SUBMITTED TO IT, SEPT 76 (PAPER #1858)
(RCVD 9/10/76)

September 1979 (DATE MADE INTO LIDS REPORT)
Under Revision
(BUT NEVER REVISED

LIDS-P-937

SOURCE CODING WITH SIDE INFORMATION AND UNIVERSAL CODING*

by

Robert G. Gallager**

ABSTRACT

Two problems concerning noiseless source coding with side information are considered. The first is a problem earlier considered by Slepian and Wolf in which the decoder has access to the side information but the encoder does not. We show that not only is the maximum reliable transmission rate unaffected by whether or not the encoder has access to the side information, but also the block error probability is essentially unaffected and that, in a sense, all the encoder need know about the source is the alphabet size. The second problem considered is that where neither the encoder nor decoder knows the side information and good performance, is some sense, is required for all values of side information (i.e., universal coding). We show, for *variable length codes, that the minmax redundancy and the maxmin redundancy*, as defined by Davisson, are essentially the same. Finally we establish a similar minmax, maxmin equivalence for error probability with block codes.

---

*This research was conducted in the M.I.T. Laboratory for Information and Decision Systems with partial support provided by the National Science Foundation under Grant NSF/ENG-7719971.

**Room No. 35-206, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139.

Key Words:

This paper has been submitted to IEEE Trans. on Information Theory.

Source Coding with Side Information and Universal Coding

I. Introduction

Consider a source with a finite alphabet, say $\{0,1,\ldots, K-1\}$, whose probabilistic description depends on a side information parameter, y, from some set Y. We let $Q_N(\vec{x}|y)$ denote the probability of any particular block of N letters, $\vec{x} = (x_1,\ldots, x_N)$, from the source, conditional on y. Sometimes we shall assume that Y has a probability distribution $P(y)$ and sometimes not. Also Y might depend on the block length N of source letters under consideration.

We shall consider both block codes and variable length codes for the sources described above. An (N,R) block code for a source is defined as a mapping f from the set X^N of N-tuples of source letters to the set $(1,2,3,\ldots, \lceil 2^{NR} \rceil)$ where $\lceil z \rceil$ denotes the smallest integer greater than or equal to z. The integers $(1,\ldots, \lceil 2^{NR} \rceil)$ are called code words and can be considered to be binary sequences of length $\lceil NR \rceil$, where R is the code rate. Typically the number of code words will be smaller than the number of N-tuples from the source and thus a decoder, mapping the code words back into source N-tuples will make occasional errors. One of our main objectives is to establish bounds on the probability of such errors under various conditions.

An N-tuple variable length code for a source is defined as a mapping f from X^N to the set of binary strings. We assume throughout that each

$\vec{x} \in X^N$ is mapped into a different string (code word) and that the set of such code words is uniquely decodable. The error problem for block codes has been exchanged for the buffering problems that arise with variable length code words. Our main interest here will be to evaluate the expected length of the set of code words under various conditions. For each of the above types of codes, we shall consider two types of decoders, namely decoders using side information and decoders not using side information. A decoder using side information may be viewed as a mapping from the Cartesian product of the code word set and side information set to the set X^N of source N-tuples. Block codes with decoders using side information were first studied by Slepian and Wolf (1973). In their model, the side information was a second source and the pair of sources were memoryless but statistically dependent on each other. They showed that if (and only if) the code rate exceeds the conditional entropy of the first source given the second, then arbitrarily small error probability is achievable in the limit of large block length. This result can be interpreted as saying that lack of side information at the encoder does not effect the minimum rate at which data can be transmitted if the side information is available at the decoder. Cover (1975) generalized the Slepian-Wolf result to the case of a jointly ergodic pair of sources. Wyner (1975) and Ahlswede and Korner (1975) also generalized the result to the situation in which the side information is separately encoded at a low rate and only partial side information is available at the decoder.

In Section 2 we analyze the achievable error probability for block codes and decoders using side information, both as a function of y (the side information) and as an average over a distribution on y. We shall see that the error probability is essentially the same as if the encoder could observe the side information and generate an optimal code for each value of side information. In fact, the results seem to indicate that one can generate essentially optimal block codes for sources knowing nothing about them but the alphabet (obviously, however, the decoding is not so easy). Strangely enough, the derivation of these error probability results is extremely simple. Aside from one observation, the derivation is just a special case of a joint source-channel coding theorem given as exercise 5.16 in Gallager (1968).

A decoder not using side information is just an ordinary decoder which may be viewed as a mapping from the code word set into the set X^N of source N-tuples. It is helpful here to view the source as a collection of sources, one for each value of the side information parameter. Our problem then is to generate a single code and decoder which in some sense is good for all or most sources in the collection. Codes meeting this type of objective are called universal codes, although definitions vary. Davisson (1973) is an excellent reference on universal codes, and we follow his formulation closely. He describes a maxmin and a minmax approach to universal coding. In the maxmin approach, we view nature as

first choosing the worst probability distribution on the side information set Y and then we design an encoder and decoder for expected performance against that distribution. In the minmax approach, we first design the encoder and decoder and nature chooses the worst side information for our choice. As one might guess, the maxmin approach is analytically simpler, but the minmax approach is more important.

In Section 3 we analyze the maxmin and minmax approaches, first for variable length codes, using redundancy (expected code word length minus conditional entropy) as a performance criterion, and then for block codes, using error probability as a performance criterion. For N-tuple variable length codes, we show that the minmax redundancy (the redundancy for the worst $y \in Y$ given the best code), in bits per N-tuple, lies between C_N and C_N+1 where C_N is the capacity of the channel from Y to X^N with transition probabilities $Q_N(\vec{x}|y)$. Davisson derived the same bounds for the maxmin redundancy, so the two approaches are in a sense equivalent. For block codes, we obtain the same sort of result, first deriving a tight upper to the maxmin error probability, and then showing that the same bound applies to the minmax error probability.

We do not analyze situations where the side information is available to the encoder for the following reasons. First suppose the side information is available to both encoder and decoder. Then it suffices, for each y, to construct the best code for the source probabilities $Q_N(\vec{x}|y)$, which is just

the conventional source coding problem without side information. Second suppose the side information is available to the encoder but not the decoder. The decoder can then do no better than have a fixed mapping from code words to source sequences. Given this decoding, the encoder can do no better than map a source sequence into the code word (or shortest code word) for which it will be correctly decoded. Thus the side information in this case is of no use to the encoder.

