Multiplexing Two Information Sources over Fading Channels: A Cross-layer Design Perspective

Zhiyu Yang and Lang Tong[†]

Abstract

We consider the transmission over an unknown frequency-selective channel of two independent sources with different application-layer characteristics: one source (such as voice) has a low information rate with a strict delay constraint; the other (such as data) has a high rate but without any delay constraints. We proposed a system structure that jointly considers the different decoding requirements of the application layer and the unknown fading nature of the physical channel. In the proposed communication system, pilot symbols are not present and the low-rate information is decoded non-coherently first. The decoded low-rate codewords are then used for channel estimation to facilitate coherent decoding of the high-rate source. For a fixed detection error probability of the low-rate source, we derive achievable rate expressions for the high-rate source. We demonstrate a convergence behavior of the achievable rate of the high-rate source as the decision error probability of the low-rate source goes to zero. Numerical results show that the achievable rate of the high-rate source converges to that achievable by a training-based scheme at moderate decision error levels.

Index Terms

Achievable rates, cross-layer design, decision-directed receivers, fading channels, mixed-rate transmission, orthogonal-frequency-division-multiplexing (OFDM).

[†]Corresponding author

Z. Yang and L. Tong are with the School of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14853, USA (email: {zy26, ltong}@ece.cornell.edu).

I. INTRODUCTION

A. Motivation

We consider the transmission of a high-rate source mixed with a delay-sensitive low-rate source over an unknown frequency-selective wireless channel. One example is transmitting low-rate voice and high-rate data over the same link. Another is multiplexing protocol information along with high-rate data. In digital video broadcasting [1], for example, control information such as frame sequence numbers and modulation schemes is transmitted with video data.

We assume that there is a delay constraint imposed on the low-rate source, such as in the case of voice and other real-time traffic. Therefore the codeword length for the low-rate source cannot be made arbitrarily long. The low-rate messages are allowed to be decoded with a certain error probability. For the high-rate source such as data traffic, on the other hand, our goal is to provide the highest rate at which reliable communication is possible. Therefore, we allow its codeword length to be sufficiently long to satisfy any predetermined error probability.

We assume a time-varying ergodic block fading channel, unknown at both transmitting and receiving ends. The optimal approach in the information theoretic setting is to encode the two sources jointly (across fading blocks) and to perform non-coherent decoding that fully exploits the channel statistics. Such an approach, despite its theoretical significance, provides little guidance in practical implementation.

A layered, standard practical approach is to include pilot symbols in data packets. The channel acquisition in the physical layer is made using the pilot symbols (and possibly data symbols as well) within each block and the decoder performs coherent decoding using the estimated channel. The demultiplexing of the two sources is carried out in a higher layer. Such a layered approach, despite its simplicity, requires very high signal-to-noise ratio (SNR) in the physical layer in order to satisfy simultaneously the requirements of the application layer: the stringent delay constraint of the low-rate source and the low decoding error requirement of the high-rate source. Such an approach treats the two types of sources equally without taking into account the different decoding requirements set by the application layer. Furthermore, the use of pilot symbols reduces the number of channel uses available for data transmission and is in general suboptimal.

Can one exploit the decoding requirements from the application layer for a better physical

channel utilization? Can pilot symbols be removed to save the bandwidth? One possible approach is to use blind channel estimation techniques that acquire the common channel without requiring pilot symbols. However, the analysis of blind channel estimation coupled with coherent detection is intractable, and therefore, its performance unknown. In this paper, from a cross-layer perspective, we propose an approach that explicitly considers the decoding constraints from the application layer for a better physical layer efficiency.

B. Summary of Results

We propose a system structure in which the two sources are encoded independently in the physical layer. We use a sequential decision-directed receiver where a non-coherent detection is first used to decode the low-rate source. The decoded codewords are then used as pilots to estimate the high-rate channel response. The idea is to exploit the high redundancy in the coded low-rate source to cope with the channel uncertainty. Once the codewords of the low-rate source are obtained, they are almost as good as training for obtaining accurate channel estimates for the coherent detection of the high-rate source. Such an approach, jointly considering the application layer requirements and the physical channel uncertainty, requires lower SNR and has a higher channel utilization than the layered approach.

The problem would be trivial, of course, if we assumed that the low-rate source were always decoded correctly. In this paper, we make an explicit assumption that decoding errors may be present for the low-rate source and focus on the achievable rate for the high-rate source. We demonstrate that when the low-rate source has an error rate below 10^{-3}, which is a moderate error probability requirement for voice applications, the performance of the system for the high-rate source is close to that when the low-rate source is substituted by pilot symbols.

We first provide a lower bound on the achievable rate for the high-rate source assuming that the decoded low-rate source (with errors) provides partial channel state information. This lower bound generalizes a channel capacity bounding technique first popularized by Medard [2]. Next, we provide an explicit evaluation of the lower bound when the low-rate source uses orthogonal signaling—a natural choice for non-coherent detection. We then consider specific low-rate decoders and evaluate the performance of the high-rate source under different decoding error probabilities of the low-rate source. Numerical examples are included to illustrate the effect of the decision error on the achievable rate for the high-rate source.

C. Related Work

In the framework of transmitting mixed-rate sources as considered in this paper, one application is to model explicitly the protocol information (such as headers and other control information) as the low-rate source. The idea of considering protocol information in transmission was first considered by Gallager [3], where the amount of protocol information needed aside from data to reconstruct random arrival sources is studied from an information theoretical perspective. A practical solution that explicitly uses the low-rate source for channel estimation and equalization was considered in [4], where it was assumed that the low-rate source was decoded without error. In this work, the low-rate decoding error is taken into consideration in the analysis.

The fading channel capacity has been a popular research topic. A commonly used block fading channel model is given by

$$\mathbf{Y}_{i\times j} = \mathbf{H}_{i\times k}\mathbf{X}_{k\times j} + \mathbf{W}_{i\times j}$$

where the subscripts denote the dimension of the matrices, and $\mathbf{Y}_{i\times j}$ is the reception matrix, $\mathbf{H}_{i\times k}$ the channel matrix, $\mathbf{X}_{k\times j}$ the input matrix, $\mathbf{W}_{i\times j}$ the noise matrix. $\mathbf{H}_{i\times k}$ and $\mathbf{W}_{i\times j}$ are ergodic and independent, and independent of the input $\mathbf{X}_{k\times j}$. In the multiple-input-multiple-output (MIMO) flat fading case (see, *e.g.*, [5]), i is the number of receivers, j the coherence time during which the channel stays constant, and k the number of transmitters. In the single-input-single-output (SISO) frequency selective case (see, *e.g.*, [6]), $j = 1$, i is the receive vector length, and k the transmit vector length. In the literature, extensive research has been done on the fading channel capacity for different assumptions on the knowledge of the fading channel $\mathbf{H}_{i\times k}$. Telatar obtained a closed-form expression for the capacity of the Gaussian channel with Rayleigh fading assuming prefect fading knowledge at the receiver [7]. For the realization of $\mathbf{H}_{i\times k}$ unknown at the receiver, [8] and [9] address the properties of the capacity achieving input distribution. However, there are no closed-form expressions for the channel capacity.

Between the two extreme assumptions on the knowledge of the channel state, other work assumes that the receiver has partial information on the channel response, either provided by imperfect channel estimation and tracking or by some genie. The capacity metric with partial side information is difficult to obtain analytically. As a result, lower bounds on the channel capacity (achievable rates) are often used as the performance measure of a system with partial side information. Most of the work assumes Gaussian codebooks and an ML

decoder to compute an achievable rate (*e.g.*, see [5], [6], [10]–[18]). In their analysis, it is assumed that

$$E[\mathbf{H}_{i\times k}\big|\,\hat{\mathbf{H}}_{i\times k}] = \hat{\mathbf{H}}_{i\times k} \tag{1}$$

where $\hat{\mathbf{H}}_{i\times k}$ is the channel estimate (side information). The above assumption is valid if the MMSE channel estimator is used, which has a simple form if the channel is estimated based only on the outputs due to known pilot symbols. In this work, we propose a channel estimation structure that utilizes the decoded low-rate messages as pilots for channel estimation. The resulted channel estimate does not satisfy (1). To obtain an achievable rate expression, one can apply [19, Theorem 4.0.2], which is more general and does not require the side information to satisfy (1). In this work, we obtain an achievable rate expression that is larger than or equal to [19, Theorem 4.0.2] and use it to evaluate the performance of the decision-directed receiver structure.

D. Notations

For convenience, denote the probability density function (p.d.f.) of a circularly symmetric complex Gaussian vector $\mathbf{z}$ of length L with mean $\mathbf{a}$ and covariance matrix $\mathbf{Q}$ by

$$\mathcal{N}_{\mathbf{z}}\big(\mathbf{a}, \mathbf{Q}\big) \triangleq \frac{1}{\pi^L|\mathbf{Q}|}e^{-(\mathbf{z}-\mathbf{a})^H\mathbf{Q}^{-1}(\mathbf{z}-\mathbf{a})}.$$

When there is no ambiguity, we use $\mathcal{N}\big(\mathbf{a}, \mathbf{Q}\big)$ to denote $\mathcal{N}_{\mathbf{z}}\big(\mathbf{a}, \mathbf{Q}\big)$. Let $\mathrm{diag}(\mathbf{z})$ be the diagonal matrix with the diagonal entries equal to the entries of vector $\mathbf{z}$. Let $(\cdot)^*$ and $(\cdot)^H$ denote conjugate and conjugate transpose, respectively.

The paper is organized as follows. The system model is introduced in Section II. In Section III, we derive a lower bound on the mutual information with Gaussian codebooks and side information and apply the lower bound to obtain an achievable rate expression for the high-rate source by the decision-directed structure. Under the assumptions of binary low-rate orthogonal signaling, the computational complexity of the achievable rate is further reduced to that of a real double integral. Numerical results are presented in Section IV and the paper is concluded in Section V.

