Universal Coding with Minimum Probability of Codeword Length Overflow

Neri Merhav

Abstract—**Lossless block-to-variable length source coding is studied for finite-state, finite-alphabet sources. We aim to minimize the probability that the normalized length of the codeword will exceed a given threshold B, subject to the Kraft inequality. It is shown that the Lempel–Ziv (LZ) algorithm asymptotically attains the optimal performance in the sense just defined, *independently* of the source and the value of B. For the subclass of unifilar Markov sources, faster convergence to the asymptotic optimum performance can be accomplished by using the minimum description length (MDL) universal code for this subclass. It is demonstrated that these universal codes are also nearly optimal in the sense of minimizing buffer overflow probability, and asymptotically optimal in a competitive sense.**

Index Terms—**Universal noiseless coding, Lempel–Ziv algorithm, length overflow, buffer overflow, finite-state sources, large deviations, competitive optimality.**

I. Introduction

LOSSLESS block-to-variable length source coding schemes are usually examined under the criterion of minimizing either the expected length $E(L)$ of a codeword [1], [2] or the cost function $\lambda^{-1}\log E(2^{\lambda L})$ [2], [3], where L denotes the codeword length, $\lambda > 0$ is a given constant, and $\log(\cdot) \triangleq \log_2(\cdot)$ throughout the sequel. It is well known that the Shannon entropy and the Rényi entropy, respectively, are achievable lower bounds to these cost functions.

We introduce a different criterion for the performance of a lossless code. Given a constant $B > 0$, we seek a uniquely decipherable block code, or a class of block codes, which asymptotically minimize the probability that $L > Bn$, where n is the input block size. This criterion might be useful in applications of fixed rate coding schemes, where a block of length n is first encoded to a binary string of variable length L, which if $L > Bn$, is truncated to Bn bits, resulting in a constant rate of B bits per input letter. We would like to minimize the probability that $L > Bn$, i.e., the probability that information is lost.

For the class of finite-state, finite alphabet sources, it is shown that the Lempel–Ziv (LZ) algorithm [4] is asymptotically optimal in the above sense, *independently* of the source and the value of B. In other words, even if the source statistics are *known*, the best code, in that sense, is the LZ code, which does not use the knowledge of the source. The technique used to prove this result is similar to that of [5]. However, all the pertinent results from [5] are rederived here for the sake of completeness.

Next, it is demonstrated that for unifilar finite-state sources [6, p. 187], that is, sources for which the underlying state sequence is uniquely determined by the observations (e.g., Markov sources), the probability of length overflow associated with the LZ algorithm tends to zero exponentially fast as the size n of the block tends to infinity. If, in addition, the source is known *a priori* to be unifilar, then the convergence of $n^{-1}\log\Pr\{n^{-1}L > B\}$ to the asymptotic optimal exponent can be accelerated if one uses the minimum description length (MDL) universal code [7], [8] for the subclass of unifilar sources.

Two additional aspects of these universal codes are also discussed. First, it is shown that when buffer is used in variable length coding of fixed rate memoryless sources, the LZ algorithm, as well as the MDL universal code, nearly maximize the exponential decay rate of the buffer overflow probability. This is done by showing that these codes asymptotically attain the Rényi entropy (as well as the Shannon entropy) for every memoryless source. Second, it is shown that these universal codes are asymptotically optimal in a competitive sense [9], i.e., most of the time they provide a codeword shorter than that of any given competing code, within a vanishingly small redundancy term.

In Section II we formulate the problem and state the main theorem. The case of unifilar sources is considered in Section III. In Section IV these results, along with additional aspects, are discussed. Finally, in Section V a proof of the main result is provided.

II. Problem Formulation and Main Result

Let $\boldsymbol{x} = x_1, x_2, \cdots, x_i, \cdots, x_n$, be a sequence of observable random variables taking values in a finite set $\boldsymbol{X}$ with cardinality $|\boldsymbol{X}| = X$. Similarly, let $\boldsymbol{s} = s_1, s_2, \cdots, s_i, \cdots, s_n$, be another sequence of random variables, called states, which take values in another finite set $\boldsymbol{S}$ of size $|\boldsymbol{S}| = S$. A probabilistic source P is called finite-state (with S states) if

$$P(\boldsymbol{x}, \boldsymbol{s}) = \prod_{i=1}^{n} p(x_i, s_i | s_{i-1}), \tag{1}$$

Manuscript received October 20, 1989; revised August 29, 1990. This work was presented at the 1990 AAAI Spring Symposium on the Theory and Applications of Minimum Length Coding, Stanford University, Stanford, CA, March 27–29, 1990.

The author is with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel.

IEEE Log Number 9041980.

0018-9448/91/0500-0556$01.00 ©1991 IEEE

where $p(x_i, s_i|s_{i-1})$ is the joint probability of a letter x_i and a state s_i given the previous state s_{i-1}, $P(\boldsymbol{x}, \boldsymbol{s})$ is the joint probability of $\boldsymbol{x}$ and $\boldsymbol{s}$, and $s_0 \in S$ is a fixed initial state. Often the state sequence $\boldsymbol{s}$ is unobservable. The class of finite-state sources with no more than S states will be denoted by $\boldsymbol{P}_S$.

A length function of a binary lossless code is a mapping L_n from the set $\boldsymbol{X}^n$ of all possible observation sequences $\boldsymbol{x}$ to the set N of positive integers, which satisfies the Kraft inequality,

$$\sum_{\boldsymbol{x} \in \boldsymbol{X}^n} 2^{-L_n(\boldsymbol{x})} \le 1. \tag{2}$$

Given a finite-state source P and a constant $B > 0$, the problem is that of finding a length function of a uniquely decipherable code $L_n(\cdot)$ that minimizes the probability of length overflow, defined as

$$\Pr\{n^{-1}L_n(\boldsymbol{x}) > B\} = \sum_{\boldsymbol{x}:\, n^{-1}L_n(\boldsymbol{x}) > B} P(\boldsymbol{x}), \tag{3}$$

where $P(\boldsymbol{x}) = \sum_{\boldsymbol{s} \in S^n} P(\boldsymbol{x}, \boldsymbol{s})$ and $P(\boldsymbol{x}, \boldsymbol{s})$ is as in (1). It is assumed that

$$H < B < \log X, \tag{4}$$

where H is the Shannon entropy, which for a stationary source is given by

$$H \triangleq -\lim_{n \to \infty} n^{-1} \sum_{\boldsymbol{x} \in \boldsymbol{X}^n} P(\boldsymbol{x}) \log P(\boldsymbol{x}). \tag{5}$$

The assumption $B > H$ is necessary for the existence of a sequence of length functions $\{L_n(\cdot)\}_{n \ge 1}$ such that the probability of overflow, (3), will vanish as $n \to \infty$. On the other hand, the maximum value of B for which the problem is interesting is $\log X$, beyond which (3) can be made zero in a trivial manner.

