Pattern Matching and Lossy Data Compression on Random Fields

I. Kontoyiannis

December 16, 2002

Abstract — We consider the problem of lossy data compression for data arranged on two-dimensional arrays (such as images), or more generally on higher-dimensional arrays (such as video sequences). Several of the most commonly used algorithms are based on pattern matching: Given a distortion level D and a block of data to be compressed, the encoder first finds a D-close match of this block into some database, and then describes the data by describing the position of the match. We consider two idealized versions of this scenario. In the first one, the database is taken to be a collection of independent realizations of the same size and from the same distribution as the original data. In the second, the database is assumed to be a single long realization from the same source as the data. We show that the compression rate achieved (in either version) is no worse than $R(D/2)$ bits per symbol, where $R(D)$ is the rate-distortion function. This is proved under the assumptions that (1) the data is generated by a Gibbs distribution, and (2) the distortion measure is a metric, generalizing the corresponding one-dimensional bound of Steinberg and Gutman. Using recent large deviations results by Dembo and Kontoyiannis and by Chi, we are able to give short proofs for the present results.

Keywords — Rate-distortion theory, random fields, pattern matching, large deviations, lossy data compression.

[1]I. Kontoyiannis is with the Division of Applied Mathematics and the Department of Computer Science, Brown University, Box F, 182 George St., Providence, RI 02912, USA. Email: `yiannis@dam.brown.edu` Web: `www.dam.brown.edu/people/yiannis/`

[2]Research supported in part by NSF grants #0073378-CCR and DMS-9615444, and by USDA-IFAFS grant #00-52100-9615.

1 Introduction

Many types of data encountered in applications are naturally arranged on multi-dimensional arrays. For example, an m-by-n pixel image can be represented as a two-dimensional array $\{X_{ij} \ ; \ 1 \leq i \leq m, \ 1 \leq j \leq n\}$, with each X_{ij} denoting the pixel-value at location (i, j). More generally, we consider collections of data X_u indexed by points $u = (u_1, u_2, \ldots, u_d) \in \mathbb{Z}^d$, and arranged in a d-dimensional array $\{X_u \ ; \ u \in U\}$ for some, usually rectangular, subset $U \subset \mathbb{Z}^d$. Typical examples of applications involving multi-dimensional data include image and video processing, geostatistics, and statistical mechanics.

An important problem in applications is to develop efficient lossy data compression algorithms. For example, a huge amount of effort has been devoted to image and video compression over the past decade; see [13][4][25] and the references therein. One of the most commonly used ingredients of compression algorithms used in practice is the idea of *pattern matching*. Roughly speaking, this means that, instead of being described directly, certain parts of the data are described by pointers to locations where approximate versions of the data occur. In image compression, for example, an 8-by-8 block of an image to be compressed may be described by a pointer to an earlier part of the image (that has already been encoded) where an approximate version of the current 8-by-8 block appears. Similarly, in video compression it is common to describe an entire object in a given frame by describing the position of an approximate version of that same object in a previous frame.

The wide use of pattern matching in lossy compression is partly due to its intrinsic simplicity, and also has been motivated by the great practical success of the family of Lempel-Ziv algorithms for lossless compression. Constantinescu and Storer [6][7] have introduced extensions of the Lempel-Ziv algorithm for lossy image compression, and, more recently, Szpankowski and his collaborators [2][1] have developed numerous practical algorithms for image and video compression, producing extensive experimental results that demonstrate their performance. Moreover, one of the important features of many video compression algorithms is *motion-compensation*, an idea often implemented using pattern matching; see [25] and the texts [13][22] for details.

Despite the practical significance of pattern matching-based compression algorithms for multi-dimensional data, little has been rigorously established regarding their performance.[1] Thus motivated, we consider two idealized versions of this problem and we analyze the compression performance of two simple compression algorithms that we hope capture some of the essential elements of the corresponding practical schemes.

Suppose that data is generated by a random field $\boldsymbol{X} = \{X_u \ ; \ u \in \mathbb{Z}^d\}$ on the integer lattice $\mathbb{Z}^d$, taking values in a finite alphabet A. Let $C(n)$ denote the d-dimensional integer cube with side length n,

$$C(n) \stackrel{\triangle}{=} \{u = (u_1, u_2, \ldots, u_d) \in \mathbb{Z}^d \ ; \ 1 \leq u_j \leq n, \text{ for all } j\}. \tag{1}$$

Given a block of data $x^n \stackrel{\triangle}{=} \{x_u \ ; \ u \in C(n)\}$ generated by $\boldsymbol{X}$, the encoder's task is to find

[1] As far as we know there are no rigorous results, with the exception of [1] where a theoretical analysis is given but only for the corresponding one-dimensional model.

an efficient approximate representation of x^n by a different block $y^n \in A^{n^d}$. More precisely, we require that the distortion $\rho_n(x^n, y^n)$ between x^n and its representation y^n does not exceed some predetermined limit D (assuming that ρ_n is a single-letter distortion measure – see Section 2 for precise definitions).

