Adaptive MLSDE Using the EM Algorithm

Hossein Zamiri-Jafarian, *Member, IEEE,* and Subbarayan Pasupathy, *Fellow, IEEE*

***Abstract*—The theory of adaptive sequence detection incorporating estimation of channel and related parameters is studied in the context of maximum-likelihood (ML) principles in a general framework based on the expectation and maximization (EM) algorithm. A generalized ML sequence detection and estimation (GMLSDE) criterion is derived based on the EM approach, and it is shown how the per-survivor processing and per-branch processing methods emerge naturally from GMLSDE.**

GMLSDE is developed into a real time detection/estimation algorithm using the online EM algorithm with coupling between estimation and detection. By utilizing Titterington's stochastic approximation approach, different adaptive ML sequence detection and estimation (MLSDE) algorithms are formulated in a unified manner for different channel models and for different amounts of channel knowledge available at the receiver. Computer simulation results are presented for differentially encoded quadrature phase-shift keying in frequency flat and selective fading channels, and comparisons are made among the performances of the various adaptive MLSDE algorithms derived earlier.

***Index Terms*— Adaptive detection and estimation, estimation theory, expectation and maximization algorithm, fading channels, maximum-likelihood detection, maximum-likelihood estimation, sequence detection theory, statistical communication theory.**

I. Introduction

THE detection of a signal transmitted through a communication channel having memory and additive Gaussian noise has been widely studied for different channel models. Equalization techniques have been used in communication systems to combat the intersymbol interference (ISI) induced by dispersive channels. When the transmitted data sequences are equiprobable, maximum-likelihood sequence detection (MLSD) minimizes sequence-error probability and can, hence, be considered as an optimal equalization method. MLSD, implemented using the Viterbi algorithm for known finite channel-impulse response (CIR), is well known [1]. The MLSD algorithm has also been studied for a mobile communication channel that disperses the transmitted signal in both time and frequency domains and whose impulse response is considered as a Gaussian random process [2]–[7].

Due to unknown CIR or unknown statistical parameters of the CIR, joint data detection and channel estimation methods were proposed by combining Viterbi algorithm for data detection with adaptive methods, such as least mean square, recursive least squares (RLS), and Kalman filtering for estimating the CIR [3], [8]–[10]. However, the inherent decision delay in such procedures causes poor channel tracking in a time-variant environment. The idea of per-survivor processing (PSP) was proposed to combat the decision delay problem, where each survivor path of the trellis diagram in the MLSD structure has its own CIR estimation [11]–[13]. Although PSP is a practical way to achieve better performance in a time-variant channel, the nature and degree of optimality of such PSP-based channel estimation procedures, the influence of such estimates on the optimality of the MLSD criterion, and the coupling between estimation, detection, and channel models are not clear.

In this paper, the maximum-likelihood sequence detection and estimation (MLSDE) algorithm is considered in a general framework based on maximum-likelihood (ML) detection/estimation theory. To detect the transmitted signal, the receiver usually needs to know some other parameters. We show that the MLSDE criterion for detecting the data sequence and estimating the unknown parameters can be achieved by the expectation and maximization (EM) algorithm, which is an iterative method. The EM algorithm increases the likelihood of the detected/estimated parameters in each iteration with expectation and maximization steps until it achieves the global or a local maximum [14], [15].

Generalized MLSDE (GMLSDE) is presented as an EM-based algorithm, which alternates between detection and estimation and still satisfies the MLSDE criterion. GMLSDE is implemented based on the online EM algorithm for real time detection/estimation where in each recursion, the algorithm increases the likelihood function. It is shown that the concept of PSP and per-branch processing (PBP), which estimates a separate CIR for each branch in the trellis diagram, emerge naturally from the EM aspect of the GMLSDE algorithm as an integral part of a likelihood-increasing procedure when the previously detected/estimated parameters are used as given conditions for the next expectation step. Some adaptive MLSDE receivers are derived based on the GMLSDE framework in a unified way for different levels of knowledge that are available at the receiver. The recursive estimation proposed by Titterington [16] is employed for the estimation of time-variant/invariant unknown deterministic parameters. Each adaptive receiver uses those steps of detection and estimation of GMLSDE generated by the selected channel model. Although two new adaptive MLSDE receivers along with some previously known ones are derived as examples, the power of GMLSDE is not limited to these particular algorithms, and one can use the unified framework of GMLSDE to develop

Paper approved by Z. Kostic, the Editor for Wireless Communications of the IEEE Communications Society. Manuscript received May 7, 1998; revised August 11, 1998 and February 18, 1999. This work was supported by the Information Technology Research Center (ITRC), a center of excellence funded by the Province of Ontario, Canada. This work was presented in part at the IEEE Communications Theory Workshop, Tucson, AZ, April 1997, and at the Canadian Workshop on Information Theory, Toronto, Canada, June 1997.

The authors are with the Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S 3G4 Canada (e-mail: zamiri@comm.utoronto.ca).

Publisher Item Identifier S 0090-6778(99)06296-0.

new algorithms based on different channel models and levels of knowledge available at the receiver.

Even though the concept behind the EM algorithm was known in early statistical literature [17], [18], it was the later seminal paper by Dempster *et al.* [14] that spurred much research on many applications of the EM algorithm including communication-related ones [15]. Kaleh *et al.* [19] derived the iterative method for joint channel parameter estimation and symbol detection using an EM-based forward–backward algorithm. Georghiades and Han [20] used the EM algorithm to study sequence estimation in random-phase and fading channels. We may point to [21]–[23] for some other applications of the EM algorithm in communication systems.

This paper is organized as follows: GMLSDE is developed via the EM algorithm and the relation to PSP and PBP is explained in Section II. Different adaptive MLSDE algorithms associated with different levels of channel knowledge available at the receiver are derived in Section III. Section IV contains computer simulations, results, and comparisons for a differentially encoded quadrature phase-shift keying (DQPSK) modulation scheme in flat and selective fading channels with different fading rates. The results of simulations with time-variant fading rate for a flat fading channel with different levels of channel knowledge are also presented in Section IV. Section V presents some conclusions.

II. Generalized MLSDE Algorithm

The main goal in digital communication systems is to detect the sequence of the transmitted symbols $\mathbf{a} = \{a_0, \cdots, a_{I-1}\}$ by observing the received signal $\mathbf{y} = [y(K-1), \cdots, y(0)]^{\mathcal{T}}$, where $X^{\mathcal{T}}$ denotes the transpose of X. The optimal receiver maximizes the joint probability density function (pdf) of $\mathbf{a}$ and $\mathbf{y}$ for detecting $\mathbf{a}$. When all the sequences $\mathbf{a}$ are equiprobable, the optimal sequence receiver accomplishes MLSD with the criterion given by

$$C_{\text{MLSD}} \doteq \max_{\mathbf{a}}\{p(\mathbf{y}|\mathbf{a})\} \equiv \max_{\mathbf{a}}\{\log\, p(\mathbf{y}|\mathbf{a})\} \tag{1}$$

where $\mathbf{a}$ is selected from $\mathcal{A}$, a set of all its possibilities $\mathbf{a} \in \mathcal{A}$. The detected sequence $\hat{\mathbf{a}}$ is

$$\hat{\mathbf{a}} = \arg\,\max_{\mathbf{a}}\{\log\, p(\mathbf{y}|\mathbf{a})\}. \tag{2}$$

The received signal is a function of the transmitted symbols $\mathbf{a}$, media parameters such as channel parameters $\hbar_c$, and additive noise $\mathbf{z} = [z(K-1), \cdots, z(0)]^{\mathcal{T}}$

$$\mathbf{y} = f_{\text{ch}}(\mathbf{a}, \hbar_c) + \mathbf{z}. \tag{3}$$

When the channel is modeled as a linear system, $f_{\text{ch}}(\cdot)$ is a linear function and its arguments are $\mathbf{a}$ and CIR. To detect the sequence $\mathbf{a}$ by observing $\mathbf{y}$, the receiver should know the function $f_{\text{ch}}(\cdot)$, the pdf of $\mathbf{z}$, and the CIR (if it is deterministic) or the pdf of CIR (if it is a random variable/process).

The structure of the MLSD receiver will be different for different channel models. The linear function $f_{\text{ch}}(\cdot)$ is usually modeled as having finite memory (e.g., a finite impulse response system model for channels with ISI or multipath fading). Also, it is very common to model the additive noise $\mathbf{z}$ as a stationary white complex zero-mean circularly symmetric Gaussian random process with autocorrelation function $R_z(k) = N_0\delta(k)$. The CIR or its parameters (usually unknown to the receiver) should be estimated during the detection procedure.

The MLSD criterion (1) is suitable when the symbols $\mathbf{a}$ are the only parameters that are unknown and the received signal $\mathbf{y}$ provides complete information necessary for such a detection procedure. However, in general, other parameters that can be modeled as a set of unknown deterministic parameters, random variables/processes, or both are also needed to complete the detection procedure. In detection theory, such problems are termed as composite hypothesis testing or detection with unwanted parameters [24].

If we define $\boldsymbol{\theta}$ as the needed unknown deterministic parameters that should be estimated, the criterion of MLSDE becomes

$$C_{\text{MLSDE}} \doteq \max_{\mathbf{a},\boldsymbol{\theta}}\{\log\, p(\mathbf{y}|\mathbf{a}, \boldsymbol{\theta})\}. \tag{4}$$

Since $\mathbf{y}$ sometimes does not provide the complete information necessary to obtain the ML estimates of the parameters $\mathcal{U} = \{\mathbf{a}, \boldsymbol{\theta}\}$ from (4) directly, we present the solution to (4) using the EM algorithm [14]. By considering $\mathcal{I} = \mathbf{y}$ as incomplete data and $\mathcal{D}$ as the random variables or processes needed to complete $\mathcal{I}$ for detecting/estimating $\mathcal{U}$, the log-likelihood function is given by

$$\boldsymbol{L}(\mathcal{U}) = \log\, p(\mathcal{I}|\mathcal{U}) = \log\, p(\mathcal{C}|\mathcal{U}) - \log\, p(\mathcal{D}|\mathcal{U}, \mathcal{I}) \tag{5}$$

where $\mathcal{C} = \{\mathcal{I}, \mathcal{D}\}$ is the complete data. On taking the conditional expectation of both sides of (5) with respect to $\mathcal{D}$ given $\mathcal{I}$ and a parameter set $\hat{\mathcal{U}}^{(l)}$ (the estimation of $\mathcal{U}$ at lth iteration or initial estimate of $\mathcal{U}$), we have

$$\boldsymbol{L}(\mathcal{U}) = E\Big[\log\, p(\mathcal{C}|\mathcal{U})|\hat{\mathcal{U}}^{(l)}, \mathcal{I}\Big] - E\Big[\log\, p(\mathcal{D}|\mathcal{U}, \mathcal{I})|\hat{\mathcal{U}}^{(l)}, \mathcal{I}\Big]. \tag{6}$$

Following [14], one can show that $\boldsymbol{L}(\hat{\mathcal{U}}^{(l+1)}) \geq \boldsymbol{L}(\hat{\mathcal{U}}^{(l)})$ where

$$Q\Big(\mathcal{U}|\hat{\mathcal{U}}^{(l)}\Big) = E\Big[\log\, p(\mathcal{C}|\mathcal{U})|\hat{\mathcal{U}}^{(l)}, \mathcal{I}\Big] \tag{7}$$

and

$$\hat{\mathcal{U}}^{(l+1)} = \arg\,\max_{\mathcal{U}}\Big\{Q\Big(\mathcal{U}|\hat{\mathcal{U}}^{(l)}\Big)\Big\}. \tag{8}$$

The above iteration, (7) and (8), is repeated until the stationary estimation of $\mathcal{U}$ is achieved. In other words, the previous estimation of $\mathcal{U}$ is used as the given condition for estimating $\mathcal{U}$ based on (7) and (8) until at $\hat{l}$th iteration, $\hat{\mathcal{U}}^{(\hat{l})} = \hat{\mathcal{U}}^{(\hat{l}-1)} \doteq \hat{\mathcal{U}}$. Hence, if $\boldsymbol{L}(\mathcal{U})$ has only one maximum point, we have

$$\hat{\mathcal{U}} = \arg\,\max_{\mathcal{U}}\{\boldsymbol{L}(\mathcal{U})\}. \tag{9}$$

Otherwise, $\hat{\mathcal{U}}$ is a local maximum point. Thus, the MLSDE criterion can be achieved iteratively by following the EM algorithm.

The EM algorithm has two steps: 1) expectation (7) and 2) maximization (8). The first step (7) is to take the expectation of the log-likelihood function of the complete data given the current detected/estimated parameters $\hat{\mathcal{U}}^{(l)}$ and the incomplete (observed) data $\mathcal{I}$. The second step (8) provides a new estimate of the parameters $\hat{\mathcal{U}}^{(l+1)}$ by maximizing the expectation of the log-likelihood function (computed in the first step) over the unknown parameters $\mathcal{U}$. The EM algorithm repeats the expectation and maximization steps iteratively in order to increase the likelihood of the parameters [14], [15]. Therefore, instead of calculating the ML detection/estimation of $\mathcal{U} = \{\mathbf{a}, \boldsymbol{\theta}\}$ directly from (4) in a closed form or in one iteration that is very complex and impractical especially for a high-dimensional problem, the EM algorithm uses the iterative method that increases the likelihood of detected/estimated parameters in each iteration until it achieves the MLSDE criterion (4).