Let $U_{LZ}(\boldsymbol{x})$ be the length function of the Lempel–Ziv [4] code, that is, $U_{LZ}(\boldsymbol{x})$ is the number of bits associated with the Lempel–Ziv codeword of a sequence $\boldsymbol{x}$. The following theorem establishes the asymptotic optimality of the Lempel–Ziv algorithm in the sense of minimum probability of length overflow (3).

Theorem 1: For any length function $L_n(\cdot)$, every $B \in (H, \log X)$, every finite-state source P, and all large n,

$$\Pr\{n^{-1}U_{LZ}(\boldsymbol{x}) > B + \epsilon(n)\} \le \left(1 + n^2 2^{-n/\sqrt{\log n}}\right) \Pr\{n^{-1}L_n(\boldsymbol{x}) > B\}, \tag{6}$$

where $\epsilon(n) = O(1/\sqrt{\log n})$ is a positive sequence depending on X and S.

The proof appears in Section V.

Theorem 1 states that the Lempel–Ziv algorithm attains the best tail behavior of the distribution of codeword lengths, i.e., the best large deviations performance. In [4] Ziv and Lempel have shown their algorithm to yield the shortest length, uniformly for every sufficiently long sequence $\boldsymbol{x}$, among all information lossless *finite-state* encoders. Here the result is dual to [4] in the sense that the *source* is limited to be finite-state, but *any* competing lossless encoder is allowed, not necessarily a finite-state encoder.

III. Unifilar Sources

The convergence of $\epsilon(n)$ in (6) can be accelerated if P is known to belong to the subclass $\boldsymbol{P}_U \subset \boldsymbol{P}_S$ of *unifilar* sources. For these sources the state s_i, at time instant i, obeys the recursion

$$s_i = f(x_i, s_{i-1}), \tag{7}$$

where f: $\boldsymbol{X} \times \boldsymbol{S} \to \boldsymbol{S}$ is a known deterministic mapping. Clearly, in this case, given s_0, which is fixed, one can reconstruct the state sequence $\boldsymbol{s}$ recursively upon observing $\boldsymbol{x}$. If, in addition, f is a 1–1 map from $\boldsymbol{X}$ to $\boldsymbol{S}$ given its second argument, then $\boldsymbol{s}$ uniquely determines $\boldsymbol{x}$ as well, by the inverse map. An important special case of this model is a kth order Markov source where $s_i = (x_i, x_{i-1}, \cdots, x_{i-k+1})$.

We now demonstrate that by using the MDL universal code for the subclass $\boldsymbol{P}_U$, the term $\epsilon(n)$ can be reduced to $O(\log n / n)$. Let,

$$q_{\boldsymbol{x}}(x, s) = \frac{1}{n} \sum_{i=1}^{n} \delta(x_i = x,\, s_{i-1} = s),\ x \in \boldsymbol{X},\ s \in \boldsymbol{S}, \tag{8}$$

where $\delta(x_i = x,\ s_{i-1} = s)$ is the indicator function for $x_i = x$ jointly with $s_{i-1} = s$. Also, let $q_{\boldsymbol{x}}(s) = \sum_{x \in X} q_{\boldsymbol{x}}(x, s)$ and

$$q_{\boldsymbol{x}}(x|s) = \begin{cases} q_{\boldsymbol{x}}(x, s)/q_{\boldsymbol{x}}(s), & q_{\boldsymbol{x}}(s) > 0 \\ 0, & q_{\boldsymbol{x}}(s) = 0. \end{cases} \tag{9}$$

We denote by $Q_{\boldsymbol{x}}$ the empirical distribution $Q_{\boldsymbol{x}} \triangleq \{q_{\boldsymbol{x}}(x, s),\ x \in \boldsymbol{X},\ s \in \boldsymbol{S}\}$, and define the empirical entropy as

$$H(Q_{\boldsymbol{x}}) = -\sum_{s \in S} \sum_{x \in X} q_{\boldsymbol{x}}(x, s) \log q_{\boldsymbol{x}}(x|s), \tag{10}$$

and the Kullback–Leibler divergence between the empirical distribution $Q_{\boldsymbol{x}}$ and the source P is defined as

$$D(Q_{\boldsymbol{x}} \| P) \triangleq \sum_{s \in S} \sum_{x \in X} q_{\boldsymbol{x}}(x, s) \log \frac{q_{\boldsymbol{x}}(x|s)}{p(x|s)}, \tag{11}$$

where $p(x|s) = \Pr\{x_i = x | s_{i-1} = s\}$. Note that in the unifilar case considered here,

$$P(\boldsymbol{x}) = \exp_2\{-n[H(Q_{\boldsymbol{x}}) + D(Q_{\boldsymbol{x}} \| P)]\}. \tag{12}$$

It is not difficult to show that, for the subclass of unifilar finite-state sources, it is possible to attain an exponentially vanishing probability of codeword length overflow, that is,

$$\lim_{n \to \infty} n^{-1} \log \Pr\{n^{-1}L_n(\boldsymbol{x}) > B\} < 0. \tag{13}$$

To do this, the first step is to show, by a technique similar to that in Section V, that for an arbitrary length function $L_n(\boldsymbol{x})$ there exists another length function $L'_n(\boldsymbol{x})$, depending on $\boldsymbol{x}$ only through the empirical distribution $Q_{\boldsymbol{x}} =$

$\{q_x(x,s),\ x\in X,\ s\in S\}$, such that

$$n^{-1}\log\Pr\left\{n^{-1}L'_n(\boldsymbol{x})>B+O\left(\frac{\log n}{n}\right)\right\}$$
$$\le n^{-1}\log\Pr\left\{n^{-1}L_n(\boldsymbol{x})>B\right\}+O\left(\frac{\log n}{n}\right). \quad (14)$$

Therefore, without loss in asymptotic performance, the optimal length function can be assumed to depend on $\boldsymbol{x}$ only through Q_x. Let T_x be the type of $\boldsymbol{x}$, namely,

$$T_x=\{\boldsymbol{x}'\in X^n\colon Q_{x'}=Q_x\}. \quad (15)$$

Bounds on the probability of T_x are given in the following Lemma.

Lemma 1: For every $P\in P_U$, where f is a 1–1 mapping given its second argument, and for every $\boldsymbol{x}\in X^n$,

$$\left|\frac{1}{n}\log\Pr\{T_x\}+D(Q_x\|P)\right|\le\epsilon(n), \quad (16)$$

where $\epsilon(n)=O(n^{-1}\log n)$ is independent of $\boldsymbol{x}$.

Proof: The proof is a straightforward extension of [10, Lemma 1] where this argument was proved for the specific case of a kth order Markov source.