II Source Coding for a Decoder Using Side Information

Consider an (N,R) block code for a source with encoding function $f: X^N \to \{1,\ldots, \lceil 2^{NR} \rceil\}$. Let $g: [\{1,\ldots, \lceil 2^{NR} \rceil\}, Y] \to X^N$ be the decoder function. A decoding error occurs for source N-tuple $\vec{x}$ and side information value y if $g(f(\vec{x}), y) \neq \vec{x}$. A decoder will minimize the error probability for each $y \in Y$ if it maps (m,y) into the most likely (conditional on y) source word encoded into m. That is, g(m,y) will be some $\vec{x}$ for which $f(\vec{x}) = m$ and for which $Q_N(\vec{x}|y) \geq Q_N(\vec{x}'|y)$ for all $\vec{x}'$ such that $f(\vec{x}') = m$. In what follows we assume all decoders using side information to be optimum in this sense.

Definition: A random ensemble of (N,R) block codes for alphabet $\{0,1,\ldots, k-1\}$ is the set of all (N,R) block codes for that alphabet and a probability measure on these codes with the following properties: a) each

source N-tuple is mapped with equal probability (1/M) into each of the $M = \lceil 2^{NR} \rceil$ code words; b) (pairwise independence) each pair of different source N-tuple, $\vec{x}$, $\vec{x}'$, is mapped $\vec{x} \rightarrow m$, $\vec{x}' \rightarrow m'$ with probability $1/M^2$ for each pair of code words m,m' (not necessarily different).

It is important to observe that this ensemble of codes is independent of the probabilities associated with source words. It is that fact that will make the analysis of the effect of side information at the decoder almost trivially simple. It will also be noticed that the definition above defines a whole class of ensembles of codes, leaving unspecified higher order statistical dependencies between the mappings of different source words. One ensemble in the above class (that considered by Cover (1975)) is the ensemble in which each source word is independently mapped into $\{1,2.,...,M\}$. Another ensemble in the class, which is of greater interest for implementation purposes, is the ensemble of random coset codes (see Section 6.2, Gallager (1968)). Codes in this ensemble are generated as follows: first each source letter is mapped into its binary representation of length log K (all logs here are base 2). This maps each source N-tuple into a binary row vector of length N log K . This row vector is then multiplied (over the field of 2 elements) by a binary $\lceil NR \rceil$ by $N\lceil \log K \rceil$ binary matrix P. The result is then added (modulo 2) to an $\lceil NR \rceil$ binary row vector $\vec{z}$ to yield the code word. The ensemble of codes is formed by

choosing such binary digit in P and $\vec{z}$ independently with equiprobable 1's and 0's.*

Given a particular random ensemble of (N,R) codes, given a source with probabilities $Q_N(\vec{x}|y)$ for some value of side information y, and given an optimal decoder for each code in the ensemble, we want to investigate the ensemble average probability of decoding error, $P_e(y)$, for that value of side information.

Theorem 1:** For all ρ, $0 \leq \rho \leq 1$, and $M = \lceil 2^{NR} \rceil$

$$P_e(y) \leq M^{-\rho} \left[\sum_{\vec{x}_N} Q_N(\vec{x}|y)^{\frac{1}{1+\rho}} \right]^{1+\rho} \tag{1}$$

Proof: A source sequence $\vec{x}$ is decoded incorrectly for an encoder f if f maps some more likely $\vec{x}'$ into the same code word. Thus

$$P_e(y) \leq \sum_{\vec{x}} Q_N(\vec{x}|y) \; P_N\left[\bigcup_{\vec{x}' \neq \vec{x}:\; Q_N(\vec{x}'|y) \geq Q_N(\vec{x},y)} \{f:\; f(\vec{x}') = f(\vec{x})\} \right] \tag{2}$$

---

*The random vector $\vec{z}$ is needed only to satisfy the definition. It will be seen in retrospect that $\vec{z}$ could be omitted without changing any of the results.

**Aside from the presence of the side information y, which is really immaterial, this theorem is a special case of problem 5.16 in Gallager (1968).

Now, for any set of events $\{A_i\}$, $P_N[\bigcup_i A_i] \leq [\sum_i P_N(A_i)]^\rho$ for any ρ, $0 \leq \rho \leq 1$. This can be easily verified by considering separately the case where $\sum P_N(A_i) \leq 1$ and when $\sum P_N(A_i) > 1$. Also, from definition 2, $P_N\{f: f(\vec{x}') = f(\vec{x})\} = 1/M$. Thus,

$$P_e(y) \leq \sum_{\vec{x}} Q_N(\vec{x}|y) \left[\sum_{\vec{x}' \neq \vec{x}:\ Q_N(\vec{x}'|y) \geq Q_N(\vec{x},y)} 1/M \right]^\rho \tag{3}$$

Using the familiar Chernoff bounding technique, this is upper bounded, for any $s \geq 0$, by

$$P_e(y) \leq \sum_{\vec{x}} Q_N(\vec{x}|y) \left[\sum_{\vec{x}'} \left(\frac{Q_N(\vec{x}'|y)}{Q_N(\vec{x}|y)} \right)^s \frac{1}{M} \right]^\rho \tag{4}$$

Choosing $s = 1/(1+\rho)$, this reduces to (1).

It is convenient now to rewrite (1) in the form

$$\log P_e(y) \leq -\rho NR + NE_s(\rho,N,y) \quad ; \quad 0 \leq \rho \leq 1 \tag{5}$$

$$E_s(\rho,N,y) \triangleq \frac{(1+\rho)}{N} \log \sum_{\vec{x}} Q_N(\vec{x}|y)^{\frac{1}{1+\rho}} \tag{6}$$

Theorem 2: $E_s(\rho,N,y)$ is a convex $\cup$ function of ρ for $\rho \geq 0$, with strict convexity unless $Q_N(\vec{x}|y)$ is constant for all $\vec{x}$ such that $Q_N(\vec{x}|y) > 0$. Furthermore

$$E_s(\rho,N,y)\Big|_{\rho=0} = 0 \tag{7}$$

$$\frac{\partial E_s(\rho,N,y)}{\partial \rho} = \frac{1}{N} H(X_\rho^N|y) \tag{8}$$

where $H(X_\rho^N|y)$ is the entropy of the tilted distribution

$$Q_{N,\rho}(\vec{x}|y) = \frac{Q_N(\vec{x}|y)^{1/(1+\rho)}}{\sum_{\vec{x}} Q_N(\vec{x}|y)^{1/(1+\rho)}} \tag{9}$$

$$H(X_\rho^N|y) = \sum_{\vec{x}} -Q_{N,\rho}(\vec{x}|y) \log_2 Q_{N,\rho}(\vec{x}|y) \tag{10}$$

Proof: The convexity follows from Holder's inequality, and is a trivial variation of lemma 5B1. in Gallager (1968). Equations (7) and (8) follow from straightforward calculation and differentiation.

