Universal Delay Estimation for Discrete Channels

Jeremy Stein, Jacob Ziv and Neri Merhav

Department of Electrical Engineering
Technion - Israel Institute of Technology
Haifa 32000, Israel

April 10, 1998

Abstract

The use of information theory concepts for universal estimation of delay for classes of discrete channels is discussed. The problem is presented as one of hypothesis testing. Although the channel statistics are not known, for large enough signal duration, the exponent of the average error probability is equal to that associated with the optimal maximum likelihood (ML) decision procedure which utilizes full knowledge of the channel parameters. Two categories of problems are discussed: The single channel problem, where the random transmitted signal is known to the receiver, and the two sensor problem, where the random signal is unknown.

Index Terms: delay estimation, hypothesis testing, maximum likelihood, entropy, mutual information, finite-state source, Lempel-Ziv compression, large deviations theory, memoryless types, Markov types.

I Introduction

The use of information theory to solve radar related problems is not new. In the early 1950's Woodward and Davies [1, 2, 3, 4] examined the use of information theoretic principles to obtain the *a posteriori* radar receiver. They discussed the case of a known signal in white Gaussian noise which results in a correlation receiver. A summary of their results, with a tutorial on information theory, appeared in a book by Woodward [5]. Besides dealing with the detection problem, they also considered estimation of delay. In [1] they defined the quantity of information from a radar observation as the difference in entropies of the *a priori* and *a posteriori* probability distributions for delay; and calculated it for the Gaussian channel using approximations for high and low degrees of ambiguity (existence of several distinct peaks in the *a posteriori* delay probability function). This quantity was compared to the Shannon formula for maximum information transmission over the Gaussian channel [6] and shown to have some similar threshold properties and asymptotic values. In [2] Woodward pointed out that, although some of his assumptions (e.g. echo of known strength) were artificial, he hoped to "set the ball rolling" for application of information theory to radar problems.

Since Woodward's and Davies' work few researchers have considered the connection between information theory and radar detection and estimation problems. Ziv and Zakai [7] compared their lower bound, known as the Ziv-Zakai lower bound, for the example of average mean square delay error of a rectangular pulse transmitted over the Gaussian channel, with the bound obtained from Shannon's rate-distortion theory [6, Th. 21]. They also showed [8] that using a generalized rate-distortion theory instead of Shannon's rate-distortion theory, tighter bounds on the average mean square delay error may be achieved. Zeoli [9] used rate distortion theory to obtain a lower bound on the data rate for processing synthetic aperture radar signals, and showed that conventional analog to digital methods give rates close to this lower bound for a given distortion. Frost and Shanmugan [10] demonstrated the use of mutual information to measure the information content of synthetic aperture radar signals; they used the mutual information measure to illustrate the tradeoff between spatial and radiometric resolution. A recent paper by Bell [11] describes use of maximizing the mutual information between a random extended target ensemble and the received radar signal to design radar waveforms. These waveforms were reported to be optimal in a certain sense.

In this paper, we use information-theoretic techniques for estimation of delay for cases where the charac-

teristics of the medium (referred to as channel) are unknown, which is typically the situation. Two problems are addressed. The first problem is range estimation by a single radar or sonar unit. The transmitted waveform is known to the receiver, which has to decide on its correct shift in order to estimate the delay. This problem is solved for discrete memoryless channels (DMC's) and modular additive finite-state channels (MAFSC's). The second problem is estimation of the relative delay of a signal at two sensors, where the transmitted or reflected signal from the target is unknown. This is a common problem of bearing estimation and is solved for DMC's.

Since discrete-time signals are considered, the delay problem here is treated as one of hypothesis testing, and not as an estimation problem as customarily done. We consider only discrete finite alphabet time signals and refer to them as vectors. Thus the appropriate error criterion is the probability of error. If we presume that the delay is uniformly distributed over a finite number of signal shifts, the optimal decoder is the maximum likelihood (ML) decoder. We further assume that the transmitted signal is random. This assumption is legitimate for both problems. For the single channel problem, the optimal signal depends on the channel parameters which are unknown. Even if the channel parameters are known, the optimal choice of a signal is still an unsolved problem, although some kinds of signals have been proved to be better than others. For the well studied Gaussian case, the optimal signal, measured in terms of the ambiguity function, is still unknown [12]. Pseudo random sequences have been shown to have good properties for delay estimation [13]. For the two sensor problem, the signal emitted or reflected from the target is unknown and thus these signals can be considered random. Therefore our error criterion is the average probability of error, where the averaging is over all possible transmitted and received signals. For the above mentioned cases, a universal decision rule, that has no knowledge of channel parameters, is proposed that accomplishes the exponent of the average error probability associated with the optimal ML decision rule.

The problem formulated here can be interpreted as a channel coding problem, where the code words are simply shifts of a single randomly chosen vector. It must be noted that our problem has less degrees of freedom than the random coding problem, as only one code vector is chosen instead of a number as large as the codebook size. Thus our problem can be viewed as a random coding pulse position modulation (PPM) problem. The proposed universal decision rule for the DMC cases of delay estimation is based on minimum

joint empirical entropy. The shift that obtains minimum joint empirical entropy of the channel input and output determines the delay. Ignoring end-effects, the empirical entropy of a signal and its time shift are equal. Thus, the decision rule can be seen as one of maximum empirical mutual information (MMI) as was proposed by Goppa [14] for universal decoding for DMC's. This decoder, known as the MMI receiver, selects an input signal that maximizes the empirical mutual information with a given output vector. Goppa showed that, for large enough signal duration, the channel capacity is achieved. Csiszár and Körner [15] demonstrated that this universal decoder yields the random coding exponent given by the optimal ML decoder. Ziv [16] extended Csiszár's and Körner's result to finite-state channels by using a decoder based on the Lempel-Ziv (LZ) compression algorithm [17]. In the case of a DMC this decoder can be replaced by one that minimizes joint empirical entropy. Our decision rule for the MAFSC's also utilizes the LZ algorithm.

In Section II, the two problems discussed for the class of DMC's. The single channel problem is also solved for the class of MAFSC's in Section III. Finally, Section IV contains the discussion and conclusions.

II Discrete Memoryless Channel (DMC)

In this Section, the two delay estimation problems will be solved for the class of DMC's. We will first define the delay estimation problem for the single channel, where the transmitted signal is known to the receiver. The optimal ML decision rule will be given, and a universal decision rule is proposed. This universal decision rule will be proved to attain the exponent of the average error probability associated with the optimal ML decision procedure. Subsequently, the double channel delay estimation problem, where the transmitted signal is unknown to the two receivers, is presented. This problem, for the DMC case, is solved and shown to be similar to the single channel problem.

Before continuing, some notation will be needed: Let $\mathbf{x}$ denote the channel input vector, $\mathbf{x} = (x_1, x_2, ..., x_n)$ where $x_i \in \mathcal{X}$ is the input to the channel at the ith time instant and $\mathcal{X}$ is a finite alphabet of size $|\mathcal{X}|$. The vector $\mathbf{x}$ takes on values in $\mathcal{X}^n$ which is the set of all channel input vectors. Similarly, let $\mathbf{y}$ denote the channel output vector, $\mathbf{y} = (y_1, y_2, ..., y_n)$ where $y_i \in \mathcal{Y}$ is the output of the channel at the ith time instant and $\mathcal{Y}$ is a finite alphabet of size $|\mathcal{Y}|$. The vector $\mathbf{y}$ takes on values in $\mathcal{Y}^n$ which is the set of all channel output vectors. In the case of two channels, the second channel output $\mathbf{z} \in \mathcal{Z}^n$, is defined similarly over

the finite single-letter alphabet $\mathcal{Z}$ of cardinality $|\mathcal{Z}|$. We adopt the rule that boldface letters denote vectors while the usual font is used for scalars. A vector with a subscript i will denote a shifted version of the vector by i time units, i.e., $\mathbf{x}_i = (x_{1+i}, x_{2+i}, ..., x_{n+i})$. For simplicity, we assume in the proof cyclic shifts. Thus the index i is equal to i mod n. Consequently, $\mathbf{x}_0 \equiv \mathbf{x}$.

