Acknowledgment

The authors wish to acknowledge the advice and criticism of Prof. Toby Berger of Cornell University, Ithaca, NY.

References

[1] S. S. Lam, "Satellite packet communication—Multiple access protocols and performance," *IEEE Trans. Comm.*, vol. COM-27, pp. 1456–1466, Oct. 1979.

[2] F. Tobagi, "Multiaccess protocols in packet communication systems," *IEEE Trans. Comm.*, vol. COM-28, pp. 468–488, April 1980.

[3] J. I. Capetanakis, "Tree algorithms for packet broadcast channels," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 505–515, Sept. 1979.

[4] J. L. Massey, "Collision resolution algorithms and random-access communications," Tech. Rep., UCLA-ENG-8016, Apr. 1980.

[5] R. G. Gallager, "Conflict resolution in random access broadcast networks," in *Proc. of the AFOSR Workshop in Communications Theory and Applications*, Provincetown, MA, pp. 74–76, Sept. 17–20, 1978.

[6] B. S. Tsybakov and V. A. Mikhailov, "Slotted multiaccess packet-broadcasting feedback channel," *Probl. Peredach. Inform.*, vol. 14, no. 4, pp. 32–59, Oct.–Dec., 1978.

[7] G. Ruget, "Some channel-sharing algorithms," Lecture notes, Udine, Italy, 1979.

[8] T. Berger, "The Poisson multiple-access conflict resolution problems," in *Multi-User Communications*, G. Longo, Ed., CISM Courses and Lecture Series, no. 265. New York: Springer Verlag, 1981.

[9] J. C. Huang, "Pretested random tree protocols for packet communication systems," M.S. thesis, South China Institute of Technology, Guangzhou, China, Aug., 1981.

The Throughput Time Delay Function of an M / M / 1 Queue

AUREL A. LAZAR, MEMBER, IEEE

***Abstract*—A study of the design requirements of the data link control protocol leads us to consider a throughput time delay criterion as a performance measure for optimal flow control of computer communication networks. A closed network model consisting of a ·/ M / 1 queuing system (the receiver node) controlled by an exponential queuing system with variable rate (the source node) is considered. The flow control maximizing the throughput of the M / M / 1 queuing system in equilibrium under a bounded average time delay criterion is shown to be a window flow control mechanism. The window size L can be easily derived from the maximum tolerated average time delay T and the maximum offered load c. The dependence of the maximum throughput on the average time delay is also analyzed.**

I. Introduction

The throughput time delay trade-off as a performance criterion for optimal control of simple queuing systems (that are used as network models) has been recently studied. In investigating this trade-off Kleinrock [3], Bharath–Kumar [1], and Schwartz [7] have considered the "power" as an optimality criterion. This criterion, however, does not display the direct dependence between the maximum throughput and the average time delay. Schwartz mentions in [7] the problem of finding a state dependent control to a queuing system with an exponential server. The realizable control should maximize the throughput under a maximum constraint on the average time delay. This optimization criterion has, however, for the simple case of an M/M/1 queuing system a trivial solution (see Section II).

Manuscript received October 6, 1981; revised September 10, 1982. This work was supported by the National Science Foundation under Grants ECS-81-06476 and ECS-81-08776.

The author is with the Department of Electrical Engineering, Columbia University, New York, NY 10027.

The main idea behind the latter criterion is very "appealing" for designing efficient communication protocols. In the case of the data link protocol, for example, a source node is transmitting packets to an adjacent receiving node [9]. The flow control function of this protocol consists in dropping a packet if the receiving node is not ready [8]. Up to L packets (L is called the window size) may be transmitted without waiting for an acknowledgment. After a certain timeout, unacknowledged packets are retransmitted. Therefore, in order to design an efficient protocol it is imperative to have knowledge of the maximum time delay a packet requires to be transmitted and acknowledged under normal conditions.

