provide a way of predicting the performance of the Viterbi algorithm and gaining some insight into what machines will be improved with the use of context.

References

[1] E. M. Riseman and R. W. Ehrich, "Contextual word recognition using binary digrams," *IEEE Trans. Comput.*, vol. C-20, pp. 397–403, Apr. 1971.

[2] E. J. Sitar, "Machine recognition of cursive script: The use of context for error detection and correction," Bell Telephone Laboratories, Inc., Murray Hill, N.J., unpublished memorandum, Sept. 12, 1961.

[3] J. Raviv, "Decision making in Markov chains applied to the problem of pattern recognition," *IEEE Trans. Inform. Theory*, vol. IT-13, pp. 536–551, Oct. 1967.

[4] G. T. Toussaint and R. W. Donaldson, "Some simple contextual decoding algorithms applied to recognition of handprinted text," in *Proc. Annual Canadian Computer Conference*, Session '72, June 1–3, 1972, Montreal, pp. 422101–422116.

[5] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Inform. Theory*, vol. IT-13, pp. 260–269, Apr. 1967.

[6] G. D. Forney, Jr., "MAP sequence estimation of Markov sequences via the Viterbi algorithm, with application to text recognition," unpublished paper, Stanford Electronics Laboratories, Stanford, Calif., Aug. 9, 1971.

[7] ——, "The Viterbi algorithm," *Proc. IEEE*, vol. 61, pp. 268–278, Mar. 1973.

[8] F. Pratt, *Secret and Urgent*. Blue Ribbon Books, 1939.

[9] M. S. Mayzner and M. E. Tresselt, "Tables of single-letter and digram frequency counts for various word-length and letter-position combinations," *Psychonomic Monograph Supplements*, vol. 1, no. 2, 1965.

[10] A. D. Baddeley and R. Conrad, "Letter structure of the English Language," *Nature*, vol. 186, Apr. 30, 1960.

[11] J. Raviv, "Tables of N-gram frequencies," unpublished.

[12] R. G. Casey and G. Nagy, "Advances in pattern recognition," *Sci. Amer.*, vol. 224, pp. 56–71, Apr. 1971.

[13] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968, pp. 77–78.

A Proof of the Data Compression Theorem of Slepian and Wolf for Ergodic Sources

THOMAS M. COVER, FELLOW, IEEE

***Abstract*—If $\{(X_i,Y_i)\}_{i=1}^{\infty}$ is a sequence of independent identically distributed discrete random pairs with $(X_i,Y_i) \sim p(x,y)$, Slepian and Wolf have shown that the X process and the Y process can be separately described to a common receiver at rates R_X and R_Y bits per symbol if $R_X + R_Y > H(X,Y)$, $R_X > H(X \mid Y)$, $R_Y > H(Y \mid X)$. A simpler proof of this result will be given. As a consequence it is established that the Slepian–Wolf theorem is true without change for arbitrary ergodic processes $\{(X_i,Y_i)\}_{i=1}^{\infty}$ and countably infinite alphabets. The extension to an arbitrary number of processes is immediate.**

I. Introduction

Very roughly and somewhat incorrectly speaking, a sequence of n rainy and fair days in New York City can be compressed without loss of information into a binary sequence of length $nH(X)$, where $H(X)$ is the entropy of the random process. Similarly, Boston weather can be compressed from n bits to $nH(Y)$ bits.

Now suppose that both these compressed sequences are noiselessly made available to San Francisco. Can $X = (X_1,\cdots,X_n)$ and $Y = (Y_1,\cdots,Y_n)$ be reconstructed? The answer is yes, but a total of $n(H(X) + H(Y))$ bits have been transmitted. Clearly the weather reports in Boston and New York are correlated. It appears that only $nH(X,Y)$ bits should be needed.

Manuscript received June 10, 1974; revised October 16, 1974. This work was supported by the National Science Foundation under Grant GK-34363.

The author is with the Departments of Electrical Engineering and Statistics, Stanford University, Stanford, Calif. 94305.

