AT&T Bell Laboratories Technical Journal
Vol. 63, No. 6, July–August 1984
Printed in U.S.A.

Coding for a Write-Once Memory

By J. K. WOLF,* A. D. WYNER,† J. ZIV,‡ and J. KÖRNER§

(Manuscript received October 21, 1983)

A *write-once memory* (WOM) is a binary storage medium in which the individual bit positions can be changed from the 0 state to the 1 state only once. Examples of WOMs are paper tapes, punched cards, and, most importantly, optical disks. For the latter storage medium, the 1's are marked by a laser that burns away a portion of the disk. In a recent paper, Rivest and Shamir showed that it is possible to update or rewrite a WOM to a surprising degree, and that the total amount of information which can be stored in an N-position WOM in many write/read "generations" or "stages" can be much larger than N.[1] In this paper we extend their results in several directions. Let $C(T, N)$ be the total number of bits of information that can be stored in an N-position WOM using T write/read generations. We consider the four cases that result when the writer (encoder) and/or reader (decoder) know the state of the memory at the previous generation. For three of these cases, when either the encoder and/or decoder knows the previous state, we show that $C(T, N) \sim N \log(T + 1)$, with T held fixed, as $N \to \infty$. For the remaining case, when neither the encoder nor the decoder knows the previous state, we show that $C(T, N) < N\pi^2/(6 \ln 2) \approx N(2.37)$ and that this bound can be approached arbitrarily closely with T, N sufficiently large.

I. INTRODUCTION

A *write-once memory* (WOM) is a binary storage medium in which the individual bit positions can be changed from the 0 state to the 1 state only once. Examples of WOMs are paper tapes, punched cards,

* University of Massachusetts, Amherst, Massachusetts. † AT&T Bell Laboratories. ‡ Technion-Israel Institute of Technology, Haifa, Israel. § Mathematical Institute of the Hungarian Academy of Sciences, Budapest, Hungary.

and, most importantly, optical disks. For the latter storage medium, the 1's are marked by a laser that burns away a portion of the disk. In a recent paper, Rivest and Shamir showed that it is possible to update or rewrite a WOM to a surprising degree, and that the total amount of information that can be stored in an N-position WOM in many write/read "generations" or "stages" can be much larger than N.[1] In this paper we extend their results in several directions. (See Section II, "Discussion on previous work.")

To fix ideas, consider an N-position WOM which we use successively for T write/read generations. Assume that N is large. Assume that initially all memory positions are in the 0 state, and that at the t-th write/read stage ($1 \leq t \leq T$), the writer (encoder) and the reader (decoder) are aware of the state of the memory after the previous [i.e., $(t-1)$-th] write.

At the first write stage, the encoder writes N independent and uniformly distributed bits, of which about half ($N/2$) will be 0. At the second write stage, the encoder writes about $N/2$ independent uniformly distributed bits using only the positions that were in the 0 state after the first write stage. The reader will be able to read the second generation information since we are assuming that it knows the state of the memory after the first generation. We continue in this way, storing $N2^{-(t-1)}$ bits at the t-th generation, for $t \leq T$. Thus, the total number of bits of information that is stored in T generations is about

$$N + \frac{N}{2} + \frac{N}{2^2} + \frac{N}{2^3} + \cdots + \frac{N}{2^{T-1}} = 2(1 - 2^{-T})N \sim 2N,$$

when T is large. Thus, we see that a total of more than N bits can be stored in the N-bit-position WOM. Actually, we can do somewhat better.

Let $\{q_t\}_{t=1}^{T}$, $0 < q_t < 1$, be arbitrary. At the first generation, write 1's on q_1N-bit positions. This can be done in $\binom{N}{q_1N}$ ways, so that we can store

$$B_1 = \log_2 \binom{N}{q_1N}$$

bits at the first generation. (All logarithms in this paper are taken to the base two.) Prior to the second write stage, there are $(1 - q_1)N$ positions in the 0 state. At the second stage, write on a fraction q_2 of these positions, storing

$$B_2 = \log \binom{N(1 - q_1)}{N(1 - q_1)q_2}$$

bits. Continuing in this way for successive stages—writing on a frac-

tion q_t of the $(1 - q_1)(1 - q_2) \cdots (1 - q_{t-1})N$ positions, which are in the 0 state prior to the t-th write—we can store

$$B_t = \log \binom{N \prod_{j=1}^{t-1} (1 - q_j)}{Nq_t \prod_{j=1}^{t-1} (1 - q_j)}$$

$$= \log \binom{Np_1 \cdots p_{t-1}}{Np_1 \cdots p_{t-1}q_t}, \tag{1a}$$

bits at the t-th generation ($1 \leq t \leq T$), where

$$p_t = 1 - q_t, \qquad 1 \leq t \leq T. \tag{1b}$$

Using the Stirling formula for the factorial, we see that when N is large, we can store about $Nh(p_t) \prod_{j=1}^{t-1} p_j$ bits at the t-th generation, where $h(\lambda) = -\lambda \log \lambda - (1 - \lambda) \log (1 - \lambda)$ $(0 \leq \lambda \leq 1)$ is the binary entropy function,* and $\prod_{j=1}^{t-1} p_j = 1$ for $t = 1$.

Suppose that we define the rate R_t as $1/N$ times the number of bits which are stored at the t-th generation. Our principal problem is to find the family of achievable $R_1, R_2, \cdots, R_t$, for the four situations that arise when, at the t-th generation, the encoder and/or decoder is, or is not, informed of the state of the memory after the previous [$(t - 1)$-th] write generation. (We hold T fixed, and let $N \to \infty$.) In particular we are interested in the total rate

$$C_T = \sum_{t=1}^{T} R_t, \tag{2}$$

at which information can be stored in the memory after T generations. For the case considered above (with the encoder and decoder informed),

$$C_T = \sum_{t=1}^{T} h(p_t) \prod_{j=1}^{t-1} p_j. \tag{3}$$

The choice of the $\{p_j\}$, or alternately the $\{q_j\}$, which maximizes C_T is given by

Lemma 1: Let $0 \leq p_t \leq 1$, *for* $t = 1, 2, \cdots, T$. *Then*

$$\sum_{t=1}^{T} h(p_t) \prod_{j=1}^{t-1} p_j \leq \log(T + 1), \tag{4a}$$

* Take $h(0) = h(1) = 0$. It follows immediately from the Stirling formula that $\lim_{N \to \infty} \frac{1}{N} \log \binom{N}{\lambda N} = h(\lambda)$, $0 < \lambda < 1$.

with equality when

$$p_t = \frac{T - t + 1}{T - t + 2}, \qquad 1 \le t \le T. \tag{4b}$$

Proof: For $T = 1, 2, \cdots$, define

$$F_T(p_1, \cdots, p_T) = \sum_{t=1}^{T} h(p_t) \prod_{j=1}^{t-1} p_j, \tag{5}$$

$0 \le p_t \le 1$, $1 \le t \le T$. Observe that

$$\begin{aligned} F_T(p_1, \cdots, p_T) &= h(p_1) + p_1 \sum_{t=2}^{T} h(p_t) \prod_{j=2}^{t-1} p_j \\ &= h(p_1) + p_1 F_{T-1}(p_2, \cdots, p_T). \end{aligned} \tag{6}$$

We now prove the lemma by induction on T. When $T = 1$, $F_1(p_1) = h(p_1) \le \log(2)$, with equality when $p_1 = 1/2$. Assume that the lemma holds for $T = T_0 - 1$. We will show (a) $F_{T_0} \le \log(T_0 + 1)$, and (b) with $\{p_t\}$ given by (4b) with $T = T_0$, $F_{T_0}(p_1, \cdots, p_{T_0}) = \log(T_0 + 1)$.

To show (a), invoke (6) and the induction hypothesis, yielding

$$F_{T_0}(p_1, p_2, \cdots, p_{T_0}) \le h(p_1) + p_1 \log T_0. \tag{7}$$

Setting the derivative of the right member of (7) (which is a concave function of p_1) with respect to p_1 equal to zero, we see that the right member is maximized at $p_1 = T_0/(T_0 + 1)$, so that

$$\begin{aligned} F_{T_0}(p_1, \cdots, p_{T_0}) &\le h\left(\frac{T_0}{T_0 + 1}\right) + \frac{T_0}{T_0 + 1} \log T_0 \\ &= \log(T_0 + 1), \end{aligned}$$

which is (a).

To show (b) let

$$\lambda_t = \frac{(T_0 - 1) - t + 1}{(T_0 - 1) - t + 2}, \qquad t = 1, 2, \cdots, T_0.$$

The induction hypothesis implies that

$$F_{T_0 - 1}(\lambda_1, \cdots, \lambda_{T_0 - 1}) = \log T_0.$$

Further, for p_t given by (4b) with $T = T_0$, $p_{t+1} = \lambda_t$. Thus, (6) yields

$$\begin{aligned} F_{T_0}(p_1, \cdots, p_{T_0}) &= h\left(\frac{T_0}{T_0 + 1}\right) + \frac{T_0}{T_0 + 1} \log T_0 \\ &= \log(T_0 + 1), \end{aligned}$$

establishing (b), and the lemma.