A *Single DMC*

Consider the class of finite alphabet DMC's with the transition probability function of the form

$$Q(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{n} Q(y_i|x_i). \tag{1}$$

Let the memoryless (zero order Markov) type of a sequence [15] $\mathbf{x} \in \mathcal{X}^n$ be the empirical probability mass function (PMF) $q_{\mathbf{x}}$ defined by the relative frequencies

$$q_{\mathbf{x}}(a) = \frac{1}{n}|\{i \in \{1, ..., n\} : \quad x_i = a\}| \qquad \forall\ a \in \mathcal{X}. \tag{2}$$

The set of sequences having the same empirical PMF $q_{\mathbf{x}}$ is called a typical set and is denoted $T_{\mathbf{x}}$. Throughout, capital italics (Q,P,M,K,R and J) will denote probability measures and q with a subscript, will denote the empirical PMF of the respective vector. Let $P(\mathbf{x})$ be the PMF governing the input vector $\mathbf{x}$. This PMF must have the property that vectors of the same memoryless type $q_{\mathbf{x}}$ have equal probability. The PMF $P(\mathbf{x})$ could be, for example, an identically independent distribution (i.i.d.) or be uniformly distributed over a memoryless type or group of memoryless types.

The optimum ML decision rule for the above problem is the shift of the vector $\mathbf{x}$ for which maximum conditional probability is obtained:

$$\hat{\tau}_o = \arg\max_i Q(\mathbf{y}|\mathbf{x}_i). \tag{3}$$

A universal decision rule is now proposed, i.e., it is a function of $\mathbf{x}$ and $\mathbf{y}$ only and is independent of the unknown probability function $Q(\cdot|\cdot)$. This rule is based on joint empirical entropy. Let $q_{\mathbf{x},\mathbf{y}}$ denote the joint memoryless empirical PMF of $\mathbf{x}$ and $\mathbf{y}$,

$$q_{\mathbf{x},\mathbf{y}}(a,b) = \frac{1}{n}|\{i \in \{1, ..., n\} : \quad (x_i, y_i) = (a,b)\}| \qquad \forall a \in \mathcal{X}, b \in \mathcal{Y}. \tag{4}$$

The joint empirical entropy is defined as

$$\hat{H}(\mathbf{x},\mathbf{y}) \stackrel{\triangle}{=} -\sum_{a\in\mathcal{X}}\sum_{b\in\mathcal{Y}} q_{\mathbf{x},\mathbf{y}}(a,b)\log q_{\mathbf{x},\mathbf{y}}(a,b). \tag{5}$$

Similarly $\hat{H}(\mathbf{x}_i,\mathbf{y})$ will denote the joint empirical entropy of $\mathbf{x}_i$ (the ith shift of $\mathbf{x}$) and $\mathbf{y}$. The universal decision rule selects the shift of the vector $\mathbf{x}$ for which minimum joint empirical entropy with the received vector $\mathbf{y}$ is obtained, namely:

$$\hat{\tau}_u = \arg\min_i \hat{H}(\mathbf{x}_i,\mathbf{y}). \tag{6}$$

This decision rule can be interpreted as one of MMI. If we assume cyclic shifts then $\hat{H}(\mathbf{x}_i) = \hat{H}(\mathbf{x})$, and therefore from information equalities (mutual information between two variables is equal to the sum of the marginal entropies minus their joint entropy), a rule of minimum joint empirical entropy is equal to one of maximum empirical mutual information.

The intuitive reasoning for this decision rule is that if we have the wrong shift, the memoryless joint probability of the input and output is of two independent variables, and thus the memoryless joint entropy is equal to the sum of their marginal entropies. For the correct shift, the memoryless probability distributions of the input and output vectors are dependent. A joint entropy of dependent variables is less than the sum of the entropies for each variable. From the law of large numbers empirical entropies converge to the true corresponding entropies.

Let $\mathbf{1_e}(\mathbf{x},\mathbf{y})$ denote the indicator function of an error for a decision rule given the input and output vectors. The average probability of error $\overline{P}_e$ averaged over all input vectors $\mathbf{x}$ and output vectors $\mathbf{y}$ is thus

$$\overline{P}_e = \sum_{\mathbf{x}\in\mathcal{X}^n}\sum_{\mathbf{y}\in\mathcal{Y}^n} P(\mathbf{x})Q(\mathbf{y}|\mathbf{x})\mathbf{1_e}(\mathbf{x},\mathbf{y}). \tag{7}$$

Henceforth, $\overline{P}_{e,o}$ and $\overline{P}_{e,u}$, will denote the average probability of error for the ML and for the universal decision rule, respectively.

The Theorem for this case is now stated:

Theorem:

For the above defined problem and for a fixed number of shifts,

a) The average error of the optimal ML decision rule $\overline{P}_{e,o}$ vanishes exponentially with n.

b) The asymptotic error exponent of the universal rule is equal to that of the optimal ML rule, i.e.,

$$\lim_{n\to\infty} \frac{1}{n} \log \overline{P}_{e,u} = \lim_{n\to\infty} \frac{1}{n} \log \overline{P}_{e,o}. \tag{8}$$

The first part of the Theorem asserts that, for the delay problem presented here, as for random channel coding, it is possible to achieve an average error probability that decreases exponentially with signal duration, provided the channel statistics are known and the optimal ML decision rule is used. Part b of the Theorem claims that, for the above mentioned case and its corresponding universal decision rule, the exponent of the average error probability is equal to that associated with the optimal ML decision rule, where the channel statistics are fully known.

Before giving the proof of the above Theorem for this case, some preliminaries are needed. The proof is based on the well known method of memoryless types [15] and the less common Markov type method [18, 19, 20]. The essentials of the Markov type method needed for the proof are given.

Let the first order Markov type of a sequence $\mathbf{x} \in \mathcal{X}^n$ be the empirical PMF $q_{\mathbf{x}_0,\mathbf{x}_1}$ defined by the relative frequencies

$$q_{\mathbf{x}_0,\mathbf{x}_1}(a_0,a_1) = \frac{1}{n}|\{i \in \{1,...,n\}: \quad (x_i,x_{i+1}) = (a_0,a_1)\}| \qquad \forall\ (a_0,a_1) \in \mathcal{X}^2 \tag{9}$$

where $\mathcal{X}^2$ is the Cartesian product $\mathcal{X} \times \mathcal{X}$. The empirical PMF $q_{\mathbf{x}_0,\mathbf{x}_1}(a_0,a_1)$ defines two marginal empirical PMF's $q_{\mathbf{x}_0}(a)$ and $q_{\mathbf{x}_1}(a)$ which are equal for a cyclic shift of $\mathbf{x}$ ($x_{n+1} = x_1$). The set of sequences of the type $q_{\mathbf{x}_0,\mathbf{x}_1}$ is called the first order Markov typical set denoted $T_{\mathbf{x}_0,\mathbf{x}_1}$. This set is obviously a subset of the typical set defined by its marginal empirical PMF $q_{\mathbf{x}_0}(a)$, i.e., $T_{\mathbf{x}_0,\mathbf{x}_1} \subseteq T_{\mathbf{x}}$. The exact expression for the size of the typical set $T_{\mathbf{x}_0,\mathbf{x}_1}$ is given by Whittle's formula [21, 22],

$$|T_{\mathbf{x}_0,\mathbf{x}_1}| = \prod_{a_0\in\mathcal{X}} \frac{(nq_{\mathbf{x}_0}(a_0))!}{\prod_{a_1\in\mathcal{X}}(nq_{\mathbf{x}_0,\mathbf{x}_1}(a_0,a_1))!} F^* \tag{10}$$

where the factor F^* is bounded from below by $(n+1)^{-|\mathcal{X}|}$ and from above by 1. Using the exponential bounds for the multinomial coefficients [15, Lemma 1.2.3], we get

$$(n+1)^{-|\mathcal{X}|^2-|\mathcal{X}|} \cdot \ 2^{n(\hat{H}(\mathbf{x}_0,\mathbf{x}_1)-\hat{H}(\mathbf{x}))} \leq |T_{\mathbf{x}_0,\mathbf{x}_1}| \leq 2^{n(\hat{H}(\mathbf{x}_0,\mathbf{x}_1)-\hat{H}(\mathbf{x}))} \tag{11}$$

where $\hat{H}(\mathbf{x}_0,\mathbf{x}_1)$ and $\hat{H}(\mathbf{x})$ are the empirical entropies defined as

$$\hat{H}(\mathbf{x}_0,\mathbf{x}_1) \stackrel{\triangle}{=} -\sum_{a_0\in\mathcal{X}}\sum_{a_1\in\mathcal{X}} q_{\mathbf{x}_0,\mathbf{x}_1}(a_0,a_1)\log q_{\mathbf{x}_0,\mathbf{x}_1}(a_0,a_1) \tag{12}$$

and

$$\hat{H}(\mathbf{x}) \stackrel{\triangle}{=} -\sum_{a\in\mathcal{X}} q_{\mathbf{x}_0}(a)\log q_{\mathbf{x}_0}(a). \tag{13}$$

Let the joint first order Markov type of a pair of sequences $\mathbf{y}\in\mathcal{Y}^n, \mathbf{x}\in\mathcal{X}^n$ be the empirical PMF $q_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}}$ defined by the relative frequencies

$$q_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}}(a_0,a_1,b) = \frac{1}{n}|\{i\in\{1,...,n\}: \quad (x_i,x_{i+1},y_i)=(a_0,a_1,b)\}| \qquad \forall\ (a_0,a_1)\in\mathcal{X}^2,\ b\in\mathcal{Y}. \tag{14}$$

The above empirical PMF defines marginal PMF's $q_{\mathbf{x}_0,\mathbf{x}_1}(a_0,a_1)$ (eq. 9) and $q_{\mathbf{x},\mathbf{y}}(a_0,b)$ (eq. 4) where $q_{\mathbf{x}_0,\mathbf{y}}$ $\equiv q_{\mathbf{x},\mathbf{y}}$.

