A SUBOPTIMAL LOSSY DATA COMPRESSION BASED ON APPROXIMATE PATTERN MATCHING*†

August 5, 1996

Tomasz Łuczak‡
Mathematical Institute
Polish Academy of Science
60-769 Poznań
Poland
tomasz@math.amu.edu.pl

Wojciech Szpankowski§
Department of Computer Science
Purdue University
W. Lafayette, IN 47907
U.S.A.
spa@cs.purdue.edu

Abstract

A practical suboptimal (variable source coding) algorithm for lossy data compression is presented. This scheme is based on approximate string matching, and it naturally extends the lossless Lempel-Ziv data compression scheme. Among others we consider the typical length of *approximately* repeated pattern within the first n positions of a stationary mixing sequence where $D\%$ of mismatches is allowed. We prove that there exists a constant $r_0(D)$ such that the length of such an approximately repeated pattern converges in probability to $1/r_0(D)\log n$ (pr.) but it *almost surely* oscillates between $1/r_{-\infty}(D)\log n$ and $2/r_1(D)\log n$, where $r_{-\infty}(D) > r_0(D) > r_1(D)/2$ are some constants. These constants are natural generalizations of Rényi entropies to the lossy environment. More importantly, we show that the compression ratio of a lossy data compression scheme based on such an approximate pattern matching is asymptotically equal to $r_0(D)$. We also establish the asymptotic behavior of the so called *approximate waiting time* N_ℓ which is defined as the time until a pattern of length ℓ repeats approximately for the first time. We prove that $\log N_\ell/\ell \to r_0(D)$ (pr.) as $\ell \to \infty$. In general, $r_0(D) > R(D)$ where $R(D)$ is the rate distortion function. Thus, for stationary mixing sequences we settle in the negative the problem recently investigated by Steinberg and Gutman by showing that a lossy extension of Wyner-Ziv scheme cannot be optimal.

Index Terms: Lossy data compression, approximate pattern matching, generalized Lempel-Ziv scheme, rate distortion, generalized Rényi entropy, mixing probabilistic model.

---

*Preliminary versions of the paper were presented at *Combinatorial Pattern Matching* conference, Asilomar, California, 1994, and 1995 *IEEE International Symposium on Information Theory*, Whistler, Canada.

†This research started off while both authors were visiting INRIA, Rocquencourt, France, and then continued during the first author visit at Purdue University and the second author visit in Poznań, Poland. The authors wish to thank INRIA (project ALGO) and NSF Grant NCR-9206315 for a generous support.

‡On leave from Department of Discrete Mathematics, Adam Mickiewicz University, Poznań, Poland.

§Additional support was provided by NSF Grants NCR-9415491, CCR-9201078 and INT-8912631, and in part by NATO Collaborative Grants 0057/89 and CGR.950060.

1. INTRODUCTION

Data compression is an important and much-studied area, and therefore fairly mature. It could be traced in the past at least to the seminal papers of Shannon. On the one hand, today many powerful trends are converging to make data compression even more crucial: The rapid growth of multimedia, of genetic and other huge on-line databases, and especially the convergence of computing and communications that has been accelerating since the triumph of digital HDTV over analog HDTV. On the other hand, recent theoretical developments (cf. [14, 20, 24, 25, 30, 32]) bring to light unexplored so far new areas of research. This was initiated by a marvelous paper of Wyner and Ziv [32], and continued by its followers (cf. [14, 20, 25, 28, 30, 31, 34]) who brought into play "stringology", i.e., algorithms on strings. For example, a suffix tree was used in [30] to solve an open problem posed by Wyner and Ziv [32] (cf. see also [25, 31]), while recently digital search trees (and analytical analysis of algorithms on words) were used in [14] and [20] to obtain the limiting distribution of the number of phrases in the lossless Lempel-Ziv parsing scheme and its redundancy.

In this paper, we plan to adopt approximate pattern matching to lossy data compression. An approximate pattern matching searches for an *approximate* occurrence of a given pattern in a text string, where the "approximation" is measured by some distance between the pattern and the text strings (e.g., Hamming distance, edit or Levenshtein distance, squared error, etc.). In information theory, in particular in data compression, the distance is measured by distortion. Thus, we first briefly review some aspects of the rate distortion theory to put our results in proper perspective. The reader is referred to [6] for more details.

Consider a stationary and ergodic sequence $\{X_k\}_{k=-\infty}^{\infty}$ taking values in a finite alphabet $\mathcal{A}$. For simplicity of presentation, we consider only the binary alphabet $\mathcal{A} = \{0, 1\}$. We write X_m^n to denote $X_m X_{m+1} \dots X_n$. The fundamental problem of data compression can be presented as follows: Imagine a source of information generating a block $x_1^n = (x_1, \dots, x_n)$ which is a realization of a stochastic process X_1^n. We encode x_1^n into a compression code c_n, and the decoder produces an estimate $\hat{x}_1^n$ of x_1^n. We assume for simplicity that the reproduction alphabet $\hat{\mathcal{A}} = \mathcal{A}$. More precisely, a code c_n is a function $\phi: \mathcal{A}^n \to \{0,1\}^*$, thus, $c_n = \phi(x_1^n)$. On the decoding side, the decoder function $\psi: \{0,1\}^* \to \mathcal{A}^n$ is applied to find $\hat{x}_1^n = \psi(c_n)$. Let $\ell(c_n)$ be the length of a code representing x_1^n. Then, the *compression ratio* is defined as $r(x_1^n) = \ell(c_n)/n$ (e.g., for image compression $r(x_1^n)$ is expressed in bits per pixel), and the *average* compression ratio is $E(r(X_1^n)) = E\ell(c_n(X_1^n))/n$. What are the achievable values of the average compression ratio? The answer depends on whether lossless (i.e., exact reconstruction of the original code is possible) or lossy (i.e., one assumes some degradation relative to the original code) is considered.

It is well known [6, 15, 38] that the average compression ratio in a lossless data compression can asymptotically reach the entropy rate, h. For a lossy transmission, one needs to introduce a measure of *fidelity*. We restrict our discussion to the Hamming distance (but subadditive distortion measures such as the ones adopted in [28] can be easily accommodated into our main results as shown in [4]) defined as

$$d_n(x_1^n, \widetilde{x}_1^n) = \frac{1}{n}\sum_{i=1}^{n} d_1(x_i, \widetilde{x}_i) \tag{1}$$

where $d_1(x, \widetilde{x}) = 0$ for $x = \widetilde{x}$ and 1 otherwise $(x, \widetilde{x} \in \mathcal{A})$. Let us now fix $D > 0$. Then, a code c_n is D-semifaithful (e.g., for lossy compression) if $d_n(x_1^n, \hat{x}_1^n) = d_n(x_1^n, \psi(c_n(x_1^n))) \le D$ (cf. [24] for a more precise definition).

The optimal compression ratio depends on the rate-distortion function $R(D)$. This is defined as follows (we give the definition of the operational rate-distortion function): Let $B_D(w_n)$ be the set of all sequences of length n whose distance from the center w_n is smaller or equal to D, that is, $B_D(w_n) = \{x_1^n :\ d_n(x_1^n, w_n) \le D\}$. We call the set $B_D(w_n)$ a D-ball with the center at w_n. Consider now the set $\mathcal{A}^n$ of all sequences of length n, and let $\mathcal{S}_n$ be a subset of $\mathcal{A}^n$. We define $N(D, \mathcal{S}_n)$ as the *minimum* number of D-balls needed to cover $\mathcal{S}_n$. Then[1]

$$R_n(D, \varepsilon) = \min_{\mathcal{S}_n:\ P(\mathcal{S}_n) \ge 1-\varepsilon} \frac{\log N(D, \mathcal{S}_n)}{n} ,$$

and the operational rate-distortion is defined as $R(D) = \lim_{\varepsilon\to 0}\lim_{n\to\infty} R_n(D, \varepsilon)$ (cf. [15, 24]). Kieffer [15], and Ornstein and Shields [24] proved that the optimal compression ratio in a lossy data compression is asymptotically equal to $R(D)$ (a.s.), and this cannot be improved. (Observe that $R(0) = h$ where h is the entropy of the underlying sequence.)

In this paper, we propose a *practical* (i.e., of a polynomial complexity) suboptimal lossy data compression scheme that extends the Lempel-Ziv scheme [38]. It achieves rate $r_0(D)$ which is asymptotically optimal only for $D \to 0$ and symmetric memoryless source. Although in general $r_0(D) \ge R(D)$, the quantity $r_0(D)$ is close to $R(D)$ for small values of D, at least for memoryless sources (cf. Figure 1 in Section 2). Our scheme reduces to the following approximate pattern matching problem: Let the "training sequence" or "database sequence" x_1^n be given. Find the largest L_n such that there exists $1 \le i_0 \le n - L_n + 1$ of the database satisfying $d(x_{i_0}^{i_0-1+L_n}, x_{n+1}^{n+L_n}) \le D$. This naturally extends Wyner and Ziv [32] idea to the lossy situation (cf. also [28]).

[1] All logarithms in this paper are with base 2 unless otherwise explicitly stated.

For $D = 0$ Wyner and Ziv [32] proposed the following data compression scheme based on L_n: The encoder sends the position i_0 in the database, the length L_n and possibly one more symbol, namely x_{n+L_n+1}. Using this information the decoder reconstructs the original message, and both the encoder and the decoder enlarge the database. Based on a probabilistic analysis the authors of [32] (cf. also [25, 30]) concluded that with high probability the compression ratio of such an algorithm is equal to the entropy, thus it is asymptotically optimal. An efficient algorithm based on a suffix tree (cf. [30]) can find L_n in $O(n)$ steps in the worst case and in $\mathcal{O}(\log n)$ steps on average (we shall use $\mathcal{O}(\cdot)$ to denote average case complexity while $O(\cdot)$ is reserved for the worst case complexity).

The situation is more complicated in the lossy case considered in this paper (cf. also [28, 34]) since one cannot use suffix trees to find the approximate longest prefix L_n, and a decoder at any time might have as a database a sample of the *distorted* process not the original process. We propose, however, an algorithm that finds the approximate prefix of length L_n in $O(n^2)$ steps in the worst case. We only briefly address algorithmic issues at the end of Section 2.2, and the reader is referred to Atallah, Genin and Szpankowski [4] for a detailed discussion. It is worth mentioning here that the authors of [4] applied a significantly enhanced version of the lossy scheme described above to image compression. In [4] some promising results for pattern matching image compression are reported, especially when variable (adaptive) D is used. Similar conclusions for image compression were drawn by Constantinescu and Storer [8] who implemented a lossy extension of another Lempel-Ziv scheme, namely the parsing scheme LZ78 [39]. However, no theoretical justifications were provided in [8] (cf. Remark 1(iv)).

