Compressing Tabular Data via Latent Variable Estimation

Andrea Montanari [1] **Eric Weiner** [1]

Abstract

Data used for analytics and machine learning often take the form of tables with categorical entries. We introduce a family of lossless compression algorithms for such data that proceed in four steps: (i) Estimate latent variables associated to rows and columns; (ii) Partition the table in blocks according to the row/column latents; (iii) Apply a sequential (e.g. Lempel-Ziv) coder to each of the blocks; (iv) Append a compressed encoding of the latents. We evaluate this approach on several benchmark datasets, and study optimal compression in a probabilistic model for tabular data, whereby latent values are independent and table entries are conditionally independent given the latent values. We prove that the model has a well defined entropy rate and satisfies an asymptotic equipartition property. We also prove that classical compression schemes such as Lempel-Ziv and finite-state encoders do not achieve this rate. On the other hand, the latent estimation strategy outlined above achieves the optimal rate.

1. Introduction

Classical theory of lossless compression (Cover & Thomas, 2006; Salomon, 2004) assumes that data take the form of a random vector $\boldsymbol{X}^N = (X_1, X_2, \dots, X_N)$ of length N with entries in a finite alphabet $\mathcal{X}$. Under suitable ergodicity assumptions, the entropy per letter converges to a limit $h := \lim_{N\to\infty} H(\boldsymbol{X}^N)/N$ (Shannon-McMillan-Breiman theorem). Universal coding schemes (e.g. Lempel-Ziv coding) do not requite knowledge of the distribution of $\boldsymbol{X}^N$, and can encode such a sequence without information loss using (asymptotically) h bits per symbol.

While this theory is mathematically satisfying, its modeling assumptions (stationarity, ergodicity) are unlikely to be satisfied in many applications. This has long been recognized by practitioners. The main objective of this paper is to investigate this fact mathematically in the context of tabular data, characterize the gap to optimality of classical schemes, and describe an asymptotically optimal algorithm that overcomes their limitations.

We consider a data table with m rows and n columns and entries in $\mathcal{X}$, $\boldsymbol{X}^{m,n} \in \mathcal{X}^{m\times n}$ $\boldsymbol{X}^{m,n} := (X_{ij})_{i\le m, j\le n}$. The standard approach to such data is: (i) Serialize, e.g. in row-first order, to form a vector of length $N = mn$, $\boldsymbol{X}^N = (X_{11}, X_{12}, \dots, X_{1n}, X_{21}, \dots, X_{mn})$; (ii) Apply a standard compressor (e.g., Lempel-Ziv) to this vector.

We will show, both empirically and mathematically, that this standard approach can be suboptimal in the sense of not achieving the optimal compression rate. This happens even in the limit of large tables, as long as the number of columns and rows are polynomially related (i.e. $n^{\varepsilon} \le m \le n^M$ for some small constant ε and large constant M).

We advocate an alternative approach:

1. Estimate row/column latents $\boldsymbol{u}^m = (u_1, \dots, u_m) \in \mathcal{L}^m$, $\boldsymbol{v}^n = (v_1, \dots, v_n) \in \mathcal{L}^n$, with $\mathcal{L}$ a finite alphabet.

2. Partition the table in blocks according to the row/column latents, Namely, for $u, v \in \mathcal{L}$, define

$$\boldsymbol{X}(u, v) = \mathsf{vec}\big(X_{ij} : \ u_i = u, v_j = v\big)\,. \qquad (1.1)$$

   where $\mathsf{vec}(\boldsymbol{M})$ denote the serialization of matrix $\boldsymbol{M}$ (either row-wise or column-wise).

3. Apply a base compressor (generically denoted by $\mathsf{Z}_{\mathcal{X}} : \mathcal{X}^* \to \{0,1\}^*$) to each block $\boldsymbol{X}(u, v)$

$$\boldsymbol{z}(u, v) = \mathsf{Z}_{\mathcal{X}}(\boldsymbol{X}(u, v))\,, \quad \forall u, v \in \mathcal{L}\,. \qquad (1.2)$$

4. Encode the row latents and column latents using a possibly different compressor $\mathsf{Z}_{\mathcal{L}} : \mathcal{X}^* \to \{0,1\}^*$, to get $\boldsymbol{z}_{\text{row}} = \mathsf{Z}_{\mathcal{L}}(\boldsymbol{u})$, $\boldsymbol{z}_{\text{col}} = \mathsf{Z}_{\mathcal{L}}(\boldsymbol{v})$. Finally output the concatenation (denoted by $\oplus$)

$$\mathsf{Enc}(\boldsymbol{X}^{m,n}) = \mathsf{header} \oplus \boldsymbol{z}_{\text{row}} \oplus \boldsymbol{z}_{\text{col}} \oplus \bigoplus_{u,v\in\mathcal{L}} \boldsymbol{z}(u, v)\,. \qquad (1.3)$$

   Here header is a header that contains encodings of the lengths of subsequent segments.

[1]Project N, Mountain View, CA, United States. Correspondence to: Andrea Montanari <am@projectn.co>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

Note that encoding the latents can in general lead to a suboptimal compression rate. While this can be remedied with techniques such as bits-back coding, we observed in our applications that using such techniques yields limited improvement. Our analysis shows that the rate improvement afforded by bits-back coding is only significant in certain special regimes. We refer to Sections 5 and 6 for further discussion.

The above description leaves several design choices undefined, namely: (a) The latents estimation procedure at point 1; (b) The base compressor $\mathsf{Z}_{\mathcal{X}}$ for the blocks $\boldsymbol{X}(u,v)$; (c) The base compressor $\mathsf{Z}_{\mathcal{L}}$ for the latents.

We will provide details for a specific implementation in Section 2, alongside empirical evaluation in Section 3. Section 4 introduces a probabilistic model for the data $\boldsymbol{X}^{m,n}$, and Section 5 establishes our main theoretical results: standard compression schemes are suboptimal on this model, while the above latents-based approach is asymptotically optimal. Finally we discuss extensions in Section 6.

1.1. Related work

The use of latent variables is quite prevalent in compression methods based on machine learning and probabilistic modeling. Hinton and Zemel (1993) introduced the idea that stochastically generated codewords (e.g., random latents) can lead to minimum description lengths via bits back coding. This idea was explicitly applied to lossless compression using arithmetic coding in (Frey & Hinton, 1996), and ANS coding in (Townsend et al., 2019a;b).

Compression via low-rank approximation is closely-related to our latents-based approach and has been studied in the past. An incomplete list of contributions includes (Cheng et al., 2005) (numerical analysis), (Li & Li, 2010) (hyperspectral imaging), (Yuan & Oja, 2005; Hou et al., 2015) (image processing), (Taylor, 2013) (quantum chemistry), (Phan et al., 2020) (compressing the gradient for distributed optimization), (Chen et al., 2021) (large language models compression).

The present paper contributes to this line of work, but departs from it in a number of ways. (i) We study lossless compression while earlier work is mainly centered on lossy compression. (ii) Most of the papers in this literature do not precisely quantify compression rate: they do not 'count bits.' (iii) We show empirically an improvement in terms of lossless compression rate over state of the art.

Another related area is network compression: simple graphs can be viewed as matrices with entries in $\{0,1\}$. In the case of graph compression, one is interested only in such matrices up to graph isomorphisms. The idea of reordering the nodes of the network and exploiting similarity between nodes has been investigated in this context, see e.g. (Boldi & Vigna, 2004; Chierichetti et al., 2009; Lim et al., 2014; Besta & Hoefler, 2018) However, we are not aware of results analogous to ours in this literature.

To the best of our knowledge, our work is the first to prove that classical lossless compression techniques do not achieve the ideal compression rate under a probabilistic model for tabular data. We characterize this ideal rate as well as the one achieved by classical compressors, and prove that latents estimation can be used to close this gap.

Algorithm 1 Latent-based Tabular Compressor

Input: Data matrix $\boldsymbol{X}^{m,n} \in \mathcal{X}^{m\times n}$, range $k = |\mathcal{L}|$
Output: Compressed data $\mathsf{Enc}(\boldsymbol{X}^{m,n}) \in \{0,1\}^*$

Estimate latents $\boldsymbol{u}^m \in [k]^m$, $\boldsymbol{v}^n \in [k]^n$ using Algorithm 2, with inputs $\boldsymbol{X}^{m,n}$, k
for $u, v \in [k]$ **do**
  $\boldsymbol{X}(u,v) = \mathsf{vec}\big(X_{ij} : \ u_i = u, v_j = v\big)$
  $\boldsymbol{z}(u,v) = \mathsf{Z}_{\mathcal{X}}(\boldsymbol{X}(u,v))$
end for
Compute $\boldsymbol{z}_{\text{row}} = \mathsf{Z}_{\mathcal{L}}(\boldsymbol{u})$, $\boldsymbol{z}_{\text{col}} = \mathsf{Z}_{\mathcal{L}}(\boldsymbol{v})$
return concatenation of $\{\boldsymbol{z}(u,v) : u,v \in [k]\}$, $\boldsymbol{z}_{\text{row}}$,$\boldsymbol{z}_{\text{col}}$, metadata

1.2. Notations

We generally use boldface for vectors and uppercase boldface for matrices, without making any typographic distinction between numbers and random variables. When useful, we indicate by superscripts the dimensions of a matrix or a vector: $\boldsymbol{u}^m$ is a vector of length m, and $\boldsymbol{X}^{m,n}$ is a matrix of dimensions $m \times n$. For a string $\boldsymbol{v}$ and $a \le b$, we use $\boldsymbol{v}_a^b = (v_a, \dots, v_b)$ to denote the substring of $\boldsymbol{v}$.

If X, Y are random variables on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we denote by $H(X)$, $H(Y)$ their entropies, $H(X,Y)$ their joint entropy, $H(X|Y)$ the conditional entropy of X given Y. We will overload this notation: if p is a discrete probability distribution, we denote by $H(p)$ its entropy. Unless stated otherwise, all entropies will be measured in bits. For $\varepsilon \in [0,1]$, $\mathsf{h}(\varepsilon) := -\varepsilon \log_2 \varepsilon - (1-\varepsilon)\log_2(1-\varepsilon)$.

2. Implementation

The overall structure of the compression algorithm was already described in the introduction. Algorithm 1 summarizes it. In this section we provide further details about the two basic components it relies on: the base compressor $\mathsf{Z}_\cdot$, and the latents estimation algorithm. In both cases, the choice is in no-way unique and we only describe what we used in our implementation.

2.1. Base compressors

We implemented the following two options for the base compressors $\mathsf{Z}_{\mathcal{X}}$ (for data blocks) and $\mathsf{Z}_{\mathcal{L}}$ (for latents).

Dictionary-based compression (Lempel-Ziv, LZ). For this we used Zstandard (ZSTD) Python bindings to the C implementation using the library zstd, with level 12. While ZSTD can use run-length encoding schemes or literal encoding schemes, we verified that in in this case ZSTD always use its LZ algorithm.

The LZ algorithm in ZSTD is somewhat more sophisticated than the plain LZ algorithm used in our proofs. In particular it includes (Collet & Kucherawy, 2018) Huffman coding of literals 0-255 and entropy coding of the LZ stream. Experiments with other (simpler) LZ implementations yielded similar results. We focus on ZSTD because of its broad adoption in industry.

Frequency-based entropy coding (ANS). For each data portion (i.e each block $\boldsymbol{X}(u,v)$ and each of the row latents $\boldsymbol{u}$ and column latents $\boldsymbol{v}$) compute empirical frequencies of the corresponding symbols. Namely for all $u, v \in \mathcal{L}$, $x \in \mathcal{X}$, we compute

$$\widehat{Q}(x|u,v) := \frac{1}{N(u,v)} \sum_{i:u_i=u} \sum_{j:v_j=v} \mathbf{1}_{x_{ij}=x} \,,$$

$$\hat{q}_{\mathsf{r}}(u) := \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{u_i=u} \,, \quad \hat{q}_{\mathsf{c}}(v) := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{v_i=v} \,,$$

where $N(u,v)$ is the number of $i \le m$, $j \le n$ such that $u_i = u$, $v_j = v$. We then apply ANS coding (Duda, 2009) to each block $\boldsymbol{X}(u,v)$ modeling its entries as independent with distribution $\widehat{Q}(\,\cdot\,|u,v)$, and to the latents $\boldsymbol{u}^m$, $\boldsymbol{v}^n$ using the distributions $\hat{q}_{\mathsf{r}}(\,\cdot\,)$, $\hat{q}_{\mathsf{c}}(\,\cdot\,)$. We separately encode these counts as long integers.

Since our main objective was to study the impact of learning latents, we did not try to optimize these base compressors.

2.2. Latent estimation

We implemented latents estimation using a spectral clustering algorithm outlined in Algorithm 2.

In words, the algorithm encodes the data matrix $\boldsymbol{X}^{m,n}$ as an $m \times n$ real-valued matrix $\boldsymbol{M}^{m,n} \in \mathbb{R}^{m\times n}$ using a map $\psi : \mathcal{X} \to \mathbb{R}$. It then computes the top $k-1$ left and right singular vectors of $\boldsymbol{M}$, and stores them as matrices $\boldsymbol{A} \in \mathbb{R}^{m\times(k-1)}$, $\boldsymbol{B} \in \mathbb{R}^{m\times(k-1)}$. The rows $\boldsymbol{a}_i, \boldsymbol{b}_j \in \mathbb{R}^{k-1}$ of these matrices are used as embedding of the rows and columns indices in $k-1$ dimensions. Finally, we run KMeans on these vectors to construct k clusters of rows/columns.

A few remarks are in order. The algorithm encodes the data matrix $\boldsymbol{X}^{m,n}$ as an $m \times n$ real-valued matrix $\boldsymbol{M}^{m,n} \in \mathbb{R}^{m\times n}$ using a map $\psi : \mathcal{X} \to \mathbb{R}$. In our experiments we did not optimize this map and encoded the elements of $\mathcal{X}$ as $0, 1, \dots, |\mathcal{X}| - 1$ arbitrarily, cf. also Section 5.3

The singular vector calculation turns out to be the most time consuming part of the algorithm. Computing approximate singular vectors via power iteration requires in this case of the order of $\log(m \wedge n)$ matrix vector multiplications for each of k vectors. This amounts to $mnk \log(m \wedge n)$ operations, which is larger than the time needed to compress the blocks or to run KMeans. A substantial speed-up is obtained via row subsampling, cf. Section 6

For the clustering step we use the scikit-learn implementation via sklearn.cluster.KMeans, with random initialization.

Algorithm 2 Spectral latents estimation

Input: Data matrix $\boldsymbol{X}^{m,n} \in \mathcal{X}^{m\times n}$
latents range $k = |\mathcal{L}|$; map $\psi : \mathcal{X} \to \mathbb{R}$
Output: Factors $\boldsymbol{u}^m \in \mathcal{L}^m$, $\boldsymbol{v}^n \in \mathcal{L}^n$

Compute top $(k-1)$ singular vectors of $\boldsymbol{M}^{m,n} = \psi(\boldsymbol{X}^{m,n})$, $(\tilde{\boldsymbol{a}}_i)_{i\le k-1}$, $(\tilde{\boldsymbol{b}}_i)_{i\le k-1}$
Stack singular vectors in matrices $\boldsymbol{A} = [\tilde{\boldsymbol{a}}_1|\dots|\tilde{\boldsymbol{a}}_{k-1}] \in \mathbb{R}^{m\times(k-1)}$, $\boldsymbol{B} = [\tilde{\boldsymbol{b}}_1|\cdots|\tilde{\boldsymbol{b}}_{k-1}] \in \mathbb{R}^{n\times(k-1)}$;
Let $(\boldsymbol{a}_i)_{i\le m}$, $\boldsymbol{a}_i \in \mathbb{R}^{k-1}$ be the rows of $\boldsymbol{A}$; $(\boldsymbol{b}_i)_{i\le n}$, $\boldsymbol{b} \in \mathbb{R}^{k-1}$ the rows of $\boldsymbol{B}$
Apply KMeans to $(\boldsymbol{a}_i)_{i\le m}$; store the cluster labels as vector $\boldsymbol{u}^m$
Apply KMeans to $(\boldsymbol{b}_i)_{i\le n}$; store the cluster labels as vector $\boldsymbol{v}^n$
return $\boldsymbol{u}^m$, $\boldsymbol{v}^n$

3. Empirical evaluation

We evaluated our approach on tabular datasets with different origins. Our objective is to assess the impact of using latents in reordering columns and rows, so we will not attempt to achieve the best possible data reduction rate (DRR) on each dataset, but rather to compare compression with latents and without in as-uniform-as-possible fashion.

Since our focus is on categorical variables, we preprocess the data to fit in this setting as described in Section A.2. This preprocessing step might involve dropping some of the columns of the original table. We denote the number of columns after preprocessing by n.

We point out two simple improvements we introduce in the implementation: (i) We use different sizes for rows latent alphabet and column latent alphabet $|\mathcal{L}_r| \neq |\mathcal{L}_c|$; (ii) We choose $|\mathcal{L}_r|$, $|\mathcal{L}_c|$ by optimizing the compressed size .

3.1. Datasets

More details on these data can be found in Appendix A.1:

Taxicab. A table with $m = 62,495$, $n = 18$ (NYC.gov, 2022). LZ: $|\mathcal{L}_r| = 9$, $|\mathcal{L}_c| = 15$. ANS: $|\mathcal{L}_r| = 5$, $|\mathcal{L}_c| = 14$.

Network. Four social networks from (Leskovec & Krevl, 2014) with $m = n \in \{333, 747, 786, 1187\}$. LZ and ANS: $|\mathcal{L}_r| = 5$, $|\mathcal{L}_c| = 5$.

Card transactions. A table with $m = 24,386,900$ and $n = 12$ (Altman, 2019). LZ and ANS: $|\mathcal{L}_r| = 3$, $|\mathcal{L}_c| = n$.

Business price index. A table with $m = 72,750$ and $n = 10$ (stats.govt.nz, 2022). LZ: $|\mathcal{L}_r| = 6$, $|\mathcal{L}_c| = 7$. ANS: $|\mathcal{L}_r| = 2$, $|\mathcal{L}_c| = 6$.

Forest. A table from the UCI data repository with $m = 581,011$, $n = 55$ (Dua & Graff, 2017). LZ and ANS: $|\mathcal{L}_r| = 6$, $|\mathcal{L}_c| = 17$.

US Census. Another table from (Dua & Graff, 2017) with $m = 2,458,285$ and $n = 68$. LZ and ANS: $|\mathcal{L}_r| = 9$, $|\mathcal{L}_c| = 68$.

Jokes. A collaborative filtering dataset with $m = 23,983$ rows and $n = 101$ (Goldberg et al., 2001; Goldberg et al.). LZ: $|\mathcal{L}_r| = 2$, $|\mathcal{L}_c| = 101$. ANS: $|\mathcal{L}_r| = 8$, $|\mathcal{L}_c| = 8$.