When the unknown parameter set $\mathcal{U}$ contains different types of unknown variables/vectors (i.e., in our problem $\mathbf{a}$ and $\boldsymbol{\theta}$ are discrete and continuous, respectively), the maximization step of the EM algorithm is a challenging job and often intractable. In order for the maximization step to become tractable, we partition the set of unknown parameters into disjoint sets and find the maximum of each partitioned set separately. Therefore, in this method, each iteration contains more than one expectation and maximization step. In the following theorem, we show how this GMLSDE procedure increases the likelihood at each iteration using the framework of the EM algorithm.

Theorem 1: Let the unknown parameter set $\mathcal{U}$ be divided into $\mathbf{i}$ separate sets $\mathcal{U} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_{\mathbf{i}}\}$. $\boldsymbol{L}(\hat{\mathcal{U}}^{(l+1)}) \geq \boldsymbol{L}(\hat{\mathcal{U}}^{(l)})$ where the estimation of $\mathcal{U}$ at $(l+1)$th iteration $\hat{\mathcal{U}}^{(l+1)} = \{\hat{\boldsymbol{\theta}}_1^{(l+1)}, \hat{\boldsymbol{\theta}}_2^{(l+1)}, \cdots, \hat{\boldsymbol{\theta}}_{\mathbf{i}}^{(l+1)}\}$ is estimated from the following steps using the EM algorithm:

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_1^{(l+1)} &= \max_{\boldsymbol{\theta}_1}\left\{Q\left(\boldsymbol{\theta}_1, \hat{\boldsymbol{\theta}}_2^{(l)}, \cdots, \hat{\boldsymbol{\theta}}_{\mathbf{i}}^{(l)} | \hat{\mathcal{U}}^{(l)}\right)\right\}\\
\hat{\boldsymbol{\theta}}_2^{(l+1)} &= \max_{\boldsymbol{\theta}_2}\left\{Q\left(\hat{\boldsymbol{\theta}}_1^{(l+1)}, \boldsymbol{\theta}_2, \hat{\boldsymbol{\theta}}_3^{(l)}, \cdots, \hat{\boldsymbol{\theta}}_{\mathbf{i}}^{(l)} | \hat{\mathcal{U}}^{(l)}\right)\right\}\\
&\vdots\\
\hat{\boldsymbol{\theta}}_{\mathbf{i}}^{(l+1)} &= \max_{\boldsymbol{\theta}_{\mathbf{i}}}\left\{Q\left(\hat{\boldsymbol{\theta}}_1^{(l+1)}, \hat{\boldsymbol{\theta}}_2^{(l+1)}, \cdots, \hat{\boldsymbol{\theta}}_{\mathbf{i}-1}^{(l+1)}, \boldsymbol{\theta}_{\mathbf{i}} | \hat{\mathcal{U}}^{(l)}\right)\right\}.
\end{aligned} \tag{10}$$

Proof: See Appendix A. Iteration between some maximization steps is similar to the Gauss–Seidel method [25] and is also called the cyclic coordinate ascent method [26]. Meng and Rubin [27] proposed a similar model reduction to achieve simple conditional maximization and called it expectation-conditional maximization.

By following Theorem 1 and noting that $\mathbf{a}$ is a vector of discrete parameters, GMLSDE (10) satisfies the ML criterion[1] by dividing $\mathcal{U}$ into $\mathbf{a}$ and $\boldsymbol{\theta}$ and estimation of $\boldsymbol{\theta}$ (continuous unknown parameter set) and detection of $\mathbf{a}$ (discrete unknown parameter set) are alternated. The estimation and detection steps at the $(l+1)$th iteration for the GMLSDE are:[2]

1) Estimation part
   a) 1-E step

$$\begin{aligned}
Q_1\left(\boldsymbol{\theta}|\hat{\mathbf{a}}^{(l)}, \hat{\boldsymbol{\theta}}^{(l)}\right) &= Q\left(\hat{\mathbf{a}}^{(l)}, \boldsymbol{\theta}|\hat{\mathbf{a}}^{(l)}, \hat{\boldsymbol{\theta}}^{(l)}\right)\\
&= E\left[\log p\left(\mathcal{C}|\hat{\mathbf{a}}^{(l)}, \boldsymbol{\theta}\right)|\hat{\mathbf{a}}^{(l)}, \hat{\boldsymbol{\theta}}^{(l)}, \mathcal{I}\right]
\end{aligned} \tag{11}$$

   b) 2-M step

$$\hat{\boldsymbol{\theta}}^{(l+1)} = \arg\max_{\boldsymbol{\theta}}\left\{Q_1\left(\boldsymbol{\theta}|\hat{\mathbf{a}}^{(l)}, \hat{\boldsymbol{\theta}}^{(l)}\right)\right\} \tag{12}$$

2) Detection part
   a) 3-E step

$$\begin{aligned}
Q_2\left(\mathbf{a}|\hat{\mathbf{a}}^{(l)}, \hat{\boldsymbol{\theta}}^{(l)}\right) &= Q\left(\mathbf{a}, \hat{\boldsymbol{\theta}}^{(l+1)}|\hat{\mathbf{a}}^{(l)}, \hat{\boldsymbol{\theta}}^{(l)}\right)\\
&= E\left[\log p\left(\mathcal{C}|\mathbf{a}, \hat{\boldsymbol{\theta}}^{(l+1)}\right)|\hat{\mathbf{a}}^{(l)}, \hat{\boldsymbol{\theta}}^{(l)}, \mathcal{I}\right]
\end{aligned} \tag{13}$$

   b) 4-M step[3]

$$\hat{\mathbf{a}}^{(l+1)} = \arg\max_{\mathbf{a}}\left\{Q_2\left(\mathbf{a}|\hat{\mathbf{a}}^{(l)}, \hat{\boldsymbol{\theta}}^{(l)}\right)\right\}. \tag{14}$$

Dividing the parameters as discrete and continuous decreases the dimension of the parameter space and usually increases the convergence of the algorithm [29]. Also, it allows us to use different methods in the maximizing step for discrete and continuous parameters. GMLSDE shows a natural coupling between estimation and detection, which should allow us to get some insight about the influence of either one on the other. This has not been possible with earlier approaches to the combined detection/estimation problem, since channel estimation has been traditionally uncoupled from the detection problem when developing the detection algorithm.

The proposed GMLSDE algorithm (11)–(14) has been developed using the entire received signal $\mathbf{y}$, or in other words, the EM algorithm is offline. Since the main purpose of an adaptive algorithm is to detect the sequentially emitted $\mathbf{a}$ in real time, we are interested in an online (recursive) version of the EM algorithm. By defining $\mathcal{C}_k$ and $\mathcal{I}_k$ as the complete data and the incomplete data available up to time k, respectively, the steps of GMLSDE algorithm at time k using the online scheme for estimating and detecting are given as:

1) Estimation part
   a) 1-E step

$$\begin{aligned}
&Q_{1,k}\left(\boldsymbol{\theta}_k|\tilde{\mathbf{a}}_{k|k-1}, \tilde{\boldsymbol{\theta}}_{k|k-1}\right)\\
&\quad= E\left[\log p(\mathcal{C}_k|\tilde{\mathbf{a}}_{k|k-1}, \boldsymbol{\theta}_k)|\tilde{\mathbf{a}}_{k|k-1}, \tilde{\boldsymbol{\theta}}_{k|k-1}, \mathcal{I}_k\right]
\end{aligned} \tag{15}$$

   b) 2-M step

$$\tilde{\boldsymbol{\theta}}_{k|k} = \arg\max_{\boldsymbol{\theta}_k}\left\{Q_{1,k}\left(\boldsymbol{\theta}_k|\tilde{\mathbf{a}}_{k|k-1}, \tilde{\boldsymbol{\theta}}_{k|k-1}\right)\right\} \tag{16}$$

[1] Similar to the EM algorithm, GMLSDE may only achieve a local maximum point if $\boldsymbol{L}(\mathcal{U})$ has many maxima.

[2] In general, dividing the parameter set into two separate sets (see Theorem 1) does not guarantee achieving ML criterion; however, since $\mathbf{a}$ is discrete and the detection part considers all the possibilities of $\mathbf{a}$, the algorithm is guaranteed to achieve ML criterion when there are no local maxima.

[3] The relationship between (11)–(14) and the space-alternating generalized EM (SAGE) algorithm [28] is shown in Appendix B when $\hat{\boldsymbol{\theta}}^{(l+1)}$ is used instead of $\hat{\boldsymbol{\theta}}^{(l)}$ as a given condition for computing the expectation in (13).

2) Detection part

a) 3-E step

$$Q_{2,k}\Big(\mathbf{a}_k|\tilde{\mathbf{a}}_{k|k-1}, \tilde{\boldsymbol{\theta}}_{k|k-1}\Big) = E\Big[\log p\Big(\mathcal{C}_k|\mathbf{a}_k, \tilde{\boldsymbol{\theta}}_{k|k}\Big)|\tilde{\mathbf{a}}_{k|k-1}, \tilde{\boldsymbol{\theta}}_{k|k-1}, \mathcal{I}_k\Big] \quad (17)$$

b) 4-M step

$$\tilde{\mathbf{a}}_{k|k} = \arg\max_{\mathbf{a}_k}\Big\{Q_{2,k}\Big(\mathbf{a}_k|\tilde{\mathbf{a}}_{k|k-1}, \tilde{\boldsymbol{\theta}}_{k|k-1}\Big)\Big\} \quad (18)$$

where $\{\mathbf{a}_k, \boldsymbol{\theta}_k\}$ is the parameter set up to time k. Also, $\tilde{\mathbf{a}}_{k|l}$ and $\tilde{\boldsymbol{\theta}}_{k|l}$ are the detected values of $\mathbf{a}_k$ and the estimates of $\boldsymbol{\theta}_k$ based on the received data up to time l, respectively. Similar to Theorem 1, since$Q_{1,k}(\tilde{\boldsymbol{\theta}}_{k|k}|\tilde{\mathbf{a}}_{k|k-1}, \tilde{\boldsymbol{\theta}}_{k|k-1}) \geq Q_{1,k}(\tilde{\boldsymbol{\theta}}_{k|k-1}|\tilde{\mathbf{a}}_{k|k-1}, \tilde{\boldsymbol{\theta}}_{k|k-1})$ and $Q_{2,k}(\tilde{\mathbf{a}}_{k|k}|\tilde{\mathbf{a}}_{k|k-1}, \tilde{\boldsymbol{\theta}}_{k|k-1}) \geq Q_{2,k}(\tilde{\mathbf{a}}_{k|k-1}|\tilde{\mathbf{a}}_{k|k-1}, \tilde{\boldsymbol{\theta}}_{k|k-1})$, one can show that

$$\boldsymbol{L}_k\Big(\tilde{\mathbf{a}}_{k|k}, \tilde{\boldsymbol{\theta}}_{k|k}\Big) \geq \boldsymbol{L}_k\Big(\tilde{\mathbf{a}}_{k|k-1}, \tilde{\boldsymbol{\theta}}_{k|k}\Big) \geq \boldsymbol{L}_k\Big(\tilde{\mathbf{a}}_{k|k-1}, \tilde{\boldsymbol{\theta}}_{k|k-1}\Big) \quad (19)$$

where $\boldsymbol{L}_k(\cdot) = \log p(\mathbf{y}_k|.)$ is the log-likelihood function based on the received signal up to time k.[4] Therefore, the likelihood function is increased at each recursion[5] (increasing time). Appendix C shows the relation between log-likelihood function at time k and previous log-likelihood functions. Meanwhile, the time-update vector $\tilde{\boldsymbol{\theta}}_{k+1|k}$ is given by

$$\tilde{\boldsymbol{\theta}}_{k+1|k} = \begin{bmatrix} \tilde{\boldsymbol{\varphi}}_{k+1|k} \\ \tilde{\boldsymbol{\theta}}_{k|k} \end{bmatrix}$$

where $\tilde{\boldsymbol{\theta}}_{k|k} = [\tilde{\boldsymbol{\varphi}}^{\mathcal{T}}_{k|k}, \tilde{\boldsymbol{\varphi}}^{\mathcal{T}}_{k-1|k}, \cdots, \tilde{\boldsymbol{\varphi}}^{\mathcal{T}}_{0|k}]^{\mathcal{T}}$, and $\tilde{\boldsymbol{\varphi}}_{k+1|k}$ is obtained generally from the dynamic evolution of the $\boldsymbol{\varphi}_{k+1}$ process

$$\boldsymbol{\varphi}_{k+1} = f_\theta(\boldsymbol{\theta}_k, \mathbf{a}_k). \quad (20)$$

The elements of the transmitted symbols $\mathbf{a} = \{a_k\}_{k=0}^{I-1}$ are selected from a finite set [i.e., $a_k \in \{\pm 1 \pm j\}$ in quadrature phase-shift keying (QPSK)] and generally (assuming no coding), the symbols are independent of each other. In other words [unlike (20)], there is no dynamic relation between the present symbol and the previous symbols. Therefore, the time-update $\tilde{\mathbf{a}}_{k+1|k}$ is defined by

$$\tilde{\mathbf{a}}_{k+1|k} = \begin{bmatrix} a_{k+1} \\ \tilde{\mathbf{a}}_{k|k} \end{bmatrix}$$

where a_{k+1} is selected from all the possibilities of the alphabet set (i.e., $\{\pm 1 \pm j\}$ in QPSK). Meanwhile, in general, the estimation part of GMLSDE may contain more than one step of expectation and maximization. Based on Theorem 1, the unknown parameter set $\boldsymbol{\theta}$ may be divided into different sets and each set will have its own expectation and maximization steps. This idea is very useful in increasing the convergence of the algorithm especially when the complete data is different for each separate set [28]. Meanwhile, from the viewpoint of increasing the likelihood, estimation and detection can be done in a different order in GMLSDE. However, in order to decrease the complexity or due to some other practical issues, it may be that one order (i.e., estimation after detection) is preferable to the other order. Doing the estimation part before the detection part needs to estimate $\boldsymbol{\theta}_k$ at time k for all possibilities of $\tilde{\mathbf{a}}_{k|k-1} = [a_k, \tilde{\mathbf{a}}^{\mathcal{T}}_{k-1|k-1}]^{\mathcal{T}}$, since $\tilde{\mathbf{a}}_{k|k-1}$ is a given condition in the estimation part at time k. In other words, $\boldsymbol{\theta}_k$ is estimated for all the branch metrics in the trellis diagram. This method is called per-branch processing (PBP) [30] which estimates different channel parameters for each branch metric. However, when the detection part is done before the estimation part, since $\tilde{\mathbf{a}}_{k|k}$ is a given condition in the estimation part, channel parameter set $\boldsymbol{\theta}_k$ is estimated for all $\tilde{\mathbf{a}}_{k|k}$ at time k where $\tilde{\mathbf{a}}_{k|k}$ indicates all survivor paths up to time k. In other words, $\boldsymbol{\theta}_k$ is estimated only for survivor paths, and this method is called per-survivor processing (PSP) [11].

