Joint Noncoherent Demodulation and Decoding for the Block Fading Channel: A Practical Framework for Approaching Shannon Capacity

Rong-Rong Chen, Ralf Koetter, *Member, IEEE*, Upamanyu Madhow, *Senior Member, IEEE*, and Dakshi Agrawal

***Abstract*—This paper contains a systematic investigation of practical coding strategies for *noncoherent* communication over fading channels, guided by explicit comparisons with information-theoretic benchmarks. Noncoherent reception is interpreted as joint data and channel estimation, assuming that the channel is time varying and *a priori* unknown. We consider iterative decoding for a serial concatenation of a standard binary outer channel code with an inner modulation code amenable to noncoherent detection. For an information rate of about 1/2 bit per channel use, the proposed scheme, using a quaternary phase-shift keying (QPSK) alphabet, provides performance within 1.6–1.7 dB of Shannon capacity for the block fading channel, and is about 2.5–3 dB superior to standard differential demodulation in conjunction with an outer channel code. We also provide capacity computations for noncoherent communication using standard phase-shift keying (PSK) and quadrature amplitude modulation (QAM) alphabets, comparing these with the capacity with unconstrained input provides guidance as to the choice of constellation as a function of the signal-to-noise ratio. These results imply that QPSK suffices to approach the unconstrained capacity for the relatively low information and fading rates considered in our performance evaluations, but that QAM is superior to PSK for higher information or fading rates, motivating further research into efficient noncoherent coded modulation with QAM alphabets.**

***Index Terms*—Capacity, coding, fading channels, noncoherent detection, wireless communications.**

I. Introduction

WHILE THE problem of reliable digital communication over time-varying wireless channels has a rich history, it is increasingly important to find power- and bandwidth-efficient solutions to this problem, in view of the rapid growth of commercial cellular and personal communication systems. There are two main approaches currently employed for this purpose.

1) Use pilot symbols or codes to estimate and track the time-varying channel, and then do *coherent* demodulation and decoding using the estimated channel. This is the approach used in most current cellular communication systems.
2) Use *noncoherent* demodulation, which does not require explicit knowledge or estimation of the channel phase. A special case of this is differential modulation, in which the channel is approximated as roughly constant over at least two consecutive symbols, and information is encoded in symbol transitions. Another special case is orthogonal modulation, whose bandwidth efficiency, however, is too low for consideration in most commercial systems.

Since a time-varying channel is *a priori* unknown, the noncoherent paradigm, interpreted in its broadest sense as joint data and channel estimation, is most appropriate in this setting. Indeed, pilot-based systems can be viewed as suboptimal implementations within this paradigm, rather than as coherent communication systems (in which the channel estimate should be essentially ideal). A standard pilot-based system is inherently suboptimal because it uses only the energy of the pilot symbols for channel estimation, rather than also exploiting the (typically larger) energies in the unknown data symbols. Moreover, it requires excessive overhead for rapidly time-varying channels or for multiple-antenna communication (the latter is not considered here).

However, existing noncoherent methods such as standard differential demodulation are far from optimal, and are often inferior compared to pilot-based schemes. Block demodulation of differentially modulated signals can improve performance for uncoded systems [1], [2], but such methods must be integrated with sophisticated coding techniques to obtain an efficient noncoherent system.

The goal of this paper is to obtain systematic approaches for designing practical noncoherent coding and modulation schemes for fading channels, with a view to approaching information-theoretic limits. There are two reasons why this effort is particularly timely. First, Shannon capacity of a reasonable (albeit idealized) block Rayleigh fading model has been recently derived by Marzetta and Hochwald [3], thus providing a benchmark for comparing with practical schemes. Second, efficient iterative decoding methods and turbo-like

Paper approved by I. Lee, the Editor for Wireless Communication Theory of the IEEE Communications Society. Manuscript received January 21, 2002; revised December 5, 2002 and April 22, 2003. This work was supported in part by the National Science Foundation under Grant NSF NCR96-24008 (CAREER), Grant CCR 99-79381, and Grant ITR 00-85929, and in part by the Army Research Office under Grant DAAG55-98-1-0219 and Grant DAAD19-00-1-0567. This paper was presented in part at the Conference on Information Sciences and Systems (CISS), Princeton, NJ, March 2000.

R.-R. Chen was with the Coordinated Science Laboratory, Department of Electrical and Computer Engineering, University of Illinois, Urbana, IL 61801 USA. She is now with the Department of Electrical and Computer Engineering, University of Utah, Salt Lake City, UT 84112 USA (e-mail: rchen@eng.utah.edu).

R. Koetter is with the Coordinated Science Laboratory, Department of Electrical and Computer Engineering, University of Illinois, Urbana, IL 61801 USA (e-mail: koetter@uiuc.edu).

U. Madhow is with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106 USA (e-mail: madhow@ece.ucsb.edu).

D. Agrawal is with the IBM T.J. Watson Research Center, Yorktown Heights, NY 10598 USA (e-mail: agrawal@us.ibm.com).

Digital Object Identifier 10.1109/TCOMM.2003.818087

codes, used typically on the additive white Gaussian noise (AWGN) channel [4], hold the hope for approaching capacity over more complicated channels.

As a practical approach to noncoherent reception, we propose a turbo-like serial concatenation of a standard outer channel code with an inner modulation code amenable to noncoherent detection, together with iterative joint decoding and demodulation. We wish to determine whether this class of schemes can approach the Shannon capacity for noncoherent communication. Hence, while the proposed methods apply to continuously time-varying channels, performance evaluation is restricted to the block fading channel model, since exact information-theoretic benchmarks are available only for this model.

The main results of this paper are summarized as follows.

1) We study concatenated coding schemes which consist of an outer channel code, a random interleaver, and an inner modulation code amenable to noncoherent demodulation. The encoding and decoding techniques are tailored to phase-shift keying (PSK) alphabets. It is shown that, for information rates of about 1/2 bit per channel use and bit-error rate (BER) of 10^{-4}, the proposed methods are within 1.6–1.7 dB of Shannon capacity (ergodic capacity), and are about 2.5–3 dB superior to the system composed of a standard differential phase-shift keying (DPSK) demodulation scheme in conjunction with a powerful outer channel code.
2) We compute the capacity of the noncoherent block fading channel for both PSK and quadrature amplitude modulation (QAM) alphabets, comparing it with the capacity with unconstrained input (except an average power constraint) [3] in order to determine the appropriate alphabet for a given channel coherence length and transmission rate. QPSK is a suitable choice for transmission rates of approximately 1/2 bit per channel use and coherence lengths larger than or equal to 10, which are the settings emphasized in the performance evaluation of our practical strategies. For higher transmission rates and/or smaller coherence lengths, our computations show that QAM signaling is superior to PSK signaling for approaching unconstrained capacity. This motivates future work on extending our code constructions and decoding techniques to QAM alphabets.
3) We consider two different types of inner modulation codes based on differential encoding. M-ary DPSK and block-based M-ary differential phase-shift keying (B-DPSK). In order to achieve better performance, inner modulation codes should be matched to the outer channel codes. Our simulations suggest that the following three combinations of outer channel codes and inner modulation codes lead to excellent performance: convolutional code with DPSK, repeat-accumulate (RA) code with B-DPSK, and turbo code with B-DPSK.
4) We propose a suboptimal noncoherent demodulator which has linear complexity with respect to the coherence length. This demodulator employs a simple, robust, averaging estimator for the channel fading amplitude, and the technique of phase quantization for the phase shift caused by the channel. By considering several different hypotheses regarding the channel phase shift, we reduce a noncoherent channel to a set of coherent channels, thus greatly reducing the complexity. Simulations show that our simple estimator for the fading amplitude induces only about 0.3–0.5 dB performance degradation compared to the case when the fading amplitude is perfectly known.

The literature most relevant to this paper is summarized as follows. The work most closely related to ours is [5], which employed a similar iterative noncoherent receiver for an AWGN channel with unknown phase shift. They also pointed out that a certain differential encoder (RDE code) similar to the B-DPSK code considered in this paper, is superior to DPSK, when concatenated with turbo code and decoded iteratively. The differences and similarities between our work and [5] will be highlighted appropriately in later sections of the paper.

A key paper in uncoded noncoherent communication is [1], which showed that block demodulation of DPSK alleviates performance loss from standard demodulation based on a two-symbol block. Peleg and Shamai [6] first proposed a new receiver for coded and interleaved DPSK based on multiple-symbol differential detection (MSDD) and iterative decoding for an AWGN channel with an unknown phase shift. Iterative structures for joint noncoherent demodulation and decoding have also been considered in [7]–[9] for AWGN channels, and in [10]–[15] for fading channels. In all of these papers, iterative processing is applied to a serially concatenated system consisting of an outer channel code, an interleaver, and an inner modulation code, with soft information exchange between the demodulator and the outer channel decoder proceeding in an iterative fashion as described in [16]. Since the optimal noncoherent demodulator has a complexity exponential in the memory, a key component of the design effort in many of these papers is in devising soft-input–soft-output (SISO) noncoherent detectors with reasonable complexity. In the following, we classify some of the major approaches to complexity reduction.

Explicit estimation of the channel followed by coherent detection was considered in [11], which uses linear prediction and per-survivor processing. Another approach is to employ Bahl–Cocke–Jelinek–Raviv (BCJR)-type algorithms [17] with a truncated channel memory. Colavolpe *et al.* [8] derived a modified BCJR algorithm for both PSK and QAM signaling over an AWGN channel with an unknown phase shift. Similar detectors for PSK alphabets were also studied in [9] and [10]. The complexity of this approach, however, is exponential in the channel memory. An alternative implementation based on noncoherent sequence detection [18], [19] was proposed in [8], where state-reduction techniques were used to reduce complexity. In contrast, our approach in this paper and in [20], and the approach of [5], is to quantize the unknown channel state. This approach is particularly advantageous for channels with large memory, since the complexity per demodulated symbol is independent of the channel memory. A detailed comparison of the performance and complexity of these various approaches is beyond the scope of this paper.

The phase-quantization approach for reducing complexity was considered previously for noncoherent demodulation alone (see [2], [21], [22], and references in [5]). In other related

work, [23] uses the tangential sphere bound to show that a serial concatenation of a convolutional code with differential encoding outperforms the stand-alone convolutional code. Turbo coding for coherent fading channels has been studied in [24] and [25].

The remainder of this paper is organized as follows. Section II contains information-theoretical results for the noncoherent block fading channel. The proposed capacity-approaching coding and modulation schemes are discussed in Section III. Simulation results are presented in Section IV. Finally, Section V contains conclusions and discusses directions for future work.