3.2. Results

Given a lossless encoder $\phi : \mathcal{X}^{m\times n} \to \{0,1\}^*$, we define its compression rate and data reduction rate (DRR) as

$$\mathsf{R}_\phi(\boldsymbol{X}^{m,n}) := \frac{\mathsf{len}(\phi(\boldsymbol{X}^{m,n}))}{mn\log_2|\mathcal{X}|},$$
$$\mathsf{DRR}_\phi(\boldsymbol{X}^{m,n}) := 1 - \mathsf{R}_\phi(\boldsymbol{X}^{m,n}). \tag{3.1}$$

(Larger DRR means better compression.)

The DRR of each algorithm is reported in Table 1. For the table of results, **LZ** refers to row-major order ZSTD, **LZ (c)** refers to column-major order ZSTD. We run KMeans on the data 5 times, with random initializations finding the DRR each time and reporting the average.

We make the following observations on the empirical results of Table 1. First, Latent + ANS encoder achieves systematically the best DRR. Second, the use of latent in several cases yields a DRR improvement of 5% (of the uncompressed size) or more. Third, as intuitively natural, this improvement appears to be larger for data with a large number of columns (e.g. the network data).

The analysis of the next section provides further support for these findings.

4. A probabilistic model

In order to better understand the limitations of classical approaches, and the optimality of latent-based compression, we introduce a probabilistic model for the table $\boldsymbol{X}^{m,n} \in \mathcal{X}^{m\times n}$. We assume the true latents $(u_i)_{i\le m}$, $(v_j)_{j\le n}$ to be independent random variables with

$$\mathbb{P}(u_i = u) = q_{\mathsf{r}}(u)\,, \quad \mathbb{P}(v_i = v) = q_{\mathsf{c}}(v)\,. \tag{4.1}$$

We assume that the entries $(X_{ij})_{i\le m, j\le n}$ are conditionally independent given $\boldsymbol{u}^m = (u_i)_{i\le m}$ $\boldsymbol{v}^n = (v_j)_{j\le n}$, with

$$\mathbb{P}\big(X_{ij} = x\big|\boldsymbol{u}^m, \boldsymbol{v}^n\big) = Q(x|u_i, v_j)\,. \tag{4.2}$$

The distributions $q_{\mathsf{r}}, q_{\mathsf{c}}$, and conditional distribution Q are parameters of the model (a total of $2(|\mathcal{L}| - 1) + |\mathcal{L}|^2(|\mathcal{X}| - 1)$ real parameters). We will write $(\boldsymbol{X}^{m,n}, \boldsymbol{u}^m, \boldsymbol{v}^n) \sim \mathcal{T}(Q, q_{\mathsf{r}}, q_{\mathsf{c}}; m, n)$ to indicate that the triple $(\boldsymbol{X}^{m,n}, \boldsymbol{u}^m, \boldsymbol{v}^n)$ is distributed according to the model.

Remark 4.1. Some of our statements will be non-asymptotic, in which case m, n, $\mathcal{X}$, $\mathcal{L}$, Q, q_{r}, q_{c} are fixed. Others will be of asymptotic. In the latter case, we have in mind a sequence of problems indexed by n. In principle, we could write m_n, $\mathcal{X}_n$, $\mathcal{L}_n$, Q_n, $q_{\mathsf{r},n}$, $q_{\mathsf{c},n}$ to emphasize the fact that these quantities depend on n. However, we will typically omit these subscripts.

Example 4.2 (Symmetric Binary Model)**.** As a toy example, we will use the following Symmetric Binary Model (SBM) which parallels the symmetric stochastic block model for community detection (Holland et al., 1983). We take $\mathcal{L} = [k] := \{1, \dots, k\}$, $\mathcal{X} = \{0, 1\}$, $q_{\mathsf{r}} = q_{\mathsf{c}} = \mathsf{Unif}([k])$ (the uniform distribution over $[k]$) and

$$Q(1|u, v) = p_1 \text{ if } u = v, \quad Q(1|u, v) = p_0 \text{ if } u \neq v. \tag{4.3}$$

We will write $(\boldsymbol{X}^{m,n}, \boldsymbol{u}^m, \boldsymbol{v}^n) \sim \mathcal{T}_{\mathsf{SBM}}(p_0, p_1, k; m, n)$ when this distribution is used.

Figure 1 reports the results of simulations within the SBM, for ZSTD and ANS base compressors. In this case $m = n = 1000$, $k = 3$, and we average DRR values over 4 realizations. Appendix B reports additional simulations under the same model for $k \in \{5, 7\}$: the results are very similar to the ones of Figure 1. As expected, the use of latents is irrelevant along the line $p_1 \approx p_0$ (in this case, the latents do not impact the distribution of X_{ij}). However, it becomes important when p_1 and p_0 are significantly different.

The figures also report contour lines of the theoretical predictions for the asymptotic DRR of various compression algorithms (cf. Example 5.4). The agreement is excellent.

5. Theoretical analysis

In this section we present our theoretical results on compression rates under the model $\mathcal{T}(Q, q_{\mathsf{r}}, q_{\mathsf{c}}, k; m, n)$ introduced above. We first characterize the optimal compression rate in Section 5.1, then prove that standard compression methods fail to attain this goal in Section 5.2, and finally show that

Table 1: Data reduction rate (DRR) achieved by classical and latent-based compressors on real tabular data.

Data	Size	LZ	LZ (c)	ANS	Latent + LZ	Latent + ANS
Taxicab	380 KB	0.41	0.44	0.43	0.48	**0.54**
FB Network 1	13.6 KB	0.63	0.63	0.76	0.58	**0.78**
FB Network 2	68.1 KB	0.44	0.44	0.57	0.64	**0.75**
FB Network 3	75.4 KB	0.59	0.59	0.75	0.69	**0.80**
GP Network 1	172 KB	0.46	0.46	0.65	0.58	**0.70**
Forest (s)	6.10 MB	0.29	0.38	0.47	0.41	**0.49**
Card Transactions (s)	123 MB	0.03	0.21	0.29	0.20	**0.30**
Business price index (s)	153 KB	−0.03	0.20	0.28	0.25	**0.32**
US Census	43.9 MB	0.38	0.31	0.47	0.52	**0.62**
Jokes	515 KB	−0.21	−0.15	0.07	−0.03	**0.14**

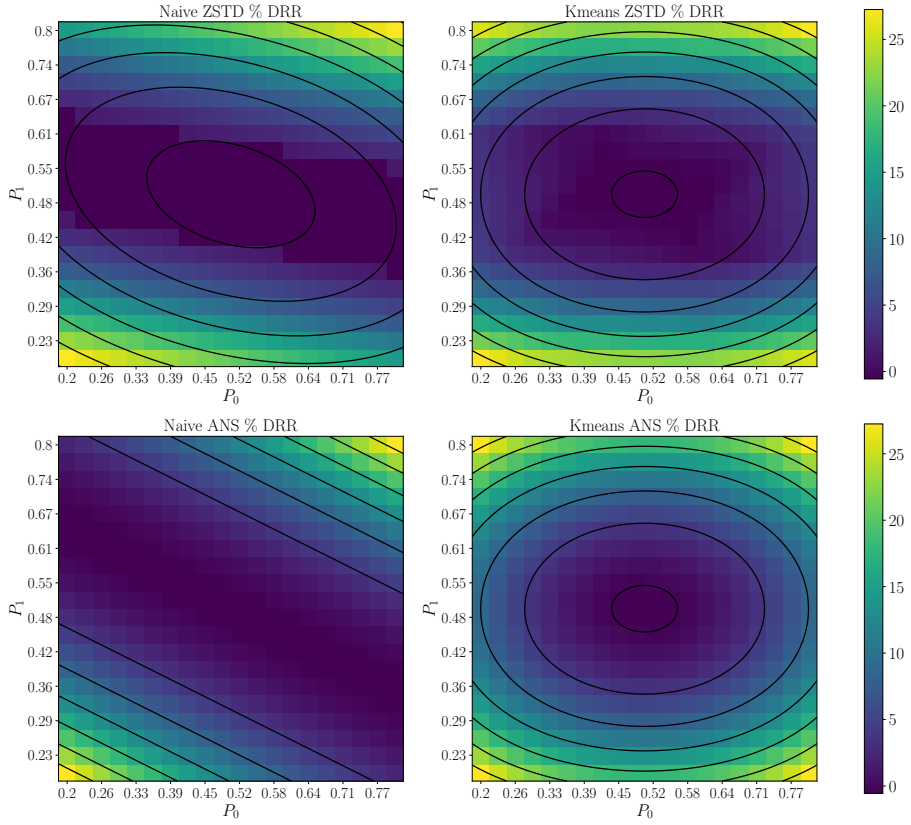

Figure 1: Comparing data reduction rate of naive coding and latent-based coding for synthetically generated data. Top: ZSTD base compressor. Bottom: ANS base compressor. Contour lines correspond to the compression rate predicted by the theorems of Section 5 (coinciding with optimal rate for latent-based encoders).

latent-based compression does in Section 5.3. Proofs are deferred to Appendices D, E, F, G.

Throughout, we denote by (X, U, V) a triple with joint distribution $\mathbb{P}(X = x, U = u, V = v) = Q(x|u, v)q_{\mathrm{r}}(u)q_{\mathrm{c}}(v)$ (this is the same as the joint distribution of (X_{ij}, u_i, v_j) for fixed i, j).

5.1. Ideal compression

Our first lemma provides upper and lower bounds on the entropy per symbol $H(\boldsymbol{X}^{m,n})/mn$.

Lemma 5.1. *Defining* $H^{+}_{m,n}(X|U, V) := H(X|U, V) + \frac{1}{n}H(U) + \frac{1}{m}H(V)$, *we have*

$$H(X|U, V) \le \frac{1}{mn}H(\boldsymbol{X}^{m,n}) \le H^{+}_{m,n}(X|U, V)\,. \quad (5.1)$$

Further, for any estimators $\widehat{\boldsymbol{u}} : \mathcal{X}^{m\times n} \to \mathcal{L}^m$, $\widehat{\boldsymbol{v}} : \mathcal{X}^{m\times n} \to \mathcal{L}^n$, *let* $\mathsf{Err}_U := \min_{\pi\in\mathfrak{S}_\mathcal{L}} \sum_{i=1}^m \mathbf{1}_{\hat{u}_i\neq\pi(u_i)}/m$, $\mathsf{Err}_V := \min_{\pi\in\mathfrak{S}_\mathcal{L}} \sum_{i=1}^n \mathbf{1}_{\hat{v}_i\neq\pi(v_i)}/n$ (min *over permutations of* $\mathcal{L}$), *letting* $\varepsilon_U := \mathbb{E}\,\mathsf{Err}_U$, $\varepsilon_V := \mathbb{E}\,\mathsf{Err}_V$, *we have*

$$H^+_{m,n}(X|U,V) - \delta_{m,n} \le \frac{1}{mn} H(\boldsymbol{X}^{m,n}) \le H^+_{m,n}(X|U,V)\,. \tag{5.2}$$

where $\delta_{m,n} := \delta(\varepsilon_U)/n + \delta(\varepsilon_V)/m$ *and* $\delta(\varepsilon) := \mathsf{h}(\varepsilon) + \varepsilon\log(|\mathcal{L}|-1)$.

Corollary 5.2. *There exists a lossless compressor* ϕ *whose rate (cf.Eq. (3.1)) is*

$$\mathbb{E}\,\mathsf{R}_\phi(\boldsymbol{X}^{m,n}) \le \frac{1}{\log_2|\mathcal{X}|}\Big\{H^+_{m,n}(X|U,V) + \frac{1}{mn}\Big\}\,. \tag{5.3}$$

Further, for any lossless compressor ϕ, $\mathbb{E}\,\mathsf{R}_\phi(\boldsymbol{X}^{m,n}) \ge H^+_{m,n}(X|U,V) - \delta_{m,n} - 2\log_2(mn)/mn$.

Remark 5.1. The simpler bound (5.1) implies that the entropy per entry is $H(X|U,V) + O(1/(m\wedge n))$. The operational interpretation of this result is that we should be able to achieve the same compression rate per symbol *as if* the latents were given to us.

The additional terms $\frac{1}{n}H(U)+\frac{1}{m}H(V)$ in Eq. (5.2) account for the additional memory required for the latents. The lower bound in Eq. (5.2) implies that, if the latents can be accurately estimated from the data $\boldsymbol{X}^{m,n}$ (that is if ε_U, ε_V are small), then this overhead is essentially unavoidable.

The nearly ideal compression rate in Eq. (5.3) can be achieved by Huffmann or arithmetic coding, and requires knowledge of the probability distribution of $\boldsymbol{X}^{m,n}$. Under the these schemes, the length of the codeword associated to $\boldsymbol{X}^{m,n}$ is within constant number of bits from $-\log_2 \mathrm{P}(\boldsymbol{X}^{m,n})$, where $\mathrm{P}(\boldsymbol{X}_0) := \mathbb{P}(\boldsymbol{X}^{m,n} = \boldsymbol{X}_0)$ is the probability mass function of the random table $\boldsymbol{X}^{m,n}$ (Cover & Thomas, 2006; Salomon, 2004). The next lemma implies that the length concentrates tightly around the entropy.

Lemma 5.3 (Asymptotic Equipartition Property)**.** *For* $\boldsymbol{X}_0 \in \mathcal{X}^{m\times n}$, *let* $\mathrm{P}(\boldsymbol{X}_0) = \mathrm{P}_{Q,\mathsf{q_r},\mathsf{q_c};m,n}(\boldsymbol{X}_0)$ *the probability of* $\boldsymbol{X}^{m,n} = \boldsymbol{X}_0$ *under model* $\boldsymbol{X}^{m,n} \sim \mathcal{T}(Q,\mathsf{q_r},\mathsf{q_c};m,n)$. *Assume there exists a constant* $c>0$ *such that* $\min_{x\in\mathcal{X}}\min_{u,v\in\mathcal{L}} Q(x|u,v) \ge c$. *Then there exists a constant* C *(depending on* c*) such that the following happens.*

For $\boldsymbol{X}^{m,n} \sim \mathcal{T}(Q,\mathsf{q_r},\mathsf{q_c};m,n)$ *and any* $t\ge 0$ *with probability at least* $1-2\,e^{-t}$:

$$\big|-\log \mathrm{P}(\boldsymbol{X}^{m,n}) - H(\boldsymbol{X}^{m,n})\big| \le C\sqrt{mn(m+n)}\,t\,. \tag{5.4}$$

For the sake of simplicity, in the last statement we assume a uniform lower bound on $Q(x|u,v)$. While such a lower bound holds without loss of generality when Q is independent of m,n (symbols with zero probability can be dropped), it might not hold in the n-dependent case. Appendix D gives a more general statement.

5.2. Failure of classical compression schemes

We analyze two types of codes: finite-state encoders and Lempel-Ziv codes. Both operate on the serialized data $\boldsymbol{X}^N = \mathsf{vec}(\boldsymbol{X}^{m,n})$, $N=mn$, obtained by scanning the table in row-first order (obviously column-first yields symmetric results).

5.2.1. Finite state encoders

A finite state (FS) encoder takes the form of a triple (Σ, f, g) with Σ a finite set of cardinality $M=|\Sigma|$ and $f : \mathcal{X}\times\Sigma \to \{0,1\}^*$, $g : \mathcal{X}\times\Sigma\to\Sigma$.

We assume that Σ contains a special 'initialization' symbol s_{init}. Starting from state $s_0 = s_{\text{init}}$, the encoder scans the input $\boldsymbol{X}^N$ sequentially. Assume after the first ℓ input symbols it is in state s_ℓ, and produced encoding $\boldsymbol{z}_1^{k(\ell)}$. Given input symbol $X_{\ell+1}$, it appends $f(X_{\ell+1}, s_\ell)$ to the codeword, and updates its state to $s_{\ell+1} = g(X_{\ell+1}, s_\ell)$.

With an abuse of notation, denote by $f_\ell(\boldsymbol{X}^\ell, s_{\text{init}}) \in \{0,1\}^*$ the binary sequence obtained by applying the finite state encoder to $\boldsymbol{X}^\ell = (X_1,\dots,X_\ell)$ We say that the FS encoder is information lossless if for any $\ell\in\mathbb{N}$, $\boldsymbol{X}^\ell \mapsto f_\ell(\boldsymbol{X}^\ell, s_{\text{init}})$ is injective.

Theorem 5.4. *Let* $\boldsymbol{X} = \boldsymbol{X}^{m,n} \sim \mathcal{T}(Q,\mathsf{q_r},\mathsf{q_c};m,n)$ *and* $\phi := (\Sigma, f, g)$ *be an information lossless finite state encoder. Define the corresponding compression rate* $\mathsf{R}_\phi(\boldsymbol{X})$, *as per Eq. (3.1). Assuming* $m > 10$, $|\Sigma| \ge |\mathcal{X}|$, *and* $\log_2|\Sigma| \le n\log_2|\mathcal{X}|/9$,

$$\mathbb{E}\,\mathsf{R}_\phi(\boldsymbol{X}) \ge \frac{H(X|U)}{\log_2|\mathcal{X}|} - 10\sqrt{\frac{\log|\Sigma|}{n\log|\mathcal{X}|}}\cdot\log(n\log|\Sigma|)\,. \tag{5.5}$$

The asymmetry between U and V in the last statement (and below) arises because we assume that the table is serialized in row-major order. Of course the roles of U and V are exchanged if we use column major.

Remark 5.2. The leading term of the above lower bound is $H(X|U)/\log_2|\mathcal{X}|$. Since conditioning reduces entropy, this is strictly larger than the ideal rate which is roughly $H(X|U,V)/\log_2|\mathcal{X}|$, cf. Eq. (5.3).

The next term is negligible provided $\log|\Sigma| \ll n\log|\mathcal{X}|$. This condition is easy to interpret: it amounts to say that the finite state machine does not have enough states to memorize a row of the table $\boldsymbol{X}^{m,n}$.

The gap between $H(X|U,V)$ (appearing in the ideal rate of Lemma 5.1) and $H(X|U)$ (appearing in the last statement

and in the analysis of LZ encoders) can be illustrated by a toy example. Assume u_i, v_j are uniform in $\{0,1\}$ and $X_{ij} = u_i + v_i (\mod 2)$. Then $H(X|U,V) = 0$ while $H(X|U) = H(X|V) = \log 2$. It is easy to compress very well by identifying the latents (just store the latents), but if we scan a single row (or column), we will only see a random sequence of bits.

5.2.2. LEMPEL-ZIV

The pseudocode of the Lempel-Ziv algorithm that we will analyze is given in Appendix F.

In words, after the first k characters of the input have been parsed, the encoder finds the longest string $\boldsymbol{X}_k^{k+\ell-1}$ which appears in the past. It then encodes a pointer to the position of the earlier appearance of the string T_k, and its length L_k. If a simbol X_k never appeared in the past, we use a special encoding, cf. Appendix F.

We encode the pointer T_k in plain binary using $\lceil \log_2(N + |\mathcal{X}|) \rceil$ bits (note that $T_k \in \{-|\mathcal{X}|+1, \dots, 1, \dots, N\}$), and L_k using an instantaneous prefix-free code, e.g. Elias δ-coding, taking $2\lfloor \log_2 L_k \rfloor + 1$ bits.