As the estimation and detection parts of GMLSDE (15)–(18) show, the idea of using the previous decisions and estimates as tools for detecting and estimating the future arises naturally in implementing MLSDE based on the EM algorithm. The decision-based receiver was used in [31] for an ISI channel with infinite impulse response. Later, it was proposed in many such joint detection and estimation methods and is generally known as PSP [11]. The PSP was originally proposed as a practical way to implement joint detection and estimation; however, in GMLSDE due to the inherent coupling between the estimation and detection parts, the temporary decision of the data $\tilde{\mathbf{a}}_{k|k}$ at time k is a given condition in the estimation part when the detection part is done before the estimation part. Therefore, the PSP (estimating different CIR or other channel parameters for different survivor paths) comes up naturally as an integral part of the EM-based ML detection/estimation procedure. Also, when the estimation part is done before the detection part in the GMLSDE, $\tilde{\mathbf{a}}_{k|k-1}$ is a given condition in the estimation part at time k. Since $\tilde{\mathbf{a}}_{k|k-1}$ is the temporary decision of the data for all branch metrics, the PBP (estimating different channel parameters for all branch metrics) also appears naturally from the GMLSDE procedure. Thus, one can say that the EM algorithm provides a solid theoretical foundation for using PSP and PBP in ML-based receivers due to the inherent embedding of decision feedback in the EM approach. Meanwhile, if all survivor paths at time k have the same root in the trellis diagram at time $k - L_d$, since there is only one detected sequence $\tilde{\mathbf{a}}_{k-L_d}$, there will be only one estimation for $\boldsymbol{\theta}_{k-L_d}$ at time k, $\tilde{\boldsymbol{\theta}}_{k-L_d|k}$. If we assume that $\tilde{\boldsymbol{\theta}}_{k|k} \simeq \tilde{\boldsymbol{\theta}}_{k-L_d|k}$, the detection/estimation procedure leads to the MLSDE receiver that was proposed by Qureshi [8] before the idea of PSP emerged in the research literature.

The convergence of the EM algorithm is inversely related to the dimension of its complete data space. Less necessary and less informative complete data improves the asymptotic convergence rate. The online EM algorithm deals with unknown parameters and complete data only up to the process

[4] We will show in Section III that the maximization step of the online EM algorithm becomes more tractable in some special cases such as Gaussian model.

[5] In general, each recursion (15)–(18) can be implemented with some iterations (11)–(14). In this case, $\{\tilde{\mathbf{a}}_{k|k-1}, \tilde{\boldsymbol{\theta}}_{k|k-1}\}$ is the set of initial values for the first iteration at time k and $\{\tilde{\mathbf{a}}_{k|k}, \tilde{\boldsymbol{\theta}}_{k|k}\}$, which is the detection/estimation of the unknown parameter set at recursion k (or the result of the first iteration at time k), and is the initial value for the second iteration at time k. This procedure is able to increase the likelihood.

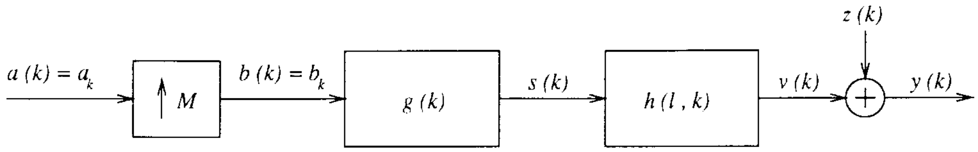


Fig. 1. A discrete model of the communication system for a linear channel, where $\uparrow M$ represents up-sampling by a factor M.

time, and their dimensions increase linearly with time, thus causing faster convergence. This idea is similar to the SAGE method, which achieves faster convergence by partitioning the parameters and the complete data [28]. In the online EM algorithm, the parameters and complete data are naturally partitioned through time. Meanwhile, due to the parameter space-time coupling, the online EM algorithm is a recursive algorithm (each recursion can generally be done by more than one iteration [32]), whereas the offline EM algorithm is only an iterative algorithm in general.

The estimation part in the online EM algorithm can be approximately implemented by a recursive formula based on Titterington's approach [16]. Using a stochastic approximation in order to consider only three elements of a Taylor series expansion of $Q_{1,k}(\boldsymbol{\theta}|\tilde{\mathcal{U}}_{k|k-1})$ (15) at point $\boldsymbol{\theta}_k = \tilde{\boldsymbol{\theta}}_{k|k-1}$, one can show (16) becomes

$$\tilde{\boldsymbol{\theta}}_{k|k} \simeq \tilde{\boldsymbol{\theta}}_{k|k-1} - \left(\left. \frac{\partial^2 Q_{1,k}\left(\boldsymbol{\theta}_k|\tilde{\mathcal{U}}_{k|k-1}\right)}{\partial^2 \boldsymbol{\theta}_k} \right|_{\boldsymbol{\theta}_k = \tilde{\boldsymbol{\theta}}_{k|k-1}} \right)^{-1} \cdot \left(\left. \frac{\partial Q_{1,k}\left(\boldsymbol{\theta}_k|\tilde{\mathcal{U}}_{k|k-1}\right)}{\partial \boldsymbol{\theta}_k^*} \right|_{\boldsymbol{\theta}_k = \tilde{\boldsymbol{\theta}}_{k|k-1}} \right), \quad \text{for all } \tilde{\mathcal{U}}_{k|k-1} \tag{21}$$

where $\tilde{\mathcal{U}}_{k|k-1} = \{\tilde{\mathbf{a}}_{k|k-1}, \tilde{\boldsymbol{\theta}}_{k|k-1}\}$ and $\boldsymbol{\theta}^*$ denotes conjugate of $\boldsymbol{\theta}$. If $\boldsymbol{\varphi}_k$ is an $L{+}1$ vector, $\boldsymbol{\theta}_k$ and $(\partial Q_{1,k}(\boldsymbol{\theta}_k|\tilde{\mathcal{U}}_{k|k-1})/\partial\boldsymbol{\theta}_k^*)$ are $k(L+1)$ vectors, and also $(\partial^2 Q_{1,k}(\boldsymbol{\theta}_k|\tilde{\mathcal{U}}_{k|k-1})/\partial^2\boldsymbol{\theta}_k)$ is a $k(L{+}1)\times k(L{+}1)$ matrix. It should be noted that when the third and higher derivatives of $Q_{1,k}(.|.)$ are zero, as is usually true for the Gaussian case, the recursive formula (21) is exact.

The detection part, due to the finite alphabet of the transmitted symbols, can be implemented with a dynamic programming (Viterbi) algorithm. The expectation step in the detection part (17) is obtained for all possibilities of $\{\tilde{\mathbf{a}}_{k|k-1}, \tilde{\boldsymbol{\theta}}_{k|k}\}$ at time k. When a_k is selected from a q-ary signaling scheme, there are q possible estimates for $\tilde{\mathcal{U}}_{k|k-1}$ and $\tilde{\boldsymbol{\theta}}_{k|k}$ in each estimated $\tilde{\mathcal{U}}_{k-1|k-1} = \{\tilde{\mathbf{a}}_{k-1|k-1}, \tilde{\boldsymbol{\theta}}_{k-1|k-1}\}$.[6] When the length of the channel memory is L, based on the trellis structure, the number of detection/estimation of $\tilde{\mathcal{U}}_{k|k}$ is q^L and the number of the possibilities of $\tilde{\mathcal{U}}_{k|k-1}$ is q^{L+1} [e.g., $\tilde{\mathcal{U}}_{k|k-1}^{(\vartheta)}$ for $\vartheta = 1, \cdots, q^{L+1}$]. Therefore, each $\tilde{\mathcal{U}}_{k|k}$ is maximized over q values of $\tilde{\mathcal{U}}_{k|k-1}$. To be more precise, each $\tilde{\mathcal{U}}_{k|k-1}$ should correspond with each hypothesis of $\boldsymbol{a}_k$, the sequence of transmitted symbols affecting $y(k)$. For easy presentation, however, we avoid using the $\tilde{\mathcal{U}}_{k|k-1}^{(\vartheta)}$ notation.

[6] In PBP method, there are q possible estimates for channel parameters $\boldsymbol{\theta}_k$ for each state. However, since the detection is done before the estimation in PSP, there is only one estimate for each state, and the estimation of $\boldsymbol{\theta}_k$ is computed only for survivor paths.

Meanwhile, although the EM algorithm is a method to achieve ML criterion with expectation and maximization steps, the algorithm does not tell us how to do these steps. We only considered some very general implementation aspects of the detection and estimation steps in this section. More details of the GMLSDE implementation by adaptive algorithms will be explored in the next section based on different channel models.

III. Adaptive MLSDE Based on Channel Models

The GMLSDE algorithm in Section II was derived in a general framework without specifying any channel model. In this section, we show how some adaptive MLSD/MLSDE algorithms developed previously in the literature, along with some new adaptive MLSDE algorithms, can be derived from the GMLSDE. The main goal of this section is to show the power of the proposed GMLSDE algorithm in deriving joint detection and estimation algorithms in a unified way based on the different channel models and available channel knowledge.

We consider three different model categories for a linear channel:

1) known CIR;
2) unknown deterministic CIR (time-invariant/variant);
3) stochastic CIR (random vector/process)

$$\begin{cases} \text{known statistical parameters} \\ \text{unknown statistical parameters.} \end{cases}$$

In all above models, the additive noise $z(k)$ in (3) is considered as a circularly symmetric zero-mean white complex Gaussian random process whose variance is N_0. The main step for developing an adaptive MLSDE algorithm is to define $\boldsymbol{\theta}_k$, $\mathcal{C}_k$, and $\mathcal{I}_k$ in association with the channel model. Based on the definitions of these parameters, the procedure to implement adaptive MLSDE may contain only a detection part, an estimation part, or both. Also, it may need to do only the maximization step in the detection/estimation part. We focus on the statistical CIR model, which is suitable for mobile communications, and briefly mention known CIR and unknown deterministic CIR models whose MLSD/MLSDE algorithms are well known in the literature. Meanwhile, in order to reduce the complexity and implement the algorithm in a causal manner (as we explain later), adaptive MLSDE may not achieve the maximum likelihood, but increases the likelihood function in each recursion.

A discrete model of the communication system for a linear channel is shown in Fig. 1. $\mathbf{a} = \{a_k\}_{k=0}^{I}$ is the set of the transmitted symbols and b_k is upsampling of a_k by a factor $J = \lceil T/T_s \rceil$, where T is the symbol period and T_s is the sample period.[7] The received signal is

$$y(k)=\sum_{l=0}^{L} h(l,k)s(k-l)+z(k)=\boldsymbol{s}(k)\boldsymbol{h}(k)+z(k) \quad (22)$$

where the duration of $h(l, k)$ with respect to l is $L+1$. $s(k)$ is the output of the transmitter filter with impulse response $g(k)$, $z(k)$ is additive noise, $\boldsymbol{s}(k)=[s(k),\cdots,s(k-L)]$, and $\boldsymbol{h}(k)=[h(0,k),\cdots,h(L,k)]^{\mathcal{T}}$.

Model A—Known CIR: The MLSD receiver for known CIR was derived by Forney [1]. From the GMLSDE point of view, the complete data is available in the receiver or $\mathcal{C}_k=\mathcal{I}_k=\mathbf{y}_k=[y(k),\cdots,y(0)]^{\mathcal{T}}$, and there is no unknown parameter set $\boldsymbol{\theta}$. Therefore, only the maximization step of the detection part (18) is needed, and it can be done recursively by the Viterbi algorithm [1].