From (12), (16), and the fact that all sequences in T_x are equally likely, it follows that

$$\left|\frac{1}{n}\log|T_x|-H(Q_x)\right|=\left|\frac{1}{n}\log\frac{\Pr\{T_x\}}{P(\boldsymbol{x})}-H(Q_x)\right|\le\epsilon(n). \quad (17)$$

Following (14), let all members of T_x have the same codeword length $L_n(\boldsymbol{x})$. Clearly, for a uniquely decipherable code, this length must be at least as large as the base 2 logarithm of the cardinality of T_x, namely,

$$L_n(\boldsymbol{x})\ge\log|T_x|\ge nH(Q_x)-n\epsilon(n), \quad (18)$$

where we have used (17). Note, that (18) holds even if $L_n(\cdot)$ does not satisfy the Kraft inequality (2). The only assumption is that $L_n(\cdot)$ is a length function associated with a 1–1 mapping from X^n to a set of codewords. From (18), we have

$$\begin{aligned}&n^{-1}\log\Pr\{n^{-1}L_n(\boldsymbol{x})\ge B\}\\&\quad\ge n^{-1}\log\Pr\{H(Q_x)\ge B+\epsilon(n)\}\\&\quad=n^{-1}\log\sum_{T_x\subset\{\boldsymbol{x}\colon H(Q_x)\ge B+\epsilon(n)\}}\Pr\{T_x\}\\&\quad\ge n^{-1}\log\max_{\{Q_x\colon H(Q_x)\ge B+\epsilon(n)\}}\\&\qquad\cdot\exp_2\{-n[D(Q_x\|P)+\epsilon(n)]\}\\&\quad=-\min_{\{Q_x\colon H(Q_x)\ge B+\epsilon(n)\}}D(Q_x\|P)-\epsilon(n), \quad (19)\end{aligned}$$

where again we have used (16) for a lower bound on $\Pr\{T_x\}$. Since $D(\cdot\|P)$ and $H(\cdot)$ are continuous functions and since the set of rational empirical distributions $\{Q_x\}$ becomes dense in P_U as $n\to\infty$, the right-most side of (19) tends to a constant D (see also [14]) given by,

$$D\triangleq\min_{\{Q\colon H(Q)\ge B\}}D(Q\|P). \quad (20)$$

An alternative expression [11], [12] for D in the memoryless case ($S=1$) is

$$D=\max_{\lambda>0}\lambda(B-H_\lambda), \quad (21)$$

where H_λ is the Rényi entropy [2] given by

$$H_\lambda\triangleq\frac{\lambda+1}{\lambda}\log\sum_{x\in X}p(x)^{1/(1+\lambda)}, \quad (22)$$

$\{p(x)\}_{x\in X}$ being the letter probabilities. Clearly, to attain the optimal overflow exponent D, $L_n(\boldsymbol{x})$ should be as close as possible to $nH(Q_x)$, which is essentially the lower bound (18). In particular, consider a simple universal code which consists of $\log(n+1)$ bits (neglecting roundoff terms) to encode each one of the $S(X-1)$ source parameters $\{p(x|s)\}$ estimated from $\boldsymbol{x}$, followed by $-\log Q_x(\boldsymbol{x})$ bits for Huffman coding of $\boldsymbol{x}$ with respect to the source estimate Q_x. The resulting length function is

$$\begin{aligned}L_n^*(\boldsymbol{x})&=-\log Q_x(\boldsymbol{x})+S(X-1)\log(n+1)\\&=nH(Q_x)+S(X-1)\log(n+1). \quad (23)\end{aligned}$$

This code is similar to the well known MDL universal code (see e.g., [8]), which was shown to be asymptotically optimal in the sense of uniformly minimizing the expected redundancy. A Bayesian approach [7], in which a Huffman code is designed with respect to a (uniform) mixture of all sources in the class, also yields a length function which is asymptotically equivalent to the MDL. Similarly to (19), by using (17) for an upper bound on the cardinality of T_x, and the fact that the number of distinct types is smaller than $(n+1)^{XS}$, one arrives at the upper bound

$$\begin{aligned}&n^{-1}\log\Pr\{n^{-1}L_n^*(\boldsymbol{x})\ge B\}\\&\quad=n^{-1}\log\sum_{T_x\colon L_n^*(\boldsymbol{x})\ge Bn}\Pr\{T_x\}\\&\quad\le n^{-1}\log\sum_{Q_x\colon H(Q_x)\ge B-\epsilon'(n)}\\&\qquad\cdot\exp_2\{-n[D(Q_x\|P)-\epsilon(n)]\}\\&\quad\le n^{-1}\log(n+1)^{XS}\\&\qquad\cdot\exp_2\left\{-n\left[\min_{\{Q_x\colon H(Q_x)\ge B-\epsilon'(n)\}}D(Q_x\|P)-\epsilon(n)\right]\right\}\\&\quad\le-\min_{\{Q_x\colon H(Q_x)\ge B-\epsilon'(n)\}}D(Q_x\|P)\\&\qquad+\epsilon(n)+\frac{1}{n}XS\log(n+1), \quad (24)\end{aligned}$$

where $\epsilon'(n)=n^{-1}S(X-1)\log(n+1)$. Clearly, the right-hand side of (24) also tends to D as $n\to\infty$. This establishes the asymptotic optimality of $L_n^*(\boldsymbol{x})$ in the sense of (3).

IV. Discussion

Note that no matter whether the true source $P\in P_U$ is given or not, the optimal length function $L_n^*(\boldsymbol{x})$ depends neither on P nor on B. In other words, knowledge of P or B *cannot* improve the asymptotic overflow expo-

nent D. Only a *second-order* improvement, associated with the rate of convergence to D, can be attained if it is known *a priori* that P belongs to a certain subset of $\boldsymbol{P}_U$. For instance, if the source P is known to be memoryless, then the second term in (23) can be reduced to $(X-1)\log(n+1)$, resulting in a slightly faster convergence to D.

It should be pointed out that, rather than truncating variable length codewords $L_n(\boldsymbol{x})$ into a fixed length of Bn bits, one can consider a codebook for the set G_n of the 2^{Bn} most likely input n-tuples $\boldsymbol{x}$, and use Bn bits to represent the codeword index. In this case, information is lost whenever $\boldsymbol{x} \in G_n^c$. It is shown in [13, Theorem 2.15, p. 37]) that the best achievable exponential rate of the probability of information loss, $\lim_{n\to\infty}[-n^{-1}\log \Pr\{G_n^c\}]$, again equals D as defined in (20). (See also Marton [14, (4)] for the distortionless case.) This means that there is no loss in asymptotic optimality in that sense, when using the truncation approach.

