On the Minimum Description Length Principle for Sources with Piecewise Constant Parameters

Neri Merhav, *Senior Member, IEEE*

Abstract—**Universal lossless coding in the presence of finitely many abrupt changes in the statistics of the source, at unknown points, is investigated. The minimum description length (MDL) principle is derived for this setting. In particular, it is shown that for any uniquely decipherable code, for almost every combination of statistical parameter vectors governing each segment, and for almost every vector of transition instants, the minimum achievable redundancy is composed from $0.5 \log n/n$ bits for each unknown segmental parameter and $\log n/n$ bits for each transition, where n is the length of the input string. This redundancy is shown to be attainable by a strongly sequential universal encoder, i.e., an encoder that does not utilize the knowledge of a prescribed value of n.**

Key Words—**Minimum description length, universal coding, sequential coding, segmentation, edge detection.**

I. Introduction

Universal lossless coding schemes are normally developed for classes of stationary or asymptotically stationary sources, ranging from parametric classes such as memoryless, Markov, and finite-state sources (see, e.g., [1]–[10]) to nonparametric classes, like the class of all stationary and ergodic sources over a given alphabet (see, e.g., [11]–[13]). In a nonstationary regime, a common approach is to estimate the current statistical parameters at every moment and to perform *dynamic* or *adaptive* Huffman coding (see, e.g., [14]–[17]).

In this correspondence, we adopt a simple parametric model for a class of nonstationary sources, and we are concerned with second-order optimality of universal coding schemes with respect to this class. Specifically, we assume an information source whose unknown statistical parameter vector is subject to jumps, i.e., abrupt changes, at *a priori* unknown time instants. In other words, the parameter vector of the source is piecewise constant in time. The main result here is an extension of Rissanen's minimum description length (MDL) principle to this model.

As an example, consider a $(k+1)$-ary sequence $x_1, \cdots, x_n$ drawn from a memoryless source whose vector of letter probabilities is held fixed at $\theta = \theta_1$ until time instant $t = m$ $(1 \le m \le n)$, but then jumps to θ_2, where it again remains constant until $t = n$, that is, a single transition in θ. This source can be characterized by the triplet $\psi = (\theta_1, \theta_2, \alpha)$ where $\alpha = m/n$ is the normalized transition point. Alternatively, one can think of α as a continuous-valued parameter taking values between 0 and 1, and $m = \lfloor n\alpha \rfloor$. We first show that under a suitable regularity condition, for every uniquely decipherable coding scheme [18] operating on length n input strings, the expected codeword length is essentially never less than

$$n\alpha H(\theta_1) + n(1-\alpha)H(\theta_2) + \frac{k}{2} \cdot \log(\alpha n) + \frac{k}{2} \cdot \log[(1-\alpha)n] + \log n, \quad (1)$$

except for a set of points ψ whose volume vanishes as n grows.

Manuscript received September 10, 1992; revised March 20, 1993.

The author is with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel.

IEEE Log Number 9212889.

Here, $H(\theta_1)$ and $H(\theta_2)$ are the per-letter entropies associated with the two segments. The first two terms form the least achievable compression ratio, even when ψ is completely known. The next two terms represent the extra redundancy that stem from universality in θ_1 and in θ_2, respectively (see, e.g., [7]). Note that, if $\alpha \in (0,1)$, i.e., m grows linearly with n, then $\log(\alpha n)$ and $\log[(1-\alpha)n]$ are both asymptotically equivalent to $\log n$ since $\log \alpha$ and $\log(1-\alpha)$ become negligible compared to $\log n$. Thus, these two terms can be essentially merged to $k \cdot \log n$. The last $\log n$ term expresses the penalty for not knowing m or, equivalently α. A point to observe here is that while each unknown segmental parameter θ_1 and θ_2 contributes essentially $0.5 \log n$ bits per component (as is well known from the MDL principle), the unknown transition point α contributes $\log n$ bits *without* the factor 0.5. The intuitive reason for this phenomenon is that the likelihood function $P_\psi(x_1, \cdots, x_n)$ is much more sensitive to perturbations (errors) in α than in the segmental parameters, and hence the former should be encoded in full resolution.

As an evidence of the special sensitivity of P_ψ to the transition instant, it will be shown that α can be estimated with an error that decays essentially as fast as n^{-1}, while the segmental parameters can normally be estimated at the rate of $n^{-1/2}$. In fact, this will be a key step in proving the above result (see Lemma 3 below).

It is easy to show that this lower bound on the expected codeword length is achievable. For instance, consider the following coding scheme. For each possible division of $x_1, \cdots, x_n$, i.e., for each possible value of m, encode $x_1, \cdots, x_m$ and $x_{m+1}, \cdots, x_n$ separately, each by a universal code for memoryless sources (see, e.g., [1]), and find which value of m yields the shortest codeword. Then, to encode the optimal m, use $\log n$ bits as m can take only n possible values. While the above-described scheme requires a prescan in order to find the best value of m, we will demonstrate a sequential encoder that attains (1) even without needing to prescribe n in advance, i.e., a *strongly sequential* scheme. Moreover, the redundancy term of (1) is attained in a pointwise manner for every n-tuple, and not merely on the average. It is interesting that the proposed scheme does not involve an explicit estimation of m.

These results extend to parametric classes of sources that are more general than the class of memoryless sources (within each segment) and to any fixed number q of segments. The extension of (1) will consist of the appropriate convex combination of segmental entropies, plus $0.5 \log n$ bits for each one of the k components of the segmental parameter and for each one of the q segments, plus $\log n$ bits for every one of the $q-1$ transitions.

From the lower bound and its achievability, it is apparent that this extension of (1) is the MDL for sources with piecewise fixed parameters, and as such, may serve as a guideline for data segmentation in certain applications, such as speech signal analysis (see, e.g., [19]–[21]), curve segmentation and edge detection in image processing (see, e.g., [22]–[26]), DNA segmentation in molecular analysis (see, e.g., [27], [28]), and others. The MDL criterion can be applied for simultaneously deciding how many different segments q there are (if *a priori* unknown), and determining the segment endpoints.

The outline of this correspondence is as follows. In Section II, we provide some notation and definitions. In Section III, we state and prove the lower bound on the expected codeword

0018-9448/93$03.00

length for sources with piecewise constant parameters. Finally, in Section IV, we show a few ways of achieving this bound and discuss their properties.

