A Complete Sufficient Statistic for Finite-State Markov Processes with Application to Source Coding

Laurence B. Wolfe and Chein-I Chang, *Senior Member, IEEE*

***Abstract*—A complete sufficient statistic is presented in this paper for the class of all finite-state, finite-order stationary discrete Markov processes. This sufficient statistic is complete in the sense that it summarizes in entirety the whole of the relevant information supplied by any process sample. The sufficient statistic has application to source coding problems such as source matching and calculation of the rate distortion function.**

***Index Terms*—Complete sufficient statistic, markov chain, source coding.**

I. Introduction

The purpose of this correspondence is to present a complete sufficient statistic for the class of stationary finite-order, discrete-time Markov processes with a discrete state-space (i.e., discrete Markov chains). A sufficient statistic is said to be complete if it summarizes in entirety the whole of the relevant information supplied by any process sample.

Complete sufficient statistics have a well known role in estimation theory [1] and have also found application in source coding problems such as source matching [2] and calculation of the rate distortion function [3], [4]. A complete sufficient statistic is presented in Section II, examples are given in Section III, and conclusions are drawn in Section IV.

II. A Complete Sufficient Statistic for Markov Chains

A complete sufficient statistic is presented in this Section for the class of stationary finite-order, discrete Markov chains with a finite state-space A of size J. We first show the existence of a complete sufficient statistic for first-order Markov chains and then extend the result to finite-order chains.

Consider the class of stationary first-order Markov chains with a finite state space A of size J; that is $|A| = J$. The stochastic transition matrix is given by

$$\begin{bmatrix} \theta_{11} & \theta_{12} & \cdots & \theta_{1J} \\ \cdot & & & \\ \cdot & & & \\ \cdot & \cdot & & \\ \theta_{J1} & \theta_{J2} & \cdots & \theta_{JJ} \end{bmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \cdots & 1-\sum_{i\neq J} & \theta_{1i} \\ \cdot & & & & \\ \cdot & \cdot & & & \\ \theta_{J1} & \theta_{J2} & \cdots & 1-\sum_{k\neq J} & \theta_{Jk} \end{bmatrix}. \quad (1)$$

θ_{ij} represents the transition probability of j given a preceding i for all $i, j \in A$. Therefore, any Markov chain can be represented by an array of J vectors $\boldsymbol{\theta} = [\boldsymbol{\theta}_j^T : \boldsymbol{\theta}_j^T = (\theta_{j1}, \cdots, \theta_{jJ})]$, where $\boldsymbol{\theta} \in \Theta$ takes values in the J-dimensional hypercube $[0, 1]^J$.

Manuscript received March 9, 1992; revised September 18, 1992. This work was presented at the 26th Annual Conference for Information Sciences and Systems, Princeton University, Princeton, NJ, March 1992.

L. B. Wolfe is with the U.S. Government (GSA), Washington DC 20405.

C.-I. Chang is with the Department of Electrical Engineering, University of Maryland, Baltimore County Campus, Baltimore, MD 21228.

IEEE Log Number 9207881.

Let $\underline{x}^N = (x_1, \cdots, x_N)$ and define π_α as the given probability that $X_1 = \alpha$. Also, let the transition count statistic $\{s_{ij}: i, j \in (A \times A)\}$ be defined as the number of m: $0 \leq m \leq N-1$ such that $x_m = i$ and $x_{m+1} = j$. Thus, there is a function T such that for each specific realization of the Markov process $\underline{X}^N = \underline{x}^N$; $T = (x_1, \{s_{ij}\})$ when $X_1 = x_1$ and $\{S_{ij}\} = \{s_{ij}\}$. Let $\delta_{ij} = 1$ if $i = j$ and zero, otherwise. Given any parameter $\boldsymbol{\theta} \in \Theta$, the probability that $\underline{X}^N = \underline{x}^N$ is thus given by

$$\begin{aligned} p^{\underline{\theta}}(\underline{X}^N = \underline{x}^N) &= \pi_{x_1} \cdot \theta_{x_1,1}^{\delta_{x_2,1}} \cdots \theta_{x_1,J}^{\delta_{x_2,J}} \cdots \theta_{x_{N-1},1}^{\delta_{x_N,1}} \cdots \theta_{x_{N-1},J}^{\delta_{x_N,J}} \\ &= \pi_{x_1} \cdot \prod_{k=1}^{N-1} \left(\prod_{j\in A} \theta_{x_k,j}^{\delta_{x_{k+1},j}} \right) \\ &= \pi_{x_1} \cdot \prod_{i,j\in A} \theta_{ij}^{s_{ij}}, \end{aligned} \quad (2)$$

