Universal Graph Compression: Stochastic Block Models

Alankrita Bhatt*†, Ziao Wang*‡, Chi Wang§, and Lele Wang‡
†University of California San Diego, La Jolla, CA 92093, USA, a2bhatt@eng.ucsd.edu
‡University of British Columbia, Vancouver, BC V6T1Z4, Canada, {ziaow, lelewang}@ece.ubc.ca
§Microsoft Research, Redmond, WA 98052, USA, wang.chi@microsoft.com

2021 IEEE International Symposium on Information Theory (ISIT) | 978-1-5386-8209-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISIT45174.2021.9517737

***Abstract*—Motivated by the prevalent data science applications of processing large-scale graph data such as social networks, web graphs, and biological networks, as well as the high I/O and communication costs of storing and transmitting such data, this paper investigates universal compression of data appearing in the form of a labeled graph. In particular, we consider a widely used random graph model, stochastic block model (SBM), which captures the clustering effects in social networks. A universal graph compressor is proposed, which achieves the optimal compression rate for a wide family of SBMs with edge probabilities from $O(1)$ to $\Omega(1/n^{2-\epsilon})$ for any $0 < \epsilon < 1$.**

Existing universal compression techniques are developed mostly for stationary ergodic one-dimensional sequences with entropy linear in the number of variables. However, the adjacency matrix of SBM has complex two-dimensional correlations and sublinear entropy in the sparse regime. These challenges are alleviated through a carefully designed transform that converts two-dimensional correlated data into *almost* i.i.d. blocks. The blocks are then compressed by a Krichevsky–Trofimov compressor, whose length analysis is generalized to *arbitrarily correlated* processes with identical marginals.

I. Introduction

Consider a discrete-time stochastic process $Z_1, \ldots, Z_N$ belonging to a class of processes $\mathscr{P}$, such as independent and identically distributed (i.i.d.) processes or stationary ergodic processes. The problem of *universal compression* deals with finding compression schemes that, without depending on the statistics of the process $Z_1, \ldots, Z_N$, achieve a per-symbol length converging to the entropy rate (when it exists) for every process in $\mathscr{P}$. Universal compression for sequences is a well-studied problem with asymptotically optimal compression schemes known for a large class of discrete processes such as m-ary i.i.d. processes [1, 2], stationary ergodic processes [3, 4, 5] and finite memory processes [6].

Consider now the problem of universal compression of a random *graph*. For an n-vertex simple random graph with adjacency matrix A_n, we wish to compress the $\binom{n}{2}$ entries in the upper triangle of the adjacency matrix of this graph without knowledge of the underlying distribution generating the graph, as long as it belongs to a certain class of distributions. This problem is inspired by emerging data science applications where data to be stored and processed appears in the form of a graph. Some of these applications include the study of social and biological networks.

In light of the previous discussion on universal compression, a natural question thus arises: can we convert the two-dimensional adjacency matrix of the graph into a one-dimensional sequence in some order and apply a universal compressor for the sequence? For some simple graph model such as Erdős–Rényi graph, where each edge is generated i.i.d. with probability p, this would indeed work, and we could use the optimal compressor for binary i.i.d. processes. However, the problem becomes more challenging when the graph is generated by complex distributions. In particular, when the graph is drawn from a stochastic block model (SBM), a canonical random graph model that captures cluster behavior, the bits in the upper triangle of A_n exhibit complex two-dimensional correlations. For general SBMs, it is unclear whether there is an ordering of the entries that results in a stationary process. We show in the full version of this paper [7] several orders including row-by-row, column-by-column, and diagonal-by-diagonal fail to produce a stationary process. Thus, we need to construct novel compression schemes for this problem with correlated two-dimensional data.

In this paper, we construct universal compression schemes for stochastic block models, i.e. compression schemes that do not depend on the parameters generating the SBM and achieve the entropy of the graph (defined as the entropy of A_n) asymptotically. We define the problem formally in the next subsection.

A. Problem Setup

For simplicity, we focus on simple (undirected, unweighted, no self-loop) graphs with labeled vertices in this paper. Let $\mathcal{A}_n$ be the set of all labeled simple graphs on n vertices. Let $\{0,1\}^i$ be the set of binary sequences of length i, and set $\{0,1\}^* = \cup_{i=0}^{\infty}\{0,1\}^i$. A lossless graph compressor $C\colon \mathcal{A}_n \to \{0,1\}^*$ is a one-to-one function that maps a graph to a binary sequence. Let $\ell(C(A_n))$ denote the length of the output sequence. When A_n is generated