Slepian and Wolf [1] argue that the Y information can be sent as before with a message of length $nH(Y)$ and that the X information can then be sent with $nH(X \mid Y)$ bits by conditioning the encoding of X on the actual value of the Y sequence. They then go on to show the very interesting result that this compression of X to $nH(X \mid Y)$ bits can be achieved even if the value of Y is unknown, and conditional compression is consequently precluded. They show this for processes $\{(X_i,Y_i)\}_{i=1}^{n}$ of independent identically distributed random pairs $(X_i,Y_i) \sim p(x,y)$, $i = 1,2,\cdots,n$, for finite-alphabet sizes. The proof involves $2^{nH(X|Y)}$ separate codes. They prove many related theorems, but it is fair to say that this theorem is the heart of their work.

We shall consider the following encoding-decoding scheme based on the asymptotic equipartition property. The idea involves a single random code. Each y sequence is randomly assigned an index $j(y)$ in the set $\{1,2,\cdots,2^{n(H(Y)+\varepsilon)}\}$, and each x sequence is randomly assigned an index $i(x)$ in the set $\{1,2,\cdots,2^{n(H(X|Y)+\varepsilon)}\}$. The assignment functions $i(\cdot),j(\cdot)$ are known to the receiver. A pair of sequences (x,y) is said to be typical if $p(x) \approx 2^{-nH(X)}$, $p(y) \approx 2^{-nH(Y)}$, $p(x,y) \approx 2^{-nH(X,Y)}$. Given (i,j), the receiver decides (x,y) was sent if: i) (x,y) is typical, ii) $i(x) = i$, $j(y) = j$, and iii) there is no other typical pair (x',y') satisfying ii). The proof that the probability of decoding error is small comes from the observation that the expected number of typical sequences having a given index (i,j) is $\ll 1$.

Ahlswede and Körner [7] have found another proof of the result of Slepian and Wolf using the code construction of Feinstein.

II. Preliminaries; Typical Sequences

Let $\{(X_i,Y_i)\}_{i=1}^{\infty}$ be an ergodic process of discrete random variables. Then $\{X_i\}_{i=1}^{\infty}$ and $\{Y_i\}_{i=1}^{\infty}$ are necessarily ergodic processes, since they are projections of $\{(X_i,Y_i)\}_{i=1}^{\infty}$. If $\{(X_i,Y_i)\}_{i=1}^{\infty}$ is ergodic, we shall say that $\{X_i\}_{i=1}^{\infty}$ and $\{Y_i\}_{i=1}^{\infty}$ are jointly ergodic. By the asymptotic equipartition property due to Shannon [2], McMillan [3], and Breiman [4], we know that

$$
\begin{aligned}
&-\frac{1}{n}\log p(X_1,X_2,\cdots,X_n,Y_1,Y_2,\cdots,Y_n) \to H(X,Y)\\
&-\frac{1}{n}\log p(X_1,\cdots,X_n) \to H(X)\\
&-\frac{1}{n}\log p(Y_1,\cdots,Y_n) \to H(Y)
\end{aligned}
\tag{1}
$$

with probability one, where

$$
\begin{aligned}
H(X,Y) &= \lim_{n\to\infty}\frac{1}{n}H(X_1,X_2,\cdots,X_n,Y_1,Y_2,\cdots,Y_n)\\
H(X) &= \lim_{n\to\infty}\frac{1}{n}H(X_1,X_2,\cdots,X_n)\\
H(Y) &= \lim_{n\to\infty}\frac{1}{n}H(Y_1,Y_2,\cdots,Y_n)\\
H(X \mid Y) &= H(X,Y) - H(Y).
\end{aligned}
\tag{2}
$$

The proof of convergence with probability one was extended to countably infinite alphabets by Chung [5]. Consequently, we have the following lemma (see also [8]).