II. SYSTEM MODEL

A. Channel Model

We assume an ISI channel of length m and it experiences identical independent distributed (i.i.d.) Rayleigh block fading with coherence interval T_c. The channel response, unknown at

either the transmitter or the receiver, remains constant during one coherence interval T_c and changes to an independent value in the next interval. Denote $\mathbf{h}^k = [h_1^k, h_2^k, \ldots, h_m^k]^T$ the channel response at the k-th coherence interval. The channel responses $\mathbf{h}^1, \mathbf{h}^2, \ldots$ are i.i.d. $\mathcal{N}(\mathbf{0}, \frac{1}{m}\mathbf{I})$.

We assume that the system uses orthogonal-frequency-division-multiplexing (OFDM) modulation with inverse-discrete-Fourier-transform (IDFT) block size T and cyclic prefix length $T_c - T$. Further assume $T_c - T \geq m - 1$. The channel is then converted into T parallel flat fading channels:

$$\underbrace{\begin{pmatrix} y_1^k \\ y_2^k \\ \vdots \\ y_T^k \end{pmatrix}}_{\mathbf{y}^k} = \underbrace{\begin{pmatrix} g_1^k & 0 & \cdots & 0 \\ 0 & g_2^k & & \\ \vdots & & \ddots & \\ 0 & & & g_T^k \end{pmatrix}}_{\mathbf{G}^k} \underbrace{\begin{pmatrix} x_1^k \\ x_2^k \\ \vdots \\ x_T^k \end{pmatrix}}_{\mathbf{x}^k} + \mathbf{w}^k$$

where the superscript k represents the OFDM block index. $\mathbf{x}^k$ and $\mathbf{y}^k$ are the input and output vectors in the k-th block, respectively. We assume a power constraint on each input position of $\mathbf{x}^k$: $E\big[|x_i^k|^2\big] \leq P$. The white additive noise $\mathbf{w}^k$ is i.i.d. $\mathcal{N}(\mathbf{0}, \mathbf{I})$ across blocks. The i-th diagonalized channel response g_i^k is the i-th Fourier coefficient of the T-point discrete-Fourier-transform (DFT) of $\mathbf{h}^k$, *i.e.*, $\big[g_1^k, \ldots, g_T^k\big]^T = \mathbf{F}\mathbf{h}^k$, where $\mathbf{F}$ is the truncated DFT matrix of size $T \times m$ with

$$[\mathbf{F}]_{il} = \exp \frac{-j2\pi(i-1)(l-1)}{T}.$$

B. System Structure

We assume that there are two independent information sources to be communicated over the unknown frequency-selective fading channel. One source has a low information rate R_l and an encoding delay constraint of N blocks, the other a high information rate R_h and no delay constraints. The low-rate source with the delay constraint may be voice or protocol information that is delay-sensitive. The high rate source without delay constraints may be delay-insensitive data.

Since the two sources have different delay requirements, we assume that the two sources are encoded separately and transmitted over dedicated tones (Fig. 1). Suppose that T_l tones are allocated for the low-rate transmission and T_h tones for the high-rate transmission. Denote $\mathbf{\Pi}_l \in \{0,1\}^{T_l \times T}$ and $\mathbf{\Pi}_h \in \{0,1\}^{T_h \times T}$ the tone selection matrices for the low-rate and high-rate transmissions respectively, where a tone selection matrix is part of a permutation matrix

that connects each input position to a distinct tone. Thus we have $\mathbf{x}_l^k = \mathbf{\Pi}_l \mathbf{x}^k$ and $\mathbf{x}_h^k = \mathbf{\Pi}_h \mathbf{x}^k$ as the channel inputs for the two sources respectively. And the channel is given by

$$\mathbf{y}_l^k = \mathbf{G}_l^k \mathbf{x}_l^k + \mathbf{w}_l^k \tag{2}$$

$$\mathbf{y}_h^k = \mathbf{G}_h^k \mathbf{x}_h^k + \mathbf{w}_h^k \tag{3}$$

where $\mathbf{G}_l^k = \mathrm{diag}(\mathbf{\Pi}_l \mathbf{F} \mathbf{h}^k)$ and $\mathbf{G}_h^k = \mathrm{diag}(\mathbf{\Pi}_h \mathbf{F} \mathbf{h}^k)$.
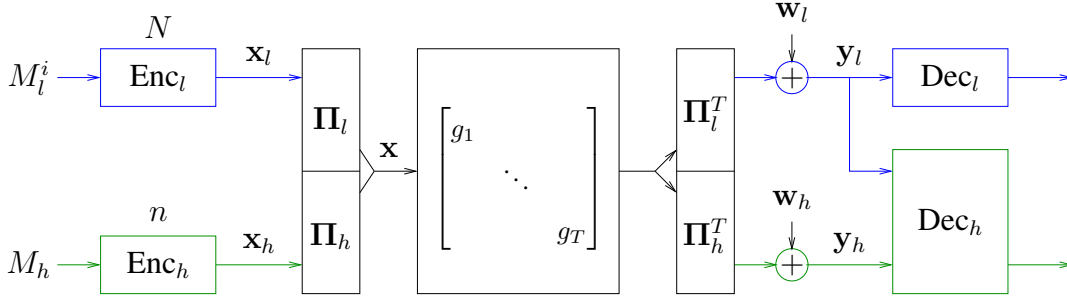

Fig. 1. The system structure.

Because of the different delay requirements, the information bit flows from the two sources are collected as messages and encoded into codewords based on different time scales. Thus, multiple low-rate messages are sent during the transmission of one high-rate message, which, of course, has much more information bits packed due to a higher rate and a longer duration. Suppose that the low-rate and high-rate codewords have block lengths N and n, respectively. Denote $M_h \in \{1, \ldots, 2^{nR_h}\}$ the high-rate message and $M_l^i \in \{1, \ldots, 2^{NR_l}\}$, $1 \le i \le \lfloor \frac{n}{N} \rfloor$, the i-th low-rate message during the transmission of M_h. Let $\{\mathbf{z}\}_i^j$ denote $(\mathbf{z}^i, \ldots, \mathbf{z}^j)$. Then M_l^i is encoded into codeword $\{\mathbf{x}_l\}_{1+(i-1)N}^{iN}$ and M_h is encoded into codeword $\{\mathbf{x}_h\}_1^n$ (Fig. 2). The low-rate symbols $\mathbf{x}_l$'s and the high-rate symbols $\mathbf{x}_h$'s are multiplexed through the tone selection matrices $\mathbf{\Pi}_l$ and $\mathbf{\Pi}_h$, obtaining the input sequence $\{\mathbf{x}\}_1^n$ to the OFDM modulation.

On the receiver side, due to the delay constraint, we assume that M_l^i is decoded based on $\{\mathbf{y}_l\}_{1+(i-1)N}^{iN}$ only. Without any delay constraints, the optimal high-rate receiver decodes M_h based on both $\{\mathbf{y}_h\}_1^n$ and $\{\mathbf{y}_l\}_1^n$ with a maximum likelihood (ML) decoder.

When the low-rate transmission is predetermined, *i.e.*, known pilot symbols are transmitted, Fig. 1 reduces to a training-based scheme.

C. Decision-directed Receiver

Due to the complexity of the optimal high-rate receiver, in this paper, we consider a receiver structure with a decision-directed channel estimator (Fig. 3). For simplicity, we assume that
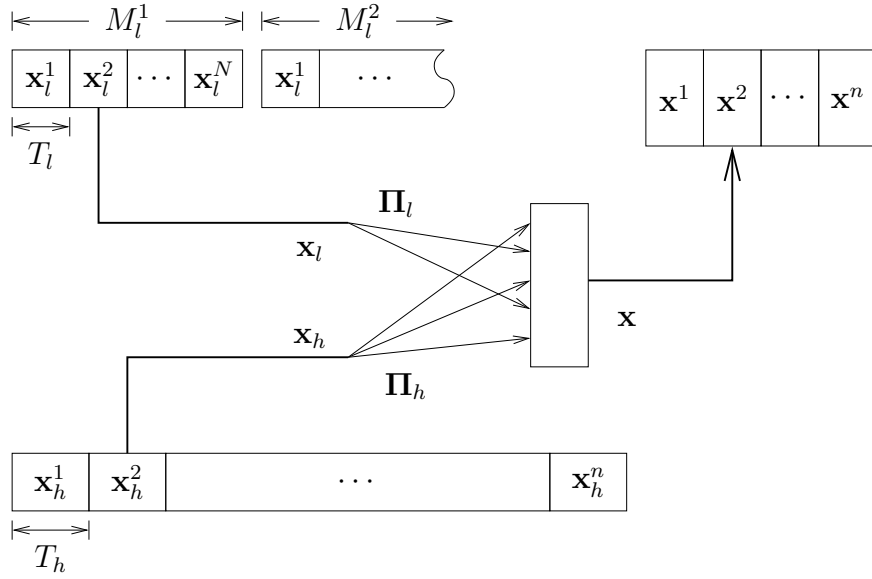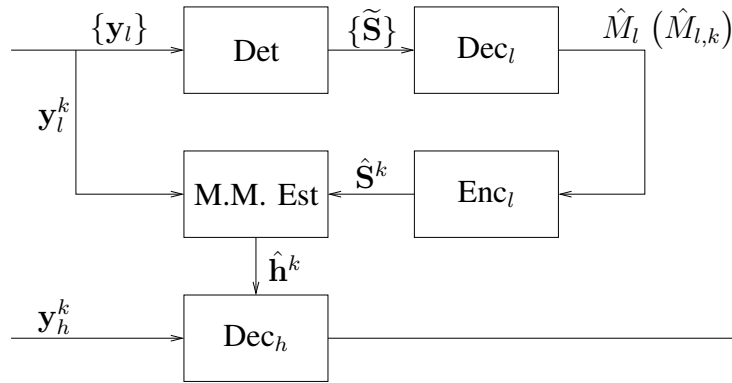
Fig. 2.   The encoding structure.

the low-rate encoder uses binary constant modulus modulation, *i.e.,* $\mathbf{x}_l^k \in \{\mathbf{u}_1, \mathbf{u}_2\}$ where $[\mathbf{u}_{i,j}] = 1, 2$ and $j = 1, \ldots, T_l$. Rewrite the low-rate channel (2) as

$$\mathbf{y}_l^k = \underbrace{\mathrm{diag}(\mathbf{x}_l^k)\mathbf{\Pi}_l \mathbf{F}}_{\mathbf{S}^k}\, \mathbf{h}^k + \mathbf{w}_l^k. \tag{4}$$

The symbol vector $\mathbf{x}_l^k \in \{\mathbf{u}_1, \mathbf{u}_2\}$ has a one-to-one correspondence with the symbol matrix $\mathbf{S}^k \in \{\mathbf{U}_1, \mathbf{U}_2\}$, where $\mathbf{U}_i = \mathrm{diag}(\mathbf{u}_i)\mathbf{\Pi}_l\mathbf{F}$ for $i = 1, 2$.