Model B—Unknown Deterministic CIR: It is common to consider the CIR as a vector of time-variant/invariant deterministic parameters in channels with ISI. In this model for a time-invariant CIR, the vector of unknown continuous parameters is defined as $\boldsymbol{\varphi}=\boldsymbol{h}=[h(0),\cdots,h(L)]^{\mathcal{T}}$, and at time k, the complete and incomplete data are equal to $\mathcal{C}_k=\mathcal{I}_k=\mathbf{y}_k$. Thus, only the maximization steps of estimation and detection parts (16), (18) are needed in this model. It can be shown that (16) leads to the RLS algorithm by using (21) [32], [33]. If $\tilde{\boldsymbol{h}}_{|k}$ is the estimation of CIR at time k, (18) becomes

$$\begin{aligned}\tilde{\mathbf{a}}_{k|k}&=\arg\max_{\mathbf{a}_k}\left\{Q_{2,k}\left(\mathbf{a}_k|\tilde{\mathbf{a}}_{k|k-1},\tilde{\boldsymbol{h}}_{|k-1}\right)\right\}\\&=\arg\max_{\mathbf{a}_k}\left\{\log p\left(\mathbf{y}_k|\mathbf{a}_k,\tilde{\boldsymbol{h}}_{|k}\right)\right\}\\&=\arg\max_{\mathbf{a}_k}\left\{\sum_{l=0}^{k}-|y(l)-\boldsymbol{s}(l)\tilde{\boldsymbol{h}}_{|k}|^2\right\},\\&\qquad\text{for all }\left\{\tilde{\mathbf{a}}_{k|k-1},\tilde{\boldsymbol{h}}_{|k}\right\}. \end{aligned}\quad (23)$$

Using the new estimation of $\boldsymbol{h}$ at time k, $\tilde{\boldsymbol{h}}_{|k}$ in order to detect $\mathbf{a}_k=\{a_k,\cdots,a_0\}$ from (23) is a very complex and noncausal process. In order to avoid complexity, a causal detection procedure[8] is considered. Therefore, instead of $\tilde{\boldsymbol{h}}_{|k}$, $\tilde{\boldsymbol{h}}_{|l}$ is used in calculating $|y(l)-\boldsymbol{s}(l)\tilde{\boldsymbol{h}}_{|k}|^2$ for $0\le l\le k$ in (23). Thus, (23) becomes

$$\tilde{\mathbf{a}}_{k|k}=\arg\max_{\mathbf{a}_k}\left\{\sum_{l=0}^{k}-|y(l)-\boldsymbol{s}(l)\tilde{\boldsymbol{h}}_{|l}|^2\right\}. \quad (24)$$

Since $\tilde{\mathbf{a}}_{k-1|k-1}$ maximizes $\sum_{l=0}^{k-1}-|y(l)-\boldsymbol{s}(l)\tilde{\boldsymbol{h}}_{|l}|^2$, (24) can be implemented in a recursive manner using the Viterbi algorithm. It should be noted that the causal detection procedure does not guarantee achieving the maximum of the likelihood function (global or local); however, it guarantees to increase the likelihood function in each recursion.[9] Meanwhile, the estimation and detection procedures show how the idea of PSP [11] emerges from the GMLSDE algorithm as a recursive approach to increase the likelihood function.

For a time-variant deterministic CIR, the estimation part leads to a Kalman-type algorithm [32] where the causal detection procedure needs only a causal estimation procedure.[10]

Model C—Stochastic CIR with Known Parameters: The CIR is often modeled as a random vector or random process in a mobile environment. For example, Rayleigh multipath fading, whose impulse response is considered as a complex Gaussian random vector (for very slow fading) or process (for fast fading), is a very common model in mobile communication systems. In Model "C," we focus on the random process CIR. The received signal is obtained from (22) with $h(l,k)=h_l(k)$ being considered a Gaussian random process for $0\le l\le L$. By defining $\mathbf{h}_k=\mathbf{h}(k)=[\boldsymbol{h}(k)^{\mathcal{T}},\boldsymbol{h}(k-1)^{\mathcal{T}},\cdots,\boldsymbol{h}(k-M+1)^{\mathcal{T}}]^{\mathcal{T}}$, the received signal $y(k)$ and the dynamic changing of the CIR represented by the state-space formulation is given by

$$y(k)=\mathbf{s}(k)\mathbf{h}(k)+z(k) \quad (25)$$

$$\mathbf{h}(k)=\boldsymbol{F}\mathbf{h}(k-1)+G\boldsymbol{w}(k) \quad (26)$$

where $\mathbf{s}(k)=[\boldsymbol{s}(k),\mathbf{0}]$, $\mathbf{0}$ is an $(M-1)(L+1)$ zero row vector, $\boldsymbol{w}(k)=[w_0(k),\cdots,w_L(k)]^{\mathcal{T}}$ is a zero-mean white complex stationary Gaussian vector process which is independent of $z(k)$ and its autocorrelation matrix is $R_{\boldsymbol{w}}(k)=I_{(L+1)}\delta(k)$, where $I_{(L+1)}$ is an $(L+1)\times(L+1)$ identity matrix. $\boldsymbol{F}$ and G matrices are defined by

$$\boldsymbol{F}=\begin{bmatrix}F_1 & F_2 & \cdots & F_M\\ I_{(L+1)} & 0_{(L+1)} & \cdots & 0_{(L+1)}\\ & \ddots & 0_{(L+1)} & \\ & 0_{(L+1)} & I_{(L+1)} & 0_{(L+1)}\end{bmatrix}$$

$$G=\begin{bmatrix}\mathbf{g}\\ 0_{(L+1)}\\ \vdots\\ 0_{(L+1)}\end{bmatrix} \quad (27)$$

where $0_{(L+1)}$ is an $(L+1)\times(L+1)$ zero matrix.

In this channel model, the complete data and the incomplete data at time k are $\mathcal{C}_k=\{\mathbf{y}_k,\overline{\mathbf{h}}_k\}$ and $\mathcal{I}_k=\mathbf{y}_k$, respectively, where $\overline{\mathbf{h}}_k=[\mathbf{h}_k^{\mathcal{T}},\cdots,\mathbf{h}_0^{\mathcal{T}}]^{\mathcal{T}}$. It is easy to show that the unknown continuous parameters at time k is $\boldsymbol{\theta}_{k|k}=[\boldsymbol{\varphi}_{k|k}^{\mathcal{T}},\boldsymbol{\varphi}_{k-1|k}^{\mathcal{T}},\cdots,\boldsymbol{\varphi}_{0|k}^{\mathcal{T}}]^{\mathcal{T}}$ where $\boldsymbol{\varphi}_{l|k}=\boldsymbol{\mu}_{l|k}=E[\mathbf{h}_l|\mathbf{y}_k]$. Meanwhile, it can be shown that the estimation of $\sum_{l|k}=\mathrm{cov}(\mathbf{h}_l|\mathbf{y}_k)$ is necessary to estimate $\boldsymbol{\mu}_{l|k}$ [32], [34]. In this model, all the steps in detecting and estimating parts are necessary. The expectation step of detection part (17) becomes

$$\begin{aligned}&Q_{2,k}\left(\mathbf{a}_k|\tilde{\mathbf{a}}_{k|k-1},\tilde{\boldsymbol{\theta}}_{k|k-1}\right)\\&\quad=E\left[\log p\left(\mathbf{y}_k,\overline{\mathbf{h}}_k|\mathbf{a}_k,\tilde{\boldsymbol{\theta}}_{k|k}\right)|\tilde{\mathbf{a}}_{k|k-1},\tilde{\boldsymbol{\theta}}_{k|k-1},\mathbf{y}_k\right]\\&\quad=\sum_{l=0}^{k}\Big\{\log p(y(l)|\mathbf{a}_k,\tilde{\boldsymbol{\varphi}}_{l|k-1},\mathbf{y}_{l-1})\\&\qquad+E\Big[\log p(\boldsymbol{h}(l)|\mathbf{a}_k,\tilde{\boldsymbol{\varphi}}_{l|k},\mathbf{h}(l-1),\mathbf{y}_k)\\&\qquad\cdot|\tilde{\mathbf{a}}_{k|k-1},\tilde{\boldsymbol{\theta}}_{l|k-1},\mathbf{y}_k\Big]\Big\}. \end{aligned}\quad (28)$$

[7] When $J>1$, each recursion in the detection part needs J recursions in the estimation part.

[8] Detecting a_k from $y(l)$ for $l\le k$ is defined as a causal detecting procedure.

[9] From Appendix C, it is easy to show that $L_l(\tilde{\boldsymbol{h}}_{|l})\ge L_l(\tilde{\boldsymbol{h}}_{|l-1})$.

[10] Estimating $\mathbf{h}(k)=[\boldsymbol{h}(k)^{\mathcal{T}},\cdots,\boldsymbol{h}(k-M+1)^{\mathcal{T}}]^{\mathcal{T}}$ from $y(l)$ for $l\le k$ is defined as a causal estimation procedure.

Due to the noncausal estimation of $\boldsymbol{\varphi}_{l|k-1}$ for $0 \leq l \leq k$, detecting $\mathbf{a}_k$ from (28) is very complex. Considering causal estimation of $\boldsymbol{\mu}_{l|k-1} = E[\mathbf{h}_l|\mathbf{y}_{k-1}]$ and $\sum_{l|k-1} = \mathrm{cov}(\mathbf{h}_l|\mathbf{y}_{k-1})$, the detecting procedure at time k becomes

$$\tilde{\mathbf{a}}_{k|k} = \arg\max_{\mathbf{a}_k} \left\{ \sum_{l=0}^{k} -\beta(l) \right\} \tag{29}$$

such that the branch metric $\beta(l)$ is given by

$$\begin{aligned}\beta(l) =& \frac{|y(l) - \tilde{\mathbf{s}}(l)\tilde{\boldsymbol{\mu}}_{l|l-1}|^2}{\tilde{\mathbf{s}}(l)\tilde{\Sigma}_{l|l-1}\tilde{\mathbf{s}}(l)^{\mathcal{H}} + N_0} \\ &+ \log\left(\tilde{\mathbf{s}}(l)\tilde{\Sigma}_{l|l-1}\tilde{\mathbf{s}}(l)^{\mathcal{H}} + N_0\right) \\ &+ \log\left(\det\left(\overline{\boldsymbol{F}}\tilde{\Sigma}_{\boldsymbol{x}_{l|l}}\overline{\boldsymbol{F}}^{\mathcal{H}}\right)\right)\end{aligned} \tag{30}$$

where $\boldsymbol{x} = [\boldsymbol{h}(k)^{\mathcal{T}}, \cdots, \boldsymbol{h}(k-M)^{\mathcal{T}}]^{\mathcal{T}}$, $\overline{\boldsymbol{F}} = [I_{(L+1)}, -F_1, \cdots, -F_M]$, and $\tilde{\Sigma}_{\boldsymbol{x}_{l|l}} = \mathrm{cov}(\boldsymbol{x}_l|\mathbf{y}_l)$. By using the Viterbi algorithm, (29) can be implemented in a recursive manner where $\beta(l)$ is the branch metric of the trellis diagram at time l. It can be shown that causal estimation leads the estimating part of Model "C" to the Kalman algorithm by using (21), where $\tilde{\boldsymbol{\mu}}_{l+1|l}$ and $\sum_{l+1|l}$ are updated for all possibilities of $\{\tilde{\mathbf{a}}_{l|l-1}, \tilde{\boldsymbol{\theta}}_{l|l-1}\}$ using the recursions given by [32], [35]

$$\begin{aligned}\tilde{\boldsymbol{\mu}}_{l+1|l} =& \boldsymbol{F}\tilde{\boldsymbol{\mu}}_{l|l-1} + \boldsymbol{F}\tilde{\Sigma}_{l|l-1}\tilde{\mathbf{s}}(l)^{\mathcal{H}} \\ &\cdot \left(N_0 + \tilde{\mathbf{s}}(l)\tilde{\Sigma}_{l|l-1}\tilde{\mathbf{s}}(l)^{\mathcal{H}}\right)^{-1}\left(y(l) - \tilde{\mathbf{s}}(l)\tilde{\boldsymbol{\mu}}_{l|l-1}\right)\end{aligned} \tag{31}$$

$$\tilde{\Sigma}_{l+1|l} = \boldsymbol{F}\tilde{\Sigma}_{l|l}\boldsymbol{F}^{\mathcal{H}} + GG^{\mathcal{H}} \tag{32}$$

where

$$\begin{aligned}\tilde{\Sigma}_{l|l} =& \tilde{\Sigma}_{l|l-1} + \tilde{\Sigma}_{l|l-1}\tilde{\mathbf{s}}(l)^{\mathcal{H}} \\ &\cdot \left(N_0 + \tilde{\mathbf{s}}(l)\tilde{\Sigma}_{l|l-1}\tilde{\mathbf{s}}(l)^{\mathcal{H}}\right)^{-1}\tilde{\mathbf{s}}(l)\tilde{\Sigma}_{l|l-1}.\end{aligned} \tag{33}$$

Meanwhile, it is straightforward to compute $\tilde{\Sigma}_{\boldsymbol{x}_{l|l}}$ from $\tilde{\Sigma}_{l|l}$ using (26).

The branch metric measure derived in (30) is different from the branch metric measure proposed in [2], [3], [6], and [36]. Although the same state-space model was chosen in the above references, the last term in (30) is extra in comparison with the branch metric proposed in [2], [3], [6], and [36], which maximizes the logarithm of the pdf of $\mathbf{y}_k$ at time k to compute the branch metrics. In GMLSDE using Model "C," however, the expectation of the logarithm of the joint pdf of $\mathbf{y}_k$ and $\overline{\mathbf{h}}_k$ over $\overline{\mathbf{h}}_k$ is maximized. Thus, the last term in (30) can be interpreted as the contribution of the error in estimating $\boldsymbol{h}(k)$ given $y(k)$ to the branch metric at time k. Meanwhile, since $\overline{\boldsymbol{F}}\tilde{\Sigma}_{\boldsymbol{x}_{l|l-1}} = 0$, if $\tilde{\Sigma}_{\boldsymbol{x}_{l|l}}$ is replaced with $\tilde{\Sigma}_{\boldsymbol{x}_{l|l-1}}$ (using PSP method instead of PBP method), the last term of (30) vanishes in computing the branch metric.