Lemma 1: For $\{(X_i,Y_i)\}_{i=1}^{\infty}$ ergodic and $\varepsilon > 0$, there exists an integer n and a set A of typical n-sequences $x = (x_1,\cdots,x_n)$, $y = (y_1,\cdots,y_n)$ such that

$$
\Pr\{A\} \equiv \Pr\{(X,Y) \in A\} = \sum_{A} p(x,y) \geq 1 - \varepsilon \tag{3}
$$

and

$$\left|-\frac{1}{n}\log p(x,y) - H(X,Y)\right| \le \frac{\varepsilon}{2}$$
$$\left|-\frac{1}{n}\log p(x) - H(X)\right| \le \frac{\varepsilon}{2}$$
$$\left|-\frac{1}{n}\log p(y) - H(Y)\right| \le \frac{\varepsilon}{2}, \quad \text{for } (x,y) \in A. \tag{4}$$

Definition: For each y, let the set of the x that are *jointly typical* with y be defined by

$$T_y = \{x: (x,y) \in A\}. \tag{5}$$

Lemma 2: For n satisfying Lemma 1

$$|T_y| \le 2^{n(H(X|Y)+\varepsilon)}. \tag{6}$$

Proof: $x \in T_y$ implies $(x,y) \in A$. Thus for each y,

$$\begin{aligned} 1 &= \sum_x p(x \mid y) = \sum_x \frac{p(x,y)}{p(y)} \\ &\ge \sum_{x \in T_y} \frac{p(x,y)}{p(y)} \\ \text{a)} \quad &\ge \sum_{x \in T_y} \frac{2^{-n(H(X,Y)+\varepsilon/2)}}{2^{-n(H(Y)-\varepsilon/2)}} \\ \text{b)} \quad &= 2^{-n(H(X|Y)+\varepsilon)}|T_y| \end{aligned} \tag{7}$$

where inequality a) follows from (4) and b) follows from (2).

Q.E.D.

This is the uniform bound that we require on the size of the set T_y of conditionally typical x-sequences for a given y and is the primary reason that $\{X_i\}$ and $\{Y_i\}$ are assumed to be jointly ergodic.

III. Theorem and Proof

Theorem 1: Let $\{(X_i,Y_i)\}_{i=1}^{\infty}$ be an ergodic process, where $X_i \in \mathscr{X}$, $Y_i \in \mathscr{Y}$, and $\mathscr{X},\mathscr{Y}$ are countably infinite. Then, for any

$$R_X > H(X \mid Y), \quad R_Y > H(Y \mid X), \quad R_X + R_Y > H(X,Y) \tag{8}$$

there exists an integer n and mappings

$$\begin{aligned} i&: \mathscr{X}^n \to I = \{1,2,\cdots,2^{nR_X}\} \\ j&: \mathscr{Y}^n \to J = \{1,2,\cdots,2^{nR_Y}\} \\ g&: I \times J \to \mathscr{X}^n \times \mathscr{Y}^n \end{aligned} \tag{9}$$

such that

$$\Pr\{g(i(X_1,\cdots,X_n),j(Y_1,\cdots,Y_n)) = (X_1,\cdots,X_n,Y_1,\cdots,Y_n)\} \ge 1-\varepsilon. \tag{10}$$

Remark: This theorem implies that $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$ can be independently compressed from n symbols each to nR_X and nR_Y bits, respectively, with negligible probability of error in the reconstruction of $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$, where the reconstruction is based on the knowledge of both compressed sequences.

Remark: For the finite-alphabet case, the probability of error of this encoding can be made zero while achieving the preceding compression by the standard method of mapping the nontypical sequences one-to-one into long binary strings that have not been used for the encoding of $i(x)$ and $j(y)$. Since the nontypical strings have total probability less than ε, the change in the compression in bits per symbol can be made arbitrarily small, and all $(x,y) \in \mathscr{X}^n \times \mathscr{Y}^n$ will be perfectly recovered.

