notion of a neighborhood set of a sequence, and the degree γ to which the length of a codeword approaches its minimal value K_Ψ. The concern with individual messages is perhaps best founded with respect to sources for which we do not entertain a relative-frequency characterization of uncertainty.

References

[1] R. J. Solomonoff, "Formal theory of inductive inference, I," *Inform. Contr.*, vol. 7, pp. 1–22, 1964.
[2] G. J. Chaitin, "Information-theoretic computational complexity," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 10–15, Jan. 1974.
[3] ——, "Information-theoretic limitations of formal systems," *J. Ass. Comput. Mach.*, vol. 21, pp. 403–424, July 1974.
[4] A. N. Kolmogorov, "Three approaches to the quantitative definition of information," *Probl. Inform. Transmission* (Russian), vol. 1, pp. 4–7, 1965.
[5] ——, "Logical basis for information theory and probability theory," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 662–664, Sept. 1968.
[6] T. L. Fine, *Theories of Probability: An Examination of Foundations.* New York: Academic, 1973, ch. V.
[7] J. Hartmanis and J. Hopcroft, "An overview of the theory of computational complexity," *J. Ass. Comput. Mach.*, vol. 18, pp. 444–475, 1971.
[8] H. Rogers, Jr., *Theory of Recursive Functions and Effective Computability.* New York: McGraw-Hill, 1967.
[9] L. L. Campbell, "Kraft inequality for decoding with respect to a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 68–73, Jan. 1973.

On Tree Coding with a Fidelity Criterion

CHARLES R. DAVIS, STUDENT MEMBER, IEEE, AND MARTIN E. HELLMAN, MEMBER, IEEE

***Abstract*—This paper reexamines Jelinek's proof that tree codes can be used to approach the rate-distortion bound. It is shown that the branching process used in Jelinek's proof is not a (strict-sense) branching process (SSBP) when the source is asymmetric. Branching processes with random environments (BPWRE) are introduced and used to extend the proof to general discrete-time memoryless sources. The theory developed indicates why a particular metric used in experiments performed better than another suggested by the original proof.**

I. Introduction

The existence of random tree codes, suitable for encoding discrete-time sources with independent identically distributed (i.i.d.) outputs with respect to a distortion measure was first proved by Jelinek [1] for a fairly general class of source distributions. More recently, algorithms have been developed by Anderson and Jelinek [2] and Gallager [3] that further demonstrate the applicability of random tree codes. Their results, however, apply only to a class of sources and distortion measures that are symmetric (i.e., the distribution on source letters is uniform, and the columns of the distortion matrix can be partitioned in such a way that within each partitioning all rows are permutations of one another and all columns are permutations of one another).

Recent work by Viterbi and Omura [4] has demonstrated the applicability of time-varying trellis codes to this encoding problem, and their results constitute an alternative proof of the tree coding theorem.

Manuscript received November 11, 1974; revised January 20, 1975. This work was supported in part by the U.S. Air Force Office of Scientific Research under Contract F44620-73-C-0065 and in part by the Joint Services Electronics Program (U.S. Army, Navy, and Air Force) under Contract N00014-67-A-0044.
The authors are with the Department of Electrical Engineering, Stanford University, Stanford, Calif. 94305.

In the original proof of the tree coding theorem by Jelinek [1] the existence of suitable tree codes is dependent on the probability of extinction of a related branching process being strictly less than one. As will be presently demonstrated, this branching process has probability of extinction strictly less than one only for symmetric sources previously defined. However, with a modification to the branching process and a reevaluation of the extinction criteria, it will be possible to reestablish the tree coding theorem.

The remainder of the paper is organized as follows. Section II contains some rate-distortion relationships we will need and some results concerning branching processes. In Section III several problems concerning the extinction criteria of a branching process are discussed along with an example. In Section IV the tree coding theorem is proved using the correct extinction criteria, and Section V is a summary.