II. Block Fading Channel

In this section, we describe the block fading channel model, followed by capacity computations for both unconstrained input (except for a power constraint) and constrained input from finite constellations.

A. Block Fading Channel Model

In a block fading channel model, the complex-baseband transmitted signals undergo fading, characterized by multiplication by a complex Gaussian random variable. This Gaussian random variable is called the *fading coefficient*, which is assumed to remain constant over a block of T symbols. The constant T is also called the *coherence length*. Fading coefficients for different blocks are modeled as independent random variables. The independence condition requires sufficient separation in time or frequency between blocks. Thus, block fading channels can provide good approximations for time-division multiple access (TDMA), frequency hopping, or block-interleaved channels.

For a block fading model, we have

$$\mathbf{y} = h\mathbf{s} + \mathbf{n} \tag{1}$$

where $\mathbf{y}$, $\mathbf{s}$, and $\mathbf{n}$ are $T\times 1$ complex-valued column vectors; $\mathbf{y}$ is the received signal, $\mathbf{s}$ is the transmitted signal with power constraint $E(\mathbf{s}^H\mathbf{s}) \leq T$, and $(\cdot)^H$ denotes the Hermitian operator; h is the fading coefficient, which is complex Gaussian $\mathcal{CN}(0,1)$ distributed with distribution $p(h) = 1/\pi \exp\{-\|h\|^2\}$; and $\mathbf{n}$ is the white Gaussian additive vector with distribution $\mathcal{CN}(0, 2\sigma^2)$.

B. Capacity With Unconstrained Input

In this section, we briefly review capacity results obtained in [3] for noncoherent block fading channels.

Let $H(\cdot)$ denote the differential entropy of a continuous random variable. Also let $\mathrm{tr}(\cdot)$ and $\det(\cdot)$ denote the trace and determinant, respectively, of a square matrix.

The noncoherent channel capacity C, by definition, is equal to

$$\begin{aligned} C &= \max_{p(\mathbf{s}):E(\mathbf{s}^H\mathbf{s})\leq T} [H(\mathbf{y}) - H(\mathbf{y}|\mathbf{s})] \\ &= \max_{p(\mathbf{s}):E(\mathbf{s}^H\mathbf{s})\leq T} \int p(\mathbf{s}) \int p(\mathbf{y}|\mathbf{s}) \log \frac{p(\mathbf{y}|\mathbf{s})}{p(\mathbf{y})} d\mathbf{y}\, d\mathbf{s} \end{aligned} \tag{2}$$

where $p(\mathbf{s})$ is the probability density function (pdf) for the input signal $\mathbf{s}$ and $p(\mathbf{y}|\mathbf{s})$ is the conditional density function.

Any input signal $\mathbf{s}$ can be written as $\mathbf{s} = \mathbf{u}v$, where the complex vector $\mathbf{u} \in \mathbb{C}^T$ is a unit vector and v is a scalar representing the signal amplitude that satisfies the power constraint $E[v^2] \leq T$. It is shown in [3] that $\mathbf{u}$ and v can be assumed independent for the capacity-achieving distribution

$$p(\mathbf{s})|_{\mathbf{s}=\mathbf{u}v} = p(\mathbf{u})p(v). \tag{3}$$

Furthermore, the distribution $p(\mathbf{u})$ is uniform over the T-dimensional complex sphere $\mathbb{S}^T$.

Since $p(\mathbf{u})$ is known, the task of finding the optimal distribution $p(\mathbf{s})$ reduces to a simpler task of finding the optimal distribution $p(v)$. Substituting the independence condition (3) into the expression of the channel capacity (2), we have

$$\begin{aligned} C = \max_{p(v):E(v^2)\leq T} \int p(v) \int p(\mathbf{u}) \int p(\mathbf{y}|\mathbf{s}=\mathbf{u}v) \\ \cdot \log \frac{p(\mathbf{y}|\mathbf{s}=\mathbf{u}v)}{p(\mathbf{y})} d\mathbf{y}\, d\mathbf{u}\, dv. \end{aligned} \tag{4}$$

Based on (4), the capacity C is computed [3] numerically using a modified version of the Blahut–Arimoto algorithm [26]. However, no details of this algorithm were given in [3]. Appendix A gives some of the necessary details in order to compute the unconstrained capacity.

C. Capacity With Constrained Input From Finite Constellations

In this paper, we are interested in computing the noncoherent capacity achieved by constrained inputs from finite constellations and compare them with the unconstrained capacity. In designing practical coded modulation schemes, the rule of thumb is to choose the smallest constellation size which, for the information rate of interest, gives a capacity "close enough" to the capacity with unconstrained input.

Assume that each element s_i of the input signal vector $\mathbf{s}$ is chosen independently from a *finite* constellation set $\mathcal{S}$ with equal probabilities. The channel capacity can then be written as $C = H(\mathbf{y}) - H(\mathbf{y}|\mathbf{s})$.

To compute $H(\mathbf{y}|\mathbf{s})$, we use the fact that $\mathbf{y}$ conditioned on $\mathbf{s}$ is Gaussian. It is known that the differential entropy of a complex Gaussian random variable with the covariance matrix Q is equal to $\log\det(\pi e Q)$. Here, we have $Q = (2\sigma^2)I_T + \mathbf{s}\mathbf{s}^H$, where I_T is a $T \times T$ identity matrix. It follows that

$$H(\mathbf{y}|\mathbf{s}) = T\left[\log(2\pi e\sigma^2) + \log\frac{(1+\mathrm{tr}(\mathbf{s}\mathbf{s}^H)}{(2\sigma^2)}\right].$$

For the first term $H(\mathbf{y})$, we have

$$H(\mathbf{y}) = -\int p(\mathbf{y}) \log[p(\mathbf{y})] d\mathbf{y}. \tag{5}$$

Evaluation of (5) cannot be done in a closed form. A practical approach to numerically evaluate it is by Monte–Carlo integration. A detailed implementation of this method is described in Appendix B.

In Fig. 1, we plot the noncoherent channel capacity computed using various PSK/QAM modulation schemes for coherence

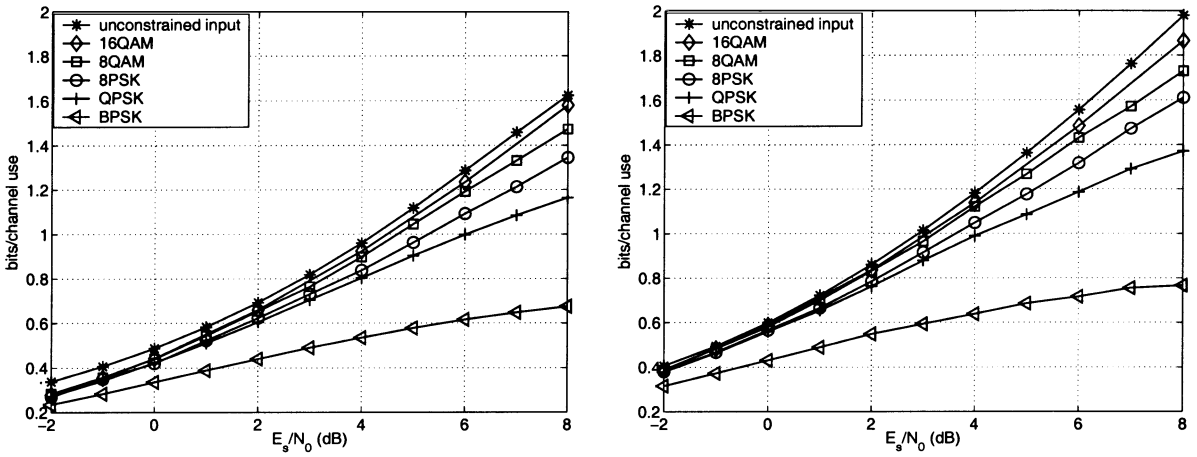


Fig. 1. Noncoherent channel capacity for various constellations.

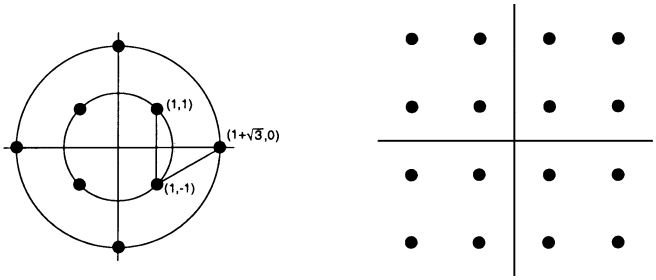


Fig. 2. QAM constellations.

lengths 5 and 10. We use the 8-QAM and 16-QAM constellations shown in Fig. 2. The 8-QAM constellation is known to be the best eight-point QAM constellation because it requires the least power for a given minimum distance between signal points [27]. Also, for the most frequently used rectangular 16-QAM constellation shown in Fig. 2, the average transmitted power required to achieve a minimum distance is only slightly greater than the average power required for the best 16-point QAM constellation [27]. Capacity with unconstrained input is also plotted, serving as benchmarks.

Fig. 1 shows that, at the coherence length $T = 10$ and a transmission rate of about 1/2 information bit/channel use, QPSK is a good choice, being the smallest constellation whose capacity comes close to that of the unconstrained input. However, for higher transmission rates exceeding one information bit/channel use, approaching the capacity with the unconstrained input requires the use of amplitude/phase modulation (e.g., 8-QAM is appreciably better than 8-PSK) and larger constellations. This effect is more pronounced for smaller coherence lengths. Similar conclusions were obtained by Abou-Faycal *et al.* [28] for the extreme case of coherence length $T = 1$.

III. Capacity-Approaching Coding and Modulation Schemes

This section describes our results for designing capacity-approaching coding and modulation schemes. It is organized as follows. In Section III-A, the system model is introduced. The joint demodulation and iterative decoding schemes are discussed in Sections III-B–III-D. Section III-B discusses the structures for the proposed suboptimal noncoherent demodulator, assuming the prior information for the transmitted bits are available from the outer channel codes. Section III-C studies two different types of inner modulation codes: DPSK and B-DPSK. Section III-D gives a brief description of the iterative decoding algorithms for the outer channel codes.

A. System Model

Fig. 3 shows a schematic block diagram of the system. The channel encoder encodes a sequence of information bits $\mathbf{b}$ into a sequence of coded bits $\mathbf{c}$. The resulting coded sequence is interleaved by a random permutation and then mapped to a sequence of MPSK symbols $\mathbf{w}$ using Gray mapping. The sequence $\mathbf{w}$ is then divided into blocks of $T-1$ symbols, where T is equal to the coherence length of the channel. The function of the inner modulation coder is to code blocks of $T-1$ input symbols into blocks of T output symbols. Each block of T output symbols is then passed to the transmit filter, and sent through the block fading channel.

