OPTIMAL CAUSAL CODING-DECODING PROBLEMS

by

J. Walrand and P. Varaiya

Memorandum No. UCB/ERL M82/3

1 January 1982

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

OPTIMAL CAUSAL CODING-DECODING PROBLEMS

J. Walrand[1] and P. Varaiya[2]

Abstract

The symbols produced by a finite Markov source have to be causally encoded so as to be transmitted through a noisy memoryless channel. The encoder is assumed to have channel feedback information and the decoder to be causal.

The feedback information is shown to be useful in general.

Separation results are derived and used to prove that encoding is useless for a class of symmetric channels.

[1] School of Electrical Engineering, Cornell University, Ithaca, NY 14853. Present address: see below (#2).

[2] Department of Electrical Engineering and Computer Science and the Electronics Research Laboratory, University of California, Berkeley, CA 94720

1. Introduction

The coding theorem for discrete memoryless channels and ergodic sources asserts the possibility of reliable communication under the well known rate constraints (see [1]). This result is made possible by the observation that long ergodic sequences are asymtotically typical and can therefore be coded into sequences having the distribution which achieves the capacity of the channel.

The need for encoding arbitrarily long sequences introduces undesirable delays in the communication systems. It is an important problem to measure the trade-off between reliability and excessive delays. One approach is to consider the rate of decrease of the average probability of decoding error as the length n of the encoded sequences increases. The usual results state that this error is bounded by exp $\{-n\ E\}$ for some constant E depending on the channel and source parameters. (see e.g. [2] chap. 5, 6). These bounds are asymptotically tight but of limited value for short sequences.

In this paper we adopt a different approach for the problem of reliable communication with finite delays. Instead of deriving bounds indicating the improvement obtained by increasing the delays we consider the optimization of a system with given delays.

In the basic model both the encoding and the decoding have to be performed causally. This formulation can be motivated by control applications in which the decoder has to control a system in real time.

Our objectives are to examine the usefulness of the channel feedback information and the structures of the optimal encoder and decoder.

The structure of real time encoders for a Markov source and a noiseless channel was discussed by Witsenhausen in [3]. Some properties of the decoders were analyzed in [4], [5].

The paper is organized as follows. In section 2 the basic model is introduced. Section 3 discusses two simple examples that will illustrate some features of those problems. In section 4 the separation results are established. Those results are applied to a class of symmetric channels in section 5, where it will be proved that causal encoding is useless for such channels.

2. Optimal causal coding and decoding

The model under investigation is described below. The situation is pictured on figure 1.

Definition 2.1

Let $\{(x_n, u_n, y_n, z_n), n \geq 1\}$ be a stochastic process taking values in a finite product space $X \times U \times Y \times Z$.

For $n \geqslant 1$, let $x^n := (x_1, \ldots, x_n)$ and similarly for u^n and y^n.

Let also $z_0 = 0 = y^0 = y_0$.

The probability law P of the process is assumed to be such that for $n \geq 1$

$$P\{x_{n+1} | x^n, u^n, y^n\} = M_n(x_{n+1} | x_n),$$

$$u_n = c_n(x^n, y^{n-1}),$$

$$P\{y_n | x^{n'}, u^n, y^{n-1}\} = Q_n(y_n | u_n),$$

$$z_n = h_n(z_{n-1}, y_n),$$

where Q_n, M_n are given transition matrices and c_n, h_n are given functions.

The interpretation is that (x_n) is a *Markov source* with transition matrices (M_n), $c = (c_n, n \geqslant 1)$ is a *code* and z_n is the memory contents of the receiver at time n. This model of memory updating is borrowed from [3]. (See also [1], §8.)

Alternatively, one can think of the probability law P as being a functional of the code c for given source (M_n), channel (Q_n) and receiver (h_n). The dependence of P on c is not indicated explicitly to simplify the notation.

Definition 2.2

Let N be some fixed integer.

