is appropriate for turbo-coded deep space signals. When the source-interference separation is less than an array beamwidth (approximately 10 $\mu$rad), the capacity is significantly reduced from its maximum value of approximately 0.4 bits per channel use. Note that the regular array suffers additional drops in capacity at separations of 1, 6, and 30 milliradians, due to grating lobes in the array beam pattern admitting more interference. The irregular array does not suffer from this defect, having been designed to minimize grating lobes.

While it is clear that the irregular spacing of antenna elements will mitigate large losses from interference admitted though grating lobes, it is less clear how different irregular arrays compare with each other. In Fig. 3(a) we show three different irregular array layouts, each comprising 50 Ka-band antenna elements. The first set of antennas are selected randomly from the irregular layout shown in Fig. 2(a), the second set is chosen from the "eastern" antennas, and the third set is chosen from the "northern" antennas. A spacecraft trajectory in a 5000 km orbit around Mars is assumed (with occultation periods ignored), and capacity is calculated as a function of time over this track for a system with an ideal combined symbol SNR of 2 dB over 50 array elements. Mars is assumed to have an angular diameter of 30 microradians and is modeled as a cluster of 37 Gaussian point sources whose total summed interference power is nominally one-fifth of the signal power, based on a simple planetary noise model. In addition, elevation-angle-dependent atmospheric noise is included and set equal to the thermal noise (the small contribution from cosmic background noise is ignored). Fig. 3(b) shows the resulting capacity for the three different antenna sets. The change in capacity over the track is due to spacecraft elevation profile and atmospheric noise that increases as elevation decreases. The periodic dips in capacity are due to the changes in separation between the signal and the planetary interferer as the spacecraft orbits around Mars. We observe that there are differences in capacity for the three sets, with the uniform set 1 performing better when Mars is directly behind the spacecraft and sets 2 and 3 generally performing better than set 1 when the spacecraft is near zenith and is maximally separated from Mars. These effects are due to the quality of the beam formed by the various array configurations, and demonstrate that selection of antenna elements to support a particular spacecraft track can affect the achievable data throughput by 0.3–0.4 dB.

We next consider the effect of various types of processing upon capacity in Fig. 4, using the same spacecraft orbit as in the previous example. For the irregular 100-element Ka-band
array shown in Fig. 2(a), capacity is shown as a function of time, with Mars at maximum and intermediate ranges, and an ideal combined symbol SINR of 2 dB. At maximum range, the angular diameter of Mars as seen from Earth is about 15 microradians, resulting in an interference model consisting of only one point. At an intermediate range corresponding to about 45 microradians angular diameter, Mars is modeled as a cluster of 37 Gaussian point sources.

We consider four types of signal processing. The first is optimal array processing, which maximizes combined output SINR and hence maximizes capacity. Optimal processing for the Gaussian channel uses the Wiener array combining weights as explained in Section III. The second type of array processing is a suboptimal method in which the steering vector for the desired signal is used to point the array elements in the proper direction. Third, we compute the capacity of a compact array in which the antennas are placed very close to each other (but ignore shadowing effects). We assumed use of one hundred 12-m antennas, so that the total array diameter is approximately 120 meters. Finally, we compute the capacity of a single antenna with steering capability but no beam-shaping, whose collecting area is equivalent to that of the entire array. As in Fig. 3, we see that capacity oscillates with spacecraft-Mars separation, but that dips in capacity are much greater when Mars is at intermediate range due to higher interference power. Note that the peak capacity values with optimal array processing are similar in the two cases, demonstrating that this array has similar interference rejection capability at those points for both Mars distances. The suboptimal steering vector array combining method performs almost identically to the optimal array processing when Mars is at maximum range (Fig. 4a), but is noticeably worse when Mars is closer (Fig. 4b), since it does not perform interference nulling. Finally, we observe a large array processing gain relative to the performance of...
a single antenna equal in aperture to the sum of the array elements. Note that the single antenna has the same ideal SNR as the array, but lower SINR, because it lacks the resolution and degrees of freedom to perform interference rejection. Fig. 4b shows that optimal array processing provides up to 5 times (7 dB) improvement in capacity over performance with a single antenna of equal collecting aperture, as well as significant gains over compact array and steering-vector array processing. The compact array performs better than the single antenna, because although its resolution is merely that of the single large antenna, it is able to perform some degree of interference rejection due to its beamforming capability.

V. CONCLUSIONS

The channel capacity of an array system in a Gaussian channel was investigated. An analytical framework was developed to incorporate the effects of array geometry, interference, and atmospheric attenuation and phase into the capacity calculations. Numerical results were presented showing the impact of array configuration, spacecraft orbit, and planetary interference upon achievable data throughput. Of particular significance was the demonstration of distinct improvement in achievable throughput when using the optimal Wiener array combining weights in the presence of interference rather than combining elements based only upon knowledge of the signal steering vector.

REFERENCES

Generalized LDPC Codes and Generalized Stopping Sets

Nenad Miladinovic and Marc P. C. Fossorier, Fellow, IEEE

Abstract—A generalized low-density parity check code (GLDPC) is a low-density parity check code in which the constraint nodes of the code graph are block codes, rather than single parity checks. In this paper, we study GLDPC codes which have BCH or Reed-Solomon codes as subcodes under bounded distance decoding (BDD). The performance of the proposed scheme is investigated in the limit case of an infinite length (cycle free) code used over a binary erasure channel (BEC) and the corresponding thresholds for iterative decoding are derived. The performance of the proposed scheme for finite code lengths over a BEC is investigated as well. Structures responsible for decoding failures are defined and a theoretical analysis over the ensemble of GLDPC codes which yields exact bit and block error rates of the ensemble average is derived. Unfortunately this study shows that GLDPC codes do not compare favorably with their LDPC counterpart over the BEC. Fortunately, it is also shown that under certain conditions, objects identified in the analysis of GLDPC codes over a BEC and the corresponding theoretical results remain useful to derive tight lower bounds on the performance of GLDPC codes over a binary symmetric channel (BSC). Simulation results show that the proposed method yields competitive performance with a good decoding complexity trade-off for the BSC.

Index Terms—BCH codes, BEC, BSC, generalized stopping sets, GLDPC, LDPC, RS codes, stopping sets.

I. INTRODUCTION

LDPC codes were first proposed in [1], [2] and later resurrected in [3], [4]. A regular LDPC code is defined as the null space of a sparse binary $M \times N$ matrix $H$ with the following properties: constant row weight $L$, constant column weight $J$ and no two rows/columns have more than one position in common in their support. This definition of a LDPC code readily translates into an undirected bipartite graph with $N$ symbol nodes and $M$ single parity check (SPC) constraint nodes, where the matrix $H$ is the graph adjacency matrix. Tanner generalized the conventional LDPC code construction by keeping the sparse graph representation but replacing the SPC constraint nodes with error-correcting block codes which he referred to as subcodes, thus creating generalized LDPC (GLDPC) codes [5]. An advantage of GLDPC codes is that more powerful decoders can be employed at the constraint nodes during decoding, therefore yielding better performance. Recently GLDPC codes were investigated in [7]–[10]. Hamming codes were proposed as component codes in all these papers and soft-input soft-output (SISO) decoders in constraints were considered in [7], [9], [10]. The weighted bit-flipping voting algorithm proposed in [8] uses the error correcting capability of the subcodes at constraint nodes to determine positions of unreliable symbols more accurately and therefore improves the performance upon the existing bit-flipping algorithms. Notable exception in the study of GLDPC codes can be found in [10], [14] where BCH codes where first considered as potential component codes. However, the authors considered maximum-likelihood (ML) decoding on a BSC and were concerned with obtaining a decoding threshold and code minimum distance using the ensemble average word weight distribution. An idea similar to that of GLDPC codes was also explored in [11]–[13], with serial and parallel concatenated codes. It is proved that such constructions attain capacity for a BSC under a linear-time iterative decoding. However, codes constructed on expander graphs and (ML) decoding of the component codes are considered. Two interesting approaches at GLDPC codes have been presented in [15], [16]. In [15] the authors propose constructions of structured GLDPC codes in which local codes themselves are LDPC or GLDPC codes. The proposed construction concentrates on maximizing the girth and minimum distance of the global code to maximize the minimum distance of the concatenated code [5]. The proposed codes achieved coding gain up to 11dB over 40Gb/s optical channel. In [16] the authors propose doubly generalized LDPC codes. These codes employ local codes at both variable and check nodes.

In this paper we are concerned with iterative decoding of GLDPC codes over a binary symmetric channel (BSC) and are interested in algebraic bounded distance decoding (BDD) only in constraint nodes. In [10], [14] BCH codes were first considered in context of GLDPC codes and ML decoding. This paper also includes RS codes as possible component codes and investigates performance of GLDPC codes under iterative decoding. The BCH and RS codes achieve higher error correcting capability at a small expense in complexity...
compared to Hamming codes, as BDD of these codes remains relatively simple and fast. Although most of the research on GLDPC codes focused on AWGN channel and soft decision decoding using different flavors of MAP BP decoding we believe that implementation simplicity makes this codes competitive. These codes could find potential implementation in free-space optical communication as required BER is not as low as for the long-haul optical communication.

II. CODE CONSTRUCTION AND DECODING

In [5], Tanner generalized LDPC codes by allowing constraint nodes to be block codes. Namely, all symbol nodes connected to the constraint node must form a valid codeword in a specified block code, rather than satisfying a SPC. Therefore, the code graph of a GLDPC code can be defined with the same sparse adjacency matrix \( H \) as the code graph of a regular LDPC code, but with a different interpretation of the constraint nodes. Hence, each constraint node of a GLDPC code can be broken into a number of SPC nodes corresponding to the parity checks of the block code in the constraint node. This new code graph defines a GLDPC code in which the parity check matrix can be obtained in the following manner. In every row of \( H \), each “1” is replaced with a column from the parity check matrix defining the block code in the constraint node and each zero is replaced with a zero column vector. The assignment of columns of the subcode parity check matrix to the symbol node positions was studied by Tanner [5], who showed that random assignments produce good codes. Consider the construction of a GLDPC with \((n, k, d) = (2^m - 1, d = 2t + 1)\) primitive BCH codes as subcodes. We start with a \( M \times N \) parity check matrix \( H \) of a LDPC code with column weight \( J \) and row weight \( n \). Using a random assignment of positions, we obtain \( M(n - k) \times N \) parity check matrix \( H_{\text{glldpc}} \). In order to define a nontrivial code, \( M(n - k) \) should be less than \( N \). Therefore, \( H_{\text{glldpc}} \) defines a code with rate \( r \geq 1 - \frac{m+1}{n} \). Since we are interested in high rate codes we consider only codes with column weight \( J = 2 \). The construction of GLDPC codes with RS \((n, k, d) = (2^m - 1, d = 2t + 1)\) codes as components is similar. The only difference is that the resulting matrix \( H_{\text{glldpc}} \) is no longer binary but a matrix over \( \text{GF}(2^m) \) and therefore defines a GLDPC code over \( \text{GF}(2^m) \) with rate \( r \geq 1 - \frac{2t+1}{2t} \). Generalized LDPC codes can also be decoded with message passing algorithms. Depending on the decoder employed in constraint nodes, the decoder can either have the form of a belief propagation decoder, (if an APP decoding is used to decode subcodes [7]), or the form of a bit flipping decoder if algebraic decoding is performed on subcodes [8]. In this paper we are concerned with the BSC channel and therefore we use the latter type of decoding with BDD. The decoder implemented in this paper has the following message update rules. Suppose that symbol node \( i \) is connected to the constraint nodes \( j \) and \( k \). The constraint node \( j \) sends two messages to symbol node \( i \). It sends message \( c_{ji} \) which is the value of the symbol node \( i \) obtained from the subcode decoder, and it sends message \( C_{ji} \) that contains information on whether the decoding at the constraint node has succeeded. This is obtained at no extra cost. Therefore, \( C_{ji} \) is a binary valued indicator that takes value 1 if the decoded word is a codeword or 0 if the decoded word is not a codeword. The same rule applies to the messages coming from node \( k \). The symbol node message \( v_{ij} \) depends on the value received from the channel \( r_i \), and the messages received from constraint nodes other than node \( j \). Since we are concerned only with \( J = 2 \) codes, there is only one other constraint node. Hence, the updating rule in the symbol node can be expressed as:

\[
v_{ij} = r_i \cdot \overline{C_{ki}} + c_{ki} \cdot C_{ki}
\]

where \( \overline{C_{ki}} \) is the complement of \( C_{ki} \). The decoding is performed until all constraints are satisfied or until the maximum number of iterations is reached. At the end, the value of each symbol is determined as the value proposed by a satisfied constraint. If both constraints are satisfied one which is closer in Hamming distance to the received symbol. If both constraints are unsatisfied, the received value is retained.

III. INFINITE LENGTH ANALYSIS

The infinite length analysis is usually the first step in analyzing of iterative decoding algorithms for codes constructed over graphs. Within this framework, determination of a threshold and convergence analysis are the two most important characteristics associated with an iterative decoding algorithm. We intend to develop a convergence model for GLDPC codes with RS and BCH codes as subcodes under the assumption of an infinite length linear code (code with a tree like code graph) when used on a BSC and a BEC. To make the density evolution analysis we also assume that all-zero codeword is transmitted.

A. Infinite Length Analysis on a BSC

Consider the message passing decoding algorithm described in Section II, performed on a tree like code graph with \( t \) error correcting RS codes as component codes and let \( p_s^i \) be the average symbol error rate at iteration \( i \). Then at iteration \( i+1 \), a symbol is in error if decoding in the constraint node failed and the symbol was received in error, or if the decoder made an error and the value sent is the incorrect one. Since we consider a nonbinary (RS) code as component code over a BSC, we use a uniform discrete symmetric channel (UDSC) model to find the corresponding decoder failure and error rates. Consider a RS code over \( \text{GF}(q) (q = 2^m) \) and a BSC with crossover probability \( \epsilon \). Then the parameters of the UDSC are \( s = (1 - \epsilon)^m \) and \( p = \frac{1 - s}{q - 1} \). The corresponding error rates are:

\[
P_E(p_i, s_i) = \sum_{j = 2t+1}^{n} A_i \sum_{k = 0}^{\lfloor t \rfloor} P_k^j(p_i, s_i),
\]

\[
P_E(p_i, s_i) = \sum_{j = 2t+1}^{n} A_i \sum_{k = 0}^{\lfloor t \rfloor} P_k^j,
\]

\[
P_k^j(p_i, s_i) = \sum_{r = 0}^{\lfloor j \rfloor} \binom{j}{k-r} \binom{n-j}{r} p_i^{j-k+r}(1-p_i)^{k-r}(1-s_i)^{j-r},
\]

\[
P_F(p_i, s_i) = 1 - \sum_{j = 0}^{n-1} (1-s_i)^{j}s_i^{n-1-j} - P_E(p_i, s_i),
\]
TABLE I
THRESHOLD VALUES FOR SOME GLDPC CODES WHEN USED ON BSC

<table>
<thead>
<tr>
<th>Subcode</th>
<th>Rate</th>
<th>$\epsilon^*$ [dB]</th>
<th>$\epsilon_{\text{opt}}$ [dB]</th>
<th>$\Delta$ [dB]</th>
<th>$\Delta_{\text{max}}$ − $\epsilon^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>RS(31,31,3)</td>
<td>0.067</td>
<td>0.0582 [4.41]</td>
<td>0.0428 [4.04]</td>
<td>2.17</td>
<td>0.64</td>
</tr>
<tr>
<td>RS(31,25,9)</td>
<td>0.498</td>
<td>0.0478 [4.58]</td>
<td>0.1195 [3.70]</td>
<td>2.67</td>
<td>0.24</td>
</tr>
<tr>
<td>RS(31,25,7)</td>
<td>0.613</td>
<td>0.0432 [4.33]</td>
<td>0.0789 [2.23]</td>
<td>2.10</td>
<td>0.102</td>
</tr>
<tr>
<td>RS(31,27,5)</td>
<td>0.742</td>
<td>0.0182 [4.70]</td>
<td>0.0405 [2.97]</td>
<td>1.73</td>
<td>0.144</td>
</tr>
<tr>
<td>RS(31,53,11)</td>
<td>0.603</td>
<td>0.0232 [6.65]</td>
<td>0.0755 [2.60]</td>
<td>2.02</td>
<td>0.151</td>
</tr>
<tr>
<td>RS(31,55,9)</td>
<td>0.746</td>
<td>0.0186 [4.64]</td>
<td>0.0405 [2.94]</td>
<td>1.70</td>
<td>0.124</td>
</tr>
<tr>
<td>RS(63,57,7)</td>
<td>0.809</td>
<td>0.0332 [4.83]</td>
<td>0.0820 [3.44]</td>
<td>1.39</td>
<td>0.098</td>
</tr>
<tr>
<td>RS(63,59,5)</td>
<td>0.873</td>
<td>0.0090 [4.51]</td>
<td>0.0175 [4.04]</td>
<td>0.97</td>
<td>0.072</td>
</tr>
<tr>
<td>RS(127,121,7)</td>
<td>0.906</td>
<td>0.0855 [5.53]</td>
<td>0.0175 [4.04]</td>
<td>0.97</td>
<td>0.072</td>
</tr>
<tr>
<td>RS(127,123,3)</td>
<td>0.937</td>
<td>0.0927 [6.80]</td>
<td>0.0145 [3.00]</td>
<td>1.16</td>
<td>0.106</td>
</tr>
<tr>
<td>RS(255,249,7)</td>
<td>0.953</td>
<td>0.0024 [4.22]</td>
<td>0.0202 [3.88]</td>
<td>0.84</td>
<td>0.053</td>
</tr>
<tr>
<td>RS(255,251,3)</td>
<td>0.966</td>
<td>0.0800 [5.80]</td>
<td>0.0172 [3.21]</td>
<td>0.98</td>
<td>0.078</td>
</tr>
<tr>
<td>RS(255,251,5)</td>
<td>0.955</td>
<td>0.0969 [4.46]</td>
<td>0.0166 [2.14]</td>
<td>2.92</td>
<td>0.075</td>
</tr>
<tr>
<td>RS(31,25,7)</td>
<td>0.428</td>
<td>0.0776 [5.80]</td>
<td>0.1372 [1.49]</td>
<td>2.31</td>
<td>0.103</td>
</tr>
<tr>
<td>RS(63,51,5)</td>
<td>0.619</td>
<td>0.0397 [5.91]</td>
<td>0.0741 [2.50]</td>
<td>1.65</td>
<td>0.099</td>
</tr>
<tr>
<td>RS(127,106,7)</td>
<td>0.669</td>
<td>0.0645 [5.37]</td>
<td>0.0648 [2.53]</td>
<td>1.04</td>
<td>0.051</td>
</tr>
<tr>
<td>RS(127,111,5)</td>
<td>0.779</td>
<td>0.0284 [4.81]</td>
<td>0.0395 [3.23]</td>
<td>0.58</td>
<td>0.037</td>
</tr>
<tr>
<td>RS(255,231,7)</td>
<td>0.812</td>
<td>0.0202 [4.13]</td>
<td>0.0264 [3.47]</td>
<td>0.66</td>
<td>0.062</td>
</tr>
<tr>
<td>RS(255,239,9)</td>
<td>0.875</td>
<td>0.0331 [4.51]</td>
<td>0.0722 [4.08]</td>
<td>0.45</td>
<td>0.057</td>
</tr>
</tbody>
</table>

where $P_{E}(p_i, s_i)$ is the decoder word error rate, $P_{E}(p_i, s_i)$ is the decoder symbol error rate, $P_{E}^{\text{b}}(p_i, s_i)$ is the probability that a received word is at distance $k$ from a codeword of weight $j$. $P_{E}^{\text{b}}|p_0^{\text{a}}$ is the decoder failure rate conditioned on the fact that one symbol is in error, $s_i$ is the probability that a symbol is correct and $p_i$ is the probability that a symbol is mistaken for another particular symbol. Therefore the convergence equations for the case of RS component codes are:

$$p_{i+1}^{a} = p_0^{a} P_{E}^{\text{b}}|p_0^{\text{a}}(s_i) + P_{E}^{\text{b}}(p_i, s_i)$$

where $P_{E}(p_i, s_i)$ is the decoder word error rate, $P_{E}^{\text{b}}(p_i, s_i)$ is the decoder bit error rate. Note that equations (2) and (6) hold only in case of cyclic codes [22]. In general, an input-output weight enumerators of a code are needed to obtain an exact symbol ($P_{E}^{\text{b}}$) or bit ($P_{E}^{\text{b}}$) error rates. The threshold values for some RS and BCH GLDPC codes are given in Table I with following notation: $\epsilon^*$ represents the threshold of a GLDPC code, $\epsilon_{\text{opt}}$ represents capacity of a BSC and $\Delta$(dB) represents capacity gap in dB.

In order to gain insight about the performance of GLDPC codes compared to that of LDPC codes when used over a BSC, we compare the thresholds of GLDPC codes to the thresholds of the best regular LDPC codes. The code rates can be matched only for few LDPC codes. However, it is enough to illustrate the general trend. It is readily verified that the thresholds of GLDPC codes with RS codes as components are comparable to the thresholds of the best regular LDPC codes suggesting that similar performance could be expected. However, the thresholds of GLDPC codes with BCH codes as components are significantly improved, especially for high rates. First we compared the GLDPC code with the BCH(63,45,7) code as component to a (4,7) LDPC code with threshold $\epsilon^* = 0.061$. The GLDPC code achieves 0.66dB gain in threshold. Similarly, thresholds 0.027(4.46dB), 0.0142(4.9dB), 0.0113(5.07dB), 0.0077(5.27dB) are obtained for (4,12), (4,18), (4,21) and (5,40) LDPC codes, respectively. These codes compare to the bottom four codes in Table I, respectively. Although irregular codes with better thresholds can probably be constructed it is unlikely they can significantly outperform the best regular codes given the high rate. Therefore, GLDPC codes achieve considerable gain in threshold compared to the best regular LDPC codes for the BSC.

B. Infinite Length Analysis on a BEC

We also considered the same decoding algorithm performed on a tree like code graph with $\epsilon$ error correcting RS codes as component codes over a BEC with erasure probability $\epsilon$. Let $p_i^e$ be the average symbol erasure rate at iteration $i$. Then at iteration $i + 1$, a symbol is an erasure if decoding in the constraint node failed and the symbol was received as an erasure. Since, we consider a nonbinary (RS) code as component code over a BEC we use a Q-ary erasure channel (QEC) model. The corresponding QEC has an erasure probability $\epsilon_{\text{QEC}} = 1 - (1 - \epsilon)^m$, where $m$ is the order of GF used to construct the RS code. In other words, if at least one bit is erased from the vector representation of the code symbol, the symbol is erased; hence $p_0^e = \epsilon_{\text{QEC}}$. Both RS and BCH codes can correct up to $d_{\text{min}} - 1$ erasures, which gives the following equation:

$$p_{i+1}^e = p_0^e \left(1 - \sum_{j=0}^{d_{\text{min}}-2} \left[egin{array}{c} n-1 \\ j \end{array}\right] p_i^e (1 - p_i^e)^{n-1-j}\right) \quad (9)$$

This equation is easily simplified for BCH codes. The thresholds for some GLDPC codes over a BEC are given in Table II where $R$ represents the rate of a GLDPC code, $\epsilon^*$ represents the threshold of a GLDPC code, $\epsilon_{\text{opt}}$ represents the capacity of a BEC and $\frac{1}{1-R}$ represents the threshold for (2, L) LDPC codes with rate $R$ when used on a BEC, according to [25]. As opposed to the BSC case, a similar analysis reveals that GLDPC codes seriously underperform LDPC codes on a BEC. This difference can be explained by the difference between BDD and ML decoding for the component codes considered [17].
TABLE II

<table>
<thead>
<tr>
<th>Subcode</th>
<th>( R )</th>
<th>( e^* )</th>
<th>( \frac{1-R}{1+R} )</th>
<th>( \epsilon_{opt} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>RS(15,11,5)</td>
<td>0.467</td>
<td>0.1383</td>
<td>0.363</td>
<td>0.413</td>
</tr>
<tr>
<td>RS(31,23,9)</td>
<td>0.484</td>
<td>0.0968</td>
<td>0.348</td>
<td>0.373</td>
</tr>
<tr>
<td>RS(31,25,7)</td>
<td>0.613</td>
<td>0.0668</td>
<td>0.240</td>
<td>0.343</td>
</tr>
<tr>
<td>RS(31,27,5)</td>
<td>0.742</td>
<td>0.0482</td>
<td>0.148</td>
<td>0.317</td>
</tr>
<tr>
<td>RS(63,53,11)</td>
<td>0.683</td>
<td>0.0452</td>
<td>0.180</td>
<td>0.304</td>
</tr>
<tr>
<td>RS(63,55,9)</td>
<td>0.746</td>
<td>0.0365</td>
<td>0.145</td>
<td>0.293</td>
</tr>
<tr>
<td>RS(63,57,7)</td>
<td>0.809</td>
<td>0.0278</td>
<td>0.106</td>
<td>0.284</td>
</tr>
<tr>
<td>RS(63,59,5)</td>
<td>0.873</td>
<td>0.0188</td>
<td>0.068</td>
<td>0.275</td>
</tr>
<tr>
<td>RS(127,121,7)</td>
<td>0.906</td>
<td>0.0133</td>
<td>0.049</td>
<td>0.253</td>
</tr>
<tr>
<td>RS(127,123,5)</td>
<td>0.937</td>
<td>0.0091</td>
<td>0.032</td>
<td>0.249</td>
</tr>
<tr>
<td>RS(255,249,7)</td>
<td>0.953</td>
<td>0.0060</td>
<td>0.024</td>
<td>0.233</td>
</tr>
<tr>
<td>RS(255,251,5)</td>
<td>0.968</td>
<td>0.0040</td>
<td>0.016</td>
<td>0.231</td>
</tr>
<tr>
<td>BCH(31,21,5)</td>
<td>0.355</td>
<td>0.2169</td>
<td>0.476</td>
<td>0.645</td>
</tr>
<tr>
<td>BCH(63,45,7)</td>
<td>0.428</td>
<td>0.1556</td>
<td>0.400</td>
<td>0.572</td>
</tr>
<tr>
<td>BCH(63,51,5)</td>
<td>0.619</td>
<td>0.1079</td>
<td>0.235</td>
<td>0.381</td>
</tr>
<tr>
<td>BCH(127,106,7)</td>
<td>0.669</td>
<td>0.0775</td>
<td>0.198</td>
<td>0.331</td>
</tr>
<tr>
<td>BCH(127,113,5)</td>
<td>0.779</td>
<td>0.0580</td>
<td>0.124</td>
<td>0.221</td>
</tr>
<tr>
<td>BCH(255,231,7)</td>
<td>0.812</td>
<td>0.0385</td>
<td>0.104</td>
<td>0.188</td>
</tr>
<tr>
<td>BCH(255,239,5)</td>
<td>0.875</td>
<td>0.0266</td>
<td>0.067</td>
<td>0.125</td>
</tr>
</tbody>
</table>

IV. FUTURE LENGTH ANALYSIS OF GLDPC CODES OVER BEC

LDPC codes on a BEC have been studied in [18] where the authors identified a special type of objects within a codegraph they named “stopping sets”. They proved that stopping sets are the only type of cycles that cause decoder failure on the BEC. This enabled the authors of [18] to develop a combinatorial approach to determine the exact average performance of the LDPC codes over their ensemble. The main result of [18] was stated in the form of the following set of recursive equations:

**Theorem 1:** [18]

\[
\begin{align*}
T(v, c, d) &= \left( \frac{d + cr}{vl} \right)_{vl}, \\
N(v, c, d) &= T(v, c, d) - M(v, c, d), \\
M(v, c, d) &= \sum_{s \geq 0} \binom{v}{s} O(v, s, c, d), \\
O(v, s, c, d) &= \sum_k \frac{\epsilon}{k} (1 + x)^r - 1 - rz)^k (1 + x)^s, x^d) \\
&\quad \text{subject to } s \leq 0 \text{ or } vl > cr + d
\end{align*}
\]

According to the notations introduced in [18] \( T(v, c, d) \) represents the number of constellations that can be constructed using \( v \) symbol nodes, \( c \) check nodes, and a super check node of degree \( d \). \( N(v, c, d) \) is the number of stopping set free constellations, \( M(v, c, d) \) is the number of constellations containing at least one stopping set, \( O(v, s, c, d) \) is the number of constellations having a stopping set of size \( s \) as a unique and maximal stopping set and \( \epsilon(f(x), x^d) \) is the coefficient \( f_i \) of the power series \( f(x) = \sum_{i=0}^n f_i x^i \). Hence, the exact average bit error performance \( E_{C(n,x^l-1,x^r-1)}[P_b^T(G, \epsilon)] \) and block error performance \( E_{C(n,x^l-1,x^r-1)}[P_B^T(G, \epsilon)] \) for the ensemble \( C(n, x^l-1, x^r-1) \) are given by

\[
E_{C(n,x^l-1,x^r-1)}[P_b^T(G, \epsilon)] = \sum_{e} \frac{n}{e} \epsilon^e (\tau)^{n-e} \frac{s}{n} O(e, s, n x) T(e, n x),
\]

\[
E_{C(n,x^l-1,x^r-1)}[P_B^T(G, \epsilon)] = \sum_{e} \frac{n}{e} \epsilon^e (\tau)^{n-e} \frac{1 - N_e (n x) T(e, n x)}{T(e, n x)},
\]

where \( \tau = 1 - \epsilon \), and if \( f(x) \) is a power series, by \( \epsilon(f(x), x^d) \) we denote its \( t \)-th coefficient \( f_t \). Consider a GLDPC code with \( t \) error correcting BCH or RS component codes used on a BEC and message passing decoding described in Section II. The decoder in the constraint node is capable of determining whether it can decode the received word correctly based on the number of erasures and error correcting capability \( t \) of the component codes and sends the indicator message \( C \) accordingly. Therefore, there is an analogy between the indicator message \( C \) passed from a constraint node of a GLDPC code and the message passed from check node of a LDPC code used over BEC. Both are different than 0 only if the corresponding node can deliver a correct value to the symbol node. We intend to exploit the analogy of GLDPC codes to LDPC codes when used over BEC and adopt the combinatorial approach developed in [18]. Hence, we introduce a generalized stopping set as a structure analogous to the stopping set structure in a LDPC.

**Definition 1 (Generalized stopping set):** A generalized stopping set of order \( m \) is a set of variable nodes, such that all neighbors of \( S^m \) are connected to \( S^m \) at most \( m \) times.

All properties of stopping sets (i.e. \( S^2 \)) derived in [18] are readily extended to generalized stopping sets. It is also noteworthy that \( S^2 \supset S^3 \supset S^4 \ldots \supset S^m \). Although this contradicts with the name generalized stopping sets because \( S^2 \) is the superset of all stopping sets and therefore the most "general", the name generalized stopping set came from generalized LDPC codes where this type of object was first identified. The following theorem puts generalized stopping sets in the focal point of our theoretical analysis and enables us to extend the combinatorial approach of [18] to generalized stopping sets. This provides us with the exact average performance of the ensemble of GLDPC codes on the BEC with algebraic decoding in constraints.

**Theorem 2:** Consider a GLDPC code \( G \in C(n, x^l-1, x^r-1) \) which has a linear block code with error correcting capability \( t \) as subcode. Assume that \( G \) is transmitted over a BEC and suppose iterative message passing decoding with bounded distance algebraic decoding in constraint nodes. Let \( E \) denote the set of variable nodes received as erasures. The set of variable nodes which remains incorrect after decoding is equal to the unique maximal generalized stopping set \( S^{2t+1} \) of \( E \).

The Theorem 2 is analogous to the Lemma 1.1 in [18] and the proof can be carried out in similar fashion. It gives a foundation to carry out the theoretical analysis and obtains
the exact average performance of the ensemble of GLDPC codes. A careful consideration of Theorem 1 verifies that all definitions are readily extended to the generalized stopping set $S_m$. The only difference is the way we count the number of $O(v, s, c, d)$ sets, because of the different number of ways to choose check node sockets to be connected to $sl$ variable node sockets. Therefore, the combinatorial approach and recursive equations derived in [18] can be generalized to accommodate generalized stopping set order $m$ by changing the coefficient term $\binom{(1 + x)^r - 1 - xx}{k} \binom{(1 + x)^d + x^n}{s}$ in the fourth equation with the generalized coefficient term $\binom{(r, k, s, d, m)}{C(r, k, s, d, m)}$ that is also a function of the generalized stopping set order $m$. The coefficient can be obtained as follows. We label a check-node (constraint node) socket $x$ if it is connected to generalized stopping set and 1 if it is not connected. Consider a check node of degree $r$ in the ensemble $C(n, x^{d-1}, x^{r-1})$. The number of ways in choosing $k$ of its sockets is $\binom{(1 + x)^r - 1 - xx}{k} \binom{(1 + x)^d + x^n}{s}$. The number of ways in choosing $c$ check node sockets from all sockets of $C$ is $\binom{(1 + x)^r - 1 - xx}{c} \binom{(1 + x)^d + x^n}{s}$. Since check node is connected to generalized stopping set $S_m$ if and only if at least $m$ of its sockets are connected to it, we should subtract those instances from possible socket selections. Therefore, the number of generalized stopping set constellations of order $m$ is $\binom{(1 + x)^r - \sum_{i=0}^{m-1} (\binom{x}{i} \binom{x}{d})}{s}$. Hence, the generalized coefficient term $\binom{(r, k, s, d, m)}{C(r, k, s, d, m)}$ is given with the following equation

$$\binom{(r, k, s, d, m)}{C(r, k, s, d, m)} = \binom{(1 + x)^r - \sum_{i=0}^{m-1} (\binom{x}{i} \binom{x}{d})}{s} \binom{(1 + x)^d}{s}$$

The set of recursive equations proposed in [18] was proved to be extremely complex to compute as the block length increases. Hence, the authors proposed a simplified set of recursions for the efficient computation. The efficient recursions were proposed only for $l = 2$ and $l = 3$ but the result was generalized for large left degrees in [19]. They proposed to build a set of stopping set free constellations adding a variable node at the time. We used the same concept to build constellations free from generalized stopping set of order $m = 3$. We also suppose that $d = 0$ and there is no super check node, as in [18], since general case is a straightforward extension. Since the graph structures of interest are different method of pruning the constellations is different. Fix an element from $\mathcal{A}(v, u, s_1, s_2)$ and for each variable node call it $\nu, \nu \in [v]$ let $m_1 = m_1(\nu)$ and $m_2 = m_2(\nu)$ represent the number of neighboring check node sockets of degree one or two, respectively. We will call $m_1$ and $m_2$ multiplicities of $\nu$ and let denote the neighbors by $c_1, \ldots, c_{m_1}$ and $C_1, \ldots, C_{m_2}$. To prune an element of $\mathcal{A}(v, u, s_1, s_2)$, pick a variable node $\nu$ of multiplicity at least one and delete all branches emanating from node $v$. If a neighboring check node is left unconnected, it is also deleted. This is different when compared to the [18] where neighboring check nodes are always deleted, because there exist only multiplicity $m_1$. The pruned element belongs to the new constellation $\mathcal{A}(v - 1, u - m_1, s_1, s_2)$ and labelling is done in the same way as in [18]. The above procedure can be inverted, i.e., if we start with this pruned constellation and add a variable node $v$ as well as $m_1$ check nodes with labels $c_1, \ldots, c_{m_1}$ we can recover the original constellation. If $\mathcal{A}^{c_1, \ldots, c_{m_1}}; C_1, \ldots, C_{m_2}(v, u, s_1, s_2)$ denotes the subset of $\mathcal{A}(v, u, s_1, s_2)$ which contains the variable node $\nu$ of multiplicity $m_1$ and $m_2$ each element in $\mathcal{A}^{c_1, \ldots, c_{m_1}}; C_1, \ldots, C_{m_2}(v, u, s_1, s_2)$ can be reconstructed in a unique way from an element of $\bigcup_{s_1, s_2} \mathcal{A}(v - 1, u - m_1, s_1, s_2)$ by adding variable node $\nu$, check nodes $c_1, \ldots, c_{m_1}$ and $m_2$ branches. It follows that a given element of $\mathcal{A}(v, u, s_1, s_2)$ can be reconstructed in exactly as many ways as the number of its variable nodes which have multiplicity at least one. Note that by definition, the sum of the multiplicities of all variable nodes is equal to $s_1 + 2s_2$. Therefore, the above statement can be rephrased to a similar one as in [18] as follows: If we weight each reconstruction by the multiplicity of the inserted variable node, then its weighted sum of reconstructions equals $s_1 + 2s_2$. Now, we can develop the recursions:

$$\alpha(v, u, s_1, s_2, d)(s_1 + 2s_2) = \alpha(v - 1, u - 1, s_1, 1, s_2, d)vur + \alpha(v - 1, u - 1, s_1, 2, s_2 + 1, d)vur(s_2 + 1)(r - 2) + \sum_{s_1 = 0}^{s_1 + 2s_2} \alpha(v - 1, u - 1, s_1, s_2, d)vur(r - 1)(r - 1) \cdot (ur - s_1 r - s_2 r - (v - 1)|s_1 + 1 + 2s_2 - 1)$$

If $v \geq 1$, $u \geq 0$, $s_1 \geq 0$, $s_2 \geq 0$, $s_1 + s_2 \geq 1$, $u \geq s_1 + s_2$ with the boundary condition $\alpha(v, u, s_1, s_2, d) = \begin{cases} 1 & \text{if } u = 0, s_1 = 0, s_2 = 0, d \geq 0 \\ 0 & \text{otherwise} \end{cases}$

Then we can express $N(v, c, d)$ as

$$N(v, c, d) = (l!)^v \sum_{u=0}^{c} \sum_{s_1=0}^{u} \sum_{s_2=0}^{u-s_1} \binom{c}{u} \alpha(v, u, s_1, s_2, d)$$

and thus calculate (16) to obtain the average block error rate. In order to calculate the average bit error rate, according to (13), we need to compute $\alpha(r, k, s, d, m)$ as explained earlier. Therefore, the set of recursive equations (10)-(16) is extended to generalized stopping sets of arbitrary order $m$. However, efficient evaluation is not readily generalized. The guaranteed erasure correcting capability of the coding scheme using a $t$ error correcting subcode is related to generalized stopping sets in the following theorem.

**Theorem 3 (Erasure correcting capability of GLDPC codes):** The erasure correcting capability of the GLDPC code with $t$ error correcting subcodes when used over a BEC with algebraic decoding in the constraints is $T - 1$, where $T$ is the cardinality of the smallest generalized stopping set $S^{2t+1}$.

This theorem follows directly from Theorem 2, but does not give a relation between the error correcting capability of the subcode $t$, and the error correcting capability of the GLDPC code. The following theorem addresses that.

**Theorem 4:** For the code graphs with variable node degree $J = 2$ and no 4-cycles let $S^{2t+1}$ be a generalized stopping set,
TABLE III
GUARANTEED ERROR AND ERASURE CORRECTING CAPABILITIES OF GLDPC CODE

<table>
<thead>
<tr>
<th>$t_{\text{subcode}}$</th>
<th>$T_{\text{error}}$</th>
<th>$T_{\text{error}}$</th>
<th>$T_{\text{error}}$</th>
<th>$T_{\text{error}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\text{GLDPC}$</td>
<td>$\text{product}$</td>
<td>$\text{GLDPC}$</td>
<td>$\text{product}$</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>8</td>
<td>14</td>
<td>24</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>15</td>
<td>27</td>
<td>48</td>
</tr>
<tr>
<td>4</td>
<td>14</td>
<td>24</td>
<td>44</td>
<td>80</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>35</td>
<td>65</td>
<td>120</td>
</tr>
<tr>
<td>6</td>
<td>27</td>
<td>48</td>
<td>90</td>
<td>168</td>
</tr>
<tr>
<td>7</td>
<td>35</td>
<td>63</td>
<td>120</td>
<td>224</td>
</tr>
</tbody>
</table>

$|S^{2t+1}|$ its cardinality, and let $C$ be the number of check nodes neighboring $S^{2t+1}$. Then, the following inequalities hold for valid code constellations:

$$C \leq \frac{2|S^{2t+1}|}{2t+1} \quad \text{and} \quad \binom{C}{2} \geq |S^{2t+1}|$$

Proof: There are $2|S^{2t+1}|$ branches emanating from $|S^{2t+1}|$ variable nodes which terminate in $C$ check nodes. Since $S^{2t+1}$ is a generalized stopping set, check nodes are connected to it at least $2t+1$ times. Hence, $C \leq \frac{2|S^{2t+1}|}{2t+1}$ and because $C$ is an integer, $C \leq \left\lfloor \frac{2|S^{2t+1}|}{2t+1} \right\rfloor$. As we consider code graphs with no 4-cycles, the generalized stopping sets in such code graph are also free from 4-cycles. Therefore, no two variable nodes are connected to the same pair of check nodes. There are $\binom{C}{2}$ ways to choose pairs of check nodes to be assigned to variable nodes. Hence, $\binom{C}{2} \geq |S^{2t+1}|$. Using these two inequalities we can find the guaranteed erasure correcting capabilities of the GLDPC codes, which are presented in Table III.

A. Asymptotic Behavior of Generalized Stopping Sets

In [18] stopping sets were recognized as objects that determine the performance of LDPC codes when used on a BEC. The authors also showed that performance of LDPC codes on a BEC depends on the stopping set distribution in the same way as the performance of a linear code on a BSC depends on codes’ weight distribution. Stopping set distribution and the asymptotic behavior of stopping sets was studied in [20]. We further developed some of the results to accommodate GLDPC codes and generalized stopping sets. The reader is referred to [20] for complete proofs of lemmas and theorems that will be used here without proof. In [20] the authors considered Tanner-graph ensembles of LDPC codes and defined the following parameters. The stopping number $s^*$ is the size of the smallest nonempty stopping set. The stopping ratio $\sigma^*$ of the graph is defined as $\frac{s^*}{n}$. The average stopping set distribution is defined as $E(s) = E(\{|S| = s\})$ and normalized average stopping set distribution as $\gamma(\alpha) = \lim_{n \to \infty} \frac{1}{n} \ln E(\{|S| = \alpha n\})$. Finally critical exponent stopping ratio is defined as $\alpha^* = \inf\{\alpha > 0 : \gamma(\alpha) \geq 0\}$. These definitions are extended to generalized stopping sets by replacing each reference to the stopping set by a reference to the a generalized stopping set of order $m$.

1) Normalized Average Distribution of Generalized Stopping Sets: For the ensemble of GLDPC codes probability that a randomly chosen set of symbol nodes $U$ with $e$ edges incident upon it, is a generalized stopping set of order $m$, $S^m$, is

$$Pr(U \in S^m) = \frac{\operatorname{coef}(1 + x)^r - \sum_{i=1}^{m-1} \binom{r}{i} x^i}{e^r}$$

and

$$E^m(s) = E(\{|S| = s\}) = \sum_{U \subseteq V, |U| = s} Pr(U \in S^m).$$

Therefore,

$$E^m(s) = \frac{\operatorname{coef}(1 + x)^r - \sum_{i=1}^{m-1} \binom{r}{i} x^i}{e^r}, \quad (16)$$

To evaluate the normalized average distribution of stopping sets of order $m$, we need the asymptotic coefficients of the polynomials in the equations. We used the same argument as in [20] to obtain the upper bound on the value of the coefficient $((1 + x)^r - \sum_{i=0}^{m-1} \binom{r}{i} x^i)$, Lemma 2. The following lemma is analogous to the Lemma 1 in [20] and proof can be obtained in a similar way.

Lemma 1: Let $c, 2 \leq r \in \mathbb{N}$ and $0 < \alpha < 1$. Then for every strictly increasing sequence $n_k$, $k \in \mathbb{N}$ satisfying the integer constraints $\frac{1}{n_k}, \alpha n_k l \in \mathbb{N}$, the limit

$$\lim_{k \to \infty} \frac{1}{n_k} \ln \frac{1}{r} \ln \left(1 + (1 - \alpha)x^{r - \sum_{i=1}^{m-1} \binom{r}{i} x^i} \right) = (1 - 1)h(\alpha),$$

where $x_0$ is the only positive solution of $x_0^{(1+x)^{r-\sum_{i=1}^{m-1} \binom{r}{i} x^i}} = \alpha$.

Now, we can derive normalized average distribution of generalized stopping set of order $m$, $\gamma^m(\alpha)$. In the following $h(x) = -x \ln x - (1 - x) \ln(1 - x)$ is an entropy function.

Theorem 5: For the regular Tanner-graph ensemble $C(n, \ell^{1-1}, r^{x-1})$,

$$\gamma^m(\alpha) = \frac{1}{r} \ln \left(\frac{1 + \alpha x_0^{r - \sum_{i=1}^{m-1} \binom{r}{i} x^i}}{x_0^{(1+r-x_0^{r-\sum_{i=1}^{m-1} \binom{r}{i} x^i})}}\right) = (l - 1)h(\alpha),$$

where $x_0 = \frac{e^\lambda}{x_0^{(1+x)^{r-\sum_{i=1}^{m-1} \binom{r}{i} x^i}}}$ is a unique positive solution for $x_0^{(1+x)^{r-\sum_{i=1}^{m-1} \binom{r}{i} x^i}} = \alpha$.

The proof follows from (16) and Lemma 1.

2) Generalized Stopping Number: Analogous to the definition of stopping number in [20], we define a generalized stopping number of order $m$, $s^m$ as the size of the smallest nonempty generalized stopping set $S^m$. From the definitions of critical exponent stopping ratio $\alpha^m$ and normalized average distribution of generalized stopping sets of order $m$, $\gamma^m(\alpha)$ it is clear that for $0 < \alpha < \alpha^m$, probability that there exist a generalized stopping set $S^m$ of size $\alpha n$ in a graph selected randomly from the ensemble decreases exponentially with $n$. However, this is not sufficient to show that with high probability there does not exist a generalized stopping set of size $\alpha^m n$ for $\alpha < \alpha^m$ as $\gamma^m(\alpha)$ does not address generalized stopping sets of size $\alpha^m$. Similar to [20] we prove that for
all \( \alpha < \alpha^*_m \) and regular GLDPC codes with \( l > \frac{m}{m-1} \) and \( m > 2 \) the probability of generalized stopping sets \( S^m \) with size \( \alpha n \) goes down polynomially in \( n \). As in [20] we use the following Markov inequality

\[
Pr(s^*_m \leq \alpha n) = Pr(\{ |S| \leq \alpha n \}) \leq \frac{1}{\alpha} E(\{|S| \leq \alpha n\}) \leq \frac{1}{\alpha} E(\{|S^m| \leq \alpha n\}) \tag{18}
\]

**Proof:** For the Tanner-graph ensemble of GLDPC codes of interest and \( m > 2 \) it is readily verified that \( s^*_m \geq \lceil \frac{m}{T} \rceil \).

The proof that \( Pr(s^*_m \leq \alpha n) = \Omega \left( n^{-\lceil \frac{m}{T} \rceil (l-1)} \right) \) follows the one in [20]. Next we prove that \( Pr(s^*_m \leq \alpha n) = \Omega \left( n^{-\lceil \frac{m}{T} \rceil (l-1)} \right) \). Since \( s^*_m \geq \lceil \frac{m}{T} \rceil \), for \( m > 2 \) it follows that \( Pr(s^*_m \leq \alpha n) \geq Pr(s^*_m = \lceil \frac{m}{T} \rceil ) \)

\[
\geq \sum_{1 \leq i_1 \leq i_2 \leq \ldots \leq i_{\lceil \frac{m}{T} \rceil}} Pr(\{v_{i_1}, v_{i_2}, \ldots, v_{i_{\lceil \frac{m}{T} \rceil}} \} \in S^m) \tag{20}
\]

Since the union of generalized stopping sets is a generalized stopping set,

\[
Pr(\{v_{i_1}, v_{i_2}, \ldots, v_{i_{\lceil \frac{m}{T} \rceil}} \} \in S^m) \leq Pr(\{v_{i_1}, v_{i_2}, \ldots, v_{i_{\lceil \frac{m}{T} \rceil}} \} \in S^m) \tag{21}
\]

and therefore,

\[
Pr(s^*_m \leq \alpha n) \geq Pr(s^*_m = \lceil \frac{m}{T} \rceil ) \geq \sum_{1 \leq i_1 \leq i_2 \leq \ldots \leq i_{\lceil \frac{m}{T} \rceil}} Pr(\{v_{i_1}, v_{i_2}, \ldots, v_{i_{\lceil \frac{m}{T} \rceil}} \} \in S^m) \tag{21}
\]

Combining this with (25) in the appendix, we obtain:

\[
Pr(s^*_m \leq \alpha n) = \Omega \left( n^{-\lceil \frac{m}{T} \rceil (l-1)} \right) \tag{22}
\]

The generalization of the result is stated in the following theorem which is equivalent to the Theorem 5.1 in [20].

**Theorem 7:** For the ensemble of GLDPC codes using a \( t \geq 1 \) error correcting component codes defined over the Tanner-graph ensemble \( C(n, x^{l-1}, x^{r-1}) \) if \( \epsilon < \epsilon_{ef} \) following is true

\[
\begin{aligned}
\mathbf{E}(P_B^{IT}(C, \epsilon)) &\leq \frac{n^{-\epsilon}}{\epsilon} \Omega \left( n^{-\lceil \frac{m}{T} \rceil (l-1)} \right) \\
\mathbf{E}(P_B^{IT}(C, \epsilon)) &\leq \frac{n^{-\epsilon}}{\epsilon} \Omega \left( n^{-\lceil \frac{m}{T} \rceil (l-1)} \right) \tag{23}
\end{aligned}
\]

The proof follows the steps of [20] follows. The advantage of GLDPC codes over LDPC codes when used over a BEC is that the number of objects that determines the performance of the code ensemble is considerably reduced, from \( S^2 \) to \( S^{2t+1} \). This is expected because generalized stopping sets \( S^{2t+1} \) are subsets of stopping sets \( S^2 \). The downside of GLDPC codes is that the code rate is small compared to that of LDPC codes constructed over the same Tanner-graph. The question is which dominates performance? The GLDPC codes with rates of interest are those constructed over the Tanner-graph ensemble \( C(n, x^{1}, x^{r-1}) \). However, their asymptotic performance given in Theorem 7 is poor compared to the performance of LDPC codes with the same rate. This is illustrated in Table II, where \( \epsilon^* \) represents threshold for GLDPC codes and \( \frac{1}{1+R} \) represents the threshold for \( (2, L) \) LDPC codes with rate \( R \), according to [25]. It is readily verified that thresholds for GLDPC codes are far from the capacity of the channel and that rate dominates performance. Since the rate of GLDPC code constructed over \( J = 3 \) Tanner-graphs is very small those ensembles are not considered. Therefore GLDPC codes are not competitors of LDPC codes on BEC.

In the following section we show how the developed theory can be successfully used to analyze ensembles of GLDPC codes when used on BSC. In this case, fortunately, the codes obtained can approach capacity.

**V. GLDPC OVER BSC WITH ALGEBRAIC BOUNDED DISTANCE DECODING IN CONSTRAINT NODES**

The theoretical analysis of the performance of the proposed coding scheme was enabled by recognizing codelograph structures that are responsible for decoding failures when the code is used on BEC. In [21], GLDPC codes were considered in a broader sense as verification of the constraint node is generic. The authors made the observation that when RS codes are considered as component codes in GLDPC over a BSC, a BSC behaves as an equivalent BEC channel. The analysis in this section is aimed to further elaborate this observation and relate it to the theoretical analysis of Section IV. Therefore, we consider the performance of GLDPC codes when used on a BSC with bounded distance algebraic decoding in constraints such that only decoding failures are allowed. Although this assumption may seem unrealistic, it serves as a good model to introduce the theory developed in Section IV to BSC.
The following reasoning shows that the performance of this model is a lower bound on the performance of the real system where the decoder makes both failures and errors. Note, that LDPC codes and GLDPC codes having Hamming codes as components are excluded from the analysis. The LDPC codes are excluded as failure only mode cannot be defined when constraints are SPCs. The GLDPC codes with Hamming codes as components are omitted since Hamming codes are perfect codes and BDD makes no decoding failures. Assume that a proposed scheme is used over a BSC with crossover probability $p_0$ and that the all zero word has been transmitted. Then in failure only mode, the average probability of bit/symbol error after the first iteration of decoding is given by the following equation

$$P_F = p_0 \left( 1 - \sum_{j=0}^{t} \binom{n}{j} p_0^j (1-p_0)^{n-j} \right).$$

According to the decoding algorithm introduced in Section II, the bit in the next iteration is in error if decoding failed in the constrained node failed and the bit was received in error. Similarly, for the failure and error mode, the bit in the next iteration is in error if decoding failed and the bit was in error or the decoder made a decoding error and a particular bit is incorrect. This probability is given by

$$P_{FE} = p_0 \left( 1 - \sum_{j=0}^{t} \binom{n}{j} p_0^j (1-p_0)^{n-j} - P_E \right) + P_E^b,$$

where $P_E = \sum_{j=d_{min}}^{n} A_j \sum_{k=0}^{t} p_k^j$ and $P_E^b = \sum_{j=d_{min}}^{n} A_j \sum_{k=0}^{t} p_k^j$ are the decoder word error rate and the decoder exact bit error rate, respectively, and $P_{FE}$ is the probability that the received word is exactly at distance $k$ from weight $j$ codeword [23]. Since, $P_{FE} \geq P_{min} \sum_{j=d_{min}}^{n} A_j \sum_{k=0}^{t} p_k^j = \frac{d_{min}}{n} P_E$.

$$P_{FE} \geq p_0 \left( 1 - \sum_{j=0}^{t} \binom{n}{j} p_0^j (1-p_0)^{n-j} - P_E \right) + P_E \frac{d_{min}}{n}$$

By induction, this analysis is readily applied to an arbitrary iteration step $i$, so that the relation is preserved. Therefore, the exact analysis of the failure only mode provides a lower bound on the performance of the realistic decoder with both failures and errors. It is shown in Table I that the condition $\frac{d_{min}}{n} \geq p_0$ is satisfied for $p_0$ smaller than threshold.

### A. Constraint Node Decoding With Failures Only

The performance of the proposed scheme when used on a BSC such that only decoding failures are possible is determined by the result of the following theorem.

**Theorem 8:** Consider a GLDPC code $\mathcal{G} \in \mathcal{C}(n, x^{t-1}, x^{r-1})$ which has a linear block code with error correcting capability $t$ as subcode. Assume that $\mathcal{G}$ is transmitted over a BSC and suppose iterative message passing decoding with bounded distance algebraic decoding in constraint nodes. Assume also that the algebraic decoder does not make decoding errors, but only decoding failures. Let $E$ denote the set of variable nodes received in error. Then the set of variable nodes which remains in error after decoding is equal to the unique maximal generalized stopping set $S^{t+1}$ of $E$.

The analogy between Theorem 2 and Theorem 8 is straightforward. Since in the case of GLDPC codes over a BSC decoder is allowed only failures, this can be easily verified and the corresponding indicator message can be sent. This is analogous to verifying the number of erasures in the constraint and setting a corresponding indicator message accordingly. Since there is a complete equivalence between messages passed from constraint nodes to symbol nodes when constraint is not satisfied the proof of Theorem 8 follows the proof of Theorem 2. Therefore, the set of recursive equations (10)-(16) can be used to obtain the exact average performance of the ensemble of GLDPC codes when used over a BSC with failures only in the decoder. The afore mentioned analogy also allows us to extend the results of Section IV-A. Theorem 7 provides the asymptotic behavior of the coding scheme when stopping sets of order $m = t+1$ are considered.

As in [18] there are two reasons for the relatively poor performance of the code ensembles. First, the average performance of the ensemble is dominated by a few very bad codes as argued in [18]. Second, the model proposed in [18] itself introduces many such codes.

The guaranteed error correcting capability of the coding scheme using a $t$ error correcting subcode is related to generalized stopping sets in the theorem analogous to the Theorems 3 and 4.

**Theorem 9 (Error correcting capability of GLDPC codes):**

The error correcting capability of a GLDPC code with $t$ error correcting subcodes when used over a BSC with bounded distance algebraic decoding and failures only in the node constraints is $T-1$, where $T$ is the cardinality of the smallest generalized stopping set $S^{t+1}$.

**Proof:** It follows from Theorem 8 that no error pattern containing a generalized stopping set $S^{t+1}$ as subset can be corrected. If $T$ is the cardinality of the smallest generalized stopping set $S^{t+1}$, no error pattern of weight $T - 1$ can have a generalized stopping set $S^{t+1}$ as a subset and therefore can be corrected.

This theorem relates the error correcting capability of the GLDPC code with the generalized stopping set $S^{t+1}$, but does not provide a relation between the error correcting capability $t$ of a subcode, and the error correcting capability of the GLDPC code. The following theorem establishes such a relation.

**Theorem 10:** For code graphs with variable node degree $J = 2$ and no 4-cycles, let $S^{t+1}$ be a generalized stopping set, let $|S^{t+1}|$ be its cardinality, and let $C$ be the number of check nodes neighboring $S^{t+1}$. Then, the following inequalities hold for valid code constellations:

$$C \leq \left\lfloor \frac{2 |S^{t+1}|}{t+1} \right\rfloor$$

and

$$\left( \frac{C}{2} \right) \geq |S^{t+1}|$$

It is readily verified that the model allows multiple edges between symbol and check nodes. Therefore, the model includes many non valid bad codes in the code ensemble. On the other hand, allowing these codes seems the only tractable approach.
Using these two inequalities we can find the guaranteed error correcting capability of the GLDPC codes. These results are reported in Table III and compared to the error correcting capability of the corresponding product code.

B. Constraint Node Decoding With Both Failures and Errors

Decoding errors generally occur in algebraic decoding and the assumption of a decoder that makes failures only is unrealistic in general. However, our analysis (24) shows that the performance of the proposed scheme is lower bounded by the performance of the scheme assuming a failure only decoder. If the probability of decoding failure dominates the decoding performance for a given component code at SNR values of interest, then we can use the results of Theorem 1 and the asymptotic analysis as a very tight lower bound on the exact average performance of the ensemble. This is the case for RS and BCH codes with $t > 2$, especially as the block length $n$ and the error correcting capability $t$ get larger.

VI. SIMULATION RESULTS

The lower bounds on code ensemble obtained in Section V do not provide a practical insight in the performance of the proposed scheme. Therefore, performance of the proposed coding scheme was evaluated through simulation of the proposed codes over BSC. The GLDPC codes used in simulations were constructed using a randomly constructed LDPC code as a base, and corresponding RS or BCH code as a component code. We considered only component codes with error correcting capability $t \geq 2$, since our theoretical analysis suggests that codes with $t = 1$ would have a large number of small stopping sets and therefore poor performance.

Furthermore, GLDPC codes with a single error correcting code as subcode were already studied [7]–[10]. We considered only codes with $J = 2$, as we were interested in codes with rate close to or larger than 0.5. Since, GLDPC codes can be viewed as an extension of the product codes all figures include a product code and its corresponding GLDPC counterpart. First we considered GLDPC code with RS(15,11,5) code as subcode and rate $r = 0.467$. The RS(15,11,5) is the shortest RS component code that enables construction of GLDPC codes under proposed constraints. Simulation results are presented in Fig. 3 for various lengths. It is interesting to note that RS(15,11,5) product code outperforms its GLDPC counterpart, mainly due to the better rate. As block length of the GLDPC increased performance steadily improved especially in the error floor region.

Next, we considered GLDPC codes with RS(31,27,5) codes as component and rate $r = 0.742$, depicted in Fig. 4. Note that GLDPC (961,713) now outperforms the corresponding product code. The product code no longer has the advantage of higher rate and due to its regular structure suffers from large number of small generalized stopping sets that dominate performance. Similarly, performance improves as block length increases. Performance of the GLDPC code with RS(31,27,5) and rate $r = 0.613$ as subcode is depicted in Fig. 1. The performance of the codes with RS(63,59,5) subcodes has a similar behavior in comparison to the one depicted in Fig. 2. However, signs of the error floor were observed for the RS(63,59,5) product code. The RS(63,55,9) product code and its GLDPC counterpart showed almost identical performance.
Finally, figure Fig. 5 presents an interesting comparison of a RS(255,239)x(255,223) product code and a GLDPC code of comparable rate. The figure shows that GLDPC code are strong competitors to well established coding schemes under iterative decoding as bit error performance at $10^{-6}$ is only 1.67dB away from Shannon limit. The simulation results presented in Fig. 3 - Fig. 2 shows that a proposed coding scheme offers great flexibility in meeting the code rate and required performance. There are three parameters that define a GLDPC code with RS subcodes, length $N$, order of Galois field $m$ or equivalently length of the RS subcode $n$, and error correcting capability $t$. Although, $n$ and $t$ cannot be chosen completely independent of one another, there exist a great deal of possibilities in choosing parameters that define a code of a certain rate.

These three degrees of freedom in construction of the code enables various decoding complexity-performance trade off possibilities. The performance of GLDPC codes with BCH codes as subcodes was simulated for BCH (31,21), BCH(63,51) and BCH(63,45) as subcodes. Results are depicted in Fig. 6 - 8 respectively. Codes in Fig. 8 and 3 and similarly Fig. 7 and 1 can be compared as they have similar rate. We observe that GLDPC codes with BCH codes as component codes outperform its counterpart with RS codes as components by approximately 0.5 dB. Nonetheless, both of the proposed schemes achieve good performance compared to other hard decision decoding schemes. Moreover, the two approaches complement each other with respect to code rate of the constructed GLDPC codes, as GLDPC codes with RS components offer a wider range of possible code rates. Although there exist hard decision decoding algorithms that can achieve similar BER performance [24], proposed schemes offer much better WER performance at a small increase in decoding complexity.

VII. CONCLUSION

In this paper we investigated GLDPC codes on a BEC and identified code graph structures that cause decoding failures. These structures were recognized as subsets of the stopping sets introduced by [18] and were defined more precisely so that they can be related to the erasure correcting capability of
possible check set/socket sequence pairs that have the desired property. Since each selected check node contributes at least \( m \) and at most \( r \) sockets, the size of the check set is at least \( \left\lceil \frac{n}{m} \right\rceil \) and at most \( \left\lfloor \frac{n}{m} \right\rfloor \). The number of ways to choose such a check set is

\[
|k: \exists 1 \leq i \leq j, c_i = k| = \sum_{i=\left\lfloor \frac{n}{m} \right\rfloor}^{\left\lceil \frac{n}{m} \right\rceil - \left\lfloor \frac{n}{m} \right\rfloor} \left( \begin{array}{c} n_i \vspace{0.2cm} \end{array} \right) \left( \frac{1}{m} \right)^i \left( \frac{m-1}{m} \right)^{n_i-i}.
\]

The socket sequence \( \{ s_{i+1} - s_i + r\delta_{c_i+1,c_i} : 0 \leq i \leq j-1 \} \), describes the distance between the selected sockets \( Sock_{c_i} \) and \( Sock_{c_{i+1}} \). If \( c_{i+1} = c_i \), the distance is given by \( s_{i+1} - s_i \leq r - 1 \). If \( c_{i+1} \neq c_i \), the distance is given by \( s_{i+1} - s_i + r \leq 2r - 2m + 1 \). Therefore we can upper bound the number of such socket sequences by \( (2r - 2m + 1)^j \).

ACKNOWLEDGMENT
The authors wish to thank Changyan Di for useful comments and insight into the combinatorial approach developed in [18]. They are also grateful to Toru Fujiiwara for providing the weight distribution of the BCH codes of Table I, and to Junan Zhang and Alon Orlitsky for the preprint of [20].

**APPENDIX**

**Lemma 2:** For every \( r, j, m \in \mathbb{N} \),

\[
\text{coef} \left( \left(1 + x \right)^r - \sum_{i=0}^{m-1} \binom{r}{i} x^i \right)^{n_i/2} \leq \left( \begin{array}{c} n_i \vspace{0.2cm} \\
\frac{1}{2} \end{array} \right) \left( \frac{1}{m} \right)^i \left( \frac{m-1}{m} \right)^{n_i-i}. \]

**Proof:** Proof of the lemma is similar to the proof of the corresponding lemma (Lemma 6.1) in [20]. We use the same ordering of check nodes and check-node sockets to obtain a bound on the coefficient. For detailed description readers are referred to [20]. As in Lemma 6.1 any choice of \( j \) sockets out of \( nr \) can be represented with a pair: check set given by \( \{ s_{i+1} - s_i + r\delta_{c_i+1,c_i} : 0 \leq i \leq j-1 \} \), where \( \delta_{c_i+1,c_i} \) is one if \( c_{i+1} \neq c_i \) and zero otherwise. The expression \( \text{coef}((1 + x)^r - \sum_{i=0}^{m-1} \binom{r}{i} x^i)^{n_i/2} \) gives the number of ways of selecting \( j \) check node sockets from \( n_i/2 \) sockets such that no check node is selected less than \( m \) times. This number can be bounded above by counting the number of


M. P. C. Fossorier received the B.E. degree from the National Institute of Applied Sciences (L.N.S.A.) Lyon, France in 1987, and the M.S. and Ph.D. degrees from the University of Hawai’i at Manoa, Honolulu, USA in 1991 and 1994, all in electrical engineering.

In 1996, he joined the Faculty of the University of Hawai’i, Honolulu, as an Assistant Professor of Electrical Engineering. He was promoted to Associate Professor in 1999 and to Professor in 2004. His research interests include decoding techniques for linear codes, communication algorithms and statistics.

Dr. Fossorier is a recipient of a 1998 NSF Career Development award and became an IEEE Fellow in 2006. He has served as Editor for IEEE TRANSACTIONS ON INFORMATION THEORY from 2003 to 2006, as Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS from 1996 to 2003, and as Treasurer of the IEEE Information Theory Society from 1999 to 2003. Since 2002, he has also been an elected member of the Board of Governors of the IEEE Information Theory Society, where he is currently serving as First Vice-President. He was Co-Chairman of the IEEE International Symposium of Information Theory (ISIT) held in Nice, France in June 2007, Program Co-Chairman for the 2000 International Symposium on Information Theory and Its Applications (ISITA), and Editor for the Proceedings of the 2006, 2003 and 1999 Symposium on Applied Algebra, Algebraic Algorithms and Error Correcting Codes (AAECC).

Nenad Miladinovic received the BSEE degree from the Faculty of Electrical Engineering, University of Belgrade, Serbia in 1999, and the Ph.D. degree from the University of Hawai’i at Manoa, Honolulu, USA in 2005, all in electrical engineering.

In 2005, he joined the LSI Corporation, Milpitas, CA, where he currently works a Senior Design Engineer.
Soft Handover Overhead Reduction by RAKE Reception with Finger Reassignment

Seyeong Choi, Member, IEEE, Mohamed-Slim Alouini, Senior Member, IEEE, Khalid A. Qaraqe, Senior Member, IEEE, and Hong-Chuan Yang, Senior Member, IEEE

Abstract—We propose and analyze in this paper a new finger assignment technique that is applicable for RAKE receivers when they operate in the soft handover (SHO) region. This scheme employs a new version of generalized selection combining (GSC). More specifically, in the SHO region, the receiver uses by default only the strongest paths from the serving base station (BS) and only when the combined signal-to-noise ratio (SNR) falls below a certain pre-determined threshold, the receiver uses more resolvable paths from the target BS to improve the performance. Hence, relying on some recent results on order statistics we attack the statistics of two correlated GSC stages and provide the approximate but accurate closed-form expressions for the statistics of the output SNR. By investigating the tradeoff among the error performance, the path estimation load, and the SHO overhead, we show through numerical examples that the new scheme offers commensurate performance in comparison with more complicated GSC-based diversity systems while requiring a smaller estimation load and SHO overhead.

Index Terms—Fading channels, diversity techniques, RAKE receiver, generalized selection combining (GSC), performance analysis.

I. INTRODUCTION

In wideband code division multiple access (WCDMA) systems and ultra wideband (UWB) systems, the diversity branches correspond to the different resolvable multi-paths and RAKE reception, with several baseband correlators called fingers, is used to combine these paths in order to increase the overall signal-to-noise ratio (SNR) and to lower the probability of deep fades [1, Section 9.5.1]. If there are \( j \) resolvable paths, the optimal number of fingers is \( j \), but due to receiver complexity and processing power constraints (especially for mobile units), we assume that \( i \) fingers are employed by the RAKE receiver. Usually, the mobile unit receiver is limited to 3 fingers while the base station (BS) receiver can use 4 or 5 fingers depending on the equipment manufacturer [2]. Note that in the handover (HO) region the number of available resolvable paths can be quite high since they can come from the serving BS as well as the target BS. Hence, it is natural to consider how to judiciously select a subset of paths for RAKE reception in the soft HO (SHO) region in order for the receivers to achieve the required performance while (i) maintaining a low complexity and low processing power consumption and (ii) using a minimal amount of additional network resources.