Model D—Stochastic CIR with Unknown Parameters: In the estimation part of Model "C," it was assumed that $\boldsymbol{F}$ and $GG^{\mathcal{H}}$ matrices were known in updating (31) and (32). Generally, these matrices are also unknown and should be estimated from the received signal. In Model "D," we divide the unknown parameters into two separate sets $\boldsymbol{\theta}_{1,k|k} = [\boldsymbol{\varphi}_{1,k|k}^{\mathcal{T}}, \boldsymbol{\varphi}_{1,k-1|k}^{\mathcal{T}}, \cdots, \boldsymbol{\varphi}_{1,0|k}^{\mathcal{T}}]^{\mathcal{T}}$ and $\{\boldsymbol{F}, GG^{\mathcal{H}}\}$ at time k where $\boldsymbol{\varphi}_{1,l|k} = \boldsymbol{\mu}_{l|k}$. The complete and incomplete data at time k are $\mathcal{C}_k = \{\mathbf{y}_k, \overline{\mathbf{h}}_k\}$ and $\mathcal{I}_k = \mathbf{y}_k$, respectively. Since there are two separate unknown continuous parameter sets, the estimation part contains four steps where, based on Theorem 1, $\boldsymbol{\theta}_{1,k|k}$ is estimated in the first two steps of expectation and maximization. In the second two steps, $\{\boldsymbol{F}, GG^{\mathcal{H}}\}$ is estimated. The detection part and estimation part of $\boldsymbol{\theta}_{1,k|k}$ in this model are similar to those of Model "C" based on the estimation of $\{\boldsymbol{F}, GG^{\mathcal{H}}\}$ parameters.

It is more convenient to estimate $\{\boldsymbol{F}, GG^{\mathcal{H}}\}$ based on an autoregressive moving average (ARMA) model of $h_l(k) = h(l,k)$. Also, without loss of generality and only for easier presentation, we assume $\mathbf{g}$ is a diagonal matrix $\mathrm{diag}(\mathbf{g}) = \{\mathbf{g}^0, \mathbf{g}^1, \cdots, \mathbf{g}^L\}$. The ARMA model of $h_l(k)$ is

$$h_l(k) = \mathbf{f}^l\mathbf{h}_{k-1} + \mathbf{g}^l w_l(k), \qquad 0 \leq l \leq L \tag{34}$$

where $\mathbf{f}^l$ is the lth row of $\boldsymbol{F}$ matrix. By defining $\boldsymbol{\varphi}_2 = [\mathbf{f}^L, \cdots, \mathbf{f}^0, R_{g^L}, \cdots, R_{g^0}]^{\mathcal{T}}$ where $R_{g^l} = \mathbf{g}^l\mathbf{g}^{l^{\mathcal{H}}}$ for $l = 0, \cdots, L$, the E-step of estimating $\boldsymbol{\varphi}_2$ at time k is given by

$$\begin{aligned}&Q_{2,k}\left(\boldsymbol{\varphi}_2|\tilde{\mathbf{a}}_{k|k-1}, \tilde{\boldsymbol{\theta}}_{1,k|k-1}, \tilde{\boldsymbol{\varphi}}_{2|k-1}\right) \\ &\quad= \log p\left(\mathbf{y}_k|\overline{\mathbf{h}}_k, \tilde{\mathbf{a}}_{k|k-1}, \tilde{\boldsymbol{\theta}}_{1,k|k}, \boldsymbol{\varphi}_2\right) \\ &\qquad+ \sum_{l=0}^{L}\sum_{j=0}^{k} E\Big[\log p\left(h_l(j)|\overline{\mathbf{h}}_{j-1}, \tilde{\mathbf{a}}_{k|k-1}, \tilde{\boldsymbol{\theta}}_{1,k|k}, \boldsymbol{\varphi}_2\right) \\ &\qquad\qquad \cdot |\tilde{\mathbf{a}}_{k|k-1}, \tilde{\boldsymbol{\theta}}_{1,k|k-1}, \tilde{\boldsymbol{\varphi}}_{2|k-1}, \mathbf{y}_k\Big].\end{aligned} \tag{35}$$

Taking the first and the second derivatives of $Q_{2,k}(\boldsymbol{\varphi}_2|.)$ with respect to $\mathbf{f}^l$ and R_{g^l} and using (21) along with a causal estimation procedure with some approximations and manipulations (see Appendix D), the estimation of $\mathbf{f}^l$ and R_{g^l} at time k becomes

$$\begin{aligned}\tilde{\mathbf{f}}^l_{|k} \simeq& \tilde{\mathbf{f}}^l_{|k-1} + \tilde{\boldsymbol{\mu}}^{\mathcal{H}}_{k-1|k}P^{\diamond}_{k-1|k}\left(1 + \tilde{\boldsymbol{\mu}}^{\mathcal{H}}_{k-1|k}P^{\diamond}_{k-1|k}\tilde{\boldsymbol{\mu}}_{k-1|k}\right)^{-1} \\ &\cdot\left(\tilde{\mu}^l_{k|k} - \tilde{\mathbf{f}}^l_{|k-1}\tilde{\boldsymbol{\mu}}_{k-1|k}\right), \qquad 0 \leq l \leq L\end{aligned} \tag{36}$$

$$\begin{aligned}\tilde{R}_{g^l|k} \simeq& \tilde{R}_{g^l|k-1} + \frac{1}{k}\left(\tilde{R}_{g^l|k-1} - |\tilde{\mu}^l_{k|k} - \tilde{\mathbf{f}}^l_{|k-1}\tilde{\boldsymbol{\mu}}_{k-1|k}|^2\right), \\ &0 \leq l \leq L\end{aligned} \tag{37}$$

where $P^{\diamond}_{k-1|k} = (\sum_{j=1}^{k}\tilde{\boldsymbol{\mu}}_{j-1|j}\tilde{\boldsymbol{\mu}}^{\mathcal{H}}_{j-1|j})^{-1}$, and $P^{\diamond}_{k|k+1}$ is given by

$$\begin{aligned}&P^{\diamond}_{k|k+1} \\ &\quad= P^{\diamond}_{k-1|k} - P^{\diamond}_{k-1|k}\tilde{\boldsymbol{\mu}}_{k-1|k}\left(1 + \tilde{\boldsymbol{\mu}}^{\mathcal{H}}_{k-1|k}P^{\diamond}_{k-1|k}\tilde{\boldsymbol{\mu}}_{k-1|k}\right)^{-1} \\ &\qquad\cdot\tilde{\boldsymbol{\mu}}^{\mathcal{H}}_{k-1|k}P^{\diamond}_{k-1|k}.\end{aligned} \tag{38}$$

$\mathbf{f}^l$ and R_{g^l}, computed from (36) and (37), are used to find $\tilde{\boldsymbol{\mu}}_{k+1|k}$ and $\tilde{\Sigma}_{k+1|k}$ from (31) and (32) at time $k+1$ in the framework of GMLSDE. Meanwhile, as can be seen in (36), the estimation procedure of $\mathbf{f}^l_{|k}$ is similar to the RLS algorithm.

Hart and Taylor have recently proposed a method to estimate the unknown statistical parameters of fading channels [6]. The method in [6] estimates the mean vector and autocovariance matrix of CIR based on computing the mean and autocovariance of the received signal; however, it is complex and nonrecursive [37]. The method proposed in Model "D"

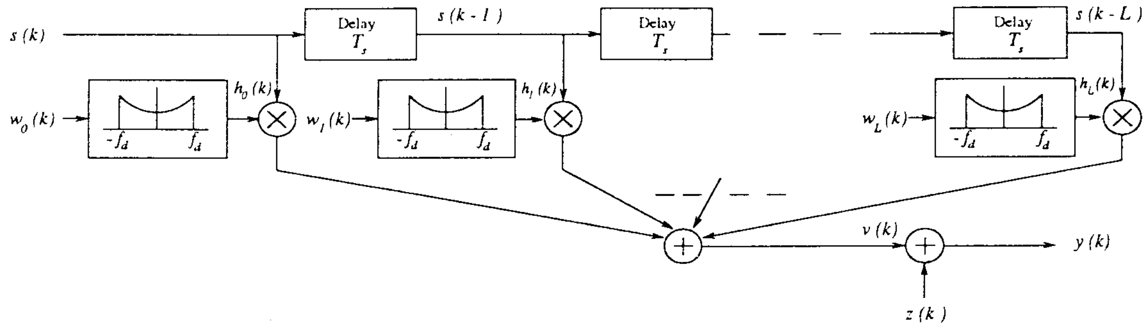


Fig. 2. The baseband-tapped delay line model of the multipath fading channel.

estimates the unknown parameters based on increasing the likelihood function in a recursive manner.

IV. Computer Simulations and Comparisons

Computer simulations have been done for different channel models to evaluate the performances of the adaptive MLSDE algorithms. The channel model is selected as flat fading and frequency-selective fading with three fading rates $f_dT = 0.1$, 0.01, and 0.001. The autocorrelation function of the CIR is modeled as

$$R_h(l_1, l_2; j) = \sum_{l_1=0}^{L} \exp(-bl_1T_s)J_0(2\pi f_d jT_s)\delta(l_1 - l_2), \quad 0 \le l_1, l_2 \le L - \infty \le j \le \infty \tag{39}$$

where J_0 is the zeroth-order Bessel function, and f_d is the maximum Doppler frequency. The delay rate in (39) was chosen as $b^{-1} = 2T_s$, $L = 0$, and $L = 2$ in (39) for flat fading and selective fading channels, respectively (i.e., the channel was simulated with three paths in the selective channel) (see Fig. 2). The impulse response of the transmitter filter is a raised-cosine pulse

$$g(t) = \text{sinc}\left(\frac{t}{T}\right)\left(\frac{\cos(\nu\pi t/T)}{1-(2\nu t/T)^2}\right) \tag{40}$$

where the symbol duration $T = 1$ and $\nu = 0.35$, $g(k) = g(t)|_{t=kT_s}$ where $J = (T/T_s) = 2$. The DQPSK modulation scheme was chosen; therefore, the number of states in the trellis diagram is 4 and 16 for flat fading and selective fading channels, respectively. The Bessel fading filter for omnidirectional antenna is approximated by an all-pole third-order filter [38]; therefore, the $\boldsymbol{F}$ in (27) becomes 3×3 and 9×9 matrices in flat fading and selective fading channels, respectively. The data sequence is divided into a sequence of frames of length L_f, where the overhead of each frame known by the receiver is two and four symbols for flat and selective fading, respectively. The length of each frame was chosen as $L_f = 160$ data.

The bit-error rate (BER) performance of the adaptive MLSDE algorithm based on the online EM algorithm is considered for different levels of channel knowledge available at the receiver. Although the impulse response of the fading channel was simulated as a stochastic random process, we consider four levels of available channel knowledge at the receiver: a) known CIR; b) unknown deterministic CIR; c) stochastic CIR with known channel parameters $\boldsymbol{F}$ and $GG^{\mathcal{H}}$ matrices; and d) stochastic CIR with unknown $\boldsymbol{F}$ and $GG^{\mathcal{H}}$ matrices.

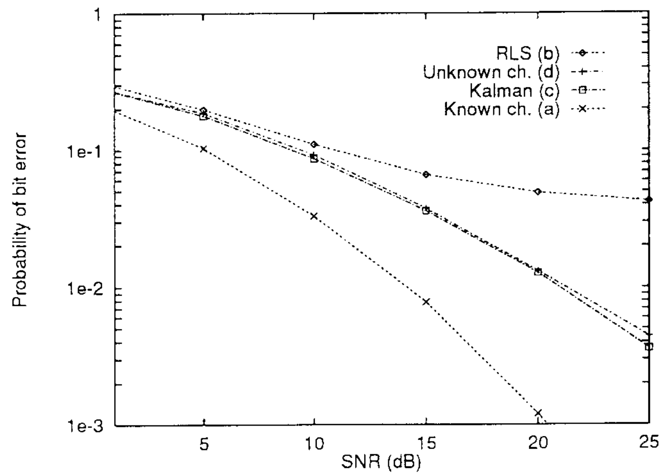


Fig. 3. BER performance for flat fading with $f_dT = 0.1$ and DQPSK signaling. (a) CIR is known. (b) Estimation of CIR using RLS algorithm. (c) Estimation of stochastic CIR with known parameters using Kalman filtering. (d) Stochastic CIR with unknown parameters.

Figs. 3–5 show the error performances of the adaptive MLSDE in the flat fading channel for $f_dT = 0.1$, 0.01, and 0.001, respectively. Also, the performances of the adaptive MLSDE for the selective fading channel are shown in Figs. 6–8 for $f_dT = 0.1$, 0.01, and 0.001, respectively.