For a given code c and a given sequence of functions $d = (d_n, n \geq 1)$, called a *decoding rule*, one defines the cost

$$J(c,d) = E \sum_{n=1}^{N} \ell_n(x_n, d_n (z_n)),$$

where the real valued functions $\ell_1, \ldots, \ell_N$ are fixed and E denotes expectation with respect to P. The functions d_n take values in a finite set D.

A code c* is said to be *optimal* if $J(c^*, d^*) \leq J(c,d)$ for some decoding rule d* and for all (c,d). A decoding rule d_0 is optimal for a given code c if $J(c,d_0) \leq J(c,d)$ for all d.

Remark 2.1

Since only finitely many values are involved the expectation is always well defined. Similarly, only finitely many codes and decoding rules exist so that the optimal c* and d_0 always exist.

The cases of perfect receiver memory ($z_n = y^n$) and that of finite receiver memory ($z_n = (y_{(n-m)+1}, \ldots, y_n)$) are covered by the model.

The causality restriction is explicit for the code c and is reflected in the cost structure for the decoding rule d.

One could have generalized the model by considering randomized codes or decoding rules. However it is clear that they could not achieve a lower cost. (See e.g. [6], T. 1.6.)

3. Two Examples

The first example shows that causal coding can be useful; the second one shows the feedback information can help the encoder and that a one-step optimal strategy may not be optimal.

Coding a single bit

Definition 3.1

Let x, u, y be {0, 1} - random variables with (see figure 2)

$$P(x=1) = \xi \varepsilon [0,1],$$

$$P(y|u) = Q(y|u) \text{ with } Q(1|1) = \beta = 1 - Q(0|1),\ Q(0|0) = \alpha = 1 - Q(1|0),$$

$$u = c(x), \text{ for some code } c: \{0,1\} \rightarrow \{0,1\}.$$

The cost to be minimized is

$$J(c,d) = P(x \neq d(u)),$$

where d: $\{0,1\} \rightarrow \{0,1\}$ is the decoding rule.

For this problem it is immediate to verify the following facts.

Facts 3.1

a) For an arbitrarily fixed code c the optimal decoding rule d_0 is given by

$$d_0(j) = \arg\max_{i \varepsilon \{0,1\}} P(x=i|y=j), \text{ for } j \varepsilon \{0,1\},$$

where $\arg\max_{i \varepsilon S} f(i)$ denotes an arbitrary $i^* \varepsilon S$ maximizing $f: S \rightarrow \mathbb{R}$.

b) As a consequence,

$$\min_d J(c,d) = 1 - \sum_{j \varepsilon \{0,1\}} \max_{i \varepsilon \{0,1\}} P(x=i, y=j)$$

$$= 1 - \sum_{j \varepsilon \{0,1\}} \max_{i \varepsilon \{0,1\}} P(x=i) Q(j|c(i)). \qquad (3.1)$$

It is then easy to compare the costs corresponding to the four possible codes c: $\{0,1\} \rightarrow \{0,1\}$. One finds the following conclusions.

Proposition 3.1

a) An optimal code c* is given by

$$u = c^*(x) \equiv \begin{cases} x \text{ if } (\xi,\alpha,\beta)\varepsilon A \\ 1-x \text{ otherwise,} \end{cases}$$

where $(\xi,\alpha,\beta)\varepsilon A$ if

$$\xi \geq \frac{1}{2} \text{ and } |\alpha - \frac{1}{2}| \leq |\beta - \frac{1}{2}|$$

or

$$\xi < \frac{1}{2} \text{ and } |\alpha - \frac{1}{2}| > |\beta - \frac{1}{2}|.$$

b) There exist some $(\xi,\alpha,\beta)\varepsilon A$ for which $J(c^*, d^*) < J(c,d)$ for all d, with $c^*(x) = 1-x$ and $c(x) = x$.

Therefore, coding can be strictly preferable to sending $u = x$. The result of this proposition can be understood as follows. If $|\alpha - \frac{1}{2}| > |\beta - \frac{1}{2}|$, then the channel is less noisy when its input is the symbol 0 than when it is 1 (see figure 2). Thus one should code in such a way that $P(u=0) > P(u=1)$.

