Journal of the Faculty of Engineering, Shinshu University, No. 37, 1974
信州大学工学部紀要　第37号

On Markov Channels

Yatsuka NAKAMURA*
(Received October 29, 1974)

1. Introduction

The information theory originated by C. E. Shannon was extended to the case of finite memory channels by Hinchin[3]. Moreover the case of infinite memory channels was also studied by Adler[1] and others. However, among these generalized channels, there are not many which are useful for concrete models. In this paper, we define a Markov channel which is a natural generalization of the Markov chain and useful to make examples of generalized channels. We show that it is determined by a pair of vector valued mappings, and give two representative examples of Markov channels and simulate how errors occur in these channels. We also show that a stationary stochastic state machine is a Markov channel and we can compute the capacity of this channel.

2. Notations

The *input space* and the *output space* are represented by the alphabet spaces $X=A^I$ and $Y=B^I$ respectively, where $A=\{a_1, a_2, \cdots\cdots, a_k\}$ and $B=\{b_1, b_2, \cdots\cdots, b_l\}$ are sets of finite letters (or states), and $I=\{0, \pm 1, \pm 2, \cdots\}$ is a set of all integers which represents the time, i. e., X and Y are the infinite dimensional product spaces of the countable copies of finite measurable spaces A and B respectively. Let $\mathscr{X}$ and $\mathscr{Y}$ be Borel fields generated by all cylinder sets of X and Y respectively. For the element x in X, x_n represents the n' th letter in A, and for y in Y, y_n represents the n' th letter in B. In the followings, we omit the explanation for y as it is same as x. The notation $[x_s^0, x_{s+1}^0, \cdots, x_t^0]$ represents a thin cylinder on the time interval $(s, s+1, \cdots, t)$, i. e.,

$$[x_s^0, x_{s+1}^0, \cdots, x_t^0]=\{x \in X\colon\ x_n = x_n^0, \quad n = s, s+1, \cdots, t\},$$

where $s, t \in I$ $(s<t)$. A σ-subfield of $\mathscr{X}$ generated by all thin cylinders of the form $[x_s, x_{s+1}, \cdots, x_t]$ on the time interval $(s, s+1, \cdots, t)$ is denoted by $\otimes_{i=s}^{t} \mathscr{X}_i$, and a σ-subfield generated by all thin cylinders of the form $[x_s]$ is denoted by $\mathscr{X}_s$ and is called the *time-s partition*, as it is also seen as a finite partition of X. Let 2_X be a *trivial partition* of X, i. e., $2_X = \{X\}$. A *shift transformation* S

---

* Assistant Professor, Department of Information Engineering

(or T) on X (or Y) represents time advance and is defined by the formula $(Sx)_n = x_{n+1}$ $((Ty)_n = y_{n+1})$. The shift transformation S (or T) is a measurable transformation on the measurable space X (or Y). The *input source* is an S-invarariant probability measure p on the measurable space X, i. e., which satisfies $p(S^{-1}E) = p(E)$ for all $E \in \mathscr{X}$. The *output source* is defined similarly. We denote a set of all input sources by Π.

A *channel* from an input space X to an output space Y is a two variable numerical function $\nu_x(F)$ of $x \in X$ and $F \in \mathscr{Y}$ and satisfies the followings:

I. If we fix $x \in X$, then $\nu_x(\cdot)$ is a probability measure on $(Y, \mathscr{Y})$,

II. if we fix $F \in \mathscr{Y}$, then $\nu_.(F)$ is a measurable function on X,

and III. for all $x \in X$ and $F \in \mathscr{Y}$, $\nu_{Sx}(F) = \nu_x(S^{-1}F)$.

An input source $p \in \Pi$ and a channel ν determine an output source $q(\cdot)$;

$$q(F) = \int_X \nu_x(F) p(dx) \qquad (F \in \mathscr{Y}).$$

Let $(X \otimes Y, \mathscr{X} \otimes \mathscr{Y})$ be the direct product measurable space of $(X, \mathscr{X})$ and $(Y, \mathscr{Y})$, then we call it a *compound space*, which is also an alphabet space $(A \otimes B)^I$. A shift transformation $S \otimes T$ on the compound space is defined by $(S \otimes T)(x, y) = (Sx, Ty)$. A *compound source* is an $S \otimes T$-invariant probability measure on the compound space. An input source p and a channel ν also determine a compound source r;

$$r(G) = \int_X \nu_x(G_x) p(dx) \qquad (G \in \mathscr{X} \otimes \mathscr{Y}),$$

where G_x is an x-section of the set G. The sources $q(\cdot)$ and $r(\cdot)$ determined by an input source p and a channel ν are sometimes denoted by $q(\cdot; p, \nu)$ and $r(\cdot; p, \nu)$, respectively. We sometimes abbreviate $\nu_x([y_s, y_{s+1}, \cdots, y_t])$ to $\nu_x(y_s y_{s+1} \cdots y_t)$.

