On the Capacity of Permanent Memory

CHRIS HEEGARD, MEMBER, IEEE

Abstract—Many forms of digital memory have been developed for the permanent storage of information. These include keypunch cards, paper tapes, PROMs, photographic film and, more recently, digital optical disks. All these "write-once" memories have the property that once a "one" is written in a particular cell, this cell becomes irreversibly set at one. Thus, the ability to rewrite information in the memory is hampered by the existence of previously written ones.

The problem of storing temporary data in permanent memory is examined here. Consider storing a sequence of t messages $W_1, W_2, \cdots, W_t$ in such a device. Let each message W_i consist of k_i bits and let the memory contain n cells. We say that a rate t-tuple $(R_1 = k_1/n,\ R_2 = k_2/n, \cdots, R_t = k_t/n)$ is achievable if we can store a sequence of messages at these rates for some n. The capacity $C_t^* \subset \boldsymbol{R}_+^t$ is the closure of the set of achievable rates. The capacity C_t^* for an optical disk-type memory is determined. This result is related to the work of Rivest and Shamir.

A more general model for permanent memory is introduced. This model allows for the possibility of random disturbances (noise), larger input and output alphabets, more possible cell states, and a more flexible set of state transitions. An inner bound on the capacity region C_t^* for this model is presented. It is shown that this bound describes C_t^* in several instances.

I. Introduction

We ARE interested in the temporary storage capacity of memories that have been developed to store permanent information. This class of "write-once" memories (WOMs) includes keypunch cards, paper tape, PROMs, photographic film, and, more recently, digital optical (or video) disks. A keypunch card is a generic example of this type of memory. Binary data is represented on a card by associating the numbers zero and one with a blank or mark, respectively. Once a one (mark) is written on the card, that location becomes permanently associated with a one. Thus the ability to write future data on the card is hampered by the existence of previously written ones.

In a recent paper [1], Rivest and Shamir consider the possibility of rewriting information in permanent memory. They note the potential cost/performance of the new class of digital optical disks. The disks cost on the order of \$100.00, have on the order of 10^{11} memory cells (the equivalent of 40 reels of magnetic tape), and have a high access rate. Rivest and Shamir ask the following question: If we would like to store a sequence of t messages $W_1, W_2, \cdots, W_t$, each of which consists of k bits, how many binary WOM cells $n^*(k,t)$ would be required? Obviously $n^*(k,t) \geq k$, but the interesting fact is that $n^*(k,t)$ can be much less than the product kt. The motivating example is a code that can store two bits ($k = 2$), in three binary WOM cells ($n = 3$), twice ($t = 2$). Rivest and Shamir determine $n^*(k,t)$ for modest values of k and t, determine

$$C_0^*(t) \approx \lim_{k\to\infty} k/n^*(k,t)$$

for small t ($C_0^*(2) \approx 0.7728$), and show

$$C_0^*(t) \approx \frac{1}{t}\log_2(1+t)$$

for large t. The authors also show

$$\lim_{t\to\infty} t/n^*(k,t) = 1$$

for every k and conjectured $n^*(k,t) \approx \max(t, kt/\log_2(1+t))$ for large k and t.

In this paper we expand on some of the notions introduced in [1]. In the process, we answer many of the open questions suggested in the conclusion of [1]. We put these problems into a coding and information theory framework. In many ways, these questions are related to the problem of storing messages in defective computer memory [2]–[4]. References [2]–[4] concern the capacity of defective memory (typically, "stuck-at" cells) when information concerning the locations of the defects is available to the writer (encoder). The difference in our problem lies in the fact that the "defects" are introduced by the storage of previous messages in memory. When we store message W_i, we first read the memory to obtain this defect information. This "side" information is then used to encode the data in such a manner that the creation of new defects is minimized. In this way we better utilize the potential storage capacity of the memory.

We begin our discussion by looking at the Rivest–Shamir problem in a more general setting. Suppose that we would like to store a sequence of t messages $W_1.\ W_2, \cdots, W_t$, where W_i is a message consisting of k_i bits. We say that a rate t-tuple $(R_1 = k_1/n,\ R_2 = k_2/n, \cdots,\ R_t = k_t/n)$ is *achievable* if we can store a sequence of messages at these rates for some n. The capacity $C_t^* \subset \boldsymbol{R}_+^t$ is the closure of the set of achievable rates. We determine C_t^* for the Rivest–Shamir model and relate this to $C_0^*(t)$. We also establish that the maximum average rate achievable

$$C^*(t) = \max_{(R_1, R_2, \cdots, R_t)\in C_t^*} \frac{1}{t}\sum_{j=1}^{t} R_j$$

is equal to $(1/t)\log_2(1+t)$ for all t.

We then develop a more general WOM model. This model allows for the possibility of random disturbances

Manuscript received November 18, 1982; revised July 19, 1984. This research was supported by the National Science Foundation under grant ECS-8204886. The material in this paper was presented in part at the 17th Annual Conference on Information Sciences and Systems, John Hopkins University, March 23–25, 1983.

The author is with the School of Electrical Engineering, Phillips Hall, Cornell University, Ithaca, NY 14853, USA.

0018-9448/85/0100-0034\$01.00 ©1985 IEEE

(noise), larger input and output alphabets, more possible cell states, and a more flexible set of state transitions. An inner bound on the capacity C_t^* is established for this general WOM model. Although it is unlikely that the inner bound is the true capcity, the bound is tight in several special cases. It is shown that when the output and the next state coincide, and when the output is a deterministic function of the current state and input, the achievable rate region is optimum. It is also demonstrated that for the Rivest–Shamir model with binary symmetric noise at the input, the bound describes the capacity region C_t^*.

In a recent paper by Wolf, Wyner, Ziv, and Körner [5], other generalizations of the Rivest–Shamir problem are studied. In a framework derived from Heegard and El Gamal [4], and Wolf *et al.* consider the ϵ-error capacity of deterministic, binary WOMs. As in [4], they study the WOM problem by considering the present state of the memory as side information available to the writer (encoder) and/or reader (decoder). Four cases are considered: 1) both encoder and decoder informed, 2) only encoder informed, 3) only decoder informed and 4) both encoder and decoder uninformed. Case (2) corresponds to our Theorem 4 when $\alpha = \beta = 0$. The most difficult case derived in [5] seems to be the last, where it is shown that

$$C^*(t) \leq \frac{1}{t}\frac{\pi^2}{6\ln(2)}$$

and

$$\lim_{t\to\infty} tC^*(t) = \frac{\pi^2}{6\ln(2)} \approx 2.37.$$

II. The Capacity of the Deterministic Binary WOM

Suppose that we are interested in storing a sequence $W_1, W_2, \cdots, W_t$ of independent messages on a nonerasable disk. We assume that when we store W_i, the values of the previous messages $W_1, W_2, \cdots, W_{i-1}$ need not be retained. Let the disk initially consist of n blank cells, and let the message W_i have rate R_i, that is, let it belong to the set $\{1, 2, \cdots, 2^{nR_i}\}$. When the sequence has length $t = 1$, we can store W_1 at any rate $R_1 \leq 1$ bit per cell. For two writes, $t = 2$, we could simply divide the disk into two pieces, the first consisting of np cells $(0 \leq p \leq 1)$ and the second containing the remaining $n(1-p)$ cells. Then we can first store W_1 followed by W_2 for any rate pair (R_1, R_2) in the set

$$\{(R_1, R_2) \in \boldsymbol{R}_+^2 | R_1 + R_2 \leq 1\}. \tag{1}$$

The purpose of this paper is to show that we can do better than this by constructing a "partition" code.