II. Notation and Definitions

Let $\{p_\theta\}$ be a parametric family of stationary probability mass functions (PMF's) of vectors whose components take on values in a finite set $\mathscr{A}$ with $|\mathscr{A}| = A$ letters. It is assumed that θ is a k-dimensional parameter vector taking on values in a compact set $\Theta \subset \mathbb{R}^k$. Let $x_1,\cdots,x_n$, $x_i \in \mathscr{A}$ be a sequence drawn from a PMF whose parameter θ takes on a particular value θ_1 from $t = 1$ to $t = m_1$; then $\theta = \theta_2$ from $t = m_1 + 1$ until $t = m_2$, and so on. Finally, from $t = m_{q-1}$ to $t = n$, θ is held at θ_q. The vectors $\{x_1,\cdots,x_{m_1}\}$, $\{x_{m_1+1},\cdots,x_{m_2}\},\cdots,\{x_{m_{q-1}+1},\cdots,x_n\}$ will be referred to as *segments*, and correspondingly, $\theta_1,\theta_2,\cdots,\theta_q$ will be called the *segmental parameters*. The extended vector $(\theta_1,\cdots,\theta_q)$ will be denoted by $\boldsymbol{\theta}$. It will be assumed that the different segments are statistically independent.

The regime of the asymptotics will be such that all segments grow linearly with n, that is, $\lim_{n\to\infty} m_i/n = \alpha_i \in (0,1)$ and $\alpha_{i+1} > \alpha_i$, for all i, as segments with an asymptotically vanishing relative length have a small effect. An asymptotically equivalent formulation is one for which, given $\alpha_1,\cdots,\alpha_{q-1}$, the transition instants are given by $m_i = \lfloor \alpha_i n \rfloor$, $i = 1,\cdots,q-1$. The parameters $\alpha_1,\cdots,\alpha_{q-1}$ will be referred to as the *asymptotic normalized transition instants* or simply the *transition parameters*, and the vector $(\alpha_1,\cdots,\alpha_{q-1})$ will be denoted by $\boldsymbol{\alpha}$. For convenience, we shall sometimes use $\alpha_0 \triangleq 0$ and $\alpha_q \triangleq 1$. The PMF of $x_1,\cdots,x_n$ is now completely defined by the combined parameter vector $\psi \triangleq (\boldsymbol{\theta}, \boldsymbol{\alpha})$, i.e.,

$$P_\psi(x_1,\cdots,x_n) = \prod_{i=1}^{q} p_{\theta_i}(x_{m_{i-1}+1},\cdots,x_{m_i}) \tag{2}$$

where $m_0 \triangleq 0$ and $m_q \triangleq n$.

The Cartesian product of two generic sets $\mathscr{U}$ and $\mathscr{V}$ and the rth Cartesian power of $\mathscr{U}$ will be denoted $\mathscr{U} \times \mathscr{V}$ and $\mathscr{U}^r$, respectively. $\mathscr{U}^c$ is the complement of $\mathscr{U}$. For a generic vector w, $\|w\|$ will denote the Euclidean norm in the appropriate space. The space $\Theta^q \times (0,1)^{q-1}$ of the extended parameter vector ψ will be denoted Ψ, where it should be kept in mind that $\alpha_1 < \alpha_2 < \cdots < \alpha_{q-1}$. Ψ_δ, for $\delta > 0$, will denote the subset of Ψ with $\|\theta_i - \theta_{i+1}\| \geq \delta$, for every $i = 1,\cdots,q-1$, and $\alpha_{i+1} - \alpha_i \geq \delta$ for all $i = 0,\cdots,q-1$. For $i < j$, the string $\{x_i,\cdots,x_j\}$ will be henceforth denoted by x_i^j. For a measurable event F, $1\{F\}$ will denote the indicator function, and $p_\theta\{F\}$ and $P_\psi\{F\}$ will denote probabilities of F under the segmental PMF p_θ and under $P_\psi(\cdot)$, respectively. Similarly, $E_\theta\{\cdot\}$ and $E_\psi\{\cdot\}$ will denote expectations under the two PMF's. The per-letter lth-order entropy associated with p_θ is defined as

$$H_l(\theta) = -\frac{1}{l} E_\theta \log p_\theta(x_1^l). \tag{3}$$

Finally, a *length function* $L_n(x_1^n)$ of a uniquely decipherable lossless code (see, e.g., [18]) is a map from $\mathscr{A}^n$ to the positive integers that satisfies Kraft's inequality

$$\sum_{x_1^n \in \mathscr{A}^n} 2^{-L_n(x_1^n)} \leq 1. \tag{4}$$

III. The Lower Bound

Throughout the paper, we shall assume the following regularity condition about the parametric family of segmental PMF's $\{p_\theta, \theta \in \Theta\}$.

(A) There exists an estimator $\hat{\theta} = f(x_1^l)$ such that for every positive integer r, there is a constant $K(r) > 0$ such that for every $\theta \in \Theta$, and all large enough l,

$$E_\theta\{\|\hat{\theta} - \theta\|^r\} \leq \frac{K(r)}{l^{r/2}}. \tag{5}$$

Condition (A) requires a fairly good estimator for the segmental parameter θ. To a certain extent, it is a stronger requirement than that of [7] where, in fact, only $\sqrt{n}$-consistency was required, namely, a uniform $O(1/\sqrt{n})$ decay rate of the estimation error. The reason for the more demanding condition here is that the identifiability of θ plays a role here, not only in the universal coding within each segment as in [7], but also for distinguishing between different segments and reliably estimating the segment boundaries. Nevertheless, it is not difficult to see that this condition holds, at least in the case where $\{p_\theta\}$ is the class of memoryless sources, where the components of θ are the letter probabilities. Here, the estimator given by the relative frequencies of the letters satisfies condition (A). This can be seen by recalling the well-known fact that for the Bernoulli process $\{y_i\}_{i=1}^l$, $y_i = 1\{x_i = a\}$, $a \in \mathscr{A}$, the rth moment $E_\theta|\Sigma_{i=1}^l y_i - lp_\theta(a)|^r$ is given by a polynomial in l whose degree is $r/2$, where the coefficient of the leading term is uniformly bounded in the simplex Θ. The same holds true for Markov sources, finite-state source, and other classes $\{p_\theta\}$ of practical interest.

The following is an extension of [7, Theorem 1, part (a)].