Following the well-known Factorization Theorem as given in [1], a function $T = t(\underline{X}^N)$ of a random vector $\underline{X}^N$, which depends on a parameter $\boldsymbol{\theta} \in \Theta$ is said to be sufficient, if and only if the frequency function factors into a product of a function of $t(\underline{X}^N)$ and $\boldsymbol{\theta}$ and a function of $\underline{X}^N$ alone; that is $f(\underline{x}^N \,|\, \boldsymbol{\theta}) = g(t(\underline{x}^N), \boldsymbol{\theta}) \cdot h(\underline{x}^N)$. The sufficiency of $T = (x_1, \{s_{ij}\})$ clearly follows from (2), which satisfies the Factorization Theorem with $h(\underline{x}^N) = 1$.

Furthermore, a sufficient statistic T is said to be complete according to [1] if for every real-valued function g, $\boldsymbol{E}^{\boldsymbol{\theta}}(g, (T)) = 0$ implies for $\theta \in \Theta$ that $\boldsymbol{P}^{\boldsymbol{\theta}}(g(T) = 0) = 1$.

Theorem 1: A complete sufficient statistic for the class of finite-state, finite-order Markov chains is given by $T = (x_1, \{s_{ij}\})$.

Proof: To see that T is indeed complete for first-order chains, we first observe that

$$\begin{aligned} E^{\underline{\theta}}(g(T)) &= \sum_{t\in T} g(t) \cdot P^{\underline{\theta}}(t) \\ &= \sum_{t\in T} g(t) \cdot \pi_{x_1} \cdot \prod_{i,j\in A} \theta_{ij}^{s_{ij}} \cdot h(t) \\ &= \sum_{t\in T} a(t) \cdot \prod_{i,j\in A} \theta_{ij}^{s_{ij}}, \qquad 0 \leq s_{ij} \leq N-1, \end{aligned} \quad (3)$$

where $h(\cdot)$ is the number of vectors $\underline{X}^N$ with $(x_1, \{s_{ij}\})$ and $a(\cdot) = g(\cdot) \cdot h(\cdot) \cdot \pi_{(\cdot)}$. An expression for $h(\cdot)$ will be explicitly derived following the proof of this theorem. Proceeding, if a polynomial, or more generally, a power series $\sum a_n \cdot y^n$, is zero for y in some open interval, it is well known that each of the coefficients a_n must be zero. Clearly, (3) is the sum of J^2 polynomials in $\{\theta_{ij}\}$, of degree $N-1$. Since the θ_{ij} are nonnegative by definition, each coefficient $a(\cdot)$ must be zero. Since $h(\cdot)$ and $\pi_{(\cdot)}$ are not zero in general, $g(\cdot)$ must be equal to zero for all t thus proving that T is complete. Apparently, the result holds in general for finite-order chains since they also factor into the sum of polynomials in $\{\theta_{(\cdot)}\}$. □

We now explicitly derive $h(t)$ for first-order, finite-state Markov chains. Observe that for every $j \in A$ it is apparent that after accounting for boundary conditions, the magnitude of the transition statistics s_{ij} when summed over i must balance with the sum over k of all s_{jk}. We conclude that for a given $X_1 = x_1$, the permutations of the s_{ij} and s_{jk} subchains ij and jk respectively, determine the number of $\underline{x}^N$ with a fixed total number of subchains ijk; that is the total number of all m such that $X_m = i$, $X_{m+1} = j$ and $X_{m+2} = k$ where $0 \leq m \leq N-2$.

More precisely, let $X_1 = x_1$ and define r_{ijk} as the number of occurrences in $\underline{x}^N$ where i precedes j, which is succeeded by k; that is the intersection of the events defined by the transition statistics s_{ij} and s_{jk}. For example, $\underline{x}^6 = [011\,201]$ has $r_{011} = 1$. Clearly, $(x_1, \{r_{ijk}\})$ is a function of the sufficient statistic $(x_1, \{s_{ij}\})$. For simplicity of notation, define the total instances where a fixed i is succeeded by any j that precedes a fixed k as $R_{ik} \equiv \sum_{j \in A} r_{ijk}$.