Fix $0 < \epsilon < 1/2$ and let $C_1 = B_\epsilon$, where the set

$$B_\epsilon \equiv \{x \in \{0,1\}^n | \|x\| \leq n\epsilon\}$$

is the ϵ Hamming ball centered at $\mathbf{0}$ (i.e., the set of binary n-tuples with Hamming weight less than or equal to $n\epsilon$). Define a partition of size 2^{nR_1} for C_1

$$\{A_1^1, A_2^1, \cdots, A_{2^{nR_1}}^1\},$$

where $A_i^1 \cap A_j^1 = \varnothing$, $i \neq j$, and

$$C_1 = \bigcup_{w=1}^{2^{nR_1}} A_w^1.$$

Similarly, let $C_2 = \{0,1\}^n - B_\epsilon$ be the set of binary vectors with Hamming weight greater than $n\epsilon$, and define a partition

$$\{A_1^2, A_2^2, \cdots, A_{2^{nR_2}}^2\}$$

of size 2^{nR_2} for C_2. To store $W_1 \in \{1, 2, \cdots, 2^{nR_1}\}$, simply choose any vector $x_1 \in A_{W_1}^1$ and write it on the disk. The value of W_1 can easily be obtained by recognizing that the vector written on the disk belongs to subset $A_{W_1}^1$. Of course, this requires that $A_{W_1}^1$ not be empty. Since [6, p. 310]

$$|B_\epsilon| = \sum_{i=1}^{n\epsilon}\binom{n}{i} \geq \frac{1}{\sqrt{8\pi n\epsilon(1-\epsilon)}}2^{nh(\epsilon)}$$

(where $h(x) = -x\log_2 x - (1-x)\log_2(1-x)$ is the binary entropy function), we can argue that for sufficiently large n, there exists a partitioning of C_1 with $A_w^1 \neq \varnothing$ for every $w \in \{1, 2, \cdots, 2^{nR_1}\}$, provided $R_1 < h(\epsilon)$.

To store message $W_2 \in \{1, 2, \cdots, 2^{nR_2}\}$, we choose

$$x_2 \in A_{W_2}^2 \cap C(x_1),$$

where

$$C(y) \equiv \{z \in \{0,1\}^n | y_i = 1 \Rightarrow z_i = 1\}$$

is the set of vectors compatible with y, i.e., $C(y)$, are those vectors that can be written on a disk that currently reads y. Note that such a choice of x_2 can be written on the disk undistorted and can be correctly decoded by recognizing its membership in $A_{W_2}^2$. The encoding of W_2 is successful whenever $A_{W_2}^2 \cap C(x_1)$ is nontrivial. We shall now use a random partition argument to show that when $R_2 < 1 - \epsilon$ and n is sufficiently large, there exists a partitioning of C_2 for which

$$A_w^2 \cap C(y) \neq \varnothing \quad \text{for every } w \in \{1, 2, \cdots, 2^{nR_2}\} \text{ and } y \in B_\epsilon.$$

Randomly partition C_2 into 2^{nR_2} subsets of equal size,

$$|A_w^2| = |C_2| 2^{-nR_2}.$$

Fix $w \in \{1, 2, \cdots, 2^{nR_2}\}$ and $y \in B_\epsilon$. Then

$$P(A_w^2 \cap C(y) = \varnothing) = \prod_{i=0}^{|A_w^2|-1} \frac{|C_2| - |C(y) \cap C_2| - i}{|C_2| - i} \leq \left(\frac{|C_2| - |C(y) \cap C_2|}{|C_2|}\right)^{|A_w^2|}.$$

Since $\|y\| \leq n\epsilon$, $|C(y) \cap C_2| \geq 2^{n(1-\epsilon)} - 1$, thus

$$P(A_w^2 \cap C(y) = \varnothing) \leq \left(1 - |C_2|^{-1}(2^{n(1-\epsilon)} - 1)\right)^{|C_2|2^{-nR_2}} < e^{-(2^{n(1-\epsilon-R_2)} - 1)},$$

where the last inequality follows from the fact that $(1-x)^y < e^{-xy}$ for $y > 0$. Note that this probability quickly vanishes for increasing n when $R_2 < 1 - \epsilon$. Furthermore,

$$\begin{aligned} &P\big(\text{there exists a } w \in \{1,2,\cdots,2^{nR_2}\} \\ &\qquad\qquad \text{and} \quad y \in B_\epsilon \text{ with } A_w^2 \cap C(y) = \varnothing\big) \\ &\qquad = P\left(\bigcup_{w=1}^{2^{nR_2}} \bigcup_{y \in B_\epsilon} \{A_w^2 \cap C(y) = \varnothing\} \right) \\ &\qquad \le \sum_{w=1}^{2^{nR_2}} \sum_{y \in B_\epsilon} P(A_w^2 \cap C(y) = \varnothing) \\ &\qquad \le 2^{n(R_2 + h(\epsilon))} e^{-(2^{n(1-\epsilon-R_2)}-1)}, \end{aligned}$$

where we use the fact that $|B_\epsilon| \le 2^{nh(\epsilon)}$ [6, p. 310]. We see that for $R_2 < 1 - \epsilon$ and large n, this probability becomes negligible. This shows not only the existence of the desired partitions of C_2 but also the fact that for large n, almost every partition satisfies the requirement that

$$A_w^2 \cap C(y) \ne \varnothing \quad \text{for every } w \in \{1,2,\cdots,w^{nR_2}\} \quad \text{and} \quad y \in B_\epsilon.$$

Combining these results, we see that there exists a partition code that allows us to store W_1 followed by W_2 for any rate pair (R_1, R_2) in the set

$$\{(R_1, R_2) \in \boldsymbol{R}_+^2 | R_1 < h(\epsilon),\ R_2 < 1 - \epsilon,\ 0 \le \epsilon \le 1/2\}. \tag{2}$$

Let C_2^* denote the closure of this set (2). Note that the set (1) defined by dividing the disk is strictly contained in C_2^* (see Fig. 1). We will presently show that it is not possible to do better than (2); thus we refer to C_2^* as the zero-error capacity region for a sequence of two writes. Similar reasoning can be used to determine the zero-error capacity region C_t^* for any finite value for the sequence length t. Before we state this result as a theorem, consider the following definitions.

The binary "or" operator $\vee$ is defined by

$$x \vee y \equiv \begin{cases} 0, & \text{if } x = y = 0 \\ 1, & \text{otherwise.} \end{cases}$$

A deterministic binary WOM (e.g., a nonerasable disk) can be modeled by $y = x \vee s$, where $y, x, s \in \{0,1\}^n$, and the or operator is performed component-wise. We can interpret the x vector as the WOM input and the s vector as the present WOM state (the vector already written on the disk). The y vector is both the WOM output and the next state of the disk.

A binary $(n, R_1, R_2, \cdots, R_t)$ code consists of t encoding functions

$$f_i\colon \{1,2,\cdots,2^{nR_i}\} \times \{0,1\}^n \to \{0,1\}^n \qquad 1 \le i \le t,$$

and t decoding functions

$$g_i\colon \{0,1\}^n \to \{1,2,\cdots,2^{nR_i}\} \qquad 1 \le i \le t.$$

(Note that f_i need only be defined for those values of the second argument that lie in the range of f_{i-1}.) A code has no errors if for every sequence $w_1, w_2, \cdots, w_t$,

$$w_i = g_i(f_i(w_i, s_{i\text{-}1}) \vee s_{i-1}) \qquad \text{for } i = 1,2,\cdots,t$$

where $s_0 = \mathbf{0}$, $s_i = f_i(w_i, s_{i\text{-}1}) \vee s_{i-1}$.