Theorem 1: Assume that condition (A) holds, and let $\{L_n(x_1^n)\}_{n\geq 1}$ be a sequence of uniquely decipherable lossless codes. Then, for every $\epsilon > 0$ and all large n,

$$E_\psi L_n(x_1^n) \geq \sum_{i=1}^{q} (m_i - m_{i-1}) H_{m_i - m_{i-1}}(\theta_i) + (1-\epsilon)\left(\frac{1}{2}kq + q - 1\right)\log n \tag{6}$$

for all points ψ, except for points in a set $A_\epsilon(n) \subset \Psi$ whose volume tends to zero as $n \to \infty$.

It has been explained in the Introduction that each transition parameter α_i contributes essentially $\log n$ rather than $0.5 \log n$ bits to the redundancy since $P_\psi(\cdot)$ is more sensitive to $\boldsymbol{\alpha}$ than to $\boldsymbol{\theta}$, and hence more encoding accuracy is required. The intuitive explanation for this difference in sensitivity is fairly simple. First, observe that for a typical sequence $x_1,\cdots,x_n$, the segmental likelihood function $\log p_\theta(\cdot)$, and hence also $\log P_\psi(\cdot)$, is in the vicinity of its maximum, and therefore, perturbations in $\boldsymbol{\theta}$ affect the likelihood function only via the *second*-order derivatives because the first-order derivatives are normally close to zero. In contrast, it is not difficult to see that the (one-sided) first-order derivatives of $\log P_\psi$ with respect to (w.r.t.) $\boldsymbol{\alpha}$ are *not* necessarily negligible near the maxima of the likelihood function, and therefore small perturbations in $\boldsymbol{\alpha}$ have a *first*-order rather than a second-order effect on the likelihood function. This raises the sensitivity to the transition parameters from $n^{-1/2}$ to n^{-1}. Indeed, as we show at the beginning of the proof of the above theorem, the transition parameters can be estimated with an error that decays almost as fast as n^{-1}, unlike the convergence rate of the segmental parameters, which is essentially $n^{-1/2}$ in most cases.

The remaining part of this section is devoted to the proof of the theorem.

Proof of Theorem 1: We first prove the joint existence of estimators for $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ with the convergence rates just described. This will be done in several steps, where in the first step

(Lemma 1), it is proved that under condition (A), it is possible to classify vectors according to whether they were drawn from the same segment, such that the error probability is sufficiently small. In the second stage (Lemma 2), it is shown that by chopping the data into small phrases and classifying these phrases into segments (using the above classification rule), an estimate of $\boldsymbol{\alpha}$ with the above-mentioned convergence rate is obtained. In the last phase (Lemma 3), it is demonstrated that once $\boldsymbol{\alpha}$ has been estimated this way, one can extract an appropriate estimate for $\boldsymbol{\theta}$ from the estimated segments. Having proved Lemma 3, the proof of the Theorem is similar to that of [7].

Lemma 1: If condition (A) holds, then there exists a sequence $\{\Omega_l\}_{l\geq 1}$ of subsets of $\mathscr{A}^l \times \mathscr{A}^l$ with the following two properties.

a) For every joint PMF $\mu_\theta(\boldsymbol{x}, \boldsymbol{y})$, whose marginals are $p_\theta(\boldsymbol{x})$ and $p_\theta(\boldsymbol{y})$, $\boldsymbol{x}, \boldsymbol{y} \in \mathscr{A}^l$,

$$\sum_{(\boldsymbol{x},\boldsymbol{y})\in\Omega_l^c} \mu_\theta(\boldsymbol{x}, \boldsymbol{y}) \leq \frac{2K(r)}{l^{r/4}}.$$

b) If $\theta_x, \theta_y \in \Theta$, then $\|\theta_x - \theta_y\| \geq 4l^{-1/4}$ implies that for every positive integer r and all large l,

$$\sum_{(\boldsymbol{x},\boldsymbol{y})\in\Omega_l} p_{\theta_x}(\boldsymbol{x}) p_{\theta_y}(\boldsymbol{y}) \leq \frac{2K(r)}{l^{r/4}}.$$

Proof: Let $\hat{\theta}_x = f(\boldsymbol{x})$, $\hat{\theta}_y = f(\boldsymbol{y})$, where $f(\cdot)$ is as in condition (A), and consider the decision rule given by $\Omega_l = \{(\boldsymbol{x}, \boldsymbol{y}): \|\hat{\theta}_x - \hat{\theta}_y\| \leq 2l^{-1/4}\}$. As for part a), if $\|\hat{\theta}_x - \hat{\theta}_y\| > 2l^{-1/4}$, then by the triangle inequality, either $\|\hat{\theta}_x - \theta\| > l^{-1/4}$ or $\|\hat{\theta}_y - \theta\| > l^{-1/4}$. Thus, by the union bound,

$$\sum_{(\boldsymbol{x},\boldsymbol{y})\in\Omega_l^c} \mu_\theta(\boldsymbol{x}, \boldsymbol{y}) \leq p_\theta\{\|\hat{\theta}_x - \theta\| > l^{-1/4}\} + p_\theta\left\{\|\hat{\theta}_y - \theta\| > l^{-1/4}\right\} = 2p_\theta\{\|\hat{\theta}_x - \theta\| > l^{-1/4}\}. \quad (7)$$

Now, by Markov's inequality and condition (A), the rightmost side of (7) is upper bounded by $2l^{r/4} \cdot K(r)/l^{r/2} = 2K(r)/l^{r/4}$, completing the proof of part a). Regarding part b), since $\|\theta_x - \theta_y\|$ is assumed larger than $4l^{-1/4}$, but $\|\hat{\theta}_x - \hat{\theta}_y\| \leq 2l^{-1/4}$ whenever $(\boldsymbol{x}, \boldsymbol{y}) \in \Omega_l^c$, then by the triangle inequality,

$$\|\hat{\theta}_x - \theta_x\| + \|\hat{\theta}_y - \theta_y\| \geq \|\theta_x - \theta_y\| - \|\hat{\theta}_x - \hat{\theta}_y\| \geq 4l^{-1/4} - 2l^{-1/4} = 2l^{-1/4},$$

which implies that either $\|\hat{\theta}_x - \theta_x\| \geq l^{-1/4}$ or $\|\hat{\theta}_y - \theta_y\| \geq l^{-1/4}$. Thus, part b) now follows similarly to part a) and completes the proof of Lemma 1.