Since Rényi's entropy is attained by the length function

$$L_n(\boldsymbol{x}) = \left\lceil -\log \frac{P(\boldsymbol{x})^{1/(1+\lambda)}}{\sum_{\boldsymbol{x}\in X^n} P(\boldsymbol{x})^{1/(1+\lambda)}} \right\rceil, \tag{25}$$

where $P(\boldsymbol{x})$ in turn depends on $\boldsymbol{x}$ only through $Q_{\boldsymbol{x}}$, the same consideration as in (14) can be used to show that $L_n^*(\boldsymbol{x})$, defined in (23), asymptotically minimizes $(n\lambda)^{-1}\log E(2^{\lambda L_n(\boldsymbol{x})})$, uniformly for any unifilar finite state source, and any $\lambda > 0$ (see Appendix B). Following [15] (see also Appendix A), the chain of inequalities

$$\begin{aligned} U_{LZ}(\boldsymbol{x}) &\le -\log \max_{P\in \boldsymbol{P}_S} P(\boldsymbol{x}) + n\delta(n) \\ &\le -\log \max_{P\in \boldsymbol{P}_U} P(\boldsymbol{x}) + n\delta(n) \\ &= nH(Q_{\boldsymbol{x}}) + n\delta(n) \\ &\le L_n^*(\boldsymbol{x}) + n\delta(n), \end{aligned} \tag{26}$$

which holds for any $\boldsymbol{x} \in X^n$ with $\delta(n) \to 0$ uniformly for every $\boldsymbol{x}$, implies that the LZ algorithm is asymptotically optimal as well, in that sense. It has been shown in [15] that $\delta(n) = O(\log\log n/\log n)$, thus, the rate of convergence to Rényi's entropy, in the LZ case, might be slower than the $O(n^{-1}\log n)$ rate of the MDL universal code. These facts might have application to universal coding with minimum buffer overflow probability. Specifically, suppose that n is fixed and we use a buffer of size K to convert the variable rate coding to a fixed rate of R bits per input symbol, in the following manner. For each input block $\boldsymbol{x} \in X^n$, a codeword of length $L_n(\boldsymbol{x})$ is stored sequentially in the buffer and at the same time, a previously stored block of size Rn is removed from the buffer and transmitted. For the memoryless case, it has been shown [16], [17] that if $E\{L_n(\boldsymbol{x})\} < Rn$, then the probability of buffer overflow P_{bo} can be made exponentially small as $K \to \infty$, and the maximum buffer overflow exponent $\lim_{K\to\infty}(-K^{-1}\log P_{bo})$ is attained by the length function (25), with $\lambda = \lambda_0$ being the unique positive solution of the equation $H_\lambda = R$, where H_λ is the Rényi entropy given in (22). This solution λ_0 also equals the maximum attainable overflow exponent. Clearly, this length function depends on $P(\cdot)$ and R. However, it is easy to show (see Appendix B) that, if, one instead uses the universal code $L_n^*(\boldsymbol{x})$ (with $S=1$), which depends neither on $P(\cdot)$ nor on R, the overflow exponent degrades but only by a quantity which vanishes as fast as $n^{-1}\log n$.

Another aspect of the MDL and the LZ algorithm is asymptotic competitive optimality. Cover [9] has shown that for a known memoryless source, the Huffman code $L_n^H(\boldsymbol{x})$ is within one bit optimal in a competitive sense, namely, for any competing code $L_n(\boldsymbol{x})$,

$$\Pr\{L_n^H(\boldsymbol{x}) > L_n(\boldsymbol{x}) + 1\} \le \Pr\{L_n^H(\boldsymbol{x}) < L_n(\boldsymbol{x}) + 1\}. \tag{27}$$

This means that, within a normalized redundancy term of the order of $1/n$, the Huffman code provides most of the time a codeword shorter than that of any other competing code. We now show that this result extends to universal coding, again, by using $L_n^*(\boldsymbol{x})$ (or the MDL universal code), at the expense of increasing the redundancy term to $O(n^{-1}\log n)$. Specifically, we claim that for any unifilar finite-state source and any competing code $L_n(\cdot)$,

$$\begin{aligned} &\Pr\{L_n^*(\boldsymbol{x}) > L_n(\boldsymbol{x}) + n\epsilon(n)\} \\ &\qquad \le \Pr\{L_n^*(\boldsymbol{x}) < L_n(\boldsymbol{x}) + n\epsilon(n)\}, \end{aligned} \tag{28}$$

where $\epsilon(n) = n^{-1}[L_n^*(\boldsymbol{x}) - \lfloor nH(Q_{\boldsymbol{x}})\rfloor] = O(n^{-1}\log n)$. To show that (28) holds, we use a technique similar to [9]. First observe that,

$$\begin{aligned} &\Pr\{L_n^*(\boldsymbol{x}) > L_n(\boldsymbol{x}) + n\epsilon(n)\} \\ &\quad - \Pr\{L_n^*(\boldsymbol{x}) < L_n(\boldsymbol{x}) + n\epsilon(n)\} \\ &\quad = E\{\operatorname{sgn}[L_n^*(\boldsymbol{x}) - L_n(\boldsymbol{x}) - n\epsilon(n)]\}. \end{aligned} \tag{29}$$

Hence, it is sufficient to prove that the expectation on the right-hand side of (29) is nonpositive. To see this, we use the inequality $\operatorname{sgn}(k) \le \exp_2(k) - 1$, ($k$-integer), and obtain

$$\begin{aligned} &E\{\operatorname{sgn}[L_n^*(\boldsymbol{x}) - L_n(\boldsymbol{x}) - n\epsilon(n)]\} \\ &\quad \le E\{\exp_2[L_n^*(\boldsymbol{x}) - L_n(\boldsymbol{x}) - n\epsilon(n)]\} - 1 \\ &\quad = \sum_{\boldsymbol{x}} P(\boldsymbol{x}) \exp_2[L_n^*(\boldsymbol{x}) - L_n(\boldsymbol{x}) - n\epsilon(n)] - 1. \end{aligned} \tag{30}$$

Finally, since $P(\boldsymbol{x}) \le \exp_2[-nH(Q_{\boldsymbol{x}})]$ for every $P \in \boldsymbol{P}_U$ by (12), we get

$$\begin{aligned} &E\{\operatorname{sgn}[L_n^*(\boldsymbol{x}) - L_n(\boldsymbol{x}) - n\epsilon(n)]\} \\ &\quad \le \sum_{\boldsymbol{x}} \exp_2[-nH(Q_{\boldsymbol{x}})] \\ &\qquad \cdot \exp_2[L_n^*(\boldsymbol{x}) - L_n(\boldsymbol{x}) - n\epsilon(n)] - 1 \\ &\quad \le \sum_{\boldsymbol{x}} \exp_2[-nH(Q_{\boldsymbol{x}})] \exp_2\{\lfloor nH(Q_{\boldsymbol{x}})\rfloor - L_n(\boldsymbol{x})\} - 1 \\ &\quad \le \sum_{\boldsymbol{x}} \exp_2[-L_n(\boldsymbol{x})] - 1 \le 0, \end{aligned} \tag{31}$$

where the last step follows from the assumption that $L_n(\cdot)$ satisfies the Kraft inequality. Note that this result extends easily to any parametric class of sources, under

fairly mild regularity conditions, where $L_n^*(x)$ should be interpreted as a Huffman code with respect to the maximum likelihood estimator of the source, followed by $O(\log n)$ bits to encode a quantized version of the parameter estimate. Again, inequality (26) implies that for the class of finite state sources, $L_n^*(x)$ can be replaced by $U_{LZ}(x)$ at the expense of increasing the normalized redundancy to $O(\log\log n/\log n)$ bits per source letter.