With this theorem it is easy to optimize the bound over ρ. Figure 1 shows the optimization geometrically. Analytically, the solution is most

conveniently represented giving R and the bound on (1/N) log $P_e(y)$ parametrically as functions of $(0 \leq \rho \leq 1)$.

$$NR = H(X_\rho^N|y) \tag{11}$$

$$\log P_e(y) \leq - H(X_\rho^N||X^N|y) \tag{12}$$

where $H(X_\rho^N||X^N|y)$ is the generalized entropy given by

$$H(X_\rho^N||X^N|y) = \sum_{\vec{x}} Q_{N,\rho}(\vec{x}|y) \log \frac{Q_{N,\rho}(\vec{x}|y)}{Q_N(\vec{x}|y)} \tag{13}$$

This parametric form can be used only for

$$H(X^N|y) \leq NR \leq H(X_\rho^N|y)_{\rho=1} \tag{14}$$

For larger values of R, (5) is optimized by $\rho = 1$; for smaller values of R, we have the unsurprising bound $P_e(y) \leq 1$.

With our current level of generality, it is difficult to make any statements about the tightness of the bound, so we now consider the special case considered by Slepian and Wolf (1973) in which the source is memoryless and the side information is another discrete memoryless source

correlated with the first. We take $Q(x|y)P(y)$ as the probability that the first source emits letter x and the second source emits y. Successive pairs of letters are independent. With this model, we represent the side information as an N-sequence, $\vec{y} = (y_1, y_2, \ldots, y_N)$ of letters from an alphabet $\{0,1,\ldots, J\}$, and we have

$$Q_N(\vec{x}|\vec{y}) = \prod_{n=1}^{N} Q(x_n|y_n) \tag{15}$$

The tilted source of (9) then becomes

$$Q_{N,\rho}(\vec{x}|\vec{y}) = \prod_{n=1}^{N} Q_\rho(x_n|y_n) \tag{16}$$

$$Q_\rho(x|y) = \frac{Q(x|y)^{1/(1+\rho)}}{\sum_x Q(x|y)^{1/(1+\rho)}} \tag{17}$$

Equations (11) and (12) then become

$$NR = \sum_{n=1}^{N} H(X_\rho|y_n) \tag{18}$$

$$\log P_e(\vec{y}) \leq \sum_{n=1}^{N} - H(X_\rho||X|y_n) \tag{19}$$

where

$$H(X_\rho|y) = \sum_x - Q_\rho(x|y) \log Q_\rho(x|y) \tag{20}$$

$$H(X_\rho||X|y) = \sum_x Q_\rho(x|y) \log \frac{Q_\rho(x|y)}{Q(x|y)} \tag{21}$$

Theorem 3: For any set of transition probabilities $Q(x|y)$ and all $\rho > 0$, there exists $A < \infty$ such that for every (N,R) source code and every $\vec{y}$, if

$$NR < \sum_{n=1}^{N} H(X_\rho|y_n) - A\sqrt{N} \tag{22}$$

then

$$\log P_e(\vec{y}) \geq \sum_{n=1}^{N} - H(X_\rho||X|y_n) - A\sqrt{N} \tag{23}$$

Note that theorem 3 asserts that for R in the range given by (14), the exponent to error probability (in the limit of large N) over the random ensemble of codes is the same as the exponent for the best code constructed using knowledge of $\vec{y}$ at the encoder.

Proof: We use theorem 5 from Shannon, Gallager, and Berlekamp (1967), which states that if $P_1(\vec{x})$ and $P_2(\vec{x})$ are two probability assignments on a space, if X, is some subset of this space, and if

$$P_{e,1} = \sum_{\vec{x} \in X_1^c} P_1(\vec{x}) \quad ; \quad P_{e,2} = \sum_{\vec{x} \in X_1} P_2(\vec{x}) \tag{24}$$

then for all s, $0 < s < 1$, either

$$P_{e,1} \geq \frac{1}{4} \exp\left[\mu(s) - s\mu'(s) - s\sqrt{2\mu''(s)}\right] \tag{25}$$

or

$$P_{e,2} \geq \frac{1}{4} \exp\left[\mu(s) + (1-s)\mu'(s) - (1-s)\sqrt{2\mu''(s)}\right] \tag{26}$$

where $\mu(s) = \ln \sum_{\vec{x}} P_1(\vec{x})^{1-s} P_2(\vec{x})^s$

For our application, $P_1(\vec{x}) = Q_N(\vec{x}|\vec{y})$ and $P_2(\vec{x}) = K^{-N}$. We also take X_1 to be the set of source sequences that are correctly decoded, so that $P_{e,1} = P_e(\vec{y})$. Finally since only 2^{NR} source sequences can be correctly decoded, $P_{e,2} \leq \lceil 2^{NR} \rceil K^{-N}$. We can rewrite $\mu(s)$ as

$$\mu(s) = \sum_{n=1}^{N} \ln \sum_{x} Q(x|y_n)^{1-s} K^{-s} \tag{27}$$

Since the two source alphabets are finite, $\mu''(s)$ can be upper bounded by N times some constant that is independent of $\vec{y}$. Taking $s = \frac{\rho}{1+\rho}$, and evaluating (25) and (26) (except for the bound on $\mu''(s)$) we get (22) and

(23), where the multiplicative factors of 1/4 have been incorporated into the constant A.

There is a more intuitive way of seeing the result of theorem 3. Consider the class of source sequences $\vec{x}$ for which $Q_\rho(x|y)$ approximately specifies the fraction of appearances of letter x in positions where $\vec{y}$ has letter y. There are approximately $2^{\sum_n H(X_\rho|y_n)}$ such sequences, and for the R given in the theorem, most of these will be incorrectly decoded. Each of these sequences will have a probability close to $2^{\sum_n \sum_x Q_\rho(x|y_n) \log Q(x|y_n)}$, and thus these sequences alone lead to the error probability predicted in (23).