Applying Lemma 1 and (3) we see that for the case where the encoder and decoder are informed, we can achieve $C_T = \log(T + 1)$. Thus, we can store a total of about $N \log(T + 1)$ bits on an N-position WOM in T generations (with T held fixed as $N \to \infty$). In the sequel we will show that the simple scheme outlined above is essentially optimal. Quite surprisingly for two of the other cases—encoder or decoder informed—we can do just as well, i.e., achieve $C_T = \log(T + 1)$. For the fourth case—neither the encoder nor decoder informed—we show that, as $T \to \infty$, the maximum achievable C_T is $(\pi^2/6)\log e \cong 2.37$, which is considerably less than $\log(T + 1)$ but nevertheless significantly greater than unity.

II. FORMAL STATEMENT OF THE PROBLEM AND RESULTS

The memory consists of N cells or bit positions that can be in either the 0 or 1 state. Assume initially that all cells are in the state 0. At time (or generation) $t = 1, 2, \cdots, T$, data $\mathbf{S}^t$ is stored in the memory. Assume that $\{\mathbf{S}^t\}$ is a set of independent random K_t-vectors, and that $\mathbf{S}^t$ is uniformly distributed in binary K_t-space, $1 \leq t \leq T$. Denote the state of the memory at time t by $\mathbf{Y}^t = (Y_{t1}, Y_{t2}, \cdots, Y_{tN})$, where $Y_{tn} = 0$ or 1 and $\mathbf{Y}^0 = (0, 0, \cdots, 0)$.

At time t, the encoder inputs into the memory a binary N-vector $\mathbf{X}^t = (X_{t1}, \cdots, X_{tN})$ (which is a function of $\mathbf{S}^t$ and perhaps $\mathbf{Y}^{t-1}$) and the state of the memory changes to $\mathbf{Y}^t$, where

$$Y_{tn} = X_{tn} \vee Y_{t-1,n} = \begin{cases} 0 \text{ if } X_{tn} = Y_{t-1,n} = 0, \\ 1 \text{ otherwise.} \end{cases} \tag{8}$$

The contents of the memory may now be read, and an estimate $\hat{\mathbf{S}}^t$ of the data $\mathbf{S}^t$ obtained. The error rate is

$$P_e^t = \frac{1}{K_t} E d_H(\mathbf{S}^t, \hat{\mathbf{S}}_t), \tag{9}$$

where $E(\ )$ is expectation and where $d_H(\mathbf{u}, \mathbf{v})$ is the number of positions in which the binary N-vectors $\mathbf{u}$ and $\mathbf{v}$ differ (*Hamming distance*). $\hat{\mathbf{S}}^t$ is a function of $\mathbf{Y}^t$ and perhaps $\mathbf{Y}^{t-1}$.

We now consider four cases.

Case 1 (encoder and decoder informed):

$$\mathbf{X}^t = f_E^t(\mathbf{S}^t, \mathbf{Y}^{t-1}),$$

$$\hat{\mathbf{S}}^t = f_D^t(\mathbf{Y}^t, \mathbf{Y}^{t-1}), \tag{10}$$

$1 \leq t \leq T$. The functions f_E^t and f_D^t are the encoder and decoder functions, respectively. In this case both the encoder and the decoder are informed of the state of the memory at the previous generation.

Case 2 (encoder informed, decoder uninformed):

$$\mathbf{X}^t = f_E^t(\mathbf{S}^t, \mathbf{Y}^{t-1}),$$
$$\hat{\mathbf{S}}^t = f_D^t(\mathbf{Y}^t), \tag{11}$$

$1 \le t \le T$.

Case 3 (encoder uninformed, decoder informed):

$$\mathbf{X}^t = f_E^t(\mathbf{S}^t),$$
$$\hat{\mathbf{S}}^t = f_D^t(\mathbf{Y}^t, \mathbf{Y}^{t-1}), \tag{12}$$

$1 \le t \le T$.

Case 4 (encoder and decoder uninformed):

$$\mathbf{X}^t = f_E^t(\mathbf{S}^t),$$
$$\hat{S}^t = f_D^t(\mathbf{Y}^t), \tag{13}$$

$1 \le t \le T$.

For a given Case (1 through 4) and a given $T \ge 1$, we say that a (rate-) vector $\mathbf{r} = (r_1, r_2, \cdots, r_T)$ $0 \le r_i \le 1$, is *achievable* if, for arbitrary $\epsilon > 0$, there exists an encoder/decoder with parameters T, N, $\{K_t\}_{t=1}^T$ such that, for $1 \le t \le T$,

$$\frac{K_t}{N} \ge r_t - \epsilon, \tag{14a}$$

$$P_e^t \le \epsilon. \tag{14b}$$

Similarly, a (total rate) R is achievable if, for arbitrary $\epsilon > 0$, there exists an encoder/decoder with parameters T, N, $\{K_t\}_1^T$ such that

$$\frac{1}{N} \sum_{t=1}^{T} K_t \ge R - \epsilon, \tag{15a}$$

$$\frac{\sum_{t=1}^{T} P_e^t K_t}{\sum_{t=1}^{T} K_t} \le \epsilon. \tag{15b}$$

The left member of (15b) is the expected fraction of the total of $\sum_1^T K_t$ bits which are decoded in error. The *capacity* C_T of the WOM is the supremum of the achievable total rates.

In each Case (1 through 4) we seek to find the family of achievable rate vectors. We now summarize our results.

Let $\mathbf{p} = (p_1, p_2, \cdots, p_T)$ be a T vector for which $0 \le p_t \le 1$, $1 \le t \le T$, and let

$$\mathscr{R}_T(\mathbf{p}) = \left\{ \mathbf{r} = (r_1, r_2, \cdots, r_T) : 0 \le r_t \le h(p_t) \prod_{j=1}^{t-1} p_j \right\}, \tag{16}$$

where, as in Section I, $h(\lambda)$ is the binary entropy function. Finally, define

$$\mathscr{R}_T = \bigcup_{\mathbf{p}} \mathscr{R}_T(\mathbf{p}). \tag{17}$$

In the sequel, we will establish the following two theorems, which assert that for Cases 1 through 3 (encoder and/or decoder informed), $\mathscr{R}_T$ is the family of achievable rates.

Theorem 1 (direct half): *For Cases 1 through 3, let $T \geq 1$ be given. If $\mathbf{r} \in \mathscr{R}_T$, then $\mathbf{r}$ is achievable.*

Theorem 2 (converse half): *For Cases 1 through 3, and any encoder/decoder with parameters T, N, $\{K_t\}$, and error probabilities $\{P_e^t\}$, there exists a member $\mathbf{r} = (r_1, r_2, \cdots, r_T)$ of $\mathscr{R}_T$ such that*

$$\frac{K_t}{N} \leq r_t + h(P_e^t), \qquad 1 \leq t \leq T. \tag{18}$$

It follows from the discussion in Section I that the capacity (for Cases 1 through 3) is $C_T = \log(T + 1)$. It is also a consequence of our proof of Theorem 1 for Case 2, that for the codes constructed, $P_e^t = 0$, $1 \leq t \leq T$.

For Case 4 (encoder and decoder uninformed), we cannot completely characterize the family of achievable rate vectors. We do, however, establish the following theorems.

Let $\mathbf{p} = (p_1, p_2, \cdots, p_T)$ be a T-vector for which $0 \leq p_t \leq 1$ $(1 \leq t \leq T)$. Let

$$Q_0 = 0, \tag{19a}$$

$$Q_t = \prod_{j=1}^{t} p_j, \qquad 1 \leq t \leq T. \tag{19b}$$

Let $\mathscr{R}'_T(\mathbf{p})$ be the set of $\mathbf{r} = (r_1, \cdots, r_T)$ for which

$$r_t \leq h(Q_t) - p_t h(Q_{t-1}), \tag{20}$$

for $1 \leq t \leq T$. For $\mathbf{r} \in \mathscr{R}'_T(\mathbf{p})$,

$$\begin{aligned} \sum_{t=1}^{T} r_t &\leq \sum_{t=1}^{T} h(Q_t) - p_t h(Q_{t-1}) \\ &= h(Q_T) + \sum_{t=1}^{T} (1 - p_t) h(Q_{t-1}). \end{aligned} \tag{21}$$

We now state

Theorem 3 (existence): *For Case 4, let $T \geq 1$ be given. If $\mathbf{r} \in \mathscr{R}'_T(\mathbf{p})$, for some $\mathbf{p}$, then $\mathbf{r}$ is achievable.*