*Alankrita Bhatt and Ziao Wang contributed equally to this work. This work was supported in part by the National Science Foundation under Grant CCF-1911238, in part by the NSERC Discovery Grant No. RGPIN-2019-05448, and in part by the NSERC Collaborative Research and Development Grant CRDPJ 543676-19.

from a distribution, it is known that the entropy $H(A_n)$ is a fundamental lower bound on the expected length of any lossless compressor [8, Theorem 8.3]

$$H(A_n) - \log(e(H(A_n)+1)) \le \mathsf{E}[\ell(C(A_n))], \quad (1)$$

and therefore

$$\liminf_{n\to\infty} \frac{\mathsf{E}[\ell(C(A_n))]}{H(A_n)} \ge 1.$$

Thus, a graph compressor is said to be *universal* for the family of distributions $\mathscr{P}$ if for all distribution $\mathsf{P} \in \mathscr{P}$ and $A_n \sim \mathsf{P}$, we have

$$\limsup_{n\to\infty} \frac{\mathsf{E}[\ell(C(A_n))]}{H(A_n)} \le 1. \quad (2)$$

We note that most existing studies of universal compression on sequences assume the random process $Z_1, \ldots, Z_N$ has entropy $H(Z^N)$ linear in N. Under this assumption, a compressor C is universal if

$$\limsup_{N\to\infty} \frac{\mathsf{E}[\ell(C(Z^N))]}{N} \le \mathcal{H}, \quad (3)$$

where $\mathcal{H}$ is the entropy rate of the process. We use the definition (2) rather than (3) since, in many practical applications where the graph is sparse, the entropy of the graph $H(A_n)$ grows slower than n^2. Sometimes, the first order term in the entropy is not even clear.

In addition to universality, many other desirable properties in standard studies of universal compression for sequences such as low redundancy, good finite-length performance, horizon-free implementation, low computational complexity, among others, are also goals for universal graph compression. Due to space limitations, we focus on establishing universality in this paper; the analysis for other properties, such as the redundancy of the compression scheme, can be found in [7].

A stochastic block model $\mathrm{SBM}(n, L, \mathbf{p}, \mathbf{W})$ defines a probability distribution over $\mathcal{A}_n$. Here n is the number of vertices, L is the number of communities. Each vertex $i \in [n] \coloneqq \{1, 2, \ldots, n\}$ is associated with a community assignment $X_i \in [L]$. The length-L column vector $\mathbf{p} = (p_1, p_2, \ldots, p_L)^T$ is a probability distribution over $[L]$, where p_i indicates the probability that any vertex is assigned community i. $\mathbf{W}$ is an $L \times L$ symmetric matrix, where W_{ij} represents the probability of having an edge between a vertex with community assignment i and a vertex with community assignment j. We say $A_n \sim \mathrm{SBM}(n, L, \mathbf{p}, \mathbf{W})$ if the community assignments $X_1, X_2, \ldots, X_n$ are generated i.i.d. according to $\mathbf{p}$ and for every pair $1 \le i < j \le n$, an edge is generated between vertex i and vertex j with probability W_{X_i,X_j}, conditionally independent of other edges given X_i, X_j. In other words, in the adjacency matrix A_n of the graph, $A_{ij} \sim \mathrm{Bern}(W_{X_i,X_j})$ for $i < j$; the diagonal entries $A_{ii} = 0$ for all $i \in [n]$; and $A_{ij} = A_{ji}$ for $i > j$. We write $\mathbf{W} = f(n)\mathbf{Q}$, where $\mathbf{Q}$ is an $L \times L$ symmetric matrix with $\max_{i,j} Q_{ij} = \Theta(1)$. We assume $L = \Theta(1)$ and all entries in $\mathbf{p}$ are $\Theta(1)$. We will consider two families of stochastic block models: For $0 < \epsilon < 1$,

$$\mathscr{P}_1(\epsilon)\text{: SBM with } f(n) = O(1), f(n) = \Omega\left(\tfrac{1}{n^{2-\epsilon}}\right), \quad (4)$$

$$\mathscr{P}_2(\epsilon)\text{: SBM with } f(n) = o(1), f(n) = \Omega\left(\tfrac{1}{n^{2-\epsilon}}\right). \quad (5)$$

Clearly, $\mathscr{P}_2(\epsilon)$ is a strict subset of $\mathscr{P}_1(\epsilon)$, as it does not contain the constant regime $f(n) = 1$.