Lemma 2: Fix $\epsilon \in (0, 1)$ and set $\delta_n = 4n^{-\epsilon/4}$. Under condition (A), there exists an estimator $\hat{\boldsymbol{\alpha}}$ such that for every $c > 0$, every sufficiently large n, and every $\psi \in \Psi_n \triangleq \Psi_{\delta_n}$,

$$P_\psi\{x_1^n: n^{1-\epsilon}\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\| > c\} \leq \frac{K(r_\epsilon)}{n} \quad (8)$$

where r_ϵ depends solely on ϵ.

Proof: We prove Lemma 2 by selecting a particular estimator $\hat{\boldsymbol{\alpha}}$. Given ϵ, parse the sequence x_1^n into $n^{1-\epsilon}$ phrases of length n^ϵ (assuming, without loss of generality, that these numbers are integers). Let $\boldsymbol{x}_i \in \mathscr{A}^{n^\epsilon}$, $i = 1, \cdots, n^{1-\epsilon}$ denote the ith phrase. Since $\psi \in \Psi_n$, and $\delta_n > n^{-\epsilon/4}$, it is clear that for all large n, the entire phrase $\boldsymbol{x}_1$ belongs to the first segment, and similarly, the last phrase $\boldsymbol{x}_{n^{1-\epsilon}}$ is in the last segment. For $i = 2, \cdots, n^{1-\epsilon} - 1$, we shall first classify $\{\boldsymbol{x}_i\}$ in accordance to their segments. To this end, we use a decision rule Ω_{n^ϵ}, that satisfies Lemma 1 with $l = n^\epsilon$; for example, the decision rule that is described in the proof of the above lemma. Every phrase $\boldsymbol{x}_i$ is marked by $b_i = 0$ (i.e., no transition) unless either $(\boldsymbol{x}_{i-1}, \boldsymbol{x}_i) \in \Omega_{n^\epsilon}^c$, or $(\boldsymbol{x}_i, \boldsymbol{x}_{i+1}) \in \Omega_{n^\epsilon}^c$, or $(\boldsymbol{x}_{i-1}, \boldsymbol{x}_{i+1}) \in \Omega_{n^\epsilon}^c$, in which case $\boldsymbol{x}_i$ is marked by $b_i = 1$ (i.e., a transition occurs). The reason for checking all combinations is associated with the fact that a transition might take place at an arbitrary point within a phrase, and hence this phrase does not belong entirely to any particular segment.

Consider now the resulting binary sequence of transition marks $\{b_i\}$. If one or several consecutive phrases are marked by $b_i = 1$, this will be interpreted as a *single* transition. The jth component of $\boldsymbol{\alpha}$ will be estimated as $\hat{\alpha}_j = \hat{m}_j/n$, where $\hat{m}_j$ corresponds to the midpoint of the string formed by the jth group of successive phrases, all marked by $b_i = 1$. If the number of such groups is larger than $q - 1$, then the excess is ignored. If it is smaller, the missing components of $\hat{\boldsymbol{\alpha}}$ are all set to an arbitrary value, say, 1.

We now show that the above estimator satisfies the assertion of the lemma. First observe that if all phrases are classified correctly, i.e., $b_i = 0$ for all internal phrases $\boldsymbol{x}_i$ that belong to the same segment as their two neighboring phrases $\boldsymbol{x}_{i-1}$ and $\boldsymbol{x}_{i+1}$, and at the same time, there is at least one and no more than three successive occurrences of $b_i = 1$ for phrases surrounding a true transition, then $\hat{m}_j$ cannot deviate from the true m_j by more than $1.5n^\epsilon$, which corresponds to a maximum estimation error of $1.5n^{-(1-\epsilon)}$ in α_j. (The constant 1.5 is obviously unimportant as it can be absorbed in ϵ.) Thus, it suffices to show that the probability of the event F of correct classification in the above-defined sense is eventually larger than $1 - K(r_\epsilon)/n$. Suppose that the real transitions occur in phrases $\boldsymbol{x}_{i_1}, \boldsymbol{x}_{i_2}, \cdots, \boldsymbol{x}_{i_{q-1}}$, and for convenience, also define $i_0 \triangleq 0$. Now, since $\|\theta_{i+1} - \theta_i\| \geq \delta_n = 4n^{-\epsilon/4}$, $i = 1, \cdots, q-1$ when $\psi \in \Psi_n$, then by applying Lemma 1 with $l = n^\epsilon$ and the union bound,

$$\begin{aligned} P_\psi\{F^c\} &= P_\psi\left\{\left[\bigcup_{j=1}^{q-1}\left(\bigcup_{i=i_{j-1}+2}^{i_j-2}\{b_i = 1\}\right) \cup \left(\bigcap_{i=i_j-1}^{i_j+1}\{b_i = 0\}\right)\right] \cup \{b_{i_{q-1}+2} = 1\}\right\} \\ &\leq \sum_{j=1}^{q-1}\sum_{i=i_{j-1}+2}^{i_j+2} P_\psi\{b_i = 1\} + \sum_{j=1}^{q-1} P_\psi\{b_{i_j} = 0\} + P_\psi\{b_{i_{q-1}+2} = 1\} \\ &\leq n^{1-\epsilon} \cdot 3 \cdot \frac{2K(r)}{n^{r\epsilon/4}} + (q-1) \cdot \frac{2K(r)}{n^{r\epsilon/4}} + 3 \cdot \frac{2K(r)}{n^{r\epsilon/4}} \\ &\leq \frac{K(r)}{n^{r\epsilon/4-1}}, \end{aligned} \quad (9)$$

provided that n is sufficiently large. Now, by selecting $r = r_\epsilon = \lceil 8/\epsilon \rceil$, the assertion of Lemma 2 is proved.

Lemma 3: Assume that condition (A) holds, and fix $c > 0$ and $\epsilon \in (0, 1)$. Then, there exists an estimator $\hat{\psi} = (\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\alpha}})$ such that for every $\psi \in \Psi_n$, and all large n,

$$P_\psi\left\{x_1^n: n^{0.5(1-\epsilon)}\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\| \geq c \cup n^{1-\epsilon}\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\| \geq c\right\} \leq \frac{G}{n} \quad (10)$$

for some constant G that depends only on ϵ, c, and q.