Theorem 4 (partial converse): *For Case 4, and any encoder/decoder with parameters T, N, $\{K_t\}$, and error probabilities $\{P_e^t\}$, then*

$$\left(\sum_{t=1}^{T} \frac{K_t}{N}\right) \left(1 - h\left(\frac{\sum_{t=1}^{T} P_e^t K_t}{\sum_{t=1}^{T} K_t}\right)\right)$$

$$\leq \sup_{\mathbf{r} \in \bigcup_{\mathbf{p}} \mathscr{R}_T(\mathbf{p})} \sum_{t=1}^{T} r_t$$

$$= \sup_{\mathbf{p}} \left\{h(Q_T) - \sum_{t=1}^{T} (1 - p_t) h(Q_{t-1})\right\} \triangleq \rho_T. \tag{22}$$

It follows from Theorem 4 that if R is an achievable total rate, then $R \leq \rho_T$. Furthermore, we show in Section III that

$$\rho_T \leq \frac{\pi^2}{6} \log_2 e \approx 2.37318, \tag{23a}$$

and that

$$\lim_{T \to \infty} \rho_T = \frac{\pi^2}{6} \log_2 e. \tag{23b}$$

Discussion of previous work

An information theoretic treatment of coding for memories of this type was given by Kusnetsov and Tsybatov[2] in 1974. They studied binary memories with defective cells—typically cells that are "stuck at 1". Their work was extended and generalized considerably by Heegard and El Gamal.[3] Rivest and Shamir[1] originated the concept of rewriting on WOMs. Their problem is similar to that in previous models with the "stuck at 1" state of Ref. 3 and Ref. 2 being the result of writing on the WOM in previous generations. In the new problem, the system designer must balance the needs of memory users at all generations. In a very recent paper, Heegard[4] generalized the Rivest-Shamir results in several ways.

The models in all of the above papers correspond to our Case 2, in that it is always assumed that the encoder can read the memory before writing, and that the decoder is unaware of the state of the memory before the present write. Heegard and El Gamal[3] proved a coding theorem for the memory with "stuck at 1" defects which can be adapted to our rewriting on WOM's problem. Essentially this was done by Rivest and Shamir, although they apparently were not aware of the earlier work. Concerning Case 2, our results represent an extension of

previous results in that our converse theorem (Theorem 2, Case 2) holds for codes with a small error probability, and not a zero error probability as in Ref. 4 and Ref. 1. Our results for Cases 1, 3, and 4 are new.

III. PROOF OF CONVERSES

In this section we establish the converse Theorems 2 and 4.

Proof of Theorem 2: It suffices to establish Theorem 2 for Case 1 (both encoder and decoder informed). Let $\{f_E^t, f_D^t\}_{t=1}^T$ define an encoder/decoder for Case 1 [defined by (10)] with parameters T, N, $\{K_t\}$ and error probabilities $\{P_e^t\}$. Consider the t-th generation. Since $\mathbf{S}^t$, $\mathbf{X}^t$, $\mathbf{Y}^t$, $\hat{\mathbf{S}}^t$ is a Markov chain given $\mathbf{Y}^{t-1}$, the data processing theorem yields*

$$I(\mathbf{S}^t; \hat{\mathbf{S}}^t | \mathbf{Y}^{t-1}) \leq I(\mathbf{X}^t; \mathbf{Y}^t | \mathbf{Y}^{t-1}). \tag{24a}$$

Now

$$\begin{aligned} I(\mathbf{S}^t; \hat{\mathbf{S}}^t | \mathbf{Y}^{t-1}) &= H(\mathbf{S}^t | \mathbf{Y}^{t-1}) - H(\mathbf{S}^t | \hat{\mathbf{S}}^t, \mathbf{Y}^{t-1}) \\ &\geq H(\mathbf{S}^t) - H(\mathbf{S}^t | \hat{\mathbf{S}}^t), \end{aligned} \tag{24b}$$

where the inequality follows from the independence of $\mathbf{S}^t$ and $\mathbf{Y}^{t-1}$, which implies $H(\mathbf{S}^t | \mathbf{Y}^{t-1}) = H(\mathbf{S}^t) = K_t$, and from the fact that conditioning decreases entropy. Further, from Fano's inequality, $1/N\, H(\mathbf{S}^t | \hat{\mathbf{S}}^t) \leq h(P_e^t)$, so that (24) yields

$$\frac{1}{N} I(\mathbf{X}^t; \mathbf{Y}^t | \mathbf{Y}^{t-1}) \geq \frac{K_t}{N} - h(P_e^t). \tag{25}$$

Now, writing $\mathbf{Y}^t = (Y_{t1}, Y_{t2}, \cdots, Y_{tN})$, we have for $0 \leq t \leq T$,

$$\begin{aligned} \frac{1}{N} I(\mathbf{X}^t; \mathbf{Y}^t | \mathbf{Y}^{t-1}) &\overset{(1)}{\leq} \frac{1}{N} H(\mathbf{Y}^t | \mathbf{Y}^{t-1}) \\ &\overset{(2)}{\leq} \frac{1}{N} \sum_{n=1}^{N} H(Y_{tn} | Y_{t-1,n}) \\ &\overset{(3)}{=} \frac{1}{N} \sum_{n=1}^{N} H(Y_{tn} | Y_{t-1,n} = 0) \Pr\{Y_{t-1,n} = 0\}. \end{aligned} \tag{26}$$

Step 1 is a standard inequality; step 2 follows from the fact that the entropy of a vector is no greater than the sum of the entropies of its components, and conditioning decreases entropy; and step 3 follows from $H(Y_{tn} | Y_{t-1,n} = 1) = 0$. Setting

* Remember $\mathbf{Y}^0 = (0, 0, \cdots, 0)$.

$$Q_{tn} = \Pr\{Y_{tn} = 0\},$$

$$Q_t = \frac{1}{N} \sum_{n=1}^{N} Q_{tn},$$

$1 \le n \le N$, $1 \le t \le T$, we have $Q_{tn} \ge Q_{t+1,n}$ and

$$\Pr\{Y_{tn} = 0 \mid Y_{t-1,n} = 0\} = \frac{\Pr\{Y_{tn} = 0,\ Y_{t-1,n} = 0\}}{\Pr\{Y_{t-1,n} = 0\}}$$

$$= \frac{\Pr\{Y_{tn} = 0\}}{\Pr\{Y_{t-1,n} = 0\}} = \frac{Q_{tn}}{Q_{t-1,n}}.$$

Hence $H(Y_{tn} \mid Y_{t-1,n} = 0) = h\left(\dfrac{Q_{tn}}{Q_{t-1,n}}\right)$, and (26) become

$$\frac{1}{N} I(\mathbf{X}^t, \mathbf{Y}^t \mid \mathbf{Y}^{t-1}) \le \frac{1}{N} \sum_{n=1}^{N} Q_{t-1,n} h\left(\frac{Q_{tn}}{Q_{t-1,n}}\right)$$

$$= Q_{t-1} \frac{1}{N} \sum_{n=1}^{N} \frac{Q_{t-1,n}}{Q_{t-1}} h\left(\frac{Q_{tn}}{Q_{t-1,n}}\right)$$

$$\le Q_{t-1} h\left(\frac{1}{Q_{t-1}} \frac{1}{N} \sum_{n=1}^{N} Q_{t-1,n} \frac{Q_{tn}}{Q_{t-1,n}}\right)$$

$$= Q_{t-1} h\left(\frac{Q_t}{Q_{t-1}}\right). \tag{27}$$

The second inequality in (27) follows from the concavity of $h(\cdot)$. Combining (25) and (27) we have, for $1 \le t \le T$,

$$Q_{t-1} h\left(\frac{Q_t}{Q_{t-1}}\right) \ge \frac{K_t}{N} - h(P_e^t). \tag{28}$$

Now let us define $p_t = \dfrac{Q_t}{Q_{t-1}} \le 1$, $1 \le t \le T$. Since $Q_0 = 1$, we have $Q_t = \prod_{j=1}^{t} p_j$, so that (28) is

$$\frac{K_t}{N} \le \prod_{j=1}^{t} p_j h(p_t) + h(P_e^t), \tag{29}$$

$1 \le t \le T$. Comparison of (29) with (16) yields (18) and Theorem 2.

We now turn our attention to Theorem 4. Let $f_E^{(t)}(\cdot)$ and $f_D^{(t)}(\cdot)$, $1 \le t \le T$ define an encoder/decoder for Case 4 with parameters T, N, $\{K_t\}_{t=1}^{T}$, and error probabilities $\{P_e^t\}_{t=1}^{T}$. Then

$$\sum_{t=1}^{T} K_t = \sum_{t=1}^{T} H(\mathbf{S}^t) \overset{(1)}{=} H(\mathbf{S}^1, \mathbf{S}^2, \cdots, \mathbf{S}^T)$$

$$\overset{(2)}{=} H(\mathbf{S}^1, \mathbf{S}^2, \cdots, \mathbf{S}^T, \mathbf{Y}^1, \mathbf{Y}^2, \cdots, \mathbf{Y}^T)$$

$$= H(\mathbf{Y}^T) + H(\mathbf{S}^1, \mathbf{S}^2, \cdots, \mathbf{S}^T, \mathbf{Y}^1 \cdots \mathbf{Y}^{T-1} | \mathbf{Y}^T)$$

$$= H(\mathbf{Y}^T) + U_T, \tag{30}$$

where

$$U_t = H(\mathbf{S}^1, \cdots, \mathbf{S}^t, \mathbf{Y}^1, \cdots, \mathbf{Y}^{t-1} | \mathbf{Y}^t), \qquad 1 \le t \le T. \tag{31}$$

Step 1 in eq. (30) follows from the independence of the $\{\mathbf{S}^t\}_1^T$, and step 2 from the fact that $\mathbf{Y}^t$ is functionally determined by $\mathbf{S}^1, \mathbf{S}^2, \cdots, \mathbf{S}^t$, $1 \le t \le T$. (Take $U_0 = 0$.)