If there exists an integer m such that $f(\cdot) = \nu_.(y_1 y_2 \cdots y_n)$ is a $\otimes_{i=1}^{n} \mathscr{X}_i$-measurable function for all positive integer n and for all letters $y_1, y_2, \cdots, y_n$, and if $f(\cdot)$ is constant for n smaller than $1-m$, then the channel ν is called a *finite memory channel* or an *m-memory channel.*

For any finite partitions $\mathscr{A}_1$ and $\mathscr{A}_2$ of X, we write

$$\mathscr{A}_1 \vee \mathscr{A}_2 = \{A_1 \cap A_2 : A_1 \in \mathscr{A}_1, \ A_2 \in \mathscr{A}_2\},$$

which is a new partition of X. Then the *entropy* $h_p(S)$ of the transformation S *relative to an input source* $p \in \Pi$ is defined by

$$h_p(S) = \lim_n \frac{1}{n} H(\mathscr{X}_0 \vee S^{-1} \mathscr{X}_0 \vee \cdots \vee S^{-n+1} \mathscr{X}_0),$$

where $\mathscr{X}_0$ is the time-0 partition, and $H(\mathscr{A})$ is the sum $-\sum_{A \in \mathscr{A}} p(A) \log p(A)$.

If q and r an output source and a compound source respectively, then we can also define the entropy $h_q(T)$ and $h_r(S\otimes T)$ of the transformation T relative to an output source q and of the transformation $S\otimes T$ relative to a compound source r. Then the *transmission rate* $R_p(\nu)$ of a channel ν relative to an input source p is given by

$$R_p(\nu) = h_p(S) + h_q(T) - h_r(S\otimes T)$$

where both $q(\cdot) = q(\cdot; p, \nu)$ and $r(\cdot) = r(\cdot; p, \nu)$ are determined by p and ν. The (*stationary*) *capacity* $C_s(\nu)$ of a channel ν is defined by

$$C_s(\nu) = \sup_{p\in \Pi} R_p(\nu).$$

3. Markov Channels

We begin with the definiton of a Markov channel.

Definition. A channel ν from the input space X to the output space Y is called an *m-fold Markov channel* if there exists a positive number m such that

IV. $$\nu_x(y_n \mid y_0\cdots y_{n-1}) = \nu_x(y_n \mid y_{n-m}\cdots y_{n-1})$$

for all $x\in X$, for all $y_1, \cdots, y_n \in B$ and for all positive integer n ($n\geq m$), where

$$\nu_x(y_s\cdots y_{s+l} \mid y_t\cdots y_{s-1}) = \begin{cases} \nu_x(y_t\cdots y_{s+l}) / \nu_x(y_t\cdots y_{s-1}) & \cdots \text{ if } \nu_x(y_t\cdots y_{s-1}) \neq 0, \\ 0 & \cdots \text{ if } \nu_x(y_t\cdots y_{s-1}) = 0. \end{cases}$$

The errors which occur in the m-fold Markov channel depend on the only m steps in the past.

Next let us see that the Markov channel is characterized by a pair of measurable mappings. Now for every element x in X, we consider a stochastic vector on B^m

$$Q(x) = (q_{\bar{b}}(x)) \qquad (\bar{b} \in B^m)$$

and we also consider a stochastic matrix

$$\widetilde{Q}(x) = (\tilde{q}_{b\mid\bar{b}}(x)),$$

whose row is indexed by b in B and whose column is indexed by $\bar{b}$ in B^m. Moreover we assume that $q_{\bar{b}}(x)$ and $\tilde{q}_{b\mid\bar{b}}(x)$ are measurable functions of x on $(X, \mathscr{X})$, and assume that

$$q_{y_{n-m+1}\cdots y_n}(Sx) = \sum_{y_{n-m}\in B} \tilde{q}_{y_n\mid y_{n-m}\cdots y_{n-1}}(x) q_{y_{n-m}\cdots y_{n-1}}(x). \tag{1}$$

The pair $(Q(x), \widetilde{Q}(x))$ is called a *stochastic pair*.