While our data compression scheme is suboptimal, it is only of a polynomial complexity, thus having a chance to be of some practical importance. The trade-off between optimality and implementability is a common issue in engineering, and often optimal algorithms are either NP-hard or too expensive to construct. Optimal lossy data compression algorithms so far proposed (cf. [15, 24, 33, 36, 37]) are expensive. However, recently proposed locally (suboptimal) lossy data compression schemes are of reasonable complexity (cf. [7, 9, 18, 19]). Actually, one can envision an *optimal* data compression scheme based on approximate pattern matching (cf. [11, 27]). It is an interesting and challenging theoretical problem that needs to be addressed. But one may wonder whether a practical (i.e., of good computational complexity) and optimal lossy compression exists at all? Yang and Kieffer in their recent paper [33] expressed the following opinion: "... it is our belief that a universal lossy source coding scheme with attractive computational complexity aspects will never be found." We share this view, and we believe that further investigations of suboptimal and practical heuristics for lossy compression are needed.

We further generalize our problem and we search for largest $L_n^{(b)}$ such that there exist *at least* b substrings in the database within distance D, that is, for some $i_1, i_2, \ldots, i_b$ where $1 \le i_1 \le i_2 - L_n^{(b)} \le \ldots \le i_b - (b-1)L_n^{(b)} \le n - bL_n^{(b)} + 1$, we have $d(x_{i_1}^{i_1-1+L_n^{(b)}}, x_{n+1}^{n+L_n^{(b)}}) \le D, \ldots,$ $d(x_{i_b}^{i_b-1+L_n^{(b)}}, x_{n+1}^{n+L_n^{(b)}}) \le D$, where b is a parameter (cf. [31] for lossless equivalent of this scheme and its implementation through the so called b-suffix trees). Observe that $b = 1$ corresponds to the original problem. A recent work of Louchard *et. al.* [21] pointed out that the average redundancy rate can slightly decrease for $b > 1$.

Actually, the real engine behind this study is a probabilistic analysis of an approximate pattern matching problem, which we discuss next. Our probabilistic results are confined to the stationary mixing model in which two random events defined on two σ-algebra separated by g symbols behave almost like independent events as $g \to \infty$; thus memoryless, stationary and ergodic Markov, and finite-state sources are included (cf. [34]). We first introduce the *generalized Rényi entropies* denoted as $r_b(D)$ which we prove to exist in our mixing model, where $-\infty \le b \le \infty$ is a parameter. We show that $L_n/\log n \to 1/r_0(D)$ in probability (pr.) where $r_0(D)$ represents the rate distortion. Observe that $\lim_{D\to 0} r_0(D) = \lim_{D\to 0} R(D) = h$. Surprisingly enough, $L_n/\log n$ does not converge almost surely (a.s.) but rather fluctuates between two *different* constants, namely $1/r_{-\infty}(D) < 2/r_1(D)$ (cf. Theorem 1). This kind of behavior was already observed in the lossless case (cf. [30, 31]). Finally, for memoryless source (i.e., Bernoulli model) we compute explicitly the entropies $r_b(D)$ (cf. Theorem 3). In passing, we should add that our $r_0(D)$ is related to the ε-entropy (cf. [26]) and/or r-entropy (cf. [10]), however, we define $r_0(D)$ with respect to the source distribution instead of the optimal one. Such an entropy seems to have other applications outside the data compression area (cf. Remark 1(iv)).

It turns out that the fluctuation of L_n is related to the probabilistic behavior of two other interesting parameters that we call *shortest path* s_n and *height* H_n due to an analogy between these parameters and similar ones studied in [30, 31] for the lossless case. Roughly speaking, s_n is the largest K such that *all* strings of length K occur approximately somewhere in the training sequence of length n, while H_n is the length of the longest substring that can be approximately recopied, that is, occurs twice. We prove that $s_n/\log n \to 1/r_{-\infty}(D)$ (a.s.) and $H_n/\log n \to 2/r_1(D)$ (a.s.) (cf. Theorem 2). Observe that $s_n \le L_n \le H_n$.

In a related paper Steinberg and Gutman [28][2] analyzed the so called waiting time N_ℓ

[2] We should point out that the model of [28] (introduced in Wyner-Ziv [32]) differs slightly from ours. In [28, 32] the database is counted backward and a substring is compressed always at position $n = 0$ (in [30] it was called the *left domain asymptotics* model). This describes well the finiteness of the database but fails

which is defined as length of the shortest string that contains approximately a string of length ℓ at the beginning and at the end (or equivalently string of length ℓ reoccurs approximately for the first time after N_ℓ symbols). The authors of [28] proved that for a stationary and ergodic sequence $\limsup_{\ell\to\infty} \log N_\ell/\ell \leq R(D/2)$ (pr.). As a corollary of one of our results we show that in the mixing model $\lim_{\ell\to\infty} \log N_\ell/\ell = r_0(D)$ (pr.), and this settles the problem of [28] (cf. Corollary 1) at least for the mixing model.[3] This also implies that a lossy extension of Wyner-Ziv scheme cannot be optimal. We should also mention here some recent results of Shields [29] who analyzed the waiting time N_ℓ but only for $D \to 0$ which differs significantly from the lossy situation with $D > 0$. Finally, rate of convergence for lossy source coding are discussed in [18, 19].

There is a substantial literature on probabilistic analysis of problems on pattern matching (cf. [2, 3, 14, 29, 28, 30, 31] but with exception of [2, 3, 28] (cf. also [29] for $D \to 0$ case) only lossless case (i.e., exact pattern matching) is discussed. The two papers [2, 3] on the approximate pattern matching explore only the height H_n which is not of prime interest to data compression. Thus, to the best of our knowledge our results are novel not only in the context of data compressions.

2. MAIN RESULTS

This section contains our main results. After presenting some definitions, we formulate the probabilistic model, and we introduce *generalized Rényi entropies* that are proved to exist in our probabilistic model (cf. Section 2.1). Finally, we present our main theoretical results (cf. Section 2.2) together with algorithmic results and applications (cf. Section 2.3).

2.1 Probabilistic Model and Preliminary Results

Let $\{X_k\}_{k=-\infty}^{\infty}$ be a stationary and ergodic sequence generated over a binary alphabet $\mathcal{A} = \{0,1\}$. Throughout the paper we shall work only with the one-sided sequence $\{X_k\}_{k=1}^{\infty}$. We write x_1^n for a realization of $X_1^n = X_1X_2\ldots X_n$ and call it a training sequence or a database sequence. For a partial sequence $x_m^n = (x_m,\ldots,x_n)$ with $m \leq n$ we define the

to capture a dynamic nature of the sliding window mechanism. In our model, which was introduced in [30] and called the *right domain asymptotics* model, this dynamic nature of data compression schemes is well captured, but it does not describe well the finiteness of the database.

[3] During the revision of this paper, we have learned that Yang and Kieffer [34] have recently analyzed N_ℓ in the Wyner-Ziv model (i.e., for the left domain asymptotics) under a similar mixing model, and the authors of [34] proved – under stronger assumptions regarding mixing coefficients – that $\log N_\ell/\ell \to r_0(D)$ (a.s.).

$(n-m)$-order probability distribution as $P(x_m^n) = \Pr\{X_k = x_k \ , \ m \le k \le n, \ \ x_k \in \mathcal{A}\}$. We also use $P(X_m^n)$ as a random variable defined on the Borel sets generated by X_m^n.

We start with a precise definition of some parameters, namely: depth L_n, the M-th depth $L_n(M)$, height H_n, shortest path s_n, and waiting time N_ℓ. As in (1) we write $d(x_1^n, \tilde{x}_1^n)$ for the relative Hamming distance, that is, the ratio of the number of mismatches between x_1^n and $\tilde{x}_1^n$, and the length n. The depth L_n is defined as follows:

> Let L_n be the largest K such that a prefix of X_{n+1}^∞ of length K is within distance D from X_i^{i-1+K} for $1 \le i \le n-K+1$, that is, $d(X_i^{i-1+K}, X_{n+1}^{n+K}) \le D$.

Thus, it is the longest prefix of $\{X_k\}_{k=n+1}^\infty$ which is within distance D of a substring in the database X_1^n. On the other hand, the M-th depth, $L_n(M)$, is the longest prefix of X_M^∞ for a given M which is within distance D of a substring in the database. That is:

> For *fixed* $M \le n$, let $L_n(M)$ be the length K of the longest prefix of X_M^∞ for which there exists $M+K \le i \le n+1$ such that $d(X_i^{i+K}, X_M^{M+K}) \le D$.

The probabilistic behavior of L_n is related to two other parameters, namely the height H_n and the shortest path s_n. The height H_n is the length of the longest substring in the database X_1^n for which there exists another substring in the database within distance D. More precisely:

> The height H_n is equal to the largest K for which there exist $1 \le i < j \le n+1$ such that $d(X_i^{i-1+K}, X_j^{j-1+K}) \le D$.

In order to define s_n, we let $\mathcal{A}^k$ to be the set of *all* words of length k, and $w_k \in \mathcal{A}^k$. Then:

> The shortest path s_n is the largest k such that for *every* $w_k \in \mathcal{A}^k$ there exists $1 \le i \le n+1$ such that $d(X_i^{i-1+k}, w_k) \le D$.

The waiting time N_ℓ is also of interest to data compression, and it was already studied in [28, 32]. It is the length of the shortest sequence for which the first ℓ symbols repeats approximately for the first time. That is:

> The waiting time N_ℓ is the smallest $N \ge 2\ell$ such that $d(X_1^\ell, X_{N-\ell+1}^N) \le D$.

We observe that the waiting time is related to the first depth $L_n(1)$. In the **lossless** case $D = 0$ it was shown in [30] that

$$L_{N_\ell - 1}(1) < \ell \le L_{N_\ell}(1)$$

which directly implies probabilistic behavior of N_ℓ once we know characteristics of $L_n(1)$. The situation is more intricate in the **lossy** case $D > 0$ where the above should be replaced by the following two implications:

$$\{N_\ell \le n\} \subset \{L_n(1) \ge \ell\} \qquad \text{and} \qquad \{L_n(1) \ge \ell\} \subset \bigcup_{k \ge \ell} \{N_k \le n\} \ . \tag{2}$$

Our plan is to investigate the behavior of the above parameters in a general probabilistic framework. We assume that $\{X_k\}_{k=1}^{\infty}$ is a *stationary* and *ergodic* sequence of symbols generated from a finite alphabet $\mathcal{A}$ satisfying a mixing condition as defined below. We should point out that our results cannot hold in a general stationary and ergodic model due to some negative results of Shields discussed in [30, 31] for the lossless case.