Fig. 3.   The decision-directed high-rate receiver. In the modified version, $\hat{M}_{l,k}$ is decoded based on $\{\widetilde{\mathbf{S}}\}_{-}^k$ only.

As shown in Fig. 3, the low-rate receiver first performs the symbol ML detection, and then decodes the low-rate messages. The decoded low-rate messages are re-encoded to facilitate the channel estimation, and the channel estimates are used to assist the high-rate decoding.

More specifically, the j-th symbol matrix $\mathbf{S}^j$, $1 \leq j \leq n$, is detected by

$$\widetilde{\mathbf{S}}^j = \arg\max_{\mathbf{S} \in \{\mathbf{U}_1, \mathbf{U}_2\}} p_j(\mathbf{y}_l \,\big|\, \mathbf{S}) \tag{5}$$

where $p_j(\mathbf{y}_l \,\big|\, \mathbf{S})$ is the conditional p.d.f. of $\mathbf{y}_l^j$ given $\mathbf{S}^j$. The message M_l^i, $1 \leq i \leq \lfloor \frac{n}{N} \rfloor$, is then decoded based on $\{\widetilde{\mathbf{S}}\}_{1+(i-1)N}^{iN}$. Denote $\hat{M}_l^i$ the decoded i-th low-rate message. To estimate the channel in the k-th block $\mathbf{h}^k$, where $1 + (i-1)N \leq k \leq iN$ for some i, the decoded low-rate message $\hat{M}_l^i$ is re-encoded to obtain $\hat{\mathbf{S}}^k$, the k-th symbol decision after error correction. The channel output $\mathbf{y}_l^k$ and the decision $\hat{\mathbf{S}}^k$ are plugged into the mismatched minimum-mean-squared-error (MMSE) estimator to obtain the k-th channel estimate $\hat{\mathbf{h}}^k$, which assumes the decision $\hat{\mathbf{S}}^k$ is always correct,

$$\hat{\mathbf{h}}^k(\hat{\mathbf{S}}^k, \mathbf{y}_l^k) = E\left[\mathbf{h}^k \,\big|\, \mathbf{S}^k = \hat{\mathbf{S}}^k, \mathbf{y}_l^k\right] = (\hat{\mathbf{S}}^k)^H \left(\hat{\mathbf{S}}^k(\hat{\mathbf{S}}^k)^H + m\mathbf{I}\right)^{-1} \mathbf{y}_l^k. \tag{6}$$

The high-rate message M_h is then decoded based on $\{\mathbf{y}_h\}_1^n$ and $\{\hat{\mathbf{h}}\}_1^n$.

Due to the delay constraint, the decoding error of the low-rate source is lower bounded from zero. However, if the information rate of the high-rate source is below some threshold, the decoding error of the high-rate source can be made arbitrarily small by applying arbitrarily long high-rate codewords. In the next section, we study the rate at which the high-rate source can be reliably communicated with the decision-directed receiver structure.

III. HIGH-RATE CHANNEL ACHIEVABLE RATE

We consider the achievable rate for the high-rate channel in this section. We first prove a lower bound on the mutual information with Gaussian codebooks and side information. The lower bound is applied to obtain an expression for the achievable rate with a Gaussian codebook and the decision-directed receiver. Under the assumption of binary orthogonal low-rate signaling and a modified channel estimator, the achievable rate is further reduced to an expression that involves only a real double integral of an algebraic function.

A. *Lower Bound on Mutual Information with Gaussian Codebooks for the High-rate Source*

Since the low-rate messages, unknown to the receiver, are encoded into codewords of length N, we consider the N-th extension of the channel and omit the low-rate message index i. Let $\{\mathbf{z}\}$ denote $\{\mathbf{z}\}_1^N$. For any high-rate input distribution $p(\{\mathbf{x}_h\})$, the rate $\frac{1}{N} I(\{\mathbf{x}_h\}; \{\mathbf{y}_h\}, \{\hat{\mathbf{h}}\})$ is achievable by the decision-directed receiver structure. The optimal input distribution $p(\{\mathbf{x}_h\})$, however, is difficult to obtain. In this paper, we are interested in the achievable rate by

an i.i.d. Gaussian codebook, *i.e.,* when $\mathbf{x}_h^1, \ldots, \mathbf{x}_h^N$ are i.i.d. $\mathcal{N}(\mathbf{0}, P\mathbf{I})$, and denote the mutual information when using a Gaussian codebook as $I_g(\cdot; \cdot)$. The mutual information $\frac{1}{N} I_g(\{\mathbf{x}_h\}; \{\mathbf{y}_h\}, \{\hat{\mathbf{h}}\})$, however, is still difficult to evaluate. In this section, we derive a simple achievable rate by lower bounding the mutual information.

For convenience, define a function F on the domain $\mathcal{D} \triangleq \{(c, r, P) : c \in \mathcal{C}, r \in \mathcal{R}, r \geq |c|^2, P \in \mathcal{R}^+\}$:

$$F(c, r, P) \triangleq \log\Big(1 + \frac{P|c|^2}{1 + P(r - |c|^2)}\Big). \tag{7}$$

Lemma 1: Let $x \sim \mathcal{N}(0, P)$ and let $w \in \mathcal{C}$ with zero mean and unit variance be independent of x. Let $g \in \mathcal{C}$ and G_{SI} belongs to an arbitrary domain, and let (g, G_{SI}) be independent of (x, w). Let $y = gx + w$. Let f be a measurable function of G_{SI}, and $z = f(G_{SI})$. Then

$$I_g(x; y, G_{SI}) \geq E_z\Big[F\big(E[g \,|\, z], E\big[|g|^2 \,\big|\, z\big], P\big)\Big]. \tag{8}$$

The lower bound is maximized by letting $z = G_{SI}$.

Proof: See Appendix I. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ $\square$

A lower bound expression on the mutual information with Gaussian codebooks and side information was obtained in [19] by using $z = E[g \,|\, G_{SI}]$. However, (8) is more general and the maximal is achieved when $z = G_{SI}$.

Let $x_{h,i}^k$ and $y_{h,i}^k$ denote the i-th entries of $\mathbf{x}_h^k$ and $\mathbf{y}_h^k$ respectively. Let $g_{h,i}^k$ denote the i-th diagonal entry of the matrix $\mathbf{G}_h^k = \text{diag}(\mathbf{\Pi}_h \mathbf{F} \mathbf{h}^k)$. Since $x_{h,i}^k$, $1 \leq i \leq T_h$ and $1 \leq k \leq N$, are i.i.d., the Gaussian mutual information is lower bounded as follows,

$$\begin{aligned}
\frac{1}{N} I_g\big(\{\mathbf{x}_h\}; \{\mathbf{y}_h\}, \{\hat{\mathbf{h}}\}\big) &= \frac{1}{N} h\big(\{\mathbf{x}_h\}\big) - \frac{1}{N} h(\{\mathbf{x}_h\} \,|\, \{\mathbf{y}_h\}, \{\hat{\mathbf{h}}\}) \\
&= \frac{1}{N}\Big(\sum_{k=1}^N \sum_{i=1}^{T_h} h(x_{h,i}^k)\Big) - \frac{1}{N} h\big(\{\mathbf{x}_h\} \,|\, \{\mathbf{y}_h\}, \{\hat{\mathbf{h}}\}\big) \\
&\geq \frac{1}{N} \sum_{k=1}^N \sum_{i=1}^{T_h} \Big(h(x_{h,i}^k) - h(x_{h,i}^k \,|\, y_{h,i}^k, \hat{\mathbf{h}}^k)\Big) \qquad (9) \\
&= \frac{1}{N} \sum_{k=1}^N \sum_{i=1}^{T_h} I_g\big(x_{h,i}^k; y_{h,i}^k, \hat{\mathbf{h}}^k\big) \\
&\geq \frac{1}{N} \sum_{i=1}^{T_h} \sum_{k=1}^N E_{\hat{\mathbf{h}}^k}\Big[F\big(E[g_{h,i}^k \,|\, \hat{\mathbf{h}}^k], E\big[|g_{h,i}^k|^2 \,\big|\, \hat{\mathbf{h}}^k\big], P\big)\Big] \qquad (10)
\end{aligned}$$

where (9) is because conditioning reduces entropy, (10) by viewing $\hat{\mathbf{h}}^k$ as the side information and applying Lemma 1.

Thus we have obtained a lower bound on $\frac{1}{N}I_g(\{\mathbf{x}_h\}; \{\mathbf{y}_h\}, \{\hat{\mathbf{h}}\})$. Denote $p_k(\hat{\mathbf{h}}\big|\,\mathbf{h})$ the conditional p.d.f. of $\hat{\mathbf{h}}^k$ given $\mathbf{h}^k$. If we know $p_k(\hat{\mathbf{h}}\big|\,\mathbf{h})$ for all $k \in \{1, \ldots, N\}$, it is possible to evaluate (10). However, in some situations, we may only be able to access the average of $p_k(\hat{\mathbf{h}}\big|\,\mathbf{h})$,

$$\underline{p}(\hat{\mathbf{h}}\big|\,\mathbf{h}) \triangleq \frac{1}{N}\sum_{k=1}^{N} p_k(\hat{\mathbf{h}}\big|\,\mathbf{h}). \tag{11}$$

The next proposition summarizes an achievable rate expression that does not depend on individual $p_k(\hat{\mathbf{h}}\big|\,\mathbf{h})$'s.

