Combined Coding and Training for Unknown ISI Channels

Orhan Coskun and Keith M. Chugg, *Member, IEEE*

***Abstract*—The traditional method of sending a training signal to identify a channel, followed by data, may be viewed as a simple code for the unknown channel. Results in blind sequence detection suggest that performance similar to this traditional approach can be obtained without training. However, for short packets and/or time-recursive algorithms, significant error floors exist due to the presence of sequences that are indistinguishable without knowledge of the channel. In this paper, we reconsider training-signal design in light of recent results in blind sequence detection. Specifically, we consider the tradeoff between the complexity of receiver processing and the amount of training overhead required. More generally, we design *training codes* which combine modulation and training. In order to design these codes, we find an expression for the pairwise error probability of the joint maximum-likelihood (JML) channel and sequence estimator. This expression motivates a pairwise distance for the JML receiver based on principal angles between the range spaces of data matrices. The general code-design problem (generalized sphere packing) is formulated as the clique problem associated with an unweighted, undirected graph. We provide optimal and heuristic algorithms for this clique problem. For both long and short packets, we demonstrate that significant improvements are possible by jointly considering the design of the training, modulation, and receiver processing.**

***Index Terms*—Adaptive estimation, blind acquisition, clique algorithm, intersymbol interference (ISI), training codes.**

I. Introduction

DATA DETECTION in channels with intersymbol interference (ISI) is a well-studied problem, with maximum-likelihood sequence detection (MLSD) providing an optimal strategy if the channel is known at the receiver [2]. In many practical applications where MLSD is desirable, such as time-varying ISI channels in mobile radio systems, one must account for uncertainty in the channel. Most investigations and system designs fall into one of two extreme catergories: 1) *trained algorithms/systems*, which insert sufficient training signals to allow the receiver to reliably identify the channel; or 2) *blind algorithms/systems*, which attempt to identify the channel without the aid of training. The trained approach is a conservative design that allows relatively simple receiver processing, but sacrifices throughput or spectral efficiency. The blind approach is an aggressive design that relies on complicated receiver processing and may be susceptible to false-acquisition phenomena.

For trained systems, previous results have focused on the design of good training sequences [3], [4]. Such sequences can, for example, be used to initialize a Viterbi algorithm (VA) to implement MLSD under the assumption that the channel estimate is accurate or to initialize an adaptive approximation to MLSD, such as per-survivor processing (PSP)-based algorithms (e.g., [5]). Such approaches have been suggested for trained mobile radio systems such as GSM and IS-136 [5]–[7].

Improvement in throughput or spectral efficiency ostensibly gained by blind approaches may, in reality, be lost due to poor data-detection performance (or lost data) during a "pull-in" period. This acquisition period is typically hundreds to thousands of symbols for traditional blind linear equalizer structures (e.g., [8] and [9]). Similarly, many blind equalization/channel-identification approaches suggested in the literature (e.g., [10]–[15]) require the collection of relatively large data blocks, which limits their applicability to systems which use short data bursts and/or operate in time-varying channels.

Blind approaches based on joint maximum-likelihood (JML) sequence and channel estimation,[1] have shown the potential for rapid channel acquisition and excellent performance [15]–[17]. Strictly speaking, JML requires an exhaustive search process so that its complexity increases exponentially with sequence length [18]. While the performance of the suboptimal approximations of this process has been shown to approach that of the known channel case for large block lengths [15], application to short-burst communication has shown that misacquisitions are problematic [17]. Misacquisition for short packets is inherent in the JML optimality criterion [17]. Thus, an important challenge has been to understand the structure of the exhaustive search space associated with JML. In previous investigations, this structure has been characterized by the ability to distinguish sequences in the absence of observation noise [16], [17], [19]. Intuitively, one may expect that if sequences that have identifiability problems are disallowed for transmission, reliable acquisition and low training overhead could be simultaneously achieved. This leads to the combined training/coding approach described below.[2]

We consider the training to be part of the general signal-design problem. For example, consider transmission of N bits through an unknown channel. A trained system using $K < N$ bits of training may be considered a rate-$(N-K)/N$ code. However,

Paper approved by W. E. Ryan, the Editor for Modulation, Coding, and Equalization of the IEEE Communications Society. Manuscript received June 8, 2000; revised January 18, 2005. This paper was presented in part at the Allerton Conference on Communications, Control, and Computing, Allerton, IL, October 2000.

O. Coskun is with the Izmir Institute of Technology, Urla Izmir 35430, Turkey (orhancoskun@iyte.edu.tr).

K. M. Chugg is with the Communication Sciences Institute, Electrical Engineering Department, Viterbi School of Engineering, University of Southern California, Los Angeles, CA 90089-2565 USA (chugg@usc.edu).

Digital Object Identifier 10.1109/TCOMM.2005.852847

[1]We use JML as a short abbreviation for joint maximum-likelihood channel and sequence estimation.

[2]The authors wish to acknowledge the fact that after this paper was submitted for publication, we have become aware of related work, subsequently published in [20].

TABLE I
EFFECT OF TRAINING SEQUENCE SELECTION ON CHANNEL ESTIMATION ERROR VARIANCE FOR $L = 3$. VALUES SHOWN FOR THE WORST FULL-RANK SEQUENCE AND THE BEST TRAINING SEQUENCE

Training length	variance (worst)	variance (best)	lower bound	worst/best (dB)
6	1.25	3/4	3/4	2.2 dB
10	1.083	3/8	3/8	4.6 dB
15	1.045	0.2333	3/13	6.5 dB

since this is typically accomplished with the *segregated* approach of sending K consecutive training bits, followed by $(N-K)$ information-bearing bits, the code is actually a rate-zero code of block size K, followed by a rate-one code with block size $(N-K)$. The blind approach may be viewed as using a rate-one code. In this paper, we consider how one should select the 2^{N-K} possible N-bit sequences to best identify the channel and communicate the $(N-K)$ bits. We refer to such designs as *training codes*.

In Section II, the model and notation, together with traditional training-sequence design, is described. The tradeoff between receiver complexity and training overhead for long packets is explored via numerical experiments in Section III. In Section IV, we give a simple pairwise separability test for the noiseless case and a pairwise distance for the JML receiver in noise. These measures are used to determine the minimum amount of training overhead required to eliminate ambiguities in the absence of noise and to design *training codes* optimized for the JML receiver, respectively. Both of these are general code-design (packing) problems. We formulate this as a clique problem on a graph and provide an optimal algorithm to solve it, along with a greedy heuristic. These algorithms are used to design training codes for short-packet communications, which yield significant performance improvements relative to the traditional segregated approach. The three appendixes contain: 1) unification of the identifiability results in [19] and [17]; 2) the clique algorithm; and 3) the derivation of the distance measure and related quantities.

II. SIGNAL MODEL, TRADITIONAL TRAINING, AND JML DETECTION

After appropriate front-end filtering and sampling, the received signal $\{z_i\}_{i=1}^{k}$ can be modeled as

$$\mathbf{z}_k = A_k\mathbf{f} + \mathbf{w}_k \tag{1}$$

$$\mathbf{f} = [f_L f_{L-1} \cdots f_1]^T \tag{2}$$

$$A_k = \begin{bmatrix} a_{k-L+1} \cdots a_k \\ \vdots \quad \vdots \\ a_{-L+2} \cdots a_1 \end{bmatrix} \tag{3}$$

where $(\cdot)^T$ denotes transposition and $\mathbf{y}_k = [y_k\ y_{k-1}\ \cdots\ y_1]^T$ is the notation for a signal vector to time k (i.e., $\mathbf{z}_k$ and $\mathbf{w}_k$). The additive white Gaussian noise (AWGN) $\mathbf{w}_k$ is zero mean with covariance matrix $\sigma_w^2\mathbf{I}_k$. Convolution of the independent identically distributed (i.i.d.), uniformly distributed, digital sequence a_i with the finite support channel is represented by multiplication of the $(L \times 1)$ channel vector $\mathbf{f}$ by the Toeplitz $(k \times L)$ data matrix A_k. In the examples presented in this paper, a_k is drawn from $\{-1, +1\}$, which models binary phase-shift keying (BPSK), but the development is applicable to arbitrary M-ary constellations.

Except for the length of the impulse response,[3] the continuous-time channel is assumed to be unknown in this paper, so that the model of (1) arises by some form of filtering and sampling of a continuous-time observation. This may consist of an anti-aliasing filter or a pulse-matched filter, followed by a sampler and, possibly, a discrete-time noise-whitening filter. In order for this conversion to obtain an approximate set of sufficient statistics, fractionally spaced sampling is generally required. However, it has been shown that reasonable front-end processing with N_s samples per symbol leads to a model of the form in (1), where the components of $\mathbf{z}_k$, $\mathbf{w}_k$, and $\mathbf{f}$ are $(N_s \times 1)$ vectors [18], [21]. For simplicity of the presentation and reduced simulation effort, we work with the $N_s = 1$ simplified version. All results can be directly generalized to the oversampled case, and the qualitative results will be similar. Furthermore, the primary issues addressed in this paper are due to the structure of the data matrix which is independent of the sampling rate.[4]

A traditional approach to communicating over the channel is to first send a training sequence to estimate the channel $\mathbf{f}$, and then use this channel estimate to perform MLSD under the assumption of perfect estimation. With perfect knowledge of the channel, MLSD can be implemented via the VA [2]. Good training sequences are those that optimize an associated least-squares (LS) channel estimator. The unbiased LS channel estimate and the associated LS error for a length-N training sequence is

$$\hat{\mathbf{f}}_N = \left[A_N^T A_N\right]^{-1} A_N^T \mathbf{z}_N \tag{4}$$

$$\mathbb{E}\left\{\|\hat{\mathbf{f}}_N - \mathbf{f}\|^2\right\} = \sigma_w^2 \sum_{i=1}^{L} \frac{1}{\sigma_i^2} \geq \frac{\sigma_w^2 L}{N} \tag{5}$$

where $\sigma_1, \ldots, \sigma_L$ are the singular values of the training sequence[5] A_k (assumed to be rank L) and $\mathbb{E}\{\cdot\}$ denotes ensemble averaging. The lower bound in (5) is obtained if and only if (iff) all singular values are equal. Because the training signal is typically drawn from the same modulation alphabet as the data, the lower bound in (5) may not be obtainable. Table I shows the variation of the estimation-error variance with training sequences length N and singular-value spread. Note that for a three-tap channel, the error variance in (5) for the best sequence is reduced by 3 dB when increasing the training sequence length from 6 to 10 b, while the corresponding reduction is 5.1 dB when increasing the length from 6 to 15 b.

[3]The approaches discussed throughout are robust to overestimation of L, but sensitive to underestimation of L.

[4]If one attempts to exploit the structure of $\mathbf{f}$ induced by the known pulse shaping [i.e., not just a fractionally spaced version of (1)], then oversampling may alter qualitative conclusions.

[5]We will refer to matrices of the form in (3) as sequences, since there is a one-to-one mapping between the sequence and matrix representation. Also, A_k will denote the actual transmitted sequence with hypothesized or conditional versions denoted by $\bar{A}_k$ and or $A_k^{(i)}$.