(A) Mixing Model

Let $\mathcal{F}_m^n$ be a σ-field generated by $\{X_k\}_{k=m}^n$ for $m \le n$. There exists a function $\alpha(\cdot)$ of g such that: (i) $\lim_{g\to\infty} \alpha(g) = 0$, (ii) $\alpha(1) < 1$, and (iii) for any m, and two events $A \in \mathcal{F}_{-\infty}^m$ and $B \in \mathcal{F}_{m+g}^{\infty}$ the following holds

$$(1 - \alpha(g))\Pr\{A\}\Pr\{B\} \le \Pr\{AB\} \le (1 + \alpha(g))\Pr\{A\}\Pr\{B\}. \tag{3}$$

In some statements of our results we have to restrict the mixing model either to the Markovian model or to the Bernoulli model as defined below:

(M) Markovian Model

The sequence $\{X_k\}$ forms a stationary, aperiodic and irreducible Markov chain where the $(k+1)$st symbol in $\{X_k\}$ depends on the previously selected symbol. The transition probability of the Markov chain is $p_{i,j} = \Pr\{X_{k+1} = j \in \mathcal{A} | X_k = i \in \mathcal{A}\} > 0$ with the transition matrix denoted by $\mathbf{P} = \{p_{i,j}\}_{i,j=1}^2$.

(B) Bernoulli Model

The sequence $\{X_k\}$ forms an i.i.d. sequence with $\Pr\{X_1 = 0\} = p$ and $\Pr\{X_1 = 1\} = q = 1 - p$.

As expected, probabilistic behaviors of the above parameters depend on some kind of entropies, which we define next. We first need some additional notation. By a D-ball $B_D(w_k)$ with center $w_k \in \mathcal{A}^k$ we mean a set of all strings of length k that are within distance D from w_k, that is, $B_D(w_k) = \{x_1^k : \ d(w_k, x_1^k) \le D\}$. We simple write $P(B_D(X_1^n))$ for the probability measure of the set of all sequences of length n within distance D from a random sequence X_1^n.

Definition: GENERALIZED b-ORDER RÉNYI ENTROPY. For any $-\infty \le b \le \infty$

$$r_b(D) = \lim_{k\to\infty} \frac{-\log EP^b(B_D(X_1^k))}{bk} = \lim_{k\to\infty} \frac{-\log\left(\sum_{w_k\in\mathcal{A}^k} P^b(B_D(w_k))P(w_k)\right)}{bk}\ , \qquad (4)$$

where for $b = 0$ we understand $r_0(D) = \lim_{b\to 0} r_b(D)$, that is,

$$r_0(D) = \lim_{k\to\infty} \frac{-E\log P(B_D(X_1^k))}{k}\ , \qquad (5)$$

provided the above limits exist.

Remark 1. (i) *Special Cases.* For $b = -\infty$ and $b = \infty$ we obtain

$$r_{-\infty}(D) = \lim_{k\to\infty} \frac{-\log\left(\min_{w_k\in\mathcal{A}^k}\{P(B_D(w_k)),\ P(w_k) > 0\}\right)}{k} \qquad (6)$$

$$r_{\infty}(D) = \lim_{k\to\infty} \frac{-\log\left(\max_{w_k\in\mathcal{A}^k}\{P(B_D(w_k)),\ P(w_k) > 0\}\right)}{k}\ . \qquad (7)$$

The above follows from the *inequality on means* (cf. [13]) by taking the appropriate limits with respect to b.

(ii) *Lossless case* $D = 0$. In the lossless case $D = 0$, the generalized b-order Rényi's entropies $r_b(D)$ reduces to the b-order Rényi entropies $h^{(b)}$ studied in Szpankowski [31].

(iii) *Related Entropies.* To the best of our knowledge the entropies $r_b(D)$ were not previously used or studied in the information theory community. However, the entropy $r_0(D)$ is close in spirit to the so called ε-entropy of Posner and Rodemich [26] or Feldman's r-entropy [10]. Observe that ε-entropy corresponds to an optimal cover of the space $\mathcal{A}^n$ with D-balls, while from Lemma 1 below we conclude that using $r_0(D)$ the space is covered with typical D-balls centered at the source distribution. In other words, the probability of a typical D-ball centered at the source distribution is asymptotically equal to $2^{-nr_0(D)}$, while from Ornstein and Shields [24] one concludes that the probability of a typical D-ball centered at the optimal output distribution is asymptotically equal to $2^{-nR(D)}$ (see also the end of Section 2.2).

(iv) *Other Applications of Generalized Rényi Entropies.* Generalized Rényi entropies $r_b(D)$ find other applications in approximate pattern matching problems. The entropy $r_1(D)$ was used by Arratia and Waterman [2] to study similarities between molecular sequences, while $r_{-\infty}(D)$ might be used to analyze an approximate "signature" of a sequence (cf. [31]). We believe we can prove that $r_0(D)$ is also the asymptotic compression ratio for a lossy extension of another Lempel-Ziv scheme, namely the Lempel-Ziv (LZ78) incremental parsing scheme

[39] (i.e., in this case the next phrase is the longest phrase that is within distance D from a previous phrase). Finally, a lossy extension of the so called *Shortest Common Superstring* problem (i.e., for a given set of strings find the shortest string that contains approximately all of the original strings as substrings) brings into light the entropy $r_1(D)$ and possibly $r_0(D)$ (cf. [12, 35]). □

Lemma 1. (i) *Under assumption (A), the generalized b-th order entropy $r_b(D)$ is well defined (i.e., the limit in (4) exists) for any $-\infty \le b \le \infty$. In addition,*

$$r_0(D) = \lim_{k\to\infty} \frac{-\log P(B_D(X_1^k))}{k} \qquad \text{(a.s.)} \ . \tag{8}$$

(ii) *The entropies $r_b(D)$ are non-increasing functions of D for all $-\infty \le b \le \infty$. In addition, the following property holds*

$$b < c \qquad \Rightarrow \qquad r_b(D) \ge r_c(D) \tag{9}$$

for any $-\infty \le b < c \le \infty$.

Proof. We only consider $0 \le b < \infty$ leaving the proof of other cases to the interested reader who can follow our line of arguments. For (i) it suffices to show that for some constant $c > 0$

$$EP^b(B_D(w_{n+m})) \ge cEP^b(B_D(w_n))EP^b(B_D(w_m)) \ . \tag{10}$$

Provided (10) is true we simply use the *Subadditive Theorem* [16] applied to the sequence $a_n = -c\log EP^b(X_1^n)$ to establish our claim. In the course of proving (10) we shall see that $P^b(B_D(X_{n+m})) \ge cP^b(B_D(X_n))P^b(B_D(X_m))$ which by *Subadditive Ergodic Theorem* [16] will imply (8).

Let us now wrestle with (10). Observe that for any string w_{n+m} of length $n+m$ we have

$$\begin{aligned}
P^b(B_D(w_{n+m})) &= \left(\sum_{z_{n+m}\in B_D(w_{n+m})} P(z_{n+m})\right)^b \ge c\left(\sum_{z_{n+m}\in B_D(w_{n+m})} P(z_n)P(z_m)\right)^b \\
&\ge c\left(\sum_{z_n\in B_D(w_n)} P(z_n)\right)^b \left(\sum_{z_m\in B_D(w_m)} P(z_m)\right)^b \\
&= cP^b(B_D(w_n))P^b(B_D(w_m))
\end{aligned}$$

where c is an universal constant whose value can change from line to line. The first inequality of the above follows from the mixing condition of (A) (observe that we actually require only that $\alpha(\cdot)$ in (3) is bounded away from 1), and the second one is a simple consequence of the Hamming distance property. In fact, this inequality is satisfied by any fidelity measure having subadditivity property (cf. conditions in [28]). To complete the proof we use again the mixing condition to get

$$\begin{aligned} EP^b(B_D(w_{n+m})) &= \sum_{w_{n+m}\in\mathcal{A}^{n+m}} P^b(B_D(w_{n+m}))P(w_{n+m}) \\ &\geq c\sum_{w_n\in\mathcal{A}^n} P^b(B_D(w_n))P(w_n)\sum_{w_m\in\mathcal{A}^m} P^b(B_D(w_m))P(w_m) \\ &= cEP^b(B_D(w_n))EP^b(B_D(w_m)) \ . \end{aligned}$$

Thus (10) is proved. The proof for $b=-\infty$ and $b=\infty$ (cf. Remark 1(i)) follows the same line of arguments as above.

Part (ii) is a direct consequence of the increase of $B_D(w_k)$ and its probability with the increase of D. Clearly, (9) follows directly from the *inequality on means* [13]. This completes the proof. ■

2.2 Theoretical Results

Throughout we assume that $0 < r_\infty(D) \leq r_{-\infty}(D) < \infty$. The main result presented below describes a probabilistic behavior of L_n under the mixing model assumption (A). Its proof can be found in Section 3.1.

Theorem 1. DEPTH AND M-TH DEPTH. *Assume the mixing model (A), and* $0 < r_\infty(D) \leq r_{-\infty}(D) < \infty$.

(i) `Convergence in Probability.` *For any given M the following holds*

$$\lim_{n\to\infty}\frac{L_n(M)}{\log n} = \lim_{n\to\infty}\frac{L_n}{\log n} = \frac{1}{r_0(D)} \qquad \text{(pr.)} \tag{11}$$

provided $\alpha(g)\to 0$ as $g\to\infty$, and the rate of convergence of $\log P(B_D(X_1^k))/k$ in Lemma 1 is at least $O(1/k^{1+\delta})$ for some $\delta>0$.

(ii) `Almost Sure Convergence.` *Assume additionally that the rate of convergence of $\log P(B_D(X_1^k))/k$ in Lemma 1 is exponential. Then, for any fixed M*

$$\lim_{n\to\infty}\frac{L_n(M)}{\log n} = \frac{1}{r_0(D)} \qquad \text{(a.s.)} \tag{12}$$

provided $\sum_{g=1}^\infty \alpha(g) < \infty$. Nonetheless, one can claim only that the following is true for L_n

$$\liminf_{n\to\infty}\frac{s_n}{\log n} \leq \liminf_{n\to\infty}\frac{L_n}{\log n} \leq \limsup_{n\to\infty}\frac{L_n}{\log n} \leq \limsup_{n\to\infty}\frac{H_n}{\log n} \qquad \text{(a.s.)}\ . \tag{13}$$

Let us add that the limits $s_n/\log n$ and $H_n/\log n$ exist under some stronger assumptions on the convergence of $\alpha(g)$ (cf. Theorem 2 below) and, in general, do not coincide.