At the complex baseband receive filter, for each block of received samples, the block demodulator computes *a posteriori* probabilities (APPs) of the input PSK symbols. Subsequently, these symbol APPs are used to compute bit-wise APPs (a precise description follows in Section III-B). The "extrinsic" part of these bit-wise APPs is then deinterleaved and passed to the channel decoder. The channel decoder performs one decoding iteration and generates bit-wise extrinsic information $\{\eta^b\}$. Assuming that the bits which constitute a PSK symbol are statistically independent, the interleaved bit-wise extrinsic information $\{\lambda^b\}$ is fed back to the demodulator, which updates the prior symbol probabilities. For the next iteration, the demodulator computes symbol APPs using updated prior probabilities. In this manner, the demodulation and channel decoding proceed iteratively. After a fixed number of iterations, decisions are made at the output of the channel decoder to generate the decoded bits.

We introduce some notation that we will use throughout this paper.

- Denote the size of the MPSK constellation by M and the signal set by $\mathcal{S} = \{1, e^{j2\pi/M}, \ldots, e^{j2\pi/M(M-1)}\}$. Alternatively, define the index set for $\mathcal{S}$ as $\mathcal{M} = \{0, 1, \ldots, M-1\}$, where each index $i \in \mathcal{M}$ represents

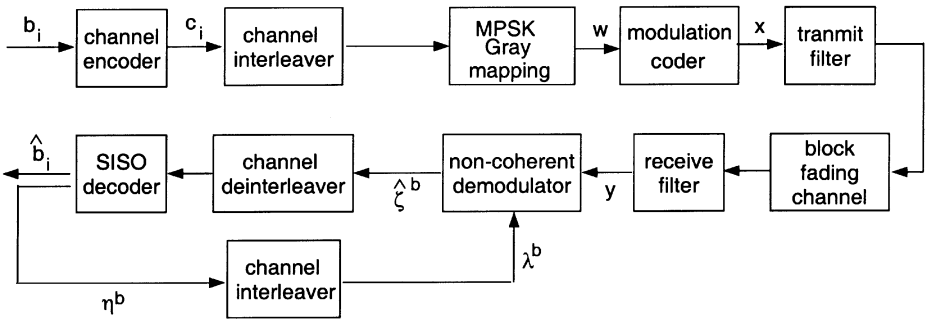


Fig. 3. Schematic block diagram of the system.

a constellation point $s_i = e^{j2\pi/Mi}$. We restrict attention to the case where M is a power of two, and define $m = \log_2 M$ as the number of bits in each MPSK symbol.

- Denote the input vector for the modulation coder by $\mathbf{w} = (w_1, \ldots, w_{T-1})$. For each $i = 1, \ldots, T-1$, $w_i \in \mathcal{M}$ represents an input MPSK symbol encoding m bits. The bits encoded in w_i are denoted by $\{w_i^j, j \in 1, \ldots, m\}$.
- Denote the output vector of the modulation coder corresponding to the input vector $\mathbf{w}$ by $\mathbf{x}(\mathbf{w}) = (x_0(\mathbf{w}), x_1(\mathbf{w}), \ldots, x_{T-1}(\mathbf{w}))$. Each $x_i(\mathbf{w}) \in \mathcal{M}$ is the index for the MPSK symbol $e^{j2\pi/Mx_i(\mathbf{w})}$. We also write $\mathbf{x} = (x_0, x_1, \ldots, x_{T-1})$ instead of $\mathbf{x}(\mathbf{w}) = (x_0(\mathbf{w}), x_1(\mathbf{w}), \ldots, x_{T-1}(\mathbf{w}))$.

For the modulation codes considered in this paper, x_0 is a reference symbol inserted by the modulation coder and is independent of $\mathbf{w}$. To simplify notation, we let $w_0 = x_0$. Even though we only consider the block fading channel here, the block modulation codes proposed are designed to accommodate a block approximation of the continuously time-varying channel. Therefore, in general, we do not assume that x_0 is a pilot symbol known by the demodulator. More details are given in Section III-C.

B. Noncoherent Block Demodulator

In this section, we give a description of the noncoherent block demodulator employed in our iterative decoding scheme. Due to the statistical independence of the fading coefficients in adjacent blocks and the block encoding scheme, the noncoherent demodulator operates blockwise.

For each bit w_i^j, $1 \le i \le T$, $1 \le j \le m$, let $\lambda_{ij}^b(a) = P(w_i^j = a)$, $a = 0,1$, denote the input bit-wise priors provided by the channel decoder. We assume that these bit-wise priors associated with the same block are independent due to the presence of the channel interleaver. Based on this independency assumption, the demodulator takes the following three steps to compute the output bit-wise extrinsic information to be passed back to the channel decoder.

1) For each symbol w_i, the demodulator generates symbol-wise priors $\{\lambda_i^s(w_i)\}$ from input bit-wise priors as

$$\lambda_i^s(w_i) = \prod_{j=1}^{m} \lambda_{ij}^b(w_i^j).$$

2) For each symbol w_i, the demodulator computes symbol-wise APPs $\{\zeta_i^s(u), u = 0, \ldots, M-1\}$ as

$$\begin{aligned}\zeta_i^s(u) &= P(w_i = u|\mathbf{y}) \\ &= \sum_{\mathbf{x}:w_i=u} P[\mathbf{x}|\mathbf{y}] = \sum_{\mathbf{x}:w_i=u} \frac{p[\mathbf{y}|\mathbf{x}]P[\mathbf{x}]}{p[\mathbf{y}]} \\ &\propto \sum_{\mathbf{x}:w_i=u} p[\mathbf{y}|\mathbf{x}]P[\mathbf{w}]P[w_0] \\ &= \sum_{\mathbf{x}:w_i=u} p[\mathbf{y}|\mathbf{x}] \left[\prod_{j=0}^{T-1} \lambda_j^s(w_j)\right]. \end{aligned} \tag{6}$$

To make our notation consistent, we let $\lambda_0^s(w_0) = P(w_0)$ denote the prior probability for the reference symbol w_0. Computing $p[\mathbf{y}|\mathbf{x}]$ turns out to be the bottleneck in terms of computational efficiency. We focus on this problem in Sections III-B.1 and III-B.2.

3) For each bit w_i^j, the demodulator generates updated bit-wise APPs $\{\zeta_{ij}^b(a), a = 0,1\}$ and the bit-wise extrinsic information $\{\hat{\zeta}_{ij}^b(a), a = 0,1\}$. The latter will be passed back to the channel decoder.

Using the symbol-wise APPs computed in step 2, we have

$$\zeta_{ij}^b(a) = \sum_{w_i:w_i^j=a} \zeta_i^s(w_i), \quad a = 0,1.$$

Then the bit-wise extrinsic information $\{\hat{\zeta}_{ij}^b(a)\}$ is obtained by removing bit-wise priors $\{\lambda_{ij}^b(a)\}$ from $\{\zeta_{ij}^b(a)\}$

$$\hat{\zeta}_{ij}^b(a) = \frac{\frac{\zeta_{ij}^b(a)}{\lambda_{ij}^b(a)}}{\frac{\zeta_{ij}^b(a)}{\lambda_{ij}^b(a)} + \frac{\zeta_{ij}^b(1-a)}{\lambda_{ij}^b(1-a)}}, \quad a = 0,1.$$

Next, we discuss how to efficiently compute the summation in (6).

1) Optimal Noncoherent Demodulator: When a PSK constellation is used, we simplify the noncoherent pdf given in [3] to obtain the following:

$$\begin{aligned} p[\mathbf{y}|\mathbf{x}] = \frac{1}{(2\pi\sigma^2)^T} e^{-\left[\left(\sum_{i=0}^{T-1} \|y_i\|^2\right)/2\sigma^2\right]} \\ \cdot \frac{1}{1+\frac{T}{2\sigma^2}} e^{\left[\left(\left\|\sum_{i=0}^{T-1} (e^{j(2\pi/M)x_i})^H y_i\right\|^2\right)/(2\sigma^2(2\sigma^2+T))\right]}. \end{aligned} \tag{7}$$

For any $0 \leq a \leq M-1$, define a new vector $\mathbf{x}'$ by $x'_i = (x_i + a) \bmod M$ for every $0 \leq i \leq T-1$. In other words, x' is obtained from $\mathbf{x}$ by a rotation in the signal space. Note that the noncoherent pdf satisfies the *rotational invariance* property: $p(\mathbf{y}|\mathbf{x}) = p(\mathbf{y}|\mathbf{x}')$.

For the inner modulation codes considered in this paper, the number of terms in the summation of (6) is equal to M^{T-1}, which increases exponentially with the coherence length. In contrast to the coherent setting where h is known, the noncoherent conditional density $p[\mathbf{y}|\mathbf{x}]$ shown in (7) cannot be computed recursively, because the second exponential in $p[\mathbf{y}|\mathbf{x}]$ does not decompose into a product of suitable individual terms. Therefore, when (7) is used to compute symbol APPs, the resulting complexity is prohibitive, even for moderate coherence lengths.

2) Suboptimal Noncoherent Demodulator: In the following section, we derive an approximate approach close to the optimal solution, while requiring complexity that grows only linearly in the coherence length. The main idea of the proposed noncoherent demodulator is to approximate the noncoherent channel by a set of coherent channels. This is achieved by discretizing the unknown channel fading coefficient $h = Ae^{j\theta}$ with the amplitude A and the phase shift θ. We treat these separately.

We derive an averaging estimator $\hat{A}$ for the fading amplitude A

$$\hat{A}^2 = \max\left(\frac{1}{T}\sum_{i=0}^{T-1} \|y_i\|^2 - 2\sigma^2, 0\right). \tag{8}$$

Once $\hat{A}$ is computed, we keep it fixed for all iterations of joint demodulation and iterative decoding.

This estimator is simple because it does not require any prior knowledge of the transmitted symbols. As shown in the simulation results, $\hat{A}$ is rather robust. For moderate-to-small coherence lengths, this estimator incurs only 0.3–0.5 dB performance degradation, compared with a genie-aided system in which the fading amplitude is *perfectly* known. Thus, as expected, the performance for PSK alphabets is more sensitive to the channel phase than to the amplitude. However, we anticipate that our coarse amplitude estimator will need to be improved for good performance with multiamplitude signaling such as QAM.

We assume that θ takes only L discrete values in the interval $[0, 2\pi]$

$$\theta \in \Phi = \left\{0, \frac{1}{L}2\pi, \frac{2}{L}2\pi, \ldots, \frac{L-1}{L}2\pi\right\}. \tag{9}$$

Define $\phi_l = l/L2\pi$ for every $0 \leq l < L$.