Next we describe two idealized compression algorithms based on pattern matching.

I. *Shannon Codebook.* Suppose that the data $X^n = \{X_u \ ; \ u \in C(n)\}$ from $\boldsymbol{X}$ is to be compressed with distortion D or less, and let P_n denote the distribution of X^n. We assume that a codebook

$$Y^n(i), \quad i \geq 1$$

consisting of independent and identically distributed (i.i.d.) realizations from the distribution P_n is available to both the encoder and decoder. Let W_n denote the position i of the first element $Y^n(i)$ of the codebook that matches X^n with distortion D or less,

$$W_n \ = \ \inf\{i \geq 1 \ : \ \rho_n(X^n, Y^n(i)) \leq D\}, \tag{2}$$

with the convention that $W_n = \infty$ if no such match exists. Then the encoder can describe X^n to the decoder with distortion no more than D by describing the position W_n of this match. This is a simple variation on Shannon's original random code [23], with the important difference that the codebook distribution is taken to be the same as the source distribution.

II. *Fixed Database Coding.* Next we describe a d-dimensional lossy version of the Fixed-Database Lempel-Ziv (FDLZ) algorithm; cf. [27]. Here we assume that an infinitely long database $Y^\infty \stackrel{\triangle}{=} \{Y_u \ ; \ u_j \geq 1 \text{ for all } j\}$ is available to both the encoder and decoder, and that Y^∞ has the same distribution as the source. Given X^n to be compressed, the encoder looks for the smallest index m such that X^n appears with distortion D or less as a (contiguous) sub-block of Y^m. Let W'_n denote the smallest such m,

$$W'_n \ = \ \inf\{m \geq n \ : \ \rho_n(X^n, \widetilde{Y}^n) \leq D \ \text{ for some } \widetilde{Y}^n \subset_c Y^m\}, \tag{3}$$

where $\widetilde{Y}^n \subset_c Y^m$ means that $\widetilde{Y}^n$ is a contiguous sub-block of Y^m, and again with the convention that $W'_n = \infty$ if no such m exists. The encoder can then give a D-accurate description of X^n to the decoder by describing the coordinates of the location where the above match first occurs.

Because the database distribution in I and II above is taken to be the source distribution, we naturally expect that the compression performance of both of these schemes will be strictly suboptimal: In order to achieve the rate-distortion function of $\boldsymbol{X}$, the database should have the optimal reproduction distribution; see [28] and [9] for extensive discussions on this issue.

Our main result (Theorem 1 in Section 3) states that, although suboptimal, the compression rate achieved by either of these schemes in no worse than $R(D/2)$ bits per symbol. That is, to achieve the optimal rate we may need to allow for twice as much distortion, but no more. This

result generalizes the corresponding bound of Steinberg and Gutman [24] for one-dimensional sources. Theorem 1 is shown to hold for the class of all ergodic Gibbs fields – see Section 3 for the precise definition, and a discussion of this class of random fields. The proof we give here is completely different from that of the one-dimensional result in [24], and it uses recent large deviations results from [5] and [9].

For comparison, we recall that Steinberg and Gutman's result [24] was stated in the case of a coding scenario similar to II above, except that matches were found in *non*-overlapping blocks si that their waiting time W_n'' was the smallest m such that $\rho_n(X_1^n, Y_{mn+1}^{mn+n}) \leq D$. In [24] it is shown that, assuming the source is "totally ergodic," the asymptotic rate achieved in this scenario is bounded above by $R(D/2)$, in probability. Methods similar to those used in [24] can also be applied in the case of random fields to obtain corresponding bounds, but our purpose here is to show that conceptually simple proofs can be given if one relies on the recent large deviations results of [5] and [9], following the trend of a lot of the recent literature in rate-distortion theory; see [28][30][9] and the references therein.

In Section 2 we collect several background results about lossy data compression of random fields, and Section 3 contains our main result described above, together with its proof.

2 Lossy Data Compression on Random Fields

As the data source, we consider a random field $\boldsymbol{X} = \{X_u \ ; \ u \in \mathbb{Z}^d\}$ indexed by points $u = (u_1, u_2, \ldots, u_d)$ on the integer lattice $\mathbb{Z}^d$, $d \geq 2$, and taking values in a finite set, the *source alphabet* A. For any subset $U \subset \mathbb{Z}^d$ of size $|U|$, we write X_U for the block of random variables $\{X_u \ ; \ u \in U\}$, and similarly $x_U = \{x_u \ ; \ u \in U\} \in A^{|U|}$ for a realization of X_U. We denote the distribution of X_U by P_U, and write $\mathbb{P}$ for the measure describing the distribution of the entire random field $\boldsymbol{X}$. For $v, w \in \mathbb{Z}^d$ with $v_i \leq w_i$ for all i, we let $[v, w]$ denote the rectangle

$$[v, w] \stackrel{\triangle}{=} \{u \in \mathbb{Z}^d \ : \ v_i \leq u_i \leq w_i \ \text{ for all } i\},$$

and we write $C(n)$ for the d-dimensional cube $C(n) = [(1, \ldots, 1), (n, \ldots, n)]$ defined as in (1).