Remark 2. *Blowing-up Property.* Using recent results of Marton and Shields [23] one can prove that the exponential rate of convergence in Lemma 1 follows from the so called *blowing-up property.* To recall: a stationary and ergodic process $\{X_k\}_{k=1}^{\infty}$ has the blowing-up property if for any $\varepsilon > 0$ there exists a $\delta > 0$ and integer N such that for any $n \geq N$ and any $\mathcal{B} \subset \mathcal{A}^n$

$$\Pr\{\mathcal{B}\} \geq e^{-n\delta} \qquad \Longrightarrow \qquad \Pr\{[\mathcal{B}]_\varepsilon\} \geq 1 - \varepsilon$$

where $[\mathcal{B}]_\varepsilon = \{y_1^n : \ d(y_1^n, x_1^n) \leq \varepsilon \text{ for some } \ x_1^n \in \mathcal{B}\}$. The case $D = 0$ was analyzed in Marton and Shields [23]. One can follow their proof to show a similar conclusion for $D > 0$. For example, in order to see how the blowing-up property implies the exponential convergence in Lemma 1, let us consider for simplicity of presentation only a subset of "bad" states, namely $\widetilde{\mathcal{B}} = \{X_1^n : \ P(B_D(X_1^n)) \geq 2^{-n(r_0(D)-\theta)}\}$ for some $\theta > 0$. From Lemma 1 we know that $\Pr\{\widetilde{\mathcal{B}}\} \to 0$ as $n \to \infty$. Due to continuity of $r_0(D)$ (cf. [34]), one proves that for sufficiently small $\varepsilon > 0$ the following holds $\Pr\{[\widetilde{\mathcal{B}}]_\varepsilon\} \to 0$ as $n \to \infty$. Assume now contrary that the rate of convergence in Lemma 1 is not exponential. Then, for all $\delta > 0$ we must have $\Pr\{\widetilde{\mathcal{B}}\} \geq e^{-n\delta}$. By blowing-up property this would imply that $\Pr\{[\widetilde{\mathcal{B}}]_\varepsilon\} > 1 - \varepsilon$, which is the desired contradiction. A proof of exponential convergence for $\bar{\mathcal{B}} = \{X_1^n : \ P(B_D(X_1^n)) \leq 2^{-n(r_0(D)+\theta)}\}$ for some $\theta > 0$ is a little bit more intricate but follows the line of arguments as in Shields and Marton [23], so we omit it here. (One should first establish the exponential convergence for the empirical distribution of frequency, and then translate it into exponentiality of $\Pr\{\bar{\mathcal{B}}\}$). In passing, we should mention that in [23] Marton and Shields shown that aperiodic Markov sources, finite-state sources, and m-dependent processes have the blowing-up property. For details the reader is referred to [23]. □

The bounds for $L_n/\log n$ in the almost sure convergence of the above theorem follows directly from the simple observation that $s_n \leq L_n \leq H_n$. Furthermore, one can follow ideas of [30, 31] and show that a.s. the value of $L_n/\log n$ does **not** tend to a limit, i.e. almost surely we have $\liminf L_n/\log n < \limsup L_n/\log n$. As a matter of fact we conjecture that in the first and the last inequality of (13) the equality holds. This was proved for the lossless case $D = 0$ in [30, 31]).

Now we are in position to present our second main result concerning the height and the shortest path. The height was previously studied by Arratia and Waterman [2] and we use their result to set the issue with the height. The proof for s_n is presented in Section 3.2, while a discussion of H_n can be found in Section 3.3.

Theorem 2. SHORTEST PATH AND THE HEIGHT. *Assume that (A) holds, and* $0 < r_{\infty}(D) \leq r_{-\infty}(D) < \infty$.

(i) *If for every* $\kappa \geq 0$ *we have*

$$\lim_{g\to\infty} g^{\kappa}\alpha(g) = 0 \tag{14}$$

then

$$\lim_{n\to\infty} \frac{s_n}{\log n} = \frac{1}{r_{-\infty}(D)} \qquad \text{(a.s.)} \tag{15}$$

(ii) *For the Bernoulli model (B)*

$$\lim_{n\to\infty} \frac{H_n}{\log n} = \frac{2}{r_1(D)} \qquad \text{(a.s.)} \; . \tag{16}$$

In addition, the above holds for the Markovian model (M) if only non-overlapping substrings are considered.

Steinberg and Gutman [28] following the idea of Wyner and Ziv [32] proposed a suboptimal block source coding scheme for a lossy data compression based on the analysis of the waiting time N_ℓ. The authors of [28] were able to establish an upper bound on $\log N_\ell/\ell$, namely they proved that asymptotically $\log N_\ell/\ell \leq R(D/2)$ (pr.) for any stationary and ergodic sequence $\{X_k\}$. A more refined bound was obtained for the memoryless source (the so called Bernoulli model). Corollary 1 (cf. also [34]) below gives a precise limiting behavior of $\log N_\ell$ in terms of $r_0(D)$, and in our next finding we compute – among others – explicit formula for $r_b(D)$ for the Bernoulli model. The lower bound in Corollary 1 (which in fact is also true for the almost sure convergence) follows directly from (2), while for the upper bound we must use some arguments from the proof of the lower bound for $L_n(1)$. Hence, we delay the proof of Corollary 1 until Section 3.4.

Corollary 1. WAITING TIME. *The following holds*

$$\lim_{\ell\to\infty} \frac{\log N_\ell}{\ell} = r_0(D) \qquad \text{(pr.)} \tag{17}$$

under assumptions of Theorem 1(i), that is, $\lim_{g\to\infty}\alpha(g) = 0$ *and the rate of convergence in Lemma 1 is* $O(1/n^{1+\delta})$ *for some* $\delta > 0$.

One may wonder how the generalized Rényi's entropies depend on the parameters of the underlying model. Can we derive explicit formulas for $r_b(D)$ in the Bernoulli and/or Markovian model, as it was accomplished for the lossless case $D = 0$ (cf. [2, 30, 31])? The theorem below provides an explicit formula for $r_b(D)$ in the Bernoulli model (B). The proof can be found in Section 3.5.

Theorem 3. BERNOULLI MODEL. *Define* $h(D,x) = (1-D)\log((1-D)/x) + D\log(D/(1-x))$ *for* $0 < D, x < 1$. *Then:*
(i) *Let* $p_{\min} = \min\{p,q\}$ *and* $p_{\max} = \max\{p,q\}$, *then*

$$r_{-\infty}(D) = \begin{cases} h(D, p_{\min}) & \textit{for } D \le p_{\max} \\ 0 & \textit{for } D > p_{\max} \end{cases}.$$

and

$$r_{\infty}(D) = \begin{cases} h(D, p_{\max}) & \textit{for } D \le p_{\min} \\ 0 & \textit{for } D > p_{\min} \end{cases}.$$

In addition, $r_{-\infty}(D)$ *and* $r_\infty(D)$ *are convex functions of* D.

(ii) *If* $p = q = 1/2$ *then, for every* $-\infty \le b \le \infty$ *and* $D \le p_{\min}$, *we have* $r_b(D) = h(D, 1/2)$.

(iii) *Let* $p \ne q$ *and* $-\infty < b < \infty$. *Then,* $r_b(D) = 0$ *whenever* $D > 2pq$, *while for* $0 \le D \le 2pq$ *and* $b \ne 0$

$$\begin{aligned} r_b(D) &= (1/b) \min_{0 \le x \le 1} \Big\{ x\log(x/p) + (1-x)\log((1-x)/q) - b\Big(D\log(p/q) \\ &+ x\log(px) + (1-x)\log(q(1-x)) - x\log(x - F(x)) \\ &- D\log(D - F(x)) - (1-x-D)\log(1-x-D+F(x))\Big)\Big\}, \end{aligned} \tag{18}$$

where $F(x)$ *is defined as*

$$F(x) = \frac{x+D}{2} + \frac{\sqrt{(p^2 + (x+D)(q-p))^2 + 4xq^2 D(p-q)} - p^2}{2(p-q)}. \tag{19}$$

In particular, we have

$$r_1(D) = \begin{cases} h(D, P) & \textit{for } \ D \le 1 - P = 2pq \\ 0 & \textit{for } \ D > 1 - P = 2pq \end{cases}$$

where $P = p^2 + q^2$. *The function* $r_1(D)$ *is convex with respect to* D.

(iv) *If* $p \ne q$ *then* $r_0(D) = 0$ *for* $D > 2pq$, *and for* $0 \le D \le 2pq$

$$\begin{aligned} r_0(D) &= -\Big(D\log(p/q) + 2p\log p + 2q\log q - p\log(p - F(p)) \\ &\quad - D\log(D - F(p)) - (q-D)\log(q - D + F(p))\Big), \end{aligned} \tag{20}$$

where F *is the function defined by (19). In addition,* $r_0(D)$ *is convex with respect to* D.

Remark 3. *Degenerate Behaviors.* It is interesting to see how the functions s_n, H_n and L_n grow with D for the Bernoulli model. Theorem 3 states that for $D > 2pq$ we have $r_b(D) = 0$, so $H_n/\log n \to \infty$ and $L_n/\log n \to \infty$. It is not hard to find a reason behind such a behavior. If $D > 2pq = 1 - P$, then with probability tending to 1 as $k \to \infty$ the distance between every two randomly chosen strings of length k is less than D. Thus, for such a large D (a.s.) both H_n and L_n are of the order of n. Similarly, one can easily see that s_n is of the order of n whenever $D > p_{\max}$. $\Box$

A second-order improvement in the data compression scheme can be obtained if one implements a simple generalization of our construction (cf. [31] for $D = 0$) that might lead to a better compression code redundancy. The main idea is to search for the longest prefix of X_{n+1}^{∞} that occurs at least b times in the database, where b is a parameter. More precisely:

> Let $L_n^{(b)}$ be the largest K such that a prefix of X_{n+1}^{∞} of length K is within distance at most D from at least b disjoint substrings of X_1^n, i.e. there exist $i_1, i_2, \ldots, i_b$ such that $1 \le i_1 \le i_2 - K \le \ldots \le i_k - (b-1)K \le n - bK + 1$, and $d(X_{i_1}^{i_1-1+K}, X_{n+1}^{n+K}) \le D, \ldots, d(X_{i_k}^{i_k-1+K}, X_{n+1}^{n+K}) \le D$.

In a similar manner we define $L_n^{(b)}(M)$, $s_n^{(b)}$ and $H_n^{(b)} = \max_{1\le M\le n}\{L_n^{(b)}(M)\}$.

We can prove the following generalization of Theorem 1 and Theorem 3:

Generalized depth, height and shortest path. *Under appropriate assumptions of Theorems 2 we have*

$$\lim_{n\to\infty} \frac{s_n^{(b)}}{\log n} = \frac{1}{r_{-\infty}(D)} \qquad \text{(a.s.)} \qquad \lim_{n\to\infty} \frac{H_n^{(b)}}{\log n} = \left(1 + \frac{1}{b}\right)\frac{1}{r_b(D)} \ . \tag{21}$$

Furthermore, under hypotheses of Theorem 1, $L_n^{(b)}$ and $L_n^{(b)}(M)$ behave as expressed in (11)–(13). For example:

$$\lim_{n\to\infty} \frac{L_n^{(b)}(M)}{\log n} = \lim_{n\to\infty} \frac{L_n^{(b)}}{\log n} = \frac{1}{r_0(D)} \qquad \text{(pr.)} \tag{22}$$

for all bounded b.