Many newly proposed low complexity combining approaches can be used for our problem of interest (i.e., combining in the SHO region) [3]–[12]. Among them is generalized selection combining (GSC) [3]–[6] which is a generalization of selection combining (SC) and which chooses a fixed number of paths with the largest instantaneous SNR from all available diversity paths and then combines them as per the rules of maximal ratio combining (MRC). As a power-saving implementation of GSC, minimum selection GSC (MS-GSC) [7]–[9], minimum estimation and combining GSC (MEC-GSC) [10], and output-threshold GSC (OT-GSC) [11], [12] were recently proposed. While these combining schemes can be applicable for our problem of interest, the way they operate does not make them distinguish the resolvable paths coming from the serving and the target BS. As such, if they are used without any modification or adaptation to the SHO, they end up using continuously the hardware/transmission resources of the serving and the target BS and result therefore in a considerable increase in overhead on the network (known as SHO overhead [13, Section 9.3.1.4]).

In this paper, we propose and study a new finger reassignment-based scheme that is specifically applicable for RAKE reception in the SHO region. With this scheme, we assume that the \( L_c \) out of total \( L \) resolvable paths from the serving BS are by default assigned to the RAKE fingers of the mobile unit in the SHO region following \( L_c/L\)-GSC type of combining. Only when the output SNR falls below a predetermined fixed SNR threshold (known also as a target SNR), the receiver asks for the additional resources from the target BS. More specifically, the receiver scans the additional \( L_a \) resolvable paths from the target BS and selects again the strongest \( L_c \) paths but now among the \( L + L_a \) available paths (i.e., the receiver uses \( L_c/(L+L_a)\)-GSC). Unlike MS-GSC and OT-GSC, our proposed scheme always uses a fixed number of fingers, i.e., \( L_c \), but as we will show in the performance results section, it can reduce the unnecessary path estimations and the SHO overhead compared to the conventional GSC.
The main contribution of this paper is to derive the statistics of the receiver output SNR for our newly proposed scheme, including its probability density function (PDF), cumulative distribution function (CDF), and moment generating function (MGF). We provide not only the analytical framework that leads to exact but complicated expressions but also an alternative approximate approach which yield relatively simple expressions that come close to the exact solutions. These results are then used (i) to analyze the performance in terms of the average probability of error and (ii) to investigate the tradeoff between complexity and performance. Some selected numerical results show that in poor channel conditions our scheme can essentially give the same performance as the GSC scheme while it offers in good channel conditions a smaller path estimation load and considerable reduction in the SHO overhead. To simplify our analysis and make it tractable, we assume that the receiver operates over a “perfect” uniform propagation delay profile provided by a multi-path searcher in a way that the multi-path components are correctly assigned to the RAKE fingers.

The remaining of this paper is organized as follows. In Section II, we present the system and channel model under consideration as well as the mode of operation of the proposed scheme. Based on this mode of operation, we derive the expressions for the statistics of the combined SNR in Section III. These results are next applied to the performance analysis of the proposed system in Section IV. This section also illustrates the tradeoff of complexity versus performance by comparing the number of path estimations and the SHO overhead of our proposed systems to that of conventional GSC and MRC. Finally, Section V provides some concluding remarks.

II. FINGER REASSIGNMENT-BASED RAKE COMBINING

A. System and Channel Model

We consider a multi-cell CDMA system with universal frequency reuse. Each cell uses different sets of spreading codes to control the intercell interference. We focus on the receiver operation when the mobile unit is moving from the coverage area of its serving BS to that of a target BS. We assume that the mobile unit is equipped with an $L_c$ finger RAKE receiver and is capable of despreading signals from different BSs using different fingers, and thus facilitating SHO. The RAKE receiver also implements a GSC-based path selection mechanism to select the $L_c$ best paths for RAKE combining among all the resolvable paths. Note that in the SHO region the mobile unit is of roughly the same long distance from the serving and the target BSs. We further assume that the average signal strength on a path from both BSs is the same. As such, we assume that the received signals on all the resolvable paths from the serving and the target BSs experience independent and identically distributed (i.i.d.) Rayleigh fading\(^1\). Let $\gamma_i$ denote the instantaneous received SNR of the $i$th resolved path. Then, $\gamma_i$ follows the same exponential distribution, with common PDF and CDF given as [1, Eq. (6.5)]

$$f_{\gamma_i}(x) = \frac{1}{\gamma} \exp \left( -\frac{x}{\gamma} \right), \quad x \geq 0$$

and

$$F_{\gamma_i}(x) = 1 - \exp \left( -\frac{x}{\gamma} \right), \quad x \geq 0,$$

respectively, where $\gamma$ is the common average faded SNR.

B. Mode of Operation

We assume without loss of generality that in the SHO region, the mobile unit resolves $L$ multi-paths from the serving BS and $L_a$ additional paths from the target BS. As the mobile unit enters the SHO region, the RAKE receiver relies at first on the $L$ resolvable paths gathered from the serving BS and as such starts with $L_c/L$-GSC. If we let $\Gamma_{i:j}$ be the sum of the $i$ largest SNRs among $j$ ones, i.e., $\Gamma_{i:j} = \sum_{k=1}^i \gamma_{k:j}$, then the total received SNR after GSC is given by $\Gamma_{L_c:L}$. At the beginning of every time slot, the receiver compares the received SNR, $\Gamma_{L_c:L}$, with a certain target SNR, denoted by $\gamma_T$. If $\Gamma_{L_c:L}$ is greater than or equal to $\gamma_T$, a one-way SHO\(^2\) is used and no finger reassignment is needed. On the other hand, whenever $\Gamma_{L_c:L}$ falls below $\gamma_T$, a two-way SHO\(^3\) is attempted. In this case, the RAKE reassigns its $L_c$ fingers to the $L_a$ strongest paths among the $L + L_a$ available resolvable paths (i.e., the RAKE receiver uses $L_c/(L + L_a)$-GSC). Now the total received SNR is given by $\Gamma_{L_c:L+L_a}$.

Based on the above mode of operation, we can see that the final combined SNR, denoted by $\gamma_t$, is mathematically given by

$$\gamma_t = \begin{cases} \Gamma_{L_c:L+L_a}, & 0 \leq \Gamma_{L_c:L} < \gamma_T; \\ \Gamma_{L_c:L}, & \Gamma_{L_c:L} \geq \gamma_T. \end{cases}$$

III. STATISTICS OF THE COMBINED SNR

Although the mode of operation in (3) describes a scheme that essentially switches between $L_c/L$-GSC and $L_c/(L + L_a)$-GSC depending on the channel conditions, we cannot obtain the statistics of $\gamma_t$ directly from the statistics of the output SNR with conventional GSC. Hence, in this section, we rely on some recent results on order statistics [8], [12] to derive the statistics of the combined SNR, $\gamma_t$.

A. CDF

From (3), the CDF of $\gamma_t$, $F_{\gamma_t}(x)$, can be written as

$$F_{\gamma_t}(x) = \Pr [\gamma_t < x]$$

$$= \Pr [\gamma_T \leq \Gamma_{L_c:L} < x] + \Pr [\Gamma_{L_c:L+L_a} < x, \Gamma_{L_c:L} < \gamma_T].$$

Since it is clear that $\Gamma_{L_c:L} \leq \Gamma_{L_c:L+L_a}$, we can rewrite

$$\Pr [\Gamma_{L_c:L+L_a} < x, \Gamma_{L_c:L} < \gamma_T]$$

in (4) as shown in (5). Substituting (5) into (4), we can express the CDF of $\gamma_t$, $F_{\gamma_t}(x)$, as shown in (6). To obtain a closed-form expression for

\(^{1}\)In [14], more practical channel environments, such as non-identical/correlated fading channels and outdated channel estimation, are considered.

\(^{2}\)One-way SHO refers to the scenario in which the mobile unit is connected only to the serving BS while being in the SHO region.

\(^{3}\)Two-way SHO refers to the scenario in which the mobile unit is connected to the serving and the target BSs while being in the SHO region.
\[
F_{\gamma_t}(x) = \begin{cases} 
\Pr[\Gamma_{L_c:L+L_a} < x], & 0 \leq x < \gamma_T; \\
\Pr[\Gamma_{L_c:L+L_a} < \gamma_T] + \Pr[\Gamma_{L_c:L+L_a} < x, \Gamma_{L_c:L} < \gamma_T], & x \geq \gamma_T
\end{cases}
\]

By recursively performing the following integration
\[
f_{\Gamma_{L_c:L+2}, \Gamma_{L_c:L+1}}(y_0, y_1)dy_0dy_1,
\]
we can express the conditional PDF in (7), for the general value of \(L_a\) (\(\geq 2\)), as shown in (9). Even though the joint PDFs and the conditional PDF in (9) are available in closed-form using some results that will be shown in what follows, the resulting expressions are complicated and quite tedious to obtain. Here, we rather use in what follows another approximate approach which leads to results that are very close to the exact solutions as we will demonstrate it by computer simulations in Section IV.

Going back to Eq. (7) and based on the derivation in the Appendix, we can show that \(\Pr[\gamma_T \leq \Gamma_{L_c:L+L_a} < x, \Gamma_{L_c:L} < \gamma_T]\) can be expressed as shown in (10). Substitution (10) into (6) gives the CDF of \(\gamma_t\), \(F_{\gamma_t}(x)\), as shown in (11), where
\[
\mathcal{J}(x) = \Pr[\gamma_T \leq \Gamma_{L_c:L+L_a} < x, \Gamma_{L_c:L+L_a-1} < \gamma_T].
\]

Although (11) looks more complicate than (6), it actually leads to the desired final result, as we show in what follows. Since for i.i.d. Rayleigh fading channels, all probabilities, \(\Pr[\cdot]\), in (11) can be easily be obtained by using the well-known CDF of the GSC output SNR [15, Eq. (9.440)], we just need to derive a closed-form expression for \(\mathcal{J}(x)\) in (12). This joint probability can be expressed as shown in (13).

Since all branch SNRs are i.i.d. random variables, \(\gamma_{L_c:L+a}\) is independent of both \(\Gamma_{L_c:L+L_a-1}\) and \(\gamma_{L_c:L+L_a-1}\). As such, we can compute the joint probability in (13) by using the joint PDF of \(\Gamma_{L_c:L+L_a-1}\) and \(\gamma_{L_c:L+L_a-1}\) denoted by
\[
f_{\Gamma_{L_c:L+L_a-1}, \gamma_{L_c:L+L_a-1}}(y, z),
\]
and the single-branch CDF of \(\gamma_{L_c:L+L_a-1}\), \(F_{\Gamma_{L_c:L+L_a-1}}(\cdot)\), given in (2), as shown in (14). For i.i.d. Rayleigh fading channels, it has been shown in [8, Eq. (9)] that the joint PDF in (14) is given by (15). After substitution (15) into (14) and integrations, we can obtain the closed-form expression for \(\mathcal{J}(x)\) as shown in (16), where \((a_1, a_2, \ldots, a_n)\) is the multinomial coefficient, defined as \(\binom{A}{a_1, a_2, \ldots, a_n} = \frac{A!}{a_1!a_2!\cdots a_n!}\). Hence, we can obtain the closed-form expression for the CDF of \(\gamma_t\) by substituting (16) in (11).

### B. PDF

Differentiation of (11) gives the PDF of \(\gamma_t\), \(f_{\gamma_t}(x)\), as shown in (17), which leads to (18). For i.i.d. Rayleigh fading channels, \(f_{\Gamma_{L_c:L}}(x)\) and \(F_{\Gamma_{L_c:L}}(x)\) are the well-known PDF and CDF of \(i/j\)-GSC output SNR which can be found in [15, Eqs. (9.433)(9.440)], respectively.

### C. MGF

Substituting (18) into (17) leads to the desired closed-form expression for the PDF of the proposed scheme. With this PDF in hand, the MGF of \(\gamma_t\), \(M_{\gamma_t}(s) = \int_0^\infty e^{sx}f_{\gamma_t}(x)dx\), can be obtained in closed-form after lengthy and tedious calculations as
\[
M_{\gamma_t}(s) = \frac{A(L_c : L + L_a, 0, s) + A(L_c : L, \gamma_T, s)}{1 - B(L_c : L, \gamma_T)} - \frac{1 - B(L_c : L + L_a - 1, \gamma_T)}{1 - B(L_c : L + L_a - 1, \gamma_T)} \times (A(L_c : L + L_a, \gamma_T, s) - C(\gamma_T, s)),
\]
which leads to (20), (21), and (22), and where \(\Gamma[\cdot, \cdot, \cdot] = \Gamma[\cdot, \cdot] - \Gamma[\cdot, \cdot, \cdot]\) are the upper and lower incomplete gamma functions, respectively, defined as [16, Eq. (8.350)]
\[
\Gamma(\alpha, \beta) = \int_0^\beta e^{-t}t^{\alpha-1}dt, \quad \gamma(\alpha, \beta) = \int_0^\beta e^{-t}t^{\alpha-1}dt.
\]

### IV. PERFORMANCE RESULTS

In this section, we apply the closed-form results from the previous section for the performance analysis of our proposed combining scheme over Rayleigh fading channels. More specifically, we first examine its average bit error rate (BER) by using the well-known MGF-based approach [15, Sec. 9.2.3]. We then look into the average number of path estimations and the SHO overhead it requires.

#### A. Average BER Comparison with MRC and GSC

First, we consider the relationship between the number of resolvable paths from the serving BS and the average BER performance. In Fig. 1, the average BER of binary phase shift keying (BPSK) versus the average SNR per path, \(\gamma\), of the proposed scheme for various values of \(L\) over i.i.d.
\[
\begin{align*}
\Pr \{ \gamma_T \leq \Gamma_{Le:L+La} < x, \Gamma_{Le:L} < \gamma_T \} &= \Pr \{ \gamma_T \leq \Gamma_{Le:L+La} < x \} - \frac{1 - \Pr[\Gamma_{Le:L} < \gamma_T]}{1 - \Pr[\Gamma_{Le:L+La} < x] - \Pr[\gamma_T \leq \Gamma_{Le:L+La} < x, \Gamma_{Le:L+La} < x, \Gamma_{Le:L+La} - 1 < \gamma_T]} \\
F_{\gamma}(x) &= \begin{cases} 
\Pr[\Gamma_{Le:L+La} < x], & 0 \leq x < \gamma_T; \\
\Pr[\gamma_T \leq \Gamma_{Le:L+La} < x] + \Pr[\Gamma_{Le:L+La} < \gamma_T] + \Pr[\gamma_T \leq \Gamma_{Le:L+La} < x] \times (\Pr[\gamma_T \leq \Gamma_{Le:L+La} < x] - J(x)), & x \geq \gamma_T
\end{cases}
\end{align*}
\]

\[
\begin{align*}
\Pr \{ \gamma_T \leq \Gamma_{Le:L+La} < x, \Gamma_{Le:L+La} - 1 < \gamma_T \} &= \Pr \{ \gamma_T \leq \Gamma_{Le:L+La} < x, \Gamma_{Le:L+La} < x, \Gamma_{Le:L+La} - 1 < \gamma_T \}
\end{align*}
\]

\[
\begin{align*}
\frac{\gamma_T}{\gamma_c} &\leq \frac{\gamma_T - y}{\gamma_c - y} \\
&= \left( e^{-\frac{\gamma_T}{\gamma_c}} - e^{-\frac{y}{\gamma_c}} \right) \left( \frac{L_c L_c + L_L - L_L - 1}{(L_c - u - 1)!} \right)^{L_c - 1} \\
&\sum_{u=0}^{L_c - 1} \left( L_c \sum_{u=0}^{L_c - 1} \frac{(-1)^{L_c + 1}}{(L_c - u - 1)! (t + 1)! (t + 1)! (\gamma_T / \gamma_c)^{u+1}} \right) \\
&= f_{\gamma_{Le:L+La}}(x), \\
J(x) &= \left( e^{-\frac{\gamma_T}{\gamma_c}} - e^{-\frac{y}{\gamma_c}} \right) \left( \frac{L_c L_c + L_L - L_L - 1}{(L_c - u - 1)!} \right)^{L_c - 1} \\
&\sum_{u=0}^{L_c - 1} \left( L_c \sum_{u=0}^{L_c - 1} \frac{(-1)^{L_c + 1}}{(L_c - u - 1)! (t + 1)! (t + 1)! (\gamma_T / \gamma_c)^{u+1}} \right) \\
&= \left[ 1 - e^{-\frac{(t+1)\gamma_T}{\gamma_c}} \right] \left[ \frac{(t+1)! \gamma_T}{\gamma_c} \right]^{u+1}
\end{align*}
\]

\[
\begin{align*}
\gamma \gamma_T &\leq \frac{\gamma_T - y}{\gamma_c - y} \\
&= \left( e^{-\frac{\gamma_T}{\gamma_c}} - e^{-\frac{y}{\gamma_c}} \right) \left( \frac{L_c L_c + L_L - L_L - 1}{(L_c - u - 1)!} \right)^{L_c - 1} \\
&\sum_{u=0}^{L_c - 1} \left( L_c \sum_{u=0}^{L_c - 1} \frac{(-1)^{L_c + 1}}{(L_c - u - 1)! (t + 1)! (t + 1)! (\gamma_T / \gamma_c)^{u+1}} \right) \\
&= f_{\gamma_{Le:L+La}}(x), \\
J(x) &= \left( e^{-\frac{\gamma_T}{\gamma_c}} - e^{-\frac{y}{\gamma_c}} \right) \left( \frac{L_c L_c + L_L - L_L - 1}{(L_c - u - 1)!} \right)^{L_c - 1} \\
&\sum_{u=0}^{L_c - 1} \left( L_c \sum_{u=0}^{L_c - 1} \frac{(-1)^{L_c + 1}}{(L_c - u - 1)! (t + 1)! (t + 1)! (\gamma_T / \gamma_c)^{u+1}} \right) \\
&= \left[ 1 - e^{-\frac{(t+1)\gamma_T}{\gamma_c}} \right] \left[ \frac{(t+1)! \gamma_T}{\gamma_c} \right]^{u+1}
\end{align*}
\]

\[
\begin{align*}
A(i : j, k, s) &= \int_{k}^{\infty} e^sx f_{\Gamma_{Le:L}}(x) dx = \frac{(j+1)}{i} \Gamma \left[ (i, (1-s\gamma )k / \gamma ) \right] + \frac{(j-1)}{i} \frac{1}{(i-1)! (1-s\gamma )^i} \\
&\sum_{l=1}^{j-1} \frac{(-1)^{j+l-1}}{j-l} \frac{(j-l)^i}{i} \\
&\left( e^{k(s-1)} \frac{1}{1+(s-1)^2} \right) + \sum_{m=0}^{i-2} \frac{1}{i (s\gamma - 1)} \frac{(l+1)! \gamma (m+1, (1-s\gamma )k / \gamma )}{m!}
\end{align*}
\]
Rayleigh fading channels is plotted. For comparison purpose, we also plot the average BER of BPSK with La/L−GSC. This graph, we set \( L_c = 3, L_a = 2, \) and \( \gamma_T = 5 \) dB. The simulation result for the case of \( L = 4 \) shows that our alternative simple approach is indeed a good approximation\(^4\). It is clear from this figure that our proposed scheme always outperforms MRC. Also it is very interesting to note that when the channel condition is poor, i.e., \( \gamma \) is relatively small compared to \( \gamma_T \), our scheme has the same error performance as GSC. This behavior can be explained as follows. When \( \gamma \) is small compared to \( \gamma_T \), our proposed scheme acts most of the times as \( L_c/(L + L_a) \)-GSC since \( L_c/L \)-GSC output SNR has a high chance of not exceeding the required target SNR. On the other hand, in good channel conditions, our scheme shows a higher error probability. This is because when \( \gamma \) becomes larger, the combined SNR of \( L_c/L \)-GSC has a higher chance to exceed the target SNR, \( \gamma_T \), and as such does not need to rely on the additional resolvable paths from the target BS. Hence, we can conclude that our proposed combiner relies on the additional resources provided by the target BS only in poor channel conditions. For a better understanding of our scheme, we study when \( L \) is fixed and \( L_a \) is variable in what follows.

Fig. 2 shows the average BER of BPSK with MRC, GSC, and the proposed combining scheme versus the average SNR per path, \( \gamma_T \), for various values of \( L_a \) over i.i.d. Rayleigh fading channels when \( L = 4, L_c = 3, \) and \( \gamma_T = 5 \) dB. Similar trends to those observed in Fig. 1 can also be seen in this figure, but since \( L \) is fixed, as one expects intuitively, all the curves of our proposed scheme are converging to the case of \( L_c/4 \)-GSC in the higher average SNR region.

We now study the average BER dependence on the threshold SNR, \( \gamma_T \). Fig. 3 represents the average BER of BPSK versus the average SNR per path, \( \gamma \), with MRC, GSC, and the proposed scheme for various values of \( \gamma_T \) over i.i.d. Rayleigh fading channels when \( L = 4, L_c = 3, \) and \( \gamma_T = 5 \) dB from this figure, it is clear that the higher the threshold, the better the performance, as one expects. However, high thresholds increase the path estimation load. We examine in what follows

\[ B(i : j, k) = F_{\gamma, j}(k) = \int_0^k f_{\gamma, j}(x)dx = \left( \frac{j}{i} \right) \frac{\gamma [i, k/\gamma]}{(i - 1)!} + \sum_{l=1}^{j-i-1} \frac{(-1)^{j-l} (j-i-1)}{l} \left( \frac{i}{l} \right) \frac{\gamma [m+1, k/\gamma]}{m!} \]

\[ C(k, s) = \int_0^\infty e^{x} \mathcal{I}(x)dx = \left\{ \begin{array}{ll} \frac{e^{\gamma/\gamma_T}}{s^{\gamma/\gamma_T}} & \text{if } s > \gamma_T \\ 1 - e^{-\gamma/\gamma_T} \sum_{u=0}^{\infty} \left( \frac{(t+1)\gamma_T}{\gamma_T - (t+1)\gamma_T} \right)^u & \text{if } s = \gamma_T \end{array} \right. \]
Fig. 3. Average BER of BPSK versus the average SNR per path, $\gamma$, with MRC, GSC, and the proposed scheme for various values of $\gamma T$ over i.i.d. Rayleigh fading channels when $L = 4, L_c = 3$, and $L_a = 2$.

Fig. 4. Average number of path estimations versus the output threshold, $\gamma_T$, with MRC, GSC, and the proposed scheme for various values of $L_a$ over i.i.d. Rayleigh fading channels when $L = 4, L_c = 3$, and $\gamma = 0$ dB.

Fig. 5. Average BER of BPSK versus the output threshold, $\gamma_T$, with MRC, GSC, and the proposed scheme for various values of $L_a$ over i.i.d. Rayleigh fading channels when $L = 4, L_c = 3$, and $\gamma = 0$ dB.

the issue in details.

B. Average Number of Path Estimations

With the proposed scheme, the RAKE receiver estimates $L$ paths in the case of $\Gamma_{L_c:L} \geq \gamma_T$ or $L + L_a$ in the case of $\Gamma_{L_c:L} < \gamma_T$. Hence, we can easily quantify the average number of path estimations, denoted by $N_E$, as

$$N_E = L \cdot \Pr \left[ \Gamma_{L_c:L} \geq \gamma_T \right] + \left( L + L_a \right) \cdot \Pr \left[ \Gamma_{L_c:L} < \gamma_T \right], \quad (24)$$

which reduces to

$$N_E = L + L_a \cdot F_{\Gamma_{L_c:L}}(\gamma_T), \quad (25)$$

where $F_{\Gamma_{L_c:L}}(\gamma_T)$ can be calculated from (21) for i.i.d. Rayleigh fading channels. Note that $L_c$-MRC and $L_c/(L+L_a)$-GSC always require $L_c$ and $L + L_a$ estimations, respectively. Fig. 4 shows the average number of path estimations versus the output threshold, $\gamma_T$, with MRC, GSC, and the proposed scheme for various values of $L_a$ over i.i.d. Rayleigh fading channels when $L = 4, L_c = 3$, and $\gamma = 0$ dB. For a better illustration of the tradeoff between complexity and performance, Fig. 5 shows the average BER of BPSK versus the output threshold, $\gamma_T$, with MRC, GSC, and the proposed scheme. As we can see, the error rate of the proposed scheme decreases to that of $L_c/(L+L_a)$-GSC when the output threshold increases. Considering Figs. 4 and 5 together, we observe that the proposed scheme can save a certain amount of estimation load with a slight performance loss compared to GSC if the required threshold is 2 to 6 dB above $\gamma$ for our chosen set of parameters.

C. SHO Overhead

In this section, we investigate the probability of the SHO attempt and the SHO overhead. In our proposed scheme, the SHO is attempted whenever $\Gamma_{L_c:L}$ is below $\gamma_T$. Hence, the probability of the SHO attempt is same as the outage probability of $L_c/L$-GSC evaluated at $\gamma_T$, i.e., $F_{\Gamma_{L_c:L}}(\gamma_T)$.

The SHO overhead, denoted by $\beta$, is commonly used to quantify the SHO activity in a network and is defined as [13, Eq. (9.2)]

$$\beta = \sum_{n=1}^{N} n P_n - 1, \quad (26)$$

where $N$ is the number of active BSs and $P_n$ is the average probability that the mobile unit uses $n$-way SHO.

1) $L_a < L_c$: Based on the mode of operation in Section II-B, $P_1$ and $P_2$ can be defined as

$$P_1 = \Pr \left[ \Gamma_{L_c:L} \geq \gamma_T \right] + \Pr \left[ \Gamma_{L_c:L} < \gamma_T, \Gamma_{L_c:L} \geq \gamma_1: L_a \right], \quad (27)$$
and

$$P_2 = 1 - P_1,$$  \hspace{1cm}  \hspace{1cm} (28)

where $\gamma_{L_c:L}$ is the $L_c$-th strongest path among $L$ ones from the serving BS and $\gamma_{1:L_a}$ is the strongest path among $L_a$ ones from the target BS. Substituting (27) and (28) into (26), we can express the SHO overhead, $\beta$, as

$$\beta = P_1 + 2P_2 - 1 = F_{\Gamma_{L_c:L}}(\gamma_T) \Pr[\gamma_{L_c:L} < \gamma_{1:L_a} \mid \Gamma_{L_c:L} < \gamma_T].$$  \hspace{1cm} (29)

Since $\gamma_{1:L_a}$ is independent to $\gamma_{L_c:L}$ and $\Gamma_{L_c:L}$, we can calculate the conditional probability, $\Pr[\gamma_{L_c:L} < \gamma_{1:L_a} \mid \Gamma_{L_c:L} < \gamma_T]$, as

$$\Pr[\gamma_{L_c:L} < \gamma_{1:L_a} \mid \Gamma_{L_c:L} < \gamma_T] = \int_0^{\gamma_T} F_{\gamma_{1:L_a} \mid \gamma_{L_c:L} < \gamma_T}(x)f_{\gamma_{1:L_a}}(x)dx.$$  \hspace{1cm} (30)

The conditional CDF in (30) can be written as shown in (31), where $f_{\Gamma_{L_c:L} \mid \gamma_{L_c:L} < \gamma_T}(y,z)$ can be obtained from (15). After successive substitutions from (31) to (29), we can express the SHO overhead, $\beta$, as shown in (32). Finally, integrating (32), we can obtain the exact closed-form expression for the SHO overhead, $\beta$ [17].

2) $L_a \geq L_c$: In this case, we need to consider the probability that a call is completely handed over to the target BS. Hence, the joint probability, $\Pr[\Gamma_{L_c:L} < \gamma_T, \gamma_{1:L} \leq \Gamma_{L_c:L}]$, should be added to $P_1$ in (27) where $\gamma_{1:L}$ is the strongest path among $L$ ones from the serving BS and $\gamma_{1:L_a}$ is the $L_c$-th strongest path among $L_a$ ones from the target BS. This joint probability is given by the analytical expression shown in (33), where $f_{\Gamma_{L_c:L} \mid \gamma_{L_c:L} < \gamma_T, \gamma_{1:L} < \gamma_{1:L_a}, \gamma_{1:L_a} < \gamma_{1:L_b}, \cdots}$ is the joint PDF of the first $L_c$ order statistics out of $L$ ones [15, Eq. (9.420)]. Unfortunately, it seems difficult to obtain a simple close-form expression for this nested $L_c + 1$ multi-fold integral.

Fig. 6 shows the SHO overhead versus the output threshold, $\gamma_T$, with GSC and the proposed scheme for various values of $L_a$ over i.i.d. Rayleigh fading channels when $L = 4, L_c = 3,$ and $\gamma = 0$ dB. The SHO overhead of $L_c/(L + L_a)$-GSC is plotted by calculating $P_1$ and $P_2$ as

$$P_1 = \begin{cases} \Pr[\gamma_{L_c:L} \geq \gamma_{1:L_a}], & L_a < L_c; \\ \Pr[\gamma_{L_c:L} \geq \gamma_{1:L_a}] + \Pr[\gamma_{1:L} \leq \gamma_{L_c:L}], & L_a \geq L_c, \end{cases}$$  \hspace{1cm} (34)

and

$$P_2 = 1 - P_1.$$  \hspace{1cm} (35)

It is clear from this figure that we have a higher chance to use 2-way SHO, as the number of additional paths from the target BS increases. Note that our proposed scheme acts as $L_c/(L + L_a)$-GSC when the output threshold is very high. Hence, we can observe that the SHO overhead of the proposed scheme converges to that of GSC as $\gamma_T$ increases. From this figure together with Fig. 5, we can see the SHO overhead reduction of our proposed scheme. For example, if the required threshold is 6 dB above $\gamma$, our scheme shows for $L_a = 2$ around 0.55 SHO overhead while maintaining the same error rate as GSC which requires 0.8 SHO overhead.

V. CONCLUSION

In this paper, we proposed a new finger assignment scheme for RAKE receivers in the SHO region. In this scheme, the receiver checks the GSC output SNR from the serving BS against a certain pre-determined output threshold. If the output SNR is below this threshold, the receiver performs a finger reassignment after using GSC on the paths coming from the serving BS and the target BS. We derived the statistics of the output SNR of the proposed scheme in accurate approximate closed-form, based on which we carried out the performance analysis of the resulting systems. We showed through numerical examples that the new scheme offers commensurate performance in comparison with more complicated GSC-based diversity systems while requiring a smaller estimation load and SHO overhead.

APPENDIX

DERIVATION OF EQ. (10)

The joint probability $\Pr[\gamma_T \geq \Gamma_{L_c:L} + L_a < x, \Gamma_{L_c:L} < \gamma_T]$ in (7) can be written as

$$\Pr[\gamma_T \geq \Gamma_{L_c:L} + L_a < x, \Gamma_{L_c:L} < \gamma_T] = \Pr[\Gamma_{L_c:L} < \gamma_T] \Pr[\gamma_T \geq \Gamma_{L_c:L} + L_a < x \mid \Gamma_{L_c:L} < \gamma_T].$$  \hspace{1cm} (36)

For simplicity, if we define the events $A, B,$ and $C$ as

$$A = \gamma_T \leq \Gamma_{L_c:L} + L_a < x,$$  \hspace{1cm} (37)

$$B = \Gamma_{L_c:L} < \gamma_T,$$  \hspace{1cm} (38)

$$C = \Gamma_{L_c:L} + L_a - 1 < \gamma_T,$$  \hspace{1cm} (39)

then (36) can be rewritten as (40), where $\overline{C}$ is the complementary set of event $C$, i.e., $\overline{C} = \Gamma_{L_c:L} + L_a - 1 \geq \gamma_T$. Since event $B$ includes event $C$, we have $\Pr[A \mid B, C] = \Pr[A \mid C]$. Note also that when event $B$ and event $\overline{C}$ happened, $\Gamma_{L_c:L}$ and $\Gamma_{L_c:L} + L_a - 1$ are sums of different set of exponential random variables. Based on the memoryless property of exponential random variables and noting that given that $\overline{C}$ happened after
\[ F_{\gamma_{Lc:L}|\Gamma_{Lc:L} < \gamma_T}(x) = \frac{\Pr[\gamma_{Lc:L} < x, \Gamma_{Lc:L} < \gamma_T]}{\Pr[\Gamma_{Lc:L} < \gamma_T]} = \frac{\Pr[\gamma_{Lc:L} < x, \Gamma_{Lc:L-1} + \gamma_{Lc:L} < \gamma_T]}{\Pr[\Gamma_{Lc:L} < \gamma_T]} = \frac{1}{\Pr[\Gamma_{Lc:L} < \gamma_T]} \int_0^{\gamma_T} f_{\gamma_{Lc:L}}(y) f_{\Gamma_{Lc:L-1} \gamma_{Lc:L}}(y, z) dy dz, \quad 0 \leq x \leq \frac{\gamma_T}{L_c} \]  

(31)

\[ \beta = \int_0^{\gamma_T} \left( f_{\gamma_{Lc:L}}(x) \int_0^{\gamma_T-y} f_{\Gamma_{Lc:L-1} \gamma_{Lc:L}}(y, z) dy \right) dx + \left[ 1 - F_{\gamma_{Lc:L}}(\frac{\gamma_T}{L_c}) \right] \int_0^{\gamma_T} f_{\Gamma_{Lc:L-1} \gamma_{Lc:L}}(y, z) dy 
\]

(32)

\[
\Pr[\Gamma_{Lc:L} < \gamma_T, \gamma_{Lc:L} \leq \gamma_{Lc:L}] = \int_0^{\gamma_T} \int_0^{\gamma_T-y} \int_0^{\gamma_T-x} \int_0^{\gamma_T-y} \int_0^{\gamma_T-x} \cdots \int_0^{\gamma_T-x} \Pr[\gamma_{Lc:L} < x, \Gamma_{Lc:L} < \gamma_T] \]

(33)

\[
\Pr[\Gamma_{Lc:L} \leq \gamma_T | \Gamma_{Lc:L+1} - 1 < \gamma_T] = \frac{1}{1 - \Pr[\Gamma_{Lc:L} < \gamma_T]} - \frac{1 - \Pr[\Gamma_{Lc:L} < \gamma_T]}{1 - \Pr[\Gamma_{Lc:L} < \gamma_T]},
\]

(45)

we finally arrive at the desired result given in (10).

REFERENCES


Pr \[ A, B \] = Pr \[ B \] (Pr \[ A, C \] Pr \[ C | B \] + Pr \[ A | C \] Pr \[ C | B \])

\[ = Pr \[ B \] \left( \frac{Pr \[ A, C \]}{Pr \[ C \]} Pr \[ B | C \] + \frac{Pr \[ A, C \] - Pr \[ B | C \]}{Pr \[ C \] - Pr \[ B | C \]} \right) \]

(41)

Pr \[ \gamma_T \leq \Gamma_{L_c:L} x \] \[ \Gamma_{L_c:L} \leq \gamma_T \]

(42)

\[ \begin{align*}
&= Pr \[ \gamma_T \leq \Gamma_{L_c:L} x \] \[ \Gamma_{L_c:L} = 1 \] \leq \gamma_T \] Pr \[ \Gamma_{L_c:L} \leq 1 \] \leq \gamma_T \]

\[ + Pr \[ \gamma_T \leq \Gamma_{L_c:L} x \] \[ \Gamma_{L_c:L} = 1 \geq \gamma_T \] Pr \[ \Gamma_{L_c:L} \leq 1 \] \geq \gamma_T \]

Seyeong Choi (S’03, M’08) was born in Seoul, Korea. He received his B.S. and M.S. degrees from Hanyang University, Seoul, Korea, in 1996 and 1998, respectively, and Ph.D. degree in Electrical and Computer Engineering from the Texas A&M University, College Station, TX, in 2007. From 1998 through 2001 he worked as a researcher at LGTeleCom’s Network R&D Center, Seoul, Korea, operating the CDMA mobile communication networks. His research interests include wireless communications, MIMO fading channels, diversity techniques, and system performance evaluation.

Mohamed-Slim Alouini (S’94, M’98, SM’03) was born in Tunis, Tunisia. He received the Ph.D. degree in Electrical Engineering from the California Institute of Technology (Caltech), Pasadena, CA, USA, in 1998. He was an Associate Professor with the department of Electrical and Computer Engineering of the University of Minnesota, Minneapolis, MN, USA. Since July 2006, he has been with the Electrical Engineering program of Texas A&M University at Qatar, Education City, Doha, Qatar. His current research interests include the design and performance analysis of wireless communication systems.

Khalid A. Qaraqe (M’97, SM’00) was born in Bethlehem. Dr. Qaraqe received the B.S. degree in EE from the University of Technology, Baghdad in 1986, with honors, he received the M.S. degree in EE from the University of Jordan, Jordan, in 1989, and he earned his Ph.D. degree in EE from Texas A&M University, College Station, TX, in 1997. From 1989 to 2004 Dr. Qaraqe has held a variety positions in many companies and he has over 12 years of experience in the telecommunication industry. Dr. Qaraqe has worked for Qualcomm, Enad Design Systems, Cadence Design Systems/Tality Corporation, STC, SBC and Ericsson. He has worked on numerous GSM, CDMA, WCDMA projects and he has experience in product development, design, deployments, testing, and integration. Dr. Qaraqe joined the department of Electrical Engineering of Texas A&M University at Qatar, in July 2004, where he is now a visiting associate professor. Dr. Qaraqe research interests include performance analysis of the 3G UMTS wireless communication, WCDMA estimation theories, Fading channels, Frequency hopping, and STTD diversity.

Hong-Chuan Yang (S’00, M’03, SM’07) was born in Liaoning, China. He received the Ph.D. degree in Electrical Engineering from the University of Minnesota, MN, USA in 2003. Dr. Yang is now an assistant professor of the Electrical and Computer Engineering Department of the University of Victoria, Victoria, B.C., Canada. His research focuses on different aspects of wireless and mobile communications, with special emphasis on diversity techniques, cross-layer design, energy-efficient communications, and system performance evaluation. Dr. Yang is a recipient of the Doctoral Dissertation Fellowship (DDF) Award from the graduate school of the University of Minnesota for the 2002-2003 academic year.
Novel GLRT Packet-Data Receivers

Haoli Qian, Stella N. Batalama, and Bruce W. Suter

Abstract — In this paper we design novel generalized likelihood ratio test (GLRT)-type packet-data detectors for general multi-access/multiuser digital communication systems and we develop analytical performance evaluation tools for finite data packet sizes. For the known channel case, we derive a coherent GLRT packet-data detector while for the unknown channel case we derive both a coherent pilot assisted GLRT packet-data detector and a differential phase-shift-keying (DPSK) GLRT packet-data detector. Efficient suboptimum implementations of the above schemes that exhibit complexity linear in the packet size are also considered. Simulation studies evaluate the performance of the proposed schemes in the context of packet-data code-division multiple access (CDMA) communications.

Index Terms — Demodulation, finite data analysis, generalized likelihood ratio test, multiuser channels, packet radio, probability of error.

I. INTRODUCTION

The concept of packet data receivers emerged to facilitate packet data communications where the statistics of the environment during the transmission of a packet remain almost unchanged while they may significantly vary between two different packets. In this context, offline receiver training [1] or training between packets become unreliable and an adaptive packet data receiver may have to be redesigned/reevaluated as often as a new data packet is received. In mobile packet data communications, for example, the size of the data record that is available for receiver adaptation and redesign is limited by the coherence time of the communication link and may be in the order of 300 symbols or less in practical situations [2]. In the same spirit, assessment of receiver performance is realistic only when it relies on finite (and often short) rather than asymptotically long data records. This work was motivated primarily by the above considerations.

The optimum rule for the detection of a transmitted packet of digital data under perfectly known parameters of the received data joint probability density function is the well known likelihood ratio test (LRT) that selects the most likely data combination among a finite set of alternatives. When, however, there are unknown parameters in the received signal model/density function and a uniformly most powerful (UMP) test does not exist [3], the design of a detection scheme becomes a coupled optimization process that involves joint detection (hypothesis testing) and parameter estimation. As such, we can either solve the estimation part first (i.e., maximize the likelihood of each hypothesis with respect to the unknown parameters) and then solve the detection part (i.e., choose the most likely hypothesis) or execute the optimization sequence in the opposite order. Under certain general conditions, the above double maximization problems are equivalent and result to what is known as the generalized likelihood ratio test (GLRT). In particular, the estimation-detection sequence of optimization is the most intuitive and results in a likelihood ratio test that utilizes maximum likelihood (ML) estimates of the unknown parameters. On the other hand, for specific applications, the detection-estimation optimization sequence, although a more difficult optimization problem in general, may lead to computationally simpler test implementations than the estimation-detection sequence [4]. In any case, the overall statistical optimality of GLRT tests is difficult to be determined theoretically, if at all possible. An alternative, ad-hoc but frequently used approach to the design of a detection scheme in the presence of unknown parameters in the distribution of the received data is to proceed by directly substituting parameter estimates in the LRT formula.

GLRT has been a rather popular methodology in the past for radar signal detection problems (as seen for example in [5]–[8] and references therein) while it has been given limited consideration in the context of multiaccess/multiuser digital communications [9]–[13]. The binary nature of the radar hypothesis testing problem as well as the availability of secondary data (pure disturbance observations) in addition to primary data (data that include both the signal of interest and disturbance) facilitates greatly the design of the GLRT test. On the other hand, we may argue that GLRT approaches to multiaccess/multiuser digital communications are not so straightforward because of the usual non-binary nature of the detection problem, the absence of secondary data, and/or the non-Gaussian characteristics of the disturbance (e.g. multiaccess interference).

In this paper we study GLRT packet-data detectors for general multiaccess/multiuser digital communication systems. We consider a system where the transmitted signal of the user of interest experiences multipath quasi-static fading, and is received in the presence of multiple access interference and additive noise. The overall effect of the disturbance is assumed to be Gaussian distributed with unknown covariance matrix. While this assumption facilitates receiver design as we will see later in this paper, it causes insignificant performance

0090-6778/08$25.00 © 2008 IEEE
degradation with respect to ideal designs that are usually unrealizable. Given a received data packet, our approach aims at simultaneous estimation of the unknown disturbance covariance matrix and multipath channel and detection of the digital data within the same packet.

In particular, we develop: a) a coherent GLRT packet-data detector for the known channel case, b) a coherent GLRT pilot assisted detector for the unknown channel case (the channel is estimated implicitly as part of the GLRT formulation while short pilot signaling is used to resolve phase ambiguity), and c) a differential GLRT detector for differentially encoded packet-data. In view of the exponential complexity in the size of the data packet of the above GLRT schemes, we also propose suboptimum implementations that exhibit linear complexity. Last but not least, we develop analytical performance evaluation tools for finite data packet sizes. The importance/novelty of our analytical performance evaluation tools lies in the fact that they deviate from the conventional and convenient yet inaccurate performance analysis assumption of sufficiently (or infinitely) long packet sizes, i.e. long enough to provide good estimates of the environment statistics [3], [14], usually without taking into account the restrictions imposed by the dynamic nature of the environment which may limit substantially the size of the data record that is available for adaptation and redesign. Instead, our formulas provide the bit error rate (BER) that can be reached by a GLRT test for any given finite data packet size as well as the size of a data packet that is necessary for the test to reach any pre-specified BER level. Comparative studies included in this paper establish analytically the BER performance merits of the new GLRT tests relative to the popular practice of directly substituting sample average estimates of the unknown parameters in the LRT formula. The proposed GLRT detectors are evaluated in the context of packet-data CDMA communications and state-of-the-art performance is demonstrated.

The rest of the paper is organized as follows. Our signal model and some background information are presented in Section II. The known and unknown channel cases are treated in Section III and IV, respectively. Therein, the proposed packet-data detectors are derived, their suboptimum implementations of linear complexity are outlined, and the relationship between the size of the available data record and the BER performance that can be reached by each GLRT scheme is identified analytically. Section V is devoted to simulation studies and comparisons. A few concluding remarks are drawn in Section VI.

II. SIGNAL MODEL AND BACKGROUND

We consider the following general discrete-time signal model for a transmitted data packet of interest of size $N$:

$$x_i = \sqrt{E} b_i s, \quad i = 1, 2, \ldots, N,$$

where the information bits $b_i$, $i = 1, \ldots, N$, take values $\pm 1$ with equal probability, are independent from each other, and modulate a known discrete-time complex signal waveform of unit norm, $s \in \mathbb{C}^2$, $||s|| = 1$, where $G$ denotes the number of sample periods that $s$ occupies. With this setup, $E$ represents total transmitted energy per bit. We assume that the transmitted signal experiences multipath quasi-static fading of maximum delay $M < G$ sampling periods with negligible inter-symbol interference (ISI) effects [9]-[13]. If $(y_i \in \mathbb{C}^L)_{i=1}^N$, $L = G + M$, denotes the corresponding received data packet, then

$$y_i = b_i \sqrt{E} S a + n_i, \quad i = 1, \ldots, N,$$

where $S$ is the $L \times M$ known signal waveform delay matrix, $a \in \mathbb{C}^M$ is the vector of the multipath channel coefficients that are assumed to remain constant during the transmission of the data packet, and $n_i$, $i = 1, \ldots, N$, is a sequence of independent identically distributed zero mean Gaussian vector with unknown covariance matrix $R_n$, that represent comprehensively channel interference and noise$^1$ that is independent of the data sequence $b_i$, $i = 1, \ldots, N$.

The probability density function (pdf) of the observations $Y \triangleq [y_1, y_2, \ldots, y_N]$ conditioned on the transmitted bits $b = [b_1, b_2, \ldots, b_N]^T$ can be expressed in the following compact form:

$$f(Y | b; E, S, a, R_n) = \frac{1}{\pi^L N} \text{trace}\left[ -R_n^{-1} \text{trace}(Y - \sqrt{E} S a)^H (Y - \sqrt{E} S a)^T \right].$$

When $a$ and $R_n$ are perfectly known, the optimum rule for the detection of $b_i$, $i = 1, \ldots, N$, simplifies to the familiar one-shot tests

$$\hat{b}_{i\text{ML}} = \text{sgn} \left\{ \text{Re}(R_n^{-1} S a y_i) \right\}, \quad i = 1, \ldots, N,$$

where $\text{sgn}[x]$ extracts the sign of the real variable $x$ and $\text{Re}(y)$ extracts the real part of the complex scalar $y$. In other words, the optimum maximum likelihood (ML) detector of $b \in \{\pm 1\}^N$ reduces to $N$ individual applications of the linear filter $R_n^{-1} S a$ followed by an output sign detector and the detection of all bits in the packet has computational cost linear in the packet size (detection of $b_i$ depends only on $y_i$ and joint detection of all bits in the packet is equivalent to disjoint bit-by-bit detection)$^2$.

On the other hand, when a priori knowledge of the parameters $a$ and/or $R_n$ cannot be assumed, we may proceed in one of two different ways: (i) We can use again the parametrically described test in (4) and substitute the unknown parameters/statistics by corresponding estimates (usually sample-average estimates) which results in a scheme that maintains linear complexity in the packet size; or (ii) we may carry out joint detection and parameter estimation which results in superior performance GLRT schemes at the expense of increased complexity. In the next section, we propose a novel packet-data GLRT test for the case of known multipath channels (parameter $a$) and colored Gaussian disturbance of unknown correlation matrix (parameter $R_n$) and we establish analytically how the new GLRT test compares to the common sample-average LRT parameter substitution approach.

---

$^1$The effective non-zero correlation between the transmitted signal waveforms that correspond to different users shift the second order statistics of the received signal pertinent to each user away from the “white” assumption.

$^2$In the presence of severe ISI, (3) and (4) should be considered, respectively, as an approximate pdf and as a suboptimum decision rule. Then all subsequent developments should be treated in this context. Incorporation of ISI directly in (2) will result in mathematically cumbersome and less intuitive expressions.
Alongside, we develop analytical tools that provide the BER performance of the proposed test for any given data packet size as well as the data packet size that is necessary for the test to reach any given BER level.

III. GLRT DETECTION: KNOWN CHANNEL
A. Algorithmic Development

For convenience, define $v^\Delta = \sqrt{E}Sa$ where $a$ is the known channel coefficient vector and $S$ is the known signal waveform matrix. The GLRT packet-data detector is given by the following Proposition.

Proposition 1: The GLRT test for the detection of the data packet $b$ of size $N$ in the presence of complex Gaussian disturbance with unknown covariance matrix $R_n$ is

$$
\hat{b}_{GLRT} = \arg \max_b \left\{ \frac{1}{N} f(Y | b, v, R_n) \right\}
$$

$$
= \arg \max_b \left\{ l_1(b) \right\}
$$

(5)

where

$$
l_1(b) = \frac{1}{N} b^T Y^H R_{SA}(N)^{-1} v + N v^H R_{SA}(N)^{-1} Yb
+ (b^T Y^H R_{SA}(N)^{-1} Yb) (v^H R_{SA}(N)^{-1} v)
- (b^T Y^H R_{SA}(N)^{-1} v) (v^H R_{SA}(N)^{-1} Yb)
$$

(6)

and $R_{SA}(N)^{-1} = \frac{1}{N} YY^H$ is the sample average received data correlation matrix.

Proof: For a given bit combination $b$, the maximum of $f(Y | b, v, R_n)$ is reached when $R_n$ is the ML estimate, i.e., $R_n = \hat{R}_{ML}(b) = \frac{1}{N} (Y - vb^T)^T (Y - vb^T)^H$. Thus, (5) reduces readily to

$$
\hat{b}_{GLRT} = \arg \min_b \left\{ \frac{1}{N} (Y - vb^T)^T (Y - vb^T)^H \right\}
$$

$$
= \arg \min_b \left\{ R_{SA}(N) - \frac{vb^T Y^H}{N} - \frac{Yvb^H}{N} + vv^H \right\},
$$

(7)

where $| |$ denotes the determinant of the matrix operand. The test in (7) can be further reduced to (5) using the identity

$$
(\hat{R}_{SA} - v \frac{Yb}{N}) H^H = \frac{Yb}{N} H^H + vv^H
$$

$$
(\hat{R}_{SA} - v \frac{Yb}{N}) H^H = \frac{v^H R_{SA}^{-1} Yb}{N} - \frac{(Yb)^H}{N} R_{SA}^{-1} v
+ v^H R_{SA}^{-1} \frac{Yb}{N} R_{SA}^{-1} v - \frac{Yvb^H}{N} + vv^H
$$

(8)

that holds true for any Hermitian positive definite matrix $R_{SA}$ and vectors $v$ and $\frac{Yb}{N}$. The identity in (8) can be derived by repeated determinant calculations of rank one updates of a nonsingular matrix.

We note that direct implementation of test in (5) has complexity exponential in the packet size $N$.

The GLRT test in (5) can be contrasted with the standard “sample-matrix-inversion” (SMI) detection scheme [15]-[16] that replaces the unknown parameter $R_n$ of the LRT expression in (4) by a sample average estimate $R_{nSA}(K) = \frac{1}{K} \sum_{k=1}^{K} n_k n_k^H$ based on pure disturbance observations $n_k$, $k = 1, \ldots, K$ that are independent from $y_i$, $i = 1, \ldots, N$:

$$
b_{SMI} = \text{sgn} \left\{ v^H R_{nSA}^{-1}(K) y_i \right\}, \quad i = 1, \ldots, N.
$$

(9)

When pure disturbance observations (secondary data) $n_k$, $k = 1, \ldots, K$ are not available, a popular version of the test in (9) utilizes directly the sample-average correlation matrix of the (desired-signal-present) received data, $R_{SA}(N)$, evaluated using the same received data $y_i$, $i = 1, \ldots, N$, that are processed by the detector [17]. We denote this test by

$$
b_{SMI-dsp} = \text{sgn} \left\{ v^H R_{SA}^{-1}(N) y_i \right\}, \quad i = 1, \ldots, N,
$$

(10)

where the subscript part “dsp” stands for desired-signal-present. While it is understood that (9) and (10) converge with probability 1 to the test in (4) as the $K, N \to \infty$, recent analytical results on short data record adaptive filtering [18]-[20] indicate that for finite sample support of equal size ($K = N$) the test in (9) outperforms the test in (10) in terms of BER. This performance degradation in (10) was also documented in the array processing literature [1], [17]. Yet, as Theorem 1 shows below, if we utilize the new packet-data GLRT detector in (5), we can achieve approximately the same average BER performance as with the test in (9) without the need for pure disturbance observations (secondary data) that are independent of the received data packet. The proof is given in the Appendix.

Theorem 1: (i) Let $b$ be the transmitted data packet and $\hat{b}$ a data packet decision that differs from $b$ in $m$ bits (i.e., contains $m$ bits in error). Then, the pairwise error probability is given by

$$
P \left[ l_1(b) > l_1(\hat{b}) \left| b \right. \right] = \int_0^1 Q(\sqrt{2m\gamma}x) \frac{e^{N-1}(1-x)^{L-2}}{B(N-L+1, L-1)} dx
$$

(11)

where $Q(x) = \frac{1}{\sqrt{\pi}} \int_x^\infty e^{-u^2/2} du$, $\gamma = \frac{\Delta}{\sqrt{v^H R^{-1} v}}$, and $l_1(\cdot)$ is given by (6).

(ii) For sufficiently large transmitted energy per bit $E$, the average BER of the GLRT detector in (5) is approximately equal to the average BER of the scheme in (9) that utilizes $R_{nSA}$ evaluated based on $K = N - 1$ pure disturbance observations

$$
\lim_{\gamma \to \infty} \frac{\text{BER}_{GLRT}(N)}{\text{BER}_{SMI}(N - 1)} \approx 1.
$$

(12)

Using Theorem 1, we can derive an approximation of the average BER of the GLRT detector and evaluate the packet size that is necessary for the GLRT detector to achieve any given BER performance level. Our findings are summarized in the following theorem whose proof is in the Appendix.

Theorem 2: (i) The average BER of the GLRT detector that operates on a data packet of size $N \geq L + 2$ is given by (13) and (14), where $M(a, b, z)$ is Kummer’s confluent hypergeometric function, $\gamma = \frac{\Delta}{\sqrt{v^H R^{-1} v}}$, $\mu = \frac{N-L+1}{N^2(N+1)}$, and $\sigma^2 = \frac{(N-L+1)(L-1)}{N^2(N+1)}$.
steps as follows. At each step, only one bit in the data packet is updated (which is a common practice for recursive non-convex optimization problems). We also note that the ad hoc parallel search algorithm using the best sequence estimate among the current \( P \) alternatives. The performance of this algorithmic method is examined in the simulation studies of Section V.

### IV. GLRT DETECTION: UNKNOWN CHANNEL

#### A. Algorithmic Development

In this section we investigate the case of unknown channel, that is unknown \( E \) and \( a \) according to the signal model in (2). The following proposition provides the GLRT detection scheme for this case.

**Proposition 2:** The GLRT test for the detection of the data packet \( b \) of size \( N \) transmitted over an unknown linear channel in the presence of complex Gaussian disturbance of unknown covariance matrix \( R_n \) is given by (16) and (17).

**Proof:** For a given bit combination \( b \) and channel coefficient vector \( a \), \( f (Y | b, S, a, R_n) \) is maximized for

\[
R_n = R_{n,ML} (b, a) \triangleq \frac{1}{N} (Y - Sab^T) (Y - Sab^T)^H
\]

and (16) becomes

\[
\hat{b}_{GLRT} = \arg \max_{b} \max_{a} \left| \frac{1}{N} (Y - Sab^T) (Y - Sab^T)^H \right|^{-1}.
\]

The inner maximization is solved by finding the stationary point with respect to \( a \). Using the identity in (8) and after some simplifications, we obtain (20). Substituting (20) into (19) leads to the detection rule in (16).

The following theorem evaluates the asymptotic pairwise probability of error of the above detection rule in the high SNR region. The proof can be found in the Appendix.

**Theorem 3:** Let \( b \) be the transmitted data packet and \( \hat{b} \) a data packet decision that differs from \( b \) in \( m \) bits. Then,

\[
\lim_{E \to \infty} \int_{0}^{1} Q \left( \frac{2m(N-m)}{\gamma^2} \right)^{\left( \frac{N-1}{2} \right)} \left( \frac{2m(N-m)}{\gamma^2} \right)^{\left( \frac{N-1}{2} \right)} dx = 1
\]

where \( \gamma = \sqrt{\nu} R_{c}^{-1} v \).

We note that exploiting the structure of the signal waveform matrix \( S \) leads to significantly reduced dimension of the effective optimization/search space. We also note that the
Suboptimal GLRT algorithm

Initialization: Number of parallel search paths $P$; search depth $D$;
initial decision vector $\hat{b}^{(p)}(0) := \left[\hat{b}_{1}^{(0)}, \hat{b}_{2}^{(0)}, \ldots, \hat{b}_{N}^{(0)}\right]^{T}$, $p = 1, 2, \ldots, P$;
search index sequences $\left\{\pi_{p}(n)\right\}_{n=1}^{\bar{N}}$, $p = 1, 2, \ldots, P$.

For step $d = 1, 2, \ldots, D$
For path $p = 1, 2, \ldots, P$
\quad $i := \pi_{p}(d \text{ mod } N)$
\quad $\hat{b}_{i}^{(p)}(d) := \arg\max_{b_{i}} \left\{\max_{R_{n}} f\left(Y \left| \left\{\hat{b}_{j}^{(p)}(d-1)\right\}_{j \neq i}, \hat{b}_{i}^{(p)}, v, R_{n}\right\}\right\}$
\quad $\hat{b}_{j}^{(p)}(d) := \hat{b}_{j}^{(p)}(d-1)$, \quad $j \neq i$.
end
\end{verbatim}
\[
\hat{b}_{\text{GLRT}} := \arg\max_{b} l_{2}(b).
\]

where
\[
l_{2}(b) = \frac{Nb^{T}Y^{H}[YY^{H}]^{-1}S(S^{H}[YY^{H}]^{-1}S)^{-1}S^{H}[YY^{H}]^{-1}Yb}{N^{2} - Nb^{T}Y^{H}[YY^{H}]^{-1}Yb}
\] (17)

and thus eliminates the phase ambiguity problem. This phase ambiguity problem is also present in Theorem 3. In practice, phase ambiguity is resolved either by using a pilot sequence or by employing differential modulation at the transmitter; the rest of this section deals exactly with these two approaches.

A.1 Pilot Assisted GLRT Detection

Proposition 3: Let $\left\{b_{j}\right\}_{j=1}^{J}$ and $\left\{b_{i}\right\}_{i=J+1}^{N}$ denote, respectively, J known pilot bits and $N - J$ unknown information bits within the data packet b of size N that is transmitted over an unknown linear channel in the presence of complex Gaussian disturbance of unknown covariance. Then, the pilot assisted GLRT detector of $\left\{b_{i}\right\}_{i=J+1}^{N}$ is given by
\[
\hat{b}_{\text{GLRT}} = \arg\max_{b} \left\{\max_{a, R_{n}} f\left(Y \left| b, S, a, R_{n}\right\}\right\} = \arg\max_{b} l_{2}(b)
\]
(16)

Proof: We note that the joint conditional pdf of the observations is given by
\[
f\left(\left\{y_{i}\right\}_{i=1}^{N} \left| b_{\left\{j\right\}_{j=1}^{J}}, b_{\left\{i\right\}_{i=J+1}^{N}}, S, a, R_{n}\right\}\right) = \frac{R_{n}^{-N} \exp \left\{-\sum_{i=1}^{N} (y_{i} - b_{i}S_{a})^{H}R_{n}^{-1}(y_{i} - b_{i}S_{a})\right\}}{\pi^{L-N}}.
\]
(23)

Thus, the GLRT detection algorithm is as in (16) with a difference only in the support of the outer optimization.

It is interesting to note that in (22) the pilot sequence $\left\{b_{i}\right\}_{i=1}^{J}$ is not used to directly estimate the phase in an explicit manner but is rather incorporated implicitly in the GLRT rule. It is also interesting to observe that the GLRT test expression in (22) maintains the same structure as in (17). For a reasonably long pilot sequence, e.g. $J \geq 2$, Theorem 3 implies that we can safely neglect the pairwise probability of error in the high SNR region for $m \geq 2$ (we note that the above pilot assisted GLRT detection rule ensures that $1 \leq m \leq N - J$ and thus eliminates the phase ambiguity problem). We conclude the treatment of the pilot assisted GLRT detection problem by deriving an approximation of the BER performance of the test in (22) and evaluating the size of the data packet that is necessary for the detector to achieve any given BER performance level (the proof utilizes Theorems 2 and 3 and is omitted due to lack of space).

Corollary 1: (i) The average BER of the pilot assisted GLRT detector for a data packet of size $N \geq L + 3$ is given by
\[
\text{BER}_{\text{GLRT-pilot}}(N) \approx \text{BER}_{\text{SMI}}(N - 2)
\]
\[
= \frac{1}{\pi} \int_{0}^{\pi/2} M \left(\frac{N - L}{N - 1} - \frac{(N - 1)\gamma}{N \sin^{2}\theta}\right) d\theta
\]
\[
\approx \frac{2}{3} Q\left(\sqrt{2\mu} + \frac{1}{6} Q\left(\sqrt{2\mu + 2\sqrt{3}\sigma}\right)\right)
\]
(24)

where $\mu \triangleq \frac{-N - L}{N} \gamma$, $\sigma^{2} \triangleq \frac{(N - L)(L - 1)}{N} \gamma^{2}$, and $\text{BER}_{\text{SMI}}(N - 2)$ is the BER of the coherent SMI detector in (9) that would require perfect knowledge of $a$ and utilize $K = N - 2$ independent pure disturbance observations.

(ii) For any given $\nu$, the smallest packet size $N_{\nu}$ that guarantees that the BER performance of the GLRT pilot assisted packet-data detector is within $\nu$ dB from the BER performance of the optimum coherent ML detector in (4)
(i.e., $BER_{GLRT-pilot}(N) \leq Q(\sqrt{2\gamma 10^{-\nu/10}})$) is given by the ceiling of the maximum real root of the cubic equation
\[
N^3 - \frac{2L}{1 - 10^{-\frac{\nu}{10}}} N^2 + \frac{L^2 - 3(L - 1)}{(1 - 10^{-\frac{\nu}{10}})^2} N + \frac{3L(L - 1)}{(1 - 10^{-\frac{\nu}{10}})^2} = 0.
\]
(26)

Corollary 1 implies that the pilot assisted GLRT detector performs closely to the coherent SMI detector in (9) that requires perfect knowledge of $a$ and assumes availability of pure disturbance observations. We can modify in a straightforward manner the algorithm outlined in Section III.B to obtain a suboptimum implementation of the unknown-channel pilot GLRT scheme in (22) that exhibits linear complexity.

A.2 DPSK GLRT Detection

As an alternative to pilot signaling, phase ambiguity of the GLRT detector in (17) can be resolved by employing differential encoding at the transmitter. To avoid redundancy in our presentation, in this section we keep the size of the transmitted packet equal to $N$ while the number of the information bits embedded in the differentially encoded packet is $N - 1$, $\{b_i\}_{i=1}^{N-1}$. The differentially encoded bits themselves are $d_0 = +1$ and $d_i = d_{i-1} b_i$, $i = 1, 2, \ldots, N - 1$. The transmitted vector $y_i$ is still of the form of (2) with $d_i$ in place of $b_i$. Given the transmitted bits $d_i$, $i = 0, 1, \ldots, N - 1$, the information bits can be uniquely determined by $b_i = d_{i-1} d_i$, $i = 1, 2, \ldots, N - 1$.

We recall [22] that under ideal conditions (i.e., perfectly known interference-plus-noise statistics and channel impulse response) the ideal optimum (ML) differential detector of a packet of $N - 1$ information bits consists of the ideal linear filter $R_y^+ S_a$ followed first by a sign detector and then by the 2-symbol block differential decoder. On the other hand, when interference-plus-noise statistics are perfectly known but the channel is known only within a phase ambiguity $\theta$, the optimum (ML) differential detector of a block (packet) of $N - 1$ information bits consists of the ideal linear filter $R_y^+ S_a e^{i\theta}$ followed by an $N$-symbol differential decoder that operates on a block of complex, in general, scalar outputs of the optimum linear filter [23] (the linear filter $R_y^+ S_a e^{i\theta}$ provides the sufficient statistics for differential decoding). Direct implementation of the optimum block differential decoder according to the likelihood metric requires exponential complexity in the block (packet) size $N$ (fast approximate algorithms with complexity $N \log N$ can be used instead [24]-[27]). It is well understood that the BER performance of the phase-ambiguity-optimum block differential detector is lower bounded by the BER performance of the all-known differential detector and approaches this lower bound as $N \to \infty$. A popular suboptimum receiver for differentially encoded data has been the 2-symbol differential detector that utilizes a 2-symbol only differential decoder and detects one bit at a time by evaluating the sign of the real part of the product of the current filter output with the previous conjugated filter output.

Under unknown input statistics and channel coefficients, the common approach has been to produce estimates of the unknown quantities and insert the estimates in the N-symbol (or 2-symbol) block differential detector. Instead, what we propose herein is a GLRT-type scheme that combines into a single optimization effort estimation of interference-plus-noise covariance matrix and channel coefficients and detection of packet data. The following proposition identifies our DPSK GLRT scheme.

**Proposition 4:** The DPSK GLRT detector of differentially encoded packet data $\{b_i\}_{i=1}^{N-1}$ transmitted over an unknown linear channel in the presence of complex Gaussian disturbance of unknown covariance is given by
\[
\hat{d}_{GLRT} = \arg \max_{d_i \in \{0, 1\}} l_2(d),
\]
(27)

\[
\hat{b}_{GLRT} = \hat{d}_{1-GLRT} \hat{a}_{GLRT}, \quad i = 1, 2, \ldots, N - 1,
\]
(28)
where $d \triangleq [d_0, \ldots, d_{N-1}]^\tau$.

**Proof:** It suffices to observe that the one-to-one mapping of information bits to differentially encoded bits implies that maximization of the generalized likelihood function with respect to the information bits is equivalent to maximization with respect to the differentially encoded (transmitted) bits.

Using Theorem 3 and the observation that both a single-bit error and an $(N - 1)$-bit error in $d$ results in a 2-bit error in the bit sequence $\{b_i\}_{i=1}^{N-1}$, we can approximate the BER performance of the DPSK GLRT detector and obtain the packet size that is necessary for the detector to achieve a certain BER performance level as follows.

**Corollary 2:** (i) The average BER of the DPSK GLRT detector that operates on a data packet of size $N \geq L - 3$ is given by
\[
BER_{GLRT-DPSK}(N) \approx BER_{SMI,DPSK}(N - 2)
\]
\[
\approx \frac{2}{\pi} \int_0^{\pi/2} \frac{M}{M - L, N - 1, -\frac{(N - 1)\gamma}{N \sin^2 \theta}} d\theta
\]
(29)

\[
\approx \frac{\gamma}{3} Q\left(\sqrt{2\mu}\right) + \frac{1}{3} Q\left(\sqrt{2\mu + 2\sqrt{3}\sigma}\right)
\]
additionally,
\[
\frac{1}{3} Q\left(\sqrt{2\mu - 2\sqrt{3}\sigma}\right),
\]
(30)
where
\[
\mu \triangleq \frac{N - L - \gamma}{N}, \quad \sigma^2 \triangleq \frac{(N - L)(L - 1)\gamma^2}{N^3},
\]
and $BER_{SMI,DPSK}(N - 2)$ is the BER of a detection scheme that would utilize the coherent SMI detector of $\{d_i\}_{i=1}^{N}$ built on $K = N - 2$ independent pure disturbance observations under perfect knowledge of $a$, followed by the differential decoder $b_i = \hat{d}_{i-1} \hat{d}_i$, $i = 1, 2, \ldots, N - 1$.

(ii) For any given $\nu$, the smallest packet size $N_\nu$ that guarantees that the BER performance of the DPSK GLRT packet-data detector is within $\nu$ dB from the BER performance of the optimum ML DPSK detector (i.e., $BER_{GLRT-DPSK}(N) \leq 2Q(\sqrt{2\gamma 10^{-\nu/10}})$) is given by the ceiling of the maximum real root of the cubic equation
\[
N^3 - \frac{2L}{1 - 10^{-\frac{\nu}{10}}} N^2 + \frac{L^2 - 3(L - 1)}{(1 - 10^{-\frac{\nu}{10}})^2} N + \frac{3L(L - 1)}{(1 - 10^{-\frac{\nu}{10}})^2} = 0.
\]
(31)

In the following section, simulation studies demonstrate that the proposed DPSK GLRT detector that combines estimation of the unknown parameters and packet-data detection into
one process outperforms the common estimate-and-plug-in approach where we first take the optimum $N$-symbol (or the popular suboptimum 2-symbol) block differential detector formula and then substitute therein unknown statistics and channel coefficients by estimates obtained separately.