In this correspondence an optimality criterion is introduced that takes into account these design constraints. The criterion displays the direct dependence between the throughput and the average time delay: the resulting optimal control maximizes the throughput under the condition that the average time delay does not exceed a preassigned value T. *In addition*, the offered load (or equivalently the maximum capacity in packets/s) in this optimization criterion is set to be a fixed parameter. Our investigations have shown that the optimal control of an M/M/1 system is a "window" flow control type mechanism. The window size L can be easily derived from the maximum tolerated average time delay T and offered load c.

The results obtained in the M/M/1 case suggest that our present views on the congestion control of computer communication networks need additional clarifications. It is a widely accepted in the computer communication networks literature that the total throughput of a network increases with the offered load from zero to a peak value and then decreases sharply to a low value (see, for example, [4] and the references therein). This rather qualitative assessment implies the necessity of flow control algorithms in computer networks. In finding the maximum attainable throughput by employing a flow control mechanism, however, not only the offered load but also the maximum time delay has to be taken into consideration. The main contribution of this correspondence is to show for the simple case of a M/M/1 system in what way the maximum obtainable throughput depends on the offered load and the average time delay.

This correspondence is organized as follows. In Section II the optimization criterion is defined. It is shown that maximizing the throughput under a maximum time delay criterion is equivalent in the M/M/1 case with minimizing the probability that the system is empty. In Section III the maximum time delay that can be achieved with L packets (maximum buffer size) in the M/M/1 system is found. It is shown that the maximum achievable time delay is strictly increasing with the number of packets. The optimum control that maximizes the throughput with respect to a bounded average time delay criterion is given in *Theorem 1*. The maximum obtainable throughput is also specified as a function of the existing time delay in the system. It is further shown that the so called throughput time delay function is continuously monotonically increasing and piecewise convex.

II. The Optimization Criterion

Let us consider the queuing system depicted in Fig. 1 in equilibrium. From a total of N packets of the closed system, k are assumed to be in the upper queue. The upper queue which has an exponential server μ has to be controlled according to a suitable optimality criterion. The lower or feedback queue has an exponential server (λ_k), $1 \leqslant k \leqslant N$. Since there are a maximum of N packets in the above system, the upper queue can be seen to be an M/M/1 queuing system with a finite buffer size N. Without any loss of generality, therefore, the feedback queue models the input stream to a M/M/1 system with a finite buffer size. The above setting represents a simple model for the data link flow control between two adjacent nodes in a computer network were

0018-9448/83/1100-0914$01.00

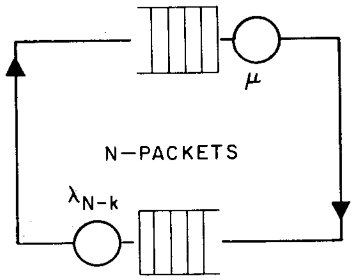

Fig. 1. Model for data link flow control.

each packet is acknowledged upon its arrival [9]. The upper queue models the receiving node and the lower queue the source node. The time delay incurred on the acknowledgments is assumed to be negligible. For modeling purposes it is also assumed that the source node has an infinitely large pool of packets.

The average throughput and the average time delay of the M/M/1 system (upper queue) are given by [2], [6]:

$$E\gamma_N = \mu \sum_{k=1}^{N} p_k$$

and

$$E\tau_N = \frac{\sum_{k=1}^{N} k p_k}{\mu \sum_{k=1}^{N} p_k},$$

respectively, where

$$p_k = \prod_{l=0}^{k-1} \frac{\lambda_{N-l}}{\mu} p_0$$

and

$$p_0 = \frac{1}{1 + \sum_{k=1}^{N} \prod_{l=0}^{k-1} \frac{\lambda_{N-l}}{\mu}}$$

denote the probabilities that the upper queue contains k packets, $1 \leqslant k \leqslant N$.

Definition 1: $\lambda = (\lambda_k)$, $1 \leqslant k \leqslant N$, will hereafter denote the control.

Following [5] we formally introduce now the class of admissible controls.

Definition 2: The class of controls $\lambda = (\lambda_k)$, $1 \leqslant k \leqslant N$, satisfying the peak constraint

$$0 \leqslant \lambda_k \leqslant c,$$

for all k, $1 \leqslant k \leqslant N$, where c, $c \in R_+$, is a constant is called admissible.