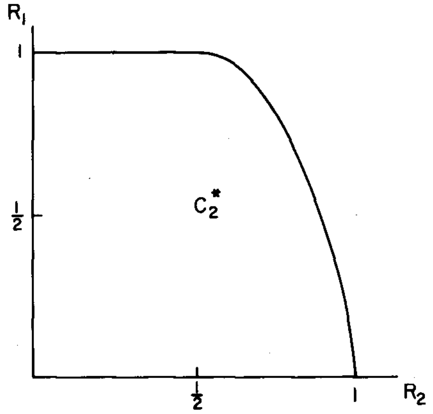

Fig. 1. The capacity region for the deterministic binary WOM when $t = 2$.

We now argue that we cannot do better than C_2^* when $t = 2$. Suppose that we have an encoder $(f_1(w_1), f_2(w_2, f_1(w_1)))$ that maps a message pair (w_1, w_2) onto a pair of binary n-vectors and a decoder (g_1, g_2) that maps these binary vectors onto an estimate of the message pair. We can assume without loss of generality that $f_2(w_2, f_1(w_1))$ is compatible with $f_1(w_1)$ (i.e., $f_2(w_2, f_1(w_1)) \in C(f_1(w_1))$). Let

$$n\epsilon = \max_{w_1 \in \{1,2,\cdots,2^{nR_1}\}} \|f_1(w_1)\|$$

be the maximum Hamming weight of a vector produced by the encoder f_1. Since f_1 must be one-to-one,

$$2^{nR_1} \le \sum_{i=0}^{n\epsilon} \binom{n}{i} \le 2^{nh(\epsilon)}$$

or $R_1 \le h(\epsilon)$. The second encoder, f_2, must also be one-to-one, in the first argument. Let w be any message achieving $\|f_1(w)\| = n\epsilon$. Then

$$2^{nR_2} \le |C(f_1(w))| = 2^{n(1-\epsilon)}$$

or $R_2 \le 1 - \epsilon$. Thus $(R_1, R_2) \in C_2^*$.

A rate t-tuple $(R_1, R_2, \cdots, R_t)$ is said to be zero-error achievable if and only if there exists an error-free $(n, R_1, R_2, \cdots, R_t)$ code for some n. The closure of the set of zero-error achievable rates is called the zero-error capacity region C_t^*.

Theorem 1: The zero-error capacity region for the deterministic binary WOM is the convex region

$$\begin{aligned} C_t^* = \{(R_1, R_2, \cdots, R_t) \in \boldsymbol{R}_+^t | \\ R_1 &\le h(\epsilon_1), \\ R_2 &\le (1 - \epsilon_1) h(\epsilon_2), \\ &\ \vdots \\ R_{t-1} &\le (1 - \epsilon_1)(1 - \epsilon_2) \cdots (1 - \epsilon_{t-2}) h(\epsilon_{t-1}), \\ R_t &\le (1 - \epsilon_1)(1 - \epsilon_2) \cdots (1 - \epsilon_{t-1}), \end{aligned}$$

where $0 \le \epsilon_1, \epsilon_2, \cdots, \epsilon_{t-1} \le 1/2\}$. □

Up to this point, we have assumed that the generation number, i.e., the time index of the current message is known during both the encoding and decoding stages of storage, e.g., for $t = 2$ we know if W_1 or W_2 is stored. The problem as originally formulated [1] did not have this assumption.

We define $C_0^*(t)$ as the zero-error, fixed-rate capacity when the generation number is not explicitly known by the encoder and decoder. The following argument will show

$$C_0^*(t) \equiv \max_{\substack{(R_1, R_2, \cdots, R_t) \in C_t^* \\ R_1 = R_2 = \cdots = R_t = R}} R.$$

This implies that knowledge of the generation number cannot improve the storage rate.

Let $t = 2$ and consider the previously described partition code when $R_1 = R_2 = R$. Assume that the code has zero probability of error. Since $C_1 \cap C_2 = \varnothing$, we can obtain a partition

$$\{A_1, A_2, \cdots, A_{2^{nR}}\}$$

of the set of binary n-tuples $\{0,1\}^n$ by letting

$$A_w = A_w^1 \cup A_w^2.$$

Consider the following encoding algorithm. Given a message $w \in \{1, 2, \cdots, 2^{nR}\}$ and a vector $s \in \{0,1\}^n$ on the disk (initially $s = \mathbf{0}$), store any vector $x \in A_w \cap C(s)$ of minimum Hamming weight. Since we begin with an initially blank disk, we are guaranteed that the first vector written on the disk will have Hamming weight less than or equal to $n\epsilon$. This in turn implies that the second encoding will be successful. Constructing such a code requires $R \leq \min(h(\epsilon), 1 - \epsilon)$. This bound is maximized when $h(\epsilon) = 1 - \epsilon$. Thus $C_0^*(2) = \text{root}\{h(z) - z\} \approx 0.773$. Similar arguments show the following corollary (see Fig. 2).

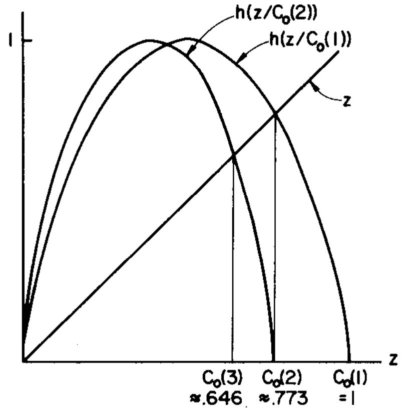

Fig. 2. $C_0(t)$ for the deterministic, binary WOM.

Corollary: For a deterministic binary WOM, $C_0^*(t)$ is obtained recursively

$$C_0^*(1) = 1$$

$$C_0^*(t+1) = \text{root}\{h(z/C_0^*(t)) - z\}. \quad \square$$

The proof of this result involves finding $\epsilon_1, \epsilon_2, \cdots, \epsilon_{t-1}$, which are the solution to

$$h(\epsilon_1) = (1 - \epsilon_1)h(\epsilon_2)$$
$$= \cdots = (1 - \epsilon_1)(1 - \epsilon_2) \cdots (1 - \epsilon_{t-1}).$$

The answer is to set $\epsilon_{t-i} = 1 - Z_i/Z_{i-1}$, where $Z_0 = 1$, $Z_i = \text{root}\{h(z/Z_{i-1}) - z\}$. This corollary is consistent with [1], where Rivest and Shamir also show $C_0^*(t+1) \approx \log_2(1+t)/t$ for large t. Another interesting parameter

$$C^*(t) \equiv \max_{(R_1, R_2, \cdots, R_t) \in C_t^*} \frac{1}{t} \sum_{j=1}^{t} R_j$$

is the maximum average capacity. We have the following corollary:

Corollary: For a deterministic binary WOM,

$$C^*(t) = \frac{1}{t}\log_2(1+t). \quad \square$$

Proving this result involves finding $\epsilon_1, \epsilon_2, \cdots, \epsilon_{t-1}$ that maximize

$$h(\epsilon_1) + (1 - \epsilon_1)h(\epsilon_2) + \cdots$$
$$+ (1 - \epsilon_1)(1 - \epsilon_2) \cdots (1 - \epsilon_{t-1}).$$

The solution is to set $\epsilon_{t-i} = 1/(2+i)$. Thus, for example, in the $t = 2$ case, we find that $\epsilon = 1/3$ will maximize $R_1 + R_2$ in C_2^* and that $C^*(2) = 1/2 \log(3)$. Note that $C_0^*(t) < C^*(t)$ for $t > 1$.