Theorem 1. *Let $(Q(x), \widetilde{Q}(x))$ be a stochastic pair. Then it defines an m-fold Markov channel. Conversely every m-fold Markov channel defines a stochastic pair.*

Proof. For a stochastic pair $(Q(x), \widetilde{Q}(x))$, we can construct an m-fold Markov channel by the following proceedure:
We put

$$\nu_x(y_1\cdots y_n) = \begin{cases} q_{y_1\cdots y_m}(x)\prod_{i=1}^{n-m} \tilde{q}_{y_{m+i}\mid y_i\cdots y_{m+i-1}}(S^{i-1}x) & \cdots \text{ if } n-m>0, \\ \sum_{y_n\cdots y_m} q_{y_1\cdots y_m}(x) & \cdots \text{ if } n-m\leq 0. \end{cases}$$

Assumming $n-m>0$, we see

$$\sum_{y_n\in B}\nu_x(y_1\cdots y_n) = \sum_{y_n\in B} q_{y_1\cdots y_m}(x)\prod_{i=1}^{n-m}\tilde{q}_{y_{m+i}\mid y_i\cdots y_{m+i-1}}(x)$$

$$= q_{y_1\cdots y_m}(x)\prod_{i=1}^{n-m-1}\tilde{q}_{y_{m+i}\mid y_i\cdots y_{m+i-1}}(S^{i-1}x)$$

$$= \nu_x(y_1\cdots y_{n-1}).$$

As a similar equality also holds in case of $n-m\leq 0$, we conclude that $\nu_x(\cdot)$ can be extended to a probability measure on $\otimes_{i=1}^{\infty}\mathscr{Y}_i$ by the Kolmogorov extension theorem. Next, for any thin cylinder $[y_l, \cdots, y_n]$ in Y $(n\geq l\geq 1,\ n>m)$,

$$\nu_x(y_l\cdots y_n) = \sum_{y_1\cdots y_{l-1}}\nu_x(y_1\cdots y_m)$$

$$= \sum_{y_1\cdots y_{l-1}} q_{y_1\cdots y_m}(x)\prod_{i=1}^{n-m}\tilde{q}_{y_{m+i}\mid y_i\cdots y_{m+i-1}}(S^{i-1}x)$$

$$= \sum_{y_2\cdots y_{l-1}} q_{y_2\cdots y_{m+1}}(Sx)\prod_{i=2}^{n-m}\tilde{q}_{y_{m+i}\mid y_i\cdots y_{m+i-1}}(S^{i-1}x)$$

$$= \cdots\cdots\cdots$$

$$= \begin{cases} q_{y_l\cdots y_{m+l-1}}(S^l x)\prod_{i=l}^{n-m}\tilde{q}_{y_{m+i}\mid y_i\cdots y_{m+i-1}}(S^{i-1}x) & \cdots\text{if } n-m\geq l, \\ \sum_{y_{n-m}\cdots y_{l-1}} q_{y_{n-m}\cdots y_{n-1}}(S^{n-m}x) & \cdots\text{if } n-m<l. \end{cases}$$

In both cases of $n>m$ and $n\leq m$, we get

$$\nu_x(y_l\cdots y_n) = \nu_{S^l x}(T^l[y_l\cdots y_n]). \tag{2}$$

Replacing l in (2) by $l-1$, we get

$$\nu_{Sx}(T[y_l\cdots y_n]) = \nu_{S^{l-1}(Sx)}(T^{l-1}T[y_l, \cdots, y_n])$$

$$= \nu_{S^l x}(T^l[y_l, \cdots, y_n])$$

as $T[y_l, \cdots, y_n]$ is a thin cylinder on the time interval $(l-1, \cdots, n-1)$. Hence

$$\nu_x(y_l \cdots y_n) = \nu_{Sx}(T[y_l \cdots y_n])$$

holds. We extend $\nu_x(\cdot)$ from $\otimes_{i=1}^{\infty} \mathscr{Y}_i$ to $\mathscr{Y} = \otimes_{i=-\infty}^{\infty} \mathscr{Y}_i$, by the formula