Since $(\mathbf{x}_h^k, \mathbf{h}^k, \mathbf{w}_h^k)$ are i.i.d. across block index $k \in \{1, \ldots, N\}$, there is no lost of generality in considering $k = 1$ for the next analysis. Introduce a new random vector $\underline{\hat{\mathbf{h}}}$ with $\underline{p}(\hat{\mathbf{h}}\big|\,\mathbf{h})$ as the p.d.f. conditioning on $\mathbf{h}^1$. Let $(\mathbf{h}^1, \underline{\hat{\mathbf{h}}})$, $\mathbf{w}_h^1$, and $\mathbf{x}_h^1$ be independent.

Proposition 2: The Gaussian mutual information is lower bounded by

$$\frac{1}{N}I_g\big(\{\mathbf{x}_h\}; \{\mathbf{y}_h\}, \{\hat{\mathbf{h}}\}\big) \geq \sum_{i=1}^{T_h} E_{\underline{\hat{\mathbf{h}}}}\Big[F\big(E[g_{h,i}^1\big|\,\underline{\hat{\mathbf{h}}}], E\big[|g_{h,i}^1|^2\big|\,\underline{\hat{\mathbf{h}}}\big], P\big)\Big] \triangleq R_d. \tag{12}$$

Proof: See Appendix II. □

B. Achievable Rate under Binary Orthogonal Low-rate Signaling

Since $g_{h,i}^1$ is a function of $\mathbf{h}^1$, if $\underline{p}(\mathbf{h}^1, \underline{\hat{\mathbf{h}}})$ is known, it is possible to evaluate the achievable rate R_d by (12). However, the computation involves an m-dimensional complex integration and the calculations of the conditional mean $E\big[g_{h,i}^1\big|\,\underline{\hat{\mathbf{h}}}\big]$ and the conditional second moment $E\big[|g_{h,i}^1|^2\big|\,\underline{\hat{\mathbf{h}}}\big]$ for all $\underline{\hat{\mathbf{h}}}$. Next, under the assumptions of binary orthogonal low-rate signaling and a modified channel estimator, we reduce R_d to an expression that involves only a double integral.

We assume that $T_l \geq 2m$. The binary low-rate signaling set $\{\mathbf{u}_1, \mathbf{u}_2\}$ and the low-rate tone selection matrix $\boldsymbol{\Pi}_l$ are designed such that for $i, j = 1, 2$,

$$\mathbf{U}_i^H\mathbf{U}_j = \begin{cases} PT_l\mathbf{I} & \text{if } i = j \\ \mathbf{0} & \text{if } i \neq j \end{cases} \tag{13}$$

where $\mathbf{U}_i = \text{diag}(\mathbf{u}_i)\boldsymbol{\Pi}_l\mathbf{F}$. Examples of orthogonal signaling designs include equally spaced low-rate tones and $\{\mathbf{u}_1, \mathbf{u}_2\}$ being columns of a Fourier transform matrix. For instance, if $T = qT_l$ where q is an integer, allocate to the low-rate channel the tones $\{p, p+q, \ldots, p+q(T_l-1)\}$ where $1 \leq p \leq q$. Let $u_{i,l} = \exp\frac{-j2\pi(i-1)(l-1)m}{T_l}$ for $i = 1, 2$ and $l = 1, \ldots, T_l$. The resulting $\mathbf{U}_1$ and $\mathbf{U}_2$ have the orthogonal property (13).

Because of the orthogonality, we have $(\hat{\mathbf{S}}^k)^H \hat{\mathbf{S}}^k = PT_l \mathbf{I}$. Thus

$$\left(\mathbf{I} + \frac{1}{m} \hat{\mathbf{S}}^k (\hat{\mathbf{S}}^k)^H \right)^{-1} = \mathbf{I} - \frac{1}{m + PT_l} \hat{\mathbf{S}}^k (\hat{\mathbf{S}}^k)^H. \tag{14}$$

Therefore, the channel estimator (6) reduces to

$$\hat{\mathbf{h}}^k = \frac{1}{m + PT_l} (\hat{\mathbf{S}}^k)^H \mathbf{y}_l^k \tag{15}$$

$$= \frac{1}{m + PT_l} (\hat{\mathbf{S}}^k)^H (\mathbf{S}^k \mathbf{h}^k + \mathbf{w}_l^k), \tag{16}$$

where (16) is obtained by substituting the low-rate channel (4) into (15).

In the analysis, we consider a modified channel estimator Fig. 3. Consider estimating the k-th channel $\mathbf{h}^k$, $1 \le k \le N$. Let $\{\mathbf{z}\}_-^k$ denote $\{\mathbf{z}^1, \ldots, \mathbf{z}^{k-1}, \mathbf{z}^{k+1}, \ldots, \mathbf{z}^N\}$. In the modified version, the low-rate message is decoded based on the observation $\{\widetilde{\mathbf{S}}\}_-^k$ only. Denote $\hat{M}_{l,k}$ the output of the modified low-rate decoder when $\widetilde{\mathbf{S}}^k$ is discarded. The rest of the operation of the modified estimator remains the same as the original version: $\hat{M}_{l,k}$ is re-encoded to obtain $\hat{\mathbf{S}}^k$; $\hat{\mathbf{S}}^k$ and $\mathbf{y}_l^k$ are plugged into the mismatched MMSE estimator (15) to obtain the estimate $\hat{\mathbf{h}}^k$. Since one of the inputs to the low-rate decoder is discarded in the modified estimator, we expect that the original version has a better performance.

As described above, $\hat{\mathbf{S}}^k$ in the modified estimator is obtained based on the observation of $\{\mathbf{y}_l\}_-^k$, a function of $(\{\mathbf{h}\}_-^k, \{\mathbf{w}_l\}_-^k, \{\mathbf{S}\}_-^k)$. Since $(\mathbf{h}^k, \mathbf{w}_l^k)$ is independent of $(\{\mathbf{h}\}_-^k, \{\mathbf{w}_l\}_-^k, \{\mathbf{S}\})$, $(\mathbf{h}^k, \mathbf{w}_l^k)$ is independent of $(\mathbf{S}^k, \hat{\mathbf{S}}^k)$. Therefore, from (16) and the orthogonality of the low-rate symbols (13), the conditional p.d.f. of $\hat{\mathbf{h}}^k$ conditioning on $\mathbf{h}^k$, $\mathbf{S}^k$ and $\hat{\mathbf{S}}^k$ is given by

$$p_k(\hat{\mathbf{h}} \,|\, \mathbf{h}, \mathbf{S}, \hat{\mathbf{S}}) = \begin{cases} \mathcal{N}_{\hat{\mathbf{h}}} \big(\frac{PT_l}{m + PT_l} \mathbf{h}, \frac{PT_l}{(m + PT_l)^2} \mathbf{I} \big) & \text{if } \mathbf{S} = \hat{\mathbf{S}} \\ \mathcal{N}_{\hat{\mathbf{h}}} \big(\mathbf{0}, \frac{PT_l}{(m + PT_l)^2} \mathbf{I} \big) & \text{if } \mathbf{S} \ne \hat{\mathbf{S}} \end{cases}.$$

Denote $P_k(\mathbf{S}, \hat{\mathbf{S}})$ the probability mass function (p.m.f.) of $(\mathbf{S}^k, \hat{\mathbf{S}}^k)$. Apply the independence of $(\mathbf{S}^k, \hat{\mathbf{S}}^k)$ and $\mathbf{h}^k$ and rewrite (11) as

$$\underline{p}(\hat{\mathbf{h}} \,|\, \mathbf{h}) = \frac{1}{N} \sum_{k=1}^N \Big(\sum_{\mathbf{S}, \hat{\mathbf{S}}} P_k(\mathbf{S}, \hat{\mathbf{S}}) p_k(\hat{\mathbf{h}} \,|\, \mathbf{h}, \mathbf{S}, \hat{\mathbf{S}}) \Big)$$

$$= (1 - \theta) \mathcal{N}_{\hat{\mathbf{h}}} \Big(\frac{PT_l}{m + PT_l} \mathbf{h}, \frac{PT_l}{(m + PT_l)^2} \mathbf{I} \Big) + \theta \mathcal{N}_{\hat{\mathbf{h}}} \Big(\mathbf{0}, \frac{PT_l}{(m + PT_l)^2} \mathbf{I} \Big), \tag{17}$$

where the average decision error probability

$$\theta \triangleq \frac{1}{N} \sum_{k=1}^N \mathcal{P}_r(\mathbf{S}^k \ne \hat{\mathbf{S}}^k).$$

It is worth noting that the decision error probability at different blocks $\mathcal{P}_r(\mathbf{S}^k \ne \hat{\mathbf{S}}^k)$ may not be the same, depending on the low-rate codebook. However, $\underline{p}(\hat{\mathbf{h}} \,|\, \mathbf{h})$ depends only on θ.

Applying (17) to (12) gives the following proposition:

Proposition 3: Under the assumption of the orthogonal low-rate signaling and the modified channel estimator, the achievable rate is given by

$$R_d = T_h\Big((1-\theta)f\big(\frac{PT_l}{m(m+PT_l)}\big) + \theta f\big(\frac{PT_l}{(m+PT_l)^2}\big)\Big), \tag{18}$$

where

$$f(\lambda^2) \triangleq \begin{cases} \int_0^\infty \int_0^\infty \frac{v_2^{m-2}\exp(-(v_1+v_2)/2)}{2^m(m-2)!} f_1\Big(\frac{m\lambda^2 v_1}{2}, \frac{\lambda^2(v_1+v_2)}{2}\Big)\mathrm{d}v_1\mathrm{d}v_2 & \text{if } m \geq 2, \\ \int_0^\infty \frac{\exp(-v_1/2)}{2} f_1\Big(\frac{\lambda^2 v_1}{2}, \frac{\lambda^2 v_1}{2}\Big)\mathrm{d}v_1 & \text{if } m = 1, \end{cases}$$

$$f_1(a^2, b^2) \triangleq \log\Big(1 + \frac{Pf_2(a^2, b^2)}{1 + P\big(f_3(a^2, b^2) - f_2(a^2, b^2)\big)}\Big),$$

$$f_2(a^2, b^2) \triangleq a^2\Big(\frac{1}{f_4(b^2)}(1-\theta)f_5\big(b^2, \frac{PT_l}{m(m+PT_l)}\big)\Big)^2,$$

$$f_3(a^2, b^2) \triangleq \frac{1}{f_4(b^2)}\Big((1-\theta)f_5\big(b^2, \frac{PT_l}{m(m+PT_l)}\big)\big(\frac{m}{m+PT_l} + a^2\big) + \theta f_5\big(b^2, \frac{PT_l}{(m+PT_l)^2}\big)\Big),$$

$$f_4(b^2) \triangleq (1-\theta)f_5\Big(b^2, \frac{PT_l}{m(m+PT_l)}\Big) + \theta f_5\Big(b^2, \frac{PT_l}{(m+PT_l)^2}\Big),$$

$$f_5(b^2, c^2) \triangleq \frac{\exp(-b^2/c^2)}{(\pi c^2)^m}.$$

Proof: See Appendix III. □

Equation (18) involves only a double integral function, which can be readily numerical evaluated. Applying Dominant Convergence Theorem to (18) gives the following corollary:

Corollary 4: The achievable rate R_d has the following convergence,

$$\lim_{\theta\to 0} R_d = T_h \int_0^\infty e^{-z}\log\Big(1 + \frac{P^2 T_l z}{m + P(m+T_l)}\Big)\mathrm{d}z \triangleq R_t. \tag{19}$$

The rate R_t is achievable by the training-based scheme [6]. Corollary 4 shows that the achievable rate by the decision-directed receiver converges to that by the training-based scheme as the decision error probability goes to zero.