III. On the Capacity of a Noisy WOM

We have developed the notions of coding and capacity for the storage of a sequence of messages on a nonerasable disk. The operation of the individual cells of the disk could be described as follows. Initially, a binary letter applied to the input of the cell is faithfully reproduced at the output. However, once the value of the stored letter is a one, the operation of the cell is altered. From this point forward, the output of the cell becomes fixed at one. Thus we may describe the cell as a binary device with two states (the output equal the input state and the stuck-at-one state). Each cell begins in the first state and permanently transfers to the second state once a one is stored.

We now introduce a general memory cell model to allow for the possibility of random disturbances (noise), larger input and output alphabets, more cell states, and a more flexible set of state transitions. The addition of noise makes it likely that the zero-error capacity is not a useful notion (i.e., it is trivial). It is, in these cases, more meaningful to determine the ϵ-error capacity of the memory (i.e., "At which rates can the probability of error be made arbitrarily small?").

Consider the following definitions. An $(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{Y}, p_0(s,y), p(s^+, y|x,s))$ generalized discrete memoryless WOM consists of three alphabets, $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{S}$; a probability distribution $p_0(s,y)$ on the letters of $\boldsymbol{S} \times \boldsymbol{Y}$; and a conditional probability distribution $p(s^+, y|x,s)$ on the letters of $\boldsymbol{S} \times \boldsymbol{Y}$ conditioned on the letters of $\boldsymbol{X} \times \boldsymbol{S}$. We may interpret the cells as having an input space $\boldsymbol{X}$, a state space $\boldsymbol{S}$ and an output space $\boldsymbol{Y}$.

For a memory consisting of n cells, an initial state vector $\mathscr{S}_0 \in \boldsymbol{S}^n$ and output vector $\mathscr{Y}_0 \in \boldsymbol{Y}^n$ are generated according to the product distribution

$$P(\mathscr{S}_0 = s, \mathscr{Y}_0 = y) = \prod_{j=1}^{n} p_0(s_j, y_j).$$

Thus, initially each cell of the memory independently chooses a state $s \in \boldsymbol{S}$ and output $y \in \boldsymbol{Y}$ with probabiity $p_0(s, y)$. When an input vector $\boldsymbol{x} \in \boldsymbol{X}^n$ is stored in a memory with state vector $\boldsymbol{s} \in \boldsymbol{S}^n$, a new state vector $\mathcal{S}^+ \in \boldsymbol{S}^n$ and output vector $\mathcal{Y} \in \boldsymbol{Y}^n$ are obtained according to the product distribution

$$P(\mathcal{S}^+ = \boldsymbol{s}^+, \mathcal{Y} = \boldsymbol{y} | \boldsymbol{x}, \boldsymbol{s}) = \prod_{j=1}^{n} p(s_j^+, y_j | x_j, s_j).$$

Thus, for each cell with state $s \in \boldsymbol{S}$ and input $x \in \boldsymbol{X}$, the next state $s^+ \in \boldsymbol{S}$ and output $y \in \boldsymbol{Y}$ are independently selected with probability $p(s^+, y | x, s)$.

For example, the deterministic binary WOM has alphabets $\boldsymbol{X} = \boldsymbol{Y} = \boldsymbol{S} = \{0, 1\}$, initial distribution $p_0(0,0) = 1$, and a transition and output distribution

$$\begin{aligned} p(0,0|0,0) &= p(1,1|1,0) = p(1,1|0,1) \\ &= p(1,1|1,1) = 1. \end{aligned}$$

In this case, the output is the next state and a function of the current state and input.

An $(n, R_1, R_2, \cdots, R_t, \delta)$ code consists of t encoding functions

$$f_i \colon \{1, 2, \cdots, 2^{nR_i}\} \times \boldsymbol{Y}^n \to \boldsymbol{X}^n$$

and t decoding functions

$$g_i \colon \boldsymbol{Y}^n \to \{1, 2, \cdots, 2^{nR_i}\}.$$

Let $W_1, W_2, \cdots, W_t$ be a sequence of independent messages with W_i uniformly distributed over the set $W_i \in \{1, 2, \cdots, 2^{nR_i}\}$. A sequence of input, state, and output vectors $(\mathcal{S}_0, \mathcal{Y}_0)$, $(\mathcal{X}_1, \mathcal{S}_1, \mathcal{Y}_1)$, $(\mathcal{X}_2, \mathcal{S}_2, \mathcal{Y}_2), \cdots$, $(\mathcal{X}_t, \mathcal{S}_t, \mathcal{Y}_t)$ is obtained, where for $1 \le i \le t$, $\mathcal{X}_i = f_i(W_i, \mathcal{Y}_{i-1})$. The ith probability of error is defined as

$$P_e^i \equiv P(g_i(\mathcal{Y}_i) \neq W_i).$$

The (worst case) probability of error is

$$\delta \equiv \max_{1 \le i \le t} P_e^i.$$

A rate t-tupe $(R_1, R_2, \cdots, R_t)$ is said to be ϵ-achievable if for any $\epsilon > 0$ there exists an $(n, R_1, R_2, \cdots, R_t, \delta)$ code for some n with $\delta < \epsilon$. The closure of the set of ϵ-achievable rates C_t^* is called the ϵ-error capacity region. $C_0^*(t)$ is defined as the least upper bound on the set of ϵ-achievable rates for a fixed encoder and decoder

$$f \colon \{1, 2, \cdots, 2^{nR}\} \times \boldsymbol{Y}^n \to \boldsymbol{X}^n$$

$$g \colon \boldsymbol{Y}^n \to \{1, 2, \cdots, 2^{nR}\}.$$

(Note that these functions are not allowed to depend on the generation number of the message.) Thus $C_0^*(t)$ is referred to as the fixed-rate capacity. Similarly, $C^*(t)$ is defined as the maximum average rate that is ϵ-achievable:

$$C^*(t) = \max_{(R_1, R_2, \cdots, R_t) \in C_t^*} \frac{1}{t} \sum_{j=1}^{t} R_j.$$

Let us consider how we might extend the idea of a partition code to a memory consisting of n $(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{Y}, p_0(s, y), p(s^+, y | x, s))$ cells. Before we begin, we will need the notion of ϵ-typical sets. Fix a small $\epsilon > 0$, and let $(\mathcal{X}, \mathcal{Y})$ be a pair of independent, identically distributed (i.i.d.) random vectors

$$P(\mathcal{X} = \boldsymbol{x}, \mathcal{Y} = \boldsymbol{y}) = \prod_{j=1}^{n} p(x_j, y_j).$$

Then the set of ϵ-typical $\boldsymbol{x}$ vectors is defined as

$$T_\epsilon(X) = \left\{ \boldsymbol{x} \in \boldsymbol{X}^n \,\middle\|\, \left| \frac{1}{n} \sum_{j=1}^{n} 1_x(x_j) - p(x) \right| < \epsilon \quad \text{for every } x \in \boldsymbol{X} \right\},$$

where $1_x(x_j)$ is 1 if $x_j = x$ and 0 otherwise. This is the set of sequences for which the empirical frequency is within ϵ of the probability $p(x)$ for every letter $x \in \boldsymbol{X}$. We can similarly define the set of jointly ϵ-typical vectors $T_\epsilon(X, Y)$ and the set $T_\epsilon(Y|\boldsymbol{x})$ of vectors $\boldsymbol{y} \in T_\epsilon(Y)$ that are jointly ϵ-typical with a given vector $\boldsymbol{x} \in \boldsymbol{X}^n$. (A complete discussion of ϵ-typical vectors can be found in [7, 8].) We shall need the following facts:

1) If $\mathcal{X}$ is randomly chosen, then $P(\mathcal{X} \in T_\epsilon(X)) \to 1$ as $n \to \infty$

2) If $\boldsymbol{x} \in T_\epsilon(X)$ and $\mathcal{Y}$ is independently chosen according to the marginal distribution for $\mathcal{Y}$, then

$$2^{-n(I(X;Y)+\lambda)} \le P(\mathcal{Y} \in T_\epsilon(Y|\boldsymbol{x})) \le 2^{-n(I(X;Y)-\lambda)}$$

for some $\lambda(\epsilon) > 0$ with $\lambda \to 0$ as $\epsilon \to 0$ (note that $I(X;Y)$ is the mutual information).