II. Preliminaries

A. Coding with Respect to Fidelity Criterion

The minimum rate necessary to encode an i.i.d. information source with additive single letter distortion measure so that the average distortion is no more than some target distortion D_0 is given by [6]

$$R(D_0) = \inf I(Z;Y) \tag{1}$$

where $I(Z;Y)$ is the mutual information between a source output Z and a reproduction symbol Y. The source outputs have probability density $q(z)$, and the infimum is over all conditional probability assignments $w(y \mid z)$ such that

$$\int q(z)\, dz \int w(y \mid z)\, d(z,y)\, dy \le D_0. \tag{2}$$

We shall assume that the single letter distortion is bounded above by $D_m < \infty$. Although we write integrals for expectations we can include discrete and mixed distributions merely by using Lebesque–Stieltjes integrals. For discrete distributions this is equivalent to replacing integrals by sums.

The following equivalent definition of the rate-distortion function is readily derived from Haskell [5]

$$R(D_0) = \max_{\rho \geq 0} \min_{w(y)} - \int q(z)\, dz \log \int w(y) \cdot \exp\left[-\rho(d(z,y) - D_0)\right] dy \quad (3)$$

where $w(y)$ is a probability density defined on the reproduction alphabet. Assume that the optimizing values of ρ and $w(y)$ in (3) are known and given by ρ_0 and $w_0(y)$, respectively. ρ_0 and $w_0(y)$ depend implicitly on D_0. Letting $w_0(y \mid z)$ denote the optimal "test channel" we then have the following relations (see [6, pp. 37, 90]):

$$w_0(y \mid z) = w_0(y) \exp\left[-\rho_0\, d(z,y)\right]\lambda(z) \quad (4)$$

$$[\lambda(z)]^{-1} = \int w_0(y) \exp\left[-\rho_0\, d(z,y)\right] dy \quad (5)$$

$$w_0(y) = \int w_0(y \mid z) q(z)\, dz \quad (6)$$

$$R(D_0) = -\rho_0 D + \int q(z) \log \lambda(z)\, dz. \quad (7)$$

We denote a sequence of source letters of length l by z^l, a like sequence from the reproduction alphabet by y^l, and cumulative distortion between the two by $d(z^l,y^l)$.

B. Branching Processes

A generalized branching process may be defined as follows. Beginning with a single "particle" at discrete time $k = 0$, let each particle give rise to new particles (offspring) with probability distribution $p_k(v) = \Pr$ (a single particle at time k gives rise to v new particles). The total number of particles given birth to at time $k + 1$, $N(k + 1)$ is termed the generation size at time $k + 1$. After a parent particle gives rise to its offspring, it is no longer counted. If, for some k, none of the $N(k)$ parent particles given rise to offspring, the branching process is said to become extinct. Then $N(k + t) = 0$, for $t \geq 1$. Given a generation time k, each parent particle is assumed to give rise to offspring independently of other parent particles both in that generation and otherwise. Note that here we allow the offspring distribution $p_k(v)$ to vary from generation to generation, but require that within a generation each parent particle give rise to offspring according to the same distribution $p_k(v)$. The reason for this more general model will soon become apparent.

If the generalized branching process has the same distribution $p_k(v)$ for all generations, the process is a strict-sense branching process (SSBP). This is the process used in [1] to establish the tree coding theorem. A standard result from the theory of SSBP [7]–[10] concerns ϕ, the probability of eventual extinction of the process. Letting m denote the expected number of offspring from a single particle we have the following.

Theorem 1: For the SSBP defined previously

$$\phi = 1, \qquad \text{if } m \leq 1 \quad (8)$$

$$\phi < 1, \qquad \text{if } m > 1. \quad (9)$$

Thus we see that if the mean number of offspring per parent particle is strictly greater than one, there is a nonzero probability that the process survives forever.