C. Low-rate Decoder

As shown in (18), the achievable rate R_d is affected by the low-rate decision error θ. Here we study the performance of the modified low-rate decoder. We first derive the crossover probability of the equivalent binary symmetric channel (BSC), and then obtain the error probability of the modified decoder with a block code or a random code.

Rewrite the ML low-rate symbol detector (5) as

$$\widetilde{\mathbf{S}}^j = \arg\max_{\mathbf{S}\in\{\mathbf{U}_1, \mathbf{U}_2\}} \mathcal{N}_{\mathbf{y}_l^j}\Big(\mathbf{0}, \frac{1}{m}\mathbf{SS}^H + \mathbf{I}\Big)$$

$$= \underset{\mathbf{S} \in \{\mathbf{U}_1, \mathbf{U}_2\}}{\arg\min} (\mathbf{y}_l^j)^H \Big(\frac{1}{m} \mathbf{S}\mathbf{S}^H + \mathbf{I} \Big)^{-1} \mathbf{y}_l^j \tag{20}$$

$$= \underset{\mathbf{S} \in \{\mathbf{U}_1, \mathbf{U}_2\}}{\arg\max} \Big| \mathbf{S}^H \mathbf{y}_l^j \Big|^2 \tag{21}$$

where (21) is because of (14) and (20) because of the following equality,

$$\det\Big(\frac{1}{m} \mathbf{S}\mathbf{S}^H + \mathbf{I} \Big) = \det\Big(\frac{1}{m} \mathbf{S}^H \mathbf{S} + \mathbf{I} \Big) = \Big(\frac{m + PT_l}{m} \Big)^m.$$

The ML detector (21) converts the low-rate channel into a BSC. The crossover probability of the BSC, $\epsilon \triangleq \mathcal{P}_r(\widetilde{\mathbf{S}}^j \neq \mathbf{S}^j)$, is given by

$$\epsilon = \sum_{i=0}^{m-1} \frac{(m-1+i)!}{(m-1)!i!} \frac{1}{(\frac{m}{m+PT_l}+1)^i} \frac{1}{(1+\frac{m+PT_l}{m})^m}. \tag{22}$$

The derivation of (22) is given in Appendix IV.

The decision error probability θ depends on the error correction performance of the low-rate code. We first consider using an (N, k, t) code for the low-rate source, where (N, k, t) denotes block codes with length N, k input bits, and t bits error correction capability. Recall that in the modified high-rate receiver (Fig. 3), $\hat{M}_{l,k}$ is decoded based on $\{\widetilde{\mathbf{S}}\}_-^k$ only. We replace $\widetilde{\mathbf{S}}^k$ with an independent fair coin flip output and use the decoder that corrects up to t bits of error. The decision error probability is then bounded by

$$\theta \leq \frac{1}{N} \sum_{k=1}^{N} \mathcal{P}_r(\hat{M}_{l,k} \neq M_l) \tag{23}$$

$$= \frac{1}{2} \sum_{i=t+1}^{N-1} \binom{N-1}{i} \epsilon^i (1-\epsilon)^{N-1-i} + \frac{1}{2} \sum_{i=t}^{N-1} \binom{N-1}{i} \epsilon^i (1-\epsilon)^{N-1-i} \tag{24}$$

where (23) is because $\{\hat{\mathbf{S}}^k \neq \mathbf{S}^k\}$ implies $\{\hat{M}_{l,k} \neq M_l\}$, (24) due to the random replacement.

Next, we consider a random code for the low-rate source. Denote $\mathbb{C}$ the random codebook, with 2^{NR_l} messages and N symbols for each message, generated with equal probability over the binary alphabet. By the random coding argument [20], the modified low-rate decoder that discards $\widetilde{\mathbf{S}}^k$ has the ensemble error probability, averaged over the codebook ensemble,

$$E_{\mathbb{C}}\Big[\mathcal{P}_r(\hat{M}_{l,k} \neq M_l \big| \mathbb{C}) \Big] \leq 2^{-(N-1)E_r(\frac{N}{N-1}R_l, \epsilon)} \tag{25}$$

where the random coding exponent

$$E_r(R, \epsilon) = \max_{0 \leq \rho \leq 1} \Big(\rho - (1+\rho)\log_2\Big(\epsilon^{\frac{1}{1+\rho}} + (1-\epsilon)^{\frac{1}{1+\rho}} \Big) - \rho R \Big).$$

Since the right hand side of (25) does not depend on k, we have

$$E_{\mathbb{C}}\Big[\frac{1}{N} \sum_{k=1}^{N} \mathcal{P}_r(\hat{M}_{l,k} \neq M_l \big| \mathbb{C}) \Big] \leq 2^{-(N-1)E_r(\frac{N}{N-1}R_l, \epsilon)},$$

which, together with (23), implies that there exists a codebook such that

$$\theta \le 2^{-(N-1)E_r\left(\frac{N}{N-1}R_l, \epsilon\right)}. \tag{26}$$

D. Optimal Decoding

In this subsection, we investigate the rate achievable by decoders without the explicit channel estimation structure. We derive an achievable rate for the optimal decoding structure, expressed as a function of the low-rate decoding error probability. We compare the achievable rates by the decision-directed and the optimal-decoding receivers in Section IV and illustrate the sub-optimality of the decision-directed structure at large decision error probabilities.

Denote $\gamma \triangleq \mathcal{P}_r(\hat{M}_l \ne M_l)$ the low-rate decoding error probability. When the explicit channel estimation constraint is removed, the rate $\frac{1}{N}I\big(\{\mathbf{x}_h\}; \{\mathbf{y}_h\}, \{\mathbf{y}_l\}\big)$ is achievable. By the chain rule of mutual information,

$$I\big(\{\mathbf{x}_h\}; \{\mathbf{y}_h\}, \{\mathbf{y}_l\}, \{\mathbf{x}_l\}\big) - I\big(\{\mathbf{x}_h\}; \{\mathbf{y}_h\}, \{\mathbf{y}_l\}\big) = I\big(\{\mathbf{x}_h\}; \{\mathbf{x}_l\}\big|\{\mathbf{y}_h\}, \{\mathbf{y}_l\}\big)$$

$$\le H\big(\{\mathbf{x}_l\}\big|\{\mathbf{y}_h\}, \{\mathbf{y}_l\}\big) \tag{27}$$

$$\le H\big(\{\mathbf{x}_l\}\big|\{\mathbf{y}_l\}\big) \tag{28}$$

$$\le H(\gamma) + \gamma N R_l \tag{29}$$

where (27) is due to the positivity of entropy, (28) due to the fact that conditioning decreases entropy, and (29) due to Fano's inequality. The achievable rate by the optimal decoding structure is therefore lower bounded by

$$\frac{1}{N}I\big(\{\mathbf{x}_h\}; \{\mathbf{y}_h\}, \{\mathbf{y}_l\}\big) \ge \frac{1}{N}I\big(\{\mathbf{x}_h\}; \{\mathbf{y}_h\}, \{\mathbf{y}_l\}, \{\mathbf{x}_l\}\big) - \frac{H(\gamma)}{N} - \gamma R_l. \tag{30}$$

As γ goes to zero, the optimal receiver achieves the performance of the training based scheme $\frac{1}{N}I\big(\{\mathbf{x}_h\}; \{\mathbf{y}_h\}, \{\mathbf{y}_l\}, \{\mathbf{x}_l\}\big)$. Even if γ is large, the difference of the two achievable rates is less than $\frac{1}{N} + R_l$, which is small since R_l is small.

Since R_t (19) is achievable by the training-based scheme with a Gaussian codebook, we have

$$\frac{1}{N}I_g\big(\{\mathbf{x}_h\}; \{\mathbf{y}_h\}, \{\mathbf{y}_l\}, \{\mathbf{x}_l\}\big) \ge R_t. \tag{31}$$

Applying (31) to (30) gives the achievable rate with the optimal decoder and a Gaussian codebook

$$\frac{1}{N}I_g\big(\{\mathbf{x}_h\}; \{\mathbf{y}_h\}, \{\mathbf{y}_l\}\big) \ge R_t - \frac{H(\gamma)}{N} - \gamma R_l \triangleq R_o. \tag{32}$$

IV. NUMERICAL RESULTS

In the first example, we assume that the low-rate channel uses the binary orthogonal signaling and an (N, k, t) error correction code. The low-rate detector converts the low-rate channel into a BSC with crossover probability ϵ (22). The decision error probability θ is bounded by (24).