Two-step coding

Definition 3.2

Let now (see figure 3) $x_1 = x_2 = x$, u_1, u_2, y_1, y_2 be {0,1} valued and such that $P(x=1) = \xi\varepsilon[0,1]$, $y_1 = c_1(x)$, $P(y_1|u_1) = Q(y_1|u_1)$,

$$u_2 = c_2(x,y_1),\ P(y_2|u_1,\ u_2,\ x,\ y_1) = Q(y_2|u_2),$$

where Q is as in definition 3.1.

Two cost functions will be considered.

$$J_1\ (c_1,\ d_1) = P(x \neq d_1(y_1))$$

$$J_2\ (c_1,\ c_2,\ d_2) = P\ (x \neq d_2(y_1,\ y_2)).$$

The problem is to find the codes and decoding rules minimizing those costs. Define, for codes c_1 and c_2,

$$J_1(c_1) = \min_{d_1} J_1(c_1,\ d_1),\ J_2(c_1,c_2) = \min_{d_2} J_2(c_1,\ c_2,\ d_2).$$

As in the previous example one can check the following fact.

Fact 3.2

$$J_2(c_1, c_2) = 1 - \sum_{j_1, j_2} \max_{i \in \{0,1\}} P(x_2=i, y_1=j_1, y_2=j_2). \tag{3.2}$$

This can be used to compare the following codes.

Definition 3.3

The codes c^i, $i=1,\ldots, 6$ are defined by the values

$u_1^i = c_1^i(x_1)$ and $u_2^i = c_2^i(x_1, x_2, y_1)$ given by

$$(u_1^1, u_2^1) = (x_1, y_1(1-x_2) + (1-y_1)x_2),$$
$$(u_1^2, u_2^2) = (x_1, y_1x_2 + (1-y_1)(1-x_2)),$$
$$(u_1^3, u_2^3) = (x_1, x_2), (u_1^4, u_2^4) = (x_1, 1-x_2),$$
$$(u_1^5, u_2^5) = (1-x_1, x_2), (u_1^6, u_2^6) = (1-x_1, 1-x_2).$$

Substitutions in (3.1), (3.2) then yield the next conclusions.

Proposition 3.2

a) There is some (ξ, α, β) such that

$$J_2(c^1) < J_2(c^i), \; i = 3,\ldots, 6.$$

b) There is some (ξ, α, β) such that

$$J_2(c^6) < J_2(c^i), \; i = 1,\ldots, 4$$
$$J_1(c^3) < J_1(c^5).$$

The first part of this proposition shows that feedback can be strictly useful. (For c^i, $i = 3,\ldots, 6$ are all the codes which do not use feedback.) This fact can be contrasted with the well known fact that feedback cannot increase the capacity of a memoryless channel and is, in that sense, useless in the classical information theoretic formulation. (e.g. [2], p. 520).

The second part shows that $c_1(x_1) = x_1$ can be optimal for estimating x_1 on the basis of y_1 alone, while $c_1(x_1) = 1-x_1$ is better for estimating x_1 on the basis of y_1 and y_2. This shows that one step optimality is not optimal. In the same line of ideas one can construct examples in which

$c_1(x_1) = x_1$ maximizes $\max_{d_1} P(x_1 = d_1(y_1))$ but does not maximize $\max_{d_2} P(x_2 = d_2(y_1))$. Thus, improving the knowledge about the initial state of a Markov chain may not improve the knowledge about subsequent states, a somewhat a priori counter-intuitive fact.

The above negative results motivate the next section which attempts to characterize the usefulness of the available information.

4. Separation results

The model is that of section 2 (figure 1).

Observe that z_{n-1} is available to the encoder at time $n(n \geq 1)$.

Theorem 4.1

There is an optimal code c^* of the form

$$c_n^* (x^n, y^{n-1}) = \gamma_n(x_n, z_{n-1}), \ n \geq 1. \qquad (4.1)$$

Proof: Fix an arbitrary decoding rule d. Then the process $(v_n = (x_n, z_{n-1}), n \geq 1)$ is conditionally Markov given the u_n's, i.e., for $n \geq 1$,

$$P\{v_{n+1} | v^n, u^n\} = P\{v_{n+1} | v_n, u_n\}.$$

This Markov property implies that

$$J(c,d) = E\{ \sum_{n=1}^{N} E[\ell_n(x_n, d_n(x_n)) | z^{n-1}, x^n, u^n]\}$$

$$= E \sum_{n=1}^{N} k_n(x_n, z_{n-1}, u_n) = E \sum_{n=1}^{N} k_n(v_n, u_n).$$

for some functions k_n.