Now let us allow a slight generalization in the process. We will consider the $p_k(v)$ to be time varying according to an underlying i.i.d. stochastic process $\{\xi_k\}_{k=0}^{\infty}$, where $\xi_k \in \Xi$, and Ξ is the space of possible environments for each generation, and $p_k(v) = \mathscr{P}_{\xi_k}(v)$, where $\mathscr{P}_\xi(v)$ is a collection of distributions for $\xi \in \Xi$. Thus at each generation, a distribution $\mathscr{P}_\xi(v)$ is picked according to an underlying distribution $f(\xi)$ and for that generation all parent particles give rise to offspring according to the $\mathscr{P}_\xi(v)$ that was chosen. Conditioned on the environments, the number of offspring are independent both within a generation and from generation to generation. Such a process is called a branching process with random environments (BPWRE) and was first studied by Smith and Wilkinson [11]. The random environments, of course, refer to the randomization of the offspring distribution from generation to generation.

To establish the optimality of tree codes we shall need the extinction criteria for a BPWRE. Let m_ξ be the expected number of offspring due from a parent particle given environment ξ. Let $\sigma_\xi = \mathscr{P}_\xi(0)$ be the probability that a parent particle has no offspring, given the environment ξ. Then as shown by Smith and Wilkinson [11], the following theorem governs the probability of extinction. We denote expectation with respect to $f(\xi)$ by E_ξ.

Theorem 2: Let the probability of eventual extinction be ϕ, and assume $E_\xi|\log m_\xi| < \infty$. Then $\phi = 1$ if

$$E_\xi \log (m_\xi) \leq 0 \quad (10)$$

and $\phi < 1$ if

$$E_\xi \log (m_\xi) > 0 \quad (11)$$

and in addition

$$E_\xi|\log (1 - \sigma_\xi)| < \infty. \quad (12)$$

We see that for a BPWRE we are mostly concerned with whether or not the geometric mean $\exp\{E_\xi \log (m_\xi)\}$ is greater than one. The second criterion (12), known as the catastrosphe criterion, requires that the probability of almost catastrophic environments (which cause all parent particles to produce no offspring with high probability) be "small." We shall show that except when the source is symmetric the branching process used by Jelinek [1] is not a SSBP but is instead a BPWRE. Consequently the extinction criteria used in [1] is valid only for symmetric sources, and the proof in [1] applies only to symmetric sources.

III. Tree Codes and Branching Processes

In Jelinek's paper [1] a random tree code of rate $R \triangleq \ln(g)/n_0 \geq R(D_0)$ is constructed with g branches emanating from each node. Along each branch are placed n_0 symbols, drawn at random from the reproduction alphabet according to the density $w_0(y)$. The object is to find a path through the tree that has a time averaged distortion close to D_0.

Associated with such a tree code, Jelinek defines the following stochastic process, which we will call the tree coding process. Beginning at the origin node, extend all paths to a depth of l symbols. Keep as survivors all paths whose associated length l reproduction sequences are in the set $S(\delta, z_1{}^l)$, where $z_1{}^l$ represents the first l source outputs, and $S(\delta, z^l)$ is defined by

$$S(\delta,z^l) \triangleq \left\{ y^l: d(z^l,y^l) \leq l(D_0 + \delta), \log \frac{w_0(y^l \mid z^l)}{w_0(y^l)} \leq l(R(D_0) + \tfrac{2}{3}\gamma) \right\} \quad (13)$$

where

$$\gamma \triangleq R - R(D_0). \quad (14)$$

This constitutes the first cycle in the process. Now extend all paths that survived the first cycle to a further depth l and compare with the second l source symbols, keeping those in $S(\delta, z_2{}^l)$. This is the second cycle of the process. Similar cycles are continued unless and until the process terminates due to no paths in a cycle being extended.

Note that the process may never terminate, and we shall be vitally interested in the probability of this event. Also note that any path that survives satisfies $d(z^l,y^l) \leq l(D_0 + \delta)$ along each of its length l intervals. Hence its time average distortion is at most $D_0 + \delta$. Let us further adopt the convention that should the process terminate we start a new process by extending some path in the terminating cycle that is an extension of a node that survived all previous cycles. Starting from this node we are dealing with a new tree coding process. If it terminates we start yet another, etc.