We assume the channel length $m = 4$, the OFDM block size $T = 64$, the tone allocation $(T_l, T_h) = (8, 56)$, and the $(7, 4, 1)$ Hamming code for low-rate error correction. Fig. 4 shows the achievable rate versus P, the signal-to-noise ratio (SNR): R_d (18) for the decision-directed structure, R_t (19) for the training-based scheme, and R_o (32) for the optimal-decoding structure. Also shown in Fig. 4 is the decision error probability θ. The SNR value of the vertical line that a θ value points at is the SNR value at which the decision error probability equals that θ value. In Fig. 4 and the following figures, achievable rates are normalized by the OFDM block size T.
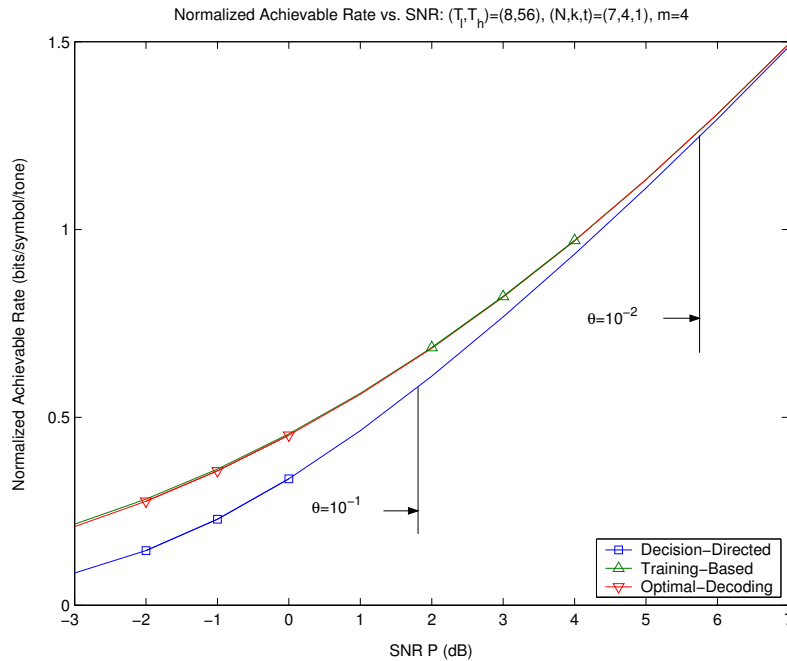

Fig. 4. Normalized achievable rate versus SNR.

As shown in Fig. 4, when SNR increases, the low-rate decoding error probability decreases, and the achievable rate by the decision-directed receiver converges to the achievable rate by the training-based scheme. The convergence happens at moderate decision error probability levels, around $\theta = 10^{-2} \sim 10^{-3}$. The achievable rate by the optimal-decoding receiver, on

the other hand, is very close to that by the training-based scheme regardless of the SNR level.

In the second example, we fix SNR and the low-rate information rate R_l and vary the low-rate code length N. The decision error probability θ decreases as the code length N increases. Fig. 5 shows the achievable rates versus the decision error probability θ when $(T_l, T_h) = (8, 56)$ and $m = 4$ under two SNR levels, $P = 15dB$ and $P = 7dB$. When the decision error probability θ goes to zero, the achievable rate R_d for the decision-directed structure converges to R_t for the training-based scheme. The convergence happens at moderate decision error probabilities, around $\theta = 10^{-3}$. The relationship between θ and N is given by the random coding bound (26). Fig. 6 shows the achievable rate R_d versus the coding length N when $(T_l, T_h) = (8, 56)$, $m = 4$, and $R_l = 1/2$ bits per OFDM block under two SNR levels, $P = 15dB$ and $P = 7dB$. Fig. 6 is very similar to the mirror image of Fig. 5 along the vertical axis. Fig. 6 shows how long the low-rate message should be encoded with a random code in order to consider the low-rate decoding is error free. In Fig. 6, the curves converge when $N \geq 20$. Also shown in Fig. 5 and Fig. 6 is the achievable rate by the optima-decoding receiver, which is very close to that by the training-based scheme, indicating the sub-optimality of the decision-directed scheme when the decision error probability is large.
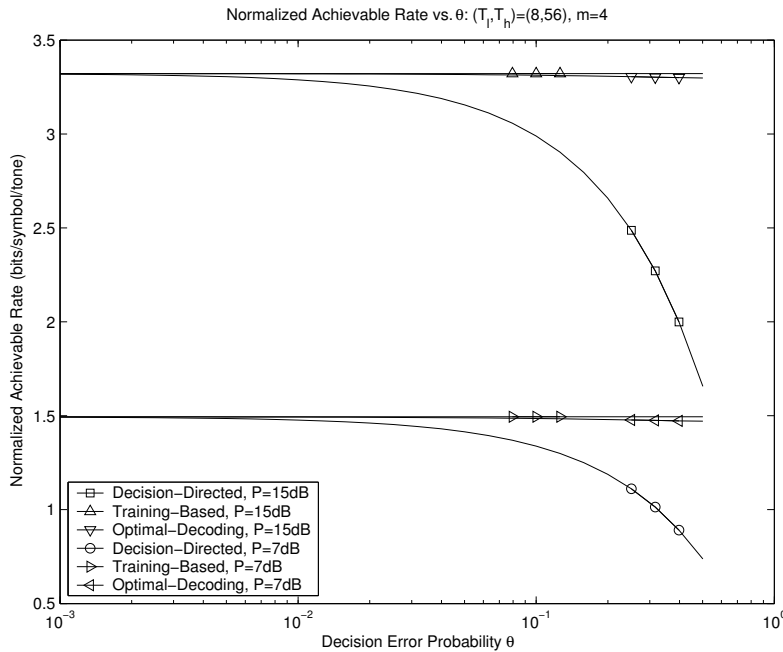

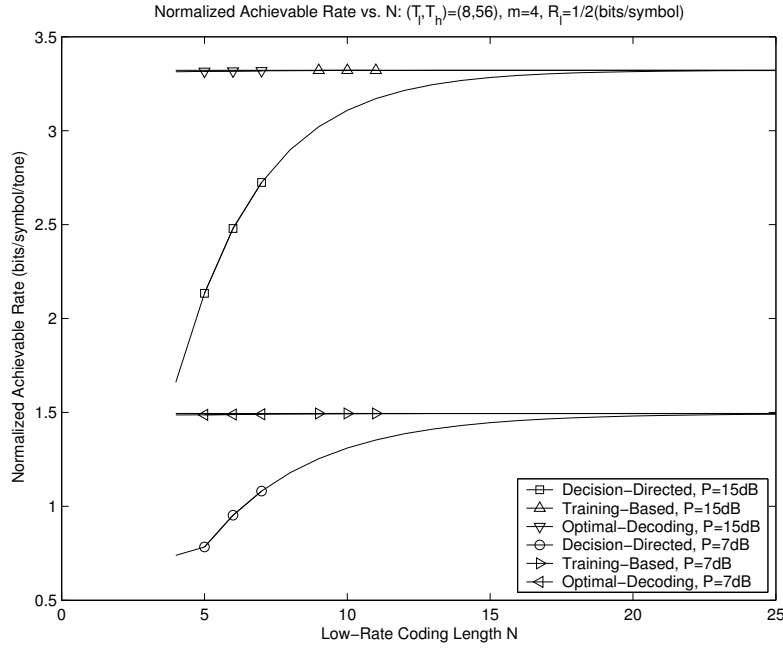Fig. 5. Normalized achievable rate versus decision-error probability.

Fig. 6.   Normalized achievable rate versus low-rate coding length.

V. SUMMARY

We have studied the communication of two independent sources over unknown fading channels. We have proposed, and analyzed the performance of, a system structure (Fig. 1) that jointly considers the different decoding requirements of the application layer and the unknown fading nature of the physical channel for a better channel utilization. Two receiver structures are considered. In the decision-directed structure, the low-rate transmission is utilized to estimate the channel response, which assists the high-rate decoding. In the optimal decoding structure, the explicit channel estimation constraint is removed.

We have derived a lower bound on the mutual information with Gaussian codebooks and side information. The lower bound generalizes previous published results [2], [5], [6], [10]–[19] in that it does not require the side information to satisfy (1) and it is tighter than the result in [19]. We have used the lower bound to obtain high-rate achievable rate expressions for both receiver structures. The calculation of the achievable rate by the decision-directed receiver is reduced to a double integral under the assumption of binary orthogonal low-rate signaling.

We have shown that the achievable rates by the decision-directed receiver and the optimal receiver converge to that by the training-based scheme as the low-rate decision error

probability goes to zero. By simulation, the convergence happens at moderate low-rate error probabilities. Compared with the training-based scheme, the rate achievable using optimal decoding is lower by at most $\frac{1}{N} + R_l$ regardless of the low-rate error probability. When the low-rate error probability is large, the decision-directed structure is suboptimal. When the low-rate decoding error probability is small, both decoding schemes provide an additional low-rate channel while maintaining the performance of the high-rate channel close to that of the training-based system.

APPENDIX I
PROOF OF LEMMA 1

Recall $y = gx + w$. We have $E[xy^* | z] = PE[g^* | z]$ and $E[|y|^2 | z] = PE[|g|^2 | z] + 1$. Defining $\operatorname{cov}(v | u) \triangleq E[|v - E[v]|^2 | u]$, we have

$$\operatorname{cov}\Big(x - \frac{E[xy^* | z]}{E[|y|^2 | z]}y \Big| z\Big) = E[|x|^2 | z] - \frac{\big|E[xy^* | z]\big|^2}{E[|y|^2 | z]} = P - \frac{P^2\big|E[g^* | z]\big|^2}{PE[|g|^2 | z] + 1}. \qquad (33)$$

The Gaussian mutual information is lower bounded as follows,

$$I_g(x; y, G_{SI}) \geq I_g(x; y, z) \qquad (34)$$

$$= h(x) - h(x | y, z)$$

$$= h(x) - h\Big(x - \frac{E[xy^* | z]}{E[|y|^2 | z]}y \Big| y, z\Big) \qquad (35)$$

$$\geq h(x) - h\Big(x - \frac{E[xy^* | z]}{E[|y|^2 | z]}y \Big| z\Big) \qquad (36)$$

$$\geq h(x) - E_z \log\Big(\pi e \operatorname{cov}\Big(x - \frac{E[xy^* | z]}{E[|y|^2 | z]}y \Big| z\Big)\Big) \qquad (37)$$

$$= E_z\Big[F\big(E[g | z], E[|g|^2 | z], P\big)\Big] \qquad (38)$$

where (34) is because of the data processing inequality, (35) because $yE[xy^* | z]/E[|y|^2 | z]$ is a function of (y, z), (36) because conditioning reduces entropy, (37) because the Gaussian distribution maximizes entropy among distributions with the same covariance, (38) because of (33) and the fact that x is complex Gaussian distributed with variance P. Thus we have proved the first part of Lemma 1. Next, we show the optimality of $z = G_{SI}$.