Let $t = 2$ and consider the following random partition argument. We will show that for certain values of (R_1, R_2), we can randomly construct an encoder (f_1, f_2) and decoder (g_1, g_2) that will, on the average, have a probability of error that vanishes with increasing n. This will prove the ϵ-achievability of these rates.

Fix $\epsilon > 0$ and let $U_1 \in \boldsymbol{U}$ be an auxiliary random variable. Choose a conditional distribution $p_1(u, x|y)$ on $\boldsymbol{U} \times \boldsymbol{X}$ conditioned on $\boldsymbol{Y}$. Let $p_1(u)$ be the marginal distribution of U_1 under the joint distribution $p_0(s, y) p_1(u, x|y)$ of the random variables (S_0, Y_0, U_1, X_1). Independently choose a set of 2^{nQ_1} vectors according to the uniform distribution over the set $T_\epsilon(U_1)$. Call this new set C_1. Next, randomly partition C_1 into 2^{nR_1} equal size subsets

$$\{A_1^1, A_2^1, \cdots, A_{2^{nR_1}}^1\}.$$

Let $W_1 \in \{1, 2, \cdots, 2^{nR_1}\}$ be the first message to be stored. An initial output $\mathcal{Y}_0$ is read from the memory. Since $\mathcal{Y}_0$ is correlated with the initial state of the memory $\mathcal{S}_0$, the $\mathcal{Y}_0$ vector can be useful in storing W_1. This is done by choosing a vector

$$\mathcal{U}_1 \in A_{W_1}^1 \cap T_\epsilon(U_1 | \mathcal{Y}_0).$$

If such a vector exists, then randomly choose a vector from the set $T_\epsilon(X | \mathcal{U}_1, \mathcal{Y}_0)$ and write this vector into the memory. The encoding is successful whenever

$$A_{W_1}^1 \cap T_\epsilon(U_1 | \mathcal{Y}_0)$$

is not empty. Now

$$\begin{aligned}P\big(A^1_{W_1}\cap T_\epsilon(U_1|\mathscr{Y}_0)=\phi\big) &\le P(\mathscr{Y}_0\notin T_\epsilon(Y_0))\\ &\quad+P\big(A^1_{W_1}\cap T_\epsilon(U_1|\mathscr{Y}_0)=\phi|\mathscr{Y}_0\in T_\epsilon(Y_0)\big).\end{aligned}$$

Since $P(\mathscr{Y}_0\in T_\epsilon(Y_0))\to 1$ for $\epsilon>0$ and large n, we need to find a bound on the second term:

$$\begin{aligned}&P\big(A^1_{W_1}\cap T_\epsilon(U_1|\mathscr{Y}_0)=\phi|\mathscr{Y}_0\in T_\epsilon(Y_0)\big)\\ &\quad=P\Bigg(\bigcap_{u\in A^1_{W_1}}\{u\notin T_\epsilon(U_1|\mathscr{Y}_0)\}|\mathscr{Y}_0\in T_\epsilon(Y_0)\Bigg)\\ &\quad=\big(1-P(u\in T_\epsilon(U_1|\mathscr{Y}_0)|\mathscr{Y}_0\in T_\epsilon(Y_0)\big)^{|A^1_{W_1}|}\\ &\quad\le\big(1-2^{-n(I(U_1;Y_0)+\lambda)}\big)^{2^{n(Q_1-R_1)}}\\ &\quad\le\exp\{-2^{(Q_1-R_1-I(U_1;Y_0)-\lambda)}.\}\end{aligned}$$

Thus if $Q_1-R_1>I(U_1;Y_0)$, ϵ is sufficiently small, and n is large, then the encoding will be successful with high probability.

To decode W_1, we first read $\mathscr{Y}_1$ from the memory. Then we look for a unique estimate $\hat{\mathscr{U}}_1\in C_1\cap T_\epsilon(U_1|\mathscr{Y}_1)$ of the vector $\mathscr{U}_1$. If such an estimate exists and it belongs to the A^1_k subset of C_1, then we set $\hat W_1=k$. Then

$$\begin{aligned}P(\hat W_1\ne W_1)\le P(\hat{\mathscr{U}}_1\ne\mathscr{U}_1)&\le P((\mathscr{U}_1,\mathscr{Y}_1)\notin T_\epsilon(U_1,Y_1))\\ &\quad+P\big(\text{there exists a } u\ne\mathscr{U}_1,\\ &\qquad u\in C_1\cap T_\epsilon(U_1|\mathscr{Y}_1)|\\ &\qquad(\mathscr{U}_1,\mathscr{Y}_1)\in T_\epsilon(U_1,Y_1)\big).\end{aligned}$$

Again the first term will approach zero for large n and positive ϵ. The second term can be expressed as

$$\begin{aligned}&P\Bigg(\bigcup_{\substack{u\in C_1\\ u\ne\mathscr{U}_1}}\{u\in T_\epsilon(U_1|\mathscr{Y}_1)\}|(\mathscr{U}_1,\mathscr{Y}_1)\in T_\epsilon(U_1,Y_1)\Bigg)\\ &\quad\le\sum_{\substack{u\in C_1\\ u\ne\mathscr{U}_1}}P(u\in T_\epsilon(U_1|\mathscr{Y}_1)|(\mathscr{U}_1,\mathscr{Y}_1)\in T_\epsilon(U_1,Y_1))\\ &\quad\le 2^{n(Q_1-I(U_1;Y_1)+\lambda)}.\end{aligned}$$

This probability will become negligible for large n when $Q_1<I(U_1;Y_1)$ and ϵ is small. Combining this bound with the previous $Q_1-R_1>I(U_1;Y_0)$, we conclude that rates $R_1<I(U_1;Y_1)-I(U_1;Y_0)$ are ϵ-achievable.