When a process terminates we incur distortion at most lD_m for the terminating cycle and at most $l(D_0 + \delta)$ for each preceding cycle (D_m is the maximum per letter distortion). Should we, after M such restarts, start a process that survives forever we have, for N overall cycles, an average overall distortion of at most

$$\frac{l}{Nl}[MlD_m + (N - M)l(D_0 + \delta)].$$

Clearly, as N becomes large, the average distortion will tend to $D_0 + \delta$, if $E(M)$ is finite. As the following theorem shows, $E(M)$ will be finite if ϕ, the probability of eventual termination of the process, is strictly less than one.

Theorem 3: Let ϕ be the probability of eventual termination of the tree coding process. Let the process be restarted whenever it becomes extinct. Let M represent the number of times we restart the process. If $\phi < 1$, then $E(M) < \infty$.

Proof: Since the termination of a process that terminates at cycle k is independent of the number of potential offspring in cycles $k + 1$ and beyond, the probability of termination of a restarted branching process is independent of the preceding processes. Therefore,

$$\Pr(M \text{ restarts followed by survival}) = \phi^M(1 - \phi) \quad (15)$$

$$E(M) = \sum_{i=0} i\phi^i(1 - \phi) = \frac{\phi}{1 - \phi} \quad (16)$$

which is finite if $\phi < 1$.

Thus in order to establish the optimality of random tree codes for source coding we need only demonstrate that the probability of termination of the tree coding process is strictly less than one.

We now associate the probability of termination of this process with the probability of extinction of a related branching process defined as follows. Each cycle is considered to be a generation. Each starting node in a cycle is a parent particle and each surviving path a descendant. Clearly the probability of termination of the tree coding process and the probability of extinction of the branching process are one and the same. We now are concerned with finding criteria for which this branching process has probability of extinction strictly less than one.

In [1] the noncertain extinction of the branching process is established by demonstrating that the expected number of offspring, averaged over all source and reproduction alphabet sequences, is greater than one for some sufficiently large l. From Section II we know that this establishes noncertain extinction if the branching process is a SSBP. However, except for symmetric sources, the branching process is not a SSBP. This is because in a SSBP the distribution on offspring from a parent particle remains the same from generation to generation. However, note that in the tree coding process, at a given cycle or generation, all paths are compared to the same source sequence. If that source sequence is an atypical one, then all paths in that generation are adversely affected (are less likely to encode within $D_0 + \delta$). Since l is finite, atypical source sequences will occur with probability one (except for symmetric sources, where all sequences are typical in this sense). This dependence of the offspring distribution on the source sequence thus violates one of the hypotheses of a SSBP. Since the source outputs are independent, the offspring distributions vary from generation to generation in an i.i.d. fashion, and we have a BPWRE. The environments process is simply the sequence of source outputs taken l at a time. We, therefore, can identify $\xi = z^l$.

It would be nice at this point if we could simply apply the BPWRE survival criteria to the tree coding process. If this were to yield a probability of extinction less than one then the tree coding theorem would be proved. Unfortunately, the discard criterion $S(\delta,z^l)$ is too severe in general. This is because the two conditions for y^l to be in $S(\delta,z^l)$ place upper and lower bounds on $d(z^l,y^l)$. For asymmetric sources and certain values of z^l these bounds may define $S(\delta,z^l)$ to be the null set (i.e., the upper bound is smaller than the lower bound). Whenever such a z^l occurs the tree coding process definitely terminates, and since such a z^l will eventually occur with probability one it follows that $\phi = 1$. However, as we shall show, a modified discard criterion results in $\phi < 1$ and allows a proof of the optimality of tree codes. The following example demonstrates that $S(\delta,z^l)$ can be empty.

Example: Let the source be binary with $p = \Pr(z = 1) = 0.1$ and $q = \Pr(z = 0) = 0.9$. Take $d(z,y)$ to be the Hamming distortion measure (zero if $z = y$, one if $z \neq y$) and take $D_0 = 0.05$. Then from Berger [6, p. 37] we have

$$w_0(1) = (p - D_0)/(1 - 2D_0) = 0.05556 \quad (17a)$$

$$w_0(0) = (q - D_0)/(1 - 2D_0) = 0.94444 \quad (17b)$$

$$\exp(-\rho_0) = D_0/(1 - D_0) = 0.05263 \quad (17c)$$

$$\rho_0 = -\log(0.05263) = 2.94444 \quad (17d)$$

$$R(D_0) = H(p) - H(D_0) = 0.12657 \text{ nats.} \quad (17e)$$

For y^l to be in $S(\delta,z^l)$ (13) requires both

$$d(z^l,y^l) \leq l(D_0 + \delta) \quad (18)$$

and

$$\log[w_0(y^l \mid z^l)/w_0(y^l)] \leq l[R(D_0) + \tfrac{2}{3}\gamma]. \quad (19)$$