Consider Case 1 first where the process transitions to distinct states, that is; $i \neq j$; $j \neq k$. Clearly, the unique number of arrangements of the individual set element $\{r_{ijk}\}$ out of R_{ik} can be determined in an obvious way. Now consider Case 2 where the $\{r_{ijk}\}$ are given by $i = j$ and (or) $j = k$. In this case, the process remains in the same state for a specific number of transitions; that is the chain has subchains (e.g., runs) of i of a specific length. Let D_i be defined as the total number of subchains of i and define n_i as the number of i's in $\underline{x}^N$. For example, $\underline{x}^6 = [011\,201]$ has one subchain (run) of 1's of length 2 and one of length 1. Thus, $D_i = 2$ and $n_i = 3$. Clearly, D_i and n_i are uniquely determined by $(x_1, \{r_{ijk}\})$. For each $i \in A$, the number of ways to begin D_i runs of n_i (nondistinct) i's is determined in an obvious way, $h(t)$ is given by combining Cases 1 and 2:

$$\begin{aligned} h(t = (x_1, \{r_{ijk}\})) &= \left|\{\underline{x}^N \in A^N\colon\ t(\underline{x}^N) = (x_1, \{r_{ijk}\})\}\right| \\ &= \prod_{i,\,k \in A} \frac{R_{ik}!}{\prod_{j \neq i,\,k} r_{ijk}!} \cdot \prod_{m \in A} \binom{n_m - 1 + \delta_{n_m,\,0}}{D_m - 1 + \delta_{D_m,\,0}}, \end{aligned} \tag{4}$$

where $|\cdot|$ denotes the size of set $\{\cdot\}$ and $\delta_{(\cdot)}$ accounts for boundary conditions. Clearly,

$$\sum_{\substack{t = (x_1, \{r_{ijk}\}) \\ i,\,j,\,k \in A}} h(t) = |A|^N. \tag{5}$$

Consequently, the desired probability mass function is given by

$$P(T = t \,|\, \boldsymbol{\theta}) = h(t) \cdot \pi_{x_1} \cdot \prod_{i,\,j \in A} \theta_{ij}^{s_{ij}}. \tag{6}$$

Extension To Higher Order Markov Chains: Extension of this result can be readily accomplished either by transfer of the higher order chain to a first-order Markov chain or by utilizing the higher order stochastic transition matrix $\boldsymbol{\theta}$ to construct a suitable definition of the sets $\{s_{(\cdot)}\}$ and $\{r_{(\cdot)}\}$. The first method will be demonstrated in Example 3 in the next section. The latter method will be described in the following for a m-order chain with state-space A.

In analogy to the first-order case, for every $j \in A$ we define a transition statistic $s_{\underline{i}j}$ where j is preceded by the subchain $\underline{i}$ of length m. That is, $\underline{i} = [x_{p-m-1}, \cdots, x_{p-1}]$ and $x_p = j$ for some $m + 1 \leq p \leq N$. After accounting for boundary conditions, the $s_{\underline{i}j}$, when summed over $\underline{i}$, must balance with the sum over $\underline{k}$ of all s_{jk}. We can now define $\{R_{ik}\}$ and $\{r_{\underline{i}jk}\}$ to derive $(x_1, \{r_{\underline{i}jk}\})$ and $T = (x_1, \{s_{\underline{i}j}\})$. For example, a second-order J-state Markov chain would determine h by using the sets $\{s_{(abj)}\}$ and $\{s_{(jcd)}\}$ to define $\{r_{(abjcd)}\}$ where $a, b, c, d, j \in A$.

The difficulty of calculating $h(t)$ for a first-order Markov chain is shown by (4). In general, calculating $h(t)$ increases in difficulty for higher order chains. However, some reduction in difficulty can be achieved as shown in Examples 1 and 3 in Section III.