Our findings concerning L_n and N_ℓ can be used to predict the performance of a lossy data compression scheme based on Wyner-Ziv algorithm (cf. [32]). In the **lossless case**, a data compression scheme works as follows (cf. [32]): After identifying the largest prefix of length L_n of X_{n+1}^{∞}, we encode it by a position of its occurrence in the database sequence X_1^n which costs $\log n$ bits, and its length which costs $\log L_n = O(\log\log n)$). Thus, the compression ratio

$$C = \frac{\text{length of the compression code in bits}}{\text{uncoded length of } L_n} \tag{23}$$

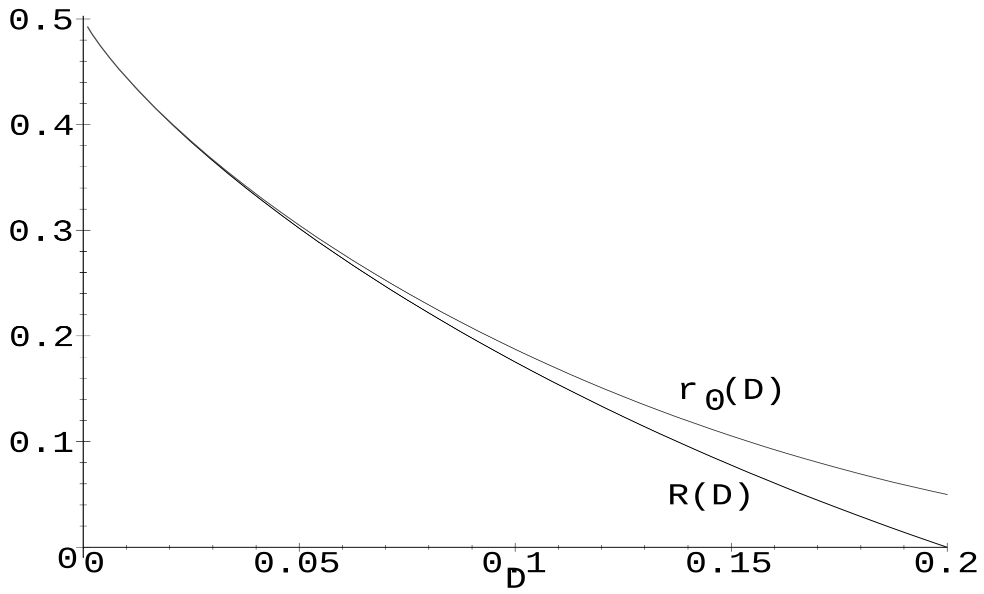

Figure 1: Comparison of compression rates for $p = 0.2$

becomes $C = h(1 + O(\log\log n/\log n)$ (pr.), which is asymptotically optimal for lossless compression.

In view of the above, one may ask how close is the rate (compression ratio) $r_0(D)$ of our scheme to the optimal compression ratio $R(D)$. For a memoryless source (i.e., Bernoulli model) it can be proved that $R(D) = h - h(D)$ where $h = -p\log p - q\log q$ is the entropy of the memoryless source, and $h(D) = -D\log D - (1-D)\log(1-D)$. Note that $R(0) = h$. From Theorem 3 we conclude that the scheme is:

- asymptotically optimal in the limiting case, namely

$$\lim_{D\to 0} R(D) = \lim_{D\to 0} r_0(D) = h \ , \tag{24}$$

- asymptotically optimal in the symmetric Bernoulli case ($p = q = 0.5$) since

$$r_0(D) = R(D) = \log 2 - h(D) \ . \tag{25}$$

In general, $r_0(D) > R(D)$. In Figure 1 the rate distortion function $R(D)$ and $r_0(D)$ are plotted versus D for a memoryless source with $p = 0.2$. While $r_0(D)$ is close to the optimal rate $R(D)$ for small D, one would still like to know whether an optimal scheme based on an approximate pattern matching exists. Recently, there have been some attempts in this direction (cf. [11, 17, 27]), though a definite answer was not yet established.

We end up this section with some remarks concerning algorithmic issues and applications of our scheme to lossy data compressions. With respect to the latter problem, one should observe that in the lossy case the decoder and encoder have different database (i.e., the decoder has as a database a sample of the distorted process), and this might lead to some complications. We can identify two general approaches to overcome this problem:

- **Adaptive Update of Database** (AUD) in which we keep the *same* database on both sides of a transmission channel that is updated either periodically or adaptively (e.g., still image compression) The undistorted (reference) database is transmitted faithfully (e.g., by Lempel-Ziv scheme). As a compression code we use the position in the database (cost: $\log n$ bits) and the length of approximately repeated pattern (cost: $\log L_n = O(\log\log n)$), thus according to Theorem 1 and (23) the compression ratio is $C = r_0(D)(1 + O(\log\log n/\log n)$ (pr.).

- **Distorted Version of Database** (DVD) in which we modify the original database X_1^n to a distorted one, say $\widetilde{X}_1^n$ that replaces substrings of length $O(\log n)$ of X_1^n by centers of a D-ball. The distorted database $\widetilde{X}_1^n$ is transmitted faithfully, say by Lempel-Ziv scheme or Wyner-Ziv scheme. We discuss this scheme in more details below. It should be pointed out that the idea of this scheme was first suggested by Steinberg and Gutman [28].

As mentioned above, when the database is varying quickly, Distorted Version Database (DVD) scheme is more appropriate. The algorithm of [28] can be used, and it is based on N_ℓ for a block source coding: Fix block length ℓ. Let y_{-1}^{-V} where $V = |\mathcal{A}|^\ell$ be a sequence generated by the information source. We append it at the left end, so the new sequence y_{-1}^{-2V} is of length $2V$. We first partition X_1^∞ into blocks of size ℓ, say $X_1(\ell), X_2(\ell)$, etc. For given block, say $X_i(\ell)$ we find the smallest $J(i)$ such that $X_i(\ell) \in B_D(y_{J(i)}(\ell))$ where $y_j(\ell) = y_{-\ell(j-1)-1}^{-\ell j}$. The distorted database is: $\widetilde{X} = \widetilde{X}_1(\ell)\widetilde{X}_2(\ell)\ldots$ where $\widetilde{X}_i(\ell) = y_{J(i)}(\ell)$. The distorted sequence $\widetilde{X}$ is send faithfully.

If we want to design a DVD version of our scheme based on L_n we need only some modifications. We use the sequence $\{y_i\}$ but we re-number it from 1 to $2V$, that is, our reference is y_1^{2V}. The sequence X_1^∞ is parsed into distorted variable blocks of length $l = O(\log n)$ as before. More precisely: We find the longest prefix of X that occurred approximately in $\{y_i\}$, say at position J, that is, $X_1^l \in B_D(y_J^{J-1+l})$ Then, we replace X_1^l by $\widetilde{X}_1^l = y_J^{J-1+l}$. And so on. The distorted sequences $\widetilde{X}$ is send faithfully. By our construction and Theorem 1 we conclude that the compression ratio is $r_0(D)$ with high probability.

Finally, we briefly address the algorithmic issues. In the lossless case, the prefix of length L_n can be found in $O(n)$ time-complexity (and $\mathcal{O}(\log n)$ on average) by a simple application of the suffix tree structure (cf. [30]). In the lossy case, the situation is more intricate. To simplify our discussion, let $Y_1^m = X_{n+1}^{n+m}$ be an uncompressed subsequence of length $m \le n$. An (intelligent) brute force algorithm can find the longest prefix of Y_1^m that approximately occurs in X_1^n in $O(nm)$ steps in the worst case. (Indeed, for every prefix of Y_1^m we check if it approximately occurs in X_1^n by comparing symbols of the involved substrings.) In the worst case, one should set $m = n$ which leads to $O(n^2)$ algorithm. However, based on Theorem 1 we can restrict $m = O(\log n)$, and then the average complexity of the algorithm is $\mathcal{O}(n \log n)$. While it seems to be an algorithmic challenge to be beat the $O(n^2)$ worst case complexity, Atallah, Genin and Szpankowski [4] reported several approximate algorithms of as good worst case complexity as $O(n \log^2 n)$ (cf. also [5, 22]).

3. ANALYSIS AND PROOFS

In this section we present proofs of Theorem 1 for the depth (cf. Section 3.1), Theorem 2 for the shortest path (cf. Section 3.2) and the height (cf. Section 3.3), Corollary 1 (cf. Section 3.4), and Theorem 3 for evaluating the Rényi entropies in the Bernoulli model (cf. Section 3.5).

Throughout we use the *first moment method* and the *second moment method* which we briefly review now. If Z_n is a sequence of nonnegative random variables, then for every n from Markov's and Chebyshev's inequalities we have (cf. [1])

$$\Pr\{Z_n > 0\} \le EZ_n \,, \tag{26}$$

$$\Pr\{Z_n = 0\} \le \frac{\mathrm{Var}\ Z_n}{(EZ_n)^2} \,. \tag{27}$$

In applications, Z_n is a function of a parameter k (e.g., length k of the depth L_n) such that for appropriately chosen k the average $EZ_n \to 0$ in the case (26), while for (27) one requests that Var $Z_n/(EZ_n)^2 \to 0$.

3.1 The Depth

We now prove Theorem 1(i) concerning the probabilistic behavior of the depth L_n. We start with an upper bound, and use the first moment method. Let Z_n be the number of positions $1 \le i \le n-k+1$ such that the prefix of X_i^{i+k-1} is within distance D from X_{n+1}^{n+k}, i.e.

$$Z_n = |\{1 \le i \le n-k+1 : \ d(X_i^{i+k-1}, X_{n+1}^{n+k}) \le D\}| \,.$$

The main idea behind our argument will be to condition on the structure of X_{n+1}^{n+k}. First observe that (8) of Lemma 1(i) can be translated into a generalization of the *Asymptotic Equipartition Property* (AEP) as follows: *For a mixing sequence X_1^n satisfying (A) with $\alpha(g)$ bounded and any fixed $\varepsilon > 0$, the state space $\mathcal{A}^n$ can be partitioned into two subsets $\mathcal{B}_n^\varepsilon$ ("bad set") and $\mathcal{G}_n^\varepsilon$ ("good set") such that sufficiently large n we have $P(\mathcal{B}_n^\varepsilon) \le \varepsilon$, and $2^{-nr_0(D)(1+\varepsilon)} \le P(B_D(x_1^n)) \le 2^{-nr_0(D)(1-\varepsilon)}$ for $x_1^n \in \mathcal{G}_n^\varepsilon$.* Thus, $\mathcal{B}_n^\varepsilon$ is the set of "bad" centers x_1^n for which either $P(B_D(x_1^n)) \ge 2^{-nr_0(D)(1-\varepsilon)}$ or $P(B_D(x_1^n)) \le 2^{-nr_0(D)(1+\varepsilon)}$.