Assuming that the estimated fading amplitude is $\hat{A}$, we have

$$\begin{aligned} p[\mathbf{y}|\mathbf{x}, \|h\| = \hat{A}] &\approx \sum_{l=0}^{L-1} p[\theta = \phi_l, \mathbf{y}|\mathbf{x}, \|h\| = \hat{A}] \\ &= \sum_{l=0}^{L-1} p[\theta = \phi_l|\mathbf{x}, \|h\| = \hat{A}] p[\mathbf{y}|\mathbf{x}, \|h\| = \hat{A}, \theta = \phi_l] \\ &\propto \sum_{l=0}^{L-1} p[\mathbf{y}|\mathbf{x}, \|h\| = \hat{A}, \theta = \phi_l] \\ &\propto \sum_{l=0}^{L-1}\prod_{k=0}^{T-1} \exp\left\{-\frac{\|y_k - \hat{A}e^{j(\phi_l + (2\pi/M)x_k)}\|^2}{2\sigma^2}\right\}. \end{aligned} \tag{10}$$

We now substitute (10) into (6) to approximate $p(\mathbf{y}|\mathbf{x})$. It follows that for every $i \geq 1$, we get the result shown in (11) at the bottom of the page.

Based on the last equality of (11), we can compute $\zeta^s(w_i)$ recursively using the BCJR algorithm [17] or the sum–product algorithm [29]. The overall computation requirement is equivalent to L times the complexity of a BCJR algorithm applied to the coherent demodulation of the block modulation code.

C. Block Modulation Codes

Block modulation codes suitable for noncoherent communication were considered by Sun and Leib [30], Warrier and Madhow [2], and Peleg and Shamai [7]. Codes considered in [30] and [2] are especially effective for higher SNR, in which case they provide significant coding gains. Codes considered in [7] are effective for lower SNR. In our system, coding gains are realized by employing powerful outer channel codes in conjunction with very simple inner modulation codes with iterative information exchange between the inner demodulator and the outer decoder. For the relatively low SNRs considered in this paper, this approach appears to suffice for approaching channel capacity. Next, we consider two examples of such modulation codes based on differential encoding. In particular, we concentrate on the form that (11) takes in this case.

1) M-ary DPSK: The first modulation code structure we consider is the standard M-ary DPSK. The encoding enforces the rule

$$x_i = \begin{cases} w_0, & \text{if } i = 0 \\ (x_{i-1} + w_i) \bmod M, & \text{if } 1 \leq i \leq T-1. \end{cases}$$

$$\begin{aligned} \zeta_i^s(u) &= P(w_i = u|\mathbf{y}) \\ &\propto \sum_{\mathbf{x}: w_i = u}\left(\sum_{l=0}^{L-1}\prod_{k=0}^{T-1} \exp\left\{-\frac{\|y_k - \hat{A}e^{j(\phi_l + (2\pi/M)x_k)}\|^2}{2\sigma^2}\right\}\right)\left[\prod_{j=1}^{T-1}\lambda_j^s(w_j)\right] \\ &= \sum_{l=0}^{L-1}\left(\sum_{\mathbf{x}: w_i = u}\prod_{k=0}^{T-1}\left[\exp\left\{-\frac{\|y_k - \hat{A}e^{j((2\pi/L)l + (2\pi/M)x_k)}\|^2}{2\sigma^2}\right\}\lambda_k^s(w_k)\right]\right) \end{aligned} \tag{11}$$

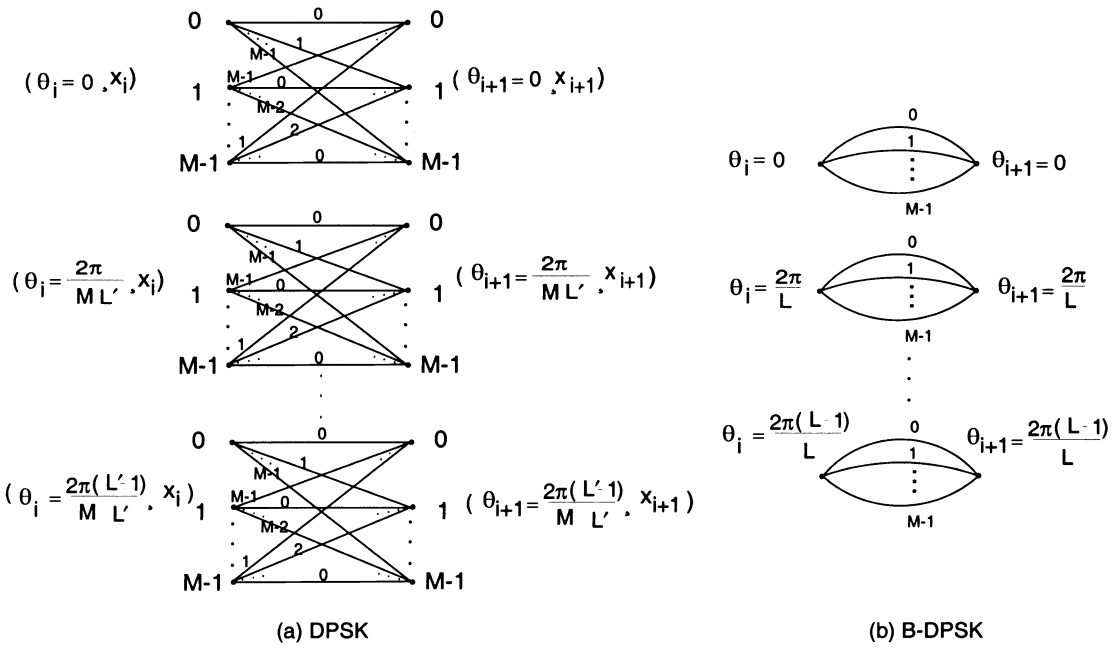


Fig. 4. Trellis sections for the noncoherent demodulator.

The first symbol w_0 is the reference symbol. For the block fading channel, it can be a known pilot symbol inserted by the modulation coder; thus, the transmission rate is reduced by a factor of $(T-1)/T$. Such a transmission rate loss can be avoided when using a block fading approximation for a continuously fading channel, by overlapping successive blocks by one symbol, using the last symbol of the previous block as the reference symbol for the encoding of the current block. In this case, the demodulator does not know the value of w_0. Applying the rotational invariance property of $p(\mathbf{y}|\mathbf{x})$ to (6) and using the fact that w_0 is independent of the other symbols w_i in the same block, one can see that the output APPs do not depend on the priors of w_0.

Without loss of generality, therefore, we assume that w_0 takes any value in $\mathcal{M}$ with equal probability. Next, we show that, when the quantization level L is an integer multiple of M, (11) can be simplified. Instead of quantizing the phase shift θ in the interval $[0, 2\pi]$ as shown in (9), we quantize it in the subinterval $[0, 2\pi/M]$, thus reducing the number of quantization levels. The reduced phase quantization level L' is equal to L/M. Writing any number l, $0 \le l < L$ in the form $l = m'L' + l'$, where $0 \le m' < M$ and $0 \le l' < L'$, we can simplify (11) in the last equality of (12), shown at the bottom of the page.

We next give a graphical description of the demodulation scheme suitable for an efficient BCJR algorithm. A trellis is a graphical structure illustrating the dynamical behavior of a system. We define the trellis structure for the DPSK demodulator as follows. There are a total of T trellis sections and $T+1$ trellis state classes. Let $\mathcal{S}_i$ denote the ith state class and let $s \in \mathcal{S}$ be a particular state. Each state $s \in \mathcal{S}_i$ represents a 2-tuple (θ_i, x_i), where $\theta_i \in \{0, (2\pi/L'M), \ldots, ((L'-1)2\pi/L'M)\}$ and $x_i \in \mathcal{M}$. The number of possible values of each state equals $M \cdot L'$. Branch transitions between state $s_i = (\theta_i, x_i) \in \mathcal{S}_i$ and $s_{i+1} = (\theta_{i+1}, x_{i+1}) \in \mathcal{S}_{i+1}$ are allowed if and only if $\theta_i = \theta_{i+1}$ and $x_{i+1} = (x_i + w_i) \bmod M$. Hence, it can be seen that the trellis is composed of complete bipartite graphs of $2M$ vertices, as depicted in Fig. 4(a). In accordance with the last equality of (12), the branch

$$
\begin{aligned}
\zeta_i^s(u) &\propto \sum_{l=0}^{L-1}\left(\sum_{\mathbf{x}:w_i=u}\prod_{k=0}^{T-1}\left[\exp\left\{-\frac{\|y_k-\hat{A}e^{j((2\pi/L)l+(2\pi/M)x_k)}\|^2}{2\sigma^2}\right\}\lambda_k^s(w_k)\right]\right)\\
&= \sum_{l'=0}^{L'-1}\left[\sum_{m'=0}^{M-1}\left(\sum_{\mathbf{x}:w_i=u}\prod_{k=0}^{T-1}\left[\exp\left\{-\frac{\|y_k-\hat{A}e^{j((2\pi/M)(m'+(l'/L'))+(2\pi/M)x_k)}\|^2}{2\sigma^2}\right\}\lambda_k^s(w_k)\right]\right)\right]\\
&= \sum_{l'=0}^{L'-1}\left[\sum_{m'=0}^{M-1}\left(\sum_{\mathbf{x}:w_i=u}\prod_{k=0}^{T-1}\left[\exp\left\{-\frac{\|y_k-\hat{A}e^{j((2\pi/M)(l'/L')+(2\pi/M)(x_k+m'))}\|^2}{2\sigma^2}\right\}\lambda_k^s(w_k)\right]\right)\right]\\
&= M\cdot\sum_{l'=0}^{L'-1}\left(\sum_{\mathbf{x}:w_i=u}\prod_{k=0}^{T-1}\left[\exp\left\{-\frac{\|y_k-\hat{A}e^{j((2\pi/M)(l'/L')+(2\pi/M)x_k)}\|^2}{2\sigma^2}\right\}\lambda_k^s(w_k)\right]\right) \qquad (12)
\end{aligned}
$$

metric for each of the transitions defined above is equal to $\lambda_i^s(w_i) \cdot \exp\left\{ -(\|y_i - \hat{A}e^{j(\theta_i+(2\pi/M)x_i)}\|^2)/2\sigma^2 \right\}$, $i \in 0,\ldots,T-1$.

In the definition above, state transitions are not allowed between states with different values of θ. Therefore, the total number of branches within each trellis section is $M^2 \cdot L'$.