$$\nu_x(y_{-l} \cdots y_n) = \nu_{S^{-l}x}(T^{-l}[y_{-l}, \cdots, y_n]).$$

Then ν satisfies the condition III of the channel. For any $[y_l, \cdots, y_n]$, $\nu_{\cdot}(y_l \cdots y_n)$ is $\mathscr{X}$-measurable on X, as this is defined by the product of the measurable functions $q_{\bar{b}}(x)$ and $\tilde{q}_{b|\bar{b}}(x)$. And for any measurable set F $(\in \mathscr{Y})$, $\nu_{\cdot}(F)$ is also measurable, as F can be approximated by a countable sequence of cylinder sets with the measure $\nu_x(\cdot)$. Therefore the condition II is also clear.

Conversely let ν be an m-fold Markov channel. We define a stochastic pair $(Q(x), \widetilde{Q}(x))$ by the following;

$$q_{y_1}(x) = \nu_x(y_1)$$

and

$$\tilde{q}_{y_{m+1}|\bar{b}}(x) = \nu_x(y_{m+1}|\bar{b})$$

where $\bar{b} = [y_1, y_2, \cdots y_m]$. Clearly the pair is well defined by the condition III of the channel.

$$\sum_{y_1 \in B} \tilde{q}_{y_{m+1}|y_1 \cdots y_m}(x)\, q_{y_1 \cdots y_m}(x) = \sum_{y_1 \in B} \nu_x(y_1 \cdots y_m)\ \nu_x(y_{m+1}|y_1 \cdots y_m)$$

$$= \sum_{y_1 \in B} \nu_x(y_1 \cdots y_{m+1}) = \nu_x(y_2 \cdots y_{m+1}) = \nu_{Sx}(T(y_2 \cdots y_{m+1}))$$

$$= q_{y_2 \cdots y_{m+1}}(Sx).$$

The above chain of formulae shows that the pair satisfies the condition of the stochastic pair. Q. E. D.

This theorem tells us that any Markov channel is determined by a suit of mappings, which resembles to the fact that a memoryless channel (cf. 2) p. 154) is determined by a stochastic matrix. Let (c_{ab}) be a stochastic matrix which determines a memoryless channel, then taking $q_b(x) = c_{x_1 b}$ and $\tilde{q}_{b|\bar{b}}(x) = c_{x_2 b}$, we see that the memoryless channel is a kind of Markov channels.

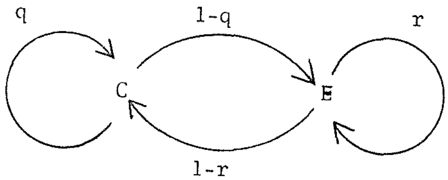
4. Examples

We put $A = B = \{0, 1\}$, i. e., we consider a binary case. For a stochastic pair, let us choose

$$Q(x) = \begin{cases} \begin{pmatrix} p \\ 1-p \end{pmatrix} = Q(0) & \cdots \text{ if } x_0 = 0, \\ \begin{pmatrix} 1-p \\ p \end{pmatrix} = Q(1) & \cdots \text{ if } x_0 = 1, \end{cases}$$

$$\widetilde{Q}(x) = \begin{cases} \begin{pmatrix} q & 1-r \\ 1-q & r \end{pmatrix} = \widetilde{Q}(00) & \cdots \text{ if } x_0 = 0,\ x_1 = 0, \\ \begin{pmatrix} 1-q & r \\ q & 1-r \end{pmatrix} = \widetilde{Q}(01) & \cdots \text{ if } x_0 = 0,\ x_1 = 1, \\ \begin{pmatrix} 1-r & q \\ r & 1-q \end{pmatrix} = \widetilde{Q}(10) & \cdots \text{ if } x_0 = 1,\ x_1 = 0, \\ \begin{pmatrix} r & 1-q \\ 1-r & q \end{pmatrix} = \widetilde{Q}(11) & \cdots \text{ if } x_0 = 1,\ x_1 = 1, \end{cases}$$

and assume that $\widetilde{Q}(00)Q(0) = Q(0)$ i. e., $p = (r-1)/(r+q-2)$. Then we find that $\widetilde{Q}(11)Q(1) = Q(1)$, $\widetilde{Q}(01)Q(0) = Q(1)$ and $\widetilde{Q}(10)Q(1) = Q(0)$. Hence $(Q(x), \widetilde{Q}(x))$ is a stochastic pair, so that we can define a Markov channel, which we call an additive noise channel. Errors in this channel occur as shown in Fig. 1. Fig. 2 is a result of a simple simulation in the case of $q = 9/10$, $r = 1/2$ and $p=5/7$.