Proof: Again, the proof is constructive. For estimating $\boldsymbol{\alpha}$, we use any estimator that satisfies Lemma 2, say, the estimator described in the proof of that lemma. The estimator for each segmental parameter θ_i, $i = 1,\cdots,q$ will be given by the function $f(\cdot)$ as defined in (A), where the argument is a substring of the estimated segment starting at $\hat{m}_{i-1}$ and ending at $\hat{m}_i$. Specifically, let $v_i = (\hat{m}_{i-1} + \hat{m}_i)/2$ be the midpoint of the ith estimated segment. Then, $\hat{\theta}_i$ will be defined as $f(x_{v_i-0.5n^{1-\epsilon/2}},\cdots,x_{v_i+0.5n^{1-\epsilon/2}})$. At first glance, this way of estimating $\boldsymbol{\theta}$ might seem somewhat obscure. However, it is good enough to satisfy Lemma 3, and it appears easier to analyze than an estimator that uses the entire estimated segment from $\hat{m}_{i-1}$ to $\hat{m}_i$, as the latter operates on a random number of observations.

Let $\boldsymbol{m}$ and $\hat{\boldsymbol{m}}$ denote the vectors $(m_1,\cdots,m_{q-1})$ and $(\hat{m}_1,\cdots,\hat{m}_{q-1})$, respectively. Let $S(\boldsymbol{m})$ denote the set all $(q-1)$-dimensional vectors $\tilde{\boldsymbol{m}} = (\tilde{m}_1,\cdots,\tilde{m}_{q-1})$ with integer-valued components such that $\|\tilde{\boldsymbol{m}} - \boldsymbol{m}\| < cn^{\epsilon}$. Similarly, denote by $S(\boldsymbol{\theta})$ the Euclidean sphere of radius $c/n^{0.5(1-\epsilon)}$ centered at $\boldsymbol{\theta}$. Finally, to stress the dependency of $\hat{\boldsymbol{\theta}}$ on $\hat{\boldsymbol{m}}$, it will be denoted $\hat{\boldsymbol{\theta}} = g(x_1^n, \hat{\boldsymbol{m}})$, whereas $g(x_1^n, \tilde{\boldsymbol{m}})$ is understood similarly to $g(x_1^n, \hat{\boldsymbol{m}})$ but with v_i defined as the midpoint of the string with the *deterministic* endpoints $\tilde{m}_{i-1}$ and $\tilde{m}_i$. Now, the probability of the event of (10) is upper bounded as follows. First, by the union bound,

$$P_\psi\{g(x_1^n, \hat{\boldsymbol{m}}) \in S^c(\boldsymbol{\theta}) \cup \hat{\boldsymbol{m}} \in S^c(\boldsymbol{m})\} \le P_\psi\{g(x_1^n, \hat{\boldsymbol{m}}) \in S^c(\boldsymbol{\theta})\} + P_\psi\{\hat{\boldsymbol{m}} \in S^c(\boldsymbol{m})\} \quad (11)$$

where the second term is less than $K(r_\epsilon)/n$ by Lemma 2. As for the first term,

$$P_\psi\{g(x_1^n, \hat{\boldsymbol{m}}) \in S^c(\boldsymbol{\theta})\} \le P_\psi\{g(x_1^n, \tilde{\boldsymbol{m}}) \in S^c(\boldsymbol{\theta}) \cap \hat{\boldsymbol{m}} \in S(\boldsymbol{m})\} + P_\psi\{\hat{\boldsymbol{m}} \in S^c(\boldsymbol{m})\} \quad (12)$$

where, again, the second term does not exceed $K(r_\epsilon)/n$. The first term on the right-hand side of (12) can be upper bounded as follows:

$$\begin{aligned} &P_\psi\{g(x_1^n, \hat{\boldsymbol{m}}) \in S^c(\boldsymbol{\theta}) \cap \hat{\boldsymbol{m}} \in S(\boldsymbol{m})\} \\ &\quad = \sum_{\tilde{\boldsymbol{m}} \in S(\boldsymbol{m})} P_\psi\{g(x_1^n, \hat{\boldsymbol{m}}) \in S^c(\boldsymbol{\theta}) \cap \hat{\boldsymbol{m}} = \tilde{\boldsymbol{m}}\} \\ &\quad = \sum_{\tilde{\boldsymbol{m}} \in S(\boldsymbol{m})} P_\psi\{g(x_1^n, \tilde{\boldsymbol{m}}) \in S^c(\boldsymbol{\theta}) \cap \hat{\boldsymbol{m}} = \tilde{\boldsymbol{m}}\} \\ &\quad \le \sum_{\tilde{\boldsymbol{m}} \in S(\boldsymbol{m})} P_\psi\{g(x_1^n, \tilde{\boldsymbol{m}}) \in S^c(\boldsymbol{\theta})\}. \end{aligned} \quad (13)$$

Since the estimator $\hat{\theta}_i$ of each segmental parameter θ_i operates on a string, whose length $n^{1-\epsilon/2}$ is very small compared to the segment length (which, in turn, is never shorter than $n\delta_n > n^{1-\epsilon/4}$, provided that $\psi \in \Psi_n$), then for all large n, for $\tilde{\boldsymbol{m}} \in S(\boldsymbol{m})$, it is guaranteed that each estimator $\hat{\theta}_i$ is computed from a vector that is entirely within the appropriate segment. Furthermore, since p_θ is assumed stationary, it then follows that $P_\psi\{g(x_1^n, \tilde{\boldsymbol{m}}) \in S^c(\boldsymbol{\theta})\}$ is independent of $\tilde{\boldsymbol{m}}$ [as long as $\tilde{\boldsymbol{m}} \in S(\boldsymbol{m})$]. In particular, it is equal to $P_\psi\{g(x_1^n, \boldsymbol{m}) \in S^c(\boldsymbol{\theta})\}$. Thus, the right side of the inequality in (13) is given by

$$\sum_{\tilde{\boldsymbol{m}} \in S(\boldsymbol{m})} P_\psi\{g(x_1^n, \tilde{\boldsymbol{m}}) \in S^c(\boldsymbol{\theta})\} = |S(\boldsymbol{m})| \cdot P_\psi\{g(x_1^n, \boldsymbol{m}) \in S^c(\boldsymbol{\theta})\}. \quad (14)$$

Now, since $|S(\boldsymbol{m})| \le (cn^{\epsilon})^q$, we have by the union bound, Markov's inequality, assumption (A), and the fact that $\hat{\theta}_i$ is calculated from $l = n^{1-\epsilon/2}$ observations:

$$\begin{aligned} &|S(\boldsymbol{m})| \cdot P_\psi\{g(x_1^n, \boldsymbol{m}) \in S^c(\boldsymbol{\theta})\} \\ &\quad \le c^q n^{\epsilon q} \cdot \sum_{i=1}^{q} p_\theta\left\{n^{0.5(1-\epsilon)}\|\hat{\theta}_i - \theta_i\| > \frac{c}{\sqrt{q}}\right\} \\ &\quad \le c^q n^{\epsilon q} \cdot q \cdot \frac{q^{0.5r} n^{0.5r(1-\epsilon)}}{c^r} \cdot \frac{K(r)}{n^{0.5(1-\epsilon/2)r}} \end{aligned} \quad (15)$$

which, for $r \ge \lceil 4(q + 1/\epsilon)\rceil$, decays faster than $1/n$. By combining (11), (12), and (15), and by taking r at least as large as max $\{r_\epsilon, \lceil 4(q + 1/\epsilon)\rceil\}$, we complete the proof of Lemma 3.