Now, for $1 \le t \le T$,

$$U_t = H(\mathbf{S}^t, \mathbf{Y}^{t-1} | \mathbf{Y}^t)$$

$$+ H(\mathbf{S}^1, \cdots, \mathbf{S}^{t-1}, \mathbf{Y}^1, \cdots, \mathbf{Y}^{t-2} | \mathbf{Y}^{t-1}, \mathbf{S}^t, \mathbf{Y}^t)$$

$$\overset{(1)}{=} H(\mathbf{Y}^{t-1} | \mathbf{S}^t, \mathbf{Y}^t) + H(\mathbf{S}^t | \mathbf{Y}^t)$$

$$+ H(\mathbf{S}^1, \cdots, \mathbf{S}^{t-1}, \mathbf{Y}^1, \cdots, \mathbf{Y}^{t-2} | \mathbf{Y}^{t-1})$$

$$\overset{(2)}{=} H(\mathbf{Y}^{t-1} | \mathbf{S}^t, \mathbf{X}^t, \mathbf{Y}^t) + H(\mathbf{S}^t | \mathbf{Y}^t) + U_{t-1}, \tag{32}$$

where step 1 follows from the fact that $(\mathbf{S}^1, \cdots, \mathbf{S}^{t-1}, \mathbf{Y}^1, \cdots, \mathbf{Y}^{t-2})$, $(\mathbf{S}^t, \mathbf{Y}^t)$ are conditionally independent given $\mathbf{Y}^{t-1}$, and step 2 follows from $\mathbf{X}^t = f_E^t(\mathbf{S}^t)$.

Now from Fano's inequality,

$$H(\mathbf{S}^t | \mathbf{Y}^t) \le K_t h(P_e^t),$$

and since conditioning decreases entropy,

$$H(\mathbf{Y}^{t-1} | \mathbf{S}^t, \mathbf{X}^t, \mathbf{Y}^t) \le H(\mathbf{Y}^{t-1} | \mathbf{X}^t, \mathbf{Y}^t).$$

Thus, from (32), for $1 \le t \le T$,

$$(U_t - U_{t-1}) \le H(\mathbf{Y}^{t-1} | \mathbf{X}^t, \mathbf{Y}^t) + K_t h(P_e^t).$$

Summing on t, we obtain (noting that $U_0 = 0$)

$$U_T \le \sum_{t=1}^{T} H(\mathbf{Y}^{t-1} | \mathbf{X}^t, \mathbf{Y}^t) + \sum_{t=1}^{T} K_t h(P_e^t).$$

Substituting into (30) we have

$$\sum_{k=1}^{T} K_t \le H(\mathbf{Y}^T) + \sum_{t=1}^{T} H(\mathbf{Y}^{t-1}|\mathbf{X}^t\mathbf{Y}^t) + \sum_{t=1}^{T} K_t h(P_e^t)$$

$$\le \sum_{n=1}^{N} H(Y_{Tn}) + \sum_{n=1}^{N}\sum_{t=1}^{T} H(Y_{t-1,n}|X_{tn}Y_{tn}) + \sum_{t=1}^{T} K_t h(P_e^t). \quad (33)$$

Applying the concavity of $h(\cdot)$ and Jensen's inequality, we have

$$\sum_{t=1}^{T} K_t h(P_e^t) = \left(\sum_t K_t\right)\frac{\sum K_t h(P_e^t)}{\sum K_t}$$

$$\le (\sum K_t) h\left(\frac{\sum K_t P_e^t}{\sum K_t}\right),$$

so that (33) yields

$$\left(\sum_{t=1}^{T} K_t\right)\left(1 - h\left(\frac{\sum K_t P_e^t}{\sum K_t}\right)\right)$$

$$\le \sum_{n=1}^{N}\left[H(Y_{Tn}) + \sum_{t=1}^{T} H(Y_{t-1,n}|X_{tn}Y_{tn})\right]. \quad (34)$$

Now fix n ($1 \le n \le N$) and write $X_{tn} = X_t$, $Y_{tn} = Y_t$, $Y_{t-1,n} = Y_{t-1}$. The random variables X_t, Y_t, Y_{t-1} are binary. Consider, for $2 \le t \le T$,

$$H(Y_{t-1}|X_tY_t) = H(Y_{t-1}, X_t, Y_t) - H(X_t, Y_t)$$

$$\overset{(1)}{=} H(Y_{t-1}, X_t) - H(X_t, Y_t)$$

$$= H(Y_{t-1}) + H(X_t|Y_{t-1}) - H(X_t) - H(Y_t|X_t)$$

$$\overset{(2)}{=} H(Y_{t-1}) - H(Y_t|X_t), \quad (35)$$

where step 1 follows from $Y_t = X_t \vee Y_{t-1}$, and step 2 from the independence of X_t and Y_{t-1}. Now put back the n dependence. Letting $p_{tn} = \Pr\{X_{tn} = 0\}$, $1 \le t \le T$, $1 \le n \le N$, we have from the independence of the $\{X_{tn}\}_{t=1}^{T}$,

$$\Pr\{Y_{tn} = 0\} = \Pr\{X_{1n} = X_{2n} = \cdots = X_{tn} = 0\}$$

$$= \prod_{j=1}^{t} p_{tn} \triangleq Q_{tn},$$

and

$$\Pr\{Y_{tn} = 0|X_{tn} = 0\} = \Pr\{Y_{t-1,n} = 0\} = Q_{t-1,n},$$

and

$$\Pr\{Y_{tn} = 0 \mid X_{tn} = 1\} = 0.$$

Thus (35) is

$$\begin{aligned} H(Y_{t-1,n} \mid X_{tn} Y_{tn}) &= h(Q_{t-1,n}) + p_{tn} h(Q_{t-1,n}) \\ &= (1 - p_{tn}) h(Q_{t-1,n}), \end{aligned}$$

and the term in brackets in (34) is

$$\begin{aligned} & H(Y_{Tn}) + \sum_{t=1}^{T} H(Y_{t-1,n} \mid X_{tn} Y_{tn}) \\ &= h(Q_{Tn}) + \sum_{t=1}^{T} (1 - p_{tn}) h(Q_{t-1,n}) \\ &\leq \rho_T. \end{aligned}$$

Thus (34) is

$$\sum K_t \left[1 - h\left(\frac{\sum P_e^t K_t}{\sum K_t} \right) \right] \leq N \rho_T,$$

which is Theorem 4.

Our final task in this section is to establish (23a) and (23b). We begin by establishing the following:

Proposition 1: Let $0 \leq a < b < \infty$. Then

$$\int_a^b h(e^{-x}) dx \geq (1 - e^{-(b-a)}) h(e^{-a}).$$

Proof: Let $y = e^{-x}$, $y_1 = e^{-a}$, $y_0 = e^{-b}$. Then $0 \leq y_0 < y_1 \leq 1$ and we must show

$$\begin{aligned} \int_a^b h(e^{-x}) dx = \int_{y_0}^{y_1} \frac{h(y)}{y} dy &\geq \left(1 - \frac{y_0}{y_1} \right) h(y_1) \\ &= (y_1 - y_0) \frac{h(y_1)}{y_1}. \end{aligned}$$

But $\dfrac{h(y)}{y}$ is nonincreasing $\left(\dfrac{d}{dy} \dfrac{h(y)}{y} = y^{-2} \log(1 - y) \leq 0 \right)$, so that $\dfrac{h(y)}{y}$ can be underbounded in the integral by $h(y_1)/y_1$, establishing the proposition.