Let $\hat{A}$ denote a finite *reproduction alphabet.* Given $\rho : A \times \hat{A} \to [0, \infty)$, for every finite $U \subset \mathbb{Z}^d$ the single-letter distortion measure ρ_U on $A^{|U|} \times \hat{A}^{|U|}$ is defined by

$$\rho_U(x_U, y_U) \stackrel{\triangle}{=} \frac{1}{|U|} \sum_{u \in U} \rho(x_u, y_u), \qquad x_U \in A^{|U|}, \ y_U \in \hat{A}^{|U|}. \tag{4}$$

For simplicity we write $X^n = X_{C(n)}$ and $\rho_n(x^n, y^n) = \rho_{C(n)}(x_{C(n)}, y_{C(n)})$. We also make the usual assumption that for all $x \in A$ there is a $y \in \hat{A}$ such that $\rho(x, y) = 0$.

For every rectangle $U \subset \mathbb{Z}^d$, the *rate-distortion function* of X_U, is defined, for all $D \geq 0$, as

$$R_U(D) \stackrel{\triangle}{=} \inf I(X_U; Y_U),$$

where $I(X_U; Y_U)$ denotes the mutual information (in bits) between X_U and Y_U, and the infimum is taken over all jointly distributed (X_U, Y_U) with values in $A^{|U|} \times \hat{A}^{|U|}$, such that $X_U \sim P_U$ and $E[\rho_U(X_U, Y_U)] \leq D$.

Let $\{U_n\}$ be an increasing sequence of rectangles in $\mathbb{Z}^d$ such that $U_n \to \mathbb{Z}^d$, i.e., $U_n \subset U_{n+1}$ for all n, and $\cup_{n\geq 1} U_n = \mathbb{Z}^d$. The rate-distortion function of $\boldsymbol{X}$ is defined by

$$R(D) = \lim_{n\to\infty} \frac{1}{|U_n|} R_{U_n}(D)$$

whenever this limit exists and is independent of the choice of the sequence $\{U_n\}$.

For $U \subset \mathbb{Z}^d$ and $v \in \mathbb{Z}^d$ we let $v + U$ denote the translate

$$v + U \stackrel{\triangle}{=} \{v + U \ : \ u \in U\}.$$

The random field $\boldsymbol{X}$ is said to be *stationary* if X_U and X_{v+U} have the same distribution, for all $v \in \mathbb{Z}^d$ and all finite $U \subset \mathbb{Z}^d$. The rate-distortion function exists for all stationary random fields, as can be easily seen by an application of the multi-dimensional subadditivity lemma; see, e.g., [29, Lemma 5.2.1]. A direct proof is also given in [16].

The operational significance of $R(D)$ comes from the fact that it characterizes the best achievable performance of lossy compression algorithms. Specifically, we consider the general class of *variable-rate/variable-distortion codes.* Such a code for X_U is defined as a triplet (B_U, q_U, ψ_U) where:

(*a*) B_U is subset of $\hat{A}^{|U|}$, called the *codebook*;

(*b*) $q_U : A^{|U|} \to B_U$ is the *quantizer*;

(*c*) $\psi_U : B_U \to \{0,1\}^*$ is a uniquely decodable representation of the elements of B_U by finite-length binary strings.

The compression performance of such a code (B_U, q_U, ψ_U) is described by its length function

$$\ell_U(x_U) = \text{ length of } [\psi_U(q_U(x_U))] \quad \text{bits.}$$

Coding Theorem. Let $\boldsymbol{X}$ be a stationary random field.