The set of sequences $\mathbf{y}$ having the type $q_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}}$ given $\mathbf{x}$ is called the first order conditional Markov type denoted by $T_{\mathbf{y}|\mathbf{x}_0,\mathbf{x}_1}$,

$$T_{\mathbf{y}|\mathbf{x}_0,\mathbf{x}_1} = \{\mathbf{y}\in\mathcal{Y}^n: \quad q_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}}(a_0,a_1,b) = q_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}}\}. \tag{15}$$

This set is obviously a subset of the typical set $T_{\mathbf{y}|\mathbf{x}}$ defined by the set of $\mathbf{y}$ vectors having the empirical PMF $q_{\mathbf{x},\mathbf{y}}(a,b) = q_{\mathbf{x},\mathbf{y}}$ for a given $\mathbf{x}$. The size of this type $T_{\mathbf{y}|\mathbf{x}_0,\mathbf{x}_1}$ can be calculated by permuting y_i with y_j having the same Markov condition $(x_i,x_{i+1})=(x_j,x_{j+1})$, giving

$$|T_{\mathbf{y}|\mathbf{x}_0,\mathbf{x}_1}| = \prod_{a_0,a_1\in\mathcal{X}} \frac{(nq_{\mathbf{x}_0,\mathbf{x}_1}(a_0,a_1))!}{\prod_{b\in\mathcal{Y}}(nq_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}}(a_0,a_1,b))!}. \tag{16}$$

Using the exponential bounds for the multinomial coefficients [15, Lemma 1.2.3] as before,

$$(n+1)^{-|\mathcal{X}|^2|\mathcal{Y}|} \cdot \ 2^{n(\hat{H}(\mathbf{x}_0,\mathbf{x}_1,\mathbf{y})-\hat{H}(\mathbf{x}_0,\mathbf{x}_1))} \ \leq \ |T_{\mathbf{y}|\mathbf{x}_0,\mathbf{x}_1}| \ \leq \ 2^{n(\hat{H}(\mathbf{x}_0,\mathbf{x}_1,\mathbf{y})-\hat{H}(\mathbf{x}_0,\mathbf{x}_1))} \tag{17}$$

where $\hat{H}(\mathbf{x}_0, \mathbf{x}_1, \mathbf{y})$ is defined

$$\hat{H}(\mathbf{x}_0, \mathbf{x}_1, \mathbf{y}) \stackrel{\triangle}{=} -\sum_{a_0 \in \mathcal{X}} \sum_{a_1 \in \mathcal{X}} \sum_{b \in \mathcal{Y}} q_{\mathbf{x}_0, \mathbf{x}_1, \mathbf{y}}(a_0, a_1, b) \log q_{\mathbf{x}_0, \mathbf{x}_1, \mathbf{y}}(a_0, a_1, b). \tag{18}$$

The number of types of this kind is bounded by $(n+1)^{|\mathcal{X}|^2|\mathcal{Y}|}$ (A tighter bound may be obtained, but as we are interested in exponential behaviour it is of no significance as it is polynomial with n). The above type is a special case of conditional finite-state types derived by Satt [23].

Proof of the Theorem for the DMC:

The proof of the Theorem will first be given for simple hypotheses, where there are only two possible relative shifts of the vectors. The proof will then be extended to the case of finite multiple hypotheses. For the given channel model, where the delay is not part of the channel characteristic, the correct delay is obtained for $i = 0$. The error probability is calculated for the wrong decision $i = 1$. The direction of shift for this problem is of no significance as will become apparent. If the decision rule is inconclusive, the expressions are equal, it will be considered an error.

1) Proof of part a for simple hypotheses:

Assume that the channel is not altogether noisy, that is,

$$H(Y|X) = H(Y) - \delta \tag{19}$$

for some $\delta > 0$ where $H(Y|X)$ and $H(Y)$ are per-letter entropies of the channel output given the channel input and of the channel output respectively. For the memoryless channel and $\mathbf{x}$ a memoryless random vector, there is no memoryless dependency between the received vector $\mathbf{y}$ and the shifted vector $\mathbf{x}_1$. Therefore,

$$H(Y|X_1) = H(Y). \tag{20}$$

The average error probability associated with the universal decision rule can be bounded as follows:

$$\begin{aligned} \overline{P}_{e,u} &= Pr\{\hat{H}(\mathbf{x}_1, \mathbf{y}) \leq \hat{H}(\mathbf{x}, \mathbf{y})\} \\ &\stackrel{a}{=} Pr\{\hat{H}(\mathbf{y}|\mathbf{x}_1) \leq \hat{H}(\mathbf{y}|\mathbf{x})\} \\ &\stackrel{b}{\leq} Pr\{\hat{H}(\mathbf{y}|\mathbf{x}_1) \leq \lambda\} + Pr\{\hat{H}(\mathbf{y}|\mathbf{x}) \geq \lambda\} \end{aligned}$$

$$\stackrel{c}{=} \quad Pr\{\hat{H}(\mathbf{y}|\mathbf{x}_1) \leq H(Y|X_1) - \frac{\delta}{2}\} + Pr\{\hat{H}(\mathbf{y}|\mathbf{x}) \geq H(Y|X) + \frac{\delta}{2}\} \tag{21}$$

where the equality (a) is obtained by subtracting $\hat{H}(\mathbf{x}) = \hat{H}(\mathbf{x}_1)$ from both sides, inequality (b) is given by determining an arbitrary λ, and (c) is obtained by defining $\lambda \stackrel{\triangle}{=} H(Y|X) + \frac{\delta}{2}$ and use of (19) and (20). The two probabilities in (21c) are probabilities of rare events and therefore exponentially small, as is well known from large deviations theory (see e.g. [24, Th. 12.4.1 and Lemma 12.6.1]). Thus $\overline{P}_{e,u} \leq 2^{-n\epsilon}$ for some $\epsilon > 0$. Seeing that the optimal ML decision rule cannot be worse than the universal rule, part *a* of the Theorem is proved for the case of two hypotheses.

2) Proof of part b for simple hypotheses:

The average probability of error $\overline{P}_e$ for both the ML and the universal decision rules (3) and (6) can be expressed by type counting. As the vectors are i.i.d. (random coding and memoryless channel), vectors belonging to the same typical set have equal probability. From (15), vectors from the same conditional Markov type $T_{\mathbf{y}|\mathbf{x}_0,\mathbf{x}_1}$ have equal empirical probabilities $q_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}}(a_0,a_1,b)$ and therefore identical marginal probabilities $q_{\mathbf{x}_0,\mathbf{y}}(a_0,b)$ and $q_{\mathbf{x}_1,\mathbf{y}}(a_1,b)$. Therefore, for these decision rules, vectors belonging to the same type set $T_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}}$, give the same decision. Thus, the average error probability is expressed as sums of types of this kind. First, we count the number of $\mathbf{y}$ vectors, for a given $\mathbf{x}$ vector, that give a decision error $|\{\mathbf{y} \in T_{\mathbf{y}|\mathbf{x}} : error\}|$. As all $\mathbf{y}$ vectors belonging to the same type $T_{\mathbf{y}|\mathbf{x}}$ have equal probability $Pr\{\mathbf{y} \in T_{\mathbf{y}|\mathbf{x}}|\mathbf{x}\}$,

$$\begin{aligned}
\overline{P}_e &= \sum_{\mathbf{x}\in\mathcal{X}^n} \sum_{\mathbf{y}\in\mathcal{Y}^n} P(\mathbf{x})Q(\mathbf{y}|\mathbf{x})\mathbf{1}_{\mathbf{e}}(\mathbf{x},\mathbf{y}) = \sum_{\mathbf{x}\in\mathcal{X}^n} P(\mathbf{x}) \sum_{T_{\mathbf{y}|\mathbf{x}}} Pr\{\mathbf{y} \in T_{\mathbf{y}|\mathbf{x}}|\mathbf{x}\} \; |\{\mathbf{y} \in T_{\mathbf{y}|\mathbf{x}} : error\}| \\
&= \sum_{\mathbf{x}\in\mathcal{X}^n} P(\mathbf{x}) \sum_{T_{\mathbf{y}|\mathbf{x}}} Pr\{\mathbf{y} \in T_{\mathbf{y}|\mathbf{x}}|\mathbf{x}\} \sum_{\substack{T_{\mathbf{y}|\mathbf{x}_0,\mathbf{x}_1}: \\ T_{\mathbf{y}|\mathbf{x}_0,\mathbf{x}_1} \subseteq T_{\mathbf{y}|\mathbf{x}},\ error}} |T_{\mathbf{y}|\mathbf{x}_0,\mathbf{x}_1}|.
\end{aligned} \tag{22}$$