Considering the resulting Markovian decision problem of controlling the transition probabilities of v_n with complete observation to minimize an additive cost in (v_n, u_n) then yields the result. $\square$

This result was given in [3] in the case where the channel is noiseless, i.e., when the entries of the Q_n are in $\{0,1\}$.

The case of perfect receiver memory $(z_n = y^n)$ leads to a sharpening of that result. It is considered next.

Definition 4.1

Let $M(X)$ denote the set of probability measures on X.

For $\xi = (\xi(x), x \in X) \in M(x)$ and $n \geq 1$ define

$$\hat{\alpha}_n(\xi) = \arg \min_{\alpha \in D} \sum_{x \in X} \ell_n(x, \alpha) \xi(x).$$

For any $x \in X$ and $n \geq 1$ let

$$L_n(x, \xi) = \ell_n(x, \alpha_n(\xi)).$$

For every code c define the conditional probability of x_n given y^n as

$$\xi_n^c(y^n)(x) = P\{x_n = x | y^n\}, \; y^n \varepsilon Y^n, \; x \varepsilon X.$$

Let also $L_0 = \ell_0 = 0$.

With those definitions one can state the following immediate fact.

Lemma 4.1

Let c be any given code. The optimal decoding rule d for c is given by

$$d_n(y^n) = \hat{\alpha}_n(\xi_n^c(y^n)), \; n \geq 1.$$

Theorem 4.2

Assume that $z_n = y^n$, $n \geq 1$. Then there is an optimal code c* of the form

$$c_n^*(x^n, y^{n-1}) = \psi_n^*(x_n, \xi_{n-1}^*(y^{n-1})), \; n \geq 1,$$

for some functions ψ_n^*, where $\xi_n^* = \xi_n^{c^*}$.

Proof:

The main idea of the proof is to consider that the *receiver* chooses the function $\gamma_n(\cdot, y^{n-1})$ (see (4.1)) to be used next by the encoder. The receiver is then faced with a control problem with partial observation to which one can apply the dynamic programming techniques.

We now proceed with a formal proof which will be given as a succession of lemmas.

For $\xi \varepsilon M(X)$, $x \varepsilon X$, $y \varepsilon Y$, $n \geq 1$ and $w: X \to U$ let

$$F_n(\xi, y, w)(x) = \frac{Q_n(y|w(x))\Sigma_{x'} M_{n-1}(x|x')\xi(x')}{\Sigma_{\tilde{x}} Q_n(y|w(\tilde{x}))\Sigma_{x'} M_{n-1}(\tilde{x}|x')\xi(x')} \qquad (4.2)$$

where the sums extend over $x' \varepsilon X$ and $\tilde{x} \varepsilon X$, and M_0 is the identity matrix. Observe that by Theorem 4.1 one can restrict attention to codes c of the form

$$c_n(x^n, y^{n-1}) = \gamma_n(x_n, y^{n-1}), \; n \geq 1. \qquad (4.3)$$

Lemma 4.2

Fix an arbitrary code c of the form (4.1). The rule for updating conditional probabilities is

$$\xi_n(y^n) = F_n(\xi_{n-1}(y^{n-1}), y_n, \gamma_n(\cdot, y^{n-1})), \ n \geqslant 1,$$

where $\xi_0(y^0)(x) := P(x_1 = x)$, $x \in X$.