Assumption 5.5. There exist a constant $c_0 > 0$ such that

$$\max_{x \in \mathcal{X}} \max_{u,v \in \mathcal{L}} Q(x|u,v) \le 1 - c_0 \,.$$

Further $Q, q_{\mathrm{r}}, co, \mathcal{X}, \mathcal{L}$ are fixed and $m, n \to \infty$ with $m = n^{\alpha + o(1)}$, i.e.

$$\lim_{n\to\infty} \frac{\log m}{\log n} = \alpha \in (0, \infty) \,. \tag{5.6}$$

Theorem 5.6. *Under Assumption 5.5, the asymptotic Lempel-Ziv rate is*

$$\lim_{m,n\to\infty} \mathbb{E}\, \mathsf{R}_{\mathsf{LZ}}(\boldsymbol{X}^{m,n}) = \mathsf{R}^{\infty}_{\mathsf{LZ}} := \sum_{u \in \mathcal{L}} \frac{q_{\mathrm{r}}(u) \mathsf{R}^{\infty}_{\mathsf{LZ}}(u)}{\log_2 |\mathcal{X}|} \,, \tag{5.7}$$

$$\mathsf{R}^{\infty}_{\mathsf{LZ}}(u) := H(X|U=u) \wedge \Big(\frac{1+\alpha}{\alpha}\Big) H(X|U=u,V) \,.$$

Remark 5.3. The asymptotics of the Lempel-Ziv rate is given by the minimum of two expressions, which correspond to different behaviors of the encoder. For $u \in \mathcal{L}$, define $\alpha_*(u) := H(X|U=u,V)/(H(X|U=u) - H(X|U=u,V))$ (with $\alpha_*(u) = \infty$ if $H(X|U=u) = H(X|U=u,V)$). Then:

If $\alpha < \alpha_*(u)$, then we are a 'skinny table' regime. The algorithm mostly deduplicates segments in rows with latent u by using strings in different rows but aligned in the same columns. If $\alpha > \alpha_*(u)$, then we are a 'fat table' regime. The algorithm mostly deduplicates segments on rows with latent u by using rows and columns that are not the same as the current segment.

Example 5.4 (Symmetric Binary Model, dense regime)**.** Under the Symmetric Binary Model $\mathcal{T}_{\mathrm{SBM}}(p_0, p_1, k; m, n)$ of Example 4.2, we can compute the optimal compression rate of Corollary 5.2, the finite state compression rate of Theorem 5.4, the Lempel-Ziv rate of Theorem 5.6.

If p_0, p_1 are of order one, and $m = n^{\alpha + o_n(1)}$ as $m, n \to \infty$, letting $\overline{p} := ((k-1)/k)p_0 + (1/k)\, p_1$, $\overline{\mathsf{h}}(p_0, p_1) := ((k-1)/k)\mathsf{h}(p_0) + (1/k)\, \mathsf{h}(p_1)$, we obtain:

$$\begin{aligned}
\mathbb{E}\, \mathsf{R}_{\text{opt}}(\boldsymbol{X}) &= \Big(1 - \frac{1}{k}\Big)\, \mathsf{h}(p_0) + \frac{1}{k}\, \mathsf{h}(p_1) + o_n(1) \,, \\
\mathbb{E}\, \mathsf{R}_{\text{fin. st.}}(\boldsymbol{X}) &\ge \mathsf{h}(\overline{p}) + o_n(1) \,, \\
\mathbb{E}\, \mathsf{R}_{\mathsf{LZ}}(\boldsymbol{X}) &= \mathsf{h}(\overline{p}) \wedge \Big(\frac{1+\alpha}{\alpha}\Big) \overline{\mathsf{h}}(p_0, p_1) + o_n(1) \,.
\end{aligned}$$

These theoretical predictions are used to trace the contour lines in Figure 1. (ANS coding is implemented as a finite state code here.)

5.3. Practical latent-based compression

Achieving the ideal rate of Corollary 5.2 via arithmetic or Huffmann coding requires to compute the probability $\mathrm{P}(\boldsymbol{X}^{m,n})$, which is intractable. We will next show that we can achieve a compression rate that is close to the ideal rate via latents estimation.

We begin by considering general latents estimators $\widehat{\boldsymbol{u}} : \mathcal{X}^{m\times n} \to \mathcal{L}^m$, $\widehat{\boldsymbol{v}} : \mathcal{X}^{m\times n} \to \mathcal{L}^n$. We measure their accuracy by the error (cf. Lemma 5.1)

$$\mathsf{Err}_U(\boldsymbol{X}; \widehat{\boldsymbol{u}}) := \frac{1}{m} \min_{\pi \in \mathfrak{S}_{\mathcal{L}}} \sum_{i=1}^{m} \Big\{ \mathbf{1}_{\hat{u}_i(\boldsymbol{X}) \neq \pi(u_i)} \Big\}$$

and the analogous $\mathsf{Err}_V(\boldsymbol{X}; \widehat{\boldsymbol{v}})$. Here the minimization is over the set $\mathfrak{S}_{\mathcal{L}}$ of permutations of the latents alphabet $\mathcal{L}$.

We can use any estimators $\widehat{\boldsymbol{u}}, \widehat{\boldsymbol{v}}$ to reorder rows and columns and compress the table $\boldsymbol{X}^{m,n}$ according to the algorithm described in the introduction. We denote by $\mathsf{R}_{\text{lat}}(\boldsymbol{X})$ the compression rate achieved by such a procedure.

Our first result implies that, if the latent estimators are consistent (namely, they recover the true latents with high probability, up to permutations), then the resulting rate is close to the ideal one.

Lemma 5.7. *Assume data distributed according to model $\boldsymbol{X}^{m,n} \sim \mathcal{T}(Q, q_{\mathrm{r}}, q_{\mathrm{c}}; m, n)$, with $m, n \ge \log_2 |\mathcal{L}|$. Further assume there exists $c_0 > 0$ such that $q_{\mathrm{r}}(u), q_{\mathrm{c}}(v) \ge c_0$ for all $u, v \in \mathcal{L}$. Let $\mathsf{R}_{\text{lat}}(\boldsymbol{X})$ be the rate achieved by the latent-based scheme with latents estimators $\widehat{\boldsymbol{u}}$, $\widehat{\boldsymbol{v}}$, and base encoders $\mathsf{Z}_{\mathcal{X}} = \mathsf{Z}_{\mathcal{L}} = \mathsf{Z}$. Then*

$$\begin{aligned}
\mathbb{E}\, \mathsf{R}_{\text{lat}}(\boldsymbol{X}) \le{}& \frac{H(\boldsymbol{X}^{m,n})}{mn \log_2 |\mathcal{X}|} + 2P_{\text{err}}(m,n) + \frac{4\log(mn)}{mn} \\
&+ |\mathcal{L}|^2 \Delta_{\mathsf{Z}}(c \cdot mn; \mathscr{Q}) + 2\Delta_{\mathsf{Z}}(m \wedge n; \{q_{\mathrm{r}}, q_{\mathrm{c}}\}) \,.
\end{aligned} \tag{5.8}$$

Here $P_{\text{err}}(m,n) := \mathbb{P}(\mathsf{Err}_U(\boldsymbol{X}^{m,n};\widehat{\boldsymbol{u}}) > 0) + \mathbb{P}(\mathsf{Err}_V(\boldsymbol{X}^{m,n};\widehat{\boldsymbol{v}}) > 0)$, $\Delta_{\mathsf{Z}}(N;\mathscr{P}_*)$ *is the worst-case redundancy of encoder* Z *over i.i.d. sources with distributions in* $\mathscr{P}_*$ *(see comments below),* $\mathscr{Q} := \{Q(\cdot\,|u,v)\}_{u,v\in\mathcal{L}}$.

The redundancies of Lempel-Ziv, frequency-based arithmetic coding and ANS coding can be upper bounded as (in the last bound $Q, q_{\mathsf{r}}, q_{\mathsf{c}}$ *need to be be independent of* N*)*

$$\Delta_{\mathsf{LZ}}(N;\mathscr{P}_*) \le 40 c_*(\mathscr{P}_*)\Big(\frac{\log\log N}{\log N}\Big)^{1/2}, \tag{5.9}$$

$$\Delta_{\mathsf{AC}}(N;\mathscr{P}_*) \le \frac{2|\mathcal{X}|}{\log|\mathcal{X}|}\cdot\frac{\log N}{N}, \tag{5.10}$$

$$\Delta_{\mathsf{ANS}}(N;\mathscr{P}_*) \le \frac{2|\mathcal{X}|\log N + C_{|\mathcal{X}|}}{N}. \tag{5.11}$$

Here Eq. (5.9) holds for $N \ge \exp\{\sup_{q\in\mathscr{P}_*}(4\log(2/H(q)))^2\}$, *and* $c_*(\mathscr{P}_*) := \sup_{q\in\mathscr{P}_*}\sum_{x\in\mathcal{X}}(\log q(x))^2/|\mathcal{X}|$.

The proof of this lemma is given in Appendix G.1. The main content of the lemma is in the general bound (5.8) which is proven in Appendix G.1.1.

Remark 5.5. We define the worst case redundancy $\Delta_{\mathsf{Z}}(N_0;\mathscr{P}_*) := \max_{N\ge N_0}\widehat{\Delta}_{\mathsf{Z}}(N;\mathscr{P}_*)$, where

$$\widehat{\Delta}_{\mathsf{Z}}(N;\mathscr{P}_*) := \max_{q\in\mathscr{P}_*}\frac{\mathbb{E}_q\mathsf{len}(\mathsf{Z}(\boldsymbol{Y}^N)) - H(\boldsymbol{Y}^N)}{N\log_2 k}, \tag{5.12}$$

where $\mathscr{P}_* \subseteq \mathscr{P}([k]) := \{(p_i)_{i\le k}\in\mathbb{R}^k : p_i \ge 0\,\forall i \text{ and } \sum_{i\le k}p_i = 1\}$ is a set of probability distributions over $[k]$ and $\boldsymbol{Y}^N$ is a vector with i.i.d. entries $Y_i \sim q$.

While Eqs. (5.9)—(5.11) are closely related to well known facts, there are nevertheless differences with respect to statements in the literature. We address them in Section G.1.2. Perhaps the most noteworthy difference is in the bound (5.9) for the LZ algorithm. Existing results, e.g. Theorem 2 in (Savari, 1998), assume a single, N-independent, distribution q and are asymptotic in nature. Equation (5.9) is a non-asymptotic statement and applies to a collection of distributions $\mathscr{P}_*$ that could depend on N.

Lemma 5.7 can be used in conjunction with any latent estimation algorithm, as we next demonstrate by considering the spectral algorithm of Section 2.2. Recall that the algorithm makes use of a map $\psi : \mathcal{X}\to\mathbb{R}$. For $(X,U,V)\sim Q(\cdot\,|\cdot,\cdot)q_{\mathsf{r}}(\cdot)q_{\mathsf{c}}(\cdot)$, we define $\overline{\psi}(u,v) := \mathbb{E}[\psi(X)|U=u,V=v]$, $\boldsymbol{\Psi} := \big(\overline{\psi}(u,v)\big)_{u,v\in\mathcal{L}}$ and the parameters:

$$\mu_n := \sigma_{\min}(\boldsymbol{\Psi}), \quad \nu_n := \max_{u,v\in\mathcal{L}}|\overline{\psi}(u,v)|, \tag{5.13}$$

$$\sigma_n^2 := \max_{u,v\in\mathcal{L}}\mathrm{Var}\big(\psi(X)|U=u,V=v\big). \tag{5.14}$$

We further will assume, without loss of generality $\max_{x\in\mathcal{X}}|\psi(x)|\le 1$.

Finally, we need to formally specify the version of the KMeans primitive in the spectral clustering algorithm. In fact, we establish correctness for a simpler thresholding procedure. Considering to be definite the row latents, and for a given threshold $\theta>0$, we construct a graph $G_\theta = ([m], E_\theta)$ by letting (for distinct $i,j\in[m]$)

$$\frac{\|\boldsymbol{a}_i-\boldsymbol{a}_j\|_2}{(\|\boldsymbol{a}_i\|_2+\|\boldsymbol{a}_j\|_2)/2} \le \theta \;\;\Leftrightarrow\;\; (i,j)\in E_\theta. \tag{5.15}$$

The algorithm then output the connected components of G_θ.

Theorem 5.8. *Assume data* $\boldsymbol{X}^{m,n}\sim\mathcal{T}(Q,q_{\mathsf{r}},q_{\mathsf{c}};m,n)$, *with* $m,n\ge\log_2|\mathcal{L}|$ *and* $\min_{u\in\mathcal{L}}(q_{\mathsf{r}}(u)\wedge q_{\mathsf{c}}(u))\ge c_0$ *for a constant* $c_0>0$. *Let* $\mathsf{R}_{\text{lat}}(\boldsymbol{X})$ *be the rate achieved by the latent-based scheme with spectral latents estimators* $\widehat{\boldsymbol{u}}$, $\widehat{\boldsymbol{v}}$, *base compressors* $\mathsf{Z}_{\mathcal{X}} = \mathsf{Z}_{\mathcal{L}} = \mathsf{Z}$, *and thresholding algorithm as described above. Then, assuming* $\sigma_n \ge c\sqrt{(\log n)/n}$, $\nu_n/\sigma_n \le c\sqrt{\log n}$ $\mu_n \ge C(\sigma_n\sqrt{(\log n)/m}\vee(\log n)/\sqrt{mn})$, $\theta\le\sqrt{c_0}/100$, *we have*

$$\mathbb{E}\,\mathsf{R}_{\text{lat}}(\boldsymbol{X}) \le \frac{H(\boldsymbol{X}^{m,n})}{mn\log_2|\mathcal{X}|} + \frac{10\log(mn)}{mn} + |\mathcal{L}|^2\Delta_{\mathsf{Z}}(c\cdot mn;\mathscr{Q}) + 2\Delta_{\mathsf{Z}}(m\wedge n;\{q_{\mathsf{r}},q_{\mathsf{c}}\}).$$

We focus on the simpler thresholding algorithm of Eq. (5.15) instead of KMeans in order to avoid technical complications that are not the main focus of this paper. We expect it to be relatively easy to generalize this result, e.g. using the results of (Makarychev et al., 2020) for KMeans++.

Example 5.6. Consider the Symmetric Binary Model $\mathcal{T}_{\text{SBM}}(p_0,p_1,k;m,n)$ of Example 4.2, with $p_0 = p_{0,n}$, $p_1 = p_{1,n}$ potentially dependent on n. Since in this case $\mathcal{X} = \{0,1\}$ the choice of the map ψ has little impact and we set $\psi(x)=x$. We assume, to simplify formulas, $|p_{1,n}-p_{0,n}|\le kp_{0,n}$, $p_{1,n}\vee p_{0,n}\le 9/10$. It is easy to compute $\mu_n = |p_{1,n}-p_{0,n}|$, $\nu_n = p_{0,n}\vee p_{1,n}$, $\sigma_n^2 \asymp p_{1,n}\vee p_{0,n}$:

Theorem 5.8 implies nearly optimal compression rate under the following conditions on the model parameters:

$$p_{1,n}\vee p_{0,n} \gtrsim \sqrt{\frac{\log n}{n}}, \qquad |p_{1,n}-p_{0,n}| \gtrsim \frac{\log n}{\sqrt{mn}},$$
$$\frac{|p_{1,n}-p_{0,n}|}{p_{1,n}\vee p_{0,n}} \gtrsim \sqrt{\frac{\log n}{n}}.$$

Here $\gtrsim$ hides factors depending on k, c_0. The last of these condition amounts to requiring that the signal-to-noise ratio is large enough to consistently reconstruct the latents. In the special case of square symmetric matrices ($m=n$), sharp constants in these bounds can be derived from (Abbe, 2017).

Method	DRR ($k=4$)	($k=6$)	($k=10$)
Oracle	0.27	0.27	0.27
No Latent	0.06	0.11	0.17
$k_{\text{alg}} = k-2$	0.14	0.21	0.25
$k_{\text{alg}} = k-1$	0.21	0.25	0.26
$k_{\text{alg}} = k$	0.27	0.27	0.26
$k_{\text{alg}} = k+1$	0.27	0.27	0.26
$k_{\text{alg}} = k+2$	0.25	0.27	0.23

Method	Taxi	Census	Forest	Jokes
ZSTD 22	0.53 sec	146 sec	17.2 sec	2.8 sec
Latents + LZ	1.2 sec	96 sec	5.2 sec	3.8 sec
Subsampling	0.7 sec	30 sec	3.7 sec	1.5 sec

Table 2: Left: DRR of the latent-based compressor inthe SBM with miss-specified latent space size k_{alg}. Right: average runtime (seconds) of latent-based compressors on several datasets, compared with state of the art.

6. Discussion and extensions

We proved that classical lossless compression schemes, that serialize the data and then apply a finite state encoder or a Lempel-Ziv encoder to the resulting sequence are sub-optimal when applied to tabular data. Namely, we introduced a simple model for tabular data, and made the following novel contributions:

1. We characterized the optimal compression rate under this model.

2. We rigorously quantified the gap in compression rate suffered by classical compressors.

3. We showed that a compression scheme that estimates the latents performs well in practice and provably achieves optimal rate on our model.

The present work naturally suggests several questions and directions for future work.

Model miss-specification. The numerical simulations of Section 4 were carried our within the Symmetric Binary Model, under the assumption that the correct cardinality of the latents space, k, is known at the encoder. While the rigorous guarantees of Section 5.3 are more general, it is important to understand the effect of model misspecification.

Table 2 presents initial evidence of the robustness of the proposed approach. We generate tables $\boldsymbol{X}^{m,n} \sim \mathcal{T}_{\text{SBM}}(p_0, p_1, k; m, n)$, with $p_0 = 0.2$, $p_1 = 0.8$, and dimensions $m = n = 1000$. We vary $k \in \{4, 6, 8\}$ and run the latent-based compressor in a miss-specified setting by using a latentss-space size $k_{\text{alg}} \neq k$. We observe that the resulting DRR is fairly robust to miss-specification in the the number of latents.

It is also worth pointing out that parameters of the latents model, such as k_{alg}, can be chosen at compression time to optimize DRR. However this implies severe performance losses, and therefore further investigation into model misspecification is warranted.

Computational efficiency. While we did not attempt to develop a highly optimized compressor, we believe that the present approach is amenable to such a development. The largest overhead over standard compressors is in the clustering step, and in particular computing the singular value decomposition (SVD) of the data. This can be accelerated using methods from randomized linear algebra. We implemented row subsampling SVD (Drineas et al., 2006), and observed essentially no loss in DRR by using 10% of the rows as compared to full SVD.

Table 2 reports running time experiments to compare the original latent based compressor, its subsampling version and ZSTD at its strongest compression level. Runtimes were averaged over 5 runs on a Macbook Pro single-threaded with a 2 GHz 4-core Intel i5 chip.

We observe that row-subsampling yields significant performance improvements, and runtimes that compare well with the industry state of the art.