B. Related Work

Due to space limitations, we only present the most related literature. See [7] for a more extensive literature survey. Lossless compression for graphs in an information-theoretic framework has been studied by [9] and [10]. In [9], universal compression of *unlabeled* graphs (isomorphism classes) generated from Erdős–Rényi models is investigated. In [10], lossless compression of labeled graphs generated from SBMs is studied, but it does not consider universal compression schemes. Recently, universal compression of graphs with marked edges and vertices is studied by Delgosha and Anantharam [11, 12]. They focus on the *sparse* graph regime, where the number of edges is in the same order as the number of vertices n. They employ the framework of local weak convergence, which provides a technique to view a sequence of graphs as a sequence of distributions on neighbourhood structures. Built on this framework, they propose an algorithm that compresses graphs by describing the local neighbourhood structures. Moreover, they introduce a universality/optimality criterion through a notion of entropy for graph sequences under the local weak convergence framework, known as the *BC entropy* [13]. This universality criterion is stronger than the one used in this paper. It requires the asymptotic length of the compressor to match the constants in both first and second order terms in Shannon entropy, whereas the universality criterion we use only requires to match the first order term. As a consequence of the stronger criterion, the compressor in [11] is universal over a smaller random graph family. In comparison, we expand the range of edge numbers from $\Theta(n)$ in the sparse regime to $\Theta(n^\alpha)$ for every $0 < \alpha \le 2$ and propose a single universal compressor for the whole family under the weaker universality criterion. In [7, Section 5], we evaluate the proposed compressor under the criterion in [11] for the family of *symmetric* SBMs. The proposed compressor achieves a similar performance in terms of BC entropy in the sparse regime.

C. Notation

We use the standard order notation: $f(n) = O(g(n))$ if $\lim_{n\to\infty} \frac{|f(n)|}{g(n)} < \infty$; $f(n) = \Omega(g(n))$ if $\lim_{n\to\infty} \frac{f(n)}{g(n)} > 0$; $f(n) = \Theta(g(n))$ if $f(n) = O(g(n))$ and $f(n) = \Omega(g(n))$; $f(n) = o(g(n))$ if $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0$; $f(n) = \omega(g(n))$ if $\lim_{n\to\infty} \frac{|f(n)|}{|g(n)|} = \infty$; and $f(n) \sim g(n)$ if $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 1$. We denote $\log(\cdot) = \log_2(\cdot)$.

II. Main Results

We present our compression scheme in Section II-A and state its performance guarantee in Theorems 1 and 2. The key idea in our design is a decomposition of the adjacency matrix into blocks, which, with a carefully chosen parameter, converts the two-dimensional correlated entries in the adjacency matrix into a sequence of *almost* i.i.d. blocks. We then compress the blocks using a Krichevsky–Trofimov or Laplace compressor and generalize their length analyses from i.i.d. processes to *arbitrarily correlated* but identically distributed processes.

A. Compression Scheme

For each integer $k \le n$, the graph compressor $C_k \colon \mathcal{A}_n \to \{0,1\}^*$ is described in the following steps.

Block decomposition. Let $n' = \lfloor n/k \rfloor$ and $\tilde{n} = \lfloor n/k \rfloor k$. We first represent the top-left $\tilde{n} \times \tilde{n}$ submatrix of A_n in the block-matrix form as

$$\begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} & \cdots & \mathbf{B}_{1,n'} \\ \mathbf{B}_{21} & \mathbf{B}_{22} & \cdots & \mathbf{B}_{2,n'} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{B}_{n',1} & \mathbf{B}_{n',2} & \cdots & \mathbf{B}_{n',n'} \end{bmatrix}, \tag{6}$$

where $\mathbf{B}_{ij}$ is the $k \times k$ submatrix formed by the rows $(i-1)k+1, (i-1)k+2, \ldots, ik$ and the columns $(j-1)k+1, (j-1)k+2, \ldots, jk$, for each $i, j \in [n']$. Denote

$$\begin{aligned} \mathbf{B}_{\mathrm{ut}} \coloneqq \mathbf{B}_{12}, &\mathbf{B}_{13}, \mathbf{B}_{23}, \mathbf{B}_{14}, \mathbf{B}_{24}, \mathbf{B}_{34}, \ldots \\ &\ldots, \mathbf{B}_{1,n'}, \cdots, \mathbf{B}_{n'-1,n'} \end{aligned} \tag{7}$$

as the sequence of off-diagonal blocks in the upper triangle and

$$\mathbf{B}_{\mathrm{d}} \coloneqq \mathbf{B}_{11}, \mathbf{B}_{22}, \ldots, \mathbf{B}_{n',n'} \tag{8}$$

as the sequence of diagonal blocks.

Binary to m-ary conversion. Let $m \coloneqq 2^{k^2}$. Each $k \times k$ block with binary entries in the two block sequences $\mathbf{B}_{\mathrm{ut}}$ and $\mathbf{B}_{\mathrm{d}}$ is converted into a symbol in $[m]$.