Proof:

This is a direct consequence of definition 2.1, (4.2) and Bayes' rule. □

For $\xi \in M(X)$ define recursively for $n = N+1, N, N-1, \ldots, 1$

$$V_{N+1}(\xi) = 0$$
$$V_{n-1}(\xi) = \min_{w:X\to U} \Sigma_x \xi(x)\{L_{n-1}(x,\xi) + \Sigma_y V_n(F_n(\xi,y,w)) \Sigma_{\tilde{x}} Q_n(y|w(\tilde{x})) M_{n-1}(\tilde{x}|x)\} \tag{4.4}$$

and denote the minimizer w in (4.4) by $\psi_n^*(\cdot,\xi)$ and the corresponding ξ_n^c by ξ_n^*.

Lemma 4.3

For any code c of the form (4.3) and any decoding rule d one has

$$V_n(\xi_n^c) \leq E^c[\sum_{m=n}^{N} \ell_m(x_m, d_m(y^m))|y^n] \quad , \ n = 0,\ldots, N+1, \tag{4.5}$$

where E^c is the expectation with respect to the law induced by c. (In (4.5) we define $\sum_{m=N+1}^{N} \ell_m = 0$.)

Proof:

Assume that (4.5) holds for some $n \leq N+1$ (it trivially does for n=N+1). We show that it then holds for n-1.

Choosing $w(\cdot) = \gamma_n(\cdot, y^{n-1})$ and $\xi = \xi_{n-1}^c$ shows that (4.4) implies

$$V_{n-1}(\xi_{n-1}^c) \leq \Sigma_x \xi_{n-1}^c(x)\{L_{n-1}(x, \xi_{n-1}^c)$$
$$+ \Sigma_y V_n(\xi_n(y^{n-1}, y)) \Sigma_{x'} Q_n(y|\gamma_n(x', y^n)) M_{n-1}(x'|x)\},$$

where we used (4.4). This is

$V_{n-1}(\xi^c_{n-1}) \le \Sigma_x\xi^c_{n-1}(x)[L_{n-1}(x,\xi^c_{n-1}) + \Sigma_y E^c[V_n(\xi^c_n)|y^{n-1},y_n=y]P\{y_n=y|x_{n-1}=x\}],$

i.e.,

$$V_{n-1}(\xi^c_{n-1}) \le \Sigma_x\xi^c_{n-1}(x)\; L_{n-1}(x,\xi^c_{n-1}) + E^c[V_n(\xi^c_n)|y^{n-1}]. \qquad (4.6)$$

Now, using definition 4.1 we find that for all $\alpha\varepsilon D$

$$\Sigma_x\xi^c_{n-1}(x)\; L_{n-1}(x,\xi^c_{n-1}) \le \Sigma_x\ell_{n-1}(x,\alpha)\xi^c_{n-1}(x).$$

In particular,

$$\Sigma_x\xi^c_{n-1}(x)\; L_{n-1}(x,\xi^c_{n-1}) \le \Sigma_x\ell_{n-1}(x,d_{n-1}(y^{n-1}))\xi^c_{n-1}(x), \text{ i.e.,}$$

$$\Sigma_x\xi^c_{n-1}(x)\; L_{n-1}(x,\xi^c_{n-1}) \le E^c[\ell_{n-1}(x_{n-1},\, d_{n-1}(y^{n-1}))|y^{n-1}].$$

Introducing this inequality in (4.6) gives

$$V_{n-1}(\xi^c_{n-1}) \le E^c[\ell_{n-1}(x_{n-1},\, d_{n-1}(y^{n-1})) + V_n(\xi^\gamma_n)|y^{n-1}].$$

Substituting this inequality into the induction hypothesis proves that (4.3) must indeed hold with n replaced by n-1. □

Similar calculations with inequalities replaced by equalities show that the definition of ψ^* and ξ^* yields the following

<u>Lemma 4.4</u>

$$V_n(\xi^*_n) = E^*[\sum_{m=n}^{N} L_m(x_m,\xi^*_m)|y^n],\; n = 0,\dots,\, N + 1, \qquad (4.7)$$

where E* is the expectation with respect to the law induced by the code $c^*_n(x^n,\, y^{n-1}) = \psi^*_n(x_n,\, \xi^*_{n-1}(y^{n-1}))$.

We now conclude the proof of Theorem 4.2.