Next let us assume a probability assignment P(y) on the space of side information values. Clearly the probability of error, over the ensemble of codes and the ensemble of side information values, is

$$P_e = \sum_y P(y)\, P_e(y)$$

$$\leq 2^{-\rho NR} \sum_y P(y) \left[\sum_{\vec{x}} Q_N(\vec{x}|y)^{1/(1+\rho)} \right]^{1+\rho} ; \quad 0 \leq \rho \leq 1 \tag{28}$$

It appears that one might get a tighter bound in (28) by using a different value of ρ for each y, but as we shall soon see, the improvement is unimportant. The logarithm of the sum over y in (28) is convex (as can be seen by applying Holder's inequality and using the convexity of $E_s(\rho,N,y)$ given in (6)). Thus the right hand side of (28) can be minimized over ρ. The result, after some manipulation, is the following set of parametric equations.

$$NR = H(X_\rho^N | Y_\rho) \tag{29}$$

$$\log P_e \leq - H(X_\rho^N \; Y \; || \; X^N \; Y \;) \tag{30}$$

where the conditional entropy in (29) is

$$H(X_\rho^N | Y_\rho) = \sum_{\vec{x}} - P_\rho(y) \; Q_{N,\rho}(\vec{x}|y) \; \log \; Q_{N,\rho}(\vec{x}|y) \tag{31}$$

and the generalized entropy in (30) is

$$H(X_\rho^N \;\; Y_\rho \; || X^N \; Y) = \sum_{\vec{x},\vec{y}} P_\rho(y) Q_{N,\rho}(\vec{x}|y) \; \log \; \frac{P_\rho(y) Q_{N,\rho}(\vec{x}|y)}{P(y) Q_N(\vec{x}|y)} \tag{32}$$

The tilted probability $Q_{N,\rho}$ is given by (9) and P_ρ is given by

$$P_\rho(y) = \frac{P(y) \sum_{\vec{x}} Q_N(\vec{x}|y)^{1/(1+\rho)}}{\sum_y P(y) \sum_{\vec{x}} Q_N(\vec{x}|y)^{1/(1+\rho)}}^{1+\rho}$$

These equations are valid for R in the range

$$H(X^N|Y) \leq NR \leq H(X_\rho^N|Y_\rho)\Big|_{\rho=1} \tag{33}$$

and for larger values of R, (28) is optimized by $\rho = 1$.

For the correlated memoryless sources of Slepian and Wolf, all of the probabilities above factor, and the entropies become N times the corresponding single letter entropies. This proves the first half of the following theorem and also shows that P_e decays exponentially with block length N for all $R > H(X|Y)$.

Theorem 4: a) For a correlated memoryless source with single letter probabilities $P(y)Q(x/y)$ and with the side information source output $\vec{y}$ available at the decoder, the probability of error for a random ensemble of (N,R) codes, averaged over both sources, satisfies the parametric equations

$$R = H(X_\rho|Y) \quad ; \quad H(X|Y) \leq R \leq H(X_\rho|Y_\rho)_{\rho=1}$$

$$\log P_e \leq -N\, H(X_\rho Y_\rho || X\, Y) \tag{34}$$

$$\log P_e \leq -NR + N \log \sum_y P(y) \left[\sum_x Q(x\, y)^{1/(1+\rho)} \right]^{1+\rho} \quad ; \quad R > H(X_\rho|Y_\rho)_{\rho=1} \tag{35}$$

b) For any such source there exists a finite constant A such that for all (N,R) encoders (including encoders using the side information $\vec{y}$) and for all $\rho \geq 0$, if $R \leq H(X_\rho|Y_\rho) - A/\sqrt{N}$, then

$$\log P_e \geq -N\, H(X_\rho Y_\rho\, ||X\, Y) - A\sqrt{N} \tag{36}$$

Proof of part b: The proof is a minor modification of that of theorem 3. In place of $P_1(\vec{x})$ in the Shannon, Gallager, and Berlekamp theorem, we use $P_N(\vec{y})Q_N(\vec{x}|\vec{y})$, and in place of $P_2(\vec{x})$ we use $K^{-N}P_{N,\rho}(\vec{y})$. Finally we take the region X_1 to be the set of $\vec{x},\vec{y}$ such that with side information $\vec{y}$, $\vec{x}$ is decoded from some code word. Since at most 2^{NR} source sequences can be decoded for each $\vec{y}$, we again have $P_{e,2} \leq 2^{NR} K^{-N}$. This leads, after some straightforward calculation to (36).

Next we make the trivial observation that at least one code in an (N,R) random ensemble has an error probability as small as the ensemble

average. This code encodes source sequences into code words without use of the side information, and from theorem 4 we see that this code is substantially as good as the best code using the side information at the encoder if N is large and R is the range of (34). For rates larger than the range of (34), the exponents in the upper and lower bounds of theorem 4 differ. It can be shown, by a slight modification of the argument in Chapter 5 of Jelinek (1968), that the best code using side information at the encoder has the exponent given by the lower bound. The exponent for the best codes not using side information at the encoder is unknown at these high rates and might be inferior to the lower bound exponent.

A more important open question is whether codes exist (not using side information at the encoder) which are uniformly good in the sense of almost satisfying (11) and (12) for each value of side information $\vec{y}$. There is one example, in which X and Y are both binary, with $P(0) = P(1) = 1/2$ and $Q(1|0) = 0$, $Q(1|1) = 1/2$, for which it can be shown that no code in the coset random ensemble previously discussed is uniformly good. It is not clear, however, whether less structured codes can be uniformly good.

Since we have now seen that side information is of very little use to an encoder for block codes, it is reasonable to ask whether the same type of result is true for variable length codes. The answer, surprisingly enough, is no. The reason for this is quite simple and depends on our assumption that variable length codes must be error free (i.e. uniquely

decodable). This means that the encoder must provide a different code word for each source sequence,* and thus the side information is of no use to the decoder.

III Source Coding For Decoders Not Using Side Information

At first glance this problem seems uninteresting; if the side information is unavailable, one should simply average over it and encode for the average source. For the Slepian and Wolf type correlated memoryless sources, this approach certainly makes sense. However, if the side information is unchanging or very slowly changing in time, then one wants a source code that is in some sense universally good for all values of the side information

Variable Length Codes

We first consider variable length codes, both because of their inherent interest and because of their potential for practical applications. An N-tuple variable length encoder (encoder for short) is a mapping from X^N (the set of source sequences of length N) into the set of finite length binary strings. For each N-tuple $\vec{x}$ from the source, let $\ell(\vec{x})$ be the length of the binary string that $\vec{x}$ is encoded into. The Kraft inequality,

---

*One exception to this is where some source sequences have zero probability for some values of side information.

$$\sum_{\vec{x}} 2^{-\ell(\vec{x})} \leq 1 \qquad (37)$$

must be satisfied by any uniquely decodable code and a binary prefix condition code can be constructed for any set of non-negative integer lengths satisfying (37) (see Gallager (1968)). Because of these facts, we can consider only the set of lengths in a code and ignore the actual encoded strings. Thus a code can be considered as a non-negative integer valued function ℓ satisfying (37).