The upper bound on $d(z^l,y^l)$, for $y^l \in S(\delta,z^l)$, is easily evaluated

$$d_{\max} = l(D_0 + \delta) = l(0.05 + \delta). \quad (20)$$

The lower bound follows from the second condition (19) and extensions of (4) and (5)

$$w_0(y^l \mid z^l) = w_0(y^l) \exp[-\rho_0\, d(z^l,y^l)]\lambda(z^l) \quad (21)$$

$$[\lambda(z^l)]^{-1} = \int w_0(y^l) \exp[-\rho_0\, d(z^l,y^l)]\, dy^l. \quad (22)$$

Taken together (19) and (21) yield

$$d(z^l,y^l) \geq (1/\rho_0)[\ln \lambda(z^l) - lR(D_0) - (2/3)\gamma l] \quad (23)$$

as the lower bound on $d(z^l,y^l)$, for $y^l \in S(\delta,z^l)$. Entering numerical values (17) in (22) yields for $z^l = \mathbf{1}^l$, the sequence of l ones,

$$[\lambda(\mathbf{1}^l)]^{-1} = [w_0(1) + w_0(0)\exp(-\rho_0)]^l = (0.10526)^l \quad (24)$$

so that the lower bound on $d(z^l,y^l)$, for $z^l = \mathbf{1}^l$, is

$$d_{\min}(\mathbf{1}^l) = l[0.72161 - 0.22642\gamma]. \quad (25)$$

Therefore, $S(\delta,\mathbf{1}^l)$ is empty if

$$d_{\min}(\mathbf{1}^l) \geq d_{\max} \quad (26)$$

or equivalently

$$l(0.72161 - 0.22642\gamma) \geq l(0.05 + 8) \quad (27)$$

or

$$0.67161 \geq \delta + 0.22642\gamma. \quad (28)$$

It is easily seen that even for moderate values of $\gamma = R - R(D_0)$ and δ = target distortion $- D_0$, the set $S(\delta,\mathbf{1}^l)$ is empty. Thus whenever the source outputs l ones in a row during a cycle time the tree coding process terminates. Also since

$$\Pr(z^l = \mathbf{1}^l) = (0.1)^l \quad (29)$$

this happens in a finite time with probability one. Indeed $E(T) = 10^l$.

The preceding example demonstrates the impossibility of approaching the $R(D)$ curve using tree codes with the $S(\delta,z^l)$ discard criterion. It is not difficult to show that a similar problem arises whenever the binary source is asymmetric. With the extinction criteria for a BPWRE in mind we shall define a new discard criterion and proceed to prove the tree coding theorem.

IV. Proof of Tree Coding Theorem

In order to prove that random tree codes are suitable for source encoding with respect to a fidelity criterion at rates and average distortions arbitrarily close to the $R(D)$ curve, it suffices to show that the associated branching process has probability of extinction strictly less than one. However, as the example in Section III illustrates, the discard criterion $S(\delta,z^l)$ is, in general, too stringent to ensure this. Thus we define a new discard criterion $S'(\delta,z^l)$ by

$$S'(\delta,z^l) \triangleq \{y^l: d(z^l,y^l) \leq D(z^l) + l\delta\} \quad (30)$$

