The Minimum Description Length Principle for Modeling Recording Channels

Aleksandar Kavčić and Murari Srinivasan

A. Kavčić is with the Division of Engineering and Applied Sciences, Harvard University, Cambridge, MA.
M. Srinivasan is with Flarion Technologies, Bedminster, NJ.
This work was supported by the National Science Foundation under Grant No. CCR-9904458 and by the National Storage Industry Consortium.

Abstract

Modeling the magnetic recording channel has long been a challenging research problem. Typically the tradeoff has been simplicity of the model for its accuracy. For a given family of channel models, the accuracy will grow with the model size, at a price of a more complex model. In this paper, we develop a formalism that strikes a balance between these opposing criteria. The formalism is based on Rissanen's notion of minimum required complexity - the minimum description length (MDL). The family of channel models in this study is the family of signal-dependent autoregressive channel models chosen for its simplicity of description and experimentally verified modeling accuracy. For this family of models, the minimum description complexity is directly linked to the minimum required complexity of a detector. Furthermore, the minimum description principle for autoregressive models lends itself for an intuitively pleasing interpretation. The description complexity is the sum of two terms: 1) the entropy of the sequence of uncorrelated Gaussian random variables driving the autoregressive filters, which decreases with the model order (i.e., model size) and 2) a penalty term proportional to the model size. We exploit this interpretation to formulate the minimum description length criterion for the magnetic recording channel corrupted by nonlinearities and signal-dependent noise. Results on synthetically generated data are presented to validate the method. We then apply the method to data collected from the spin stand to establish the model's size and parameters that strike a balance between complexity and accuracy.

Keywords

magnetic recording channel, minimum description length, autoregressive processes, signal-dependent noise, nonstationary noise, maximum likelihood estimation.

I. Introduction

As areal densities of magnetic recording products continue to increase, magnetic recording channels are becoming dominated by magnetic media noise (also referred to as jitter noise, transition noise or signal-dependent noise) [1], [2], [3], [4], [5]. Predictive studies of future magnetic recording media indicate that his trend is not likely to reverse [6]. Another effect that often accompanies magnetic recording channels is nonlinearity, either due to partial signal erasure or magnetoresistive head nonlinearities [7].

In recent years, considerable research efforts have been devoted to modeling the magnetic recording channel, i.e. channel nonlinearities and media noise [8], [9], [10], [11], [12]. The benefits expected from these efforts are 1) to have a realistic model for creating waveforms in channel simulations and 2) to aid the design of detectors matched to the media noise characteristics and channel nonlinearities. Recently, an autoregressive model for the magnetic recording channel

was proposed in [13]. This model has been the basis for designing the maximum likelihood sequence detector (Viterbi detector) for signal-dependent noise [14]. Recent comparative studies conducted on spin-stand waveforms and waveforms generated by the autoregressive model show that the model is realistic not only in its capability to match the second order statistics of the spin-stand waveforms, but also to match, 90% of the error event types [15] on the average.

The importance of accurate modeling of spin-stand waveforms is further underlined by the advances made in soft-output symbol detection coupled with iterative decoding strategies for turbo codes [16], [17], [18], [19], [20] and Gallager codes (low-density parity-check codes) [21], [22], [23], [24]. It has been shown in [25] that the autoregressive magnetic recording channel model is also suited for symbol-wise soft-output detection (BCJR detection [26]). In [27], it was experimentally shown that, when coupled with an iterative decoder of low-density parity-check codes, the gain obtained from matching the detector to the characteristic of the signal-dependent noise is independent of the coding gain. The total detection gain over signal-*independent* detectors is the sum of the two gains, underlining equal importance of channel matching (i.e., modeling) and code design.

This paper presents a method for modeling the magnetic recording channel. The chosen channel model is essentially the model presented in [13]. Here, however, we use the extended model where the intersymbol interference window may be shifted with respect to the filter-dependence window of the noise-coloring filter. In [13], the window size and the noise memory length were vastly overestimated to ensure accurate parameter estimation. Here we take a different approach - we want to accurately estimate both the model parameters (intersymbol interference levels and the coefficients of the noise-coloring filter) and the model size (signal-dependence window sizes and the noise memory length). While the former can be solved with standard maximum likelihood parameter estimation techniques, the latter can not be. To find an estimate for the model size, we modify Rissanen's minimum description length (MDL) principle [28] to fit the signal-dependent autoregressive channel model.

Applying Rissanen's MDL principle to modeling *stationary* autoregressive processes leads to an intuitively pleasing interpretation of the stochastic complexity. The stochastic complexity is the sum of two terms. 1) The first term is the entropy (uncertainty) of the estimated white

Gaussian noise that drives the autoregressive filter

$$\frac{1}{2}\sum_{k=1}^{N}\ln\left[2\pi\mathrm{e}\cdot\hat{\sigma}_{ML}^2\right], \tag{1}$$

where $\hat{\sigma}_{ML}^2$ is the maximum likelihood estimate of the variance of the noise and N is the length of the sequence from which $\hat{\sigma}_{ML}^2$ is estimated; the term in (1) falls with an increase in estimated model size. 2) The second term is a complexity penalty factor that grows proportionally to the model size.

The difficulty in finding the stochastic complexity formulation for the signal-*dependent* (i.e., nonstationary) channel model comes from the intractability of the maximum likelihood equations for the model estimate, as we show in this paper. We therefore propose an approximate maximum-likelihood method to estimate the parameters of the model. To formulate the stochastic complexity, we rely on these approximate maximum likelihood estimates and produce an entropy term for the noise driving the signal-dependent autoregressive filter, while the penalty term is proportional to the model size. Since several approximations are used, we cannot claim to have developed a rigorous analog of Rissanen's MDL criterion for the signal-dependent case, but we present simulation results on synthetically generated waveforms that provide experimental verification of the proposed method. We then apply the method to spin stand data to find the minimum description of the recording channel.

The paper is organized as follows. In Section II, we briefly describe the signal-dependent autoregressive channel model. In Section III, the equations for maximum likelihood parameter estimates are derived. Due to the intractability of these equations, we propose an approximate maximum-likelihood (ML) parameter estimation technique based on the exact ML equations. A brief exposition of Rissanen's minimum description length (MDL) principle is presented in Section IV, with an emphasis on applying the principle for *stationary* autoregressive model size estimation. We show that the model size estimate is related to the entropy of the noise driving the autoregressive filter [29], [30], the number of parameters in the model and the number of data windows used to estimate each parameter. With this interpretation, in Section V we formulate the approximate MDL criterion for the signal-dependent autoregressive channel model. Section VI presents the estimation results for both artificially created waveforms with known parameters and for spin-stand waveforms where the parameters are not known. The conclusions follow in Section VII.

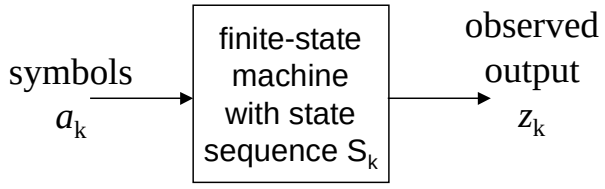


Fig. 1. Block diagram of the discrete-time magnetic recording channel. Binary input symbols a_k determine the real-valued random sequence z_k that is the output of a finite-state machine.

The following notation is used throughout the paper. Underlined characters (i.e., $\underline{z}$) represent column vectors. Boldface characters ($\mathbf{C}$) represent matrices (typically square matrices in this paper). The superscript $^{\mathrm{T}}$ stands for matrix and vector transposition. Further notation is introduced in the paper as needed.

II. The Signal-Dependent Autoregressive Channel Model

We consider a discrete-time model of a finite-state machine recording channel, Figure 1. The input sequence to the channel consists of binary symbols $a_k \in \{0, 1\}$, where k is an integer denoting discrete time. A subsequence of symbols a_k, ranging from indices $k = i$ to $k = j$ $(j \geq i)$, is denoted by the column vector

$$\underline{a}_j^i = \begin{bmatrix} a_i \\ a_{i+1} \\ \vdots \\ a_j \end{bmatrix}.$$

The output sequence of the channel consists of symbol-interval samples $z_k \in \mathbb{R}$. A subsequence of symbols z_k, ranging from indices $k = i$ to $k = j$ $(j \geq i)$, is denoted by the column vector

$$\underline{z}_j^i = \begin{bmatrix} z_i \\ z_{i+1} \\ \vdots \\ z_j \end{bmatrix}.$$

Here we use a very specific model for the finite-state machine. First, we represent the channel output as

$$z_k = y\left(\underline{a}_{k+I_2}^{k+I_1}\right) + n_k, \tag{2}$$

denoting that the deterministic part of the channel output $y\left(\underline{a}_{k+I_2}^{k+I_1}\right) \in \mathbb{R}$ depends on the input symbol window $\underline{a}_{k+I_2}^{k+I_1}$, while n_k is an additive noise component. Here, we assume that $I_2 \geq I_1$. The function $y\left(\underline{a}_{k+I_2}^{k+I_1}\right)$ is a table look-up, with $I_2 - I_1 + 1$ binary variables determining a single real-valued output. A common practice in *linear* channel modeling is to specify $y\left(\underline{a}_{k+I_2}^{k+I_1}\right)$ as a convolution between the input sequence and a channel polynomial $y\left(\underline{a}_{k+I_2}^{k+I_1}\right) = \sum_{i=0}^{I_2-I_1} h_i a_{k+I_2-i}$, where h_i are the coefficients of the linear time-invariant channel response. Since a magnetic recording channel is typically a *nonlinear* channel, we do not use the convolution sum, but rather stay with the general table look-up form $y\left(\underline{a}_{k+I_2}^{k+I_1}\right)$.