The trellis structure proposed here resembles that of [5]. The main distinction is that the trellis structure proposed here is the product [31] of two trellises, one corresponding to the modulation coder, and the other one corresponding to the phase shift of the channel. This is a general structure that can be modified to accommodate time-varying channels, and to support modulation coders beyond differential coders with PSK signaling. The complexity of the proposed trellis structure is the same as that of [5].

2) Block-Based M-ary DPSK: An alternative block modulation code relates all symbols directly to the first reference symbol in each block. We refer to this modulation code as block-based M-ary DPSK (B-DPSK).

For each input $\mathbf{w}$, the B-DPSK encoder generates output $\mathbf{x}$ as

$$x_i = \begin{cases} w_0, & \text{if } i = 0 \\ (w_0 + w_i) \bmod M, & \text{if } 1 \le i \le T-1. \end{cases}$$

When the reference symbol is placed in the center of each block and all other symbols are differentially encoded to it, one obtains the RDE code studied by Peleg *et al.* [5]. For the block fading channel, B-DPSK and RDE are equivalent because the location of the reference symbol is irrelevant, as the channel remains constant within each block. The advantage of B-DPSK comes when the block fading model is used as an approximation for a continuously fading channel, with the reference symbol taken to be the last symbol of the previous block to overcome the transmission rate loss of the factor $(T-1)/T$. The RDE code can also overcome such rate loss using similar techniques described in [5] and [7].

Using the rotational invariance property of the noncoherent density function, the demodulator can assume that $w_0 = 0$, even though for the B-DPSK encoder w_0 may take any values between 0 and $M-1$. Under this assumption, as far as the demodulator is concerned, the encoding rule becomes

$$x_i = \begin{cases} 0, & \text{if } i = 0 \\ w_i, & \text{if } 1 \le i \le T-1. \end{cases}$$

In other words, the demodulator treats the first output symbol as a pilot symbol and the remaining $T-1$ output symbols as uncoded information symbols. As shown in later sections, despite its simplicity, B-DPSK works surprisingly well with certain outer channel codes.

The trellis structure of the noncoherent demodulator with B-DPSK is described as follows. There are a total of T trellis sections and $T+1$ trellis state classes. Let $\mathcal{S}_i$ denote the ith state class. Here each $\mathcal{S}_i$ is equal to the quantization set $\Phi = \{0, (2\pi/L), \ldots, ((L-1)2\pi/L)\}$. Each particular state $\theta_i \in \mathcal{S} = \Phi$ represents a given quantization level. The possible values of each state equal to L. Branch transitions between state $\theta_i \in \mathcal{S}_i$ and $\theta_{i+1} \in \mathcal{S}_{i+1}$ are allowed if and only if $\theta_i = \theta_{i+1}$. It follows that the trellis is composed of complete bipartite graphs of two vertices, as depicted in Fig. 4(b). There are a total of M branches corresponding to the M possible inputs for the information symbol w_i. The branch metric for each of the branches defined above is equal to $\lambda_i^s(w_i) \cdot \exp\left[-(\|y_i - \hat{A}e^{j(\theta_i+(2\pi/M)w_i)}\|^2)/2\sigma^2 \right]$, $i = 0,\ldots,T-1$. Note that since the first symbol w_0 is treated as a pilot symbol, we have $\lambda_0^s(w_0 = 0) = 1$. The number of valid branches for a given trellis section equals to $M \cdot L = M^2 \cdot L'$, which is the same as that of the DPSK demodulator. It is known that the complexity of a BCJR algorithm is proportional to the number of trellis branches [32]. Hence, the complexity of running a BCJR algorithm on the trellis of the B-DPSK demodulator is about the same as that of the DPSK demodulator.

D. Iterative Decoding Algorithms for the Outer Channel Codes

In previous sections, we discussed the trellis structures of the noncoherent demodulator for different modulation codes. Based on such structures, the demodulator can be implemented as a SISO module using a BCJR-type decoding algorithm. The outer channel decoder can also be regarded as a SISO module. The input to the SISO module for the decoder is fed by the output of the SISO demodulator. However, only one iteration is performed inside the SISO decoding module for the channel decoder during each joint demodulation and decoding cycle.

Decoding algorithms for various outer channel codes considered in this paper are well investigated in the literature. For turbo decoding, we refer to [4], and for decoding of RA codes, we refer to [33]. A general description of SISO decoding of serial concatenated modulus is given in [16].

IV. Simulation Results

In this section, we evaluate the performance of the proposed joint demodulation and iterative decoding algorithm by simulating the system shown in Fig. 3. Various choices of outer channel codes and inner modulation codes are considered. The block fading channel defined in Section II-A is used in all simulations.

A. Coding and Modulation Strategies

Three types of outer channel codes are considered: an RA code, a convolutional code, and a turbo code, each with a rate of 1/4 and a codeword length of 64 000 bits. We use the regular rate 1/4 RA code as defined in [33]. For the convolutional code, we consider two different systematic codes, a recursive and a nonrecursive one. For the turbo code, a systematic code using parallel concatenation of two component codes of rates 1/3 and 1/2 is considered. The overall rate of the turbo code is also 1/4. For the inner modulation codes, we consider both types, DPSK and B-DPSK, and each has a rate of $(T-1)/T$. We always use QPSK modulation in this paper.

The code parameters for the convolutional codes and the turbo code are shown in Table I.

TABLE I
PARAMETERS OF OUTER CHANNEL CODES

outer channel codes		rate	generator polynomial
convolutional	recursive	1/4	[25 27 33 37]
code	non-recursive	1/4	[20 25 27 33]
turbo	first component code	1/3	[25 33 37]
code	second component code	1/2	[23 35]

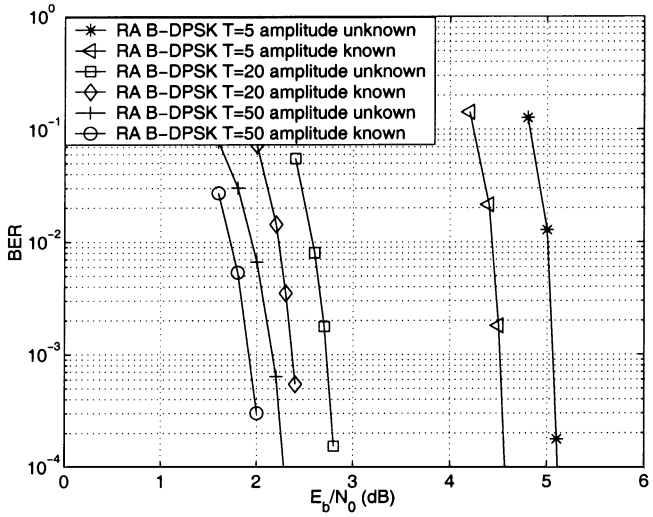


Fig. 5. Performance of the amplitude estimator.

1) Noncoherent Demodulator: Amplitude Estimation and Phase Quantization: In this section, we look at the performance loss due to the simple averaging amplitude estimator of (8) and the finite number of quantization levels of (9).

In Fig. 5, we compare the performance of the noncoherent demodulator with unknown phase and unknown amplitude with the genie-aided noncoherent demodulator with unknown phase but known amplitude. The latter is implemented by replacing $\hat{A}$ by the true amplitude $||h||$ in (10). This gives an upper bound on the performance degradation due to the crudeness of the amplitude estimator. We use the example of the RA code of a codeword length 64 000 with the B-DPSK modulation code. The number of iterations is set to 20. At $T = 50$, the performance degradation inflicted by the amplitude estimator is less than 0.3 dB. Performance degrades slightly for smaller T. Even for a very small coherence length $T = 5$, the performance is still within 0.5 dB of the case when the amplitude is perfectly known.

The number of quantization levels also affects the system performance. Our simulation results show that when B-DPSK is used, it is sufficient to use 20 quantization levels. Increasing the number of quantization levels beyond 20 does not lead to significant performance improvement. For DPSK, the number of quantization levels needed is roughly $1/M$ of that of the B-DPSK. Hence, four or five quantization levels are sufficient.

2) Different Combinations of Outer Channel Codes and Inner Modulation Codes: Our results show that the best choice of the modulation code is closely tied to the choice of the outer channel code.

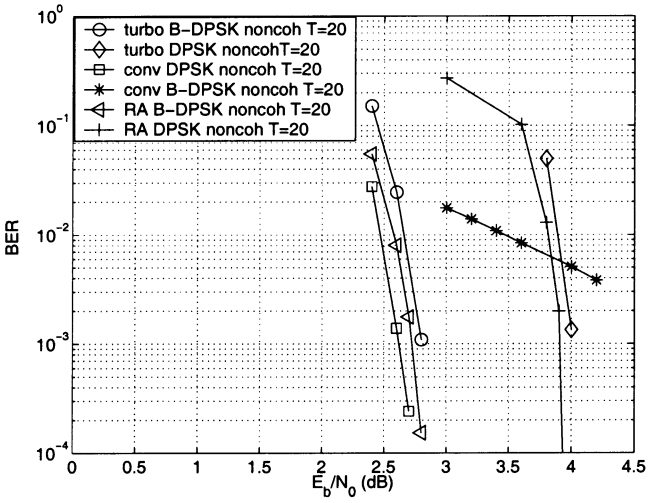


Fig. 6. Performance of different combinations of outer channel codes and inner modulation codes.

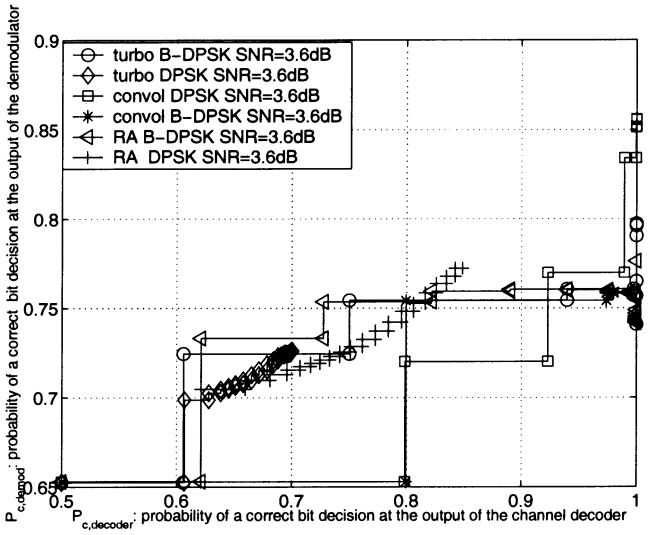


Fig. 7. Probability of a correct bit decision through iterations.