Remark: The following analysis shows the need for introducing the class of admissible controls. Let us assume that the rate of the lower queuing system in Fig. 1 is infinitely large. Since the lower queue is a "short circuit" in this case, packets leaving the upper queuing system rejoin instantaneously its queue. With N packets in the closed system the time delay amounts to N/μ and the throughput is equal with the maximum achievable value μ. By using only one packet in the system the time delay can be reduced to its minimum value $1/\mu$ and the value of the throughput remains the same.

Definition 3: The control $\lambda = (\lambda_k)$, $1 \leqslant k \leqslant N$, is said to be optimum over the class of admissible controls for a given $T, T \in R_+$, if the maximum

$$\max_{E\tau_N \leqslant T} E\gamma_N$$

is achieved.

Let x_k denote the expression

$$x_k = \prod_{l=0}^{k-1} \frac{\lambda_{N-l}}{\mu} \left(= \frac{p_k}{p_0} \right),$$

for all k, $1 \leqslant k \leqslant N$. Thus

$$\sum_{k=1}^{N} x_k = \sum_{k=1}^{N} \frac{p_k}{p_0} = \frac{1 - p_0}{p_0} = \frac{1}{p_0} - 1$$

and

$$p_0 = \frac{1}{1 + \sum_{k=1}^{N} x_k}.$$

Intuitively, maximizing the throughput by using a bounded time delay criterion is the same as minimizing the probability that the system is empty. This assertion is proven in the following lemma.

Lemma 1: $\lambda = (\lambda_k)$, $1 \leqslant k \leqslant N$, is optimum in the class of admissible controls for a given $T, T \in R_+$, if it achieves the maximum

$$\max_{\sum_{k=1}^{N} (k - \mu T) x_k \leqslant 0} \left\{ \sum_{k=1}^{N} x_k = \frac{1}{p_0} - 1 \right\}, \tag{1}$$

where $0 \leqslant x_k \leqslant (c/\mu)^k = \rho^k$, for all k, $1 \leqslant k \leqslant N$.

Proof: Since the condition $E\tau_N \leqslant T$ is equivalent with

$$\sum_{k=1}^{N} (k - \mu T) x_k \leqslant 0$$

and

$$E\gamma_N = p_0 \cdot \sum_{k=1}^{N} \mu x_k = \mu \left(1 - \frac{1}{1 + \sum_{k=1}^{N} x_k} \right) = \mu (1 - p_0)$$

the optimum control $\lambda = (\lambda_k)$, $1 \leqslant k \leqslant N$, satisfies (1). Q.E.D.

Definition 4: The mapping $F: R_+ \to R_+$ given by

$$F(T) = \max_{E\tau_N \leqslant T} E\gamma_N$$

is called the throughput time delay function.

III. Optimal Flow Control of the M/M/1 Queue

In order to simplify the proof of the main theorem, a series of necessary results will be first presented. In *Lemma 2* we derive the expression for the maximum time delay and an achievable upper bound to it. In *Lemma 3* it is shown that the maximum achievable time delay in an M/M/1 system is increasing with the number of packets.

Lemma 2: The maximum average time delay $T_{\max}^{(L)}$ achieved with L packets in the system is

$$T_{\max}^{(L)} = \frac{1}{\mu} \cdot \frac{\sum_{k=1}^{L} k \rho^k}{\sum_{k=1}^{L} \rho^k} < \frac{L}{\mu}. \tag{2}$$

Proof: Let us first show that

$$\frac{\sum_{k=1}^{L} k x_k}{\sum_{k=1}^{L} x_k} \leqslant \frac{\sum_{k=1}^{L} k \rho^k}{\sum_{k=1}^{L} \rho^k},$$

or

$$\sum_{k=1}^{L}\sum_{l=1}^{L} x_k \rho^l (k-l) \leqslant 0.$$

Due to the symmetry in k and l and the fact that

$$\frac{x_k}{x_l} \leqslant \rho^{k-l},$$

for all $k > l$, the expression on the left-hand side is negative. Note also that

$$T_{\max}^{(L)} < \frac{L}{\mu}$$

since

$$\sum_{k=1}^{L} (L-k)\rho^k > 0.$$

Q.E.D.