The remaining part of the proof of Theorem 1 is almost a straightforward extension of the proof of [7, Theorem 1, part (a)], but it will be presented here for the sake of completeness. Let $\psi \in \Psi_n$ and denote

$$E_n(\psi) = \{\tilde{\psi} = (\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\alpha}}) : \|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\| \le cn^{-0.5(1-\epsilon)}, \|\tilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\| \le cn^{-(1-\epsilon)}\}. \quad (16)$$

Let $\hat{\psi}$ be an estimator of ψ that satisfies Lemma 3. Define the set of "typical" sequences

$$X_n(\psi) \triangleq \{x_1^n : \hat{\psi} \in E_n(\psi)\}, \quad (17)$$

and denote $P_\psi\{X_n(\psi)\}$ by $P_n(\psi)$. Lemma 3 guarantees that if $\psi \in \Psi_n$ and $\epsilon > 0$, then $P_n(\psi) > 1 - \epsilon$ for all large n. Let $L_n(\cdot)$ be a length function that satisfies Kraft's inequality (4), and denote

$$Q_n(\psi) \triangleq \sum_{x_1^n \in X_n(\psi)} 2^{-L_n(x_1^n)}. \quad (18)$$

Now, by Jensen's inequality and the nonnegativity of the Kullback–Leibler informational divergence,

$$T_n(\psi) \triangleq \sum_{x_1^n \in X_n(\psi)} P_\psi(x_1^n) \log \frac{P_\psi(x_1^n)}{2^{-L_n(x_1^n)}} \ge P_n(\psi) \log \frac{P_n(\psi)}{Q_n(\psi)}. \quad (19)$$

Now, let $B_\epsilon(n)$ be the set of ψ's in Ψ_n such that

$$T_n(\psi) < (1 - \epsilon) \log n^{(1-\epsilon)(0.5kq+q-1)}. \quad (20)$$

Let N_n denote the maximum number of *disjoint* neighborhoods $E_n(\psi)$ with centers at $B_\epsilon(n)$, and let C_n denote the set of centers. Because of the triangle inequality, it is easy to see that by doubling the radius of each neighborhood $E_n(\psi)$ [i.e., by replacing c by $2c$ in (16)], we obtain a cover of $B_\epsilon(n)$. Therefore, the volume V_n of $B_\epsilon(n)$ is bounded by

$$V_n \le DN_n\left[\frac{2c}{n^{0.5(1-\epsilon)}}\right]^{k^\theta} \cdot \left[\frac{2c}{n^{1-\epsilon}}\right]^{q-1} \quad (21)$$

where D is a constant depending only on k and q. From (19) and (20), we conclude that for every $\psi \in B_\epsilon(n)$ and all large n,

$$-\log Q_n(\psi) < \left[\frac{1-\epsilon}{P_n(\psi)} - \frac{\log P_n(\psi)}{\log n^{(1-\epsilon)(0.5kq+q-1)}}\right] \cdot \log n^{(1-\epsilon)(0.5kq+q-1)}. \quad (22)$$

By Lemma 3, for all large n, the expression in the brackets is less than $1 - \epsilon/2$, provided that $\psi \in \Psi_n$. This implies that

$$Q_n(\psi) > n^{-(1-\epsilon/2)(1-\epsilon)(0.5kq+q-1)}. \quad (23)$$

Since the sets $E_n(\psi)$ are disjoint by construction, then by Kraft's

inequality,

$$1 \geq \sum_{\psi \in C_n} Q_n(\psi) \geq N_n \cdot n^{-(1-\epsilon/2)(1-\epsilon)(0.5kq+q-1)} \quad (24)$$

which implies that $N_n \leq n^{(1-\epsilon/2)(1-\epsilon)(0.5kq+q-1)}$. Thus, by (21),

$$V_n \leq Dn^{(1-\epsilon/2)(1-\epsilon)(0.5kq+q-1)} \cdot \left[\frac{2c}{n^{0.5(1-\epsilon)}}\right]^{kq} \cdot \left[\frac{2c}{n^{1-\epsilon}}\right]^{q-1} \quad (25)$$

which tends to zero as $n \to \infty$. Next, observe that for every ψ in Ψ_n but outside $B_\epsilon(n)$, we have by (20)

$$\begin{aligned}(1-\epsilon)^2\left(\frac{1}{2}kq+q-1\right)\log n &\leq \sum_{x_1^n \in X_n(\psi)} P_\psi(x_1^n)\log\frac{P_\psi(x_1^n)}{2^{-L_n(x_1^n)}} \\ &\leq E_\psi L_n(x_1^n) - nH_n(\psi) \\ &\quad + \sum_{x_1^n \in X_n^c(\psi)} P_\psi(x_1^n)\log\frac{1}{P_\psi(x_1^n)} \quad (26)\end{aligned}$$

where $nH_n(\psi)$ is the unnormalized nth-order entropy associated with ψ given by the first summation on the right-hand side of (6). The second term on the rightmost side of the last expression is upper bounded as follows:

$$\begin{aligned}&\sum_{x_1^n \in X_n^c(\psi)} P_\psi(x_1^n)\log\frac{1}{P_\psi(x_1^n)} \\ &= [1-P_n(\psi)]\cdot\sum_{x_1^n \in X_n^c(\psi)}\frac{P_\psi(x_1^n)}{1-P_n(\psi)}\cdot\log\frac{1-P_n(\psi)}{P_\psi(x_1^n)} \\ &\quad + [1-P_n(\psi)]\log\frac{1}{1-P_n(\psi)} \\ &\leq [1-P_n(\psi)]\log|X_n^c(\psi)| + [1-P_n(\psi)]\log\frac{1}{1-P_n(\psi)} \\ &\leq [1-P_n(\psi)]n\log A + [1-P_n(\psi)]\log\frac{1}{1-P_n(\psi)} \\ &\leq G\log A + o(1) \quad (27)\end{aligned}$$

where the last step follows from Lemma 3. Thus, the last term on the rightmost side of (26) is absorbed in its leftmost side, and the assertion of Theorem 1 is proved (with ϵ replaced by 2ϵ) for $\psi \in \Psi_n \cap B_\epsilon^c(n)$ and all large n. The volume of the complementary set $A_\epsilon(n) \triangleq \Psi_n^c \cup B_\epsilon(n)$ clearly does not exceed the sum of the volumes of Ψ_n^c and $B_\epsilon(n)$, which both vanish with n. This completes the proof of Theorem 1.