Probability assignment. Apply the Krichevsky–Trofimov (KT) sequential probability assignment for the two m-ary sequences $\mathbf{B}_{\mathrm{ut}}$ and $\mathbf{B}_{\mathrm{d}}$ respectively. Given an m-ary sequence $x_1, x_2, \ldots, x_N$, *KT sequential probability assignment* defines N conditional probability distributions over $[m]$ as follows. For $i \in [m], j = 0, 1, \ldots, N-1$, assign conditional probability

$$q_{\mathrm{KT}}(X_{j+1} = i | X^j = x^j) = \frac{N_i(x^j) + 1/2}{j + m/2}, \tag{9}$$

where $x^j \coloneqq (x_1, \ldots, x_j)$ and $N_i(x^j) \coloneqq \sum_{k=1}^{j} \mathbb{1}\{x_k = i\}$ counts the number of appearances of symbol i in x^j.

Adaptive arithmetic coding. With the KT sequential probability assignments, compress the two sequences $\mathbf{B}_{\mathrm{ut}}$ and $\mathbf{B}_{\mathrm{d}}$ separately using adaptive arithmetic coding [14]. In case $k = 1$, the diagonal sequence $\mathbf{B}_{\mathrm{d}}$ becomes an all-zero sequence since we assume the graph is simple. So we will only compress the off-diagonal sequence $\mathbf{B}_{\mathrm{ut}}$.

Encoding the remaining bits. The above process compressed the top-left $\tilde{n} \times \tilde{n}$ block of the adjacency matrix. For the remaining $(n-\tilde{n})\tilde{n} + \binom{n-\tilde{n}}{2}$ entries in the upper diagonal of the adjacency matrix, we simply use $2\lceil \log n \rceil$ bits to encode the row and column number of each ones.

The total length of C_k for a realization A_n is

$$\log\left(\frac{1}{q_{\mathrm{KT}}(\mathbf{B}_{\mathrm{ut}})}\right) + \log\left(\frac{1}{q_{\mathrm{KT}}(\mathbf{B}_{\mathrm{d}})}\right) + 2\lceil \log n \rceil N_{\mathrm{r}} + O(1),$$

where N_{r} is the number of ones in the remaining entries in the upper diagonal of the adjacency matrix and $O(1)$ accounts for constant costs due to arithmetic coding and headers to separate the three parts.

The complexity of the proposed algorithm is $O(2^{k^2} n^2)$. For the choice of k that achieves universality over $\mathscr{P}_1(\epsilon)$ family in Theorem 1, $O(2^{k^2} n^2) = O(n^{2+\delta})$ for $\delta < \epsilon$. For the choice of k that achieves universality over $\mathscr{P}_2(\epsilon)$ family in Theorem 2, $O(2^{k^2} n^2) = O(n^2)$.

Remark 1 (Laplace probability assignment). As an alternative to the KT probability assignment, one can also use the Laplace sequential probability assignment. Given an m-ary sequence $x_1, x_2, \ldots, x_N$, *Laplace sequential probability assignment* defines N conditional probability distributions over $[m]$ as follows. For $i \in [m], j = 0, 1, \ldots, N-1$, we assign conditional probability

$$q_{\mathrm{L}}(X_{j+1} = i | X^j = x^j) = \frac{N_i(x^j) + 1}{j + m}. \tag{10}$$

While both methods can be shown to be universal, using the Laplace probability assignment substantially simplifies the proofs and more transparently conveys the main idea. However, the KT probability assignment produces a better empirical performance. For this reason, we keep both in the paper.

B. Performance of the Universal Graph Compressor

We now state the main result of this paper: the proposed compressor C_k, for a carefully chosen k, is universal over the classes $\mathscr{P}_1(\epsilon)$ and $\mathscr{P}_2(\epsilon)$ respectively for every $0 < \epsilon < 1$.

Theorem 1 (Universality over $\mathscr{P}_1$)**.** *For every $0 < \epsilon < 1$, the graph compressor C_k defined in Section II-A is universal over the family $\mathscr{P}_1(\epsilon)$ provided that*

$$0 < \delta < \epsilon, \quad k \le \sqrt{\delta \log n}, \quad \text{and} \quad k = \omega(1).$$

Theorem 2 (Universality over $\mathscr{P}_2$)**.** *For every $0 < \epsilon < 1$, the graph compressor C_1 defined in Section II-A is universal over the family $\mathscr{P}_2(\epsilon)$.*

Remark 2. Recall that $\mathscr{P}_1(\epsilon)$ is the family of SBMs with edge probability in the regime $\Omega(\frac{1}{n^{2-\epsilon}})$ and $O(1)$.