A similar construction can be used to encode and decode W_2, since for large n with high probability $(\mathscr{U}_1,\mathscr{S}_1,\mathscr{Y}_1)\in T_\epsilon(U_1,S_1,Y_1)$. However, in this case, we may be able to get a better estimate of the state $\mathscr{S}_1$ by using the fact that the estimate $\hat{\mathscr{U}}_1$ is equal to $\mathscr{U}_1$ with high probability. In this case, we choose a conditional distribution $p_2(u^+,x|u,y)$ on $\boldsymbol{U}\times\boldsymbol{X}$ conditioned on $\boldsymbol{U}\times\boldsymbol{Y}$. The marginal $p_2(u)$ of the random variable U_2 under the joint distribution

$$p_0(s,y)p_1(u,x|y)p(s^+,y^+|x,s)p_2(u^+,x^+|u,y^+)$$

of the random variables $(S_0,Y_0,U_1,X_1,S_1,Y_1,U_2,X_2)$ is used to choose 2^{nQ_2} vectors $C_2\subset T_\epsilon(U_2)$. Then C_2 is partitioned into 2^{nR_2} subsets of equal size

$$\{A^2_1,A^2_2,\cdots,A^2_{2^{nR_2}}\}.$$

To encode W_2, choose $\mathscr{U}_2\in A^2_{W_2}\cap T_\epsilon(U_2|\hat{\mathscr{U}}_1,\mathscr{Y}_1)$ and store a vector $\mathscr{X}_2\in T_\epsilon(X_2|\mathscr{U}_2,\hat{\mathscr{U}}_1,\mathscr{Y}_1)$. This will be possible, with high probability, for $Q_2-R_2>I(U_2;U_1,Y_1)$, small ϵ, and large n. To decode W_2, find a unique $\hat{\mathscr{U}}_2\in C_2\cap T_\epsilon(U_2|\mathscr{Y}_2)$. If $\hat{\mathscr{U}}_2\in A^2_k$, then $\hat W_2=k$. The probability of a decoder error will be small for $Q_2<I(U_2;Y_2)$. Thus rates pairs (R_1,R_2) are ϵ-achievable for

$$R_1<I(U_1,Y_1)-I(U_1;Y_0)$$
$$R_2<I(U_2;Y_2)-I(U_2;U_1,Y_1).$$

The following theorem extends this argument to any finite t.

Theorem 2: An achievable rate region. Fix t, $(\boldsymbol{X},\boldsymbol{S},\boldsymbol{Y},p_0(s,y),\ p(s^+,y|x,s))$, and an auxiliary alphabet U. Choose t conditional distributions

$$p_1(u,x|y),\ p_2(u^+,x|u,y),\cdots,p_t(u^+,x|u,y)$$

for $\boldsymbol{U}\times\boldsymbol{X}$ conditioned on $\boldsymbol{U}\times\boldsymbol{Y}$. Let the joint distribution of the random variables $(S_0,Y_0,U_1,X_1,S_1,Y_1,U_2,X_2,S_2,Y_2,\cdots,U_t,X_t,S_t,Y_t)$ take the form

$$p_0(s_0,y_0)p_1(u_1,x_1|y_0)p(s_1,y_1|x_1,s_0)\cdot\prod_{i=2}^{t}p_i(u_i,x_i|u_{i-1},y_{i-1})p(s_i,y_i|x_i,s_{i-1}).$$

Then $(R_1,R_2,\cdots,R_t)\in C_t^*$ if

$$R_1<I(U_1;Y_1)-I(U_1;Y_0),$$

and for $i=2,3,\cdots t$

$$R_i<I(U_i;Y_i)-I(U_i;U_{i-1},Y_{i-1}).\qquad\square$$

We note that the region described by Theorem 2 may not be convex; it is easy to show that convex combinations of rates described by Theorem 2 are also achievable.

Although it is unlikely that the region described by Theorem 2 is the capacity region C_t^*, the following theorems demonstrate that for nontrivial cases this region is optimum. First we note that if we choose conditional distributions $p_i(u^+,x|u,y)$ that do not depend on u we get the following.

Corollary: Fix t, $(\boldsymbol{X},\boldsymbol{S},\boldsymbol{Y},p_0(s,y),p(s^+,y|x,s))$, and an auxiliary alphabet U. Choose t conditional distributions

$$p_1(u,x|y),\ p_2(u^+,x|y),\cdots,p_t(u^+,x|y)$$

for $\boldsymbol{U}\times\boldsymbol{X}$ conditioned on $\boldsymbol{Y}$. Let the joint distribution of the random variables $(S_0,Y_0,U_1,X_1,S_1,Y_1,U_2,X_2,S_2,Y_2,\cdots,U_t,X_t,S_t,Y_t)$ take the form

$$p_0(s_0,y_0)\prod_{i=1}^{t}p_i(u_i,x_i|y_{i-1})p(s_i,y_i|x_i,s_{i-1}).$$

Then $(R_1,R_2,\cdots,R_t)\in C_t^*$ if for $i=1,2,\cdots t$

$$R_i<I(U_i;Y_i)-I(U_i;Y_{i-1}).\qquad\square$$

We note that in general the (convex hull of the) rate region described by the corollary is a subset of the (convex hull of the) rate region of Theorem 2. However, when the output and the state are the same (i.e., $\boldsymbol{S} = \boldsymbol{Y}$ and $S_i = Y_i$, $0 \le i \le t$), the rates achieved by the corollary are equivalent. This is intuitively reasonable. (To see that this is the case, use the chain rule for mutual information to write

$$\begin{aligned} I(U_i; Y_i) - I(U_i; U_{i-1}, Y_{i-1}) \\ = I(U_i; Y_i) - I(U_i; Y_{i-1}) - I(U_i; U_{i-1}|Y_{i-1}). \end{aligned}$$

Then note that $I(U_i; Y_i) - I(U_i; Y_{i-1})$ is only a function of

$$p(u_i, x_i|y_{i-1}) = \sum_{u_{i-1} \in U} p(u_{i-1}|y_{i-1}) p(u_i, x_i|u_{i-1}, y_{i-1})$$

and not $p(u_i, x_i|u_{i-1}, y_{i-1})$. Finally, note that $I(U_i; U_{i-1}|Y_{i-1}) \ge 0$.)

If, in addition to $S_i = Y_i$, the output is a deterministic function of the input and state (i.e., $Y_i = e(X_i, Y_{i-1})$, then the corollary characterizes C_t^*.

Theorem 3: The capacity of deterministic WOMs. Fix t and let $(\boldsymbol{X}, \boldsymbol{Y}, p_0(y), e(x, y))$ describe a deterministic, memoryless WOM. Then

$$\begin{aligned} C_t^* = \{(R_1, R_2, \cdots, R_t) \in R_+^t| \\ R_i \le H(Y_i|Y_{i-1}), 1 \le i \le t\}, \end{aligned}$$

where the joint distribution of the random variables $(Y_0, X_1, Y_1 = e(X_1, Y_0), X_2, Y_2 = e(X_2, Y_1), \cdots, X_t, Y_t = e(X_t, Y_{t-1}))$ is described by

$$p_0(y_0) p_1(x_1|y_0) p_2(x_2|y_1) \cdots p_t(x_t|y_{t-1}). \qquad \square$$

The region described by Theorem 3 is convex since the conditional entropy $H(Y_i|Y_{i-1})$ is a concave function of the joint distribution of (Y_i, Y_{i-1}). The achievability of the interior of this region follows by setting $U_i = Y_i$ in the corollary to Theorem 2. Next we will sketch the converse.

Fix a small $\epsilon > 0$, and suppose there exists an $(n, R_1, R_2, \cdots, R_t, \delta)$ code with $\delta < \epsilon$. Let $\mathscr{Y}_i \equiv e(f_i(W_i, \mathscr{Y}_{i-1}), \mathscr{Y}_{i-1})$, where $W_1, W_2, \cdots, W_t$ is a sequence of independent messages with W_i uniformly distributed over the set $W_i \in \{1, 2, \cdots, 2^{nR_i}\}$. By Fano's inequality,

$$H(W_i|\mathscr{Y}_i) \le h(\epsilon) + n\epsilon R_i \equiv n\theta_n,$$

where $\theta_n \equiv (1/n)h(\epsilon) + \epsilon R_i \to 0$ as $\epsilon \to 0$. Then

$$\begin{aligned} nR_i = H(W_i) &= H(W_i|\mathscr{Y}_{i-1}) \\ &\le I(W_i; \mathscr{Y}_i|\mathscr{Y}_{i-1}) + n\theta_n \\ &\le \sum_{j=1}^{n} H(Y_i^j|Y_{i-1}^j) + \theta_n, \end{aligned}$$

where Y_i^j is the jth component of $\mathscr{Y}_i$, etc. Finally, by the concavity of entropy, we can find random variables (Y_i^*, Y_{i-1}^*) satisfying

$$R_i \le H(Y_i^*|Y_{i-1}^*) + \theta_n.$$

By making ϵ small, we see the rates approach the region described by Theorem 3.