Remark: The right-hand side of inequality (2) can be achieved when $\rho \to \infty$. This corresponds to the situation when the served packets rejoin instantaneously the queue. L/μ is then exactly the average time delay of a M/M/1 queue having constantly L packets in the system. A simple equivalent expression for $T_{\max}^{(L)}$ is derived in the *corollary* to the *Theorem 1*.

Lemma 3: The maximum average time delay in the M/M/1 system is strictly increasing with the number of packets L, i.e.,

$$T_{\max}^{(L)} < T_{\max}^{(L+1)},$$

for all L, $L < N$.

Proof: We have to show that

$$\frac{\sum_{k=1}^{L} k\rho^k}{\sum_{k=1}^{L} \rho^k} < \frac{\sum_{k=1}^{L+1} k\rho^k}{\sum_{k=1}^{L+1} \rho^k},$$

or

$$\sum_{k=1}^{L} (L+1-k)\rho^k > 0.$$

The latter inequality is true since $k < L+1$. Q.E.D.

We can now give the following theorem.

Theorem 1: Given that $T_{\max}^{(L-1)} < T \leqslant T_{\max}^{(L)}$, $2 \leqslant L \leqslant N$, the optimum control of a M/M/1 queue with a maximum of N packets in the system is given by

$$\lambda_k = \begin{cases} c, & N-L+2 \leqslant k \leqslant N; \\ \dfrac{1}{L/\mu - T} \displaystyle\sum_{l=1}^{L-1} (\mu T - l)\rho^{l+1-L}, & k = N-L+1; \\ 0, & 1 \leqslant k \leqslant N-L. \end{cases}$$

Finally, the maximum throughput amounts to

$$F(T) = \mu \left(1 - \frac{1}{1 + \dfrac{1}{L-\mu T} \displaystyle\sum_{k=1}^{L-1} (L-k)\rho^k} \right).$$

Proof: Let us assume that

$$T_{\max}^{(L-1)} < T \leqslant T_{\max}^{(L)},$$

where $2 \leqslant L \leqslant N$. The condition

$$\sum_{k=1}^{N} (k-\mu T) x_k \leqslant 0,$$

can be written as

$$(L-\mu T)x_L \leqslant -\sum_{\substack{k=1 \\ k \neq L}}^{N} (k-\mu T)x_k$$

or since $L - \mu T > 0$

$$x_L \leqslant -\frac{1}{L-\mu T} \sum_{\substack{k=1 \\ k \neq L}}^{N} (k-\mu T)x_k.$$

Thus

$$\begin{aligned} \sum_{k=1}^{N} x_k &\leqslant -\frac{1}{L-\mu T} \sum_{k=L+1}^{N} (k-L)x_k \\ &\quad + \frac{1}{L-\mu T} \sum_{k=1}^{L-1} (L-k)x_k \\ &\leqslant \frac{1}{L-\mu T} \sum_{k=1}^{L-1} (L-k)\rho^k. \end{aligned}$$

Equality can be achieved if and only if

$$x_k = \begin{cases} 0, & L+1 \leqslant k \leqslant N; \\ \dfrac{1}{L-\mu T} \displaystyle\sum_{l=1}^{L-1} (\mu T - l)\rho^l, & k = L; \\ \rho^k, & 1 \leqslant k \leqslant L-1. \end{cases}$$

It remains only to show that

$$0 < x_L \leqslant \rho^L,$$

or

$$0 < \sum_{k=1}^{L-1} (\mu T - k)\rho^k \leqslant (L-\mu T)\rho^L,$$

which is equivalent to

$$\frac{1}{\mu} \cdot \frac{\sum_{k=1}^{L-1} k\rho^k}{\sum_{k=1}^{L-1} \rho^k} < T \leqslant \frac{1}{\mu} \cdot \frac{\sum_{k=1}^{L} k\rho^k}{\sum_{k=1}^{L} \rho^k},$$

or

$$T_{\max}^{(L-1)} < T \leqslant T_{\max}^{(L)}.$$

Q.E.D.