Bits back coding. As mentioned several times, encoding the latents is sub-optimal, unless these can be estimated accurately in the sense of $\mathbb{E}\,\mathsf{Err}_U(\boldsymbol{X}; \widehat{\boldsymbol{u}}) \to 0$, $\mathbb{E}\,\mathsf{Err}_V(\boldsymbol{X}; \widehat{\boldsymbol{v}}) \to 0$, cf. Lemma 5.1. If this is not the case, then optimal rates can be achieved using bits-back coding.

Continuous latents. Of course, using discrete latents is somewhat un-natural, and it would be interesting to consider continuous ones, in which case bits-back coding is required.

Multi-way tables (tensors). A natural extension of the current work is to order-s multiway tables or (equivalently for our purposes) tensors $\boldsymbol{X} \in \mathcal{X}^{n_1 \times \cdots \times n_s}$. The basic scheme would essentially be the same: estimate s vector of latents $\boldsymbol{v}_i \in \mathcal{L}^{n_i}$, $i \le s$ and use them to partition rows/columns.

Acknowledgements

This work was carried out while Andrea Montanari was on leave from Stanford and a Chief Scientist at Ndata Inc dba Project N. The present research is unrelated to AMs Stanford activity.

References

Abbe, E. Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531, 2017.

Abbe, E., Fan, J., Wang, K., and Zhong, Y. Entrywise eigenvector analysis of random matrices with low expected rank. *Annals of statistics*, 48(3):1452, 2020.

Altman, E. R. Synthesizing credit card transactions, 2019. https://arxiv.org/abs/1910.03033.

Besta, M. and Hoefler, T. Survey and taxonomy of lossless graph compression and space-efficient graph representations. *arXiv preprint arXiv:1806.01799*, 2018.

Boldi, P. and Vigna, S. The webgraph framework i: compression techniques. In *Proceedings of the 13th international conference on World Wide Web*, pp. 595–602, 2004.

Chen, P., Yu, H.-F., Dhillon, I., and Hsieh, C.-J. Drone: Data-aware low-rank compression for large nlp models. *Advances in neural information processing systems*, 34: 29321–29334, 2021.

Cheng, H., Gimbutas, Z., Martinsson, P.-G., and Rokhlin, V. On the compression of low rank matrices. *SIAM Journal on Scientific Computing*, 26(4):1389–1404, 2005.

Chierichetti, F., Kumar, R., Lattanzi, S., Mitzenmacher, M., Panconesi, A., and Raghavan, P. On compressing social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 219–228, 2009.

Collet, Y. and Kucherawy, M. Zstandard compression and the application/zstd media type. Technical report, 2018.

Cover, T. M. and Thomas, J. A. *Elements of information theory*. Wiley, 2006.

Drineas, P., Kannan, R., and Mahoney, M. W. Fast monte carlo algorithms for matrices ii: Computing a low-rank approximation to a matrix. *SIAM Journal on computing*, 36(1):158–183, 2006.

Dua, D. and Graff, C. UCI machine learning repository, 2017. http://archive.ics.uci.edu/ml.

Duda, J. Asymmetric numeral systems. *arXiv:0902.0271*, 2009.

Duda, J. Asymmetric numeral systems: entropy coding combining speed of huffman coding with compression rate of arithmetic coding. *arXiv preprint arXiv:1311.2540*, 2013.

Frey, B. J. and Hinton, G. E. Free energy coding. In *Proceedings of Data Compression Conference-DCC'96*, pp. 73–81. IEEE, 1996.

Goldberg, K., Roeder, T., Gupta, D., and Perkins, C. Jester Datasets for Recommender Systems and Collaborative Filtering Research. https://eigentaste.berkeley.edu/dataset/.

Goldberg, K., Roeder, T., Gupta, D., and Perkins, C. Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval*, 4(2):133–151, 2001. https://doi.org/10.1023/A:1011419012209.

Hinton, G. E. and Zemel, R. Autoencoders, minimum description length and helmholtz free energy. *Advances in neural information processing systems*, 6, 1993.

Holland, P. W., Laskey, K. B., and Leinhardt, S. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

Hou, J., Chau, L.-P., Magnenat-Thalmann, N., and He, Y. Sparse low-rank matrix approximation for data compression. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(5):1043–1054, 2015.

IBM. Stats NZ Business Price indexes: March 2022 quarter. https://www.stats.govt.nz/information-releases/business-price-indexes-march-2022-quarter/, March 2022.

Kosolobov, D. The efficiency of the ans entropy encoding. *arXiv preprint arXiv:2201.02514*, 2022.

Leskovec, J. and Krevl, A. SNAP Datasets: Stanford large network dataset collection, June 2014. http://snap.stanford.edu/data.

Li, N. and Li, B. Tensor completion for on-board compression of hyperspectral images. In *2010 IEEE International Conference on Image Processing*, pp. 517–520. IEEE, 2010.

Lim, Y., Kang, U., and Faloutsos, C. Slashburn: Graph compression and mining beyond caveman communities. *IEEE Transactions on Knowledge and Data Engineering*, 26(12):3077–3089, 2014.

Makarychev, K., Reddy, A., and Shan, L. Improved guarantees for k-means++ and k-means++ parallel. *Advances in Neural Information Processing Systems*, 33:16142–16152, 2020.

NYC.gov. TLC Trip Record Data: NYC Taxi & Limousine Commission, January 2022.

Phan, A.-H., Sobolev, K., Sozykin, K., Ermilov, D., Gusak, J., Tichavskỳ, P., Glukhov, V., Oseledets, I., and Cichocki, A. Stable low-rank tensor decomposition for compression

of convolutional neural network. In *European Conference on Computer Vision*, pp. 522–539. Springer, 2020.

Salomon, D. *Data compression: the complete reference*. Springer Science & Business Media, 2004.

Savari, S. A. Redundancy of the lempel-ziv string matching code. *IEEE Transactions on Information Theory*, 44(2): 787–791, 1998.

stats.govt.nz. Stats NZ Business Price indexes: March 2022 quarter. https://www.stats.govt.nz/information-releases/business-price-indexes-march-2022-quarter/, March 2022.

Taylor, P. R. Lossless compression of wave function information using matrix factorization: A gzip for quantum chemistry. *The Journal of Chemical Physics*, 139(7): 074113, 2013.

Townsend, J., Bird, T., and Barber, D. Practical lossless compression with latent variables using bits back coding. In *7th International Conference on Learning Representations, ICLR 2019*, volume 7. International Conference on Learning Representations (ICLR), 2019a.

Townsend, J., Bird, T., Kunze, J., and Barber, D. Hilloc: lossless image compression with hierarchical latent variable models. In *International Conference on Learning Representations*, 2019b.

Yuan, Z. and Oja, E. Projective nonnegative matrix factorization for image compression and feature extraction. In *Scandinavian Conference on Image Analysis*, pp. 333–342. Springer, 2005.

A. Details on the empirical evaluation

A.1. Datasets

We used the following datasets:

- *Taxicab.* A table with $m = 62,495$ rows, $n_0 = 20$ columns comprising data for taxi rides in NYC during January 2022 (NYC.gov, 2022). After preprocessing this table has $n = 18$ columns. For the LZ (ZSTD) compressor we used $|\mathcal{L}_r| = 9$ row latents and 15 column latents, for the ANS compressor we used $|\mathcal{L}_r| = 5$ row latents and 14 column latents.
- *Network.* Four social networks from SNAP Datasets, representing either friends as undirected edges for Facebook or directed following relationships on Google Plus (Leskovec & Krevl, 2014). We regard these as four distinct tables with $0-1$ entries, with dimensions, respectively $m = n \in \{333, 747, 786, 1187\}$. For each table we used 5 row latents and 5 column latents.
- *Card transactions.* A table of simulated credit card transactions containing information like card ID, merchant city, zip, etc. This table has $m = 24,386,900$ rows and $n_0 = 15$ columns and was generated as described in (Altman, 2019) and downloaded from (IBM, 2022). After preprocessing the table has $n = 12$ columns. For this table we used 3 row latents and n column latents.
- *Business price index.* A table of the values of the consumer price index of various goods in New Zealand between 1996 and 2022. This is a table with $m = 72,750$ rows and $n_0 = 12$ columns from the Business price indexes: March 2022 quarter - CSV file from (stats.govt.nz, 2022). After preprocessing this table has $n = 10$ columns. Due to the highly correlated nature of consecutive rows, we first shuffle them before compressing. For the LZ method we used 6 row latents and 7 column latents, for the ANS method we used 2 row latents and 6 column latents.
- *Forest.* A table from the UCI data repository comprising $m = 581,011$ cartographic measurements with $n_0 = 55$ attributes, to predict forest cover type based on information gathered from US Geological Survey (Dua & Graff, 2017). It contains binary qualitative variables, and some continuous values like elevation and slope. After preprocessing this data has $n = 55$ columns. For the LZ method we used 6 row latents and 17 column latents, for the ANS method we used 6 row latents and 17 column latents.
- *US Census.* Another table from the UCI Machine Learning Repository (Dua & Graff, 2017) with $m = 2,458,285$ and $n_0 = 68$ categorical attributes related to demographic information, income, and occupation information. After preprocessing this data has $n = 68$ columns. For this data we used 9 row latents and n column latents.
- *Jokes.* A table containing ratings of a series of jokes by 24,983 users collected between April 1999 and May 2003 (Goldberg et al., 2001; Goldberg et al.). These ratings are real numbers on a scale from -10 to 10, and a value of 99 is given to jokes that were not rated. There are $m = 23,983$ rows and $n_0 = 101$. The first column identifies how many jokes were rated by a user, and the rest of the columns contain the ratings. After preprocessing this data has $n = 101$ columns, all quantized. For the LZ method we used 2 row latents and n column latents, for the ANS method we used 8 row latents and 8 column latents.

A.2. Preprocessing

We preprocessed different columns as follows:

- If a column comprises $K \leq 256$ unique values, then we map the values to $\{0, \ldots, K-1\}$.
- If a column is numerical and comprises more than 256 unique values, we calculate the quartiles for the data and map each entry to its quartile membership (0 for the lowest quartile, 1 for the next largest, 2 for the next and 3 for the largest).
- If a column does not meet either of the above criteria, we discard it.

Finally, in some experiments we randomly permuted before compression. The rationale is that some of the above datasets have rows already ordered in a way that makes nearby rows highly correlated. In these cases, row reordering is –obviously– of limited use.

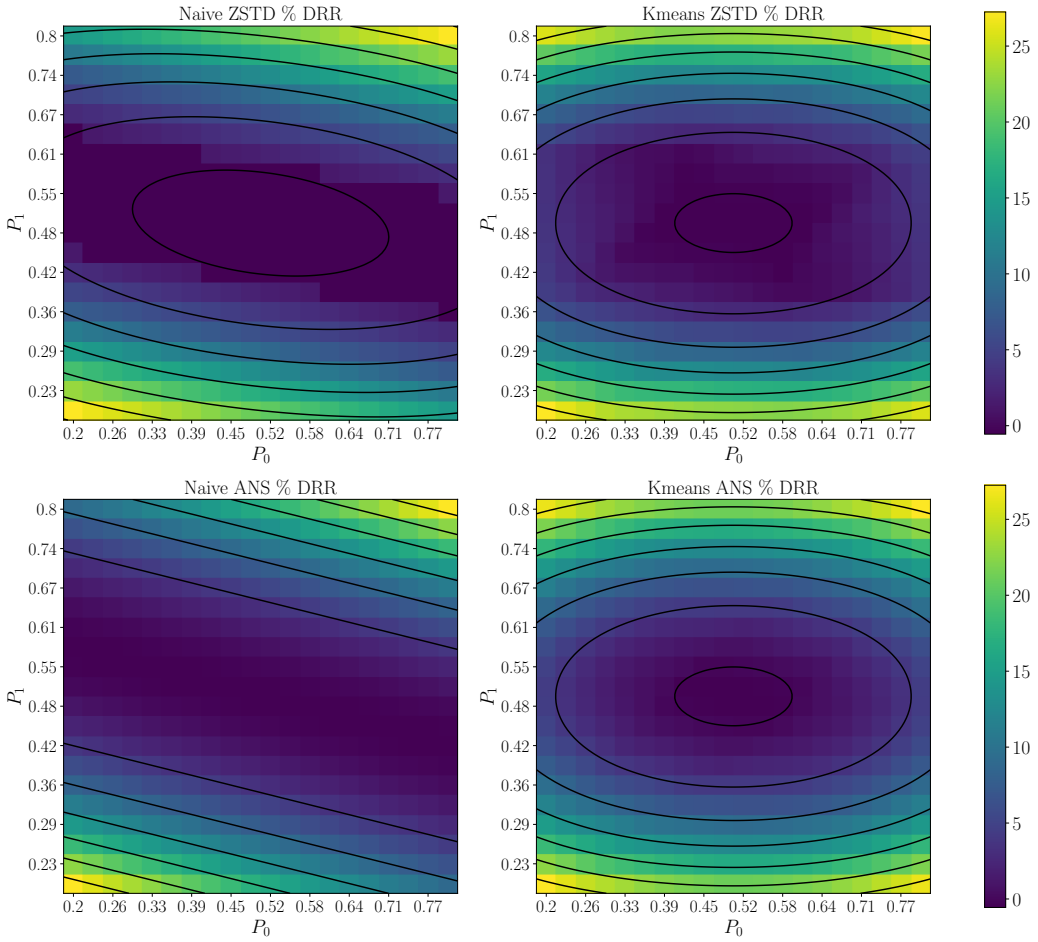

Figure 2: Comparing data reduction rate of naive coding and latent-based coding for data from SBM with $k = 5$ latents. Top row: ZSTD base compressor. Bottom row: ANS based compressor. Contour lines correspond to the theoretical predictions for various compression algorithms (cf. Example 5.4).

B. Further simulations

Figures 2 and 3 report empirical DRR values for ZSTD and ANS coding, for data generated according to the symmetric model $\mathcal{T}_{\text{SBM}}(p_0, p_1, k; m, n)$ of Section 4. We use $m = n = 1000$ as before, but now $k \in \{5, 7\}$. Results confirm the conclusions of Section 4.

C. A basic fact

Lemma C.1. *Let $\mathcal{A}$ be a finite set and $F : \mathcal{A} \to \{0,1\}^*$ be an injective map. Then, for any probability distribution p over $\mathcal{A}$,*

$$\sum_{a \in \mathcal{A}} p(a)\mathsf{len}(F(a)) \geq H(p) - \log_2 \log_2(|\mathcal{A}| + 2)\,. \tag{C.1}$$

Proof. Assume without loss of generality that $\mathcal{A} = \{1, \dots, M\}$, with $|\mathcal{A}| = K$, and that the elements of $\mathcal{A}$ have all non-vanishing probability and are ordered by decreasing probability $p_1 \geq p_2 \geq \cdots \geq p_K > 0$. Let $N_\ell := 2 + 4 + \cdots + 2^\ell = 2\ell + 1 - 2$. Then the expected length is minimized any map F such that $\mathsf{len}(F(a)) = \ell$ for $N_{\ell-1} \leq a \leq N_\ell$ with the

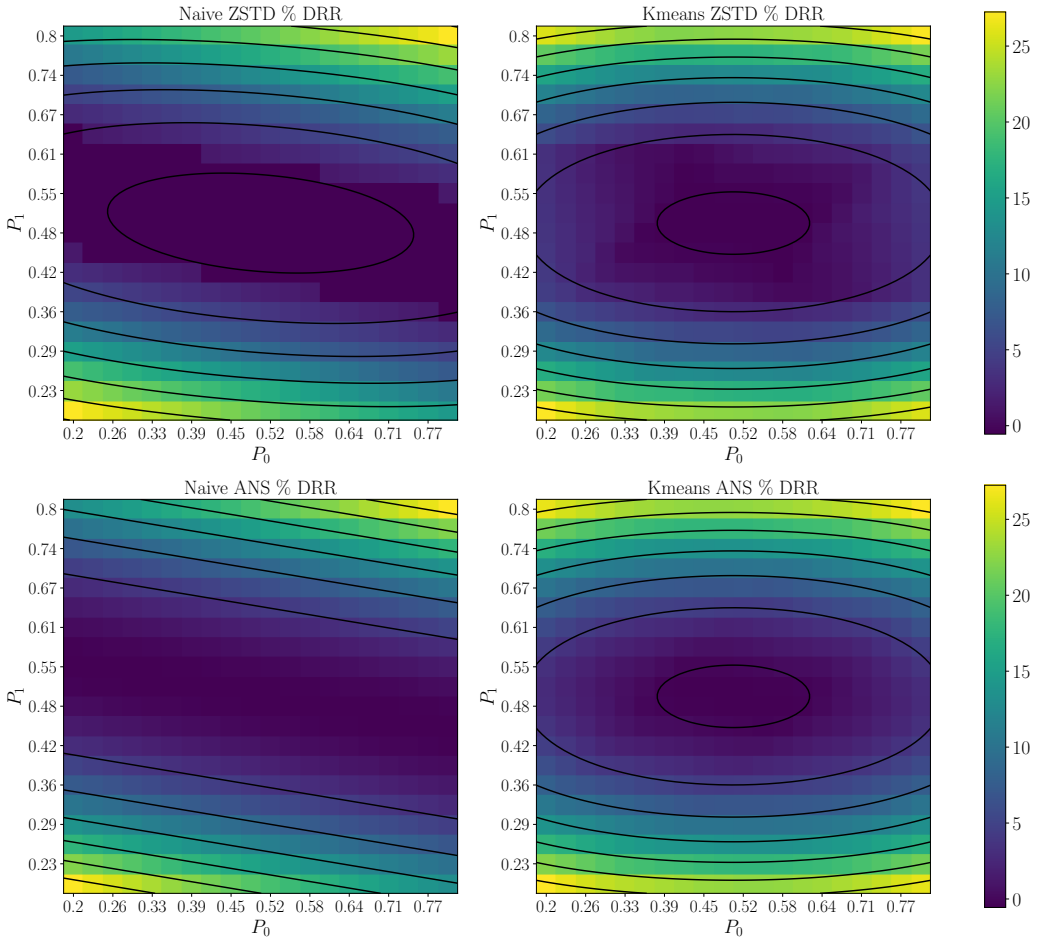

Figure 3: Same as Figure 2 with $k = 7$.

maximum length ℓ_K being defined by $N_{\ell_K - 1} < K \leq N_{\ell_K}$. For $A \sim p$, $L := \mathsf{len}(F(A))$, we have

$$\begin{aligned} H(p) := H(A) &\overset{(a)}{\leq} H(L) + H(A|L) \\ &\leq \log_2 \ell_K + \sum_{\ell=1}^{\ell_K} \mathbb{P}(L = \ell) H(A|L = \ell) \\ &\overset{(b)}{\leq} \log_2 \ell_K + \sum_{\ell=1}^{\ell_M} \mathbb{P}(L = \ell)\, \ell \\ &\leq \log_2 \log_2 (K+2) + \sum_{a \in \mathcal{A}} p(a) \mathsf{len}(F(a))\,, \end{aligned}$$

where (a) is the chain rule of entropy and (b) follows because by injectivity, given $\mathsf{len}(F(A)) = \ell$, A can take at most 2^ℓ values. □

D. Proofs of results on ideal compression

D.1. Proof of Lemma 5.1

We begin by claiming that

$$\frac{1}{mn} H(\boldsymbol{X}^{m,n}) = H(X|U,V) + \frac{1}{mn} I(\boldsymbol{X}^{m,n}; \boldsymbol{U}^m, \boldsymbol{V}^n)\,. \tag{D.1}$$

Indeed, by the definition of mutual information, we have $H(\boldsymbol{X}_{m,n}) = H(\boldsymbol{X}_{m,n}|\boldsymbol{U}_m, \boldsymbol{V}_n) + I(\boldsymbol{X}_{m,n}; \boldsymbol{U}_m, \boldsymbol{V}_n)$. Equation (D.1) follows by noting that

$$\begin{aligned}
H(\boldsymbol{X}^{m,n}|\boldsymbol{U}^m, \boldsymbol{V}^n) &= \sum_{\boldsymbol{u}\in\mathcal{L}^m} \sum_{\boldsymbol{v}\in\mathcal{L}^n} \mathbb{P}(\boldsymbol{U}^m = \boldsymbol{u}, \boldsymbol{V}^n = \boldsymbol{v})\, H(\boldsymbol{X}^{m,n}|\boldsymbol{U}^m = \boldsymbol{u}, \boldsymbol{V}^n = \boldsymbol{v}) \\
&\overset{(a)}{=} \sum_{i=1}^{m}\sum_{j=1}^{n} \sum_{\boldsymbol{u}\in\mathcal{L}^m} \sum_{\boldsymbol{v}\in\mathcal{L}^n} \mathbb{P}(\boldsymbol{U}^m = \boldsymbol{u}, \boldsymbol{V}^n = \boldsymbol{v})\, H(X_{i,j}|U_i = u, V_j = v) \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n} \sum_{\boldsymbol{u}\in\mathcal{L}} \sum_{\boldsymbol{v}\in\mathcal{L}} \mathbb{P}(U_i = u_i, V_j = v_j) H(X_{i,j}|U_i = u_i, V_j = v_j) \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n} H(X_{i,j}|U_i, V_j) \overset{(b)}{=} mnH(X_{1,1}|U_1, V_1)\,,
\end{aligned}$$

where (a) follows from the fact that the $(X_{i,j})$ are conditionally independent given $\boldsymbol{U}^m$, $\boldsymbol{V}^n$, ad since the conditional distribution of $X_{i,j}$ only depends on $\boldsymbol{U}^m$, $\boldsymbol{V}^n$ via U_i, V_j; (b) holds because the triples $(X_{i,j}, U_i, V_j)$ are identically distributed.