The redundancy of a code ℓ, for a particular value of side information y, is defined to be

$$r_N(\ell,y) = \frac{1}{N}\left[\sum_{\vec{x}} Q_N(\vec{x}|y)\ \ell(\vec{x}) - \frac{1}{N} H(X^N|y)\right] \qquad (38)$$

This is just the expected length of the code, given y, minus the entropy of the source (conditional on that y), normalized by the source block length N. It is well known from the elementary source coding theorem that $r_N(y,\ell) \geq 0$, with the value 0 only if $\ell(\vec{x}) = -\log Q_N(\vec{x}|y)$ for all $\vec{x}$, and thus the smallness of $r_N(y,\ell)$ is a reasonable measure of how effective the code is for a particular y.

Davisson now defines maxmin redundancy $\mathcal{R}_N^-$ and minmax redundancy $\mathcal{R}_N^+$ by

$$\mathcal{R}_N^- = \sup_P \min_{\ell \in \mathcal{L}} \sum_y P(y)\, r_N(y,\ell) \tag{39}$$

$$\mathcal{R}_N^+ = \min_{\ell \in \mathcal{L}} \sup_{y \in Y} r_N(y,\ell) \tag{40}$$

The minimization in (39) is over the set $\mathcal{L}$ of non negative integer functions ℓ satisfying the Kraft inequality (37). The supremum is over the set of probability measures on Y. Here we are tacitly assuming Y to be a finite set, although in the appendix, it is shown that the results are valid for an arbitrary measurable set Y.

The maxmin redundancy, $\mathcal{R}_N^-$ is the expected redundancy that results if nature first perversely picks a distribution P(y) to maximize the redundancy, and then we observe nature's choice and choose a code to minimize redundancy. For $\mathcal{R}_N^+$, we must choose the code first, and then nature chooses the most unfavorable value of y. Davisson (1973) has shown that $\mathcal{R}_N^+ \geq \mathcal{R}_N^-$ and also that $C_N \leq N\mathcal{R}_N^- \leq (C_N+1)$ where C_N is the capacity of the channel with input alphabet Y, output alphabet X^N and transition probabilities $Q_N(\vec{x}|y)$. We shall show, in addition that $C_N \leq N\mathcal{R}_N^+ \leq (C_N+1)$. One consequence of this is that if a source with side information has $\lim_{N\to\infty} C_N/N = 0$, then the redundancy can be made uniformly arbitrarily small for all values of side information by using variable length encoders of sufficiently large block length. Asymptotically, knowledge of the side information at the encoder and decoder is not of any use.

It turns out that the major difficulty in evaluating $\mathcal{R}_N^-$ and $\mathcal{R}_N^+$ comes from the integer constraint in the code word lengths. Thus our strategy will be to define two new quantities $\hat{\mathcal{R}}_N^-$ and $\hat{\mathcal{R}}_N^+$ without the integer constraint. We will relate these to $\mathcal{R}_N^-$ and $\mathcal{R}_N^+$ and then we will evaluate them. Let $\hat{\mathcal{L}}$ be the class of all functions $\ell(\vec{x})$ mapping X^N into the non-negative real numbers satisfying the Kraft inequality (37). Then define

$$\hat{\mathcal{R}}_N^- = \sup_P \min_{\ell \in \hat{\mathcal{L}}} \sum_y P(y)\, r_N(y,\ell) \tag{41}$$

$$\hat{\mathcal{R}}_N^+ = \min_{\ell \in \hat{\mathcal{L}}} \sup_{y \in Y} r_N(y,\ell) \tag{42}$$

The key to relating the constrained and unconstrained redundancies lies in the observation that if ℓ is any non-integer length function satisfying the Kraft inequality (37), then the function $\lceil \ell \rceil$ formed by increasing each length to an integer (i.e. $\lceil \ell \rceil(\vec{x}) = \lceil \ell(\vec{x}) \rceil$) also satisfies the Kraft inequality. Thus $\ell \in \hat{\mathcal{L}}$ implies $\lceil \ell \rceil \in \mathcal{L}$. From (38) we see that for all $y \in Y$, $\ell \in \hat{\mathcal{L}}$,

$$r_N(\ell,y) \leq r_N(\lceil \ell \rceil, y) < r_N(\ell,y) + 1/N \tag{43}$$

Lemma:

$$\hat{\mathcal{R}}_N^+ \leq \mathcal{R}_N^+ \leq \hat{\mathcal{R}}_N^+ + 1/N \tag{44}$$

$$\hat{\mathcal{R}}_N^- \leq \mathcal{R}_N^- \leq \hat{\mathcal{R}}_N^- + 1/N \tag{45}$$

Proof: The left hand inequalities are almost obvious consequences of the fact that $\mathcal{L} \subset \hat{\mathcal{L}}$; when one minimizes over a larger set, one gets a smaller or equal result. The right hand inequalities come from (43). We demostrate $\mathcal{R}_N^+ \leq \hat{\mathcal{R}}_N^+ + 1/N$ in detail; minor variations establish the others. For any given $\ell \in \hat{\mathcal{L}}$ let $\{y_i\}$ be a sequence of elements in Y such that

$$\lim_{i\to\infty} r_N(\lceil \ell \rceil, y_i) = \sup_y r_N(\lceil \ell \rceil, y) \tag{44}$$

From (43),

$$r_N(\lceil \ell \rceil, y_i) < r_N(\ell, y_i) + 1/N \leq \sup_y r_N(\ell, y) + 1/N \tag{45}$$

Combining (44) and (45)

$$\sup_y r_N(\lceil \ell \rceil, y) \leq \sup_y r_N(\ell, y) + 1/N \tag{46}$$

Next let ℓ' minimize the right hand side of (46) over $\ell \in \hat{\mathcal{I}}$.

$$\mathcal{R}_N^+ \leq \sup_y r_N(\lceil \ell' \rceil, y) \leq \sup_y r_N(\ell', y) + 1/N$$

$$= \min_{\ell \in \hat{\mathcal{I}}} \sup_y r_N(\ell, y) + 1/N = \hat{\mathcal{R}}_N^+ + 1/N \tag{47}$$

Theorem 5:

$$C_N/N = \hat{\mathcal{R}}_N^- = \hat{\mathcal{R}}_N^+ \tag{48}$$

$$C_N/N \leq \mathcal{R}_N^- \leq (C_N + 1)/N \tag{49}$$

$$C_N/N \leq \mathcal{R}_N^+ \leq (C_N + 1)/N \tag{50}$$

where C_N is the capacity of the channel from Y to X^N with transition probabilities $Q_N(\vec{x}|y)$.