V. SIMULATION STUDIES

We prepare a communication system simulation study where packet-data are received in the presence of Gaussian noise of unknown covariance. The covariance matrix used to generate received data is taken directly from the literature [28]. The dimension of the received data vectors is $L = 9$. The eigenvalues of this covariance matrix are 4.26, 1.83, 0.66, 0.49, 0.46, 0.45, 0.33, 0.31, and 0.21, respectively. We note that, as in case of most real systems, in the above correlation structure there is no distinct white noise subspace that possesses identical eigenvalues. The channel processed signal waveform $\mathbf{Sa}$ is chosen arbitrarily and is assumed to be known (known channel case). We prepare this hypothetical communication system study primarily to examine the accuracy of our analytical BER expression in (14) when the disturbance is exactly colored Gaussian distributed with unknown covariance. The performance of the proposed packet-data GLRT detectors operating in other communication environments will be examined later in this section. We study the BER performance of the proposed GLRT packet-data detector as a function of the data packet size $N$. The GLRT detector is implemented in its linear cost form as presented in Section III.B with $P = 16$, $D = 6N$, and arbitrary initial bit estimates.

Fig. 1 presents our findings and comparisons with SMI-dsp (desired-signal-present) in (10), SMI (disturbance only observations) in (9), and ideal ML detection. In Fig. 1 we also include the performance curve of the constraint-LMS where the update step size is set equal to $10^{-3}$ [29], [30], as well as the performance curve of the RLS algorithm for which the initialization parameter and the forgetting ratio are set equal to 100 and 1, respectively. The training data record size for both the constraint-LMS and RLS algorithms is chosen to be equal to the packet size $N$. In view of the nearly overlapping analytical and simulated GLRT BER curves, we may claim that our linear cost GLRT implementation performs very close to full GLRT and (14) provides an accurate approximation of the BER of the GLRT packet-data detector. Furthermore, the GLRT packet-data detector outperforms significantly the SMI-dsp, constraint-LMS, and RLS detectors and performs nearly the same as the SMI detector in (9) that requires $N - 1$ additional pure disturbance observations.

In the rest of this section, we use as an illustration vehicle for the proposed GLRT schemes a packet-data DS-CDMA communication system$^4$. At all times, the GLRT detectors are implemented via the linear complexity algorithm of Section III.B with $P = 16$ and $D = 6N$. Initial bit estimates are taken either by conventional matched-filter (MF) outputs (Case-study #1) or are arbitrarily set (Case-studies #2 and #3).

$^4$The combined effect of DS-CDMA multiple access interference (MAI) and AWGN is Gaussian-mixture distributed and not plain Gaussian. It is interesting to examine how the newly developed GLRT detectors perform in such an environment.

DS-CDMA Case-study #1 Synchronous multiuser system and single-path channel

We consider a system with 10 synchronous users with Gold signatures of length $G = 31$. We select a “user of interest” and have the SNR’s of the interfering users fixed in the range $[0dB, 11dB]$. In this study we assume exact knowledge of the channel of the user of interest. We compare the BER of the GLRT detector with the BER of the MF, SMI-dsp, constraint-LMS (step size $10^{-4}$), RLS (initialization parameter 100), and SMI detector in (9) that assumes availability of $N - 1$ additional pure disturbance observations.

In Figs. 2 and 3, we plot the BER as a function of the SNR of the user of interest ($N = 127$).
SNR region, the pairwise probability of error (11) with more than 1 bit error is negligible. Thus, based on (34), (37), and (46) (which hold regardless of the disturbance distribution), we conclude that the BER of the GLRT packet-data detector is approximately equal to the BER of the SMI detector in (9). In addition, it was observed in [19], that the distribution of the output SINR of the SMI detector (9) operating in a non-Gaussian disturbance environment can be very well approximated by the Beta distribution; hence the analytical result in (14) continues to provide accurate BER evaluation.

**DS-CDMA Case-study #2** Asynchronous multipath fading channel: Pilot-assisted signaling

We consider the same setup as in Case-study #1, except that now the users transmit asynchronously and each user channel has 3 resolvable paths. The path coefficients are modeled as independent complex Gaussian random variables all of unit variance. The length of the pilot sequence is fixed at $J = 10$. We compare the BER of the GLRT detector with the BER of the RAKE-MF, the SMI and SMI-dsp detectors in (9) and (10), and the constraint-LMS and RLS detectors. We note that in this study the GLRT detector assumes no knowledge of the channel while all other detectors assume exact knowledge of the channel. In addition, the SMI detector in (9) uses $N - 2$ extra pure disturbance observations that are assumed to be available. It is worth noting that the pilot sequence is incorporated and processed internally and elegantly by the GLRT algorithm without the need for a separate phase estimation stage. Our simulation findings given in Figs. 4 and 5 are self-explanatory and make a strong case in favor of the new GLRT developments.

**DS-CDMA Case-study #3** Asynchronous multipath fading channel: DPSK signaling

We consider the same setup as in Case-study #2, except that the transmitter now uses DPSK encoding instead of pilot signaling; hence, at the receiver end a differential decoder is needed to recover the information bits. We compare the BER of our DPSK GLRT detector with the BER of the DPSK version of the following detectors: RAKE-MF, SMI-dsp, constraint-LMS, RLS, and ideal MMSE. A 2-symbol differential decoder is used in all cases. The coherent SMI detector that is described in Corollary 2, Part (i) is also included as a reference. We note that the DPSK GLRT detector assumes no knowledge of the channel while the RAKE-MF, constraint-LMS, RLS, SMI-dsp, and ideal MMSE detectors assume perfectly known path coefficients up to an unknown phase (phase ambiguity is resolved by differential encoding/decoding). In addition, the ideal MMSE detector assumes perfectly known interference-plus-noise covariance matrix and the coherent SMI detector in Corollary 2, Part (i) requires perfect knowledge of the path coefficients (including the phase) and $N - 2$ additional pure disturbance observations. In Figs. 6 and 7 we repeat the studies of Figs. 4 and 5. The superiority of the new DPSK GLRT detector is striking. In fact, for data packets of size $N = 250$ and higher the DPSK GLRT detector outperforms even the ideal MMSE (2-symbol decoder) detector.
The SNR of the user of interest is fixed at 9 dB.

**VI. CONCLUSIONS**

We considered the problem of packet-data detection for general multiaccess/multiserver digital communication systems. We proposed novel GLRT-type detection schemes that perform joint estimation of the unknown system parameters and detection of the packet-data and we designed suboptimum implementations of linear complexity in the packet size. In particular, for the known channel case we developed a coherent GLRT detector, while for the unknown channel case we developed a pilot assisted GLRT detector (the pilot signal is implicitly used to resolve phase ambiguity) and a DPSK GLRT detector. We established analytically the performance of each proposed GLRT-type scheme relative to the corresponding conventional estimate-and-plug-in detector that replaces unknown parameters in its ideal formula by estimates obtained through a separate estimation process. In all cases, the GLRT schemes maintain the same elegant core structure regardless of known or unknown channels and pilot or DPSK signaling. Finally, we developed analytical performance evaluation tools that provide the BER that can be reached by each proposed GLRT scheme for any given finite data record size, as well as the data record size that is necessary for each GLRT detector to perform within a certain neighborhood of the optimal performance point (without the need to know the latter).

**APPENDIX I**

**PROOF OF THEOREM 1**

(i) We can write $Y = v b^T + N$ where $N \overset{\Delta}{=} [n_1, \ldots, n_N] \sim \mathcal{N}(0, I_N \otimes R_n)$. W.l.o.g. assume $b = [1, \ldots, 1]^T$ and $b = [-1, \ldots, -1, 1, \ldots, 1]^T$. By (5) and (7), $l_1(b) > l_2(b)$ is equivalent to

$$\left| N N^H + 2v \left( \sum_{i=1}^m n_i \right)^H + 2 \left( \sum_{i=1}^m n_i \right) v^H + 4mvv^H \right| < |NN^H| \quad (32)$$

where $| \cdot |$ denotes the matrix determinant calculation. Let $U \overset{\Delta}{=} [u_1, \ldots, u_N]$ be a unitary matrix and define $N^{'} \overset{\Delta}{=} NU = [n_1^{\prime}, N^{\prime}]$ where $n_1^{\prime} = Nu_1$ and $u_1 = [1, \ldots, 1, 0, \ldots, 0]^T/\sqrt{m}$. Then, $N^{\prime} \sim \mathcal{N}(0, I_N \otimes R_n)$, $n_1^{\prime} = \frac{1}{\sqrt{m}} \sum_{i=1}^m n_i \sim \mathcal{N}(0, R_n)$, $N^{\prime} \sim \mathcal{N}(0, I_{N-1} \otimes R_n)$, and $N N^H = N^{\prime} N^{\prime H} = N^{\prime} N^{\prime H} + n_1^{\prime} n_1^{\prime H}$. Thus, (32) can be reduced to

$$\left| R_{n^{\prime} n^{\prime}} + \left( n_1^{\prime} + 2\sqrt{mv} \right) \left( n_1^{\prime} + 2\sqrt{mv} \right)^H \right| < \left| R_{n^{\prime} n^{\prime}} + n_1^{\prime H} \right| \quad (33)$$

where $R_{n^{\prime} n^{\prime}} \overset{\Delta}{=} N^{\prime} N^{\prime H} \sim \mathcal{C}W_L(N-1, R_n)$ is complex Wishart distributed with $N-1$ degrees of freedom and independent of $n_1^{\prime}$ [32]. Using identity (8), we simplify (33) to

$$\text{Re} \left( v^H R_n^{-1} n_1^{\prime} \right) < -\sqrt{m} \left( v^H R_n^{-1} v \right). \quad (34)$$

Given $R_{n^{\prime} n^{\prime}}$, $v^H R_n^{-1} n_1^{\prime}$ is circular complex Gaussian distributed with zero mean and variance $v^H R_n^{-1} R_n^{-1} v$. Thus, given $R_{n^{\prime} n^{\prime}}$ and $b$

$$P \left( l_1(b) > l_2(b) \mid R_{n^{\prime} n^{\prime}}, b \right) = Q \left( \frac{2m \left( v^H R_n^{-1} v \right)^2}{v^H R_n^{-1} R_n^{-1} v} \right). \quad (35)$$

Using a similar reasoning as in [15], $\eta = \frac{\left( v^H R_n^{-1} v \right)^2}{\gamma} \frac{\gamma}{B(a, b)} \frac{B(a, b)}{\gamma}$ has a Beta distribution $f_\eta(x) = \frac{x^a - 1}{B(a, b)} \frac{B(a, b)}{\gamma}$, $0 \leq x \leq 1$, where $a = N - L + 1$, $b = L - 1$, and $B(a, b)$ is the Beta function [21]. Then, (11) is just the expected value of (35) with respect to $R_{n^{\prime} n^{\prime}}$ (or equivalently $\eta$).

(ii) For $m \geq 2$ we can show that

$$\lim_{\gamma \to \infty} \int_0^1 Q \left( \frac{v^H R_n^{-1} v}{\gamma \sqrt{2m}} \right) f_\eta(x) dx \bigg/ \int_0^1 Q \left( \frac{v^H R_n^{-1} v}{\gamma \sqrt{2m}} \right) f_\eta(x) dx = \lim_{\gamma \to \infty} \frac{\sqrt{m}^{\Gamma(a+b+0.5)}}{\Gamma(b)} (\gamma)^{-a-0.5} \frac{1 + O(|\gamma|^{-1})}{1 + O(|\gamma|^{-1})} \quad (36)$$

$$= m^{-(N-L+1.5)}$$
where $\Gamma(\cdot)$ is the complete Gamma function. The sequence of operations that lead to this result in (36) includes applications of L'Hospital’s rule, use of the fact that the moment generating function of a Beta distributed variable with parameters $a$ and $b$ is a Kummer’s confluent hypergeometric function [21], i.e.,

$$
\int_0^\infty e^{xy} f_P(x) dx = M(a, a + b, s),
$$

and use of the asymptotic expression of $M(a, a + b, s)$ [21]. Expression (36) implies that the pairwise probability of error in (11) is negligible in the high received SNR region for $\nu < \gamma$. Thus, the BER of the detector in (9) is lower bounded by the BER of the following: [NOVEL GLRT PACKET-DATA RECEIVERS 231]

$$
\hat{b}_N = \arg \max_{b_N} \left\{ Q \left( \frac{\sqrt{2} Q^2}{N} \right) f_\eta(x) dx \right\}
$$

Following similar calculations used to obtain (7) and then applying (8), we have

$$
\hat{b}_N = \arg \max_{b_N} \left\{ \text{Re} \left( e^{\phi(b_N)} \right) \right\}
$$

where $R_{nSA}(N - 1)$ is the sample average noise correlation matrix evaluated using the first $N - 1$ noise components. We see that (45) is the detector in (9) with $K = N - 1$. Hence, the BER of the detection scheme in (44) is equal to $BER_{SMI}(N - 1)$, and

$$
BER_{GLRT}(N) \geq BER_{SMI}(N - 1).
$$

If we combine (41), (43) and (46), then (12) follows.

**APPENDIX II**

**Proof of Theorem 2**

(i) Expression (13) can be proved by virtue of Theorem 1, expression (5.3) of [33], and the fact that the moment generating function of a Beta distributed random variable is a Kummer’s confluent hypergeometric function [21]. Alternatively, we can express $BER_{GLRT}(N)$ as $E \{ g(\theta) \}$ for $g(\theta) \triangleq Q(\sqrt{2} h)$, $h \triangleq \eta \gamma$ and $\eta$ a Beta distributed random variable with $a = N - L + 1$ and $b = L - 1$ [21], and then approximate by $\frac{1}{2} g(\mu) + \frac{1}{2} g(\mu + \sqrt{3}) + \frac{1}{2} g(\mu - \sqrt{3})$ [34] for $\mu \triangleq \frac{N - L + 1}{N \gamma}$, $\sigma^2 \triangleq \frac{(N - L + 1)(L - 1)}{N^2(N + 1)} \gamma^2$, and any

$$
\frac{N - L + 1}{N} \geq \frac{3(N - L + 1)(L - 1)}{N^2(N + 1)}.
$$

The latter inequality holds for all $L$ and $N > L + 2$ which also satisfies the condition for the existence of the inverse of $R_{nSA}(N - 1)$ with probability 1.

(ii) Let $h(N) \triangleq \frac{N - L + 1}{N} - \frac{3(N - L + 1)(L - 1)}{N^2(N + 1)}$. The monotonicity of $Q(x)$ and (14) imply that

$$
\frac{1}{2} Q \left( \frac{\sqrt{2} h}{N} \right) < BER_{GLRT}(N) < Q \left( \frac{\sqrt{2} h}{N} \right).
$$

Thus, $Q \left( \frac{\sqrt{2} h}{N} \right)$ is an asymptotically tight upper bound on $BER_{GLRT}(N)$. The smallest $N_v$ that guarantees $BER_{GLRT}(N) \leq Q \left( \frac{\sqrt{2} h}{N} \right)$ for all $\gamma$ is the ceiling of the solution of the equation $h(N) = 10^{-\nu/10}$ which can be found as the maximum real root of (15).

---

5We recall that $N - 1 \geq L$ is the necessary condition to guarantee the existence of the inverse of $R_{nSA}(N - 1)$ with probability 1.

---

6The approximation in [34] uses the Stirling formula to replace the Taylor expansion of $g(\theta)$ around the mean value of $\theta$ and then takes the expectation with respect to the random variable $\theta$. 
\[
(YY^H)^{-1} = A^{-1} - A^{-1}\gamma (D^{-1} + \gamma^H A^{-1}\gamma)^{-1}\gamma^H A^{-1},
\]

\[
[S^H (YY^H)^{-1} S]^{-1} = B^{-1} + B^{-1} Z \left[ D^{-1} + \gamma^H A^{-1}\gamma - Z^H B^{-1} Z \right]^{-1} Z^H B^{-1},
\]

\[
D^{-1} - \gamma^H (YY^H)^{-1} \gamma = D^{-1} \left( D^{-1} + \gamma^H A^{-1}\gamma \right)^{-1} D^{-1},
\]

\[
\gamma^H (YY^H)^{-1} S [S^H (YY^H)^{-1} S]^{-1} S^H (YY^H)^{-1} \gamma
\]

\[
= D^{-1} \left[ D^{-1} + \gamma^H A^{-1}\gamma - Z^H B^{-1} Z \right]^{-1} Z^H B^{-1} Z \left[ D^{-1} + \gamma^H A^{-1}\gamma \right]^{-1} D^{-1},
\]

\[
D^{-1} - \gamma^H (YY^H)^{-1} \gamma + \gamma^H (YY^H)^{-1} S [S^H (YY^H)^{-1} S]^{-1} S^H (YY^H)^{-1} \gamma
\]

\[
= D^{-1} \left[ D^{-1} + \gamma^H A^{-1}\gamma - Z^H B^{-1} Z \right]^{-1} D^{-1}.
\]

**APPENDIX III**

**Proof of Theorem 3**

W.l.o.g. assume \( b = \{1, \ldots, 1\} \) and \( \hat{b} = \)
\[
[-1, \ldots, -1, 1, \ldots, 1].
\]

Let \( \gamma_1 \triangleq \frac{1}{m} \sum_{i=1}^{m} \gamma_i, \quad \gamma_2 \triangleq \)
\[
\frac{1}{N-m} \sum_{i=m+1}^{N} \gamma_i, \quad A_1 \triangleq \sum_{i=1}^{m} (\gamma_i - \gamma_1)(\gamma_i - \gamma_1)^H,
\]

\[
A_2 \triangleq \sum_{i=m+1}^{N} (\gamma_i - \gamma_2)(\gamma_i - \gamma_2)^H, \quad \text{and} \quad A = A_1 + A_2.
\]

Then \( \gamma_1 \sim N(\mu, R_n/m) \) is independent of \( A_1 \) and \( \gamma_2 \sim N(\mu, R_n/(N-m)) \) is independent of \( A_2 \) [32]. We also note that \( \gamma_1 \) is independent of \( A_2 \) and \( \gamma_2 \) is independent of \( A_1 \). Thus, \( \gamma \triangleq [\gamma_1, \gamma_2] \) is independent of \( A \), which has a complex Wishart distribution with \( N-2 \) degrees of freedom [32]. Set \( D \triangleq \text{diag}(m, N-m) \). Then \( YY^H = A + m\gamma\gamma^H + (N-m)\gamma_2\gamma_2^H = A + \gamma \gamma^H \).

Through straightforward–yet tedious–matrix algebra [31], we have (48), (49), (50), (51), and (52), where \( Z \triangleq S^H A^{-1}\gamma \) and \( B \triangleq S^H A^{-1} S \) are introduced for notation simplicity. Using (48)-(52), inequality \( l_2(b) > l_2(b) \) can be reduced to

\[
\begin{align*}
& \begin{bmatrix} 1, 1 \end{bmatrix} \left[ Z^H B^{-1} Z \right]^{-[1, 1]^T} [1, 1]^{T} \\
& \begin{bmatrix} 1, 1 \end{bmatrix} \left[ D^{-1} + \gamma^H A^{-1}\gamma \right]^{-[1, 1]^T} [1, 1]^{T} \\
& > \begin{bmatrix} -1, 1 \end{bmatrix} \left[ Z^H B^{-1} Z \right]^{-[-1, 1]^T} [-1, 1]^{T} \\
& \begin{bmatrix} -1, 1 \end{bmatrix} \left[ D^{-1} + \gamma^H A^{-1}\gamma \right]^{-[-1, 1]^T} [-1, 1]^{T}
\end{align*}
\]

(53)

where the positivity of \( [Z^H B^{-1} Z] \) and \( [D^{-1} + \gamma^H A^{-1}\gamma - Z^H B^{-1} Z] \) can be easily verified. Let \( \gamma \triangleq \begin{bmatrix} \gamma_1, \gamma_2 \end{bmatrix} \) and \( \gamma \triangleq \gamma^H A^{-1}\gamma \) and \( \gamma \triangleq Z^H B^{-1} Z \). Using the explicit expression of the inverse of a 2 x 2 matrix, we can simplify (53) to

\[
(\alpha_{12} + \alpha_{21}) (\beta_{11} + \beta_{22}) > (\beta_{12} + \beta_{21}) \frac{1}{m} + \frac{1}{N-m} + \alpha_{11} + \alpha_{22}.
\]

(54)

We note that \( \gamma^H A^{-1}\gamma \) and \( Z^H B^{-1} Z \) is semi-positive definite since it can be written as \( \gamma^H A^{-1/2} \left[ I - A^{-1/2} S [S^H A^{-1/2} A^{-1/2} S]^{-1} S^H A^{-1/2} \right] A^{-1/2} \).

Then (21) follows by [15] and the observation that \( (\gamma^H A^{-1}\gamma)^2 = \gamma \) is a scaled Beta distributed variable with parameters \( a = N-L \) and \( b = L-1 \).

**REFERENCES**


\[ \alpha_{11} + \alpha_{22} = \zeta + z_H^t A^{-1} z_1 + z_H^t A^{-1} z_2, \]  
(55)

\[ \alpha_{12} + \alpha_{21} = \zeta + 2 \text{Re} [z_H^t A^{-1} z_2], \]  
(56)

\[ \beta_{11} + \beta_{22} = \zeta + z_H^t A^{-1} S (S^H A^{-1} S)^{-1} S^H A^{-1} z_1 + z_H^t A^{-1} S (S^H A^{-1} S)^{-1} S^H A^{-1} z_2, \]  
(57)

\[ \beta_{12} + \beta_{21} = \zeta + 2 \text{Re} [z_H^t A^{-1} S (S^H A^{-1} S)^{-1} S^H A^{-1} z_2], \]  
(58)

\[ \zeta = 2 v^H A^{-1} v + 2 \text{Re} [v^H A^{-1} z_1] + 2 \text{Re} [v^H A^{-1} z_2], \]  
(59)


Haoli Qian (M’2004) was born in Sichuan, China in 1977. He received the B.S in Electronic Engineering from Tsinghua University, Beijing, China in 1998, and the Ph.D. degree in Electrical Engineering from the State University of New York at Buffalo in 2004. Dr. Qian is currently a Senior Staff Design Engineer with Marvell Semiconductor Inc., Santa Clara, CA. His research interests are in the areas of signal processing and wireless communications.

Stella N. Batalama (S’91, M’94) received the Diploma degree in computer engineering and science (5-year program) from the University of Patras, Greece in 1989 and the Ph.D. degree in electrical engineering from the University of Virginia, Charlottesville, VA, in 1994.

From 1989 to 1990 she was with the Computer Technology Institute, Patras, Greece. In 1995 she joined the Department of Electrical Engineering, State University of New York at Buffalo, Buffalo, NY, where she is presently a Professor. During the summers of 1997-2002 she was Visiting Faculty in the U.S. Air Force Research Laboratory (AFRL), Rome, NY. From Aug. 2003 to July 2004 she served as the Acting Director of the AFRL Center for Integrated Transmission and Exploitation (CITE), Rome NY.

Her research interests include small-sample-support adaptive filtering and receiver design, adaptive multiuser detection, robust spread-spectrum communications, supervised and unsupervised optimization, distributed detection, sensor networks, covert communications and steganography.

Bruce W. Suter (S’69-M’72-SM’92) received the B.S. and M.S. degrees in electrical engineering in 1972 and the Ph.D. degree in computer science in 1988, all from the University of South Florida, Tampa, FL.

Since 1998, he has been with the Information Directorate of the Air Force Research Laboratory, Rome, NY, where he was the Founding Director of the Center for Integrated Transmission and Exploitation. Dr. Suter has authored over a hundred publications. He is also the author of the book Multirate and Wavelet Signal Processing (Academic Press, 1998). His research interests include multiscale signal and image processing, cross layer optimization, networking of unmanned aerial systems, and wireless communications. His professional background includes industrial experience with Honeywell Inc., St. Petersburg, FL and with Litton Industries, Woodland Hills, CA, and academic experience at the University of Alabama, Birmingham, AL, and the Air Force Institute of Technology, Wright-Patterson AFB, OH.
Abstract—This paper proposes a new peak-to-average power ratio (PAPR) reduction scheme in multiple transmit antenna environments. Instead of applying individual PAPR reduction on each antenna, a joint PAPR reduction technique is proposed. By applying a relevant unitary rotation over the transmit antennas, overall PAPR of the multiple transmit antenna system is reduced. This scheme does not require any side information to decode the signal in the receiver, enabling throughput-lossless PAPR reduction. Furthermore, there is no increase in the complexity of the receiver.

Index Terms—Peak power reduction, PAPR, MIMO systems, OFDM, unitary rotation.

I. INTRODUCTION

MULTIPLE-input multiple-output (MIMO) radio systems have seen a great deal of attention since Telatar [1], and Foschini and Gans [2] showed that exploiting the multiple transmit/receive antenna increases the outage capacity. For high data rate wireless applications, MIMO combined with orthogonal frequency division multiplexing (OFDM) is being considered in a large number of current technology applications. While OFDM has a great advantage of having simple equalization, it has an inherent drawback of high peak-to-average power ratio (PAPR).

Researchers have extensively examined PAPR reduction in single-input single-output (SISO) systems. Clipping [3] is the simplest way to limit the maximum magnitude of transmit signals. Clipping causes distortion resulting in increased bit error rate (BER) and out-of-band spectral radiation. Selective mapping (SLM) [4] and partial transmit sequence (PTS) [5] transmit the lowest peak power signal among several candidates generated using different data phases. These methods, however, require redundant bits to decode the information bits in the receiver, which results in a spectral efficiency decrease. PAPR reduction using coding [6], [7] was proposed, but it also sacrifices the data rate. Tone reservation [8], [9] reduces the PAPR using a small set of pre-allocated tones, that wastes the bandwidth. Nonbijective mapping [9], [10], [11] changes the constellations to combat large signal peaks. It does not need any redundant information, but results in increased average power and high complexity (some trellis shaping designs [11] reduce the average power as well). All aforementioned schemes involve some compromises between bandwidth, peak power and complexity. So far none of the SISO PAPR reduction schemes dominates the others.

PAPR reduction in MIMO transmission has rarely been considered yet. If the PAPR reduction schemes in SISO are directly applied to the MIMO, the complexity and redundancy increase proportional to the number of transmit antennas. Two approaches have been taken to solve the PAPR problem in MIMO case. In [12], the MIMO-SLM is applied through overall transmit antennas, i.e., the sequence with the lowest PAPR over all transmit antennas is selected instead of applying individual SLM to each transmit antenna. Using the MIMO-SLM, redundancy can be reduced or the reliability of the side information can be increased at the sacrifice of reduced PAPR reduction gain. In [13], the reduced PAPR signals are generated by scrambling the multiple transmit signals over the antennas (which is called cross-antenna rotation and inversion (CARI)) at the portion of subcarriers. Both MIMO-SLM and CARI, however, require the side information and multiple IFFT operations.

In this paper a new paradigm of PAPR reduction in multiple transmit antenna environments is proposed. By applying a unitary rotation through transmit antennas, the maximum peak power over all transmit antennas can be decreased. The unitary rotation plays a role in dispersing the maximum power in one transmit antenna into the other transmit antennas. This scheme is not limited to an OFDM system, but it can be combined with any system using multiple transmit antennas. The unitary rotation can be made transparent at the receiver so that no loss in bandwidth is needed to specify which transformation was used at the transmitter. Hence this technique offers reduced PAPR without any loss in bandwidth efficiency. Since it is more advantageous with short frame length, the PAPR reduction gain is evaluated with short packet structures in the paper. The influence of the frame length on PAPR reduction is investigated subsequently.

The rest of the paper is organized as follows. Section II defines the system model and shows the example packet structures used throughout the paper, and then introduces the new PAPR reduction idea, which is called a unitary PAPR reduction. Section III presents an optimization problem formulation for finding the unitary matrix and addresses the nonconvexity of the problem. The algorithms to find a sub-optimal unitary matrix for 2 transmit antennas are presented in Section IV and extended to the general number of transmit antenna case in Section V. Simulation results are provided in Section VI. Section VII confirms that beamforming effect due to the unitary rotation does not affect the outage capacity and performance for open loop coding techniques. Finally, Section VIII gives concluding remarks.


H. Lee was with the Department of Electrical Engineering, University of California Los Angeles. He is now with Qualcomm Incorporated (e-mail: heechoon@gmail.com).

M. P. Fitz was with the Department of Electrical Engineering, University of California Los Angeles. He is now with Northrop Grumman Corporation (e-mail: michael.fitz@ngc.com).

Digital Object Identifier 10.1109/TCOMM.2008.050395.
II. UNITARY PEAK POWER REDUCTION

A. System Model

The MIMO OFDM system with $L_t$ transmit and $L_r$ receive antennas and $D$ subcarriers is considered. $D_p$ payload subcarriers out of the $D$ subcarriers are used for pilot and data transmission. For all the examples considered in the paper $D = 64$ and $D_p = 52$ are considered, as is often the case with IEEE 802.11x systems. The OFDM signal transmitted from the $l$th transmit antenna is given as

$$X_i(t) = \sum_{k=1}^{N_p} \sum_{n=0}^{N-1} X_i(l,k) u(l, t - (k - 1)T_s)$$

where $X_i(l,k)$ is the symbol transmitted at the $l$th subcarrier in the $k$th symbol time; $u(l, t - (k - 1)T_s)$ is the subcarrier pulse shape for the $l$th subcarrier and $k$th symbol; $T_s$ is the OFDM symbol time; $N$ is the frame length in symbol. It should be noted that the subcarrier pulse is typically a time-limited pulse of a complex sinusoid. The PAPR of the MIMO OFDM system can be defined as

$$\text{PAPR} = \frac{\max_{t \in [1:T_s]} |X_i(t)|^2}{E_{i,t}[|X_i(t)|^2]}$$

where $E_{i,t}[\cdot]$ denotes the expectation (over antennas and time).

There is a relatively simple model for the output of the MIMO OFDM matched filters. Assuming that a cyclic prefix of an appropriate length is used, then the matched filter outputs for the $m$th symbol time and the $n$th subcarrier are given as an $L_r \times 1$ vector

$$\tilde{Y}(m,n) = \mathbf{H}(m,n) \sqrt{E_s} \tilde{X}(m,n) + \tilde{N}(m,n)$$

where $E_s$ is the energy per transmitted symbol; $H_{ji}(m,n)$ is the complex path gain from transmit antenna $i$ to receive antenna $j$ at time $mT_s$ on subcarrier $n$; $\tilde{X}(m,n)$ is the $I_t \times 1$ vector of symbols transmitted at symbol time $m$ and subcarrier $n$; $\tilde{N}(m,n)$ is the additive white Gaussian noise vector of dimension $L_r \times 1$. The noise samples are independent samples of circularly symmetric zero-mean complex Gaussian random variables with variance $N_0/2$ per dimension.

B. Packet Structure

Channel state information (CSI) is required for the OFDM receiver to perform coherent detection, or diversity combining in MIMO environments. In practice, a channel estimate is acquired in the receiver using known pilot symbols [14], [15]. The channel estimates are obtained by the interpolation from the observations at the pilot symbol locations. Performance with finite samples in time and frequency is aided by having more pilot symbols at the edges of the packet [16]. The simplest and most efficient way to estimate the CSI in a MIMO environment is to use the orthogonal modulation on each of the transmit antennas [15]. With this orthogonal pilot modulation linear filters work well as channel estimators. Therefore, the Wiener filter coefficients can be pre-computed and the resulting channel estimator is an open loop structure which has no acquisition time [17].

Short packets are very important for MAC layer communication; RTS, CTS and TCP/IP ACKs are examples of short packets. Even though they include little information compared to data packet, their reliability is critical to the overall performance of network. In this paper, a set of example short packet structures for the MIMO OFDM system containing pilot symbols and space-time block codes (STBCs) are employed that can use the unitary PAPR reduction technique. The STBC of $L_t$ transmit antennas and $N_f$ time duration is considered [18]. The packet length of the presented short packet structure is $N_f$ OFDM symbols, i.e., one STBC block per subcarrier composes one packet. Simple block codes are used to capture the essence of coded MIMO OFDM modulation without overly complicating the system so to keep the focus on the PAPR techniques but in a realistic environment.

The detailed packet structures are described in [19] for $L_t = N_f = 2$ and $L_t = N_f = 4$. Since channel estimation takes a larger percentage of bandwidth in a short packet, data and pilot are mixed together instead of preamble-based structure. However, it should be emphasized that unitary PAPR reduction can also be used with a preamble-based long packet structure. The applicability to the long packet structure is explored in Section VI. Based on the structure in [19], randomized pilots are designed to exhibit low PAPR so that the PAPR of transmit signal is not dominated by pilots. These particular pilot symbol placements in the frequency-time grid enables the system to track a frequency roll across the short packet. The work in the paper is compliant to IEEE 802.11x systems, so 52 effective subcarriers are used out of 64 subcarriers. As per the 802.11a standard the DC subcarrier is not used.

C. Unitary Rotation

Since all the existing SISO PAPR reduction schemes involve some drawbacks, it is very attractive to reduce PAPR over all transmit antennas jointly, in place of separate SISO reduction on each transmit antenna. Most PAPR reduction schemes in SISO systems try to eliminate the occurrences of the peak by controlling the frequency domain signals before the IFFT. Since they control the frequency domain signals, the IFFT processing is required during PAPR reduction algorithms.

The idea of unitary PAPR reduction exploits the characteristics of the multiple antennas. Applying a relevant unitary rotation to the multiple streams of the time domain signals, the peak power in one transmit antenna can be dispersed into signals in the other transmit antennas. That is, instead of avoiding peak generation, unitary PAPR reduction trims down the peaks by spreading out the energy over all antennas at each time instance. As a result, the peak power over all the transmit antennas can be reduced. In the operation of the MIMO OFDM system this unitary rotation can be absorbed into an effective channel. For example if $U$ represents the unitary rotation then the received signal is given as

$$\tilde{Y}(m,n) = \mathbf{H}(m,n) U \sqrt{E_s} \tilde{X}(m,n) + \tilde{N}(m,n)$$

where the effective channel matrix is denoted by $\tilde{\mathbf{H}}(m,n) \equiv \mathbf{H}(m,n) U$. The effective channel matrix can be retrieved from a standard channel estimator, if the unitary matrix is constant within the duration of channel estimation.
It should be noted that there is a tradeoff between achieving low PAPR and achieving good channel estimation. The best PAPR would be achieved by applying a unitary PAPR reduction to each OFDM symbol. In that case, the channel estimator should work at every symbol time independently, since the effective channel experiences the abrupt change at each symbol time. On the contrary, channel estimation can be improved by smoothing over multiple symbols, but the PAPR reduction gain will be diminished since the unitary matrix needs to be constant during the period of channel estimation. In addition, if an STBC is used, the unitary matrix also should be constant for the duration of space-time block length if the low complexity decoders for block codes are to be used.

In order to reduce the continuous time PAPR, oversampling should be used when PAPR reduction is performed [20]. In this paper, oversampling rate $L = 5$ is used. The resulting packet length is $LN_F(D + D_g)$ samples per each transmit antenna, where $D_g = 16$ is the cyclic prefix length. Note that the cyclic prefix does not affect the PAPR performance, because it is just a cyclic extension of original signal. Considering the oversampling factor in discrete domain, the optimal unitary matrix for PAPR reduction can be written as

\begin{equation}
U_{o} = \arg \min_{U} \left[ \max \{U\mathbf{X}\} \right] = \arg \min_{U} \left[ \max \{U\mathbf{X}\} \right] \tag{5}
\end{equation}

where $\mathbf{X}$ represents an $L_t \times LN_f(D + D_g)$ matrix of the transmitted packet signals and $\max \{U\mathbf{X}\}$ denotes the maximum magnitude element of matrix $U\mathbf{X}$. That is, the unitary PAPR reduction reduces the peak power using the unitary matrix that minimizes the maximum peak power over all transmit antennas through the packet.

The block diagram of the STBC MIMO OFDM system adopting the unitary PAPR reduction is depicted in Fig. 1. The unitary PAPR reduction is located just before the D/A converter and high power amplifier (HPA) in the MIMO OFDM system. By reducing peak power using the unitary rotation, the distortion due to the nonlinearity of HPA can be decreased. As shown in Fig. 1, the unitary PAPR reduction does not require any additional processing in the receiver. The original channel estimator can retrieve the channel information (the effective channel matrix) including the unitary matrix generated in the transmitter. Therefore, this scheme does not have any redundant information to transmit, resulting in throughput-lossless PAPR reduction. Furthermore, it can be combined with any multiple antenna systems regardless of modulation or coding structure. An algorithm to find the optimal unitary matrix is a tough optimization problem. The techniques to find out a sub-optimal unitary matrix are provided in Section IV.

### III. NONCONVEX OPTIMIZATION

To make a unitary PAPR reduction algorithm feasible, the optimization problem for selecting the unitary matrix should be solved. Optimization in (5) can be rewritten as

\begin{equation}
\text{Minimize} \quad \max_{n \in [0, LN_F, D-1]} \| U\bar{x}_n \|_\infty \\
\text{Subject to} \quad U^H U = I \tag{6}
\end{equation}

where $\bar{x}_n$ is the transmit signal vector at time instance $n$; $U$ is a variable to optimize; superscript $H$ denotes a complex conjugate transpose. In this paper the problem for 2 transmit antennas is dealt with first, and then the results will be extended to a general number of transmit antennas in Section V.

Any $2 \times 2$ unitary matrix $V$ can be represented using a special unitary matrix $\mathbf{W}$ (SU(2) [21]: $2 \times 2$ unitary matrix with determinant 1) as

\begin{equation}
\mathbf{V} = \begin{bmatrix} 1 & 0 \\ 0 & z \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1/z \end{bmatrix} \mathbf{V} = \mathbf{w} \\
\text{where} \quad z = \text{det}(\mathbf{V}) = e^{j\theta}, |\alpha|^2 + |\beta|^2 = 1 \quad \text{and} \quad \text{superscript} * \text{denotes a complex conjugate. The general unitary matrix is the product of a phase rotating matrix (}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\text{) and special unitary matrix (}\begin{bmatrix} \alpha & \beta \\ -\beta^* & \alpha^* \end{bmatrix}\text{). Therefore, the amplitude of a unitary matrix rotated signal is equivalent to that of a unitary rotated signal given that the difference exists only in the phase due to $z$. Since the PAPR is the concern, a special unitary matrix is only considered for the $2 \times 2$ unitary matrix from now on.}

Exploiting the characteristics that the phase is independent of the PAPR, the degrees of freedom in unitary matrix can be further reduced. Given that $\alpha \triangleq A_{\alpha}e^{j\theta_{\alpha}}$ and $\beta \triangleq A_{\beta}e^{j\theta_{\beta}}$, the special unitary matrix $\mathbf{W}$ can be decomposed as

\begin{equation}
\mathbf{W} = \begin{bmatrix} \alpha & \beta \\ -\beta^* & \alpha^* \end{bmatrix} = \begin{bmatrix} A_{\alpha}e^{j\theta_{\alpha}} & A_{\beta}e^{j\theta_{\beta}} \\ -A_{\beta}e^{-j\theta_{\beta}} & A_{\alpha}e^{-j\theta_{\alpha}} \end{bmatrix} \\
\text{where} \quad A_{\alpha} = \sqrt{z} \quad \text{and} \quad A_{\beta} = \frac{1}{\sqrt{z}} \\
(8)
\end{equation}

Again the phase rotating matrix (\begin{bmatrix} e^{j\theta_{\alpha}} & 0 \\ 0 & e^{-j\theta_{\alpha}} \end{bmatrix}) does not affect the PAPR. Hence, $\alpha$ in (8) can be assumed to be real and $\beta$ to be complex. Therefore, the unitary matrix $\mathbf{U}$ in (6) can be constrained as

\begin{equation}
\mathbf{U} = \begin{bmatrix} \alpha & \beta \\ -\beta^* & \alpha \end{bmatrix} \tag{9}
\end{equation}

for 2 transmit antennas without loss of generality. Assuming that

\begin{equation}
\bar{x}_n \triangleq \begin{bmatrix} a_n \\ b_n \end{bmatrix}
\end{equation}

Fig. 1. Overall block diagram of STBC MIMO OFDM using unitary PAPR reduction.
the problem (6) can be equivalently reformulated as

\[
\begin{align*}
\text{Minimize} & \quad \max_{n \in [0, LN_f D - 1]} \| X_n \tilde{u} \|_\infty \\
\text{Subject to} & \quad \tilde{u}^H \tilde{u} = 1
\end{align*}
\]

(10)

where

\[
X_n = \begin{bmatrix} a_n & b_n \\ b_n & -a_n^* \end{bmatrix} \quad \text{and} \quad \tilde{u} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}.
\]

Then, the complex problem (10) is transformed into real problem as

\[
\begin{align*}
\text{Minimize} & \quad t \\
\text{Subject to} & \quad \tilde{s}^T A_n^T A_n \tilde{s} \leq t \\
& \quad \tilde{s}^T B_n^T B_n \tilde{s} \leq t \\
& \quad \tilde{s}^T \tilde{s} = 1
\end{align*}
\]

(11)

where

\[
\tilde{s} \triangleq \begin{bmatrix} \alpha \\ \beta \end{bmatrix},
\]

\[
A_n \triangleq \begin{bmatrix} \text{Re}(a_n) & \text{Re}(b_n) \\ \text{Im}(a_n) & \text{Im}(b_n) \end{bmatrix},
\]

\[
B_n \triangleq \begin{bmatrix} \text{Re}(b_n) & -\text{Re}(a_n) \\ \text{Im}(b_n) & -\text{Im}(a_n) \end{bmatrix},
\]

and superscript \( T \) denotes a transpose. In (11) variables are \( t \) and \( \tilde{s} \).

Owing to the quadratic equality constraint \( \tilde{s}^T \tilde{s} = 1 \), the optimization problem in (11) cannot be formulated as a quadratically-constrained quadratic programming (QCQP) [22]. Instead the problem to find out an optimal unitary matrix is shown to be nonconvex and cannot be solved via a global nonconvex solver such as the one provided by YALMIP [23]. YALMIP uses a spatial branch and bound technique [24] based on linear relaxations of bilinear terms to solve this nonconvex problem (11). Unfortunately the complexity of this technique is too high for real-time implementation and other alternatives should be explored. Semidefinite programming (SDP) relaxation of this nonconvex QCQP problem is also tried [25], but it is shown that the relaxation fails to obtain any feasible solution. Since the optimization problem to determine the unitary matrix is shown to be nonconvex and cannot be economically solved using the conventional optimization package, sub-optimal and practical methods are presented in Section IV.

IV. IMPLEMENTATION

This section deals with practical implementation techniques for 2 transmit antenna unitary PAPR reduction. Since maximization over all \( n \) in (6) is numerically very expensive, at first it is better to reduce the cardinality of maximization domain. According to the definition of PAPR in MIMO system in (2), peak occurrences after unitary PAPR reduction might depend mostly on the time instances when both transmit signals are big. Therefore, maximization over all \( n \) can be reduced into maximization over reduced set of \( n \). Random searches using only \( N_c \) vectors corresponding to \( N_c \) largest \( \| \tilde{x}_n \|_2 \) are evaluated with various \( N_c \)'s in Fig. 2. Using only 16 out of 640 \( (L = 5, D = 64 \text{ and } N_f = 2) \) samples,\(^1\) Note that the cyclic prefix is excluded, when considering the PAPR problem.

about the same performance as using entire packet can be achieved in 2 transmit antennas. Therefore, maximization over \( N_c \triangleq \{ n \mid N_c \text{ largest } \| \tilde{x}_n \|_2 \text{ for all } n \in [0, LN_f D - 1] \} \) will be used in place of maximization over \( n \in [0, LN_f D - 1] \).

When handling \( \| U \tilde{x}_n \|_\infty \), unitary matrix \( U \) needs to be parameterized. Considering (9), let \( U \) and \( \tilde{x}_n \) defined by

\[
U = \begin{bmatrix} \alpha & -\beta^* \\ \beta & \alpha \end{bmatrix} = \begin{bmatrix} r \cos(\theta) & -r^2 e^{j\theta} \\ -r^2 e^{-j\theta} & r \end{bmatrix}
\]

and

\[
\tilde{x}_n = \begin{bmatrix} A_{n,1} e^{j\theta_{n,1}} \\ A_{n,2} e^{j\theta_{n,2}} \end{bmatrix}
\]

(12)

respectively, where \( 0 \leq r \leq 1 \) and \(-\pi \leq \theta \leq \pi \). Then, the objective function \( \| U \tilde{x}_n \|_\infty^2 \) is given by (note that minimizing \( \| U \tilde{x}_n \|_\infty^2 \) is equivalent to minimizing \( \| U \tilde{x}_n \|_\infty \))

\[
\| U \tilde{x}_n \|_\infty^2 = \max \{ f_{n,1}(r, \theta), f_{n,2}(r, \theta) \} \triangleq F_n(r, \theta)
\]

(13)

where

\[
f_{n,1}(r, \theta) = r A_{n,1} e^{j\theta_{n,1}} + \sqrt{1 - r^2} A_{n,2} e^{j(\theta_{n,1} + \theta_{n,2})}
\]

\[
= r^2 A_{n,1}^2 + (1 - r^2) A_{n,2}^2 + 2r \sqrt{1 - r^2} A_{n,1} A_{n,2} \cos(\theta_{n,1} - \theta_{n,2} - \theta)
\]

and

\[
f_{n,2}(r, \theta) = -r \sqrt{1 - r^2} A_{n,1} e^{j(\theta_{n,1} - \theta)} + r A_{n,2} e^{j\theta_{n,2}}
\]

\[
= (1 - r^2) A_{n,1}^2 + r^2 A_{n,2}^2 + 2r \sqrt{1 - r^2} A_{n,1} A_{n,2} \cos(\theta_{n,1} - \theta_{n,2} - \theta).
\]

It should be noted that \( \sqrt{1 - r^2} \) in (12) is only used instead of \( \pm \sqrt{1 - r^2} \) due to the anti-symmetry between \( f_{n,1}(r, \theta) \) and \( f_{n,2}(r, \theta) \). Again for the same reason, \( 0 \leq r \leq 1 \) is considered as opposed to \(-1 \leq r \leq 1 \).

Now the problem in (6) can be simply rewritten as

\[
\begin{align*}
\text{Minimize} & \quad \max_{n \in N_c} F_n(r, \theta)
\end{align*}
\]

(15)
where the variables are $r$ and $\theta$. The objective function $F_n(r, \theta)$ has two local maxima for each $n$ at

$$r = \frac{A_{n,1}}{\sqrt{A_{n,1}^2 + A_{n,2}^2}}, \quad \theta = \theta_{n,1} - \theta_{n,2}$$

and

$$r = \frac{A_{n,2}}{\sqrt{A_{n,1}^2 + A_{n,2}^2}}, \quad \theta = \theta_{n,1} - \theta_{n,2} + \pi$$

and minima valleys around two maximal points. Hence, the overall objective function ($\max_{n \in \mathbb{N}_c} F_n(r, \theta)$) has nonconvex characteristics with a few local maxima and minima. An example contour plot for this objective function with $N_c = 16$ is depicted in Fig. 3, where the centers of the ovals designate local maxima and two global minimal points are pointed out. Because of the nonconvexity and irregularity of the objective function, it is very tough to locate global minimum using conventional searching methods. Two suboptimal methods to determine near-minimal point are presented in the following subsections.

### A. Multiple Gradient Search (MGS)

Gradient descent method with multiple starting points is considered to combat the nonconvexity. Multiple gradient search procedure is described as:

**Given** multiple starting points

$$(r_{ni}, \theta_{ni}) \in \{0 \leq r \leq 1, -\pi \leq \theta \leq \pi\}$$

**For each starting point**

1. Repeat
   1. $\Delta r = -\nabla_r (\max_{n \in \mathbb{N}_c} F_n(r, \theta))$
   2. $\Delta \theta = -\nabla_{\theta} (\max_{n \in \mathbb{N}_c} F_n(r, \theta))$
2. Line search.
   - Choose step size $t$ via backtracking line search.
3. Update.
   - $r = r + t \Delta r$
   - $\theta = \theta + t \Delta \theta$

**Until** stopping criterion is satisfied.

**Determine** $(r_{opt}, \theta_{opt})$ with minimum objective function.

The number of starting points to obtain a reasonable suboptimal solution depends on the cardinality of maximization domain $N_c$ and packet length $N_f$. 4 starting points are used for 2 transmit antenna case ($N_c = 16$ and $N_f = 2$).

### B. Parameterized Grid Search (PGS)

Since the gradient computation for the objective function ($\max_{n \in \mathbb{N}_c} F_n(r, \theta)$) is expensive (note that the objective function consists of $2N_c$ subfunctions), exhaustive grid search over $(r, \theta)$ plane is considered instead. When considering the characteristics of objective function, more dense grids are required when $r \geq 0.5$. Consequently $N_r$ partitions along $0 \leq r \leq 0.5$, $2N_r$ partitions along $0.5 \leq r \leq 1$, and $N_\theta$ partitions along $-\pi \leq \theta \leq \pi$ are used. Therefore, nonuniform grid pattern with total $3N_rN_\theta$ grids is employed. The objective function value for each grid center is computed and the center with minimum objective function value is declared as the parameters for a sub-optimal unitary matrix. $N_r = 2$ and $N_\theta = 18$ are used for 2 transmit antenna case (i.e., $3N_rN_\theta = 108$) grids are searched over entire $(r, \theta)$ plane. Further gradient search can be applied after grid search, but additional gain is negligible. On the other hand, gradient search combined with a reduced number of grid searches can also be considered, but the performance is worse and complexity is higher compared to the original grid search.

### V. Extension to General Number of Transmit Antennas

Unitary PAPR reduction for 2 transmit antennas has been discussed so far. This section will extend this algorithm to a general number of transmit antennas. When $L_t$ transmit antennas are employed, an $L_t \times L_t$ unitary matrix needs to be parameterized for the optimization. Generally $L_t^2$ number of parameters are required to represent an $L_t \times L_t$ unitary matrix [26]. When $L_t > 2$, a couple of parameters can be omitted similarly as for the $2 \times 2$ unitary matrix, but it is still hard to solve the optimization problem practically. In this paper iterative extension methodology based on the $2 \times 2$ unitary matrix formulation is proposed.

At each iteration, an independent $2 \times 2$ unitary matrix is applied on each set of paired-off transmit antennas. The pairings should be distinct from iteration to iteration until all the combinations are considered. When the number of transmit antenna is odd, one unpaired antenna at each iteration remains unchanged by skipping the PAPR reduction. The iteration continues up until achieving the desired PAPR reduction or reaching the stopping criterion. In the following unitary PAPR reduction for 4 transmit antennas is described, but this can be generalized to any number of transmit antennas.

The overall procedure for 4 transmit antennas is summarized in Table I. At the 1st iteration, one $2 \times 2$ unitary matrix is applied onto antenna 1 and antenna 2, and the other $2 \times 2$ unitary matrix onto antenna 3 and antenna 4. At the 2nd iteration, one onto antenna 1 and 4, the other onto antenna 2 and 3, at the 3rd iteration, one onto antenna 1 and 3, the other onto antenna 2 and 4, and so on. For each $2 \times 2$ unitary matrix the optimization discussed in Section IV can be applied individually. Assuming 3 iterations, the resulting overall unitary matrix would be $U_3 U_2 U_1$.

### VI. Performance and Complexity

The performance of the proposed unitary PAPR reduction with short packet structure (52 payload subcarriers out of 64
TABLE I
UNITARY PAPR REDUCTION FOR 4 TRANSMIT ANTENNAS.

\[
\begin{pmatrix}
  r_{1,1} & 0 & 0 & 0 \\
  0 & r_{1,1} & 0 & 0 \\
  0 & 0 & \sqrt{1 - r_{1,2}^2}e^{j \theta_{1,2}} & r_{1,2} \\
  0 & 0 & -\sqrt{1 - r_{1,2}^2}e^{-j \theta_{1,2}} & r_{1,2}
\end{pmatrix}
\]

(12, 34): \( U_1 = \)

\[
\begin{pmatrix}
  r_{2,1} & 0 & 0 & 0 \\
  0 & r_{2,2} & \sqrt{1 - r_{2,2}^2}e^{j \theta_{2,2}} & 0 \\
  0 & -\sqrt{1 - r_{2,2}^2}e^{-j \theta_{2,2}} & r_{2,2} & 0 \\
 -\sqrt{1 - r_{2,2}^2}e^{-j \theta_{2,2}} & 0 & 0 & r_{2,1}
\end{pmatrix}
\]

\( \downarrow \)

(14, 23): \( U_2 = \)

\[
\begin{pmatrix}
  r_{3,1} & 0 & 0 & 0 \\
  0 & r_{3,2} & \sqrt{1 - r_{3,2}^2}e^{j \theta_{3,2}} & 0 \\
  0 & -\sqrt{1 - r_{3,2}^2}e^{-j \theta_{3,2}} & r_{3,1} & 0 \\
 -\sqrt{1 - r_{3,2}^2}e^{-j \theta_{3,2}} & 0 & 0 & r_{3,2}
\end{pmatrix}
\]

(13, 24): \( U_3 = \)

subcarriers) [19] is provided via computer simulation. The use of MIMO OFDM system with Alamouti STBC [27] in 2 transmit antennas and with super-orthogonal STBC [18] in 4 transmit antennas are considered. This PAPR reduction scheme is compatible with more powerful space-time modulations but this paper considers simple block codes to keep the focus on the achieved PAPR reduction and performance improvement.

First, unitary PAPR reduction in Section IV is compared with active constellation extension (ACE) [10]. ACE is one of the nonbijective constellation techniques. By intelligently extending outer constellation points it minimizes the PAPR of an OFDM system. Since it is one of the PAPR reduction techniques that do not lose spectral efficiency, a fair spectral efficiency comparison can be made with the unitary PAPR reduction. ACE is implemented using a smart gradient-project method [10] with 3 iterations and a clip level of 4.86 dB above the average power is used. ACE is only operated whenever the PAPR of the symbol is greater than 6 dB to reduce the complexity. Unitary PAPR reduction is executed using both MGS and PGS with \( N_c = 16 \). MGS uses 4 starting points and PGS uses 108 grids (\( N_r = 2 \) and \( N_\theta = 18 \)). For fair comparison unitary PAPR reduction also operates only when the PAPR of the signal is greater than 6 dB.

The complementary cumulative distribution function (CCDF) using MGS, PGS and ACE in 2 transmit antennas employing Alamouti STBC with QPSK and 64 QAM are compared in Fig. 4. The CCDF using a sub-optimal search [19] out of 200 randomly generated unitary matrices over the entire packet is also presented for the reference\(^2\). Additional gains obtained by using more than 200 unitary matrices were shown to be negligible in experimentation. MGS and PGS exhibit almost the same performance as the random search, and both are superior to the ACE especially when 64QAM is used. While ACE has less PAPR reduction gain when higher constellation is used\(^3\), unitary PAPR reduction shows robust performance regardless of the constellation size. The amount of PAPR reduction of the proposed scheme is relatively small compared to other SISO techniques, but it should be emphasized that these gains can be obtained without sacrificing spectral efficiency at all and it can be used additionally in conjunction with SISO PAPR techniques in the MIMO case.

Complexity comparison is also performed for ACE and unitary PAPR reduction via MGS and PGS. Average computational complexity is measured in the number of real floating point operations (which include multiplication, addition, division and comparison) in Fig. 5. PGS has the least complexity and then ACE and MGS in overall. While MGS and PGS show about the same performance in Fig. 4, PGS is much less complex than MGS. By using PGS instead of ACE the complexity can be reduced by about 26% in 64 QAM and by about 43% in QPSK, even though PGS shows better performance than ACE especially in 64 QAM. It should be noted that the overall complexity in Fig. 5 just added all the number of operations. ACE, in fact, requires 284 and 190 divisions in QPSK and 64 QAM respectively, while PGS does not have any division at all. Divisions tend to be more complex than additions and multiplications.

It is seen in Fig. 4 and Fig. 5 that the complexity and performance of unitary PAPR reduction is not affected by constellation size, whereas ACE is critically dependent on the

\(^2\)Random search is used for the reference instead of the global nonconvex solution due to the extreme complexity of global solution. The CCDF difference between two is about 0.2 dB from the preliminary result.

\(^3\)The larger constellation results in less PAPR reduction with ACE, since the smaller ratio of exterior points to interior points gives less degrees of freedom to ACE.
constellation size due to its algorithmic characteristics. Fig. 4 and Fig. 5 also confirm the superiority of unitary PAPR reduction compared to the ACE in terms of both performance and complexity in given condition. The performance of the ACE can be improved when a larger number of data subcarriers are used and/or a bigger constellation is used, but it will result in relatively large complexity increase with the ACE.

The CCDF for 4 transmit antennas is presented in Fig. 6. PGS using 108 grids with $N_c = 32$ is used for a $2 \times 2$ unitary matrix in each iteration. After 2 iterations iterative grid search achieves comparable performance with random search over 200 unitary matrices, and after 3 iterations it outperforms the random search. To reduce the complexity, smaller number of iterations and/or smaller $N_c$ can be used.

To evaluate the overall system performance using unitary PAPR reduction, RAPP nonlinear amplifier model [28] is used for the HPA. The RAPP model is developed for a solid-state power amplifier and it is recommended for use in the IEEE 802.11 TGn comparison criteria document [29]. The RAPP model is given by

$$V_{out} = \frac{V_{in}}{1 + \left(\frac{V_{in}}{V_{sat}}\right)^{2p}}^{1/2p}$$  \hspace{1cm} (16)

where $V_{in}$ and $V_{out}$ is the input and output magnitude of the amplifier, $V_{sat}$ is the saturation magnitude and $p$ is the smoothness factor. $p = 3$ is used in the simulation [29]. The backoff of the amplifier $P_{bo}$ is defined as the input power backoff from the full saturation, i.e.,

$$P_{bo} = -10 \log_{10} \left( \frac{P_{tx}}{P_{sat}} \right)$$  \hspace{1cm} (17)

where $P_{tx}$ is the average power per transmit antenna and $P_{sat}$ is the saturation power of the amplifier.

The BER performance is evaluated in Fig. 7 with respect to different $P_{bo}$ as a function of $E_b/N_0$ per receive antenna. Additive white Gaussian noise (AWGN) is used to measure

---

**Fig. 4.** Performance comparison of PAPR reductions in terms of PAPR CCDF in 2 transmit antennas with Alamouti STBC.

**Fig. 5.** Complexity comparison of PAPR reductions in 2 transmit antennas with Alamouti STBC.

**Fig. 6.** Iterative unitary PAPR reduction using PGS for 4 transmit antennas.
the BER. PGS using 108 grids with $N_c = 16$ is employed for 2 transmit antennas and iterative PGS using 108 grids with $N_c = 32$ and 3 iterations is employed for 4 transmit antennas. ‘NoAmp:PCSI’ and ‘NoAmp:PSAM’ denote the BER without nonlinear HPA using perfect CSI and estimated CSI respectively, which assumes the use of ideal HPA. All other curves shown in Fig. 7 use the estimated CSI and nonlinear HPA. Apparently, the lower $P_{bo}$ is, the bigger gain is obtained in both cases. The unitary PAPR reduction enables to use the HPA with lower $P_{bo}$ requirement. Furthermore, it is also seen that high throughput cannot be achieved without PAPR reduction due to high error floor when the HPA with low $P_{bo}$ is deployed in the transmitter. This nonlinearity effect from the HPA is more dominant in 2 transmit antennas due to lower spatial diversity. It also should be noted that Fig. 7 shows the BER gain in uncoded system, but this gain has been preserved when a strong channel coding such as bit interleaved coded modulation is employed.

So far all the simulation results are obtained with short packet structures. It is obvious that the longer packet length degrades the performance of the PAPR reduction. In order to understand how the packet length influences the PAPR reduction, the CCDF of PAPR is shown with respect to different frame length $N_f$ in Fig. 8, where a sub-optimal unitary matrix is chosen over 200 random unitary matrices, and all 52 subcarriers are allocated for the data without any pilot. As expected, the PAPR reduction gain is reduced when the frame length increases. Interestingly most degradations happen in high probability region and the highest PAPRs are not much affected. To maximize the PAPR reduction gain in long packets it would require multiple transformations in one packet. In this case the side information representing the unitary matrices would be transmitted or separate channel estimation per each transformation is required.

VII. BEAMFORMING EFFECT

A unitary matrix in the proposed PAPR reduction is a beamformer at the MIMO transmitter. It is worthwhile to show that beamforming effect due to the unitary rotation does not change the system performance for open loop coding techniques. This claim will be proved in terms of unconstrained capacity and the constrained capacity for the Alamouti code. BER performance is simulated for the Alamouti code to further emphasize this point.

1) Unconstrained Capacity: Assuming that (1) the transmitted signal vector is a zero-mean circularly symmetric complex Gaussian with equal power components, (2) the channel is independent of both transmit signal and noise, and (3) channel is known to the receiver, the average channel capacity of the MIMO system with $L_t$ transmit and $L_r$ receive antennas is given by [1], [2]

$$C = E_H \left[ \log_2 \det \left( I_{L_r} + \frac{\rho}{L_t} H H^H \right) \right] \text{bps/Hz} \quad (18)$$

where $H$ is an $L_r \times L_t$ channel matrix; $\rho$ denotes the average signal-to-noise ratio per receive antenna; $E_H$ represents the expectation over different channel realizations.

The mutual information between the transmitted signal $\tilde{X}$ with unitary rotation $U$ and received signal $\tilde{Y}$ given channel
transmit power, $R_U$ is upper-bounded by

$$I(\tilde{X};\tilde{Y}|H) = H(\tilde{Y}|H) - H(\tilde{N}) \leq \log_2 \det \pi e \left( N_0 I_{L_r} + HE_{\tilde{X}} \left[U \tilde{X} \tilde{X}^H U^H\right] H^H\right) - \log_2 \det \pi e N_0 I_{L_r} = \log_2 \det \left(I_{L_r} + \frac{1}{N_0} HR \tilde{X} \tilde{X}^H H^H\right)$$ (19)

where $\tilde{X} = U \tilde{X}$ and $R_\tilde{X}$ is an autocovariance of $\tilde{X}$. In (19), $U$ is included inside an expectation operator since it is a function of $\tilde{X}$. For the simplicity argument of $U$ is omitted. $U$ is a nonlinear function of $\tilde{X}$ and thus $R_\tilde{X}$ is very complex to analyze. Assuming $R_\tilde{X} = P/L_t I_{L_r}$ where $P$ is the average transmit power, $R_\tilde{X}$ is numerically verified and turned out to be

$$R_\tilde{X} \simeq R_X = \frac{P}{L_t} I_{L_r}$$ (20)

where $U$ is computed with the PGS. Inserting (20) into (19) and taking the expectation over the channel, the capacity with unitary rotation becomes same as the original channel capacity (18).

2) Constrained Capacity: Channel capacity in (18) assumes an i.i.d Gaussian inputs [1], [2]. However, Gaussian-distributed input is not feasible due to implementation complexity. Instead, a constrained modulation, such as finite constellation (e.g., QPSK, 16 QAM, and 64 QAM etc.) and/or STBC, is employed. In the case of some STBCs, channel capacity was computed still assuming indefinite constellation. Channel capacity of Alamouti STBC was shown to achieve the open-loop capacity of a $2 \times 1$ system despite its decoupled nature [30].

The constrained channel capacity of Alamouti STBC with $L_r$ receive antenna is given by

$$C_A = \log_2 \left(1 + \beta \frac{\tilde{h}_1^H \tilde{h}_1 + \tilde{h}_2^H \tilde{h}_2}{2}\right)$$ (21)

where $H = [\tilde{h}_1 \tilde{h}_2]$; $\tilde{h}_1$ and $\tilde{h}_2$ are $L_r \times 1$ vectors. It is shown in Appendix that the constrained capacity is also invariant with the unitary rotation.

3) BER Performance: BER performance is simulated in Fig. 9 with and without unitary PAPR reduction assuming no nonlinear distortion from the HPA, which means that the PAPR could not impact the performance. The use of a $2 \times 2$ system with Alamouti STBC employing a 64 QAM is considered in AWGN and channel F (the worst fading channel model with non-line-of-sight (NLOS), K-factor = $-\infty$, rms delay spread = 150 ns, and 6 clusters) defined in IEEE 802.11 TGn channel models document [31]. In Fig. 9, ‘PCSI’ stands for perfect CSI and ‘PSAM’ for estimated CSI. Unitary PAPR reduction is executed via a random search of the unitary matrix [19]. It is shown that the BER is not related with the unitary rotation when assuming no nonlinear distortion.

This BER performance can be also predicted from the pairwise error probability (PWEPE) of the space-time codes. In the case of both small [32] and large [33], [34], [35] number

of independent channels, the upper-bound of the PWEPE in Rayleigh fading channels is shown to be unchanged with the unitary PAPR reduction [25]. The unitary matrix $U$ is absorbed into the unitary part of the singular value decomposition (SVD) in the codeword difference matrix regardless of its dependency on the transmitted signal. As a result, the average PWEPE can be represented only with the eigenvalues of the codeword difference matrix.

VIII. CONCLUSIONS

In this paper, a new PAPR reduction scheme in the multiple transmit antenna environments was proposed. PAPR can be reduced by a unitary rotation over the transmit antennas and it achieves the throughput-lossless PAPR reduction. Furthermore, this technique does not require any processing in the receiver. Finding the optimal unitary matrix was shown to be nonconvex optimization problem, and alternate sub-optimal solutions were provided. The drawback of this scheme is that it is more advantageous in short packet structure for the costless channel estimation. When it is applied to a long packet, the PAPR reduction gain decreases, but the complexity can be further saved compared to the SISO PAPR reduction techniques.