The lower bound in Eq. (5.1) holds because mutual information is non-negative, and the upper bound because $I(\boldsymbol{X}^{m,n}; \boldsymbol{U}^m, \boldsymbol{V}^n) \le H(\boldsymbol{U}^m, \boldsymbol{V}^n) = mH(U_1) + nH(V_1)$.

Finally, to prove Eq. (5.2), define

$$\pi_{\boldsymbol{U},\boldsymbol{X}} := \arg\min_{\pi\in\mathfrak{S}_{\mathcal{L}}} \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}_{\hat{u}_i \neq \pi(u_i)}\,, \quad \pi_{\boldsymbol{V},\boldsymbol{X}} := \arg\min_{\pi\in\mathfrak{S}_{\mathcal{L}}} \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\hat{v}_i \neq \pi(v_i)}\,, \tag{D.2}$$

If the minimizer is not unique, one can be chosen arbitrarily. We then have holds because

$$\begin{aligned}
I(\boldsymbol{X}^{m,n}; \boldsymbol{U}^m, \boldsymbol{V}^n) &= H(\boldsymbol{U}^m, \boldsymbol{V}^n) - H(\boldsymbol{U}^m, \boldsymbol{V}^n|\boldsymbol{X}^{m,n}) \\
&\ge mH(U_1) + nH(V_1) - H(\boldsymbol{U}^m|\boldsymbol{X}^{m,n}) - H(\boldsymbol{V}^n|\boldsymbol{X}^{m,n}) \\
&\ge mH(U_1) + nH(V_1) - \big[H(\boldsymbol{U}^m|\boldsymbol{X}^{m,n}, \pi_{\boldsymbol{U},\boldsymbol{X}}) + I(\boldsymbol{U}^m; \pi_{\boldsymbol{U},\boldsymbol{X}}|\boldsymbol{X}^{m,n})\big] \qquad \text{(D.3)} \\
&\qquad - \big[H(\boldsymbol{V}|\boldsymbol{X}^{m,n}, \pi_{\boldsymbol{V},\boldsymbol{X}}) + I(\boldsymbol{V}^n; \pi_{\boldsymbol{V},\boldsymbol{X}}|\boldsymbol{X}^{m,n})\big] \\
&\ge mH(U_1) + nH(V_1) - H(\boldsymbol{U}^m|\boldsymbol{X}^{m,n}, \pi_{\boldsymbol{U},\boldsymbol{X}}) - H(\boldsymbol{V}|\boldsymbol{X}^{m,n}, \pi_{\boldsymbol{V},\boldsymbol{X}}) - 2|\mathcal{L}|\log_2(|\mathcal{L}|)\,, \qquad \text{(D.4)}
\end{aligned}$$

where in the last inequality we used the fact that $I(\boldsymbol{U}^m; \pi_{\boldsymbol{U},\boldsymbol{X}}|\boldsymbol{X}^{m,n}) \le H(\pi_{\boldsymbol{U},\boldsymbol{X}}) \le \log_2(|\mathcal{L}|!)$.

Now, consider the term $H(\boldsymbol{U}^m|\boldsymbol{X}^{m,n}, \pi_{\boldsymbol{U},\boldsymbol{X}})$. Letting $\boldsymbol{Y} := (\boldsymbol{X}^{m,n}, \pi_{\boldsymbol{U},\boldsymbol{X}})$ the stated accuracy assumption implies that there exists an estimators $\widehat{\boldsymbol{u}}^+ = \widehat{\boldsymbol{u}}^+(\boldsymbol{Y})$ such that

$$\varepsilon_U = \frac{1}{m}\sum_{i=1}^{m} \varepsilon_{U,i}\,, \qquad \varepsilon_{U,i} := \mathbb{P}\big(\hat{u}_i^+(\boldsymbol{Y}) \neq U_i\big)\,.$$

By Fano's inequality

$$\begin{aligned}
H(\boldsymbol{U}^m|\boldsymbol{X}^{m,n}, \pi_{\boldsymbol{U},\boldsymbol{X}}) &\le \sum_{i=1}^{m} H(U_i|\boldsymbol{X}^{m,n}, \pi_{\boldsymbol{U},\boldsymbol{X}}) \\
&\le \sum_{i=1}^{m} \big[\mathsf{h}(\varepsilon_{U,i}) + \varepsilon_{U,i}\log(|\mathcal{L}| - 1)\big] \\
&\le m\big[\mathsf{h}(\varepsilon_U) + \varepsilon_U \log(|\mathcal{L}| - 1)\big]\,,
\end{aligned}$$

where the last step follows by Jensen's inequality. The claim (5.2) follows by substituting this bound in Eq. (D.4) and using a similar bound for $H(\boldsymbol{V}^n|\boldsymbol{X}^{m,n}, \pi_{\boldsymbol{V},\boldsymbol{X}})$.

D.2. Proof of Lemma 5.3

We begin with a technical fact.

Lemma D.1. *Let $\boldsymbol{\xi} = (\xi_{ij})_{i\le m, j\le n}$, $\boldsymbol{\sigma} = (\sigma_i)_{i\le m}$, $\boldsymbol{\tau} = (\tau_j)_{j\le n}$ be collections of mutually independent random variables taking values in a measurable space $\mathcal{Z}$. $x : \mathcal{Z}^3 \to \mathcal{Z}$, $F : \mathcal{Z}^{m\times n} \to \mathbb{R}$. Define $\boldsymbol{x}(\boldsymbol{\xi}, \boldsymbol{\sigma}, \boldsymbol{\tau}) \in \mathbb{R}^{m\times n}$ via $\boldsymbol{x}(\boldsymbol{\xi}, \boldsymbol{\sigma}, \boldsymbol{\tau})_{ij} = x(\xi_{ij}, \sigma_i, \tau_j)$.*

Given a vector of independent random variables $\boldsymbol{z}$, we let $\mathrm{Var}_{z_i}(f(\boldsymbol{z})) := \mathbb{E}_{z_i}[(f(\boldsymbol{z}) - \mathbb{E}_{z_i} f(\boldsymbol{z}))^2]$. Define the quantities

$$B_* := \max_{\substack{\boldsymbol{x},\boldsymbol{x}'\in\mathcal{Z}^{m\times n}\\ d(\boldsymbol{x},\boldsymbol{x}')\le 1}} \left|F(\boldsymbol{x}) - F(\boldsymbol{x}'')\right|, \tag{D.5}$$

$$B_1 := \max_{\substack{\boldsymbol{\sigma},\boldsymbol{\sigma}'\in\mathcal{Z}^{m}\\ d(\boldsymbol{\sigma},\boldsymbol{\sigma}')\le 1}} \max_{\boldsymbol{\tau}\in\mathcal{Z}^n} \left|\mathbb{E}_{\boldsymbol{\xi}} F(\boldsymbol{x}(\boldsymbol{\xi},\boldsymbol{\sigma},\boldsymbol{\tau})) - \mathbb{E}_{\boldsymbol{\xi}} F(\boldsymbol{x}(\boldsymbol{\xi},\boldsymbol{\sigma}',\boldsymbol{\tau}))\right|, \tag{D.6}$$

$$B_2 := \max_{\boldsymbol{\sigma}\in\mathcal{Z}^{m}} \max_{\substack{\boldsymbol{\tau},\boldsymbol{\tau}'\in\mathcal{Z}^{n}\\ d(\boldsymbol{\tau},\boldsymbol{\tau}')\le 1}} \left|\mathbb{E}_{\boldsymbol{\xi}} F(\boldsymbol{x}(\boldsymbol{\xi},\boldsymbol{\sigma},\boldsymbol{\tau})) - \mathbb{E}_{\boldsymbol{\xi}} F(\boldsymbol{x}(\boldsymbol{\xi},\boldsymbol{\sigma},\boldsymbol{\tau}'))\right|, \tag{D.7}$$

$$V_* := \sup_{\boldsymbol{\xi},\boldsymbol{\tau},\boldsymbol{\sigma}} \sum_{i\le m, j\le n} \mathrm{Var}_{\xi_{ij}}\left\{F(\boldsymbol{x}(\boldsymbol{\xi},\boldsymbol{\sigma},\boldsymbol{\tau}))\right\}, \tag{D.8}$$

$$V_1 := \sup_{\boldsymbol{\tau},\boldsymbol{\sigma}} \sum_{i\le m} \mathrm{Var}_{\sigma_i}\left\{\mathbb{E}_{\boldsymbol{\xi}} F(\boldsymbol{x}(\boldsymbol{\xi},\boldsymbol{\sigma},\boldsymbol{\tau})\right\}, \tag{D.9}$$

$$V_2 := \sup_{\boldsymbol{\tau},\boldsymbol{\sigma}} \sum_{j\le n} \mathrm{Var}_{\sigma_j}\left\{\mathbb{E}_{\boldsymbol{\xi}} F(\boldsymbol{x}(\boldsymbol{\xi},\boldsymbol{\sigma},\boldsymbol{\tau}))\right\}. \tag{D.10}$$

Then, for any $t \ge 0$, the following holds with probability at least $1 - 8e^{-t}$:

$$\left|F(\boldsymbol{x}(\boldsymbol{\xi},\boldsymbol{\sigma},\boldsymbol{\tau})) - \mathbb{E}F(\boldsymbol{x}(\boldsymbol{\xi},\boldsymbol{\sigma},\boldsymbol{\tau}))\right\| \le 2\max(\sqrt{2V_* t} + \sqrt{2V_1 t} + \sqrt{2V_2 t}; (B_* + B_1 + B_2)t). \tag{D.11}$$

Proof. Let $\boldsymbol{z} \in \mathcal{Z}^N$ be a vector of independent random variables and $f : \mathcal{Z}^N \to \mathbb{R}$. Define the martingale $X_k := \mathbb{E}[f(\boldsymbol{z})|\mathcal{F}_k]$ (where $\mathcal{F}_k := \sigma(z_1, \dots, z_k)$). Then we have

$$\operatorname{ess\,sup} |X_k - X_{k-1}| \le B_0 := \sup_{d(\boldsymbol{z},\boldsymbol{z}')\le 1} |f(\boldsymbol{z}) - f(\boldsymbol{z}')|, \tag{D.12}$$

$$\sum_{k=1}^N \mathbb{E}[(X_k - X_{k-1})^2|\mathcal{F}_{k-1}] = \sum_{k=1}^N \mathbb{E}\left[(\mathbb{E}[f|\boldsymbol{z}_{<k}, z_k] - \mathbb{E}_{z_k'}\mathbb{E}[f|\boldsymbol{z}_{<k}, z_k'])^2\middle|\boldsymbol{z}_{<k}\right] \tag{D.13}$$

$$\le V_0 := \sup_{\boldsymbol{z}\in\mathcal{Z}^N} \sum_{k=1}^N \mathrm{Var}_{z_k}\big(f(\boldsymbol{z})\big). \tag{D.14}$$

By Freedman's inequality, with probability at least $1 - 2e^{-t}$, we have

$$\left|f(\boldsymbol{z}) - \mathbb{E}f(\boldsymbol{z})\right| \le \max\left(\sqrt{2V_0 t} : \frac{2B_0 t}{3}\right). \tag{D.15}$$

Define $E(\boldsymbol{\sigma},\boldsymbol{\tau}) := \mathbb{E}_{\boldsymbol{\xi}} F(\boldsymbol{x}(\boldsymbol{\xi},\boldsymbol{\sigma},\boldsymbol{\tau}))$, $L(\boldsymbol{\tau}) := \mathbb{E}_{\boldsymbol{\sigma},bxi} F(\boldsymbol{x}(\boldsymbol{\xi},\boldsymbol{\sigma},\boldsymbol{\tau}))$. Applying the above inequality, each of the following holds with probability at least $1 - 2e^{-t}$

$$|F(\boldsymbol{x}(\boldsymbol{\xi},\boldsymbol{\sigma},\boldsymbol{\tau})) - E(\boldsymbol{\sigma},\boldsymbol{\tau})| \le \max\left(\sqrt{2V_* t} : \frac{2B_* t}{3}\right), \tag{D.16}$$

$$|E(\boldsymbol{\sigma},\boldsymbol{\tau}) - L(\boldsymbol{\tau})| \le \max\left(\sqrt{2V_1 t} : \frac{2B_1 t}{3}\right), \tag{D.17}$$

$$|L(\boldsymbol{\tau}) - \mathbb{E}F(\boldsymbol{x})| \le \max\left(\sqrt{2V_2 t} : \frac{2B_2 t}{3}\right), \tag{D.18}$$

and the claim follows by union bound. □

We next state and prove a more stronger version of Lemma 5.3.

Lemma D.2. *For $\boldsymbol{X} \in \mathcal{X}^{m\times n}$, let $\mathrm{P}(\boldsymbol{X}) = \mathrm{P}_{Q,q_{\mathrm{r}},q_{\mathrm{c}};m,n}(\boldsymbol{X})$ the probability of table $\boldsymbol{X}$ under the model $\mathcal{T}(Q, q_{\mathrm{r}}, q_{\mathrm{c}}; m, n)$, i.e.*

$$\mathrm{P}(\boldsymbol{X}) = \sum_{\boldsymbol{u}\in\mathcal{L}^m}\sum_{\boldsymbol{v}\in\mathcal{L}^n}\prod_{(i,j)\in[m]\times[n]} Q(X_{ij}|u_i,v_i)\prod_{i\in[m]} q_{\mathrm{r}}(u_i)\prod_{j\in[n]} q_{\mathrm{c}}(v_j)\,. \tag{D.19}$$

Define the following quantities:

$$M_* := \max_{x,x'\in\mathcal{X}}\max_{u,v\in\mathcal{L}}\Big|\log\frac{Q(x|u,v)}{Q(x'|u,v)}\Big|, \tag{D.20}$$

$$M_1 := \max_{\tau,\sigma,\sigma'}\|Q(\,\cdot\,|\sigma,\tau)-Q(\,\cdot\,|\sigma',\tau)\|_{\mathrm{TV}}\max_{u,v,x,x''}\Big|\log\frac{Q(x|u,v)}{Q(x'|u,v)}\Big|, \tag{D.21}$$

$$M_2 := \max_{\tau,\tau',\sigma}\|Q(\,\cdot\,|\sigma,\tau)-Q(\,\cdot\,|\sigma,\tau'')\|_{\mathrm{TV}}\max_{u,v,x,x''}\Big|\log\frac{Q(x|u,v)}{Q(x'|u,v)}\Big|, \tag{D.22}$$

$$s_* := \frac{1}{2}\max_{u_0,v_0\in\mathcal{L}}\sum_{x,x'\in\mathcal{X}}Q(x|u_0,v_0)Q(x'|u_0,v_0)\max_{u,v\in\mathcal{L}}\Big(\log\frac{Q(x|u,v)}{Q(x'|u,v)}\Big)^2, \tag{D.23}$$

$$s_1 := \frac{1}{2}\max_{u_0,u_0',v_0\in\mathcal{L}}\|Q(\,\cdot\,|u_0,v_0)-Q(\,\cdot\,|u_0',v_0)\|_{\mathrm{TV}}\max_{x,x'\in\mathcal{L}}\max_{u,v\in\mathcal{L}}\Big(\log\frac{Q(x|u,v)}{Q(x'|u,v)}\Big)^2, \tag{D.24}$$

$$s_2 := \frac{1}{2}\max_{u_0,v_0,v_0'\in\mathcal{L}}\|Q(\,\cdot\,|u_0,v_0)-Q(\,\cdot\,|u_0,v_0')\|_{\mathrm{TV}}\max_{x,x'\in\mathcal{L}}\max_{u,v\in\mathcal{L}}\Big(\log\frac{Q(x|u,v)}{Q(x'|u,v)}\Big)^2. \tag{D.25}$$

Then, for $\boldsymbol{X}\sim\mathcal{T}(Q,q_{\mathrm{r}},q_{\mathrm{c}};m,n)$ and any $t\geq 0$ the following bound holds with probability at least $-2\,e^{-t}$:

$$\big|-\log\mathrm{P}(\boldsymbol{X})-H(\boldsymbol{X})\big| \leq 3\max\Big(\sqrt{s_*mnt}+\sqrt{s_1mn^2t}+\sqrt{s_2m^2nt}, M_*+M_1n+M_2m\Big)\,. \tag{D.26}$$

Proof. Let $\boldsymbol{\sigma}=(\sigma_i)_{i\leq m}\sim_{iid} r$, $\boldsymbol{\tau}=(\tau_i)_{i\leq n}\sim_{iid} c$, $\boldsymbol{\xi}=(\xi_{ij})_{i\leq m,j\leq n}\sim_{iid}\mathsf{Unif}([0,1])$, and $x:[0,1]\times\mathcal{L}\times\mathcal{L}\to\mathcal{X}$ be such that $x(\xi_{ij},\sigma_i,\tau_j)|_{\sigma_i,\tau_j}\sim Q(\,\cdot\,|\sigma_i,\tau_j)$. We define $F(\boldsymbol{x})=-\log\mathrm{P}(\boldsymbol{x})$, and will apply Lemma D.1 to this function. Using the notation from that lemma, we claim that $B_*\leq M_*$, $B_1\leq M_1n$, $B_2\leq M_2m$, and $V_*\leq mns_*$, $V_1\leq mn^2s_1$, $V_2\leq m^2ns_2$.