Proof: It suffices to establish (48) since (49) and (50) then follow from the previous lemma. First we show that $N\hat{\mathcal{R}}_N^- = C_N$. Combining (41) and (38),

$$N\hat{\mathcal{R}}_N^- = \sup_P \min_{\ell \in \hat{\mathcal{I}}} \sum_y P(y) \sum_{\vec{x}} Q_N(\vec{x}|y)\, \ell(\vec{x}) - H(X^N|y) \tag{51}$$

For a given P, $\sum_{y} P(y) \sum_{\vec{x}} Q_N(\vec{x}|y)\ell(\vec{x})$ is minimized (subject to the Kraft inequality by

$$\ell(\vec{x}) = -\log w_p(\vec{x}) \tag{52}$$

$$w_p(\vec{x}) = \sum_{y} P(y)\, Q_N(\vec{x}|y)$$

This minimization is precisely the same as that used in the elementary source coding theorem to show that the expected length of a variable length code exceeds the source entropy (Section 3.3, Gallager (1968). Substituting (52) into (51) and writing $H(X^N|y)$,

$$N\hat{R}_N = \sup_{P} \sum_{y,\vec{x}} P(y)\, Q_N(\vec{x}|y) \log \frac{Q_N(\vec{x}|y)}{\sum_{y'} P(y')Q_N(\vec{x}|y')} \tag{53}$$

The right hand side of (53) is, by the definition of capacity, C_N. For the moment, assume Y to be a finite set. Then necessary and sufficient conditions on the P, say P_0, that maximizes the average mutual information on the right hand side of (53) are:

$$\sum_{\vec{x}} Q_N(\vec{x}|y) \log \frac{Q_N(\vec{x}|y)}{w_0(\vec{x})} \leq C_N \quad ; \quad \text{all } y \tag{54}$$

with equality for all y such that $P_0(y) > 0$ and where $w_0(\vec{x}) = \sum_y P_0(y) Q_N(\vec{x}|y)$ (theorem 4.5.1, Gallager (1968)). The appendix provides an equivalent version of this result for the general case where Y is an arbitrary measurable set. From (42) and (38) we have

$$N\hat{\mathcal{R}}_N^+ = \min_{\ell \in \hat{\mathcal{L}}} \sup_{y \in Y} \sum_{\vec{x}} Q_N(\vec{x}|y)\ell(\vec{x}) - H(X^N|y) \tag{55}$$

Now let $\ell(x)$ satisfy (52) for the P_0 that satisfies (54). For this $\ell(x)$,

$$\sup_y \sum_{\vec{x}} Q_N(\vec{x}|y)\ell(\vec{x}) - H(X^N|y) = \sup_y \sum_{\vec{x}} Q_N(\vec{x}|y) \log \frac{Q_N(\vec{x}|y)}{w_0(\vec{x})}$$

$$\leq C_N \tag{56}$$

where we have used (54). Comparing (55) and (56), we have $N\hat{\mathcal{R}}_N^+ \leq C_N$. Finally let ℓ' minimize (55) and let P(y) be an arbitrary probability distribution on y. Then

$$\hat{\mathcal{R}}_N^+ = \sup_{y \in Y} r_N(y,\ell') \geq \sum_y P(y)\, r_N(y,\ell')$$

$$\geq \min_{\ell \in \hat{\mathcal{L}}} \sum_y P(y)\, r_N(y,\ell) \quad ; \quad \text{all } P(y) \tag{57}$$

From the definition of $\hat{\mathcal{R}}_N^-$, then, $\hat{\mathcal{R}}_N^+ \geq \hat{\mathcal{R}}_N^-$, completing the proof.

One can make a number of conjectures about the actual values of $\mathcal{R}_N^-$ and $\mathcal{R}_N^+$ within the limits of (49) and (50), but most of them turn out to be false. For example, it is not true in general that $\mathcal{R}_N^- = \mathcal{R}_N^+$. It is also not true in general that if one finds the input probabilities P(y) that lead to capacity, and then finds the corresponding output probabilities $w(\vec{x}) = \sum_y P(y) Q_N(\vec{x}|y)$, that the Huffman code generated from $w(\vec{x})$ will achieve either $\mathcal{R}_N^-$ or $\mathcal{R}_N^+$.

Block Codes

We now study block codes where neither the encoder nor decoder observe the side information. As in Section 2, we shall be concerned with bounds on the error probability as a function of the rate R, the block length N, and the value of side information y. This problem was studied earlier by Ziv (1972) who developed an encoding strategy and showed that for any particular value of side information, if the code rate exceeds the source entropy conditional on that value of side information, then the error probability goes to zero as block length increases. Our results will indicate that the choice of encoder is not critical (being based on random coding) and also will provide rather tight bounds on error probability.

To begin, let us assume a probability assignment P(y) on the side information. We also use the same ensemble of codes as in Section 2. The decoder will minimize error probability, over the side information ensemble,

by decoding code word m into the source word $\vec{x}$ that maximizes $w_p(\vec{x}) = \sum_y P(y) Q_N(\vec{x}|y)$ over all $\vec{x}$ encoded into m.

Let $P_e(y,P)$ be the average error probability, over the ensemble of codes, when y is the value of side information and P is the probability measure on y assumed by the decoder. By repeating the steps of the proof of Theorem 1, we find that, for all ρ, $0 \leq \rho \leq 1$,

$$P_e(y,P) \leq 2^{-NR\rho} \sum_{\vec{x}} Q_N(\vec{x}|y)\, w_p(\vec{x})^{-\frac{\rho}{1+\rho}} \left[\sum_{\vec{x}'} w_p(\vec{x}')^{\frac{\rho}{1+\rho}} \right]^{\rho} \tag{58}$$

Averaging over y according to P(y), this simplifies to

$$P_e(P) \leq 2^{-NR\rho} \left[\sum_{\vec{x}} w_p(\vec{x})^{\frac{1}{1+\rho}} \right]^{1+\rho} \tag{59}$$

It is important to note that the measure P in (58) appears only because of the decoder's decision rule; (58) is valid for that decision rule for any probability assignment (or none) on y.