The additive noise component n_k in (2) is modeled as the output of an L-th order signal-dependent autoregressive (AR) filter. The L-th order autoregressive stochastic difference equation guiding the evolution of n_k is assumed to be

$$n_k = \underline{b}\left(\underline{a}_{k+D_2}^{k+D_1}\right)^{\mathrm{T}} \cdot \underline{n}_{k-1}^{k-L} + \sigma\left(\underline{a}_{k+D_2}^{k+D_1}\right) \cdot w_k. \tag{3}$$

Here the vector $\underline{b}\left(\underline{a}_{k+D_2}^{k+D_1}\right)$ is a collection of $L \geq 0$ autoregressive coefficients

$$\underline{b}\left(\underline{a}_{k+D_2}^{k+D_1}\right) = \begin{bmatrix} b_L\left(\underline{a}_{k+D_2}^{k+D_1}\right) \\ b_{L-1}\left(\underline{a}_{k+D_2}^{k+D_1}\right) \\ \vdots \\ b_1\left(\underline{a}_{k+D_2}^{k+D_1}\right) \end{bmatrix}. \tag{4}$$

The coefficients in the vector $\underline{b}\left(\underline{a}_{k+D_2}^{k+D_1}\right)$ are dependent on a window of input symbols $\underline{a}_{k+D_2}^{k+D_1}$. We refer to this window as the noise shaping filter dependence window. In (3), w_k is a sequence of zero-mean independent identically distributed (iid) random variables, each with variance $\mathrm{E}\left[w_k^2\right] = 1$. The coefficient $\sigma\left(\underline{a}_{k+D_2}^{k+D_1}\right)$ that multiplies w_k in (3), referred to as the signal-dependent standard deviation, depends on the same noise shaping filter dependence window $\underline{a}_{k+D_2}^{k+D_1}$ as the autoregressive coefficients. The dependence of the noise shaping filter coefficients on the input sequence a_k assures modeling of the signal-dependent character of media noise in magnetic recording. The full channel model is depicted in Figure 2. Note that the signal-dependence window $\underline{a}_{k+I_2}^{k+I_1}$ and the noise shaping filter dependence window $\underline{a}_{k+D_2}^{k+D_1}$ are not the same to allow for a general signal-dependent channel with correlated noise.

If we require the driving noise w_k to be Gauss-distributed, in addition to being a zero-mean unit-variance iid process, the noise process n_k defined in (3) is a Gauss-*Markov* process. Gauss-

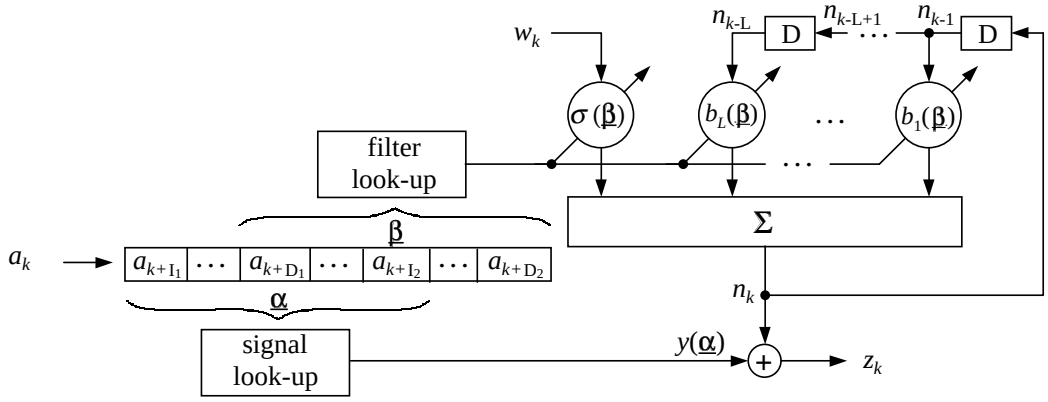


Fig. 2. Autoregressive channel model. The symbol window $\underline{a}_{k+I_2}^{k+I_1}$ determines the channel output $y\left(\underline{a}_{k+I_2}^{k+I_1}\right)$, while a different symbol window $\underline{a}_{k+D_2}^{k+D_1}$, determines the noise shaping filter coefficients $\sigma\left(\underline{a}_{k+D_2}^{k+D_1}\right)$, $b_1\left(\underline{a}_{k+D_2}^{k+D_1}\right)$, ..., $b_L\left(\underline{a}_{k+D_2}^{k+D_1}\right)$.

Markov processes are highly structured, with several advantageous properties. First, notice that the model in (3) is *acausal*, suggesting that this model may not be general enough. However, if the process is Gauss-*Markov*, then it can be shown that even an *acausal* process can be represented as a *causal*, but non-stationary Gauss-Markov process [31]. With that in mind, equation (3) is in fact the most general representation of a signal-dependent Gauss-Markov noise process. Second, it was shown in [14] that the most general channel model for which a finite-complexity optimal detector can be formulated is in fact a channel with a finite Markov memory. Due to the Gauss assumption for w_k, the channel represented in (2) and (3) is also *Markov*, and therefore admits a finite-complexity implementation of the Maximum likelihood sequence detector (MLSD). This detector is a Viterbi detector with $2^{C_2-C_1}$ states, where $C_2 = \max[I_2, D_2]$ and $C_1 = \min[I_1 - L, D_1]$, as demonstrated in [14]. The trellis structure of this detection approach can be adapted towards a soft-output detector (BCJR detector [26]) as shown in [25] and in [27] where the soft-output detector is used in concatenation with an iterative decoder of Gallager codes [21].

The remaining question is whether this model is a good representation of the magnetic recording channel. The model in (2) and (3) captures nonlinearities and the signal-dependent character of noise correlation. However, it does not capture non-Gaussianity of the noise. Depending on the degree of non-Gaussianity in the channel noise, the model will misrepresent the channel. In that case, the model will only capture the first and second order statistics of the noise. Usually,

a match up to the second order statistics is good enough because detectors are typically built to match these statistics [14], [27]; matching higher order statistics usually overly increases the detector complexity for only a marginally small gain in performance. Indeed, experimental results show that the autoregressive channel model is capable of representing a recording channel with an accuracy that matches not just the second order statistics, but also the error event statistics at the output of a Viterbi detector, as reported in [15].

III. Parameter Estimation

Fitting an autoregressive (AR) channel model of the kind in (2) and (3) can be viewed as a two-step process. First, we need to find the model size, that is, the signal-dependent window size variables I_1 and I_2, the noise filter shaping window size variables D_1 and D_2, and the Markov memory length L. Second, for a determined model size (I_1, I_2, D_1, D_2, L), we need to determine the model parameters $y\,(\underline{\alpha})$, $\sigma\big(\underline{\beta}\big)$, and the $L \times 1$ vectors $\underline{b}\big(\underline{\beta}\big)$, for every binary column vector $\underline{\alpha}$ of size $(I_2 - I_1 + 1) \times 1$ and for every binary column vector $\underline{\beta}$ of size $(D_2 - D_1 + 1) \times 1$.

While these two steps may seem independent, we show in Section IV that they are very much interdependent. For the presentation purposes, however, we first deal with the problem of estimating the model parameters given that a suitable model size has been chosen. We formulate the problem as follows. Assume that the model size parameters I_1, I_2, D_1, D_2 and L are known. For these model-size parameters, find the optimal model parameter estimates $\hat{y}(\underline{\alpha})$, $\hat{\sigma}\big(\underline{\beta}\big)$, and $\hat{\underline{b}}\big(\underline{\beta}\big)$. Depending on the optimality criterion we choose, we get different estimates.

A. Maximum likelihood model parameter estimation

We next use the maximum-likelihood criterion to derive the equations for the optimal parameter estimates. For brevity of notation, denote by F the joint conditional probability density function of a vector of N (N very large) channel-outputs $\underline{z}_N^1$, conditioned on the knowledge of a sequence of input symbols $\underline{a}_{1+C_1}^{N+C_2}$ and a sequence of L prior outputs $\underline{z}_0^{-L+1}$

$$F = f\left(\underline{z}_N^1 \;\mid\; \underline{a}_{N+C_2}^{1+C_1}\,,\; \underline{z}_0^{-L+1}\right), \tag{5}$$

where $C_2 = \max\left[I_2, D_2\right]$, $C_1 = \min\left[I_1 - L, D_1\right]$, while the model size parameters I_1, I_2, D_1, D_2 and L are assumed to be known. The maximum likelihood criterion finds those values $y(\underline{\alpha})$, $\sigma\big(\underline{\beta}\big)$, and $\underline{b}\big(\underline{\beta}\big)$ that maximize the density function F in (5), or equivalently, that minimize the negative logarithm of the function F in (5). Due to the Gauss-Markov assumption of the channel

noise, we have that the negative logarithm of F is given by

$$-\ln F = \frac{1}{2}\sum_{k=1}^{N}\left\{\ln\left[2\pi\cdot\sigma^2\left(\underline{a}_{k+D_2}^{k+D_1}\right)\right] + \frac{\left(\left[\underline{z}_k^{k-L} - \underline{y}\left(\underline{a}_{k+I_2}^{k+I_1-L}\right)\right]^{\mathrm{T}}\begin{bmatrix} -\underline{b}\left(\underline{a}_{k+D_2}^{k+D_1}\right) \\ 1 \end{bmatrix}\right)^2}{\sigma^2\left(\underline{a}_{k+D_2}^{k+D1}\right)}\right\}, \quad (6)$$

where $\underline{y}\left(\underline{a}_{k+I_2}^{k+I_1-L}\right)$ is short notation for

$$\underline{y}\left(\underline{a}_{k+I_2}^{k+I_1-L}\right) = \begin{bmatrix} y\left(\underline{a}_{k+I_2-L}^{k+I_1-L}\right) \\ y\left(\underline{a}_{k+I_2-L+1}^{k+I_1-L+1}\right) \\ \vdots \\ y\left(\underline{a}_{k+I_2}^{k+I_1}\right) \end{bmatrix}.$$