Note that, if (x_{ij}), (x'_{ij}) differ only for entry i,j, then

$$F(\boldsymbol{x})-F(\boldsymbol{x}') = -\log\mathrm{E}_{\boldsymbol{u},\boldsymbol{v}|\boldsymbol{x}}\Big\{\frac{Q(x'_{ij}|u_i,v_j)}{Q(x_{ij}|u_i,v_j)}\Big\}, \tag{D.27}$$

where $\mathrm{E}_{\boldsymbol{u},\boldsymbol{v}|\boldsymbol{x}}$ denotes expectation with respect to the posterior measure $\mathrm{P}(\boldsymbol{u},\boldsymbol{v}|\boldsymbol{X}=\boldsymbol{x})$. This immediately implies $B_*\leq M_*$.

Next consider the constant B_1 defined in Eq. (D.6). Using the exchangeability of the $(\xi_{i,\cdot},\sigma_i)$, we get

$$\begin{aligned}
B_1 &= \max_{\boldsymbol{\tau}}\big|\mathbb{E}_{\boldsymbol{\xi}}\log\mathrm{E}_{\boldsymbol{u},\boldsymbol{v}|\boldsymbol{x}}\Big\{\prod_{j=1}^n\frac{Q(x(\xi_{1,j},\sigma_1',\tau_j)|u_1,v_j)}{Q(x(\xi_{1,j},\sigma_1,\tau_j)|u_1,v_j)}\Big\}\big|\\
&\leq \max_{\boldsymbol{\tau}}\mathbb{E}_{\boldsymbol{\xi}}\max_{\boldsymbol{u},\boldsymbol{v}}\Big|\log\Big\{\prod_{j=1}^n\frac{Q(x(\xi_{1,j},\sigma_1',\tau_j)|u_1,v_j)}{Q(x(\xi_{1,j},\sigma_1,\tau_j)|u_1,v_j)}\Big\}\Big|\\
&\leq \max_{\boldsymbol{\tau}}\sum_{j=1}^n\mathbb{E}_{\boldsymbol{\xi}}\max_{\boldsymbol{u},\boldsymbol{v}}\Big|\log\Big\{\frac{Q(x(\xi_{1,j},\sigma_1',\tau_j)|u_1,v_j)}{Q(x(\xi_{1,j},\sigma_1,\tau_j)|u_1,v_j)}\Big\}\Big|\\
&\leq n\max_{\tau,\sigma,\sigma'}\mathbb{E}_{\xi}\max_{u,v}\Big|\log\frac{Q(x(\xi,\sigma_1',\tau_j)|u,v)}{Q(x(\xi,\sigma_1,\tau_j)|u,v)}\Big|\\
&\leq n\max_{\tau,\sigma,\sigma'}\|Q(\,\cdot\,|\sigma,\tau)-Q(\,\cdot\,|\sigma',\tau)\|_{\mathrm{TV}}\max_{u,v,x,x'}\Big|\log\frac{Q(x|u,v)}{Q(x'|u,v)}\Big| = M_1\,.
\end{aligned}$$

The bound $B_2\leq M_2m$ is proved analogously.

Consider now the quantity V_* of Eq. (D.8). Denote by $\boldsymbol{\xi}_{(ij)}(t)$ the array obtained by replacing entry (i,j) in $\boldsymbol{\xi}$ by t, and by $\boldsymbol{x}(t) = \boldsymbol{x}(\boldsymbol{\xi}_{(ij)}(t), \boldsymbol{\sigma}, \boldsymbol{\tau})$. Then we have

$$\begin{aligned}
\mathrm{Var}_{\xi_{ij}}(F(\boldsymbol{x})) &= \frac{1}{2}\mathbb{E}_{\xi',\xi''}\Big\{\big(F(\boldsymbol{x}(\boldsymbol{\xi}_{(ij)}(\xi'),\boldsymbol{\sigma},\boldsymbol{\tau})) - F(\boldsymbol{x}(\boldsymbol{\xi}_{(ij)}(\xi''),\boldsymbol{\sigma},\boldsymbol{\tau}))\big)^2\Big\} \\
&= \frac{1}{2}\mathbb{E}_{\xi',\xi''}\Big\{\Big(\log \mathrm{E}_{\boldsymbol{u},\boldsymbol{v}|\boldsymbol{x}(\xi')}\Big\{\frac{Q(x(\xi'',\sigma_i,\tau_j)|u_i,v_j)}{Q(x(\xi',\sigma_i,\tau_j)|u_i,v_j)}\Big\}\Big)^2\Big\} \\
&\le \frac{1}{2}\mathbb{E}_{\xi',\xi''}\max_{u,v}\Big(\log\Big\{\frac{Q(x(\xi'',\sigma_i,\tau_j)|u,v)}{Q(x(\xi',\sigma_i,\tau_j)|u,v)}\Big\}\Big)^2 \\
&= \frac{1}{2}\sum_{x,x'} Q(x|\sigma,\tau)Q(x'|\sigma,\tau)\max_{u,v}\Big(\log\Big\{\frac{Q(x|u,v)}{Q(x'|u,v)}\Big\}\Big)^2 .
\end{aligned}$$

We then have, as claimed

$$\begin{aligned}
V_* &\le \max_{\boldsymbol{\xi},\boldsymbol{\sigma},\boldsymbol{\tau}} \sum_{i\le m, j\le n} \mathrm{Var}_{\xi_{ij}}\big\{F(\boldsymbol{x})\big\} \\
&\le mn \max_{\boldsymbol{\xi},\boldsymbol{\sigma},\boldsymbol{\tau}} \mathrm{Var}_{\xi_{ij}}\big\{F(\boldsymbol{x})\big\} \le mns_* .
\end{aligned}$$

Finally consider the quantity V_1 of Eq. (D.9) (the argument is similar for V_2). Denote by $\boldsymbol{\sigma}_{(i)}(t)$ the vector obtained by replacing entry i in $\boldsymbol{\sigma}$ by t. Proceeding as above, we have

$$\begin{aligned}
\mathrm{Var}_{\sigma_i}(\mathbb{E}_{\boldsymbol{\xi}}F(\boldsymbol{x})) &= \frac{1}{2}\mathbb{E}_{\sigma',\sigma''}\Big\{\big(\mathbb{E}_{\boldsymbol{\xi}}F(\boldsymbol{x}(\boldsymbol{\xi},\boldsymbol{\sigma}_{(i)}(\sigma'),\boldsymbol{\tau})) - \mathbb{E}_{\boldsymbol{\xi}}F(\boldsymbol{x}(\boldsymbol{\xi},\boldsymbol{\sigma}_{(i)},\boldsymbol{\tau}))\big)^2\Big\} \\
&= \frac{1}{2}\mathbb{E}_{\sigma',\sigma''}\Big\{\Big(\mathbb{E}_{\boldsymbol{\xi}}\log \mathrm{E}_{\boldsymbol{u},\boldsymbol{v}|\boldsymbol{x}(\sigma')}\Big\{\prod_{j=1}^n \frac{Q(x(\xi_{ij},\sigma'',\tau_j)|u_i,v_j)}{Q(x(\xi_{ij},\sigma',\tau_j)|u_i,v_j)}\Big\}\Big)^2\Big\} \\
&\le \frac{1}{2}\mathbb{E}_{\sigma',\sigma''}\Big\{\Big(\mathbb{E}_{\boldsymbol{\xi}}\log\Big\{\prod_{j=1}^n \max_{u,v}\frac{Q(x(\xi_{ij},\sigma'',\tau_j)|u,v)}{Q(x(\xi_{ij},\sigma',\tau_j)|u,v)}\Big\}\Big)^2\Big\} \\
&\le \frac{1}{2}\mathbb{E}_{\sigma',\sigma''}\Big\{\Big(\sum_{j=1}^n E_\xi \log\Big\{\max_{u,v}\frac{Q(x(\xi,\sigma'',\tau_j)|u,v)}{Q(x(\xi,\sigma',\tau_j)|u,v)}\Big\}\Big)^2\Big\} \\
&\le \frac{n^2}{2}\max_{\tau}\mathbb{E}_{\sigma',\sigma''}\Big\{\Big(\mathbb{E}_\xi \log\Big\{\max_{u,v}\frac{Q(x(\xi,\sigma'',\tau)|u,v)}{Q(x(\xi,\sigma',\tau)|u,v)}\Big\}\Big)^2\Big\} \\
&\le \frac{n^2}{2}\max_{\tau,\sigma,\sigma'}\|Q(\,\cdot\,|\sigma,\tau) - Q(\,\cdot\,|\sigma',\tau)\|_{\mathrm{TV}}\max_{x,x'}\Big(\mathbb{E}_\xi\log\Big\{\max_{u,v}\frac{Q(x'|u,v)}{Q(x|u,v)}\Big\}\Big)^2 = n^2 s_1 .
\end{aligned}$$

Therefore

$$V_1 = \max_{\boldsymbol{\sigma},\boldsymbol{\tau}}\sum_{i=1}^m \mathrm{Var}_{\sigma_i}(\mathbb{E}_{\boldsymbol{\xi}}F(\boldsymbol{x})) \le mn^2 s_1 .$$

This finishes the proof. □

E. Proofs for finite state encoders

Recall from Section 5.2 that a finite-state encoder is defined by a triple (Σ, f, g). Formally we can define the action of f, g on $\boldsymbol{X}^n \in \mathcal{X}^n$ recursively via (recall that $\oplus$ denotes concatenation)

$$f_{m+1}(\boldsymbol{X}^{m+1}, s_0) = f_m(\boldsymbol{X}^m, s_0) \oplus f(X_{m+1}, g(\boldsymbol{X}^m, s_0)) , \tag{E.1}$$

$$g_{m+1}(\boldsymbol{X}^{m+1}, s_0) = g(X_{m+1}, g(\boldsymbol{X}^m, s_0)) , \tag{E.2}$$

and the encoder is thus given by $E(\boldsymbol{X}^n) = f_n(\boldsymbol{X}^n, s_{\text{init}})$.

We say that the state space Σ is non-degenerate if, for each $s_1 \in \Sigma$ there exists m, $\boldsymbol{X}^m \in \mathcal{X}^m$ such that $g_m(\boldsymbol{X}^m, s_{\text{init}}) = s_1$. Notice that if state space is degenerate, we could always remove one or more symbols from Σ without changing the encoder, and making the state-space non-degenerate. For this reason, we will hereafter assume non-degeneracy without mentioning it.

We say that the FS encoder is information lossless (IL) if for any $n \in \mathbb{N}$, $\boldsymbol{X}^n \mapsto f_n(\boldsymbol{X}^n, s_{\text{init}})$ is injective.

Remark E.1. An information-lossless encoder satisfies a stronger condition: for any $m \in \mathbb{N}$ and any $s_* \in \Sigma$, the map $\boldsymbol{X}^m \mapsto f_m(\boldsymbol{X}^m, s_*)$ is injective.

Indeed, assume this were not the case. Then there would exist two distinct inputs $\boldsymbol{X}^m$, $\tilde{\boldsymbol{X}}_1^m \in \mathcal{X}^m$ and a state $s_* \in \Sigma$ such that $f_m(\boldsymbol{X}^m, s_*) = f_m(\tilde{\boldsymbol{X}}^m, s_*)$. By non-degeneracy, there exists $a_1^\ell \in \mathcal{X}^\ell$ such that $s_* = g_\ell(a_1^\ell, s_{\text{init}})$, Defining $n = \ell + m$, $\boldsymbol{Y}^n = a_1^\ell \oplus \boldsymbol{X}^m$, $\tilde{\boldsymbol{Y}}^n = a_1^\ell \oplus \tilde{\boldsymbol{X}}^m$, it is not hard to check that these inputs are distinct but $f_n(\boldsymbol{Y}^n, s_{\text{init}}) = f_n(\tilde{\boldsymbol{Y}}_1^n, s_{\text{init}})$.

Proposition E.1. *Define the compression rate on input x_1^n as $\mathsf{R}(\boldsymbol{X}^n) = \mathsf{len}(f_n(\boldsymbol{X}^n, s_{\text{init}}))/(n \log_2 |\mathcal{X}|)$. Then for any $\ell \geq 1$, the following holds (where $n' := n - 2\ell$ and we recall that $M := |\Sigma|$):*

$$\mathsf{R}(\boldsymbol{X}^n) \geq \frac{n - 2\ell}{n\ell \log_2 |\mathcal{X}|} H(\hat{p}^\ell_{\boldsymbol{X}_1^{n'}}) - \frac{1}{\ell \log_2 |\mathcal{X}|} \left(\log_2(|\Sigma|\ell) + \log_2 \log_2 |\mathcal{X}| \right). \tag{E.3}$$

Proof. We will denote by $L(\boldsymbol{X}^m; s_*)$ the length of the encoding of $\boldsymbol{X}^m$ when starting in state s_*:

$$L(\boldsymbol{X}^m; s_*) := \mathsf{len}(f_n(\boldsymbol{X}^m, s_*)). \tag{E.4}$$

We then have, for any $b \in \{0, \dots, \ell - 1\}$, and setting by convention $s_0 = s_{\text{init}}$, we get

$$\mathsf{R}(\boldsymbol{X}^n) \geq \frac{1}{n \log_2 |\mathcal{X}|} \sum_{k=0}^{\lfloor n/\ell \rfloor - 2} L(\boldsymbol{X}_{k\ell+b+1}^{(k+1)\ell+b}; s_{k\ell+b}). \tag{E.5}$$

By averaging over b, and introducing the shorthand $n' := n - 2\ell$, we get

$$\begin{aligned}
\mathsf{R}(\boldsymbol{X}^n) &\geq \frac{1}{n\ell \log_2 |\mathcal{X}|} \sum_{m=1}^{(\lfloor n/\ell \rfloor - 1)\ell} L(\boldsymbol{X}_m^{m+\ell-1}; s_{m-1}) && \text{(E.6)} \\
&\geq \frac{n - 2\ell}{n\ell \log_2 |\mathcal{X}|} \sum_{s \in \Sigma} \sum_{u_1^\ell \in \mathcal{X}^\ell} \hat{p}^\ell_{\boldsymbol{X}_1^{n'}}(u_1^\ell, s)\, L(u_1^\ell; s) && \text{(E.7)} \\
&\overset{(a)}{\geq} \frac{n - 2\ell}{n\ell \log_2 |\mathcal{X}|} \sum_{s \in \Sigma} \left\{ \hat{p}^\ell_{\boldsymbol{X}_1^{n'}}(s)\, H(\hat{p}^\ell_{\boldsymbol{X}_1^{n'}}(\,\cdot\,|s)) - \log_2 \log_2(|\mathcal{X}|^\ell) \right\}, && \text{(E.8)}
\end{aligned}$$

where (a) holds by Lemma C.1. By the chain rule of entropy (recalling that $M := |\Sigma|$), we have:

$$\begin{aligned}
\sum_{s \in \Sigma} \hat{p}^\ell_{\boldsymbol{X}_1^{n'}}(s)\, H(\hat{p}^\ell_{\boldsymbol{X}_1^{n'}}(\,\cdot\,|s)) &= H(\boldsymbol{X}_1^\ell | S) = H(\boldsymbol{X}_1^\ell) + H(S|\boldsymbol{X}_1^\ell) - H(S) \\
&\geq H(\boldsymbol{X}_1^\ell) - \log_2 M = H(\hat{p}^\ell_{\boldsymbol{X}_1^{n'}}) - \log_2 M.
\end{aligned}$$

The claim (E.3) follows by using the last inequality in Eq. (E.8). □

Theorem E.2. *Let $\boldsymbol{X}^{m,n} \sim \mathcal{T}(Q, q_{\mathsf{r}}, q_{\mathsf{c}}; m, n)$ and (Σ, f, g) be an information lossless finite state encoder. With an abuse of notation, denote $f_{mn}(\boldsymbol{X}^{m\times n}, s_{\text{init}}) \in \{0,1\}^*$ the binary sequence obtained by applying the finite state encoder to the vector $\mathsf{vec}(\boldsymbol{X}^{m\times n}) \in \mathcal{X}^{mn}$ obtained by scanning $\boldsymbol{X}^{m\times n}$ in row-first order. Define the compression rate by*

$$\mathsf{R}(\boldsymbol{X}^{m,n}) := \frac{\mathsf{len}(f_{mn}(\boldsymbol{X}^{m\times n}, s_{\text{init}}))}{mn \log_2 |\mathcal{X}|}. \tag{E.9}$$

Assuming $m > 10$, $|\Sigma| \geq |\mathcal{X}|$, and $\log_2 |\Sigma| \leq n \log_2 |\mathcal{X}|/9$, the expected compression rate is lower bounded as follows

$$\mathbb{E}\, \mathsf{R}(\boldsymbol{X}^{m,n}) \geq \frac{H(X|U)}{\log_2 |\mathcal{X}|} - 10 \sqrt{\frac{\log |\Sigma|}{n \log |\mathcal{X}|}} \cdot \log(n \log |\Sigma|). \tag{E.10}$$

Proof. We let $N := mn$, $N' = mn - 2\ell$ where we $\ell \le n/3$ will be selected later. We write $\boldsymbol{X}^N := \mathsf{vec}(\boldsymbol{X}^{m,n})$ for the vectorization $\boldsymbol{X}^{m,n}$, $\boldsymbol{X}^{N'}$ for the vector comprising its first N' entries. Recall the definition of empirical distribution. For any fixed $\boldsymbol{w} \in \mathcal{X}^\ell$

$$\hat{p}^\ell_{\boldsymbol{X}^{N'}}(\boldsymbol{w}) := \frac{1}{N' - \ell + 1} \sum_{i=1}^{N'-\ell+1} \mathbf{1}_{\boldsymbol{X}_i^{i+\ell-1} = \boldsymbol{w}} \,.$$

Let $S := \{i \in [N' - \ell + 1] : [i, i + \ell - 2] \cap n\mathbb{N} = \emptyset\}$. In words, these are the subset of blocks of length ℓ that do not cross the end of a line in the table. Since for each line break there are at most $\ell - 1$ such blocks, we have $|S| \ge N' - \ell + 1 - (m-1)(\ell-1)$. We will consider the following modified empirical distribution

$$\overline{p}^\ell_{\boldsymbol{X}^{N'}}(\boldsymbol{w}) := \frac{1}{|S|} \sum_{i \in S} \mathbf{1}_{\boldsymbol{X}_i^{i+\ell-1} = \boldsymbol{w}} \,.$$

Then by construction

$$\begin{aligned} \hat{p}^\ell_{\boldsymbol{X}^{N'}}(\boldsymbol{w}) &= (1 - \eta_\ell)\overline{p}^\ell_{\boldsymbol{X}^{N'}}(\boldsymbol{w}) + \eta_\ell q^\ell_{\boldsymbol{X}^{N'}}(\boldsymbol{w}) \,, \\ \eta_\ell &:= 1 - \frac{|S|}{N' - \ell + 1} = \frac{(m-1)(\ell-1)}{N' - \ell + 1} \,, \end{aligned}$$

where $q^\ell_{\boldsymbol{X}^{N'}}$ is the empirical distribution of blocks that do cross the line. By concavity of the entropy, we have

$$H(\hat{p}^\ell_{\boldsymbol{X}^{N'}}) \ge (1 - \eta_\ell) H(\overline{p}^\ell_{\boldsymbol{X}^{N'}}) + \eta_\ell H(q^\ell_{\boldsymbol{X}^{N'}}) \ge (1 - \eta_\ell) H(\overline{p}^\ell_{\boldsymbol{X}^{N'}}) \,. \tag{E.11}$$

Further, since $\ell \le n/3$,

$$\begin{aligned} \eta_\ell &= \frac{(m-1)(\ell-1)}{mn - 3\ell + 1} \\ &\le \frac{(m-1)\ell}{mn - 3\ell} \le \frac{(m-1)\ell}{(m-1)n} \le \frac{\ell}{n} \,. \end{aligned} \tag{E.12}$$

Now let the row latents $\boldsymbol{u} := (u_i)_{i \le m}$ be fixed, and denote by $\hat{r}^S_{\boldsymbol{u}}$ their weighted empirical distribution, defined as follows:

$$\hat{r}^S_{\boldsymbol{u}}(s) := \sum_{i=1}^m \frac{|S \cap [(i-1)n + 1, in]|}{|S|} \mathbf{1}_{u_i = s} \,.$$

In words, $\hat{r}^S_{\boldsymbol{u}}$ is the empirical distribution of the latents $(u_i)_{i \le m}$ where row i is weighted by its contribution to S. Note that all the weights are equal to $(n - 2(\ell-1))/|S|$ except, potentially, for the last one.