IV. Achievability

For the sake of simplicity, we shall hereafter confine attention to the case of a single transition, namely, $q = 2$. The extension to a general value of q will be straightforward.

The conceptually simplest approach to achieving the lower bound, as given by Theorem 1, relies on the existence of a universal prefix code for the class of *segmental* PMF's $\{p_\theta\}$ that is optimal in the sense of [7], namely, a universal code with a length function $L_l(x_1^l)$, such that

$$E_\theta L_l(x_1^l) \leq lH_l(\theta) + (\tfrac{1}{2}+\epsilon)k\log l \quad (28)$$

for every $\theta \in \Theta$, $\epsilon > 0$, and l sufficiently large. Once equipped with such a code, we can find for each x_1^n the best value m^* of m in the sense of minimizing $L_m(x_1^m) + L_{n-m}(x_{m+1}^n)$. The code will be constructed from $\log n$ bits specifying m^*, followed by $\min_m(L_m(x_1^m) + L_{n-m}(x_{m+1}^n))$ bits for encoding the data themselves. This scheme obviously attains the lower bound for a single transition. However, an inherent limitation of this method is that it cannot be implemented in a sequential manner because one must first see the entire vector x_1^n before deciding on its optimal partition.

In certain situations of practical interest, the need of such a two-pass procedure can be avoided by forming a universal probability measure, which is a mixture of all PMF's in the class, and constructing a code that is optimal w.r.t. this measure (see, e.g., [29], [30]). Consider, for example, the class $\{p_\theta\}$ of memoryless sources where θ denotes the vector of $k = A-1$ freely chosen letter probabilities (which has immediate extensions to Markov sources and finite-state sources). For a given string x_r^l, $r \leq l$, let $n_t(a)$, $a \in \mathcal{A}$, $t = r,\cdots,l$ denote the count of occurrences of $x_j = a$, along $j = r,\cdots,t$. Let

$$\tilde{p}(x_t \mid x_r^{t-1}) = \frac{n_{t-1}(x_t) + 1/2}{t-r+A/2} \quad (29)$$

where $n_{r-1}(a) \triangleq 0$ and

$$\tilde{p}(x_r^l) = \prod_{t=r}^{l}\tilde{p}(x_t \mid x_r^{t-1}). \quad (30)$$

Finally, let

$$\bar{p}(x_1^n) = \frac{1}{n}\sum_{j=1}^{n}\tilde{p}(x_1^j)\tilde{p}(x_{j+1}^n) \quad (31)$$

where x_{n+1}^n is interpreted as the "empty" string whose probability under $\tilde{p}(\cdot)$ is defined as 1. Now, consider the Shannon code (see, e.g., [18]) w.r.t. $\bar{p}(\cdot)$, whose codeword length for each n-tuple is upper bounded as follows:

$$\begin{aligned}L_n(x_1^n) &= \lceil -\log\bar{p}(x_1^n)\rceil \\ &\leq -\log\bar{p}(x_1^n) + 1 \\ &\leq -\log\left[\frac{1}{n}\tilde{p}(x_1^m)\tilde{p}(x_{m+1}^n)\right] + 1 \\ &= [-\log\tilde{p}(x_1^m)] + [-\log\tilde{p}(x_{m-1}^n)] + \log n + 1. \quad (32)\end{aligned}$$

It is well known (see, e.g., [3]) that for any string x_1^m,

$$-\log\tilde{p}(x_1^m) \leq m\hat{H}(x_1^m) + \tfrac{1}{2}k\log m + O(1) \quad (33)$$

(and a similar relation for x_{m+1}^n), where $\hat{H}(x_1^m)$ is the empirical entropy associated with x_1^m, defined as

$$\hat{H}(x_1^m) = -\sum_{a \in \mathcal{A}}\frac{n_m(a)}{m}\log\frac{n_m(a)}{m}. \quad (34)$$

Since $E_{\theta_1}\hat{H}(x_1^m) \leq H(\theta_1)$, as can easily be seen, and similarly, $E_{\theta_2}\hat{H}(x_{m+1}^n) \leq H(\theta_2)$, the bound is attained. To implement this code sequentially, one may calculate the conditional measures $\bar{p}(x_t \mid x_1^{t-1})$, $t = 1,\cdots,n$, and use an arithmetic code w.r.t. $\{-\log\bar{p}(x_t \mid x_1^{t-1})\}_{t=1}^n$. However, since $\{\bar{p}\{x(\mid x_1^{t-1})\}_{t=1}^n$ all depend on the prescribed value of the block length n, then the code is only *weakly* sequential in the sense that n has to be known in advance.

To relax the necessity of knowing n *a priori*, one may use a slowly decaying nonuniform weighting on j rather than the uniform weighting $1/n$ in (31). For instance, let $\pi(j) = 1/j^{1+\epsilon}$, $C_n = \sum_{j=1}^n\pi(j)$, and $C_\infty = \sum_{j=1}^\infty\pi(j)$. Then, (31) can be modified

as follows:

$$\bar{p}(x_1^n) = C_\infty^{-1} \sum_{j=1}^{n-1} \pi(j)\bar{p}(x_1^j)\bar{p}(x_{j+1}^n) + \left(1 - \frac{C_{n-1}}{C_\infty}\right)\tilde{p}(x_1^n). \quad (35)$$

This can be interpreted as a mixture of PMF's with a prior on m given by $C_\infty^{-1}\pi(m)$. Thus, with probability $(1 - C_{n-1}/C_\infty)$, m might be at least as large as n, which means that no transition occurs in the first n symbols. It is easy to see that now the conditional probabilities associated with $\bar{p}$, as defined in (35), do not depend on n, and hence the block length need not be prescribed in advance. Again, one can show in a manner similar to (32) that an arithmetic code w.r.t. (35) attains the lower bound, where the redundancy term $\log n$, associated with the transition, is replaced by $(1 + \epsilon)\log m$, which is essentially as large as $(1 + \epsilon)\log n$. This extra redundancy can be eliminated by letting $\epsilon = \epsilon_j$ vanish with j sufficiently slowly such that $\{\pi(j)\}_{j \geq 1}$ remains summable, e.g., $\epsilon_j = O(\log\log j/\log j)$. Alternatively, one may use the universal prior on the integers as a weighting sequence (see, e.g., [30]).