The last equality is obtained by counting the conditional Markov types $T_{\mathbf{y}|\mathbf{x}_0,\mathbf{x}_1}$ for which all the vectors $\mathbf{y}$ belonging to this type give an error. Note that $Pr\{\mathbf{y} \in T_{\mathbf{y}|\mathbf{x}}|\mathbf{x}\}$ is dependent on $\mathbf{x}$ only via its zero order Markov type $T_{\mathbf{x}}$. Now, by counting the $\mathbf{x}$ vectors by their first order Markov types $T_{\mathbf{x}_0,\mathbf{x}_1}$ and noting that the probability of a $\mathbf{x}$ vector in this type $Pr\{\mathbf{x} \in T_{\mathbf{x}_0,\mathbf{x}_1}\}$ is equal to the probability of it being in its marginal

type $T_{\mathbf{x}}$, namely $Pr\{\mathbf{x} \in T_{\mathbf{x}}\}$, the following arises:

$$\overline{P}_e = \sum_{T_{\mathbf{x}_0,\mathbf{x}_1}} Pr\{\mathbf{x} \in T_{\mathbf{x}_0,\mathbf{x}_1}\}|T_{\mathbf{x}_0,\mathbf{x}_1}| \sum_{T_{\mathbf{y}|\mathbf{x}}} Pr\{\mathbf{y} \in T_{\mathbf{y}|\mathbf{x}}|\mathbf{x} \in T_{\mathbf{x}}\} \sum_{\substack{T_{\mathbf{y}|\mathbf{x}_0,\mathbf{x}_1}: \\ T_{\mathbf{y}|\mathbf{x}_0,\mathbf{x}_1} \subseteq T_{\mathbf{y}|\mathbf{x}},\ error}} |T_{\mathbf{y}|\mathbf{x}_0,\mathbf{x}_1}|$$

$$= \sum_{T_{\mathbf{x}}} Pr\{\mathbf{x} \in T_{\mathbf{x}}\} \sum_{\substack{T_{\mathbf{x}_0,\mathbf{x}_1}: \\ T_{\mathbf{x}_0,\mathbf{x}_1} \subseteq T_{\mathbf{x}}}} |T_{\mathbf{x}_0,\mathbf{x}_1}| \sum_{T_{\mathbf{y}|\mathbf{x}}} Pr\{\mathbf{y} \in T_{\mathbf{y}|\mathbf{x}}|\mathbf{x} \in T_{\mathbf{x}}\} \sum_{\substack{T_{\mathbf{y}|\mathbf{x}_0,\mathbf{x}_1}: \\ T_{\mathbf{y}|\mathbf{x}_0,\mathbf{x}_1} \subseteq T_{\mathbf{y}|\mathbf{x}},\ error}} |T_{\mathbf{y}|\mathbf{x}_0,\mathbf{x}_1}|. \quad (23)$$

Rearranging the above expression,

$$\overline{P}_e = \sum_{T_{\mathbf{x}}} \sum_{T_{\mathbf{y}|\mathbf{x}}} Pr\{\mathbf{x} \in T_{\mathbf{x}}\} Pr\{\mathbf{y} \in T_{\mathbf{y}|\mathbf{x}}|\mathbf{x} \in T_{\mathbf{x}}\} \sum |T_{\mathbf{x}_0,\mathbf{x}_1}||T_{\mathbf{y}|\mathbf{x}_0,\mathbf{x}_1}| \quad (24)$$

where the right summation is over all types $T_{\mathbf{x}_0,\mathbf{x}_1}$ and $T_{\mathbf{y}|\mathbf{x}_0,\mathbf{x}_1}$ such that $T_{\mathbf{x}_0,\mathbf{x}_1} \subseteq T_{\mathbf{x}}$, $T_{\mathbf{y}|\mathbf{x}_0,\mathbf{x}_1} \subseteq T_{\mathbf{y}|\mathbf{x}}$ and give a decision error. As typical sets are defined by their empirical PMF, the summation over typical sets is exchanged with summation over the appropriate empirical PMF's. Defining $V_o(q_{\mathbf{x},\mathbf{y}})$ and $V_u(q_{\mathbf{x},\mathbf{y}})$ as the number of vectors with the joint PMF $q_{\mathbf{x},\mathbf{y}}$ that give an error (i.e. the right summation in the above expression) for the ML and universal decision rules respectively,

$$\begin{aligned}
\overline{P}_{e,u} &= \sum_{q_{\mathbf{x},\mathbf{y}}} \prod_{a \in \mathcal{X}} \prod_{b \in \mathcal{Y}} (P(a)Q(b|a))^{nq_{\mathbf{x},\mathbf{y}}(a,b)} V_u(q_{\mathbf{x},\mathbf{y}}) \\
&= \sum_{q_{\mathbf{x},\mathbf{y}}} \prod_{a \in \mathcal{X}} \prod_{b \in \mathcal{Y}} (P(a)Q(b|a))^{nq_{\mathbf{x},\mathbf{y}}(a,b)} V_o(q_{\mathbf{x},\mathbf{y}}) \frac{V_u(q_{\mathbf{x},\mathbf{y}})}{V_o(q_{\mathbf{x},\mathbf{y}})} \\
&\leq \sum_{q_{\mathbf{x},\mathbf{y}}} \prod_{a \in \mathcal{X}} \prod_{b \in \mathcal{Y}} (P(a)Q(b|a))^{nq_{\mathbf{x},\mathbf{y}}(a,b)} V_o(q_{\mathbf{x},\mathbf{y}}) \max_{q_{\mathbf{x},\mathbf{y}}} \left\{ \frac{V_u(q_{\mathbf{x},\mathbf{y}})}{V_o(q_{\mathbf{x},\mathbf{y}})} \right\} \\
&= \overline{P}_{e,o} \max_{q_{\mathbf{x},\mathbf{y}}} \left\{ \frac{V_u(q_{\mathbf{x},\mathbf{y}})}{V_o(q_{\mathbf{x},\mathbf{y}})} \right\}. \quad (25)
\end{aligned}$$

Lemma 1

$$\max_{q_{\mathbf{x},\mathbf{y}}} \left\{ \frac{V_u(q_{\mathbf{x},\mathbf{y}})}{V_o(q_{\mathbf{x},\mathbf{y}})} \right\} \leq 2^{n\epsilon_n}. \quad (26)$$

where $\lim_{n\to\infty} \epsilon_n = 0$.

The proof of the Lemma will be given in Appendix A.

With Lemma 1 and (25) the second part of the Theorem (8) is proved for the single DMC case, when considering simple hypotheses.

3) Extension of the proof to finite multiple hypotheses:

Previously the proof of the Theorem was given for only one relative shift of the two vectors. For the case of a finite number of possible shifts, the proof can easily be updated as follows: Suppose we would like to calculate the probability of error for two shifts (only two hypotheses are considered: no shift and two shifts). By rearranging the two vectors $\mathbf{x}$ and $\mathbf{y}$, taking the odd indices first and then the even vector indices, we obtain the one shift problem except for an extra end effect in the middle of the new vectors. ¿From [15, Lemma 1.2.7] explained at the end of Appendix A, this does not affect the error exponent for large enough n. Similarly, for any finite shift the asymptotic error exponent is equal to that for one shift. Upper bounding $\overline{P}_{e,u}$ by using the union bound, and lower bounding the ML average error by considering only one possible shift, the probability exponents are still asymptotically equal for any finite number of shifts.