where

$$D(z^l) \triangleq \int d(z^l,y^l) w_0(y^l \mid z^l)\, dy^l. \quad (31)$$

Note that $\int D(z^l)q(z^l)\, dz^l \triangleq lD_0$, so that if the branching process survives then by the law of large numbers the average target distortion $(1/Nl)\sum_{k=1}^{N} D(z_k^l) + \delta$ will tend to a value less than $D_0 + \delta + \Delta$, for any $\Delta > 0$, as the number of depth l encoding cycles $N \to \infty$. For probability of extinction $\phi < 1$ the average additional distortion due to restarting is upperbounded by $E(M)D_m/N$, which tends to zero as $N \to \infty$. Therefore, proving that $\phi < 1$, for the BPWRE associated with the new discard criterion, will establish the existence of random tree codes with rates arbitrarily close to the $R(D)$ curve. Also since $E(M)$ is the expected number of restarts, averaged over all codes and possible source sequences, there must exist at least one deterministic code for which the expected number of restarts, averaged over all source sequences, is no greater than $E(M)$. The encoding algorithm used for the random code can, therefore, also be used with such a deterministic code. First let us establish the following lemma that will prove useful in the proof of the main theorem.

Lemma 1: Define

$$P(z^l) = \int_{S'(\delta,z^l)} w_0(y^l)\, dy^l \quad (32)$$

which is implicitly a function of δ. Then

$$P(z^l) \geq [\lambda(z^l)]^{-1}(1 - D_m{}^2/\delta^2 l) \exp[\rho_0 D(z^l) - \rho_0 l\delta]. \quad (33)$$

Proof: Define

$$S''(\delta,z^l) = \{y^l: |d(z^l,y^l) - D(z^l)| \leq l\delta\} \quad (34)$$

and note that $S''(\delta,z^l) \subseteq S'(\delta,z^l)$. Therefore,

$$P(z^l) \le \int_{S''} w_0(y^l)\, dy^l \tag{35}$$

where for notational convenience we use $S'' = S''(\delta,z^l)$ and $S' = S'(\delta,z^l)$. Then using (4) or (21) we obtain

$$P(z^l) \ge [\lambda(z^l)]^{-1} \int_{S''} w_0(y^l \mid z^l) \exp[\rho_0 d(z^l,y^l)]\, dy^l \tag{36}$$

$$= [\lambda(z^l)]^{-1} \exp[\rho_0 D(z^l)] \int_{S''} w_0(y^l \mid z^l) \cdot \exp\{\rho_0[d(z^l,y^l) - D(z^l)]\}\, dy^l \tag{37}$$

$$\ge [\lambda(z^l)]^{-1} \exp[\rho_0 D(z^l)] \cdot \exp(-\rho_0 l\delta) \int_{S''} w_0(y^l \mid z^l)\, dy^l \tag{38}$$

where the last step follows from (34). Defining

$$P_{y\mid z}(A) = \int_A w_0(y^l \mid z^l)\, dy^l \tag{39}$$

we have

$$P(z^l) \ge [\lambda(z^l)]^{-1} \exp(\rho_0 D(z^l) - \rho_0 l\delta) P_{y|z}(S'') \tag{40}$$

where $P_{y|z}$ depends implicitly on z^l. Note that

$$E_{y\mid z} d(z^l,y^l) = D(z^l) = \sum_{i=1}^{l} D(z_i) \tag{41}$$

and

$$\operatorname{var}_{y\mid z} d(z^l,y^l) = \sum_{i=1}^{l} \operatorname{var} d(z_i,y_i) \le l D_m^2 \tag{42}$$

since $d(z_i,y_i) \in [0,D_m]$ and the $\{d(z_i,y_i)\}_{i=1}^{l}$ are independent when conditioned on z^l. Therefore, from (34) and Chebyshev's inequality

$$1 - P_{y\mid z}(S'') \le \frac{\operatorname{var}_{y\mid z} d(z^l,y^l)}{l^2\delta^2} \le \frac{D_m^2}{l\delta^2} \tag{43}$$

and for all z^l

$$P_{y|z}(S'') \ge 1 - (D_m^2/\delta^2 l). \tag{44}$$

Combining (40) and (44) we obtain (33) which completes the proof of the lemma and allows us to proceed to the following crucial theorem.