Encoding Technique for Theorem

Let A be the set of typical (x,y) n-sequences of Section II. We shall prove that $(R_X,R_Y) = (H(X \mid Y),H(Y))$ is achievable. The achievability of the pair $(H(X),H(Y \mid X))$ follows an identical argument. The remainder of the boundary follows by time-sharing these two schemes. Let $B_1,B_2,B_3,\cdots,B_M$, $M = 2^{nR_X}$, be a random partition of $\mathscr{X}^n$. Specifically, let $\Pr\{x \in B_i\} = 1/M$, $\forall i$, $\forall x \in \mathscr{X}^n$, and independently assign each $x \in \mathscr{X}^n$ to some B_i, $i = 1,2,\cdots,M$. (Actually all that is needed is a random uniform partition of the finite set of typical x sequences $\{x: (x,y) \in A$, for some $y\} = \cup T_y$.)

Then the encoding is as follows. Let $j(y)$ be the index of y in a fixed enumeration of the (not more than) $2^{n(H(Y)+\varepsilon)}$ typical sequences $y \in S$. Let $j(y) = 0$, $y \notin S$. Let $i(x) \in \{1,2,\cdots,2^{nR_X}\}$ be the index of the set B_i containing x.

Decoding Algorithm

Given (i,j), let $\hat{y}$ be the solution of $j(\hat{y}) = j$. Let $\hat{x}$ be the unique element in $B_i \cap T_{\hat{y}}$ if this set contains one and only one element; otherwise declare an error. Thus $g(i,j) = (\hat{x},\hat{y})$ is the decoding.

The decoding consists of choosing the only jointly typical pair $(x,y) \in A$ consistent with (i,j), unless there are none or more than one.

Proof of Theorem

For $H(X,Y) = \infty$, the theorem is vacuously true. Assume $H(X,Y) < \infty$. Consider the following exhaustive error events:

$$E_0: (X,Y) \notin A, \text{ i.e., } (X,Y) \text{ is not typical.} \tag{11}$$

$$E_1: (X,Y) \in A, \text{ and there exists } x' \neq X,\ x' \in B_{i(X)} \cap T_Y,$$
$$\text{i.e., some other conditionally typical}$$
$$x' \in T_Y \text{ has the same index } i \text{ as } X. \tag{12}$$

The random variables in these events are X,Y, and the random partitioning $(B_1,\cdots,B_m)$ of $\mathscr{X}^n$. Setting $(\hat{X},\hat{Y}) = g(i(X),j(Y))$, let

$$\bar{P}_e = \Pr\{(\hat{X},\hat{Y}) \neq (X,Y)\} \tag{13}$$

denote the probability of a decoding error, where the bar over P_e indicates that the probability also includes the random choice of partition. We see that

$$\bar{P}_e = \Pr\{E_0 \cup E_1\} \le \Pr\{E_0\} + \Pr\{E_1\}. \tag{14}$$

Clearly, $\Pr\{E_0\} \le \varepsilon$, by the definition of A. It remains only to prove that $\Pr\{E_1\} \le \varepsilon$. However,

$$\begin{aligned} \Pr\{E_1\} &= \Pr\{(X,Y) \in A, \exists x' \neq X,\ x' \in T_Y,\ x' \in B_{i(X)}\} \\ &= \sum_{(x,y)\in A} p(x,y) \Pr\{\exists x' \neq x,\ x' \in T_y,\ x' \in B_{i(x)}\} \\ \text{a)} \quad &\le \sum_{(x,y)\in A} p(x,y) \sum_{\substack{x' \neq x \\ x' \in T_y}} \Pr\{x' \in B_{i(x)}\} \\ \text{b)} \quad &= \sum_{(x,y)\in A} p(x,y) \sum_{\substack{x' \neq x \\ x' \in T_y}} 2^{-nR_X} \\ &\le \sum_{(x,y)\in A} p(x,y)|T_y|2^{-nR_X} \\ \text{c)} \quad &\le \sum_{(x,y)\in A} p(x,y) 2^{n(H(X|Y)+\varepsilon)-nR_X} \\ \text{d)} \quad &\le 2^{n(H(X|Y)+\varepsilon)-nR_X} \end{aligned} \tag{15}$$