When the channel is unknown, application of the VA using an estimate of the channel is not MLSD [22], and an exhaustive search of possible sequences is required to implement the JML receiver, which attempts to minimize

$$\Lambda_k(\tilde{A}_k;\mathbf{z}_k)=\left\|\left(I-\tilde{A}_k\tilde{A}_k^I\right)\mathbf{z}_k\right\|^2=\left\|(I-\tilde{P}_k)\mathbf{z}_k\right\|^2 \quad (6)$$

over all possible allowable sequences $\tilde{A}_k$. The matrix $\tilde{A}_k^I$ is the pseudoinverse of $\tilde{A}_k$, which is $(\tilde{A}_k^T\tilde{A}_k)^{-1}\tilde{A}_k^T$ for full-rank $\tilde{A}_k$, and $\tilde{P}_k=\tilde{A}_k\tilde{A}_k^I$ is the matrix[6] that projects onto the range of $\tilde{A}_k$. Thus, the JML receiver may be interpreted as a deterministic version of the estimator-correlator receiver [23], since it attempts to find the combination of $\tilde{A}_k$ and $\hat{\mathbf{f}}(\tilde{A}_k)=\tilde{A}_k^I\mathbf{z}_k$ that best aligns with $\mathbf{z}_k$. Note that even with training, the JML receiver should perform an exhaustive search of all possible sequences. The initial training sequence simply eliminates a subset of sequences from the search. With sufficient training, however, the channel is well estimated after the training, and the performance difference between the exhaustive search and a Viterbi or PSP-based approach should be small.

Practical, forward-only approximations to the JML receiver can be based on PSP and its generalized version (e.g., [5] and [24]). Specifically, the JML metric in (6) can be computed recursively [21, eq. (3)], converting the problem into a tree-search with per-sequence recursive LS (RLS) channel estimation. Any search algorithm can be applied to this exponentially growing tree. We refer to the case when the VA with M^{L-1} states is applied as a suboptimal search strategy as PSP, and use the term generalized PSP (G-PSP) for other search algorithms. It was shown in [17] that significant improvements in blind acquisition performance can be obtained by increasing the tree-search complexity.

III. Overhead Complexity Tradeoff for Long Packets

In this section, we compare the performance of several training and receiver processing strategies illustrated in Fig. 1. In all cases, a burst of N symbols is sent at the beginning of the packet to aid with channel identification, followed by data symbols and $L-1$ tail symbols to terminate the channel state. We consider two types of training sequences, a traditional length-N sequence, and a *split-training* sequence with K of the N symbols fixed for training, and the other $(N-K)$ freely used to convey data. Thus, the traditional approach is a special case of the split-training approach with $K=N$. Therefore, $K+L-1$ of the P symbols are overhead, and $D=P-K-L+1$ are data symbols.

Three types of receivers are also considered: 1) the VA with standard trained initialization; 2) the PSP algorithm with standard trained initialization; and 3) the PSP/G-PSP algorithm with split training and exhaustive initialization. With standard trained initialization, the initial channel estimate and the initial trellis state are determined by the known, length-N training sequence. Receiver 1) uses this initial channel estimate throughout the packet, while the receiver 2) updates this estimate in the standard PSP manner. In receiver 3), an exhaustive JML search is performed over the first Q symbols to initialize at the G-PSP receiver, with M^{L_t-1} states with $L_t\geq L$ and $Q\geq N$. This

[6]For compactness, we do not show the dependence of the projection matrix $\tilde{P}_k$ on $\tilde{A}_k$.

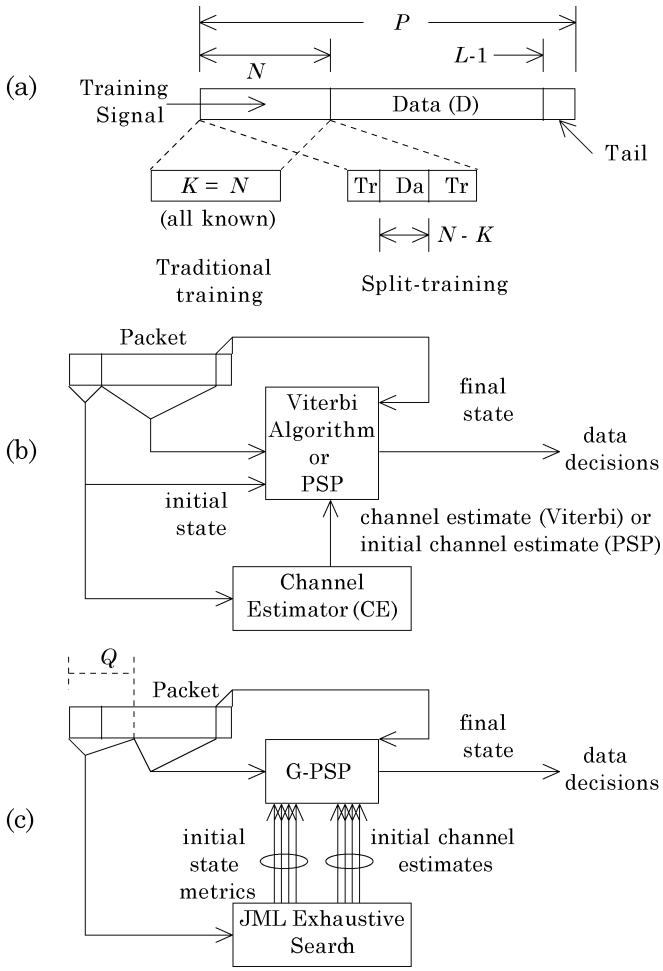

Fig. 1. Configuration of long-packet investigation. (a) Traditional approach and the split-training approach. (b) Standard initialization for a PSP or Viterbi processor. (c) Exhaustive JML initialization up to time Q for a generalized PSP receiver.

receiver considers M^{Q-K} different sequences during the initialization process, and for each state in the G-PSP trellis, selects the best of the M^{Q-K-L_t-1} sequences entering that state at time $Q+1$.

Simulations were run to assess the relative effectiveness of these approaches. Unless otherwise stated, all simulations in this paper were run under the condition that $\|\mathbf{f}\|=1$, $a_k\in\{-1,+1\}$, and $\sigma_w^2=(N_0R/2E_b)=(N_0/2E_s)$, where R is the rate of the system accounting for overhead, and E_b (E_s) is the energy per bit (symbol) accounting for this overhead. From the above description, the rate is $R=D/P=(P-K-L+1)/P$ for the current example. Care must be taken when making such rate comparisons over non-AWGN channels. For example, if the transmission bandwidth is doubled through the use of a rate-1/2 code, the length of the ISI, measured in terms of channel symbols, is also doubled. In this paper, we assume that the length of the ISI remains fixed when the rate of the training code is varied. We adopt this convention mainly as a convenient way to include the effect of different overhead rates, and other conventions may be more appropriate, based on the specific application. In general, this is a valid assumption if R is nearly one. For rates significantly lower than unity, this assumption holds when the rate change is accomplished by some means other than bandwidth

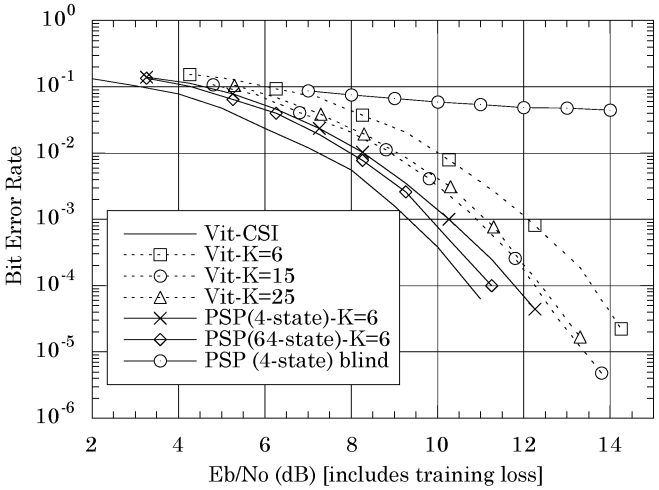

Fig. 2. Tradeoff between training overhead and receiver complexity for relatively long packets. Simulation parameters: $\mathbf{f} = [1\ 2\ 1]^T/\sqrt{6}$, $D = 60$ information-bearing bits, and RLS forgetting factor $\rho = 0.9$.

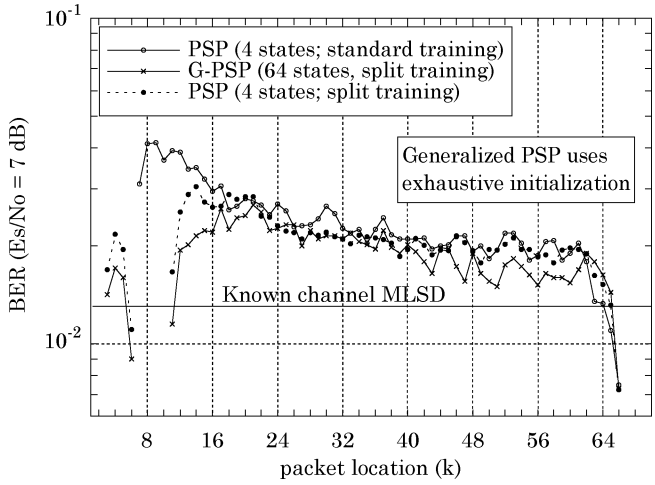

Fig. 3. BER versus packet location for split training and standard training. Simulation parameters: $\mathbf{f} = [1\ 2\ 1]^T/\sqrt{6}$, $D = 60$ information-bearing bits, RLS forgetting factor $\rho = 0.9$, and $E_s/N_0 = 7$ dB.

variation. Consider, for example, a time-division multiple-access (TDMA)-based system that uses eight time slots, each with 20 b of training and 40 b of data. If one could effectively enable the use of 8 b of training overhead for the same 40 data bits, then the eight 60-b slots could be replaced by ten 48-b slots. Thus, the reduction in training overhead yields a 25% increase in throughput or capacity without changing the channel bandwidth, and thus, the ISI channel.

Results are shown in Figs. 2 and 3 for various systems simulated with $L = 3$, $\mathbf{f} = [1\ 2\ 1]^T/\sqrt{6}$, and $D = 60$. For the system of Fig. 1(a), we simulated standard training and initialization with $K = N = 6$, 10 (not shown), 15, 20 (not shown), 25. Also shown in Fig. 2 is the case where PSP is started with an all-zero initial channel estimate. Larger training sequences provide better channel estimation, and thus bit-error rate (BER), but the resulting overhead reduces the rate R. The results in Fig. 2 suggest that, under this scenario, the best performance is achieved with $N = 15$. With the standard four-state PSP receiver using standard training initialization and an RLS forgetting factor of $\rho = 0.9$, a length-6 training sequence yields an improvement of approximately 0.5 dB in E_b/N_0 with respect to the best trained system using the VA ($K = 15$). Thus, additional receiver complexity yields improved performance with less overhead. The final performance curve in Fig. 2 corresponds to the G-PSP approach with exhaustive initialization and split training. Specifically, the split-training format was $N = 10$ and $K = 6$, with two training bits followed by four data bits and another four training bits.[7] The training bursts were designed separately according to the traditional methodology described in Section II. A 64-state ($L_t = 7$) trellis was used for the G-PSP processor [17] with $Q = 16$. So, $\{z_i\}_{i=1}^{16}$ were observed, and $2^{(N-K)+(Q-N)} = 2^{4+6} = 1024$ sequences were considered for the exhaustive JML initialization. One of $2^{10}/64 = 16$ sequences entering each state was selected according to the JML metric in (6). Compared with the system with standard training and PSP, this significant increase in complexity yields an improvement of approximately 0.5 dB. Note that this system performs within 0.5 dB of the perfect channel state information (CSI) case, and 0.4 dB of this degradation is irreducible, due to the $K = 6$ overhead.