Remark: Since the lower queue will contain at all times at least $N-L$ packets, the optimum control can also be achieved by a control scheme using a total of $N = L$ packets. We require that

$$\lambda_k = \begin{cases} c, & 2 \leqslant k \leqslant L; \\ \dfrac{1}{L/\mu - T} \displaystyle\sum_{l=1}^{L-1} (\mu T - l)\rho^{l+1-L}, & k = 1. \end{cases}$$

Therefore the optimal control is a window type mechanism. The window size L, the maximum number of unacknowledged packets, can be easily derived from the maximum time delay of the system. Naturally the number of packets will depend on the

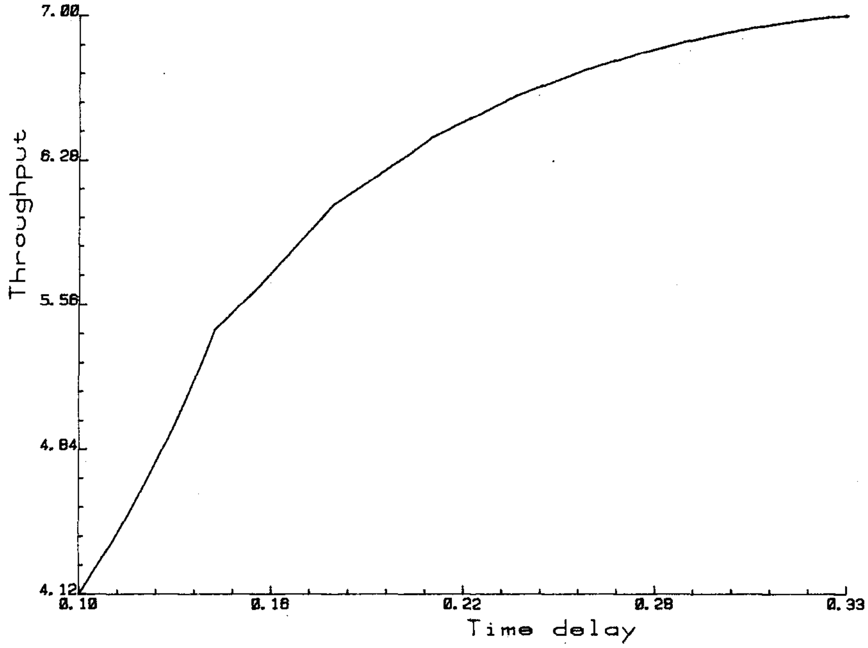

Fig. 2. Throughput time delay function ($\mu = 10$, $c = 7$).

maximum offered load (or line capacity into the system). The result obtained is known to be commonly implemented in practice [9].

Corollary: If $T = T_{\max}^{(L)}$ then

$$F\left(T_{\max}^{(L)}\right) = c \cdot \frac{1-\rho^{L}}{1-\rho^{L+1}}$$

and

$$T_{\max}^{(L)} = \frac{1}{\mu}\left(\frac{1}{1-\rho} + \frac{L}{1-\rho^{-L}}\right).$$

Proof: From $T = T_{\max}^{(L)}$ we have

$$\sum_{k=1}^{L-1} (\mu T - k)\rho^{k} = (L - \mu T)\rho^{L}$$

and therefore

$$\lambda_1 = \frac{\mu}{L-\mu T}\rho^{1-L} \sum_{k=1}^{L-1} (\mu T - k)\rho^{k} = c.$$

The corresponding throughput amounts to

$$F\left(T_{\max}^{(L)}\right) = \mu\left(1 - \frac{1}{1 + \sum_{k=1}^{L} \rho^{k}}\right),$$

or

$$F\left(T_{\max}^{(L)}\right) = c \cdot \frac{1-\rho^{L}}{1-\rho^{L+1}}.$$

The computation of the time delay can be similarly done. Q.E.D.