Moreover, $\frac{1}{n^2}$ is the threshold for a random graph to contain an edge with high probability [15]. Thus, the family $\mathscr{P}_1(\epsilon)$ covers most non-trivial SBM graphs.

Remark 3. Any compressor universal over the class $\mathscr{P}_1(\epsilon)$ is also universal over the class $\mathscr{P}_2(\epsilon)$, but our compressor designed specifically for the class $\mathscr{P}_2(\epsilon)$ has a lower computational complexity.

C. Experiments

We implement the proposed universal graph compressor (UGC) in four widely used benchmark graph datasets: protein-to-protein interaction network (PPI) [16], LiveJournal friendship network (Blogcatalog) [17], Flickr user network [17], and YouTube user network [18]. We compare the our algorithm to four practical compressors: CSR (compressed sparse row), Ligra+ [19, 20], PNG image compressor, and Lempel–Ziv compressor for two-dimensional data [21]. The proposed compressor outperforms all competing algorithms in all datasets. The compression ratios from competing algorithms are 2.4 to 27 times that of the universal graph compressor (see Tables I and II).

III. Proof of Theorem 1

In this section, we outline the main ideas in establishing Theorem 1 and state key intermediate results. Full details and the proof of the rest parts are in [7].

A. Graph Entropy

We first calculate the entropy of the random graph A_n, which, recall, is the fundamental lower bound on the expected compression length for any lossless compression scheme. Since to establish optimality we need to show that $\limsup_{n\to\infty} \frac{\mathsf{E}[\ell(C(A_n))]}{H(A_n)} \le 1$, we will only be concerned with the first order term in $H(A_n)$.

Proposition 1. *Let $A_n \sim \mathrm{SBM}(n, L, \mathbf{p}, f(n)\mathbf{Q})$ with $f(n) = O(1), f(n) = \Omega\left(\frac{1}{n^2}\right)$, and $L = \Theta(1)$. For $0 \le p \le 1$, let $h(p) \triangleq -p\log(p) - (1-p)\log(1-p)$ denote the binary entropy function. For a matrix W with entries in $[0,1]$, let $h(W)$ be a matrix of the same dimension whose (i,j) entry is $h(W_{ij})$. Then*

$$H(A_n) = \binom{n}{2} H(A_{12}|X_1, X_2)(1 + o(1)) \tag{11}$$

$$= \binom{n}{2} \mathbf{p}^T h\big(f(n)\mathbf{Q}\big)\mathbf{p} + o\left(n^2 h\big(f(n)\big)\right). \tag{12}$$

In particular, when $f(n) = \Omega\left(\frac{1}{n^2}\right)$ and $f(n) = o(1)$, expression (12) can be further simplified as

$$H(A_n) = \binom{n}{2} f(n) \log\left(\frac{1}{f(n)}\right)(\mathbf{p}^T\mathbf{Q}\mathbf{p} + o(1)). \tag{13}$$

Remark 4. In the regime $f(n) = \Omega\left(\frac{1}{n}\right)$ and $f(n) = O(1)$, the above result has been established in [10]. We extend the analysis to the regime $f(n) = o\left(\frac{1}{n}\right)$ and $f(n) = \Omega(\frac{1}{n^2})$.

Proposition 1 can be used to calculate the entropy of the graph for certain important regimes of $f(n)$, in which the SBM displays characteristic behavior. For $f(n) = 1$, we have $H(A_n) = \binom{n}{2}\left(\mathbf{p}^T h(\mathbf{Q})\mathbf{p} + o(1)\right)$; for $f(n) = \frac{\log n}{n}$ (the regime where the phase transition for exact recovery of the community assignments occurs [22, 23]), we have $H(A_n) = \frac{n\log^2 n}{2}(\mathbf{p}^T\mathbf{Q}\mathbf{p} + o(1))$; when $f(n) = \frac{1}{n}$ (the regime where the phase transition for detection between SBM and the Erdős–Rényi model occurs [24]), we have $H(A_n) = \frac{n\log n}{2}(\mathbf{p}^T\mathbf{Q}\mathbf{p} + o(1))$; when $f(n) = \frac{1}{n^2}$ (the regime where the phase transition for the existence of an edge occurs), we have $H(A_n) = \log n(\mathbf{p}^T\mathbf{Q}\mathbf{p} + o(1))$.

B. Block Decomposition of A_n

To compress the matrix A_n, we wish to decompose it into a large number of components that have little correlation between them. This leads to the idea of block decomposition described previously. Since the sequence of blocks are used to compress A_n, the next theorem claims these blocks are identically distributed and asymptotically independent in a precise sense described as follows.