($\Leftarrow$) For any finite rectangle $U \subset \mathbb{Z}^d$ and any code (B_U, q_U, ψ_U) with expected distortion $D = E[\rho_U(X_U, q_U(X_U))]$, we have

$$E[\ell_U(X_U)] \geq R_U(D) \geq |U|\, R(D) \quad \text{bits.}$$

($\Rightarrow$) If $\boldsymbol{X}$ is also ergodic,[2] then for every $D > 0$ and any $\epsilon > 0$ there is a sequence of codes $(B_{C(n)}, q_{C(n)}, \psi_{C(n)})$ on A^{n^d}, $n \geq 1$, such that

$$E[\rho_n(X^n, q_{C(n)}(X^n))] \to D \quad \text{as } n \to \infty, \tag{5}$$

[2]Ergodicity here means that the group of translations $\{T_u \ ; \ u \in \mathbb{Z}^d\}$ acts on $(A^{\mathbb{Z}^d}, \mathcal{A}^{\mathbb{Z}^d}, \mathbb{P})$ in an ergodic manner, where $\mathcal{A}$ is the set of subsets of A and $\mathcal{A}^{\mathbb{Z}^d}$ denotes the product σ-field generated by finite-dimensional cylinders; see [17] for details.

and also

$$\frac{\ell_{C(n)}(x^n)}{n^d} \leq R(D) + \epsilon, \quad \text{bits per symbol},$$

for all $x^n \in A^{n^d}$ and all n.

The proof of this theorem is a rather technical but straightforward extension of the corresponding one-dimensional argument. A proof outline is given in [16], where a somewhat stronger statement is given for the direct coding theorem: Instead of (5), it is shown that

$$\Pr\{\rho_n(X^n, q_{C(n)}(X^n)) > D\} \to 0, \quad \text{as } n \to \infty.$$

Although the existence of $R(D)$ and the validity of the coding theorem have been implicitly assumed for a while by various authors, it appears that these statements have not explicitly appeared in the literature before. As far as we know, the most complete account of what is known in this area is summarized by Berger, Shen and Ye in [3] and in Ye and Berger's recent monograph [29].

3 Pattern Matching: Main Results

In this section we state and prove our main result described in the introduction.

From now on we assume that the source and reproduction alphabets are the same, $A = \hat{A}$, and that the distortion measure ρ is a *metric* on A, i.e., $\rho(x, y) = \rho(y, x)$, it satisfies the triangle inequality, and $\rho(x, y) = 0$ if and only if $x = y$. This of course implies that, for each finite $U \subset \mathbb{Z}^d$, the corresponding single-letter distortion measure ρ_U defined in (4) is also a metric.

For the data source $\boldsymbol{X} = \{X_u \,;\, u \in \mathbb{Z}^d\}$ we assume that it is a stationary and ergodic *Gibbs field*. The class of Gibbs fields contains most of the random field models used in applications, including essentially all Markov random fields [14, Section 2.2]. See [14][26] for examples of applications in the areas of image processing and image analysis. Moreover, Gibbs fields are, in a certain sense, "dense" within the class of all stationary random fields [18].

Formally, Gibbs fields are defined in terms of stationary, summable, interaction potentials. An *interaction potential* is a collection of functions $\{\mathcal{H}_U\}$, where, for each finite $U \subset \mathbb{Z}^d$, $\mathcal{H}_U$ is a function $\mathcal{H}_U : A^{|U|} \to \mathbb{R}$. The interaction potential $\{\mathcal{H}_U\}$ is called *stationary* if, for all U and all $v \in \mathbb{Z}^d$, the functions $\mathcal{H}_U$ and $\mathcal{H}_{v+U}$ coincide on $A^{|U|}$. And $\{\mathcal{H}_U\}$ is called *summable* if the following series is finite,

$$\sum_{U\,:\,\mathbf{0}\in U} \max_{x_U \in A^{|U|}} |\mathcal{H}_U(x_U)| < \infty,$$

where the sum is taken over all finite subsets U of $\mathbb{Z}^d$ containing the origin $\mathbf{0} = (0, \ldots, 0)$. The random field $\boldsymbol{X}$ is a *Gibbs field with interaction potential* $\{\mathcal{H}_U\}$, if for every finite $U \subset \mathbb{Z}^d$ the conditional distribution $P_{U|U^c}$ of X_U given X_{U^c} can be written as

$$P_{U|U^c}(x_U|x_{U^c}) = Z^{-1} \exp\left\{ - \sum_{V\,:\,V\cap U\neq\emptyset} \mathcal{H}_V(x_V) \right\},$$

for any configuration $(x_u \ ; \ u \in \mathbb{Z}^d) \in A^{\mathbb{Z}^d}$, where the sum is over all finite $V \subset \mathbb{Z}^d$ such that $V \cap U \neq \emptyset$, and $Z = Z(U, x_{U^c})$ is simply the normalization constant. The existence of Gibbs fields for a given interaction potential $\{\mathcal{H}_U\}$ is well-known; see, e.g., [12][14].

Next we recall the two pattern-matching compression algorithms from the introduction. Suppose that $\boldsymbol{X}$ is a stationary and ergodic Gibbs field with distribution $\mathbb{P}$ on the alphabet A. We write, as before, X^n for the block of random variables $\{X_u \ ; \ u \in C(n)\}$, and let P_n denote the distribution of X^n.