Fig. 3 shows the performance of the PSP and G-PSP systems described above versus packet location at an $E_s/N_0 = 7$ dB. The PSP with split-training system is a standard four-state PSP algorithm with the same $N = 10$, $K = 6$ split-training sequence used for the G-PSP scheme, but the initialization is simplified. Specifically, $Q = 10$ is used, so that $\{z_i\}_{i=1}^{10}$ were used to compute the JML metric of $2^{N-K} = 16$ sequences. Each of these 16 sequences terminate in the same state, so the best is selected to initialize the PSP state and channel estimate. In [17], where no training was used and G-PSP algorithms were started blindly, increasing L_t was found to improve the performance, but performance was relatively poor for the first 15–20 b. Using training, of course, improves the performance at the beginning of the packet. However, using split training is superior to traditional training for two reasons. First, detection of the $N - K$ bits between the training bursts can be performed reliably, since the symbols on either side of this burst are known. Second, assuming that the $2^{N-K} = 16$ different sequences can be reliably distinguished, then the effective training length is $N = 10$. However, even for the split-training approach, the BER of the standard PSP is large for symbols at locations just after N. This is presumably caused by decision errors on the 4 b in between the training bursts, which causes the PSP algorithm to be initialized with a poor channel estimate. In contrast, the G-PSP receiver that uses the same split-training format, but performs exhaustive search up to time 16 and 64 states, alleviates this effect, and has approximately constant BER over the packet length.[8] This BER is slightly worse than the known channel MLSD performance, due to steady-state estimation error associated with PSP-RLS channel estimators.

In this section, we have demonstrated that receiver complexity can be traded for training overhead and that different training schemes with the same overhead can yield different performance. In the next section, we formalize this optimization process and demonstrate it for short-packet systems.

[7]These 4 b are counted in the total of D information bits.

[8]The final bits are more reliably detected due to the tail bits.

IV. Training Code Design Using Separability and Distance Conditions

In this section, we describe a simple check to determine whether a pair of sequences can be distinguished in the absence of noise under any channel. This can be used to determine the minimum amount of training required for a JML receiver. Similarly, for the case that includes observation noise, we develop a distance measure to be used for training-code design. In Appendix I, we summarize and relate the tests to characterize separability (i.e., without noise) suggested in [17] and [19]. Neither of these identifiability tests has all the desirable properties for our training-code design methodology. Before describing the separability check and distance, we describe this methodology and the type of design criteria that are most appropriate.

A. Code-Design Methodology

We propose to design length-N training codes by starting with all possible 2^N sequences, and discarding sequences until the largest set of sequences meeting a certain pairwise condition is obtained. We then select the largest integer D such that 2^D is less than or equal to the size of this largest set, yielding an (N, D) code. In the noise-free case, the pairwise condition is separability, while with noise, it is a minimum distance requirement. Toward this goal, the condition used to discard sequences should 1) account for the digital nature of allowable sequences, 2) be a pairwise property, 3) be symmetric, 4) be relatively simple to check, and 5) be independent of the actual channel coefficients $\mathbf{f}$. The property 1) allows for larger code rates, since potential codewords are not discarded due to possible confusion with an "analog" sequence (i.e., see Appendix I). Properties 2)–4) allow one to discard sequences one at a time in a simple manner. The last property ensures that the design is robust to the unknown channel.

Given some condition with the above properties, we can formalize the code-design problem as finding the largest complete subgraph (or clique) on an undirected, unweighted graph, i.e., the so-called *clique problem* [25]. Specifically, consider a graph representation of the code-design problem with each of the 2^N potential codewords defining a vertex. An edge exists between two vertices if the associated sequences satisfy the condition. Due to the above properties, the resulting graph $G = (V, E)$ (i.e., vertices V and edges E) is undirected and unweighted. The design problem is to find the largest subset $\mathcal{C}$ of the $|V| = 2^N$ vertices for which the pairwise condition is satisfied for each pair in $\mathcal{C}$.

The general class of clique problems is known to be NP-hard [25]. However, in Appendix II, we present an optimal algorithm which enumerates all cliques in an efficient manner when the maximum number of edges for any vertex is much smaller than $|V|$. We also describe a (suboptimal) greedy heuristic to approximate the solution to the clique problem for cases when the optimal algorithm is prohibitively complex.

B. Noise-Free Separability

It has been noted that there are sequences that cannot be separated from some other sequences by the JML receiver in (6), even in the absence of noise [17], [19]. Specifically, we say that A_k is *separable or distinguishable* from A'_k under channel $\mathbf{f}$ if when A_k is transmitted, $\Lambda_k(A'_k; \mathbf{z}_k) > \Lambda_k(A_k; \mathbf{z}_k)$, i.e., that the correct sequence is decided by the rule in (6). The existence of indistinguishable sequences induces an error floor for JML, so that a BER of zero is not achieved, even when $E_b/N_0 \to \infty$. Based on all evidence available [15]–[17], [19], this error floor decreases rapidly with the length of the observation interval. Thus, an interesting issue is to characterize the minimum overhead required to reliably identify the channel when no noise is present. For example, for BPSK signaling, at least one bit of redundancy is required to eliminate the sign ambiguity.

More generally, addressing this issue requires a simple check for pairwise separability under any channel $\mathbf{f}$. Specifically, we define two sequences to be *pairwise separable* if they are each separable from the other under any nonzero channel $\mathbf{f}$.[9] A necessary and sufficient condition for pairwise separability is that the range spaces of $A^{(i)}$ and $A^{(j)}$ have dimension[10] L and share only the origin, i.e., $\dim[\mathcal{R}(A^{(i)}) \cap \mathcal{R}(A^{(j)})] = 0$ [17]. As discussed in Appendix I, such a check is not available from the related previous research [17], [19]. A simple test is provided by the following theorem.

Theorem 4.1: Two digital sequences $A^{(i)}$ and $A^{(j)}$ are pairwise separable iff the matrix $C = [A_k^{(i)}\ A_k^{(j)}]$ has rank $2L$.

Proof: If C has rank $2L$, the set of $2L$ columns of C must be linearly independent, which implies that the columns of $A^{(i)}$ and the columns of $A^{(j)}$ must be linearly independent sets. Thus, no point in $\mathcal{R}(A^{(i)})$ other than the zero vector can be expressed in terms of the columns of $A^{(j)}$, and vice-versa. So if C has rank $2L$, then $\dim[\mathcal{R}(A^{(i)}) \cap \mathcal{R}(A^{(j)})] = 0$, and the two sequences are pairwise separable. To show the converse, since the ranges are disjoint, excluding the zero vector, the dimension of their union (i.e., the range space of C) must be equal to their sum, namely, $2L$. ■

Example of Noise-Free Design Methodology: As an example of the applicability of the above result, consider transmission of a BPSK sequence of length $N = 8$ through an $L = 3$ channel. First, consider the approach where only 1 b of overhead is used to remove the sign ambiguity by fixing the first bit to 1. Considering 400 different channels equally spaced on the unit hemisphere,[11] the BER associated with this technique is 0.04 conditioned on the "best sequence" being sent, and as high as 0.5 conditioned on the "worst" sequences. Thus, even in the absence of noise, a significant floor for the average BER of approximately 0.3 is reached, because not all of the $2^7 = 128$ sequences are pairwise separable.

The optimal clique algorithm described in Appendix II was run on a graph with 128 vertices[12] to determine the maximum number of sequences that can be decoded without any error in the absence of noise. Two vertices were connected if the associated sequences were pairwise separable. The

[9]This was called "completely distinguishable" in [17].

[10]If the rank of $A^{(i)}$ was less than L, then there would be nonzero $\mathbf{f}$ such that $A^{(i)}\mathbf{f} = \mathbf{0}$. Thus, when $A^{(i)}$ is transmitted with this channel and no noise, both metrics are zero and the two sequences cannot be distinguished.

[11]Henceforth referred to as "all channels" when describing simulation results.

[12]One bit location was fixed to reduce the number of vertices from 256 to 128.

maximum training-code size consisting of all pairwise-separable sequences was determined to be eight. Moreover, 882 size $2^D = 8$ training codes were found. Since the largest pairwise-separable training code has $2^3 = 8$ codewords, $(N-D) = 5$ b of overhead are required to uniquely determine which one of the length-8 sequences is sent without any restriction on the channel coefficients. In the absence of noise, such a system will perform without error (i.e., the error floor has been removed). The BER in the presence of noise, however, for each of the 882 pairwise-separable codes can be expected to vary. We simulated several of these 882 possible $R = 3/8$ training codes in the presence of noise, and although the results are not shown here, we observed variations in performance of up to 1 dB in SNR.

C. A Pairwise Distance for JML Channel and Sequence Estimation

Next, we introduce a distance criterion for JML that accounts for AWGN. The distance is related with the conditional second-order statistics of the random variable[13]

$$\begin{aligned} r_k(i,j) &= \Lambda_k\left(A_k^{(j)};\mathbf{z}_k\right) - \Lambda_k\left(A_k^{(i)};\mathbf{z}_k\right) \\ &= \mathbf{z}_k^T\left(P_k^{(i)} - P_k^{(j)}\right)\mathbf{z}_k. \end{aligned} \tag{7}$$

In a pairwise JML decision between $A_k^{(i)}$ and $A_k^{(j)}$, the former is selected iff $r_k(i,j) \geq 0$. In Appendix III, it is shown that conditioned on $A_k^{(i)}$ being sent and a particular $\mathbf{f}$, $r_k(i,j)$ can be represented as a difference of two sums, each sum consisting of L mutually independent *noncentral chi-square* random variables. It follows that the conditional probability density function (pdf) of $r_k(i,j)$ is the convolution of $2L$ densities of the χ^2 form. Therefore, for a given channel $\mathbf{f}$ and $A_k^{(i)}$, the pairwise error probability (PEP) with $A_k^{(j)}$ can be found by numerical integration of this pdf function. A reasonable distance between $A_k^{(i)}$ and $A_k^{(j)}$ could be defined based on such PEPs. However, this is numerically intensive, and still should be made independent of the channel by averaging over all channels, or considering the worst case.

We define a distance that is relatively simple to compute, and its minimum value over unit norm channels is easily found. In Appendix III, the conditional mean and the variance of $r_k(i,j)$ are shown to be

$$\begin{aligned} m_r\left(A_k^{(j)}|A_k^{(i)};\mathbf{f}\right) &\triangleq \mathbb{E}\left\{r_k(i,j)|A_k^{(i)};\mathbf{f}\right\} \\ &= \sum_{l=1}^{L} c_l^2 \sin^2\theta_l \end{aligned} \tag{8}$$

$$\begin{aligned} \sigma_r^2\left(A_k^{(j)}|A_k^{(i)};\mathbf{f}\right) &\triangleq \mathrm{var}\left[r_k(i,j)|A_k^{(i)};\mathbf{f}\right] \\ &= 4\sum_{l=1}^{L}\sin^2\theta_l\sigma_w^4 \\ &\quad + 4\sigma_w^2 m_r\left(A_k^{(j)}|A_k^{(i)};\mathbf{f}\right) \end{aligned} \tag{9, 10}$$

where $\{\theta_i\}_{i=1}^{L}$ is the set of *principal angles* [26] between the range spaces of $A_k^{(i)}$ and $A_k^{(j)}$, and c_l is the coefficient vector $A_k^{(i)}\mathbf{f}$ represented with respect to a particular basis (see Appendix III).