We have

$$p^\ell_*(\boldsymbol{w}) := \mathbb{E}[\overline{p}^\ell_{\boldsymbol{X}^{N'}}(\boldsymbol{w})] = \sum_{u \in \mathcal{L}} \hat{r}^S_{\boldsymbol{u}}(u) \prod_{i=1}^\ell Q_{x|u}(w_i|u) \,, \quad Q_{x|u}(w|u) := \sum_{v \in \mathcal{L}} Q(w|u,v)\, q_{\mathsf{c}}(v) \,.$$

Using Eq. (E.11), (E.12) and concavity of the entropy, we get

$$\mathbb{E}\big[H(\hat{p}^\ell_{\boldsymbol{X}^{N'}})|\boldsymbol{u}\big] \ge \Big(1 - \frac{\ell}{n}\Big) H(p^\ell_*) \,. \tag{E.13}$$

By Proposition E.1, we get

$$\begin{aligned} \mathbb{E}\left[\mathsf{R}(\boldsymbol{X}^{m,n})|\boldsymbol{u}\right] \ge& \frac{mn - 2\ell}{mn\ell \log_2 |\mathcal{X}|} \Big(1 - \frac{\ell}{n}\Big) H(p^\ell_*) - \frac{1}{\ell \log_2 |\mathcal{X}|} \big(\log_2(|\Sigma|\ell) + \log_2 \log_2 |\mathcal{X}|\big) \\ \ge& \frac{1}{\ell \log_2 |\mathcal{X}|} H(p^\ell_*) - \frac{2\ell}{n} - \frac{1}{\ell \log_2 |\mathcal{X}|} \big(\log_2(|\Sigma|\ell) + \log_2 \log_2 |\mathcal{X}|\big) \,, \end{aligned}$$

Algorithm 3 Lempel-Ziv

Input: Data $\boldsymbol{X}^N \in \mathcal{X}^N = \{0, \cdots, |\mathcal{X}|-1\}^N$
Output: Binary string $\boldsymbol{Z} \in \{0,1\}^*$
for $k = 1$ **to** N **do**
  if $\exists j < k : X_j = X_k$ **then**
    $L_k \leftarrow \max\{\ell \geq 1 : \ \exists j \in \{1, \dots, k-1\} \text{ s.t. } \boldsymbol{X}_j^{j+\ell-1} = \boldsymbol{X}_k^{k+\ell-1}\}$
    $T_k \leftarrow \max\{j \in \{1, \dots, k-1\} \text{ s.t. } \boldsymbol{X}_j^{j+L_k-1} = \boldsymbol{X}_k^{k+L_k-1}\}$
  else
    $L_k \leftarrow 1$
    $T_k \leftarrow (-X_k)$
  end if
  $\boldsymbol{Z} \leftarrow \boldsymbol{Z} \oplus \mathsf{plain}(T_k) \oplus \mathsf{elias}(L_k)$
  $k \leftarrow k + L_k$
  if $\mathsf{len}(\boldsymbol{Z}) \leq \mathsf{len}(\mathsf{plain}(\boldsymbol{X}^N))$ **then**
    return $\boldsymbol{Z}$
  else
    return $\mathsf{plain}(\boldsymbol{X}^N)$
  end if
end for

where in the last inequality we used the fact that $H(p_*^\ell) \leq \ell \log_2 |\mathcal{X}|$. We choose

$$\ell = \sqrt{\frac{n \log_2 |\Sigma|}{\log_2 |\mathcal{X}|}} \leq \frac{n}{3}\,, \tag{E.14}$$

Substituting and simplifying, we get

$$\mathbb{E}\big[\mathsf{R}(\boldsymbol{X}^{m,n})|\boldsymbol{u}\big] \geq \frac{H(p_*^\ell)}{\ell \log_2 |\mathcal{X}|} - \frac{10}{\sqrt{n}} \cdot \sqrt{\frac{\log |\Sigma|}{\log |\mathcal{X}|}} \cdot \log(n \log |\Sigma|)\,. \tag{E.15}$$

Finally, letting $(W_1, \dots, W_\ell, U) \in \mathcal{X}^\ell \times \mathcal{L}$ be random variables with joint distribution $\hat{r}^S_{\boldsymbol{u}}(u) \prod_{i=1}^\ell Q_{x|u}(w_i|u)$. Then

$$H(p_*^\ell) \geq \sum_{u \in \mathcal{L}} \hat{r}^S_{\boldsymbol{u}}(u) H\big(Q_{x|u}^{\otimes \ell}(\,\cdot\,|u)\big) \tag{E.16}$$

$$\geq \ell \sum_{u \in \mathcal{L}} \hat{r}^S_{\boldsymbol{u}}(u) H(X|U=u)\,, \tag{E.17}$$

and therefore $\mathbb{E}H(p_*^\ell) \geq H(X|U)$, finishing the proof. □

F. Proofs for Lempel-Ziv coding

The pseudocode of the Lempel-Ziv algorithm that we will analyze is given here. For ease of presentation, we identify $\mathcal{X}$ with a set of integers.

Note that if a simbol X_k never appeared in the past, we point to $T_k = -X_k$ and set $L_k = 1$. This is essentially equivalent to prepending a sequence of distinct $|\mathcal{X}|$ symbols to $\boldsymbol{X}^N$.

It is useful to define for each $k \leq N$,

$$L_k(\boldsymbol{X}^N) := \max\big\{\ell \geq 1 : \ \exists j \in \{1, \dots, k-1\} \text{ s.t. } \boldsymbol{X}_j^{j+\ell-1} = \boldsymbol{X}_k^{k+\ell-1}\big\}\,, \tag{F.1}$$

$$T_k(\boldsymbol{X}^N) := \min\big\{j \in \{1, \dots, k-1\} \text{ s.t. } \boldsymbol{X}_j^{j+L_k-1} = \boldsymbol{X}_k^{k+L_k-1}\big\}\,. \tag{F.2}$$

F.1. Proof of Theorem 5.6

Lemma F.1. *Under Assumption 5.5, there exists a constant C such that the following holds with probability at least $1 - N^{-10}$:*

$$\max_{k \le N} L_k(\boldsymbol{X}^N) \le C \log N \,. \tag{F.3}$$

Proof. We begin by considering a slightly different setting, and will then show that our question reduces to this setting. Let $(Z_i)_{i \ge 1}$ be independent random variables with $Z_i \sim q_i$ a probability distribution over $\mathcal{X}$. Further assume $\max_{x \in \mathcal{X}} q_i(x) \le 1 - c$ for all $i \ge 1$. Then we claim that, for any $t, \ell \ge 1$, we have

$$\mathbb{P}\big(Z_1^\ell = Z_{t+1}^{t+\ell}\big) \le (1-c)^\ell \,. \tag{F.4}$$

Indeed, condition on the event $Z_1^t = x_1^t$ for some $x_1, \dots, x_t \in \mathcal{X}$. Then the event $Z_1^\ell = Z_{t+1}^{t+\ell}$ implies that, for $i \in \{t+1, \dots, t+\ell\}$, $Z_i = x_{\pi(i)}$ where $\pi(i) = i \mod t$, $\pi(i) \in [1, t]$. Then

$$\begin{aligned}
\mathbb{P}\big(Z_1^\ell = Z_{t+1}^{t+\ell}\big) &\le \max_{x_1^t \in \mathcal{X}^t} \mathbb{P}\big(Z_1^\ell = Z_{t+1}^{t+\ell} | Z_1^t = x_1^t\big) \\
&\le \max_{x_1^t \in \mathcal{X}^t} \mathbb{P}\big(Z_i = x_{\pi(i)} \forall i \in \{t+1, \dots, t+\ell\} | Z_1^t = x_1^t\big) \\
&\le \max_{x_1^t \in \mathcal{X}^t} \prod_{i=t+1}^{t+\ell} \mathbb{P}\big(Z_i = x_{\pi(i)}\big) \le (1-c)^\ell \,.
\end{aligned}$$

This proves claim (F.4).

Let us now reconsider our original setting:

$$\begin{aligned}
\mathbb{P}\big(\max_{k \le N} L_k(\boldsymbol{X}^N) \ge \ell\big) &= \mathbb{P}\big(\exists i < j \le N : \; X_i^{i+\ell-1} = X_j^{j+\ell-1}\big) \\
&\le N^2 \max_{i < j \le N} \mathbb{P}\big(X_i^{i+\ell-1} = X_j^{j+\ell-1}\big) \\
&\le N^2 \max_{\boldsymbol{u}^m \in \mathcal{L}^m, \boldsymbol{v}^n \in \mathcal{L}^n} \max_{i < j \le N} \mathbb{P}\big(X_i^{i+\ell-1} = X_j^{j+\ell-1} \big| \boldsymbol{u}^m, \boldsymbol{v}^n\big) \\
&\le N^2 (1-c)^\ell \,,
\end{aligned}$$

where the last inequality follows from claim (F.4), since the $(X_i)_{i \le N}$ are conditionally independent given the latents $\boldsymbol{u}^m, \boldsymbol{v}^n$, with probability mass function upper bounded by $1 - c$. The thesis follows by taking $\ell = 12 \log N / \log(1/(1-c))$. □

For $i \in [m]$, $j \in [n]$, we define $\langle ij \rangle := (i-1)n + j$. In words, $k = \langle ij \rangle$ is the of entry at row i column j when the table $\boldsymbol{X}^{m,n}$ is scanned in row first order. For $\ell \ge 1$, define the events

$$\mathcal{E}_{i,j}(\ell) := \Big\{ \exists i' \in [m], j' \in [n] \;:\; \langle i'j' \rangle < \langle ij \rangle,\; |j' - j| \ge \ell,\; \boldsymbol{X}_{\langle i'j' \rangle}^{\langle i'j' \rangle + \ell - 1} = \boldsymbol{X}_{\langle ij \rangle}^{\langle ij \rangle + \ell - 1} \Big\} \,, \tag{F.5}$$

$$\mathcal{F}_{i,j}(\ell) := \Big\{ \exists i' \in [m], j' \in [n] \;:\; \langle i'j' \rangle < \langle ij \rangle,\; |j' - j| < \ell,\; \boldsymbol{X}_{\langle i'j' \rangle}^{\langle i'j' \rangle + \ell - 1} = \boldsymbol{X}_{\langle ij \rangle}^{\langle ij \rangle + \ell - 1} \Big\} \,. \tag{F.6}$$

Then we have

$$\mathbb{P}\big(L_{\langle ij \rangle}(\boldsymbol{X}^N) \ge \ell\big) \le \mathbb{P}\big(\mathcal{E}_{i,j}(\ell)\big) + \mathbb{P}\big(\mathcal{F}_{i,j}(\ell)\big) \,. \tag{F.7}$$

The next two lemmas control the probabilities of these events.

Lemma F.2. *Let $\ell(\delta, u) := \lceil (1+\delta)(\log N)/H(X|U=u) \rceil$, $n' = n - \max_{u \in \mathcal{L}} \ell(\delta, u)$, and $m_0 = m^{1 - o_n(1)}$. Under Assumption 5.5, for any $\delta > 0$, there exist constants $C, \varepsilon > 0$ independent of $\boldsymbol{u} \in \mathcal{L}^m$, such that the following hold*

$$\max_{i \le m, j \le n'} \mathbb{P}\big(\mathcal{E}_{i,j}(\ell(\delta, u_i))\big) \le C N^{-\varepsilon} \,, \tag{F.8}$$

$$\min_{m_0 \le i \le m, j \le n'} \mathbb{P}\big(\mathcal{E}_{i,j}(\ell(-\delta, u_i))\big) \ge 1 - C N^{-\varepsilon} \,. \tag{F.9}$$

Lemma F.3. *Let $\ell_c(\delta,u) := \lceil (1+\delta)(\log m)/H(X|U=u,V)\rceil$, $n_c' = n - \max_{u\in\mathcal{L}} \ell_c(\delta,u)$, and $m_0 = m^{1-o_n(1)}$. Under Assumption 5.5, for any $\delta>0$, there exist constants $C,\varepsilon>0$ independent of $\boldsymbol{u}\in\mathcal{L}^m$, such that the following hold*

$$\max_{i\le m, j\le n_c'} \mathbb{P}\big(\mathcal{F}_{i,j}(\ell_c(\delta,u_i))\big) \le C\, m^{-\varepsilon}\,, \tag{F.10}$$

$$\min_{m_0\le i\le m, j\le n_c'} \mathbb{P}\big(\mathcal{F}_{i,j}(\ell_c(-\delta,u_i))\big) \ge 1 - C\, m^{-\varepsilon}\,. \tag{F.11}$$

We are now in position to prove Theorem 5.6.

Proof of Theorem 5.6. We denote by $(k(1),\dots,k(M))$ the values taken by k in the while loop of the Lempel-Ziv pseudocode. In particular

$$k(1) = 1\,, \tag{F.12}$$

$$k(\ell+1) = k(\ell) + L_{k(\ell)}(\boldsymbol{X}^N)\,, \tag{F.13}$$

$$k(M) = N\,. \tag{F.14}$$

Therefore the total length of the code is

$$\mathsf{len}(\mathsf{LZ}(\boldsymbol{X}^{m,n})) = M\lceil \log_2(N+|\mathcal{X}|)\rceil + \sum_{\ell=1}^{M} \mathsf{len}(\mathsf{elias}(L_{k(\ell)})) \tag{F.15}$$

By Lemma F.1 (and recalling that $\mathsf{len}(\mathsf{elias}(L)) \le 2\log_2 L + 1$) we have, with high probability, $\max_{\ell\le m} \mathsf{len}(\mathsf{elias}(L_{k(\ell)})) \le 2\log_2(C\log N)$. Letting $\mathcal{G}$ denote the 'good' event that this bound holds, we have, on $\mathcal{G}$

$$M\log_2 N \le \mathsf{len}(\mathsf{LZ}(\boldsymbol{X}^{m,n})) \le M\lceil \log_2(N+|\mathcal{X}|)\rceil + 2M\log_2(C\log N) \tag{F.16}$$

Since $|\mathcal{X}|$ is a constant, this means that for any $\eta>0$, there exists $N_0(\eta)$ such that, for all $N\ge N_0(\eta)$, with probability at least $1-\eta$:

$$M\cdot \mathbf{1}_{\mathcal{G}} \log_2 N \le \mathsf{len}(\mathsf{LZ}(\boldsymbol{X}^{m,n})) \le (1+\eta) M\cdot \mathbf{1}_{\mathcal{G}} \log_2 N + N\cdot \mathbf{1}_{\mathcal{G}^c}\log_2|\mathcal{X}|\,, \tag{F.17}$$

where on the right $\mathsf{len}(\mathsf{LZ}\boldsymbol{X}^{m,n})) \le N\log_2|\mathcal{X}|$ by construction. We thus have

$$\mathbb{E}\big\{M\cdot\mathbf{1}_{\mathcal{G}}\big\}\frac{\log_2 N}{N\log_2|\mathcal{X}|} \le \mathbb{E}\,\mathsf{R}_{\mathsf{LZ}}(\boldsymbol{X}^{m,n}) \le (1+\eta)\mathbb{E}\big\{M\cdot\mathbf{1}_{\mathcal{G}}\big\}\frac{\log_2 N}{N\log_2|\mathcal{X}|} + \eta\,, \tag{F.18}$$

that is

$$\liminf_{m,n\to\infty} \mathbb{E}\,\mathsf{R}_{\mathsf{LZ}}(\boldsymbol{X}^{m,n}) \ge \liminf_{m,n\to\infty} \mathbb{E}\big\{M\cdot\mathbf{1}_{\mathcal{G}}\big\}\cdot\frac{\log_2 N}{N\log_2|\mathcal{X}|}\,, \tag{F.19}$$

$$\limsup_{m,n\to\infty} \mathbb{E}\,\mathsf{R}_{\mathsf{LZ}}(\boldsymbol{X}^{m,n}) \le \limsup_{m,n\to\infty} \mathbb{E}\big\{M\cdot\mathbf{1}_{\mathcal{G}}\big\}\cdot\frac{\log_2 N}{N\log_2|\mathcal{X}|}\,. \tag{F.20}$$

We are therefore left with the task of bounding $\mathbb{E}\big\{M\cdot\mathbf{1}_{\mathcal{G}}\big\}$

We begin by the lower bound. Define the set of 'bad indices' $B(\boldsymbol{X}^{m,n},\delta)\subseteq[m]\times[n]$,

$$B(\boldsymbol{X}^{m,n},\delta) := \Big\{(i,j)\in[m]\times[n]:\ \mathcal{E}_{i,j}(\ell(\delta,u_i)) \text{ or } \mathcal{F}_{i,j}(\ell_c(\delta,u_i))\Big\} \tag{F.21}$$

We will drop the arguments $\boldsymbol{X}^{m,n},\delta$ for economy of notation, and write $B := B(\boldsymbol{X}^{m,n},\delta)$. We further define

$$S(u) = S(u;\boldsymbol{X}^{m,n}) := \{(i,j)\in[m]\times[n]:\ u_i = u \text{ and } \exists \ell\le M:\ \langle ij\rangle = k(\ell)\}\,. \tag{F.22}$$

In words, $S(u)$ is the set of positions (i,j) of the table $\boldsymbol{X}^{m,n}$ where words in the LZ parsing begin.