In these figures, the simulation results in curves (a)–(d) correspond to Models "A," "B," "C," and "D" in Section III, respectively. Curves (a) show the performance of the receiver when the CIR is known. Curves (b) show how the receiver performs when it assumes a deterministic CIR and estimates it by the modified RLS algorithm with a proper forgetting factor λ for different fading rates (in order to optimize the RLS performance [35]). The performance of the receiver in curves (c) are achieved based on the assumption of a stochastic CIR, whose statistical parameters are estimated by Kalman filtering, with known channel parameters $\boldsymbol{F}$ and $GG^{\mathcal{H}}$. Finally, curves (d) indicate the receiver performance with a stochastic model for CIR and unknown channel parameters $\boldsymbol{F}$ and $GG^{\mathcal{H}}$, where both Kalman and RLS algorithms are used in the estimation part.

As can be seen, the performances of the receiver in (b) is close to the performance of (c) only for slow flat fading and

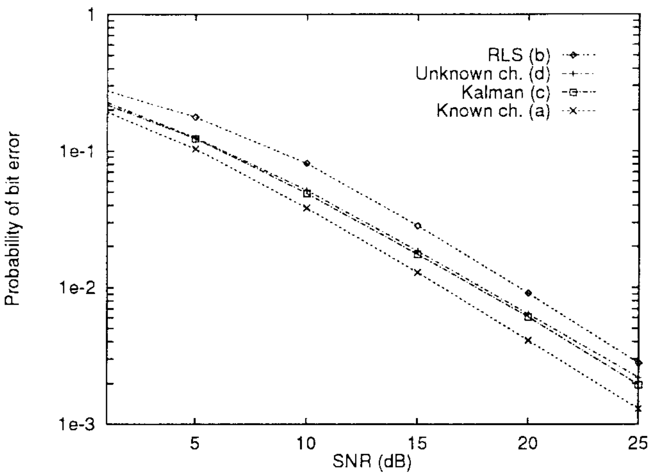


Fig. 4. BER performance for flat fading with $f_dT = 0.01$ and DQPSK signaling. (a) CIR is known. (b) Estimation of CIR using RLS algorithm. (c) Estimation of stochastic CIR with known parameters using Kalman filtering. (d) Stochastic CIR with unknown parameters.

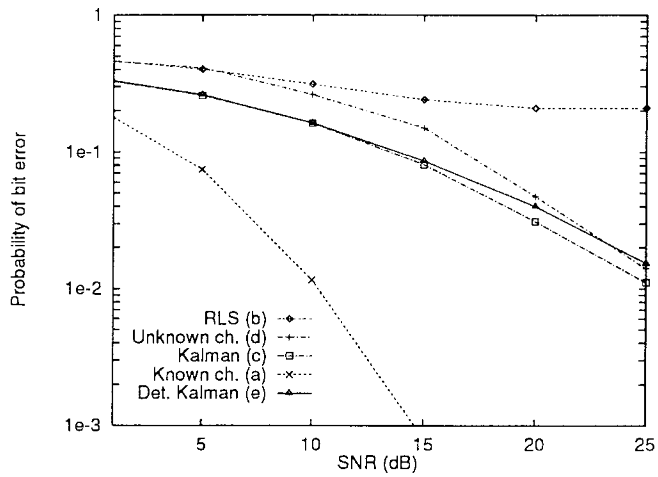


Fig. 6. BER performance for selective fading with $f_dT = 0.1$ and DQPSK signaling. (a) CIR is known. (b) Estimation of CIR using RLS algorithm. (c) Estimation of stochastic CIR with known parameters using Kalman filtering. (d) Stochastic CIR with unknown parameters. (e) Assuming time-variant deterministic CIR and estimating CIR with a Kalman-type algorithm.

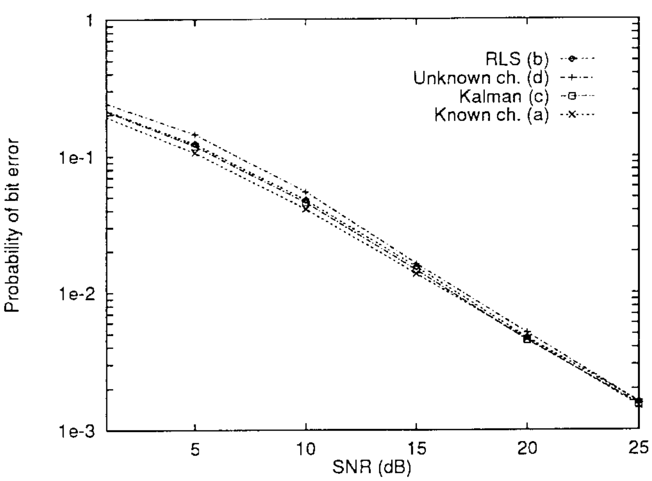


Fig. 5. BER performance for flat fading with $f_dT = 0.001$ and DQPSK signaling. (a) CIR is known. (b) Estimation of CIR using RLS algorithm. (c) Estimation of stochastic CIR with known parameters using Kalman filtering. (d) Stochastic CIR with unknown parameters.

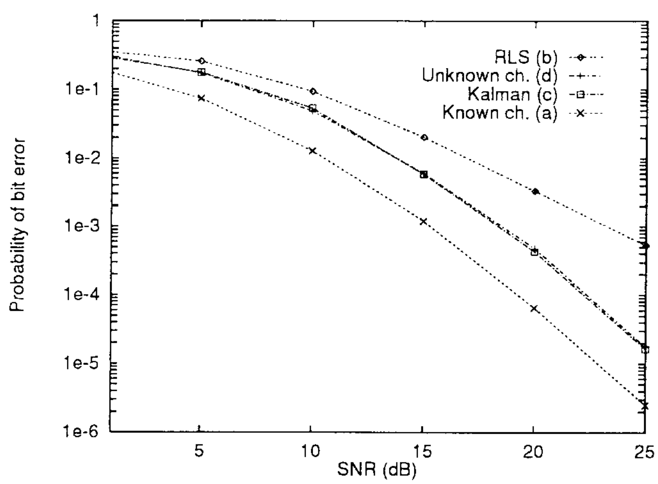


Fig. 7. BER performance for selective fading with $f_dT = 0.01$ and DQPSK signaling. (a) CIR is known. (b) Estimation of CIR using RLS algorithm. (c) estimation of stochastic CIR with known parameters using Kalman filtering. (d) Stochastic CIR with unknown parameters.

very slow selective channels. By increasing the fading rate, the difference in performance between (b) and (c) increases due to an insufficient number of degrees of freedom in the RLS updating equation to track the dynamic channel changes.

Meanwhile, the performances of (c) and (d) are close for $f_dT = 0.01$ and $f_dT = 0.001$. It should be mentioned that in order to limit the complexity of the algorithm and time of simulations in (d), the $\boldsymbol{F}$ and $GG^{\mathcal{H}}$ matrices were updated in each frame of data instead of each recursion. However, in (b)–(d) cases, the CIR is estimated in each recursion accompanied by PSP, which estimates different CIR for each survivor path in the trellis diagram. Meanwhile, the convergence speed of the algorithm in Model "D" for estimating the elements of unknown $\boldsymbol{F}$ and $GG^{\mathcal{H}}$ matrices depends on the fading rate. The length of data for converging is about five frames, one frame, and less than one frame for $f_dT = 0.001$, 0.01, and 0.1, respectively.

Fig. 9 shows the BER's of a flat fading channel for cases (a)–(d) when the fading rate is changing periodically between $f_dT = 0.1$ and $f_dT = 0.01$. One period of $f_dT(t)$ is

$$f_dT(t) = \begin{cases} 0.1, & 0 \le t \le T_f/4 \\ -\dfrac{1}{80000T}\left(t - \dfrac{T_f}{4}\right) + 0.1, & T_f/4 \le t \le T_f/2 \\ 0.01, & T_f/2 \le t \le 3T_f/4 \\ \dfrac{1}{80000T}\left(t - \dfrac{3T_f}{4}\right) + 0.01, & 3T_f/4 \le t \le T_f \end{cases} \tag{41}$$

where the period time $T_f = 28800T$. The linear change in fading rate corresponds to a linear change in vehicle speed. In this situation, the modified RLS algorithm with forgetting factor $\lambda = 0.999$ is used in estimating the $\boldsymbol{F}$ and $GG^{\mathcal{H}}$ matrices. As Fig. 9 shows, the performance of (c) with known

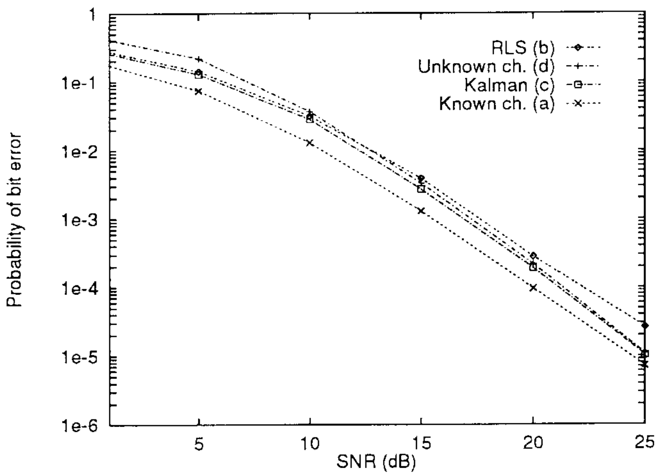


Fig. 8. BER performance for selective fading with $f_dT = 0.001$ and DQPSK signaling. (a) CIR is known. (b) Estimation of CIR using RLS algorithm. (c) Estimation of stochastic CIR with known parameters using Kalman filtering. (d) Stochastic CIR with unknown parameters.

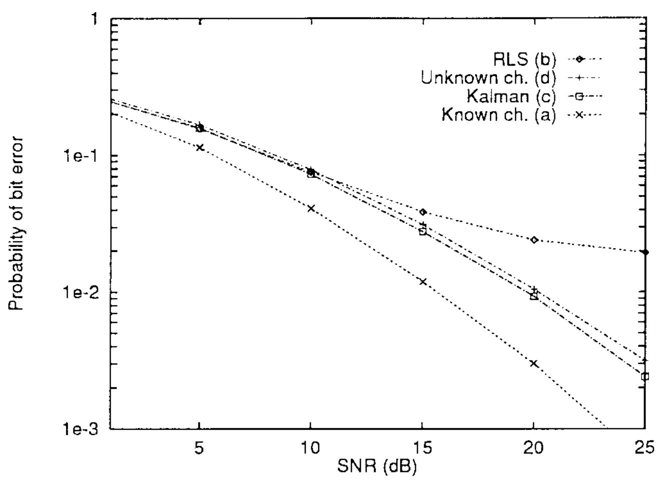


Fig. 9. BER performance of flat fading for linear changing in the normalized fading rate between $f_dT = 0.1$ and 0.01 for DQPSK signaling. (a) CIR is known. (b) Estimation of CIR using RLS algorithm. (c) Estimation of stochastic CIR with known parameters using Kalman filtering. (d) Stochastic CIR with unknown parameters.

$\boldsymbol{F}$ and $GG^{\mathcal{H}}$ and (d) with unknown $\boldsymbol{F}$ and $GG^{\mathcal{H}}$ are close. However, the performance of (b) with the assumption of unknown deterministic CIR is far from (c). Therefore, the maximum fading rate dominates the performance when the fading rate is time-variant. Meanwhile, in very slow flat fading channels, the performance of the differentially coherent detection methods, which are very low in complexity, is close to MLSD receiver with known CIR (Model "A") [2], [4], [39]. However, for selective fading and fast flat fading channels, the performance of the differentially coherent detection methods is very poor [2], [4], [40].

When the CIR is modeled as a random process (Model "C") or time-variant deterministic process (Model "B"), the estimation part can be done through a Kalman-type algorithm. However, the branch metrics are computed in a different manner. For a CIR modeled as a Gaussian random process, the branch metric is computed from (30) as shown in Model "C" or from (42), which was derived in [2], [3], [6], and [36], by maximizing $\log p(\mathbf{y}_k|\mathbf{a}_k)$

$$\beta(l) = \frac{|y(l) - \mathbf{s}(l)\tilde{\boldsymbol{\mu}}_{l|l-1}|^2}{\mathbf{s}(l)\tilde{\Sigma}_{l|l-1}\mathbf{s}(l)^{\mathcal{H}} + N_0} + \log\Big(\mathbf{s}(l)\tilde{\Sigma}_{l|l-1}\mathbf{s}(l)^{\mathcal{H}} + N_0\Big) \tag{42}$$

For deterministic time-variant CIR similar to Model "B," the branch metric becomes

$$\beta(k) = |y(l) - \mathbf{s}(l)\tilde{\mathbf{h}}_{l|l}|^2. \tag{43}$$

Simulation results (which have not been shown) for flat and selective fading do not show a significant difference in performance between computing branch metrics from (30) and (42). Also, for flat and selective fading with $f_dT = 0.01$ and 0.001, difference in performance between computing branch metrics from (43) and (30) or (42) is negligible; however, for flat fading with $f_dT = 0.1$, this difference is around 5% and the branch metric computed from (42) achieves better performance than the branch metric computed from (43) (not shown in Fig. 3). For selective fast fading ($f_dT = 0.1$), the performance of the receiver that computes the branch metrics from (43) is shown as curve (e) in Fig. 6.

These results show that when the variance of estimation error is small, computing the branch metrics from (43) is sufficient. Moreover, in this situation, by doing detection before estimation in the GMLSDE, where $\tilde{\mathbf{h}}_{l|l-1}$ is replaced by $\tilde{\mathbf{h}}_{l|l}$ in (43), the complexity of estimation is decreased since the estimation procedure needs to compute only q^L branches instead of q^{L+1}.