Observe first that, unlike the lossless case ($D = 0$), in the lossy case the following can be claimed

$$\{L_n \ge k\} \quad \Longrightarrow \quad \exists_{\ell \ge k} \exists_{1 \le i \le n-\ell+1} \;\; d(X_{i+1}^{i+\ell}, X_{n+1}^{n+\ell}) \le D \; . \tag{28}$$

Then,

$$\begin{aligned}
\Pr\{L_n \ge k\} &\le \sum_{\ell \ge k} \sum_{i=1}^{n-\ell} \Pr\{d(X_{n+1}^{n+\ell}, X_i^{i+\ell-1}) \le D, \; X_{n+1}^{n+\ell} \in \mathcal{G}_\ell^{\varepsilon/2}\} + \sum_{\ell \ge k} P(\mathcal{B}_\ell^{\varepsilon/2}) \\
&= \sum_{\ell \ge k} \sum_{i=1}^{n-\ell} \sum_{w_\ell \in \mathcal{G}_\ell^{\varepsilon/2}} \Pr\{d(X_{n+1}^{n+\ell}, X_i^{i+\ell-1}) \le D, \; X_{n+1}^{n+\ell} = w_\ell\} + \sum_{\ell \ge k} P(\mathcal{B}_\ell^{\varepsilon/2}) \\
&\le \sum_{\ell=k}^{\infty} \sum_{i=1}^{n-\ell} (1 + \alpha(n+2-\ell-i)) 2^{-\ell r_0(D)(1-\varepsilon/2)} + \sum_{\ell \ge k} P(\mathcal{B}_\ell^{\varepsilon/2}) \\
&\le nC 2^{-kr_0(D)(1-\varepsilon/2)} + \sum_{\ell \ge k} P(\mathcal{B}_\ell^{\varepsilon/2}).
\end{aligned} \tag{29}$$

where $C > 0$ is a constant. Set $k = \lfloor (1+\varepsilon) r_0^{-1}(D) \log n \rfloor$, and assume $\sum_{\ell \ge k} P(\mathcal{B}_\ell^{\varepsilon/2}) \le \varepsilon$ for sufficiently large n (which holds for example when $P(\mathcal{B}_\ell^{\varepsilon/2}) = O(1/\ell^{1+\delta})$ for some $\delta > 0$), we finally arrive at

$$\Pr\{L_n \ge \lfloor (1+\varepsilon) r_0^{-1}(D) \log n \rfloor\} \le \frac{c}{n^{\varepsilon/2(1-\varepsilon)}} + \varepsilon$$

for some constant c. This completes the prove of an upper bound.

For the lower bound, we use the second moment method. Let $k = \lfloor (1-\varepsilon) r_0^{-1}(D) \log n \rfloor$, and define

$$Z_n' = |\{1 \le i \le n/(k+g) : \; d(X_{i(k+g)+1}^{(i+1)k+ig}, X_{n+1}^{n+k}) \le D\}|$$

and $g = \Theta(\log n)$ is a *gap* between $\lfloor n/(k+g) \rfloor = \Theta(n/\log n)$ *non-overlapping* substrings of length k. In words, instead of looking at all strings of length k we consider only $m = \lfloor n/(k+g) \rfloor$ strings with gaps of length g among them. These gaps are used to "weaken" dependency between the substrings of length k. Observe now that $\Pr\{L_n < k\} \le \Pr\{Z_n' = 0\}$. Indeed, if $Z'_n > 0$ then by definition $L_n \geq k$. Note also that

$$\begin{aligned}
\Pr\{Z'_n = 0\} &= \Pr\{Z'_n = 0, X_{n+1}^{n+k} \in \mathcal{G}_k^{\varepsilon/2}\} + \Pr\{Z'_n = 0, X_{n+1}^{n+k} \in \mathcal{B}_k^{\varepsilon/2}) \\
&\leq \sum_{w_k \in \mathcal{G}_k^{\varepsilon/2}} \Pr\{Z'_n = 0 | X_{n+1}^{n+k} = w_k\} \Pr\{w_k \in \mathcal{G}_k^{\varepsilon/2}\} + \Pr\{\mathcal{B}_k^{\varepsilon/2}\} \\
&\leq \sum_{w_k \in \mathcal{G}_k^{\varepsilon/2}} \Pr\{Z'_n(w_k) = 0\} \Pr\{w_k \in \mathcal{G}_k^{\varepsilon/2}\} + \Pr\{\mathcal{B}_k^{\varepsilon/2}\} \ , \qquad (30)
\end{aligned}$$

where

$$Z'_n(w_k) = |\{1 \leq i \leq n/(k+g) : \ d(X_{i(k+g)+1}^{(i+1)k+ig}, w_k) \leq D\}| \ ,$$

and $\Pr\{Z'_n(w_k) = 0\} = \Pr\{Z'_n = 0 | X_{n+1}^{n+k} = w_k\}$. Thus, it is suffices to show that $\Pr\{Z'_n(w_k) = 0\} \to 0$ *uniformly* for all $w_k \in \mathcal{G}_k^{\varepsilon/2}$. Hereafter, we assume that $w_k \in \mathcal{G}_k^{\varepsilon/2}$.

Let now $m = n/k = \Theta(n/\log n)$. From the definition of the set $\mathcal{G}_k^{\varepsilon/2}$ for every $w_k \in \mathcal{G}_k^{\varepsilon/2}$ we have

$$EZ'_n(w_k) = mP(B_D(w_k)) \geq m2^{-kr_0(D)(1+\varepsilon/2)} \geq c\frac{n^{\varepsilon/2(1+\varepsilon)}}{\log n} \qquad (31)$$

for a constant c. We now compute the variance Var $Z'_n(w_n)$ for $w_k \in \mathcal{G}_k^{\varepsilon/2}$. Let $Z_n^i = 1$ if w_k occurs approximately at position $i(k+g)$, otherwise $Z_n^i = 0$. Certainly, $Z'_n(w_n) = \sum_{i=1}^m Z_n^i$, and Var $Z'_n(w_n) = \sum_{i=1}^m \mathrm{Var}\ Z_n^i + \sum_{i=1}^m \sum_{i,j=1}^m \mathrm{Cov}(Z_n^i, Z_n^j)$. A simple algebra reveals that

$$\sum_{i=1}^m \mathrm{Var}\ Z_n^i \leq mEZ_n^i = mP(B_D(w_k)) = EZ'_n(w_k) \ . \qquad (32)$$

To compute the second term in the sum above, we split it as $\sum_{i,j=1}^n \mathrm{Cov}(Z_n^i, Z_n^j) = S_1 + S_2$ where

$$\begin{aligned}
S_1 &= \sum_{i=1}^m \sum_{|i-j| \leq n^{\varepsilon/4}} \mathrm{Cov}(Z_n^i, Z_n^j) \\
S_2 &= \sum_{i=1}^m \sum_{|i-j| \geq n^{\varepsilon/4}} \mathrm{Cov}(Z_n^i, Z_n^j) \ .
\end{aligned}$$

Observe that

$$\mathrm{Cov}(Z_n^i, Z_n^j) = \Pr\{Z_n^i Z_n^j = 1\} - \Pr\{Z_n^i = 1\}\Pr\{Z_n^j = 1\} \leq \Pr\{Z_n^i = 1\} = EZ_n^i \ .$$

Hence $S_1 \leq 2n^{\varepsilon/4} EZ'_n(w_k)$.

On the other hand, proceeding as the above and using the mixing condition from (A) we also have $\mathrm{Cov}(Z_n^i, Z_n^j) \leq \alpha(g)\Pr\{Z_n^i = 1\}\Pr\{Z_n^j = 1\}$ where $g = O(n^{\varepsilon/4})$. Thus, $S_2 \leq 2\alpha(g)(EZ'_n(w_k))^2$. Consequently, for every $w_k \in \mathcal{G}_k^{\varepsilon/2}$ we have ($\varepsilon < 1$)

$$\Pr\{Z'_n(w_k) = 0) \leq \frac{\mathrm{Var}\ Z'_n(w_k)}{(EZ'_n(w_k))^2} \leq 2\alpha(g) + O\left(\frac{n^{\varepsilon/4}}{EZ'_n(w_k)}\right) \leq 2\alpha(g) + O\left(\frac{\log n}{n^{\varepsilon/4}}\right) \ , \qquad (33)$$

and finally by (32) and the above we obtain

$$\Pr\{L_n < \lfloor (1-\varepsilon) r_0^{-1}(D) \log n \rfloor\} \to 0$$

as $n \to \infty$ which completes the lower bound.

The proof of Theorem 1(i) concerning $L_n(M)$ follows exactly the same path as above. To establish Theorem 1(ii) dealing with the almost sure convergence of $L_n(M)$ we first observe that $L_n(M)$ is a nondecreasing sequence of n (in contrast to L_n), that is, $L_{n+1}(M) \geq L_n(M)$. Taking into account the definition of $L_n(M)$ for fixed M, and using our rate of convergence for the upper and the lower bounds proved above, together with the Borel-Cantelli Lemma along an exponentially increasing skeleton such as $n_k = 2^k$, we obtain the almost sure convergence as in [2, 30, 31] provided $\sum_{n=1}^{\infty} \sum_{\ell \geq n} P(\mathcal{B}_\ell^\varepsilon) < \infty$. For example, the latter condition holds for sequences for which $P(\mathcal{B}_\ell^\varepsilon)$ decays exponentially with ℓ (e.g., sequences satisfying the blowing-up property discussed in Remark 2).

3.2 Shortest Path

We now deal with the shortest path s_n and establish Theorem 2(i). The proof is along the lines suggested in [30, 31]. Therefore, we only briefly sketch it.