Finding the partial derivatives of (6) with respect to $y(\underline{\alpha})$, $\sigma\left(\underline{\beta}\right)$ and $\underline{b}\left(\underline{\beta}\right)$, setting them equal to zero and substituting $y(\underline{\alpha})$, $\sigma\left(\underline{\beta}\right)$ and $\underline{b}\left(\underline{\beta}\right)$ with their maximum-likelihood estimates $\hat{y}_{\mathrm{ML}}(\underline{\alpha})$, $\hat{\sigma}_{\mathrm{ML}}\left(\underline{\beta}\right)$, and $\hat{\underline{b}}_{\mathrm{ML}}\left(\underline{\beta}\right)$, we get the following equations for the maximum-likelihood parameter estimates for our AR channel model

$$\begin{aligned} 0 \;=\; & \sum_{k\,:\,\underline{a}_{k+I_2}^{k+I_1}=\underline{\alpha}} \frac{1}{\hat{\sigma}^2_{\mathrm{ML}}\left(\underline{a}_{k+D_2}^{k+D_1}\right)}\left[\underline{z}_k^{k-L} - \hat{\underline{y}}_{\mathrm{ML}}\left(\underline{a}_{k+I_2}^{k+I_1-L}\right)\right]^{\mathrm{T}}\begin{bmatrix} -\hat{\underline{b}}_{\mathrm{ML}}\left(\underline{a}_{k+D_1}^{k+D_2}\right) \\ 1 \end{bmatrix} \\ & - \sum_{i=1}^{L}\sum_{k\,:\,\underline{a}_{k+I_2-i}^{k+I_1-i}=\underline{\alpha}} \frac{\hat{b}_{i_{\mathrm{ML}}}\left(\underline{a}_{k+D_1}^{k+D_2}\right)}{\hat{\sigma}^2_{\mathrm{ML}}\left(\underline{a}_{k+D_2}^{k+D_1}\right)}\left[\underline{z}_k^{k-L} - \hat{\underline{y}}_{\mathrm{ML}}\left(\underline{a}_{k+I_2}^{k+I_1-L}\right)\right]^{\mathrm{T}}\begin{bmatrix} -\hat{\underline{b}}_{\mathrm{ML}}\left(\underline{a}_{k+D_2}^{k+D_1}\right) \\ 1 \end{bmatrix} \quad (7) \end{aligned}$$

$$0 \;=\; \hat{\sigma}^2_{\mathrm{ML}}\left(\underline{\beta}\right) - \frac{1}{N_{\underline{\beta}}}\cdot\sum_{k\,:\,\underline{a}_{k+D_2}^{k+D_1}=\underline{\beta}}\left(\left[\underline{z}_k^{k-L} - \hat{\underline{y}}_{\mathrm{ML}}\left(\underline{a}_{k+I_2}^{k+I_1-L}\right)\right]^{\mathrm{T}}\begin{bmatrix} -\hat{\underline{b}}_{\mathrm{ML}}\left(\underline{a}_{k+D_2}^{k+D_1}\right) \\ 1 \end{bmatrix}\right)^2 \quad (8)$$

$$\underline{0} \;=\; \sum_{k\,:\,\underline{a}_{k+D_2}^{k+D_1}=\underline{\beta}}\left[\underline{z}_{k-1}^{k-L} - \hat{\underline{y}}_{\mathrm{ML}}\left(\underline{a}_{k+I_2-1}^{k+I_1-L}\right)\right]\left[\underline{z}_k^{k-L} - \hat{\underline{y}}_{\mathrm{ML}}\left(\underline{a}_{k+I_2}^{k+I_1-L}\right)\right]^{\mathrm{T}}\begin{bmatrix} -\hat{\underline{b}}_{\mathrm{ML}}\left(\underline{a}_{k+D_2}^{k+D_1}\right) \\ 1 \end{bmatrix}. \quad (9)$$

Here, the notation

$$\sum_{k\,:\,\underline{a}_{k+D_2}^{k+D_1}=\underline{\beta}}$$

denotes that the summation is done over all k such that the signal-dependent window of symbols $\underline{a}_{k+D_2}^{k+D_1}$ equals the binary vector $\underline{\beta}$. The number $N_{\underline{\beta}}$ denotes the number of occurrences of a binary

pattern $\underline{\beta}$ within the sequence $\underline{a}_{N+C_2}^{1+C_1}$.

Equations (8) and (9) can be further manipulated. Define the $(L+1)\times(L+1)$ empirical maximum-likelihood covariance matrix as

$$\hat{\mathbf{C}}_{\mathrm{ML}}\left(\underline{\beta}\right) = \frac{1}{N_{\underline{\beta}}} \cdot \sum_{k\,:\,\underline{a}_{k+D_2}^{k+D_1}=\underline{\beta}} \left[\underline{z}_k^{k-L} - \underline{\hat{y}}_{\mathrm{ML}}\left(\underline{a}_{k+I_2}^{k+I_1-L}\right)\right]\left[\underline{z}_k^{k-L} - \underline{\hat{y}}_{\mathrm{ML}}\left(\underline{a}_{k+I_2}^{k+I_1-L}\right)\right]^{\mathrm{T}}. \tag{10}$$

Next, partition the empirical maximum-likelihood covariance matrix as

$$\hat{\mathbf{C}}_{\mathrm{ML}}\left(\underline{\beta}\right) = \begin{bmatrix} \hat{\mathbf{c}}_{\mathrm{ML}}\left(\underline{\beta}\right) & \underline{\hat{c}}_{\mathrm{ML}}\left(\underline{\beta}\right) \\ \underline{\hat{c}}_{\mathrm{ML}}\left(\underline{\beta}\right)^{\mathrm{T}} & \hat{c}_{0_{\mathrm{ML}}}\left(\underline{\beta}\right) \end{bmatrix}, \tag{11}$$

where $\hat{\mathbf{c}}_{\mathrm{ML}}\left(\underline{\beta}\right)$ is the $L\times L$ principal minor of $\hat{\mathbf{C}}_{\mathrm{ML}}\left(\underline{\beta}\right)$. With this notation, after some manipulation, it can be shown that equations (8) and (9) can be expressed as

$$\hat{\sigma}^2_{\mathrm{ML}}\left(\underline{\beta}\right) = \frac{\det \hat{\mathbf{C}}_{\mathrm{ML}}\left(\underline{\beta}\right)}{\det \hat{\mathbf{c}}_{\mathrm{ML}}\left(\underline{\beta}\right)} \tag{12}$$

$$\underline{\hat{b}}_{\mathrm{ML}}\left(\underline{\beta}\right) = \hat{\mathbf{C}}_{\mathrm{ML}}\left(\underline{\beta}\right)^{-1} \cdot \underline{\hat{c}}_{\mathrm{ML}}\left(\underline{\beta}\right). \tag{13}$$

These are exactly the solutions to the nonstationary Yule-Walker equations (normal equations) when instead of the true covariance matrices, we use the empirical covariance matrices.

Finding the maximum-likelihood estimates involves solving equations (7), (12) and (13) simultaneously. These equations are interdependent and separating the solutions has thus far evaded us. We have not been able to solve these equations analytically, which is why we propose a less rigorous, but straight-forward method for parameter estimation.

B. A practical approach to model parameter estimation

Notice that if we have "good" guesses for the signal-dependent noiseless channel outputs $y(\underline{\alpha})$, we could use them in the expressions for the empirical covariance matrix in (10) and thus obtain the maximum likelihood estimates for the noise filter parameters in (12) and (13). However, as shown in the previous subsection, obtaining universally accepted "good" guesses - the maximum likelihood estimates for $\hat{y}_{\mathrm{ML}}(\underline{\alpha})$ - is very hard due to the complexity of equation (7) and its dependence on $\hat{\sigma}^2_{\mathrm{ML}}\left(\underline{\beta}\right)$ and $\underline{\hat{b}}_{\mathrm{ML}}\left(\underline{\beta}\right)$, which cannot be computed independent of $\hat{y}_{\mathrm{ML}}(\underline{\alpha})$. To circumvent this problem, we use the empirical mean (obtained under the noise energy minimization

criterion) as our estimates for the signal-dependent noiseless channel outputs

$$\hat{y}_{\mathrm{EM}}(\underline{\alpha}) = \frac{1}{N_{\underline{\alpha}}} \sum_{k\,:\,\underline{a}_{k+I_2}^{k+I_1}=\underline{\alpha}} z_k. \tag{14}$$

If instead of $\hat{y}_{\mathrm{ML}}(\underline{\alpha})$, we use $\hat{y}_{\mathrm{EM}}(\underline{\alpha})$ in (10), we can proceed to find the maximum-likelihood estimates $\hat{\sigma}^2_{\mathrm{ML}}\big(\underline{\beta}\big)$ and $\hat{\underline{b}}_{\mathrm{ML}}\big(\underline{\beta}\big)$ using equations (12) and (13). Rigorously speaking, the so obtained estimates for $\sigma\big(\underline{\beta}\big)$ and $\underline{b}\big(\underline{\beta}\big)$ should *not* be called maximum likelihood estimates since they were obtained using only 2 out of 3 maximum likelihood equations (equation (7) was circumvented by (14)). However, it is easy to show that the estimate in (14) is unbiased. Further, for a large sample size N, the estimate $\hat{y}_{\mathrm{EM}}(\underline{\alpha})$ in (14) will be very close to the true value $y(\underline{\alpha})$. In that sense the estimates for $\sigma\big(\underline{\beta}\big)$ and $\underline{b}\big(\underline{\beta}\big)$, which we obtain by substituting $\hat{y}_{\mathrm{EM}}(\underline{\alpha})$ for $\hat{y}_{\mathrm{ML}}(\underline{\alpha})$ in (10) can be regarded as approximately maximum likelihood. This way of computing the estimates is computationally more efficient, because we can break the problem into two steps: 1) compute $\hat{y}_{\mathrm{EM}}(\underline{\alpha})$ and 2) compute $\hat{\sigma}_{\mathrm{ML}}\big(\underline{\beta}\big)$ and $\hat{\underline{b}}_{\mathrm{ML}}\big(\underline{\beta}\big)$; this is opposed to simultaneously solving equations (7), (12) and (13).