Finally, it should be pointed out that the latter coding scheme is not only strongly sequential in the sense that n need not be specified in advance, but it also attains the minimum description length in a *pointwise* manner, i.e., the redundancy term coincides with the lower bound for *any* n-tuple and not merely on the average, while the leading term of the code length is given by the empirical entropy.

ACKNOWLEDGMENT

Useful discussions with M. Weinberger are greatly appreciated.

REFERENCES

[1] L. D. Davisson, "Universal noiseless coding," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 783–795, Nov. 1973.
[2] F. Jelinek and K. S. Schneider, "Variable-length encoding of fixed-rate Markov sources for fixed-rate channels," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 750–755, 1974.
[3] R. E. Krichevsky and V. K. Trofimov, "The performance of universal encoding," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 199–207, Mar. 1981.
[4] L. D. 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.
[5] J. Rissanen and G. G. Langdon, "Universal modeling and coding," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 12–23, Jan. 1981.
[6] L. D. Davisson, "Minimax noiseless universal coding for Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 211–215, Mar. 1983.
[7] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 629–636, July 1984.
[8] A. C. Blumer, "Minimax universal noiseless coding for unifilar and Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 925–930, Nov. 1987.
[9] T. J. Tjalkens and F. M. J. Willems, "Variable to fixed-length codes for Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 246–257, Mar. 1987.
[10] M. J. Weinberger, A. Lempel, and J. Ziv, "A sequential algorithm for the universal coding of finite memory sources," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1002–1014, May 1992.
[11] 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.
[12] J. C. Kieffer, "A unified approach to weak universal source coding," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 674–682, Nov. 1978.
[13] A. D. Wyner and J. Ziv, "Some properties of the entropy of a stationary ergodic data source with applications to data compression," *IEEE Trans. Inform. Theory*, vol. 35, pp. 1250–1258, Nov. 1989.
[14] R. G. Gallager, "Variations on a theme by Huffman," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 668–674, Nov. 1978.
[15] D. E. Knuth, "Dynamic Huffman coding," *J. Algorithms*, vol. 6, pp. 163–180, June 1985.
[16] J. S. Vitter, "Design and analysis of dynamic Huffman codes," *J. Ass. Comput. Mach.*, vol. 34, pp. 825–845, Oct. 1987.
[17] H. Yokoo, "An improvement of dynamic Huffman coding with a simple repetition finder," *IEEE Trans. Commun.*, vol. 39, pp. 8–10, Jan. 1991.
[18] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[19] T. Svendsen and F. K. Soong, "On the automatic segmentation of speech signals," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1987, pp. 77–80.
[20] F. K. Soong and B.-H. Juang, "A segment model based approach to speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1988, pp. 501–504.
[21] J. G. Wilpon and L. R. Rabiner, "Application of hidden Markov models to automatic speech endpoint detection," *Comput. Speech Language*, vol. 2, pp. 321–341, 1988.
[22] M. A. Fischler and R. C. Bolles, "Perceptual organization of curve partitioning," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-8, pp. 100–105, Jan. 1986.
[23] N. Katzir, M. Lindenbaum, and M. Porat, "Planar curve segmentation for recognition of partially occluded shapes," in *Proc. Int. Conf. Pattern Recognition*, June 1990, pp. 842–846.
[24] R. C. Dubes, A. K. Jain, S. G. Nadaban, and C. C. Chen, "MRF model based algorithms for image segmentation," in *Proc. Int. Conf. Pattern Recognition*, June 1990, pp. 804–814.
[25] K. Keeler, "Minimum-length encoding of planar subdivision topologies with application to image segmentation," in *Proc. 1990 Spring Symp. Ser. Theory and Appl. of Minimal-Length Encoding*, Mar. 1990, pp. 95–99.
[26] A. Pentland and T. Darrell, "Part segmentation for object recognition," in *Proc. 1990 Spring Symp. Ser. Theory and Appl. of Minimal-Length Encoding*, Mar. 1990, pp. 105–109.
[27] M. S. Babcock, W. K. Olson, and E. P. D. Pednault, "The use of the minimum description length principle to segment DNA into structural and functional domains," in *Proc.* 1990 *Spring Symp. Ser. Theory and Appl. of Minimal-Length Encoding*, Mar. *1990*, pp. 40–44.
[28] A. Milosavljevic, D. Haussler, and J. Jurka, "Clustering of macromolecular sequences by minimal length encoding," in *Proc. 1990 Spring Symp. Ser. Theory and Appl. of Minimal-Length Encoding*, Mar. 1990, pp. 45–49.
[29] Yu. M. Shtar'kov, "Universal sequential coding of single messages," *Probl. Inform. Transmission*, pp. 175–186, July–Sept. 1988.
[30] B. Ya. Ryabko, "Prediction of random sequences and universal coding," *Probl. Inform. Transmission*, pp. 87–96, Apr.–June 1988.

Limits of Conditional Expectations

Eimear Goggin

***Abstract*— If (X^N, Y^N) on a probability space $(\Omega^N, \mathcal{F}^N, P^N)$ converge in distribution to (X, Y) on $(\Omega, \mathcal{F}, P)$, it is not necessarily true that the conditional expectations $E^{P^N}\{F(X^N)|Y^N\}$ converge in distribution to $E^P\{F(X)|Y\}$, even for bounded, continuous functions F. The limits of the conditional expectations can be determined if it is possible to make**

Manuscript received March 3, 1992; revised December 1, 1992. This work was supported in part by the NSF under Grant DMS-8913222.

The author is with the Department of Mathematics, Iowa State University, Ames, IA 50011.

IEEE Log Number 9213025.