I. *Shannon Codebook.* Suppose that an infinite codebook $\{Y^n(i) \ ; \ i \geq 1\}$ is available to both the encoder and decoder, where the $Y^n(i)$ are i.i.d. blocks of random variables, each with distribution P_n, and generated independently from X^n. Given the data X^n to be compressed, and given a distortion level D, the encoder searches the codebook for the first $Y^n(i)$ that matches X^n with distortion D or less. Let W_n denote the position of this first match (see (2)). Then the encoder describes X^n to the decoder (with distortion no more than D) by describing W_n; the decoder can read $Y^n(W_n)$ from the database, obtaining a D-close version of X^n. This description can be given using

$$\ell_n(X^n) = \log W_n + O(\log \log W_n) \quad \text{bits}, \tag{6}$$

cf. [11], where 'log' denotes the logarithm taken to base 2.

II. *Fixed Database Coding.* Alternatively, suppose that an infinitely long database $Y^\infty = \{Y_u \ ; \ u_j \geq 1 \text{ for all } j\}$ is available to the encoder and decoder, where Y^∞ has the same distribution as the source $\boldsymbol{X}$ and is independent of $\boldsymbol{X}$. Here, the encoder looks for the side-length m of the smallest cube $C(m)$ such that $Y^m = Y_{C(m)}$ contains a D-close version of X^n as a contiguous sub-block. Let W'_n denote this smallest m (see (3)). The encoder then describes the position (within $C(m)$) where X^n appears with distortion no greater than D. Since each coordinate of this position is no larger than W'_n, this can be done using

$$\ell'_n(X^n) = d \log W'_n + O(\log \log W'_n) \quad \text{bits}. \tag{7}$$

Our main result, given next, states that in both of these versions the asymptotic compression rate achieved is no worse that $R(D/2)$ bits per symbol. Recall that the rate-distortion function $R(D)$ is equal to zero for D greater than

$$D_{\max} \stackrel{\triangle}{=} \min_{y \in A} E[\rho(X_{\mathbf{0}}, y)].$$

Theorem 1. Let $\boldsymbol{X}$ be a stationary ergodic Gibbs field. For any $D \in (0, D_{\max})$ the compression rate achieved by algorithm I satisfies

$$\limsup_{n\to\infty} \frac{1}{n^d}\, \ell_n(X^n) \le R(D/2) \quad \text{bits per symbol,} \tag{8}$$

with probability one. If, in addition, the ϕ-mixing coefficients of $\boldsymbol{X}$ satisfy

$$\limsup_{n\to\infty} \sum_{j=1}^{\infty} (j+1)^{d-1} \phi_{n^d}(jn) < \infty,$$

then for any $D \in (0, D_{\max})$ the compression rate achieved by algorithm II also satisfies

$$\limsup_{n\to\infty} \frac{1}{n^d}\, \ell'_n(X^n) \le R(D/2) \quad \text{bits per symbol,} \tag{9}$$

with probability one.

Recall that the *(non-uniform) ϕ-mixing coefficients* of a stationary random field $\boldsymbol{X}$ with distribution $\mathbb{P}$ are defined by

$$\begin{aligned}\phi_\ell(k) \;=\; \sup\{|\mathbb{P}(B|A) - \mathbb{P}(B)| \;:\; B \in \sigma(X_U),\; A \in \sigma(X_V),\; \mathbb{P}(A) > 0 \\ |U| \le \ell,\; |V| < \infty,\; d(U,V) \ge k\}\end{aligned}$$

where $\sigma(X_U)$ denotes the σ-field generated by the random variables X_U, $U \subset \mathbb{Z}^d$, and the distance $d(U,V)$ between two subsets of $\mathbb{Z}^d$ is defined as

$$d(U,V) \stackrel{\triangle}{=} \inf_{u\in U,\, v\in V} \max_{1\le i\le d} |u_i - v_i|.$$

Note that the mixing condition (9) is satisfied by a large class of Gibbs fields. For example, it can be easily verified that (9) holds for all Markov random fields that satisfy Dobrushin's uniqueness condition; see [12, Section 8.2] or [14, Section 2.1], and also [10] or [20, Chapter 6] for more detailed discussions of the coefficients $\phi_\ell(k)$ and their properties.

3.1 Proof

First we consider the algorithm of version I. The result in (8) will be proved in two steps.