In the remainder of the paper, when we refer to the maximum likelihood parameter estimation, it should be understood that we are actually referring to the practical method described in this subsection. Next, we turn our attention next to estimating the model-size parameters I_1, I_2, D_1, D_2 and L, which until now we have assumed to be known a priori.

IV. The Minimum Description Length (MDL) Principle

A. Basic MDL theory

In this section, we provide an overview of the Minimum Description Length (MDL) principle for model selection and statistical inference. Consider a vector of observations $\underline{z}_N^1 = [z_1, z_2, ..., z_N]^{\mathrm{T}}$. A fundamental question in statistical modeling is to determine an appropriate statistical model for these observations. There is a well-established tradeoff between the complexity of the model chosen and the goodness-of-fit that results. In the case of the ubiquitous maximum-likelihood principle, optimal model parameters may be estimated only when the size of the model is known. Akaike [32] was the first to establish criteria that allowed the estimation of the model size along with the model parameters. The MDL principle was developed independently by Schwartz [33] and Rissanen [34], [35], where Rissanen also showed that the classical problem of optimal model fitting could be placed in the context of universal coding. The MDL principle does not assume

any "true" data generating distribution, but searches for good probability models for the data. There are two central ideas of this principle. One is to represent an entire family of probability distributions by a single universal representative model. The other is to use the notion of codeword length relative to a model as the metric that determines the choice of the optimal model. These two ideas facilitate the selection of the best model for the observed data from a universal class. In particular, Rissanen showed that the optimal model chosen from a parametric family of models is that model which permits the shortest codeword length description of the available data. This has an elegant and intuitive interpretation in the light of the duality between probability distributions and optimal code lengths, as was demonstrated by Shannon [36].

Consider a family of models parameterized by a vector $\underline{\theta}(L) \in \mathbb{R}^{L+1}$, where L is an integer $0 \leq L < \infty$. Denote by Θ a family of models under consideration. Let $\Theta(L)$ represent the class of all models in the family Θ, such that the $(L+1)$-dimensional vector $\underline{\theta}(L)$ is a member of the class $\Theta(L)$, i.e., $\Theta(L)$ is the set $\Theta(L) = \left\{\underline{\theta}(L) : \underline{\theta}(L) \in \mathbb{R}^{L+1}\right\}$. Consequently, Θ is the union of all $\Theta(L)$, i.e., $\Theta = \bigcup_{L \geq 0} \Theta(L)$.

To each vector $\underline{\theta}(L)$, we assign a probability density function of the observed sequence $\underline{z}_N^1$, $f\left(\underline{z}_N^1 \mid \underline{\theta}(L)\right)$, conditioned on the assumption that the vector $\underline{z}_N^1$ is created by the model parameterized by the vector $\underline{\theta}(L)$. That is, we assume that the vector observation $\underline{z}_N^1$ is drawn from the probability density function parameterized by the parameter vector $\underline{\theta}(L)$. The maximum likelihood estimate $\underline{\hat{\theta}}_{\mathrm{ML}}(L)$ of $\underline{\theta}(L) \in \Theta(L)$ captured from the observation vector $\underline{z}_N^1$ is

$$\underline{\hat{\theta}}_{\mathrm{ML}}(L) = \arg \min_{\underline{\theta}(L) \in \Theta(L)} \left\{ -\ln f\left(\underline{z}_N^1 \mid \underline{\theta}(L)\right)\right\} \tag{15}$$

Given the knowledge of $\underline{\hat{\theta}}_{\mathrm{ML}}(L)$, an optimal code for the sequence can be constructed from $f\left(\underline{z}_N^1 \mid \underline{\hat{\theta}}_{\mathrm{ML}}(L)\right)$. Shannon [36] proved that an optimal code could be constructed that has a codeword length

$$-\ln f\left(\underline{z}_N^1 \mid \underline{\hat{\theta}}_{\mathrm{ML}}(L)\right).$$

Therefore, one way of communicating the observation vector to a receiver is to first transmit a preamble that contains a quantized description of $\underline{\hat{\theta}}_{\mathrm{ML}}(L)$ and then the codeword of length $-\ln f\left(\underline{z}_N^1 \mid \underline{\hat{\theta}}_{\mathrm{ML}}(L)\right)$. The total cost (codeword length) of this two-part description was shown

by Rissanen [37] to be

$$\mathcal{C}(L) = -\ln f\left(\underline{z}_N^1 \mid \hat{\underline{\theta}}_{\mathrm{ML}}(L) \right) + \frac{L+1}{2} \ln N \tag{16}$$

where the second term represents the penalty associated with communicating the quantized maximum-likelihood estimate $\hat{\underline{\theta}}_{\mathrm{ML}}(L)$ to the receiver. While different penalty terms can be used [32], [28], the penalty term $\frac{L+1}{2} \ln N$ in (16) was shown by Rissanen [37] to be optimal for large observation lengths N.

Model selection according to the MDL principle now simply becomes a matter of evaluating the complexity cost in Equation (16) for different model orders. Therefore, the MDL estimate of the model order is

$$\hat{L}_{\mathrm{MDL}} = \arg \min_{L \geq 0} \mathcal{C}(L) \,. \tag{17}$$

Rissanen showed that model selection carried out through the MDL principle results in the most efficient coding of the observation sequence among all universal codes [37].

B. MDL principle for stationary autoregressive processes

The MDL principle can be applied in a straight-forward way to *stationary* autoregressive (AR) processes [33], [34]. We replicate this derivation for reference in Section V where we present an MDL criterion for *signal-dependent* autoregressive processes.

Consider an observation vector of length N, $\underline{z}_N^1 = [z_1, z_2, ..., z_N]^{\mathrm{T}}$. We are required to choose an AR model from a parametric family of models

$$\begin{aligned} \Theta &= \{\underline{\theta}(L), L \geq 0\} \\ \text{where } \underline{\theta}(L) &= \begin{bmatrix} \sigma(L) \\ \underline{b}(L) \end{bmatrix} = \begin{bmatrix} \sigma(L) \\ b_L(L) \\ b_{L-1}(L) \\ \vdots \\ b_1(L) \end{bmatrix} \end{aligned} \tag{18}$$

that defines the stationary zero-mean autoregressive process z_k powered by a zero-mean unit-variance white Gaussian noise w_k

$$z_k = \underline{b}(L)^{\mathrm{T}} \cdot \underline{z}_{k-1}^{k-L} + \sigma(L) \cdot w_k. \tag{19}$$

In the family Θ in (18), each member $\underline{\theta}(L)$ represents a *stationary* AR model of order L. For an AR model of order L, the stochastic complexity may be derived from Equation (16). The first step in this process is to obtain the maximum-likelihood (ML) estimates of the $(L+1)$-dimensional parameter $\hat{\underline{\theta}}_{\mathrm{ML}}(L) = \left[\begin{array}{cc} \hat{\sigma}_{\mathrm{ML}}(L) & \hat{\underline{b}}_{\mathrm{ML}}(L)^{\mathrm{T}} \end{array}\right]^{\mathrm{T}}$ The ML estimates are related to the ML covariance matrix estimate

$$\hat{\mathbf{C}}_{\mathrm{ML}}(L) = \left[\begin{array}{cc} \hat{\mathbf{c}}_{\mathrm{ML}}(L) & \hat{\underline{c}}_{\mathrm{ML}}(L) \\ \hat{\underline{c}}_{\mathrm{ML}}(L)^{\mathrm{T}} & \hat{c}_{0_{\mathrm{ML}}}(L) \end{array}\right] = \frac{1}{N}\sum_{k=1}^{N} \underline{z}_k^{k-L} \cdot \left(\underline{z}_k^{k-L}\right)^{\mathrm{T}} \tag{20}$$

via the familiar Yule-Walker equations from Section III

$$\begin{aligned} \hat{\underline{b}}_{\mathrm{ML}}(L) &= \hat{\mathbf{C}}_{\mathrm{ML}}(L)^{-1}\hat{\underline{c}}_{\mathrm{ML}}(L), \\ \hat{\sigma}^2_{\mathrm{ML}}(L) &= \frac{\det \hat{\mathbf{C}}_{\mathrm{ML}}(L)}{\det \hat{\mathbf{c}}_{\mathrm{ML}}(L)} = \frac{1}{N}\sum_{k=1}^{N}\left(z_k - \hat{\underline{b}}_{\mathrm{ML}}(L)^{\mathrm{T}} \cdot \underline{z}_{k-1}^{k-L}\right)^2 \end{aligned} \tag{21}$$

From the expression for complexity in Equation (16),

$$\begin{aligned} \mathcal{C}(L) &= -\ln f\left(\underline{z}_N^1 \mid \hat{\underline{\theta}}_{\mathrm{ML}}(L)\right) + \frac{L+1}{2}\ln N \\ &= \frac{N}{2}\ln\left[2\pi\cdot\hat{\sigma}^2_{\mathrm{ML}}(L)\right] + \frac{1}{2\hat{\sigma}^2_{\mathrm{ML}}(L)}\sum_{k=1}^{N}\left(z_k - \hat{\underline{b}}_{\mathrm{ML}}(L)^{\mathrm{T}}\underline{z}_{k-1}^{k-L}\right)^2 + \frac{L+1}{2}\ln N \\ &= \frac{N}{2}\ln\left[2\pi\cdot\hat{\sigma}^2_{\mathrm{ML}}(L)\right] + \frac{N}{2} + \frac{L+1}{2}\ln N \\ &= \underbrace{\frac{N}{2}\ln\left[2\pi\mathrm{e}\cdot\hat{\sigma}^2_{\mathrm{ML}}(L)\right]}_{\text{entropy}} + \underbrace{\frac{L+1}{2}\ln N}_{\text{penalty}} \end{aligned} \tag{22}$$