B *Double DMC*

Consider the problem of passive tracking. It is of interest to estimate the relative delay between a signal received at two sensors. Assume that the signal $\mathbf{x}$, radiated from the target, is i.i.d. Let the two channels be memoryless but not necessarily identical or independant, i.e.,

$$R(\mathbf{y},\mathbf{z}|\mathbf{x}) = \prod_{i=1}^{n} R(y_i, z_i|x_i). \tag{27}$$

Taking a Bayesian approach to the decision rule, calculating the expectation of $R(\mathbf{y},\mathbf{z}|\mathbf{x})$ with respect to $\mathbf{x}$ and substituting (27), we get

$$\begin{aligned} ER(\mathbf{y},\mathbf{z}|\mathbf{x}) &= \sum_{\mathbf{x}\in\mathcal{X}^n} P(\mathbf{x})R(\mathbf{y},\mathbf{z}|\mathbf{x}) = \sum_{\mathbf{x}\in\mathcal{X}^n}\prod_{i=1}^{n} P(x_i)R(y_i,z_i|x_i) \\ &= \prod_{i=1}^{n}\left(\sum_{x\in\mathcal{X}} P(x)R(y_i,z_i|x)\right). \end{aligned} \tag{28}$$

The last line is obtained by changing the order of summation and multiplication and we get an i.i.d. probability function of $\mathbf{y}$ and $\mathbf{z}$. The ML decision rule is therefore,

$$\hat{\tau}_o = \arg\max_i J(\mathbf{y}|\mathbf{z}_i) \tag{29}$$

where $J(\cdot|\cdot)$ is the conditional probability of $\mathbf{y}$ given $\mathbf{z}$.

The universal decision rule is similar to the single DMC case, but now the vector $\mathbf{x}$ is replaced by a noisy version, namely $\mathbf{z}$, the output of the second channel. This test is based on the joint empirical entropy between $\mathbf{y}$ and $\mathbf{z}$ where the joint empirical entropy of $\mathbf{y}$ and $\mathbf{z}$ is defined similar to (4) and (5) by substituting $\mathbf{x}$ with $\mathbf{z}$,

$$\hat{\tau}_u = \arg\min_i \hat{H}(\mathbf{y}, \mathbf{z}_i). \tag{30}$$

Note the similarity between the decision rules for this case (29) (30) and the first case (3) (6). In this case, after averaging over $\mathbf{x}$, one can think of the two channels as a single channel from $\mathbf{z}$ to $\mathbf{y}$. Thus the intuition behind this rule is the same as for the single DMC problem. For this problem the average probability of error $\overline{P}_e$ is similar to (7) by changing $\mathbf{x}$ with $\mathbf{z}$. For this problem the above Theorem is valid.

The proof of the Theorem for the double DMC delay estimation problem is the same as that of the case of a single DMC, where the vector $\mathbf{x}$ is exchanged with a noisy vector $\mathbf{z}$ of the second channel. Changing $\mathbf{x}$ with $\mathbf{z}$ will give the proof of both parts of the Theorem.

III Modular Additive Finite-State Channel (MAFSC)

In this Section the single channel delay estimation problem will be solved for the class of MAFSC's. Consider the class of finite-state noise vectors characterized by a probability function of the form

$$M(\mathbf{w}) = \prod_{i=1}^{n} K(w_i|s_i) \tag{31}$$

where $\mathbf{w} = (w_1, w_2, ..., w_n)$, w_i is the channel noise at the ith time instant, $w_i \in \mathcal{W}$, $\mathbf{w} \in \mathcal{W}^n$, $|\mathcal{W}| < \infty$; s_i is the state of the noise source at the ith instant, $s_i \in \mathcal{S}$, $|\mathcal{S}| < \infty$; $s_{i+1} = f(w_i, s_i)$ where f is the next-state function, that maps $\mathcal{S} \times \mathcal{W}$ into $\mathcal{S}$, and $s_1 \in \mathcal{S}$ is the initial state. The class of finite alphabet MAFSC's is

described by

$$\mathbf{y} = \mathbf{x} \oplus \mathbf{w} \tag{32}$$

where w, y and x have the same alphabet size $|\mathcal{W}|$. Representing the alphabet letters by integers 0 to $|\mathcal{W}|-1$ the modulo addition operation $y = x \oplus w$ is defined as $y = (x + w) \bmod |\mathcal{W}|$. Similarly, the modulo subtraction operation $w = y \ominus x$ is equal to $w = (y - x) \bmod |\mathcal{W}|$. An interesting example of this class of channels is the binary symmetric channel, where the noise is a finite-state process. Another possible application for this case is to approximate, for large signal-to-noise ratio, an additive finite-state channel. The transition probability is therefore equal to the probability of the noise vector, i.e.,

$$Q(\mathbf{y}|\mathbf{x}) = M(\mathbf{w}). \tag{33}$$

In this case it is assumed that the input vector $\mathbf{x}$ is governed by uniform distribution.

The ML decision rule for the above problem is the shift of the vector $\mathbf{x}$ for which maximum probability is obtained. Let $\mathbf{w}_i$ denote the vector obtained by the modulo subtraction of the shifted vector $\mathbf{x}_i$ from $\mathbf{y}$,

$$\mathbf{w}_i = \mathbf{y} \ominus \mathbf{x}_i = \mathbf{w} \oplus \mathbf{x}_0 \ominus \mathbf{x}_i. \tag{34}$$

The ML decision rule is therefore

$$\hat{\tau}_o = \arg\max_i M(\mathbf{w}_i). \tag{35}$$

The proposed universal decision rule is based on the LZ compression algorithm [17]. A brief description of the algorithm is given. The sequence $\mathbf{w}$ is divided into phrases. As we move along the vector $\mathbf{w}$ we parse it into the shortest subsequences that differ from all previous phrases. Each phrase (except maybe the last one) differs from all the others and is a concatenation of a previous phrase with one extra symbol. The encoding of each phrase consists of sending the index of this past phrase and coding the extra character. Let $C_l(\mathbf{w})$ denote the number of phrases of length l in the incremental parsing of the vector $\mathbf{w}$. The LZ complexity is now defined as

$$U_{LZ}(\mathbf{w}) \triangleq \frac{1}{n} \sum_{l=1}^{l_{max}} C_l(\mathbf{w}) \log(C_l(\mathbf{w})) \tag{36}$$

where l_{max} is the length of the longest phrase. The universal decision rule is the vector $\mathbf{w}_i$ with the smallest

LZ complexity,

$$\hat{\tau}_u = \arg\min_i U_{LZ}(\mathbf{w}_i). \tag{37}$$

As will be shown in Appendix B, the LZ complexity $U_{LZ}(\mathbf{w})$ can be seen as the number of bits per symbol needed to compress the sequence $\mathbf{w}$. For the wrong shift, as will be shown presently, the vector $\mathbf{w}_i$ obtained from (34) is of uniform distribution and thus on average incompressible. For the correct shift, we obtain a vector $\mathbf{w}$ characterized by the additive noise distribution and thus compressible to the entropy of the noise source. The compressibility of a sequence, using the efficient LZ compression algorithm, can be used to distinguish between a uniformly distributed sequence (incompressible) and a random sequence from a finite state source which is compressible.

For this problem, and $\overline{P}_e$ given by (7), the Theorem holds.

Proof of the Theorem for the MAFSC:

As before the proof of the Theorem will first be given for simple hypotheses. Extension of the proof to multiple hypotheses is similar to the previous case and will thus be omitted. The proof of this case is similar to the proof by Ziv for universal decoding [16]. The probability of error will be calculated for a given noise vector $\mathbf{w}$, and then averaged over all the input vectors $\mathbf{x}$. ¿From (34), we have $\mathbf{w}_1 = \mathbf{w} \oplus \mathbf{x}_0 \ominus \mathbf{x}_1$. As $\mathbf{x}$ is chosen randomly with uniform probability, ignoring end effects, the vector $\mathbf{x}_0 \ominus \mathbf{x}_1$ is also of uniform distribution. Thus, for a given noise vector $\mathbf{w}$ added modulo to a uniformly distributed vector, gives a vector $\mathbf{w}_1$ of uniform distribution (with probability $M(\mathbf{w}_1) = 2^{-n \log |\mathcal{W}|}$ for all $\mathbf{w}_1$).

For both parts of the proof the following Lemma, which is proved in Appendix B, is needed. This Lemma is similar to [16, Lemma 2] which was given for joint parsing of a pair of sequences.