Theorem 4: Let an i.i.d. source have marginal density $q(z)$, let D_0 be a desired average distortion and let $d(z,y)$ be a single letter additive distortion measure bounded above by D_m (and below by zero). Then for any $\gamma > 0$ there exists a tree code of rate $R = R(D_0) + \gamma$, an $l_0 < \infty$, and a $\delta_0 > 0$ such that for $l > l_0$ and $0 < \delta < \delta_0$ the associated BPWRE induced by the tree coding process with discard criteria $S'(\delta,z^l)$ has probability of extinction ϕ strictly less than one.

Proof: Remember that we can identify $\xi_k = z_k{}^l$, the kth block of source outputs taken l at a time. Now define

$$m(z^l) = E\begin{pmatrix}\text{number of nodes } y^l \text{ at depth } l \text{ that}\\ \text{are in } S'(\delta,z^l)\end{pmatrix} \tag{45}$$

and

$$\sigma(z^l) = \Pr\begin{pmatrix}\text{none of the } g^{l/n_0} \text{ nodes at depth } l\\ \text{are in } S'(\delta,z^l)\end{pmatrix} \tag{46}$$

where the expectation and probability are over the ensemble of random tree codes of depth l and rate $R = \ln(g)/n_0$, and with symbols drawn i.i.d. according to $w_0(y)$. Then the conditions of Theorem 2 that ensure $\phi < 1$ can be written as

$$\int q(z^l)|\log m(z^l)|\, dz^l < \infty \tag{47}$$

$$\int q(z^l) \log[m(z^l)]\, dz^l > 0 \tag{48}$$

and

$$\int q(z^l) \log[1 - \sigma(z^l)]\, dz^l > -\infty. \tag{49}$$

Since $m(z^l) \le g^{l/n_0}$ (47) will follow if we can establish (48), which we now proceed to do. Since all g^{l/n_0} nodes at depth l have the y^l identically distributed (although not independent due to the tree structure)

$$m(z^l) = g^{l/n_0} P(z^l). \tag{50}$$

Therefore, from Lemma 1

$$m(z^l) \ge g^{l/n_0}[\lambda(z^l)]^{-1}(1 - D_m^2/\delta^2 l) \exp[\rho_0 D(z^l) - \rho_0 l\delta] \tag{51}$$

and

$$\int q(z^l) \log[m(z^l)]\, dz^l \ge l[\ln(g)/n_0] + \log(1 - D_m^2/\delta^2 l) - \rho_0 l\delta + \int q(z^l)[-\log\lambda(z^l) + \rho_0 D(z^l)]\, dz^l \tag{52}$$

$$\ge l(R - \rho_0\delta) + \log\tfrac{1}{2} + l\rho_0 D_0 - \int q(z^l) \log\lambda(z^l)\, dz^l \tag{53}$$

provided l is large enough so that

$$1 - D_m^2/\delta^2 l \ge \tfrac{1}{2} \tag{54}$$

or

$$l \ge 2D_m^2/\delta^2. \tag{55}$$

Now using

$$R(D) = -\rho_0 D + \int q(z) \log\lambda(z)\, dz \tag{7}$$

we have

$$\int q(z^l) \log[m(z^l)]\, dz^l \ge l[R - \rho_0\delta - R(D_0)] + \log\tfrac{1}{2} = l(\gamma - \rho_0\delta) + \log\tfrac{1}{2} > 0 \tag{56}$$

for δ small enough and l large enough provided $\gamma > 0$ is fixed.

Having thus established (47) and (48) we now turn to (49) to complete the proof of the theorem. First note that if all g^{l/n_0} nodes are discarded (i.e., not in S'), then $y_1{}^l$, the first node (lexicographically in the code) is also not in $S'(\delta,z^l)$. Therefore, using (32),

$$\sigma(z^l) \le \Pr(y_1{}^l \notin S'(\delta,z^l)) = 1 - P(z^l) \tag{57}$$

since the components of $y_1{}^l$ are i.i.d. according to $w_0(y)$. However, from Lemma 1

$$\sigma(z^l) \le 1 - [\lambda(z^l)]^{-1}(1 - D_m{}^2/\delta^2 l) \exp[\rho_0 D(z^l) - \rho_0 l\delta] \tag{58}$$

and

$$\begin{aligned}\int q(z^l) \log[1 - \sigma(z^l)]\, dz^l &\ge \int q(z^l)[-\log \lambda(z^l) + \rho_0 D(z^l) - \rho_0 l\delta + \log \tfrac{1}{2}]\, dz^l \\ &= -R(D_0) - \rho_0 l\delta + \log \tfrac{1}{2} \\ &> -\infty\end{aligned}$$

provided l is at least as large as in (55). Q.E.D.