Since ρ_T is nondecreasing in T, we can establish (23a) and (23b) by showing that

$$\sup \sum_{t=1}^{\infty} (1 - p_t) h(Q_{t-1}) = \frac{\pi^2}{6} \log e, \tag{36}$$

where the supremum is with respect to sequences $\{p_t\}_{t=1}^{\infty}$, where $0 \le p_t \le 1$, and $Q_t = \prod_{j=1}^{t} p_j$. Let $\{p_t\}$ be given, and consider

$$\psi = \sum_{t=1}^{\infty} (1 - p_t)h(Q_{t-1}). \tag{37}$$

For $1 \le t < \infty$, let $\alpha_t = -\ln p_t$, so that $p_t = e^{-\alpha_t}$, and

$$Q_t = \prod_{j=1}^{t} p_j = \prod_{j=1}^{t} e^{-\alpha_j} = \exp\left\{-\sum_{j=1}^{t} \alpha_j\right\} = e^{-x_t},$$

where $x_t = \sum_{j=1}^{t} \alpha_j$. (Take $x_0 = 0$.) Thus,

$$p_t = e^{-\alpha_t} = e^{-(x_t - x_{t-1})},$$

and

$$\begin{aligned}
\psi &= \sum_{t} (1 - p_t)h(Q_{t-1}) \\
&= \sum_{t=1}^{\infty} (1 - e^{-(x_t - x_{t-1})})h(e^{-x_{t-1}}) \\
&\le \sum_{t=1}^{\infty} \int_{x_{t-1}}^{x_t} h(e^{-x})dx = \lim_{t\to\infty} \int_{0}^{x_t} h(e^{-x})dx,
\end{aligned}$$

where the inequality follows from the Proposition 1. Since $h(\cdot) \ge 0$ and

$$\begin{aligned}
\int_0^{\infty} h(e^{-x})dx &= \int_0^1 \frac{h(y)}{y}\,dy \\
&= (\log_2 e) \int_0^1 \left(-\ln y - \frac{(1-y)}{y} \ln(1-y)\right) dy \\
&= (\log_2 e)\,\frac{\pi^2}{6},
\end{aligned}$$

we have shown that

$$\sup \psi = \sup \sum (1 - p_t)h(Q_{t-1}) \le (\log_2 e)\,\frac{\pi^2}{6}. \tag{38}$$

Furthermore, ψ can be made arbitrarily close to the right member of (38) by setting $p_t = e^{-\delta}$ for sufficiently small $\delta > 0$. In other words,

$$\psi = \sum_{t=1}^{\infty} (1 - e^{-\delta})h(e^{-t\delta}) \to \int_0^{\infty} h(e^{-x})dx = (\log e)\,\frac{\pi^2}{6},$$

as $\delta \to 0$. This completes the verification of (23a) and (23b).

IV. PROOFS OF (DIRECT) THEOREMS 1 and 3

In this section we give proofs of the "direct" coding theorems (Theorems 1 and 3). Actually, we need two proofs (for Cases 2 and 3) for Theorem 1. We give these in Sections 4.1 and 4.3, respectively, and prove Theorem 3 (for Case 4) in Section 4.4.

4.1 Case 2 (encoder informed, decoder uninformed)

We begin with some definitions. The *weight*, $|\mathbf{u}|$, of a binary N-vector $\mathbf{u}$ is the number of nonzero entries in $\mathbf{u}$. Let $B_N(w)$ be the set of binary N vectors with weight w. We say that binary N-vector $\mathbf{u}$ *covers* the binary N-vector $\mathbf{v}$, denoted $\mathbf{u} > \mathbf{v}$, if $\mathbf{u}$ has 0 entries only in positions in which $\mathbf{v}$ has 0 entries. Thus, for example, when $N = 4$, $(1010) > (1000)$, but (1010) does not cover (1100).

Now consider the encoder for Case 2 with the parameters N, T, $\{K_t\}_1^T$ given. Let $M_t = 2^{K_t}$, $1 \le t \le T$. We will specify an ad hoc encoder as follows. Let $\{w_t\}_{t=0}^T$ satisfy

$$0 = w_0 < w_1 \cdots < w_T \le N.$$

The encoder will see to it that $|\mathbf{Y}^t| = w_t$, $1 \le t \le T$. It does this by setting $\mathbf{X}^t$ equal to an N-vector which covers $\mathbf{Y}^{t-1}$ (so that $\mathbf{X}^t = \mathbf{Y}^t$) and for which $|\mathbf{X}^t| = |\mathbf{Y}^t| = w_t$. The encoding is done as follows. For $1 \le t \le T$, let $\{A_m^t\}$, $1 \le m \le M_t$, be a partition of $B_N(w_t)$. Thus, for $1 \le t \le T$,

$$A_m^t \subseteq B_N(w_t), \qquad 1 \le m \le M_t,$$

$$A_m^t \cap A_{m'}^t = \phi, \qquad m \ne m',$$

$$\sum_{m=1}^{M_t} A_m^t = B_N(w_t).$$

At the t-th write, the encoder observes $\mathbf{Y}^{t-1}$, and if $\mathbf{S}^t$ corresponds to message m, it searches A_m^t to find a vector that covers $\mathbf{Y}^{t-1}$. If it finds such a vector, $\mathbf{y}$, it sets $\mathbf{X}^t = \mathbf{Y}^t = \mathbf{y}$. The decoder can recover the message m by observing that $\mathbf{Y}^t \in A_m^t$. Also $|\mathbf{Y}^t| = w_t$. An error will occur if and only if no $\mathbf{y}$ which covers $\mathbf{Y}^{t-1}$ can be found in A_m^t.

For $1 \le t \le T$, $1 \le m \le M_t$, $\mathbf{u} \in B_N(w_{t-1})$, let $F(\mathbf{u}, A_m^t) = 0$ or 1 according as A_m^t contains a vector that covers $\mathbf{u}$. Clearly, we make no error for $1 \le t \le T$, $1 \le m \le M_t$, if

$$\psi \triangleq \sum_{t=1}^{T} \sum_{m=1}^{M_t} \sum_{\mathbf{u} \in B_N(w_{t-1})} F(\mathbf{u}, A_m^t) = 0. \tag{39}$$

Now turn to Theorem 2. Let T, $\epsilon > 0$, $\mathbf{r} \in \mathscr{R}_T$ be given. We will show that with N sufficiently large and with $\{w_t\}$ suitably chosen, and with $K_t/N = \log M_t = r_t - \epsilon$, that there exists a family of partitions

$\{A_m^t\}$ for which $\psi = 0$. Thus we will have shown that not only is $\mathbf{r} \in \mathscr{R}_T$ achievable, but that P_e^t can be made equal to 0. We do this by choosing the partitions $\{A_m^t\}$ at random (according to a probability law which we will specify later) and computing the expectation $E\,\psi$. We will show that $E\,\psi \to 0$ as $N \to \infty$. Since ψ is integer valued, when $E\,\psi < 1$, there must be a family of partitions for which $\psi = 0$.

Here is how the random partitions are chosen: For $1 \le t \le T$, pick a $\mathbf{v} \in B_N(w_t)$ and place it in class A_m^t with probability $1/M_t$ $(1 \le m \le M_t)$. Do this independently for each of the members of $B_N(w_t)$, and each t. Under this random experiment, ψ is a random variable and

$$E\,\psi = \sum_t \sum_m \sum_{\mathbf{u} \in B_N(w_{t-1})} E\,F(\mathbf{u}, A_m^t). \tag{40}$$

For fixed t, m, $\mathbf{u} \in B_N(w_{t-1})$,

$$E\,F(\mathbf{u}, A_m^t) = \Pr\left\{\begin{array}{l}\text{for all } \mathbf{v} \in B_N(w_t) \text{ such that } \mathbf{v} > \mathbf{u}, \\ \mathbf{v} \notin A_m^t\end{array}\right\}$$

$$= \left(1 - \frac{1}{M_t}\right)^{\nu_t} \le \exp\left\{\frac{-\nu_t}{M_t}\right\}, \tag{41}$$

where ν_t is the number of vectors $\mathbf{v} \in B_N(w_t)$ which cover $\mathbf{u} \in B_N(w_{t-1})$. Since in choosing $\mathbf{v}$ to cover $\mathbf{u}$ we must place w_{t-1} 1's in those positions in which $\mathbf{u}$ is 1, and we can put the remaining $w_t - w_{t-1}$ 1's in any of the remaining $N - w_{t-1}$ positions, we have

$$\nu_t = \binom{N - w_{t-1}}{w_t - w_{t-1}}.$$

We now choose the $\{w_t\}$. Since $\mathbf{r} \in \mathscr{R}_T$, there must be a vector $\mathbf{p} = (p_1, \cdots, p_T)$ such that $\mathbf{r} \in \mathscr{R}_T(\mathbf{p})$. Let

$$w_t = N - Q_t N,$$

where $Q_t = \prod_{j=1}^t p_j$ (and $Q_0 = 1$). Then, as $N \to \infty$,

$$\nu_t = \binom{Q_{t-1}N}{Q_{t-1}(1 - p_t)N} = 2^{NQ_{t-1}h(1-p_t)+O(\log N)}. \tag{42}$$

Substituting (41) and (42) into (40) and using $M_t \le 2^N$, $|B_N(w_t)| \le 2^N$ we have

$$E\,\psi \le T2^{2N}\exp\left\{\frac{-1}{M_t}\,2^{NQ_{t-1}h(p_t)+O(\log N)}\right\}.$$

Setting $K_t/N = 1/N \log M_t = r_t - \epsilon \le Q_{t-1}h(p_t) - \epsilon$, we have

$$E\,\psi \le T\exp\{-2^{N\epsilon+o(N)}\} \to 0,$$

which is what we had to prove.