V. Proof of Theorem 1

Assume that l divides n and parse x into n/l vectors of length l:

$$x = x_1^l, x_2^l, \cdots, x_{n/l}^l, \qquad x_i^l \in X^l. \tag{32}$$

Let s^l denote the sequence of initial states of the resulting phrases $\{x_i^l\}$,

$$s^l = s_0^l, s_1^l, \cdots, s_{n/l}^l. \tag{33}$$

Fix s^l and let $K(x|s^l)$ denote the set of n-vectors x' obtained from x by permuting phrases x_i^l and x_j^l of x, for which $s_i^l = s_j^l$ and $s_{i+1}^l = s_{j+1}^l$. Clearly, for every $x' \in K(x|s^l)$we have from (1) that $P(x, s^l) = P(x', s^l)$, and hence $K(x|s^l)$ can be thought of as a conditional type.

Given an arbitrary length function $L_n(\cdot)$ satisfying the Kraft inequality, we first find a modified length function $L_n'(\cdot|s^l)$, depending on a given s^l, which satisfies Kraft's inequality as well, and which has an additional property of mapping "large portions" of each conditional type $K(x|s^l)$ onto the same length. This property will be useful later in the derivation of the desired lower bound on the length overflow probability. We first define some notation.

Let $L_n(\cdot)$ be as above and $L_{\max} \triangleq \max_x L_n(x)$. Next, let $I \triangleq \{1, 2, \cdots, \lfloor Bn \rfloor, L_{\max}\}$, where $\lfloor Bn \rfloor$ is the largest integer not exceeding Bn (floor function), and define a partition $\Omega = \{\Omega_j\}_{j \in I}$ of X^n by $\Omega_j = \{x\colon L_n(x) = j\}$ for $j = 1, \cdots, \lfloor Bn \rfloor$ and $\Omega_{L_{\max}} = \{x\colon L_n(x) > Bn\}$. Let $\{\rho_n\}_{n \geq 1}$ be a positive sequence with the following properties:

$$\rho_n \leq 1, \tag{34a}$$

$$\lim_{n \to \infty} \rho_n = 0, \tag{34b}$$

$$\lim_{n \to \infty} \frac{1}{n} \log \rho_n = 0. \tag{34c}$$

The exact form of $\{\rho_n\}_{n \geq 1}$ will be chosen later. Define

$$j^* = \max\left\{ j \in I\colon \left|\Omega_j \cap K(x|s^l)\right| = \max_{i \in I} \left|\Omega_i \cap K(x|s^l)\right| \right\}. \tag{35}$$

As the number of classes $\{\Omega_j\}$ never exceeds $(Bn+1)$, clearly,

$$\left|K(x|s^l) \cap \Omega_{j^*}\right| \geq \frac{1}{Bn+1} \left|K(x|s^l)\right|. \tag{36}$$

It follows from (34a) and (36) that the set

$$V_{xs^l} \triangleq \left\{ j \in I\colon \left|K(x|s^l) \cap \Omega_j\right| \geq \frac{\rho_n}{Bn+1} \left|K(x|s^l)\right| \right\} \tag{37}$$

is nonempty. In particular, $j^* \in V_{xs^l}$. Henceforth the subscript xs^l will be omitted whenever clear from the context and we define $V_* = V - \{j^*\}$.

Next, define a modification $\Omega' = \{\Omega_j'\}_{j \in I}$ of Ω as follows. If $x \in \Omega_j$ with $j \in V$ then also $x \in \Omega_{j^*}'$. Otherwise, if $x \in \Omega_j$ with $j \in V^c$ then $x \in \Omega_{j^*}'$. Finally, define the modified length function as follows. $L_n'(x|s^l) = j$ for all $x \in \Omega_j'$, $j \neq j^*$. For $x \in \Omega_{j^*}'$, $L_n'(x|s^l) = j^* + \lceil \log(Bn+1) \rceil$, where $\lceil a \rceil$ is the least integer not less than a (ceiling function). Thus, sequences in Ω_j, for any $j \in V_*$, are unaffected by the modification and all other sequences will have modified codewords of length $j^* + \lceil \log(Bn+1) \rceil$ bits.

Comment: Note, that even for a fixed s^l, not all sequences in Ω_j' have modified codewords of the same length $L_n'(x|s^l)$. The reason is that j^* depends not only on s^l, but also on the conditional type $K(x|s^l)$, and therefore, within a given Ω_j', for some of the types, j might coincide with j^*, in which case, $L_n'(x|s^l) = j + \lceil \log(Bn+1) \rceil$, and for other types, $L_n'(x|s^l) = j$. However, all sequences from the same type within given Ω_j' and s^l, are assigned to modified codewords of the same length.

We now show that $L_n'(\cdot|s^l)$ satisfies Kraft's inequality for every $s^l \in S^{n/l+1}$. Since the original length function $L_n(\cdot)$ satisfies Kraft's inequality by assumption and since for every s^l, the set of all conditional types $\{K(x|s^l)\}$ form a partition of X^n, we obtain

$$\begin{aligned}
1 &\geq \sum_{x \in X^n} 2^{-L_n(x)} = \sum_{\{K(x|s^l)\} \subseteq X^n} \sum_{x' \in K(x|s^l)} 2^{-L_n(x')} \\
&\geq \sum_{\{K(x|s^l)\}} \left[\sum_{x' \in \Omega_{j^*} \cap K(x|s^l)} 2^{-j^*} + \sum_{j \in V_*} \sum_{x' \in \Omega_j \cap K(x|s^l)} 2^{-j} \right] \\
&= \sum_{\{K(x|s^l)\}} \left[\left|\Omega_{j^*} \cap K(x|s^l)\right| 2^{-j^*} + \sum_{j \in V_*} \sum_{x' \in \Omega_j \cap K(x|s^l)} 2^{-j} \right] \\
&\geq \sum_{\{K(x|s^l)\}} \left[\frac{1}{Bn+1} \left|K(x|s^l)\right| 2^{-j^*} + \sum_{j \in V_*} \sum_{x' \in \Omega_j \cap K(x|s^l)} 2^{-j} \right] \\
&\geq \sum_{\{K(x|s^l)\}} \left[\frac{1}{Bn+1} \sum_{j \in V_*^c} \left|\Omega_j \cap K(x|s^l)\right| 2^{-j^*} + \sum_{j \in V_*} \sum_{x' \in K(x|s^l) \cap \Omega_j} 2^{-j} \right] \\
&\geq \sum_{\{K(x|s^l)\}} \left[\sum_{j \in V_*^c} \sum_{x' \in \Omega_j \cap K(x|s^l)} 2^{-(j^* + \lceil \log(Bn+1) \rceil)} + \sum_{j \in V_*} \sum_{x' \in K(x|s^l) \cap \Omega_j} 2^{-j} \right] \\
&= \sum_{x \in X^n} 2^{-L_n'(x|s^l)}.
\end{aligned} \tag{38}$$

Let $\Pr\{L_n(x) > Bn, s^l\}$ denote the joint probability under P (1) of the event $\{L_n(x) > Bn\}$ and the occurrence of a state sequence s^l. For every fixed s^l, we now lower bound