APPENDIX

CONSTRAINED CAPACITY OF ALAMOUTI STBC

The unitary rotated received signal of Alamouti STBC can be represented as

$$\begin{align*}
\tilde{Y} &= \begin{bmatrix} \tilde{h}_1 \\ \tilde{h}_2 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ -\beta^* \\ \alpha^* \end{bmatrix} \begin{bmatrix} a \\ b \\ -b^* \\ a^* \end{bmatrix} + \begin{bmatrix} \tilde{n}_1 \\ \tilde{n}_2 \end{bmatrix} \\
&= \begin{bmatrix} \alpha \tilde{h}_1 - \beta^* \tilde{h}_2 \\ \beta \tilde{h}_1 + \alpha^* \tilde{h}_2 \end{bmatrix} \begin{bmatrix} a \\ b \\ -b^* \\ a^* \end{bmatrix} + \begin{bmatrix} \tilde{n}_1 \\ \tilde{n}_2 \end{bmatrix} \\
&= \begin{bmatrix} \Re \tilde{c} \\ \Im \tilde{c} \end{bmatrix} + \begin{bmatrix} \Re \tilde{q} \\ \Im \tilde{q} \end{bmatrix} + \begin{bmatrix} \tilde{n}_1 \\ \tilde{n}_2 \end{bmatrix} \\
&= \begin{bmatrix} a \Re \tilde{c} + b \Im \tilde{c} \\ -b^* \Re \tilde{c} + a^* \Im \tilde{c} \end{bmatrix} + \begin{bmatrix} \tilde{n}_1 \\ \tilde{n}_2 \end{bmatrix}
\end{align*}$$

(22)
All the vectors in (22) are the column vectors of dimension $L_r \times 1$. A special unitary matrix $SU(2)$ [21] is only considered for the unitary matrix $U$ in (22), since the phase does not affect the PAPR. By taking the conjugate in the 2nd column,
\[
\vec{Y}' = \begin{bmatrix} \alpha \vec{p}' + \beta \vec{q}' \end{bmatrix} = \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} \vec{p}' \vec{q}' \end{bmatrix} + [n_1 \ n_2^*].
\]

Applying the maximum ratio combining after changing the dimension of $\vec{Y}'$ as a column vector,
\[
\vec{Y}' = \begin{bmatrix} \vec{p}' \vec{q}' - \vec{p}' \vec{q}' \end{bmatrix} = \begin{bmatrix} p_1^H \vec{q}_1 \vec{p}_2 \vec{q}_2 \\ q_1^H \vec{p}_1 \vec{q}_2 \vec{p}_2 \end{bmatrix} = \begin{bmatrix} (\vec{h}_1^H \vec{h}_1 + \vec{h}_2^H \vec{h}_2) a \\ (\vec{h}_1^H \vec{h}_1 + \vec{h}_2^H \vec{h}_2) b \end{bmatrix} + \begin{bmatrix} \vec{p}' \vec{q}_1 \vec{q}_2 \\ \vec{q}_1^H \vec{h}_1 \vec{q}_2 \end{bmatrix},
\]

since
\[
\begin{align*}
\sigma_N^2 & \triangleq E \left[ |p_1^H \vec{q}_1 + q_1^H \vec{q}_2|^2 \right] = E \left[ |q_1^H \vec{h}_1 + \vec{h}_2^H \vec{h}_2|^2 \right] \\
& = E \left[ p_1^H \vec{q}_1 \vec{q}_2 \vec{p}_2 \vec{q}_2 \right] + E \left[ q_1^H \vec{p}_1 \vec{q}_2 \vec{p}_2 \vec{q}_2 \right] \\
& = \sigma_n^2 \left( \vec{h}_1^H \vec{h}_1 + \vec{h}_2^H \vec{h}_2 \right). 
\end{align*}
\]

In (24) noise is a function of the data $a$ and $b$ (recalling that $\alpha$ and $\beta$ is a function of $a$ and $b$), but it is uncorrelated with data. From (24) and (26), the channel capacity of unitary rotated Alamouti STBC can be derived as
\[
\begin{align*}
\bar{C}_A &= \log_2 \left( 1 + \frac{(\vec{h}_1^H \vec{h}_1 + \vec{h}_2^H \vec{h}_2)^2 \sigma_n^2}{\vec{h}_1^H \vec{h}_1 + \vec{h}_2^H \vec{h}_2 \sigma_n^2} \right) \\
& = \log_2 \left( 1 + \frac{2\sigma_n^2 \vec{h}_1^H \vec{h}_1 + \vec{h}_2^H \vec{h}_2}{\sigma_n^2} \right) \\
& = \log_2 \left( 1 + \frac{2\sigma_n^2 \vec{h}_1^H \vec{h}_1 + \vec{h}_2^H \vec{h}_2}{\sigma_n^2} \right) \\
& = \bar{C}_A.
\end{align*}
\]

REFERENCES


**Heechoon Lee** received his B.S. and M.S. degrees in electrical engineering from the Seoul National University, Korea in 1997 and 2002, respectively, and his Ph.D. degree in electrical engineering from the University of California, Los Angeles in 2007. He currently works on WCDMA systems as a senior engineer for Qualcomm Inc., San Diego. His research interests are in the design of physical layer communication systems including MIMO signal processing, peak-to-average power ratio reduction, equalization and synchronization. He was a recipient of the IT scholarship from the Ministry of Information and Communication, Republic of Korea in 2002.

**Michael P. Fitz** received his B.E.E. degree (summa cum laude) from the University of Dayton, Dayton, Ohio, in 1983 and his MS and Ph.D. degrees in electrical engineering from the University of Southern California in 1984 and 1989, respectively. From 1983-1989 he worked as a communication systems engineer for Hughes Aircraft and TRW Inc. In 1989 he ventured into academia and was faculty at Purdue University, the Ohio State University (OSU), and the University of California Los Angeles. Dr. Fitz is currently employed at the Northrop Grumman Corporation as a senior systems engineer working on satellite communications. Dr. Fitz’s research is in the broad area of statistical communication theory and experimentation. Dr. Fitz is the author of the forthcoming textbook *The Fundamentals of Communications Systems*. He was a recipient of the 2001 IEEE Communications Society Leonard G. Abraham Prize Paper Award in the Field of Communications Systems.
Abstract—In this paper, we propose an original detection scheme for high rate short-range impulse radio ultra-wideband systems. The proposed receiver relies on both the introduction of the cyclic prefix at the transmitter and the use of a frequency domain multiuser detector at the receiver. Zero Forcing (ZF) and Minimum Mean Square Error (MMSE) detection strategies have been investigated and compared with the classical RAKE, considering a scenario where several mobile terminals communicate with a base station in an indoor environment characterized by severe multipath propagation. The results show that the MMSE receiver achieves the best performance, irrespective of the number of active terminals, both in the uplink and in the downlink communications. Hence, the proposed approach is well suited in indoor wireless environments where the multipath propagation tends to increase the effects of both the inter-path and the inter-user interference.

Index Terms—Ultra-wideband, impulse radio, frequency domain detection, RAKE receiver, dense multipath propagation.

I. INTRODUCTION

Ulra-wideband (UWB) techniques are generally considered as promising technologies for future broadband short-range wireless communications, mainly because of their inherent capabilities in providing solutions for spectrum management and radio system engineering issues [1]. UWB applications can naturally be foreseen in high data-rate wireless personal area networks, in sensor, positioning and identification networks, in imaging and vehicular radar systems, in peer-to-peer communications [2]. Nonetheless, the interest in UWB communication systems before 2001 was mainly concentrated on radar applications until the FCC released a spectral mask allowing operation of UWB systems at the noise floor, over an enormous bandwidth [3].

Impulse Radio (IR) communications [4], which are based on the use of baseband pulses of very short duration, typically on the order of a nanosecond, have been recently massively studied as one of the most interesting UWB techniques [2]; the reasons of the attention devoted to the UWB-IR communications can be identified in the very low power spectral density of the signals involved and in the strongly reduced interference that is caused upon narrow band systems operating in the licensed bands [5].

IR multiuser communication systems can be implemented by resorting to the use of Time Hopping (TH) or Direct Sequence (DS) spread-spectrum signals [2]: both techniques rely on the use of short duration pulses and the introduction of spreading signatures for the users, even if different spreading techniques are used. Moreover, it can be observed that the transmitter which is used in TH-based IR communications is less complex.

IR systems based on TH signals generally use impulsive modulation techniques such as Pulse Amplitude Modulation (PAM) and, mainly, Pulse Position Modulation (PPM) [4], [6]–[10]; in these systems the same symbol is repeated many times, according to a specific random code, so providing a very high processing gain.

For what concerns channel effects on IR signals, it is well known that a pulse having a very short duration strongly reduces the minimum discernible interval between individual components and, as a consequence, considerably increases the multipath resolution; this phenomenon is particularly evident in a indoor environment [11]. As a result, UWB communications based on IR signals have to face an extremely frequency-selective channel and the received signal contains a significant number of resolvable multipath components [12]. Conversely, modulation resolution down to a nanosecond in differential path delay leads to the elimination of significant multipath fading [13] and may considerably reduce fading margins in link budget.

The multipath diversity inherent in the received IR signals and the high processing gain have led most of researchers to consider correlation or RAKE receivers as the most suitable solution for this kind of communications (see the references in [14]); nevertheless, RAKE receivers are known to be vulnerable to Multiple Access Interference (MAI) with remarkable performance and system capacity loss even for a moderate number of active interfering users [15]. This phenomenon is also evident when the transmitted signals can be assumed synchronous and coordinated, e.g., when downlink indoor communications between the Access Point (AP) and the Mobile Terminals (MTs) are considered, since a dense multipath channel ends up destroying the orthogonality among the users.

These impairments are even more harmful in the case of asynchronous systems, e.g., when considering the uplink between the MTs and the AP [6]. In this case, many alternative strategies can be adopted for facing both channel and interference impairments in UWB-IR communication systems. In [16], a multistage block spreading (MS-BS) approach is proposed to achieve resilience to MAI and inter-symbol interference: this scheme aims at rendering the multiple access channel equivalent to a set of independent parallel single-user communications.
frequency-selective channels with AWGN.

As to the multiuser detection strategies, several receivers have been proposed [17]–[24]: in all of these receivers the detection is performed in the time domain, as in the classical MUD approach for Direct Sequence Code Division Multiple Access (DS-CDMA) [15]. As a consequence, a remarkable number of operation has to be performed in the pulse interval, making the implementation of these detectors unfeasible in the near future. Moreover, the receivers proposed in [17]–[21] have been defined for an AWGN channel and their generalization to a frequency selective channel should be derived. On the other hand, the receivers [22]–[24] are designed for a frequency-selective channel scenario but are characterized by a prohibitive complexity.

Another possible strategy to combat multipath in a single-carrier transmission is the adaptive equalization at the receiver [25], [26]: anyway, since adaptive equalizers require one or more filters for which the number of adaptive tap coefficients is on the order of the number of data samples spanned by the multipath, they are not suitable for UWB indoor communications where more than 100 channel resolvable replicas have to be taken into account.

Recently, Frequency Domain Equalization (FDE), which has been previously proposed and studied for a single-carrier single-user environment [27], has been considered for short-range IR communications. Based on this approach, a simple frequency domain detection scheme affording a good complexity/performance trade-off has been derived for UWB-IR downlink multiuser communications in [28]. Similar schemes were also proposed and analyzed for single user communications in [29] and [30]. Nonetheless, the encouraging results which are afforded in terms of performance and complexity pave the way for its generalization to the uplink multiuser environment.

In this paper, Frequency Domain Detection (FDD) schemes for UWB-IR short-range multiuser communications will be proposed and simulated in an extremely frequency selective environment [11], in order to highlight how the orthogonality loss and the rise of both inter-path interference (IPI) and MAI can be effectively coped with. With respect to [28], an original Frequency Domain Multiuser Detection (FDMUD) scheme will be derived and simulated in an uplink multiuser environment, where the effects of the MAI are more harmful. Moreover, we will introduce a novel design strategy regarding both the cyclic prefix length and the TH sequences whose effects are twofold: on the one hand, this strategy eliminates the need for transmit redundancy; on the other hand, since this design strategy permits to consider short signal blocks, it affords a sensible reduction in the number of operations to be performed and, therefore, a simplification in the detection task.

Particular attention will be dedicated to the maximum number of users that can be correctly detected by the proposed detectors in both the downlink and the uplink, and to the comparison with a conventional RAKE receiver. The proposed receivers are based on the use of an analog correlation as the front end, followed by an Analog-to-Digital Converter (ADC) [31] [32]: this hybrid architecture affords looser sampling rate requirements, e.g., down to the inverse of the pulse duration, and allows for less complex system implementations.

This paper is organized as follows. The system model together with the block vectorial representation is presented in Section II. The frequency domain approach and the proposed detectors are discussed in Section III while the simulation results and the complexity considerations are presented in Section IV and V, respectively. The concluding remarks are given in Section VI.

II. SYSTEM MODEL

A. UWB signal model

In an UWB-IR communication system, the signal relative to the $i$th user can be expressed as [6] [33]

$$s_i(t) = \sum_{m=-\infty}^{+\infty} w_{tx}(t-mT_f - \tau b_i([m/N_f])), \quad (1)$$

where $w_{tx}(t)$ represents the shape of the impulse, $T_f$ and $T_c$ are the frame and the chip periods, respectively, and $b_i(t) = \pm 1$ is the $i$th binary symbol transmitted by the $i$th user. In particular, since $|x|$ stands for the integer part of $x$, eq. (1) indicates that the same bit is transmitted over $N_f$ consecutive frame periods. We assume that $N_c$ chips exactly fit in one frame period, i.e., $T_f = N_cT_c$. Each active user is associated with a time-hopping pattern $c_i(m)$. In the most general case, $c_i(m)$ can be modeled as a periodic pseudo-random sequence with period $N_f$. Pulse position modulation is implemented by means of an additional pulse shift $\tau(b)$. In the binary case, we have $\tau(b) = \{0, T_w\}$ depending on $b = \{1, -1\}$, where $T_w = T_c/2$ (see Fig. 1).

The signal relative to the $i$th user can be represented more conveniently as

$$s_i(t) = \sum_{k=-\infty}^{+\infty} w_{tx}(t-kT_w)x_i(k), \quad (2)$$

where we define $x_i(k) \triangleq [q_t(k)b_i([k/N_w)] + p_t(k)], N_w = 2N_cN_f$, and the sequences $q_t(k)$ and $p_t(k)$ can be expressed as

$$q_t(k) = \begin{cases} \frac{1}{2} & \text{if } k = 2[mN_c + c_i(m)] \\ -\frac{1}{2} & \text{if } k = 2[mN_c + c_i(m)] + 1 \\ 0 & \text{elsewhere} \end{cases}$$

and

$$p_t(k) = |q_t(k)|, \quad (3)$$

respectively. In particular, both $q_t(k)$ and $p_t(k)$ are periodic with period equal to $N_w$.

B. Downlink

If we consider an access point transmitting to a set of $N_u$ active users $\ell_u = \{\ell_1, \ell_2, \ldots, \ell_{N_u}\}$, the signal which is transmitted by the access point can be expressed as

$$s(t) = \sum_{\ell \in I_u} s_\ell(t). \quad (5)$$

Recalling (2), we can express the received waveform as

$$r_D(t) = \sum_{k=-\infty}^{+\infty} \phi(t-kT_w) \sum_{\ell \in I_u} x_\ell(k) + \eta(t), \quad (6)$$
where $\phi(t)$ models the effects of the channel, the antennas and both the transmit and receive filters, and $\eta(t)$ models the thermal noise. By assuming that the channel characteristics are constant over the entire block of samples and sampling period $T_w$, we obtain the digital transmission model as

$$y_D(n) \triangleq r_D(nT_w) = \sum_k h(n-k) \sum_{\ell \in I_u} x_\ell(k) + e(n),$$

where $h(n) \triangleq \phi(nT_w)$ represents the equivalent discrete time channel impulse response of the UWB-IR system and $e(n) \triangleq \eta(nT_w)$.

### C. Uplink

An analogous representation can apply also to the case of a set of $N_u$ active users transmitting to an access point. In this case, we can express the received waveform as

$$r_U(t) = \sum_{\ell \in I_u} \sum_{k=-\infty}^{+\infty} \phi_\ell(t - kT_w - \tau_\ell)x_\ell(k) + \eta(t),$$

where $\phi_\ell(t)$ models the overall transfer function of the $\ell$th user, and we suppose that each user transmits with a different delay $\tau_\ell$. Let us express the delay of the $\ell$th user as a function of the sampling period $T_w$, as $\tau_\ell = d_\ell T_w + \delta_\ell$. In particular, $d_\ell$ expresses the delay of the $\ell$th user in term of received samples, whereas $\delta_\ell < T_w$ can be thought of as a residual delay. If we sample $r(t)$ with period $T_w$, then a full digital transmission model can be obtained as

$$y_U(n) \triangleq r_U(nT_w) = \sum_{\ell \in I_u} \sum_k h_\ell(n-k-d_\ell)x_\ell(k) + e(n),$$

where $h_\ell(n) \triangleq \phi_\ell(nT_w - \delta_\ell)$ represents the equivalent discrete time channel impulse response of the UWB-IR system relative to the $\ell$th user. The above equation can be expressed in a more convenient form as

$$y_U(n) = \sum_{\ell \in I_u} \sum_k h'_\ell(n-k)x_\ell(k) + e(n),$$

where for each user $\ell$ we consider the equivalent channel $h'_\ell(n) \triangleq h_\ell(n - d_\ell)$. In (10), the effect of the different delays $\tau_\ell$ is modeled by the increased maximum delay spread of each equivalent channel response.

### D. Block Representation

A block vectorial representation of the above model is more convenient for a clear description of the FDE approach. Let us subdivide the discrete signal $x_\ell(n)$ in blocks of $M$ samples. In the following, we will assume $M = N_bT_w$, so that a group of $N_b$ bits is exactly spread over a block of $M$ samples. We define the vector $x_\ell(i) = [x_\ell(iM), x_\ell(iM+1), \ldots, x_\ell(iM+M-1)]^T$, consisting of the samples of the signals transmitted by the $\ell$th user: the samples are relative to the bits in the $i$th block. We can express $x_\ell(i)$ as

$$x_\ell(i) = [b_\ell(iN_b)q_\ell^{\top} + p_\ell^{\top}, \ldots, b_\ell(iN_b+N_b-1)q_\ell^{\top} + p_\ell^{\top}]^T,$$

(11)

where $q_\ell = [q_\ell(0), q_\ell(1), \ldots, q_\ell(N_b-1)]^T$ and $p_\ell = [p_\ell(0), p_\ell(1), \ldots, p_\ell(N_b-1)]^T$. If we define the vector of the bits transmitted to the $\ell$th user in the $i$th block as $b_\ell(i) = [b_\ell(iN_b), b_\ell(iN_b+1), \ldots, b_\ell(iN_b+N_b-1)]^T$, eq. (11) can be rewritten in a more compact form as

$$x_\ell(i) = Q_{\ell,M}b_\ell(i) + p_{\ell,M},$$

(12)

where $Q_{\ell,M} = I_{N_b} \otimes q_\ell$, $p_{\ell,M} = I_{N_b} \otimes p_\ell$, $\otimes$ indicates Kronecker product and $I_{N_b}$ is an all-ones vector of size $N_b$.

### E. Frequency Domain Model

In order to perform FDE [27], each block is extended by means of a cyclic prefix of length $K$, i.e., the last $K$ samples of the block are repeated at the beginning of the block. Unlike the receiver in [28], in this paper the redundancy due to the CP approach is not considered as an overhead, but as an alternative to the processing gain $N_f$. If we assume that the CP size $K$ has been fixed, the block size required by FD equalization is $M \geq K$. Since UWB systems usually allow for redundancy in terms of pulse repetition [34], it is convenient to set the block size as small as possible, so reducing the complexity of the FD equalization, and to compensate for the loss of data rate by shortening the pulse repetition factor $N_f$. In particular, if we can reduce $N_f$ by a factor $M/(M + K)$, it follows that the data rate is kept unaltered.

Obviously, this choice decreases the processing gain at the receiver. However, this loss is compensated by the better performance of FDE with respect to RAKE. Moreover, since we reduce the pulse repetition factor before CP insertion, this means that after CP insertion the proposed system transmits using the same number of pulses per time unit as the original system, i.e., the proposed system will fulfill the same peak power requirements as the original system.

In the following, the block size is set to $M = K$. Therefore, in order to have the same rate of the original system, the repetition factor of the FD system is set to $N_f^{\text{CP}} = N_f/2$. Moreover, if we choose the cyclic prefix length so that $K \geq L_c$, where $L_c$ indicates the discrete channel length in the downlink, then the channel does not cause any interference between adjacent blocks and the effect of UWB-IR channel can be modeled as a circular convolution between the channel impulse response and the block of $M$ samples [35]. It is well known that circular convolution can be represented by means of a circulant matrix $H$, whose rows are shifted versions of channel impulse response. Hence, if we define the received vector after cyclic prefix removal as $y_D(i) = [y_D(iM), y_D(iM+1), \ldots, y_D(iM+M-1)]^T$, then the input-output relation of the UWB-IR system with cyclic prefix can be expressed as

$$y_D(i) = H \sum_{\ell \in I_u} x_\ell(i) + e(i),$$

(13)
where $\mathcal{H}$ is given by [35]

$$
\mathcal{H} = \begin{bmatrix}
    h(0) & \cdots & 0 & h(L_c) & \cdots & h(1) \\
    \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
    h(L_c - 1) & \cdots & h(L_c) & 0 & \cdots & 0 \\
    h(L_c) & 0 & \cdots & \cdots & \cdots & \cdots \\
    \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
    0 & \cdots & h(L_c) & h(L_c - 1) & \cdots & h(0)
\end{bmatrix},
$$

and $e(i) = [e(iM), e(iM + 1), \ldots, e(iM + M - 1)]^T$.

A modeling similar to that depicted above can be used also to represent the signal received in the uplink. In the following, we will adopt the hypothesis that the system is quasi-synchronous, i.e., we will assume that the cyclic prefix is long enough to cover the longest channel impulse response plus the maximum delay difference between any two users. If we assume the minimum delay equal to zero, this can be expressed as $K \geq L'_\ell$ for each $\ell$, where $L'_\ell$ indicates the length of the equivalent digital channel $h'_\ell(n)$. Such an assumption can be justified considering that the maximum delay difference between any two users is related to the maximum distance covered by the AP. For example, if we suppose a system operating in a circular room having a radius $r$, the maximum delay difference can be assumed equal to the maximum difference of the relative round-trip times, i.e., $d_{max} = 2r/c$, where $c$ indicates the light speed.

If we define the received vector at the AP after cyclic prefix removal as $y_U(i) = [y_U(iM), y_U(iM + 1), \ldots, y_U(iM + M - 1)]^T$, by using the above assumption the input-output relation for a quasi-synchronous UWB-IR system in the uplink is given by

$$
y_U(i) = \sum_{\ell \in I_u} \mathcal{H}_\ell x_\ell(i) + e(i),
$$

where $\mathcal{H}_\ell$ are circulant matrices that can be obtained from (14) by replacing $h(n)$ with $h'_\ell(n)$.

### III. Receiver Schemes

#### A. RAKE

The RAKE receiver [14] has been massively studied for application in UWB-IR systems, both for its ability in exploiting the multipath diversity as well as for its low complexity. If we apply the Maximum Ratio Combining (MRC) algorithm to the model considered herein, the decision variable can be expressed as

$$
v_\ell^{\text{RAKE}}(i) = \sum_{k \in I_{\ell\text{p}}} h^*(k) z_\ell(i, k),
$$

where with $I_{\ell\text{p}}$, we indicate the set of the resolvable channel paths and $z_\ell(i, k)$ is the contribute of the $k$th RAKE finger, given by $z_\ell(i, k) = \sum_{r=0}^{N_f-1} \{y(2[iN_cN_f + rN_c + c_r(r)] + k) - y(2[iN_cN_f + rN_c + c_r(r)] + k + 1)\}$. The above equation can be applied also in the uplink by substituting $h^*(k)$ with $h'_\ell(k - d_\ell)$.

#### B. Frequency Domain Detection

Usually the UWB-IR channel shows a severe delay spread, resulting in a high number of resolvable paths in the RAKE receiver. Therefore, the implementation of a receiver collecting all resolvable channel replicas may prove expensive. Moreover, the channel replicas are closely spaced in time, and, therefore, a sensible inter-path interference arises: even if TH sequences may be adopted in UWB-IR communications, they fail in mitigating this effect because of the non ideal characteristics of their autocorrelation functions. As a result, the long and dense multipath causes a remarkable level of self-interference between the replicas of the signals.

Channel equalization in the frequency domain [27], [28] is a possible solution to the problem of the inter-path interference. In the following, we will describe how to apply FDE to the proposed UWB-IR system.

1) Downlink: Consider the block model in (13). Since the matrix $\mathcal{H}$ is circulant, it can be diagonalized by using a Discrete Fourier Transform (DFT) as $\mathcal{H} = \mathcal{W}_M^H \Lambda_H \mathcal{W}_M$, where $\mathcal{W}_M$ is an $M \times M$ Fourier transform matrix and $\Lambda_H$ is a $M \times M$ diagonal matrix whose entries represent the channel frequency response. Therefore, (13) can be expressed as a function of $\Lambda_H$ as

$$
y(i) = \mathcal{W}_M^H \Lambda_H \mathcal{W}_M \sum_{\ell \in I_u} [\mathcal{Q}_{\ell,M} b_\ell(i) + p_{\ell,M}] + e(i),
$$

where $\mathcal{D}$ and $\mathcal{Q}_{\ell,M}^T$ represent the frequency domain equalization and the correlation with the sequence of the $\ell$th user, respectively.

In this paper, we will focus on two linear receiver techniques, namely Zero Forcing (ZF) and Minimum Mean Square Error (MMSE) equalization, due to their good tradeoff between performance and complexity. In the down-link scenario, the ZF detector is implemented by letting $\mathcal{D}$ equal to the inverse of channel frequency response, i.e.,

$$
\mathcal{D}^{\text{ZF}} = \Lambda_H^{-1}.
$$

In this case, the effect of channel is exactly compensated and self-interference is totally avoided. Moreover, if we use orthogonal time hopping sequences, also the MAI can be completely eliminated. Nevertheless, it is well known that this solution amplifies the noise at the receiver, and hence a performance degradation for low SNR values is expected.

The expression of $\mathcal{D}$ for the MMSE detector is given by

$$
\mathcal{D}^{\text{MMSE}} = \Lambda_H^H \left( \Lambda_H \Lambda_H^H + \frac{N_w \sigma^2_r}{N_w} I_M \right)^{-1},
$$

where $\sigma^2_r$ is the noise variance and $\sigma^2_r$ indicates the power of transmitted symbols. This solution avoids noise amplification at the detector when the SNR is low. However, this algorithm depends on the value of both the noise variance and the power of transmitted symbols, as well as the number of active users and requires the knowledge of the ratio of these parameters.
times \(N_w\), i.e., it requires the knowledge of the inverse of the SNR per sample. This parameter can be estimated during the synchronism acquisition phase by a proper spectral analysis.

Let us consider the MMSE estimator of the overall transmitted signal \(x(i) = \sum_\ell x_\ell(i)\) given by [36]

\[
\hat{x}(i) = C_{xx} H^H \left( C_{xx} H^H + C_{ee} \right)^{-1} y_D(i),
\]

where \(C_{xx}\) and \(C_{ee}\) are the covariance matrices of \(x(i)\) and \(e(i)\), respectively. By comparing (20) with (21), it can be seen that the proposed MMSE detector relies on the approximation \(C_{xx} \approx N_u \sigma_w^2 / N_u I_M\). Actually, this assumption does not hold due to pulse repetition and PPM, but it allows us to derive a diagonal \(\mathbf{D}^{MMSE}\). Moreover, the MMSE detector (20) does not require any knowledge of the time-hopping sequences of the interfering users.

We remark that both the proposed FDD receivers lead to a diagonal matrix \(\mathbf{D}\) and, therefore, the involved complexities are sustainable.

2) Uplink: Consider now the block model for the uplink in (15). By using the DFT factorization \(H_\ell = W_M^H A_\ell W_M\), (15) can be expressed as a function of \(A_\ell\) as

\[
y(i) = W_M^H \sum_{\ell \in I_u} A_\ell W_M \left[ Q_{\ell,M} b_\ell(i) + p_{\ell,M} \right] + e(i).
\]

In the single user case, FDD alone is able to eliminate inter-path interference. However, when several active users are considered, the different channel responses cause a loss of orthogonality among users and a more sophisticated multiuser detection (MUD) approach is needed in order to face inter-user interference. Let us rewrite (22) as

\[
y(i) = W_M^H \Phi b(i) + \omega + e(i),
\]

where \(\Phi \triangleq \left[ A_1 W_M Q_{1,M}, \ldots, A_{N_u} W_M Q_{N_u,M} \right], \omega \triangleq \sum_{\ell \in I_u} A_\ell W_M p_{\ell,M} \) and \(b(i) \triangleq \left[ b_1^T(i), \ldots, b_{N_u}^T(i) \right]^T\).

Relying on the model in (17), a frequency domain MUD (FDMUD) approach can be derived. The vector of the decision variables for all active users, indicated as \(v(i) = [\mathbf{v}_1^T(i), \ldots, \mathbf{v}_{N_u}^T(i)]^T\), can be expressed in a general form as

\[
v(i) = \mathbf{A} \Phi^H \left[ W_M y(i) - \omega \right],
\]

where \(\mathbf{A}\) represents a decorrelating block that is designed according to the selected criterion (see Fig. 3). We will keep on focusing on the two linear decorrelating criteria which have been considered in the downlink scenarios.

In the uplink context, the ZF detector is implemented by letting \(\mathbf{A}\) equal to the inverse of the users’ autocorrelation matrix, i.e.,

\[
\mathbf{A}^{ZF} = (\Phi^H \Phi)^{-1}.
\]

It is worth underlining that in the uplink system the ZF is derived as the result of the estimation of all the transmitted bits of all the users. Therefore, this detector is expected to be robust against the noise enhancement phenomena especially when a moderate load is foreseen, i.e., when \(N_u \ll N_f N_c\). In particular, the noise enhancement is caused by the inversion of matrix \(\mathbf{A}\). Since the dimension of matrix \(\mathbf{A}\) depends on the number of active users, unless the number of active users is very close to the maximum number of detectable users, the FDMUD receiver performs a sort of averaging of the noise variables at the receiver, resulting in a limited noise enhancement. Nonetheless, the ZF performance will be more and more impaired as the number of users increases: when a high system load is considered, a sensible performance degradation is expected.

For what concerns the MMSE detector in the uplink scenario, the expression of \(\mathbf{A}\) is given by

\[
\mathbf{A}^{MMSE} = \left( \Phi^H \Phi + \frac{\sigma_w^2}{\sigma_b^2} I_{N_u \times N_u} \right)^{-1}.
\]

Note that this approach aims at directly estimating and detecting the bits of all the different users. Since for each bit period we observe \(N_f N_c\) chips, this means that both the above receivers are expected to correctly detect up to \(N_f N_c\) bits, or, equivalently, a single bit from \(N_f N_c\) distinct users.
We want to remark that the FDMUD receivers could be used also in the downlink by considering the same channel for all active users. However, this would present some shortcomings: in particular, each user would require the knowledge of the TH sequences of all the active users, with heavy concerns regarding the system complexity and security.

IV. Simulation Results

The proposed receivers have been tested by simulating a short-range UWB-IR link between a variable number of MTs and an AP. The information bits are modulated by means of a 2-PPM and repeated over $N_f = 4$ frames each consisting of $N_c = 4$ chips, resulting in an uncoded rate of 15.6 Mbit/s. As to the downlink, we have considered generic random generated orthogonal time hopping sequences which are designed so that up to four orthogonal users can be allocated. Since they are transmitted from the same AP, in this case all the users can be assumed synchronous. Conversely, when considering the uplink no particular constraint has been imposed on the properties of the time hopping sequences.

A slow fading scenario has been considered, i.e., the channel coefficients have been assumed constant over a single block of samples. The channel has been simulated according to the model proposed by Cassioli et al. in [11]. Therefore, we considered only the small scale fading statistics in [11], assuming no shadowing and a reference pathloss of 0 dB. For what concerns the power delay profile, we have set the power ratio between the line-of-sight replica and the reflected ones as 0.4. The decaying constant has been chosen so as to yield a rms delay spread of about 50 ns, which is a typical value for indoor environments. This resulted in a digital channel model with 100 sample-spaced resolvable replicas.

Moreover, all users are assumed to be not farther than 7 meters from the AP. Hence, the maximum round trip delay is equal to 50 ns: since the pulse duration $T_w$ is equal to 2 ns, this delay corresponds to 25 samples of the received signal.

According to the operational condition depicted above, when using the proposed frequency domain approach each block of $M = 128$ samples is extended by means of a cyclic prefix of 128 samples, so that the channel and the lack of synchronism do not cause any interference between adjacent blocks. We remark that in this case the actual pulse repetition is halved, that is $N_f^{CP} = N_f / 2 = 2$, so as to maintain the same redundancy as the systems without CP.

Note that the performance of a correlation receiver for a single-user UWB-IR system in an AWGN channel is reported for comparison in all the figures.

A. Performance of FDD Receivers

The bit error rate (BER) for the system using RAKE receiver and the systems using FDD with ZF equalization (FDD-ZF) and MMSE equalization (FDD-MMSE) has been evaluated in the downlink by averaging over 10000 independent channel realizations. Perfect knowledge of the channel parameters has been assumed.

In Fig. 4 we show the comparison of BER performance versus $E_b/N_0$ ratio for a single user communication: though no multiple access interference has been introduced, the long delay spread of the multipath components causes a remarkable level of self-interference between the replicas of the signals. Hence, even if the RAKE receiver exploits all the channel replicas, its performance is bounded by an irreducible error floor that is clearly visible for high values of $E_b/N_0$ ratio\footnote{A lower value of the error floor could be achieved by reducing the self-interference, i.e., increasing the value of $N_u$. This approach, anyway, would reduce on the same time the efficiency of the communication.}. On the other hand, even if FDD-ZF compensates channel effects, and, therefore, does not show any error floor, the noise enhancement caused by zero forcing equalization greatly impairs system performance with a loss of about 10 dB. As it can be clearly seen, FDD-MMSE results to be the best solution since it does not increase the effects of thermal noise while suppressing self-interference and eliminating the error floor.

If we consider a multiuser environment as in Fig. 5 and in Fig. 6, where UWB-IR systems with 2 and 4 users are simulated, respectively, the abilities of FDD-MMSE are even more evident. The RAKE receiver is not able to cope with MAI whose effects are increased by the long multipath spread: as a result, performance is greatly impaired and the error floor is evident also for medium to low $E_b/N_0$ values. On the contrary, both FDD strategies are able to restore the orthogonality between users since they compensate the effects of the channel. While FDD-ZF performance is impaired by the...
noise enhancement, FDD-MMSE performance shows a limited degradation with respect to the single user case; particularly, the performance gain which is afforded by the FDD approach is remarkable for the 4-users system.

Finally, in Fig. 7 the BER performance of the considered systems is shown in the case of system with 8 active users. In this case, the performance of both FDD detectors is impaired by the MAI. Since the number of users is greater than the number of chips $N_c$, no orthogonality can be imposed between the transmitted data and some more advanced detection scheme is needed.

B. Performance of FDMUD Receivers

The BER for the system using the RAKE receiver and the systems using FDMUD with ZF equalization (FDMUD-ZF) and MMSE equalization (FDMUD-MMSE) has been evaluated in the uplink by averaging over 10000 independent channel realizations for each active user. Perfect knowledge of the channel parameters of all users has been assumed.

In Fig. 8 we report the comparison of BER performance versus $E_b/N_0$ ratio for a single user communication. As expected, the RAKE behavior is exactly the same of the downlink case, because of the self-interference which is caused by the multipath components. On the other hand, both FDMUD receivers are effective in facing channel impairments. It is important to highlight that the ZF performance in the uplink system is much better than in the downlink context: as highlighted in Section III, the FDMUD-ZF receiver (see (25)) is based on the estimation of all the transmitted bits and presents a negligible noise enhancement for small system loads.

In Fig. 9 we consider a multiuser environment where a UWB-IR system with 4 users is simulated. While the RAKE receiver performance is greatly impaired and the error floor is evident even for low $E_b/N_0$ values, both FDMUD strategies are effective in compensating the effects of the channel and suppressing the MAI. On the other hand, the performance of the FDMUD-ZF receiver becomes substantially impaired by the MAI – about 2 dB loss with respect to the FDMUD-MMSE – when the system load increases, as reported in Fig. 10, which is relative to 8 active users.

Finally, in Fig. 11 the BER performance of the considered systems is shown in the case of 16 active users, i.e., the number of active users is greater than $N_f^CP N_c$. As expected, when the system load is so high, the performance of both FDMUD receivers is greatly impaired by the MAI. Even if the FDMUD-MMSE receiver is able to suppress the error floor, the large performance degradation confirms the theoretical limitations...
of the FDMUD approach which have been addressed in Section III-B.2.

V. COMPLEXITY ISSUES

A. Sampling Requirements

In the proposed receivers, we have adopted a specific reception chain whose front end, after the receiving antenna, is composed by an analog correlator, namely a pulse deshaper, followed by an Analog-to-Digital Converter (ADC) which provides the samples to form the data-blocks: in particular, followed by an Analog-to-Digital Converter (ADC) which provides the samples to form the data-blocks: in particular, the pulse duration $T_w$, is equal to $2 \text{ ns}$. Therefore, in order to recover all the information, we only need to sample the output of the pulse correlator at 500 MSPS.

The requirements in terms of ADC performance of the proposed receiver architecture are not too demanding, so that this hybrid solution can be seriously considered for implementation. On the other hand, this architecture will suffer from circuit mismatches and other nonidealities. The effects of these impairments on the performance of the proposed receiver could be taken into account by introducing more sophisticated channel and system models. However, this topic is out of the scope of the present paper, which aims at presenting a new detection technique for IR-UWB systems.

B. Computational Complexity

For what concerns the complexity involved, the RAKE receiver appears to be the most simple since its computational load is proportional to the number of finger in the receiver. If we rely on equation (16), the computational complexity of the RAKE receiver in the downlink can be expressed as

$$C_{\text{RAKE}} = L_{\text{RAKE}} \text{ multiplications/bit},$$

where $L_{\text{RAKE}}$ indicates the number of resolvable paths. On the other hand, both FDD detectors are characterized by a higher complexity. Let us consider (18). Its implementation requires two Fourier transforms and $2M$ multiplications every $N_b$ detected bits. Hence, if a fast Fourier transform algorithm is used, the computational load of these detectors can be derived as

$$C_{\text{FDD}} = N_w (\log_2 M + 2) \text{ multiplications/bit.}$$

With the parameters we considered in our system, we would have $C_{\text{RAKE}} = 100$ and $C_{\text{FDD}} = 288$. Although the FDD system is about three times as complex as the RAKE one, this value does not seem prohibitive for future implementations. Moreover, this increasing complexity is well justified by the performance gain of FDD with respect to the RAKE.

As to the uplink, if we consider the complexity per user, the RAKE has the same complexity as in the downlink, since the receiver has to implement one RAKE for each active user. When considering the FDMUD receiver, the complexity can be derived from equation (24). In particular, (24) requires one Fourier transform and $MN_b N_u$ multiplications every block of $N_b N_u$ bits ($N_b$ bits for each of the $N_u$ active users). Therefore, its computational complexity can be expressed as

$$C_{\text{FDMUD}} = N_u (N_b + \frac{\log_2 M}{2N_u}) \text{ multiplications/bit/user.}$$

In the proposed system, we would have $C_{\text{FDMUD}} = 368$ for $N_u = 1$ and $C_{\text{FDMUD}} = 270$ for $N_u = 8$, i.e., a complexity that is comparable to that of the FDD. Moreover, the simulations results show that the RAKE would be totally useless in the uplink, whereas the proposed FDMUD approach offers very good performance even for a high number of active users.

VI. CONCLUSIONS

In this paper, we proposed an original frequency domain detection strategy for high rate short-range impulse radio ultra-wideband systems. The proposed receivers are based on both the introduction of the cyclic prefix and the use of a frequency domain multiuser detector. Detection strategies based on either the ZF or the MMSE criteria have been investigated. The proposed detectors have been compared with the classical RAKE, in a scenario where several mobile terminals communicate with an access point through a severe multipath channel. Simulation results have shown that both the FDMUD strategies are able to suppress the MAI. We found that the MMSE receiver achieves the best performance for any configuration of active terminals, both in the uplink and in the downlink communications. Moreover, the complexity of the proposed approach is proved sustainable. Hence, the proposed
approach is well suited in indoor wireless environments where the multipath propagation tends to increase the effects of both the inter-path and the inter-user interference.

REFERENCES


Proportional QoS Adjustment for Achieving Feasible Power Allocation in CDMA Systems

Rudolf Mathar and Anke Schmeink

Abstract—Resource management is the general topic of the present paper, particularly, we deal with capacity sharing for interference limited wireless networks by power control. Proportional reduction of the signal-to-interference ratio (SIR) requirements is suggested as the control mechanism to accommodate users in the case of overload. For this purpose, we carefully describe the geometrical structure and the asymptotic behavior of the set of feasible power vectors as a proportionality factor tends to its boundaries. In the case that there is no feasible power adjustment, the minimum proportional SIR reduction is determined under general power constraints. We conclude with developing a locally quadratic convergent algorithm for numerical computation of the optimum power assignment. The investigations provide both insight into the theoretical structure of optimum power allocation as well as a practical method for call admission control.

Index Terms—Cellular networks, code division multiple access, resource management, optimal power control, power region, call admission control.

I. INTRODUCTION

POWER control is one of the major ingredients for code division multiple access (CDMA) mobile networks to achieve the potential capacity. The quality-of-service (QoS) performance of users depends on the power assignment in the whole network and usually becomes better with increasing sum power (see [1]). However, in order to save sparse energy for handheld devices, and to keep interference to other stations low, it is desirable that stations transmit with the minimum power such that a required QoS level is just guaranteed.

The existence of some feasible power allocation for a community of transmitters and related problems have been extensively investigated over the last years. The sheer existence of a solution, assuming unlimited power is clarified by Perron-Frobenius theory, as we briefly outline in Section II, and has been used, e.g., in [2]. If the power budget is limited, additional constraints arise.

Three important questions are directly connected to power control. First, for practical applications individual power settings must be computed, favorably in a decentral manner using only local information. In [2] a convergent algorithm is presented which solves this task and simultaneously allocates mobiles to base stations. In an elegant setup, the author [3] develops a general framework for proving convergence of a whole class of power assignment algorithms. For improving efficiency different approaches to dimensionality reduction are used in [4], [5], [6]. Probabilistic aspects of the channel are included in [4], [7], [8].

The second class of problems is concerned with the set of QoS requirements which can be supported by a feasible power assignment. This leads to the concept of user capacity which is investigated by [9], [10], [11] by considering the optimum linear receiver jointly with signature sequences. Convexity, monotonicity and asymptotic properties of the capacity region are themes of the works [12], [13], [14], [15].

Access control by power adjustment is the third type of problem to be solved when operating CDMA mobile radio. In [16], active links are protected when new users are admitted to the network. A fast algorithm to decide if new users can be accommodated while maintaining the required QoS is given in [17]. A novel game theoretic approach to admission control is used in [18]. How this approach relates to the point of least power adjustment via monotonic functionals is shown in [19].

In this work, we approach the problem of admitting new users by proportionally reducing the QoS parameters of all users whenever there is need for. The idea behind this concept is that each user sacrifices a proportional part of his transmission capacity to admit further subscribers to the network. To apply this strategy a graceful degradation of service quality in terms of higher bit error or lower transmission rates must be acceptable to the involved users.

After introducing the system model and some basic preliminaries in Section II we deal with the geometry of the power region. It turns out that the shifted power region is a closed convex cone containing a componentwise minimum power assignment. This element increases monotonically as the proportionality factor does. In Section III, we investigate the orbit of the optimum power assignment by determining derivatives, and also the direction of divergence as the proportionality factor approaches the boundary of the interval where a feasible power allocation exists.

For practical applications power restrictions must be taken into account. In Section IV, we consider the case that power constraints can be described by a certain functional. We present a convergent algorithm for determining the largest proportional QoS vector which allows for a feasible power adjustment. The most common cases such as total and componentwise power constraints are contained as special cases. We conclude with a short summary and possible future extensions in Section V.
II. System Model and Preliminaries

In a synchronous multiuser CDMA communication system with \( K \) users and processing gain \( N \) let \( s_i \in \mathbb{R}^N \), \( i = 1, \ldots, K \), denote the \( N \)-dimensional signature sequence of user \( i \). Let \( G_{ij} \) denote the fixed path gain from user \( j \) to the assigned base station of user \( i \). Usually \( G_{ij} \) is subject to slow fading effects which are assumed to be known to the transmitter. Suppose the symbol of user \( i \) is decoded using a linear receiver represented by some vector \( c_i \in \mathbb{R}^N \). The signal-to-interference ratio of user \( i \) is then given as

\[
\text{SIR}_i(p) = \frac{G_{ii} (c_i^T s_i)^2 p_i}{\sum_{j \neq i} G_{ij} (c_i^T s_j)^2 p_j + \sigma^2 (c_i^T c_i)^2},
\]

where \( \sigma^2 \) denotes the variance of the additive Gaussian noise and \( p = (p_1, \ldots, p_K)^T \) the vector of transmit powers. In the following we assume that the receiver sequences \( c_i \) are fixed. Combining the known channel and receiver effects into \( A_{ij} = G_{ij} (c_i^T s_j)^2 \) we obtain \( \text{SIR}_i(p) \) of user \( i \) as

\[
\text{SIR}_i(p) = \frac{A_{ii} p_i}{\sum_{j \neq i} A_{ij} p_j + C_{ii} \sigma^2},
\]

with \( C_{ii} = (c_i^T c_i)^2 \). Now given QoS requirements \( \gamma_1, \ldots, \gamma_K \) for each user, we define the power region \( P_{\text{SIR}}(\gamma) \) as the set of power settings \( p \in \mathbb{R}^K \) such that each user \( i \) meets his SIR requirement \( \gamma_i \), i.e.,

\[
P_{\text{SIR}}(\gamma) = \{ p \in \mathbb{R}^K | \text{SIR}_i(p) \geq \gamma_i, \ i = 1, \ldots, K \}. \tag{1}
\]

Here and in the following orderings ‘<’ and ‘\( \leq \)’ between vectors are always meant componentwise. Obviously it may happen that not all requirements \( \gamma_i \) can be simultaneously satisfied, in which case \( P_{\text{SIR}}(\gamma) \) is empty.

For convenience of notation we quote the following result from [15]. It deals with solutions of the equation

\[
[I - A]x = c
\]

when \( A \) is a nonnegative but not necessarily irreducible matrix. The proof given in [15] is direct and self-contained, and does not rely on the Perron-Frobenius theory. In the irreducible case the result is well known from [20]. Let \( \rho(A) \) denote the spectral radius of a square matrix \( A \), defined as \( \rho(A) = \max \{|\lambda_i(A)|\} \), where \( \lambda_i(A) \) denotes the complex eigenvalues of \( A \).

**Proposition 1:** Let \( A \in \mathbb{R}^{n \times n} \) be non-negative.

a) If there are \( x > 0 \), \( c > 0 \) satisfying (2), then \( \rho(A) < 1 \).

b) If \( \rho(A) < 1 \), then \( I - A \) is non-singular and for every \( c > 0 \), the unique solution \( x \in \mathbb{R}^n \) of (2) is positive.

c) If \( \rho(A) < 1 \), then for every \( c \geq 0 \), the unique solution \( x \in \mathbb{R}^n \) of (2) is non-negative.

d) If \( c > 0 \) and there exists \( y > 0 \) such that \( (I - A)y \geq c \), then (2) has a unique solution \( x \) and \( 0 < x \leq y \).

The above is now applied to \( P_{\text{SIR}}(\gamma) \). The inequalities defining (1) can be rewritten as a system of linear inequalities. For this purpose write \( B = (b_{ij})_{K \times 1} \), with

\[
b_{ij} = \begin{cases} A_{ij}/A_{ii}, & i \neq j, \\ 0, & i = j, \end{cases}
\]

and \( \tau = (\tau_1, \ldots, \tau_K)^T \), where \( \tau_i = C_{ii} \sigma^2 / A_{ii} \). Then for every \( p > 0 \) it holds that \( p \in P_{\text{SIR}}(\gamma) \) if and only if

\[
[I - GB]p \geq \Gamma \tau,
\]

where \( \Gamma = \text{diag}(\gamma) \) denotes the matrix with diagonal entries \( \gamma_i \) and nondiagonal entries equal to zero.

If \( \Gamma \tau > 0 \) and system (3) has a solution \( p > 0 \), then there is a unique solution \( p^* \leq p \) satisfying

\[
[I - GB]p^* = \Gamma \tau,
\]

as follows from Proposition 1. Moreover, for any given \( \gamma > 0 \), the equation \( [I - GB]p = \Gamma \tau \) has a positive solution \( p \) if and only if the spectral radius \( \rho(GB) < 1 \), and in that case, the solution is unique. Denote it by \( p^*(\gamma) = (\gamma_1(\gamma), \ldots, \gamma_K(\gamma))^T \). Thus

\[
p^*(\gamma) = [I - GB]^{-1} \Gamma \tau \tag{4}
\]

with all components positive.

Summarizing the above arguments, we see that there is a unique componentwise minimum power allocation in \( P_{\text{SIR}}(\gamma) \), provided the power region is nonempty, see also Theorem 2 in [2].

**Proposition 2:** If \( P_{\text{SIR}}(\gamma) \neq \emptyset \), then there is a unique power allocation \( p^* = p^*(\gamma) \) such that \( \text{SIR}_i(p^*) = \gamma_i \) for all \( i = 1, \ldots, K \) and \( p^* \leq p \) for all \( p \in P_{\text{SIR}}(\gamma) \).

The shifted power region \( P_{\text{SIR}}(\gamma) - p^*(\gamma) \) has a nice geometrical structure which is important for finding, e.g., the projection of inadmissible points onto \( P_{\text{SIR}}(\gamma) \). For completeness we recall the following definition. A set \( C \) in a linear vector space is said to be a cone if \( c \in C \) implies that \( \alpha c \in C \) for all \( \alpha \geq 0 \) (see, e.g., [21]).

**Proposition 3:** \( P_{\text{SIR}}(\gamma) - p^*(\gamma) \) is a closed convex cone. Consider the sets (eq. (19))

\[
P_i = \{ p | A_{ii} p_i - \gamma_i \sum_{j \neq i} A_{ij} p_j \geq \gamma_i C_{ii} \sigma^2, \ i = 1, \ldots, K \},
\]

which are closed convex affine halfspaces in \( \mathbb{R}^K \). Obviously, \( P_{\text{SIR}}(\gamma) = \bigcap_{i=1}^K P_i \), and from Theorem C in Section III of [22] it follows that \( P_{\text{SIR}}(\gamma) \) is a closed and convex polytope.

To prove that \( P_{\text{SIR}}(\gamma) - p^*(\gamma) \) is a cone, we show that \( \text{SIR}_i(p^*(\gamma) + \alpha [p - p^*(\gamma)]) \geq \gamma_i \), and hence \( p^*(\gamma) + \alpha [p - p^*(\gamma)] \in P_{\text{SIR}}(\gamma) \) for any \( p \in P_{\text{SIR}}(\gamma) \) and \( \alpha \geq 0 \). By assumption \( \text{SIR}_i(p) \geq \gamma_i \) for all \( i = 1, \ldots, K \), in detail,

\[
A_{ii} p_i - \gamma_i \sum_{j \neq i} A_{ij} p_j \geq \gamma_i C_{ii} \sigma^2, \ i = 1, \ldots, K.
\]

Denote by \( a_{i(\gamma)} = (-\gamma_i A_{ii}, \ldots, -\gamma_i A_{iK})^T \) and \( c_i = \gamma_i C_{ii} \sigma^2 \). Then \( a_{i(\gamma)}^T p \geq c_i \) and \( a_{i(\gamma)}^T p^*(\gamma) = c_i \), for all \( i = 1, \ldots, K \). It follows that

\[
a_{i(\gamma)}^T [p^*(\gamma) + \alpha [p - p^*(\gamma)]] \geq a_{i(\gamma)}^T p^*(\gamma) = c_i,
\]

for all \( i = 1, \ldots, K \) and all \( \alpha \geq 0 \), and hence \( p^*(\gamma) + \alpha [p - p^*(\gamma)] \in P_{\text{SIR}}(\gamma) \) for all \( \alpha \geq 0 \).

A related result, however, not including convexity and without shifting \( P_{\text{SIR}}(\gamma) \) to the origin is derived in [23].

The uniformly minimal point \( p^*(\gamma) \in P_{\text{SIR}}(\gamma) \) is of particular interest since it requires minimal power while maintaining the SIR demands \( \gamma = (\gamma_1, \ldots, \gamma_K)^T \) of all users. In the
following we deal with the behavior of \( p^*(\gamma) \) as a function of \( \gamma \).

**Proposition 4:** The function \( p^*(\gamma) \) is monotonically increasing, i.e., if \( \mathcal{P}_{\text{SIR}}(\gamma^{(2)}) \neq \emptyset \) and \( \gamma^{(1)} \leq \gamma^{(2)} \), then \( p^*(\gamma^{(1)}) \leq p^*(\gamma^{(2)}) \). Furthermore, \( p^*(\gamma) \to 0 \) as \( \gamma \to 0 \).

From Proposition 1 it follows that \( \rho(\Gamma^2 B) < 1 \). Hence, expanding representation (4) in a von Neumann series gives

\[
p^*(\gamma^{(1)}) = [I - \Gamma^1 B]^\dagger \Gamma^1 \tau = \sum_{i=0}^\infty (\Gamma^1 B)^i \Gamma^1 \tau \leq \sum_{i=0}^\infty (\Gamma^2 B)^i \Gamma^2 \tau = [I - \Gamma^2 B]^\dagger \Gamma^2 \tau = p^*(\gamma^{(2)}),
\]

where \( \Gamma^i = \text{diag}(\gamma^i) \).

It is immediate from (4) that \( p^*(\gamma) \to 0 \) as \( \gamma \to 0 \). Observe that \( [I - \Gamma B]^{-1} \) exists in a sufficiently small neighborhood of 0.

Further conditions for strict monotonicity of \( p^*(\gamma) \) are given in [15]. It should be mentioned that for a different, but related model the generic concept of \( p^*(\gamma) \) as a one-dimensional manifold and its monotonicity are also considered in [24].

### III. PROPORTIONAL POWER ADJUSTMENT

As we have seen in the previous section, \( \mathcal{P}_{\text{SIR}}(\gamma) \neq \emptyset \) whenever \( \gamma \) is sufficiently small. If the requirements \( \gamma \) of a community of users are such that there is no power allocation, i.e., \( \mathcal{P}_{\text{SIR}}(\gamma) = \emptyset \), access control becomes inevitable. A rational concept is to request that each user sets aside a proportional fraction of his requirement until a feasible power allocation can be found. This concept quantifies to a certain extent the notion of graceful degradation of CDMA as discussed in [25] and leads to investigating the behavior of

\[
p^*(\alpha) = [I - \alpha \Gamma B]^{-1} \alpha \Gamma \tau, \quad \alpha > 0.
\]

By Proposition 1, the point \( p^*(\alpha) \) exists whenever \( \alpha < 1/\rho(\Gamma B) \). This breakpoint can be described as the solution of the following max-min problem.

**Proposition 5:** Let \( \gamma > 0 \) be fixed and \( B \) be irreducible. A proportional SIR requirement vector \( \alpha \gamma \), \( \alpha > 0 \), allows for a feasible power assignment if and only if

\[
\alpha < \sup_{p > 0} \min_{i=1, \ldots, K} \frac{\text{SIR}_i(p)}{\gamma_i}.
\]

The only point remaining is to show that the right hand side of (5) coincides with \( 1/\rho(\Gamma B) \). For this purpose we use Corollary 8.1.31 in [26], stating that for any irreducible nonnegative matrix \( C = (c_{ij})_{i,j=1,...,K} \) the spectral radius is given by

\[
\rho(C) = \min_{p > 0} \max_{i=1, \ldots, K} \frac{1}{p_i} \sum_{j=1}^K c_{ij} p_j.
\]

Some elementary algebra gives that

\[
\frac{\gamma_i}{\text{SIR}_i(p)} = \frac{\sum_j c_{ij} p_j + \gamma_i \tau_i}{p_i},
\]

with \( C = \Gamma B \). Hence,

\[
\inf_{p > 0} \max_i \frac{\gamma_i}{\text{SIR}_i(p)} = \min_{p > 0} \max_i \frac{1}{p_i} \sum_{j=1}^K c_{ij} p_j = \rho(C) = \rho(\Gamma B),
\]

which yields the assertion by considering the reciprocal value.

To describe the set \( \{ p^*(\alpha) > 0 < \alpha < 1/\rho(\Gamma B) \} \) the componentwise derivative with respect to \( \alpha \) is important.

**Proposition 6:** For any \( 0 < \alpha < 1/\rho(\Gamma B) \), the derivative of \( p^*(\alpha) = [I - \alpha \Gamma B]^{-1} \alpha \Gamma \tau \) is given by

\[
\frac{d}{d\alpha} p^*(\alpha) = [I - \alpha \Gamma B]^{-2} \Gamma \tau.
\]

First write \( p^*(\alpha) = \left[\frac{1}{\alpha} I - \Gamma B \right]^{-2} \Gamma \tau \). The derivative of the inverse is given by (see, e.g., [27])

\[
\frac{d}{d\alpha} \left[\frac{1}{\alpha} I - \Gamma B \right]^{-1} = \frac{1}{\alpha^2} \left[\frac{1}{\alpha} I - \Gamma B \right]^{-2} = [I - \alpha \Gamma B]^{-2}.
\]

By linearity, (6) follows from multiplication by \( \Gamma \tau \).

Since \( [I - \alpha \Gamma B]^{-1} \) consists of nonnegative entries, \( \frac{d}{d\alpha} p^*(\alpha) \) in (6) is nonnegative for each component, in accordance with the componentwise monotonicity property in Proposition 4.

If power is not the limiting factor, as may be assumed approximately true for the downlink in cellular networks, it is relevant to determine how the power adjustment \( p^*(\alpha) \) diverges as the quality profile \( \alpha \gamma \) tends to its limit at \( \alpha = 1/\rho(\Gamma B) \). This is important if power consumption is of no concern and the main objective is to achieve best possible performance.

Since \( \gamma \) is fixed in the sequel we may reparametrize \( \beta = 1/\alpha \) and write

\[
p^*(\beta) = [\beta I - \Gamma B]^{-1} \Gamma \tau,
\]

provided that \( \beta > \rho(\Gamma B) \).

If \( B \) is irreducible, in [28], p. 315, and [29], (2.3), the following spectral representation of \( [\beta I - A]^{-1} \) for the nonnegative matrix \( A = \Gamma B \) is shown. Let \( \{\lambda_1 = \rho(A), \lambda_2, \ldots, \lambda_K\} \) denote the spectrum of \( A \) and \( m(\lambda) = (\lambda - \rho(A)) \prod_{k=2}^K (\lambda - \lambda_k)^{m_k} \) the minimal polynomial of \( A \). Then for \( \beta > \rho(A) \) it holds that

\[
[\beta I - A]^{-1} = \frac{1}{\beta - \rho(A)} xy^T + \sum_{k=2}^K \sum_{j=1}^{m_k} \left( j - 1 \right) \frac{1}{(\beta - \lambda_k)^j} Z_{kj}.
\]

where \( x \) and \( y \) denote the right and left Perron vectors of \( A \), respectively, and \( Z_{kj} \) are fixed matrices, known as principal component matrices. The right and left Perron vector are defined as the the right and left eigenvectors of \( A \) with positive components corresponding to eigenvalue \( \rho(A) \), the spectral radius.

From representation (7) the order and direction of divergence can be easily deduced. As \( \beta \to \rho(A) \) the first term \( \frac{1}{\beta - \rho(A)} xy^T \) tends componentwise to infinity while \( R(\beta) \) tends to a fixed matrix \( R(\rho(C)) \). Note that \( \rho(A) > \lambda_k \) for
all $k = 2, \ldots, K$. Hence, multiplication by $\Gamma \tau$ yields in the limit as $\beta \to \rho(A)$ that

$$p^*(\beta) - \frac{y^\tau \Gamma \tau}{\beta - \rho(A)} x - R(\rho(A)) \Gamma \tau \to 0.$$  

The interpretation is that $p^*(\beta)$ diverges to infinity along direction $x$ up to the constant shift vector $R(\rho(A)) \Gamma \tau$. In summary we obtain the following result.

**Proposition 7:** It holds that

$$\lim_{\beta \to \rho(B)} \left( p^*(\beta) - \frac{y^\tau \Gamma \tau}{\beta - \rho(B)} x \right) = R(\rho(B)) \Gamma \tau.$$  

Hence, $p^*(\beta)$ diverges as $\beta \to \rho(\Gamma B)$. The factor $\frac{y^\tau \Gamma \tau}{\beta - \rho(B)}$ represents the order of divergence, the right Perron vector $x$ is the direction of divergence, and $R(\rho(\Gamma B)) \Gamma \tau$ is a constant deviation between $p^*(\beta)$ and the limiting direction.

Spectral representation (7) is also employed in [12] for analyzing the asymptotic behavior of the minimum total power.

Our results so far are visualized in Figure 1 for $K = 2, A_{11} = 0.6, A_{22} = 1, A_{12} = 0.1, A_{21} = 0.3, \gamma_i = 3.16 W, C_i = \sigma^2 = 10^{-7} W, i = 1, 2$. The orbit $p^*(\beta)$ is shown as a parametric plot of $\beta \in (\rho(\Gamma B), \infty)$. The convex cones $\mathcal{P}_{\text{SIH}}(\frac{1}{2} \gamma)$ are depicted for three cases $\beta = 0.9, 1.2, 2$. Divergence along a fixed direction can be clearly recognized as $\beta \to \rho(\Gamma B)$.

**IV. ACCESS CONTROL WITH LIMITED POWER BUDGET**

In practice, power is limited, particularly for small handsets as are used for the uplink. Hence, mobiles may select their power adjustment only from a bounded set $P$, say. To be quite general, we describe power constraints by help of a convex function $g : \mathbb{R}^K \to \mathbb{R}$ satisfying $g(0) = 0$. The set of feasible power allocations is then defined as

$$\mathcal{P}_{\text{feas}} = \{ p \geq 0 \mid g(p) \leq m \}$$

for some positive threshold $m > 0$. It is clear that $\mathcal{P}_{\text{feas}}$ is a convex set which contains with each power assignment $p > 0$ any componentwisely smaller $q > 0$.

Important special cases are covered by this approach. Individual peak power constraints $p_i \leq m_i$, $i = 1, \ldots, K$, evolve from choosing $g(p) = \max_i \frac{p_i}{m_i}$ and $m = 1$. Total power restrictions like $\sum_i p_i \leq m$ follow by setting $g(p) = \sum_i p_i$. Both are special cases of general restrictions of the form $g(p) = \| p \|$ for some norm $\| \cdot \|$. We will deal with this case later under numerical aspects. The intersection of peak and sum power constraints is also covered by the present general approach.

Let $\gamma > 0$ be given and assume that $p^*(\gamma) \notin \mathcal{P}_{\text{feas}}$, as visualized in Figure 2 for $K = 2$ and parameters according to Figure 1. We search for a minimum proportional reduction of $\gamma$ by some $0 \leq \alpha \leq 1$ such that $p^*(\alpha \gamma) \in \mathcal{P}_{\text{feas}}$. With $\beta = 1/\alpha$ it is clear from the above that a solution of the system

$$(\beta I - \Gamma B)p = \Gamma \tau$$  

$$g(p) = m$$

(8)

with variables $p$ and $\beta$ is sought. Observe that a solution exists and is unique, however, in general it is hard to determine. For this purpose rewrite (8) as

$$F : (\rho(\Gamma B), \infty) \to \mathbb{R} : \beta \mapsto g([\beta I - \Gamma B]^{-1} \Gamma \tau) - m. \quad (9)$$

Seeking the roots of $F$ yields the solution $\beta^* \text{ of (8)}$.

In the case that $g$ is continuously differentiable Newton’s Method is a favorite candidate for finding a solution $\beta^*$ of (9) as follows,  

$$\beta_{n+1} = \beta_n + \Delta \beta_n, \quad \Delta \beta_n = -\frac{F(\beta_n)}{F'(\beta_n)}. \quad (10)$$

Algorithm (10) converges locally quadratic, see, e.g., [30].

Using the chain rule for multivariate functions and setting
\( p(\beta) = [\beta I - GB]^{-1} \Gamma \) gives
\[
\frac{d}{d\beta} F(\beta) = \frac{d}{d\beta} g(u) \bigg|_{u=p(\beta)} \frac{d}{d\beta} p(\beta) \\
= -\frac{d}{du} g(u) \bigg|_{u=p(\beta)} [\beta I - GB]^{-2} \Gamma ,
\]

(11)

We evaluate (11) for the \( \ell_q \)-norms, \( 1 \leq q \leq \infty \), i.e.,
\( \|u\|_q = \left( \sum_{i=1}^{K} u_i^q \right)^{1/q}, \ u > 0 \). The derivative of \( g(u) = \|u\|_q \) is given by
\[
\frac{d}{du} g(u) = \frac{d}{du} \|u\|_q = \|u\|^{1-q} \left( u_1^{q-1}, \ldots, u_K^{q-1} \right).
\]

Abbreviating componentwise exponentiation as
\( \left(u_1^{q-1}, \ldots, u_K^{q-1} \right) = (u^q) \), Newton's iteration (10) becomes
\[
\beta_{n+1} = \beta_n + \frac{\|p(\beta_n)\|_q - m}{\|p(\beta_n)\|_q} \beta_n - m \frac{\|p(\beta_n)\|_q - m}{\|p(\beta_n)\|_q} \beta_n + \frac{\|p(\beta_n)\|_q - m}{\|p(\beta_n)\|_q} \beta_n,
\]

(12)
where \( p'(\beta) = \frac{d}{d\beta} p(\beta) \) means the column vector of componentwise derivatives w.r.t. \( \beta \).

Important special cases are \( q = 1 \), describing a total limited power budget (cf. [12]), and \( q = 2 \). In the case \( q = 1 \) iteration (12) reads as
\[
\beta_{n+1} = \beta_n - \frac{\|p(\beta_n)\|_q - m}{(1, \ldots, 1)^T p'(\beta_n) = \beta_n - \frac{\|p(\beta_n)\|_q - m}{(1, \ldots, 1)^T p'(\beta_n)}.
\]
since \( p'(\beta_n) \leq 0 \) by Proposition 4. For \( q = 2 \) we obtain from (12)
\[
\beta_{n+1} = \beta_n - \frac{\|p(\beta_n)\|_q - m}{(1, \ldots, 1)^T p'(\beta_n)}.
\]
A numerical problem in implementing algorithms (10), or (12), respectively, lies in computing \( p(\beta_n) = [\beta_n I - \Gamma B]^{-1} \Gamma \) and \( p'(\beta_n) = -[\beta_n I - \Gamma B]^{-2} \Gamma \). The following algorithm converges towards \( p(\beta_n) \). The i-th component in the iteration \( p_n(i) \) is given by
\[
p_{n,i}(k) = \frac{\gamma_n}{\beta_n} (\xi_i + \sum_{j \neq i} A_{ij})(p_{n,j}(k - 1)) - 1,
\]
(13)
i = 1, \ldots, K, k \in \mathbb{N}. If \( \beta_n > \rho(\Gamma B) \), then it holds that \( \lim_{k \to \infty} p_n(k) = p(\beta_n) \) for any initial value \( p_n(0) \), as is shown in the appendix.

Similarly, the following sequence \( p'_n(k) \) converges towards \( p'(\beta_n) \),
\[
p'_n(k) = \left( \frac{2}{\beta_n} GB - \frac{1}{\beta_n^2} (GB)^2 \right) p'_n(k-1) + \frac{1}{\beta_n^2} \Gamma \tau,
\]
(14)
k \in \mathbb{N}. If \( \rho(\frac{2}{\beta_n} GB - \frac{1}{\beta_n^2} (GB)^2) < 1 \), then \( \lim_{k \to \infty} p'_n(k) = p'(\beta_n) \) holds for any \( n \) and any initial value \( p'_n(0) \).

We conclude this section by evaluating the proposed algorithm on the example depicted in Figure 1 and 2 for sum power constraints. Table 1 shows the results, comparison is fast giving an optimum value \( \beta^* = 1.1469 \) with corresponding power allocation \( p^*(\beta^*) = 10^{-6}(0.944, 1.056)^T \) after 6 iterations with a relative error of \( 10^{-6} \). The results \( p(\beta_n) \) of the iterations \( n = 0, 1, \ldots, 5 \) are also indicated by open circles in Figure 2.

<table>
<thead>
<tr>
<th>( n )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_n )</td>
<td>0.9</td>
<td>1.0084</td>
<td>1.1032</td>
<td>1.1426</td>
<td>1.1468</td>
</tr>
<tr>
<td>( 10^6 p(\beta_n) )</td>
<td>(2.061)</td>
<td>(1.348)</td>
<td>(1.041)</td>
<td>(0.953)</td>
<td>(0.944)</td>
</tr>
</tbody>
</table>

V. Conclusions

This paper has introduced the concept of proportional SIR reduction for a community of users if there exists no feasible power adjustment for given transmission rate requirements. We have derived an algorithm to determine the point of minimal reduction numerically for a rather general class of power restrictions, including the case of limited total power. To achieve these results we have investigated the geometrical structure of the set of admissible power assignments, and furthermore, monotonicity and asymptotic properties when the proportionality factor tends to its boundaries.

Interesting open problems are the development of algorithms for a nonsmooth boundary of the power restrictions, and decentralizing the computation such that power assignments can be determined locally with only small global information exchange.

Appendix

We complement the convergence proofs of (13) and (14). As \( \beta_n I - GB \) is an M-matrix, we can apply Jakobi's method which is convergent for all initial vectors, see [31]. Equation (13) is the i-th component of
\[
p_n(k) = \frac{1}{\beta_n} GB p_n(k-1) + \frac{1}{\beta_n} \Gamma \tau.
\]
Successive application yields
\[
p_n(k) = \left( \frac{1}{\beta_n} GB \right)^k p_n(0) + \sum_{i=0}^{k-1} \left( \frac{1}{\beta_n} GB \right)^i \Gamma \tau.
\]
(15)
As \( \beta_n > \rho(\Gamma B) \) by assumption, using the von Neumann Series yields \( (I - \frac{1}{\beta_n} GB)^{-1} = \sum_{i=0}^{\infty} \left( \frac{1}{\beta_n} GB \right)^i \). In particular, it holds that
\[
\lim_{k \to \infty} \left( \frac{1}{\beta_n} GB \right)^k = 0 \quad \text{and} \quad \lim_{k \to \infty} \left( \sum_{i=0}^{\infty} \left( \frac{1}{\beta_n} GB \right)^i \Gamma \tau = p_n.
\]
(16)
Thus, we get \( \lim_{k \to \infty} p_n(k) = p(\beta_n) \) which concludes the proof of (13). Equation (14) is shown in a similar way noting that \( (\beta_n I - GB)^{-2} = (\beta_n^2 I - (2\beta_n GB - (GB)^2))^{-1} \).