4.2 Random coding

In this section we state the well-known random channel-coding theorem[5] in a form that will enable us to establish our direct theorems for Cases 3 and 4 with little difficulty. Consider a discrete memoryless channel with input and output alphabets $\mathscr{X}$, $\mathscr{Y}$, respectively, and transition probability $P_c(y|x)$, $y \in \mathscr{Y}$, $x \in \mathscr{X}$. A *code* $\mathscr{C}$ with parameters N, M is a subset $\mathscr{C} = \{\mathbf{x}_1, \cdots, \mathbf{x}_M\} \subseteq \mathscr{X}^N$ with cardinality $|\mathscr{C}| = M$. The *maximum-likelihood* decoder is a mapping $F_D: Y^N \rightarrow \{1, 2, \cdots, M\}$ for which $f_D(\mathbf{y})$ = the smallest m such that

$$P_c^{(N)}(\mathbf{y}|\mathbf{x}_m) \geq P_c^{(N)}(\mathbf{y}|\mathbf{x}_{m'}), \qquad m' \neq m, \tag{43a}$$

where

$$P_c^{(N)}(\mathbf{y}|\mathbf{x}) = \prod_{n=1}^{N} P_c(y_n|\mathbf{x}_n), \tag{43b}$$

$\mathbf{y} = (y_1, \cdots, y_N) \in \mathscr{Y}^N$, $\mathbf{x} = (x_1, \cdots, x_N) \in \mathscr{X}^N$. Let $\Phi_m(\mathbf{y}, \mathscr{C}) = 0$ or 1 according as $f_D(\mathbf{y}) = m$ or $\neq m$. When each of the M-code vectors in $\mathscr{C}$ are used with equal probability, the "word" error probability is

$$P_e = \frac{1}{M} \sum_{m=1}^{M} \sum_{y \in \mathscr{Y}^N} P_c^{(N)}(\mathbf{y}|\mathbf{x}_m)\Phi_m(\mathbf{y}, \mathscr{C}). \tag{44}$$

Now let $p_0(x)$, $x \in \mathscr{X}$, be a probability distribution on $\mathscr{X}$, and let I_0 be the mutual information corresponding to the distribution $p_0(x)P_c(y|x)$ on $\mathscr{X} \times \mathscr{Y}$. A random-code ensemble is constructed as follows. Let the M-code vectors $\mathbf{X}_m$ in $\mathscr{C}$ be drawn independently with $\Pr\{\mathbf{X}_m = (x_1, \cdots, x_N)\} = \prod_{n=1}^{N} p_0(x_n)$. The quantity P_e in (44) is now a random variable which depends on the choice of $\mathscr{C}$. Write it as $P_e(\mathscr{C})$, and write its expectation $E\,P_e(\mathscr{C}) = g(N, M)$. Of course, $g(\cdot)$ depends on $p_0(\cdot)$ and $P_c(\cdot|\cdot)$ too. We now state the well-known random-coding theorem.[5]

Theorem 5: Let $P_c(\cdot|\cdot)$ and $p_0(\cdot)$ be given, and let $g(N, M)$ and I_0 be as defined above. Then, with $\mathscr{R} > 0$, held fixed,

$$g(N, 2^{NR}) \rightarrow 0, \text{ as } N \rightarrow \infty,$$

provided $R < I_0$.

We conclude from this theorem that provided N is sufficiently large, there exists at least one code $\mathscr{C}$ with parameters N and $M = 2^{RN}$ such that $P_e(\mathscr{C})$ is arbitrarily small.

4.3 Case 3 (encoder uninformed, decoder informed)

For a given N, T and encoder/decoder as defined in Section II, we can think of the information K_t-vector $\mathbf{S}^t$ as an integer in $\{1, 2, \cdots, M_t\}$, where $M_t = 2^{K_t}$, and set

$$\mathbf{x}_m^t = f_E^t(m), \quad 1 \le t \le T, \quad 1 \le m \le M_t. \tag{45}$$

Thus at the t-th generation, when the message is m, the encoder writes $\mathbf{x}_m^t$. Let $\mathscr{C}_t = \{\mathbf{x}_m^t\}_{m=1}^{M_t}$ be the "code" for the t-th generation, $1 \le t \le T$. The proofs for this theorem and the next depend on a random choice of $\{\mathscr{C}_t\}_{t=1}^T$. Here is a rough and imprecise sketch of the main idea.

Let $\{p_t\}_{t=1}^T$ satisfy $0 \le p_t \le 1$, $1 \le t \le T$. The codes $\{\mathscr{C}_t\}$ are chosen randomly and independently, according to the following probability law. Each of the M_t code vectors in $\mathscr{C}_t$ is chosen independently, with the probability that the mth code vector be $\mathbf{x} = (x_1, \cdots, x_N) \in \mathscr{X}^N$ is equal to $\prod_{n=1}^N p^{(t)}(x_n)$, where

$$p^{(t)}(0) = p_t, \qquad p^{(t)}(1) = 1 - p_t. \tag{46}$$

Now let us consider the t-th write/read generation. Prior to the t-th write, the n-th bit position will be a 0, i.e., $Y_{t-1,n} = 0$, if it was not written in *each* of the $(t-1)$ previous generations. In some "average" sense, this happens with probability $\prod_{j=1}^{t-1} p_t = Q_{t-1}$. Since the decoder at the t-th generation knows $Y_{t-1,n}$ and Y_{tn}, and it is impossible for $Y_{t-1,n} = 1$, $Y_{tn} = 0$, there are essentially three possible "outputs" $(Y_{t-1,n}, Y_{tn}) \triangleq Z_{tn}$. Z_{tn} can take the values: $a \triangleq (0, 0)$, $b \triangleq (0, 1)$, $c \triangleq (1, 1)$. If, for example, the channel input $X_{tn} = 0$, then

$$\Pr\{Z_{tn} = (0, 0) = a \mid X_{tn} = 0\} = \Pr\{Y_{t-1,n} = 0\} = Q_{t-1}.$$

Thus, as far as the t-th generation is concerned, $\mathbf{Z}^t$ is the output of the memoryless channel with input X, output Z and transition probability given by Fig. 1. The random-coding theorem suggests that, in the t-th generation, we can have highly reliable transmission provided that $M_t \le 2^{N(I_t - \epsilon)}$, where I_t is the $I(X; Z)$, which results when the input X to the channel in Fig. 1 has $\Pr\{X = 0\} = p_t$. Thus,

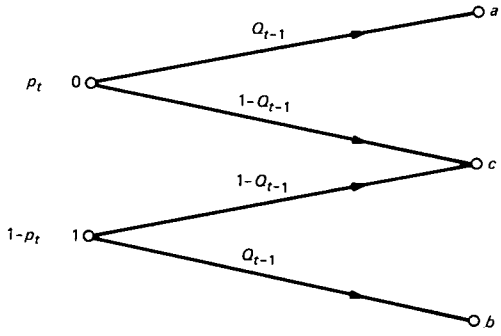

Fig. 1—Equivalent channel for Case 3.

$$I_t = I(X; Z) = H(Z) - H(Z|X)$$
$$= H(V, Z) - H(Z|X),$$

where $V = V(Z) = 1$ when $Z = a$ or b, and $V = 0$ when $Z = c$. Since $H(V) = h(Q_{t-1})$ and $H(Z|X) = h(Q_{t-1})$, we have

$$I_t = H(V) + H(Z|V) - H(Z|X)$$
$$= h(Q_{t-1}) + \Pr\{V = 0\}H(Z|V = 0)$$
$$+ \Pr\{V = 1\}H(Z|V = 1) - h(Q_{t-1})$$
$$= \Pr\{V = 1\}H(Z|V = 1) = Q_{t-1}h(p_t). \tag{47}$$

Thus we are led to conjecture that, for a given T and $\mathbf{p}$, any $\mathbf{r} \in \mathscr{R}_T(\mathbf{p})$ [given by (16)] is achievable for Case 3. We now proceed to a rigorous proof.