Remark: One can easily see that

$$\lim_{L\to\infty} F\left(T_{\max}^{(L)}\right) = \begin{cases} c, & \rho < 1; \\ \mu, & \rho \geq 1; \end{cases}$$

and

$$\lim_{L\to\infty} T_{\max}^{(L)} = \begin{cases} \dfrac{1}{\mu - c}, & \rho < 1; \\ \infty, & \rho \geq 1. \end{cases}$$

Lemma 4: The throughput time delay function is continuous nondecreasing on R_{+}. Furthermore, for all T, $T_{\max}^{(L-1)} < T \leq T_{\max}^{(L)}$, $L \leq N$, F is convex.

Proof: The proof can be easily supplied. Q.E.D.

For $\mu = 10$ and $c = 7$ the throughput time delay function is depicted in Fig. 2. Fig. 3 depicts the parametrized throughput time delay function for various values of c. The proof of the "overall concave" behavior of the throughput time delay function is left to the reader as a straightforward exercise.

IV. Conclusion

In this correspondence we have shown that the control which maximizes the throughput of a M/M/1 queuing system under a bounded time delay and maximum-offered load criterion is state dependent. The number of packets that when optimally controlled lead to the maximum throughput has been shown to be easily derivable from the preassigned upper bound on the average time delay T and the maximum admissible load c. The dependence of the maximum throughput on the average time delay and the load c has also beeen investigated. The analysis of the M/M/1 system has shown that it is appropriate to discuss congestion problems in terms of maximum throughput under a bounded average time delay criterion. The offered load then appears only as a parameter. The results obtained for the data link control model adopted here are currently implemented in practice.

Acknowledgment

The author would like to thank Prof. Mischa Schwartz for helpful discussions on this and related topics.

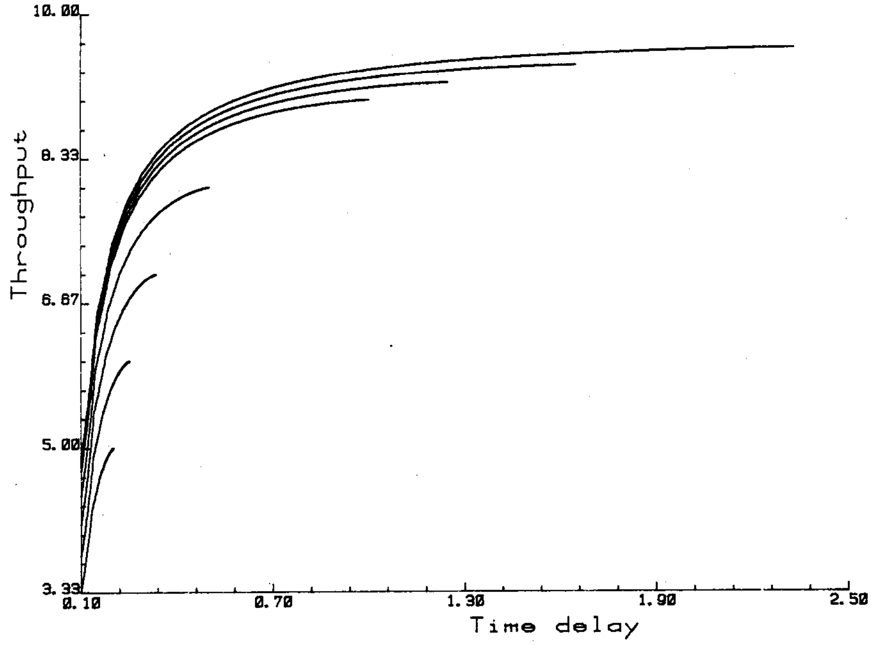

Fig. 3. Dependence of throughput time delay function on parameter c ($\mu = 10$).

References

[1] K. Bharath-Kumar, "Optimum end-to-end flow control in networks," *Conf. Rec. Int. Conf. Commun.*, pp. 23.3.1–32.3.6, June 1980.

[2] L. Kleinrock, *Queueing Systems, Volume I: Theory*. New York: John Wiley, 1975, pp. 147–160.

[3] ——, "Power and deterministic rules of thumb for probabilistic problems in computer communications," *Conf. Rec. Int. Conf. Commun.*, pp. 43.1.1–43.1.10, June 1979.