Proposition 2. *Let $A_n \sim \mathrm{SBM}(n, L, \mathbf{p}, f(n)\mathbf{Q})$ with $f(n) = \Omega\left(\frac{1}{n^{2-\epsilon}}\right)$ for some $0 < \epsilon < 1$, $f(n) = O(1)$, and $L = \Theta(1)$. Let $k \le n$ be an integer and $n' = \lfloor n/k \rfloor$. Consider the $k \times k$ block decomposition in (6). We have all*

	$k=1$	$k=2$	$k=3$	$k=4$
PPI	0.0228	**0.0226**	0.0227	0.034
Blogcatalog	0.0275	0.0270	**0.0267**	0.0288
Flickr	0.00960	0.00935	0.00915	**0.00907**
YouTube	4.51×10^{-5}	4.11×10^{-5}	$\mathbf{3.98 \times 10^{-5}}$	4.00×10^{-5}

Table I: Compression ratio of UGC under different k values.

	CSR	Ligra+	Lemple–Ziv	PNG
PPI	0.166	0.0605	0.06	0.089
Blogcatalog	0.203	0.0682	0.080	0.096
Flickr	0.0584	0.0217	0.0307	0.0262
YouTube	3.23×10^{-4}	9.90×10^{-5}	1.09×10^{-4}	1.10×10^{-3}

Table II: Compression ratios of competing algorithms. Full details of experiments can be found in [7].

the off-diagonal blocks share the same joint distribution; all the diagonal blocks share the same joint distribution. In other words, for any $1 \le i_1, i_2, j_1, j_2 \le n'$ *with* $i_1 \neq j_1, i_2 \neq j_2$ *and* $1 \le l_1, l_2 \le n'$*, we have*

$$\mathbf{B}_{i_1,j_1} \overset{d}{=} \mathbf{B}_{i_2,j_2},$$
$$\mathbf{B}_{l_1,l_1} \overset{d}{=} \mathbf{B}_{l_2,l_2}.$$

In addition, if $k = \omega(1)$ *and* $k = o(n)$*, we have*

$$\lim_{n\to\infty} \frac{H(\mathbf{B}_{\mathrm{ut}})}{\binom{n'}{2} H(\mathbf{B}_{12})} = 1. \tag{14}$$

C. Length Analysis for Correlated Sequences

Thanks to property of the block decomposition established in Proposition 2, we hope to compress these blocks as if they are independent using a Laplace probability assignment (which, recall, is universal for the class of all m-ary iid processes). However, since these blocks are still correlated (albeit weakly), we will need a result on the performance of Laplace probability assignment on correlated sequences with identical marginals, which we give next. These are generalizations of the well-known results on i.i.d. sequences [25, 26].

Proposition 3 (Laplace probability assignment for correlated sequence)**.** *Consider arbitrarily correlated* $Z_1, Z_2, \ldots, Z_N$*, where the marginal distribution of each* Z_i *is identically distributed over an alphabet of size* $m \ge 2$*. Let* $\ell_{\mathrm{L}}(z^N) = \lceil \log \frac{1}{q_{\mathrm{L}}(z^N)} \rceil + 1$ *where* $q_{\mathrm{L}}(\cdot)$ *is the marginal distribution induced by Laplace sequential probability assignment in* (10)

$$q_{\mathrm{L}}(z^N) := \frac{N_1! N_2! \cdots N_m!}{N!} \cdot \frac{1}{\binom{N+m-1}{m-1}}. \tag{15}$$

We then have

$$\mathsf{E}[\ell_{\mathrm{L}}(Z^N)] \le m \log(2eN) + NH(Z_1) + 2. \tag{16}$$

The Laplace probability assignment for m-ary i.i.d. processes achieves redundancy at most $m \log n + o(\log n)$, while the optimal minmax redundancy is known to be $\frac{m-1}{2} \log n + o(\log n)$, achieved asymptotically by the KT probability assignment. In fact, the Laplace probability assignment is known to be optimal up to constants even for very rich classes of processes [27]. Next, we provide a similar result for the KT probability assignment.

Proposition 4 (KT probability assignment for correlated sequence)**.** *Consider arbitrarily correlated* $Z_1, Z_2, \ldots, Z_N$*, where the marginal distribution of each* Z_i *is identically distributed over an alphabet of size* $m \ge 2$*. Let* $\ell_{\mathrm{KT}}(z^N) = \lceil \log \frac{1}{q_{\mathrm{KT}}(z^N)} \rceil + 1$ *where* $q_{\mathrm{KT}}(\cdot)$ *is the marginal distribution induced by KT probability assignment in* (9)

$$q_{\mathrm{KT}}(z^N) = \frac{(2N_1 - 1)!!(2N_2 - 1)!! \cdots (2N_m - 1)!!}{m(m+2)\cdots(m+2N-2)}$$

with $(-1)!! \triangleq 1$*. We then have*

$$\mathsf{E}[\ell_{\mathrm{KT}}(Z^N)] \le \tfrac{m}{2} \log\left(e\left(1 + \tfrac{2N}{m}\right)\right) + \tfrac{1}{2}\log(\pi N) + NH(Z_1) + 2. \tag{17}$$