Again, consider an encoder/decoder with parameters N, T, $\{K_t\}$. Consider the t-th write/read generation. Suppose $\mathbf{S}^t = m$ $(1 \le m \le M_t)$, and consider

$$\Pr\{\mathbf{Y}^t = \mathbf{y}^t, \mathbf{Y}^{t-1} = \mathbf{y}^{t-1}|\mathbf{S}^t = m\}$$
$$= \Pr\{\mathbf{Y}^{t-1} = \mathbf{y}^{t-1}\}f(\mathbf{y}^t|\mathbf{y}^{t-1}\mathbf{x}_m^t), \tag{48a}$$

where $f(\mathbf{y}^t|\mathbf{y}^{t-1}, \mathbf{x}_m^t) = 1$ if

$$\mathbf{y}^t = \mathbf{y}^{t-1} \vee \mathbf{x}_m^t \tag{48b}$$

("V" is bitwise "inclusive or"), and $f = 0$ otherwise. Now $\Pr\{\mathbf{Y}^{t-1} = \mathbf{y}^{t-1}\}$ depends on the codes $\mathscr{C}_j$, $1 \le j \le t-1$ (but not on $\mathscr{C}_t$). Let us make this dependence explicit and write

$$\Pr\{\mathbf{Y}^t = \mathbf{y}^t, \mathbf{Y}^{t-1} = \mathbf{y}^{t-1}|\mathbf{S}^t = m\}$$
$$= P(\mathbf{y}^t, \mathbf{y}^{t-1}|\mathbf{x}_m^t, \mathscr{C}_1, \cdots, \mathscr{C}_{t-1}). \tag{49}$$

Now, let $\mathbf{y}^t = (y_{t1}, y_{t2}, \cdots, y_{tN})$, and note that $(y_{t-1,n}, y_{t,n})$ cannot take the value $(1, 0)$. Let us define $a = (0, 1)$, $b = (1, 0)$, $c = (1, 1)$, and let $\mathbf{z}^t = (z_{t1}, z_{t2}, \cdots, z_{tn})$, where

$$z_{tn} = (y_{t-1,n}, y_{tn}) \in \{a, b, c\} \triangleq \mathscr{Z}.$$

Now write (49) as

$$\Pr\{\mathbf{Y}^t = \mathbf{y}^t, \mathbf{Y}^{t-1} = \mathbf{y}^{t-1}|\mathbf{S}^t = m\} = P(\mathbf{z}^t|\mathbf{x}_m^t, \mathscr{C}_1, \cdots, \mathscr{C}_{t-1}). \tag{50}$$

Finally, let the vector $(p_1, p_2, \cdots, p_T)$ be arbitrary, $0 \le p_T \le 1$, $1 \le t \le T$. Let the codes $\mathscr{C}_1 \cdots \mathscr{C}_T$ be chosen independently according to the presciption given above eq. (46). Then (with $\mathbf{x}_m^t$ held fixed), the expectation

$$E\, P(\mathbf{z}^t|\mathbf{x}_m^t, \mathscr{C}_1, \cdots, \mathscr{C}_{t-1}) = P_c^{(N)}(\mathbf{z}^t|\mathbf{x}_m^t), \tag{51}$$

where $P_c^{(N)}(\mathbf{z}^t|\mathbf{x}_m^t)$ corresponds to the discrete memoryless channel with input alphabet $\mathscr{X}=\{0, 1\}$, and output alphabet $\mathscr{Y}=\{a, b, c\}$ and transition probability represented by Fig. 1.

Now consider the decoder at the t-th read generation. The decoder examines $\mathbf{Y}^t$, $\mathbf{Y}^{t-1}$. Let us use the following decoding rule. When $(\mathbf{Y}^t, \mathbf{Y}^{t-1}) = \mathbf{z} \in \mathscr{Y}^N$, let $f_D(\mathbf{z})$ be the smallest m, $1 \le m \le M_t$, such that

$$P_c^{(N)}(\mathbf{z}^t|\mathbf{x}_m^t) \ge P_c^{(N)}(\mathbf{z}^t|\mathbf{x}_{m'}^t), \qquad m' \ne m.$$

Let $\psi_m^t(\mathbf{z}^t, \mathscr{C}_t)$ be 0 or 1 according as $f_D(\mathbf{z}^t) = m, f_D(\mathbf{z}^t) \ne m$. Then the error probability* at the t-th generation (given $\mathbf{S}^t = m$) is, using (50),

$$\begin{aligned} P_{em}^t &\le \sum_{y^{t-1}, y^t} \Pr\{\mathbf{Y}^{t-1} = \mathbf{y}^{t-1}, \mathbf{Y}^t = \mathbf{y}^t|\mathbf{S}^t = m\} \\ &= \sum_{\mathbf{z}^t \in \mathscr{Y}^N} P(\mathbf{z}^t|\mathbf{x}_m^t, \mathscr{C}_1, \cdots, \mathscr{C}_{t-1})\Psi_m^t(\mathbf{z}^t, \mathscr{C}_t), \end{aligned} \tag{52}$$

and the overall error probability is

$$\begin{aligned} P_e &= \sum_{t=1}^{T} \sum_{m=1}^{M_t} \frac{1}{M_t} P_{em}^t \\ &= \sum_{t=1}^{T} \sum_{m=1}^{M_t} \sum_{\mathbf{z}^t} \frac{1}{M_t} P(\mathbf{z}^t|\mathbf{x}_m^t, \mathscr{C}_1, \cdots \mathscr{C}_{t-1})\Psi_m^t(\mathbf{z}^t, \mathscr{C}_t). \end{aligned} \tag{53}$$

Taking the expectation of P_{em}^t over the random-code ensemble defined above, and noting that the random codes $\mathscr{C}_1, \cdots, \mathscr{C}_T$ are independent, we have from (51) and (53)

$$EP_e = \sum_{t=1}^{T} E\left[\frac{1}{M_t} \sum_{m=1}^{M_t} \sum_{z^t} P_c^{(N)}(\mathbf{z}^t|\mathbf{x}_m^t)\Psi_m^t(z^t, \mathscr{C}_t)\right], \tag{54}$$

where the expectation in the right number of (53) is taken with respect to $\mathscr{C}_t$. Applying Theorem 5, we conclude that for $\epsilon > 0$, this expectation $\to 0$, as $N \to \infty$, provided

$$M_t \le 2^{N(I_t - \epsilon)},$$

where $I_t = Q_{t-1}h(p_t)$. See (47). Thus for given T, $\mathbf{p} = (p_1, \cdots, p_T)$, we have established that $\mathbf{r} \in \mathscr{R}_T(\mathbf{p})$ is achievable. This is Theorem 1 for Case 3.

4.4 Case 4 (encoder and decoder uninformed)

In this section we establish Theorem 3 for the situation in Case 4 (encoder and decoder uninformed of the state of the memory at the previous generation). The proof is almost exactly the same as that of

* The right member of (52) is the so-called "word error probability", i.e., the probability that $\hat{S}^t \ne S^t$. P_e^t as defined by (9) $\le \Pr\{\hat{S}^t \ne S^t\}$.

Theorem 1 for Case 3, which was given in Section 4.3 (where only the encoder was uninformed).

Let T, $\mathbf{p} = (p_1, \cdots, p_T)(0 \le p_t \le 1)$ be given. The codes $\{\mathscr{C}_t\}_{t=1}^{T}$ are defined exactly as in Section 4.3, and we use a random-code ensemble exactly as above (46). Since the decoder at the t-th generation is uninformed of $\mathbf{Y}^{t-1}$, it must operate on $\mathbf{Y}^t$ instead of $\mathbf{Z}^t$ as in Case 3. This leads us to define the channel in Fig. 2 to replace the channel in Fig. 1, to define $P_c^{(N)}(\cdot \mid \cdot)$.

The rest of the proof parallels the proof in Section 4.3, but here

$$\begin{aligned} I_t = I(X;\, Y) &= H(Y) - H(Y|X) \\ &= h(p_t Q_{t-1}) - \Pr\{X = 0\} H(Y|X = 0) \\ &\quad - \Pr\{X = 1\} H(Y|X = 1) \\ &= h(p_t Q_{t-1}) - p_t h(Q_{t-1}) = h(Q_t) - p_t h(Q_{t-1}). \end{aligned}$$

Referring to (19a) and (19b) and (20) leads us to conclude that any $\mathbf{r} = (r_1, \cdots, r_T) \in \mathscr{R}_T'(\mathbf{p})$ is achievable, which is Theorem 3.

5. SOME AD-HOC RESULTS

Let us look at the family of achievable rates for Cases 1 through 3 when we impose the additional constraint that the rates at each generation be equal, i.e., that $R_t \equiv R$, $1 \le t \le T$. This is the case studied by Rivest and Shamir.[1] There is no closed-form expression for the maximum achievable R, but we can find it numerically as follows.