Lemma 2

The number of sequences $\mathbf{w}_1 \in \mathcal{W}^n$ such that $U_{LZ}(\mathbf{w}_1) \leq D$ is no more than $2^{n[D+O(\frac{\log\log n}{\log n})]}$.

1) Proof of part a for simple hypotheses:

Suppose that the finite-state noise vector does not have maximum entropy, i.e.,

$$H(W) = \log |\mathcal{W}| - \delta \tag{38}$$

where $H(W)$ is the normalized entropy of the finite-state noise source and $\delta > 0$. The average error probability associated with the universal decision rule can then be bounded for some arbitrary constant λ,

$$\begin{aligned}
\overline{P}_{e,u} &= Pr\{U_{LZ}(\mathbf{w}_1) \leq U_{LZ}(\mathbf{w})\} \\
&\leq Pr\{U_{LZ}(\mathbf{w}_1) \leq \lambda\} + Pr\{U_{LZ}(\mathbf{w}) \geq \lambda\} \\
&= Pr\{U_{LZ}(\mathbf{w}_1) \leq \log|\mathcal{W}| - \frac{\delta}{2}\} + Pr\{U_{LZ}(\mathbf{w}) \geq H(W) + \frac{\delta}{2}\}. \qquad (39)
\end{aligned}$$

The last equality is given by choosing $\lambda = H(W) + \frac{\delta}{2}$ and (38). For the first expression, remembering that all vectors $\mathbf{w}_1$ have probability $2^{-n\log|\mathcal{W}|}$ and using Lemma 2, it is bounded by $2^{-n[\frac{\delta}{2} - O(\frac{\log\log n}{\log n})]}$ which decreases exponentially with n. ¿From [25, Th. 2] and [26, Th. 1] $U_{LZ}(\mathbf{w}) \leq -\frac{1}{n}\log M(\mathbf{w}) + O(\frac{1}{\log n})$. Thus the second expression, for large enough n, is bounded by $Pr\{\frac{1}{n}\log M(\mathbf{w}) - H(W) \geq \delta'\}$ where $\delta' > 0$. ¿From the large deviations theory this is the probability of a non-typical set and thus exponentially small. As before, the optimal ML decision rule gives an average probability of error not greater than $\overline{P}_{e,u}$, thus part *a* is proved.

2) Proof of part b for simple hypotheses:

For a given noise vector $\mathbf{w}$, $V_o(\mathbf{w})$ and $V_u(\mathbf{w})$ are defined as the number of input vectors that give an error for ML and universal decision rules respectively,

$$\begin{aligned}
V_o(\mathbf{w}) &= |\{\mathbf{w}_1 \in \mathcal{W}^n : \ M(\mathbf{w}_1) \geq M(\mathbf{w})\}| \\
V_u(\mathbf{w}) &= |\{\mathbf{w}_1 \in \mathcal{W}^n : \ U_{LZ}(\mathbf{w}_1) \leq U_{LZ}(\mathbf{w})\}|. \qquad (40)
\end{aligned}$$

Similarly to (21),

$$\begin{aligned}
\overline{P}_{e,u} &= \sum_{\mathbf{x}\in\mathcal{X}^n} P(\mathbf{x}) \sum_{\mathbf{y}\in\mathcal{Y}^n} Q(\mathbf{y}|\mathbf{x})\mathbf{1}_{\mathbf{e}}(\mathbf{x},\mathbf{y}) \ = \ 2^{-n\log|\mathcal{W}|} \sum_{\mathbf{w}\in\mathcal{W}^n} M(\mathbf{w})V_u(\mathbf{w}) \\
&\leq 2^{-n\log|\mathcal{W}|} \sum_{\mathbf{w}\in\mathcal{W}^n} M(\mathbf{w})V_o(\mathbf{w}) \max_{\mathbf{w}}\left\{\frac{V_u(\mathbf{w})}{V_o(\mathbf{w})}\right\} \ = \ \overline{P}_{e,o} \max_{\mathbf{w}}\left\{\frac{V_u(\mathbf{w})}{V_o(\mathbf{w})}\right\}. \qquad (41)
\end{aligned}$$

Proving,

$$\max_{\mathbf{w}}\left\{\frac{V_u(\mathbf{w})}{V_o(\mathbf{w})}\right\} \leq 2^{n\epsilon_n} \qquad (42)$$

where $\lim_{n\to\infty} \epsilon_n = 0$ will complete the proof of the second half of the Theorem.

Lemma 3

$$\log V_o(\mathbf{w}) \geq n[U_{LZ}(\mathbf{w}) - O(\tfrac{1}{\log n}) \log |\mathcal{S}|^2]. \tag{43}$$

Lemma 3 is proved in Appendix B and is similar to [16, Lemma 1].

Inserting Lemma 2 into $V_u(\mathbf{w})$ defined in (40) and choosing $D = U_{LZ}(\mathbf{w})$ we obtain

$$V_u(\mathbf{w}) \leq 2^{n[U_{LZ}(\mathbf{w}) + O(\frac{\log\log n}{\log n})]}. \tag{44}$$

Combining this and Lemma 3, (42) is obtained, where $\epsilon_n = O(\frac{\log\log n}{\log n})$ and thus part b of the Theorem proved for simple hypothesis.

Extension of the proof to multiple hypothesis, as mentioned before, is similar to the DMC case.

IV Discussion and Conclusions

The issue of universal delay estimation for the discrete single channel problem, where the random reference signal is known to the receiver, was addressed and proved in the previous sections for all i.i.d. finite alphabet channels and all modulo additive finite-state channels. The modulo additive i.i.d. channel is covered by both of these cases, and it can also be proved, in a similar way to the MAFSC case, for a universal decision rule taking the minimum empirical entropy of the noise vector, i.e., $\hat{\tau}_u = \min_i H(\mathbf{w}_i)$.

Several extensions of this problem to a larger class of channels is of more practical interest. It must be noted that the delay problem may not always be properly defined. Consider the example of a Markov channel, where the transition probability is a function of the current input and a finite number of previous inputs and outputs. Taking into account delays of less than the channel memory can be ambiguous, as delay may be an intrinsic part of the probability function. If one only considers delays that are longer than l, where l grows slowly with vector length (e.g. $l = O(\log\log n)$), we can apply the Theorem and get a similar result when looking at the joint empirical entropy of blocks of length l. This can be seen intuitively, because as l grows the blocks become less dependent. For the class of autoregressive channels, where the current output is a function of the present input and a finite number of previous outputs, the question is well defined. The proposition, in this instance, was not found to be a trivial modification.

The universal delay estimation for the two channel problem was proved in the previous section for all discrete i.i.d. channels (not necessary identical). Extensions to larger classes of ergodic channels with finite memory is well defined for any number of shifts, as long as the two channels are identical. This is due to the fact that the relative delay between the two channels is of interest, so any intrinsic delay in the probability function is canceled out. We were not able to prove any of these extensions. We were always impeded by the inability to assess the size of a Markov conditional type (e.g. $T_{\mathbf{x}_0,\mathbf{x}_1|\mathbf{y}}$), where permutations are counted of a vector with a Markov condition while a joint Markov empirical probability is maintained. For two Markov channels not essentially alike and delays longer than l (where l grows with n as before and for large enough l), the problem converges to the i.i.d. case and joint empirical entropy of l-blocks is considered.

In this paper the stubborn problem of universal delay estimation was tackled from an information-theoretic point of view. For the i.i.d. cases the universal decision rule was found to be the minimum joint empirical entropy. As the vectors and their shifts have essentially the same empirical entropy, this rule can be seen as a maximum empirical mutual information rule. For the MAFSC case, where the channel has memory, the universal rule made use of the LZ data compression algorithm.