We start with the upper bound which is quite simple in this case. Let $P_{\min}(k) = \min_{w_k \in \mathcal{A}^k}\{P(B_D(w_k))\}$. From Lemma 1 we conclude that $P_{\min}(k) \leq 2^{-kr_{-\infty}(D)(1-\varepsilon)}$ (more precisely: $\log P_{\min} \sim -kr_{-\infty}(D)$). Observe that – unlike the lossless case – by definition of s_n we have

$$\{s_n > \ell\} \quad \Longrightarrow \quad \exists_{k>\ell} \forall_{w_k \in \mathcal{A}^k} \exists_{1 \leq i \leq n+1} \; d(X_i^{i-1+k}, w_k) \leq D \; . \tag{34}$$

In words, if $s_n > \ell$ then there exists $k > \ell$ such that for each $w_k \in \mathcal{A}^k$ the ball $B_D(w_k)$ must contain at least one of the string X_i^{i-1+k} where $1 \leq i \leq n+1$. Thus, in particular, $\Pr\{s_n > \ell\} \leq (n+1) \sum_{k > \ell} P(B_D(w_k^{\min}))$, where $w_k^{\min}$ is a word from $\mathcal{A}^k$ for which $\log(P(B_D(w_k^{\min})) \sim -kr_{-\infty}(D)$. Hence, for $\ell = \lfloor (1+\varepsilon) r_{-\infty}^{-1}(D) \log n \rfloor$ we have

$$\Pr\{s > \ell\} \leq (n+1) \sum_{k>\ell} P_{\min}(k) = O(1/n^\varepsilon) \; . \tag{35}$$

The lower bound requires a bit more work. Let us set $k = \lfloor (1-\varepsilon) r_{-\infty}^{-1}(D) \log n \rfloor$ and consider a set of *non-overlapping* substrings of X_1^n of length $k = O(\log n)$ between which one inserts *gaps* of length $g = O(\log n)$. Thus, there are $m = \lfloor (n+1)/(k+g) \rfloor = O(n/\log n)$ substrings $\{X_{i(k+g)+1}^{(i+1)k+ig}\}_{i=1}^m$. We show that with probability tending to 1 as $n \to \infty$ for every $w_k \in \mathcal{A}^k$ one can find among these m substrings at least one which are within the distance D from w_k and consequently $s_n \geq k$.

Indeed, from the mixing condition from (A) we get that the probability that such an event does not hold is bounded from above by

$$\begin{aligned}\Pr\{s_n < k\} &\leq \Pr\{\bigcup_{w_k\in\mathcal{A}^k}\bigcap_{i=1}^{m}\left(X_{i(k+g)+1}^{(i+1)k+ig} \neq w_k\right)\} \\ &\leq \sum_{w_k\in\mathcal{A}^k}(1+\alpha(g))^m(1-P(B_D(w_k)))^m \leq 2^k(1+\alpha(g))^m(1-P_{\min}(k))^m .\end{aligned}$$

Thus, taking into account our condition (14) we immediately prove that

$$\Pr\{s_n < \lfloor (1-\varepsilon)r_{-\infty}^{-1}(D)\log n\rfloor\} \leq O(\exp(-n^{\varepsilon/2}/\log n))$$

which completes the proof of the convergence in probability of s_n.

The rate of convergence for the upper bound does not yet justify to apply Borel-Cantelli lemma. But, as before taking exponentially increasing skeleton such as $n_l = 2^l$, we obtain almost sure convergence for the shortest path.

3.3 The Height

We establish now Theorem 2 (ii). The lower bound for the Bernoulli model for the height follows directly from Theorem 2 of Arratia and Waterman [2], while for the Markovian model we must use Theorem 6 of [2].

The upper bound is more intricate (especially that there is a minor recoverable error in [2] which has some subtlies for the upper bound proof). To show this, let us estimate the probability of $\{H_n \geq k\}$. Observe that

$$\begin{aligned}\{H_n \geq k\} &= \bigcup_{\ell\geq k}\bigcup_{1\leq i<j\leq n+1}\{d(X_i^{i+\ell}, X_j^{j+\ell}) \leq D\} \\ &= \bigcup_{\ell\geq k}\left(\bigcup_{|i-j|\leq\ell}\{d(X_i^{i+\ell}, X_j^{j+\ell}) \leq D\} \cup \bigcup_{|i-j|>\ell}\{d(X_i^{i+\ell}, X_j^{j+\ell}) \leq D\}\right). \quad (36)\end{aligned}$$

Let us first estimate the second term of (36) which we denote as $T_2(k)$. We obtain in the sequel

$$\begin{aligned}T_2(k) &\leq \sum_{\ell\geq k}\sum_{|i-j|>\ell}\Pr\{d(X_i^{i+\ell}, X_j^{j+\ell}) \leq D\} \\ &= \sum_{\ell\geq k}\sum_{|i-j|>\ell}\sum_{w_\ell\in\mathcal{A}^\ell}\Pr\{d(X_i^{i+\ell}, X_j^{j+\ell}) \leq D,\ X_j^{j+\ell} = w_\ell\} \\ &\leq (n+1)^2(1+\alpha(|i-j|))\sum_{\ell\geq k}\sum_{w_\ell\in\mathcal{A}^\ell}\Pr\{d(X_i^{i+\ell}, w_\ell) \leq D\}P(w_\ell) \\ &\leq (n+1)^2(1+\alpha(|i-j|))\sum_{\ell\geq k}EP(B_D(X_1^\ell))\end{aligned}$$

Using Lemma 1 and setting $k = 2(1+\varepsilon)r_1^{-1}(D)\log n$, one immediately sees that $T_2(k) = O(1/n^{2\varepsilon})$. This is true for *any* mixing model with bounded $\alpha(g)$.

The first term in (36) is much harder to deal with. The main contribution to the probability of this term comes form self-overlaps of substrings of X_1^n. For the Bernoulli model, using Theorem 4 of Arratia and Waterman [2] we can estimate that the contribution of the self-overlaps is smaller than $n \log n 2^{-kr_1(D)/2}$, and for $k = 2(1+\varepsilon)r_1^{-1}(D)\log n$ we obtain $O(\log n/n^{\varepsilon})$. Unfortunately, there is no equivalence of Theorem 4 in [2] for the Markovian model, and the authors of [2] gave some good reasons why this is so. We conjecture, however, that (16) holds for $D \to 0$ (cf. [30, 31] for $D = 0$ case). If self-overlaps are ignored, then the upper bound works fine for the Markovian model, and together with Theorem 6 of [2] it proves (16).

3.4 Waiting Time

To prove Corollary 1, we observe that a lower bound for N_ℓ follows directly from property (2) and Theorem 1. Indeed, from $\{N_\ell \le n\} \subset \{L_n(1) \ge \ell\}$ of (2) we conclude that for $n = 2^{(1-\varepsilon)r_0(D)\ell}$ there exists $\delta > 0$ such that

$$\Pr\{\log N_\ell \le r_0(D)(1-\varepsilon)\} \le \Pr\left\{L_n(1) \ge (1+\delta)\frac{\log n}{r_0(D)}\right\} \to 0$$

where the convergence to zero of the latter probability follows from Theorem 1 (i.e., the upper bound on L_n proved in Section 3.1; cf. (29)). In order to derive an upper bound for N_ℓ, it is enough to argue as in the proof of the lower bound for L_n of Theorem 1(i). Thus, one should consider a random variable Z_n'' that counts the number of strings lying within distance D from X_1^ℓ that occur at places of X_1^n separated by gaps of length ℓ. Then, we use the second moment method as above to show that $Z_n'' > 0$ with probability tending to 1. Since the argument and all calculations are the same as in the case of the random variable Z_n' (only now we consider substrings of length precisely ℓ) we omit the details.

3.5 Rényi Entropies in the Bernoulli Model

In this section, we present explicit formulæ for $r_b(D)$ in the Bernoulli model, that is, we prove Theorem 3.

We must compute the probability of the D-ball $P(B_D(w_k))$. Consider first $b = -\infty$. It is not hard to see the $P(B_D(w_k))$ is minimized for $w_k = w_{\min}$, where $w_{\min}$ consists of symbols that appear with probability $p_{\min}$. Then

$$P(B_D(w_{\min})) = \sum_{j=0}^{kD} \binom{k}{j} p_{\min}^{k-j}(1-p_{\min})^j \ .$$

If $D > p_{\max}$ then the above sum tends to 1 as $k \to \infty$, and, consequently, $r_{-\infty}(D) = 0$. Suppose then that $0 \le D \le p_{\max}$. Then, the last term in the sum is the largest one. Furthermore, by Stirling's formula for every $x \in (0,1)$ we have

$$\binom{k}{xk} \sim \left(\frac{1}{(1-x)^{1-x}x^x}\right)^k . \tag{37}$$

Thus, for large k and $D \le p_{\max}$

$$\left(\left(\frac{p_{\min}}{1-D}\right)^{1-D}\left(\frac{1-p_{\min}}{D}\right)^{D}\right)^k \le P(B_D(w_{\min})) \le k\left(\left(\frac{p_{\min}}{1-D}\right)^{1-D}\left(\frac{1-p_{\min}}{D}\right)^{D}\right)^k,$$

and this leads to our formula on $r_{-\infty}(D)$. The entropy $r_\infty(D)$ can be computed in a similar manner.

In order to see (ii) it is enough to notice that Lemma 1(ii) and the above imply that for $D \le 1/2$, $p = q = 1/2$ and $-\infty < b < \infty$, we have

$$h(D, 1/2) = r_\infty(D) \le r_b(D) \le r_{-\infty}(D) = h(D, 1/2) .$$

Now, let $p \ne q$ and $-\infty < b < \infty$, $b \ne 0$. From the definition of the expectation, for $EP^b(B_D(X_1^k))$ we have

$$EP^b(B_D(X_1^k)) = \sum_{i=0}^{k}\binom{k}{i}p^i q^{k-i}\left(\sum_{j=0}^{\lfloor Dk\rfloor}\sum_{\ell=\max\{0,i+j-k\}}^{\min\{i,j\}}\binom{i}{\ell}\binom{k-i}{j-\ell}p^{i+j-2\ell}q^{k-i-j+2\ell}\right)^b ,$$

where i counts the number of ones in X_1^k, j stays for the overall number of mismatches and ℓ is the number of disagreements among ones. Let us look first at the sum

$$\begin{aligned} s(k,i) &= \sum_{j=0}^{\lfloor Dk\rfloor}\sum_{\ell=\max\{0,i+j-k\}}^{\min\{i,j\}}\binom{i}{\ell}\binom{k-i}{j-\ell}p^{i+j-2\ell}q^{k-i-j+2\ell} \\ &= \sum_{j=0}^{\lfloor Dk\rfloor}\sum_{\ell=\max\{0,i+j-k\}}^{\min\{i,j\}} r(k,i,j,\ell) . \end{aligned}$$

Since there are at most k^2 terms in the sum, certainly we have

$$\max_{j,\ell} r(k,i,j,\ell) \le s(k,i) \le k^2 \max_{j,\ell} r(k,i,j,\ell) .$$

(Note that all ratios which grows polynomially with k will disappear if we divide the logarithm of $s(k,i)$ by k, thus they will not affect the value of $r_b(D)$.) Similarly,

$$\max_i \binom{k}{i}p^i q^{k-i}s^b(k,i) \le EP^b(B_D(X_1^k)) \le k\max_i \binom{k}{i}p^i q^{k-i}s^b(k,i) .$$