References

Rudolf Mathar received his Diploma and Ph.D. degree in mathematics from RWTH Aachen University in 1978 and 1981, respectively. Previous positions include a research fellowship at Augsburg University and a lecturer position at the European Business School. In 1989, he joined the Faculty of Natural Sciences at RWTH Aachen as a Professor of Stochastics and Computer Science. He has held the International IBM Chair in Computer Science at Brussels Free University in 1999. In spring 2001 and 2008, he was invited as Erskine Fellow to Canterbury University, Christchurch, New Zealand. In 2002, he was the recipient of the Vodafone D2 Innovation Award. In 2004 he was appointed head of the Institute of Theoretical Information Technology in the Faculty of Electrical Engineering and Information Technology at RWTH Aachen University. His research interests include mobile communication systems, particularly optimization, resource allocation and access control, as well as radio network information theory.

Anke Schmeink received her Diploma degree in mathematics and the Ph.D. degree in electrical engineering and information technology from RWTH Aachen University, Germany, in 2002 and 2006, respectively. From September to December 2005, she was a Visiting Researcher with the Department of Electrical and Electronic Engineering, The University of Melbourne. She received the E-plus best dissertation award 2006 and the Vodafone young scientists award 2007. She is currently a Research Scientist with the Medical Signal Processing Group at Philips Research Laboratories. Her research interests are in information theory, game theory, and communications theory with current focus on wireless communication networks and medical applications.
The Impact of Space Division Multiplexing on Resource Allocation: A Unified Treatment of TDMA, OFDMA and CDMA

Iordanis Koutsopoulos, Member, IEEE, and Leandros Tassiulas, Fellow, IEEE

Abstract—Space division multiple access (SDMA) with an antenna array at the transmitter is a promising means for increasing system capacity and supporting rate-demanding services. However, the presence of an antenna array at the physical layer raises significant issues at higher layers. In this paper, we attempt to capture the impact of SDMA on access layer channel allocation, reflected on channel reuse. This impact obtains different twists in TDMA, CDMA and OFDMA due to the different nature of co-channel and cross-channel interference and the different interaction of user spatial channel characteristics with system channels, namely time slots, codes and subcarriers. We consider these access schemes in a generalized unified framework and propose heuristic algorithms for channel allocation, downlink beamforming and transmit power control so as to increase total Provisioned system rate and provide QoS to users in the form of minimum rate guarantees. We study the class of greedy algorithms that rely on criteria such as induced or received interference and signal-to-interference ratio (SIR), and a class of SIR balancing algorithms. Results show superior performance for SIR balancing resource allocation and expose the performance benefits of cross-layer design.

Index Terms—Beamforming, SDMA, power control, cross-layer design, resource allocation, T/C/OFDMA.

I. INTRODUCTION

The fundamental problem in wireless networks is to provide diverse quality of service (QoS) guarantees to users in the inherently volatile wireless medium by using limited resources. QoS at the physical layer is perceived as an acceptable signal-to-interference and noise ratio (SINR) or bit error rate (BER) at the user receiver, while QoS at the access and network layers implies provisioning of minimum rate or maximum delay guarantees. The fulfillment of QoS requirements relies on control mechanisms at different layers, such as scheduling, channel allocation, transmission power control and modulation level adaptation [2]. Existing and envisioned wireless systems use Time, Code or Orthogonal Frequency Division Multiple Access (TDMA, CDMA, OFDMA). The work in [3] is a representative one of joint treatment of channel allocation and power control in TDMA/FDMA systems. In currently employed CDMA-based 3G UMTS networks, the channels are spreading codes that are either randomly generated or designed in a deterministic manner to maximize user capacity and minimize total squared cross-correlation (TSC) [4]. Multi-code structures and spreading gain adaptation further aid transmission rate control and provisioning of diverse rates [5].

OFDMA is included in the IEEE 802.11ad/g WLAN standards and is also considered in the IEEE 802.15.3 standard for wireless personal area networks (WPANs) and in the evolving IEEE 802.16x WiMAX standards for broadband wireless access. In OFDMA, the spectrum is divided into narrow-band subcarriers with overlapping spectra, which are orthogonal since they are appropriately spaced. The user bit stream is split into subsets, and each bit subset is called a subsymbol. Each subsymbol modulates a subcarrier and several subsymbols of a user are transmitted in parallel over subcarriers. OFDM transmission reduces effective symbol transmission rate and alleviates the effects of inter-symbol interference (ISI). A basic problem in OFDMA is to allocate subcarriers, transmission powers and rates to users so as to maximize information-theoretic rate. For a given subcarrier allocation this problem is solved by water-filling. For single-cell systems, the work [6] formulates this allocation problem as an integer programming one and finds a suboptimal solution for the continuous-valued problem. The same objective with a total power constraint over all users is achieved by assignment of each subcarrier to the user with the largest gain in it and subsequent power water-filling [7]. A continuous relaxation approach is applied in [8] for the dual problem of minimizing transmission power subject to minimum rate constraints. Heuristic algorithms for channel allocation, modulation and power control for a multicell OFDM system are presented in [9].

Adaptive antenna arrays promise to offer multiples of currently achieved data rates through Space Division Multiple Access (SDMA) [10], [11]. SDMA with a connection-oriented or connectionless access scheme allows reuse of conventional channels by spatially separable users. Within a channel, multiple beams are formed by the transmitter or receiver antenna array, with the main lobe of each beam steered to the direction of a desired user and nulls placed to directions of interferers. The objective is to separate co-channel users, that is, ensure acceptable SINRs at each user receiver. At reception, user separation can be achieved by computing one beam per user independently of others, while at transmission this issue becomes cumbersome since (i) each user beam
affects interference at all other receivers and (ii) receivers are not collocated so that all user signals are jointly detected. Even if user receivers are equipped with multiple antennas or employ multi-user detection, extraction of a user signal takes place at each receiver separately. Downlink beamforming for power minimization in a single cell is studied in [12], where beamforming and power control are decoupled. In [13], the authors study beamforming for single-user OFDM transmission with multiple antennas at the transmitter and the receiver with the objective to maximize SINR. In [14], an iterative algorithm for transmit power control and receive beamforming for the uplink is proposed for a set of co-channel links, each with a minimum SINR requirement. The algorithm converges to a feasible solution if there exists one, and this solution minimizes total transmit power. In [15], the corresponding problem for the downlink is transformed to an equivalent problem for the uplink and is solved with the method of [14]. This approach cannot a priori detect infeasible instances where target SINRs cannot be reached. In [16], [17], an iterative algorithm for downlink beamforming and power control is presented, which always converges to the maximum common scaled SINR (scaled by each link SINR target). The employment of antenna arrays introduces novel challenges at higher layers and recent works begin to address them. For an SDMA/TDMA system, heuristics for time slot assignment and user scheduling subject to deadlines are proposed in [18], [19]. In [20], subcarrier allocation, modulation control and beamforming are addressed for an OFDMA/SDMA system, where an algorithm for constructing co-channel user sets with large subcarrier rate is presented when channel reuse is allowed. Beamforming is also viewed there as an additional dimension to enhance user SINR for the case of no channel reuse. An effort to study distributed resource allocation in a multi-cell SDMA system is made in [21]. In a multi-channel system, channel allocation is coupled with physical layer beamforming and power control. Different users experience different quality in different channels, and spatial separability of users in a channel depends on beamforming, power control and user spatial channel characteristics in that channel. A given user allocation to channels constrains the choice of beamforming and power control, and vice versa. A channel allocation is efficient if it leads to beamforming and transmit power instances with good physical and access layer QoS. In this work, we investigate the impact of SDMA on access layer channel allocation for the downlink of a single cell, with the objective to increase total achievable rate while providing user minimum rate guarantees. Our contribution to the current literature is summarized as follows: (i) we adopt a generalized framework for channel allocation, out of which TDMA, CDMA and OFDMA emerge as special cases, (ii) we incorporate downlink multi-user beamforming and power control in our approach and set up the framework for inclusion of transmission rate adaptation, (iii) we present and compare two classes of heuristic algorithms for identifying spatially separable co-channel sets of users, which can be viewed as instances of cross-layer design, since they employ physical and access layer mechanisms and strive to ensure acceptable QoS at both layers. Specifically, we focus on the class of greedy algorithms with assignment criteria that rely on induced and received interference or user SIR, as well as on the class of SIR balancing channel allocation algorithms. The goals of our study are to identify the structure of algorithms in each multiple access scheme and to demonstrate the benefits of cross-layer design in terms of achievable rate. The intense interest in adaptive antenna arrays is denoted by ongoing standardization efforts in the IEEE 802.11n standard for high throughput in conjunction with OFDMA [22]. Furthermore, multi-user beamforming is incorporated in the evolving cellular standards Cdma2000 1xEV-DO (Ultra Mobile Broadband) Revision C [23] and UMTS Long-Term Evolution (LTE) [24] for achieving high data rates through intra-cell channel reuse. The rest of the paper is organized as follows. In section II we present the model and in section III we give the rationale of our approach and proposed algorithms. Section IV includes numerical results and section V concludes our study. A few words about notation. Vectors and matrices are set in boldface. The cardinality of set $\mathcal{X}$ is $|\mathcal{X}|$. Superscripts $(\cdot)^T$, $(\cdot)^*$, $(\cdot)^H$ denote transpose, complex conjugate and conjugate transpose and $||\mathbf{u}|| = \sqrt{\sum_{i=1}^{n}|u_i|^2}$ is the $\ell_2$-norm of complex vector $\mathbf{u} = (u_1, \ldots, u_n)$ respectively. The dominant generalized eigenvector of matrix pair $(\mathbf{A}, \mathbf{B})$, $\mathbf{u}_{\text{max}}(\mathbf{A}, \mathbf{B})$ is the normalized eigenvector that corresponds to the largest positive eigenvalue of problem $\mathbf{A}\mathbf{x} = \lambda \mathbf{B}\mathbf{x}$. When $\mathbf{A}, \mathbf{B}$ are symmetric and positive-definite, the above is equivalent to $\mathbf{C}y = \lambda \mathbf{y}$, with $\mathbf{C} = \mathbf{L}^{-1}\mathbf{A}(\mathbf{L}^{-1})^H$ and $\mathbf{y} = \mathbf{L}^Hx$, where $\mathbf{L}$ is a non-singular lower-triangular matrix from Cholesky decomposition of $\mathbf{B}$, $\mathbf{B} = \mathbf{LL}^H$.

II. System Model

We consider the downlink of a single-cell system with $K$ users, operating in a frequency band of certain bandwidth. Depending on the access scheme (TDMA, CDMA or OFDMA), transmission occurs in channels that can be time slots, codes or subcarrier frequencies within the specified bandwidth and time frame. The base station (BS) has a uniform linear array of $M$ antennas. Each user receiver has an omni-directional antenna. Packetized data arrive from higher layers and are decomposed into bits before transmission with an underlying slotted scheme. A fixed number of symbols, $S$ are transmitted in a slot of duration $T$, and the symbol (signaling) period is $T$. A user $k$ has minimum rate requirement of $r_k$ bits/sec over a certain time interval $(0, t)$, which is mapped to a minimum number of required channels $x_k$ for single-rate transmission. The block diagram of a generic SDMA transmitter is depicted in Figure 1(a). The channel allocation module allocates channels to each user and identifies co-channel user sets for each channel. Beamforming and power adaptation are then performed for each user allocated to a channel. The transmitter antenna array can form a unit-power beam vector $\mathbf{u}_{n,k} = (u_{n,k,1}, \ldots, u_{n,k,M})^T$ of controllable orientation, determined by complex elements $\{u_{n,k,m}\}_{m=1}^{M}$ and it can transmit with controllable power $p_{n,k}$ to user $k$ assigned to channel $n$. Since a transmit antenna array with $M$ elements can form at most $M$ linearly independent beam vectors to $M$ users in a channel, at most $M$ users can be separated in a channel. A beam is formed by a dedicated transceiver (beamformer) hardware unit, and we assume there exist at least $NM$ such hardware units. A set of
M transceivers is shown in Figure 1(b). User minimum rate requirements and channel state information (CSI) are inputs to the resource allocation algorithms.

Channel quality for a user remains constant in a time slot but may change between slots. The time-invariant (within a slot) channel between antenna \( m \) and user \( k \) has impulse response

\[
h_{k,m}(t) = \sum_{\ell=1}^{L} \beta_{k,\ell} \delta(t - \tau_{k,\ell} + \tau_{k,m}^n),
\]

where \( L \) is the number of paths, \( \beta_{k,\ell} \) and \( \tau_{k,\ell} \) are the complex gain and delay of the \( \ell \)-th path of user \( k \) with respect to a reference antenna element respectively, and \( \delta(\cdot) \) is the impulse function. Gains \( \beta_{k,\ell} \) are complex Gaussian random variables with zero mean and variance \( \sigma_{k,m}^2 \) and delays \( \tau_{k,\ell} \) are uniformly distributed in [0, \( T \)]. In general, different paths of a user are correlated. Close spacing between antennas is assumed, so that multi-path characteristics of a user are similar across antennas. The term \( \tau_{k,m}^n \) captures the delay caused by the spacing between the \( m \)-th antenna and the reference one, where \( \Delta \) is the spacing between antennas, \( \theta_{k,m} \) is the angle of the \( m \)-th path of user \( k \) with respect to the antenna array and \( c \) is the electromagnetic wave propagation speed.

We now provide models for the OFDM, TDMA and CDMA schemes. An OFDM/SDMA transmitter is shown in Figure 2. After subcarrier allocation, beamforming and power control, user bits are forwarded into \( M \) parallel modules of \( N \) modulators. A modulator modulates the corresponding subcarrier with \( b \) bits of each user assigned to that subcarrier. The complex subsymbol at the output of each modulator is formed by a QAM constellation with \( b \) bits per subsymbol. All subsymbols of each user are fed into the Inverse Discrete Fourier Transform (IDFT) module and are transformed into \( N \) time samples that make an OFDM user symbol. After cyclic prefix addition, D/A conversion and up-conversion, the continuous signals are transmitted from the \( M \) antennas.

Assuming that OFDM symbols do not overlap in time, we concentrate on one symbol. In the following, we use baseband equivalent signal models. At receiver \( k \), after down-conversion, sampling at times \( \{ \frac{i}{N}, i = 0, \ldots, N-1 \} \) and DFT on time samples, the useful signal for user \( k \) at subcarrier \( n \) is

\[
y_{n,k} = \sqrt{p_{n,k}} (a_{n,k} H_{n,k} u_{n,k}) d_{n,k},
\]

where \( d_{n,k} \) denotes a unit-power complex subsymbol. Vector

\[
a_{n,k} = \sum_{\ell=1}^{L} \xi_{k,\ell}(n) v_n(\theta_{k,\ell})
\]

is called spatial signature of user \( k \) at subcarrier \( n \). Factors \( \xi_{k,\ell}(n) = \beta_{k,\ell}^n \exp(j2\pi n/\tau_{k,\ell}) \) capture the impact of delay of path \( \ell \) of user \( k \) on channel response at subcarrier \( n \), and vector \( v_n(\theta_{k,\ell}) \), whose \( m \)-th component is \( v_{n}^{m}(\theta_{k,\ell}) = \exp(-j2\pi \tau_{k,m}^n(n)) \), is the \( M \times 1 \) antenna steering vector at subcarrier \( n \) and direction \( \theta_{k,\ell} \). Clearly \( a_{n,k} \) captures angular and multi-path properties of user \( k \) at subcarrier \( n \). The expected useful received signal power is

\[
\mathbb{E}\{ |y_{n,k}|^2 \} = p_{n,k} (u_{n,k} H_{n,k} u_{n,k})^2,
\]

where the \( M \times M \) matrix \( H_{n,k} \) is

\[
H_{n,k} = \sum_{\ell_1=1}^{L} \sum_{\ell_2=1}^{L} \mathbb{E}\{ \xi_{k,\ell_1}(n) \xi_{k,\ell_2}^*(n) \} v_n(\theta_{k,\ell_1}) v_n^H(\theta_{k,\ell_2})
\]

and is called spatial covariance matrix of user \( k \) at subcarrier \( n \). In general, \( \text{rank}(H_{n,k}) > 1 \). If paths are uncorrelated, namely if it is

\[
\mathbb{E}\{ \xi_{k,\ell_1}(n) \xi_{k,\ell_2}^*(n) \} = 0 \text{ for } \ell_1 \neq \ell_2,
\]

then \( \text{rank}(H_{n,k}) > 1 \) unless there is only a line-of-sight (LOS) path.

A note about CSI is in place here. Different forms of transmitter CSI can be captured by modeling a spatial signature \( a \) as a complex Gaussian vector random variable with mean \( \mu \) and covariance matrix \( \Sigma \), namely \( a \sim \mathcal{N}(\mu, \Sigma) \). Perfect CSI is modeled by \( \Sigma = 0 \). This arises in time duplexing with reasonably small channel variation rate implying low Doppler spread. The BS can learn the average vector channel of each user through uplink measurements and can use them to adapt the downlink beam. For perfect CSI, beamforming towards the signature vector is optimal in the sense of maximizing capacity [25]. For no CSI, transmission in orthogonal directions is optimal [26]. For rapid channel variations, the channel realization cannot be tracked. However, the relative geometry of propagation paths changes more slowly and this is reflected in the entries of the spatial covariance matrix. Then,
CSI is modeled by knowledge of $\Sigma$, and beamforming toward a direction corresponding to the largest eigenvalue of the covariance matrix is asymptotically optimal for low SNRs [25] and close to optimal in general [27]. In this work, we assume that transmitter CSI consists of estimates of spatial covariance matrices $\mathcal{H}_{n,k}$ for each user $k$ and subcarrier $n$. Each matrix $\mathcal{H}_{n,k}$ is estimated by sampling received vector signal of user $k$ in subcarrier $n$ several times with known pilot symbols and by performing sample averaging. This presupposes that the channel process corresponding to an antenna element should be ergodic and wide-sense stationary.

The average signal-to-interference ratio (SIR) at the output of the matched filter receiver of user $k$ at subcarrier $n$ is

$$S_{n,k} = \frac{p_{n,k}(u_{n,k}^H \mathcal{H}_{n,k} u_{n,k})}{\sum_{j \in \mathcal{U}^{(n)} : j \neq k} p_{n,j}(u_{n,j}^H \mathcal{H}_{n,k} u_{n,j})},$$  

(4)

where the denominator denotes co-channel interference and $\mathcal{U}^{(n)}$ is the set of users in subcarrier $n$. Our model is interference-limited in the sense that co-channel interference prevails. Apart from practical implications, this approach eliminates the need for total transmit power constraints. In the discussion above for OFDMA, we omitted time variation from $a_{n,k}(t)$ and $\mathcal{H}_{n,k}(t)$ for notational simplicity.

In TDMA, time is partitioned in time slots and the entire bandwidth is used. The spatial signature and spatial covariance matrix of a user in each slot are obtained by averaging over frequency. That is, frequency selectivity is averaged out whenever frequency dependence needs to be taken into account. These quantities depend on temporal variations of multi-path characteristics in different slots. In OFDMA and TDMA, the SIR of a user does not depend on transmissions in other channels due to channel orthogonality.

In CDMA, the entire bandwidth and time frame is used. Channels are deterministic normalized codes of processing gain $G$ that emerge from a code design or code generation method. We refer to vector $c_n = (c_{n1}, \ldots, c_{nG})$ as code $n$. A code pair $(n,m)$ has cross-correlation $\rho_{nm} = c_n^T c_m$, with $\rho_{nn} = 1$. Code $n$ is expressed as $c_n(t) = \sum_{r=1}^G c_{nr} q(t - (r - 1)T_c)$, where $q(t)$ is the chip pulse and $T_c$ is the chip duration. The signal of user $k$ carried by code $n$ is $s_{n,k}(t) = \sum_{r=1}^G d_{n,k}(t)c_{nr}q(t - iT_c)$, where $\{d_{n,k}(t)\}$ is the symbol sequence. A single symbol is denoted by $s_{n,k}(t) = d_{n,k}c(t)$, where $d_{n,k}$ is a complex symbol formed by a linear modulation scheme with $b$ bits per symbol. A code is associated with rate $b/(GT_c)$. A user $k$ that uses $n_k$ codes achieves rate $b_{nk}/(GT_c)$ bits/sec. The signal of user $k$ carried by code $n$ is multiplied by beam vector $u_{n,k}$ and is allocated power $p_{n,k}$ before transmission. A code can be reused by several users if beamforming ensures user spatial separation.

The receiver of user $k$ consists of a bank of matched filters, each of which is matched to a code used by that user. The signal at the output of the matched filter to code $n$ is given by $y_{n,k} = c_n^T y_k$, where $y_k = \sum_{m=1}^N \sum_{j \in \mathcal{U}^{(m)}} \sqrt{p_{m,j}} c_{mj}(a_{m,j}^H u_{m,j})d_{m,j}$ is the total received signal at the input of receiver $k$. The average SIR at the output of the matched filter to code $n$ of user $k$ is

$$S_{n,k} = \frac{p_{n,k}(u_{n,k}^H \mathcal{H}_{n,k} u_{n,k})}{\sum_{j \in \mathcal{U}^{(n)} : j \neq k} p_{n,j}(u_{n,j}^H \mathcal{H}_{n,k} u_{n,j}) + \sum_{m=1}^N \sum_{j \in \mathcal{U}^{(m)} : m \neq n} p_{m,j}^2(u_{m,j}^H \mathcal{H}_{n,k} u_{m,j})}$$

(5)

The two terms in the denominator are co-channel interference from other users that use code $n$ and cross-channel interference from codes other than $n$ respectively. At the receiver, multi-path components at different delays can be coherently combined with a RAKE structure and frequency selectivity is compensated. Spatial signatures $a_{nk}$ and spatial covariance matrices $\mathcal{H}_{nk}$ for user $k$ do not depend on code $n$.

In this work we adhere to a conventional matched filter receiver in order to place emphasis on the impact of SDMA on resource allocation for OFDMA, TDMA and CDMA. We note that beamforming and power control have also been considered in conjunction with advanced multi-user receiver structures in single-channel CDMA [28], [29]. A fixed modulation level with $b$ bits/symbol is assumed. The minimum required SIR (in dB) for BER $\leq \epsilon$ at the receiver is threshold $\gamma = -(\ln(5\epsilon)/1.5)(2^b - 1)$ [30]. In a later section, we discuss implications of adaptive modulation on the problem, but we defer detailed study for a future work.

III. RESOURCE ALLOCATION IN SDMA-BASED SYSTEMS

A. Problem Statement

In TDMA and OFDMA, where channels are orthogonal, a set of users is spatially separable in a channel if there exist beamforming vectors and powers for each user $i$ such that SIR$_i \geq \gamma$ for each user $i$ in the channel. In a given channel, spatial separability depends on user spatial channel characteristics (that are captured by spatial covariance matrices) and on beamforming vectors and transmit powers that affect SIRs at all receivers. In TDMA, user spatial covariance matrices vary due to temporal variations of multi-path characteristics and spatial separability is addressed for each slot. In OFDMA, spatial separability depends on the particular subcarrier due to frequency selectivity of the wide-band channel, and also on the temporal variations of spatial characteristics between slots. That is, within a time slot, angular and multi-path characteristics of a user depend on the subcarrier frequency. Large channel reuse induces high total rate in a channel but also renders spatial separability difficult due to high interference.

In CDMA, user multi-path is compensated at each receiver with a RAKE path combiner after matched filtering. The spatial covariance matrix of a user is the same across all codes. The salient feature of CDMA is cross-channel interference due to code cross-correlation. A user with an assigned code receives co-channel interference from other users that use the same code and also receives cross-channel interference from other used codes that are non-orthogonal to its code. Hence, spatial separability of a user set $\mathcal{U}$ cannot be addressed separately for each code but must be considered collectively for a set of codes $\mathcal{C}$ that are used by $\mathcal{U}$. A user set $\mathcal{U}$ is spatially separable with respect to a channel (code) set $\mathcal{C}$ if there exists
a beamforming vector and power for each user \( i \) assigned to a channel in \( \mathcal{C} \) such that SIR\(_i \geq \gamma \) for each user \( i \).

The arising problem is to perform channel allocation and user spatial separation jointly so as to increase total rate and provide QoS guarantees to users. For TDMA and OFDMA, a large co-channel set of spatially separable users needs to be identified for each channel. Spatial separability of users amounts to large angular separation (if only a LOS path exists) or to nearly-orthogonal user spatial signatures (if several paths exist), so that the joint effect of spatial covariance matrices and beams is a small induced interference. Finding the largest spatially separable co-channel user set is a hard combinatorial optimization problem. Spatial separability depends jointly on beamforming vectors and powers of all users and enumeration of all possible user assignments in a channel has exponential complexity. In CDMA, users in different channels also create cross-channel interference among themselves. Code assignment should be such that user spatial channel characteristics, code cross-correlations, beamforming vectors and powers result in a spatially separable user set. We therefore need to resort to efficient heuristic algorithms and study three of them in the sequel. The first two belong in the class of greedy heuristics and use criteria based on minimum induced or in the sequel. The first two belong in the class of greedy heuristics and use criteria based on minimum induced or in the sequel. The first two belong in the class of greedy heuristics and use criteria based on minimum induced or

\[
\Psi_{n,k} = \max_{u_{n,k}} u_{n,k}^H \sum_{j \in \mathcal{U}(n)} H_{n,j} + \sum_{m=1}^{N} \sum_{i \in \mathcal{U}(m), i \neq m} \rho_{mn}^2 H_{m,i} u_{n,k}
\]

such that \( \|u_{n,k}\| = 1 \). The denominator captures co-channel and cross-channel interference caused by beam \( u_{n,k} \) to other users. The vector \( u_{n,k}^* \) that maximizes the ratio in (6) is the dominant generalized eigenvector corresponding to the two matrices in the numerator and denominator of the fraction in (6) and is computed by the method outlined at the end of section I. We also compute the ratio \( \Psi_{n,k}^{(n,k)} \) that quantifies the impact of user \( k \)’s insertion in channel \( n \) on user \( j \) in \( \mathcal{U}(m) \) and is equal to

\[
\max_{u_{n,j}} u_{n,j}^H \sum_{\mu=1}^{N} \sum_{\mu \neq m} \rho_{mn}^2 H_{\mu,i} + \sum_{i \in \mathcal{U}(m), i \neq j} H_{m,i} + \rho_{mn}^2 H_{n,k} u_{n,j}
\]

such that \( \|u_{m,j}\| = 1 \). Note that the computed beams are the ones that maximize user SIRs in a virtual uplink system. With the computed beamforming vectors, we evaluate the SIRs for all assigned users in channels.

2. Transmit power control. If SIRs for some users do not exceed \( \gamma \), we employ power control (while keeping the computed beamforming vectors fixed) so that all SIRs exceed \( \gamma \). For channel \( n \), let \( i, j \) be indices of users in that channel. Let \( \mathcal{U} \) be the computed ensemble of beamforming vectors for all users in channels, i.e \( \mathcal{U} = \{u_{n,k} : k \in \mathcal{U}(n), n = 1, \ldots, N\} \). Define the \( (\sum_{n=1}^{N} |\mathcal{U}(n)|) \times (\sum_{n=1}^{N} |\mathcal{U}(n)|) \) block matrix

\[
A(U) = \begin{pmatrix}
A_{11}(U) & \cdots & A_{1N}(U) \\
\vdots & \ddots & \vdots \\
A_{N1}(U) & \cdots & A_{NN}(U)
\end{pmatrix}
\]

The \( [i, j] \)-th element of the \( (|\mathcal{U}(n)| \times |\mathcal{U}(n)|) \) matrix \( A_{nn}(U) \) in the diagonal of \( A(U) \) denotes co-channel interference caused by the beam of user \( j \) to user \( i \) in channel \( n \),

\[
A_{nn}(U)[i, j] = \begin{cases}
\sum_{m \neq n} \rho_{mn}^2 \sum_{k \in \mathcal{U}(n), k \neq j} u_{n,m}^H H_{n,k} u_{n,j} & \text{if } i \neq j \\
0 & \text{if } i = j.
\end{cases}
\]

The \( [i, j] \)-th element of the \( (|\mathcal{U}(n)| \times |\mathcal{U}(n)|) \) matrix \( A_{nm}(U) \), \( n \neq m \) denotes the cross-channel interference caused by the beam of user \( j \) in \( \mathcal{U}(m) \) to user \( i \) in \( \mathcal{U}(n) \) and is given by \( A_{nm}(U)[i, j] = \rho_{nm}^2 \sum_{k \in \mathcal{U}(m), k \neq j} u_{m,k}^H H_{n,k} u_{n,j} \), where \( \rho_{nm} \) is the cross-correlation between channels \( n \) and \( m \). We also define the diagonal matrix \( \Delta = \text{diag} \{\sum_{n=1}^{N} |\mathcal{U}(n)| \times |\mathcal{U}(n)| \times 1 \} \) vector \( p \) of user transmission powers in channels. Then, the condition \( \text{SIR}_{n,k} \geq \gamma \) for each user \( k \) in each channel \( n \) is written in matrix form as:

\[
p \geq \gamma \Delta A(U)p.
\]

Matrix \( \Delta A(U) \) is non-negative definite and irreducible. From the Perron-Frobenius theorem, it has exactly one positive, real eigenvalue \( \lambda^* = \max_i |\lambda_i| \), where \( \lambda_i \), for \( i = 1, \ldots, (\sum_n |\mathcal{U}(n)|) \) are the eigenvalues of \( \Delta A(U) \). Eigenvalue \( \lambda^* \) has an associated eigenvector \( p^* \) with strictly positive entries. Furthermore, the minimum real \( \lambda \) for which inequality \( \lambda p \geq \Delta A(U)p \) has solutions \( p > 0 \) is \( \lambda = \lambda^* \). We start by finding the maximum real positive eigenvalue \( \lambda^* \) of \( \Delta A(U) \). If \( \lambda^* \leq 1/\gamma \), then (10) holds and the SIR level \( \gamma \) is called feasible or equivalently users are said to form a feasible
set. The power vector for feasible $\gamma$ is the eigenvector that corresponds to $\lambda^*$. 

With the procedure above, we compute beams and powers for a specific user assignment. We need to evaluate different assignments that lead to feasible user sets and select the best one. For each possible insertion of user $k$ in channel $n$ that leads to a feasible user set, we define the assignment preference factor $\Phi_{n,k}$. This should be large if beams and powers yield strong useful signal for user $k$, low interference $I_{n,k}$ caused by user $k$ to other users and low induced interference $I'_{n,k}$ by other users on $k$. We consider the largest of these two amounts of interference and define factors

$$\Phi_{n,k} = \frac{p_{n,k}(u_{n,k}^H H_{n,k} u_{n,k})}{\max \{I_{n,k}, I'_{n,k}\}},$$

(11)

where $I_{n,k}$ and $I'_{n,k}$ are given by

$$I_{n,k} = p_{n,k}u_{n,k}^H \left( \sum_{j \in \mathcal{U}^n} H_{n,j} + \sum_{m=1, m \neq n}^N \sum_{i \in \mathcal{U}(m)} \rho_{mn}^2 H_{m,i} \right) u_{n,k}^*,$$

(12)

$$I'_{n,k} = \sum_{j \in \mathcal{U}(n)} p_{n,j} \left( u_{n,j}^H H_{n,k} u_{n,j}^* \right) + \sum_{m=1, m \neq n}^N \sum_{i \in \mathcal{U}(m)} \rho_{mn}^2 p_{m,i} \left( u_{m,i}^H H_{m,n} u_{m,i} \right).$$

(13)

When power control is not active, these expressions do not include powers. At each step, factors $\Phi_{n,k}$ are computed for all possible insertions of users $k$ that have not satisfied minimum rate requirements $x_k$ and for all channels $n$ where insertion of a user leads to a feasible user set. The assignment with maximum $\Phi_{n,k}$ is selected and the rate of user $k$ is updated. When a user $k$ reaches $x_k$, it is not considered for further assignments until all users reach minimum rate requirements.

A channel is not further considered if $M$ users are already assigned to it. The algorithm terminates if no further user insertion in any channel leads to feasible user set.

The greedy approach of least incremental interference in algorithm A aims at inserting as many users as possible. Greedy algorithm B relies on maximizing the minimum SIR of users: a user assignment in a channel is performed if it maximizes the minimum SIR of users in the system over all possible assignments. That is, algorithm B also captures the impact of an assignment on other users, so that SIRs that are closer to $\gamma$ are maximized, future assignments are facilitated and ultimately the number of users with SIRs above $\gamma$ is increased. The assignment factors for algorithm B are

$$\Phi_{n,k} = \min \{S_{n,k}, \min_{m \in \mathcal{U}^n} \min_{j \in \mathcal{U}(m)} S_{m,j}\}.$$

A large time interval of several time frames is considered, over which users need to achieve their minimum rates. In TDMA, channels are orthogonal time slots and spatial covariance matrices of users change due to temporal variations of multi-path channel characteristics between slots. Only interference from co-channel users exists. Each slot is considered separately and users are assigned sequentially based on factors $\Phi_{n,k}$. In OFDMA, channels are orthogonal subcarriers and a user $k$ has different spatial covariance matrices $H_{n,k}$ in each subcarrier $n$ and slot. Subcarriers are filled with users in each slot of the time frame. In CDMA, channels are non-orthogonal codes due to nonzero pairwise code cross-correlations. Spatial covariance matrices do not depend on codes and change only due to multi-path temporal variations. In each frame, codes are allocated to users until the algorithm terminates, and the procedure repeats in the next frame.

C. SIR Balancing Assignment Algorithm C

In algorithms A and B, beamforming and power control were decoupled. Algorithm C is based on SIR balancing and attempts to provide maximum common user SIR by employing joint beamforming and power control.

1) Single-channel algorithm: Consider channel $n$ with user set $\mathcal{U}^n$. Let $p_n$ and $\mathbf{U}_n = \{u_{n,k} : k \in \mathcal{U}^n\}$ be the transmit power vector and ensemble of beamforming vectors for users in $\mathcal{U}^n$. Consider interference matrix $A_{nn}(U)$ in (9) and call it $B(U_n)$. Define the diagonal matrix $\Delta_n = \text{diag}\{\frac{1}{\lambda_{n,k}} : k \in \mathcal{U}^n\}$. The condition $S_{n,k} \geq \gamma$ for users is channel $n$ with beamforming vectors $U_n$ and power vector $p_n$ in the downlink is written in matrix form as [16], [17]:

$$p_n \geq \gamma \Delta_n B(U_n) p_n.$$

(14)

Matrix $\Delta_n B(U_n)$ has the same properties as $\Delta A(U)$ with respect to existence of a positive eigenvalue and an eigenvector $p_n$ with positive components. With the same reasoning as before, the maximum common SIR is given by

$$\gamma_c^* = \frac{1}{\min_{U_n} \lambda^*(\Delta_n B(U_n))}.$$

(15)

For the corresponding problem in the uplink, the SIR requirements with ensemble of beamforming vectors $U_n$ and power vector $p_n$ are expressed as $p_n \geq \gamma \Delta_n B^T(U_n) p_n$ and the maximum possible common SIR $\bar{\gamma}_c^*$ is

$$\bar{\gamma}_c^* = \frac{1}{\min_{U_n} \lambda^*(\Delta_n B^T(U_n))}.$$

(16)

The following properties for the relationship between the downlink and uplink problems hold [16], [17]:

Property 1: For given set of beamforming vectors $U_n$, it is $\lambda^*(\Delta_n B(U_n)) = \lambda^*(\Delta_n B^T(U_n))$.

Property 2: The downlink and uplink problems have the same maximum achievable common SIR, namely $\gamma_c^* = \bar{\gamma}_c^*$.

Property 3: The beamforming vectors that solve downlink and uplink problems (15),(16) are the same, i.e. $U^*_n = U^*_n$.

Property 4: In algorithm I below, the sequence of eigenvalues $\lambda^*(t)$ is monotonically decreasing with iteration $t$ and converges to a minimum eigenvalue that is related to the maximum common SIR through (15) and (16).

As a side note, the properties above also hold in the presence of noise, with equal noise power level at all receivers [17]. The steps of Algorithm I are:

- **STEP 1:** Set $t = 0$. Start with arbitrary beamforming vectors $U_n^{(0)}$.
- **STEP 2:** $t \leftarrow t + 1$. For given $U_n^{(t)}$, solve the uplink eigenproblem $\Delta_n B^T(U_n^{(t)}) p_n^{(t)} = \lambda^{*(t)} p_n^{(t)}$. 

• **STEP 3:** For the computed \( p^{(t)}_n \), solve a set of *decoupled* generalized eigen-problems:

\[
\mathbf{u}^{(t)}_{n,k} = \arg \max_{\mathbf{u}_{n,k}} \quad \mathbf{u}_{n,k}^H \mathbf{H}_{n,k} \mathbf{u}_{n,k} \\
\text{subject to } \| \mathbf{u}_{n,k} \| = 1 \quad \text{for } k \in U^{(n)}, \text{ where matrix } \\
\mathbf{R}_{n,k}(p^{(t)}_n) = \sum_{j \in U^{(n)}, j \neq k} p^{(t)}_n \mathbf{H}_{n,j}. 
\]

• **STEP 4:** With the computed \( U^{(t)}_n \), go to Step 2. Continue until \( |\lambda^{(t+1)} - \lambda^{(t)}| < \epsilon \), with \( \epsilon > 0 \) a small constant.

In Step 3, the quantity to be maximized is the uplink SIR of user \( k \). Beamforming vectors \( \mathbf{U}^{(n)}_n \) at the end of the algorithm are the desired downlink beams. If \( \lambda_0 = \lambda^* (\Delta_n, \mathbf{B}(\mathbf{U}^{(n)}_n)) \) is the eigenvalue at the end of the algorithm, the downlink power vector is the eigenvector of \( \Delta_n, \mathbf{B}(\mathbf{U}^{(n)}_n) \) corresponding to \( \lambda_0 \).

If \( 1/\lambda_0 \geq \gamma \), SIR \( \gamma \) is achievable for all co-channel users.

2) **Description of Algorithm C:** Algorithm C is based on the observation that a system with \( K \) users in \( N \) non-orthogonal channels is equivalent to a single-channel system of interfering users and can be described by block matrix \( \mathbf{A}(\mathbf{U}) \) in (8).

A system where users achieve common SIR \( \gamma_c \) in the downlink is described by equations \( \mathbf{p} = \gamma_c \Delta \mathbf{A}(\mathbf{U}) \mathbf{p} \) and the matrix in Step 3 of algorithm I is

\[
\mathbf{R}_{n,k}(p^{(t)}_n) = \sum_{j \in U^{(n)}, j \neq k} p^{(t)}_n \mathbf{H}_{n,j} + \sum_{m=1}^{N} \sum_{j \in U^{(m)}, m \neq n} \rho_{nm}^2 p^{(t)}_{m,j} \mathbf{H}_{m,j}. 
\]  

(18)

Fix an assignment of users to channels and let \( \gamma_c^* \) be the maximum common SIR of users as the outcome of algorithm I. For each user \( k \in U^{(n)}, n = 1, \ldots, N \) let \( \gamma_{c,n}(k) \) be the maximum common SIR of remaining users when \( k \) is removed from channel \( n \). Again \( \gamma_{c,n}(k) \) is found by Algorithm I with an appropriately modified matrix \( \mathbf{A}(\mathbf{U}) \). Initially, all users are assigned in all channels. At each step, a user is removed from a channel, such that the highest common SIR is incurred for remaining users. The procedure continues until the desired common SIR \( \gamma_c \) is reached. The goal is to remove few users until common SIR \( \gamma_c \) is reached, so as to achieve high total rate. The steps of algorithm C are as follows:

• **STEP 0:** Start by assigning all \( K \) users in each channel \( n, n = 1, \ldots, N \).

• **STEP 1:** Run algorithm I and find the maximum common SIR \( \gamma^*_c \) for the system.

• **STEP 2:** If \( \gamma^*_c \geq \gamma_c \), the algorithm is terminated. Otherwise go to Step 3.

• **STEP 3:** For each \( k \in U^{(n)}, n = 1, \ldots, N \) compute \( \gamma_{c,n}(k) \) with Algorithm I. Select pair \((n^*, k^*)\) with the maximum \( \gamma_{c,n}(k) \) and remove user \( k^* \) from channel \( n^* \).

• **STEP 4:** Update user rates. If a user \( k \) reaches minimum rate requirements \( x_k \), do not consider it for removal. Set \( \gamma_{c,n}(k^*) = \gamma^*_c \). Go to Step 2.

Again users need to achieve some minimum rates over a time interval. In CDMA, Algorithm C runs in each frame. In TDMA and OFDMA, where matrices \( \mathbf{A}_{ij}(\mathbf{U}) = 0 \) for \( i \neq j \), a separate problem (15) is solved for each channel \( n \). In TDMA, each slot is considered separately and the single-channel version of algorithm C is applied. In OFDMA, in each slot we start by assigning all users in each subcarrier and run algorithm I for each subcarrier to get a vector of common channel SIRs, \( \gamma_c = (\gamma_{c,1}, \ldots, \gamma_{c,N}) \), where \( \gamma_{c,n} \) is the common SIR for users in subcarrier \( n \). If \( \gamma_{c,n} \geq \gamma \) for all \( n = 1, \ldots, N \), the algorithm terminates. Otherwise users must be removed from subcarriers \( n \) with \( \gamma_{c,n} < \gamma \). For each user \( k \) in such a subcarrier \( n \), let \( \gamma_{c,n}(k) \) be the common SIR in \( n \) after \( k \) is removed. At each step we remove the user \( k \) from subcarrier \( n \) so that \( \gamma_{c,n}(k) \) is maximum. Next, user rates are updated. When a user \( k \) reaches \( x_k \), it is not considered in later iterations. If \( \gamma_{c,n} \geq \gamma \) for a subcarrier \( n \) at some stage of the algorithm, no more users are removed from \( n \).

D. **Optimal solution for K=2 users and N=1 channel**

For \( K = 2 \) co-channel users and \( M \geq 2 \), let \( \mathbf{H}_i, \mathbf{u}_1, \mathbf{p}_i \) be the spatial covariance matrix, beam and power of user \( i \). We start with beams \( \{ \mathbf{u}_1 \} \). In the first iteration of algorithm I, we have \( \lambda^{(1)}(\mathbf{H}_1, \mathbf{H}_2) \) and \( \mathbf{p}_1 \) and we get power ratio \( \mu^{(1)} = p_2/p_1 \) in Step 2. In Step 3, we find beams \( \mathbf{u}_1 = \mathbf{u}_{\max}(\mathbf{H}_1, \mathbf{H}_2) \) and \( \mathbf{u}_2 = \mathbf{u}_{\max}(\mathbf{H}_2, \mathbf{H}_1) \). In second iteration, we have \( \lambda^{(2)} = \frac{\lambda_{\max}(\mathbf{H}_1, \mathbf{H}_2)}{\lambda_{\min}(\mathbf{H}_1, \mathbf{H}_2)} \) and ratio \( \mu^{(2)} = \sqrt{\frac{\lambda_{\max}(\mathbf{H}_1, \mathbf{H}_2)}{\lambda_{\min}(\mathbf{H}_1, \mathbf{H}_2)}} \), where \( \lambda_{\max}(\mathbf{H}_1, \mathbf{H}_2), \lambda_{\min}(\mathbf{H}_1, \mathbf{H}_2) \) are the maximum and minimum generalized eigenvalues of \( (\mathbf{H}_1, \mathbf{H}_2) \). These do not change in later iterations. Thus, the maximum common SIR is \( \gamma_c = 1/\lambda^{(2)} = \mathbf{u}_1, \mathbf{u}_2 \) and power ratio \( p_2/p_1 \) above.

E. **Adaptive Modulation**

When adaptive modulation is employed, the number of bits of a user is selected from a \( L_0 \)-element set \( M \) of available QAM or QPSK modulation levels with different number of bits per symbol, \( \{ b_i \} \). In OFDMA, different number of bits of a user can modulate a subcarrier depending on subcarrier quality. In TDMA, the number of bits \( b_i \) conveyed to a user in a slot is adaptable, while in CDMA the rate \( b_i/G_t \) carried by a code is controlled.

Consider \( m \leq M \) co-channel users in channel \( n \). Let \( \mathbf{b} = (b_1, b_2, \ldots, b_m)^T \) be the user modulation level vector and \( \gamma = (\gamma_1, \gamma_2, \ldots, \gamma_m)^T \) the corresponding threshold vector, with \( \gamma_i = -\ln(5e)/1.5(2^{b_i} - 1) \). Co-channel user set \( U^{(n)} \) is spatially separable with respect to modulation vector \( \mathbf{b} \) or equivalently \( \mathbf{b} \) is achievable if \( \mathbf{p} \geq \Delta_n \mathbf{B}(U_n) \mathbf{p} \), where \( \Delta_n = \text{diag}\{\gamma_{i,n}^2, \gamma_{i,n}^2, \ldots, i \in U^{(n)}\} \) and \( \mathbf{p} \) is the corresponding transmit power vector. If the maximum positive eigenvalue of matrix \( \Delta_n \mathbf{B}(U_n) \) satisfies \( \lambda^{**} \leq 1 \), the modulation vector \( \mathbf{b} \) is achievable and the power vector that achieves \( \mathbf{b} \) is the eigenvector that corresponds to \( \lambda^{**} \). High modulation levels for users in a channel imply more transmitted bits per user in a channel but do not favor large channel reuse since they are vulnerable to interference. On the other hand, low modulation levels can sustain more interference and thus more crowded co-channel sets but transmit few bits per user. Clearly, these two aspects of impact of modulation level on channel rate are conflicting. The objective of a resource allocation algorithm is to identify co-channel user sets of maximum rate. The design of such algorithms is left for future consideration.
IV. NUMERICAL RESULTS

We consider a base station (BS) with \( K = 10 \) users, \( N = 10 \) channels, and an antenna array with \( M = 4 \) or 8 elements and \( \Delta = \lambda/2 \). Minimum rate requirements in terms of minimum number of required channels are \( x_k = 3 \) for each user \( k \). Received power decays with distance \( d \) from the BS as \( d^{-4} \). For a link between an antenna element and a user, multi-path is modeled by 2 paths with angles \( \theta_1, \theta_2 \), where \( \theta_1 \) is uniformly distributed in \([0, 2\pi]\) and \( \theta_2 \) deviates from \( \theta_1 \) by a random quantity, uniform in \([0, 0.1\pi]\). The complex gain of each path is a normal random variable with standard deviation \( \sigma = 6 \) dB that accounts for shadowing. Path gains of different users are uncorrelated. Our goal is to compare the performance of proposed algorithms for TDMA, OFDMA or CDMA. We also assess the benefit of power control in algorithms A, B and thus we present results with and without power control (NPC). The performance metrics are (i) total user rate in terms of user channels and (ii) total residual rate, namely additional rate so that users reach minimum rate requirements. For CDMA, we assume that code cross-correlation is uniformly distributed in \([0, \rho_{max}^2]\) and consider low and high cross-correlation scenarios with \( \rho_{max}^2 = 0.02 \) and 0.1 respectively. Results are averaged over several experiments with different channel conditions. The observed fluctuations in the plots are mostly due to minimum rate requirements of users.

In Figure 3, the total rate is depicted as a function of SIR threshold \( \gamma \) for OFDMA. A high value of \( \gamma \) implies stringent BER requirement. Algorithm C achieves the best performance while algorithms A and B perform almost the same. For moderate values of \( \gamma \) (e.g. \( \gamma > 17 \) dB), rate improvements of 20-25\% are achieved with power control for algorithm A, while the corresponding benefit for algorithm B is only 5-10\%. The performance of algorithm B-NPC is close to that of A with power control. This suggests that method B-NPC can be adopted in situations where low complexity is needed.

In Figure 3, the total rate is depicted as a function of SIR threshold \( \gamma \) for OFDMA. A high value of \( \gamma \) implies stringent BER requirement. Algorithm C achieves the best performance while algorithms A and B perform almost the same. For moderate values of \( \gamma \) (e.g. \( \gamma > 17 \) dB), rate improvements of 20-25\% are achieved with power control for algorithm A, while the corresponding benefit for algorithm B is only 5-10\%. The performance of algorithm B-NPC is close to that of A with power control. This suggests that method B-NPC can be adopted in situations where low complexity is needed. For larger values of \( \gamma \) (e.g. \( \gamma > 17 \) dB), three of the four alternatives of algorithms A and B lead to similar performance. Similar conclusions are derived for TDMA in Figure 4. However, the performance difference between algorithm C and
other techniques is smaller than in OFDMA.

The case of CDMA is depicted in figures 5 and 6. For low code cross-correlation, algorithm C yields the highest total rate and algorithm A leads to similar performance. Algorithm A-NPC performs better than B regardless of the use of power control in B, although it has the lowest rate in TDMA and OFDMA. As code cross-correlation increases, algorithms A and C achieve similar rates and the performance gap between these and other algorithms decreases. In Figure 7 we show performance in terms of total residual rate for OFDMA, which again verifies the superiority of algorithm C. Minimum rate requirements of users are fulfilled for \( \gamma \leq 14 \) dB and a small portion of user requirements remains unsatisfied for larger \( \gamma \). For \( M = 8 \) antennas (Figure 8), algorithms A and B yield only 30 – 35% more rate than algorithm C with \( M = 4 \), while C achieves double rate for \( M = 8 \) than with \( M = 4 \). This justifies the claim that performance depends jointly on physical and access layer adaptation methods. The superiority of algorithm C over A and B is more evident in OFDMA and TDMA and is marginal in CDMA with non-orthogonal channels. This is mostly due to the joint adaptation of beams and powers which achieves SIR balancing, but also due to the machinery of allocation, which starts from an all-users-to-all-channels initial condition and proceeds by iteratively removing users with the SIR balancing criterion. While this scheme has similar effect with the one of least-interference greedy incremental user insertion for a virtual channel (encountered in CDMA), its performance is better when channels are orthogonal and separate, since then the greedy algorithm is applied in two dimensions, namely user and channel selection.

V. DISCUSSION

The impact of SDMA on resource allocation should lead to efficient channel reuse in the presence of interference and high total rate. We adhered to a unified approach for TDMA, OFDMA and CDMA and presented algorithms for joint channel allocation, beamforming and power control that capture properties of a good solution. We observed that SIR balancing with joint beamforming and power control leads to very good performance, superior to that of greedy heuristics, especially in cases of orthogonal channels. Our conclusions about superiority of SIR balancing over greedy least interference channel assignment are in line with existing findings for systems with no SDMA [3], [31].

The resource allocation problem and associated algorithms obtain an interesting twist if multi-rate transmission is employed with adaptive modulation in TDMA and OFDMA and also by spreading gain adaptation in CDMA. Then, user spatial separability will also depend on modulation levels as discussed in section III. Another interesting issue is to devise distributed algorithms for multi-cell systems, which will be executed independently by each BS as best responses to actions of other BSs, so that BS rate is maximized. Finally, an important extension in multi-cell systems is BS assignment for load balancing and interference mitigation.

VI. ACKNOWLEDGMENTS

The authors wish to thank the anonymous reviewers for their comments that improved the presentation of the paper. The authors acknowledge support from a bilateral Greece-USA collaboration grant, support of European Commission STREP WIP (FP6-IST-027402) and support of Greek Secretariat for Research and Technology through a PENED grant. Finally, the authors thank T. Ren for his help on numerical results.

REFERENCES


http://grouper.ieee.org/groups/802/11/Reports/ign_update.htm


Iordanis Koutsopoulos (S ’99, M ’03) is a Lecturer at the Department of Computer and Telecommunications Engineering, University of Thessaly, Greece. He was born in 1974 in Athens, Greece. He obtained the Diploma in Electrical and Computer Engineering from the National Technical University of Athens, Athens, Greece in 1997 and the M.S and Ph.D degrees in Electrical and Computer Engineering from the University of Maryland, College Park (UMCP) in 1999 and 2002 respectively. From 1997 to 2002 he was a Fulbright Fellow and a Research Assistant with the Institute for Systems Research (ISR) of UMCP. He has held internship positions with Hughes Network Systems, Germantown, MD, Hughes Research Laboratories LLC, Malibu, CA, and Aperto Networks Inc., Milpitas, CA, in 1998, 1999 and 2000 respectively. For the summer period of 2005 he was a visiting scholar with University of Washington, Seattle, WA. For the period 2005-2007 he was awarded a Marie Curie International Reintegration Grant (IRG).

His research interests are in the field of networking with emphasis on wireless networks, cross-layer design, sensor networks, smart antennas and wireless network security.

Leandros Tassiulas (S’89, M’91, SM’06, F’07) obtained the Diploma in Electrical Engineering from the Aristotelian University of Thessaloniki, Thessaloniki, Greece in 1987, and the M.S. and Ph.D degrees in Electrical Engineering from the University of Maryland, College Park in 1989 and 1991 respectively.

He is Professor in the Dept. of Computer and Telecommunications Engineering, University of Thessaly, since 2002. He has held positions as Assistant Professor at Polytechnic University New York (1991-95), Assistant and Associate Professor University of Maryland College Park (1995-2001) and Professor University of Ioannina Greece (1999-2001). His research interests are in the field of computer and communication networks with emphasis on fundamental mathematical models, architectures and protocols of wireless systems, sensor networks, high-speed internet and satellite communications. Dr. Tassiulas is a Fellow of IEEE. He received a National Science Foundation (NSF) Research Initiation Award in 1992, an NSF CAREER Award in 1995, an Office of Naval Research, Young Investigator Award in 1997 and a Bodossaki Foundation award in 1999. He also received the INFOCOM 1994 best paper award and the INFOCOM 2007 achievement award.
A Practical Algorithm for Power Minimisation in Wireless Networks by Means of Multi-Hop and Load Partitioning

Lorenzo Piazzo

Abstract—We study the problem of minimising the power consumption of wireless networks when the traffic is specified. Initially we consider single-hop networks that model currently existing Frequency Division Multiplex (FDM) and Code Division Multiple Access (CDMA) networks and discuss the power minimisation problem for these networks. Next we consider networks where multi-hop transmission is deployed and where a source is allowed to partition its outgoing information flow into several sub-flows each delivered through different multi-hop routes. We present a simple, versatile, suboptimum algorithm for the power minimisation in these Multi-hop, Load Partitioning (MLP) networks. Finally we carry out a simulation campaign in order to quantify the power savings that are attainable by means of the proposed algorithm.

Index Terms—Multi-hop, wireless networks, resource allocation, routing.

I. INTRODUCTION

MULTI-HOP is a well known technique exploited in wireless networks in order to extend coverage and capacity, to reduce power consumption and to achieve several other benefits e.g. [1], [2], [3], [4], [5], [6], [7], [8], [9], [10]. It can be combined with other techniques and in this paper we investigate networks where it is coupled with a load partitioning mechanism that allows a source to split its outgoing information in several links each reaching the destination through a different multi-hop path. Such Multi-hop, Load Partitioning (MLP) networks are known in the literature. For example in [8] they are analysed, together with several other types of wireless networks, and their rate regions, subject to a constraint on the transmit power of each node, are defined and computed. While [8] is important from a theoretical point of view, the computational complexity of the methods proposed there prevents their practical use. Two fundamental papers concerning MLP networks are [9] and [10]. In these works a rigorous analytical framework is set up where the optimisation of a MLP network is recognised to be the solution of a Simultaneous Routing and Resource Allocation (SRRA) problem. As a further contribution in this paper we consider the problem of minimising the total power consumption of a MLP network when the traffic is assigned. The total power is a physically meaningful parameter and its minimisation translates into reduced cost, less pollution and increased battery life among other benefits. Indeed this figure of merit was used previously in the context of wireless networks e.g. [11], [12]. The power minimisation problem for MLP networks is a SRRA problem so that it can be solved or approximately solved by means of the techniques presented in [9], [10]. However in this paper we take a different point of view that yields to an alternative solution algorithm which, despite its simplicity, seems to be unpublished and has interesting features. The algorithm is intuitive, general and independent of the network physical layer. Moreover, while the focus of the paper is on the power minimisation problem, the algorithm presented can be extended to handle other SRRA problems. A drawback is that the algorithm cannot be guaranteed to yield a global minimum but only a local one. Therefore in order to prove its usefulness and as an additional contribution we carry out a simulation campaign in order to assess what is the performance gain that can be achieved by means of the proposed algorithm under several operating conditions.

To put the paper in the right perspective let us note that our results are mainly theoretical, since a number of assumptions are made that are not realistic. Specifically, like is done in [8], [9], [10], we restrict to frequency flat channels, we assume Shannon Capacity attaining transmissions and we assume perfect channel and network knowledge at the nodes. Therefore the results of the paper put an upper bound to the performance of real networks and as such are a useful design tool. In addition several of the assumptions are made in order to keep the development and the presentation simple and can be relaxed in order to extend the analysis to more realistic cases.

The paper is organised as follows. In section II we present the formalism adopted and state the power minimisation problem. In section III we introduce two single-hop networks that model the currently existing Frequency Division Multiplex (FDM) and asynchronous Code Division Multiple Access (CDMA) networks respectively and discuss their optimisation. In section IV we introduce the MLP networks and the algorithm employed for their optimisation. In section V we introduce two simple stochastic models for the network’s nodes and traffic distribution aimed at representing a Centralised and an Ad-hoc network respectively and carry out a simulation campaign in order to quantify the power gains attainable by
means of the proposed algorithm. Finally in section VI we present our conclusions. Throughout the paper the terms being defined are typed in boldface.

II. DEFINITIONS AND PROBLEM STATEMENT

A network is made up by $N$ transceivers nodes, identified by the integers $0..N-1$. The nodes exchange radio signals and we assume that node $i$ is connected to node $j$ by means of an ideal, frequency flat channel with gain $c_{i,j}$. In each node an Additive White Gaussian Noise (AWGN) process adds to the received signal so that, upon denoting by $s_i(t)$ the signal transmitted by node $i$, the signal received at node $j$ is

$$r_j(t) = \sum_{i=0}^{N-1} c_{i,j} s_i(t) + n_j(t) \quad j = 0..N-1$$

(1)

where $n_j(t)$ is the AWGN process of node $j$. Without loss of generality\(^1\) we assume that the double sided noise power density at all nodes is $N_0/2$. The transmitted signals are assumed to be stationary processes. Furthermore the network is assigned a band of $B$ Hz where transmission takes place so that all the transmitted signals must be within this band. We also assume that $c_{i,i} = 0$, i.e. that a node does not receive its own transmitted signal\(^2\). Finally we denote by $p_i$ the power of $s_i(t)$ and by

$$P = \sum_{i=0}^{N-1} p_i$$

(2)

the total power requested by the network.

The task of a network is that of transferring information among the nodes. We assume that the network’s traffic is assigned and we denote by $l_{i,j}$ (bit/sec) the rate of information to be transferred from node $i$ to node $j$. We also introduce the matrix $L = \{l_{i,j}\}$ which is termed the load matrix of the network. In order to deliver the load the network has to deploy an adequate set of transmitted signals (STS) $\{s_0(t)...s_{N-1}(t)\}$. In real networks these signals will be constrained by the physical layer organisation and capabilities. To reflect this fact we assume that a transmission protocol \(^8\) is specified for the network, giving the rules and constraints that the signals must obey. These rules define the set of the possible STS, denoted by $T$ and containing all the STS that verify the protocol’s constraints so that we write $\{s_0(t)...s_{N-1}(t)\} \in T$ if the STS $\{s_0(t)...s_{N-1}(t)\}$ verifies the protocol’s constraints. We will see several examples of transmission protocols in the following sections.

Given a network, a load matrix and a transmission protocol a basic question is whether a STS exists in $T$ that is capable of delivering the network load\(^3\). If such a set exists we say that the network is feasible for the given load and protocol and we call any such STS an implementation of the network. Note that if a network is feasible it is possible that many implementations exist so that we denote by $F \in T$ the set of the possible implementations. For any protocol a method is needed to compute $F$ and to chose an implementation when many exist. This method reflects the operations of the higher layers of real networks and shall be designed in order to optimise the network in some sense. In the following of the paper we present methods that seek the implementation in $F$ requiring the least total power. Indeed that is a sensible implementation to choose since minimising the power yields the benefits mentioned in the introduction. Therefore we call this implementation the optimum implementation and we call its total power, given by

$$C = \min_{\{s_0(t)...s_{N-1}(t)\}\in FP}$$

the consumption of the network. Note that the consumption of a network depends on the load and on the protocol (which together determine the set $F$). Therefore given a network, for any function a protocol exists yielding the network consumption as a function of the load matrix $L$. We denote this function by $C_{\text{protocol}}(L)$ and we set $C_{\text{protocol}}(L) = +\infty$ when the network is unfeasible for the given load and protocol.

III. FDM AND CDMA NETWORKS

In this section we introduce several single-hop transmission protocols which model currently existing networks. In all the protocols that we consider the transmitted signals can be written as

$$s_i(t) = \sum_{j=0}^{N-1} s_{i,j}(t) \quad i = 0..N-1$$

(4)

where the signal $s_{i,j}(t)$ carries information from node $i$ to node $j$ and realises a link between these two nodes. We denote by $r_{i,j}$ the number of bit/sec carried by the signal $s_{i,j}(t)$ and call $R = \{r_{i,j}\}$ the rate matrix of the network. The rate matrix is not to be confused with the load matrix. While the two are equal in single-hop protocols they are different when multi-hop is deployed. Furthermore, in all the protocols considered, the modulation method employed is assumed to operate at the Shannon capacity\(^4\) e.g. [14]. In other words we assume that, given a received signal with a band of $b$ Hz and a power of $s$, if the noise power at the receiver is $n$ and the noise process is AWGN, the maximum rate $r$ that can be transferred by the signal, at an error probability as low as desired, is\(^5\) $r = b \log_2(1+s/n)$. Rearranging the last equation we obtain the received power as a function of the noise power, band and rate as:

$$s = n \left(2^{r/b} - 1\right).$$

(5)

The first protocol that we consider is a model for networks where the physical layer keeps the signals orthogonal to each other, like is done in FDM, Time Division Multiplex (TDM) and synchronous CDMA networks. Over the flat channels that we are assuming these methods require the same power for

\(^1\)We can scale the noise in node $j$ by also scaling all the gains $c_{i,j}$ so that the Signal to Noise Ratio (SNR) of the node is unchanged.

\(^2\)This condition is always verified in real networks and can be achieved in a number of ways, like employing different time or frequency slots for the transmit and receive signals or assuming one transmit and one receive antenna with ideal echo cancellation.

\(^3\)Since we consider the load matrix as given our analysis applies to guaranteed quality services and not to best effort services.

\(^4\)The results can be extended to real modulations by using the Gap approximation e.g. [13].

\(^5\)More generally we could consider implementations where $r \leq b \log_2(1+s/n)$ holds. However implementations where $r < b \log_2(1+s/n)$ holds cannot be optimal, since they waste some power, and can be discarded.
the same load so that, without loss of generality, we only consider FDM. In a single-hop FDM network each link signal \( s_{i,j}(t) \) has to be loaded exactly with the information to be transferred from node \( i \) to node \( j \) so that \( R = L \) results. In addition the signals \( s_{i,j}(t) \) are assigned disjoint frequency bands by reserving to each link a fraction \( W_{i,j} \) of the total band \( B \), so that
\[
\sum_{i,j=0}^{N-1} W_{i,j} = B \tag{6}
\]
holds and we conveniently introduce a bandwidth assignment matrix \( W = \{ W_{i,j} \} \). Due to the Shannon capacity achieving transmission, the link signals are independent Gaussian processes with a flat spectrum contained in the assigned band. The signal \( s_{i,j}(t) \) is entirely specified by its power, denoted by \( p_{i,j} \), and its band \( W_{i,j} \).

We now specialise equation (5) to each link of a FDM network. Consider link \( i, j \) and denote by \( g_{i,j} = c_{i,j}^2 \) the channel power gain. The received power is \( s = p_{i,j} g_{i,j} \). Concerning the noise note that even though the signals are transmitted at the same time they do not interfere with each other since they are transmitted in disjoint bands. Therefore the only noise source is the AWGN contained in the band \( W_{i,j} \) and the noise power is \( n = N_0 W_{i,j} \). By replacing in (5) and rearranging we obtain the following fundamental set of constraints for a FDM network:
\[
p_{i,j} = (2^{r_{i,j}/W_{i,j}} - 1) \frac{N_0 W_{i,j}}{g_{i,j}} \quad \text{for } i, j = 0...N-1. \tag{7}
\]
Based on the last equations we can characterise what is the set of the possible implementations of a FDM network for a given load matrix. To this end note that as soon as a band assignment is specified, the power of the transmitted signals can be computed from (7) thereby specifying the implementation. More precisely the assignment matrix shall verify the following additional constraints
\[
W_{i,j} \geq 0 \quad i, j = 0...N-1 \tag{8}
\]
\[
W_{i,j} > 0 \quad \text{if } r_{i,j} > 0 \tag{9}
\]
for physical reasons. Therefore the set \( F \) contains infinite possible implementations, one for each band assignment matrix verifying (6), (8) and (9).

Let us now investigate what is the consumption of a FDM network and how we can find its optimal implementation. To this end note that the total power is given by
\[
P = \sum_{i,j=0}^{N-1} p_{i,j} = \sum_{i,j=0}^{N-1} (2^{r_{i,j}/W_{i,j}} - 1) \frac{N_0 W_{i,j}}{g_{i,j}} \tag{10}
\]
as a function of the band assignments. Therefore, for any given rate matrix, the consumption of a FDM network is the solution of the following minimisation problem\(^6\)
\[
C_{\text{FDM}}(R) = \min_W \sum_{i,j=0}^{N-1} (2^{r_{i,j}/W_{i,j}} - 1) \frac{N_0 W_{i,j}}{g_{i,j}} \tag{11}
\]
subject to the constraints (6), (8) and (9). The latter turns out to be a standard convex optimisation problem which can be solved by means of a water-filling like algorithm. A solution algorithm is presented in [15]. While the water-filling algorithm yields a truly optimum solution, it has a high computational complexity, therefore we also consider a second FDM protocol, that we term the suboptimum FDM protocol. In this protocol the bandwidth of each link is simply set proportionally to the rate assigned to that link, that is
\[
W_{i,j} = B \frac{r_{i,j}}{\sum_{i,j=0}^{N-1} r_{i,j}}. \tag{12}
\]
For this protocol the set \( F \) trivially contains a single implementation and the consumption, denoted by \( C_{\text{SUFB}}(R) \), is obtained by replacing (12) in (10). It can be shown that (12) is the optimum band assignment when the gains \( g_{i,j} \) of all the links are equal. However also when the gains are not equal this simple suboptimal protocol often yields results close to the optimum, how will be verified by means of simulations.

The second transmission protocol that we consider is the CDMA protocol. In a single-hop CDMA network \( R = L \) must hold and we assume that the link signals \( s_{i,j}(t) \) have a flat power spectrum in the band \( B \). Furthermore we assume that the signals cannot be separated at the receivers and interfere with each other so that this protocol is a model for asynchronous CDMA networks. Specifically we assume that for link \( i, j \) the interference is constituted by the signals transmitted by all nodes except nodes \( i \) and \( j \). Node \( j \) does not participate to the interference since we assumed \( c_{i,j} = 0 \). Node \( i \) does not participate to interference because we assume that the signals \( s_{i,h}(t) \) for \( h \neq j \) are orthogonal\(^7\) to the signal \( s_{i,j}(t) \). Given the Shannon capacity achieving transmission, the link signals are independent Gaussian processes with a flat spectrum in the band \( B \). Therefore a signal \( s_{i,j}(t) \) is entirely specified by its power \( p_{i,j} \).