where a) follows from the union of events bound on $\{x' \in T_y, x' \in B_{i(x)}\}$, and the observation that the only random variable in this event is the random partitioning $\{B_i\}_1^M$; b) follows from $\Pr\{x' \in B_i\} = 2^{-nR_X}$, for all i and all $x' \in \mathscr{X}^n$, $x' \neq x$; c) follows from (6); and d) follows from $\sum_A p(x,y) \leq 1$. Thus for $R_X > H(X \mid Y) + \varepsilon + (\log \varepsilon)/n$, $\Pr\{E_1\} < \varepsilon$, therefore, $\bar{P}_e \leq 2\varepsilon$. Since $\varepsilon > 0$ can be chosen arbitrarily small, and n can then be chosen arbitrarily large, we conclude that $R_X > H(X \mid Y)$, $R_Y > H(Y)$ implies the existence of a sequence of block codings of length n, $n = 1,2,\cdots$, such that $\bar{P}_e \to 0$. Finally, if $\bar{P}_e \leq \varepsilon$, there must exist at least one deterministic partition $B_1, B_2, \cdots, B_M$, $M = 2^{nR_X}$, such that $P_e \leq \varepsilon$. We have now shown that there exists a deterministic encoding $i(x), j(y)$ and a decoding $g(i,j) = (\hat{x}, \hat{y})$, such that $\Pr\{(\hat{X}, \hat{Y}) \neq (X,Y)\} \leq \varepsilon$.

Finally, by precisely similar arguments, using the asymptotic partition property for the collection of M jointly ergodic stochastic processes $\{X_i^{(m)}\}_{i=1}^{\infty}$, $m = 1,2,\cdots,M$, we have Theorem 2.

Theorem 2: M jointly ergodic countable-alphabet stochastic processes can be sent separately at rates $R_1, R_2, \cdots, R_M$ to a common receiver with arbitrarily small probability of error, if and only if

$$\sum_{i \in S} R_i \geq H(X^{(i)}, i \in S \mid X^{(l)}, i \in S^c)$$
$$= H(X^{(1)}, X^{(2)}, \cdots, X^{(M)}) - H(X^{(i)}, i \in S^c) \quad (16)$$

for all subsets $S \subseteq \{1,2,\cdots,M\}$, where S^c denotes the complement of S, and H denotes the entropy of the set of processes indexed by S conditioned on the processes indexed by S^c.

We have used the obvious extension of (2) to define the entropy H. Note, in the particular case $M = 3$, that these equations coincide with the following equations exhibited by Wolf [6] for the independent identically distributed case:

$$\begin{aligned} R_1 &\geq H(X^{(1)} \mid X^{(2)}, X^{(3)}) \\ R_2 &\geq H(X^{(2)} \mid X^{(1)}, X^{(3)}) \\ R_3 &\geq H(X^{(3)} \mid X^{(1)}, X^{(2)}) \\ R_1 + R_2 &\geq H(X^{(1)}, X^{(2)} \mid X^{(3)}) \\ R_2 + R_3 &\geq H(X^{(2)}, X^{(3)} \mid X^{(1)}) \\ R_1 + R_3 &\geq H(X^{(1)}, X^{(3)} \mid X^{(2)}) \\ R_1 + R_2 + R_3 &\geq H(X^{(1)}, X^{(2)}, X^{(3)}). \end{aligned} \quad (17)$$

Acknowledgment

I would like to thank D. Hughes-Hartogs and T. Wagner for their comments. D. Forney has found a most transparent proof of Shannon's second theorem using the simultaneous AEP idea that he has independently and previously developed in unpublished course notes.

References

[1] D. Slepian and J. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 471–480, July 1973.
[2] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948.
[3] B. McMillan, "The basic theorems of information theory," *Ann. Math. Statist.*, vol. 24, pp. 196–219, 1953.
[4] L. Breiman, "The individual ergodic theorem of information theory," *Ann. Math. Statist.*, vol. 28, pp. 809–811, 1957; correction, *Ann. Math. Statist.*, vol. 31, pp. 809–810, 1960.
[5] K. L. Chung, "A note on the ergodic theorem of information theory," *Ann. Math. Statist.*, vol. 32, pp. 612–614, 1961.
[6] J. Wolf, "Data reduction for multiple correlated sources," in *Proc. 5th Colloquium Microwave Communication*, 1974, pp. 287–295.
[7] R. Ahlswede and J. Körner, "Source coding with side information and a converse for degraded broadcast channels," unpublished paper, 1974.
[8] T. M. Cover, "Some advances in broadcast channels," to appear as ch. in *Advances in Communication Theory*, A. Viterbi, Ed. Englewood, N.J.: Prentice-Hall, 1975.