$\Pr\{L_n(x) > Bn, s^l\}$ by means of the auxiliary length function $L'_n(x|s^l)$. Let

$$U = \{j: j > Bn\} \tag{39}$$

$$U' = \{j: j > Bn + \lceil \log(Bn+1) \rceil\} \tag{40}$$

$$\delta_U(j) = \begin{cases} 1, & j \in U \\ 0, & j \in U^c \end{cases} \tag{41}$$

and $\delta_{U'}(j)$ be defined similarly to $\delta_U(j)$. Clearly, $\delta_U(j) = \delta_{U'}(j + \lceil \log(Bn+1) \rceil)$. Hence,

$$\begin{aligned}
&\Pr\{L_n(x) > Bn, s^l\} \\
&\quad = \sum_{j \in U} \sum_{x \in \Omega_j} P(x, s^l) \\
&\quad = \sum_{\{K(x|s^l)\}} \sum_{j \in U} \sum_{x' \in \Omega_j \cap K(x|s^l)} P(x', s^l) \\
&\quad \geq \sum_{\{K(x|s^l)\}} \Bigg[\delta_U(j^*) \left| \Omega_{j^*} \cap K(x|s^l) \right| P(x, s^l) \\
&\qquad + \sum_{j \in U \cap V_*} \left| \Omega_j \cap K(x|s^l) \right| P(x, s^l) \Bigg].
\end{aligned} \tag{42}$$

By the construction of $L'_n(\cdot|s^l)$,

$$\begin{aligned}
&\left| K(x|s^l) \cap \Omega'_{j^*} \right| \\
&\quad = \left| K(x|s^l) \cap \Omega_{j^*} \right| + \sum_{j \in V^c} \left| K(x|s^l) \cap \Omega_j \right| \\
&\quad < \left| K(x|s^l) \cap \Omega_{j^*} \right| + (Bn+1) \cdot \frac{\rho_n}{Bn+1} \left| K(x|s^l) \right| \\
&\quad \leq \left| K(x|s^l) \cap \Omega_{j^*} \right| + \rho_n (Bn+1) \left| K(x|s^l) \cap \Omega_{j^*} \right| \\
&\quad = \left| K(x|s^l) \cap \Omega_{j^*} \right| \cdot [1 + \rho_n (Bn+1)],
\end{aligned} \tag{43}$$

or, equivalently,

$$\left| K(x|s^l) \cap \Omega_{j^*} \right| \geq \frac{|K(x|s^l) \cap \Omega'_{j^*}|}{1 + \rho_n (Bn+1)}. \tag{44}$$

By plugging (44) into (42) and using the fact that $\Omega_j \cap K(x|s^l) = \Omega'_j \cap K(x|s^l)$ for every $j \in V_*$, we can further lower bound $\Pr\{L_n(x) > Bn, s^l\}$ as follows.

$$\begin{aligned}
&\Pr\{L_n(x) > Bn, s^l\} \\
&\quad \geq \sum_{\{K(x|s^l)\}} \Bigg[[1 + \rho_n (Bn+1)]^{-1} \left| K(x|s^l) \cap \Omega'_{j^*} \right| \\
&\qquad \cdot P(x, s^l) \delta_{U'}(j^* + \lceil \log(Bn+1) \rceil) \\
&\qquad + \sum_{j \in U' \cap V_*} \left| \Omega'_j \cap K(x|s^l) \right| P(x, s^l) \Bigg] \\
&\quad \geq [1 + \rho_n (Bn+1)]^{-1} \sum_{\{K(x|s^l)\}} \Bigg[\left| K(x|s^l) \cap \Omega'_{j^*} \right| P(x, s^l) \\
&\qquad \cdot \delta_{U'}(j^* + \lceil \log(Bn+1) \rceil) \\
&\qquad + \sum_{j \in U' \cap V_*} \left| \Omega'_j \cap K(x|s^l) \right| P(x, s^l) \Bigg] \\
&\quad = [1 + \rho_n (Bn+1)]^{-1} \Pr\{L'_n(x|s^l) > Bn \\
&\qquad + \lceil \log(Bn+1) \rceil, s^l\}.
\end{aligned} \tag{45}$$

Now, by construction of Ω', for any nonempty Ω'_j, we have $|K(x|s^l) \cap \Omega'_j| \geq (Bn+1)^{-1} \rho_n |K(x|s^l)|$. Since $L'_n(\cdot|s^l)$ is a uniquely decipherable code for any fixed s^l and since all n-tuples in $K(x|s^l) \cap \Omega'_j$ have modified codewords of the same length, then, similarly to (18), this length must be at least as large as the base 2 logarithm of the cardinality of this set. Thus,

$$\begin{aligned}
L'_n(x|s^l) &\geq \log \left| K(x|s^l) \cap \Omega'_j \right| \\
&\geq \log \left| K(x|s^l) \right| + \log \frac{\rho_n}{Bn+1}.
\end{aligned} \tag{46}$$

It is shown in [5, Lemma 1], [15] (see also Appendix A) that one can choose a sequence $l_n \to \infty$ such that

$$\log \left| K(x|s^{l_n}) \right| \geq U_{LZ}(x) - n\delta(n), \tag{47}$$

where $\delta(n) = O(1/\sqrt{\log n})$ uniformly for any x. By substituting (47) into (46) with $l = l_n$, we get

$$L'_n(x|s^{l_n}) \geq U_{LZ}(x) - n\delta(n) - \log\left(\frac{Bn+1}{\rho_n} \right). \tag{48}$$

From (45) and (48) we obtain the lower bound

$$\begin{aligned}
&\Pr\{L_n(x) > Bn, s^{l_n}\} \\
&\quad \geq [1 + \rho_n (Bn+1)]^{-1} \Pr\{L'_n(x|s^{l_n}) > Bn \\
&\qquad + \lceil \log(Bn+1) \rceil, s^{l_n}\} \\
&\quad \geq [1 + \rho_n (Bn+1)]^{-1} \\
&\qquad \cdot \Pr\left\{ U_{LZ}(x) > Bn + \lceil \log(Bn+1) \rceil \right. \\
&\qquad \left. + n\left[\delta(n) + \frac{1}{n} \log\left(\frac{Bn+1}{\rho_n} \right) \right], s^{l_n} \right\}.
\end{aligned} \tag{49}$$

By choosing $\rho_n = \min\{1, (Bn+1)^{-1} n^2 2^{-n/\sqrt{\log n}}\}$, which satisfies (34), the term $n^{-1} \log[(Bn+1)/\rho_n)$ on the right-most side of (49) becomes $O(1/\sqrt{\log n})$, which is the same order as $\delta(n)$. Finally, by summing the left-most and the right-most sides of (49), over all possible state sequences s^{l_n}, we complete the proof of Theorem 1.

Acknowledgment

The author wishes to thank Jacob Ziv, Toby Berger, Aaron D. Wyner, Amir Dembo, and Esther Levin for helpful discussions and suggestions. Useful comments made by the anonymous referees are greatly appreciated.