We now specialise equation (5) to each link of a CDMA network. Consider link \( i, j \). The received power is \( s = p_{i,j} g_{i,j} \) like in a FDM network. Since all the signals are transmitted at the same time and in the same band the noise is constituted by the AWGN contained in the band \( B \) plus the mutual interference among the signals and we have \( n = N_0 B + \sum_{h,k=0}^{N-1} h \neq k, h \neq j \). By replacing in (5) and rearranging we obtain the following fundamental set of constraints for a CDMA network:
\[
p_{i,j} = (2^{r_{i,j}/B} - 1) \frac{N_0 B + \sum_{h,k=0}^{N-1} h \neq k, h \neq j \frac{p_{h,k} g_{h,j}}{g_{i,j}}}{g_{i,j}} \tag{13}
\]
and we note that, for physical reasons, the powers shall verify the following additional constraints:
\[
p_{i,j} \geq 0 \quad \text{for } i, j = 0...N-1. \tag{14}
\]
Concerning the set of possible implementations and the consumption consider that, for any given rate matrix, (13) is a system of linear equations in the unknown powers of the link signals. If the system has a solution that also verifies (14) this

\(^6\)While in general the consumption is a function of the load matrix, for single-hop systems, since \( R = L \), it is more convenient to think of it as a function of the rate matrix.

\(^7\)Since these signals all stem from the same source they are synchronised and can easily be made orthogonal if the channel is not frequency selective.
solution is the only possible implementation of the network and the network consumption \( C_{\text{CDMA}}(R) \) can be obtained by summing the solution powers. While the linear system has \( N^2 \) equations and unknowns, in the appendix we show that it is possible to reduce its solution to that of a system with only \( N \) equations and unknowns, thereby realising a reduction in the computational complexity.

Let us now discuss the feasibility of the FDM and the CDMA protocols. To this end consider a given rate matrix and note that if the system (13) is singular or when the solution vector has negative elements so that it violates (14) the CDMA network is unfeasible. On the contrary a suboptimal FDM network always exists since we can use (12) and (7) to find a band and power assignment for any rate matrix. Also an optimal FDM network exists since there are infinite bandwidth assignments verifying (6) and for any band assignment a power vector can be found using (7). Therefore FDM networks are always feasible while CDMA are not. To understand the physical reason behind this fact suppose that, from a given feasible rate matrix, we want to increase the rate carried by a given link. To this end we need to increase the power transmitted over that link. From (7) one sees that in FDM this can be done without affecting the SNR of the other links. On the contrary from (13) one notes that in CDMA the signals interfere with each other and when the power of the link is increased the other links will receive more interference and will increase their powers too. This will cause more interference to all the links and a new power increase round. The process iterates and may become instable yielding to infinite power for all nodes and yet not achieve the target rate.

IV. MLP NETWORKS

Given a single-hop transmission protocol we can extend it by allowing multi-hop. When multi-hop transmission is allowed the information transfer from the source to the destination can be realised either directly or by passing through other nodes acting as relays. For example a source node \( i \) willing to transmit a load \( l_{i,j} = x \) to a destination node \( j \) can realise the transfer either directly, by loading a rate \( r_{i,j} = x \) onto the signal \( s_{i,j}(t) \), or with a two-hop transmission, by passing through a third node \( k \), loading a rate \( r_{i,k} = x \) and \( r_{k,j} = x \) onto the signals \( s_{i,k}(t) \) and \( s_{k,j}(t) \) and setting to zero the direct signal. Similarly we can consider transmissions using three or more hops. A second way to enlarge the possibilities is that of allowing a source to partition its outgoing information in several flows, each of which is delivered through a different multi-hop path. Continuing the preceding example, when partitioning is allowed the source could split the load in two flows and deliver half of the load directly, by setting \( r_{i,j} = x/2 \), and half through node \( k \), by setting \( r_{i,k} = x/2 \) and \( r_{k,j} = x/2 \). Similarly we could partition the information in three or more flows. Any single-hop protocol can be extended in this way and in this section we investigate the consumption of such MLP extended protocols.

Beside an increased complexity of the devices, the introduction of MLP in a practical network has two main drawbacks. One is the signalling that the nodes have to exchange to schedule the multi-hop and multi-flow transmission which waste some of the network band. The second is that a node acting as a relay shall demodulate and modulate again the information thereby introducing a delay in the transmission. This delay can be compensated at the receiver when combining the different flows but the overall transmission delay will be that of the longest multi-hop path and this poses a limit to the number of hops in practical networks. In any case in our analysis we neglect these factors like was done in previous works on the topic \([8],[9],[10]\).

Consider a single-hop protocol with a known consumption \( C_{\text{single}}(R) \) and assume that the protocol is extended by allowing MLP. Let us investigate the set of the possible implementations and the consumption of the resulting MLP network. To this end note that in a MLP network we can have \( R \neq L \) and, as a preliminary step, we need to better characterise what is the set of rate matrices that can be employed by the MLP network. This set is constituted by all the rate matrices that allow the delivering of the network load. These matrices will be termed allowable and their set will be denoted by \( \Lambda \). In order to specify this set note that, obviously, \( L \in \Lambda \). Next suppose that we are given an allowable rate matrix \( R \in \Lambda \) and consider the link between two nodes, say \( i \) and \( j \), over which a rate \( r_{i,j} \) is being transferred. Exploiting MLP we can deliver the same load by delivering a piece \( a \), where \( a > 0 \) and \( a \leq r_{i,j} \), of the information transferred over that link not directly but by passing through third node, say node \( k \), instead. This mechanism will be called a link reallocation. If we realise a link reallocation we obtain a new, allowable rate matrix \( R' \in \Lambda \) which is equal to \( R \) except in three elements, namely \( ^8 \)

\[
\begin{align*}
  r'_{i,j} &= r_{i,j} - a \\
  r'_{i,k} &= r_{i,k} + a \\
  r'_{k,j} &= r_{k,j} + a.
\end{align*}
\]

(15)

The set \( \Lambda \) can now be specified: it is composed by \( L \) and by all the matrices obtained from \( L \) by means of a sequence of link reallocations. Note that in the sequence of reallocations the same information can be reallocated more than one time, thereby realising multi-hop, and that at each reallocation any valid amount of information can be reallocated, thereby allowing to partition the load as finely as desired.

Having characterised \( \Lambda \) we now discuss the MLP network consumption. To this end note that for any rate matrix \( R \in \Lambda \) the underlying single-hop network may have zero, one or many implementations. The set \( F \) of the possible implementations of the MLP network contains all these implementations. However, when looking for the MLP network consumption, we can discard all the non optimum implementations of the single-hop network, since these cannot be the optimum implementation of the MLP network. In this way, for each \( R \in \Lambda \), we are left with a single implementation, the total power of which is given by \( C_{\text{single}}(R) \). Therefore the MLP network consumption can be found by solving the following optimisation problem:

\[
C_{\text{MLP}}(L) = \min_{R \in \Lambda} C_{\text{single}}(R).
\]

(16)

The latter problem is one of the SRRR problems introduced, with a different formulation, in \([9]\). When the single-hop network supporting the MLP network is FDM this is a convex

\(^8\)We use square brackets to denote the reallocation operation so that, in the example at hand, we can write \( R' = [R]_{i,j,k,a} \).
optimisation problem and an elegant method for its solution was presented in [9]. On the contrary if the single-hop network is CDMA the problem is not convex and a suboptimum solution method was proposed in [10]. However that method assumes that the network operates at a high SNR, a situation which not often arises in practice since CDMA networks normally operate at low SNR. As an additional way to tackle this problem, in the following we propose a Greedy algorithm that achieves a local minimum and is simple and general. A Greedy approach to the problem at hand is that of starting from a rate matrix equal to the load matrix and of performing a sequence of reallocations aimed at decreasing the total power.

A possible way to realise the approach is presented in the following Delta Rate Sharing Algorithm (DRSA):

**DRSA**(δ, L):

1) Set R = L.
2) Set \( P^* = C_{\text{single}}(R) \).
3) Set \( P^+ = C_{\text{single}}(R) \).
4) For \( i, j = 0, \ldots, N - 1 \):
   a) If \( r_{ij} > 0 \) then
      b) For \( k = 0, \ldots, N - 1, k \neq i, j \):
         c) Set \( R' = [R]_{i,j,k} \).
         d) If \( P^+ > C_{\text{single}}(R') \) then let \( P^+ = C_{\text{single}}(R') \) and \( R = R' \).
         e) End for k.
      f) End if.
   g) End for i, j.
   h) If \( P^+ > P^+ \) then set \( P^* = P^+ \) and goto 4) else return R and stop.

The DRSA attempts to decrease the total power of the network by exploring the allowable set \( \Lambda \) and retaining improvements. Specifically, starting from a rate matrix \( R = L \), it checks all the links (in the cycle over \( i, j \)) and tries to reallocate a portion \( \delta \) of the rate of the link through node \( k \) for \( k \) varying over all the other nodes. It computes the power of the modified system and if it is less than the original power it keeps the reallocation; otherwise the reallocation is discarded. Once all the links have been checked the algorithm controls (in step 12) if the power was decreased. If not, no change in the rate matrix took place in the last reallocation round so that it is useless to try another one. The algorithm returns the rate matrix found, from which the required power and the implementation can be obtained, and stops. By contrast, if the power was decreased, another reallocation round is carried out since the power could be further decreased.

Several considerations can be done concerning the DRSA. Firstly note that, trivially, the algorithm converges, since the power is reduced at every rate update and it is lower limited. Next note that we expect to have better solutions when \( \delta \) is low, ideally \( \delta \rightarrow 0 \). In fact when \( \delta \approx 0 \) the difference of the power before and after the link reallocation is a way to numerically compute the derivative of the \( C_{\text{single}}(R) \) function along a direction contained in the allowable domain \( \Lambda \). In this sense the algorithm can be seen as a stepwise Gradient Descent algorithm that moves the solution point in \( \Lambda \) along a direction where the objective function decreases. Therefore if \( \delta \approx 0 \) when the algorithm stops a local minimum has been found. As a further consideration note that the set \( \Lambda \) is convex. In fact it is immediate to verify that given \( R_0, R_1 \in \Lambda \) and \( 0 \leq a \leq 1 \) we have that \( R = aR_0 + (1 - a)R_1 \in \Lambda \). Therefore, if also the objective function \( C_{\text{single}}(R) \) is convex, the DRSA converges towards a global optimum yielding the exact consumption. However we note that the \( C_{\text{single}}(R) \) function is not guaranteed to be convex. Indeed if no constraints exist on the channel gains it is not difficult to find examples of ill conditioned networks where the algorithm is trapped in a local minimum. But when the gains are based on a physically meaningful model the DRSA can be expected to yield a result close to the optimum.

The DRSA algorithm just introduced is not practical for direct implementation and a number of improvements aimed at decreasing its computational complexity can be introduced. As a first consideration note that, while a value of \( \delta \approx 0 \) would be desirable, directly using such a low value would slow down the convergence to a solution so that, like in a Gradient Descent algorithm, it is convenient that \( \delta \) is a sequence of values tending to zero. Therefore we consider the following Rate Sharing Algorithm (RSA):

**RSA**(L):

1) Set R = L.
2) Set \( \delta = \max_{i,j} r_{ij} \).
3) Set \( R = \text{DRSA}(\delta, R) \).
4) If \( \delta > \delta_{\text{min}} \) then \( \delta = \delta/\Delta \) and goto 3) else return R and stop.

The RSA algorithm simply runs the DRSA algorithm for several, decreasing values of \( \delta \), lower bounded by \( \delta_{\text{min}} \). Controlling the minimum amount of information that can be reallocated is also important for making the algorithm a practical one since the control signalling involved by a reallocation is likely to make it inconvenient in real networks if the amount of information being reallocated is too small. The pace at which \( \delta \) is decreased is controlled by the parameter \( \Delta > 1 \). A second obvious way to reduce the computational complexity is that of eliminating from the search links with a low gain since these links are unlikely to be exploited in practical networks. We obtain a slightly different version of the algorithm, termed the Fast RSA, which is identical to the RSA except that step 6) of the DRSA is changed to

**DRSA**(δ, L):

5) ........................
6) For all \( k \in I_i \)
7) ........................

The computational complexity of the algorithm is difficult

...
to compute and depends on a number of factors including the network gain and load matrices. Therefore direct running time measurements will be presented in section V to gather an idea of the complexity. To conclude this section we discuss the applicability of the RSA algorithm. Concerning this point note that, while in the paper we focus on the FDM and CDMA networks, the RSA algorithm can be used to reduce the power requirements of any single-hop network for which the function $C_{\text{single}}(R)$ can be computed, including networks where the channel model is not flat and sophisticated medium access procedures are employed. As an example we could optimise a MLP Orthogonal Frequency Division Multiple Access (OFDMA) network with frequency selective channels by using the results of [12] where a method is presented for computing the function $C_{\text{OFDMA}}(R)$. Next note that even though only the power minimisation problem was considered in this paper the RSA algorithm can be modified to handle any other SRRA problem with given load matrix simply by replacing the $C_{\text{single}}(R)$ function with an appropriate alternative function.

V. RESULTS

In this section we compare the consumption of several different protocols. For single-hop networks we use the exact consumption while for MLP networks we use the upper bound to the consumption yielded by the RSA algorithm. In order to perform a statistically meaningful comparison we introduce two simple random models for producing the network’s gains and load that we term the Ad-Hoc network and the Cellular network respectively. In both models the $N$ nodes are positioned randomly, uniformly over a square grid having a 50 m edge, enforcing a minimum distance of 2 m between them. Node 0 of the Cellular network is an exception since it is always posed at the centre of the grid. The gains are computed based on the distance $d_{i,j}$ between the nodes, according to the following equation:

$$g_{i,j} = \frac{f_{i,j}}{A_0 d_{i,j}^A}$$  \hspace{1cm} (17)

where $A_0$ is the attenuation at one meter, which is taken $A_0 = 1$ since it is a direct scale factor for the network consumption, $A_e = 4$ is the exponent of the attenuation and $f_{i,j}$ is the determination of a random variable with exponential distribution and mean one taking into account Rayleigh fading. Next the load matrix is produced based on two parameters: the number of active links, denoted by $K$, and the network spectral efficiency denoted by $\eta$ (bit/sec/Hz). Specifically the load matrix is produced by firstly selecting at random $K$ pairs of nodes, one of the two being node 0 in the Cellular model, and then by loading each of these links. For the sake of simplicity the load is set equal on all the loaded links and specifically it is set to $l_{i,j} = \frac{B_0}{K}$ so that the throughput of the network is $B_0$ bit/sec. We take the performance parameter of the network to be the Power Consumption to Noise Ratio (PCNR), given by

$$PCNR = \frac{C}{B N_0},$$  \hspace{1cm} (18)

and without loss of generality set $B = N_0 = 1$ since the PCNR is invariant to the actual noise level and bandwidth. The protocols being compared (with the labels used in the figures given between parenthesis) are the optimum FDM (FDM), the suboptimum FDM (SUB), the CDMA (CDMA), the MLP suboptimum FDM (MLP-SUB) and the MLP CDMA (MLP-CDMA).

As a first example Figure 1 reports the PCNR of the five protocols as a function of $\eta$ for a Cellular network with $N = 10$ nodes and $K = 5$ active links. The minimum value of $\eta$ is 0.06 bit/sec/Hz. The curves, as all the following curves, are obtained by averaging the results over 300 randomly generated networks. By comparing the curves one immediately sees the great PCNR benefit obtained by introducing MLP. The gain can be as high as 7 dB for FDM networks and as 14 dB for CDMA networks. The gain is higher for moderate values of $\eta$ and tends to decrease for high and low values of the spectral efficiency. The gain is higher for moderate values of $\eta$ and tends to decrease for high and low values of the spectral efficiency.
efficiency. From the curves one also sees that the optimum and suboptimum FDM have close performance within the range of $\eta$ considered in the figure, the maximum gap being a couple of dB. As an additional comment we note that FDM outperform CDMA and that the gap increases for increasing values of $\eta$. This behaviour is expected since increasing $\eta$ implies a power increase which, for CDMA, implies a higher level of interference among nodes. As a consequence CDMA and MLP-CDMA become unfeasible at high rates. Indeed only a percentage of the 300 random networks tested was feasible under these protocols and the PCNR value was averaged only on these networks. This percentage is plotted in Figure 2 as a function of the spectral efficiency for both the Cellular network and for an Ad-Hoc network having the same number of nodes and links. One sees that on the Cellular network CDMA is 100% feasible up to a spectral efficiency of 1 bit/sec/Hz after which the feasibility drops. MLP is not capable of increasing the feasibility significantly. Note that to give a feedback of the festivity on the PCNR plots we set the value of the PCNR to $\infty$ when the feasibility drops below 50% as seen in Figure 1. From Figure 2 one sees that the Ad-Hoc network is an harder environment since the feasibility of CDMA and MLP-CDMA is not 100% even at low spectral efficiency and is lower than in the Cellular network. CDMA is now 50% feasible only up to 0.25 bit/sec/Hz. Introduction of MLP is beneficial and increases the feasibility. Indeed MLP-CDMA is 50% feasible up to 0.4 bit/sec/Hz. The PCNR in the Ad-Hoc network is reported in Figure 3. One sees that similar comments to those made for the Cellular network can be repeated and that the curves are shifted up by approximately 5 dB with respect to Figure 1. The worst performance on the Ad-Hoc network with respect to the Cellular one can be explained by noting that the average link length is higher in Ad-Hoc networks.

In Figure 4 the performance of the Fast RSA is compared to that of the RSA for a Cellular and an Ad-Hoc network with $N = 10$ nodes and $K = 5$ links. Four curves are reported for each topology. The ones labelled MLP-SUB-10 and MLP-CDMA-10 give the PCNR of these networks when optimised by means of the RSA while those labelled MLP-SUB-4 and MLP-CDMA-4 give the PCNR when the Fast RSA with $V = 4$ neighbours is employed. As can be seen from the figure the performance loss incurred by the Fast RSA depends on the topology and on the protocol: Ad-hoc topology and CDMA networks suffer more but the loss is limited to a couple of dB. For FDM networks the loss is negligible. Since the loss incurred by the Fast RSA is not great while its computational complexity is much lower than that of the RSA in the following examples we will only use the Fast RSA with $V = 4$. By using the Fast RSA, networks with up to twenty nodes can easily be studied. For example Figure 5 reports the PCNR of a network with $N = 20$ nodes and $K = 5$ links.

Fig. 3. PCNR as a function of the spectral efficiency for a Ad-Hoc network with $N = 10$ nodes, $K = 5$ active links.

Fig. 4. Comparison of the performance of the RSA (10 neighbouring nodes) and of the Fast RSA (4 neighbouring nodes) in a Cellular and in an Ad-hoc network with $N = 10$ nodes and $K = 5$ links.

Fig. 5. PCNR as a function of the spectral efficiency for a Cellular network with $N = 20$ nodes, $K = 5$ active links.
by a change in the number of the networks node except that MLP networks gain one or two dB at low spectral efficiency. As an additional example Figure 6 reports the PCNR of a Cellular network with \( N = 20 \) nodes and \( K = 10 \) links. By comparing with Figure 5 we see that increasing the number of links (i.e. a more fractionated traffic) does have an impact on the PCNR. The CDMA networks are those suffering more: a notable PCNR loss is incurred. Also single-hop FDM networks suffer. On the contrary MLP FDM networks do not suffer at all the increased number of links.

In Figure 7 we study the PCNR for high values of \( \eta \). The network considered is a Cellular one with \( N = 20 \) nodes and \( K = 10 \) links. Only the FDM networks are considered in the figure since CDMA networks are unfeasible in this range of \( \eta \). The figure shows two important points. Firstly that the loss incurred by the suboptimal FDM with respect to the optimum FDM grows and becomes unacceptable. As a result the single-hop FDM network outperforms the MLP suboptimal FDM network. Secondly one sees that the gain of the MLP technique decreases with increasing \( \eta \); as a result the curves of the suboptimal FDM and of the MLP FDM tend to touch. However this happens well beyond the spectral efficiency achieved by currently existing networks.

Finally in Table 1 we present results concerning the computational complexity. The results are obtained using an efficient C language implementation of the protocols running over an average laptop computer. The first three rows of the table report the average running time (msec) required to optimize one Cellular network at a spectral efficiency of 1 bit/sec/Hz by the FDM, MLP-SUB and MLP-CDMA protocols. The last two rows of the table report the number of reallocations attempted by the MLP protocols which is equal to the number of times that the RSA had to evaluate the power of an underlying single-hop network. Comparing the first column, where \( N = 10 \) and \( K = 5 \), with the second one, where \( N = 20 \) and \( K = 5 \), one sees the dramatic impact of the number of nodes on the complexity. Comparing the second column with the third column, where the fast RSA is used, one sees that the fast RSA is capable of greatly reducing the running time. The last column, where \( K = 10 \), shows that also the number of active links has an impact on the complexity even though less dramatic than the number of nodes.

### VI. Conclusions

We presented algorithms to compute the consumption of several single-hop and MLP networks. We run Monte-Carlo simulations to compare the consumption of these networks. The results showed that the MLP technique is capable of decreasing the network consumption by more than one order of magnitude (up to 14 dB in the simulations carried out) and that it is effective for both FDM and CDMA physical layers. The greater benefits are obtained for moderate values of the spectral efficiency (around 1 bit/sec/Hz) while for very high values of the spectral efficiency MLP is less effective or useless. The simulations also showed that the MLP benefits are greater in Ad-Hoc networks than in Cellular networks indicating that the MLP is more effective in harsher propagation environments. Indeed unreported simulations showed that the MLP benefits are negligible when free space propagation is assumed.
ACKNOWLEDGEMENTS

The author is grateful to Prof. S. Barbarossa and to Dr. G. Scutari of the INFOCOM dept. of the University of Rome “La Sapienza” for fruitful discussions and suggestions. The author would also like to thank the anonymous reviewers for their fruitful comments.

APPENDIX

Let us simplify the notation by introducing $\eta_{i,j} = (2^{-\gamma_{i,j}} - 1)/g_{i,j}$ and by rearranging (13) to get

$$p_{i,j} - \eta_{i,j} \sum_{h,k=0}^{N-1} p_{h,k} g_{h,j} = \eta_{i,j} N_0 B$$

for $i, j = 0, ..., N - 1$. Next by recalling that $g_{i,i} = 0$ for any $i$ and by summing over $k$ in the last equation we obtain:

$$p_{i,j} - \sum_{h \neq i} p_{h} g_{i,j} = \eta_{i,j} N_0 B$$

Finally by introducing $\gamma_{i,h} = \sum_{j} \eta_{i,j} g_{h,j}$ and $\mu_i = \sum_{j} \eta_{i,j} N_0 B$ and by summing over $j$ in (20) one gets

$$p_i - \sum_{h \neq i} p_{h} \gamma_{i,h} = \mu_i$$

for $i = 0, ..., N - 1$. The latter system express the original one directly in the power emitted by each node. It is a system of $N$ equations in the $N$ unknown powers $p_i$. From the solution of the system we can compute the network consumption by summing the powers $p_i$ and, if desired, the network implementation can be obtained by computing the power assigned to each link from (20) as follows:

$$p_{i,j} = \eta_{i,j} (N_0 B + \sum_{h \neq i} p_{h} g_{h,j})$$

for $i, j = 0, ..., N - 1$.

REFERENCES


Lorenzo Piazzo was born in Rome, Dec. 1964. He received the “ Laurea” degree (cum laude) in Electronic Engineering from the University of Rome “La Sapienza” in 1991. From 1992 to 1995 he carried out his Ph.D. program in “Information and Communication Engineering” at the INFOCOM Dept. of the university of Rome “La Sapienza,” working on speech coding and channel coding. From 1994 to 1996 he was with Space Engineering, working on spectral analysis and SAR images processing. From 1997 he is with the INFOCOM Dpt. of the university of Rome “La Sapienza.” In 1997-1998, in the framework of the ACTS project MEDIAN, he was interested in the development of synchronization schemes for OFDM transmission systems and in the assessment of the impact of phase noise on OFDM systems. In 1999, in the framework of a EU-TMR grant, he was with the dept. of ECS of the University of Southampton, UK, where he worked on Turbo TCM over OFDM and on power and bit allocation in adaptive OFDM systems. From 2000 to 2006 he was interested in UWB transmission in the framework of the IST project WHYLESS.COM. Since 2005 he is working on multi-hop networks in the framework of a research project funded by the faculty of Engineering of the University “La Sapienza.” Since 2005 he is cooperating with IFSI (Istituto per la Fisica dello Spazio Interplanetario) in the development of the On Board Software for the ESA Herschel-Planck satellite.
Power Reduction Techniques for Multiple-Subcarrier Modulated Diffuse Wireless Optical Channels

Weiwei Kang, Student Member, IEEE, and Steve Hranilovic, Senior Member, IEEE

Abstract—In this paper, two novel techniques are proposed to reduce the average optical power in wireless optical multiple-subcarrier modulated (MSM) systems, namely in-band trellis coding and out-of-band carrier design. Data transmission is confined to a bandwidth located near DC. By expanding the signal set and coding over the increased degrees of freedom, an in-band trellis coding technique achieved an average optical power reduction up to 0.95 dB over conventional MSM systems while leaving the peak optical power nearly unaffected. With a symbol-by-symbol bias, the received DC level can be detected to provide a degree of diversity at the receiver. In this manner, an additional average optical power reduction up to 0.50 dB together with a peak power reduction of 0.46 dB is achieved. Moreover, the unregulated bandwidth available in wireless optical channels is exploited and out-of-band carrier signals are designed outside the data bandwidth to reduce the average optical power. Average optical power reduction as high as 2.56 dB is realized at the expense of 4 out-of-band carriers and an increase in the peak optical power. Finally, combining the three techniques achieves the best average optical power reduction of 2.63 dB optical.

Index Terms—Wireless infrared channel, indoor diffuse infrared communication, optical intensity modulation, multiple-subcarrier modulation, trellis-coded modulation.

I. INTRODUCTION

DIFFUSE indoor wireless optical channels are an exciting complement to existing radio frequency (RF) systems due to their low cost, high security and freedom from spectral licensing issues [1], [2]. The common problem of co-channel interference and multi-access interference in RF systems is eliminated since optical transmissions are blocked by opaque objects. Whereas the spectrum in RF channels is strictly licensed, the optical band is unlicensed worldwide, providing a potentially large bandwidth. However, multipath distortion in indoor wireless optical channels imposes a sharp bandwidth constraint and confines data transmission to a lowpass region [1]. Multiple-subcarrier modulated (MSM) wireless optical systems send data in this lowpass region directly, improving spectral efficiency at the expense of optical power efficiency [3]. However, in previous work the higher frequency bands are ignored since they are effectively filtered-out by the multipath channel. In this paper, higher frequencies are exploited and out-of-band signals are designed not to send information, but to reduce the average optical power in wireless optical systems. These out-of-band emissions are contained to a given room and do not contribute to co-channel interference, as in RF systems. Data transmission is confined to an in-band region at low frequencies around DC where the channel attenuation and noise power are low. Additional gains in average optical power are realized by coding over sequences of inputs.

In diffuse indoor wireless optical systems, data are modulated onto the instantaneous intensity of an optical carrier which is emitted over a wide solid angle and allowed to reflect from surfaces in the room. As a result, only non-negative signal amplitudes can be sent on the channel. Additionally, the peak transmitted amplitude must be limited due to the limited dynamic range of the transmitter. Eye and skin safety regulations require that the average amplitude, i.e., average optical power, also be constrained. As a result of these amplitude constraints the direct application of modulation from electrical channels is not possible.

Electrical MSM systems suffer from a high peak-to-average power ratio (PAPR) due to the addition of many independent frequency carriers. This results in nonlinear distortion due to clipping of the transmitted waveform. The reduction of the PAPR in electrical MSM systems has been studied using a variety of techniques, including the use of coding [4]–[6], reserved subcarriers [7], [8], parallel-combinatory multiple-subcarrier (PCMS) techniques [9] and tone-injection [10]. A comprehensive review of electrical PAPR reduction techniques can be found in [11]. Although these techniques reduce the PAPR effectively in electrical MSM systems, their direct application to reduce the average optical power of optical wireless MSM systems is often inefficient. It is important to note that the objective of these techniques is to reduce the variance in the squared envelope of an electrical signal rather than minimizing the amplitudes, i.e., optical intensities, directly. In order to achieve gains, the amplitude constraints of wireless optical channels must be considered explicitly in the design of algorithms.

In fiber optic systems, MSM techniques have been applied to improve the aggregate capacity of video distribution sys-
tems [12]. Carruthers and Kahn [3] first investigated MSM indoor wireless optical communications as a means to provide bandwidth efficient communication as well as multiple access. Independent data were transmitted on each in-band carrier and a fixed bias was added to each symbol to ensure non-negative amplitude. A time-varying bias signal was introduced in [13] as a means of both reducing the average optical power and improving detection. As in electrical systems, large positive and negative amplitude peaks may occur in optical MSM systems. The negative peaks require a large bias, and hence average optical power, to ensure non-negativity. This problem was first treated in [14] where optimized block codes were developed to reduce the average optical power in MSM systems with QPSK symbols. In [15], reserved subcarriers are added to the in-band BPSK signals and the transmitted amplitudes optimized to provide an average optical power reduction. The PCMS technique [9] was applied to wireless optical MSM systems by adding a symbol-by-symbol bias [16]. In [17], constellation point assignments over all carriers in a given symbol were chosen from a subset of MSM signals with low average optical power such that the minimum distance among each pair of codewords is maximized. Reference [18] provides an overview of optical intensity MSM techniques. A notable trend in previous work is that solely block coding was considered and no coding was applied over sequences of input symbols. Additionally, the reduction of both peak and average optical powers was not expressly treated.

In this work, we propose two main approaches to realize a gain in average optical power for MSM wireless optical systems: in-band coding and out-of-band carrier designs. In contrast with previous work, a trellis coding technique is developed to impose structure on sequences of in-band carrier symbols to realize a reduction in both peak and average optical power. Additionally, the DC bias used to ensure non-negativity is also used in code design and detected to provide an additional degree of diversity at the receiver. We also propose the use of carriers outside of the channel bandwidth to alleviate the average optical power constraint. Although others have considered the use of reserved subcarriers in RF and wireless optical channels, these carriers were added to the in-band carriers at a cost of reduced spectral efficiency. Our approach adds carriers at high frequencies outside of the data transmission bandwidth of the channel so that the spectral efficiency is unaffected. This approach is especially well suited to wireless optical channels since their spectra are unregulated and the inherent isolation of optical emissions does not cause interference with other users.

The paper is organized as follows. Section II describes a popular model for diffuse indoor wireless optical channels and introduces optical MSM systems. Section III presents our in-band trellis coding and DC detection technique. Two out-of-band carrier design techniques are introduced in Sec. IV and their complexities compared. Simulation results of the proposed techniques along with a comparison to existing techniques are given in Sec. V. The paper concludes in Sec. VI with some directions for future work.

II. CHANNEL MODEL AND WIRELESS OPTICAL MSM SYSTEM

A. Channel Model

Indoor diffuse wireless optical channels are well modelled by the following baseband linear system [1]

$$y(t) = r x(t) \otimes h(t) + n(t)$$  \hspace{1cm} (1)

where $\otimes$ denotes convolution, $r$ [A/W] is the photodetector responsivity, $x(t)$ [W] is the transmitted intensity-modulated signal, $h(t)$ is the channel impulse response, $n(t)$ [A] is the noise process and $y(t)$ [A] is the received photocurrent. Without loss of generality, we set $r = 1$ for the balance of this paper. The dominant noise source in indoor wireless optical channels is shot noise resulting from high-intensity ambient light. The noise is well modelled as being white, signal independent and Gaussian distributed with variance $\sigma^2$ [1].

Although multipath fading is not an issue in such channels due to the inherent spatial diversity of the receiver, multipath distortion resulting from multiple reflections within the room leads to a lowpass channel frequency response [1]. The bandwidth of the channel is on the order of 10 – 40 MHz in most configurations [19], [20]. In this work we assume that the channel is flat in the bandwidth used for data transmission as shown in Fig. 1. This flat lowpass region of spectrum is termed the in-band region while all higher frequencies are termed out-of-band. Due to multipath distortion, the attenuation in the out-of-band region is high and signals transmitted in this band are heavily distorted at the receiver.

Since $x(t)$ is an optical power signal it must satisfy

$$x(t) \geq 0.$$  \hspace{1cm} (2)

The average optical power, $P_a$, must also be limited due to eye safety requirements. From (1), a constraint is placed on
the average amplitude
\[ P_a = E[x(t)] \] (3)
rather than \( E[x^2(t)] \) as in electrical channels. Thus, a gain in optical power, in decibels, is equivalent to twice of that value in electrical dB. For example, a 1 dB improvement in average optical power corresponds to a 2 dB gain in electrical domain. Moreover, the maximum amplitude
\[ P_p = \max_t x(t) \] (4)
should also be limited due to the limited dynamic range of the transmitter.

B. Wireless Optical MSM System Definition

Consider an optical MSM system with \( W \) in-band and \( L \) out-of-band carriers, as in Fig. 1. Fig. 2 presents a diagram of the MSM system. On each of the \( W \) in-band carriers a complex information-bearing symbol, \( c_1 \ldots c_W \), is transmitted while the complex values on the \( L \) out-of-band carriers, \( c_{W+1} \ldots c_{W+L} \), are designed to reduce the transmitted average optical power. A bias, \( c_0 \), must also be transmitted to ensure non-negativity of the output signal. Sections III and IV discuss the design of the in-band encoder and out-of-band carriers respectively. It is assumed that the symbol period is limited to \( T \) seconds and that a rectangular window is employed. Since only real-valued signals can be transmitted, Hermitian, i.e., conjugate, symmetry must be preserved in the spectrum. The output waveform is generated by taking the inverse fast Fourier transform (IFFT) of the sequence and appending a cyclic prefix to ensure independence between carriers is preserved with transmission through the channel. The resulting MSM waveform transmitted is given by
\[ x(t) = \sum_k \text{Re} \left[ \sum_{i=0}^{W+L} c_i^k \exp(-j2\pi i(t - kT)/T) \right] \text{rect}((t - kT)/T) \] (5)
where
\[ \text{rect}(t) = \begin{cases} 1 & : t \in [0,1] \\ 0 & : \text{otherwise} \end{cases} \]
Notice that in order for \( x(t) \) in (5) to satisfy the non-negativity constraint (2), the bias \( c_0^k \in \mathbb{R} \) must be chosen appropriately. Consider selecting the minimum bias required to ensure non-negativity in each symbol interval, that is,
\[ c_{0}^k = -\min_{t \in [0,T]} s_{k}(t) \] where
\[ s_{k}(t) = \text{Re} \left[ \sum_{i=1}^{W+L} c_i^k \exp(-j2\pi it/T) \right] \] (6)
Notice that \( s_{k}(t) \) can be interpreted as the unbiased output from the \( W \) in-band and \( L \) out-of-band carriers. It has been shown that adding such a symbol-by-symbol bias offers significant optical power reduction [13], [14]. The average optical power (3) of the \( x(t) \) is then,
\[ P_a = E[x(t)] = E[c_0^k]. \]
Thus, the goal of an optical power reduction technique is to reduce the expected magnitude of the required DC bias. The amplitudes selected on the \( W \) in-band and \( L \) out-of-band carriers are designed to minimize \( c_0^k \) and hence the average optical power of the wireless optical MSM system.

As shown in Fig. 2, the receiver performs the inverse of the transmitter by discarding the cyclic prefix and by performing a
fast Fourier transform (FFT) on the sampled data. Again, it is assumed that the addition of the cyclic prefix has maintained the independence of signals amongst the carriers. In this work, the particular details of how the inter-carrier interference is eliminated between the carriers is not important as there are many techniques which can be applied. Notice that only the DC bin and the \( W \) data-bearing carriers are used by the decoder and that the out-of-band carriers are discarded. The out-of-band carriers do not carry independent information but are added to alleviate the amplitude constraints at the transmitter. It is important to note that the out-of-band carriers do not interfere with users in other rooms due to the inherent containment of diffuse infrared radiation.

### C. Definitions

The bandwidth measure adopted in this paper is the null-to-null bandwidth, in accordance with previous work. The bandwidth of the transmitted signal is determined by \( T \) and the number of in-band carriers, \( W \). Since the \( L \) out-of-band carriers are not required for detection and are contained to a given room, they are not considered in the bandwidth computation. Assuming \( K \) bits are sent per symbol period, the null-to-null bandwidth efficiency of the system is defined as

\[
\eta = \frac{K}{W+1} \text{[bits/sec/Hz]},
\]

(7)

This measure is adopted as a fair metric of comparison for the schemes considered here since it quantifies the data rate for a given bandwidth cost, even in the presence of coding over the symbols.

To facilitate the comparison of average and peak optical power requirement of various techniques, define a reference rectangular on-off keying (OOK) wireless optical system with bit error rate (BER) \( 10^{-6} \), average optical power \( P_{\text{aOOK}} \) and peak optical power \( P_{\text{pOOK}} \) as is conventionally done \([1, 14]\). As in \([2]\), for this BER, \( P_{\text{aOOK}} = 4.75\sqrt{R_{\text{OOK}}/g^2} \) where \( R_{\text{OOK}} \) is the bit rate of the reference system. Similarly, the peak optical power \( P_{\text{pOOK}} \) is given by \( P_{\text{pOOK}} = 2P_{\text{aOOK}} \) and \( \eta_{\text{OOK}} = 1 \text{bit/s/Hz} \). The peak and average optical powers of all wireless optical MSM schemes considered are normalized by \( P_{\text{aOOK}} \) and \( P_{\text{pOOK}} \) respectively while operating at the same bit rate and BER as the reference OOK system to yield,

\[
\rho = 10\log_{10} \frac{P_a}{P_{\text{aOOK}}} \text{[dB]}, \quad \psi = 10\log_{10} \frac{P_p}{P_{\text{pOOK}}} \text{[dB]}
\]

(8)

where \( \rho \) is termed the normalized average optical power and \( \psi \) is the normalized peak optical power.

To conform with previous work in the area, we adopt in this paper the conventional term “normal MSM system” \([14]\) to denote an optical MSM system with QPSK constellation employed on each carrier and a symbol-by-symbol bias.

### III. In-Band Optical Power Reduction Techniques

Consider the encoding and mapping blocks for the \( W \) in-band carriers in Fig. 2. This section presents an in-band coding technique to reduce the average optical power in wireless optical MSM systems. Unlike block coding considered in \([14]\), the trellis coding technique presented here imposes a structure on sequences of MSM symbols. Finally, a DC detection technique is presented which exploits the signal-space diversity to realize a reduction in the average optical power.

#### A. Design of In-Band Trellis Codes

Inspired by Ungerboeck’s work on coded modulation \([21]\) and Frenger’s work on PCMS \([9]\), we expand the signal set for each carrier and develop trellis coded modulation over the increased degrees of freedom to realize gains in average optical power. The resulting system is termed trellis-coded MSM (TCMSM) in this paper and can achieve gains in average and peak optical power. As in \([21]\), coded modulation system design includes constellation expansion, set partitioning, choice of trellis structure and a search for a mapping between constellation points and edges in the trellis.

1) MSM Constellations: Given an MSM wireless optical system with \( W \) in-band carriers, define a sub-constellation as the common constellation employed on each frequency carrier. In our scheme, each sub-constellation is expanded by adding a zero amplitude to a QPSK constellation or to an 8-PSK constellation, as is done in \([9]\). The resulting constellations are called 5-APSK and 9-APSK respectively. This constellation expansion increases the degrees of freedom available in choosing constellation points. Unlike earlier work \([16]\), we do not transmit independent information on these degrees of freedom directly, but develop coded modulation to improve the optical power efficiency.

Further, define an MSM constellation point as an assignment of \( W \) sub-constellations to each of the \( W \) in-band carriers. Trellis codes are designed in Sec. III-A.2 to exploit the added degrees of freedom to provide sequences of MSM constellation points with improved Euclidean distance. The cost of this technique is additional complexity in the form of a trellis encoder at the transmitter and a Viterbi decoder at the receiver.

Each MSM constellation point is constrained to have a zero amplitude on at least \( U \) carriers, \( 0 \leq U \leq W \), for all data symbols. For \( U > 0 \), this constraint is shown to enable a suboptimal search algorithm to find codes with lower average optical power requirements. In code design, each sub-constellation is partitioned to the lowest level where only a single constellation point is contained.

2) Code Design: A trellis structure is used to design sequences of MSM constellation points. Each trellis edge, determined by both the input bits and the current state, is mapped to an MSM constellation point. The trellis structure is fixed to have \( M \) states with at least one branch to each next state. If \( K \) bits are to be sent per symbol interval, each branch in the trellis must therefore send \( K - \log_2 M \) bits using parallel transitions.

Given the trellis structure, the mapping between MSM constellation points and branches in the trellis determines the free Euclidean distance, \( d_{\text{free}} \), of the code. The \( d_{\text{free}} \) asymptotically determines the error performance of the code and is the minimum distance for any error path leaving and returning to the all-zeros codeword. In this scheme, \( d_{\text{free}} \) can be limited by two MSM constellation points on the
same branch or two different paths that start and end at the same state. Denote these distances as \( d_p \) and \( d_s \) respectively. Motivated by Ungerboeck’s trellis code design heuristics [21], the following rules are applied when searching for a mapping:

1) Any MSM constellation point can occur only once at each stage of the trellis.
2) Maximize the distance between parallel edges, \( d_p \).
3) Pick points on edges leaving or entering the same state with maximum \( d_s \).

Rule 1 avoids confusion in decoding at each stage. Rules 2 and 3 are designed to maximize \( d_p \) and \( d_s \) respectively. Figure 3 presents a hand-designed example of a two state code \((M = 2)\) satisfying the above heuristic and sending \( K = 2 \) bits per symbol using \( W = 2 \) in-band carriers with 5-APSK sub-constellations and constraint \( U = 1 \). The labels \( S_i \) refer to MSM constellation points. Notice that for all parallel edges, \( d_p \) is maximized subject to the \( U = 1 \) constraint. The remaining MSM constellation points are chosen to maximize \( d_s \). At each stage of the trellis, given the current state, one input bit selects the next state while the other bit selects which of the MSM constellation points, i.e., edges \( S_i \), to transmit. Note that these heuristics do not necessarily lead to an optimal code in any sense, but are designed to yield good performance.

To transmit \( K \) bits per symbol, the number of MSM constellation points assigned to each stage of the trellis is

\[
N_{\text{input}} = M2^K. \tag{9}
\]

The number of possible mappings grows exponentially as \( N_{\text{input}} \) and any exhaustive search is prohibitively expensive even for small \( K \). However, the search is greatly constrained if an MSM constellation point is selected at each step. Suppose at the \( f \)th step, \( f - 1 \) MSM constellation points have been assigned. Let \( d_i, i \in [1, f - 1] \) denote the Euclidean distances between the \( f \)th point and the \( i \)th assigned point. Define \( P_f \) as

\[
P_f = \sum_{d = d_i} Q \left( \frac{d}{2\sigma} \right),
\]

where \( Q(\cdot) \) is the Gaussian tail function. The function \( P_f \) is an upper bound on the a posteriori probability of error given the \( f \)th MSM constellation point was transmitted over the \( f - 1 \) assigned points. At each step, a point is chosen to minimize this upper bound. Using this intuition, the following sub-optimal search algorithm is used to find good mappings for the TCMSM system:

**Algorithm** Search for an MSM constellation point mapping for a given constraint \( U \geq 0 \).

1) Fix the first MSM constellation point.
2) For the \( f \)th MSM constellation point \((f > 1)\), select the point with at least \( U \) zero points which minimizes \( P_f \).
3) If the number of occurrences of current point \( n > 1 \), return to step 2 and pick the point with the next smallest \( P_f \).
4) Fix current point, add 1 to the number of occurrences of this point. If assignment not finished, go to step 2.

**End.**

The search algorithm uses all possible starting points in step 1. At high rates, the number of all possible starting points is too large and 100 random starting points are selected. Note that the same search algorithm is implemented over both constrained constellations \((U > 0)\) and unconstrained constellations \((U = 0)\), despite the fact that the constrained constellation is a subset of the unconstrained constellation. In fact, the search over the constrained constellation often returned codes with a smaller average optical power requirement than search results over the more general unconstrained constellation. This is due to the sub-optimal search algorithm which improves the distance properties of the code stepwise rather than globally. Additionally, for each system the search is run using QPSK, 5-APSK and 9-APSK constellations for each non-zero in-band carrier. These constellations were chosen for their simplicity of implementation. However, as the number of phases increases, more freedom is allowed at the cost of increased time consumed by the search algorithm. The search terminates with a set of codes and the code with minimum average optical power at each bandwidth efficiency is chosen. The performance of TCMSM systems is shown in Sect. V.

### B. DC Level Detection

The symbol-by-symbol bias used in MSM wireless optical systems is clearly correlated to the transmitted signal and provides additional information about the transmitted sequence. This correlated information can be exploited to provide a degree of signal space diversity, which improves the detection performance of the receiver [13]. The additional complexity required by DC detection is that joint detection over all in-band carriers and DC bin must be implemented as opposed to independent detection for each in-band carrier in normal MSM systems.

The DC bias required to ensure an MSM constellation point satisfies the non-negativity constraint can be used to increase the distance between MSM constellation points. The DC level of each MSM constellation point is computed and MSM constellation points including their DC distance are used in the search algorithm described in Sec. III-A.2. The difference in DC biases can be used to increase the distance between MSM constellation points and the search algorithms may return codes with larger \( d_{\text{dc}} \). However, due to the sub-optimal nature of the search, the inclusion of DC distance does not always return codes with larger \( d_{\text{dc}} \). In such cases, detection using the DC bias is employed solely at the receiver. The performance of this DC detection technique applied to TCMSM systems is shown in Section V.
It should be noted that this paper assumes an additive white Gaussian noise process. However, fluorescent light noise introduces larger noise power at low frequencies near DC [22]. In this case, the performance of the DC detection technique will be degraded and, in the worst case, no improvement in detection will be achieved, assuming the receiver has knowledge of the noise distribution and spectral density. Therefore, it is desirable to use the DC detection technique when the additional complexity at the receiver can be tolerated.

IV. DESIGN OF OUT-OF-BAND CARRIER SIGNALS

The unregulated bandwidth in wireless optical channels can be exploited by adding carriers at out-of-band frequencies, as shown in Fig. 1, and optimizing their amplitudes to reduce the average optical power at the transmitter. Notice that using out-of-band carriers does not incur a penalty in spectral efficiency, as is the case with previous approaches, since these frequencies are unregulated and are contained to a given room.

Out-of-band carrier symbol design need only be performed once for each MSM constellation point chosen and stored at the transmitter. The receiver has no knowledge of the out-of-band carriers and discards them before detection. Thus, the detector using out-of-band carriers is the same as a conventional MSM wireless optical receiver and each carrier can be decoded independently, yielding significant complexity savings. Therefore, this technique shifts the complexity from the receiver to the transmitter, where generation of out-of-band carriers and a lookup table to store the out-of-band carrier amplitudes are required.

Note that the out-of-band carrier design technique cannot be applied to RF systems since, unlike wireless optical systems, strict spectral masks are defined in these channels to limit interference with other communication schemes.

A. Optimization of Amplitudes Over the Complex Plane

In this section the results in [14], [15] are extended by finding the optimum amplitudes on the out-of-band carriers. For each MSM constellation point, the problem of finding the optimum out-of-band carrier amplitudes to reduce the average optical power can be formulated as a convex optimization problem [23]. For a given MSM constellation point the real-valued symbols \( c_1, \ldots, c_W \) in Fig. 2 are specified. The minimum amplitude of \( s(t) \) (6), or equivalently the required bias to ensure non-negativity, is estimated by discretizing it into samples \( \{s_0, s_2, \ldots, s_{A-1}\} \) where \( A \) is chosen so that the error in approximation is small. The optimization problem can then be formulated as follows,

\[
\text{maximize} \quad s_{\text{min}} \\
\text{subject to} \quad s_{\text{min}} \leq s_\alpha, \quad \alpha = 0, 2, \ldots, A - 1 \\
d_i = c_i, \quad i \in [1, W] \\
d_i \in \mathbb{C}, \quad i \in [W + 1, W + L] \\
s_\alpha = \text{Re} \left[ \sum_{i=1}^{W+L} d_i \exp(-j2\pi i\alpha/A) \right].
\]

This problem can be easily cast as a linear program and can be solved effectively using standard optimization algorithms. To bound the approximation error for a given \( A \), an upper bound on the absolute value of the slope for each MSM constellation point is computed. An upper bound as well as a lower bound on the amplitude of the continuous waveform at time periods between any two samples is then computed. In simulations, \( A = 1000 \) which results in an approximation error of less than 0.01 dB.

The design is done once and the amplitudes for each MSM constellation point are stored at the transmitter. If single-precision floating point numbers are used to store the results in a table, 8 bytes are required to store the in-phase and quadrature amplitudes on each out-of-band carrier [24]. Given \( n_{\text{input}} \) possible input symbols, the memory requirement for \( L \) real-amplitude out-of-band carriers is

\[ n_{\text{input}} \times L \times 8 \quad [\text{bytes}]. \]

where \( n_{\text{input}} = 4^W \) for a normal MSM system and \( n_{\text{input}} = M \times 2^K \) as in (9) for a TCMSM system.

Unlike previous work [14], [15], using out-of-band carriers does not reduce the bandwidth efficiency of the system. Moreover, the optimum amplitudes on \( L \) out-of-band carriers are found by solving the optimization problem. Section V presents the performance of this technique applied to the normal MSM system and the TCMSM system using \( L = 1, 2, 3, 4, 10 \) out-of-band carriers.

B. Optimization of Amplitudes Over a Discrete Constellation

A limitation of selecting amplitudes for the out-of-band carriers over a continuous set is the storage requirement at the transmitter. Consider that the \( L \) out-of-band carrier amplitudes are designed over a 9-APSK constellation to alleviate the transmitter memory requirement. For each data symbol, an exhaustive search over an 9-APSK constellation is performed once to find the optimal in-phase and quadrature amplitudes on the \( L \) out-of-band carriers which minimizes the average optical power. The number of possible amplitudes on \( L \) out-of-band carriers is \( 9^L \) for each data symbol, which makes an exhaustive search practical to implement for small \( L \).

To improve the average optical power reduction capability, a real scaling factor \( \alpha \), in the range \( (0, 1] \), is introduced to scale the amplitude transmitted on all out-of-band carriers with respect to the amplitudes of the in-band carriers. Unlike the most general case of individual complex values on each out-of-band carrier considered in Sec. IV-A, \( \alpha \) is fixed for all symbols. To find a good value of \( \alpha \), the range of \( (0, 1] \) is discretized into 20 points. For each \( \alpha \), the exhaustive search was performed to minimize the average optical power. Note that introducing \( \alpha \) increases the size of the lookup table by 4 bytes.

The out-of-band 9-APSK symbols stored at the transmitter require one byte for each MSM constellation point. Thus, the memory requirement for \( L \) 9-APSK out-of-band carriers and the \( \alpha \) scaling factor is

\[ n_{\text{input}} \times L + 4 \quad [\text{bytes}]. \]

The memory required by restricting the out-of-band carriers to a scaled discrete constellation is asymptotically 1/8 of that with no restriction for large \( n_{\text{input}} \). Thus, this technique is
more suitable for applications where the complexity of the transmitter must be limited.

The scaled 9-APSK out-of-band carrier technique proposed in this section differs from [14] in three aspects. Firstly, the carriers are located at the out-of-band region and using them does not affect the bandwidth efficiency of the system. Secondly, a scaling factor $\alpha$ is introduced to further reduce the average optical power. Lastly, the out-of-band carriers are chosen from 9-APSK constellation rather than a QPSK constellation. The performance of $L = 4$ discrete-constellation out-of-band carriers applied to the normal MSM system and the TCMSM system are presented in Sec. V

V. SIMULATION RESULTS

Figures 4(a) and 4(b) plot the normalized average optical power $\rho$ and the normalized peak optical power $\psi$ defined in (8) of in-band techniques versus bandwidth efficiency $\eta$ defined in (7). TCMSM codes are designed for $W = 2, 3, 4, 5$ for which the search algorithm is able to return results within a reasonable amount of time. For a given $W$, the rate of transmission is maximized by increasing $K$ until the search algorithm fails to return codes with average optical power better than the normal system. For each pair of $W$ and $K$, all possible combinations of $U = 0, 1, 2, M = 2, 4$ and choice of QPSK, 5- or 9-APSK constellations are considered, and the best codes are summarized in Tbl. I (for $L = 0$). Notice that the sub-constellation chosen in all cases is 9-APSK, illustrating that a higher degree of freedom in choosing the constellation points improves the performance of trellis codes. In fact, an improvement on the order of 0.5 dB was seen by using 9-APSK over QPSK. Although large constellations may yield further improvement, the search time required to find such codes increases greatly. The DC detection technique is also applied to TCMSM codes. For comparison, a normal MSM system is simulated for $W = 1, \ldots, 7$ and both the minimum-power block coding technique [14] and a recent SSPS code given in [17, Tbl. II] are presented. A popular PAPR reduction technique for electrical channels, tone injection [10], was also simulated in this application and the average and peak optical powers computed. All the systems are designed to have the same bit rate and BER of $10^{-6}$.

In Fig. 4(a) it is evident that bandwidth efficiency is traded for average optical power efficiency for all power reduction schemes, as is conventional in optical signalling design [25]. All proposed systems outperform existing techniques at the same bandwidth efficiency. The in-band techniques operate well at low bandwidth efficiencies. The TCMSM code achieved a reduction up to 0.95 dB optical in average optical power while simultaneously reducing the peak optical power by 0.44 dB optical at $\eta = 1$. Notice that at higher $\eta$, the performance of in-band coding approaches that of the normal system. This is due to the fact that the degrees of freedom in code design were limited at high $\eta$ since the search was carried out over at most $W = 5$ carriers due to the computational cost of the sub-optimal search procedure. It is anticipated that better codes at higher $\eta$ could also be found using the same algorithm with $W > 5$ given additional computing resources. By applying DC detection to a TCMSM system, an additional average optical power reduction of up to 0.50 dB is achieved with a simultaneous peak power reduction of 0.46 dB at $\eta = 0.75$. It should be noted that the peak optical power of both in-band techniques are comparable to normal MSM systems but are not always reduced, as shown in Fig. 4(b). This is because the average optical power is the only criterion in the selection of codes at each bandwidth efficiency. Notice also that the average optical power of the tone injection algorithm, designed for electrical PAPR reduction, is significantly increased over TCMSM at low bandwidth efficiencies. Additionally, the peak requirement of this technique is large and comparable to the minimum power block coding technique at high $\eta$ [14]. Thus, careful design of algorithms tailored to optical intensity MSM systems are required to realize large gains in power efficiency.

Figures 4(a) and 4(b) show the normalized average and peak optical powers for TCM systems at high $\eta$. For each pair of $W$ and $K$, all possible combinations of $U = 0, 1, 2, M = 2, 4$ and choice of QPSK, 5- or 9-APSK constellations are considered, and the best codes are summarized in Tbl. I (for $L = 0$). Notice that the sub-constellation chosen in all cases is 9-APSK, illustrating that a higher degree of freedom in choosing the constellation points improves the performance of trellis codes. In fact, an improvement on the order of 0.5 dB was seen by using 9-APSK over QPSK. Although large constellations may yield further improvement, the search time required to find such codes increases greatly. The DC detection technique is also applied to TCMSM codes. For comparison, a normal MSM system is simulated for $W = 1, \ldots, 7$ and both the minimum-power block coding technique [14] and a recent SSPS code given in [17, Tbl. II] are presented. A popular PAPR reduction technique for electrical channels, tone injection [10], was also simulated in this application and the average and peak optical powers computed. All the systems are designed to have the same bit rate and BER of $10^{-6}$.

In Fig. 4(a) it is evident that bandwidth efficiency is traded for average optical power efficiency for all power reduction schemes, as is conventional in optical signalling design [25]. All proposed systems outperform existing techniques at the same bandwidth efficiency. The in-band techniques operate well at low bandwidth efficiencies. The TCMSM code achieved a reduction up to 0.95 dB optical in average optical power while simultaneously reducing the peak optical power by 0.44 dB optical at $\eta = 1$. Notice that at higher $\eta$, the performance of in-band coding approaches that of the normal system. This is due to the fact that the degrees of freedom in code design were limited at high $\eta$ since the search was carried out over at most $W = 5$ carriers due to the computational cost of the sub-optimal search procedure. It is anticipated that better codes at higher $\eta$ could also be found using the same algorithm with $W > 5$ given additional computing resources. By applying DC detection to a TCMSM system, an additional average optical power reduction of up to 0.50 dB is achieved with a simultaneous peak power reduction of 0.46 dB at $\eta = 0.75$. It should be noted that the peak optical power of both in-band techniques are comparable to normal MSM systems but are not always reduced, as shown in Fig. 4(b). This is because the average optical power is the only criterion in the selection of codes at each bandwidth efficiency. Notice also that the average optical power of the tone injection algorithm, designed for electrical PAPR reduction, is significantly increased over TCMSM at low bandwidth efficiencies. Additionally, the peak requirement of this technique is large and comparable to the minimum power block coding technique at high $\eta$ [14]. Thus, careful design of algorithms tailored to optical intensity MSM systems are required to realize large gains in power efficiency.

The performance of $L = 4$ out-of-band carriers applied to normal MSM systems is shown in Figs. 5(a) and 5(b). The average optical power minimizing amplitudes for the out-of-band carriers were found using the Sedumi 1.05 optimization toolbox [26] to solve the convex optimization problem for each symbol. Out-of-band signals chosen over the complex
plane achieved the best reduction in average optical power at the expense of greatly increased peak amplitude. Adding $L = 4$ complex-valued out-of-band carriers achieved an average optical power gain as high as 2.56 dB, while an average optical power gain of up to 1.50 dB is achieved using scaled 9-APSK. The peak optical power is increased by as much as 3.96 dB optical using $L = 4$ complex-valued out-of-band carriers, while selecting amplitudes over 9-APSK increases the peak optical power as much as 2.65 dB. Note, however, that selecting carriers over the scaled 9-APSK constellation requires 1/8 of the memory required by complex-valued out-of-band carriers. Additionally, notice that adding out-of-band carriers does not impact the detection process of normal MSM, i.e., independent detection over each in-band carrier is still possible.

Figures 5(a) and 5(b) also present the performance of $L = 4$ out-of-band carriers and in-band coding with parameters shown in Table I (for $L > 0$). The Sedumi toolbox was used once again to select out-of-band carrier amplitudes over the complex plane. When compared to a normal MSM system at the same $\eta$, combining out-of-band carriers with both in-band techniques achieved the highest gain in average optical power. An average optical power reduction of 2.63 dB optical is achieved at $\eta = 1$. Although the peak optical power is increased, this increase is smaller than that of using the out-of-band carriers without in-band trellis codes or DC detection. This is because both in-band trellis codes and the DC detection technique help to reduce the peak optical power of the selected MSM constellation points. Selecting out-of-band amplitudes from a discrete scaled 9-APSK set provides greater savings in average optical power for the same memory requirement. In our simulations, the additional memory varies between 20 to 65540 bytes for $L = 4$ scaled 9-APSK out-of-band carriers depending on $\eta$, as shown in Table I. Notice, however, that the use of in-band coding techniques does not permit independent detection of each symbol or carrier. Thus, such schemes, although providing greater average optical power efficiency, require greater complexity.

In applications where the transmitter complexity is strictly limited and little additional memory is available for the look-up table, the out-of-band techniques proposed in previous sections may become infeasible to employ. To address this problem, out-of-band carriers are applied to a subset of MSM constellation points with the largest average optical power cost. This trades average optical power reduction capabilities for a reduced memory requirement. This technique is applied to 25%, 50% and 75% of the MSM constellation points with the highest average optical power cost. It is observed that the percentage of loss in average optical power reduction and the percentage of savings in memory requirements are roughly the same in all cases. The application of this technique to other in-band systems to trade-off complexity and average optical power savings is straightforward.

Figures 6(a) and 6(b) show the average and peak optical power of a normal MSM system applied with $L = 1, 2, 3, 4, 10$ out-of-band carriers whose amplitudes are optimized over the complex plane. Notice that as $L$ increases a larger reduction in average optical power is available at every $\eta$ at the cost of greatly increased peak optical power. Thus, for a given $W$ and $K$, $L$ should be maximized to reduce $\rho$ so long as the system is within the peak optical budget. Notice also that the incremental gain in $\rho$ is small for $L > 4$ while $\psi$ increases.

**TABLE I**

<table>
<thead>
<tr>
<th>$\eta$</th>
<th>0.75</th>
<th>1.00</th>
<th>1.20</th>
<th>1.25</th>
<th>1.40</th>
</tr>
</thead>
<tbody>
<tr>
<td>$W$</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>$K$</td>
<td>3</td>
<td>5</td>
<td>6</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>$U$</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$M$</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sub-constellation</th>
<th>9-APSK</th>
<th>9-APSK</th>
<th>9-APSK</th>
<th>9-APSK</th>
<th>9-APSK</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\psi$ [dB]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$L = 0$</td>
<td>0.52</td>
<td>0.68</td>
<td>1.00</td>
<td>1.34</td>
<td>2.20</td>
</tr>
<tr>
<td>$L = 4$</td>
<td>-0.70</td>
<td>-0.22</td>
<td>-0.22</td>
<td>-0.07</td>
<td>0.91</td>
</tr>
<tr>
<td>$L = 4$</td>
<td>-1.50</td>
<td>-1.00</td>
<td>-0.97</td>
<td>-0.74</td>
<td>0.17</td>
</tr>
<tr>
<td>Memory [KByte]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$L = 4$</td>
<td>0.1</td>
<td>0.5</td>
<td>1</td>
<td>0.5</td>
<td>2</td>
</tr>
<tr>
<td>$L = 4$</td>
<td>1</td>
<td>4</td>
<td>9</td>
<td>4</td>
<td>16</td>
</tr>
</tbody>
</table>
greatly. Thus, the trade-off between the choice of \(L\) and the resulting \(\psi\) and \(\rho\) must be carefully balanced in any system design.

To compare and quantify the performance of the proposed techniques, a normal MSM system operating at \(\eta = 1\) is taken as a baseline. The gain in average optical power for each technique in Fig. 5(a) is shown in Fig. 7. Although in-band TCMSM coding provides some gain, notice large gains in average optical power are achieved by applying out-of-band carriers to normal MSM systems, at a cost of increased peak optical power. A large peak optical power can lead to nonlinear distortion at the transmitter and thus degrade the BER performance of the system. Additionally, safety limits require constraints on both the average and peak optical power of wireless optical systems, although average optical power limits dominate in nearly all cases. In general, applying out-of-band carriers is desirable when the peak optical power budget is high and in the case of a tight constraint on the peak optical power, in-band coding techniques can also be used to provide an average optical power gain for lower peak optical power. The use of the DC detection technique is also beneficial, however, less so than the previous techniques. Finally, applying all of the techniques gives the best performance at the price of additional complexity. Although the particular values of \(\rho\) will change with \(\eta\), as is evident in Fig. 5(a), the above trends hold. Notice also that the gains achieved by each of the proposed techniques are not additive, i.e., the resulting gain of applying a number of techniques is less than their sum. This suggests that jointly designing in-band systems and out-of-band carriers may yield an additional reduction in the average optical power.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, the immense bandwidth available for wireless optical MSM channels is exploited to reduce the average optical power requirement. Out-of-band carriers are added at higher frequencies and their amplitudes optimized to reduce the average optical power requirement. Gains as high as 2.56 dB over conventional MSM systems are realized using four out-of-band carriers. Coding over in-band carriers realizes a gain in average optical power of 0.95 dB over uncoded systems without affecting the peak optical power at \(\eta = 1\). Finally, the DC bias level is exploited at the receiver to further improve detection. Employing all the proposed techniques yields an average optical power gain of 2.63 dB at \(\eta = 1\), which corresponds to a 5.26 dB electrical power gain.

In terms of complexity, applying out-of-band carriers to normal MSM systems has the least complexity requirement among all proposed techniques. The receiver is not changed and independent symbol-by-symbol detection over each in-band carrier is maintained. However, additional complexity is required at the transmitter in the form of a lookup table. In contrast, in-band trellis coding techniques require the use of a trellis encoder at the transmitter and a sequence detector at the receiver, while the DC detection technique requires joint detection of all in-band carriers.

The work presented here serves as an introduction to the use of out-of-band carriers and coding over sequences of MSM points to reduce peak and average optical powers of indoor diffuse wireless optical channels. Additional work is required to generalize these results to a wider class of channels which are non-flat, have colored noise and are corrupted by
fluorescent light interference. Joint design of these techniques is also a promising avenue of research which should be pursued to maximize the available gains. Finally, the decoding complexity of in-band schemes must be addressed and sub-optimal decoders should also be explored.

REFERENCES

A Comparison of Frequency-Division Systems to Code-Division Systems in Overloaded Channels

Joon Ho Cho, Member, IEEE, Qu Zhang, and Lixin Gao, Member, IEEE

Abstract—In this paper, a frequency-division counterpart of joint power control and sequence design problem for code-division multiple-access (CDMA) systems is solved. Total transmit and receive power minimizations are considered for frequency-division multiplexing (FDM) and frequency-division multiple-access (FDMA) communications over overloaded channels. After the definition of channel overloading for CDMA systems is extended to the frequency-division systems, the user admissibility is characterized by a necessary and sufficient condition for the existence of the optimal solution under unequal signal-to-interference-plus-noise ratio constraints at the output of linear receivers and asymmetric data transmission rate constraints among users. The optimal signal power, bandwidth, transmit waveform, and receive waveform are derived for each user as the decision parameters of the optimization problem. It is shown that, if this solution is applied for the uplink users to minimize the total receive power, the optimal FDMA system performs the same as the optimal CDMA system. It is also shown that, if this solution is applied for the downlink users to minimize the total transmit power, the optimal FDM system always outperforms the code-division system that minimizes the extended total squared correlation. Numerical results suggest that the optimal FDM system and the optimal downlink code-division system achieve the same performance when the total transmit power is minimized.

Index Terms—Code-division multiple-access (CDMA), frequency-division multiplexing (FDM), frequency-division multiple-access (FDMA), power control.

I. INTRODUCTION

RECENTLY, sequence optimization problems have been solved for single-cell code-division multiple-access (CDMA) communications [1]–[6]. In particular, the interest has been in overloaded systems, where the length of code sequences is outnumbered by the active users in the channel. In [1]–[3], the sum capacity is adopted as the performance metric of the overloaded CDMA systems. The optimal signature sequences are derived for equal-power, symbol-synchronous users [1], for unequal-power, symbol-synchronous users [2], and for unequal-power, chip-synchronous users [3]. In equal-power cases, the Welch bound equality (WBE) sequences [7] are identified as the optimal signature sequences. In unequal-power cases, the orthogonal sequences are identified as the optimal sequences for oversized users, while the generalized WBE (GWBE) sequences [2] and the generalized asynchronous WBE sequences [3] are identified as the optimal sequences for non-oversized users.1

Joint power control and sequence design problems also have been solved [4]–[6] for overloaded CDMA systems, with the total signal power as the performance metric and the signal-to-interference-plus-noise ratio (SINR) at the output of linear receivers as the constraints. In [4], it is shown that, for symbol-synchronous uplink CDMA systems, the WBE sequences are the optimal signature sequences that minimize the total receive power under equal SINR constraints among users. It is also shown that, under unequal SINR constraints, the orthogonal sequences are optimal for oversized users and the GWBE sequences are optimal for non-oversized users. A generalization to symbol-asynchronous but chip-synchronous uplink systems is reported in [5], where it is shown that the user capacity of a symbol-synchronous CDMA system is the same as that of the chip-synchronous CDMA system employing matched filters, when a system-wide quantity called the total squared asynchronous correlation is minimized. In [6], a downlink symbol-synchronous code-division system is considered to minimize the total transmit power under unequal SINR constraints. In contrast to the minimization of the total receive power, this problem has an additional complexity in finding the solution because the channel gains of users must be taken into account. With the full characterization of the optimal solution being left as an open problem, a suboptimal method for the power control and sequence allocation is proposed, which utilizes the GWBE sequences as signature sequences. It is shown that this suboptimal method minimizes a system-wide quantity called the extended total squared correlation (ETSC). A Lagrangian-based search algorithm is also proposed to compare the ETSC-minimizing downlink system to the numerically-found optimal code-division system that minimizes the total transmit power.