C: correct

E: erroneous

Fig. 1

input sequence 1011001011001011100110000101110100101

$\downarrow$

output sequence $\underline{01}$110010110010111$\underline{110}$1$\underline{11110}$01110100101

Fig. 2

Next we give another example;

$$Q(x) = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix} = Q \qquad \text{for any } x \in X,$$

$$Q(x) = \begin{cases} \begin{pmatrix} p & 1-p \\ 1-p & p \end{pmatrix} = Q(0) & \text{if } x_1 = 0, \\ \begin{pmatrix} q & 1-q \\ 1-q & q \end{pmatrix} = Q(1) & \text{if } x_1 = 1. \end{cases}$$

Then we can also define a Markov channel as shown in Fig. 3, which we call a *stationary stochastic state machine channel* (*SSSM channel*). Fig. 4 is a result of simulation when p = 3/4 and q = 1/4. In Fig. 4, the output letters seem to have little correlation with the input letters, but we can show later that the capacity of this channel is not zero, therefore, the transmission of the information is possible through this noisy channel by the Shannon-Hinchin theorem.

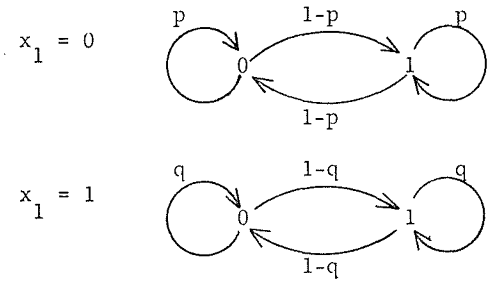


Fig. 3

input sequence 1011001011001011100110000101110100101

↓

output sequence 1101110010110010001011000010110111001

Fig. 4

5. Capacity of SSSM Channels

The second example of the previous section can be extended to a more general case, i. e., the letters need not be binary. Let $Q(x)$ be independent of $x \in X$. And we assume that $\widetilde{Q}(x)$ is only dependent on x, hence we can write $\widetilde{Q}(x) = \widetilde{Q}(x_1)$, and assume that $\widetilde{Q}(x_1)Q = Q$ for all $x \in$ A, where $Q = Q(x)$. We call this generalized one an *SSSM channel* also. We can compute the capacity

$C_s(\nu)$ of an SSSM channel ν. First, we give a proposition in a more general situation.

Proposition 1. *Let ν be a finite memory Markov channel, then*

$$R_p = h_q(T) + \int_X \sum_{y_1\cdots y_{m+1}} q_{y_1\cdots y_m}(x)\ \eta(\tilde{q}_{y_{m+1}|y_1\cdots y_m}(x))p(dx)$$

where $\eta(t) = -t\log t$.

Proof. By the definition of R_p,

$$R_p = h_p(S) + h_q(T) - h_r(S\otimes T)$$

$$= h_q(T) + h_p(S) - \lim_n \frac{1}{n} H(\mathscr{X}_0 \otimes \mathscr{Y}_0 \vee S^{-1}\mathscr{X}_0 \otimes T^{-1}\mathscr{Y}_0 \vee \cdots \vee S^{-n+1}\mathscr{X}_0 \otimes T^{-n+1}\mathscr{Y}_0)$$

$$= h_q(T) + h_p(S)$$

$$- \lim_n \frac{1}{n} H(S^{m-1}\mathscr{X}_0 \otimes 2_Y \vee S^m\mathscr{X}_0 \otimes 2_Y \vee \cdots \vee S\mathscr{X}_0 \vee 2_Y \vee \mathscr{X}_0 \otimes \mathscr{Y}_0 \vee S^{-1}\mathscr{X}_0 \otimes T^{-1}\mathscr{Y}_0 \vee \cdots \vee S^{-n+1}\mathscr{X}_0 \otimes T^{-n+1}\mathscr{Y}_0)$$

$$= h_q(T) + h_p(S) + \lim_n \frac{1}{n} \sum_{x_{1-m}\cdots x_n} \sum_{y_1\cdots y_n} \eta\Big(\int_{x_{1-m}\cdots x_n} \nu_x(y_1\cdots y_n)p(dx)\Big).$$