Step 1. We will show that

$$\ell_n(X^n) = n^d R(\mathbb{P}, \mathbb{P}, D) + o(n^d) \quad \text{a.s.,} \tag{10}$$

where, for any two stationary measures $\mathbb{P}$ and $\mathbb{Q}$ on $A^{\mathbb{Z}^d}$ (equipped with the natural product σ-field), the rate-function $R(\mathbb{P}, \mathbb{Q}, D)$ is defined as in [9] by

$$R(\mathbb{P}, \mathbb{Q}, D) = \lim_{n\to\infty} R_n(P_n, Q_n, D),$$

whenever this limit exists, where Q_n denote the $C(n)$-marginals of $\mathbb{Q}$ on A^{n^d}, and the rate-functions $R_n(P_n, Q_n, D)$ are defined as

$$R_n(P_n, Q_n, D) = \inf_{V_n} \frac{1}{n^d} H(V_n \| P_n \times Q_n),$$

with $H(V_n \| V_n')$ denoting the relative entropy (in bits) between V_n and V_n', and with the infimum taken over all joint distributions V_n on $A^{n^d} \times A^{n^d}$ such that the first marginal of V_n is P_n and $E_{V_n}[\rho_n(X^n, Y^n)] \leq D$.

From (6) we observe that, asymptotically, the main contribution to $\ell_n(X^n)$ will come from the term $(\log W_n)$. Repeating the "strong approximation" argument as in the proof of [15, Theorem 8] or [28, Lemma 1], we easily get that

$$\log W_n = -\log P_n(B(X^n, D)) + O(\log n) \quad \text{a.s.}, \tag{11}$$

where $B(X^n, D)$ denotes the distortion-ball of radius D around X^n,

$$B(x^n, D) = \left\{ y^n \in A^{n^d} \ : \ \rho_n(x^n, y^n) \leq D \right\}, \quad x^n \in A^{\mathbb{Z}^d}.$$

Further, [5, Theorem 1] states that for almost every infinite realization $(x_u \, ; \, u \in \mathbb{Z}^d)$ of $\boldsymbol{X}$, the random variables $\{\rho_n(x^n, Y^n)\}$ satisfy a large deviations principle with a deterministic, convex rate-function. In view of this, [9, Theorem 25] implies that

$$-\log P_n(B(X^n, D)) = n^d R(\mathbb{P}, \mathbb{P}, D) + o(n^d) \quad \text{a.s.} \tag{12}$$

Now let us check that D lies in the range within which we can indeed apply this theorem: Since we assumed that $D < D_{\max}$, and $D_{\max}$ is obviously bounded above by $E_{P_1 \times P_1}[\rho(X, Y)]$, we have $D < E_{P_1 \times P_1}[\rho(X, Y)]$. Moreover, taking $V_n(x^n, y^n) = P_n(x^n) \mathbb{I}_{\{y^n = x^n\}}$ in the definition of $R_n(P_n, P_n, D)$ yields

$$R_n(P_n, P_n, D) \leq \frac{1}{n^d} H(X^n)$$

where $H(X^n)$ denotes the entropy of X^n (in bits). This implies that, for all n, we have $R_n(P_n, P_n, D) \leq \log |A|$, and, therefore, if we define

$$D_{\min}^{(\infty)} \stackrel{\triangle}{=} \inf\{D \geq 0 : \sup_{n \geq 1} R_n(P_n, P_n, D) < \infty\},$$

then $D_{\min}^{(\infty)} = 0$, and so we have $D > D_{\min}^{(\infty)}$. We have thus shown that

$$D_{\min}^{(\infty)} < D < E_{P_1 \times P_1}[\rho(X, Y)],$$

so that [9, Theorem 25] applies.

Combining (11) with (12) and substituting in (6), yields (10) and completes the proof of step 1.

Step 2. Here we show that

$$R(\mathbb{P}, \mathbb{P}, D) \leq R(D/2). \tag{13}$$

Letting $R_n(D/2)$ denote $R_{C(n)}(D/2)$, an easy calculation (see equation (13) in [15]) shows that, for all $n \geq 1$,

$$R_n(D/2) = \inf_{Q_n} R_n(P_n, Q_n, D/2)$$

where the infimum is over all probability measures Q_n on A^{n^d}. This infimum is always achieved, although not necessarily uniquely, by some Q_n^*, so that $R_n(D/2) = R_n(P_n, Q_n^*, D/2)$; cf. [15, Proposition 2]. Moreover, by [15, Proposition 1] the infimum in the definition of $R_n(P_n, Q_n^*, D/2)$ is achieved by some V_n^* such that

$$E_{V_n^*}[\rho_n(X^n, Y^n)] \leq D/2, \tag{14}$$

so that, in fact,

$$R_n(D/2) = H(V_n^* \| P_n \times Q_n^*).$$

Therefore, in order to prove (13) it suffices to show that, for any n,

$$R_n(P_n, P_n, D) \leq H(V_n^* \| P_n \times Q_n^*). \tag{15}$$

We do this by constructing a triplet $(X^n, Z^n, \hat{Y}^n)$ such that $X^n \sim P^n$, and X^n and $\hat{Y}^n$ are conditionally independent given Z^n.[3] Specifically, let V_n denote the conditional distribution of Y^n given X^n induced by V_n^*, and assume that the conditional distribution of Z^n given X^n is V_n. Similarly, let $\overleftarrow{V}_n$ denote the conditional distribution of X^n given Y^n induced by V_n^*, and assume that the conditional distribution of $\hat{Y}^n$ given Z^n us $\overleftarrow{V}_n$. Then we obviously have that the three marginals of $(X^n, Z^n, \hat{Y}^n)$ are P_n, Q_n^* and P_n, respectively.