Proof of Theorem 1. Since the expected length of the KT probability assignment in (17) is upper bounded by the length of the Laplace probability assignment in (16), it suffices to show C_k with the Laplace probability assignment is universal. Recall that here we compress the diagonal blocks $\mathbf{B}_{\mathrm{d}}$ ($m = 2^{k^2}$-sized alphabet, $N = n'$ blocks), the off-diagonal blocks $\mathbf{B}_{\mathrm{ut}}$ ($m = 2^{k^2}$-sized alphabet, $N = \binom{n'}{2}$ blocks) and the remaining $(n - \tilde{n})\tilde{n} + \binom{n-\tilde{n}}{2}$ entries separately. We have,

$$\begin{aligned}
\frac{\mathsf{E}(\ell(C_k(A_n)))}{H(A_n)} &= \frac{\mathsf{E}(\ell_{\mathrm{L}}(\mathbf{B}_{\mathrm{ut}}) + \ell_{\mathrm{L}}(\mathbf{B}_{\mathrm{d}}) + 2\lceil \log n \rceil N_r)}{H(A_n)} \\
&\overset{(a)}{\le} \tfrac{1}{H(A_n)} \Big[\binom{n'}{2} H(\mathbf{B}_{12}) + 2^{k^2} \log\left(en^2\right) \\
&\qquad + nH(\mathbf{B}_{11}) + 2^{k^2} \log(2en) + 4 + \mathsf{E}(2\lceil \log n \rceil N_r) \Big] \\
&\overset{(b)}{\le} \frac{\binom{n'}{2} H(\mathbf{B}_{12})}{H(A_n)} + \frac{2^{k^2} \log\left(2e^2 n^3\right)}{H(A_n)} + \frac{nk^2 H(A_{12}) + 4}{H(A_n)} \\
&\qquad + \frac{\mathsf{E}(2\lceil \log n \rceil N_r)}{H(A_n)},
\end{aligned}$$

where in (a) we use (16), and bound $2\binom{n'}{2} \le n^2$ and $n' \le n$, and in (b) we note that $H(\mathbf{B}_{11}) \le k^2 H(A_{12})$ since there are $k^2 - k$ elements of the matrix (all apart from the diagonal elements) are distributed identically as A_{12}. We will now analyze each of these four terms separately. Firstly, using Proposition 2 yields that $\frac{\binom{n'}{2} H(\mathbf{B}_{12})}{H(A_n)} \to 1$. Next, since $f(n) = \Omega\left(\frac{1}{n^{2-\epsilon}}\right)$, we have $H(A_n) = \Omega(n^\epsilon \log n)$ and subsequently substituting $k \le \sqrt{\delta \log n}$, we have

$$\frac{2^{k^2} \log(2en^3)}{H(A_n)} = O\left(\frac{n^\delta \log n}{n^\epsilon \log n}\right) = O\left(n^{\delta - \epsilon}\right) = o(1)$$

since $\delta < \epsilon$. Moreover, we have

$$\frac{nk^2 H(A_{12}) + 4}{H(A_n)} \le \frac{nk^2 H(A_{12}) + 4}{\binom{n}{2} H(A_{12}|X_1, X_2)} = O\left(\tfrac{k^2}{n}\right) = o(1),$$

where the penultimate equality used the fact that $H(A_{12}) \sim H(A_{12}|X_1, X_2)$ (since $h(f(n)\mathbf{p}^T\mathbf{Q}\mathbf{p}) \sim \mathbf{p}^T h(f(n)\mathbf{Q})\mathbf{p}$). Finally, recall that N_r is the number of ones in the remaining $(n - \tilde{n})\tilde{n} + \binom{n-\tilde{n}}{2}$ entries, we have

$$\begin{aligned}
\frac{\mathsf{E}(2\lceil \log n \rceil N_r)}{H(A_n)} &= \frac{f(n)\mathbf{p}^T\mathbf{Q}\mathbf{p}\left((n - \tilde{n})\tilde{n} + \binom{n-\tilde{n}}{2}\right) 2\lceil \log n \rceil}{H(A_n)} \\
&= \Theta\left(\frac{nkf(n)\log n}{n^2 f(n) \log(1/f(n))}\right) = o(1).
\end{aligned}$$

This completes the proof. □

References

[1] Q. Xie and A. R. Barron, "Minimax redundancy for the class of memoryless sources," *IEEE Trans. Inf. Theory*, vol. 43, no. 2, pp. 646–657, 1997.