III. Examples

Example 1 (First-Order Binary Discrete Markov Chain): We examine a class of Markov chains where the parameter $\boldsymbol{\theta} = [\theta_0, \theta_1]$ is a two-dimensional variable in the unit square $[0, 1] \times [0, 1]$. The binary state-space is $A = \{0, 1\}$. The stochastic transition matrix is

$$\begin{bmatrix} \theta_{00} & \theta_{01} \\ \theta_{10} & \theta_{11} \end{bmatrix} = \begin{bmatrix} 1 - \theta_0 & \theta_0 \\ \theta_1 & 1 - \theta_1 \end{bmatrix}. \tag{7}$$

For simplicity of notation, transition probability θ_{01} is denoted by θ_0 and θ_{10} is denoted by θ_1. The stationary pmf is given as

$$\pi = \left(\frac{\theta_1}{\theta_0 + \theta_1}, \frac{\theta_0}{\theta_0 + \theta_1}\right) = (\pi_0, \pi_1). \tag{8}$$

Clearly, $R_{ik} = r_{ijk}$ for all $i, j, k \in A$ because A is binary. Therefore, (4) becomes

$$h(t = (x_1, \{r_{ijk}\})) = \prod_{m \in \{0,\,1\}} \binom{n_m - 1 + \delta_{n_m,\,0}}{D_m - 1 + \delta_{D_m,\,0}}. \tag{9}$$

Accordingly, $(x_1, \{r_{ijk}\})$ is completely specified by (n_1, D_1, x_1, x_N) in this example and it is apparent that the probability of each vector given by $(n_1, D_1, 0, 1)$ is identical to that of $(n_1, D_1, 1, 0)$, as can be seen by applying (2). This symmetry does not exist in general, but can be used effectively in this case to reduce the size of T by defining the following cases:

$$\begin{aligned} &\text{Case 1: } (n_1, D_1, 1) \equiv (n_1, D_1, 0, 0); \\ &\text{Case 2: } (n_1, D_1, 2) \equiv (n_1, D_1, 0, 1) \quad \text{or} \quad (n_1, D_1, 1, 0); \\ &\text{Case 3: } (n_1, D_1, 3) \equiv (n_1, D_1, 1, 1). \end{aligned} \tag{10}$$

Let i identify the specific case just described. Equation (9) can be rewritten as

$$h(t = (x_1, \{r_{ijk}\})) = (1 + \delta_{i,\,2}) \cdot \prod_{m \in \{0,\,1\}} \binom{n_m - 1 + \delta_{n_m,\,0}}{D_m - 1 + \delta_{D_m,\,0}}. \tag{11}$$

Example 2 (First-Order Ternary Discrete Markov Chain): Consider the class of ternary discrete Markov chains with stochastic transition matrix:

$$\begin{bmatrix} \theta_{00} & \theta_{01} & \theta_{02} \\ \theta_{10} & \theta_{11} & \theta_{12} \\ \theta_{20} & \theta_{21} & \theta_{22} \end{bmatrix} = \begin{bmatrix} \theta_{00} & \theta_{01} & 1 - \sum_{i=0,\,1} & \theta_{0i} \\ \theta_{10} & \theta_{11} & 1 - \sum_{j=0,\,1} & \theta_{1j} \\ \theta_{20} & \theta_{21} & 1 - \sum_{k=0,\,1} & \theta_{2k} \end{bmatrix}. \tag{12}$$

For all $i, j \in A$, θ_{ij} represents the transition probability for j given a preceding i. Unlike Example 1, the ternary state-space has no symmetry. Therefore, the complete sufficient statistic is given by direct application of (4) and (6).

Example 3 (Second-Order Binary Discrete Markov Chain): The sufficient statistic T may be derived for the class of binary discrete second-order Markov chains as suggested at the end of Section II. However, another approach is to map by a function g each chain to a first-order Markov chain by defining a new state-space $\tilde{A} = \{\alpha, \beta, \gamma, \zeta\}$ where

$$\begin{aligned} \alpha &\equiv 00 \\ \beta &\equiv 01 \\ \gamma &\equiv 10 \\ \zeta &\equiv 11. \end{aligned} \tag{13}$$