Let μ_n denote the joint distribution of all three $(X^n, Z^n, \hat{Y}^n)$, and write ν_n for the joint distribution of $(X^n, \hat{Y}^n)$. Observe that

$$\begin{aligned} E_{\nu_n}[\rho_n(X^n, \hat{Y}^n)] &\overset{(a)}{\leq} E_{\mu_n}[\rho_n(X^n, Z^n) + \rho_n(Z^n, \hat{Y}^n)] && (16)\\ &= 2\, E_{V_n^*}[\rho_n(X^n, Y^n)] && (17)\\ &\overset{(b)}{\leq} D, && (18) \end{aligned}$$

where (a) follows since ρ_n is a metric, and (b) follows from (14). Since the first marginal (in fact both marginals) of ν_n is P_n, we have

$$R_n(P_n, P_n, D) \leq H(\nu_n \| P_n \times P_n). \tag{19}$$

[3] After this paper was submitted for publication, the Associate Editor kindly pointed out to us that a very similar argument was independently discovered by Zamir and Rose in [30] and used to give an alternative proof of the Steinberg-Gutman result in one dimension.

But expanding ν_n as an average over all possible values of Z^n, we get

$$\begin{aligned} H(\nu_n \| P_n \times P_n) &= E_{P_n}[H(\nu_n(\cdot|X^n)\|P_n(\cdot))] \\ &= \sum_{x^n} P_n(x_n) \left[H\Big(\textstyle\sum_{z^n} V_n(z^n|x^n) \overleftarrow{V}_n(\cdot|z^n) \Big\| P_n(\cdot) \Big) \right] \\ &\stackrel{(a)}{\leq} \sum_{x^n, z^n} P_n(x_n) V_n(z^n|x^n) H(\overleftarrow{V}_n(\cdot|z^n)\|P_n(\cdot)) \\ &= E_{Q_n^*}[H(\overleftarrow{V}_n(\cdot|Z^n)\|P_n(\cdot))] \\ &= H(V_n^* \| P_n \times Q_n^*), \end{aligned}$$

where (a) follows from the convexity of relative entropy [8, Theorem 2.7.2]. Combining this with (19) gives (15), and, as discussed above, proves (13) and completes the proof of step 2. Finally combining steps 1 and 2 yields (8).

The proof of the corresponding result (9) for version II of the algorithm is similar, only requiring a modification to step 1 (step 2 is just a formal inequality, having nothing to do with whether the underlying compression algorithm is the one from version I or II). Proceeding as before with W_n' in place of W_n, instead of the "strong approximation" argument invoked above now we need to appeal to [9, Theorem 26], stating that

$$d \log W_n' = -\log P_n(B(X^n, D)) + O(\log n) \quad \text{a.s.}$$

This combined with (7) and with (12), gives us that

$$\ell_n'(X^n) = n^d R(\mathbb{P}, \mathbb{P}, D) + o(n^d) \quad \text{a.s.},$$

and combining this with step 2 completes the proof of (9), and hence of the theorem. □

Finally, we make a few brief remarks about the history of the method of proof above. First, in the one-dimensional case, the idea of considering distortion balls for approximate string matching in the context of lossy data compression was first employed by Łuczak and Szpankowski in [21]; soon after that, Yang and Kieffer [28] evaluated the limiting rate $R(\mathbb{P}, \mathbb{Q}, D)$ using large deviations techniques. They only consider Markov codebook distributions, a class more restricted than the "totally ergodic" distributions of Steinberg and Gutman, but their results are more precise and more general in that they actually identify the limiting rate exactly, rather than deriving an upper bound. The same strategy was followed in Step 1 of the above proof.

Approximate pattern matching for compression of random fields seems to have first been considered in [9], which also contains a review of the corresponding one-dimensional results. The strong-approximation theorem of [9] combined with the large deviations results of Chi [5] form the basis of Step 1 of our proof, and, as mentioned earlier, Step 2 is similar to an argument in [30].

Acknowledgments

We wish to thank Wojtek Szpankowski and Gershom Kutliroff for pointing out several relevant references, and Toby Berger for sending us copies of [3] and [19]. We also thank the referees for many useful comments, and the associate editor for his very close reading of the paper.

References

[1] M. Alzina, W. Szpankowski, and A. Grama. 2D-pattern matching image and video compression. *IEEE Trans. Image Processing*, 11:318–331, 2002.