A noteworthy point here is that the first term, which represents the goodness-of-fit of the model is the entropy of the estimated noise process driving the autoregressive filter. The entropy term decreases monotonically as the model order L increases (a longer model order will always reduce the entropy). The second term, the penalty, is proportional to the product of two factors: i) the number of estimated parameters $L+1$ and ii) $\ln N$ where N is the number of windows $\underline{z}_k^{k-L}$ used in (20) to estimate each one of the $L+1$ parameters. Using the complexity metric derived above, the MDL estimate of the size of the model is given by

$$\begin{aligned} \hat{L}_{\mathrm{MDL}} &= \arg\min_{L\geq 0}\{\text{entropy} + \text{penalty}\} \\ &= \arg\min_{L\geq 0}\left\{\frac{N}{2}\ln\left[2\pi\mathrm{e}\cdot\hat{\sigma}^2_{\mathrm{ML}}(L)\right] + \frac{L+1}{2}\ln N\right\}. \end{aligned} \tag{23}$$

V. Application of the MDL principle to magnetic channel modeling

In this section, we apply the MDL principle to the problem of estimating the size of the signal-dependent autoregressive (AR) channel model for the magnetic recording channel. The signal-dependent AR channel model is given by Equations (2) and (3), or equivalently by Figure 2 in Section II. Referring to Figure 2 where $\underline{\alpha}$ represents the signal-dependence window and $\underline{\beta}$ represents the filter-dependence window, we can conclude that for every vector $\underline{\alpha}$ of I_2-I_1+1 binary inputs, we have one model parameter $y(\underline{\alpha})$, while for every vector $\underline{\beta}$ of D_2-D_1+1 binary inputs, we have $L+1$ parameters describing the noise shaping filter. Therefore, the model in Figure 2 is paremeterized by a vector $\underline{\theta}\,(I_1, I_2, D_1, D_2, L)$ of $2^{I_2-I_1+1} + (L+1)2^{D_2-D_1+1}$ real-valued parameters. With this formulation of the model parameter vector $\underline{\theta}$, the goal is to apply the formalism in equation (16) to formulate the stochastic complexity that dependes now on 5 parameters – I_1, I_2, D_1, D_2 and L. However, we cannot apply the exact complexity computation as given in equation (16) for two reasons: i) due to the intractability of the exact maximum likelihood equations (7)-(9), the signal nonlinearities and the filter coefficients for the signal-dependent AR model are not estimated jointly in the maximum-likelihood sense but rather using an approximate maximum-likelihood estimate where equation (7) is substituted by (14), and ii) the penalty factor in equation (16) needs to be corrected since each individual parameter of the model is estimated from only a fraction of the waveform samples $\underline{z}_N^1$ as evidenced by equations (10) through (14).

A. An MDL principle for signal-dependent autoregressive processes

Following the interpretation in equation (22), take the sum of the entropy of the estimated uncorrelated Gaussian noise process that drives the signal-dependent autoregressive (AR) filter and a penalty factor to get the model complexity term

$$\mathcal{C}\,(I_1, I_2, D_1, D_2, L) = \text{entropy} + \text{penalty}. \tag{24}$$

First, we compute the entropy term. The entropy term is computed from the probability density function of the observed waveform $\underline{z}_N^1$ given that the chosen model is the maximum likelihood parameter estimate vector $\underline{\hat{\theta}}_{\text{ML}}$. We denote this function as $\hat{F} = f\Big(\underline{z}_N^1 \mid \underline{z}_0^{1-L},\, \underline{a}_{N+C_2}^{1+C_1},\, \underline{\hat{\theta}}_{\text{ML}}\Big)$. The entropy is then $-\ln\hat{F}$, which equals the expression in (6) when the parameters $y(\underline{\alpha})$, $\sigma\big(\underline{\beta}\big)$ and $\underline{b}\big(\underline{\beta}\big)$ are substituted by their maximum likelihood estimates $\hat{y}_{\text{ML}}(\underline{\alpha})$, $\hat{\sigma}_{\text{ML}}\big(\underline{\beta}\big)$ and $\underline{\hat{b}}_{\text{ML}}\big(\underline{\beta}\big)$.

The resulting entropy equals

$$\begin{aligned}\text{entropy} &= -\ln \hat{F} \\ &= \frac{1}{2}\sum_{k=1}^{N}\left\{\ln\left[2\pi\cdot\hat{\sigma}^2_{\mathrm{ML}}\left(\underline{a}_{k+D_2}^{k+D_1}\right)\right]+\frac{\left(\left[\underline{z}_k^{k-L}-\underline{\hat{y}}_{\mathrm{ML}}\left(\underline{a}_{k+I_2}^{k+I_1-L}\right)\right]^{\mathrm{T}}\begin{bmatrix}-\underline{\hat{b}}_{\mathrm{ML}}\left(\underline{a}_{k+D_2}^{k+D_1}\right)\\ 1\end{bmatrix}\right)^2}{\hat{\sigma}^2_{\mathrm{ML}}\left(\underline{a}_{k+D_2}^{k+D1}\right)}\right\} \quad (25)\end{aligned}$$

By summing first over all occurrences of $\underline{a}_{k+D_2}^{k+D_1}=\underline{\beta}$ and then over all $2^{D_2-D_1+1}$ values of the binary vector $\underline{\beta}$, and consequently applying equation (8), we simplify the entropy term in (25) as

$$\begin{aligned}\text{entropy} &= \frac{1}{2}\sum_{\underline{\beta}}\sum_{k\,:\,\underline{a}_{k+D_2}^{k+D_1}=\underline{\beta}}\left\{\ln\left[2\pi\cdot\hat{\sigma}^2_{\mathrm{ML}}\left(\underline{\beta}\right)\right]+\frac{\left(\left[\underline{z}_k^{k-L}-\underline{\hat{y}}_{\mathrm{ML}}\left(\underline{a}_{k+I_2}^{k+I_1-L}\right)\right]^{\mathrm{T}}\begin{bmatrix}-\underline{\hat{b}}_{\mathrm{ML}}\left(\underline{\beta}\right)\\ 1\end{bmatrix}\right)^2}{\hat{\sigma}^2_{\mathrm{ML}}\left(\underline{\beta}\right)}\right\} \\ &= \sum_{\underline{\beta}}\left\{\frac{N_{\underline{\beta}}}{2}\ln\left[2\pi\mathrm{e}\cdot\hat{\sigma}^2_{\mathrm{ML}}\left(\underline{\beta}\right)\right]\right\}. \quad (26)\end{aligned}$$

While this entropy term is exact, its computation is difficult due to the intractability of equations (7) to (9). Therefore, to compute the entropy term (26), we use the approximate maximum likelihood variance estimates $\hat{\sigma}^2_{\mathrm{ML}}\left(\underline{\beta}\right)$ resulting from the substitution of (14) for $\underline{\hat{y}}_{\mathrm{ML}}(\underline{\alpha})$ in equations (10) through (13). We also note that due to this dependence of the entropy term on the estimated values $\hat{\sigma}^2_{\mathrm{ML}}\left(\underline{\beta}\right)$, the parameter estimation covered in Section III and the model-size estimation covered in this section cannot be viewed as separate computations, but rather as an interdependent estimation that needs to be jointly computed.

To compute the penalty in (24), we rely on the interpretation given in equation (22) that the penalty factor is proportional to the number of parameters multiplied by the logarithm of the number of data windows used for the estimation of each parameter. Since the estimates for the signal-dependent mean channel outputs and for the signal-dependent filter coefficients depend on different windows $\underline{\alpha}$ and $\underline{\beta}$, the penalty factor will have two terms

$$\text{penalty} = \text{penalty}_{\alpha} + \text{penalty}_{\beta}. \quad (27)$$

Here 'penalty_{α}' represents the complexity penalty for all the parameter estimates computed

using the signal-dependent window (i.e., using equation (14)), while 'penalty$_\beta$' stands for the penalty term for all the parameters estimated using the filter-dependence window (i.e., using equation (10) to compute the covariance matrix and subsequent equations (11) through (13) to extract the parameters). For every distinct binary vector $\underline{\alpha}$, we need to determine the mean signal value $y(\underline{\alpha})$. According to (14), every value $\hat{y}_{\mathrm{EM}}(\underline{\alpha})$ is determined from $N_{\underline{\alpha}}$ different values z_k, where $N_{\underline{\alpha}}$ is the number of occurrences of the binary pattern $\underline{\alpha}$ in the known binary sequence that created the waveform. According to (23), the complexity penalty for each value $\hat{y}_{\mathrm{EM}}(\underline{\alpha})$ is therefore $\frac{1}{2}\ln N_{\underline{\alpha}}$. Summing over all binary vectors $\underline{\alpha}$ of size $2^{I_2-I_1+1}\times 1$, we get

$$\mathrm{penalty}_\alpha = \frac{1}{2}\sum_{\underline{\alpha}} \ln N_{\underline{\alpha}}. \tag{28}$$