Thus, if we use Stirling's formula (37) to estimate the binomial coefficients and set $i = xk$, $j = yk$, $\ell = zk$, we arrive at the following asymptotic formula for $EP^b(B_D(X_1^k))$

$$\max_{0\le x\le 1}\Big\{\Big(\frac{p^x q^{1-x}}{x^x(1-x)^{(1-x)}}\Big(\max_{A(x,D)}\Big\{\frac{p^{x+y-2z}q^{1-x-y+2z}x^x(1-x)^{1-x}}{z^z(x-z)^{x-z}(y-z)^{y-z}(1-x-y+z)^{1-x-y+z}}\Big\}\Big)^b\Big)^k\Big\},$$

where $A(x,D)\subset I\!R^2$ is defined as

$$A(x,D) = \{(y,z)\in I\!R^2 : 0\le y\le D,\ \max\{0, x+y-1\}\le z\le \min\{x,y\}\}\ .$$

Consequently, for the entropy

$$r_b(D) = \lim_{k\to\infty}\frac{-\log EP^b(B_D(X_1^k)))}{kb}$$

we get the following formula

$$\begin{aligned} r_b(D) &= (1/b)\min_{0\le x\le 1}\Big\{x\log(x/p) + (1-x)\log((1-x)/q) \\ &- b\max_{A(x,D)}\Big\{(x+y-2z)\log(p/q) + \log q + x\log x \\ &+ (1-x)\log(1-x) - z\log z - (x-z)\log(x-z) \\ &- (y-z)\log(y-z) - (1-x-y+z)\log(1-x-y+z)\Big\}\Big\}\ . \end{aligned} \tag{38}$$

A simple algebra reveals that (18) follows from (38). Indeed, let us assume first that $D > 2pq$. Then, the value of the maximum in (38) is 0 and is achieved for $y - z = (1-x)p$ and $z = qx$. Furthermore, the first two terms vanish for $x = p$. Hence, for such a D, we have $r_b(D) = 0$ for every $-\infty < b < \infty$.

If $D \le 2pq$ then the function which appears under the maximum in (38) grows with y, so we must put $y = D$. Furthermore, easy calculations show that to choose an optimal value of z one must solve the equation

$$(p-q)z^2 + (p^2 + (q-p)(x+D))z - xDq^2 = 0\ .$$

Thus, we should set $z = F(x)$, where F is defined by (19), and (18) follows.

In order to get $r_1(D)$ it is better to start directly from (38). As we have already observed, the maximum is achieved for $y = D$. Hence

$$\begin{aligned} r_1(D) = &- \max_{x,z}\Big\{2(x-z)\log(p/q) + D\log(p/q) + 2\log q - z\log z \\ &- (x-z)\log(x-z) - (D-z)\log(D-z) \\ &- (1-x-D+z)\log(1-x-D+z)\Big\}\ . \end{aligned} \tag{39}$$

It is convenient to maximize first with respect to x, setting $x-z=p^2(1-D)/(p^2+q^2)$, and then with respect to z, putting $z=D/2$. Then, elementary calculations give $r_1(D)=h(D,p^2+q^2)$.

Finally, let us notice that the case when $b=0$ can be easily deduced from (18). Indeed, for $b\to 0$ the sum of the first two terms $x\log(x/p)$ and $(1-x)\log((1-x)/q)$ must vanish, which is possible only for $x=p$. Thus, (20) follows.

ACKNOWLEDGEMENT.

It is our pleasure to acknowledge several discussions with Professor P. Shields that helped us to understand the difference between our model and the optimal one. Discussions with Professor J. Kieffer during the early stage of this research were very useful. We are particularly grateful to Professor E.H. Yang and one of the referees for pointing out some minor but embarrassing slips in our early arguments. We are also grateful to Professor M. Atallah for guiding us in the area of algorithms, and Mr. Y. Génin for implementing image compression scheme based on our idea. Finally, we wish to thank three anonymous referees whose remarks contributed greatly to the final version of the paper.

References

[1] N. Alon and J. Spencer, *The Probabilistic Method*, John Wiley&Sons, New York (1992).

[2] R. Arratia and M. Waterman, The Erdös-Rényi Strong Law for Pattern Matching with Given Proportion of Mismatches, *Annals of Probability*, 17, 1152-1169 (1989).

[3] R. Arratia, L. Gordon, and M. Waterman, The Erdös-Rényi Law in Distribution for Coin Tossing and Sequence Matching, *Annals of Statistics*, 18, 539-570 (1990)

[4] M. Atallah, Y. Genin, and W. Szpankowski, A Pattern Matching Approach to Image Compression, *Proc. International Conference on Image Processing*, Lausanne (1996); see also Purdue University, CSD-TR-95-083 (1995).

[5] M. Atallah, P. Jacquet and W. Szpankowski, Pattern matching with mismatches: A probabilistic analysis and a randomized algorithm, *Proc. Combinatorial Pattern Matching*, Tucson, LNCS 644, 27-40, Springer-Verlag 1992.

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

[7] P. Chou, M. Effros and R. Gray, A Vector Quantization Approach to Universal Noiseless Coding and Quantization, preprint (1995).

[8] C. Constantinescu and J. Storer, Improved Techniques for Single-Pass Adaptive Vector Quantization, *Proc. of the IEEE*, 82, 933-939 (1994).

[9] M. Effros and P. Chou, Weighted Universal Bit Allocation. *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, vol 4, 2343-2346, Detroit (1995).

[10] J. Feldman, r-Entropy, Equipartition, and Ornstein's Isomorphism Theory in R^n, *Israel J. Math.*, 86, 321-345 (1980).

[11] W. Finamore, M. Carvalho, and J. Kieffer, Lossy Compression with the Lempel-Ziv Algorithm, *11th Brasilian Telecommunication Conference*, 141-146 (1993).

[12] A. Frieze and W. Szpankowski, Greedy Algorithms for the Shortest Common Superstring That Are Asymptotically Optimal, *Proc. European Symposium on Algorithms*, Bercelona (1996).

[13] G. Hardy, J.E. Littlewood and G. Pólya, *Inequalities*, Cambridge University Press, Cambridge (1989)

[14] P. Jacquet and W. Szpankowski, Asymptotic Behavior of the Lempel-Ziv Parsing Scheme and Digital Search Trees, *Theoretical Computer Science*, 144, 161-197 (1995).

[15] J.C. Kieffer, Strong Converses in Source Coding Relative to a Fidelity Criterion, *IEEE Trans. Information Theory*, 37, 257-262 (1991).

[16] J.F.C. Kingman, *Subadditive Processes*, in Ecole d'Eté de Probabilités de Saint-Flour V-1975, Lecture Notes in Mathematics, 539, Springer-Verlag, Berlin 1976.

[17] H. Koga and S. Arimoto, Asymptotic Properties of algorithms of Data Compression with Fidelity Criterion Based on String Matching, *Proc. 1994 IEEE Information Symposium on Information Theory*, 24-25 (1994).

[18] T. Linder, G. Lugosi, and K. Zeger, Rates of Convergence in the Source Coding Theorem, in Empirical Quantizer Design, and in Universal Lossy Source Coding, *IEEE Information Theory*, 40, 1728-1740 (1994).

[19] T. Linder, G. Lugosi, and K. Zeger, Fixed-Rate Universal Lossy Source Coding and Rates of Convergence for Memoryless Sources, *IEEE Information Theory*, 41, 665-676 (1995).

[20] G. Louchard and W. Szpankowski, On the Average Redundancy Rate of the Lempel-Ziv Code, *IEEE Trans. Information Theory*, 43 (1997).

[21] G. Louchard, W. Szpankowski and J. Tang, Average Profile of Generalized Digital Search Trees and Lempel-Ziv Algorithm, Purdue University, CSD TR-96-005 (1996).

[22] T. Łuczak and W. Szpankowski, A Lossy Data Compression Based on String Matching: Preliminary Analysis and Suboptimal Algorithms, *Proc. Combinatorial Pattern Matching*, Asilomar, LNCS 807, 102-112, Springer-Verlag (1994).

[23] K. Marton and P. Shields, The Positive-Divergence and Blowing-up Properties, *Israel J. Math*, 80, 331-348 (1994).

[24] D. Ornstein and P. Shields, Universal Almost Sure Data Compression, *Annals of Probability*, 18, 441-452 (1990).

[25] D. Ornstein and B. Weiss, Entropy and Data Compression Schemes, *IEEE Information Theory*, 39, 78-83 (1993).

[26] E. Posner and E. Rodemich, Epsilon Entropy and Data Compression, *The Annals of Mathematical Statistics*, 42, 2079-2125 (1971).

[27] I. Sadeh, On Approximate String Matching, *Proc. of Data Compression Conference*, 148-157 (1993).

[28] Y. Steinberg and M. Gutman, An Algorithm for Source Coding Subject to a Fidelity Criterion, Based on String Matching, *IEEE Trans. Information Theory*, 39, 877-886 (1993).

[29] P. Shields, Waiting Times: Positive and Negative Results on the Wyner-Ziv Problem, *J. Theoretical Probability*, 6, 499-519 (1993).

[30] W. Szpankowski, Asymptotic Properties of Data Compression and Suffix Trees, *IEEE Trans. Information Theory*, 39, 1647-1659 (1993).

[31] W. Szpankowski, A Generalized Suffix Tree and Its (Un)Expected Asymptotic Behaviors, *SIAM J. Computing*, 22, 1176-1198 (1993).

[32] A. Wyner and J. Ziv, Some Asymptotic Properties of the Entropy of a Stationary Ergodic Data Source with Applications to Data Compression, *IEEE Trans. Information Theory*, 35, 1250-1258 (1989).

[33] E.H. Yang, and J. Kieffer, Simple Universal Lossy Data Compression Schemes Derived From Lempel-Ziv algorithm, *IEEE Trans. Information Theory*, 42, 239-245 (1996).

[34] E.H. Yang, and J. Kieffer, On the Performance of Data Compression Algorithms Based upon String Matching, preprint (1995).

[35] E.H. Yang and Z. Zhang, The Shortest Common Superstring Problem: Average Case Analysis for Both Exact Matching and Approximate Matching, preprint (1996).

[36] Z. Zhang and V. Wei, An On-Line Universal Lossy Data Compression Algorithm via Continuous Codebook Refinement – Part I: Basic Results, *IEEE Trans. Information Theory*, 42, 803-821 (1996).

[37] J. Ziv, Coding of Source with Unknown Statistics – Part II: Distortion Relative to a Fidelity Criterion, *IEEE Trans. Information Theory*, 18, 389-394 (1972).

[38] J. Ziv and A. Lempel, A Universal Algorithm for Sequential Data Compression, *IEEE Trans. Information Theory*, 23, 3, 337-343 (1977).

[39] J. Ziv and A. Lempel, Compression of Individual Sequences via Variable-rate Coding, *IEEE Trans. Information Theory*, 24, 530-536, 1978.