Appendix A

Proof of Lemma 1

Inserting the bounds on typical sets (11) and (17) in $V_o(q_{\mathbf{x},\mathbf{y}})$,

$$V_o(q_{\mathbf{x},\mathbf{y}}) \geq \sum_{q^o_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}}} 2^{n[\hat{H}(\mathbf{x}_0,\mathbf{x}_1,\mathbf{y})-\hat{H}(\mathbf{x})-O(\frac{\log n}{n})]} \tag{A.1}$$

where $q^o_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}}$ fulfills three conditions: the empirical PMF $q^o_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}}$ has the marginal PMF $q_{\mathbf{x}_0,\mathbf{y}}$ ($\sum_{a_1} q^o_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}}(a_0,a_1,b) = q_{\mathbf{x}_0,\mathbf{y}}(a_0,b)\ \forall a_0 \in \mathcal{X}, b \in \mathcal{Y}$), its two marginal PMF's $q^o_{\mathbf{x}_0}$ and $q^o_{\mathbf{x}_1}$ must be equal ($q^o_{\mathbf{x}_0}(a) = q^o_{\mathbf{x}_1}(a)\ \forall a \in \mathcal{X}$) and the ML decision rule must give an error ($Q(\mathbf{y}|\mathbf{x}) \leq Q(\mathbf{y}|\mathbf{x}_1)$). Consider the case of $q^o_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}}$ to be equal $q^*_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}}$ where

$$q^*_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}}(a_0,a_1,b) \stackrel{\triangle}{=} \frac{q_{\mathbf{x}_0,\mathbf{y}}(a_1,b)q_{\mathbf{x}_0,\mathbf{y}}(a_0,b)}{q_{\mathbf{y}}(b)} \tag{A.2}$$

for all $q_{\mathbf{y}}(b) > 0$. For b such that $q_{\mathbf{y}}(b) = 0$, $q^*_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}}(a_0, a_1, b) \stackrel{\triangle}{=} 0$ for all a_0 and a_1. Clearly, this probability measure satisfies each of the above conditions, where the third condition is achieved with equality. Therefore considering only this possibility, a lower bound on $V_o(q_{\mathbf{x},\mathbf{y}})$ is obtained.

We shall now upper bound $V_u(q_{\mathbf{x},\mathbf{y}})$ and show that this bound is asymptotically equal to the lower bound of $V_o(q_{\mathbf{x},\mathbf{y}})$ obtained by (A.2). By substituting the upper bounds of typical sets (11) and (17) in $V_u(q_{\mathbf{x},\mathbf{y}})$, we obtain

$$V_u(q_{\mathbf{x},\mathbf{y}}) \leq \sum_{q^u_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}}} 2^{n[\hat{H}(\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}) - \hat{H}(\mathbf{x})]}, \tag{A.3}$$

where $q^u_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}}$ fulfills three conditions: the empirical PMF $q^u_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}}$ has the marginal PMF $q_{\mathbf{x}_0,\mathbf{y}}$ ($\sum_{a_1} q^u_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}}(a_0, a_1, b) = q_{\mathbf{x}_0,\mathbf{y}}(a_0, b)\ \forall a_0 \in \mathcal{X}, b \in \mathcal{Y}$), its two marginal PMF's $q^u_{\mathbf{x}_0}$ and $q^u_{\mathbf{x}_1}$ must be equal ($q^u_{\mathbf{x}_0}(a) = q^u_{\mathbf{x}_1}(a)\ \forall a \in \mathcal{X}$) and the universal decision rule must give an error ($\hat{H}(\mathbf{x}_1, \mathbf{y}) \leq \hat{H}(\mathbf{x}, \mathbf{y})$). The number of typical sets is bounded by $(n+1)^{|\mathcal{X}|^2|\mathcal{Y}|}$, thus

$$V_u(q_{\mathbf{x},\mathbf{y}}) \leq \max_{q^u_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}}} 2^{n[\hat{H}(\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}) - \hat{H}(\mathbf{x}) + O(\frac{\log n}{n})]}. \tag{A.4}$$

This maximization need only be done on $\hat{H}(\mathbf{x}_0, \mathbf{x}_1, \mathbf{y})$ as $\hat{H}(\mathbf{x})$ is determined by the first condition. From entropy inequalities we obtain

$$\max_{q^u_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}}} \hat{H}(\mathbf{x}_0, \mathbf{x}_1, \mathbf{y}) \leq \max_{q^u_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}}} \hat{H}(\mathbf{x}_0|\mathbf{x}_1, \mathbf{y}) + \max_{q^u_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}}} \hat{H}(\mathbf{x}_1, \mathbf{y}). \tag{A.5}$$

Consider the right most term of the above inequality. This expression $\max_{q^u_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}}} \hat{H}(\mathbf{x}_1, \mathbf{y})$ can be upper bounded by $\hat{H}(\mathbf{x}, \mathbf{y})$ from the third condition. Equality is obtained when $q^u_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}}$ has marginal PMF's $q^u_{\mathbf{x}_0,\mathbf{y}}(a, b) = q^u_{\mathbf{x}_1,\mathbf{y}}(a, b) \quad \forall a \in \mathcal{X}, b \in \mathcal{Y}$. The probability function $q^*_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}}$ of (A.2) has such marginal probabilities. Thus $q^u_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}} = q^*_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}}$ achieves this maximum (while fulfilling the first two conditions as before).

Now let us analyze the first expression on the right-hand side of (A.5). Maximum is also obtained for $q^u_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}} = q^*_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}}$ and can be seen as follows: because conditioning decreases entropy

$$\max_{q^u_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}}} \hat{H}(\mathbf{x}_0|\mathbf{x}_1, \mathbf{y}) \leq \hat{H}(\mathbf{x}_0|\mathbf{y}) \tag{A.6}$$

where $\hat{H}(\mathbf{x}_0|\mathbf{y})$ is given by the first condition. Inserting $q^u_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}} = q^*_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}}$, the inequality (A.6) is achieved with equality. Therefore, $q^u_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}} = q^*_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}}$ attains equality in (A.5) and consequently an upper bound for $V_u(q_{\mathbf{x},\mathbf{y}})$ (A.4). As $q^*_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}}$ of (A.2) gives a minimum for $V_o(q_{\mathbf{x},\mathbf{y}})$ and a maximum for $V_u(q_{\mathbf{x},\mathbf{y}})$, thus inserting $q^*_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}}$ in (26), Lemma 1 is proved where $\epsilon_n = (|\mathcal{X}| + |\mathcal{X}|^2 + |\mathcal{X}|^2|\mathcal{Y}|)\frac{\log(n+1)}{n} = O(\frac{\log n}{n})$.

Remark: In the proof of Lemma 1 two points were ignored. The first is the end effects of the two marginal PMF's derived from the Markov PMF $q_{\mathbf{x}_0,\mathbf{x}_1}$. We artificially assumed a cyclic shift, and thus obtained equal marginal PMF's. The second is that the probability function $q^*_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}}$ may not be attainable for vectors of length n. This was ignored as the author H.H.Munro (pseudonym *'Saki'*) once explained "*A little inaccuracy sometimes saves tons of explanations*" (from 'The Comments of Moung Ka') [27]. The effects of these points are negligible and do not affect the result. At worst we are only able to obtain a probability distribution $\tilde{q}_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}}$ that accomplishes

$$\sum_{a_0\in\mathcal{X}}\sum_{a_1\in\mathcal{X}}\sum_{b\in\mathcal{Y}} |q^*_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}}(a_0,a_1,b) - \tilde{q}_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}}(a_0,a_1,b)| \le \frac{C}{n} \tag{A.7}$$

for some constant C, depending on alphabet size only. From [15, Lemma 1.2.7] the exponent discrepancy is at most

$$|H(q^*_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}}) - H(\tilde{q}_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{y}})| \le |\mathcal{X}|^2|\mathcal{Y}|\frac{C}{n}\log\frac{n}{C} = O(\tfrac{\log n}{n}) \overset{n\to\infty}{\longrightarrow} 0. \tag{A.8}$$

The order of convergence of the exponent is the same $O(\frac{\log n}{n})$ and was thus neglected.