We seek a set $\{p_t\}$, $0 \le p_t \le 1$, $1 \le t \le T$, such that

$$R_t = h(p_t) \prod_{j=1}^{t-1} p_j \equiv R. \tag{55}$$

Thus if p_T is chosen, $R_{T-1} = R$ implies that

$$h(p_{T-1}) = p_{T-1} h(p_T),$$

for which there is exactly one solution for p_{T-1}. Further, $1/2 \le p_{T-1} \le 1$. Define $\alpha(\lambda)$, $1/2 \le \lambda \le 1$, as the unique solution of

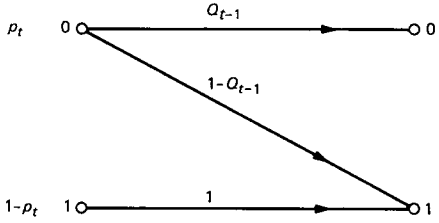

Fig. 2—Equivalent channel for Case 4.

$$h(\alpha) = \alpha h(\lambda). \tag{56}$$

We see that

$$\begin{aligned} p_{T-1} &= \alpha(p_T) \\ p_{T-2} &= \alpha(\alpha(p_T)) \\ &\vdots \\ p_1 &= \alpha^{(T-1)}(p_T), \end{aligned} \tag{57}$$

where $\alpha^{(k)}(\cdot)$ is the kth iterate of $\alpha(\cdot)$. Differentiating (56) with respect to λ yields

$$\frac{d\alpha(\lambda)}{d\lambda} = \left(\frac{d\lambda}{d\alpha}\right)^{-1} = \left(\frac{h'(\alpha) - h(\lambda)}{\alpha h'(\lambda)}\right)^{-1} > 0,$$

$1/2 \leq \lambda \leq 1$. Thus, $\alpha^{(T-1)}(\cdot)$ is monotonically increasing. Since $R_1 = h(p_1) = R$, we maximize R by minimizing $p_1 = \alpha^{(T-1)}(p_T)$. Thus choose $p_t = 1/2$.

The iterates $\alpha^{(k)}(1/2)$ can be obtained graphically or numerically in a straightforward manner. We obtained $\alpha^{(1)}(1/2) = 0.77291\cdots$, $\alpha^{(2)}(1/2) = 0.83524$, $\alpha^{(3)}(1/2) = 0.86876$, $\alpha^{(4)}(1/2) = 0.89021$. Thus, under the constraint $R_t \equiv R$, we have for $T = 4$, $C = \sum_1^4 R_t = 4h(\alpha^{(4)}(1/2)) = 1.997\cdots$, while the unconstrained total capacity is $\log(4 + 1) = 2.3219\cdots$.

Several ad-hoc coding schemes were investigated for Case 2 where the encoder is informed of the previous contents of the WOM but the decoder is uninformed. Only the results for one of these schemes is repeated here.

The simplest case of the coding scheme to be discussed is to consider that the WOM is segmented into two equal-sized sub-WOMs, one for storing data and one for directing the decoder to the newly written data. In the first generation the encoder writes in the data-storing sub-WOM and the reader reads from that sub-WOM. For subsequent generations, the encoder does two write operations. It first copies the state of the data-storing sub-WOM into the second sub-WOM. It then writes new data, only in these positions of the data-storing sub-WOM in which there are zeros. By comparing the information in the two halves of the sub-WOM, the decoder knows in which positions (of the data-storing sub-WOM) the new data have been written.

Rather than optimize and analyze this simple case, we do this for a generalized version of this scheme. We take the N-bit WOM and subdivide it into $[N/K]$ K-bit bytes. At each generation, if a given byte is the all-zero sequence, the encoder can use it to convey new data. However, if it is any other sequence, the encoder nulls it by overwriting

the all-one sequence in that portion of the memory. The decoder treats any byte other than the all-one byte as carrying new information.

The encoder uses the following scheme to write new information. On the t-th generation, $i = 1, 2, \cdots, T$, it writes the all-zero word with probability p_t and it writes any of the $(2^K - 2)$ words that are not all zero or all one with equal probability $(1 - p_t)/(2^K - 2)$. It does not use the all-one word to carry information.

The rate of information for the T generations is as follows:

$$R_1 = \frac{1}{K}\left[-p_1 \log p_1 - (1 - p_1)\log \frac{(1 - p_1)}{2^K - 2}\right] \triangleq \frac{1}{K}[k(p_1)],$$

$$R_t = \frac{p_1 p_2 \cdots p_{t-1}}{K} k(p_t), \qquad t = 2, 3, \cdots, T.$$

The total rate sum for this scheme is then

$$C_T = \frac{1}{K}[k(p_1) + p_1 k(p_2) + \cdots + (p_1 p_2 \cdots p_{T-1})k(p_T)].$$

In a manner similar to that used in Section I, one can prove that for a fixed K, the maximum C_T is obtained for

$$p_{T-i} = \frac{[(i - 1)(2^K - 2) + (2^K - 1)]}{i(2^K - 2) + (2^K - 1)},$$

resulting in a maximum C_t of

$$C_T = \frac{1}{K} \log(T(2^K - 2) + 1).$$

For $T = 2$ and 3, the largest values of C_T are obtained for $K = 3$, and for $T \geq 4$, the largest values of C_T are obtained for $K = 2$.

REFERENCES

1. R. Rivest and A. Shamir, "How to Reuse a Write-Once Memory," Inform. and Control, *55*, No. 1 (October 1982), pp. 1–19.
2. A. V. Kusnetsov and B. S. Tsybakov, "Coding in a Memory with Defective Cells," translated from Problemy Peredachi, Infromatsii, *10*, No. 2 (April–June 1974), pp. 52–60.
3. C. Heegard and A. El Gamal, "On the Capacity of a Computer Memory With Defects," IEEE Trans Inform. Theory, *IT-29*, No. 5 (September 1983), pp. 731–9.
4. C. Heegard, "On the Capacity of Permanent Memory," 1983 Conf. on Inform. Sciences and Systems, Johns Hopkins University (March 1983).
5. R. G. Gallager, *Information Theory and Reliable Communication*, New York: McGraw-Hill, 1968, Theorem 5.6.2, pp. 138.

AUTHORS

János Körner, Diploma in Mathematics, 1970, Loránd Eötvös University, Budapest. In 1970 he joined the Mathematical Institute of the Hungarian Academy of Sciences. In 1972 Mr. Körner was on leave at CISM, Udine, Italy; in the spring of 1974 he was a Visiting Professor at Ohio State University;

and in 1980 he was a Visiting Professor at Linkoping University, Sweden. From 1981–83 he was a visiting member of the Mathematics and Statistics Research Center of Bell Laboratories. He has co-authored with Imre Csiszar the book *Information Theory: Coding Theorems for Discrete Memoryless Systems* (Academic Press, 1982). His current research interests are in information theory and its interplay with combinatorics. János Körner currently is serving as an Associate Editor for the IEEE Transactions on Information Theory.

Jack K. Wolf, B.S.E.E., 1956, University of Pennsylvania; M.S.E., M.A., and Ph.D., 1957, 1958, and 1960, respectively, Princeton University; New York University 1963–65; Polytechnic Institute of Brooklyn, 1965–73; University of Massachusetts, 1973—. During the academic year, 1968–69, Mr. Wolf was a member of the Mathematics Research Center, Bell Laboratories. Presently he is a Professor of Electrical and Computer Engineering at the University of Massachusetts, Amherst. His research interests are in information theory, algebraic coding theory, communication systems, and computer networks. He is also currently International Chairman, Commission C, URSI. Editor Transactions on Information Theory, Algebraic Coding, 1969–72. Board of Governors, Information Theory Group 1970–76, 1980—.

Aaron D. Wyner, B.S., 1960, Queens College; B.S.E.E., M.S., Ph.D., 1960, 1961, and 1963, respectively, Columbia University; AT&T Bell Laboratories, 1963—. Mr. Wyner has been doing research in various aspects of information and communication theory and related mathematical problems. He is presently Head of the Communications Analysis Research Department. He spent the year 1969–70 visiting the Department of Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel, and the Faculty of Electrical Engineering, at Technion, Haifa, Israel, on a Guggenheim Foundation Fellowship. He has also been a full- and part-time faculty member at Columbia University, Princeton University, and the Polytechnic Institute of Brooklyn. Chairman of the Metropolitan New York Chapter of the IEEE Information Theory Group, Associate Editor of the Group's Transactions, and co-chairperson of two international symposia. President, IEEE Information Theory Group, 1976. Fellow, IEEE Member, AAAS, Tau Beta Pi, Eta Kappa Nu, Sigma Xi. Since September 1983, he has been Editor-in-Chief of the IEEE Transactions on Information Theory.

Jacob Ziv, B.Sc., Dipl. Eng., and M.Sc., all in Electrical Engineering, from the Technion–Israel Institute of Technology 1954, 1955, and 1957, respectively; D.Sc., 1962, Massachusetts Institute of Technology; Senior Research Engineer in the Scientific Department, Israel Ministry of Defense, 1955–9; Applied Science Division of Melpar, 1961–62. In 1962 he returned to the Scientific Department, Israel Ministry of Defense, as Head of the Communications Division and was also an Adjunct of the Faculty of Electrical Engineering, Technion–Israel Institute of Technology. Member of the Technical Staff of Bell Laboratories, 1968–70. He joined the Technion in 1970 and is a Herman Gross Professor of Electrical Engineering. Dean of the Faculty of Electrical Engineering, 1974–76, and Vice President for Academic Affairs, 1978–82. Member Israeli Academy of Science, 1981; fellow, IEEE. From 1977 to 1978, and 1982 to 1983, he was on sabbatical leave at Bell Laboratories. His research interests include general topics in information theory and statistical communication.