[4] S. S. Lam and M. Reiser, "Congestion control of store-and-forward networks by input buffer limits—An analysis," *IEEE Trans. Commun.*, vol. COM-27, no. 1, Jan. 1979.

[5] A. A. Lazar, Optimal information processing using counting point process observations," Ph.D. dissertation, Dept. of Electrical Engineering and Computer Science, Princeton University, Princeton, NJ, Oct. 1980.

[6] M. Schwartz, *Computer Communication Networks Design and Analysis*. Englewood Cliffs, NJ: Prentice Hall, 1976, pp. 242–258.

[7] ——, "Routing and flow control in data networks," IBM Res. Rep. RC 8353 (#36319), IBM T. J. Watson Research Center, Yorktown Heights, NY, 1980.

[8] M. Gerla, "Routing and flow control," in *Protocols and Techniques for Data Communication Networks*, F. F. Kuo, Ed. Englewood Cliffs, NJ: Prentice Hall, 1981, pp. 122–174.

[9] A. S. Tanenbaum, *Computer Networks*. Englewood Cliffs, NJ: Prentice Hall, 1981, pp. 137–164.

A Source Coding Problem for Sources with Additional Outputs to Keep Secret from the Receiver or Wiretappers

HIROSUKE YAMAMOTO, MEMBER, IEEE

***Abstract*—A new source coding problem is considered for a one-way communication system with correlated source outputs $\{XY\}$. One of the source outputs, i.e., $\{X\}$, must be transmitted to the receiver within a prescribed distortion tolerance as in ordinary source coding. On the other hand, the other source output, i.e., $\{Y\}$, has to be kept as secret as possible from the receiver or wiretappers. For this case the equivocation-distortion function $\Gamma^*(d)$ and the rate-distortion-equivocation function $R^*(d, e)$ are defined and evaluated. The former is the maximum achievable equivocation of $\{Y\}$ under the distortion tolerance d for $\{X\}$, and the latter is the minimum rate necessary to attain both the equivocation tolerance e for $\{Y\}$ and the distortion tolerance d for $\{X\}$. Some examples are included.**

Manuscript received December 28, 1981; revised March 9, 1983. This work was supported in part by the Sakkokai Foundation, Japan. This work was presented at the Fourth Symposium on Information and Its Applications, Kashikojima, Japan, November 1981.

The author is with the Department of Electronic Engineering, Tokushima University, 2-1 Minami-josanjima, Tokushima, 770 Japan.

I. Introduction

Let us consider a new kind of source coding problem. The system we treat is the usual one-way communication system with correlated source outputs $\{XY\}$ as in Fig. 1. Now let us assume that the proprietor of the system is an information service company and a customer wants to obtain information about $\{X\}$. When the customer pays the charge for $\{X\}$, the company has to supply the sequence $\{\hat{X}\}$ for him within a prescribed distortion tolerance. This situation corresponds to ordinary source coding. However, since $\{X\}$ and $\{Y\}$ are correlated, the customer can estimate $\{Y\}$ from $\{\hat{X}\}$ with some accuracy. Hence the company needs to keep the information about $\{Y\}$ as secret as possible from the customer, as well as from wiretappers, because the charge has been paid only for $\{X\}$. In this case, the code used in the system should keep $\{Y\}$ secret from one receiving the reproduction sequence $\{\hat{X}\}$ in addition to satisfying the distortion tolerance. The capability of keeping $\{Y\}$ secret may be evaluated by the equivocation function $H(Y|W)$, where $W = F_E(XY)$ is the encoder output for $X = (X_1, X_2, \cdots, X_n)$ and $Y = (Y_1, Y_2, \cdots, Y_n)$. It is this situation that is studied in this paper. That is, we investigate the maximum achievable equivocation and the minimum rate necessary to attain the prescribed distortion and equivocation tolerances.

Since the equivocation can be regarded as a kind of distortion measure, the above problem may be considered as a source coding problem with multiple constraints [1], [2]. However, since the equivocation is neither a per-letter distortion measure nor a weighted-average distortion measure, we cannot quote their results.

In Section II a formal and precise statement of the problem and results are given. Section III contains some examples. All the theorems and lemmas are proved in the Appendix.