V. Conclusions

In this paper, we derived GMLSDE that generated coupled estimation and detection procedures based on the EM algorithm. In each recursion, estimation and detection were done alternately in order to increase the likelihood function. The PSP and PBP methods appear naturally in GMLSDE. When the detection part is done before the estimation part in GMLSDE, the PSP method comes up. However, when the estimation part is done before the detection part, the PBP method appears. Adaptive MLSDE algorithms were derived in the framework of GMLSDE in a unified way for some important channel models. In association with the channel model and the level of knowledge available at the receiver, the adaptive MLSDE contains all or some of the steps of the GMLSDE in the estimation and detection parts. The detection part of adaptive MLSDE algorithms is implemented by the Viterbi algorithm through the trellis structure. Titterington's approach, stochastic approximation, was used to implement the estimation part of the adaptive MLSDE. Although Titterington's approach is generally an approximate method based on a Taylor series expansion, when the third and higher order elements of Taylor series are zero (as was true for the models considered), it is exact.

Adaptive MLSDE algorithm, although not achieving the maximum likelihood but increasing the likelihood function in each recursion, was simulated for frequency flat and selective fading channels with three different fading rates based on four channel model assumptions at the receiver: known CIR, unknown deterministic CIR, and Gaussian CIR with known and unknown statistical parameters. The comparison between the simulation results showed that the deterministic unknown time-invariant CIR whose estimation leads to the RLS algorithm achieves a performance close to known CIR in slow-fading channels. However, in fast or relatively fast fading channels, the deterministic unknown time-variant CIR, whose estimation can be done with Kalman filtering and RLS algorithms for obtaining impulse response and the unknown constant parameters of channels, respectively, shows good performance. Only for fast selective fading channels, the use of a Gaussian process model for CIR achieves better performance. Meanwhile, by implementing the smoothing method instead of the filtering method in many of above procedures, one can derive forward–backward versions of the corresponding estimation and detection algorithms [17], [19].

Appendix A
Proof of Theorem 1

Without loss of generality, first we assume that $\mathbf{i}=2$. From the maximization steps of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ (10), we conclude

$$Q\left(\hat{\boldsymbol{\theta}}_1^{(l+1)}, \hat{\boldsymbol{\theta}}_2^{(l)} | \hat{\mathcal{U}}^{(l)}\right) \geq Q\left(\hat{\boldsymbol{\theta}}_1^{(l)}, \hat{\boldsymbol{\theta}}_2^{(l)} | \hat{\mathcal{U}}^{(l)}\right) \tag{A.1}$$

$$Q\left(\hat{\boldsymbol{\theta}}_1^{(l+1)}, \hat{\boldsymbol{\theta}}_2^{(l+1)} | \hat{\mathcal{U}}^{(l)}\right) \geq Q\left(\hat{\boldsymbol{\theta}}_1^{(l+1)}, \hat{\boldsymbol{\theta}}_2^{(l)} | \hat{\mathcal{U}}^{(l)}\right). \tag{A.2}$$

The left-hand side of (A.1) is equal to the right-hand side of (A.2). Therefore from (A.1) and (A.2), we have

$$Q\left(\hat{\mathcal{U}}^{(l+1)} | \hat{\mathcal{U}}^{(l)}\right) \geq Q\left(\hat{\mathcal{U}}^{(l)} | \hat{\mathcal{U}}^{(l)}\right). \tag{A.3}$$

Defining $V(\mathcal{U}|\hat{\mathcal{U}}^{(l)}) = E[\log p(\mathcal{D}|\mathcal{U}, \mathcal{I})|\hat{\mathcal{U}}^{(l)}, \mathcal{I}]$ and from Jensen's inequality [14, Lemma 1], we can show

$$V\left(\hat{\mathcal{U}}^{(l+1)} | \hat{\mathcal{U}}^{(l)}\right) \leq V\left(\hat{\mathcal{U}}^{(l)} | \hat{\mathcal{U}}^{(l)}\right). \tag{A.4}$$

Thus, from (6) and considering inequalities (A.3) and (A.4), it is easy to show

$$\boldsymbol{L}\left(\hat{\mathcal{U}}^{(l+1)}\right) \geq \boldsymbol{L}\left(\hat{\mathcal{U}}^{(l)}\right). \tag{A.5}$$

By following the same procedure (A.1)–(A.4), one can achieve (A.5) for $\mathbf{i} > 2$.

Appendix B

When $\theta^{(l+1)}$ is used instead of $\theta^{(l)}$ as a given condition in (13), one can relate (11)–(14) to the SAGE algorithm [28, eqs. (5)–(7)] as follows:

For $l = 0, 1 \cdots \{$

1) Index set S^l in [28] is chosen as

$$S^l = \begin{cases} \boldsymbol{\theta}, & l \text{ even} \\ \mathbf{a}, & l \text{ odd.} \end{cases} \tag{B.1}$$

2) Admissible hidden data is chosen as $X^{S^l} = \mathcal{C}$ for all l.

3) E-step:

$$Q_{(l) \bmod 2+1}(\cdot) = \begin{cases} E\left[\log p(\mathcal{C}|\hat{\mathbf{a}}^{(l)}, \boldsymbol{\theta})|\hat{\mathbf{a}}^{(l)}, \hat{\boldsymbol{\theta}}^{(l)}, \mathcal{I}\right], & l \text{ even} \\ E\left[\log p\left(\mathcal{C}|\mathbf{a}, \hat{\boldsymbol{\theta}}^{(l)}\right)|\hat{\mathbf{a}}^{(l)}, \hat{\boldsymbol{\theta}}^{(l)}, \mathcal{I}\right], & l \text{ odd.} \end{cases} \tag{B.2}$$

4) M-step: Maximization step for l even is

$$\hat{\boldsymbol{\theta}}^{(l+1)} = \arg\max_{\boldsymbol{\theta}} \left\{Q_1\left(\boldsymbol{\theta}|\hat{\mathbf{a}}^{(l)}, \hat{\boldsymbol{\theta}}^{(l)}\right)\right\} \tag{B.3}$$

$$\hat{\mathbf{a}}^{(l+1)} = \hat{\mathbf{a}}^{(l)} \tag{B.4}$$

and for l odd it is

$$\hat{\mathbf{a}}^{(l+1)} = \arg\max_{\mathbf{a}} \left\{Q_2\left(\mathbf{a}|\hat{\mathbf{a}}^{(l)}, \hat{\boldsymbol{\theta}}^{(l)}\right)\right\} \tag{B.5}$$

$$\hat{\boldsymbol{\theta}}^{(l+1)} = \hat{\boldsymbol{\theta}}^{(l)} \tag{B.6}$$

$\}$

Appendix C

Defining $\boldsymbol{\theta}_k = [\boldsymbol{\varphi}_k^{\mathcal{T}}, \boldsymbol{\varphi}_{k-1}^{\mathcal{T}}, \cdots, \boldsymbol{\varphi}_0^{\mathcal{T}}]^{\mathcal{T}}$ where $\boldsymbol{\varphi}_l$ is the unknown parameter vector at time l and $\boldsymbol{a}_l$ as the sequence of transmitted symbols effecting $y(l)$, we have

$$\boldsymbol{L}_k\left(\tilde{\mathbf{a}}_{k|k-1}, \tilde{\boldsymbol{\theta}}_{k|k-1}\right) = \boldsymbol{L}_{k-1}\left(\tilde{\mathbf{a}}_{k-1|k-1}, \tilde{\boldsymbol{\theta}}_{k-1|k-1}\right) + L_k(\tilde{\boldsymbol{a}}_{k|k-1}, \tilde{\boldsymbol{\varphi}}_{k|k-1}) \tag{C.1}$$

where $L_k(\cdot) = \log p(y(k)|\cdot, \mathbf{y}_k - 1)$. Similar to (19), we have

$$\boldsymbol{L}_{k-1}\left(\tilde{\mathbf{a}}_{k-1|k-1}, \tilde{\boldsymbol{\theta}}_{k-1|k-1}\right) \geq \boldsymbol{L}_{k-1}\left(\tilde{\mathbf{a}}_{k-1|k-2}, \tilde{\boldsymbol{\theta}}_{k-1|k-2}\right). \tag{C.2}$$

Also, similar to (C.1) we can get

$$\boldsymbol{L}_{k-1}\left(\tilde{\mathbf{a}}_{k-1|k-2}, \tilde{\boldsymbol{\theta}}_{k-1|k-2}\right) = \boldsymbol{L}_{k-2}\left(\tilde{\mathbf{a}}_{k-2|k-2}, \tilde{\boldsymbol{\theta}}_{k-2|k-2}\right) + L_{k-1}(\tilde{\boldsymbol{a}}_{k-1|k-2}, \tilde{\boldsymbol{\varphi}}_{k-1|k-2}). \tag{C.3}$$

By substituting the right-hand side of (C.3) in (C.2) and using (19) and (C.1) and also extending (C.2) and (C.3) for $l < k-1$, one can conclude

$$\begin{aligned} &\boldsymbol{L}_k\left(\tilde{\mathbf{a}}_{k|k}, \tilde{\boldsymbol{\theta}}_{k|k}\right) \\ &\geq \boldsymbol{L}_{k-1}\left(\tilde{\mathbf{a}}_{k-1|k-1}, \tilde{\boldsymbol{\theta}}_{k-1|k-1}\right) + L_k(\tilde{\boldsymbol{a}}_{k|k-1}, \tilde{\boldsymbol{\varphi}}_{k|k-1}) \\ &\geq \boldsymbol{L}_{k-2}\left(\tilde{\mathbf{a}}_{k-2|k-2}, \tilde{\boldsymbol{\theta}}_{k-2|k-2}\right) + \sum_{l=k-1}^{k} L_l(\tilde{\boldsymbol{a}}_{l|l-1}, \tilde{\boldsymbol{\varphi}}_{l|l-1}) \\ &\geq \cdots \geq \boldsymbol{L}_0\left(\tilde{\mathbf{a}}_0, \tilde{\boldsymbol{\theta}}_0\right) + \sum_{l=1}^{k} L_l(\tilde{\boldsymbol{a}}_{l|l-1}, \tilde{\boldsymbol{\varphi}}_{l|l-1}). \end{aligned} \tag{C.4}$$

APPENDIX D

The first derivative of $Q_{2,k}(\boldsymbol{\varphi}_2|.)$ with respect to $\mathbf{f}^{l^*}$ at point $\boldsymbol{\varphi}_2 = \tilde{\boldsymbol{\varphi}}_{2|k-1}$ is

$$\left.\frac{\partial Q_{2,k}\left(\boldsymbol{\varphi}_2|\tilde{\mathbf{a}}_{k|k-1}, \tilde{\boldsymbol{\theta}}_{1,k|k-1}, \tilde{\boldsymbol{\varphi}}_{2|k-1}\right)}{\partial \mathbf{f}^{l^*}}\right|_{\boldsymbol{\varphi}_2=\tilde{\boldsymbol{\varphi}}_{2|k-1}} = \frac{1}{\tilde{R}_{g^l_{|k-1}}} \sum_{j=1}^{k} \left\{ \tilde{\boldsymbol{\sigma}}^l_{j-1,j|k} - \tilde{\mathbf{f}}^l_{|k-1}\tilde{\Sigma}_{j-1,j-1|k} + \tilde{\boldsymbol{\mu}}^{\mathcal{H}}_{j-1|k}\left(\tilde{\mu}^l_{j|k} - \tilde{\mathbf{f}}^l_{|k-1}\tilde{\boldsymbol{\mu}}_{j-1|k}\right)\right\}, \quad 0 \le l \le L \tag{D.1}$$

where $\boldsymbol{\sigma}^l_{j-1,j|k} = E[(\mathbf{h}_{j-1} - \boldsymbol{\mu}_{j-1|k})^{\mathcal{H}}(h_l(j) - \mu^l_{j-1|k})|\mathbf{y}_k]$. The second derivative of $Q_{2,k}(\boldsymbol{\varphi}_2|.)$ with respect to $\mathbf{f}^l$ becomes

$$\left.\frac{\partial^2 Q_{2,k}(\boldsymbol{\varphi}_2|\tilde{\mathbf{a}}_{k|k-1}, \tilde{\boldsymbol{\theta}}_{1,k|k-1}, \tilde{\boldsymbol{\varphi}}_{2|k-1})}{\partial^2 \mathbf{f}^l}\right|_{\boldsymbol{\varphi}_2=\tilde{\boldsymbol{\varphi}}_{2|k-1}} = -\frac{1}{\tilde{R}_{g^l_{|k-1}}} \sum_{j=1}^{k} \tilde{\Sigma}_{j-1,j-1|k} + \tilde{\boldsymbol{\mu}}_{j-1|k}\tilde{\boldsymbol{\mu}}^{\mathcal{H}}_{j-1|k}, \quad 0 \le l \le L. \tag{D.2}$$

In association with (D.1) and (D.2) and also using (21), the recursion formula for estimating the row vector $\mathbf{f}^l$ becomes

$$\tilde{\mathbf{f}}^l_{|k} = \tilde{\mathbf{f}}^l_{|k-1} + \left(\sum_{j=1}^{k} \tilde{\boldsymbol{\sigma}}^l_{j-1,j|k} - \tilde{\mathbf{f}}^l_{|k-1}\tilde{\Sigma}_{j-1,j-1|k-1} + \tilde{\boldsymbol{\mu}}^{\mathcal{H}}_{j-1|k}(\tilde{\mu}^l_{j|k} - \tilde{\mathbf{f}}^l_{|k-1}\tilde{\boldsymbol{\mu}}_{j-1|k})\right) \times \left(\sum_{j=1}^{k} \tilde{\Sigma}_{j-1,j-1|k} + \boldsymbol{\mu}_{j-1|k}\boldsymbol{\mu}^{\mathcal{H}}_{j-1|k}\right)^{-1}, \quad 0 \le l \le L. \tag{D.3}$$