Then applying Theorem 3 and the usual argument that at least one code in the ensemble is as good as the average we have the desired result.

Theorem 5: For any i.i.d. source and bounded single letter distortion measure there exist tree codes whose performance is as close to the $R(D)$ curve as desired.

As noted prior to Lemma 1, we can also infer that an algorithm based on the $S'(\delta,z^l)$ discard criterion will be usable.

V. Summary and Conclusions

We have shown that Jelinek's proof of the optimality of tree codes for encoding i.i.d. discrete-time sources with respect to a fidelity criterion is valid only for sources that are symmetric. By modifying the discard criterion and using results from the theory of branching processes with random environments, we were able to reestablish the optimality of random tree codes for asymmetric sources. The need for a discard criterion dependent on $D(z^l)$ has been encountered by Dick, Berger, and Jelinek [12] in trying to encode outputs of a discrete-time $N(0,1)$ Gaussian source. The preceding development indicates that the use of the metric $d(z,y) - D(z)$, instead of $d(z,y) - D_0$, is necessary (positive metrics being desirable). In fact, Dick, Berger, and Jelinek found that a more extreme dependence on z was needed.

It should be noted that the mean-squared error criterion $d(z_i,y_i) = (z_i - y_i)^2$ does not meet our boundedness requirement. However, it is obvious from the proof that this requirement is not necessary in general. In particular it would be sufficient if for all z^l

$$\int_{y^l \notin S'(\delta,z^l)} d(z^l,y^l) w_0(y^l)\, dy^l \le A < \infty \tag{59}$$

so that restarting a terminated process incurs finite distortion; and further for all z^l either

$$E_{y|z}[d(z,y) - D(z)]^2 \le B < \infty \tag{60a}$$

or

$$P_{y|z}(S'') = 1 - o(l) \tag{60b}$$

so that an equation similar to (44) results. While we have not yet done so, we conjecture that (59) and (60a) or (60b) hold for all reasonable sources and distortion measures.

Acknowledgment

The authors wish to thank Prof. Toby Berger for several valuable discussions and suggestions.

References

[1] F. Jelinek, "Tree encoding of memoryless time-discrete sources with a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT-15, pp. 584–590, Sept. 1969.

[2] J. Anderson and F. Jelinek, "A 2-cycle algorithm for source coding with a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 77–92, Jan. 1973.

[3] R. Gallager, "Tree encoding for symmetric sources with a distortion measure," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 65–76, Jan. 1974.

[4] A. Viterbi and J. Omura, "Trellis encoding of memoryless discrete-time sources with a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 325–332, May 1974.

[5] B. Haskell, "The computation and bounding of rate-distortion functions," *IEEE Trans. Inform. Theory*, vol. IT-15, pp. 525–531, Sept. 1969.

[6] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression.* Englewood Cliffs, N.J.: Prentice-Hall, 1971.

[7] W. Feller, *An Introduction to Probability Theory and Its Applications*, 3rd ed. New York: Wiley, 1968.

[8] E. Parzen, *Stochastic Processes.* San Francisco: Holden-Day, 1962.

[9] T. Harris, *The Theory of Branching Processes.* New York: Springer, 1963.

[10] K. Athreya and P. Ney, *Branching Processes.* New York: Springer, 1972.

[11] W. Smith and W. Wilkinson, "On branching processes in random environments," *Ann. Math. Statist.*, vol. 40, pp. 814–827, 1969.

[12] R. Dick, T. Berger, and F. Jelinek, "Tree encoding of Gaussian, sources," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 332–336, May 1974.