Writing (4.5) and (4.7) for n = 0 gives for an arbitrary code c of the form (4.3) and an arbitrary decoding rule d

$$V_0(\xi^*_0) = E^*[\sum_{n=1}^{N} L_n(x_n,\xi^*_n)] = E^*[\sum_{n=1}^{N} \ell_n(x_n,\hat{\alpha}_n(\xi^*_n))],$$

$$V_0(\xi^c_0) \le E^c[\sum_{n=1}^{N} \ell_n(s_n,\, d_n(y^n))].$$

But $\xi_0^* = \xi_0^c$, the prior distribution of x_1.

Therefore

$$J(c^*, d^*) \leq J(c, d),$$

where

$$c_n^*(x^n, y^{n-1}) = \psi_n^*(x_n, \xi_{n-1}^*(y^{n-1}))$$

and

$$d_n^*(y^n) = \hat{\alpha}_n(\xi_n^* \,(y^{n-1})) \;,$$

which completes the proof.

Remark 4.1

If the model of section 2 is modified so that $(x_n, n \geq 1)$ is no longer Markov but $(\beta_n = (x_n, \ldots, x_{n-m}), n \geq 1)$ is Markov, then similar results can be established. In the case where $z_n = y^n$, one finds that there is an optimal code c^* of the form

$$c_n^*(x^n, y^{n-1}) = \psi_n^*(x_n, \xi_{n-1}^*)$$

where ξ_{n-1}^* is now the conditional law of β_{n-1} given y^{n-1}.

A separation result such as Theorem 4.2 indicates a recursive way of updating the information ξ_n^* sufficient for the encoder. Indeed, using Lemma 4.2 shows that

$$\xi_n^*(y^n) = F_n(\xi_{n-1}^*(y^{n-1}), y_n, \psi_n^*(\cdot, \xi_{n-1}^*(y^{n-1})), n \geq 1.$$

5. Symmetric Channels

Many channel models used in communication theory possess the symmetry property defined below. It will be shown that causal encoding is useless for such channels. Again this situation should be compared with the classical information theoretic formulation.

Definition 5.1

Let U,Y be two finite sets. A transition matrix Q from U to Y is said to be of type S if it has the following property:

For every f: U → U there is a transition matrix A from Y to Y such that

$$Q(y|f(u)) = \Sigma_{y'} A(y|y') \, Q(y'|u) \ , \ u \in U, \ y \in Y \qquad (5.1)$$

A memoryless channel will be called symmetric if its transition matrices Q_n are of type S for all $n \geqslant 1$.

Examples 5.1

The following transition matrices are easily verified to be of type S. In these examples $Q_{ij}: = Q(y = j|u = i)$

a) $\begin{bmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{bmatrix}$

b) $\begin{bmatrix} 1-\varepsilon_1-\varepsilon_2 & \varepsilon_2 & \varepsilon_1 \\ \varepsilon_2 & 1-\varepsilon_1-\varepsilon_2 & \varepsilon_1 \end{bmatrix}$

c) $\begin{bmatrix} a_1 & a_2 & \cdots & a_n \\ a_n & a_{n-1} & \cdots & a_1 \end{bmatrix}$

d) $\begin{bmatrix} a_1 & a_2 & a_3 & a_4 & a_5 \\ a_2 & a_1 & a_4 & a_3 & a_5 \end{bmatrix}$

e) Q with $Q(y|u) \equiv |Y|^{-1}$, when $|Y|$ is the cardinality of Y.

It is also clear that, with compatible dimensions, if Q_1 and Q_2 are of type S, then so are $\lambda Q_1 + (1 - \lambda) Q_2$ and $[\lambda Q_1 | (1 - \lambda) Q_2]$ for $\lambda \in [0,1]$ and $Q_1 Q_2$. If Q is noiseless and one-to-one, then it is of type S.

The channel corresponding to examples a) and b) are respectively called the binary symmetric channel and the binary symmetric erasure channel. Consider once again the coding problem of Section 2.

Theorem 5.1

Assume that the channel is symmetric, that $X = U$ and $z_n = y^n$, $n \geqslant 1$. Then

$$c_n^*(x^n, y^{n-1}) = x_n, \; n \geqslant 1$$

is optimal.