As in our treatment of variable length codes, it is useful to have results independent of some assumed distribution P(y). Thus we define P_e^- as the error probability (over the code ensemble) that results if nature first chooses P(y) to maximize $P_e(P)$, but the decoder is allowed to use natures choice in its decisions. Thus

$$P_e^- \leq \sup_P 2^{-NR\rho} \left\{ \sum_{\vec{x}} \left[\sum_y P(y) Q_N(\vec{x}|y) \right]^{1/(1+\rho)} \right\}^{1+\rho} \tag{60}$$

Fortunately the term in braces is convex$\cap$ and it is easy to calculate the following necessary and sufficient conditions for P*, the maximizing P,

$$\sum_{\vec{x}} Q_N(\vec{x}|y)\, w_{P*}(\vec{x})^{-\rho/(1+\rho)} \leq \sum_{\vec{x}} w_{P*}(\vec{x})^{1/(1+\rho)} \qquad (61)$$

with equality for y such that $P^*(y) > 0$. This result (and in fact the whole concept of $\bar{P}_e$) would not be terribly interesting were it not for the striking and fortuitous resemblance between (61) and (58).

Using (61) to upper bound (58), we then get

$$P_e(y,P^*) \leq 2^{-NR\rho}\left[\sum_{\vec{x}} w_{P*}(\vec{x})^{1/(1+\rho)}\right]^{1+\rho} \qquad (62)$$

This is very nice, since it says that if the decoder assumes the worst P(y) for its decoding rule, then the resulting bound on error probability is uniformly good over all values of side information.

This bound can be written in parametric form, like Equations (11) to

(14), but we shall not repeat the Equations here. The lower limit on the rate, however, for which the bound is less than 1, is given by

$$NR > \sup_{P} \left[-w_p(\vec{x}) \log w_p(\vec{x})\right] \tag{64}$$

In other words, this uniform bound on error probability is only useful when R exceeds the unconditional source entropy per digit, maximized over P.

Next we recall that $P_e(y,P^*)$ is an average error probability over an ensemble of codes, and the question arises whether individual codes exist which are uniformly good against all values of y. Fortunately, if the set Y is finite and not too large, the answer is yes. We assume that the decoder uses probability assignment P^*, but we use a uniform probability assignment on Y. The right hand side of (63) (optimized over ρ) is an upper bound to error probability over the ensemble of codes and over this uniform distribution on Y. We pick a code from the ensemble that is as good as the average and note that for at least half the set Y, we must have

$$P_e(y,P^*) \leq 2 \cdot 2^{-NR\rho}\left[\sum_{\vec{x}} w_{p^*}(\vec{x})^{1/(1+\rho)}\right]^{1+\rho} \tag{65}$$

We next assign zero probability to all y in the code that satisfy (65), and uniform probability again to the remaining y. We then pick another code from the ensemble that is as good as the average for this new distribution on Y. Half of the remaining y must satisfy (65) for

this new code. We continue this process, exhausting the set Y after choosing $m \leq \lceil \log |Y| \rceil$ codes, where $|Y|$ is the number of elements in Y. We now combine all these codes into one code with $m\, 2^{NR}$ code words. The encoder, given $\vec{x}$, encodes it into the first of the m codes for which the decoder will decode $\vec{x}$ (if $\vec{x}$ is not decoded for any of the codes, it makes no difference what the encoder does). This new code is decoded correctly whenever any of the m codes would have decoded correctly, and thus (65) is satisfied for every $y \in Y$. The actual rate of the new code is $R' = R + (\log m)/N \leq R + [\log \lceil \log |Y| \rceil]/N$. We have thus proved the following theorem.

Theorem 6: Given any N and R, given a source with side information with the probability assignment $Q_N(\vec{x}|y)$, there exists an (N,R) block encoder mapping $X^N \to (1, \ldots, 2^{NR}) \to X^N$ with an error probability for each y satisfying

$$P_e(y) \leq \min_{0 \leq \rho \leq 1} \max_{P} 2^{-\rho(NR - \log \lceil \log |Y| \rceil) + 1} \left[\sum_{\vec{x}} \sum_{y} P(y) Q_N(\vec{x}|y)^{\frac{1}{1+\rho}} \right]^{1+\rho} \tag{66}$$

Appendix

We want to consider a channel with an arbitrary input space Y, a finite set of outputs $X = \{1, \ldots, k\}$, and a set of transition probabilities $Q(x|y)$ for all $x \in X$, $y \in Y$. X here corresponds to the set X^N in Section 3. We assume throughout that Y, along with an appropriate set of subsets, is a measurable space and that $Q(x|y)$, for each $x \in X$, is a measurable function of $y \in Y$. For any probability measure P on the space Y, we make the following definitions:

$$H(X|y) = -\sum_{x} Q(x|y) \log Q(x|y) \tag{A1}$$

$$H_P(X|Y) = \int d\,P(y)\, H(X|y) \tag{A2}$$

$$w_P(x) = \int d\,P(y)\, Q(x|y) \tag{A3}$$

$$\vec{w}_P = (w_P(1), w_P(2), \ldots, w_P(K)) \tag{A4}$$

$$\mathcal{H}(\vec{w}_P) = -\sum_{x} w_P(x) \log w_P(x) \tag{A5}$$

$$I_P(Y;X) = \mathcal{H}(\vec{w}_P) - H_P(X|Y) \tag{A6}$$

The capacity of the channel is then defined as

$$C = \sup_{P} I_P(Y;X)$$

where the supremum is over all probability measures on Y.

Theorem A: In order for a number C to be the capacity of the channel defined above, it is necessary and sufficient for a sequence $\{P_i\}$ of probability measures on Y to exist and for a probability vector $\vec{w}_0 = w_0(1), \ldots, w_0(K))$ to exist with the following properties:

$$1) \quad \lim_i I_{P_i}(Y;X) = C \tag{A7}$$

$$2) \quad \lim_i w_{P_i}(x) = w_0(x) \quad ; \quad \text{all } x \in X \tag{A8}$$

$$3) \quad \sum_x Q(x|y) \log \frac{Q(x|y)}{w_0(x)} \leq C \quad ; \quad \text{all } y \in Y \tag{A9}$$

Furthermore $\vec{w}_0$ is unique.