In Fig. 6, we compare the performance of different combinations of the three outer channel codes (convolutional, turbo, and RA) with the two modulation codes (DPSK and B-DPSK). The overall codeword length is 64 000 and $T = 20$. The RA code and the turbo code perform better with B-DPSK than with DPSK, since B-DPSK does not disturb their turbo-like structures. Similar phenomena were also observed by Peleg *et al.* [5] for AWGN channels. The convolutional code performs much better with DPSK than with B-DPSK, since serial concatenation with DPSK gives the overall code a turbo-like structure that it lacks when used with B-DPSK. Indeed, when B-DPSK is used with the convolutional code, the sharp turbo-like waterfall region seen for the other cases is no longer present.

In order to have a better understanding of the joint demodulation and iterative decoding process, we track the probability of a correct bit decision through iterations in the same spirit of EXIT charts [34]. The results are shown in Fig. 7. The x axis represents $P_{\mathrm{c,decoder}}$, the probability of a correct bit decision at the

TABLE II
CAPACITY BENCHMARKS FOR BLOCK FADING CHANNEL

Coherence length T	10	20	50
Transmission rate (information bit/channel use)	0.45	0.475	0.49
Minimum E_b/N_0 (dB) for unconstrained input	1.99	1.22	0.52

output of the decoder. The y axis represents $P_{\mathrm{c,demod}}$, the probability of a correct bit decision at the output of the demodulator. There are six curves in the plot which correspond to all possible combinations of the three outer channel codes and the two inner modulation codes. Each curve, representing a code combination, is obtained by interpolating a sequence of points. The coordinates of each point indicate the probability of a correct bit decision at the output of the demodulator and the decoder at the end of a certain iteration.

For the first iteration, since the soft input to the demodulator is uniformly distributed, the output symbol APP has the same distribution for DPSK and B-DPSK. Therefore, $P_{\mathrm{c,demod}}$ is identical. For a fixed outer channel code, $P_{\mathrm{c,decoder}}$ is also the same at the end of the first iteration. It is only after the first iteration that systems with different modulation codes start behaving differently. As seen from the plot, the RA code and the turbo code in conjunction with B-DPSK exhibit fast convergence. In four or five iterations, $P_{\mathrm{c,decoder}}$ approaches one. In contrast, when DPSK is used, it takes up to 20 iterations before the system stabilizes. Even then, there are still many incorrect bit decisions at the end of the decoder output.

Interestingly, for the convolutional code, $P_{\mathrm{c,decoder}}$ reaches as high as 0.8 after only one iteration. Note that $P_{\mathrm{c,demod}}$ increases rapidly within the next four iterations. By the time the system converges, $P_{\mathrm{c,demod}}$ equals approximately 0.85, which is higher compared to that of other code combinations.

In summary, the following code combinations lead to a superior performance:

- turbo code + B-DPSK;
- RA code + B-DPSK;
- convolutional code + DPSK.

In the remainder of this paper, we report simulation results for the three combinations above only.

B. Capacity Benchmark

Denoting the received energy per symbol by E_s and the received energy per information bit by E_b, we have the following relationship:

$$\frac{E_s}{N_0} = \frac{E_b}{N_0} \cdot R$$

where R is the transmission rate, expressed in information bit per channel use. Assume that the outer channel code has a rate of 1/4. When QPSK modulation is used, we have for both modulation codes

$$R = \frac{T-1}{T} \cdot \frac{1}{4} \cdot 2 \tag{13}$$

where the term $(T-1)/T$ takes into account the rate loss due to the reference symbol per block.

We consider three different coherence lengths: 10, 20, and 50. For each coherence length, we list in Table II the transmission rate computed from (13), and the minimum E_b/N_0 required to transmit at such a rate based on capacity computations for unconstrained input.

C. System Performance

In this section, we present various figures corresponding to different coherence lengths ($T = 10$, 20, and 50). In each figure, we plot the performance curves for three good combinations of outer channel codes and inner modulation codes mentioned in the previous section, using the proposed joint demodulation and iterative decoding algorithm. For comparison, we also provide performance curves using coherent demodulation for the same codes. The number of iterations for both coherent demodulation and noncoherent demodulation equals 20.

Fig. 8(a)–(c) contain simulation results for long outer channel codes of a codeword length 64 000. The main observations from these figures are as follows.

1) The system performance depend on the coherence length T. At BER $= 10^{-4}$, the RA code performs the best for $T = 10$. The nonrecursive convolutional code performs the best for $T = 20$. For $T = 50$, the turbo code performs the best. The gap between the Shannon limit and what is achieved by the best code in each plot is approximately 1.6 to 1.7 dB.
2) The choice of codes should depend on whether coherent or noncoherent detection is used. When coherent detection is used, the turbo code performs significantly better (by about 0.5–0.7 dB) than the RA code. However, for noncoherent detection, the RA code performs better for moderate coherence lengths of $T = 10$ and $T = 20$. Only for $T = 50$, where the longer coherence length enables a more accurate implicit channel estimate using noncoherent detection, is the turbo code superior to the RA code.
3) The convolutional code with DPSK performs well. For $T = 10$, its performance is better than that of the RA code and the turbo code for SNR smaller than 3.6 dB. At higher SNRs, it exhibits an error floor. For $T = 20$, it outperforms both the RA and turbo codes. However, its performance is appreciably worse than that of the RA and the turbo code for $T = 50$, possibly due to the long fades associated with such a large coherence length.
4) The joint demodulation and decoding algorithms proposed in this paper lead to significant gains over the traditional schemes that employ standard two-symbol SISO differential demodulation followed by channel decoding. No iterative information exchange between the demodulator and the decoder is used in these traditional

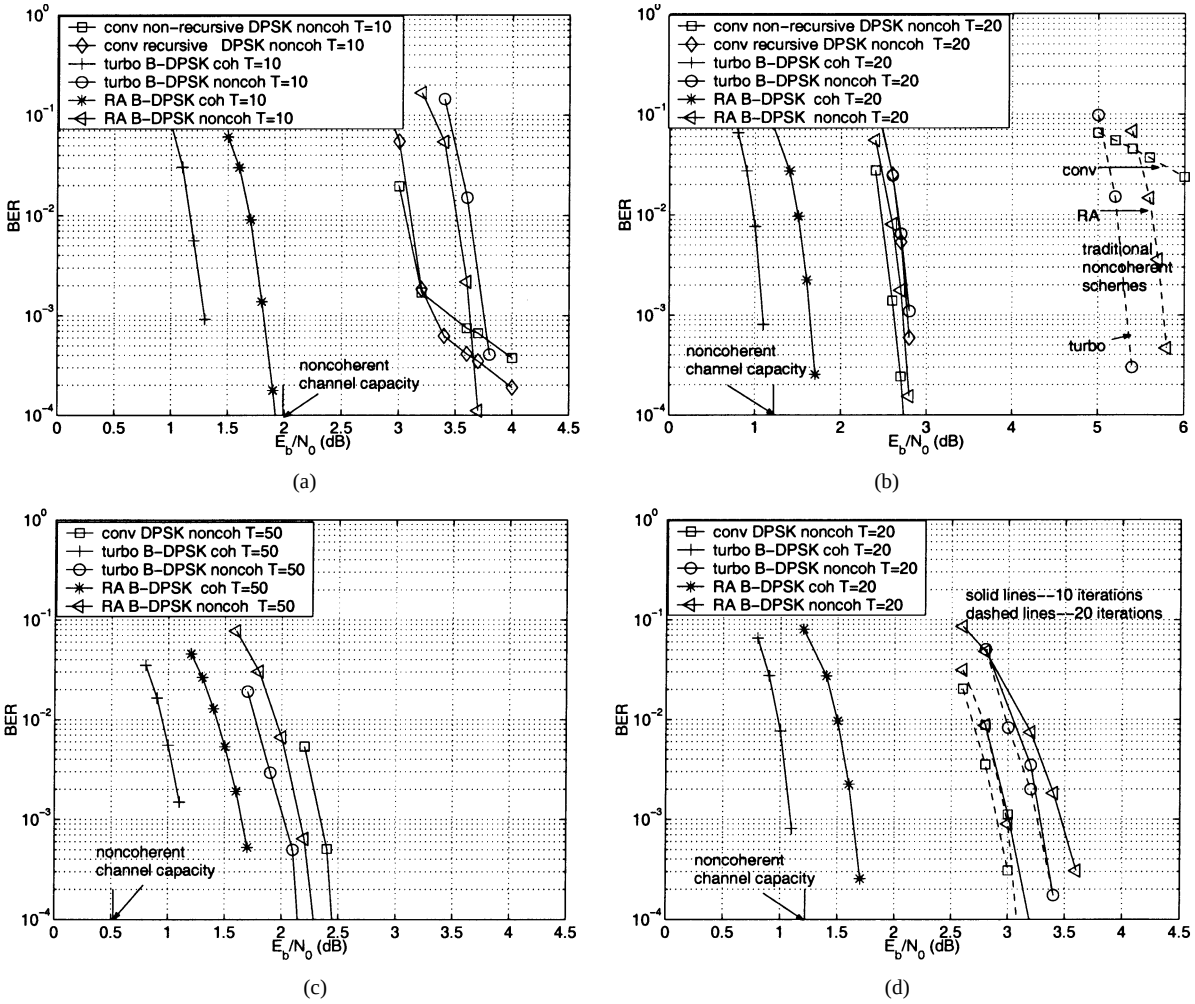


Fig. 8. Performance comparisons of different code combinations.

schemes. For $T = 20$, the performance of the traditional schemes are plotted in Fig. 8(b) using dashed lines. The performance gain of the proposed algorithm compared to the traditional schemes is about 2.5–3 dB for the RA code and the turbo code, and is even higher for the convolutional code.

So far, we have only examined the performance of long outer channel codes of a codeword length 64 000. In practice, codes of shorter lengths may be preferred because of constraints on the decoding delay. Fig. 8(d) displays simulation results for codes of a codeword length 16 000 bits. The coherence length is 20. At the end of the first 10 iterations, the convolutional code outperforms both the RA code and the turbo code. By the end of 20 iterations, the RA code performs fairly close to the convolutional code, while the performance of the turbo code barely improves after the first 10 iterations.

V. Conclusions

Our investigation shows that there are three key aspects to be considered in the design of noncoherent coded modulation systems.

1) *Information-Theoretic Aspect*

The transmitted signal constellation set $\mathcal{S}$, which consists of $T \times 1$ (T is the coherence length) signal vectors to be transmitted through the block fading channel, should be chosen such that the channel capacity achieved for this signal set is close to the unconstrained (except for the constraint on power) channel capacity. We show that independent and identically distributed (i.i.d.) input from appropriately chosen standard PSK and QAM alphabets works well for a wide range of SNRs and channel coherence lengths.