Proof

By Theorem 4.1 one can restrict attention to codes of the form

$$c_n(x^n, y^{n-1}) = \gamma_n(x_n, y^{n-1}), \; n \geqslant 1.$$

Assume that there is some $n_0 \leqslant N$ such that

$$\gamma_n(x_n, y^{n-1}) \equiv x_n \text{ for } n > n_0. \tag{5.2}$$

Define the code $\tilde{c}$ by

$$\tilde{c}_n(x^n, y^{n-1}) = \begin{cases} c_n(x^n, y^{n-1}) & , \; n \neq n_0 \\ x_n & , \; n = n_0. \end{cases}$$

Fix an arbitrary decoding rule d. We will show that there exists a decoding rule $\tilde{d}$, possible randomized, such that

$$J(\tilde{c}, \tilde{d}) = J(c, d). \tag{5.3}$$

($J(\tilde{c}, \tilde{d})$ is defined in the obvious way for a randomized $\tilde{d}$.) Since randomizing the decoder cannot possible reduce the optimal cost, this will show that one can assume that (5.2) holds with n_0 replaced by n_0 -1. By induction, this will prove the theorem, since n_0 = N trivially satisfies (5.2).

Letting $f(u) = \gamma_{n_0}(u, y^{n_0-1})$ in (5.1) shows that there exist matrices $A(y \mid y' ; y^{n_0-1})$ such that, for all (x_{n_0}, y^{n_0}),

$$Q_{n_0}(y_{n_0} | \gamma_{n_0}(x_{n_0}, y^{n_0-1})) = \Sigma_{y'} A_{n_0}(y_{n_0} | y';y^{n_0-1}) \, Q_{n_0}(y' | x_{n_0}) \qquad (5.4)$$

One then defines the randomized rule $\tilde{d}$ as follows: Let

$$\tilde{d}_n(y^n) = d_n(\tilde{y}^n), \; n = 1,\dots,N,$$

where, for $y \in Y$,

$$\tilde{y}^N = (y^{n_0-1}, y, y_{n_0+1}, \dots, y_N)$$

with probability

$$A_{n_0}(y | y_{n_0} ; y^{n_0-1})$$

Observe that this rule is causal (the die is chosen and tossed at time n_0 and its outcome is used only for $n \geq n_o$).

We claim that the law of (x^N, y^N) under c is the same as that of $(x^N, \tilde{y}^N)$ under $\tilde{c}$. This will then imply that the law of $(x^N, d_1(y), \dots, d_N(y^N))$ under c is the same as that of $(x^N, \tilde{d}_1(\tilde{y}_1), \dots, \tilde{d}_N(\tilde{y}^N))$ under $\tilde{c}$, thereby proving (5.3). To establish the claim we notice that (P^c indicates the law under c)

$$P^c(x^N, y^N) = P(x^N) \prod_{n=1}^{N} Q_m(y_m | c_n(x^n, y^{n-1}))$$

$$= P(x^N) \prod_{n=1}^{n_0-1} Q_n(y_n | c_n(x^n, y^{n-1}) \times Q_{n_0}(y_{n_0} | \gamma_{n_0}(x_{n_0}, y^{n_0-1}))$$

$$\times \prod_{m=n_o+1}^{N} Q_m(y_m | x_m)$$

Similarly,

$$P^{\tilde{c}}(x^N, \tilde{y}^N) = P(x^N) \prod_{n=1}^{n_0-1} Q_n(y_n|c_n(x_n, y^{n-1})) \times Q_{n_0}(\tilde{y}_{n_0}|x_{n_0})$$

$$\times \prod_{m=n_0+1}^{N} Q_m(y_m|x_m).$$

But, by definition of $\tilde{y}_{n_0}$,

$$Q_{n_0}(\tilde{y}_{n_0}|x_{n_0}) = \Sigma_{y'} Q_{n_0}(y'|x_{n_0}) A_{n_0}(\tilde{y}_{n_0}|y'; y^{n_\theta-1})$$

$$= Q_{n_0}(\tilde{y}_{n_o}|\gamma_{n_0}(x_{n_0} y^{n_o-1}))$$

which concludes the proof.