[13]The dependency of $r_k(i,j)$ on $\mathbf{f}$ and $\mathbf{z}_k$ is not denoted explicitly.

Using the exact expressions in (8)–(10), we will propose a pairwise distance measure, based on $m_r(\cdot)/\sigma_r(\cdot)$, that serves as a proxy for the PEP for the worst-case channel. To motivate this distance, consider that if $r_k(i,j)$ were Gaussian, the pairwise probability of error would be $\mathrm{Q}(m_r(\cdot)/\sigma_r(\cdot))$. Thus, intuitively, a large value of $m_r(\cdot)/\sigma_r(\cdot)$ should reduce the conditional PEP. To illustrate this further, the conditional pairwise error rate (PER) was simulated for a pair of separable sequences as the channel traced the unit hemisphere. For all channels of length three, two sequences of length eight were sent 4×10^5 times through an AWGN channel with $\sigma_w^2 = 0.1$. Fig. 4 shows the simulation results of the PER versus $m_r(\cdot)/\sigma_r(\cdot)$. Also shown is the function $Q(m_r(\cdot)/\sigma_r(\cdot))$, which was empirically found to be a good approximate upper bound for the conditional PER. We performed similar experiments for 20 different sequence pairs of lengths eight, nine, and ten at channel lengths three and four. In all cases, similar results were obtained; $m_r(\cdot)/\sigma_r(\cdot)$ characterized the PER, and the Q-function provided an good approximate upper bound. Although we have not shown this to be a valid upper bound, some intuitive justification is provided by the fact that the pdf of the decision statistic $r_k(i,j)$ is asymmetric around its mean, with the heavier tail in the direction away from the decision boundary.

The distance that we suggest is obtained by finding the smallest value of $m_r(\cdot)/\sigma_r(\cdot)$ considering all unit norm channels and either $A_k^{(i)}$ or $A_k^{(j)}$ as the transmitted sequence. It can be shown that $m_r(\cdot)/\sigma_r(\cdot)$ is an increasing function of $m_r(\cdot)$, so that the minimum value of $m_r(\cdot)/\sigma_r(\cdot)$ is achieved at the minimum value of $m_r(\cdot)$. While a simple lower bound on $m_r(\cdot)$ (see (49) in Appendix III) could be used to define a distance criteria, it is relatively simple to obtain the explicit minimum value of $m_r(\cdot)/\sigma_r(\cdot)$ for a unit norm channel. Let $\lambda_{\min}^{(j|i)}$ be the smallest eigenvalue of $[A_k^{(i)}t]^T(P_k^{(i)} - P_k^{(j)})A_k^{(i)}$. This leads to the minimum value of the $m_r(\cdot)/\sigma_r(\cdot)$ and the minimum conditional distance

$$d\left(A_k^{(j)}|A_k^{(i)}\right) \triangleq \min_{\mathbf{f}:\|\mathbf{f}\|=1}\left[\frac{2m_r\left(A_k^{(j)}|A_k^{(i)};\mathbf{f}\right)}{\sigma_r\left(A_k^{(j)}|A_k^{(i)};\mathbf{f}\right)}\right] \tag{11}$$

$$= \frac{\frac{\lambda_{\min}^{(j|i)}}{\sigma_w^2}}{\sqrt{\sum_{l=1}^{L}\sin^2\theta_l + \frac{\lambda_{\min}^{(j|i)}}{\sigma_w^2}}}. \tag{12}$$

Finally, we define the distance as

$$d\left(A_k^{(i)},A_k^{(j)}\right) \triangleq \min\left[d\left(A_k^{(i)}|A_k^{(j)}\right), d\left(A_k^{(j)}|A_k^{(i)}\right)\right] \tag{13}$$

so that the approximate upper bound for the pairwise error is $Q(d/2)$. The distance in (13) is an approximate minimax criterion, because the minimum value of $m_r(\cdot)/\sigma_r(\cdot)$ approximately determines the maximum error probability. Note that the distance is a function of $\lambda_{\min}^{(j|i)}/\sigma_w^2$, which may be viewed as a measure of SNR for the quadratic detector.

Computation of the distance in (13) requires the following.

- Computation of the minimum eigenvalue of $[A_k^{(i)}]^T(P_k^{(i)} - P_k^{(j)})A_k^{(i)}$ and $[A_k^{(j)}]^T(P_k^{(j)} - P_k^{(i)})A_k^{(j)}$, $\lambda_{\min}^{(j|i)}$ and $\lambda_{\min}^{(i|j)}$, respectively.

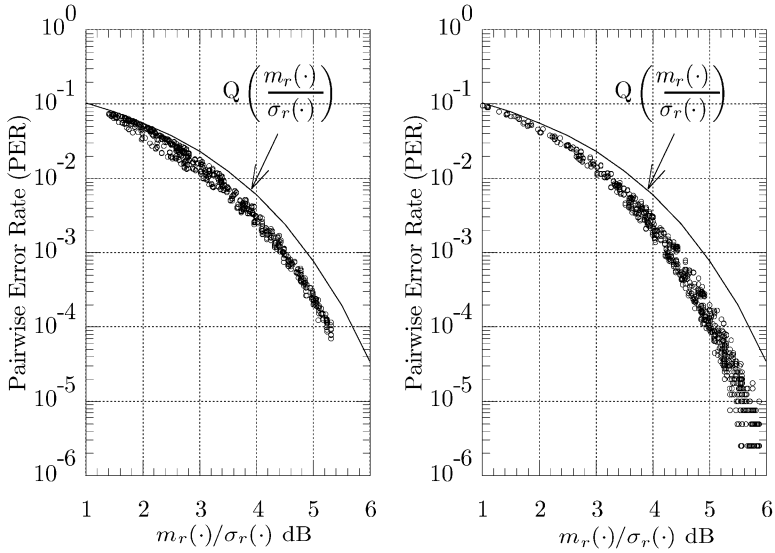

Fig. 4. Conditional PER performance of JML for 400 three-tap channels selected on the unit hemisphere as a function of $m_r(\cdot)/\sigma_r(\cdot)$. Shown under the condition of either sequence transmitted and $\sigma_w^2 = 0.1$. The empirical upper bound $\mathrm{Q}(m_r(\cdot)/\sigma_r(\cdot))$ is also shown.

- Determination of an orthonormal basis for $\mathcal{R}(A_k^{(i)})$ and for $\mathcal{R}(A_k^{(j)})$. This can be computed, for example, by performing a QR-decomposition on $A_k^{(i)}$ and $A_k^{(j)}$, respectively [26].
- A singular value decomposition (SVD) of an $(L \times L)$ matrix of inner products between the basis vectors obtained from the above step. This yields the principal angles.
- Use the above eignevalues, eigenvectors, and principal angles to compute the distance via (12) and (13).

1) Example of Noisy Design Methodology: We consider a simple example to illustrate the appropriateness of this distance. First, we design two sets of length $N = 8$ sequences, each of size three, for an $L = 3$ channel. These two sets are designed to be good or bad with respect to a known channel receiver or a JML receiver. Specifically, for a known channel $\mathbf{f}$, the pairwise error is monotonically decreasing in $\|(A_k^{(i)} - A_k^{(j)})\mathbf{f}\|$. Thus, analogous to the distance in (13) for the known-channel case, we define a known-channel distance that is the minimum of $\|(A_k^{(i)} - A_k^{(j)})\mathbf{f}\|/(2\sigma_w^2)$ over all unit norm channels, which is simply the minimum singular value of the matrix $A_k^{(i)} - A_k^{(j)}$ divided by $2\sigma_w^2$.

Set A was designed by setting the minimum value of the distance in (13) between any pair to 2.19 and the maximum known-channel distance to 3.7. Similarly, Set B was found by setting the maximum distance in (13) between any pair to 1.6 and the minimum known-channel distance to 4. All distances were computed at $\sigma_w^2 = 0.1$. The optimal clique algorithm was run on a graph of 128 vertices to determine the sets. The maximum clique size for each was found to be three, so each set contains three sequences. Fig. 5 shows the average codeword-error rate (CER) obtained by simulation of the JML receiver for the unknown-channel case, and the ML receiver for the known-channel case. As expected, Set A is superior in the unknown-channel case, and inferior in the known-channel case. This further validates the distance in (13), and distinguishes the approach from known channel signal design.

As another example, the distance criterion given in (13) was used to design a training-coded system for a packet length of

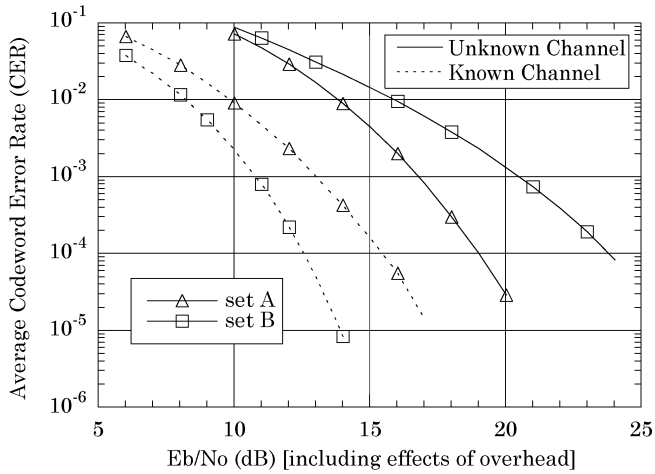

Fig. 5. CER for two sets of three sequences averaged over all three-tap channels. Set A was designed to have good unknown-channel performance, and set B was designed to have good known-channel performance.

$N = 16$. A codebook of size $2^7 = 128$ was found by running the greedy clique algorithm on a graph with 2^{15} vertices. Thus, a (16,7) training code was designed. Each sequence was sent 1000 times over all channels with the results summarized in Fig. 6. Unless specified, decoding was performed using the JML via exhaustive search.

The performance of this combined coding and training system is compared with two traditional training systems. Trained system A uses the 9-b overhead as a training sequence at the beginning of the packet. The training sequence was selected as one of the optimal training sequences, as described in Section II. This segregated design performs very poorly, as compared with the combined design, due to the two bits at the end of the packet. A second segregated design was considered, which uses 7 b of training at the beginning of the packet, with the last two symbols of the packet fixed to terminate the channel state in a known manner. This trained system B performs much better than the first segregated design, but there is a 2.8-dB loss in SNR relative to the performance of the combined design at a CER value 10^{-3}. The performance of trained system

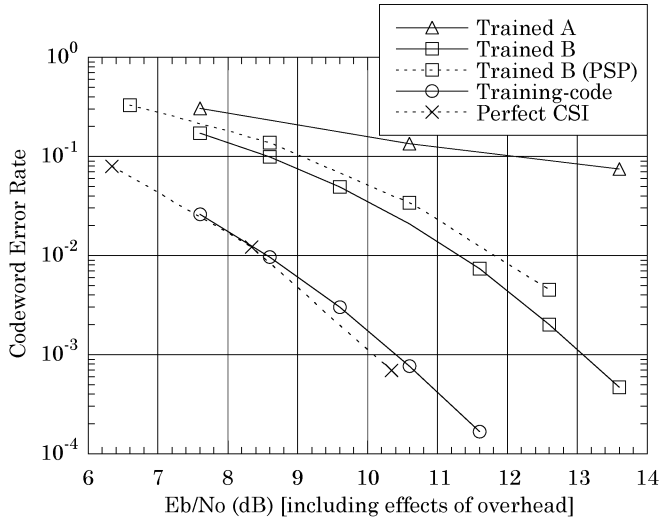

Fig. 6. Comparison of segregated and combined training and modulation. All systems use a packet length of $N = 16$, with the trained and training-code systems conveying seven information bits per packet. Performance is averaged over all three-tap channels.