[2] M. Atallah, Y. Génin, and W. Szpankowski. Pattern matching image compression: Algorithmic and empirical results. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 21:618–627, 1999.

[3] T. Berger, S.Y. Shen, and Z.X. Ye. Some communication problems of random fields. *Internat. J. Math. Statist. Sci.*, 1(1):47–77, 1992.

[4] A.C. Bovik and J.D Gibson. *Handbook of Image and Video Compression.* Academic Press, 2000.

[5] Z. Chi. Conditional large deviation principle for finite state Gibbs random fields. *Preprint*, October 2001.

[6] C. Constantinescu and J.A. Storer. On-line adaptive vector quantization with variable size codebook entries. *Information Processing and Management*, 30(6):745–758, 1994.

[7] C. Constantinescu and J.A. Storer. Application of single-pass adaptive VQ to BiLevel images. In *Proc. Data Compression Conf. – DCC 95.* IEEE, 1995.

[8] T.M. Cover and J.A. Thomas. *Elements of Information Theory.* J. Wiley, New York, 1991.

[9] A. Dembo and I. Kontoyiannis. Source coding, large deviations, and approximate pattern matching. *IEEE Trans. Inform. Theory*, 48:1590–1615, June 2002.

[10] P. Doukhan. *Mixing: Properties and Examples.* Springer-Verlag, New York, 1994.

[11] P. Elias. Universal codeword sets and representations of the integers. *IEEE Trans. Inform. Theory*, 21:194–203, 1975.

[12] H.-O. Georgii. *Gibbs Measures and Phase Transitions.* W. de Gruyter: Berlin et al, 1989.

[13] J.D. Gibson, R.L. Baker, T. Berger, T. Lookabaugh, and D. Lindbergh. *Digital Compression for Multimedia.* Morgan Kaufmann Publishers, San Fransisco, CA, 1998.

[14] X. Guyon. *Random Fields on a Network: Modeling, Statistics, and Applications.* Springer-Verlag, New York, 1995.

[15] I. Kontoyiannis. Pointwise redundancy in lossy data compression and universal lossy data compression. *IEEE Trans. Inform. Theory*, 46(1):136–152, January 2000.

[16] I. Kontoyiannis. Pattern matching and lossy data compression on random fields. Technical Report APPTS 01-6, Division of Applied Mathematics, Brown University, November 2001. [Available from www.dam.brown.edu/people/yiannis].

[17] U. Krengel. *Ergodic Theorems*. Walter de Gruyter & Co., Berlin, 1985.

[18] H. Künsch, S. Geman, and A. Kehagias. Hidden Markov random fields. *Ann. Appl. Probab.*, 5(3):577–602, 1995.

[19] T.-A. Lee. On the rate distortion function of the Ising model. Master's thesis, Dept. of Electrical Engineering, Cornell University, 1984.

[20] Z. Lin and C. Lu. *Limit Theory for Mixing Dependent Random Variables*. Kluwer Academic Publishers, Dordrecht, 1996.

[21] T. Łuczak and W. Szpankowski. A suboptimal lossy data compression algorithm based on approximate pattern matching. *IEEE Trans. Inform. Theory*, 43(5):1439–1451, 1997.

[22] K. Sayood. *Introduction to Data Compression*. Morgan Kaufmann Publishers, second edition, 2000.

[23] C.E. Shannon. Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec.*, part 4:142–163, 1959. Reprinted in D. Slepian (ed.), *Key Papers in the Development of Information Theory*, IEEE Press, 1974.

[24] Y. Steinberg and M. Gutman. An algorithm for source coding subject to a fidelity criterion, based upon string matching. *IEEE Trans. Inform. Theory*, 39(3):877–886, 1993.

[25] T. Wiegand and B. Girod. *Multi-Frame Motion-Compensated Prediction for Video Transmission*. Kluwer Academic Publishers, 2001.

[26] G. Winkler. *Image Analysis, Random Fields and Dynamic Monte Carlo Methods: A Mathematical Introduction*. Springer-Verlag, Berlin, 1995.

[27] A.D. Wyner and J. Ziv. Fixed data base version of the Lempel-Ziv data compression algorithm. *IEEE Trans. Inform. Theory*, 37(3):878–880, 1991.

[28] E.-h. Yang and J.C. Kieffer. On the performance of data compression algorithms based upon string matching. *IEEE Trans. Inform. Theory*, 44(1):47–65, 1998.

[29] Z. Ye and T. Berger. *Information Measures for Discrete Random Fields*. Science Press, Beijing, 1998.

[30] R. Zamir and K. Rose. Natural type selection in adaptive lossy compression. *IEEE Trans. Inform. Theory*, 47(1):99–111, 2001.