Appendix B

Proof of Lemma 2

Let us calculate the number of bits $L_{LZ}(\mathbf{w})$ needed to encode a vector $\mathbf{w}$ using the LZ algorithm. For each distinct phrase of length *l+1*, $\log^+ l$ ($\log^+ l \triangleq \log l + \log\log l + \ldots$ for all positive elements) bits are needed to encode the length of the prefix. Then $\lceil \log C_l(\mathbf{w}) \rceil$ bits are needed to encode the serial number of the prefix of length l. Once the prefix has been encoded, $\lceil \log|\mathcal{W}| \rceil$ bits are needed to encoded the extra symbol

of the phrase. The total length of the codeword is upper bounded by

$$L_{LZ}(\mathbf{w}) \leq \sum_{l=1}^{l_{max}} C_l(\mathbf{w})[\log C_l(\mathbf{w}) + 2\log l + 4 + \log|\mathcal{W}|]. \tag{B.1}$$

Let $C(\mathbf{w}) = \sum_{l=1}^{l_{max}} C_l(\mathbf{w})$ denote the total number phrases of the parsing of $\mathbf{w}$. Now, from the convexity of the logarithmic function

$$\sum_{l=1}^{l_{max}} C_l(\mathbf{w}) \log l = C(\mathbf{w}) \sum_{l=1}^{l_{max}} \frac{C_l(\mathbf{w})}{C(\mathbf{w})} \log l \leq C(\mathbf{w}) \log \frac{n}{C(\mathbf{w})}. \tag{B.2}$$

¿From [25, Th. 2]

$$C(\mathbf{w}) \leq O(\tfrac{n}{\log n}). \tag{B.3}$$

Thus

$$\frac{1}{n} \sum_{l=1}^{l_{max}} C_l(\mathbf{w}) \log l \leq O(\tfrac{\log\log n}{\log n}). \tag{B.4}$$

Therefore for the LZ coding scheme we have by (B.1), (B.4) and (36)

$$L_{LZ}(\mathbf{w}) \leq n[U_{LZ}(\mathbf{w}) + O(\tfrac{\log\log n}{\log n})]. \tag{B.5}$$

For $U_{LZ}(\mathbf{w}) \leq D$ the number of possible coded vectors is thus bounded by

$$2^{n[D+O(\frac{\log\log n}{\log n})]}. \tag{B.6}$$

Proof of Lemma 3

The probability $M(\mathbf{w})$ can be calculated by a finite-state machine defined by a quintuple $(\mathcal{S}, \mathcal{W}, \mathcal{D}, f, g)$, where $\mathcal{S}$ is the finite set of states, $\mathcal{W}$ is the input alphabet and and $\mathcal{D}$ is the output set. The function f is the next-state function that maps $\mathcal{S} \times \mathcal{W}$ into $\mathcal{S}$. The output function g maps $\mathcal{S} \times \mathcal{W}$ into $\mathcal{D}$. The machine is fed with a sequence $\mathbf{w}$ and emits the sequence $\mathbf{d} = (d_1, d_2, ..., d_n)$ where $d_i \in \mathcal{D}$ while going through a sequence of states $\mathbf{s} = (s_1, s_2, ..., s_n)$, $s_i \in \mathcal{S}$, where $s_{i+1} = f(s_i, w_i)$ and $d_i = -\log K(w_i|s_i)$. Therefore

$$-\log M(\mathbf{w}) = -\log \prod_{i=1}^{n} K(w_i|s_i) = \sum_{i=1}^{n} d_i. \tag{B.7}$$

¿From the incremental parsing of $\mathbf{w}$, according to the LZ compression algorithm, a set of phrases is obtained characterized by (s, s', l). Each phrase of length l has an initial state s and final state s'. Denote $C_l(\mathbf{w}|s, s')$ as the number of phrases with the same (s, s', l). By permutations of phrases, having the same initial and final states, different vectors of equal probability are obtained. Now by bounding $V_o(\mathbf{w})$ (40) by vectors that give equality

$$V_o(\mathbf{w}) \geq \prod_{l=1}^{l_{max}} \prod_{s,s'} C_l(\mathbf{w}|s, s'). \tag{B.8}$$

¿From the Stirling formula

$$\begin{aligned} \log V_o(\mathbf{w}) &\geq \sum_{l=1}^{l_{max}} \sum_{s,s'} C_l(\mathbf{w}|s, s')[\log C_l(\mathbf{w}|s, s') - \log e] \\ &= -\sum_{l=1}^{l_{max}} C_l(\mathbf{w}) \sum_{s,s'} \frac{C_l(\mathbf{w}|s, s')}{C_l(\mathbf{w})} \log \frac{C_l(\mathbf{w})}{C_l(\mathbf{w}|s, s')} + \sum_{l=1}^{l_{max}} C_l(\mathbf{w})(\log C_l(\mathbf{w}) - \log e). \end{aligned} \tag{B.9}$$

¿From the convexity of the logarithmic function and (36)

$$\begin{aligned} \log V_o(\mathbf{w}) &\geq \sum_{l=1}^{l_{max}} C_l(\mathbf{w}) \log C_l(\mathbf{w}) - \sum_{l=1}^{l_{max}} C_l(\mathbf{w}) \log |\mathcal{S}|^2 e \\ &= nU_{LZ}(\mathbf{w}) - C(\mathbf{w}) \log |\mathcal{S}|^2 e. \end{aligned} \tag{B.10}$$

Also from (B.3)

$$\log V_o(\mathbf{w}) \geq n[U_{LZ}(\mathbf{w}) - O(\tfrac{1}{\log n}) \log |\mathcal{S}|^2] \tag{B.11}$$

which completes the proof of Lemma 3.

References

[1] P. M. Woodward and I. L. Davies, "A theory of radar information", *Philosophical Mag.*, vol. 41, pp. 1001-1017, Oct. 1951.

[2] P. M. Woodward, "Information theory and the design of radar receivers", *Proc. IRE*, vol. 39, pp. 1521-1524, Dec. 1951.

[3] P. M. Woodward and I. L. Davies, "Information theory and inverse probability in telecommunications", *Proc. IEE*, vol. 99, *Part III*, pp. 37-44, Mar. 1952.

[4] I. L. Davies, "On determining the presence of signals in noise", *Proc. IEE*, vol. 99, *Part III*, pp. 45-51, Mar. 1952.

[5] P. M. Woodward, *Probability and Information Theory with Applications to Radar*, Second Ed., London, Pergamon Press, 1964.

[6] C. E. Shannon, "A mathematical theory of communication", *Bell Syst. Tech J.*, vol. 27, pp. 623-657, Oct. 1948.

[7] J. Ziv and M. Zakai, "Some lower bounds on signal parameter estimation", *IEEE Trans. Inform. Theory*, vol. IT-15, pp. 386-391, May 1969.

[8] M. Zakai and J. Ziv, "A generalization of the rate-distortion theory and applications", printed in C. Longo (Ed), *Information Theory: New Trends and Open Problems*, Wein, Springer-Verlag, 1975.

[9] G. W. Zeoli, "A lower bound on the data rate for synthetic aperture radar", *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 708-715, Nov. 1976.

[10] V. S. Frost and K. S. Shanmugan, "The information content of synthetic aperture radar images of terrain", *IEEE Trans. Aerospace Electron. Syst.*, vol. AES-19 , pp. 768-774, Sept. 1983.

[11] M. R. Bell, "Information theory and radar waveform design", *IEEE Trans. Inform. Theory*, vol. IT-39, pp. 1578-1597, Sept. 1993.

[12] R. E. Blahut, "Theory of remote surveillance algorithms", printed in R. E. Blahut, W. Miller and L. H. Wilcox (Eds), *Radar and Sonar, Part I*, New York, Springer-Verlag, 1991.

[13] H. L. Van Trees, *Detection Estimation and Modulation Theory, Part III*, New York, Wiley, 1971.

[14] V. D. Goppa, "Nonprobabilistic mutual information without memory", *Problems of Control and Information Theory*, vol. 4, pp. 97-102, 1975.

[15] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, New York, Academic Press, 1981.

[16] J. Ziv, "Universal decoding for finite-state channels", *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 453-460, July 1985.

[17] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding", *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 530-536, Sept. 1978.

[18] L. Davisson, G. Longo and A. Sgarro, "The error exponent for the noiseless encoding of finite ergodic Markov sources", *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 431-438, July 1981.

[19] I. Csiszár, T. M. Cover and B. Choi, "Conditional limit theorems under Markov conditioning", *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 788-801, Nov. 1987.

[20] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically observed statistics", *IEEE Trans. Inform. Theory*, vol. IT-35, pp. 401-408, Mar. 1989.

[21] P. Whittle, "Some distribution and moment formulae for the Markov chain", *J. Roy. Stat. Soc.*, ser. B vol. 17, pp. 235-242, 1955.

[22] P. Billingsley, "Statistical methods in Markov chains", *Ann. Math. Stat.*, vol. 32, pp. 12-40, Corrections: pp. 1343, 1961.

[23] A. Satt "Universal decoding for multi-channel systems and for finite-state channels", Ph.D. dissertation, Department of Electrical Engineering, Technion - Israel Institute of Technology, preprint.

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

[25] A. Lempel and J. Ziv, "On the complexity of finite sequences", *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 75-81, Jan. 1976.

[26] E. Plotnik, M. J. Weinberger and J. Ziv, "Upper bounds on the probability of sequences emitted by finite-state sources and on the redundancy of the Lempel-Ziv algorithm", *IEEE Trans. Inform. Theory*, vol. IT-38, pp. 66-72, Jan. 1992.

[27] cited by J. M. and M. J. Cohen (Eds), *A Dictionary of Modern Quotations*, pp. 201, Penguin Books, 1971.