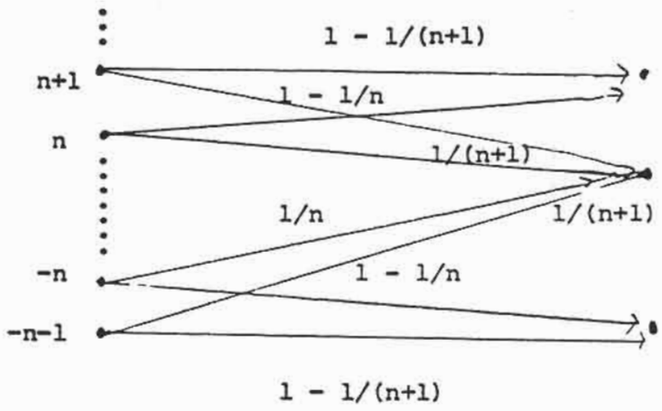
Before proving the theorem, the following erasure channel (see Figure A) will be helpful in understanding both the proof and the wording of the theorem. The input space is the set of positive and negative integers, and the output space is the three letters 0, 1, e. The transition probabilities are given by

$$Q(0|n) = 1 - 1/n \quad ; \quad n > 0$$

$$Q(e|n) = 1/|n| \quad ; \quad n > 0, \; n < 0$$

$$Q(1|n) = 1 - 1/|n| \quad ; \quad n < 0$$

The capacity of this channel is 1 bit, approached by a sequence of probability measures $P_i(n) = 1/2$ for $\pm n = i$. The resulting limiting output distribution is $\vec{w}_0 = (1/2, 1/2, 0)$. However, no input measure either leads to $\vec{w}_0$ or to capacity.



Generalized Erasure Channel

Figure A

Proof of Theorem: Necessity: since the output alphabet is finite, the capacity C clearly exists, and there is a sequence $\{P_i\}$ of input probability measures that satisfy (A7). Since $\vec{w}_{P_i}$ is an element of the compact space of K dimensional probability vectors, the sequence $\{\vec{w}_{P_i}\}$ has a cluster point, which we denote by $\vec{w}_0$. Thus there is a subsequence of $\{P_i\}$, which we redefine as a new sequence $\{P_i\}$, for which (A7) and (A8) are satisfied. Now we assume that (A9) is violated by some y, say y', and show a contradiction to the assumption that C is the capacity. Thus we have

$$\sum_x Q(x|y') \log \frac{Q(x|y')}{w_0(x)} > C \qquad \text{(A10)}$$

For any θ, $0 \leq \theta \leq 1$, let $P_{i,\theta}$ be the convex combinition of P_i with the atomic distribution on y'. That is, for any measurable Z C Y,

$$P_{i,\theta}(Z) = \begin{cases} (1-\theta)\ P_i(Z) & ;\quad y' \notin Z \\ (1-\theta)\ P_i(Z) + \theta & ;\quad y' \in Z \end{cases} \qquad \text{(A11)}$$

From (A2) and (A3) we verify that

$$H_{P_{i,\theta}}(X|Y) = (1-\theta)\ H_{P_i}(X|Y) + \theta\ H(X|y')$$

$$w_{P_{i,\theta}}(x) = (1-\theta) w_{P_i}(x) + \theta\ Q(x|y') \quad ;\ \text{all } x \in X$$

Let $\vec{Q}'$ denote the vector $(Q(1|y'), Q(2|y'), \ldots Q(k|y'))$. Then we can rewrite (A6) as

$$I_{P_{i,\theta}}(Y;X) = \mathcal{H}((1-\theta)\vec{w}_{P_i} + \theta\,\vec{Q}') - (1-\theta)H_{P_i}(X|Y) - \theta\,H(X|y') \tag{A12}$$

Using (A7) and (A8) in (A6), we see that

$$\lim_{i\to\infty} H_{P_i}(X|Y) = \mathcal{H}(\vec{w}_0) - C \tag{A13}$$

Using (A12) and (A13),

$$\lim_{i\to\infty} I_{P_{i,\theta}}(Y;X) = \mathcal{H}((1-\theta)\vec{w}_0 + \theta\vec{Q}') - (1-\theta)[\mathcal{H}(\vec{w}_0) - C] - \theta H(X|y') \tag{A14}$$

Equation (A14) is valid for each θ, $0 \leq \theta \leq 1$. Also the left hand side of (A14) cannot exceed capacity for any θ by definition. Finally the right hand side is equal to capacity at $\theta = 0$. Thus, if the derivative of the right hand side with respect to θ, is positive at $\theta = 0$, we have demonstrated our contradiction. Evaluating this derivative, we obtain

$$\sum_x Q(x|y') \log \frac{Q(x|y')}{w_0(x)} - C \tag{A15}$$

From (A10), this is positive, completing the proof of necessity.

Sufficiency: Suppose that (A7), (A8), and (A9) are satisfied for some sequence of probability measures $\{P_i\}$ and some probability vector $\vec{w}$. From (A7) the capacity can be no less than C, so all we need show is that $I_p(Y;X) \leq C$ for all P. Rewriting (A6),

$$I_p(Y;X) = \int dP(y) \sum_x Q(x|y) \log \frac{Q(x|y)}{w_P(x)}$$

$$= \int dP(y) \sum_x Q(x|y) \log \frac{Q(x|y)}{w_0(x)} + \int dP(y) \sum_x Q(x|y) \log \frac{w_0(x)}{w_P(x)}$$

Th first term above is less than or equal to C from (A9). The second is non-positive as can be seen by interchanging the sum and integral and using (A3) to get $\sum w_P(x) \log (w_0(x)/w_P(x))$, which is the negative of a generalized entropy.

In order to demonstrate the uniqueness of $\vec{w}_0$, we assume another probability vector $\vec{w}_0'$, which, along with some other sequence $\{P_i'\}$, satisfies (A7) and (A8). The sequence $\{\frac{1}{2} P_i + \frac{1}{2} P_i'\}$ then leads to the limiting output probability vector $\frac{1}{2} \vec{w}_0 + \frac{1}{2} \vec{w}_0'$. It is then an easy matter, using the strict convexity $\cap$ of $\mathcal{H}(\vec{w})$ to establish that $\lim_i I_{\frac{1}{2} P_i + \frac{1}{2} P_i'}(Y;X) > C$, which is a contradiction.