Lemma 5: For $r_1, r_2, r_3 \in \mathcal{R}$ and $r_1 > r_2^2 + r_3^2$, the function $\log\big(r_1/(r_1 - r_2^2 - r_3^2)\big)$ is convex in the triple (r_1, r_2, r_3).

Proof: Let h denote the function and let $\nabla^2 h(r_1, r_2, r_3) \in \mathcal{R}^{3 \times 3}$ denote the Hessian matrix with $\frac{\partial^2 h}{\partial r_i \partial r_j}$ as the entry in the i-th row and the j-th column. Since the domain $r_1 >$

$r_2^2 + r_3^2$ is convex, we only need to show that $\nabla^2 h(r_1, r_2, r_3)$ is non-negative definite in the domain $r_1 > r_2^2 + r_3^2$. Equivalently, we need to show that the matrix

$$\mathbf{A} = r_1^2(r_1 - r_2^2 - r_3^2)^2 \cdot \nabla^2 h(r_1, r_2, r_3)$$

is non-negative definite. Some algebra reveals that

$$\det(\lambda \mathbf{I} - \mathbf{A}) = (\lambda^2 - \alpha \lambda + \beta)(\lambda - \gamma),$$

$$\alpha = 2r_1^2(r_1 + r_2^2 + r_3^2) + (r_2^2 + r_3^2)(2r_1 - r_2^2 - r_3^2),$$

$$\beta = 2r_1^2(r_2^2 + r_3^2)^2(r_1 - r_2^2 - r_3^2),$$

$$\gamma = 2r_1^2(r_1 - r_2^2 - r_3^2).$$

Since $r_1 > r_2^2 + r_3^2$, we have $\alpha, \beta, \gamma \geq 0$. Therefore, the three eigenvalues of the symmetric matrix $\mathbf{A}$ are non-negative, proving the non-negative definiteness of $\mathbf{A}$. $\qquad\square$

Corollary 6: For a fixed $P > 0$, $F(c, r, P)$ is convex in the pair (r, c) in the domain $\mathcal{D}$.

Proof: Suppose $c = c_r + jc_i$ where $c_r, c_i \in \mathcal{R}$. Let $r_0 = r + 1/P$. We have $F = \log\big(r_0/(r_0 - c_r^2 - c_i^2)\big)$ and $r_0 > c_r^2 + c_i^2$. By Lemma 5, F is convex in (r_0, c_r, c_i), which is a linear function of (r, c), thus proving the convexity of F in (r, c). $\qquad\square$

Since z is a function of G_{SI}, we have $E_{g|z}[g] = E_{G_{SI}|z}E_{g|G_{SI},z}[g] = E_{G_{SI}|z}E_{g|G_{SI}}[g]$. Similarly, $E_{g|z}\big[|g|^2\big] = E_{G_{SI}|z}E_{g|G_{SI}}\big[|g|^2\big]$. Therefore,

$$F\big(E_{g|z}[g], E_{g|z}\big[|g|^2\big], P\big) = F\big(E_{G_{SI}|z}E_{g|G_{SI}}[g], E_{G_{SI}|z}E_{g|G_{SI}}\big[|g|^2\big], P\big)$$

$$\leq E_{G_{SI}|z}\Big[F\big(E_{g|G_{SI}}[g], E_{g|G_{SI}}\big[|g|^2\big], P\big)\Big] \qquad (39)$$

where (39) is due to the convexity of F (Corollary 6) and the fact that $E_{g|G_{SI}}\big[|g|^2\big] \geq \big|E_{g|G_{SI}}[g]\big|^2$. Taking expectation over both sides of (39) concludes the proof of the optimality of $z = G_{SI}$. $\qquad\square$

APPENDIX II

PROOF OF PROPOSITION 2

Consider a fixed $i \in \{1, \ldots, T_h\}$. Denote $p_k(\hat{\mathbf{h}}, g)$ the p.d.f. of $(\hat{\mathbf{h}}^k, g_{h,i}^k)$ for $1 \leq k \leq N$. Let $K \in \{1, \ldots, N\}$ be uniform distributed and $(\hat{\mathbf{h}}, g)$ have conditional p.d.f. $p(\hat{\mathbf{h}}, g \,|\, K = k) = p_k(\hat{\mathbf{h}}, g)$. Thus,

$$E[g_{h,i}^k \,|\, \hat{\mathbf{h}}^k = \mathbf{h}_0] = E[g \,|\, \hat{\mathbf{h}} = \mathbf{h}_0, K = k], \qquad (40)$$

$$E\big[|g_{h,i}^k|^2 \,|\, \hat{\mathbf{h}}^k = \mathbf{h}_0\big] = E\big[|g|^2 \,|\, \hat{\mathbf{h}} = \mathbf{h}_0, K = k\big]. \qquad (41)$$

The i-th term of (10) is then lower bounded as follows

$$\frac{1}{N} \sum_{k=1}^{N} E_{\hat{\mathbf{h}}^k} \left[F\left(E[g_{h,i}^k \big| \hat{\mathbf{h}}^k], E\big[|g_{h,i}^k|^2 \big| \hat{\mathbf{h}}^k\big], P \right) \right]$$

$$= E_K E_{\hat{\mathbf{h}}|K} \left[F\left(E[g \big| \hat{\mathbf{h}}, K], E\big[|g|^2 \big| \hat{\mathbf{h}}, K\big], P \right) \right] \tag{42}$$

$$= E_{\hat{\mathbf{h}}} E_{K|\hat{\mathbf{h}}} \left[F\left(E[g \big| \hat{\mathbf{h}}, K], E\big[|g|^2 \big| \hat{\mathbf{h}}, K\big], P \right) \right]$$

$$\geq E_{\hat{\mathbf{h}}} \left[F\left(E[g \big| \hat{\mathbf{h}}], E\big[|g|^2 \big| \hat{\mathbf{h}}\big], P \right) \right] \tag{43}$$

$$= E_{\underline{\hat{\mathbf{h}}}} \left[F\left(E[g_{h,i}^1 \big| \underline{\hat{\mathbf{h}}}], E\big[|g_{h,i}^1|^2 \big| \underline{\hat{\mathbf{h}}}\big], P \right) \right] \tag{44}$$

where (42) applies (40) and (41), (43) because of the convexity of F (see Corollary 6 in Appendix I), (44) because $(\underline{\hat{\mathbf{h}}}, g_{h,i}^1)$ has the same distribution as $(\hat{\mathbf{h}}, g)$. Summing over i on both sides of (44) and applying (10) conclude the proof. $\qquad\square$

APPENDIX III

PROOF OF PROPOSITION 3

Recall $\mathbf{G}_h^1 = \text{diag}(\mathbf{\Pi}_h \mathbf{F} \mathbf{h}^1)$. Denoting $\mathbf{f}_i^H$ the i-th row of $\mathbf{\Pi}_h \mathbf{F}$, we have $g_{h,i}^1 = \mathbf{f}_i^H \mathbf{h}^1$. Since $\mathbf{h}^1 \sim \mathcal{N}(\mathbf{0}, \frac{1}{m}\mathbf{I})$ and the p.d.f. of $\underline{\hat{\mathbf{h}}}$ conditioning on $\mathbf{h}^1$, $\underline{p}(\hat{\mathbf{h}} \big| \mathbf{h})$, is given by (17), it can be shown that the p.d.f. of $\underline{\hat{\mathbf{h}}}$ and the conditional p.d.f. of $g_{h,i}^1$ given $\underline{\hat{\mathbf{h}}}$ are

$$p_{\underline{\hat{\mathbf{h}}}}(\hat{\mathbf{h}}) = (1-\theta)\mathcal{N}_{\hat{\mathbf{h}}}\left(\mathbf{0}, \frac{PT_l}{m(m+PT_l)}\mathbf{I}\right) + \theta\mathcal{N}_{\hat{\mathbf{h}}}\left(\mathbf{0}, \frac{PT_l}{(m+PT_l)^2}\mathbf{I}\right), \tag{45}$$

$$p_{g_{h,i}^1|\underline{\hat{\mathbf{h}}}}(g \big| \hat{\mathbf{h}}) = \frac{1}{p_{\underline{\hat{\mathbf{h}}}}(\hat{\mathbf{h}})} \left((1-\theta)\mathcal{N}_{\hat{\mathbf{h}}}\left(\mathbf{0}, \frac{PT_l}{m(m+PT_l)}\mathbf{I}\right) \mathcal{N}_g\left(\mathbf{f}_i^H \hat{\mathbf{h}}, \frac{m}{m+PT_l}\right) \right.$$

$$\left. + \theta\mathcal{N}_{\hat{\mathbf{h}}}\left(\mathbf{0}, \frac{PT_l}{(m+PT_l)^2}\mathbf{I}\right) \mathcal{N}_g(0, 1) \right). \tag{46}$$

From (46), some calculations reveal that

$$\big|E[g_{h,i}^1 \big| \underline{\hat{\mathbf{h}}}]\big|^2 = f_2\big(|\mathbf{f}_i^H \hat{\mathbf{h}}|^2, |\hat{\mathbf{h}}|^2\big),$$

$$E\big[|g_{h,i}^1|^2 \big| \hat{\mathbf{h}}\big] = f_3\big(|\mathbf{f}_i^H \hat{\mathbf{h}}|^2, |\hat{\mathbf{h}}|^2\big).$$

Therefore, it is clear that $E\big[f_1(|\mathbf{f}_i^H \hat{\mathbf{h}}|^2, |\underline{\hat{\mathbf{h}}}|^2)\big]$ is the i-th term in the summation of (12).