Similarly, as seen from Figure 2, for every binary vector $\underline{\beta}$, we need to compute $L+1$ noise filter parameters - a variance $\hat{\sigma}_{\mathrm{ML}}\left(\underline{\beta}\right)$ and an $L\times 1$ vector $\hat{\underline{b}}_{\mathrm{ML}}\left(\underline{\beta}\right)$. Each set of $L+1$ filter parameters is computed from the covariance matrix (10). Since each covariance matrix $\hat{\mathbf{C}}_{\mathrm{ML}}\left(\underline{\beta}\right)$ is computed from $N_{\underline{\beta}}$ windows $\underline{z}_k^{k-L}$ of the observed waveform (where $N_{\underline{\beta}}$ is the number of occurrences of the binary vector $\underline{\beta}$ in the binary input sequence), the penalty factor for the estimation of the $L+1$ filter coefficients is $\frac{L+1}{2}\ln N_{\underline{\beta}}$. Summing over all binary vectors $\underline{\alpha}$ of size $2^{D_2-D_1+1}\times 1$, we get

$$\mathrm{penalty}_\beta = \frac{L+1}{2}\sum_{\underline{\beta}} \ln N_{\underline{\beta}}. \tag{29}$$

Combining equations (24) and (26) through (29), we get the stochastic complexity of the signal-dependent autoregressive model

$$\mathcal{C}\left(I_1, I_2, D_1, D_2, L\right) = \frac{1}{2}\sum_{\underline{\beta}} \left\{ N_{\underline{\beta}} \ln\left[2\pi \mathrm{e}\cdot \hat{\sigma}^2_{\mathrm{ML}}\left(\underline{\beta}\right)\right] + (L+1)\ln N_{\underline{\beta}} \right\} + \frac{1}{2}\sum_{\underline{\alpha}} \ln N_{\underline{\alpha}}. \tag{30}$$

Analogous to the stationary AR case, the optimal model order is selected by choosing the 5-tuple (I_1, I_2, D_1, D_2, L) to minimize the complexity given in Equation (30), i.e.,

$$\left(\hat{I}_{1_{\mathrm{MDL}}}, \hat{I}_{2_{\mathrm{MDL}}}, \hat{D}_{1_{\mathrm{MDL}}}, \hat{D}_{2_{\mathrm{MDL}}}, \hat{L}_{\mathrm{MDL}}\right) = \arg \min_{(I_1,I_2,D_1,D_2,L)} \mathcal{C}\left(I_1, I_2, D_1, D_2, L\right) \tag{31}$$

The difficulties with the minimization in equation (31) are that there may exist several local minima and that a comprehensive search must be conducted over a five-dimensional space. A brute-force five-dimensional search is computationally too expensive. We considered applying a simulated annealing [38] based approach to avoid local minima, but realized that this approach would also be computationally too expensive. We therefore use an essentially ad-hoc search

method that, according to the experimental evidence which we present in Section VI, performs just as well as a full five-dimensional search.

B. A pragmatic minimum-search algorithm

Finding the parameters which minimize the complexity in Equation (31) involves a five-dimensional search over the parameter space, which is computationally prohibitive. We have therefore developed an ad-hoc procedure (detailed below) which never searches over a space of more than two dimensions.

1. Find preliminary estimates I_1^* and I_2^* :

We first estimate the size of the window which determines the signal-dependent waveform means. In order to do this, we assume that the additive noise is stationary and uncorrelated as well as signal-independent. This situation can be modeled by setting ($D_1 = D_2 = I_2 + R$, where R is some large integer, and $L = 0$. We find the estimates I_1^* and I_2^* as

$$(I_1^*, I_2^*) = \arg \min_{(I_1, I_2)} \mathcal{C}\left(I_1, I_2, I_2 + R, I_2 + R, 0\right) \tag{32}$$

2. Estimation of $D_1^*, D_2^*, \hat{L}_{\mathrm{MDL}}$:

Using the values of (I_1^*, I_2^*) estimated above, we obtain approximate estimates of (D_1, D_2) and the final estimate of L. Our simulations indicate that the memory order of the noise process, L, is estimated very accurately at this stage. In order to reduce complexity at this stage, we fix the size of the noise dependence window, $d = D_2 - D_1 + 1$ and search only over (D_1, L). Operating under the premise that the noise-dependence window is located within a small neighborhood of the signal-dependence window, we search over values of D_1 that lie in a neighborhood of I_1^* and over different values of L.

$$(D_1^*, \hat{L}_{\mathrm{MDL}}) = \arg \min_{D_1, L} \mathcal{C}\left(I_1^*, I_2^*, D_1, D_1 + d - 1, L\right) \tag{33}$$

3. Estimation of $\hat{D}_{1_{\mathrm{MDL}}}$ and $\hat{D}_{2_{\mathrm{MDL}}}$:

Using the estimates obtained in the previous two steps, we refine the estimates of the noise window defined by (D_1, D_2). In this step, we search for D_1 in a neighborhood of the value D_1^* obtained in the previous step, and allow D_2 to vary over a range of values.

$$\left(\hat{D}_{1_{\mathrm{MDL}}}, \hat{D}_{2_{\mathrm{MDL}}}\right) = \arg \min_{\substack{(D_1, D_2) \\ D_1 \text{ around } D_1^*}} \mathcal{C}\left(I_1^*, I_2^*, D_1, D_2, \hat{L}_{\mathrm{MDL}}\right) \tag{34}$$

4. Find estimates $\hat{I}_{1_{\mathrm{MDL}}}$ and $\hat{I}_{2_{\mathrm{MDL}}}$:

Assuming that we have good estimates of the noise structure, we now refine the estimates of I_1 and I_2 by searching in a *small window* around (I_1^*, I_2^*)

$$\left(\hat{I}_{1_{\mathrm{MDL}}}, \hat{I}_{2_{\mathrm{MDL}}}\right) = \arg \min_{\substack{(I_1,I_2) \\ \text{around} \\ (I_1^*,I_2^*)}} \mathcal{C}\left(I_1, I_2, \hat{D}_{1_{\mathrm{MDL}}}, \hat{D}_{2_{\mathrm{MDL}}}, \hat{L}_{\mathrm{MDL}}\right) \tag{35}$$

Although there is no guarantee that this procedure will end up in the absolute minimum of the MDL curve, we tried the procedure for several artificial waveforms and obtained the same minimum point as with a full search over all five parameters I_1, I_2, D_1, D_2 and L. Thereby, the ad-hoc search proposed in (32)-(35) never searches over a space whose dimension is higher than 2. This arrangement provides computational savings of several orders of magnitude over a full 5-dimensional search, and is therefore much more practical.

VI. Estimation Results

In this section, we use the MDL principle and the search algorithm detailed in Section V-B to estimate the model order and the model parameters of magnetic recording waveforms. In Section VI-A, we first present results on synthetic waveforms, where the model parameters are known. We then apply the technique to real waveforms in Section VI-B.

A. Synthetic Data

In order to validate the use of the MDL principle in a non-stationary context, we first estimate the parameters from a synthetic waveform of $N{=}10^5$ samples that was generated from a known model. For a given 5-tuple of parameters, (I_1, I_2, D_1, D_2, L), the waveform is generated directly from Equation (2) which represents the non-stationary AR model. A synthetic waveform was generated using the following parameters

$$\begin{aligned} I_1 &= -4 \\ I_2 &= -2 \\ D_1 &= -3 \\ D_2 &= 0 \\ L &= 2 \end{aligned}$$

We then used the search procedure described in Section V-B to find the point (I_1, I_2, D_1, D_2, L) with the lowest complexity. This search resulted in the correct value of the model size. To

underscore the efficacy of the technique, we present estimation results for the parameters in Tables I and II.

$\underline{\alpha}^{\mathrm{T}} = \left(\underline{a}_{k+I_2}^{k+I_1}\right)^{\mathrm{T}}$	$y(\underline{\alpha})$	$\hat{y}_{\mathrm{ML}}(\underline{\alpha})$
000	0	-0.0002
001	1	1.0007
010	0	0.0006
011	1	1.0006
100	-1	-0.9997
101	0	0.0000
110	-1	-1.0002
111	0	0.0001

TABLE I

TRUE AND ESTIMATED VALUES OF MEAN OUTPUTS OF THE CHANNEL.

$\underline{\beta}^{\mathrm{T}} = \left(\underline{a}_{k+D_2}^{k+D_1}\right)^{\mathrm{T}}$	$\sigma(\underline{\beta})$	$\underline{b}(\underline{\beta})^{\mathrm{T}}$		$\hat{\sigma}_{\mathrm{ML}}(\underline{\beta})$	$\hat{\underline{b}}_{\mathrm{ML}}(\underline{\beta})^{\mathrm{T}}$	
0000	0.0632	0.3	0.4	0.0636	0.3088	0.3927
0001	0.0316	0.2	-0.1	0.0318	0.2006	-0.1034
0010	0.0949	0.4	-0.5	0.0948	0.3936	-0.4996
0011	0.0632	-0.2	0.4	0.0632	-0.2057	0.4088
0100	0.0949	-0.1	0.2	0.0951	-0.1136	0.2082
0101	0.0316	-0.4	0.4	0.0319	-0.4002	0.4024
0110	0.0632	0.2	0.2	0.0635	0.2064	0.1834
0111	0.0949	0.2	-0.6	0.0946	0.2142	-0.5689
1000	0.0949	0.3	0.4	0.0955	0.2978	0.3824
1001	0.0632	0.2	-0.1	0.0633	0.1983	-0.1128
1010	0.0316	0.4	-0.5	0.0319	0.3975	-0.4948
1011	0.0949	-0.2	0.4	0.0948	-0.2084	0.3866
1100	0.0632	-0.1	0.2	0.0630	-0.0962	0.1992
1101	0.0949	-0.4	0.4	0.0946	-0.4113	0.3905
1110	0.0316	0.2	0.2	0.0317	0.2012	0.2017
1111	0.0632	0.2	-0.6	0.0623	0.1973	-0.6025