□

Remark 5.1

In the case of the binary symmetric channel the above argument shows that all the codes $c_n(x^n, y^{n-1}) = \gamma_n(x_n, y^{n-1})$ with $\gamma_n(., y^{n-1})$ one-to-one are equivalent, in the sense that they all achieve the minimum cost. (The decoders must be chosen accordingly). For other work on comparing experiments see [7].

6. Conclusions

Separation results have been obtained for causal coding problems with channel feedback information. These results were used to show that causal coding is useless for symmetric channels.

We hope to extend those results to control systems in a subsequent paper.

7. Acknowledgement

This research was performed while the first author was visiting the Department of Electrical Engineering and Computer Science, at the University of California, Berkeley. His research was supported in part by the Department of Energy under Grant No. DE-AC01-80RA50419. The second author's research was supported in part by Office of Naval Research under Contract N000014-80-C-0507.

The authors wish to thank Professor T. Berger and Dr. G. Munson for suggesting this problem and Professor F.P. Kelly for stimulating discussions. Dr. G. Munson conjectured the result for the symmetric case and analyzed the problem from an information theoretic viewpoint [8].

References

[1] C. E. Shannon, "A mathematical theory of communication," Bell Syst. Tech J., 22, 379-423, 1948.

[2] R. G. Gallager: Information theory and reliable communication, J. Wiley, 1968.

[3] H. S. Witsenhausen, "On the structure of real-time source coders," Bell Syst. Tech J., 58, 1437-1451, 1979.

[4] A. W. Drake, "Observation of a Markov process through a noisy channel," Sc. D. Thesis, Department of Electrical Engineering, M.I.T., Cambridge, MA, 1962.

[5] J. L. Devore, "A note on the observation of a Markov source through a a noisy channel," IEEE Transactions in Information Theory, Vol. IT-20, 762-764, 1974.

[6] I. I. Gihman and A. V. Skorohod: Controlled Stochastic Processes, Springer Verlag, 1979.

[7] J. Marschak and K. Miyasawa, "Economic Comparability of Information Systems," Int. Economic Review, 9, No. 2, 137-174, 1968.

[8] G. Munson, "Causal information transmission with feedback," Ph.D. Thesis, School of Electrical Engineering, Cornell University, 1981.

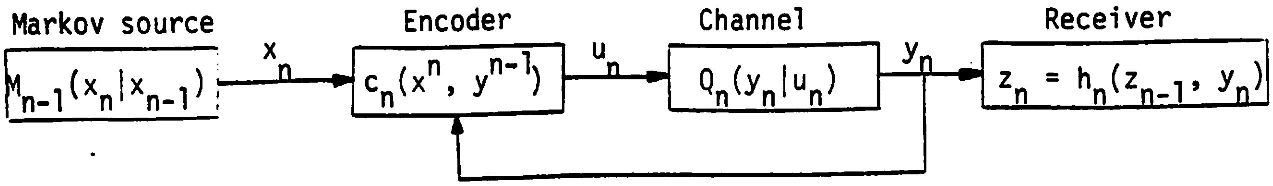

Figure 1: Causal communication system

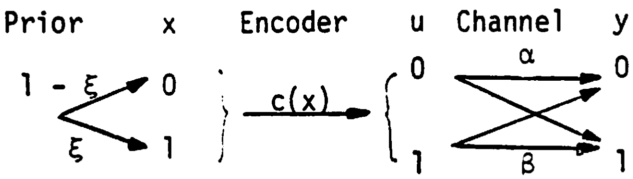

Figure 2: Coding a single bit

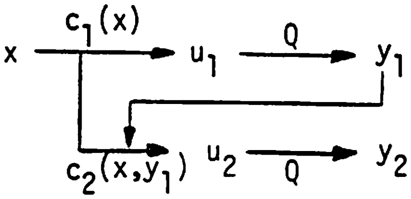

Figure 3: Two-step coding