Optimal Source Codes for Geometrically Distributed Integer Alphabets

ROBERT G. GALLAGER, FELLOW, IEEE, AND DAVID C. VAN VOORHIS, MEMBER, IEEE

***Abstract*—Let $P(i) = (1 - \theta)\theta^i$ be a probability assignment on the set of nonnegative integers where θ is an arbitrary real number, $0 < \theta < 1$. We show that an optimal binary source code for this probability assignment is constructed as follows. Let l be the integer satisfying $\theta^l + \theta^{l+1} \leq 1 < \theta^l + \theta^{l-1}$ and represent each nonnegative integer i as $i = lj + r$ when $j = \lfloor i/l \rfloor$, the integer part of i/l, and $r = [i] \bmod l$. Encode j by a unary code (i.e., j zeros followed by a single one), and encode r by a Huffman code, using codewords of length $\lfloor \log_2 l \rfloor$, for $r < 2^{\lfloor \log l + 1 \rfloor} - l$, and length $\lfloor \log_2 l \rfloor + 1$ otherwise. An optimal code for the nonnegative integers is the concatenation of those two codes.**

Introduction

The Huffman source coding algorithm [1], [2] is a well-known algorithm for encoding the letters of a finite source alphabet into a uniquely decipherable code of minimum expected codeword length. Since the algorithm operates by successively "merging" the least probable letters in the alphabet, it cannot be directly applied to infinite source alphabets. In this correspondence, we show how the Huffman algorithm can be used indirectly to prove the optimality of a code for an infinite alphabet if one can guess what the code should be first. Naturally it is not always easy to guess the structure of an optimal code, but if the structure is simple enough, and if one starts with the simplest cases, guessing often works.

The particular case that we deal with here is that of the nonnegative integers with a geometric probability assignment,

$$P(i) = (1 - \theta)\theta^i, \qquad i \geq 0 \qquad (1)$$

for some arbitrary θ, $0 < \theta < 1$. This particular distribution arises in run-length coding, where if one has an independent letter binary source, with θ being the probability of a zero, then $P(i)$ is the probability of a run of i zeros. The distribution also arises in other ways such as encoding protocol information in data networks.

Golomb [3] has derived optimal codes for the probability assignment in (1) for the special case when $\theta^l = 1/2$, for some integer l. The result in this correspondence can be interpreted as showing that Golomb's code for a particular value of l is optimal not only when $\theta^l = 1/2$, but more generally for θ satisfying

$$\theta^l + \theta^{l+1} \leq 1 < \theta^l + \theta^{l-1}. \qquad (2)$$

It is easy to see that for any θ, $0 < \theta < 1$, there is a unique positive integer l satisfying (2).

Derivation of Result

For this section of the correspondence we consider θ as fixed and l determined by (2). For this θ and l, we define an *m-reduced source*, for any $m \geq 0$, as a source with $m + 1 + l$ letters with the following probabilities:

$$P_m(i) = \begin{cases} (1 - \theta)\theta^i, & 0 \leq i \leq m \\ \dfrac{(1 - \theta)\theta^i}{1 - \theta^l}, & m < i \leq m + l. \end{cases} \qquad (3)$$

Manuscript received August 1, 1974; revised September 26, 1974. This work was supported in part by NASA under Grant NGL 22-009-013 and in part by the National Science Foundation under Grant GK-37582.

R. G. Gallager is with the Department of Electrical Engineering and the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Mass. 02139.
D. C. Van Voorhis is with the System Development Division, IBM, Los Gatos, Calif.