Appendix A

Proof of (47) and (26): We first prove (47) and then, based on this proof, we derive (26). Let $K(x|s^l)$ be defined as in the first paragraph of Section V. Define the following empirical distributions.

$$q^l_{xs^l}(u, s, s') = \frac{1}{n} \sum_{i=1}^{n/l} \delta\left(x_i^l = u, s_{i-1}^l = s, s_i^l = s'\right), \quad u \in X^l, s, s' \in S, \tag{A.1}$$

where $s^l_{n/l+1} \triangleq s_0$, $\delta(x_i^l = u, s_{i-1}^l = s, s_i^l = s')$ is the indica-

tor function for $x_i^l = u$, $s_{i-1}^l = s$ and $s_i^l = s'$. Let

$$q_x^l(u) = \sum_{s,s' \in S} q_{xs'}^l(u,s,s') \tag{A.2}$$

$$q_{s'}^l(s,s') = \sum_{u \in X^l} q_{xs'}^l(u,s,s') \tag{A.3}$$

$$q_{x|s'}^l(u|s,s') = \begin{cases} q_{xs'}^l(u,s,s')/q_{s'}^l(s,s'), & q_{s'}^l(s,s') > 0 \\ 0, & q_{s'}^l(s,s') = 0 \end{cases} \tag{A.4}$$

$$H(q_{xs'}^l) = -\sum_{s,s' \in S} \sum_{u \in X^l} q_{xs'}^l(u,s,s') \log q_{xs'}^l(u,s,s') \tag{A.5}$$

$$H(q_{s'}^l) = -\sum_{s,s' \in S} q_{s'}^l(s,s') \log q_{s'}^l(s,s') \tag{A.6}$$

$$H(q_x^l) = -\sum_{u \in X^l} q_x^l(u) \log q_x^l(u). \tag{A.7}$$

$$H(q_{x|s'}^l) = -\sum_{s,s' \in S} \sum_{u \in X^l} q_{xs'}^l(u,s,s') \log q_{x|s'}^l(u|s,s'). \tag{A.8}$$

It is easy to show that

$$\begin{aligned} H(q_{x|s'}^l) &= H(q_{xs'}^l) - H(q_{s'}^l) \geq H(q_x^l) - H(q_{s'}^l) \\ &\geq H(q_x^l) - 2\log S. \end{aligned} \tag{A.9}$$

Let $n(u,s,s') = l^{-1} n q_{xs'}^l(u,s,s')$ and $n(s,s') = \sum_{u \in X^l} \times n(u,s,s')$. By the definition of $K(\boldsymbol{x}|s^l)$, we have

$$|K(\boldsymbol{x}|s^l)| = \prod_{s,s' \in S} \frac{n(s,s')!}{\prod_{u \in X^l} n(u,s,s')!}. \tag{A.10}$$

By using the inequality $n \log n - n \log e \leq \log n! \leq n \log n$, we obtain

$$\begin{aligned} &\frac{1}{n} \log |K(\boldsymbol{x}|s^l)| \\ &\quad = \frac{1}{n} \sum_{s,s' \in S} \left[\log n(s,s')! - \sum_{u \in X^l} \log n(u,s,s')! \right] \\ &\quad \geq \frac{1}{n} \sum_{s,s' \in S} \left[n(s,s') \log n(s,s') - n(s,s') \log e \right. \\ &\qquad \left. - \sum_{u \in X^l} n(u,s,s') \cdot \log n(u,s,s') \right] \\ &\quad = \frac{1}{l} H(q_{x|s'}^l) - \frac{1}{l} \log e \\ &\quad \geq \frac{1}{l} \left[H(q_x^l) - 2\log S - \log e \right] \\ &\quad \triangleq \frac{1}{l} \left[H(q_x^l) - C \right]. \end{aligned} \tag{A.11}$$

It is known [8] that there exists a uniquely decipherable code with a length function

$$L_n(\boldsymbol{x}) \leq \frac{n}{l} \left[H(q_x^l) + 1 \right]. \tag{A.12}$$

Furthermore, this coding algorithm can be implemented by an encoder with X^l states. Hence, by [4, Theorem 1],

$$L_n(\boldsymbol{x}) \geq (c(\boldsymbol{x}) + X^{2l}) \log \frac{c(\boldsymbol{x}) + X^{2l}}{4X^{2l}}, \tag{A.13}$$

where $c(\boldsymbol{x})$ is the maximum number of distinct strings (phrases) in $\boldsymbol{x}$ [4]. On the other hand, by [4, Theorem 2],

$$U_{LZ}(\boldsymbol{x}) \leq (c(\boldsymbol{x}) + 1) \log [2X(c(\boldsymbol{x}) + 1)]. \tag{A.14}$$

Hence, by (A.11)–(A.14),

$$\begin{aligned} &\frac{1}{n} \log |K(\boldsymbol{x}|s^l)| \\ &\quad \geq \frac{1}{l} H(q_x^l) - \frac{C}{l} \\ &\quad \geq \frac{1}{n} (c(\boldsymbol{x}) + X^{2l}) \log \frac{c(\boldsymbol{x}) + X^{2l}}{4X^{2l}} - \frac{C+1}{l} \\ &\quad \geq \frac{1}{n} U_{LZ}(\boldsymbol{x}) - \frac{c(\boldsymbol{x}) + X^l}{n} \log 8X^{2l+1} - \frac{C+1}{l} \\ &\quad \geq \frac{1}{n} U_{LZ}(\boldsymbol{x}) - \frac{c(\boldsymbol{x}) + X^l}{n} \log X^{4l+2} - \frac{C+1}{l} \\ &\quad \triangleq \frac{1}{n} U_{LZ}(x) - \delta_0(n,l), \end{aligned} \tag{A.15}$$

where in the last inequality we have used the fact that $X \geq 2$ and $l \geq 1$. By [4, (6)], $c(\boldsymbol{x}) \leq n \log X / [(1 - \epsilon_n) \log n]$ for some $\epsilon_n \to 0$. Thus, for n and l sufficiently large, there exist constants C_1, C_2, and C_3 such that

$$\delta_0(n,l) \leq \frac{C_1}{l} + \frac{C_2 l}{\log n} + \frac{C_3 l X^l}{n}. \tag{A.16}$$

By choosing $l = l_n = \lceil \sqrt{\log n} \rceil$, the right-hand side of (A.16) becomes $O(1/\sqrt{\log n})$, completing the proof of (47).