B with segregated processing is also shown. Specifically, a standard (four-state) PSP receiver initialized by the 7-b training sequence and terminated by the tail bits shows only slight degradation relative to the JML receiver in this case. Also shown in Fig. 6 is the performance of a known-channel MLSD receiver averaged over all channels. This curve was generated by sending $N = 16$ b, with the first and last two bits specified to initialize and terminate the channel state, respectively. The performance of the system using the training code is similar to that of the perfect CSI system, despite the SNR penalty due to an overhead increase (i.e., 2.3 dB). This suggests that, in principle, one could obtain coding gain relative to an uncoded, known ISI channel while jointly estimating the channel.

V. Concluding Remarks

Traditional training approaches may be viewed as simple codes for unknown channels that require relatively simple receiver processing at the expense of decreased rate or throughput. In this paper, we have established a framework for the design of training codes based on a pairwise distance that is appropriate for JML channel and sequence estimation, or approximations thereof. For short data-packet lengths, the resulting training codes were demonstrated to provide significant improvements over traditional segmented approaches at the expense of increased receiver complexity. For longer packet lengths, the direct application of this framework becomes prohibitively complex. However, reliable data communication starting from the beginning of the (long) packet was demonstrated with very low overhead using split-training sequences and additional search effort during acquisition.

Clearly, the major obstacle to applying the approach developed for training-code design to larger block lengths is the exponential complexity growth in both the code design and the decoding problems. However, one could use a short training code designed using the approach suggested to initialize a PSP-based algorithm for a long packet. The practical value for such an approach, as compared against a standard training approach or a split-training approach for long packets, is questionable, because the importance of saving a small number of overhead bits diminishes with the packet length. For example, saving 3 b of overhead in a 16-b packet is substantial, but saving 3 b of overhead in a packet of 100 b has a less impressive impact on throughput. In practice, the packet length or the number of bits transmitted between training is determined by a number of factors, including the dynamics of time variation in the channel and the multiple-access method.

An interesting area for future research is the formulation of the design in a recursive manner that would allow simplified design and decoding. Also, incorporation of additional levels of error-correction coding using recent adaptive iterative detection techniques [27] is another promising research direction.

Appendix I
Previous Identifiability/Separability Tests and Their Relation

Two approaches to characterizing separability have been developed independently by Gustafsson and Wahlberg [19] and Chugg [17]. Both approaches attempt to characterize sequences which, when transmitted, may (or will) cause a blind receiver to have problems distinguishing it from another allowable sequence. We briefly describe these approaches and relate the main concepts.

In [17], only digital sequences were considered. A sequence A_k was defined to be *distinguishable from* A'_k *under channel* $\mathbf{f}$ if $A_k\mathbf{f} \notin \mathcal{R}(A'_k)$. Furthermore, a sequence A_k that is distinguishable from all sequences except those in $\mathcal{T}(A_k) \triangleq \{A'_k : A'_k = cA_k \text{ for some } c\}$ is identifiable under the definition from [19]. The *equivalence class* of a digital sequence A_k is all other such sequences with the same range space, with the trivial equivalence class defined as $\mathcal{T}(A_k)$. Sequences with a nontrivial equivalence class cannot be distinguished from some other sequence under any circumstance under the metric in (6). Such sequences were characterized as having some periodic structure in [17]. Furthermore, a sufficient condition was given [17, Lemma 7] for a BPSK sequence A_k to be distinguishable from all $A'_k \notin \mathcal{T}(A_k)$ under a class of channels that were termed "*regular*" with respect to the digital alphabet. For BPSK modulation, this condition is that all of the 2^L possible rows occur on the "interior" of A_k.

The approach in [19] is based on conditions for analog sequences X_k (i.e., a Toeplitz matrix with elements taking values on the continuum). Clearly, rank-deficient data matrices cause channel identification problems. Considering only full-rank matrices X_k, the solution of

$$\Gamma = P_{X'_k} X_k - X_k = 0 \tag{14}$$

for X'_k, where $P_{X'_k}$ represents the projection matrix of X'_k, gives the set of sequences which are equivalent to X_k. In [19], it was shown that for a channel length L, if

$$B_k(\mathbf{x}) = \begin{bmatrix} x_{k-2L+2} \cdots x_k \\ \vdots \quad \vdots \\ x_{-L+2} \cdots x_L \end{bmatrix} \tag{15}$$

is a rank-$2L-1$ matrix (*in the language of* [19] X_k *is persistently exciting (PE) of order* $2L-1$), then X'_k can always

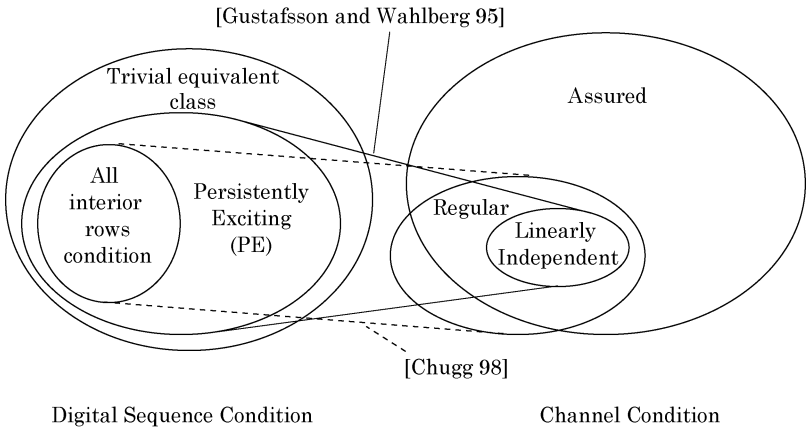

Fig. 7. Relation between the identifiability checks in [17] and [19]. Note that the relation between regular and assured channels is a conjecture.

be represented as cX_k for some constant c. Therefore, for digital sequences a_k, $B_k(\mathbf{a})$ being a full-rank matrix is a sufficient condition for A_k to have a trivial equivalence class. In general, however, this is not sufficient for A_k to be identifiable with respect to all channels. We next give the definition of a class of channels implied in [19].

Definition 1.1: Assured channels are those for which $(P_{A'_k}A_k - A_k)\mathbf{f} = 0$ iff $P_{A'_k}A_k = A_k$ holds for all A_k and A'_k rank-L digital data matrices with A_k being PE of order $2L-1$.

It follows that a trivial equivalence class is a sufficient condition for identifiability, with respect to assured channels. Thus, focusing on assured channels eliminates the possibility of the channel-dependent indistinguishabilities described in [17]. In [19], a subset of assured channels was given as "*linearly independent channels*" which has much simpler characteristics. In the remainder of this appendix, we show that the set of linearly independent channels is a subset of the set of regular channels, whereas the set of sequences satisfying the "all-rows" condition [17] is a subset of the set of sequences which are PE of order $2L-1$ for a length-L channel. These conclusions are summarized in Fig. 7. Comparing the check for identifability in [19] to that in [17], a larger class of digital sequences was considered with respect to a smaller class of channels. Note that we have not proven any relation between regular channels and assured channels, but we conjecture that the relation is as implied by Fig. 7. Finally, we note that neither reference gives a simple, *pairwise* separability check of the form desired for our code-design methodology.

Theorem 1.1: The set of linearly independent channels is a proper subset of the set of regular channels.

Proof: Let the data belong to the finite alphabet $\mathcal{A} = \{\pm 1, \pm 3, \ldots, \pm(M-1)\}$ and assume the channel is not regular, but is linearly independent. From the definition in [19], this means that the channel cannot output a zero when the input is drawn from the set

$$Q_L(\mathcal{A}) = \Big\{0, \pm 1, \pm 2, \cdots, \pm 2(M-1)\ldots \pm(2M-1)^{2L+1}L^{\frac{L}{2}}(k_0 - L + 1)^L\Big\}. \quad (16)$$

Since the channel is not regular, there exist a pair of $(L \times 1)$ vectors $\mathbf{a}$ and $\mathbf{b}$ with elements from $\mathcal{A}$, such that $\mathbf{a}^T\mathbf{f} = \mathbf{b}^T\mathbf{f}$ or $(\mathbf{a}-\mathbf{b})^T\mathbf{f} = 0$. However, the entries of the vector $\mathbf{a}-\mathbf{b}$ take values from the set $Q_L(\mathcal{A})$, therefore $\mathbf{f}$ cannot be linearly independent. Thus, every linearly independent channel is regular. On the other hand, there exist channels that are regular but not linearly independent. For example, the channel $\mathbf{f} = [2\,3\,10]^T$ is regular with respect to $\mathcal{A} = \{\pm 1\}$, but not linearly independent, since $[2\,2\,-1]\mathbf{f} = 0$. ■

Theorem 1.2: If all the rows occur on the interior of A_k with $L > 2$ and a BPSK alphabet, then A_k is PE of order $2L-1$.

Proof: A_k is PE of order $2L-1$ iff (14), where X_k is replaced by A_k, has only the trivial solution $X'_k = cA_k$ [19, Th. 3.1]. Here we show that if A_k has all the rows in its interior, then (14) has only the trivial solution $X'_k = cA_k$. Let A_k and X'_k be partitioned as follows:

$$A_k = \begin{bmatrix} \mathbf{t}_1 & \mathbf{t}_2 & \ldots & \mathbf{t}_L \\ \mathbf{a}_1 & \mathbf{a}_2 & \ldots & \mathbf{a}_L \\ \mathbf{b}_1 & \mathbf{b}_2 & \ldots & \mathbf{b}_L \end{bmatrix} \quad (17)$$

$$X'_k = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \ldots & \mathbf{u}_L \\ \mathbf{x}_1 & \mathbf{x}_2 & \ldots & \mathbf{x}_L \\ \mathbf{l}_1 & \mathbf{l}_2 & \ldots & \mathbf{l}_3 \end{bmatrix} \quad (18)$$

where $\mathbf{b}_j$ and $\mathbf{t}_j$ are $(L-1)\times 1$ vectors. The $(k+2-2L)\times 1$ vectors comprise the interior of A_k and X'_k, respectively [17]. Denoting $k+2-2L$ by m, we define the vectors $\overline{\mathbf{a}}_j$ ($\overline{\mathbf{x}}_j$) as the portion of the jth column of A_k (X'_k), starting from the second uppermost element of $\mathbf{a}_j$ down to the uppermost element of $\mathbf{b}_j$ ($\mathbf{l}_j$), e.g.,

$$\overline{\mathbf{a}}_j = [\mathbf{a}_j(2) \quad \cdots \quad \mathbf{a}_j(m) \quad \mathbf{b}_j(1)]^T. \quad (19)$$