The achievable rate region of Theorem 2 can also be shown to be optimum in nondeterministic cases. Consider the original Rivest-Shamir model with white, binary symmetric noise at the input and the output. We model the output by $\mathscr{Y} = ((\mathscr{X} + \mathscr{Z}_i) \vee \mathscr{S}) + \mathscr{Z}_0$ and the next state by $\mathscr{S}^+ = (\mathscr{X} + \mathscr{Z}_i) \vee \mathscr{S}$, where $\mathscr{Z}_i$ and $\mathscr{Z}_0$ are i.i.d. binary error vectors and addition is performed modulo 2. It would be of interest to determine C_t^* in this case. While this remains as an open problem, we can solve it in the special case of input noise only, i.e., $\mathscr{Z}_0 = 0$). Note that, in this case, the next state and the output agree $\mathscr{Y} = \mathscr{S}^+$.

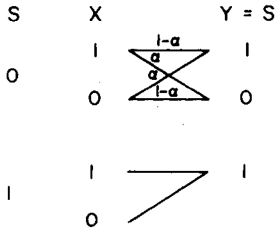

Fig. 3. A binary WOM with input noise.

Theorem 4: The ϵ-error capacity for a binary WOM with binary symmetric noise at the input. Let $\boldsymbol{X} = \boldsymbol{S} = \boldsymbol{Y} = \{0, 1\}$, $p_0(0,0) = 1 - \beta$, $p_0(1,1) = \beta$, $0 \le \alpha \le 1/2$, and $p(s^+, y|x, s)$:

x s	$s^+ = y = 0$	$s^+ = y = 1$
0 0	$1 - \alpha$	α
1 0	α	$1 - \alpha$
0 1	0	1
1 1	0	1

(see Fig. 3). Then

$$\begin{aligned} C_t^* = \{&(R_1, R_2, \cdots, R_t) \in \boldsymbol{R}_+^t| \\ &R_1 \le (1 - \beta)(h(\alpha * \epsilon_1) - h(\alpha)) \\ &R_2 \le (1 - \beta)(1 - \alpha * \epsilon_1)(h(\alpha * \epsilon_2) - h(\alpha)) \\ &\quad\vdots \\ &R_t \le (1 - \beta)(1 - \alpha * \epsilon_1) \cdots (1 - \alpha * \epsilon_{t-1})(1 - h(\alpha)), \end{aligned}$$

where

$$0 \le \epsilon_1, \epsilon, \cdots, \epsilon_{t-1} \le 1/2\}. \qquad \square$$

Note that

$$\alpha * \epsilon \equiv (1 - \alpha)\epsilon + \alpha(1 - \epsilon).$$

For $\beta = \alpha = 0$, we get the deterministic binary WOM result, in agreement with Theorem 1 and [5]. Note that the zero-error capacity and ϵ-error capacity are the same in this case (we do not get trapped by the fact that we use the same notation, C_t^*, for both).

The achievability of Theorem 4 follows from the corollary to Theorem 2 by setting $U_i = X_i$. A point on the boundary of the capacity region is obtained by setting

$$P(X_i = 1|Y_{i-1} = 0) = p_i(1|0) = \epsilon_i, \tag{3.1}$$

$$P(X_i = 1|Y_{i-1} = 1) = p_i(1|1) = \frac{\epsilon_i(1 - \alpha)}{\epsilon_i * \alpha}. \tag{3.2}$$

Note that, in general, when $U_i = X_i$

$$I(X_i; Y_i) - I(X_i; Y_{i-1}) \le I(X_i; Y_i|Y_{i-1}). \tag{3.3}$$

However, under (3.1) and (3.2) equality is achieved in (3.3).

It is for this reason that the converse, which is given in the Appendix, holds.

For this model, we can also determine $C_0^*(t)$ and $C^*(t)$.

Corollary: For a binary WOM with binary symmetric input noise, $C_0^*(t)$ can be obtained recursively,

$$C_0^*(1) = (1-\beta)(1-h(\alpha)),$$

$$C_0^*(t+1) = (1-\beta)\left(\text{root}\left\{h(z/C_0^*(t)) - z - h(\alpha)\right\}\right).$$

□

This result involves finding $\epsilon_1, \epsilon_2, \cdots, \epsilon_{t-1}$, which are the solution to

$$\begin{aligned}&h(\alpha * \epsilon_1) - h(\alpha)\\ &\quad = (1-\alpha * \epsilon_1)(h(\alpha * \epsilon_2) - h(\alpha)) = \cdots\\ &\quad = (1-\alpha * \epsilon_1)(1-\alpha * \epsilon_2)\cdots(1-\alpha * \epsilon_{t-1})(1-h(\alpha)).\end{aligned}$$

The solution is to set $\epsilon_{t-i} = (1-\alpha - Z_i/Z_{i-1})/(1-2\alpha)$, where $Z_0 = 1 - h(\alpha)$, and $Z_i = \text{root}\{h(z/Z_{i-1}) - z - h(\alpha)\}$.

Corollary: For a binary WOM with binary symmetric input noise,

$$C^*(t) = (1-\beta)\left(\frac{1}{t}\log_2\left(1 + \frac{2^{th(\alpha)}-1}{2^{h(\alpha)}-1}\right) - h(\alpha)\right). \;\square$$

Proving this result involves finding $\epsilon_1, \epsilon_2, \cdots, \epsilon_{t-1}$, which maximize

$$\begin{aligned}&(h(\alpha * \epsilon_1) - h(\alpha)) + (1-\alpha * \epsilon_1)(h(\alpha * \epsilon_2) - h(\alpha)) + \cdots\\ &\quad + (1-\alpha * \epsilon_1)(1-\alpha * \epsilon_2)\cdots(1-\alpha * \epsilon_{t-1})(1-h(\alpha)).\end{aligned}$$

The solution is

$$\epsilon_{t-i} = \frac{1-\alpha(1+\gamma_i)}{(1-2\alpha)(1+\gamma_i)},$$

where

$$\gamma_i = \left(1 + \frac{1-2^{ih(\alpha)}}{1-2^{h(\alpha)}}\right)2^{-ih(\alpha)}.$$

Acknowledgment

The author would like to thank the reviewers for their useful comments, especially the reviewer who suggested Theorem 3.

Appendix
Converse to Theorem 4

To show the converse to Theorem 4, we prove that for any $(n, R_1, R_2, \cdots, R_t, \delta)$ code, there exist $0 \le \epsilon_1, \epsilon_2, \cdots, \epsilon_t \le 1$ such that for $1 \le i \le t$

$$\begin{aligned}R_i \le (1-\beta)(1-\alpha * \epsilon_1)\cdots(1-\alpha * \epsilon_{i-1})&\\ \cdot(h(\alpha * \epsilon_i) - h(\alpha)) + \theta,&\end{aligned}$$

where $\theta(\delta) > 0$ satisfies $\theta \to 0$ as $\delta \to 0$.