TABLE II

TRUE AND ESTIMATED VALUES OF THE FILTER COEFFICIENTS

B. Real Waveforms

We now apply the algorithm described in Section V to estimate the channel model for a real waveform collected from a spin-stand. The examination was conducted on a sampled readback waveform resulting from writing a random sequence of $N = 3 \cdot 10^6$ symbols (the period of the pseudo-random number generator was $> 4 \cdot 10^6$). The following head-media assembly was used. The reader was a giant magnetoresistive (GMR) head mounted on a comb, with a read-write chip attached to the flexible cable. The magnetic spacing was between 20 and 40 nm. The read gap was around 170 nm and the read track width was approximately 1.5 μm. The disk characteristics were as follows: coercivity $H_c = 3600$ Oe; product of remanent magnetization and media thickness $M_r t = 0.45$ memu/cm^2. The data rate was 47.06 MBytes/s, the linear density

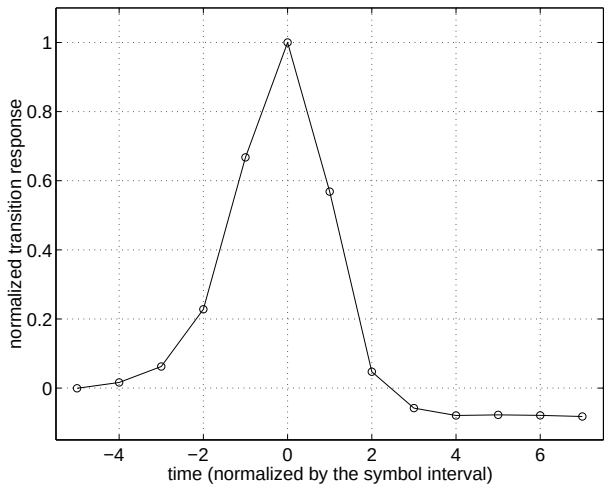


Fig. 3. The estimated response to an isolated transition normalized by the maximum amplitude. The waveform does not settle to its starting value due to a transverse magnetization component that is written at the edges of the head.

was 405.5 kbits/in and the normalized user density was 2.44 bits/PW50.

We first obtain the transition response from the waveform. This response is depicted in Figure 3. From the figure, we see that the response settles at a negative value. This effect is a result of the transverse magnetization component that is written at the edges of the head, and subsequently picked up by the magnetoresistive read head. Obtaining the model on this waveform would result in a bad estimation because the transition response in Figure 3 does not satisfy the finite memory assumption made in Section II due to the infinitely long non-zero tail.

To fix this problem, we apply a high-pass filter that removes the DC component of the signal without distorting the signal too much. In detectors implemented in practice, a high-pass digital filter is typially employed since the waveform is equalized to a target of the form $(1-D)h(D)$ or another target that has a very low DC component (see [39] and references therein). Here, we apply a similar approach, where the coefficients of $h(D)$ with indeces -4 to 4 are picked to be equal to the values depicted in Figure 3 for those same indeces. All other coefficients are chosen equal to zero. The high-pass filter is then implemented as an FIR filter, whose coefficients are found using a minimum-mean-squared-error fit. This filter does not distort the transition response very much because the target coefficients are chosen equal to the original transition response in the region of indeces -4 to 4. On the other hand, the long DC tails are removed as can be seen in Figure 4, where the high-pass-filtered transition response is depicted.

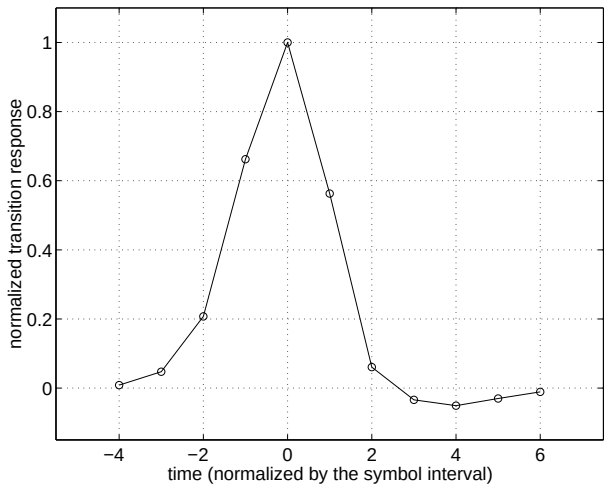


Fig. 4. The estimated transition response after high-pass filtering. The pulse eventually settles to the baseline.

Applying the MDL search for the model size to the high-pass filtered waveform, we get the following results

$$\begin{aligned} \hat{I}_{1_{\mathrm{MDL}}} &= -4 \\ \hat{I}_{2_{\mathrm{MDL}}} &= 6 \\ \hat{D}_{1_{\mathrm{MDL}}} &= -3 \\ \hat{D}_{2_{\mathrm{MDL}}} &= 4 \\ \hat{L}_{\mathrm{MDL}} &= 9 \end{aligned}$$

Obviously, the size of the model is too big to present it in a table form. Actually, the size of this model is so large that it prohibits the design of an optimal detector matched to this model. Such a detector would be a Viterbi detector with 2^{19} states, which is clearly too big for VLSI implementations. A practical approach to designing a detector for this channel is to employ equalization and design a suboptimal detector with a smaller number of states matched to the equalized waveform [27], or to revert to noise-predictive schemes [40], [41].

An interesting observation is that the signal-dependence window $(\hat{I}_{1_{\mathrm{MDL}}}, \hat{I}_{2_{\mathrm{MDL}}})$ is very large, while the filter-dependence window $(\hat{D}_{1_{\mathrm{MDL}}}, \hat{D}_{2_{\mathrm{MDL}}})$ is fully contained inside the signal-dependence window. This provides evidence that the head transfer function (responsible for the span I_1 to I_2) is longer than the span of the symbols that influence the signal-dependence length of the noise. The signal-dependence length of the noise $(D_2 - D_1)$ should not be confused with the noise

correlation length. The noise correlation length is proportional to the noise memory length L, where the constant of proportionality depends on the definition of the correlation length (e.g., distance to the 90% drop in correlation, or distance to the first zero-crossing, etc.) and the actual coefficients of the AR filter (e.g., if the AR filter polynomial has zaseros close to the unit circle, the correlation length will be long).

Unfortunately, the noise power in the analyzed waveform is very low. Therefore, there are no observable errors resulting from applying a detector either to the real waveform or to the model-generated waveform. Hence, we were not able to conduct a meaningful error-rate comparison study. For an exhaustive experimental argument showing that the error rates for the experimental and modeled waveforms match, the reader is referred to [15]. Further experiments that will test the validity of the model need to be conducted. These experiments should be conducted on very noisy media, so that the bit error rates from both the real waveforms and the modeled waveforms will be higher than 10^{-5}, which will assure reasonably short simulation periods. Clearly, the waveforms should not be collected from product-ready channels as these typically have very low error rates. A possible strategy for increasing the error rates is to collect waveforms from various off-track positions, as in [42].

VII. Conclusion

This paper presented a formal approach to modeling the magnetic recording channel. The chosen model is the signal-dependent autoregressive channel model. Prior work on this model assumed that the model order was known and concentrated on estimating the model parameters only. In this paper, we have developed a joint channel characterization that estimates both the appropriate model order (model complexity) and the corresponding model parameters.

The criterion used to strike a balance between the goodness-of-fit and the model complexity is the minimum description length (MDL) criterion. For autoregressive processes, the stochastic complexity is the sum of two terms: 1) the entropy of the estimated sequence of uncorrelated Gaussian random variables that drives the autoregressive filter (this entropy decreases monotonically as the model order increases) and 2) a penalty factor that is proportional to the model order. The MDL criterion finds the model order and the corresponding model-parameter estimates that minimize the stochastic complexity. This is a well-known method that has previously been applied to stationary autoregressive model-fitting. In this paper, we have modified the criterion for the channel with nonlinearities and signal-dependent noise typical for high density

magnetic recording channels.

While the MDL method provides a theoretically sound basis for joint parameter and order estimation, we have to overcome two difficulties in practice: 1) the maximum-likelihood parameter estimation equations are intractable and 2) the minimization of the MDL curve needs to be conducted in a 5-dimensional model-order space. In this paper, we proposed two pragmatic approaches to getting around these problems. The first problem was solved by using an approximate maximum-likelihood estimation procedure based the 'empirical mean' estimation of the channel nonlinearities, while the second was solved with an ad-hoc minimization procedure that substituted the 5-dimensional search with four 2-dimensional searches that reduce the computational load by several orders of magnitude. The method were tested on synthetically generated waveforms, showing excellent agreement between the estimated parameters and the true model-generating parameters. The pragmatic MDL method was subsequently applied to experimental waveforms collected from the spin-stand to establish the model order and the parameters of the model for the real magnetic recording channel.

The benefits of this method are three-fold. First, we now have a fully automated procedure for characterizing a given magnetic recording channel. Second, the model is computationally inexpensive and it can be easily used to generate a substantial number of waveform samples needed in extensive simulation studies such as Monte-Carlo simulations for obtaining detector error rates. Third, the model reveals the structure of the nonlinear intersymbol interference and the signal-dependent noise which can and should be used in designing detectors matched to the magnetic recording channel that is being considered.

Acknowledgments

The authors would like to thank Joost Mortelmans and Roger Wood of IBM Corporation for their help in obtaining the spin-stand waveforms used in this paper.

References Cited

References

[1] R. A. Baugh, E. S. Murdock, and B. R. Natarajan, "Measurement of noise in magnetic media," *IEEE Trans. Magn.*, vol. MAG-19, pp. 1722–1724, Sept. 1983.

[2] R. N. Belk, K. P. George, and S. G. Mowry, "Noise in high performance thin-film longitudinal magnetic recording media," *IEEE Trans. Magn.*, vol. MAG-21, pp. 1350–1355, Sept. 1985.