For example, the chain 000 110 maps to $\alpha\alpha\beta\zeta\gamma$. Observe that a chain has length N in A and length $N - 1$ in $\tilde{A}$. The stochastic transition matrix is mapped from A to $\tilde{A}$ as follows:

$$\begin{bmatrix} \theta_{000} & \theta_{001} \\ \theta_{010} & \theta_{011} \\ \theta_{100} & \theta_{101} \\ \theta_{110} & \theta_{111} \end{bmatrix} \xrightarrow{g} \begin{bmatrix} \theta_{\alpha\alpha} & \theta_{\alpha\beta} & \theta_{\alpha\gamma} & \left(\theta_{\alpha\zeta} \equiv 1 - \sum_{i \neq \zeta} \theta_{\alpha i}\right) \\ \theta_{\beta\alpha} & \theta_{\beta\beta} & \theta_{\beta\gamma} & \left(\theta_{\beta\zeta} \equiv 1 - \sum_{j + \zeta} \theta_{\beta j}\right) \\ \theta_{\gamma\alpha} & \theta_{\gamma\beta} & \theta_{\gamma\gamma} & \left(\theta_{\gamma\zeta} \equiv 1 - \sum_{k \neq \zeta} \theta_{\gamma k}\right) \\ \theta_{\zeta\alpha} & \theta_{\zeta\beta} & \theta_{\zeta\gamma} & \left(\theta_{\zeta\zeta} \equiv 1 - \sum_{p \neq \zeta} \theta_{\zeta p}\right) \end{bmatrix}. \tag{14}$$

The new stochastic matrix is a first-order Markov chain with state-space $\tilde{A}$. $h(t)$ and $p(t \mid \boldsymbol{\theta})$ can now be derived by using (4) and (6) over the new space $\tilde{A}$ and recognizing that chains containing certain subchains are forbidden. For this example, any chain containing one or more of the following subchains is not allowed: $\alpha\gamma$, $\alpha\zeta$, $\beta\alpha$, $\beta\beta$, $\beta\zeta$, $\gamma\beta$, $\gamma\gamma$, $\gamma\zeta$, $\zeta\alpha$, $\zeta\beta$. Any finite-order chain can be similarly mapped to a first-order chain.

IV. Conclusion

A sufficient statistic for finite-order, finite-state Markov chains was presented in this correspondence and shown to be complete. The sufficient statistic relies only upon the initial state and the sum of frequencies of subchain occurrence. Several examples were given and applications were cited for source coding.

References

[1] T. S. Ferguson, *Mathematical Statistics*. New York: Academic Press, 1967, pp. 113–136.

[2] L. B. Wolfe and C.-I. Chang, "Source matching problems revisited," in *Proc. Int. Conf. Signal Processing '90*, Beijing, China, Oct. 22–26, 1990, pp. 119–122.

[3] C.-I. Chang, S. C. Fan, and L. D. Davisson, "Computation of the rate distortion function for a source with uncertain statistics," in *Proc. IEEE Int. Conf. on Communications*, Oct. 31–Nov. 3, 1988, Singapore, pp. 1180–1184.

[4] L. B. Wolfe and C.-I. Chang, "A simple method for calculating the rate distortion function of a source with an unknown parameter," *Proc. IEEE Int. Symp. Commun.*, Tainan, Taiwan, R.O.C., Dec. 10–13, 1991, pp. 261–264.

[5] W. Feller, *An Introduction To Probability Theory and Its Applications*. New York: Wiley, 1968, pp. 147–150.

Quantizer Monotonicities and Globally Optimal Scalar Quantizer Design

Xiaolin Wu, *Member, IEEE,* and Kaizhong Zhang, *Member, IEEE*

***Abstract*—New monotonicity properties of optimal scalar quantizers are discovered. These monotonicities reveal a structure of a globally optimal scalar quantizer depending on the probability mass functions and on the number of quantizer levels. By incorporating the monotone quantizer structure into a dynamic programming process, the time complexities of previous algorithms for designing globally optimal scalar quantizers can be significantly reduced for very general classes of distortion measures.**

***Index Terms*—Quantization, optimization, monotonicity, dynamic programming, divide-and-conquer, matrix-search.**

I. Introduction

Two optimal quantizer design algorithms in current use, Lloyd's methods I and II [4] (the latter was also independently found by Max [5]), were proposed forty years ago. Since then the quantization research has remained active particularly after Lloyd's method I was generalized to the LBG algorithm for vector quantization [3].

Manuscript received July 18, 1990; revised August 11, 1992.