2) *Complexity Aspect*

Modulation codes should allow for efficient decoding. For a transmission rate of R bits per channel use, the total number of transmitted signals over a block equals 2^{RT}. This is usually a large number which makes an exhaustive search impossible. Therefore, efficient decoding schemes must be available for the modulation codes. Our current encoding and decoding schemes are optimized for PSK alphabets, and may need modification for QAM alphabets, including the development of more sophisticated methods for handling the unknown channel amplitude and phase, and generalization of differential modulation to QAM constellations [2], [22].

3) *Compatibility Aspect*

In order to optimize performance, the inner modulation code should be chosen to match the outer channel code, so as to introduce or preserve the"turbo effect."

For the setting considered in our simulation results, it remains an open issue as to how to account for the 1.6–1.7 dB gap to capacity, and whether it is practically feasible to further close this gap. As shown in the simulation results, the amplitude estimator can account for, at most, 0.3–0.4 dB, so there may be a few tenths of a dB to be gained by improving the amplitude estimator. Another possibility is to optimize the tradeoff between the complexity of the outer channel code and inner modulation code, using modulation codes that are more sophisticated than the rate one DPSK and B-DPSK codes. However, the results from initial experiments on such modulation codes have not been promising. Finally, the use of a finite blocklength contributes to the gap, especially because compared to the AWGN channel, a larger degree of time averaging is needed for good performance over fading channels.

Appendix A
Capacity Computation With Unconstrained Input

We state in this appendix the modified Blahut–Arimoto algorithm due to Marzetta and Hochwald [3] for computing the optimal input distribution $p^*(v)$, where v is a random scalar coming from the signal decomposition $\mathbf{s} = \mathbf{u}v$.

Theorem 1: Given the parameter $s \geq 0$, for any distribution $p^r(v)$, let $c^r(v)$ be as shown in (A.1) *at the bottom of the page. If the initial distribution $p^0(v)$ is strictly positive for any $v \geq 0$, then as $r \to \infty$, the sequence of probability distributions defined by*

$$p^{r+1}(v) = p^r(v)\frac{c^r(v)}{\int p^r(v')c^r(v')dv'} \tag{A.2}$$

satisfies

1) *$p^r(v) \to p^*(v)$ (the optimal distribution).*
2) *$\int p^r(v)v^2dv \to E_s$, where $E_s = \int p^*(v)v^2dv$ is average signal energy given the optimal distribution $p^*(v)$.*
3) *Let $I(p^r)$ denote the mutual information generated by p^r and let $C(E_s)$ denote the capacity achieved with optimal input with energy constraint E_s. Then $I(p^r) \to I(p^*) = C(E_s)$.*

A simplified expression for $c^r(v)$, defined in (A.1), is derived by [3], as shown in (A.3) at the bottom of the page, where

$$\rho = \frac{1}{2\sigma^2},\ f(\lambda) = \frac{e^\lambda}{(T-1)!}$$
$$g(\lambda, u) = \frac{1}{1+\rho u}(T-1)\frac{e^{\lambda\rho u/(1+\rho u)}}{(\frac{\lambda\rho u}{1+\rho u})^{T-1}}\gamma\left(T-1, \frac{\lambda\rho u}{1+\rho u}\right)$$
$$\gamma(a,b) = \int_0^b q^{a-1}a^{-q}dq.$$

When implementing *Theorem 1* in practice, we approximate the integral in (A.3) by finite summations. We choose a predetermined sampling constant Δ. Let J and K be the number of discrete samples used to approximate v and λ, respectively. The approximation error of the finite summation can be controlled by the choice of the parameters J, K, and Δ. Define the discrete set for v as $\mathcal{V} = \{j\Delta : j = 0, \ldots J-1\} = \{v_0, \ldots, v_{J-1}\}$, and the discrete set for λ as $\mathcal{A} = \{k\Delta : k = 0, \ldots K-1\} = \{\lambda_0, \ldots, \lambda_{K-1}\}$.

The iterative steps in *Theorem 1* can be approximated as shown in the last equation at the bottom of the page.

When the algorithm converges, we have

$$\lim_{r\to\infty} p^r(v_j) \approx p^*(v_j), \quad \lim_{r\to\infty} I(p^r) \approx I(p^*),$$
$$\lim_{r\to\infty} \sum_j p^r(v_j)v_j^2 \approx E_s$$

where $E_s = \sum_j p^*(v_j)v_j^2$. Note that, to satisfy the power constraint, the parameter s should be chosen such that $E_s = T$.

Appendix B
Capacity Computation With Constrained Input From Finite Constellations

We describe how to evaluate (5) with Monte–Carlo integration. To simplify, we assume that a uniformly distributed PSK

$$c^r(v) = \exp\left\{\int\left[\int p(\mathbf{y}|\mathbf{s}=\mathbf{u}v)\log\left(\frac{p(\mathbf{u})p(\mathbf{y}|\mathbf{s}=\mathbf{u}v)}{\int p^r(v')\int p(\mathbf{u}')p(\mathbf{y}|\mathbf{s}=\mathbf{u}'v')\,d\mathbf{u}'\,dv'}\right)d\mathbf{y}\right]d\mathbf{u} - sv^2\right\}. \tag{A.1}$$

$$c^r(v) = \exp\left[-\int(f(\lambda)e^\lambda)(e^{-\lambda}g(\lambda, v^2))\log\left(\int p^r(v')\,e^{-\lambda}g(\lambda, v'^2)\,dv'\right)d\lambda + \log(1+\rho v^2) - sv^2\right] \tag{A.3}$$

$$c^r(v_j) \approx \exp\left[-\sum_{k=0}^{K}[\Delta e^{\lambda_k}f(\lambda_k)][e^{-\lambda_k}g(\lambda_k, v_j^2)]\log\left(\sum_{l=0}^{J}p^r(v_l)\,[e^{-\lambda_k}g(\lambda_k, v_l^2)]\right) + \log(1+\rho v_j^2) - sv_j^2\right]$$
$$p^{r+1}(v_j) = p^r(v_j)\frac{c^r(v_j)}{\sum_j p^r(v_j)c^r(v_j)}.$$

constellation is used. Similar techniques can be used for other input constellations, such as QAM.

Recall (5)

$$H(\mathbf{y}) = -\int p(\mathbf{y})\log[p(\mathbf{y})]d\mathbf{y} = -\frac{1}{M^T}\sum_{\mathbf{s}}\int p(\mathbf{y}|\mathbf{s})\log[p(\mathbf{y})]d\mathbf{y}. \quad \text{(B.1)}$$

Define $\mathbf{s}_0 = (1,\ldots,1)$ to be the all-one column vector. For any input signal vector $\mathbf{s}$, we can write $\mathbf{s} = \Phi\mathbf{s}_0$, where Φ is a $T \times T$ diagonal unitary matrix representing a rotation in the signal space. Note that $p(\mathbf{y}|\mathbf{s}) = p(\Phi\mathbf{y}|\Phi\mathbf{s}) = p(\Phi\mathbf{y}|\mathbf{s}_0)$, and $p(\mathbf{y}) = p(\Phi\mathbf{y})$. The latter equality follows from the assumption that the input distribution is uniform. We have

$$\begin{aligned}\int p(\mathbf{y}|\mathbf{s})\log[p(\mathbf{y})]d\mathbf{y} &= \int p(\Phi\mathbf{y}|\mathbf{s}_0)\log[p(\Phi\mathbf{y})]d\mathbf{y}\\ &= \int p(\mathbf{y}|\mathbf{s}_0)\log[p(\mathbf{y})]d\mathbf{y}\ \text{(change of coordinates)}.\end{aligned} \quad \text{(B.2)}$$

Hence

$$\begin{aligned}H(\mathbf{y}) &= -\int p(\mathbf{y}|\mathbf{s}_0)\log[p(\mathbf{y})]\,d\mathbf{y}\\ &= -\int\Big[\int p(h)p(\mathbf{y}|\mathbf{s}_0,h)dh\Big]\log[p(\mathbf{y})]d\mathbf{y}\\ &= -\int p(h)\int p(\mathbf{y}|\mathbf{s}_0,h)\log[p(\mathbf{y})]d\mathbf{y}\,dh\\ &= -\int p(h)\int p(\mathbf{n})\log[p(\mathbf{y}(\mathbf{n},h)]d\mathbf{n}\,dh\\ &= -\int p(h)\int p(\mathbf{n})f(\mathbf{n},h)d\mathbf{n}\,dh\end{aligned} \quad \text{(B.3)}$$

where $\mathbf{y}(\mathbf{n},h) = h\mathbf{s}_0 + \mathbf{n}$ and $f(\mathbf{n},h) = \log[p(\mathbf{y}(\mathbf{n},h)]$.

Assume that we have an efficient method to compute $f(\mathbf{n},h)$. Next, we generate a total of K points, (h_k, n_k), according to the product distribution $p(h)\cdot p(\mathbf{n})$. We then approximate the last integral in (B.3) by simulating the integrand on these K sample points

$$I = \int p(h)\int p(\mathbf{n})f(\mathbf{n},h)d\mathbf{n}\,dh = \lim_{K\to\infty}\frac{1}{K}\sum_{k=1}^{K} f(n_k,h_k). \quad \text{(B.4)}$$

It remains to give an efficient way to compute $p(\mathbf{y}(\mathbf{n},h))$, or in short notation $p(\mathbf{y})$. In order to reduce complexity, we rewrite $p(\mathbf{y})$ as follows:

$$\begin{aligned}p[\mathbf{y}] &= \frac{1}{M^T}\sum_{\mathbf{s}} p(\mathbf{y}|\mathbf{s})\\ &= \frac{1}{M^T}\sum_{\mathbf{s}}\int p(\mathbf{y}|\mathbf{s},h)p(h)dh\\ &= \int\Big\{\prod_{i=0}^{T-1}\Big[\frac{1}{M}\sum_{\mathbf{s}_i} p(y_i|\mathbf{s}_i,h)\Big]\Big\}p(h)dh\\ &= \int g(\mathbf{y},h)p(h)dh\end{aligned} \quad \text{(B.5)}$$

where $g(\mathbf{y},h) = \prod_{i=0}^{T-1}\big[(1/M)\sum_{\mathbf{s}_i} p(y_i|\mathbf{s}_i,h)\big]$. The last integral in (B.5) can also be approximated by Monte–Carlo integration in the same fashion as shown in (B.4).

ACKNOWLEDGMENT

The authors would like to thank Dr. B. M. Hochwald and Dr. T. L. Marzetta for providing us with their Matlab programs to compute the noncoherent channel capacity by Blahut–Arimoto algorithm. They also thank the anonymous reviewers for their very detailed and insightful comments.