[2] ——, "Asymptotic minimax regret for data compression, gambling, and prediction," *IEEE Trans. Inf. Theory*, vol. 46, no. 2, pp. 431–445, 2000.

[3] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Trans. Inf. Theory*, vol. 23, no. 3, pp. 337–343, 1977.

[4] ——, "Compression of individual sequences via variable-rate coding," *IEEE Trans. Inf. Theory*, vol. 24, no. 5, pp. 530–536, 1978.

[5] M. Effros, K. Visweswariah, S. R. Kulkarni, and S. Verdú, "Universal lossless source coding with the burrows wheeler transform," *IEEE Trans. Inf. Theory*, vol. 48, no. 5, pp. 1061–1081, 2002.

[6] F. M. Willems, Y. M. Shtarkov, and T. J. Tjalkens, "The context-tree weighting method: basic properties," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 653–664, 1995.

[7] A. Bhatt, Z. Wang, C. Wang, and L. Wang, "Universal graph compression: Stochastic block models," 2020. [Online]. Available: https://arxiv.org/abs/2006.02643

[8] Y. Polyanskiy and Y. Wu, "Lecture notes on information theory," 2014.

[9] Y. Choi and W. Szpankowski, "Compression of graphical structures: Fundamental limits, algorithms, and experiments," *IEEE Trans. Inf. Theory*, vol. 58, no. 2, pp. 620–638, Feb 2012.

[10] E. Abbe, "Graph compression: The effect of clusters," in *Proc. 54th Ann. Allerton Conf. Commun. Control Comput.*, 2016, pp. 1–8.

[11] P. Delgosha and V. Anantharam, "Universal lossless compression of graphical data," *IEEE Trans. Inf. Theory*, vol. 66, no. 11, pp. 6962–6976, 2020.

[12] P. Delgosha and V. Anantharam, "A universal low complexity compression algorithm for sparse marked graphs," in *Proc. IEEE Internat. Symp. Inf. Theory*, June 2020.

[13] C. Bordenave and P. Caputo, "Large deviations of empirical neighborhood distribution in sparse random graphs," *Probability Theory and Related Fields*, vol. 163, no. 1-2, p. 149–222, Nov 2014.

[14] D. Marpe, H. Schwarz, and T. Wiegand, "Context-based adaptive binary arithmetic coding in the h.264/avc video compression standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, 2003.

[15] A. Frieze and M. Karoński, *Introduction to Random Graphs*. Cambridge University Press, 2015.

[16] A. Grover and J. Leskovec, "Node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, pp. 855–864.

[17] L. Tang and H. Liu, "Relational learning via latent social dimensions," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '09. New York, NY, USA: Association for Computing Machinery, 2009, pp. 817–826.

[18] S. Nandanwar and M. N. Murty, "Structural neighborhood based classification of nodes in a network," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, pp. 1085–1094.

[19] J. Shun and G. E. Blelloch, "Ligra: A lightweight graph processing framework for shared memory," in *Proceedings of the 18th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, ser. PPoPP '13. New York, NY, USA: Association for Computing Machinery, 2013, pp. 135–146.

[20] J. Shun, L. Dhulipala, and G. E. Blelloch, "Smaller and faster: Parallel processing of compressed graphs with ligra+," in *2015 Data Compression Conference*, pp. 403–412.

[21] A. Lempel and J. Ziv, "Compression of two-dimensional data," *IEEE Trans. Inf. Theory*, vol. 32, no. 1, pp. 2–8, 1986.

[22] E. Abbe, A. S. Bandeira, and G. Hall, "Exact recovery in the stochastic block model," *IEEE Trans. Inf. Theory*, vol. 62, no. 1, pp. 471–487, 2015.

[23] E. Abbe and C. Sandon, "Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery," in *IEEE 56th Annual Symposium on Foundations of Computer Science*, 2015, pp. 670–688.

[24] E. Mossel, J. Neeman, and A. Sly, "Reconstruction and estimation in the planted partition model," *Probability Theory and Related Fields*, vol. 162, no. 3-4, pp. 431–461, 2015.

[25] N. Cesa-Bianchi and G. Lugosi, *Prediction, learning, and games*. Cambridge university press, 2006.

[26] R. Krichevsky and V. Trofimov, "The performance of universal encoding," *IEEE Transactions on Information Theory*, vol. 27, no. 2, pp. 199–207, 1981.

[27] Y. Yang and A. Barron, "Information-theoretic determination of minimax rates of convergence," *Annals of Statistics*, pp. 1564–1599, 1999.