Finally, to prove (26), we return to (A.15) and observe that, for any s^l,

$$1 \geq \sum_{x' \in K(\boldsymbol{x}|s^l)} P(\boldsymbol{x}', s^l) = |K(\boldsymbol{x}|s^l)| P(\boldsymbol{x}, s^l), \tag{A.17}$$

and hence,

$$\begin{aligned} P(\boldsymbol{x}) &= \sum_{s^l} P(\boldsymbol{x}, s^l) \\ &\leq \sum_{s^l} |K(\boldsymbol{x}|s^l)|^{-1} \\ &\leq S^{n/l+1} \exp_2 \{ -U_{LZ}(\boldsymbol{x}) + n\delta_0(n,l) \} \\ &\leq \exp_2 \left\{ -U_{LZ}(\boldsymbol{x}) + n \left[\delta_0(n,l) + \left(\frac{1}{n} + \frac{1}{l} \right) \log S \right] \right\} \\ &\triangleq \exp_2 \{ -U_{LZ}(\boldsymbol{x}) + n\delta_1(n,l) \}. \end{aligned} \tag{A.18}$$

Hence, by choosing again $l_n = \lceil \sqrt{\log n} \rceil$, we get

$$U_{LZ}(\boldsymbol{x}) \leq -\log P(\boldsymbol{x}) + n\delta'(n), \tag{A.19}$$

where $\delta'(n) = \delta_1(n, l_n) = O(1/\sqrt{\log n})$. Tighter bounds

where $\delta'(n) = O(\log\log n / \log n)$ can be found in [15]. Since (A.19) holds for any finite-state source P, the right-hand side can be minimized with respect to P, resulting in

$$U_{LZ}(x) \le -\log \max_{P \in P_S} P(x) + n\delta'(n). \quad \text{(A.20)}$$

This completes the proof of the first inequality in (26). The remaining inequalities in (26) are straightforward. □

Appendix B
Universal Coding with Minimum Probability of Buffer Overflow

Let $\delta_n \triangleq n^{-1}X\log(n+1)$. We first upper bound $(n\lambda)^{-1}\log E(2^{\lambda L_n^*(x)})$.

$$\begin{aligned}
&\frac{1}{n\lambda}\log E(2^{\lambda L_n^*(x)}) \\
&\le \frac{1}{n\lambda}\log E(2^{n\lambda H(Q_x)}) + \delta_n \\
&= \frac{1}{n\lambda}\log\left[\sum_{x\in X^n} P(x)2^{\lambda n H(Q_x)}\right] + \delta_n \\
&= \frac{1}{n\lambda}\log\left[\sum_{T_x \subset X^n} |T_x| 2^{-n[H(Q_x)+D(Q_x\|P)]}2^{\lambda n H(Q_x)}\right] + \delta_n \\
&\le \frac{1}{n\lambda}\log\left[\max_{x\in X^n}(n+1)^X 2^{nH(Q_x)}\right. \\
&\quad \left.\cdot 2^{-n[H(Q_x)+D(Q_x\|P)]}2^{\lambda n H(Q_x)}\right] + \delta_n \\
&= \frac{1}{n\lambda}\log\left[\max_{x\in X^n} 2^{\lambda n H(Q_x)}2^{-nD(Q_x\|P)}\right] + \left(1+\frac{1}{\lambda}\right)\delta_n \\
&= \max_{x\in X^n}\left[H(Q_x) - \frac{1}{\lambda}D(Q_x\|P)\right] + \left(1+\frac{1}{\lambda}\right)\delta_n. \quad \text{(B.1)}
\end{aligned}$$

On the other hand, Rényi's entropy can be lower bounded as follows:

$$\begin{aligned}
H_\lambda &= \frac{\lambda+1}{n\lambda}\log\sum_{x\in X^n} P(x)^{1/1+\lambda} \\
&= \frac{\lambda+1}{n\lambda}\log\sum_{T_x\subset X^n}|T_x|P(x)^{1/1+\lambda} \\
&\ge \frac{\lambda+1}{n\lambda}\log\max_{T_x\subset X^n}\left[(n+1)^{-X}2^{nH(Q_x)}\right. \\
&\quad \left.\cdot 2^{-n/(1+\lambda)[H(Q_x)+D(Q_x\|P)]}\right] \\
&= \max_{x\in X^n}\left[H(Q_x) - \frac{1}{\lambda}D(Q_x\|P)\right] - \left(1+\frac{1}{\lambda}\right)\delta_n. \quad \text{(B.2)}
\end{aligned}$$

It now follows by (B.1) and (B.2) that

$$\frac{1}{n\lambda}\log E(2^{\lambda L_n^*(x)}) \le H_\lambda + 2\left(1+\frac{1}{\lambda}\right)\delta_n. \quad \text{(B.3)}$$

Thus, if $L_n^*(x)$ is used instead of the optimal code, then the buffer overflow exponent is lower bounded by the solution to the equation $H_\lambda = R - 2(1+\lambda^{-1})\delta_n$, while the optimal exponent is given by λ_0, the solution to $H_\lambda = R$. Since the solution $\lambda_0(R)$ is a continuous differentiable function of R in the range $H < R < \log X$, the resulting degradation in overflow exponent is also $O(\delta_n) = O(n^{-1}\log n)$.

References

[1] R. J. McEliece, *Theory of Information and Coding*. New York: Addison-Wesley, 1977.

[2] J. Aczel and Z. Daroczy, *Measures of Information and Their Characterizations*. New York: Academic Press, 1975.

[3] P. A. Humblet, "Generalization of Huffman coding to minimize the probability of buffer overflow," *IEEE Trans. Inform. Theory*, vol. IT-27, no. 2, pp. 230–237, Mar. 1981.

[4] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Trans. Inform. Theory*, vol. IT-24, no. 5, pp. 530–536, Sept. 1978.

[5] J. Ziv and N. Merhav, "Estimating the number of states of a finite-state source," submitted to *IEEE Trans. Inform. Theory*. Also summarized in *Proc. AAAI Spring Symp. Theory Applications of Minimal-Length Encoding*, Mar. 1990, pp. 80–84.

[6] R. B. Ash, *Information Theory*. New York: Wiley, 1965.

[7] L. D. Davisson, "Universal noiseless coding," *IEEE Trans. Inform. Theory*, vol. IT-19, no. 6, pp. 783–793, Nov. 1973.

[8] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inform. Theory*, vol. IT-30, no. 4, pp. 629–636, July 1984.

[9] T. M. Cover, "On the competitive optimality of Huffman codes," preprint.

[10] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically observed statistics," *IEEE Trans. Inform. Theory*, vol. 35, no. 2, pp. 401–408, Mar. 1989.

[11] F. Jelinek and K. S. Scneider, "On variable-length-to-block coding," *IEEE Trans. Inform. Theory*, vol. IT-18, no. 6, p. 762, Nov. 1972.

[12] F. Jelinek, *Probabilistic Information Theory*. New York, McGraw-Hill, 1968.

[13] I. Csiszár and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic Press, 1981.

[14] K. Marton, "Error exponent for source coding with a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT-20, no. 2, pp. 197–199, Mar. 1974.

[15] E. Plotnik, M. J. Weinberger, and J. Ziv, "Upper bounds on the probability of sequences emitted by finite-state sources and on the redundancy of the Lempel–Ziv Algorithm," submitted to *IEEE Trans. Inform. Theory*.

[16] F. Jelinek, "Buffer overflow in variable length coding of fixed rate sources," *IEEE Trans. Inform. Theory*, vol. IT-14, no. 3, pp. 490–501, May 1968.

[17] A. D. Wyner, "On the probability of buffer overflow under and arbitrary bounded input–output distribution," *SIAM J. Appl. Math.*, vol. 27, pp. 544–570, 1974.