The Toeplitz structure implies $\overline{\mathbf{a}}_j = \mathbf{a}_{j-1}$ for $j = 2, \ldots, L$ with the similar relation holding for X'_k. Since it has been assumed that A_k and X'_k have the same column space

$$\overline{\mathbf{x}}_1 = c_{11}\overline{\mathbf{a}}_1 + c_{12}\overline{\mathbf{a}}_2 + \cdots + c_{1L}\overline{\mathbf{a}}_L \quad (20)$$
$$= c_{11}\overline{\mathbf{a}}_1 + c_{12}\mathbf{a}_1 + \cdots + c_{1L}\mathbf{a}_{L-1} \quad (21)$$
$$\overline{\mathbf{x}}_2 = c_{21}\overline{\mathbf{a}}_1 + c_{22}\overline{\mathbf{a}}_2 + \cdots + c_{2L}\overline{\mathbf{a}}_L \quad (22)$$
$$= c_{21}\overline{\mathbf{a}}_1 + c_{22}\mathbf{a}_1 + \cdots + c_{2L}\mathbf{a}_{L-1} \quad (23)$$
$$\vdots$$
$$\overline{\mathbf{x}}_L = c_{L1}\overline{\mathbf{a}}_1 + c_{L2}\overline{\mathbf{a}}_2 + \cdots + c_{LL}\overline{\mathbf{a}}_L \quad (24)$$
$$= c_{L1}\overline{\mathbf{a}}_1 + c_{L2}\mathbf{a}_1 + \cdots + c_{LL}\mathbf{a}_{L-1} \quad (25)$$

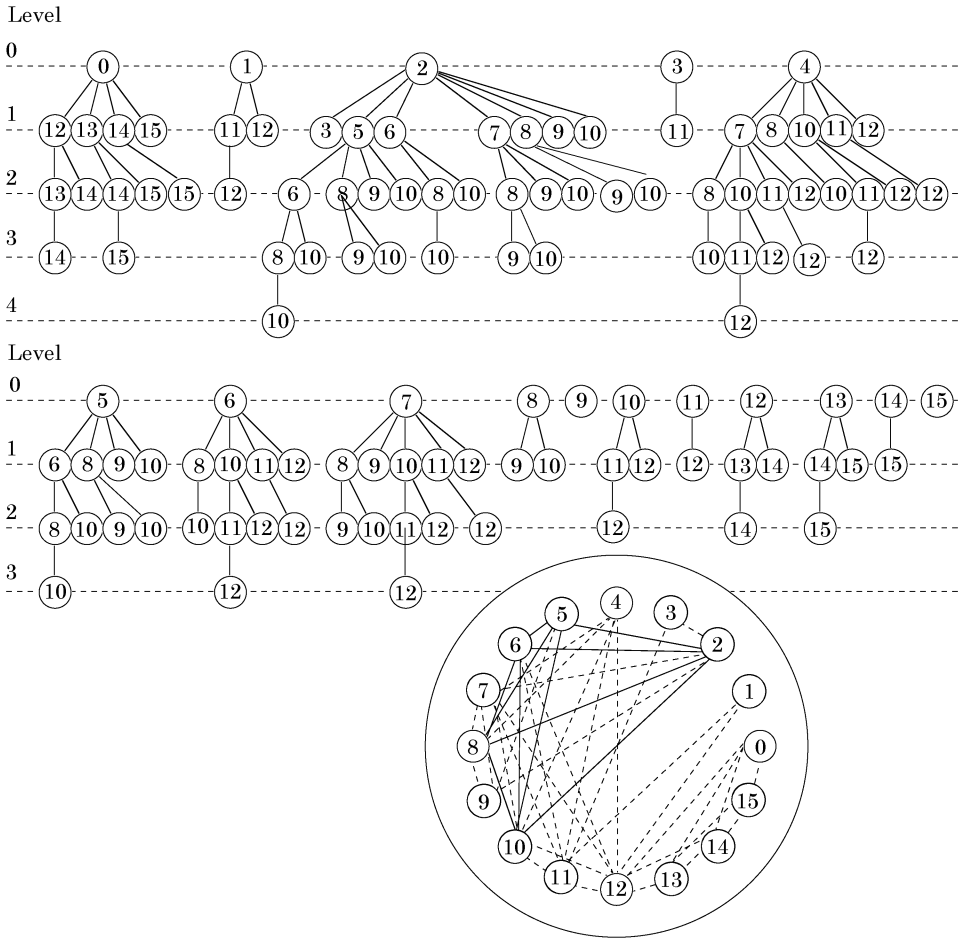

Fig. 8. Example of the execution of the clique algorithm. One of the maximum cliques of the graph (in circle) is emphasized by nondashed edges.

must hold for some choices of coefficients $\{c_{ij}\}$. Two different equations can be written for $\mathbf{x}_1$

$$\mathbf{x}_1 = c_{11}\mathbf{a}_1 + c_{12}\mathbf{a}_2 + \cdots + c_{1L}\mathbf{a}_L \tag{26}$$

$$\mathbf{x}_1 = \overline{\mathbf{x}}_2 = c_{21}\overline{\mathbf{a}}_1 + c_{22}\mathbf{a}_2 + \cdots + c_{2L}\mathbf{a}_L \tag{27}$$

where (26) follows directly from the assumption that $\mathcal{R}(A_k) = \mathcal{R}(X'_k)$, and (27) follows from (23) and the Toeplitz structure.

We now show that (27) and (23) imply that $c_{21} = 0$ by contradiction. If $c_{21} \neq 0$, (26) and (27) imply

$$\overline{\mathbf{a}}_1 = \underbrace{\frac{c_{11} - c_{22}}{c_{21}}}_{d_1}\mathbf{a}_1 + \underbrace{\frac{c_{12} - c_{23}}{c_{21}}}_{d_2}\mathbf{a}_2 \cdots + \underbrace{\frac{c_{1(L-1)} - c_{2L}}{c_{21}}}_{d_{L-1}}\mathbf{a}_{L-1} + \underbrace{\frac{c_{1L}}{c_{21}}}_{d_L}\mathbf{a}_L. \tag{28}$$

Assume that J of the d_i coefficients are nonzero, and denote these nonzero coefficients by $(\beta_1, \beta_2, \ldots, \beta_J)$ with $(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_J)$ denoting the corresponding values of the $\mathbf{a}_i$'s. The all-rows condition implies that all 2^J vectors consisting of -1 and 1 occur in $(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_J)$. Specifically, $\beta_1 + \beta_2 + \cdots + \beta_J$ and $\beta_1 + \beta_2 + \cdots - \beta_J$ both occur as entries of $\overline{\mathbf{a}}_1$. However, the entries of $\overline{\mathbf{a}}_1$ are either 1 or -1, so that a simple argument leads to the conclusion that $\beta_J = \pm 1$. Applying the same argument to β_i for $i = 1, \ldots, J-1$, we obtain $\beta_i = \pm 1$. Since $(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_J)$ has all permutations in its rows, it will have $[\beta_1, \beta_2, \ldots, \beta_J]$ as a row. Thus, $\beta_1^2 + \beta_2^2 + \cdots + \beta_J^2 = J$ is an entry of $\overline{\mathbf{a}}_1$, which is possible only if $J = 1$. So $\overline{\mathbf{a}}_1 = \beta\boldsymbol{\alpha}$, where $\beta = 1$ or $\beta = -1$ and $\boldsymbol{\alpha} = \mathbf{a}_i$ for some i. This condition, along with the Toeplitz structure of A_k, implies that $\mathbf{a}_1$ (or $\overline{\mathbf{a}}_1$) will have a periodicity of length at most $2(L-1)$. However, all rows cannot be constructed in A_k with such a column with such a period for $L \geq 3$. Thus, we conclude that the case $c_{21} = 0$ must hold, which, along with (28) and the linearly independence of $\mathbf{a}_i$, implies

$$c_{22} = c_{11},\ c_{23} = c_{12},\ \ldots,\ c_{2L} = c_{1(L-1)},\ c_{1L} = 0. \tag{29}$$

Similarly, two different equations can be written for $\mathbf{x}_2$ as in (26) and (27). With a similar reasoning

$$c_{31} = c_{32} = 0,\ c_{33} = c_{11},\ c_{34} = c_{12},\ \ldots\ c_{3L} = c_{1(L-2)},\ c_{1(L-1)} = 0. \tag{30}$$

Applying the same argument $(L-3)$ times more, we obtain $X'_k = cA_k$. ■

Appendix II
An Algorithm for Code Design (Clique Problem)

The general class of clique problems is known to be NP-hard [25]. However, we present an optimal algorithm which enumerates all cliques in an efficient manner when the maximum number of edges for any vertex is much smaller than $|V|$. We describe this algorithm via the example in Fig. 8. The graph shown in Fig. 8 has $|V| = 16$ with a maximum number of edges for a

given node of seven. A maximum clique is shown with a solid line and all other edges are shown dashed. The algorithm proceeds as follows.

1) Label the vertices of the graph from 0 to $|V|-1$.
2) Let each vertex define the root of a tree. Construct the tree associated with vertex labeled v as follows. The children of the root v are all vertices with labels $\tilde{v}$, such that there is an edge between v and $\tilde{v}$, and $v < \tilde{v}$. For example, in Fig. 8, the tree with root label 3 has no child corresponding to label 2 because $3 > 2$. The edge between 2 and 3 has been represented in the tree rooted at label 2.
3) To grow each tree from the root level zero to level one, consider the child of v with the smallest label, say $\tilde{v}$. The children of this node are determined by the intersection of all other children of root v with a larger label than $\tilde{v}$ and the children of the tree rooted at $\tilde{v}$. For example, in Fig. 8, the children of 5 under the tree rooted at 2 are determined by taking the intersection of {6,7,8,9,10} and the children of the root labeled 5, namely {6,8,9,10}; yielding {6,8,9,10}. The children of 6 under the root labeled 2 are determined by taking the intersection of {7,8,9,10} and the children of the root 6, namely, {8,10,11,12}; yielding {8,10}.
4) At level $d > 0$, the tree is extended by the same process, except the set used for intersection can be simplified. Specifically, for any of the $|V|$-rooted trees, consider extending the vertex v at level d, which has parent $P(v)$ at level $d-1$. The intersection of two sets provides the children of v. The first set is all vertices $\tilde{v}$ at level d in the same rooted tree with the same parent node $P(v)$, such that $v < \tilde{v}$. The second set is the children of the vertex label v at level $d-1$ in the same rooted tree. For example, the children of vertex 6 at level 2 in the tree rooted by vertex 2 (i.e., $2 \rightarrow 5 \rightarrow 6$) are {8,10}. This is obtained by intersecting children of 5 (i.e., $2 \rightarrow 5$) to the right of 6, namely, {8,9,10} (set one), with the children of node 6 at level 1 (i.e., $2 \rightarrow 6$) in the same rooted tree, namely, {8,10}. Note that one could also use the children of 6 at level one of the tree rooted by vertex 6 for the second set (i.e., {8,10,11,12}, but this is less efficient.
5) This process is repeated on each of the rooted trees until it terminates. The largest clique is determined by the deepest level obtained in one of the trees. Notice that all cliques are enumerated in the collection of trees. In the example, two cliques of size 5 exist: {2,5,6,8,10} from the root 2 and {4,7,10,11,12} from the root 4.

The correctness of this algorithm can be formally proven. Also, the running time of this algorithm can be shown to be upper bounded as exponential in N_E, the maximum number of edges associated with any vertex and polynomial in $|V|$. By comparison, an exhaustive search to check if a clique of size n exists has running time $N_E!/(N_E-n)!n!$. However, even though the complexity relative to an exhaustive search is small, when N_E and/or $|V|$ becomes large, this algorithm becomes prohibitively complex.