Unlike CDMA, frequency-division multiple-access (FDMA) has weaknesses in that no statistical multiplexing can be utilized and in that the bandwidth must be re-allocated whenever a new user enters the system. However, FDMA has a strength in that it can easily accommodate users with different target transmission rates. Moreover, FDMA is as optimal as CDMA under certain optimality criteria. For

1For the definitions of oversized and non-oversized users, see [2] and [3].
example, FDMA can achieve the same sum capacity as CDMA [8]. It is also recently shown that, in a single cell multiple-access system with a fixed number of overloading users, the optimal FDMA system performs the same as the optimal CDMA system when \( \text{equal-power users} \). The optimal FDMA system performs the same as a multiple-access system with a fixed number of overloading users. For example, FDMA can achieve the same sum capacity as CDMA [8].

More specifically, it is shown that the transmit waveforms of the optimal multiple-access systems are the continuous-time equivalents of WBE sequences and that the optimal FDMA and the optimal CDMA systems have the same amount of interference though they have different allocation of the total interference to multiple-access interference (MAI) and intersymbol interference (ISI).

Motivated by the above observations, a frequency-division counterpart of the power and sequence design problem for code-division systems is tackled in this paper. In particular, the optimal power and bandwidth allotment and the optimal transmit and receive waveforms are derived that jointly minimize the total signal power subject to the quality of service (QoS) requirements defined in terms of the data symbol transmission rates and the SINRs at the output of linear receivers. The channels are assumed frequency-flat.

After the definition of channel overloading for CDMA systems is extended to the frequency-division systems, the user admissibility is characterized by a necessary and sufficient condition for the existence of the optimal solution. We provide a general solution where the SINRs and the symbol rates can be different among users and each user signal can be transmitted through multiple carriers using rate splitting. Our solution is applicable to both the total transmit power minimization and the total receive power minimization. Comparisons are made to the uplink CDMA system derived in [4] and to the downlink code-division system proposed in [6]. A comparison is also made to the downlink code-division system that is obtained through the Lagrangian-based global search technique devised in [6].

The organization of this paper is as follows. In Section II, the signal and system model is described, the optimization problem is formulated, and the notion of channel overloading is extended to frequency-division systems. In Section III, a review on the optimal transmit and receive waveforms for single-user system is provided in the context of converting the optimization problem to an equivalent one. In Section IV, the necessary and sufficient condition for the user admissibility is presented and the optimal solution is derived. Comparisons are also made to code-division systems. Numerical results and discussions are provided in Section V, and concluding remarks are offered in Section VI.

II. SIGNAL AND SYSTEM MODEL

We consider FDMA systems first. Suppose that there are \( K \) users in the system. The \( k \)th user’s transmitted bandpass signal is modeled as

\[
R\left\{ \sqrt{2Q_k s_k(t)} e^{j(2\pi f_k t + \theta_k)} \right\},
\]

for \( k = 1, 2, ..., K \), where \( Q_k \) is the transmit power, \( f_k \) is the carrier frequency, and \( \theta_k \) is the carrier phase of the \( k \)th bandpass signal and \( \Re\{\cdot\} \) denotes the real part. The linearly modulated complex baseband signal \( s_k(t) \) is given by

\[
s_k(t) = \sum_{m = -\infty}^{\infty} d_k[m] g_k(t - mT_k),
\]

where \( \{d_k[m]\}_m \) is the data sequence, \( g_k(t) \) is the transmit waveform, and \( 1/T_k \) is the symbol transmission rate of the \( k \)th signal. The data symbols are modeled as zero-mean, proper-complex, independent and identically distributed (i.i.d.) random variables with unit variance.\(^2\)

The transmit waveform \( g_k(t) \) is assumed to be a complex-valued strictly band-limited waveform with bandwidth \( W_{k}/2 \) in baseband and energy \( T_k \). Note that, to properly model an FDMA system, the carrier frequencies must be selected to avoid any overlap with other user signals in the frequency domain.

We assume that each transmitted signal passes through a frequency-flat channel with channel gain \( \alpha_k(>0) \) and propagation delay \( \tau_k \), for \( k = 1, 2, ..., K \), and is received in the presence of additive white Gaussian noise. Hence, the total received signal is modeled as

\[
r(t) = \Re\left\{ \sum_{k = 1}^{K} \sqrt{2P_k} s_k(t - \tau_k) e^{j(2\pi f_k t + \phi_k)} \right\} + n(t),
\]

where \( P_k \triangleq \alpha_k Q_k \) and \( \phi_k \) are, respectively, the received signal power and the phase of the \( k \)th user’s signal, and \( n(t) \) is a white Gaussian random process with two-sided power spectral density (PSD) \( N_0/2 \). The signal is demodulated coherently using linear receivers. Hence, the decision statistic \( z_k[m] \) for \( d_k[m] \) is given by

\[
z_k[m] = \int_{-\infty}^{\infty} h_k(t - mT_k - \tau_k)^* \left( 2r(t)e^{-j(2\pi f_k t + \phi_k)} \right) dt,
\]

where \( h_k(t) \) is the \( k \)th receive waveform. We call the waveforms used by the correlators in the receiver side the receive waveforms. Thus, the time-reversed complex-conjugate of the receive waveforms become the impulse response of the linear filters. Since the in-band component is the only relevant signal portion of the received signal, without loss of generality, we assume that \( h_k(t) \) is also a complex-valued strictly band-limited waveform with bandwidth \( W_k/2 \) in baseband. Since \( h_k(t) \) rejects all the out-of-band components, we can rewrite \( z_k[m] \) as

\[
z_k[m] = \int_{-\infty}^{\infty} h_k(t - mT_k - \tau_k)^* \left( \sqrt{2P_k} s_k(t - \tau_k) + n_k(t) \right) dt,
\]

where \( n_k(t) \) is the complex baseband-equivalent of \( n(t) \) seen at the \( k \)th receiver.

Two quality of service (QoS) requirements of interest are the SINR and the transmission rate. The SINR of the \( k \)th signal at the output of the linear receiver is defined as

\[
\text{SINR}_k \triangleq \frac{\mathbb{E}[|E\{z_k[m]d_k[m]\}|^2]}{\text{Var}[z_k[m]|d_k[m]]}. \tag{6}
\]
From the signal model (2), the symbol transmission rate of the kth signal is $1/T_k$. Hence, the constraints are

$$\text{SINR}_k \geq \gamma_k \quad \text{and} \quad \frac{1}{T_k} \geq R_k, \forall k,$$

(7)

for some pre-specified target SINRs $\{\gamma_k\}_{k=1}^K$ and target rates $\{R_k\}_{k=1}^K$.

If we are to find the optimal FDMA system that occupies bandwidth $W$ in passband and minimizes the total transmit power $\sum_{k=1}^K Q_k$, the system optimization problem can be written as

**Problem 1:**

$$\begin{align*}
\text{minimize} & \quad \sum_{k=1}^K P_k \alpha_k \\
\text{subject to} & \quad \int_{-W_k/2}^{W_k/2} |G_k(f)|^2 df = T_k, \quad (8b) \\
& \quad \int_{-W_k/2}^{W_k/2} |H_k(f)|^2 df = 1, \quad (8c) \\
& \quad \text{SINR}_k \geq \gamma_k, \quad (8d) \\
& \quad \frac{1}{T_k} \geq R_k, \forall k, \quad (8e) \\
& \quad \sum_{k=1}^K W_k \leq W, \quad (8f)
\end{align*}$$

where $G_k(f)$ and $H_k(f)$ are the Fourier transforms of $g_k(t)$ and $h_k(t)$, respectively. Using the fact that $Q_k = P_k \alpha_k$, the receive powers $\{P_k\}_{k=1}^K$ are employed as the decision parameters instead of the transmit powers $\{Q_k\}_{k=1}^K$. In (8c), we normalized the energy of $h_k(t)$ just for convenience. If we are to find the optimal FDMA system that minimizes the total receive power, we just need to replace the objective function with $\sum_{k=1}^K P_k$, or set the channel gains to unity with actual channel gains being ignored.

Next, we consider FDM systems. In this case, a single transmitter transmits $K$ signals. Note that the FDMA scheme is for multi-point to point communication such as cellular uplink, while the FDM scheme is for point to multi-point communication such as cellular downlink. Now, $\alpha_k$ is re-defined as the channel gain from the transmitter to the kth user. The kth user demodulates the kth signal by using a linear receiver. Since the received $K$ signals are orthogonal in the frequency domain, it does not matter whether the $K$ signals are symbol-synchronous or asynchronous as far as the linear receivers filter out unwanted signals. The decision statistic is given the same as (5) now with $P_k$ being interpreted as the power of the kth signal at the kth receiver. Thus, the SINR at the kth receiver is given exactly the same as (6). Note that there is effectively no difference between this FDM system and the above FDMA system. Therefore, the optimization problem is also given the same as (8) and the solution to the FDMA problem is directly applicable to the FDM problem.

The optimization in **Problem 1** is conducted under the assumption that the target SINRs $\{\gamma_k\}_{k=1}^K$ and rates $\{R_k\}_{k=1}^K$ as well as the channel gains $\{\alpha_k\}_{k=1}^K$, whose estimates can be obtained, e.g., from pilot signals, are perfectly known. It is also assumed that the allocated optimal power and frequency band information is broadcasted through a control channel.

In terms of the total transmission rate and the total bandwidth, a system can be classified as an overloaded system and a non-overloaded system.

**Definition 1:** An FDMA or an FDM system is overloaded if

$$\sum_{k=1}^K R_k > W. \quad (9)$$

Note that, if the symbol transmission rates are all equal to $1/T$, then this condition reduces to $K > WT$. As shown in [9], the condition $K > WT$ is exactly the channel overloading condition for band-limited CDMA systems when the system bandwidth is not necessarily an integer multiple of the symbol rate. If the system bandwidth is an integer multiple $N$ of the common symbol rate, then the condition (9) further reduces to $K > N$, which is the well-known condition for $K$ symbol-synchronous users with processing gain or signal dimension $N$ to overload the channel.

As seen in the next section, the solution to **Problem 1** varies drastically depending on whether the system is overloaded or not. At this point, it suffices to notice that the left side of (9) is the summation of Nyquist’s minimum bandwidths required by the users for zero ISI. Consequently, at least one user signal must have the bandwidth less than Nyquist’s minimum bandwidth if the system is overloaded, while every user signal can secure Nyquist’s minimum bandwidth if the system is not overloaded.

The system model in this section assumes a single linearly modulated signal per user. After the optimal solution for this system model is derived, we also present a generalization to the systems with multiple linearly modulated signals per user.

### III. Optimal Transmit and Receive Waveforms for Each User

If the minimization in **Problem 1** is converted to an equivalent double minimization

$$\text{minimize} \left\{ (\{T_k, W_k\}_{k=1}^K) \right\}_{k=1}^K \left\{ \text{minimize} \left\{ (\{P_k, g_k(t), h_k(t)\})_{k=1}^K \right\}_{k=1}^K \sum_{k=1}^K P_k / \alpha_k \right\}, \quad (10)$$

then the inner minimization problem is given by

**Problem 2:**

$$\begin{align*}
\text{minimize} & \quad \sum_{k=1}^K P_k / \alpha_k \\
\text{subject to} & \quad \int_{-W_k/2}^{W_k/2} |G_k(f)|^2 df = T_k, \quad (11a) \\
& \quad \int_{-W_k/2}^{W_k/2} |H_k(f)|^2 df = 1, \quad (11b) \\
& \quad \text{SINR}_k \geq \gamma_k, \forall k, \quad (11c)
\end{align*}$$

which can be solved for each $k$. In this section, we provide the solution to these individual minimization problems, assuming $\{W_k\}_{k}$ and $\{1/T_k\}_{k}$ that satisfy (8e) and (8f) are given. We first consider the case with $W_k \geq 1/T_k$. In terms of the excess bandwidth $\beta_k (> -1)$ of the kth user’s signal defined by the relation $W_k T_k = 1 + \beta_k$ [14, p. 560], this is the case with non-negative excess bandwidth. If $W_k \geq 1/T_k$, Nyquist’s
minimum bandwidth requirement [14] for zero ISI is met. Thus, the solution to Problem 2 is given by
\[
P_k = \gamma_k \frac{N_0}{W_k},
\]
(12a)
\[
g_k(t) = \text{any pulse with energy } T_k \text{ and bandwidth } W_k,
\]
satisfying Nyquist’s zero-ISI condition, and (12b)
\[
h_k(t) = \frac{g_k(t)}{||g_k(t)||},
\]
(12c)
where \( || \cdot || \) denotes the 2-norm or the square root of the signal energy, i.e., \( || \cdot || \triangleq \sqrt{\int_{-\infty}^{\infty} | \cdot |^2 dt} \). This solution can be alternately derived by using Theorems 3 and 4 in [15] implies that the \( k \)th user does not need a bandwidth larger than \( 1/T_k \) to achieve the required SINR.\(^3\) This is because 1) the flat spectrum pulse with zero excess bandwidth can achieve the same maximum SINR, spending the same transmit power, that can be achieved by any Nyquist pulses with positive excess bandwidth and 2) the saved bandwidth can be spent by other users.

If \( W_k < 1/T_k \), the ISI is unavoidable and the solution to Problem 2 is given by
\[
P_k = \frac{1}{\gamma_k} + \frac{1}{W_k^2 T_k},
\]
(13a)
\[
g_k(t) = \text{any flat spectrum pulse with energy } T_k \text{ and bandwidth } W_k,
\]
(13b)
\[
h_k(t) = \frac{g_k(t)}{||g_k(t)||},
\]
(13c)
The optimality in (13b) of the flat-spectrum pulses with negative excess bandwidth can be found in [11, Example 4]. It can be alternately proved by using the results in [9] because, if we set \( K = 1 \), the problem becomes the joint transmitter and receiver optimization problem with negative excess bandwidth as considered above and because the minimized mean-squared error (MMSE) and the maximized SINR (MSINR) are monotonically related as MMSE\(_k = 1/(\text{MSINR}_k + 1) \). The result (13a) also comes from this relation and [9, Eq. (29)]. Note that the optimality in (13c) of the matched filters holds only when flat-spectrum transmit pulses are employed. For other waveforms, the matched filter receivers are not optimal due to the ISI.

By examining the solution in (13), we can see that the solution becomes the same as that for \( W_k = 1/T_k \) found in (12) as \( W_k \) tends to \( 1/T_k \) and that the optimality of a flat spectrum pulse also holds. Hence, it turns out that, to solve the optimization problem Problem 1, we only need to use the solution in (13) and consider the cases with \( W_k \leq 1/T_k \), i.e., the cases with non-positive excess bandwidth.

Generally, an increased power can save the bandwidth, and an increased bandwidth can save the power in a communication system. The precise relation among these resources in single-user communications with non-positive excess bandwidth is found from (13a), which gives the receive power \( P_k \) of an optimized system as an explicit function of the target SINR \( \gamma_k \) and the normalized bandwidth \( W_k T_k \). This relation can be rewritten in a conservation formula as
\[
\frac{1}{W_k T_k} + \frac{N_0}{P_k T_k} = \frac{1}{\gamma_k} + 1
\]
(14)
which describes how the normalized bandwidth \( W_k T_k \) \((\leq 1)\) and the symbol energy per noise density \( P_k T_k/N_0 \) \((\leq \gamma_k)\) are traded off to satisfy the SINR requirement \( \gamma_k \). Fig. 1 shows the trade-off curves between these two resources for various target SINRs. For example, to achieve 5 dB target SINR, we may choose 5 dB SNR with the unit normalized bandwidth or 12 dB SNR with the 80% of the unit normalized bandwidth.

Now that the inner minimization problem is completely solved and that the trade-off relation between the power and bandwidth of a single user is clarified, we tackle the outer minimization problem in the next section by trading off the power and bandwidth of multiple users.

IV. OPTIMAL POWER AND BANDWIDTH ALLOCATION

A. Optimal Solution

In this subsection, we find the solution to the outer minimization problem given by Problem 3:
\[
\text{minimize } \{P_k, W_k\}_{k=1}^{K} \sum_{k=1}^{K} \frac{1}{\gamma_k} + 1 - \frac{1}{W_k T_k} \cdot \frac{N_0}{T_k \alpha_k} \text{ subject to } \frac{1}{T_k} \geq R_k, \forall k, \]
(15a)
\[
\sum_{k=1}^{K} W_k \leq W, \text{ and } \frac{0 < W_k \leq \frac{1}{T_k}, \forall k, \}
(15b)
(15c)
where the last constraint (15c) is added because, as shown in the previous section, we just need to consider the cases with non-positive excess bandwidth for every user signal. As our intuition says, the optimal system satisfies the constraints (15a) and (15b) with equality. We fully justify this after solving
the outer minimization problem with equalities in these two constraints. From the relation (14), we see that there is a monotonic one-to-one correspondence between $W_k$ and $P_k$. Thus, the outer optimization problem with equalities in (15a) and (15b) can be re-formulated as

**Problem 4:**

\[
\begin{align*}
\text{minimize} & \quad \sum_{k=1}^{K} \frac{P_k}{\alpha_k} \\
\text{subject to} & \quad \frac{P_k T_k}{N_k} \geq \gamma_k, \quad \forall k, \\
& \quad \sum_{k=1}^{K} \frac{1}{T_k} \cdot \frac{1}{\gamma_k + 1} - \frac{N_k}{\gamma_k T_k} = W.
\end{align*}
\]

Once the optimal solution $\{P_k^*\}_k$ is found, the optimal transmit power $Q_k^*$ can be found from the relation $Q_k^* = \alpha_k P_k^*$.

Firstly, we examine the user admissibility by finding a necessary and sufficient condition for the solution to exist.

**Definition 2:** Users are admissible if there exists a feasible solution to Problem 4.

This is the extension of the notion of user admissibility defined in [4], where only CDMA users with equal transmission rates are considered.

**Theorem 1:** Users are admissible if and only if

\[
\sum_{k=1}^{K} \frac{1}{T_k} \cdot \frac{\gamma_k}{\gamma_k + 1} < W.
\]

**Proof:** We first show the necessity: Since $K$ users are admissible, by definition, there exists $\{P_k\}_k$ such that (16a) and (16b) are satisfied. Hence, each term in the left side of (16b) is lower bounded as

\[
\frac{1}{T_k} \cdot \frac{\gamma_k}{\gamma_k + 1} < \frac{1}{T_k} \cdot \frac{1}{\gamma_k + 1} - \frac{N_k}{\gamma_k T_k},
\]

which proves the necessity. We then show the sufficiency: The continuous function

\[
\sum_{k=1}^{K} \frac{1}{T_k} \cdot \frac{1}{\gamma_k + 1} - \frac{N_k}{\gamma_k T_k}
\]

defined on $P_k \geq \gamma_k N_k/T_k, \forall k$, tends to $\sum_{k=1}^{K} \gamma_k/(T_k(\gamma_k + 1))$ as $P_k \to \infty, \forall k$. In addition, the function evaluated at $P_k = \gamma_k N_k/T_k, \forall k$, becomes $\sum_{k=1}^{K} 1/T_k$, which is greater than $W$ by the assumption of channel overloading. Hence, if the condition (17) is met, the intermediate value theorem guarantees the existence of $\{P_k\}_k$ satisfying (16a) and (16b), which proves the sufficiency. If we adopt the definition

\[
e(\gamma_k) \triangleq \frac{\gamma_k}{\gamma_k + 1}
\]

of the effective bandwidth of the $k$th user [12], then the necessary and sufficient condition (17) can be rewritten as

\[
\sum_{k=1}^{K} \frac{e(\gamma_k)}{T_k} < W.
\]

Note that this condition is significantly different from the condition $\sum_{k=1}^{K} e(\gamma_k) < N = WT$ in [4] for overloaded CDMA systems in that the transmission rates $\{1/T_k\}_k$ now unevenly weigh the effective bandwidths.

Secondly, we find how the users in the system trade off their power and bandwidth among one another to meet the SINR requirements and at the same time to minimize the total transmit power.

**Theorem 2:** Define $W_k(\nu)$ as

\[
W_k(\nu) \triangleq \frac{e(\gamma_k)}{T_k} \left(1 + \frac{1}{\sqrt{\alpha_k \nu - \gamma_k}} + \frac{\alpha_k}{\nu} + \frac{\gamma_k}{\sqrt{\alpha_k \nu - \gamma_k}} \right)
\]

where $[x]^+$ denotes the positive part of $x$, i.e., $[x]^+ \triangleq \max(0, x)$. Then, as the solution to Problem 4, the jointly optimal receive power $P_k^*$ and bandwidth $W_k^*$ are given, respectively, by

\[
P_k^* = \frac{N_0}{T_k} \cdot \frac{1}{\gamma_k + 1} - \frac{1}{W_k(\nu^*)} T_k
\]

and $W_k^* = W_k(\nu^*)$, (23) for $k = 1, 2, ..., K$, where $\nu^*(\geq \min_k \gamma_k/\sqrt{\alpha_k})$ is the unique solution to

\[
\sum_{k=1}^{K} W_k(\nu) = W.
\]

**Proof:** Throughout this proof, we assume that

\[
\frac{\gamma_k}{\sqrt{\alpha_k}} \leq \frac{\gamma_{k-1}}{\sqrt{\alpha_{k-1}}} \leq \ldots \leq \frac{\gamma_1}{\sqrt{\alpha_1}}
\]

just for convenience. Note that the SINR requirement $\gamma_k$ is divided by the square root $\sqrt{\alpha_k}$ of the channel gain, not by the channel gain $\alpha_k$. To proceed, we define $X_k \triangleq P_k T_k/N_k - \gamma_k (\geq 0)$ for $k = 1, 2, ..., K$. Then, the Lagrangian function can be written as $l(X_1, X_2, ..., X_K, \lambda) \triangleq \sum_{k=1}^{K} X_k + \lambda \left(\sum_{k=1}^{K} X_k + \gamma_k - \frac{X_k}{\sqrt{\alpha_k}} - \frac{X_k}{\sqrt{\alpha_k}} - \gamma_k - W\right)$, where $\lambda$ is the Lagrange multiplier. By differentiating with respect to $X_k$, we obtain $X_k = e(\gamma_k) (\sqrt{\alpha_k} \nu - \gamma_k)$, where $\nu \triangleq \lambda \geq 0$. Since $P_k T_k/N_k \geq \gamma_k$, the solution $X_k^*$ must be nonnegative. Thus, we claim

\[
X_k^* = e(\gamma_k) [\sqrt{\alpha_k} \nu^* - \gamma_k]^+
\]

is the solution, where $\nu^*$ is the unique solution to $W = \sum_{k=1}^{K} \frac{e(\gamma_k)}{T_k} (1 + [\sqrt{\alpha_k} \nu^* - \gamma_k] + \gamma_k) = \sum_{k=1}^{K} W_k(\nu^*)$. The existence and the uniqueness of $\nu^*$ are proved in the Appendix. It can be straightforwardly verified that the solution (26) satisfies the Karush-Kuhn-Tucker (KKT) condition [13]. Therefore, the conclusion follows.

It is interesting to observe that the users are ordered for this water-filling-like argument by the ratio $\gamma_k/\sqrt{\alpha_k}$, not by the more intuitive ratio $\gamma_k/\alpha_k$. This is because, as seen in (26), the Lagrangian function differentiated to find the first-order necessary condition returns the variable $X_k^*$ as a piecewise linear function of $\sqrt{\lambda} (= \nu)$, not of $\lambda$.

The use of the equality constraints instead of the inequality constraints (15a) and (15b) in solving the outer minimization problem can be justified as follows: 1) As seen in Fig. 2, $W_k(\nu)$ strictly increases at every $\nu$ as $1/T_k$ increases. Thus, if at least one of $\{1/T_k\}_k$ increases then $\sum_{k=1}^{K} W_k(\nu)$ increases at every $\nu$, though not strictly increases. As seen in Fig. 3, $\sum_{k=1}^{K} W_k(\nu)$ is a strictly monotonically decreasing function

\footnote{Actually, this is a reduced Lagrangian function because the inequality constraint (16a) is not included.}
or not. To be more specific, we let

\[ P^*_k = \frac{\gamma_k N_0}{T_k} \left( 1 + \frac{\sqrt{\alpha_k \nu^*} - \gamma_k}{1 + \gamma_k} \right), \]

for \( k = 1, 2, ..., K \), which is a monotonically increasing function of \( \nu^* \). Hence, an increase in at least one of \( \{1/T_k\}_k \) leads to an increase in the total transmit power \( \sum_{k=1}^{K} P^*_k/\alpha_k \). Therefore, the optimal solution must satisfy the constraint (15a) with equality. 2) Since \( \sum_{k=1}^{K} W_k(\nu) \) is a monotonically decreasing function of \( \nu \), an increase in \( W \) results in a decrease in the solution \( \nu^* \) to (24). Hence, an increase in \( W \) leads to a decrease in the total transmit power \( \sum_{k=1}^{K} P^*_k/\alpha_k \) because \( P^*_k \) is a monotonically increasing function of \( \nu^* \). Therefore, the optimal solution must satisfy the constraint (15b) with equality.

Similar to [4], we introduce the notion of oversized and non-oversized users.

**Definition 3:** A user is oversized if its optimal bandwidth equals Nyquist’s minimum bandwidth for zero ISI, while a user is non-oversized if its optimal bandwidth is less than the minimum bandwidth.

The oversized users do not suffer any interference, while the non-oversized users suffer self-induced ISI, which makes the SINR degrade below \( P_k T_k/N_0 \). Let \( K \) be the index set of all oversized users. Then, by **Definition 3**, the number of oversized users \( |K| \) can be alternately defined as the unique integer that satisfies

\[ \frac{\gamma_{|K|+1}}{\sqrt{\alpha_{|K|}} + 1} < \nu^* \leq \frac{\gamma_{|K|}}{\sqrt{\alpha_{|K|}}}, \]

where (25) is assumed and we set \( \gamma_0/\sqrt{\alpha_0} = \infty \) for \( |K| = 0 \). Using the index set \( K \), we can rewrite \( \sum_{k \in K} W_k(\nu^*) = W \) as

\[ \sum_{k \in K} e(\gamma_k) \left( 1 + \frac{1}{\sqrt{\alpha_k \nu^*} - \gamma_k} \right). \]

Thus, the minimum total transmit power achievable by the frequency-division system is given by

\[ \sum_{k=1}^{K} P^*_k = \sum_{k \in K} \left( \frac{\gamma_k N_0}{\alpha_k T_k} \right) + \frac{\left( \sum_{k \in K} e(\gamma_k) \sqrt{\alpha_k \nu^*} \right)^2 N_0}{W - \sum_{k \in K} \frac{1}{T_k} - \sum_{k \notin K} \frac{e(\gamma_k)}{T_k}}. \]

The expression (29) of \( \nu^* \) in terms of \( K \) turns out to be very useful in Sections IV-C and IV-D, where we compare our results with some previous results in joint transmitter and receiver optimization for symbol-synchronous code-division systems.

### B. Invariance of Optimal Allocation under Rate Splitting

Up to this point, we have assumed that the data sequence of each user is transmitted in serial by using a single linearly modulated signal as (2), i.e., by using a single carrier per user. However, this is not the only way to design an FDM or an FDMA system that meets the QoS requirements. Each data sequence can be split into more than one sub-sequences and transmitted in parallel by using multiple linearly modulated signals separated in frequency, i.e., by using multiple carriers per user. As far as each received sub-sequence satisfies the SINR requirement of its original data sequence and the sum rate of the sub-sequences meets the symbol transmission rate requirement of the original sequence, the frequency-division system is a feasible solution. Thus, it is of interest to check whether this idea of rate splitting can reduce the total signal power or not. To be more specific, we let \( s_{k,l}(t) \) be the \( l \)th linearly modulated signal for the \( k \)th user under rate splitting, given by

\[ s_{k,l}(t) = \sum_{m=-\infty}^{\infty} d_{k,l}[m] p_{k,l}(t - m T_{k,l}), \]

where \( d_{k,l}[m] \) is the data sequence, \( p_{k,l}(t) \) is the symbol waveform, and \( 1/T_{k,l} \) is the symbol transmission rate, all for the \( l \)th sub-sequence of the \( k \)th user. Then, the QoS requirements (7) become

\[ \text{SINR}_{k,l} \geq \gamma_k, \forall l \quad \text{and} \quad \frac{1}{T_{k,l}} \geq R_k \]

for \( k = 1, 2, ..., K \), where \( \text{SINR}_{k,l} \) is the SINR of the \( l \)th sub-sequence of the \( k \)th user.

The following corollary shows that the solution found in the previous subsection under serial transmission assumption
is representative enough as an optimal solution to our system design problem.

**Corollary 1:** The transmit power and the amount of bandwidth optimally allocated to a user is invariant under rate splitting.

**Proof:** Consider any rate splitting that satisfies the QoS requirements (32) with equalities for all $k$. If the $k$th user sequence is split into $L_k$ sub-sequences, we may regard the $L_k$ sub-sequences as $L_k$ data sequences of distinct users having different symbol transmission rates but the same SINR requirements and the same channel gains. According to the procedure in Theorem 2 that leads to the optimal allotment, now we need to find $\nu^*$ that solves

$$\sum_{k=1}^{K} \sum_{l=1}^{L_k} W_{k,l}(\nu) = W,$$

where $W_{k,l}(\nu)$ for the $l$th sub-sequence of the $k$th user is given by

$$W_{k,l}(\nu) \triangleq \frac{e(\gamma_k)}{T_{k,l}} \left(1 + \frac{1}{\sqrt{\alpha_k \nu - \gamma_k}^+} + \gamma_k\right) = \frac{T_k}{T_{k,l}} W_k(\nu).$$

(33)

Note that $\{W_{k,l}(\nu)\}_{l=1}^{L_k}$ differ only in transmission rates $\{1/T_{k,l}\}_{l=1}^{L_k}$ and that $\sum_{l=1}^{L_k} 1/T_{k,l} = 1/T_k$ by the assumption. Thus, we obtain $\sum_{k=1}^{K} \sum_{l=1}^{L_k} W_{k,l}(\nu) = \sum_{k=1}^{K} W_k(\nu)$, where $W_k(\nu)$ is defined in (22). This tells us that the solution $\nu^*$ to (33) is the same as the solution to (24). With this $\nu^*$, it is clear that the total optimal bandwidth $\sum_{l=1}^{L_k} W_{k,l}(\nu^*)$ for the $k$th user equals the bandwidth $W_k(\nu^*)$ found in Theorem 2. It is also clear that the total optimal power $\sum_{l=1}^{L_k} P_{k,l}^*$ for the $k$th user equals the transmit power $P_k^*$ found in (27), where

$$P_{k,l}^* = \frac{\gamma_k N_0}{T_{k,l}} \left(1 + \frac{1}{\sqrt{\alpha_k \nu^* - \gamma_k}^+} + \gamma_k\right) = \frac{T_k}{T_{k,l}} P_k^*.$$

(35)

Therefore, the conclusion follows.

In summary, both the optimal power (35) and the optimal bandwidth (34) with $\nu = \nu^*$ allocated to a sub-sequence are proportional to the symbol transmission rate $1/T_{k,l}$ of the sub-sequence, and this proportionality makes the total power and the total bandwidth invariant under rate splitting. Note that this invariance under rate splitting is also obtained under the assumption of frequency-flat channels.

**C. Comparison to an Uplink CDMA**

In this and the next subsections, we compare the optimal frequency-division systems to CDMA systems having processing gain $N$ and symbol transmission rate $1/T$. These two signaling schemes are different in managing the interference in that the MAI is the only interference component in the CDMA systems, while the ISI is the only interference component in the frequency-division systems. Two comparisons are made, one for uplink or multi-point to point communication, and the other for downlink or point to multi-point communication. Note that these comparisons are in terms of the total signal power when linear receivers are used for uncoded systems.

In [4], a joint sequence optimization and power control is considered for symbol-synchronous uplink CDMA systems. The minimization of the total receive power is attempted in Gaussian noise channels, subject to unequal SINR requirements of users. The following corollary makes a comparison of this CDMA system to the FDMA system that minimizes the total receive power.

**Corollary 2:** The optimal FDMA system performs the same as the optimal CDMA system derived in [4], when the total receive power is minimized in overloaded channels.

**Proof:** By simply setting $T_k = T, \forall k$, we can find the optimal FDMA system that minimizes the total receive power for equal-rate users subject to unequal SINR requirements. To enable the comparison, we assume that the bandwidth-symbol time product $W T$ equals the processing gain $N$ of the CDMA system. We also assume that $\gamma_K \leq \gamma_{K-1} \leq \cdots \leq \gamma_1$ just for convenience. By modifying (28) and (29), we obtain the condition on $\nu^*$ as

$$\gamma_{|K|+1} < \nu^* = \frac{\sum_{k=|K|+1}^{K} e(\gamma_k)}{N - |K| - \sum_{k=|K|+1}^{K} e(\gamma_k)} \leq \gamma_{|K|}.$$

(36)

Since the function $\sum_{k=|K|+1}^{K} W_k(\nu)$ is a strictly monotonically decreasing function of $\nu \geq \gamma_K$, this condition implies $\sum_{k=|K|+1}^{K} W_k(\gamma_{|K|+1}) > N \geq \sum_{k=1}^{K} W_k(\gamma_{|K|+1}) T$. After simple manipulation, we obtain

$$(N - |K|) e(\gamma_{|K|}) \geq \sum_{k=|K|+1}^{K} e(\gamma_k) > (N - |K|) e(\gamma_{|K|+1}) + 1),$$

(37)

which is the same result as [4, Eq. (21)] after 3 dB adjustments of power and SINR for all $k$. These adjustments are required because baseband signaling with real-valued data symbols is considered in [4], while bandpass signaling with complex-valued data symbols is considered in this paper. By substituting $\nu^*$ in (29) to (27), now with $T_k = T$ and $\alpha_k = 1, \forall k$, we can obtain

$$P_k^* = \left\{ \begin{array}{ll}
\frac{\gamma_k N_0}{T} e(\gamma_k) N_0 (\nu^* + 1), & \text{for } k \in \mathcal{K}, \\
\frac{\gamma_k N_0}{T} e(\gamma_k) N_0 (\nu^* + 1), & \text{for } k \notin \mathcal{K},
\end{array} \right.$$  

(38)

which is the same result as [4, Eq. (22)]. Therefore, the conclusion follows.

In [9], the equivalence between the CDMA and the FDMA schemes is established in minimizing the TMSE subject to equal receive power requirements. Since the minimized TMSE is a monotonically decreasing function of the receive power, it also establishes the equivalence between the CDMA and the FDMA in minimizing the total receive power subject to equal SINR requirements. The equivalence of the CDMA and the FDMA with unequal SINR requirements has been just successfully established by Corollary 2.

**D. Comparison to a Downlink Code-Division System**

In [6], a joint sequence optimization and power control is considered for symbol-synchronous downlink code-division systems. The minimization of the total transmit power is attempted in flat slow-fading channels, subject to unequal SINR requirements of users. With finding the optimal solution being left as an open problem, a suboptimal solution is proposed, where orthogonal sequences are assigned to overfaded users and GWBE sequences are assigned to non-overfaded users. It is shown that this sequence allocation minimizes the ETSC.
given the proposed power allotment. It is also shown that the optimal linear receivers are the filters matched to the signature sequences.

The following corollary shows that the optimal frequency-division system always outperforms this downlink code-division system in minimizing the total transmit power. In particular, it is shown that the ETSC-minimizing code-division system performs the same as the FDM system that minimizes \( \sum_{k=1}^{K} P_k/\alpha_k^2 \), which is not the total transmit power \( \sum_{k=1}^{K} P_k/\alpha_k \).

**Corollary 3:** The optimal FDM system that minimizes the total transmit power in overloaded channels always outperforms the downlink code-division system proposed in [6] except the case with \( \alpha_k = 1, \forall k \).

**Proof:** By simply setting \( T_k = T, \forall k \), and replacing \( \alpha_k \) with \( \alpha_k^2, \forall k \), in Problem 4 and Theorem 2, we can find the frequency-division system that minimizes \( \sum_{k=1}^{K} P_k/\alpha_k^2 \) for equal-rate users. If \( \gamma_k/\alpha_k \leq \gamma_{k-1}/\alpha_{k-1} \leq \cdots \leq \gamma_1/\alpha_1 \) is assumed just for convenience, then similar to (28) combined with (29) we obtain

\[
\frac{\gamma|K|+1}{\alpha|K|+1} < \nu^* = \frac{\sum_{k=|K|+1}^{K} \frac{c(\gamma_k)}{\alpha_k}}{N - |K| - \sum_{k=|K|+1}^{K} c(\gamma_k)}/\alpha_k|K|, \quad (39)
\]

which is the same result as [6, Eq. (6)] for ETSC-minimizing downlink code-division system. In addition, (27) becomes

\[
Q_k^* = \begin{cases} 
\frac{\gamma_k N_0}{\alpha_k T}, & \text{for } k \leq |K|, \\
\frac{c(\gamma_k) N_0}{T} \left( \nu^* + \frac{1}{\alpha_k} \right), & \text{for } k > |K|,
\end{cases} \quad (40)
\]

which is the same result as [6, Eq. (7)]. This shows that the downlink code-division system proposed in [6] achieves the same performance as the frequency-division system that minimizes \( \sum_{k=1}^{K} P_k/\alpha_k^2 \). Therefore, the conclusion follows.

In the next section, we provide numerical results and discussions on the performance of the proposed frequency-division systems and that of the code-division systems. Numerical results can also be found that strongly suggest the equivalence between the optimal downlink code-division system, of which characterization is still an open problem, and the optimal downlink FDM, when the optimality criterion is to minimize the total transmit power subject to unequal SINR requirements.

**V. NUMERICAL RESULTS AND DISCUSSIONS**

The first results are to show the performance of the optimal FDM and the uplink CDMA systems that minimize the total receive power, when all the users have the same SINR requirements (and, of course, the same transmission rate requirements). As shown already, the two systems perform the same. Let the SNR and the target SINR of the users be \( PT/N_0 \) (\( \equiv P_k T_k/N_0, \forall k \)) and \( \gamma \) (\( \equiv \gamma_k, \forall k \)), respectively, and define the channel load \( \rho \) as

\[
\rho \triangleq \frac{K}{WT}. \quad (41)
\]

Then, by the conservation formula (14), the relation among these parameters is given by \( \gamma = \left( \rho - 1 + \frac{N_0}{PT} \right)^{-1} \). Fig. 4 shows the SINR versus channel load curves for various SNR conditions, where the upper bound is obtained by letting \( PT/N_0 \) tend to infinity. We can see that the upper bound on the SINR as a function of \( \rho \) rapidly decreases to 10 dB, 7 dB, 3 dB, and 0 dB as \( \rho \) increases to 1.1, 1.2, 1.5, and 2, respectively. Note that any pair \((\rho, \gamma)\) below this upper bound is achievable at the expense of the total receive power.

In addition to this comparison in minimizing the total receive power, comparisons are made in the rest of this section with the code-division systems in minimizing the total transmit power. In downlink, the total transmit power minimization has a clear practical meaning because there is only one transmitter, the base station (BS). Thus, if the total transmit power is minimized in this link, then the power consumption in the BS is directly minimized and, consequently, the adjacent cell interference is also reduced in multi-cell systems. In uplink, the total transmit power minimization again has a practical meaning under reasonable assumptions. For this, consider multiple mobile stations (MSs) moving around. Assuming very fast data transmission relative to the motion of scatterers and MSs, we may model the uplink channels of users as block fading random processes. If we further assume that these processes are ergodic and i.i.d. among the MSs, then the total transmit power of users first averaged over a long enough time-period then divided by the number of users becomes the average transmit power per MS. Since our proposed optimization scheme minimizes the total transmit power in each fading block, in the long run this scheme minimizes the average transmit power per MS.

The next results are to show that the optimal FDM system outperforms the downlink code-division system proposed in [6] in terms of the total transmit power. In Fig. 5, the approximate complementary cumulative distribution function (CCDF) of the total transmit power gain of the optimal FDM system over the code-division system is plotted for \( K = 7, 8, \) and \( 9 \), all along with \( N = 7 \). For each case, we conducted \( 10^4 \) independent realizations of the channel gains \( \{\alpha_k\}_k \), which are i.i.d. exponential random variables, to model a Rayleigh fading channel. The target SINRs are indicated in the figure. For the non-overloaded case with \( K = 7 \), the channel gain
(d) is the power ratio with the convergence error in sub-figure (c) being the same as that of the single-user system. However, in the FDM and the code-division systems induce neither multi-

is 0 dB always. This is because the user signals both in the FDM and the code-division systems induce neither multi-user interference nor ISI, so that each user’s performance is the same as that of the single-user system. However, in the overloaded cases with $K = 8$ and $K = 9$, Fig. 5 shows that the FDM system always outperforms the code-division system proposed in [6], which confirms the conclusion in Corollary 3.

The next results strongly suggest that the optimal downlink code-division system and the optimal downlink FDM system are equivalent in minimizing the total transmit power. To find the optimal downlink code-division systems, we conduct an improved version of the Lagrangian global search used in [6]. Instead of updating a spreading sequence and the associated Lagrange multiplier at the same time, we let the sequence converge with the multiplier being fixed, then we update the multiplier, which makes a single iteration. The iterations are repeated until the total transmit power does not significantly change. Fig. 6 shows the ratio of the total transmit power of the numerically-found approximately optimal downlink code-division system to that of the optimal FDM system. The simulation parameters are the spreading gain $N = 7$ and the target SINRs (a) $[\gamma_1, \gamma_2, ..., \gamma_7] = [9, 9, 7, 7, 7, 4, 4]$ dB, (b) $[\gamma_1, \gamma_2, ..., \gamma_8] = [9, 7, 7, 4, 4, 4, 4, 4]$ dB, and (c) $[\gamma_1, \gamma_2, ..., \gamma_9] = [9, 9, 7, 7, 4, 4, 4, 2, 2]$ dB, which are the same as Fig. 5. The channel gains $\{\alpha_k\}$ are i.i.d. exponential random variables with mean $10$ dB, again to model a Rayleigh fading channel.

In the non-overloaded case (a) and the lightly overloaded case (b), the code-division system found from the global search and the optimal FDM system performs almost the same. However, in the heavily overloaded case (c), it appears that the code-division system much outperforms the optimal FDM. As shown in [13, Ch. 22], such a penalty method like the Lagrangian global search returns in general infeasible solutions having less objective function values than that of the optimal solution. Thus, we compensated the errors in (c) to obtain the result (d) by re-computing the total transmit power of the optimal FDM system with the target SINRs being replaced by the resultant SINRs from the global search. Due to the convergence error, the code-division system now appears to slightly underperform the FDM system after this compensation. However, in these cases, it is found that the target SINRs are not perfectly met by the solution obtained from the algorithm. In summary, these numerical results show that the optimal downlink code-division system consumes almost the same power as the optimal FDM system within the margin of convergence error, which strongly supports the conjecture on the equivalence of the optimal FDM and the optimal downlink code-division systems in minimizing the total transmit power.

One might be also interested in comparing the optimal FDMA system to the optimal uplink CDMA system that minimizes the total transmit power. However, for this comparison, the research must be preceded on the full characterization of the optimal uplink CDMA system with unequal channel gains and asymmetric SINR requirements among users, which is still an open problem. When the research on this open problem is performed in the future, our derivation in this paper of the optimal FDMA system that minimizes the uplink total transmit power will serve as a useful reference.

VI. CONCLUSIONS

In this paper, we have derived the optimal frequency-division system that minimizes the total signal power subject to the QoS requirements defined in terms of the transmission rates and the SINRs at the output of linear receivers. The user admissibility is characterized by a necessary and sufficient condition for the existence of the optimal solution. The optimal power, bandwidth, transmit waveform, receive waveform are found for each user and the solution is shown to be invariant under any rate splitting. A resource conservation formula for a single user with non-positive excess bandwidth is also derived that describes how the normalized bandwidth and the SNR are traded off to meet the SINR requirement. It is shown that the optimal FDMA system achieves exactly the same
performance as the optimal CDMA system does in minimizing the total receive power. It is also shown that the optimal FDM system always consumes a lower total transmit power than the ETSC-minimizing downlink code-division system. Numerical results suggest that the optimal FDM system and the optimal downlink code-division system achieve the same performance.

APPENDIX

The existence and the uniqueness of the solution \( \nu^* \) in Theorem 2 can be proved as follows: As shown in Fig. 2, \( W_k(\nu) \) is a constant \( 1/T_k \) for \( 0 \leq \nu \leq \gamma_k / \sqrt{\alpha_k} \), strictly monotonically decreases for \( \nu \geq \gamma_k / \sqrt{\alpha_k} \), converges to \( e(\gamma_k) / T_k \) as \( \nu \to \infty \). Thus, as shown in Fig. 3, \( \sum_{k=1}^{K} W_k(\nu) \) is a constant \( \sum_{k=1}^{K} 1/T_k \) for \( 0 \leq \nu \leq \gamma_k / \sqrt{\alpha_k} \), strictly monotonically decreases for \( \nu \geq \gamma_k / \sqrt{\alpha_k} \), converges to \( \sum_{k=1}^{K} e(\gamma_k) / T_k \) as \( \nu \to \infty \). In addition, we have \( W < \sum_{k=1}^{K} 1/T_k \) by the assumption of channel overloading and \( \sum_{k=1}^{K} e(\gamma_k) / T_k < W \) by the existence of the solution. Therefore, due to the continuity and the monotonicity of \( \sum_{k=1}^{K} W_k(\nu) \), the intermediate value theorem guarantees the existence of a unique \( \nu = \nu^* \in [\gamma_k / \sqrt{\alpha_K}, \infty) \) satisfying \( \sum_{k=1}^{K} W_k(\nu^*) = W \). Note that this unique solution can be found by simply performing a line search starting from \( \gamma_k / \sqrt{\alpha_K} \).

REFERENCES


Joon Ho Cho (S’99–M’01) received the B.S. degree (Summa Cum Laude) in electrical engineering from Seoul National University, Seoul, Korea, in 1995 and the M.S.E.E. and Ph.D. degrees in electrical and computer engineering from Purdue University, West Lafayette, IN, in 1997 and 2001, respectively. From 2001 to 2004, he was with the University of Massachusetts at Amherst as an Assistant Professor. Since July 2004, he has been with Pohang University of Science and Technology (POSTECH), Pohang, Korea, where he is presently an Assistant Professor in the Department of Electronic and Electrical Engineering. His research interests include wideband systems, multuser communications, adaptive signal processing, packet radio networks, and information theory. Dr. Cho was an Associate Editor for the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY.

Qu Zhang received the B.S. and M.S. degrees from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 1998 and 2001, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Massachusetts (UMASS), Amherst, in 2006. From 2002 to 2006, he worked in the Wireless Systems Laboratory, UMass at Amherst, as a Research Assistant. Currently, he is with the MathWorks, Inc., Natick, MA. His research interests include wireless systems, digital communications, and signal processing.

Lixin Gao (S’93–M’96) is a professor of Electrical and Computer Engineering at the University of Massachusetts, Amherst. She received her Ph.D. degree in computer science from University of Massachusetts at Amherst in 1996. Her research interests include Internet routing, network security, network virtualization, and multimedia networking. Between May 1999 and January 2000, she was a visiting researcher at AT&T Research Labs and DIMACS. She is an Alfred P. Sloan Fellow, and received an NSF CAREER Award in 1999 and a Picker Fellowship in 1998. Dr. Gao has served on numerous technical program committees/organizing committees of prestigious conferences such as ACM SIGCOMM, ACM SIGMETRICS, and IEEE INFOCOM, and is on the Editorial Board of IEEE TRANSACTIONS ON NETWORKING.
Serial Search Code Acquisition Using Smart Antennas with Single Correlator or Matched Filter

Henri Puska, Student Member, IEEE, Harri Saarnisaari, Senior Member, IEEE, Jari Iinatti, Senior Member, IEEE, and Pekka Lilja, Student Member, IEEE

Abstract—This paper investigates three code acquisition methods for direct sequence spread spectrum systems (DS/SS) utilizing smart antennas. The methods are suitable especially for receivers that consist of a smart antenna followed by a single correlator or a matched filter (MF). The first method is the known fixed beam strategy where the whole angular uncertainty region is divided into small cells using conventional beamforming techniques. Therein, the receiver searches through all angular and delay cells via a serial search procedure. In the second method, the fixed division is made using advanced beamforming techniques, which provide improved tolerance against interference. In the third strategy, the direction-of-arrival (DOA) of incoming signals are estimated and utilized in the acquisition process. An advantage of this DOA estimation-based strategy, when compared to fixed beamforming methods, is a decrease in the region of uncertainty. Disadvantages are increased computational complexity and signal-to-noise ratio (SNR) limitations. The acquisition strategies are compared when the serial search acquisition of the code phase is made using either the correlator or the matched filter. The results indicate that a single antenna receiver gives the best acquisition performance when SNR is high. However, single antenna methods are sensitive to interference. On the other hand, DOA estimation-based methods offer shorter mean acquisition times than fixed beam methods, especially when the number of arriving signals is small.

Index Terms—Antenna arrays, array signal processing, synchronization, pseudonoise communication.

I. INTRODUCTION

All direct sequence spread spectrum (DS/SS) receivers require synchronization between the incoming and a locally generated pseudonoise (PN) sequence. If the receiver fails in code synchronization, data detection and other higher layer operations are not possible. Code synchronization is typically done in two steps; code acquisition and code tracking. The acquisition can be implemented in two ways. During active acquisition the received signal is correlated with a locally generated replica of the code and the synchronization decision is based on the correlation value. In passive acquisition, a filter matched to the spreading code is used. After acquisition, or initial synchronization, the phase of the spreading code is known within a fraction of a chip and the tracking process is started. The code tracking achieves fine synchronization and maintains it [1].

Antenna arrays for communications purposes have been investigated extensively in recent literature. An overview of different direction-of-arrival (DOA) estimation and beamforming (BF) algorithms can be found, e.g., from [2], [3]. Antenna arrays have also been studied in synchronization. A serial search concept in the delay domain is extended also to include the angular domain in [4]. The result is a two dimensional (2-D) serial search strategy. The receiver in [4] generates orthogonal fixed beams using a Butler matrix and evaluates these beams using a serial search technique that results in joint delay and DOA estimates. However, a simple beamformer like the conventional BF [2] offers only limited tolerance against interference. As a consequence, more advanced BF, like the Capon [2], must be used to form fixed beams if strong interference is a problem. A drawback is that an advanced BF requires more computation than conventional BF. Acquisition schemes in which code matched filters (MF) follow each antenna are introduced in [5]. The outputs of MFs are added noncoherently, and the result is compared to a threshold. The advantage is that neither DOA estimation nor BF is needed in synchronization. However, interference mitigation capability is not attained. In [6] and [7] the antennas are divided into groups, and the signals are combined noncoherently afterwards. Acquisition procedures where a correlator is used in each antenna are studied in [8], [9] and [10]. The outputs of the correlators are combined using BF. The weights of the beamformer are obtained via the least mean squares (LMS) adaptation using a pilot channel, i.e., the DOA information is not needed.

In this paper, a fixed beam acquisition strategy [4] is extended to use MF instead of a correlator, since MF acquisition is known to provide shorter mean acquisition times. However, the main objective is to study a method in which DOA estimation and BF are exploited before the MF serial code acquisition process. Therefore, only a single MF or correlator is needed in the proposed scheme. The mean acquisition time will be used as a measure of performance of the acquisition process. A single antenna approach and fixed beam strategies without DOA estimation are used as reference cases.

Paper approved by V. Aalo, the Editor for Diversity and Fading Channel Theory of the IEEE Communication Society. Manuscript received January 27, 2005; revised May 19, 2006 and March 28, 2007. This research was supported by the Finnish Funding Agency for Technology and Innovation (Tekes), Nokia, Elektrobit, and Finnish Defence Forces. The work of H. Puska was supported by the Graduate School in Electronics, Telecommunications and Automation (GETA), and the following Foundations: Elisa, Nokia, and Teknologian edistämisäätiö (TES). This paper was presented in part at the IEEE Military Communications Conference (MILCOM), Atlantic City, NJ, October 2005.

The authors are with the Centre for Wireless Communications (CWC), University of Oulu, 90014 Oulu, Finland (e-mail: {henri.puska, harri.saarnisaari, jari.iinatti, pekka.lilja}@ee.oulu.fi).

Digital Object Identifier 10.1109/TCOMM.2008.041041.
In fixed beam acquisition, the total uncertainty region increases linearly with an increasing number of angular cells, resulting in a longer mean acquisition time. On the other hand, signal-to-noise ratio (SNR) improvement due to BF will decrease the mean acquisition time. As a consequence, there is an optimum number of angular cells for each SNR value [4]. The SNR improvement produced by antenna array gain can be achieved also using DOA estimation followed by BF. After DOA estimation, the correct DOA and delay is searched among the found DOAs, which may result in fewer angular cells than in fixed beam strategies. It should be noted that maximal SNR improvement is obtained only if the beam is steered exactly toward the actual DOA. Therefore, fixed beam systems suffer from array gain losses due to DOA uncertainty. The DOA estimation-based system may reduce angular uncertainty, provided that the estimation accuracy is sufficient. However, the use of DOA estimation before MF requires that SNR in the channel is so large that successful DOA estimation is possible.

This paper is organized as follows. The code acquisition scheme using a smart antenna with a single correlator is presented in Section II. The fixed beam acquisition using either conventional or advanced BF techniques are considered in Section II-A. The strategy where DOA estimation is used prior to BF is considered in Section II-B. The code acquisition scheme using a smart antenna with a single MF is considered in Section III. The analysis of the mean acquisition time in MF acquisition using DOA estimation and BF is given in Section III-A. As a reference, the mean acquisition time using single antenna and MF is given in Section III-B. Finally, numerical results are shown in Section IV, and conclusions are drawn in Section V.

II. CODE ACQUISITION USING SMART ANTENNA WITH SINGLE CORRELATOR

Fig. 1 shows a block diagram of the proposed code acquisition scheme for a DS/SS receiver with a smart antenna. Therein, either a single correlator (i) or a matched filter (ii) after BF is used. If the uniform linear array (ULA) with antenna spacing of one-half of the carrier wavelength is used, the received baseband signal after chip rate sampling at the mth antenna element \( m \in [1, \ldots, M] \) can be written as

\[
x_m(n) = \sqrt{E_c} b(n - \tau) c(n - \tau) e^{j \varphi} a(\theta) + n_m(n),
\]

where \( M \) is the number of antennas, \( n \) and \( E_c \) are chip index and energy, \( c(n) \) is the PN spreading sequence with length \( LT_c \), \( T_c \) is the chip time, \( b(n) \) is information sequence with symbol time \( LT_c \), \( \varphi \) is the carrier phase, \( \tau \) is the propagation delay, \( \theta \) is the DOA of the desired signal and \( n_m(n) \) is a complex additive white Gaussian noise (AWGN) process with zero mean and variance \( N_0/2 \) per real (I) and imaginary (Q) branch. In addition, noise processes \( n_1(n), \ldots, n_M(n) \) are independent. Signals in different antennas can be expressed in a vector form as

\[
x(n) = \sqrt{E_c} b(n - \tau) c(n - \tau) e^{j \varphi} a(\theta) + n(n),
\]

where the array propagation vector (steering vector) for the ULA array can be written as

\[
a(\theta) = \begin{bmatrix} e^{j \pi \cos \theta} & \ldots & e^{j \pi (M-1) \cos \theta} \end{bmatrix}^T.
\]

The optimum BF weights which maximize array output SNR in spatially white environment are

\[
w_{opt} = a(\theta),
\]

which corresponds to the conventional beamforming technique. Therefore, the array output signal \( y(n) \) can be expressed as

\[
y(n) = w_{opt}^H x(n) = a^H(\theta) a(\theta) \sqrt{E_c} b(n - \tau) c(n - \tau) e^{j \varphi} + a^H(\theta) n(n).
\]

After antenna combining, there is a correlator (Fig. 1 (i)) which correlates the output of the smart antenna with the reference spreading code \( c(n - \tau_l) \), where \( \tau_l \) is the code phase under test. In the correlator code acquisition, the time used for correlating each code phase of the PN code sequence is denoted by \( t_d \) and called as the dwell time. After squaring \( (|\cdot|^2) \)-operation there is the first threshold comparison. The first comparator uses a threshold \( T_{H1} \) to declare the acquisition if \( Z_1(n) > T_{H1} \), where \( Z_1(n) \) is the input into the first comparator. Otherwise, the next code phase is tested. The first threshold comparison may be followed by a post detection integration (PDI) and a second threshold comparison. If the first threshold crossings are verified using PDI and another threshold comparison, then the acquisition is called as a two-dwell acquisition [11].

When PN codes are not synchronized (denoted by ’0’), the output of the correlator \( z(0) \) contains only complex AWG noise if it is assumed that the PN code has an ideal autocorrelation function (ACF). Both the real and imaginary parts of that noise are statistically independent zero mean Gaussian random variables with a variance equal to \( MLN_0/2 \). The probability density function (PDF) of the first decision variable \( Z_1(0) = |z(0)|^2 \) follows a central chi-square distribution with 2 degrees of freedom [12]. The probability of false alarm in the first threshold comparison \( P_{FA1} \) is

\[
P_{FA1} = \exp \left( -\frac{\pi T_{H1}}{\pi T_{H0}} \right), T_{H1} \geq 0.
\]

Correspondingly, the PDF of the second decision variable \( Z_V(0) = \sum_{k=0}^{K-1} |z(k)|^2 \) follows a central chi-square distribution but with 2K degrees of freedom [12], where integer \( K \) is the number of noncoherently combined samples. Thus, the false alarm in the second threshold comparison \( P_{FAV} \) is

\[
P_{FAV} = \exp \left( -\frac{\pi T_{HV}}{\pi T_{H0}} \right) \sum_{k=0}^{K-1} \left( \frac{T_{HV}}{MLN_0} \right)^k, T_{HV} \geq 0.
\]

In the synchronous code phase position (denoted by ’1’), the signal energy is captured. In this case, the output of the correlator \( z(1) \) contains both the signal and noise. The first decision variable \( Z_1(1) = |z(1)|^2 \) is characterized by the noncentral chi-square distribution with 2 degrees of freedom and \( (ML \sqrt{E_c})^2 \) as the noncentrality parameter [12]. Thus,
Correspondingly, the second decision variable loaded into memory. Therefore, a bias term has to be added techniques require a lot of computation before BF weights are already in the memory before the acquisition process starts. BF is used. Then, the BF weights for each angular cell are actually holds if the times used for BF are omitted or if a very process are presented in [4]. It should be noted that (10) results of these relationships and their impacts on the search compensate the expansion of the uncertainty region. Analytical detection in the second threshold comparison (P_{DV}) is

\[
P_{DV} = Q\left(\frac{\sqrt{2K M L E_c}}{N_0}, \frac{\sqrt{22T_{H1}}}{M L N_0}\right),
\]

where \(Q(\cdot)\) is the generalized Marcum's Q-function [12]. Consequently, the probability of detection in the second threshold comparison (P_{DV}) is

\[
P_{DV} = Q\left(\frac{\sqrt{2K M L E_c}}{N_0}, \frac{\sqrt{22T_{H1}}}{M L N_0}\right).
\]

A. Correlator Acquisition Using Fixed Beams

If DOA estimation is not used, the entire angular uncertainty region can be divided into \(B_{NDOA}\) angular cells. If the beams are orthogonal as in [4], then \(B_{NDOA} = M\). Due to DOA uncertainty, the total uncertainty region has \(L\) delay cells and \(B_{NDOA}\) angular cells resulting in a total of \(LB_{NDOA}\) cells. The a priori probability of the synchron cell is assumed to be uniformly distributed over the uncertainty region. Therefore, \(LB_{NDOA}\) cells are serially searched in the angular and delay domains. The mean acquisition time \(T_{MA}\) in this case with the single-dwell acquisition is [4]

\[
T_{MA} = \left\{\frac{2 + (2 - P_{D1})(LB_{NDOA} - 1)(1 + T_{FA}P_{F1})}{2P_{D1}}\right\}\tau_d,
\]

where \(T_{FA}\) is defined as the false alarm penalty time. The disadvantage of this strategy is that the uncertainty region increases linearly as a function of \(B_{NDOA}\) even though just one signal is present. However, the advantage is that antenna array gain is achieved if the directional beam is steered exactly toward the DS-signal. This SNR improvement will partly compensate the expansion of the uncertainty region. Analytical results of these relationships and their impacts on the search process are presented in [4]. It should be noted that (10) actually holds if the times used for BF are omitted or if a very simple BF algorithm like the Butler matrix or a conventional BF is used. Then, the BF weights for each angular cell are already in the memory before the acquisition process starts. The equation will change somewhat if the new beamforming weights have to be calculated before the next angular cell can be tested, as will be shown in Section III-A. Advanced BF techniques require a lot of computation before BF weights are loaded into memory. Therefore, a bias term has to be added to (10) if those times, consumed by smart antenna operations, are included into the acquisition process. In that case, (10) changes into the form

\[
T_{MA} = \left\{\frac{2 + (2 - P_{D1})(LB_{NDOA} - 1)(1 + T_{FA}P_{F1})}{2P_{D1}} + T_{ABF}\right\}\tau_d,
\]

where the bias term \(T_{ABF}\) is the time required for advanced digital BF operations to finalize the weights. The operations include: data collection for the correlation matrix estimation, matrix inversion and, finally, calculation of the BF weights for different directions. An advantage of the advanced BF is its ability to suppress strong interference. This is not possible if a simple BF is used. Actually, probabilities of detection (8) and (9) depend on the output signal-to-interference plus noise ratio (SINR) of the beamformer. With an advanced BF output SINR is higher in the presence of interference. However, the effect of interference is not included into the analysis of synchronization probabilities in this paper. Therefore, theoretical synchronization performance comparisons should also be studied in different interference environments.

The fixed beam method with conventional BF is not very sensitive to steering vector errors, whereas most advanced BF methods are. However, there are efficient methods to reduce the effects of uncertainties in the steering vector [13]. In this work, we have assumed error free steering vectors. Furthermore, we have assumed that each arriving signal appears only within one angular cell which is not necessarily realistic in practice. When the beam is steered towards an interfering signal, the threshold value is adjusted according to the interference plus noise power. This ensures that \(P_{FA}\) remains constant regardless of the selected DOA.

B. Correlator Acquisition Using DOA Estimation and Advanced Beamforming

Angular uncertainty may be reduced using DOA estimation, since only identified DOAs have to be searched. This means that the angular uncertainty region may be noticeably smaller than in fixed beam methods. The disadvantage compared to the fixed beam strategy is that DOA estimation requires more computations. However, DOA estimation with advanced BF does not increase the computational complexity significantly because DOA estimation and advanced BF may share computations (e.g., the correlation matrix is needed to be estimated and inverted only once). Therefore, a receiver equipped with
an advanced BF may also be used for DOA estimation. The expression for the mean acquisition time can shown to be

\[ T_{MA} = \left\{ \frac{2 + (2 - P_{D1})(LB_{DOA} - 1)(1 + T_{FA}P_{FA1})}{2P_{D1}} \right\} T_{d} + T_{DOA+ABF} \]  

(12)

where \( B_{DOA} \) is the number of arriving signals found by DOA estimation and \( T_{DOA+ABF}T_{d} \) is the time needed for the calculation of the number of signals, DOAs and corresponding BF weights.

III. CODE ACQUISITION USING SMART ANTENNA WITH SINGLE MATCHED FILTER

The principle of the noncoherent MF synchronization with a smart antenna is shown in Fig. 1 (ii). The MF based synchronization is faster than the correlator based because decision variables are generated at the chip rate (in one sample per chip processing), whereas in the correlator acquisition, they are generated at the integration time (dwell time) rate \( 1/T_{d} \) [14]. However, MF synchronization has to be started after passing one code period because of filling of the MF [15]. If MF acquisition is used in a smart antenna system, then the waiting time is a concern for all angular cells. Next, the serial search MF acquisition scheme with the DOA estimation and BF is analyzed.

A. Analysis of the Mean Acquisition Time Using DOA Estimation, Beamforming and Matched Filtering

It is assumed that there are \( L \) delay cells and \( Q \) angular cells. \( Q \) angular cells are achieved either from the DOA estimator (\( B_{DOA} \)) or by dividing the entire angle uncertainty into \( Q \) cells using fixed beam methods (\( B_{NDOA} \)). Therefore, the analyzed \( T_{MA} \) is a general result. The analysis technique used in this work is mainly the same as used in [16]. The circular diagram describing the proposed method is shown in Fig. 2. In this figure, the squares denote angular cells, i.e., \( q = 1, 2, \ldots , Q \) and the circles denote delay cells, i.e., \( l = 1, 2, \ldots , L \). Each angular cell contains \( L \) delay cells since all delays are searched for each DOA. The synchro cell \((q_{s}, l_{s})\) is, without the loss of generality, the \( l_{s} \)th delay cell of the \( q_{s} \)th angular cell, i.e., \( q_{s} = q, l_{s} = l \).

The acquisition process starts from the stage \( S \) such that the receiver calculates the number of arriving signals and their DOAs. In addition, it also calculates BF weights for different directions and saves them to memory. Therefore, different BF weights are readily available when the actual acquisition process is performed. As a consequence, the transfer function describing that process can be expressed as

\[ A(z) = z^{T_{DOA+ABF}T_{c}}. \]  

(13)

After DOA estimation, the receiver does not know which of these DOAs is the correct one. The test starts from the first angular cell \((q_{i}, l_{i}) = 1\) by focusing a radiation pattern toward that direction. The loading of the weight coefficient vector from memory consumes time \( T_{BF}T_{c} \) or alternatively, \( T_{BF}T_{c} \) may include calculation processes for the whole weight coefficient vector if these coefficients are not calculated during the DOA estimation. After BF the receiver has to wait \( T_{MF}T_{c} \) until MF fills. After that, the acquisition process can start from the first delay cell \((l_{i} = 1\). The overall graph transfer function between states \( S \) and \( ACQ \) presents the acquisition time generating function [17]. The first derivative of this function is used to derive the mean acquisition time [18]. To find the overall transfer function, a detailed state diagram shown in Fig. 3 has to be used in the analysis. Fig. 3 is actually a small detailed slice of the overall transfer function (Fig. 2) taken from the vicinity of the synchro cell. The analysis is derived using two-dwell acquisition because then the result is more general.

The transfer function from the synchro cell to the \( ACQ \) is defined as

\[ H_{D}(z) = H_{D1}(z)H_{DV}(z) = P_{D1}P_{DV}z^{((K-1)L+1)T_{c}}, \]  

(14)

where \( H_{D1}(z) = P_{D1}^{T_{c}} \) means correct detection in the first threshold comparison, \( H_{DV}(z) = P_{DV}z^{(K-1)L_{T_{c}}} \) means correct detection in the verification, where integer \( K \) is the number of samples used to make the synchronization decision. Therefore, \( K = 1 \) describes the single-dwell acquisition and \( K > 1 \) describes the two-dwell acquisition. If \( K > 1 \), then \((K-1)L \) chips have to wait until a synchronization decision can be made. If the synchro cell is not detected, then it is missed. The transfer function for that is

\[ H_{M}(z) = H_{M1}(z) + H_{D1}(z)H_{MV}(z) \]
\[ = (1 - P_{D1})z^{T_{c}} + P_{D1}(1 - P_{DV})z^{((K-1)L+1)T_{c}}, \]  

(15)

where \( H_{M1}(z) = (1 - P_{D1})z^{T_{c}} \) is a miss without verification, and \( H_{MV}(z) = (1 - P_{DV})z^{(K-1)L_{T_{c}}} \) is a miss with verification of the correct cell. The transfer function from one non-synchro delay cell to another inside each angular cell is
Circular diagram analysis methods may be used to show that the transfer function from the initial cell to the ACQ state is given by

$$u_{q_s,l_s}(z) = \frac{A(z)B_{q_s,l_s}(z)C(z)}{1 - B_{q_s,l_s}(z)D_{q_s,l_s}(z)}.$$  \( \text{(20)} \)

The generating function \( U(z) \) modeling the overall acquisition process ending to any arbitrary cell, \( q_s = 1, 2, \ldots, Q \); \( l_s = 1, 2, \ldots, L \), can be written as

$$U(z) = \sum_{q_s=1}^{Q} \sum_{l_s=1}^{L} \psi_{q_s,l_s} u_{q_s,l_s}(z).$$ \( \text{(21)} \)

where \( \psi_{q_s,l_s} \) is the probability of ending the search process to the cell \( (q_s,l_s) \). It is assumed that the acquisition process can end at any cell with the same probability, i.e., the ending point of the search process follows uniform distribution. Since there are altogether \( QL \) cells, \( \psi_{q_s,l_s} = 1/QL \) and the generating function \( (21) \) becomes

$$U(z) = \frac{1}{QL} \sum_{q_s=1}^{Q} \sum_{l_s=1}^{L} u_{q_s,l_s}(z).$$ \( \text{(22)} \)

The mean acquisition time (\( T_{MA} \)) is obtained by differentiating the generating function and evaluating the result at \( z = 1 \), i.e.,

$$T_{MA} = \frac{\partial U(z)}{\partial z} \bigg|_{z=1} = \frac{1}{QL} \sum_{q_s=1}^{Q} \sum_{l_s=1}^{L} \frac{\partial u_{q_s,l_s}(z)}{\partial z} \bigg|_{z=1}. \text{(23)}$$

The details of the derivation are given in the Appendix. The final result is

$$T_{MA} = \left\{ T_{DOA+ABF} + \frac{(2 - P_{D1}P_{DV})}{2P_{D1}P_{DV}} \{ (Q - 1)[T_{BF} + T_{MF}] + (QL - 1)[1 + P_{FA1}(K - 1)L + P_{FA1}P_{FAV}T_{FA}] + \frac{1}{P_{D1}P_{DV}}[1 + T_{BF} + T_{MF} + P_{D1}(K - 1)L] \} \right\} T_c. \text{(24)}$$

where \( Q = B_{DOA} \) if DOA estimation is used, and \( Q = B_{NDOA} \) if it is not used. Also, \( T_{DOA+ABF} \) must be replaced with \( T_{ABF} \) if the fixed beam method with the advanced BF...
is used, and $T_{DOA+ABF} = 0$ if the fixed beam method with the conventional BF is used.