Next, we consider the evaluation of $E\big[f_1(|\mathbf{f}_i^H \mathbf{r}|^2, |\mathbf{r}|^2)\big]$ where $\mathbf{r}$ has length m and distribution $\mathcal{N}(\mathbf{0}, \lambda^2\mathbf{I})$. Let $\mathbf{s} = \mathbf{U}\mathbf{r}$ where $\mathbf{U}$ is a unitary matrix and the first row of $\mathbf{U}$ is equal to $\mathbf{f}_i^H/\sqrt{m}$. Denote s_j the j-th entry of $\mathbf{s}$, $1 \leq j \leq m$. We have $\mathbf{s} \sim \mathcal{N}(\mathbf{0}, \lambda^2\mathbf{I})$, $|\mathbf{s}|^2 = |\mathbf{r}|^2$, and $|\mathbf{f}_i^H \mathbf{r}|^2 = m|s_1|^2$. Define $v_1 \triangleq \frac{2}{\lambda^2}|s_1|^2$ and $v_2 \triangleq \frac{2}{\lambda^2}\sum_{i=2}^{m}|s_i|^2$. When $m \geq 2$, since $s_1, \ldots, s_m$ are independent complex Gaussians, v_1 and v_2 are independent gamma random variables

with p.d.f. $p_{v_1}(v) = \frac{e^{-v/2}}{2}$ and $p_{v_2}(v) = \frac{v^{m-2}e^{-v/2}}{2^{m-1}(m-2)!}$, respectively [21]. When $m = 1$, we have $p_{v_1}(v) = \frac{e^{-v/2}}{2}$ and $p_{v_2}(v) = \delta(v)$. Therefore,

$$
\begin{aligned}
E\Big[f_1\big(|\mathbf{f}_i^H\mathbf{r}|^2, |\mathbf{r}|^2\big)\Big] &= E\Big[f_1\big(m|s_1|^2, |\mathbf{s}|^2\big)\Big] \\
&= E_{v_1,v_2}\Big[f_1\big(\frac{m\lambda^2 v_1}{2}, \frac{\lambda^2(v_1+v_2)}{2}\big)\Big] \\
&= f(\lambda^2).
\end{aligned}
$$

Since $\underline{\hat{\mathbf{h}}}$ is a Gaussian mixture (45), we have

$$
E\Big[f_1\big(|\mathbf{f}_i^H\underline{\hat{\mathbf{h}}}|^2, |\underline{\hat{\mathbf{h}}}|^2\big)\Big] = (1-\theta)f\Big(\frac{PT_l}{m(m+PT_l)}\Big) + \theta f\Big(\frac{PT_l}{(m+PT_l)^2}\Big).
$$

Recall that $E\big[f_1(|\mathbf{f}_i^H\underline{\hat{\mathbf{h}}}|^2, |\underline{\hat{\mathbf{h}}}|^2)\big]$ is the i-th term in the summation of (12). Summing over i gives the desired result.   □

APPENDIX IV
DETECTION ERROR PROBABILITY

In the derivation, we drop the symbol index j in the ML detector (21). Let $\mathbf{r}_i = \mathbf{U}_i^H\mathbf{y}_l$ and $v_i = |\mathbf{r}_i|^2$ for $i = 1, 2$. Suppose $\mathbf{S} = \mathbf{U}_1$ is transmitted. The detection error probability is

$$
\epsilon = \mathcal{P}_r\big(v_1 \le v_2 \big| \mathbf{S} = \mathbf{U}_1\big). \tag{47}
$$

Recall $\mathbf{y}_l = \mathbf{S}\mathbf{h} + \mathbf{w}_l$, $\mathbf{h} \sim \mathcal{N}(\mathbf{0}, \frac{1}{m}\mathbf{I})$, $\mathbf{w}_l \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $(\mathbf{U}_1, \mathbf{U}_2)$ are orthogonal (13). Given $\mathbf{S} = \mathbf{U}_1$, $\mathbf{r}_1$ and $\mathbf{r}_2$ are Gaussians with p.d.f. $\mathcal{N}\big(\mathbf{0}, \frac{PT_l(m+PT_l)}{m}\mathbf{I}\big)$ and $\mathcal{N}(\mathbf{0}, PT_l\mathbf{I})$, respectively. Because of the orthogonality of $\mathbf{U}_1$ and $\mathbf{U}_2$ (13), we have $E\big[\mathbf{r}_1\mathbf{r}_2^H \big| \mathbf{S} = \mathbf{U}_1\big] = \mathbf{0}$. Therefore, given $\mathbf{S} = \mathbf{U}_1$, $\mathbf{r}_1$ and $\mathbf{r}_2$ are independent zero mean Gaussian random vectors. Hence, given $\mathbf{S} = \mathbf{U}_1$, v_1 and v_2 are independently central chi-square distributed with

$$
p_{v_1|\mathbf{S}}\big(v \big| \mathbf{S} = \mathbf{U}_1\big) = \frac{v^{m-1}\exp\big(-\frac{vm}{PT_l(m+PT_l)}\big)}{\big(\frac{PT_l(m+PT_l)}{m}\big)^m (m-1)!} \tag{48}
$$

$$
\mathcal{P}_r\big(v_2 \ge v \big| \mathbf{S} = \mathbf{U}_1\big) = e^{-\frac{v}{PT_l}} \sum_{i=0}^{m-1} \frac{1}{i!}\Big(\frac{v}{PT_l}\Big)^i \tag{49}
$$

[21]. By the conditional independence of v_1 and v_2, the detection error (47) is expressed as

$$
\epsilon = \int_0^\infty \mathcal{P}_r\big(v_2 \ge v \big| \mathbf{S} = \mathbf{U}_1\big) p_{v_1|\mathbf{S}}\big(v \big| \mathbf{S} = \mathbf{U}_1\big)\mathrm{d}v. \tag{50}
$$

Substituting (48) and (49) into (50), switching the order of the integration and the summation, and integrating produce the formula (22).   □

REFERENCES

[1] DVB-T, "Digital video broadcasting (DVB); framing structure, channel coding and modulation for digital terrestrial television," Jan. 2001.

[2] M. Medard, "The effect upon channel capacity in wireless communication of perfect and imperfect knowledge of the channel," *IEEE Trans. Information Theory*, vol. 46, no. 3, pp. 933–946, May 2000.

[3] R. Gallager, "Basic limits on protocol information in data communication networks," *IEEE Trans. Information Theory*, vol. 22, no. 4, pp. 385–398, July 1976.

[4] Q. Bao and L. Tong, "Protocol-aided channel equalization in wireless ATM," *IEEE J. Select. Areas. Commun.*, vol. 18, no. 3, pp. 418–435, March 2000.

[5] B. Hassibi and B. Hochwald, "How much training is needed in multiple-antenna wireless links?" *IEEE Trans. Info. Theory*, vol. 49, no. 4, pp. 951–963, April 2003.

[6] S. Adireddy, L. Tong, and H. Viswanathan, "Optimal placement of known symbols for frequency-selective block-fading channels," *IEEE Trans. Info. Theory*, vol. 48, no. 8, pp. 2338–2353, August 2002.

[7] I. Telatar, "Capacity of multi-antenna gaussian channels," *European Trans. Telecomm*, vol. 10, no. 6, pp. 585–596, Nov-Dec 1999.

[8] T. Marzetta and B. Hochwald, "Capacity of a mobile multiple-antenna communication link in Rayleigh flat fading," *IEEE Trans. Information Theory*, vol. 45, no. 1, pp. 139–157, January 1999.

[9] I. Abou-Faycal, M. Trott, and S. Shamai, "The capacity of discrete-time memoryless Rayleigh-fading channels," *IEEE Trans. Information Theory*, vol. 47, no. 4, pp. 1290–1301, May 2001.

[10] H. Vikalo, B. Hassibi, B. Hochwald, and T. Kailath, "On the capacity of frequency-selective channels in training-based transmission schemes," *IEEE Trans. Signal Processing*, vol. 52, no. 9, pp. 2572–2583, Sep. 2004.

[11] S. Ohno and G. B. Giannakis, "Capacity maximizing pilots and precoders for wireless OFDM over rapidly fading channels," *IEEE Trans. Info. Theory*, vol. 50, no. 9, pp. 2138–2145, Sep. 2004.

[12] ——, "Average-rate optimal PSAM transmissions over time-selective fading channels," *IEEE Trans. Wireless Comm.*, vol. 1, no. 4, pp. 712–720, Oct. 2002.

[13] X. Ma, G. Giannakis, and S. Ohno, "Optimal training for block transmission of doubly selective wireless fading channels," *IEEE Trans. Signal Processing*, vol. 51, no. 5, pp. 1351–1366, May 2003.

[14] J. Baltersee, G. Fock, and H. Meyr, "Achievable Rate of MIMO Channels with Data-Aided Channel Estimation and Pefect Interleaving," *IEEE Journal on Select. Areas. Commun.*, vol. 19, no. 12, pp. 2358–2368, December 2001.

[15] ——, "An information theoretic foundation of synchronized detection," *IEEE Trans. Communications*, vol. 49, no. 12, pp. 2115–2123, Dec. 2001.

[16] S. Bhashyam, A. Sabharwal, and B. Aazhang, "Feedback gain in multiple antenna systems," *IEEE Trans. on Communications*, vol. 50, no. 5, pp. 785–798, May 2002.

[17] C. Wen, Y. Wang, and J. Chen, "An adaptive spatio-temporal coding scheme for indoor wireless communication," *IEEE Journal on Select. Areas. Commun.*, vol. 21, no. 2, pp. 161–170, Feb. 2003.

[18] D. Samardzija and N. Mandayam, "Pilot-assisted estimation of MIMO fading channel response and achievable data rates," *IEEE Trans. Signal Processing*, vol. 51, no. 11, pp. 2882–2890, Nov. 2003.

[19] A. Lapidoth and S. Shamai, "Fading channels: how perfect need 'perfect side information' be?" *IEEE Trans. Information Theory*, vol. 48, no. 5, pp. 1118–1134, May 2002.

[20] R. G. Gallager, *Information Theory and Reliable Communication*. New York, NY: John Wiley and Sons, Inc., 1968.

[21] J. Proakis, *Digital Communications*, 3rd ed. McGraw Hill, 1995.