A. Greedy Heuristic for Code Design

We propose the following suboptimal algorithm when the above algorithm becomes prohibitively complex. Construct the set of $|V|$ trees as described above. At each level, grow only the tree with the largest number of children. For example, in Fig. 8, the tree rooted by 2 would be selected to be extended to level 1 because it has seven children, while no other root has more than five children. Only the vertex labeled 5 under the tree rooted at 2 would be extended to level 2 because it has four children, while no other node at level 2 under the tree rooted at 2 has more than three children. When considering the next level, there is a tie condition, because both vertices 6 and 8 at level 2 have two children. In our heuristic, an arbitrary tie breaker is used. If this breaks the tie in favor of 6, then the optimal solution will be found by this heuristic for this example. However, if the tie-breaker selects 8 over 6 at level 2, then two cliques of size four are found (suboptimal), i.e., {2,5,8,9} and {2,5,8,10}.

APPENDIX III
PDF OF THE QUADRATIC DETECTOR SIGNAL IN AWGN

For compactness, we use $A^{(i)}$ in place of $A_k^{(i)}$, r in place of $r_k(i,j)$, and most quantities discussed do not explicitly denote dependence on the two data sequences ($A^{(i)}$ and $A^{(j)}$) and the actual channel ($\mathbf{f}$). Before giving the pdf of the decision statistic r in (7), we give three lemmas about the spectral properties of $P^{(i)} - P^{(j)}$, and one theorem stating that r can be represented as the sum of $2L$ independent *chi-square distributed* random variables.

Lemma 3.1: There exists a set of orthonormal eigenvectors for $P^{(i)} - P^{(j)}$ with real eigenvalues.

Proof: This follows from the symmetric structure of $P^{(i)} - P^{(j)} = [P^{(i)} - P^{(j)}]^T$. ■

The next lemma provides a coordinate system useful for describing eigenvectors of *Lemma 3.1*. First, however, we give the definition of principal angles between two subspaces $\mathcal{F}$ and $\mathcal{G}$ in $\mathcal{R}^m$ [26]. Let $\mathcal{F}$ and $\mathcal{G}$ be subspaces in $\mathcal{R}^m$ whose dimensions satisfy

$$p = \dim(\mathcal{F}) \geq \dim(\mathcal{G}) = q \geq 1. \tag{31}$$

The principal angles $\theta_1, \ldots, \theta_q \in [0, \pi/2]$ between $\mathcal{F}$ and $\mathcal{G}$ with principal vectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_q\}$ and $\{\mathbf{v}_1, \ldots, \mathbf{v}_q\}$ are defined recursively by

$$\begin{aligned} \cos(\theta_j) &= \mathbf{u}_j^T \mathbf{v}_j \\ &= \max_{\mathbf{u}\in\mathcal{F}} \max_{\mathbf{v}\in\mathcal{G}} \mathbf{u}^T \mathbf{v} \\ &\quad \text{subject to: } \|\mathbf{u}\| = \|\mathbf{v}\| = 1 \\ &\quad\quad \mathbf{u}^T \mathbf{u}_i = \mathbf{v}^T \mathbf{v}_i = 0, \quad i = 1, \ldots, j-1. \end{aligned} \tag{32}$$

Intuitively, $\mathbf{u}_1$ and $\mathbf{v}_1$ define the two directions in $\mathcal{F}$ and $\mathcal{G}$, respectively, that are closest to pointing in the same direction, and θ_1 characterizes the angle between these directions. The vectors $\mathbf{u}_j$ and $\mathbf{v}_j$ are similarly interpreted, based on the subspaces of $\mathcal{F}$ and $\mathcal{G}$ that are orthogonal to the span of $\{\mathbf{u}_i\}_{i=1}^{j-1}$ and $\{\mathbf{v}_i\}_{i=1}^{j-1}$, respectively.

The principal angles and vectors can be computed using three SVDs. Let the matrix O_i (O_j) have columns defined by orthonormal basis for the column space of $A^{(i)}$ ($A^{(j)}$), which can be obtained from an SVD of $A^{(i)}$ ($A^{(j)}$). The SVD of $O_i^T O_j$, $Y^T(O_i^T O_j)Z = \mathrm{diag}(\sigma_1, \ldots, \sigma_q)$, yields the principal vectors

$$[\mathbf{u}_1, \cdots, \mathbf{u}_q] = O_i Y \tag{33}$$

$$[\mathbf{v}_1, \cdots, \mathbf{v}_q] = O_j Z \tag{34}$$

with the principal angles as the arc-cosine of singular values [26].

Lemma 3.2: Let $\{\mathbf{u}_1,\dots,\mathbf{u}_L\}$ and $\{\mathbf{v}_1,\dots\mathbf{v}_L\}$ be the principal vectors with angles $0<\theta_1\le\theta_2<\cdots\le\theta_L$ between the column spaces of two separable data matrices, $A^{(i)}$ and $A^{(j)}$. A set of orthonormal eigenvectors of $P^{(i)}-P^{(j)}$ associated with nonzero eigenvalues can be written as

$$\mathbf{e}_l^{(+)}=\frac{1}{\sqrt{2}}\left(\frac{\mathbf{u}_l+\mathbf{v}_l}{\|\mathbf{u}_l+\mathbf{v}_l\|}-\frac{\mathbf{u}_l-\mathbf{v}_l}{\|\mathbf{u}_l-\mathbf{v}_l\|}\right) \tag{35}$$

$$\mathbf{e}_l^{(-)}=\frac{1}{\sqrt{2}}\left(\frac{\mathbf{u}_l+\mathbf{v}_l}{\|\mathbf{u}_l+\mathbf{v}_l\|}+\frac{\mathbf{u}_l-\mathbf{v}_l}{\|\mathbf{u}_l-\mathbf{v}_l\|}\right) \tag{36}$$

where $l=\{1,\dots,L\}$.

Proof: The denominators of (35) and (36) can be written in terms of principal angles[14]

$$\|\mathbf{u}_l\pm\mathbf{v}_l\|=\sqrt{2\pm2\cos\theta_l}. \tag{37}$$

The product of the given vectors $\mathbf{e}_l^{(+)}$ and $\mathbf{e}_l^{(-)}$ by $P^{(i)}-P^{(j)}$ can be shown to be

$$\left(P^{(i)}-P^{(j)}\right)\mathbf{e}_l^{(\pm)}=\frac{1}{\sqrt{2}}\left[(\mathbf{u}_l-\mathbf{v}_l)\frac{\sqrt{1+\cos\theta_l}}{\sqrt{2}}\pm(\mathbf{u}_l+\mathbf{v}_l)\frac{\sqrt{1-\cos\theta_l}}{\sqrt{2}}\right] \tag{38}$$

$$=\pm\sin\theta_l\mathbf{e}_l^{(\pm)}. \tag{39}$$

Thus, $\mathbf{e}_l^{(+)}$ and $\mathbf{e}_l^{(-)}$ are the eigenvectors of $P^{(i)}-P^{(j)}$. ∎

Corollary 3.1: If λ_l is an eigenvalue of $P^{(i)}-P^{(j)}$ corresponding to the eigenvector $\mathbf{e}_l^{(+)}$, then $-\lambda_l$ is also an eigenvalue of $P^{(i)}-P^{(j)}$ corresponding to the eigenvector $\mathbf{e}_l^{(-)}$.

Lemma 3.3: Let $\{\mathbf{u}_1,\mathbf{u}_2,\dots,\mathbf{u}_L\}$ and $\{\mathbf{v}_1,\mathbf{v}_2,\dots,\mathbf{v}_L\}$ be the principal vectors with principal angles $0<\theta_1\le\theta_2\le\dots\le\theta_L$ between the column spaces of two separable data matrices $A^{(i)}$ and $A^{(j)}$, and $\{\mathbf{e}_l^{(+)},\mathbf{e}_l^{(-)}\}_{l=1}^L$ be the orthonormal eigenvectors of $P^{(i)}-P^{(j)}$. Then the principal vectors $\{\mathbf{u}_l\}_{l=1}^L$ and $\{\mathbf{v}_l\}_{l=1}^L$ can be expressed in terms of eigenvectors of $P^{(i)}-P^{(j)}$

$$\mathbf{u}_l=\frac{1}{2}(\sqrt{1+\cos\theta_l}+\sqrt{1-\cos\theta_l})\mathbf{e}_l^{(+)}+\frac{1}{2}(\sqrt{1+\cos\theta_l}-\sqrt{1-\cos\theta_l})\mathbf{e}_l^{(-)} \tag{40}$$

$$\mathbf{v}_l=\frac{1}{2}(\sqrt{1+\cos\theta_l}-\sqrt{1-\cos\theta_l})\mathbf{e}_l^{(+)}+\frac{1}{2}(\sqrt{1+\cos\theta_l}+\sqrt{1-\cos\theta_l})\mathbf{e}_l^{(-)}. \tag{41}$$

Proof: Solving (35) and (36) for $\mathbf{u}_l$ and $\mathbf{v}_l$ yields (40) and (41). ∎

Theorem 3.1: If the transmitted sequence is $A_k^{(i)}$, the channel is $\mathbf{f}$, and the AWGN variance is σ_w^2, then r can be expressed in terms of $2L$ *independent noncentral chi-square distributed* random variables $\{r_l^{(+)},r_l^{(-)}\}_{l=1}^L$

$$r=\sum_{l=1}^{L}\left(r_l^{(+)}-r_l^{(-)}\right) \tag{42}$$

[14]For conciseness, we use $\pm$ and $\mp$ to express two equations in one using the standard convention, e.g., $x^{(\pm)}=y\mp z$ is short for $x^{(+)}=y-m$ and $x^{(-)}=y+m$.

with mean and variances

$$m_{r_l^{(\pm)}}=\frac{1}{2}c_l^2(1\pm\sin\theta_l)\sin\theta_l+\sigma_w^2\sin\theta_l \tag{43}$$

$$\sigma^2_{r_l^{(\pm)}}=2\sigma_w^4\sin^2\theta_l+(\sigma_w\sin\theta_l c_l)^2(1\pm\sin\theta_l) \tag{44}$$

where $\{c_l\}_{l=1}^L$ are the coordinates of the vector $A_k^{(i)}\mathbf{f}$ with respect to the basis $\{\mathbf{u}_l\}_{l=1}^L$.