Because of the randomness of the message sequence $W_1, W_2, \cdots, W_t$ and the noise in the memory, the sequence

$$\mathscr{S}_0 = \mathscr{Y}_0, \mathscr{X}_1, \mathscr{S}_1 = \mathscr{Y}_1, \cdots, \mathscr{X}_t, \mathscr{S}_t = \mathscr{Y}_t$$

is a random sequence where $\mathscr{X}_i \equiv f_i(W_i, \mathscr{S}_{i-1})$.

Let $\|\cdot\|$ denote the Hamming weight of a vector, then

$$E\|\mathscr{S}_0\| = \sum_{j=1}^{n} P(S_0 = 1) = \sum_{j=1}^{n} p_0(1,1) = n\beta.$$

Fix $1 \le i \le t$, and for notational convenience let $W = W_i$, $\mathscr{X} = \mathscr{X}_i$, $\mathscr{Y} = \mathscr{Y}_i$ $(= \mathscr{S}_i)$ and $\mathscr{S} = \mathscr{S}_{i-1}(= \mathscr{Y}_{i-1})$. Then

$$\begin{aligned}E\|\mathscr{Y}\| &= \sum_{j=1}^{n} P(Y_j = 1)\\ &= \sum_{j=1}^{n} P(Y_j = 1, S_j = 1) + P(Y_j = 1, S_j = 0)\\ &= \sum_{j=1}^{n} P(S_j = 1) + P(S_j = 0)P(Y_j = 1 | S_j = 0)\\ &= E\|\mathscr{S}\| + \sum_{j=1}^{n} P(S_j = 0)P(Y_j = 1|S_j = 0). \qquad (5.1)\end{aligned}$$

Since $\alpha \le P(Y_j = 1|S_j = 0) \le 1 - \alpha$, we have

$$(n - E\|\mathscr{S}\|)\alpha \le E\|\mathscr{Y}\| - E\|\mathscr{S}\| \le (n - E\|S\|)(1-\alpha).$$

Thus we may find an $0 \le \epsilon \le 1$ such that

$$E\|\mathscr{Y}\| = E\|\mathscr{S}\| + (n - E\|\mathscr{S}\|)(\epsilon * \alpha). \qquad (5.2)$$

(This can be done for each $1 \le i \le t$; thus

$$\begin{aligned}n - E\|\mathscr{S}_0\| &= n(1-\beta)\\ n - E\|\mathscr{S}_1\| &= n(1-\beta)(1-\epsilon_1 * \alpha) \qquad (5.3)\\ &\vdots\\ n - E\|\mathscr{S}_t\| &= n(1-\beta)(1-\epsilon_1 * \alpha)\cdots(1-\epsilon_t * \alpha).)\end{aligned}$$

Now take $R \hat{=} (1/n)H(W)$, $\theta_n \hat{=} (1/n)h(\delta) + \delta R$. Fano's inequality gives

$$H(W|\mathscr{Y}) \le n\theta_n.$$

Note that $\theta_n \to 0$ as $\delta \to 0$. Then

$$\begin{aligned}nR = H(W) &\le I(W;\mathscr{Y}) + n\theta_n\\ &\le I(W;\mathscr{Y}|\mathscr{S}) + n\theta_n\\ &= \sum_{j=1}^{n} I(W; Y_j|\mathscr{S}, \mathscr{Y}_j^-) + \theta_n,\end{aligned}$$

where $\mathscr{Y}_1^- = \phi$ and $\mathscr{Y}_j^- = (Y_1, Y_2, \cdots, Y_{j-1})$ for $j > 1$. Thus,

$$nR \le \sum_{j=1}^{n} I(W, \mathscr{Y}_j^-, \mathscr{S}_j^-, \mathscr{S}_j^+, X_j; Y_j|S_j) + \theta_n$$

(where $\mathscr{S}_n^+ = \phi$ and $\mathscr{S}_j^+ = (S_{j+1}, \cdots, S_n)$ for $j < n$)

$$= \sum_{j=1}^{n} I(X_j : Y_j|S_j) + \theta_n$$

since $(W, \mathscr{Y}_j^-, \mathscr{S}_j^-, \mathscr{S}_j^+) \to (X_j, S_j) \to Y_j$ form a Markov chain. Then

$$nR \le \sum_{j=1}^{n} P(S_j = 0) I(X_j; Y_j|S_j = 0) + \theta_n$$

(since $I(X_j; Y_j|S_j = 1) = 0$)

$$= \sum_{j=1}^{n} P(S_j = 0)\left[h(P(Y_j = 1|S_j = 0)) - h(\alpha)\right] + \theta_n$$

where $h(p)$ is the binary entropy function. Thus, from (5.1) and (5.2)

$$nR \le \max_q F(q),$$

where

$$\boldsymbol{q} = (q_1, q_2, \cdots, q_n)$$

$$q_j = P(Y_j = 1|S_j = 0)$$

$$F(\boldsymbol{q}) \equiv = \sum_{j=1}^{n} P(S_j = 0)\left[h(q_j) - h(\alpha)\right] + \theta_n$$

and where the maximum is over all $\boldsymbol{q}$ satisfying

$$G(\boldsymbol{q}) \equiv E\|\mathscr{Y}\| - E\|\mathscr{S}\| = \sum_{j=1}^{n} P(S_j = 0)\, q_j.$$

Using a Lagrange multiplier λ, take the partial derivatives

$$\frac{\partial[F(q) + \lambda G(q)]}{\partial q_j} = P(S_j = 0)\left[\log\left(\frac{1 - q_j}{q_j}\right) + \lambda\right].$$

To set the derivatives to zero and to satisfy the constraint, we get

$$q_j = \alpha * \epsilon \quad \text{or} \quad P(X_j = 1|S_j = 0) = \epsilon$$

independent of both the index j and the marginals $P(S_j)$. Thus we obtain the bound

$$R \leq (1 - E\|\mathscr{S}\|)(h(\alpha * \epsilon) - h(\alpha)) + \theta_n.$$

This holds for every $1 \leq i \leq t$; thus from (5.3) we see that

$$R_i \leq (1 - \beta)(1 - \alpha * \epsilon_1) \cdots (1 - \alpha * \epsilon_{i-1}) \cdot (h(\alpha * \epsilon_i) - h(\alpha)) + \theta_n$$

in accordance with Theorem 4. Finally, we note that it is always sufficient to choose $0 \leq \epsilon_i \leq 1/2$ for $1 \leq i \leq t - 1$ and $\epsilon_t = 1/2$.

References

[1] Ronald L. Rivest and Adi Shamir, "How to reuse a "write-once" memory," *Inform. Contr.*, vol. 55, 1–19, 1982.

[2] A. V. Kusnetsov and B. S. Tsybakov, "Coding in a memory with defective cells," translated from *Problemy Peredachi Informatsii*, vol. 10, no. 2, pp. 52–60, Apr.–June 1974.

[3] S. I. Gel'fand and M. S. Pinsker, "Coding for channel with random parameters," *Probl. Contr. Inform. Theory*, vol. 9, no. 1, pp. 19–31, 1980.

[4] Chris Heegard and Abbas El Gamal, "On the capacity of a computer memory with defects," *IEEE Trans. Inform. Theory*, vol. IT-29, no. 5, pp. 731–739, Sep. 1983.

[5] J. K. Wolf, A. D. Wyner, J. Ziv, and J. Körner, "Coding for a write-once memory," *AT&T Bell Lab. Tech. J.*, pp. 1089–1112, July–Aug. 1984.

[6] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*. North-Holland, 1977.

[7] Toby Berger, "Multi-terminal source coding," CISM Courses and Lectures No. 229, *The Information Theory Approach to Communications*, G. Longon, Ed. 1977.

[8] Imre Csiszár and János Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems* New York: Academic, 1982.