The authors are with the Department of Computer Science, University of Western Ontario, London, ON, N6A 5B7, Canada.

IEEE Log Number 9206222.

Lloyd's method I and the LBG algorithm are typical gradient descent approaches aiming at a local minimum of the quantization distortion function. However, they can get stuck at a local minimum far above the global minimum. Lloyd's method II also cannot guarantee a global optimum. To find the global minimum of the distortion function, we resort to search-based discrete optimization. To make the problem domain finite, the input signal amplitude density function $p(x)$ is approximated, if not given as such, by a probability mass function

$$P(x_i) = \int_{x_i-\delta}^{x_i+\delta} p(x)\,dx, \qquad 1 \le i \le N, \tag{1}$$

with $x_i = (2i-1)\delta$, $\delta = \frac{1}{2N}$. This approximation of $p(x)$ can reach any desired precision by choosing a sufficiently large N. Note that the spacing of $x_{i+1} - x_i$ does not have to be a constant as in (1).

A K-level quantization for a probability mass function $P(x_i)$ on N points where $K < N$ (shortened to $K : N$ quantization in the sequel) is a partitioning of the ordered set $\{x_i | 1 \le i \le N\}$ into K subsets, $\{x_i | q_{j-1} < i \le q_j\}$, where $0 < j \le K$, $q_0 \equiv 0$, and $q_K \equiv N$ by convention. If we define a finite set:

$$Q_n^k \equiv \{\boldsymbol{q} | 1 \le q_1 < q_2 < \cdots < q_{k-1} < n\} \subset \mathfrak{N}^{k-1}, \tag{2}$$

where $\mathfrak{N}$ is the set of all natural numbers, then there is a one-to-one map between a $\boldsymbol{q} \in Q_n^k$ and a possible partition as described above for $k : n$ quantization. Therefore, we call any $\boldsymbol{q} \in Q_N^K$ a $K : N$ quantizer. Denote the quantizer codewords by a K-vector $\boldsymbol{r} = (r_1, r_2, \cdots, r_K)$, with r_j corresponding to the representative of interval $(q_{j-1}, q_j]$. Then the expected error for a $K : N$ quantizer $\boldsymbol{q} \in Q_N^K$ is

$$E(\boldsymbol{q}, \boldsymbol{r}) = \sum_{j=1}^{K} \sum_{i=q_{j-1}+1}^{q_j} P(x_i) W(x_i, x_i - r_j), \tag{3}$$

where W is a suitable error weighting function.

Optimal $K : N$ quantization is therefore a nonlinear programming problem of minimizing $E(\boldsymbol{q}, \boldsymbol{r})$ over all $\boldsymbol{q} \in Q_N^K$. Since the problem domain Q_N^K is finite, by enumerating all possible $K : N$ quantizers, ($|Q_N^K| = \binom{N-1}{K-1}$ of them!), and then minimizing $E(\boldsymbol{q}, \boldsymbol{r})$ over $\boldsymbol{r}$ for each $K : N$ quantizer $\boldsymbol{q} \in Q_N^K$, one could arrive at the globally optimal $K : N$ quantizer. Of course, the naive exhaustive search quickly becomes computationally intractable for even modest K and N, and $K \ll N$ in practice. Bruce [2] showed that a dynamic programming technique can be used to compute the globally optimal $K : N$ quantizer in polynomial-time for general error measures. His algorithm was later improved by Sharma [6] for convex error measures. Recently, we reduced the time complexity of a dynamic programming algorithm for optimal $K : N$ quantization for the mean-square-error measure [10]. In the following sections we will generalize our previous algorithms to a considerably wider class of error measures, and also improve Bruce's algorithm for general error measures.

II. Properties of $K : N$ Quantizers

To facilitate the later development of efficient algorithms for optimal $K : N$ quantizer design, we need to explore some useful properties of globally optimal scalar quantizers. The first $t - 1$ elements of a $k : n$ quantizer $\boldsymbol{q} \in Q_n^k$ where $1 < t < k$, $k < n$, $k \le K$, $n \le N$, comprise a $t : q_t$ quantizer, namely, $(q_1, q_2, \cdots, q_{t-1}) \in Q_{q_t}^t$. Then by the definition of optimal $k : n$ quantizer and by contradiction, one can easily prove the following lemma which