Proof: Using the spectral decomposition of $P^{(i)}-P^{(j)}=\sum_{l=1}^L\sin\theta_l(\mathbf{e}_l^{(+)}[\mathbf{e}_l^{(+)}]^T-\mathbf{e}_l^{(-)}[\mathbf{e}_l^{(-)}]^T)$, (7) can be written as

$$r=\mathbf{z}_k^T\left[\sum_{l=1}^{L}\left(\sin\theta_l\mathbf{e}_l^{(+)}\left[\mathbf{e}_l^{(+)}\right]^T-\sin\theta_l\mathbf{e}_l^{(-)}\left[\mathbf{e}_l^{(-)}\right]^T\right)\right]^T\mathbf{z}_k$$

$$=\sum_{l=1}^{L}\left(\sin\theta_l\left\|\mathbf{z}_k^T\mathbf{e}_l^{(+)}\right\|^2-\sin\theta_l\left\|\mathbf{z}_k^T\mathbf{e}_l^{(-)}\right\|^2\right) \tag{45}$$

$$=\sum_{l=1}^{L}\left(\left[y_l^{(+)}\right]^2-\left[y_l^{(-)}\right]^2\right). \tag{46}$$

Since $\mathbf{z}_k$ is Gaussian and $\{\mathbf{e}_l^{(+)},\mathbf{e}_l^{(-)}\}_{l=1}^L$ is an orthonormal set, $\{y_l^{(\pm)}\}$ are independent Gaussian random variables with means $(1/2)\sqrt{\sin\theta_l(1\pm\sin\theta_l)c_l^2}$ and variance $\sigma_w^2\sin\theta_l$. Therefore, $\{r_l^{(\pm)}=[y_l^{(\pm)}]^2\}$ are *chi-square distributed* with means and variances given by (43) and (44), respectively. ∎

Corollary 3.2: The conditional pdf of r can be expressed as the following $2L$-fold convolution:

$$f_r\left(r|A^{(i)};\mathbf{f}\right)=f_{r_1^{(+)}}\left(r|A^{(i)};\mathbf{f}\right)*f_{r_1^{(-)}}\left(-r|A^{(i)};\mathbf{f}\right)\cdots *f_{r_L^{(+)}}\left(r|A^{(i)};\mathbf{f}\right)*f_{r_L^{(-)}}\left(-r|A^{(i)};\mathbf{f}\right) \tag{47}$$

with the $2L$ densities defined by

$$f_{r_l^{(\pm)}}\left(r|A^{(i)};\mathbf{f}\right)=\frac{\exp\left[\frac{-r+c_l^2(1\pm\sin\theta_l)\sin\theta_l}{2\sigma_w^2\sin\theta_l}\right]}{\sqrt{2\pi r\sigma_w^2\sin\theta_l}}\times\cosh\left(\frac{\sqrt{rc_l^2(1\pm\sin\theta_l)\sin\theta_l}}{\sigma_w^2\sin\theta_l}\right)r\ge0. \tag{48}$$

Proof: Follows directly from the fact that $\{r_l^{(+)},r_l^{(-)}\}$ is a set of independent chi-square random variables, and the density of $-r_l^{(-)}$ is $f_{r_l^{(-)}}(-r_l|A^{(i)};\mathbf{f})$. ∎

Corollary 3.3: The conditional mean and the variance of r are as stated in (8) and (10), respectively.

Proof: Follows directly from the fact that r is the sum of independent random variables with densities given in (48). ∎

Corollary 3.4: $m_r(\cdot)/\sigma_r(\cdot)$ is an monotonically increasing function of $m_r(\cdot)$.

Proof: Note that with the principal angles known, $\sigma_r(\cdot)$ is a function of $m_r(\cdot)$, as implied by (10). Using this expression, it is straightforward to verify that the derivative of $m_r(\cdot)/\sigma_r(\cdot)$ with respect to $m_r(\cdot)$ is always positive. ∎

Corollary 3.5: For a unit norm channel, the mean in (8) is bounded as

$$\sigma^2_{\min,A_k}(i)\sin^2\theta_1\le\sum_{l=1}^{L}c_l^2\sin\theta_l^2\le\sigma^2_{\max,A_k}(i)\sin^2\theta_L \tag{49}$$

where $\sigma^2_{\min,A_k}(i)$ and $\sigma^2_{\max,A_k}(i)$ are the minimum and the maximum singular values of $A_k^{(i)}$, respectively.

Proof: It follows from (32) that $\sin\theta_1 \le \sin\theta_l \le \sin\theta_L$. Also, for a unit norm $\mathbf{f}$, $\sigma^2_{\min,A_k}(i) \le \|A^{(i)}\mathbf{f}\|^2 \le \sigma^2_{\max,A_k}(i)$ and $\|A^{(i)}\mathbf{f}\|^2 = \sum_l c_l^2$. ■

Finally, we note that in [28], the characteristic function of the r is given as

$$\begin{aligned}\Psi_r(j\upsilon) &= \mathbb{E}\left\{\exp(j\upsilon r)\right\} \\ &= \left|I - 2j\upsilon\sigma^2\left(P^{(i)} - P^{(j)}\right)\right|^{\frac{1}{2}} \\ &\quad \times e^{-\frac{\mathbf{s}_i^T\left[I-\left(I-2j\upsilon\left(P^{(i)}-P^{(j)}\right)\right)^{-1}\right]\mathbf{s}_i}{2\sigma_w^2}}\end{aligned} \tag{50}$$

where $\mathbf{s}_i = A_k^{(i)}\mathbf{f}$.

REFERENCES

[1] O. Coskun and K. M. Chugg, "Codes for unknown ISI channels," in *Proc. Allerton Conf. Commun., Control, Comput.*, Oct. 2000, pp. 689–699.

[2] G. D. Forney, Jr., "Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference," *IEEE Trans. Inf. Theory*, vol. IT-18, no. 5, pp. 284–287, May 1972.

[3] S. N. Crozier, D. D. Falconer, and S. A. Mahmoud, "Least sum of squared errors (LSSE) channel estimation," *IEE Proc. Radar Signal Process.*, vol. 138, pp. 371–378, Aug. 1991.

[4] S. U. H. Qureshi, "Fast start-up equalization with periodic training sequences," *IEEE Trans. Inf. Theory*, vol. IT-23, no. 9, pp. 553–562, Sep. 1977.

[5] R. Raheli, A. Polydoros, and C. Tzou, "Per-survivor processing: A general approach to MLSE in uncertain environments," *IEEE Trans. Commun.*, vol. 43, no. 2-4, pp. 354–364, Feb.–Apr. 1995.

[6] J. Lin, F. Ling, and J. G. Proakis, "Joint data and channel estimation for TDMA mobile channels," in *Proc. PIMRC*, 1992, pp. 235–239.

[7] R. Raheli, A. Polydoros, and C. Tzou, "The principle of per-survivor processing: A general approach to approximate and adaptive MLSE," in *Proc. Globecom Conf.*, 1991, pp. 33.3.1–33.1.6.

[8] Y. Sato, "A method of self-recovering equalization for multilevel amplitude-modulation systems," *IEEE Trans. Commun.*, vol. COM-23, no. 6, pp. 679–682, Jun. 1975.

[9] D. Godard, "Self-recovering equalization and carrier tracking in two-dimensional data communication systems," *IEEE Trans. Commun.*, vol. COM-28, no. 11, pp. 1867–1875, Nov. 1980.

[10] A. Benveniste and G. Ruget, "Robust identification of a nonminimum phase system: Blind adjustment of a linear equalizer in data communication," *IEEE Trans. Autom. Control*, vol. AC-25, no. 6, pp. 385–398, Jun. 1980.

[11] H. H. Chiang and C. L. Nikias, "Adaptive deconvolution and identification of nonminimum phase FIR systems based on cumulants," *IEEE Trans. Autom. Control*, vol. 35, no. 1, pp. 36–47, Jan. 1990.

[12] G. B. Giannakis and J. M. Mendel, "Identification of nonminimum phase systems using higher order moments," *IEEE Trans. Acoust., Speech, Signal Process.*, no. 3, pp. 360–377, Mar. 1989.

[13] W. A. Gardner, "A new method of channel identification," *IEEE Trans. Commun.*, vol. 39, no. 6, pp. 813–817, Jun. 1991.

[14] L. Tong and T. K. G. Xu, "Blind idenfication and equalization based on second-order statistics: A time domain approach," *IEEE Trans. Inf. Theory*, vol. 40, no. 3, pp. 340–349, Mar. 1994.

[15] M. Ghosh and C. L. Weber, "Maximum-likelihood blind equalization," *Proc. SPIE*, pp. 181–195, Jul. 1991.

[16] N. Seshadri, "Joint data and channel estimation using blind trellis search techniques," *IEEE Trans. Commun.*, vol. 42, no. 2-4, pp. 1000–1011, Feb.-Apr. 1994.

[17] K. M. Chugg, "Blind acquisition characteristics of PSP-based sequence detectors," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 10, pp. 1518–1529, Oct. 1998.

[18] K. M. Chugg and A. Polydoros, "MLSE for an unknown channel—Part I: Optimality considerations," *IEEE Trans. Commun.*, vol. 44, no. 7, pp. 836–846, Jul. 1996.

[19] F. Gustafsson and B. Wahlberg, "Blind equalization by direct examination of the input sequences," *IEEE Trans. Commun.*, vol. 43, no. 7, pp. 2213–2222, Jul. 1995.

[20] M. Skoglund, J. Giese, and S. Parkvall, "Code design for combined channel estimation and error protection," *IEEE Trans. Inf. Theory*, vol. 48, no. 5, pp. 1162–1171, May 2002.

[21] K. M. Chugg and A. Polydoros, "MLSE for an unknown channel—Part II: Tracking performance," *IEEE Trans. Commun.*, no. 8, pp. 949–958, Aug. 1996.

[22] K. M. Chugg, "The condition for the applicability of the Viterbi algorithm with implications for fading channel MLSD," *IEEE Trans. Commun.*, vol. 46, no. 9, pp. 1112–1116, Sep. 1998.

[23] T. Kailath, "Correlation detection of signals perturbed by a random channel," *IRE Trans. Inf. Theory*, vol. IT-6, no. 6, pp. 361–366, Jun. 1960.

[24] K. M. Chugg, "Maximum-likelihood sequence estimation for unknown channels," in *IEEE Communication Theory Workshop*, Santa Cruz, CA, Apr. 1995, [CD-ROM].

[25] T. H. Cormen, C. E. Leiserson, and R. L. Rivest, *Introduction to Algorithms*. Cambridge, MA: MIT Press, 1990.

[26] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2nd ed. Baltimore, MD: Johns Hopkins Press, 1989.

[27] A. Anastasopoulos and K. M. Chugg, "Adaptive soft-input soft-output algorithms for iterative detection with parametric uncertainty," *IEEE Trans. Commun.*, vol. 48, no. 10, pp. 1638–1649, Oct. 2000.

[28] J. Omura and T. Kailath, "Some useful probability distributions," Syst. Theory Lab., Stanford Electron. Labs., Stanford Univ., Stanford, CA, Tech. Rep. 7050-6, Sep. 1965.

Orhan Coskun received the B.S. degree from Middle East Technical University, Ankara, Turkey, in 1993, and the M.S and Ph.D. degrees from the University of Southern California (USC), Los Angeles, CA, in 1998 and 2004, respectively, all in electrical engineering.

During the first half of 2001, he was with Fantasma Networks, Mountain View, CA, working on detection and link budget problems for UWB systems. From the second half of 2001 to 2004, he worked on equalization, synchronization, and error-correction techniques for fiber optical and 10 GBASE-T systems at Santel Networks. He is currently with Izmir Institute of Technology, Izmir, Turkey, where he is an Assistant Professor. He also leads the Wireless Home Portable Media Group at Vestel Corporation, Izmir, Turkey.

Keith M. Chugg (S'88-M'95) received the B.S. degree (high distinction) in engineering from Harvey Mudd College, Claremont, CA, in 1989, and the M.S. and Ph.D. degrees in electrical engineering from the University of Southern California (USC), Los Angeles, CA, in 1990 and 1995, respectively.

During the 1995–1996 academic year, he was an Assistant Professor with the Electrical and Computer Engineering Department, University of Arizona, Tucson, AZ. In 1996, he joined the Electrical Engineering Department at USC, where he is currently an Associate Professor. His research interests are in the general areas of signaling, detection, and estimation for digital communication and data-storage systems. He is also interested in architectures for efficient implementation of the resulting algorithms. Along with his former Ph.D. students, A. Anastasopoulos and X. Chen, he is co-author of the book *Iterative Detection: Adaptivity, Complexity Reduction, and Applications* (Norwell, MA: Kluwer Academic Press, 2001). He is a co-founder of TrellisWare Technologies, Inc., San Diego, CA, where he is Chief Scientist.

Dr. Chugg has served as an Associate Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS, and was Program Co-Chair for the Communication Theory Symposium at Globecom 2002. He received the Fred W. Ellersick Award for the best unclassified paper at MILCOM 2003.