Here we assume that ν is m-fold Markovian and of m-memory with the common integer m, which does not lose generality.

$$R_p = h_q(T) + h_p(S) + (\lim_n \frac{1}{n} \sum_{x_{1-m}\cdots x_n} \sum_{y_1\cdots y_n} \eta(p(x_{1-m}\cdots x_n)\ \nu_x(y_1\cdots y_n))$$

$$= h_q(T) + h_p(S) + \lim_n \frac{1}{n} \sum_{x_{1-m}\cdots x_n} \sum_{y_1\cdots y_n} \eta(p(x_{1-m}\cdots x_n)q_{y_1\cdots y_m}(x) \cdot \prod_{i=1}^{m-n} \tilde{q}_{y_{i+m}\mid y_i\cdots y_{i+m-1}}(x))$$

$$= h_q(T) + h_p(S)$$

$$+ \lim_n \frac{1}{n} \sum_{x_{1-m}\cdots x_n} \sum_{y_1\cdots y_n} p(x_{1-m}\cdots x_n)q_{y_1\cdots y_m}(x) \prod_{i=1}^{n-m} \tilde{q}_{y_{i+m}\mid y_i\cdots y_{i+m-1}}(x) \times$$

$$\times (\log p(x_{1-m}\cdots x_n) + \log q_{y_1\cdots y_m}(x) + \sum_{j=1}^{n-m} \log \tilde{q}_{y_{j+m}\mid y_j\cdots y_{j+m-1}}(x))$$

$$= h_q(T) + \lim_n \frac{1}{n} \sum_{x_{1-m}\cdots x_n} \sum_{y_1\cdots y_n} p(x_{1-m}\cdots x_n) q_{y_1\cdots y_m}(x)$$

$$\cdot \Pi_{i=1}^{n-m} \tilde{q}_{y_{i+m} \mid y_i \cdots y_{i+m-1}}(x)) \left(\sum_{j=1}^{n-m} \log \tilde{q}_{y_{j+m} \mid y_j\cdots y_{j+m-1}}(x)\right)$$

$$= h_q(T) + \lim_n \frac{1}{n} \left\{ \sum_{j=1}^{n-m} \sum_{x_{1-m}\cdots x_n} \sum_{y_1\cdots y_n} p(x_{1-m}\cdots x_n)\ \nu_x(y_1\cdots y_n) \log \tilde{q}_{y_{j+m} \mid y_j\cdots y_{m+j-1}}(x) \right\}$$

$$= h_q(T) + \lim_n \frac{1}{n} \left\{ \sum_{j=1}^{n-m} \sum_{x_{j-m}\cdots x_{j+m}} \sum_{y_j\cdots y_{j+m-1}} p(x_{j-m}\cdots x_{j+m}) \right.$$

$$\left. \cdot\ \nu_x(y_j\cdots y_{j+m-1}) \log \tilde{q}_{y_{j+m} \mid y_1\cdots y_{m+j-1}}(x) \right\}$$

$$= h_q(T) + \lim_n \frac{1}{n} \left\{ \sum_{j=1}^{n-m} H(S^{j+m} 2_X \otimes T^{j+m} \mathscr{Y}_o \right.$$

$$\left. \Big| \bigvee_{i=j-m}^{j-1} S^i \mathscr{X}_0 \otimes T^i 2_Y \vee \bigvee_{k=j}^{j+m-1} S^k \mathscr{X}_0 \otimes T^k \mathscr{Y}_o) \right\}$$

$$= h_q(T) + \int_X \sum_{y_1\cdots y_{m+1}} q_{y_1\cdots y_m}(x)\, \eta(\tilde{q}_{y_{m+1} \mid y_1\cdots y_m}(x)) p(dx),$$

which completes the proof, where $H(\mathscr{A} \mid \mathscr{B})$ means $H(\mathscr{A} \vee \mathscr{B}) - H(\mathscr{B})$.