As (D.3) shows, the recursive estimation of $\mathbf{f}^l$ at each recursion needs to calculate the inverse of a matrix. Using a causal procedure (filtering approach) in (D.3) instead of a noncausal procedure (smoothing) and assuming $\tilde{\Sigma}_{j-1,j-1|j} \ll \tilde{\boldsymbol{\mu}}_{j-1|j}\tilde{\boldsymbol{\mu}}^{\mathcal{H}}_{j-1|j}$ and after doing some manipulations, we get (36). Meanwhile, it can be seen from (D.3) that the estimation of $\mathbf{f}^l$ is independent of the estimation of R_{g^l}. The first and the second derivatives of $Q_{2,k}(\boldsymbol{\varphi}_2|.)$ with respect to $R^{-1}_{g^l}$ become

$$\left.\frac{\partial Q_{2,k}\left(\boldsymbol{\varphi}_2|\tilde{\mathbf{a}}_{k|k-1}, \tilde{\boldsymbol{\theta}}_{1,k|k-1}, \tilde{\boldsymbol{\varphi}}_{2|k-1}\right)}{\partial R^{-1^*}_{g^l|k}}\right|_{\boldsymbol{\varphi}_2=\tilde{\boldsymbol{\varphi}}_{2|k-1}} = \left.\left[kR_{g^l|k-1} - \sum_{j=1}^{k} E(h_l(j) - \mathbf{f}^l\mathbf{h}_{j-1})(h_l(j) - \mathbf{f}^l\mathbf{h}_{j-1})^{\mathcal{H}}|\mathbf{y}_k\right]\right|_{\boldsymbol{\varphi}_2=\tilde{\boldsymbol{\varphi}}_{2|k-1}} \tag{D.4}$$

$$\left.\frac{\partial^2 Q_{2,k}\left(\boldsymbol{\varphi}_2|\tilde{\mathbf{a}}_{k|k-1}, \tilde{\boldsymbol{\theta}}_{1,k|k-1}, \tilde{\boldsymbol{\varphi}}_{2|k-1}\right)}{\partial^2 R^{-1}_{g^l|k}}\right|_{\boldsymbol{\varphi}_2=\tilde{\boldsymbol{\varphi}}_{2|k-1}} = -k\tilde{R}^2_{g^l|k-1}. \tag{D.5}$$

By using a filtering approach, assuming $\tilde{\Sigma}_{j-1|j} \ll \tilde{\boldsymbol{\mu}}_{j-1|j}\tilde{\boldsymbol{\mu}}^{\mathcal{H}}_{j-1|j}$ and $\tilde{\Sigma}_{j|j} \ll \tilde{\boldsymbol{\mu}}_{j|j}\tilde{\boldsymbol{\mu}}^{\mathcal{H}}_{j|j}$ for $0 \le j \le k$ and after some manipulations, it can be shown that

$$\tilde{R}_{g^l|k} \simeq \tilde{R}_{g^l|k-1} + \frac{1}{k}\left(\tilde{R}_{g^l|k-1} - |\tilde{\mu}^l_{k|k} - \tilde{\mathbf{f}}^l_{|k-1}\tilde{\boldsymbol{\mu}}_{k-1|k}|^2\right). \tag{D.6}$$

ACKNOWLEDGMENT

The authors would like to thank an anonymous reviewer for Appendix B, which shows the relationship between their algorithm and the SAGE algorithm.

REFERENCES

[1] G. D. Forney, "Maximum-likelihood sequence estimation of digital sequences in presence of intersymbol interference," *IEEE Trans. Inform. Theory,* vol. IT-18, pp. 363–378, May 1972.

[2] J. H. Lodge and M. L. Moher, "Maximum likelihood sequence estimation of CPM signals transmitted over Rayleigh flat-fading channels," *IEEE Trans. Commun.,* vol. 38, pp. 787–794, June 1990.

[3] Q. Dai and E. Shwedyk, "Detection of bandlimited signals over frequency selective Rayleigh fading channels," *IEEE Trans. Commun.,* vol. 42, pp. 941–950, Feb. 1994.

[4] X. Yu and S. Pasupathy, "Innovations-based MLSE for Rayleigh fading channels," *IEEE Trans. Commun.,* vol. 43, pp. 1534–1544, Feb./Mar./Apr. 1995.

[5] H. Zamiri-Jafarian and S. Pasupathy, "Adaptive state allocation algorithm in MLSD receiver for multipath fading channels: Structure and strategy," *IEEE Trans. Veh. Technol.,* vol. 48, pp. 174–187, Jan. 1999.

[6] B. D. Hart and D. P. Taylor, "Maximum-likelihood synchronization, equalization, and sequence estimation for unknown time-varying frequency-selective Rician channels," *IEEE Trans. Commun.,* vol. 46, pp. 211–221, Feb. 1998.

[7] D. P. Taylor, G. M. Vitetta, B. D. Hart, and A. Mammela, "Wireless channel equalization," *European Trans. Telecommun.,* vol. 9, no. 2, pp. 117–143, Mar./Apr. 1998.

[8] S. U. H. Qureshi and E. E. Newhall, "An adaptive receiver for data transmission over time-dispersive channels," *IEEE Trans. Inform. Theory,* vol. IT-19, pp. 448–457, July 1973.

[9] W. U. Lee and F. S. Hill, Jr., "A maximum-likelihood sequence estimator with decision-feedback equalization," *IEEE Trans. Commun.,* vol. COM-25, pp. 971–979, Sept. 1977.

[10] N. Seshadri, "Joint data and channel estimation using blind trellis search techniques," *IEEE Trans. Commun.,* vol. 42, pp. 1000–1011, Feb. 1994.

[11] R. Raheli and A. Polydoros, "Per-survivor processing: A general approach to MLSE in uncertain environments," *IEEE Trans. Commun.,* vol. 43, pp. 354–364, Feb./Mar./Apr. 1995.

[12] K. M. Chugg and A. Polydoros, "MLSE for an unknown channel—Part I: Optimality considerations," *IEEE Trans. Commun.,* vol. 44, pp. 836–846, July 1996.

[13] K. M. Chugg and A. Polydoros, "MLSE for an unknown channel—Part II: Tracking performance," *IEEE Trans. Commun.,* vol. 44, pp. 949–958, Aug. 1996.

[14] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc.,* vol. 39, no. 1, pp. 1–38, 1977.

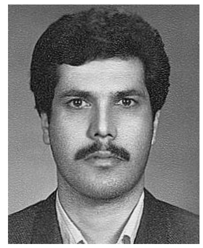
[15] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal Processing Mag.,* vol. 13, pp. 47–60, Nov. 1996.

[16] D. M. Titterington, "Recursive parameter estimation using incomplete data," *J. Roy. Statist. Soc. B,* vol. 46, no. 2, pp. 257–267, 1984.

[17] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in statistical analysis of probabilistic functions of Markov chains," *Annu. Math. Statist.,* vol. 41, pp. 164–171, 1970.

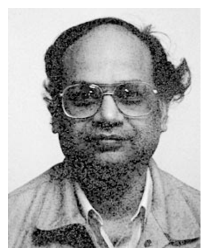
[18] L. E. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic function of Markov processes," *Inequalities,* vol. 3, pp. 1–8, 1972.

[19] G. K. Kaleh and R. Vallet, “Joint parameter estimation and symbol detection for linear or nonlinear unknown channels,” *IEEE Trans. Commun.,* vol. 42, pp. 2406–2413, July 1994.
[20] C. N. Georghiades and J. C. Han, “Sequence estimation in presence of random parameters via the EM algorithm,” *IEEE Trans. Commun.,* vol. 45, pp. 300–308, Mar. 1997.
[21] H. V. Poor, “On parameter estimation in DS/SSMA formats,” *Advances in Communication and Signal Processing.* New York: Springer-Verlag, 1988, pp. 59–70.
[22] S. M. Zabin and H. V. Poor, “Efficient estimation of class A noise parameters via the EM algorithm,” *IEEE Trans. Inform. Theory,* vol. 37, pp. 60–72, Jan. 1991.
[23] A. Ansari and R. Viswanathan, “Application of expectation-maximization algorithm to the detection of direct-sequence signal in pulsed noise jamming,” *IEEE Trans. Commun.,* vol. 41, pp. 1151–1154, Aug. 1993.
[24] H. L. Van Trees, *Detection, Estimation and Modulation Theory; Part I.* New York: Wiley, 1968.
[25] K. Sauer and C. Bouman, “A local update strategy for iterative reconstruction from projections,” *IEEE Trans. Signal Processing,* vol. 41, pp. 534–548, Feb. 1993.
[26] W. Zangwill, *Nonlinear Programming—A Unified Approach.* Englewood Cliffs, NJ: Prentice Hall, 1969.
[27] X. L. Meng and D. B. Rubin, “Maximum likelihood estimation via the ECM algorithm: A general framework,” *Biometrika,* vol. 80, no. 2, pp. 267–278, 1993.
[28] J. A. Fessler and A. O. Hero, “Space-alternating generalized expectation-maximization algorithm,” *IEEE Trans. Signal Processing,* vol. 42, pp. 2664–2677, Oct. 1994.
[29] J. A. Fessler, N. H. Clinthorne, and W. L. Rogers, “On complete-data space for PET reconstruction algorithm,” *IEEE Trans. Nucl. Sci.,* vol. 40, pp. 1055–1061, Aug. 1993.
[30] M. J. Omidi, P. G. Gulak, and S. Pasupathy, “Joint data and channel estimation using the per-branch processing method,” in *Proc. IEEE Signal Processing Workshop on Signal Processing Advances in Wireless Communications,* Paris, France, Apr. 1997, pp. 384–392.
[31] A. Duel-Hallen and C. Heegard, “Delayed decision-feedback sequence estimation,” *IEEE Trans. Commun.,* vol. 37, pp. 428–436, May 1989.
[32] H. Zamiri-Jafarian and S. Pasupathy, “EM-based recursive estimation of channel parameters,” *IEEE Trans. Commun.,* to be published.
[33] ——, “Recursive channel estimation for wireless communication via the EM algorithm,” in *Proc. Conf. ICPWC’97,* Bombay, India, Dec. 1997, pp. 33–37.
[34] H. Zamiri-Jafarian, “Adaptive MLSDE receivers for wireless communications,” Ph.D. dissertation, Univ. of Toronto, Toronto, Ontario, Canada, 1998.
[35] S. Haykin, *Adaptive Filter Theory,* 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1991.
[36] M. E. Rollins and S. J. Simmons, “Simplified per-survivor Kalman processing in fast frequency-selective fading channels,” *IEEE Trans. Commun.,* vol. 45, pp. 544–553, May 1997.
[37] B. Hart, “MLSE diversity receiver structures,” Ph.D. thesis, University of Canterbury, Christchurch, New Zealand, 1996.
[38] W. H. A. Guston and E. H. Walker, “A multipath fading simulator for radio,” *IEEE Trans. Veh. Technol.,* vol. VT-22, pp. 241–244, Nov. 1973.
[39] J. G. Proakis, *Digital Communications,* 2nd ed. New York: McGraw-Hill, 1989.
[40] A. Masoomzadeh-Fard and S. Pasupathy, “Combined equalization and differential detection using precoding,” *IEEE Trans. Commun.,* vol. 45, pp. 274–278, Mar. 1997.



Hossein Zamiri-Jafarian (M’98) was born in Mashhad, Iran, in 1960. He received the B.Sc. and M.S. degrees from Isfahan University of Technology, Isfahan, Iran, in 1985 and 1987, respectively, and the Ph.D. degree from the University of Toronto, Toronto, Canada, in 1998, all in electrical engineering.

He joined Ferdowsi University, Mashhad, Iran, in 1987. Currently, he works as a Research Associate in the Department of Electrical and Computer Engineering, University of Toronto. His current research interests include statistical signal processing, data communications, and multirate signal processing.



Subbarayan Pasupathy (M’73–SM’81–F’91) was born in Chennai (Madras), India, on September 21, 1940. He received the B.E. degree in telecommunications from the University of Madras in 1963, the M.Tech. degree in electrical engineering from the Indian Institute of Technology, Madras, India, in 1966, and the M.Phil. and Ph.D. degrees in engineering and applied science from Yale University, New Haven, CT, in 1970 and 1972, respectively.

He joined the faculty of the University of Toronto in 1973 and became a Professor of electrical engineering in 1983. He has served as the Chairman of the Communications Group and as the Associate Chairman of the Department of Electrical Engineering at the University of Toronto. His research interests are in the areas of communication theory, digital communications, and statistical signal processing.

Dr. Pasupathy is a Registered Professional Engineer in the Province of Ontario. He has served as a Technical Associate Editor for the IEEE Communications Magazine (1979–1982) and as an Associate Editor for the *Canadian Electrical Engineering Journal* (1980–1983). During 1982–1989, he was an Area Editor for *Data Communications and Modulation* for the IEEE TRANSACTIONS ON COMMUNICATIONS. Since 1984, he has been writing a regular column entitled “Light Traffic” for the *IEEE Communications Magazine.* Dr. Pasupathy was elected as a Fellow of the IEEE in 1991 “for contributions to bandwidth efficient coding and modulation schemes in digital communication.”