The $T_{MA}$ in the single-dwell acquisition is obtained from the general result (24) by inserting: $K = 1$, $P_{DV} = 1$ and $P_{FAV} = 1$. Therefore, the general $T_{MA}$ without verification is

$$T_{MA} = \left\{ T_{DOA+ABF} + \frac{(2 - P_{D1})}{2P_{D1}} \left\{ (Q - 1)[T_{BF} + T_{MF}] + (QL - 1)[1 + P_{FA1}T_{FA}] \right\} + \frac{1}{P_{D1}} \{1 + T_{BF} + T_{MF} \} \right\} T_c,$$

(25)

from which the $T_{MAS}$ with and without DOA estimation can be obtained by replacing $Q$ and $T_{DOA+ABF}$ as was discussed above.

The $T_{MA}$ in the correlator and single-dwell acquisition can be derived from (25). In the correlator acquisition, the filling time of the MF is not needed (i.e., $T_{MF} = 0$). Downloading time of the weight coefficient vector from memory ($T_{BF}$) is assumed to be insignificant (i.e., $T_{BF} = 0$). In addition, using correlator acquisition, the time $T_c$ must be replaced by $\tau_d$, i.e., the rate of decision variables. Therefore, the general $T_{MA}$ in the correlator acquisition with DOA estimation and advanced BF is

$$T_{MA} = \left\{ T_{DOA+ABF} + \frac{(2 - P_{D1})}{2P_{D1}} \left\{ (Q - 1)[T_{BF} + T_{MF}] + (QL - 1)[1 + P_{FA1}T_{FA}] \right\} + \frac{1}{P_{D1}} \{1 + T_{BF} + T_{MF} \} \right\} \tau_d,$$

(26)

which is exactly the same result obtained in (12) when $Q = B_{DOA}$. The $T_{MA}$ equations for the correlator acquisition in the fixed beam cases (10) and (11) are also obtained from (26) by replacing $Q = B_{NDOA}$. In addition, the $T_{DOA+ABF}$ must be replaced with $T_{ABF}$ if the fixed beams are generated using advanced BF and with 0 if conventional BF is used.

B. The Mean Acquisition Time Using Single Antenna and Matched Filter

It is also interesting to compare the performance of the multi antenna to single antenna receivers. Intuitively, it could be assumed that the $T_{MA}$ in the single antenna and in the single-dwell system can be obtained by setting $Q = 1$ in (25), which corresponds to the situation with only one angular cell. In addition, by setting $T_{DOA+ABF} = T_{BF} = 0$, all operations corresponding to array signal processing are removed. If these parameters are substituted into (25), it follows that

$$T_{MA} = \left\{ \frac{(2 - P_{D1})}{2P_{D1}} \left\{ (L - 1)[1 + P_{FA1}T_{FA}] \right\} + \frac{1}{P_{D1}} \{1 + T_{MF} \} \right\} T_c,$$

(27)

which is slightly different than the correct result [15]

$$T_{MA} = \frac{T_{MF}}{P_{D1}} + T_c = (L - 1)(T_c + T_{FA}P_{FA1})\frac{(2 - P_{D1})}{2P_{D1}} + \frac{1}{P_{D1}} \{1 + T_{MF} \} T_c,$$

(28)

where $T_{MF} = T_{MF} T_c$ and $T_{FA} = T_{FA} T_c$. The difference between these two equations is that in (27) there is the term $T_{MF} T_c$ (i.e., the term $\frac{T_{MF}}{P_{D1}}$) instead of the term $T_{MF}$ in (28). The difference is a consequence of the multi antenna generating function used in this paper. Therein, the filling time of the MF has to be waited at the beginning of each angular cell. However, in the real single antenna system, the delay time $T_{MF}$ only occurs once before the acquisition process starts (i.e., it is just the bias term) and does not need to be repeated, although the correct code phase would be missed. Therefore, the correct $T_{MA}$ result for the single antenna receiver can also be obtained from (25) by replacing the bias term $T_{DOA+ABF}$ with $T_{MF}$. The term $T_{MF}$ can be removed from the last term where $T_{MF}$ is weighted by $\frac{1}{P_{D1}}$, because the filling of MF does not depend on the detection probability of the single antenna receiver. After that, inserting $T_{BF} = 0$ and $Q = 1$, the result obtained in (28) follows.

IV. NUMERICAL RESULTS

In this section we provide some numerical examples that provide an overview of the performance of the studied schemes. The starting point is to consider the performance of the DOA estimation. Fig. 4 presents simulation results of the Root-MUSIC algorithm [19] when the 8 element ULA with antenna spacing of one half of the carrier wavelength is used. It shows the root mean square error (RMSE) of the DOA estimate versus $E_c/N_0$ and versus the number of sample vectors (i.e., $x(n)$ in (2)) used for DOA estimation. DOA estimation algorithms use these sample vectors for correlation estimation. The decision SNR, in decibels, is

$$\left( \frac{E_b}{N_0} \right)_d = \frac{E_c}{N_0} + PG_{ML},$$

(29)

where $E_b$ is the bit energy and $PG_{ML} = 10\log_{10}(ML)$ is the processing gain produced by antenna and DS combining. In this work, $M = 8$, $L = 63$ and, therefore, $PG_{ML} \approx 27$ dB. Since a typical requirement for data reception is positive (positive $\left( \frac{E_b}{N_0} \right)_d$), the number of sample vectors used for DOA estimation should be selected such that the DOA estimation succeeds even when $E_c/N_0 \approx -20$ dB. If the code length is larger than 63 then $E_c/N_0$ may be so low that the DOA estimation from the chip level is not possible. If the decision SNR is allowed to be larger than the minimum needed, from the data detection point of view, this problem can be avoided. On the other hand, if the radio system contains both single and multi antenna receivers, then the transmission power has to be so high that data detection in the single antenna receiver also succeeds. Then the typical requirement is that the SNR per bit per antenna ($E_b/N_0$) is $\geq 5$ dB. In that case, $E_c/N_0$ in decibels is

$$\frac{E_c}{N_0} = \frac{E_b}{N_0} - PG_L,$$

(30)

where $PG_L = 10\log_{10}(L)$ is the processing gain only produced by DS combining. $PG_L \approx 18$ dB when $L = 63$. Based on (30), the number of required sample vectors should be selected such that DOA estimation succeeds when $E_c/N_0 \approx$
$-13$ dB, which requires 500 sample vectors (Fig. 4). For the remainder of this paper following assumptions have been made: Time consumed for DOA estimation and advanced BF ($T_{DOA+ABF}$) is 1000 $T_c$; If DOA estimation is not used, then $T_{ABF} = 500 T_c$; Number of beams generated by the DOA estimation-based method ($B_{DOA}$) is equal to the actual number of arriving signals; $P_{FA} = 10^{-3}$; False alarm penalty time is 100 $LT_c$; The array gain is the maximum inside each angular cell. This may not be the situation in practice and, therefore, the obtained results reflect lower bounds of mean acquisition times. The obtained $T_{MA}$ results in Figs. 5 - 9 are normalized with respect to one code period $T_0 = LT_c$.

The mean acquisition time using the presented correlator acquisition strategies, as a function of $M$, is presented in Fig. 5. In these results, only a single signal arrives at the receiver. Therefore, the number of beams generated by the DOA estimation-based method is equal to one (i.e., $B_{DOA} = 1$). On the other hand, in Fig. 5 the number of angular cells ($B_{NDOA}$) generated by the fixed beam structure is assumed to be equal to $M$ despite of the actual number of arriving signals. It can be seen that $T_{MA}$ of the fixed beam techniques are almost the same because there is no interference. In the fixed beam cases, there is also an optimum number of antennas for which the minimum $T_{MA}$ is obtained. This is a consequence of the fact that increasing numbers of angular cells ($B_{NDOA}$) increases the uncertainty region. On the other hand, the $(E_b/N_0)_{ad}$ increases as a function of $M$, and this tends to decrease the $T_{MA}$. If the input $E_b/N_0$ is low (0 dB), then a larger number of antennas is required to obtain the minimum $T_{MA}$ and vice versa. It can be seen from Fig. 5 that the DOA estimation-based method avoids this problem.

Figs. 6, 7, 8 and 9 present $T_{MA}$ results as a function of $E_b/N_0$. Figs. 6 and 7 show the performances for the correlator and Figs. 8 and 9 for the MF acquisition, respectively. In Fig. 6 numerical results are given when $B_{DOA}$ is 1, 4, or 8 and $B_{NDOA} = M$. The single antenna receiver serves as a reference. It can be observed that the single antenna receiver performs better than multi antenna receivers with high $E_b/N_0$ and worse with low $E_b/N_0$ values. It can also be noted that $T_{MA}$ saturates very rapidly to a relative high value if a DOA estimation is not used. On the other hand, the DOA estimation-based method avoids this problem. By increasing $B_{NDOA}$, the maximum tolerable steering vector error is not exceeded. If, e.g., uniform circular array (UCA) with $M = 8$ and half wavelength antenna separation is used, its half power beam width (HPBW) is approximately $32^\circ$ [20]. If it is desired that the antenna array gain is always larger than one half of the maximum array gain, there must be at least 12 angular cells (i.e., $360^\circ/32^\circ = 11.25$). In Fig. 7 $T_{MA}$ results are given when $B_{NDOA} = 12$. It can be observed from this figure that now $T_{MA}$ results for the fixed beam methods are increased such that the DOA estimation-based method is better, although 8 signals are arriving to the receiver.
Fig. 6. The mean acquisition time as a function of $E_b/N_0$, when a single correlator after BF is used.

Fig. 7. The mean acquisition time as a function of $E_b/N_0$, when a single correlator after BF is used.

Fig. 8. The mean acquisition time as a function of $E_b/N_0$, when a single MF after BF is used.

Fig. 9. The mean acquisition time as a function of $E_b/N_0$, when a single MF after BF is used.

Fig. 8 shows $T_{MA}$ results for the investigated acquisition strategies when a single MF instead of a correlator is used after antenna combining. Otherwise the parameters are the same as they were in Fig. 6. By comparing Figs. 6 and 8 it can be noticed that the absolute $T_{MAS}$ are much smaller in Fig. 8 due to fast MF acquisition. Also, the superiority of the single antenna receiver at high $E_b/N_0$ values is shown clearly. The single antenna receiver has a clear advantage since it does not need to perform calculations prior to acquisition. However, these results are valid for the AWGN channel only, and they would be different if interfering signals were present.

The other difference between Figs. 6 and 8 emphasizes in $B_{DOA}$ value. In the correlator acquisition (Fig. 6), the fixed beam techniques and the DOA estimation-based method gave almost the same results when $B_{DOA} = 8$, but in the MF acquisition those methods almost coincide if $B_{DOA} = 4$. If $B_{DOA} = 8$ in the MF acquisition, the DOA estimation-based method is slower. This difference can be explained by the time used for DOA estimation, which is relatively larger in the MF than in the correlator acquisition.

Fig. 9 illustrates $T_{MA}$ results in the MF acquisition when $B_{NDOA} = 12$. By comparing it to the Fig. 8, where $B_{NDOA}$ was $M$, it can be seen that the results in Fig. 9 for the fixed beam methods are increased such that the DOA estimation-based method with $B_{DOA} = 8$ gives almost the same performance as fixed beam methods.

It should be kept in mind that the presented $T_{MA}$ results, especially for fixed beam acquisition, are at their lower bounds due to maximum array gain assumption. The results where DOA estimation is used are closer to true values because it is more probable that the beam is then steered almost exactly toward the signal and whole array gain is achieved.

V. CONCLUSIONS

Three different methods for code synchronization were investigated in this paper. The investigated methods are suitable, especially for receivers that consist of a smart antenna.
followed by a single correlator or MF. In two of these methods, the angular uncertainty region was divided into small cells using fixed beam techniques. Fixed beams were produced either by conventional or by advanced BF algorithms. Advanced BF requires more computation than conventional BF, but it also offers interference suppression capability. In the third method, DOAs of arriving signals and their BF weights were calculated prior to the acquisition process. This may reduce angular uncertainty. All these acquisition strategies were compared in serial search acquisition using either the correlator or the MF. A model for code acquisition using DOA estimation and BF in MF acquisition was presented and an expression for the mean acquisition time was derived. This may reduce angular uncertainty, whereas smart antennas, especially with advanced BF, are not. If $E_b/N_0$ is low, then multi antenna receivers typically outperform single antenna receivers. On the other hand, if the fixed beam and DOA estimation-based methods are compared, then the DOA estimation-based methods outperform the fixed beam techniques, especially if the number of arriving signals is low. However, DOA estimation from the chip level may be difficult if $E_c/N_0$ is very low. In that case, fixed beam methods have the best performance. By comparing the performances of the MF and correlator acquisition, it can be seen that MF acquisition will give noticeably smaller $T_{MA}$ results. The results obtained in this work can also be used to approximate the time consumed for DOA estimation in such a way that performance is still better than in strategies in which DOA estimation is not used. Future work will include investigations on the effects of steering vector errors, different acquisition strategies and interference into mean acquisition times.

**APPENDIX**

**DERIVATION OF EXPRESSION FOR THE MEAN ACQUISITION TIME**

The mean acquisition time is obtained from (23). The derivative of (20) with respect to $z$ is

$$\frac{\partial u_{q_s,l_s}(z)}{\partial z}|_{z=1} = \left\{ \left[ A(z)B_{q_s,l_s}(z)C(z) + A(z)B'_{q_s,l_s}(z) \times C(z) + A(z)B_{q_s,l_s}(z)C'(z) \times [1 - B_{q_s,l_s}(z)D_{q_s,l_s}(z)] + [B'_{q_s,l_s}(z)D_{q_s,l_s}(z)] \times \left[ A(z)B_{q_s,l_s}(z)C(z) \right] \right] \right\}$$

$$\left/ [1 - B_{q_s,l_s}(z)D_{q_s,l_s}(z)] \right|^2 ,$$

(31)

where $'$ denotes derivative. After some calculations the functions and derivatives in (31) are

$$A(z)|_{z=1} = 1,$$

(32)

$$A'(z)|_{z=1} = T_{DOA+ABF}T_c;$$

(33)

$$B_{q_s,l_s}(z)|_{z=1} = 1,$$

(34)

$$B'_{q_s,l_s}(z)|_{z=1} = (q_s - 1) \left[ H_{BF}(z) + H'_{MF}(z) + L_{H_0}(z) \right] + H'_{BF}(z) + H'_{MF}(z) + (l_s - 1)H_0'(z),$$

(35)

$$C(z)|_{z=1} = H_D|_{z=1} = P_{DL}P_{DV},$$

(36)

$$C'(z)|_{z=1} = H'_D(z)|_{z=1} = P_{DL}P_{DV}(K-1)L+1T_c;$$

(37)

$$D_{q_s,l_s}(z)|_{z=1} = 1 - C(z)|_{z=1} = (1 - P_{DL}P_{DV}),$$

(38)

and

$$D'_{q_s,l_s}(z)|_{z=1} = H'_M(z) + (L - l_s)H_0'(z)H_M(z) + (Q - q_s)H_M(z)[H'_{BF}(z) + H'_{MF}(z) + LH_0'(z)].$$

(39)

Because (31) reduces into the form

$$\frac{\partial u_{q_s,l_s}(z)}{\partial z}|_{z=1} = \frac{1}{C(z)} \left[ A'(z)C(z) + C'(z) \right]$$

$$+ \frac{1}{C(z)} \left[ B'_{q_s,l_s}(z) + D'_{q_s,l_s}(z) \right],$$

(40)

the final expression for the mean acquisition time (24) is achieved by inserting (40) into (23) and then doing some calculations.

**ACKNOWLEDGMENT**

The authors wish to thank V. Aalo, the Editor of the IEEE Transactions on Communications, and anonymous reviewers for their help in editing the paper.

**REFERENCES**


Henri Puska (S’04) was born in 1977. He received the Master of Science degree in electrical engineering from the University of Oulu, Department of Electrical and Information Engineering, Oulu, Finland, in 2002. Since then he has been a Research Scientist at the Centre for Wireless Communications (CWC) at the University of Oulu, Finland, where he is working toward the Ph.D. degree. His research interests include synchronization and array processing techniques.

Harri Saarnisaari received the M.Sc. degree in applied electrical physics from University of Kuopio in 1993 and the Dr. Tech. degree from University of Oulu in 2000. He is a member of EURASIP and a senior member of IEEE. Currently he is a senior research scientist in Centre for Wireless Communications (CWC) at the University of Oulu. His research interest lies in signal processing. Especially he is interested in direct sequence code phase synchronization in hostile environment, adaptive antenna algorithms, channel estimation and positioning. He has been participated to design and developing advanced software defined radio for Finnish defence forces. This SDR involves anti-jamming data and positioning waveforms and an adaptive antenna.

Pekka Lilja (S’05) received Master of Science and Licentiate of Science degrees in electrical engineering from the University of Oulu, Department of Electrical and Information Engineering, Oulu, Finland, in 1997 and 2005 respectively. He was with Nokia Cellular Systems in 1994-1995 working with Base Station Design. Since 1996 he has been with the Telecommunication Laboratory and the Centre for Wireless Communication. His current research interests include Software Defined Radio, adaptive array processing and delay estimation.

Jari H.J. Iinatti (Member IEEE 95, Senior Member IEEE 2005) was born in Oulu, Finland, in 1964. He received the Master of Science (in Electrical Engineering) degree in 1989, the Licentiate in Technology degree in 1993, and Doctor of Technology in 1997, all from the University of Oulu, department of Electrical Engineering, Oulu, Finland. During 1989-1997 he was a Research Scientist at the Telecommunication Laboratory at the University of Oulu. He finished his Doctoral Thesis dealing with spread spectrum code acquisition in 1997. During 1997-2002 he was acting professor of Digital Transmission Techniques at the University of Oulu, and senior research scientist, project manager and research director at Centre for Wireless Communications (CWC) at the University of Oulu. Since 2002 he has been Professor of Telecommunication Theory. His research interests are in future wireless communication systems, UWB systems and transceiver algorithms.

Prof. Iinatti has published more than 90 international journal and conference papers, holds three patents, and is a co-editor of the book, UWB Theory and Applications (Wiley & Sons, Ltd., Chichester, UK, 2004).
Performance Analysis for MIMO Systems with Lattice-Reduction Aided Linear Equalization

Xiaoli Ma, Member, IEEE, and Wei Zhang, Student Member, IEEE

Abstract—Multi-input multi-output (MIMO) systems equipped with multiple antennas have well documented merits in combating fading and enhancing data rates. MIMO V-BLAST transmission is a widely adopted method to achieve high spectral efficiency and low-complexity implementation. When the maximum likelihood (ML) or near-ML detector is employed, receive diversity is collected for MIMO V-BLAST systems to enhance the performance. However, because of its exponential complexity, ML detector may be infeasible for practical systems when the number of antennas and/or the constellation size is large. On the other hand, linear equalizers have much lower complexity but come with inferior performance. In this paper, we analytically quantify the diversity order of linear detectors for MIMO V-BLAST systems. Then, we adopt low-complexity complex lattice-reduction (LR) aided linear equalizers for V-BLAST systems to improve the performance and prove that LR-aided linear equalizers collect the same diversity order as that exploited by the ML detector but with much lower complexity. Relative to the existing real LR-aided equalizers, we illustrate that the complex LR further reduces the complexity while keeping the same performance. Simulation results corroborate our theoretical claims.

Index Terms—MIMO, diversity, linear equalizers, lattice reduction.

I. INTRODUCTION

THE multi-input multi-output (MIMO) wireless systems are motivated by two ultimate goals of wireless communications: high-data-rate and high-performance [2], [19]. During recent years, various space-time (ST) coding schemes have been proposed to collect spatial diversity and/or achieve high rates. Among them, V-BLAST (Vertical Bell Labs Layered Space-Time) transmission has been widely adopted for its high spectral efficiency and low implementation complexity [4]. When maximum-likelihood (ML) detector is employed, V-BLAST systems also enjoy receive diversity, but the decoding complexity is exponentially increased by the number of transmit-antennas. Although some (near-)ML schemes (e.g., sphere-decoding (SD), semi-definite programming (SDP)) can be used to reduce the decoding complexity, at low signal-to-noise ratio (SNR) or when a large number of transmit-antennas and/or high signal constellations are employed, the complexity of near-ML schemes is still high. Some sub-optimal detectors have been developed, e.g., successive interference cancellation (SIC), decision feedback equalizer (DFE), which are unable to collect receive diversity [20]. To further reduce the complexity, one may apply linear detectors such as zero-forcing (ZF) and minimum mean-square error (MMSE) equalizers. It is well-known that linear detectors have inferior performance relative to that of ML detector. However, unlike ML detector, the expected performance (e.g., diversity order) of linear equalizers has not been quantified directly. The mutual information of ZF equalizer has been studied in [5] with channel state information at the transmitter. The effect of transmit correlation on the performance of ZF equalizer is discussed in [6]. The first goal of our paper is to quantify the diversity orders collected by linear equalizers for V-BLAST systems based on analysis of bit error rate (BER) performance.

Recently, lattice reduction (LR) technique has been proposed to improve the performance of linear equalizers (and SIC algorithms) over MIMO systems (see e.g., [22], [24], [26]). The approaches in [22], [24] are based on real LR methods. The complex LR method is introduced in [26], but only for the two-transmit-antenna case. In [14], it is mentioned that the complex LR algorithm can be extended to any number of transmit-antennas. A more detailed algorithm is provided in [3]. In general, the existing results in [22], [23], [24], [26] show that LR improves the performance of linear detectors while the complexity does not increase much. LR-aided precoded MIMO broadcast system is shown to be able to collect the receive diversity when channel knowledge is available at the transmitter in [18]. When the lattice is infinite, [12] links the diversity of LR-aided ZF and ML equalizers where proximity factor is used. In this paper, we adopt the complex Lenstra-Lenstra-Lovász (CLLL) algorithm, compare this algorithm with its real counterpart in complexity and performance, and delineate the difference with CLLL in [3]. Based on the CLLL algorithm and orthogonality deficiency definition, we analytically prove that the complex LR-aided linear equalizers for MIMO V-BLAST systems achieve the same diversity order as that collected by the ML detector without channel knowledge at the transmitter.

The rest of this paper is organized as follows. In Section II, the MIMO system model is introduced. Section III gives the performance analysis of the linear equalizers and Section IV develops and investigates the performance of the complex LR-
aided linear detectors for MIMO systems. Simulation results are presented in Section V to corroborate our theoretical claims. Section VI concludes the paper.

Notation: Upper (lower) bold face letters will be used for matrices (column vectors). Superscript $^H$ denotes Hermitian, * conjugate, and $^T$ transpose. The real and imaginary parts are denoted as $\Re[\cdot]$ and $\Im[\cdot]$. We reserve $[a]$ for the absolute value of scalar $a$. $|a|$ for the 2-norm of vector $a$, $E[\cdot]$ for expectation, and $\det(A)$ for the determinant of $A$: $|a|$ is to round the real and imaginary parts of the complex number $a$ to the nearest integer respectively; $\text{diag}(x)$ stands for a diagonal matrix with $x$ on its main diagonal. $I_N$ denotes the $N \times N$ identity matrix. $\mathbb{Z}$ is the integer set and $\mathbb{C}$ stands for the complex field. $\mathbb{Z}[j]$ denotes the Gaussian integer ring whose elements have the form $Z + jZ$, with $j = \sqrt{-1}$.

II. System Model

Consider a multi-antenna system with $N_t$ transmit-antennas and $N_r$ receive-antennas. For V-BLAST transmissions, the data stream is divided into $N_t$ sub-streams and transmitted through $N_t$ antennas. Let $s = [s_1, s_2, \ldots, s_{N_t}]^T \in \mathbb{C}^{N_t}$ represent the $N_t \times 1$ transmitted data vector at one time slot, where $S$ is the constellation set of each entry of $s$; and let $w = [w_1, w_2, \ldots, w_{N_t}]^T$ denote the white Gaussian noise vector observed at the $N_r$ receive-antennas with zero mean and covariance matrix $E[ww^H] = \sigma_w^2 I_{N_t}$. For simplicity, we assume that the power of each transmit-antenna is normalized to one, and $E[ss^H] = I_{N_t}$.

Denote the received signal at one time slot from $N_r$ antennas as: $y = [y_1, y_2, \ldots, y_{N_r}]^T$ which can be expressed as

$$y = Hs + w,$$  

(1)

where $H$ is the channel matrix which consists of $N_r \times N_t$ independent and identically distributed (i.i.d.) complex Gaussian coefficients with zero mean and unit variance. We assume a flat-fading quasi-static environment where the channel coefficients are invariant during a frame and change independently from frame to frame. We also assume that the channel matrix $H$ is known at the receiver, but unknown at the transmitter.

Given the model in (1), a number of decoders may be used at the receiver. The maximum-likelihood (ML) detector for (1) is provided as follows:

$$\hat{s}_{ML} = \arg \min_{s \in S} \| y - Hs \|^2.$$  

(2)

The diversity order collected by this ML detector is $N_r$ (see e.g., [20, p. 82]). However, the cardinality of the searching space is $M^{N_t}$, where $M$ is the constellation size of $S$. This exponential complexity prohibits the use of ML detection in practical systems (especially for systems that have a large constellation size $M$ and/or a large number of transmit-antennas $N_t$). In the next section, we review the linear detectors, whose decoding complexity is much lower than ML and thus computationally preferable in certain situations. Furthermore, we will analytically quantify the maximum diversity that linear equalizers can collect for MIMO V-BLAST systems.

III. LINEAR EQUALIZERS AND PERFORMANCE ANALYSIS

In this section, we briefly describe the linear equalizers for V-BLAST systems and then focus on analyzing their performance. For ZF equalizer to exist, we consider $N_t \leq N_r$.

A. Linear Equalizers

The ZF detector for the input-output relationship in (1) is given as

$$x_{ZF} = H^H y = s + H^H w = s + n,$$  

(3)

where $H^H$ denotes the Moore-Penrose pseudo-inverse of the channel matrix $H$ and $n := H^H w$ is the noise after equalization. When $H$ has full column rank, the pseudo-inverse matrix can be written as

$$H^H = (H^H H)^{-1} H^H.$$  

(4)

The quantization step is then used to map each entry of $x_{ZF}$ into the symbol alphabet $S$ as

$$\hat{s}_{ZF} = Q(x_i) = \arg \min_{s \in S} |x_i - s|,$$  

(5)

where $x_i$ denotes the $i$th element of $x_{ZF}$ and $Q(\cdot)$ means the quantization operation.

Another often used linear equalizer is linear minimum-mean-square error (MMSE) detector, which takes into account the noise variance. Since we treat $H$ and $s$ deterministically and the noise is Gaussian, a linear MMSE estimator for $s$ is also an MMSE equalizer. Based on the model in (1), the MMSE equalizer is given as

$$x_{MMSE} = \left( H^H H + \sigma_w^2 I_{N_t} \right)^{-1} H^H y.$$  

(6)

The quantization step is the same as the one in (5). Here, we notice that, with the definition of an extended system (as shown in [7], [23])

$$\tilde{H} = \begin{bmatrix} H \\ \sigma_w I_{N_t} \end{bmatrix}, \quad \tilde{y} = \begin{bmatrix} y \\ 0_{N_t \times 1} \end{bmatrix},$$  

(7)

the MMSE equalizer in (6) can now be rewritten as

$$x_{MMSE} = \tilde{H}^H \tilde{y}.$$  

(8)

Thus, the MMSE equalizer in (8) has the same form as the ZF equalizer in (3). In the following, we analyze the performance of ZF and MMSE equalizers for V-BLAST systems in terms of diversity.

B. Performance Analysis

Starting from (3), we first study the performance of the ZF equalizer. Suppose that the $i$th transmitted symbol is $s_i$, and at the receiver it is erroneously decoded as $\hat{s}_i \neq s_i$. Given the channel $H$, the pairwise error probability is

$$P(s_i \rightarrow \hat{s}_i \mid H) = P(|x_i - \hat{s}_i|^2 < |x_i - s_i|^2 \mid H).$$  

(9)

If we define $e_i = s_i - \hat{s}_i$, then the pairwise error probability can be further simplified as

$$P(s_i \rightarrow \hat{s}_i \mid H) = P(|e_i + n_i|^2 < |n_i|^2 \mid H)$$

$$= P(-e_i n_i^* - e_i^* n_i > |e_i|^2 \mid H),$$  

(10)
where \( n_i \) is the \( i \)th element of \( n \) in (3). Note that after equalization, the noise vector \( n \) is no longer white and its covariance matrix depends on the equalizer \( H^H \). It is not difficult to verify that given \( H, n \) is still complex Gaussian distributed with zero mean and covariance matrix \( E[nn^H] = \sigma_w^2H^H(I + H^HCH)^{-1}HCH^H\).

\[
E[nn^H] = \sigma_w^2H^H(I + H^HCH)^{-1}HCH^H\),
\]  

(11)

where \( C := (H^HCH)^{-1} \). Define a variable \( v_i = (-e_i^n_i - e_i^n_i) \). Given the error symbol \( e_i \), \( v_i \) is real Gaussian distributed with zero mean and variance \( 2|e_i|^2E[|n_i|^2] = 2|e_i|^2\sigma_w^2C_{ii} \), where \( C_{ii} \) is the \((i, i)\)th element of \( C \) in (11). Thus, the pairwise error probability in (10) can be rewritten as

\[
P(s_i \rightarrow \hat{s}_i | H) = Q \left( \frac{|e_i|^2}{2\sigma_w^2 C_{ii}} \right),
\]  

(12)

where \( Q(a) = \frac{1}{\sqrt{2\pi}} \int_a^\infty e^{-\frac{x^2}{2}} dx \). Based on (12), the diversity order collected by ZF equalizer is \( N_r - N_t + 1 \) (see Appendix A for a proof).

Interestingly, the ZF equalizer enables the same diversity as the nulling-canceling method for V-BLAST systems [20]. However, this diversity order is much lower than the maximum diversity order \( N_r \) exploited by the ML detector.

Next, let us analyze the diversity order collected by the MMSE equalizer, which usually has better performance than ZF equalizer according to empirical results. For the MMSE equalizer, based on the matrix inversion lemma and (6), we can rewrite the output of the MMSE equalizer as

\[
x_{MMSE} = s + (H^HCH + \sigma_w^2I_{N_t})^{-1}(H^HCH - \sigma_w^2s). \tag{13}
\]

Define the equivalent noise vector \( n = (H^HCH + \sigma_w^2I_{N_t})^{-1}(H^HCH - \sigma_w^2s) \). Given a vector signal \( s \), the mean and covariance matrices of the noise vector \( n \) are

\[
\bar{n} = -\sigma_w^2(H^HCH + \sigma_w^2I_{N_t})^{-1}s,
\]

\[
\Sigma = \sigma_w^2(H^HCH + \sigma_w^2I_{N_t})^{-1} - \sigma_w^2(H^HCH + \sigma_w^2I_{N_t})^{-1}2(14)
\]

Similar to the analysis for the ZF equalizer, the pairwise error probability of the MMSE equalizer is given as

\[
P(s_i \rightarrow \hat{s}_i | H) = Q \left( \frac{|e_i|^2 + e_i^*n_i + e_i n_i^*|^2}{2|e_i|^2\Sigma_{ii}} \right),
\]  

(15)

where \( \bar{n}_i \) is the \( i \)th element of the mean noise \( \bar{n} \) and \( \Sigma_{ii} \) is the \((i, i)\)th element of the noise covariance matrix \( \Sigma \) in (14). At high SNR, i.e., when \( \sigma_w^2 \) is much smaller than \( |e_i|^2 \), Eq. (15) can be approximated as

\[
P(s_i \rightarrow \hat{s}_i | H) \approx Q \left( \frac{|e_i|^2}{2\sigma_w^2 C_{ii}} \right),
\]  

(16)

where \( C_{ii} \) is the \((i, i)\)th element of \((H^HCH + \sigma_w^2I_{N_t})^{-1} \). It is ready to verify that at high SNR, \( C_{ii} \) has the same degrees of freedom as \( C_{ii} \), in (12). This shows that the MMSE detector achieves the same diversity as the ZF detector which is \( N_r - N_t + 1 \).

Remark 1: From our analysis, we notice that for V-BLAST systems, the linear equalizers cannot exploit the full receive diversity, though they have much lower decoding complexity relative to ML or SD detectors. In general, the MMSE equalizer outperforms the ZF equalizer with larger coding advantage because \( C_{ii} \) in (16) is always less than \( C_{ii} \) in (12). Furthermore, for the same SNR, as the number of receive-antennas \( (N_r) \) increases, the performance gap between MMSE and ZF detectors decreases, because \( H^HCH \) will dominate \((H^HCH + \sigma_w^2I_{N_r}) \) and then \( C_{ii} \) becomes closer to \( C_{ii} \). Since the diversity collected by linear equalizers is only \( N_r - N_t + 1 \), there is no spatial diversity when \( N_t = N_r \). In this case, even if we increase the number of antennas, no spatial diversity can be collected by the linear equalizers to improve the performance of MIMO V-BLAST transmissions at the receiver. In the following, we show that using LR-aided linear equalizers, we can restore the maximum diversity order \( N_r \).

IV. LR- AIDED LINEAR EQUALIZERS

Given the system model in (1), if the symbols \( s \) are drawn from Gaussian integer ring \( \mathbb{Z}[j] \) (e.g., QAM, PAM constellations), then \( Hs \) belongs to a lattice spanned by the columns of \( H \) [1]. Hence, to decode \( s \) now becomes to find the nearest point to \( y \) on the lattice spanned by \( H \).

As we know, when the lattice basis \( H \) is orthogonal, i.e., \( H^HCH \) is diagonal, the decision region of linear equalizers to find the nearest point is the same as that of ML detector. Therefore, in this case, ZF equalizer has the same performance as the ML detector. However, in general \( H \) is not orthogonal, and thus linear equalizers induce performance degradation. Clearly, how close \( H \) is to an orthogonal matrix has an effect on the performance of linear equalizers.

To quantify the orthogonality of a matrix, let us introduce the following metric.

**Definition 1:** An orthogonality deficiency \((od)\) of an \( N_r \times N_t \) matrix \( H = [h_1, h_2, \ldots, h_{N_t}] \) is defined as [18]:

\[
od(H) = 1 - \frac{\det(H^HCH)}{\prod_{n=1}^{N_t} ||h_n||^2},
\]  

(17)

where \( h_n \), \( 1 \leq n \leq N_t \), is the \( n \)th column of \( H \). Note that \( 0 \leq \od(H) \leq 1, \forall H \). If \( H \) is singular, \( \od(H) = 1 \), and if the columns of \( H \) are orthogonal, \( \od(H) = 0 \). In general, when \( \od(H) \) is smaller, we say that \( H \) is closer to an orthogonal matrix. Larger \( od \) may cause possibly bigger performance gap between linear equalizers and ML. Thus, if we can find another basis \( \tilde{H} \) in the same lattice which is more orthogonal than \( H \), and use linear equalizers based on \( \tilde{H} \), intuitively the performance should be closer to that of the corresponding ML detector. Now a critical question is how to find a basis that is more orthogonal than \( H \). The process to find a more orthogonal basis is called lattice reduction (LR).

A. LR Using the Complex LLL Algorithm

Theoretically, finding an optimal set of bases in a lattice is computationally expensive [1]. Two widely used LR methods are Korkine-Zolotareff (KZ) reduction [10], [16] and Lenstra-Lenstra-Lovász (LLL) algorithm [11]. The KZ algorithm can find the optimal basis for a lattice, but it is highly complex

\[1^{st} \od \text{has also been defined as } \frac{||h_n\parallel^2}{\det(H)} \text{ in [8], which is equivalent to (17) when } H \text{ is square, but is unbounded.} \]
and thus infeasible for practical problems. The LLL algorithm does not guarantee to find the optimal basis, but it guarantees in polynomial time to find a basis within a factor to the optimal one [11]. Since complexity is one of our major concerns, we adopt the LLL algorithm for LR here.

A reduced basis for a real lattice is defined in [11]. As shown in [11, Proposition 1.26], the worst case of the number of arithmetic operations needed by the LLL algorithm to find a new basis is $O(N^4)$, where $N$ is the size of the basis. Most of existing results in [18], [22], [24] adopt the real LLL (RLLL) algorithm in [11] and use the real LR-aided equalizers. In the following, we provide a detailed complex LLL (CLLL) algorithm which reduces the RLLL’s complexity without sacrificing any performance.

Let us first extend the definition of a reduced basis in [11] to the complex field.

Definition 2: An $M \times N$ complex matrix $\tilde{H}$ is called a reduced basis of a lattice if the QR-decomposition $\tilde{H} = QR$ satisfies the following two conditions:

$$
|\tilde{R}_{i,i} - \frac{1}{2}| \leq |\tilde{R}_{i,i}|, |3\tilde{R}_{i,i}| \leq \frac{1}{2}|\tilde{R}_{i,i}|, \quad \forall i < k,
$$

$$
\delta |\tilde{R}_{i-1,i-1}|^2 \leq |\tilde{R}_{i,i}|^2 + |\tilde{R}_{i-1,i}|^2, \quad \forall i \in [2, N],
$$

where the parameter $\delta$ now is arbitrarily chosen from $(\frac{1}{2}, 1)^2$, and $\tilde{R}_{i,k}$ is the $(i, k)^{th}$ entry of $\tilde{R}$.

The detailed pseudo-code of the CLLL algorithm can be found in Table I. The parameter $\delta$ controls the complexity and performance of the LLL algorithm and the bigger $\delta$ is, the higher the complexity is. Compared with the RLLL algorithm in [23], [24], the major differences of the CLLL algorithm are: (i) at Step (8), the rounding equation is on complex numbers; and (ii) at Step (16), a complex unitary $\Theta$ is adopted. Because the CLLL algorithm does not need to split the channel into real and imaginary parts, it reduces the equivalent channel matrix size. Later by simulations, we show that the CLLL algorithm requires lower computational complexity than the RLLL algorithm while not sacrificing any performance.

Following the CLLL algorithm, we find a “better” channel matrix $\tilde{H} = HT$ from the original channel matrix $H$, where $T$ is a unimodular matrix, which means that all the entries of $T$ and $T^{-1}$ are Gaussian integers and the determinant of $T$ is ±1 or ±0. The following lemma shows the quantitative result on the od of $H$ found by the CLLL algorithm.

Lemma 1: Given a matrix $H \in \mathbb{C}^{M \times N}$ with rank $N$, $\tilde{H}$ is obtained after applying the CLLL algorithm in Table I for a given parameter $\delta \in (\frac{1}{2}, 1)$. Then, the od of $\tilde{H}$ satisfies:

$$
\sqrt{1 - \text{od}(\tilde{H})} \geq 2\delta - \frac{2}{\delta} \geq 2\delta - \frac{N(N+1)}{4} = c_{\delta}.
$$

Proof: See Appendix B.

For real $H$, Lemma 1 is consistent with the result in [11, Proposition 1.8]. Here, we extend it to the complex field according to the CLLL algorithm in Table I. Given $\delta$ and any integer $N \geq 1$, $c_{\delta}$ is always less than 1. Therefore, $\text{od}(\tilde{H})$ is bounded by $1 - c_{\delta}^2$. If $H$ is singular, i.e., rank$(H) < N$, then Lemma 1 does not hold true since $H$ is not a basis any more. In this case, we need to reduce the size of $H$ and then apply the CLLL algorithm. From Lemma 1, we can see that CLLL algorithm does not guarantee to reduce the od for every realization of $H$, but the new basis $\tilde{H}$ now has an upper bound on od which is strictly less than one. In the following, we will show that thanks to this bound on od, LR-aided linear equalizers collect receive diversity.

B. LR-aided Linear Equalizers

With the new channel matrix $\tilde{H}$ generated by the CLLL algorithm, we apply the LR-aided ZF equalizer $\tilde{H}^\dagger$ instead of $H^\dagger$, and the output can be written as [22], [24]:

$$
x = T^{-1}s + \tilde{H}^\dagger w = z + n.
$$

Since all the entries of $T^{-1}$ and the signal constellation belong to Gaussian integer ring, the entries of $z$ are also Gaussian integers. Thus, we can estimate $z$ from $x$ in (20) by rounding up to integers. After obtaining $\hat{z}$, we can recover $s$ by mapping $T\hat{z}$ to the appropriate constellation. We summarize the main steps of the LR-aided ZF equalization for MIMO V-BLAST systems in Table II. Regarding the LR-aided MMSE equalizer, since the MMSE equalizer agrees with the ZF equalizer with respect to the extended system in (8), to perform the LR-aided MMSE equalizer is equivalent to applying the LR-aided ZF equalizer in Table II but changes $ss$ according to the extended system.

C. Performance Analysis of LR-aided Linear Equalizers

The diversity order collected by the LR-aided linear detectors for MIMO V-BLAST systems is established in the following proposition.

<table>
<thead>
<tr>
<th>Table I</th>
</tr>
</thead>
<tbody>
<tr>
<td>The complex LLL algorithm (using MATLAB notation)</td>
</tr>
<tr>
<td><strong>INPUT:</strong> $H$; <strong>OUTPUT:</strong> $Q$, $R$, $T$</td>
</tr>
<tr>
<td>(1) $[Q, R] = QR$ Decomposition($H$);</td>
</tr>
<tr>
<td>(2) $\delta \in (\frac{1}{2}, 1)$;</td>
</tr>
<tr>
<td>(3) $m = \text{size}(H, 2)$;</td>
</tr>
<tr>
<td>(4) $T = I_m$;</td>
</tr>
<tr>
<td>(5) $k = 2$;</td>
</tr>
<tr>
<td>(6) while $\delta \leq m$</td>
</tr>
<tr>
<td>(7) for $n = k - 1 : -1 : 1$</td>
</tr>
<tr>
<td>(8) $u = \text{round}((\tilde{R}(n, k)/\tilde{R}(n, n)))$;</td>
</tr>
<tr>
<td>(9) if $u \sim 0$</td>
</tr>
<tr>
<td>(10) $\tilde{R}(1:n, n) = \tilde{R}(1:n, k) - u \cdot \tilde{R}(1:n, n)$;</td>
</tr>
<tr>
<td>(11) $T(:, k) = T(:, k) - u \cdot T(:, n)$;</td>
</tr>
<tr>
<td>(12) end</td>
</tr>
<tr>
<td>(13) end</td>
</tr>
<tr>
<td>(14) if $\delta</td>
</tr>
<tr>
<td>(15) Swap the (k-1)th and kth columns in $\tilde{R}$ and $T$</td>
</tr>
<tr>
<td>(16) $\Theta = \begin{bmatrix} \alpha^* &amp; \beta^* \ -\beta &amp; \alpha \end{bmatrix}$ where $\alpha = \tilde{R}(k-1,k-1);</td>
</tr>
<tr>
<td>$\beta = \tilde{R}(k,k-1) /</td>
</tr>
<tr>
<td>(17) $\tilde{R}(k-1, k-1) = \Theta \tilde{R}(k-1, k-1)$;</td>
</tr>
<tr>
<td>(18) $\tilde{Q}(:, k - 1 : k) = \tilde{Q}(:, k - 1 : k) \Theta^N$;</td>
</tr>
<tr>
<td>(19) $k = \max(k - 1, 2)$;</td>
</tr>
<tr>
<td>(20) else</td>
</tr>
<tr>
<td>(21) $k = k + 1$;</td>
</tr>
<tr>
<td>(22) end</td>
</tr>
<tr>
<td>(23) end</td>
</tr>
</tbody>
</table>
TABLE II
LATTICE-REDUCTION AIDED ZF EQUALIZATION

| S1: | Map the information bits to symbols $s$ whose constellation belongs to Gaussian integer ring; |
| S2: | Obtain (1) after transceiver operations; |
| S3: | Perform the CLLL algorithm to reduce the lattice basis of the channel matrix: $[\tilde{H}, \tilde{T}] = CLLL(\tilde{H})$; |
| S4: | Rewrite the system as $y = HT(T^{-1}s) + w = \tilde{H}z + w$; |
| S5: | Apply the ZF equalizer based on the new system to obtain $\tilde{e} = Q(\tilde{H}^H y)$; |
| S6: | Use $\tilde{e}$ and $T$ to recover the original information: $\tilde{s} = Q(T\tilde{e})$.

**Proposition 1:** Given the model in (1), if the channels are i.i.d. complex Gaussian distributed, the diversity order collected by an LR-aided linear detector (LR-aided ZF or LR-aided MMSE) is $N_r$, which is the same as that achieved by the ML detector.

**Proof:** The output of the LR-aided ZF equalizer is given in (20). For QAM symbols, the real and imaginary parts of each symbol are drawn from the set $\{-\sqrt{M}/2, \ldots, -1, 1, \ldots, \sqrt{M}/2\}$. Then, by applying $(s - (1 + j)1)/2$, we transfer the real and imaginary parts of the constellation to a consecutive integer set, which makes the real and imaginary parts of $z$ also consecutive integers. Thus, the estimate of $s$ is expressed as (by taking into account Steps S5 and S6 in Table II):

$$\tilde{s} = 2T \left[ \frac{1}{2} (x - T^{-1}(1 + j)1) \right] + (1 + j)1$$

(21)

Apparently, if $\frac{1}{2} \tilde{H}^H w = 0$, $s$ will be decoded correctly. Therefore, the symbol error probability $P_{e|H}$ for a given $H$ is upper-bounded by

$$P_{e|H} \leq 1 - P \left( \frac{1}{2} \tilde{H}^H w \right) = 0 \left| H \right).$$

(22)

Let us denote $\tilde{H}^i$ as $[a_1, a_2, \ldots, a_{N_t}]^T$, where $a^T_i, i \in [1, N_t]$ is the $i$th row of $\tilde{H}$. The upper bound can be written as

$$P_{e|H} \leq P \left( \max_{1 \leq i \leq N_t} |a_i^T w| \geq 1 \left| H \right) \right. \right)$$

(23)

From [18, Lemma 1], we obtain the following inequality:

$$\max_{1 \leq i \leq N_t} |a_i^T w| \leq \frac{1}{\sqrt{1 - od(\tilde{H}) \cdot \min_{1 \leq i \leq N_t} \|\tilde{h}_i\|}},$$

(24)

where $\tilde{h}_i, i \in [1, N_t]$ represents the $i$th column of $\tilde{H}$. Because

$$\max_{1 \leq i \leq N_t} |a_i^T w| \leq \frac{\max_{1 \leq i \leq N_t} \|a_i^T w\|}{\sqrt{1 - od(\tilde{H}) \cdot \min_{1 \leq i \leq N_t} \|\tilde{h}_i\|}}.$$

$P_{e|H}$ is further bounded by

$$P_{e|H} \leq P \left( \frac{|w|}{\sqrt{1 - od(\tilde{H}) \cdot \min_{1 \leq i \leq N_t} \|\tilde{h}_i\|}} \geq 1 \left| H \right) \right. \right)$$

(25)

Furthermore, since $\tilde{H}$ is reduced from $H$ using the CLLL algorithm with parameter $\delta$ and $H$ is full rank with probability one, according to (19) in Lemma 1, we have $\sqrt{1 - od(\tilde{H})} \geq c_\delta$. Let $h_{\min}$ represent the vector with minimum non-zero norm of all the vectors in the lattice generated by $H$. Since $T$ is unimodular, $\tilde{H}$ spans the same lattice as $H$ with an infinite coefficient set. From the definition of $h_{\min}$, we know that $\|h_{\min}\|$ is less than or equal to $\min_{1 \leq i \leq N_t} \|\tilde{h}_i\|$. In summary, we have [c.f. (23) and (25)]

$$P_{e|H} \leq P \left( \max_{1 \leq i \leq N_t} |a_i^T w| \geq 1 \left| H \right) \right. \right) \leq P \left( \frac{|w|}{c_\delta \|h_{\min}\|} \geq 1 \left| H \right) \right. \right).$$

(26)

Thus, by averaging (26) with respect to the random matrix $H$ (or $h_{\min}$), the average symbol error probability can be simplified as:

$$P_e = E_H \left[ P_{e|H} \right] \leq P \left( \frac{|w|}{c_\delta \|h_{\min}\|} \right) \leq P \left( \frac{|w|}{c_\delta} \right) \leq E_w \left[ P \left( \frac{|w|}{c_\delta} \right) \right].$$

(27)

Since $w$ is an $N_r \times 1$ complex white Gaussian noise vector with covariance matrix $\sigma_w^2 I_{N_r}$, $\|w\|^2$ is a central Chi-square random variable with $2N_r$ degrees of freedom and mean $\sigma_w^2$. To simplify the bound in (27), we need the following lemma.

**Lemma 2:** Define a lattice $L$ in $\mathbb{C}^{N_r \times 1}$ generated by a set of bases $\tilde{H} = [h_1, h_2, \ldots, h_{N_t}]$ and a complex integer coefficient set. If all the entries of $H$ are i.i.d. complex Gaussian distributed with zero mean and unit variance, then we have $P \left( \|h_{\min}\|^2 \leq \epsilon \right) \leq c_{N_r}N_t \epsilon^{N_r}$, where $c_{N_r}N_t$ is a finite constant depending on $N_r$ and $N_t$.

**Proof:** See Appendix C.

Consequently, we obtain that the average error probability in (27) is bounded as

$$P_e \leq E_w \left[ P \left( \|h_{\min}\|^2 \leq \frac{|w|^2}{c_\delta^2} \right) \right] \leq E_w \left[ c_{N_r}N_t \left( \frac{1}{c_\delta^2} \right) \|w\|^{2N_r} \right] = c_{N_r}N_t \left( \frac{1}{c_\delta^2} \right) \left( \frac{2N_r - 1}{N_r - 1} \right) \left( \frac{1}{\sigma_w^2} \right)^{-N_r},$$

(28)

where the last equality comes from the $N_r$th moment of Chi-square random variable $|w|^2$, which can be found in [17, p. 14]. Therefore, the diversity order of the LR-aided ZF equalizer is greater than or equal to $N_r$. However, as we know, the maximum diversity order for the MIMO V-BLAST system is $N_r$. Thus, the LR-aided ZF equalizer collects diversity $N_r$. Similarly, for LR-aided MMSE equalizer, we can show that it also collects diversity $N_r$. ■

**Remark 2:** As we have shown in Section III, the linear equalizers can only collect diversity $N_r - N_t + 1$ for V-BLAST transmissions. However, after introducing LR technique into the linear equalization process, the maximum diversity is recovered for V-BLAST systems. This means that the LR technique brings the performance curves of the linear equalizers parallel to that of ML decoder. However, there may still exist a performance gap between LR-aided linear equalizers and
the ML detector because of sub-optimal LR and quantization steps. How to improve the performance and thus reduce this gap is under investigation.

V. SIMULATION RESULTS

In this section, we use computer simulations to verify our theoretical claims on the diversity orders of linear equalizers and LR-aided linear equalizers. The channels are generated as i.i.d. complex Gaussian variables with zero mean and unit variance. The SNR is defined as symbol energy per transmit-antenna versus noise power spectral density.

Example 1 (Diversity collected by linear equalizers): The ZF and MMSE linear equalizers in (3) and (6) are considered for MIMO V-BLAST systems with \( N_t = 2 \) transmit-antennas, and different numbers of receive-antennas \( N_r = 2, 3, 4 \). BPSK is adopted as the modulation scheme. Bit-error rate (BER) versus SNR is depicted in Figure 1. Reading the slopes of the curves in Figure 1, we observe that the diversity orders collected by either ZF or MMSE equalizer are indeed \( N_r - N_t + 1 \), which in this example are 1, 2, and 3, respectively. Furthermore, we notice that the performance of MMSE detection is better than ZF detection and the gap between them decreases as \( N_t \) increases, which is consistent with our claims in Remark 1.

Example 2 (LR-aided linear equalizers): In this example, we fix the numbers of transmit- and receive-antennas as \( (N_t, N_r) = (3, 4) \), and use QPSK as the modulation scheme. Five detection methods are applied to the MIMO system: ZF, MMSE, complex LR-aided ZF, complex LR-aided MMSE and ML detection. Observing Figure 2, we notice that the linear detectors can only achieve diversity order \( N_r - N_t + 1 \), which is 2 in this case. The ML detector enables diversity order \( N_r = 4 \). As expected, the LR-aided linear detectors achieve the same diversity order as the ML does. The performance gap between LR-aided linear detectors and the ML detector is mainly because the CLLL algorithm does not guarantee to find the most orthogonal basis of the lattice spanned by the channel matrix and the quantization step of \( z \) in Table II does not take into account matrix \( T \).

Example 3 (Complexity of the CLLL algorithm): In this example, we compare the CLLL algorithm with the RLLL algorithm in terms of complexity and performance. The arithmetic operations we count are real additions and real multiplications while one complex addition is counted as two real additions and one complex multiplication is counted as four real multiplications and two real additions. In Figure 3, we plot the average numbers of arithmetic operations needed by the CLLL and RLLL algorithms respectively as \( N_r = N_t = n \) increases. The number of arithmetic operations that the RLLL algorithm needs is about 2 times of that of the CLLL algorithm. Therefore, the CLLL algorithm is more computationally efficient. One may doubt the complexity overhead of the RLLL algorithm, since it only adopts real operations though the matrix size is doubled. We plot the average number of basis update iterations (steps (15)-(19) in Table I) for the RLLL and CLLL algorithms in Figure 4. From the figure, we can see the average number of basis updates for the RLLL is more than three times of that for the CLLL algorithm, which also indicates that the RLLL algorithm requires more arithmetic operations than the CLLL algorithm. Different from some near-ML methods (e.g. SD), the complexity of the LR-aided ZF equalizer does not depend on SNR.

In Figure 5, we compare the complexity of the complex LR-aided ZF with the one of the real LR-aided ZF equalizer. In the same figure, we also plot the complexity of ZF equalizer as a reference. From this figure, we notice that the complexity of the complex LR-aided ZF equalizer is about 1.7 times of that of ZF equalizer, which is really low provided the significant performance improvement by the LR-aided ZF equalizer. Note that the complexity increment of the LR-aided methods has two major parts: one is the CLLL step and the other is the inverse of \( T \). In Figure 6, we compare the performance of the complex LR-aided ZF and the real LR-aided ZF equalizers. It can be seen that complex LR-aided ZF equalizer has the same
Example 4 (Performance of coded V-BLAST systems): In this example, we illustrate the performance for coded VBLAST systems with $N_t = 4$ transmit-antennas and $N_r = 4$ receive-antennas. QPSK modulation is used with a rate-1/2 convolutional code (CC) with generator $(6, 7)$. Each frame contains a total of 120 information bits. For simplicity, we adopt the hard decoder to decode the CC and apply five different detectors: ZF, MMSE, LR-aided ZF, LR-aided MMSE and ML detectors. Here we consider quasi-static channels where the channels do not change over a certain block period (e.g., 30 time-slots in this example) but independently vary from block to block. Random interleaver is used within one block. In Figure 7, we compare coded and uncoded cases with five different equalizers. We observe that: i) coding improves the performance of the LR-aided linear equalizers; ii) the performance improvement by coding for the LR-aided methods or ML is almost the same; and iii) there exists only about 2dB gap between the performance of the LR-aided MMSE detector and the one of ML detector, but the LR-aided MMSE equalizer has much lower complexity. Here the iterative/soft decoding and more general interleaving are out of the scope of the paper.

VI. CONCLUSIONS

In this paper, we analyze the performance of linear detectors for MIMO V-BLAST systems. We show that conventional linear equalizers can only collect diversity $N_r - N_t + 1$ for MIMO V-BLAST systems though they have very low complexity. However, by slightly increasing the decoding complexity, LR-aided linear equalizers are shown to achieve diversity $N_r$, which is the same as that collected by ML detector. The complexity of the CLLL algorithm is lower than that of the RLLL algorithm while they have the same
Based on the matrix inversion lemma, we know that the zero eigenvalues as those of $P$. Simulations have verified our theoretic claims.\textsuperscript{3}

\section*{Appendix A: Diversity of ZF Equalizer}

In (11), we have defined that $C = (H^H H)^{-1}$. By properly permutating $H$ to $[h_i, H_i]$, where $h_i$ is the $i$th column of $H$ and $H_i$ denotes the rest columns, we obtain that

$$ C = (H^H H)^{-1} = P \begin{bmatrix} h_i^H h_i & h_i^H H_i \\ H_i^H h_i & H_i^H H_i \end{bmatrix}^{-1} P^T, $$

where $P$ is a permutation matrix such that $PH = [h_i, H_i]$. Based on the matrix inversion lemma, we know that the $(i, i)$th element of $C$ is

$$ C_{ii} = \left(h_i^H h_i - h_i^H H_i \left(H_i^H H_i\right)^{-1} H_i^H h_i \right)^{-1}. \quad (29) $$

Suppose that the singular value decomposition (SVD) of $H_i$ is $H_i = V D U^H_i$, where $V$ is an $N_r \times (N_t - 1)$ matrix with $V V^H = I$, $D$ is an $(N_t - 1) \times (N_t - 1)$ diagonal matrix and $U$ is an $(N_t - 1) \times (N_t - 1)$ unitary matrix. Plugging this SVD result into (29), we are ready to verify that

$$ H_i (H_i^H H_i)^{-1} H_i^H = V V^H. $$

It is straightforward to see that the rank of $VV^H$ is $N_t - 1$. More specifically, $VV^H$ is an $N_r \times N_r$ matrix whose eigenvalue decomposition can be written as

$$ VV^H = \tilde{V} \begin{bmatrix} I_{N_t-1} & 0_{(N_r-N_t+1) \times (N_t-1)} \\ 0_{(N_r-N_t+1) \times (N_t-1)} & 0_{(N_r-N_t+1) \times (N_r-N_t+1)} \end{bmatrix} \tilde{V}^H, $$

where $\tilde{V}$ is an $N_r \times N_r$ unitary matrix with the first $N_t - 1$ columns same as $V$. This is because $VV^H$ has the same non-zero eigenvalues as those of $V^H V$, and we know $V^H V = I_{N_t-1}$. Therefore, we can verify

$$ C_{ii}^{-1} = h_i^T (I_{N_r} - V V^H) h_i = h_i^T \tilde{V} \begin{bmatrix} 0_{(N_t-1) \times (N_t-1)} & 0_{(N_t-1) \times (N_r-N_t+1)} \\ 0_{(N_r-N_t+1) \times (N_t-1)} & I_{N_r-N_t+1} \end{bmatrix} \tilde{V}^H h_i. $$

Since the $N_r$ entries of $h_i$ are i.i.d. complex Gaussian random variable, according to [21, Section 2.1.5], $h_i$ is bi-unitarily invariant, which means for any deterministic unitary matrix $\Phi$, $h_i$ has the same distribution as $\Phi h_i$. Here, $\tilde{V}^H$ is a random unitary matrix and independent on $h_i$. Let $\tilde{h}_i = \tilde{V}^H h_i$. With the help of the total probability theorem in [15, p. 224], we obtain that

$$ f(\tilde{h}_i) = \int f(\tilde{h}_i | \tilde{V}^H) f(\tilde{V}^H) d\tilde{V}^H = \int f(h_i | \tilde{V}^H) f(\tilde{V}^H) d\tilde{V}^H = f(h_i), \quad (30) $$

where $f(a)$ denotes the pdf of $a$. Eq. (30) means $\tilde{V}^H h_i$ still has the same distribution as $h_i$. Thus, it can be seen that $C_{ii}^{-1}$ is a chi-squared random variable with $2(N_r - N_t + 1)$ degrees of freedom, because the channel coefficients are complex Gaussian distributed. As shown in [25], if we integrate the right hand side of (12) with respect to such a random variable (i.e., $C_{ii}^{-1}$), we obtain that the diversity order is equal to $N_r - N_t + 1$. This means that the diversity order collected by the ZF detector is $N_r - N_t + 1$. \hfill $\blacksquare$

\section*{Appendix B: Proof of Lemma 1}

According to the CLLL reduction criterion in (18), for a reduced basis $\tilde{H}$, we have

$$ |\tilde{R}_{i,j}|^2 \geq \delta |\tilde{R}_{i-1,j-1}|^2 - |\tilde{R}_{i-1,j}|^2 \geq \left(\delta - \frac{1}{2}\right) |\tilde{R}_{i-1,j-1}|^2, $$

which can be generalized as $|\tilde{R}_{i,k}|^2 \leq \left(\frac{1}{2} - \frac{1}{2}\right) |\tilde{R}_{i,k}|^2$, for $1 \leq i < k \leq N_r$. Furthermore, we have

$$ |\tilde{R}_{k,k}|^2 = \sum_{i=1}^{k-1} |\tilde{R}_{i,k}|^2 \leq \left(\frac{1}{2} - \frac{1}{2}\right) |\tilde{R}_{k,k}|^2, \quad (31) $$

where $\tilde{R}_{k,k}$ is the $k$th column of $\tilde{R}$. Defining $\xi = \frac{2}{2\delta - 1}$, since $\delta \in (\frac{1}{2}, 1)$, $\xi \in (2, \infty)$. Eq. (31) can be simplified as

$$ ||\tilde{R}_{k,k}||^2 \leq \left(\frac{1}{2} + \frac{1}{2\xi} \right) |\tilde{R}_{k,k}|^2 \leq \frac{1}{2} |\tilde{R}_{k,k}|^2. \quad (32) $$

Thus, for the reduced basis $\tilde{H}$, the orthogonality deficiency $od(\tilde{H})$ satisfies

$$ od(\tilde{H}) = 1 - \frac{\det(\tilde{H}^H \tilde{H})}{\prod_{k=1}^{N_r} |\tilde{R}_{k,k}|^2} = 1 - \prod_{k=1}^{N_r} |\tilde{R}_{k,k}|^2 \prod_{k=1}^{N_r} ||\tilde{R}_{k,k}||^2 \leq 1 - \frac{\prod_{k=1}^{N_r} |\tilde{R}_{k,k}|^2}{\prod_{k=1}^{N_r} \left(\frac{1}{2} \xi^k \right) |\tilde{R}_{k,k}|^2} = 1 - 2^N \xi^{-N(N+1)/2} = 1 - 2^N \left(\frac{2}{2\delta - 1}\right)^{-N(N+1)/2}. \quad (33) $$
Therefore, we have
\[ \sqrt{1 - \od(\tilde{H})} \geq 2 \delta \left( \frac{2}{2\delta - 1} \right)^{\frac{N(N+1)}{2}} := c_\delta. \]
Similar proof for RLLA can be found in [11].

APPENDIX C: PROOF OF LEMMA 2

Let \( p_a = H a \) be an \( N_r \times 1 \) vector in the lattice \( \mathbb{L} \) spanned by \( H \) with \( a \) being a non-zero \( N_t \times 1 \) column vector with all entries belonging to the complex integer coefficient set. Based on the definition of \( h_{\min} \), we know that \( \|h_{\min}\|^2 = \arg \min_{p_a \in \mathbb{L}, \|p_a\|^2 = \epsilon} \|p_a\|^2 \). Furthermore, since all the entries of \( p_a \) are i.i.d. complex Gaussian random variables with zero mean and unit variance, we know each entry of \( p_a \) is also complex Gaussian distributed with zero mean and variance \( \|a\|^2 \). Then, it is ready to see that \( 2\|p_a\|^2/\|a\|^2 \) is central Chi-square distributed with \( 2N_r \) degrees of freedom. Thus, we can get an upper bound for the probability that \( \|p_a\|^2 \) is less than \( \epsilon \) as [9, p. 25]:

\[
P(\|p_a\|^2 \leq \epsilon) = 1 - e^{-\|a\|^2} \sum_{k=0}^{N_r-1} \frac{\epsilon^k}{k!} \leq \left( \frac{1}{\|a\|^2} \right)^{N_r} e^{\epsilon \|a\|^2}.
\]
Let us represent those different \( \|p_a\|^2 \) as \( \Theta_n, n \in [1, \infty) \), and \( \|h_{\min}\|^2 \) as \( \Theta_{\min} \). Thus, the cumulative density function (cdf) of \( \Theta_{\min} \) is

\[
P(\theta_{\min} < v) = 1 - P(\theta_{\min} \geq v) = 1 - \lim_{M \to \infty} \int_{v}^{\infty} \int_{v}^{\infty} \cdots \cdots \int_{v}^{\infty} f(\theta_1, \theta_2, \ldots, \theta_M) d\theta_1 d\theta_2 \cdots d\theta_M. \tag{35}
\]
The pdf of \( \Theta_{\min} \) is obtained by taking the derivative of the cdf in (35)

\[
f(v) = \lim_{M \to \infty} \sum_{n=1}^{M} \int_{v}^{\infty} \cdots \cdots \int_{v}^{\infty} f(\theta_1, \theta_2, \ldots, \theta_M) d\theta_1 d\theta_2 \cdots d\theta_M \leq \sum_{n=1}^{\infty} f_{\Theta_{n}}(v), \tag{36}
\]
where \( f_{\Theta_{n}}(v) \) is the pdf of \( \Theta_{n} \). Then, with the inequality in (34), we have

\[
P(\|h_{\min}\|^2 \leq \epsilon) \leq \int_{0}^{\epsilon} \sum_{n=1}^{\infty} f_{\Theta_{n}}(v) dv \leq \sum_{m=1}^{\infty} \sqrt{\frac{1}{\|a\|^2}} \left( \frac{1}{\|a\|^2} \right)^{N_r} N_r e^{\epsilon \|a\|^2}. \tag{37}
\]
Furthermore, since \( \|a\|^2 = m \) is a hyper-sphere in the \( 2N_t \)-dimension space with radius \( \sqrt{m} \), the number of integer vectors \( a \) satisfying \( \|a\|^2 = m \) is upper bounded by the surface area of the hyper-sphere. Thus, (37) can be further upper bounded as

\[
P(\|h_{\min}\|^2 \leq \epsilon) \leq \sum_{m=1}^{\infty} \frac{2\pi m N_r}{(N_t - 1)!} \left( \frac{1}{m} \right)^{N_r} \left( \frac{1}{\|a\|^2} \right)^{N_r} e^{\epsilon \|a\|^2}.
\]

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their helpful comments which improved the quality of this paper.

REFERENCES

Xiaoli Ma (M’03) received the B.S. degree in Automatic Control from Tsinghua University, Beijing, China in 1998, the M.S. degree in Electrical Engineering from the University of Virginia in 2000, and the Ph.D. degree in Electrical Engineering from the University of Minnesota in 2003. From 2003 to 2005, she was an assistant professor of Electrical and Computer Engineering at Auburn University. Since 2006, she has been with the School of Electrical and Computer Engineering at Georgia Institute of Technology. Her research interests include transceiver designs and diversity techniques for wireless fading channels, channel modeling, estimation and equalization, carrier frequency synchronization for OFDM systems, routing and cooperative designs for wireless networks. Dr. Ma serves as an Associate Editor for IEEE Signal Processing Letters, and is a member of the Signal Processing Theory and Methods Technical Committee of the IEEE Signal Processing Society.

Wei Zhang (S’05) received the B.S. degree in electrical engineering from Zhejiang University, Hangzhou, P. R. China, in 2004 and the M.Sc. degree in electrical engineering from Auburn University, Auburn, AL, in 2006. He is now working towards the Ph.D. degree in the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA. His current research interests lie in the area of signal processing for wireless communications, including signal estimation algorithms and coding theory.