Q. E. D.

Next we give another proposition:

Proposition 2. *Let ν be an SSSM channel defined by $Q = (q_b)$ and $\widetilde{Q}(x_1) = (\tilde{q}_{b \mid b'}(x_1))$. If p_0 is a Bernoulli input source, i.e., time independent probability measure on $X = A^I$, then the output source $q_0(\cdot) = q(\cdot; p_0, \nu)$ is a Markov measure on $Y = B^I$.*

Proof. Let us prove that

$$q_0([y_n] \mid [y_1\cdots y_{n-1}]) = q_0([y_n] \mid [y_{n-1}]). \tag{3}$$

We abride the brackets [ ] in the following formulae. Then,

$$q_0(y_n \mid y_1\cdots y_{n-1}) = q_0(y_1\cdots y_{n-1}\, y_n) \,/\, q_0(y_1\cdots y_{n-1})$$

$$= q(y_1)\tilde{q}_{y_2 \mid y_1}(x_2)\cdots \tilde{q}_{y_n \mid y_{n-1}}(x_n) p(x_1)\cdots p(x_n) \,/\, (q(y_1)\tilde{q}_{y_2 \mid y_1}(x_2)$$

$$\cdots \tilde{q}_{y_{n-1} \mid y_{n-2}}(x_{n-1}) p(x_1)\cdots p(x_{n-1}))$$

$$= \tilde{q}_{y_n \mid y_{n-1}}(x_n) p(x_n) = q_0(y_n \mid y_{n-1}),$$

which proves the equality (3). Q. E. D.

Theorem 2. *Let ν be an SSSM channel, then the capacity of ν is given by*

$$C_s(\nu) = \sup_{p_0}\left\{ \sum_{a_2 b_1 b_2} q_{b_1}\tilde{q}_{b_2 \mid b_1}(a_2)p_0(a_2) \log (\tilde{q}_{b_2 \mid b_1}(a_2)) / (\sum_{a_2'} \tilde{q}_{b_2 \mid b_1}(a_2')p_0(a_2')) \right\},$$

where p_0 moves on all probability measures on A.

Proof. Let p be any input source, then by Proposition 1,

$$R_p = h_q(T) + \sum_{y_1} q_{y_1}\eta(\tilde{q}_{y_2 \mid y_1}(x_2))p(x_2).$$

The following inequality and equality are well known (cf. 2) p. 82 and p. 78) :

$$h_q(T) = \lim_n H(S\ \mathscr{Y}_0 \mid \vee_{i=0}^{n} S^{-i}\ \mathscr{Y}_0) \leqq H(S\ \mathscr{Y}_0 \mid\ \mathscr{Y}_0).$$

Thus,

$$h_q(T) \leq \sum_{y_0} q(y_0)\, \eta\, (q(y_1 \mid y_0)) = \sum_{y_0} q_{y_0}\, \eta\, (\sum_{x_1} q_{y_1 \mid y_0}(x_1)p(x_1)).$$

Now let p_0 be a Bernoulli source given by

$$p_0(x_1 \cdots x_n) = p(x_1)p(x_2) \cdots p(x_n),$$

then by Proposition 2, denoting $q_0(\cdot) = q(\cdot; p_0, \nu)$,

$$H(S^{-1}\mathscr{Y}_0 \mid \mathscr{Y}_0) = h_{q_0}(T).$$

Therefore

$$\begin{aligned} R_p &= h_q(T) + \sum_{y_1} q_{y_1}\, \eta(\tilde{q}_{y_2 \mid y_1}(x_2))p(x_2) \\ &\leq h_{q_0}(T) + \sum_{y_1} q_{y_1}\, \eta(\tilde{q}_{y_2 \mid y_1}(x_2))p_0(x_2) \quad (= R_{p_0}) \\ &= \sum_{a_2 b_1 b_2} q_{b_1}\tilde{q}_{b_2 \mid b_1}(a_2)p_0(a_2) \log (\tilde{q}_{b_2 \mid b_1}(a_2)) / (\sum_{a_2'} \tilde{q}_{b_2 \mid b_1}(a_2')p_0(a_2')), \end{aligned}$$

which proves the result. Q. E. D.

Therefore we can easily compute the capacity of the channel shown in Fig. 2 of the previous section, which is about 0. 1886 bit/step.

Reference

1) ADLER, R.L., Ergodic and mixing properties of infinite memory channels. *Proc. Amer. Math. Soc.* **12**, 924–930 (1961).
2) BILLINGSLEY, P., Ergodic theory and information. John Wiley and Sons, (1965).
3) HINCHIN, A. I., Mathematical foundations of information theory. Dover, (1958). (English translation)