REFERENCES

[1] D. Divsalar and M. K. Simon, "Multiple-symbol differential detection of MPSK," *IEEE Trans. Commun.*, vol. 38, pp. 300–308, Mar. 1990.
[2] D. Warrier and U. Madhow, "Spectrally efficient noncoherent communication," *IEEE Trans. Inform. Theory*, vol. 48, pp. 651–668, Mar. 2002.
[3] T. L. Marzetta and B. M. Hochwald, "Capacity of a mobile multiple-antenna communication link in Rayleigh flat fading," *IEEE Trans. Inform. Theory*, vol. 45, pp. 139–157, Jan. 1999.
[4] G. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding: Turbo codes," in *Proc. Int. Conf. Commun.*, May 1993, pp. 1064–1070.
[5] M. Peleg, S. Shamai, and S. Galán, "Iterative decoding for coded noncoherent MPSK communications over phase-noisy AWGN channel," *Inst. Elect. Eng. Proc. Commun.*, vol. 147, no. 2, pp. 87–95, Apr. 2000.
[6] M. Peleg and S. Shamai, "Iterative decoding of coded and interleaved noncoherent multiple symbol detected DPSK," *Electron. Lett.*, vol. 33, no. 12, pp. 1018–1020, June 1997.
[7] ——, "On coded and interleaved noncoherent multiple symbol detected MPSK," *Eur. Trans. Telecommun.*, vol. 10, no. 1, pp. 65–73, Feb. 1999.
[8] G. Colavolpe, G. Ferrari, and R. Raheli, "Noncoherent iterative (turbo) decoding," *IEEE Trans. Commun.*, vol. 48, pp. 1488–1498, Sept. 2000.
[9] I. D. Marsland and P. T. Mathiopoulos, "On the performance of iterative noncoherent detection of coded *M*-PSK signals," *IEEE Trans. Commun.*, vol. 48, pp. 588–596, Apr. 2000.
[10] ——, "Multiple differential detection of parallel concatenated convolutional (turbo) codes in correlated fast Rayleigh fading," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 265–274, Feb. 1998.
[11] P. Hoeher and J. Lodge, "Turbo DPSK: Iterative differential PSK demodulation and channel decoding," *IEEE Trans. Commun.*, vol. 47, pp. 837–842, June 1999.
[12] M. C. Valenti and B. D. Woerner, "Iterative channel estimation and decoding of pilot-assisted turbo codes over flat-fading channels," *IEEE J. Select. Areas Commun.*, vol. 19, pp. 1697–1705, Sept. 2001.
[13] L. H.-J. Lampe and R. Schober, "Low-complexity iterative demodulation for noncoherent coded transmission over Ricean fading channels," *IEEE Trans. Veh. Technol.*, vol. 50, pp. 1481–1496, Nov. 2001.
[14] H. Su and E. Geraniotis, "Low-complexity joint channel estimation and decoding for pilot-symbol-assisted modulation and multiple differential detection systems with correlated Rayleigh fading," *IEEE Trans. Commun.*, vol. 50, pp. 249–261, Feb. 2002.
[15] C. Komninakis and R. D. Wesel, "Joint iterative channel estimation and decoding in flat correlated Rayleigh fading," *IEEE J. Select. Areas Commun.*, vol. 19, pp. 1706–1717, Sept. 2001.
[16] S. Benedetto, D. Divsalar, G. Montorsi, and F. Pollara, "A Soft-input Soft-output Maximum *a Posteriori* (MAP) Module to Decode Parallel and Serial Concatenated Codes," Jet Propulsion Lab., Pasadena, CA, JPL TDA Progress Rep., 1996.
[17] L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 284–287, Mar. 1974.
[18] G. Colavolpe and R. Raheli, "Noncoherent sequence detection," *IEEE Trans. Commun.*, vol. 47, pp. 1376–1385, Sept. 1999.
[19] ——, "Theoretical analysis and performance limits of noncoherent sequence detection of coded PSK," *IEEE Trans. Inform. Theory*, vol. 46, pp. 1483–1494, July 2000.
[20] R. R. Chen, D. Agrawal, and U. Madhow, "Noncoherent detection of factor-graph codes over fading channels," in *Proc. 2000 Conf. Information Sciences and Systems (CISS 2000)*, Mar. 2000.
[21] O. Macchi and L. L. Scharf, "A dynamic programming algorithm for phase estimation and data decoding on random phase channels," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 581–595, Sept. 1981.

[22] D. Warrier and U. Madhow, "Noncoherent communication in space and time," in *Proc. 1999 Conf. Informations Sciences and Systems (CISS 1999)*, Mar. 1999.
[23] M. Peleg, I. Sason, S. Shamai, and A. Elia, "On interleaved differentially encoded convolutional codes," *IEEE Trans. Inform. Theory*, vol. 45, pp. 2572–2582, Nov. 1999.
[24] K. R. Narayanan and G. L. Stüber, "A serial concatenation approach to iterative demodulation and decoding," *IEEE Trans. Commun.*, vol. 47, pp. 956–961, July 1999.
[25] E. K. Hall and S. G. Wilson, "Design and analysis of turbo codes on Rayleigh fading channels," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 160–174, Feb. 1998.
[26] R. E. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 460–473, July 1972.
[27] J. G. Proakis, *Digital Communications*. New York: McGraw-Hill, 1995.
[28] I. C. Abou-Faycal, M. D. Trott, and S. Shamai, "The capacity of discrete-time memoryless Rayleigh-fading channels," *IEEE Trans. Inform. Theory*, vol. 47, pp. 1290–1301, May 2001.
[29] F. R. Kschischang, B. J. Frey, and H. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inform. Theory*, vol. 47, pp. 498–519, Feb. 2001.
[30] F. Sun and H. Leib, "Multiple-phase codes for detection without carrier phase reference," *IEEE Trans. Inform. Theory*, vol. 44, pp. 1477–1491, July 1998.
[31] F. R. Kschischang and V. Sorokine, "On the trellis structure of block codes," *IEEE Trans. Inform. Theory*, vol. 41, pp. 1924–1937, Nov. 1995.
[32] *Handbook of Coding Theory*, V. S. Pless, W. C. Huffman, and R. A. Brualdi, Eds., Elsevier, Amsterdam, The Netherlands, 1998.
[33] D. Divsalar, H. Jin, and R. J. McEliece, "Coding theorems for turbo-like codes," in *Proc. 36th Annu. Allerton Conf. Communication, Control, Computing*, 1998, pp. 201–210.
[34] S. T. Brink, "Convergence of iterative decoding," *Electron. Lett.*, vol. 35, no. 10, pp. 806–808, May 1999.

Rong-Rong Chen received the Bachelor's degree in applied mathematics from Tsinghua University, Beijing, China, in 1994, and the Master's degree in mathematics and the Ph.D. degree in electrical engineering from the University of Illinois at Urbana-Champaign (UIUC) in 1996 and 2003, respectively.

Since August 2003, she has been an Assistant Professor in the Department of Electrical and Computer Engineering at the University of Utah, Salt Lake City. Her current research interests include wireless communication, coding and information theory, and queueing networks.

Dr. Chen was the recipient of the University Fellowship for graduate study in the Mathematics Department at UIUC and the M. E. Van Valkenburg Graduate Research Award for excellence in doctoral research in the Electrical and Computer Engineering Department at UIUC.

Ralf Koetter (S'92–M'95) was born on October 10, 1963, in Königstein, Germany. He received the Diploma in electrical engineering from the Technical University Darmstadt, Germany in 1990 and the Ph.D. degree from the Department of Electrical Engineering at Linköping University, Sweden.

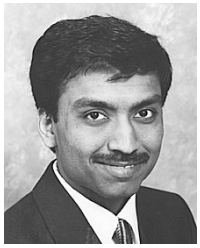
From 1996 to 1997, he was a Visiting Scientist at the IBM Almaden Research Laboratory, San Jose, CA. He was a Visiting Assistant Professor at the University of Illinois at Urbana-Champaign and Visiting Scientist at CNRS in Sophia Antipolis, France, during the years 1997-1998. He joined the faculty of the University of Illinois at Urbana-Champaign in 1999, and is currently an Assistant Professor at the Coordinated Science Laboratory at the University. His research interests include coding and information theory and their application to communication systems.

Dr. Koetter served as Associate Editor for Coding Theory and Techniques from 1999 to 2001 for the IEEE TRANSACTIONS ON COMMUNICATIONS. In 2000, he started a term as Associate Editor for Coding Theory of the IEEE TRANSACTIONS ON INFORMATION THEORY. He received an IBM Invention Achievement Award in 1997, an NSF CAREER Award in 2000, and an IBM Partnership Award in 2001.

Upamanyu Madhow (S'86–M'90–SM'96) received the bachelor's degree in electrical engineering from the Indian Institute of Technology, Kanpur, India, in 1985. He received the M.S. and Ph.D. degrees in electrical engineering from the University of Illinois at Urbana-Champaign in 1987 and 1990, respectively.

From 1990 to 1991, he was a Visiting Assistant Professor at the University of Illinois. From 1991 to 1994, he was a Research Scientist at Bell Communications Research, Morristown, NJ. From 1994 to 1999, he was with the Department of Electrical and Computer Engineering, University of Illinois, Urbana-Champaign, first as an Assistant Professor and, since 1998, as an Associate Professor. Since December 1999, he has been with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, where he is currently a Professor. His research interests are in communication systems and networking, with current emphasis on wireless communication.

Dr. Madhow is a recipient of the NSF CAREER award. He has served as Associate Editor for Spread Spectrum for the IEEE TRANSACTIONS ON COMMUNICATIONS, and as Associate Editor for Detection and Estimation for the IEEE TRANSACTIONS ON INFORMATION THEORY.



Dakshi Agrawal received the B.Tech. degree in 1993 from the Indian Institute of Technology-Kanpur, India, the M.S. degree in 1995 from Washington University, St. Louis, MO, and the Ph.D. degree in 1999 from the University of Illinois at Urbana-Champaign (UIUC), all in electrical engineering.

He worked as a Visiting Assistant Professor at UIUC in 1999–2000. Currently, he is a Research Staff Member at the T. J. Watson Research Center, IBM Corporation, Hawthorne, NY. His current research interest is to model and analyze distributed computing systems.

Dr. Agrawal received the Robert T. Chien Memorial Award for 1999 for excellence in doctoral research from the Department of Electrical and Computer Engineering at UIUC. He also received the Ross J. Martin Memorial Award for 2000 for outstanding research achievement by a graduate student in the College of Engineering at UIUC.