[3] J.-G. Zhu and N. H. Bertram, "Recording and transition noise simulations in thin film media," *IEEE Trans. Magn.*, vol. 24, pp. 2706–2708, Nov. 1988.

[4] R. D. Brandt, A. J. Armstrong, H. N. Bertram, and J. K. Wolf, "A simple statistical model of partial erasure in thin film disk recording systems," *IEEE Trans. Magn.*, vol. MAG-27, pp. 4978–4980, Nov. 1991.

[5] S. W. Yuan and H. N. Bertram, "Statistical data analysis of magnetic recording noise mechanisms," *IEEE Trans. Magn.*, vol. 28, pp. 84–92, Jan. 1992.

[6] R. Wood, "The feasibility of magnetic recording at 1 terabit per square inch," *IEEE Trans. Magn.*, vol. 36, pp. 36–42, Jan. 2000.

[7] H. N. Bertram, *Theory of Magnetic Recording.* Cambridge: Cambridge University Press, 1994.

[8] K. Fisher, J. Cioffi, and H. Thapar, "Modeling in thin film storage channels," *IEEE Trans. Magn.*, vol. 25, pp. 4081–4058, Sept. 1989.

[9] J. Moon and J.-G. Zhu, "Nonlinear effects of transition broadening," *IEEE Trans. Magn.*, vol. 27, pp. 4831–4833, Nov. 1991.

[10] S. K. Nair, H. Shafiee, and J. Moon, "Modeling and simulation of advanced read channels," *IEEE Trans. Magn.*, vol. MAG-29, pp. 4056–4058, Nov. 1993.

[11] J. Caroselli and J. K. Wolf, "Applications of a new simulation model for media noise limited magnetic recording channels," *IEEE Trans. Magn.*, vol. 32, pp. 3917–3919, Sept. 1996.

[12] A. Kavčić and J. M. F. Moura, "Expedient media noise modeling: Isolated and interacting transitions," *IEEE Trans. Magn.*, vol. 32, pp. 3875–3877, Sept. 1996.

[13] A. Kavčić and A. Patapoutian, "A signal-dependent autoregressive channel model," *IEEE Trans. Magn.*, vol. 35, pp. 2316–2318, September 1999.

[14] A. Kavčić and J. M. F. Moura, "The Viterbi algorithm and Markov noise memory," *IEEE Trans. Inform. Theory*, vol. 46, pp. 291–301, Jan. 2000.

[15] J. Stander and A. Patapoitian, "Performance of a signal-dependent autoregressive channel model," *IEEE Trans. Magn.*, vol. 36, pp. 2197–2199, Sep. 2000. to appear.

[16] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding:

Turbo-codes," in *Proc. IEEE Int. Conf. on Communications*, (Geneva, Switzerland), pp. 1064–1070, May 1993.

[17] W. Ryan, "Performance of high-rate turbo codes on PR4-equalized magnetic recording channels," in *Proc. IEEE Int. Conf. on Communications*, (Atlanta, GA), pp. 947–951, June 1998.

[18] T. Souvignier, A. Friedmann, M. Öberg, P. Siegel, R. E. Swanson, and J. K. Wolf, "Turbo codes for PR4: Parallel versus seral concatenation," in *Proc. IEEE Int. Conf. on Communications*, (Vancouver, Canada), pp. 1638–1642, June 1999.

[19] L. L. McPheters, S. W. McLaughlin, and K. R. Narayanan, "Precoded PRML, serial concatenation, and iterative (turbo) decoding for digital magnetic recording," *IEEE Trans. Magn.*, vol. 35, pp. 2325–2327, Sep. 1999.

[20] T. Duman and E. Kurtas, "Comprehensive performance investigation of turbo codes over high density magnetic recording channels," in *Proc. IEEE GLOBECOM 99*, (Rio de Janeiro), pp. 744–748, Dec. 1999.

[21] R. G. Gallager, *Low-Density Parity-Check Codes*. Cambridge, MA: MIT Press, 1962.

[22] D. J. C. MacKay, "Good error-correcting codes based on very sparce matrices," *IEEE Trans. Inform. Theory*, vol. 45, pp. 399–431, March 1999.

[23] T. Richardson and R. Urbanke, "The capacity of low-density parity check codes under message-passing decoding," *IEEE Trans. Inform. Theory*, vol. 47, pp. 599–618, February 2001.

[24] J. Fan, A. Friedmann, E. Kurtas, and S. McLaughlin, "Low density parity check codes for magnetic recording," *Proc. Allerton Conference on Communications and Control*, (Urbana, IL), October 1999.

[25] A. Kavčić, "Soft-output detector for channels with intersymbol interference and Markov noise memory," in *Proc. IEEE GLOBECOM 99*, (Rio de Janeiro), pp. 728–732, Dec. 1999.

[26] 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. 20, pp. 284–287, March 1974.

[27] J. Moon, Jan. 2000. Presentation at the NSIC quarterly review; sumbitted for publication.

[28] A. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2743–2760, Oct. 1998.

[29] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: John Wiley and Sons, 1991.

[30] A. Kavčić and J. M. F. Moura, "Matrices with banded inverses: Inversion algorithms and factorization of Gauss-Markov processes," *IEEE Trans. Inform. Theory*, vol. 46, pp. 1495–1509, July 2000.

[31] J. M. F. Moura and N. Balram, "Recursive structure of noncausal Gauss Markov random fields," *IEEE Trans. Inform. Theory*, vol. 38, pp. 334–354, March 1992.

[32] A. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automatic Control*, vol. 19, pp. 716–723, 1974.

[33] G. Schwartz, "Estimating the dimension of a model," *Ann. Stat.*, vol. 6, pp. 461–464, 1978.

[34] J. Rissanen, "Modelling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[35] J. Rissanen and G. G. Langdon, "Universal modeling and coding," *IEEE Transactions on Information Theory*, vol. 27, pp. 12–23, 1981.

[36] C. E. Shannon, "A mathematical theory of communications," *Bell Systems Technical Journal*, vol. 27, pp. 379–423 (part I) and 623–656 (part II), 1948.

[37] J. Rissanen, "Stochastic complexity and modeling," *Annals of Statistics*, vol. 14, pp. 1080–1100, 1986.

[38] S. Kirkpatrick, C. D. Gelatt Jr., and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, pp. 671–680, Sep. 1983.

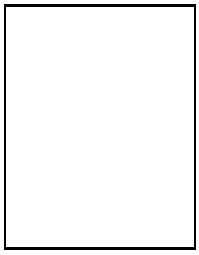
[39] N. M. Zayed and L. R. Carley, "Generalized partial response signaling and efficient MLSD using linear viterbi branch metrics," in *Proc. IEEE Global Communications Conference (GLOBECOM)*, (Rio de Janeiro), pp. 949–954, Dec. 1999.

[40] J. D. Coker, E. Eleftheriou, R. L. Galbraith, and W. Hirt, "Noise-predictive maximum likelihood NPML detection," *IEEE Trans. Magn.*, vol. 34, pp. 110–117, Jan. 1998.

[41] S. A. Altekar and J. K. Wolf, "Improvements in detectors based upon colored noise," *IEEE Trans. Magn.*, vol. 34, pp. 94–97, Jan. 1998.

[42] T. Souvignier, Z. Keirn, and C. Xu, "Turbo decoding for partial response channels using spinstand data," *IEEE Trans. Magn.*, vol. 36, pp. 2167–2169, Sep. 2000.

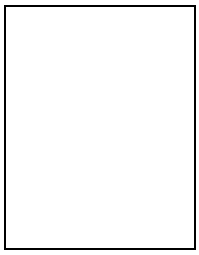
Authors' Biographies



Aleksandar Kavčić (S'93–M'98) was born in Belgrade, Yugoslavia in 1968. He received the Dipl.-Ing. degree in Electrical Engineering from Ruhr-University, Bochum, Germany in 1993, and the Ph.D. degree in Electrical and Computer Engineering from Carnegie Mellon University, Pittsburgh, Pennsylvania in 1998.

Since 1998, he has been an Assistant Professor of Electrical Engineering at Harvard University in the Division of Engineering and Applied Sciences. He held short-term research positions at Seagate Technology in 1995, Read-Rite Corporation in 1996, and Quantum Corporation from 1997 to 1998, and has served as a technical consultant for Quantum Corporation in 1999 and 2000. His research spans topics in Communications, Signal Processing, Information Theory and Magnetic Recording, with the most recent interests in magnetic recording channel modeling, multichannel signal processing, detector design, timing recovery and iterative decoding algorithms.

Dr. Kavčić received the IBM Partnership Award in 1999 and the NSF CAREER Award in 2000.



Murari Srinivasan was born in Bangalore, India in 1972. He received the B.Tech degree in Electronics and Communication Engineering from the Indian Institute of Technology, Madras in 1993 and the M.S degree and Ph.D. degrees in Electrical Engineering from the University of Maryland, College Park in 1996 and 1999. Following the completion of his Ph.D., he was a post-doctoral research fellow in the Division of Engineering and Applied Science at Harvard University from 1999 to 2000. He is currently a Member of Technical Staff at Flarion Technologies, Bedminster NJ, where he works on the design, analysis and implementation of wireless communication systems.

For his doctoral thesis, he investigated several issues related to combined source-channel coding with applications to image and video transmission over noisy channels. He has worked on digital video transmission over wireless channels and the design of ground station receivers for satellite based communication networks during internships at Bell Laboratories, Lucent Technologies and Hughes Network Systems. As part of his post-doctoral research, he worked on signal processing aspects of magnetic recording systems and information- theoretic aspects of quasi-autonomous computer vision systems. He also collaborated to the design of an image transmission protocol for the internet.

His interests broadly span the design, analysis and implementation of communication systems and networks.