We also write $N(u) = n \cdot |\{i \in [m] : \ u_i = u\}|$ for the total number of rows in $\boldsymbol{X}^{m,n}$ with row latent equal to u_i and L_i^- for the length of the first segment in row i initiated in row $i-1$:

$$\begin{aligned}
N(u) &\le \sum_{(i,j)\in S(u)} L_{\langle ij\rangle} + \sum_{i\le m: u_i = u} L_i^- \\
&\le \sum_{(i,j)\in S(u)\cap B^c} L_{\langle ij\rangle} + \sum_{(i,j)\in B} L_{\langle ij\rangle} + \sum_{i\le m: u_i = u} L_i^- \\
&\le \sum_{(i,j)\in S(u)} \ell(u;\delta) \vee \ell_c(u;\delta) + (|B| + m)\cdot C\log N \\
&\le |S(u)|\ell(u;\delta) \vee \ell_c(u;\delta) + (|B| + m)\cdot C\log N\,,
\end{aligned}$$

where the last inequality holds on event $\mathcal{G}$. By taking expectation on this event, we get

$$\mathbb{E}\{N(u)\cdot \mathbf{1}_{\mathcal{G}}\} \le \mathbb{E}\{|S(u)|\cdot \mathbf{1}_{\mathcal{G}}\}\}\cdot \ell(u;\delta) \vee \ell_c(u;\delta) + (\mathbb{E}|B| + m)\cdot C\log N\,.$$

By Lemmas F.2 and F.2,

$$\begin{aligned}
\mathbb{E}(|B|) &\le m_0 n + \sum_{m_0\le i\le m, j\le n'} \mathbb{P}\Big(\mathcal{E}_{i,j}(\ell(\delta,u_i)) \cup \mathcal{F}_{i,j}(\ell_c(\delta,u_i))\Big) + Cm\log N \\
&\le m_0 n + Cm^{1-\varepsilon} n + Cm\log N \\
&\le \frac{CN}{(\log N)^2}\,.
\end{aligned}$$

$\mathbb{E}(|B|) \le Cm^{1-\varepsilon}n + Cm\log n$. Further $\mathbb{E}\{N(u)\} = N q_{\mathrm{r}}(u)$ and $\mathbb{E}\{N(u)\cdot \mathbf{1}_{\mathcal{G}}\} \ge \mathbb{E}\{N(u)\} - N\,\mathbb{P}(\mathcal{G}^c)$, whence

$$\liminf_{m,n\to\infty} \frac{1}{N}\mathbb{E}\{|S(u)|\cdot \mathbf{1}_{\mathcal{G}}\}\}\cdot \ell(u;\delta) \vee \ell_c(u;\delta) \ge q_{\mathrm{r}}(u)\,. \tag{F.23}$$

Recalling the definition of $\ell(u;\delta)$, $\ell_c(u;\delta)$ and the fact that δ is arbitrary,n the last inequality yields

$$\liminf_{m,n\to\infty} \mathbb{E}\{|S(u)|\cdot \mathbf{1}_{\mathcal{G}}\}\frac{\log_2 N}{N} \ge q_{\mathrm{r}}(u)\Big[H(X|U=u) \wedge \Big(\frac{1+\alpha}{\alpha}\Big)H(X|U=u,V)\Big]\,. \tag{F.24}$$

Summing over u, noting that $\sum_{u\in\mathcal{L}} |S(u)| = M$, and substituting in Eq. (F.19) yields the lower bound on the rate in Eq. (5.7).

Finally, the upper bound is proved by a similar strategy as for the lower bound. Define the set of 'bad indices' $B_- = B_-(\boldsymbol{X}^{m,n},\delta) \subseteq [m]\times[n]$,

$$B_-(\boldsymbol{X}^{m,n},\delta) := \Big\{(i,j)\in[m]\times[n] : \ \mathcal{E}_{i,j}^c(\ell(-\delta,u_i)) \text{ or } \mathcal{F}_{i,j}^c(\ell_c(-\delta,u_i))\Big\} \tag{F.25}$$

We also denote by L_i^+ the length of the last segment in row i. We then have

$$\begin{aligned}
N(u) &\ge \sum_{(i,j)\in S(u)} L_{\langle ij\rangle} - \sum_{i\le m: u_i = u} L_i^+ \\
&\ge \sum_{(i,j)\in S(u)\cap B_-^c} L_{\langle ij\rangle} - \sum_{i\le m: u_i = u} L_i^+ \\
&\ge \sum_{(i,j)\in S(u)\cap B_-^c} \ell(u;-\delta) \vee \ell_c(u;-\delta) - \sum_{i\le m: u_i = u} L_i^+ \\
&\ge |S(u)|\ell(u;\delta) \vee \ell_c(u;\delta) - (|B_-| + m)\cdot C\log N\,,
\end{aligned}$$

where the last inequality holds on event $\mathcal{G}$. By taking expectation on this event, we get

$$\mathbb{E}\{N(u)\cdot \mathbf{1}_{\mathcal{G}}\} \ge \mathbb{E}\{|S(u)|\cdot \mathbf{1}_{\mathcal{G}}\}\}\cdot \ell(-\delta,u) \vee \ell_c(-\delta,u) - (\mathbb{E}|B_-| + m)\cdot C\log N\,.$$

By Lemmas F.2 and F.2,

$$\begin{aligned}\mathbb{E}(|B_-|) &\le m_0 n + \sum_{m_0 \le i \le m, j \le n'} \mathbb{P}\Big(\mathcal{E}_{i,j}^c(\ell(-\delta, u_i)) \cup \mathcal{F}_{i,j}^c(\ell_c(-\delta, u_i))\Big) + Cm \log N \\ &\le m_0 n + Cm^{1-\varepsilon} n + Cm \log N \\ &\le \frac{CN}{(\log N)^2}\,.\end{aligned}$$

The proof is completed exactly as for the lower bound. □

F.2. Proof of Lemma F.2

We will use the following standard lemmas.

Lemma F.4. *Let X be a centered random variable with $\mathbb{P}(X \le x_0) = 1$, $x_0 > 0$. Then, letting $c(x_0) = (e^{x_0} - 1 - x_0)/x_0^2$, we have*

$$\mathbb{E}(e^X) \le 1 + c(x_0)\mathbb{E}(X^2)\,. \tag{F.26}$$

Proof. This simply follows from $\exp(x) \le 1 + x + c(x_0)x^2$ for $x \le x_0$. □

Lemma F.5. *Let $(p_i)_{i\ge 1}$, $(q_i)_{i\ge 1}$, be probability distributions on $\mathcal{X}$, with $\sup_{i\ge 1} \max_{x\in\mathcal{X}} p_i(x) \le 1 - c$, and $\sup_{i\ge 1} \sum_{x\in\mathcal{X}} p_i(x)(\log p_i(x))^2 \le C$ for constants c, C.*

Let $(X_i)_{i\le \ell}$ be independent random variables with $X_i \sim p_i$, and set $\boldsymbol{X} = (X_1, \dots, X_\ell)$. Let $\boldsymbol{Y}(j) \in \mathcal{X}^\ell$, $j \ge 1$ be a sequence of i.i.d. random vectors, with $(Y_i(j))_{i\le\ell}$ independent and $Y_i(j) \sim q_i$. Finally, let $T := \min\{t \ge 1 :\ \boldsymbol{Y}(t) = \boldsymbol{X}\}$.

Then, for any $\varepsilon > 0$, there exists $\delta = \delta(\varepsilon, c, C) > 0$ such that (letting $\overline{H}(p) := \ell^{-1}\sum_{i=1}^\ell H(p_i)$)

$$\mathbb{P}(T \le e^{\ell[\overline{H}(p) - \varepsilon]}) \le e^{-\delta\ell}\,. \tag{F.27}$$

Further, the same bound holds (with a different $\delta(\varepsilon, c, C)$) $(\boldsymbol{Y}(j))_{j\ge 1}$ are independent not identically distributed, if there exist a finite set $(q_i^a)_{i\ge 1, a\in[K]}$, $K \le \ell^{C_0}$, and a map $b : \mathbb{N} \to [K]$ such that $\boldsymbol{Y}(j) \sim q_1^{b(j)} \otimes \cdots \otimes q_\ell^{b(j)}$.

Proof. We denote by $\boldsymbol{Y}$ a vector distributed as $\boldsymbol{Y}(i)$. Conditional on $\boldsymbol{X} = \boldsymbol{x}$, T is a geometric random variables with mean $1/(1 - \mathbb{P}(\boldsymbol{Y} = \boldsymbol{x}))$. Hence, for $t_\ell(\varepsilon) := e^{\ell[\overline{H}(p) - \varepsilon]}$,

$$\begin{aligned}\mathbb{P}(T \le t_\ell(\varepsilon) | \boldsymbol{X} = \boldsymbol{x}) &= 1 - (1 - \mathbb{P}(\boldsymbol{Y} = \boldsymbol{x}))^{t_\ell(\varepsilon)} \\ &\le t_{\ell(\varepsilon)} \mathbb{P}(\boldsymbol{Y} = \boldsymbol{x})\,.\end{aligned}$$

Hence

$$\mathbb{P}\big(T \le t_\ell(\varepsilon)\big) \le e^{-\ell\varepsilon/2} + \mathbb{P}\Big(\mathbb{P}(\boldsymbol{Y} = \boldsymbol{X}|\boldsymbol{X}) \ge t_\ell(\varepsilon/2)^{-1}\Big) \tag{F.28}$$

$$= e^{-\ell\varepsilon/2} + P_\ell(\varepsilon/2)\,, \tag{F.29}$$

$$P_\ell(u) := \mathbb{P}\Big(\sum_{i=1}^\ell \log\frac{1}{q_i(X_i)} < \sum_{i=1}^\ell H(p_i) - \ell\, u\Big)\,. \tag{F.30}$$

By Chernoff bound, for any $\lambda \ge 0$, $P_\ell(u) \le \exp\{-\ell\phi(\lambda, u)\}$, where

$$\phi(\lambda, u) := \lambda u - \frac{1}{\ell}\sum_{i=1}^\ell \big[\lambda H(p_i) + \log\mathbb{E}\big[q_i(X_i)^\lambda\big]\,. \tag{F.31}$$

By Hölder inequality, for $\lambda \in [0, 1]$ we have $\mathbb{E}\big[q_i(X_i)^\lambda\big] \le (\sum_x p(x)^\beta)^{1/\beta}$ where $\beta = 1/(1 - \lambda)$. Therefore

$$\begin{aligned}\psi(\lambda; p) &:= \lambda H(p) + (1 - \lambda)\log\Big(\sum_{x\in\mathcal{X}} p(x)^{1/(1-\lambda)}\Big) \\ &= (1 - \lambda)\log\mathbb{E}_{X\sim p}\exp\Big(\frac{\lambda}{1 - \lambda}\big(\log p(X) + H(p)\big)\Big)\,.\end{aligned}$$

Consider the random variable $Z_i := \frac{\lambda}{1-\lambda}\big(\log p_i(X_i) + H(p)\big)$ where $X_i \sim p_i$. Under the assumptions of the lemma, for $\lambda \in [0, 1/2]$ we have $\mathbb{E}(Z_i) = 0$ and

$$Z_i \le \log(1-c) + H(p) \le \log\big[|\mathcal{X}|(1-c)\big]\,, \tag{F.32}$$

$$\mathbb{E}[Z_i^2] \le \left(\frac{\lambda}{1-\lambda}\right)^2 \sum_{x\in\mathcal{X}} p_i(x)(\log p_i(x))^2 \le 4C\lambda^2\,, \tag{F.33}$$

Using Lemma F.4, we get

$$\psi(\lambda; p_i) = (1-\lambda)\log \mathbb{E}e^{Z_i} \tag{F.34}$$

$$\le (1-\lambda)\log\big(1 + c_0\mathbb{E}(Z_i^2)\big) \tag{F.35}$$

$$\le \log(1 + c_*\lambda^2)\,, \tag{F.36}$$

whence

$$\phi(\lambda, u) \ge \lambda u - \log(1 + c_*\lambda^2)\,.$$

By maximizing this expression over λ, we find that $P_\ell(\varepsilon/2) \le \exp(-\delta_0(\varepsilon)\ell)$ which completes the proof for the case of i.i.d. vectors $\boldsymbol{Y}(j)$.

The case of non-identically distributed vectors follows by union bound over $a \in [K]$. □

Lemma F.6. *Let $(p_i)_{i\ge1}$, be probability distributions on $\mathcal{X}$, with $\sup_{i\ge1}\max_{x\in\mathcal{X}} p_i(x) \le 1-c$, and $\sup_{i\ge1}\sum_{x\in\mathcal{X}} p_i(x)(\log p_i(x))^2 \le C$ for constants c, C.*

Let $(X_i)_{i\le\ell}$ be independent random variables with $X_i \sim p_i$, $\boldsymbol{X} = (X_1, \dots, X_\ell)$. Let $\boldsymbol{Y}(j) \in \mathcal{X}^\ell$, $j \ge 1$ be a sequence of i.i.d. copies of $\boldsymbol{X}$. Finally, let $T := \min\{t \ge 1 :\ \boldsymbol{Y}(t) = \boldsymbol{X}\}$.

Then, for any $\varepsilon > 0$, there exists $\delta = \delta(\varepsilon, c, C) > 0$, such that (letting $\overline{H}(p) := \ell^{-1}\sum_{i=1}^\ell H(p_i)$)

$$\mathbb{P}(T \ge e^{\ell[\overline{H}(p)+\varepsilon]}) \le e^{-\delta\ell}\,. \tag{F.37}$$

Proof. The proof follows the same argument as for Lemma F.5. Denote by $\boldsymbol{Y}$ a vector distributed as $\boldsymbol{Y}(i)$. and define $t_\ell(\varepsilon) := e^{\ell[\overline{H}(p)+\varepsilon]}$,

$$\begin{aligned}\mathbb{P}(T \ge t_\ell(\varepsilon)|\boldsymbol{X} = \boldsymbol{x}) &= (1 - \mathbb{P}(\boldsymbol{Y} = \boldsymbol{x}))^{t_\ell(\varepsilon)} \\ &\le \exp\Big(-t_{\ell(\varepsilon)}\mathbb{P}(\boldsymbol{Y} = \boldsymbol{x})\Big)\,.\end{aligned}$$

Hence

$$\mathbb{P}\big(T \ge t_\ell(\varepsilon)\big) \le \exp\Big\{-e^{\ell\varepsilon/2}\Big\} + \mathbb{P}\Big(\mathbb{P}(\boldsymbol{Y} = \boldsymbol{X}|\boldsymbol{X}) \le t_\ell(\varepsilon/2)^{-1}\Big) \tag{F.38}$$

$$\le e^{-\ell\varepsilon/2} + \tilde{P}_\ell(\varepsilon/2)\,, \tag{F.39}$$

$$\tilde{P}_\ell(u) := \mathbb{P}\Big(\sum_{i=1}^\ell \log\frac{1}{p_i(X_i)} \ge \sum_{i=1}^\ell H(p_i) + \ell\, u\Big)\,. \tag{F.40}$$

We claim that, for each $u > 0$, $\tilde{P}_\ell(u) \le e^{-\delta_0(u)\ell}$ for some $\delta_0(u) > 0$. Indeed, using again Chernoff's bound, we get, for any $\lambda \ge 0$, $\tilde{P}_\ell(u) \le e^{-\ell\tilde{\phi}(\lambda,u)}$, where

$$\tilde{\phi}(\lambda, u) := \lambda u - \frac{1}{\ell}\sum_{i=1}^\ell \tilde{\psi}(\lambda; p_i)\,, \tag{F.41}$$

$$\tilde{\psi}(\lambda; p_i) := \log \mathbb{E}\exp(\lambda W_i)\,, \quad W_i := \log\frac{1}{p_i(X_i)} - H(p_i)\,. \tag{F.42}$$

where in the last line $X_i \sim p_i$. Under the assumptions of the lemma, $W_i \le C$ almost surely and applying again Lemma F.4, we get $\tilde{\psi}(\lambda; p_i) \le \log(1 + c_*\lambda^2)$ for $\lambda \le 1$. The proof is completed by selecting for each $u > 0$, $\lambda > 0$ so that $\lambda u - \log(1 + c_*\lambda^2) > 0$. □

We are now in position to prove Lemma F.2.

Proof of Lemma F.2. We begin by proving the bound (F.8).

Fix $i \le m$, $j \le n'$, $\boldsymbol{u} \in \mathcal{L}^m$, $\delta > 0$, and write $\ell = \ell(\delta, u_i)$. Define $R_{ij} := \{(i, j') : \max(1, j-\ell) \le j' \le j-1\}$ and $S_{ij} := \{(i', j') : i' < i \text{ or } i' = i, j' < j - \ell\}$. Finally, for $t \in \{0, \dots, \ell-1\}$, let $S_{ij}(t) := S_{ij} \cap \{(i', j') : \langle i'j' \rangle = t \mod \ell\}$.

By union bound

$$\begin{aligned}
\mathbb{P}\big(\mathcal{E}_{i,j}(\ell)\big|\boldsymbol{u}\big) &\le A + \sum_{t=0}^{\ell-1} B(t)\,, \\
A &:= \sum_{(rs)\in R_{ij}} \mathbb{P}\big(\boldsymbol{X}_{\langle rs\rangle}^{\langle rs\rangle+\ell-1} = \boldsymbol{X}_{\langle ij\rangle}^{\langle ij\rangle+\ell-1}\big|\boldsymbol{u}\big)\,, \\
B(t) &:= \mathbb{P}\Big(\exists (r,s) \in S_{ij}(t) :\ \boldsymbol{X}_{\langle rs\rangle}^{\langle rs\rangle+\ell-1} = \boldsymbol{X}_{\langle ij\rangle}^{\langle ij\rangle+\ell-1}\Big|\boldsymbol{u}\Big)\,.
\end{aligned}$$

Now, by the bound of Eq. (F.4),

$$A \le \ell \cdot (1-c)^{\ell} \le C\, N^{-\varepsilon}\,, \tag{F.43}$$

for suitable constants C, ε.

Next, for any $t \in \{0, \dots, \ell-1\}$, the vectors $\{\boldsymbol{X}_{\langle rs\rangle}^{\langle rs\rangle+\ell-1}\}$ are mutually independent and independent of $\{\boldsymbol{X}_{\langle ij\rangle}^{\langle ij\rangle+\ell-1}\}$. Conditional on $\boldsymbol{u}$, the coordinates of $\boldsymbol{X}_{\langle rs\rangle}^{\langle rs\rangle+\ell-1} = (X_{\langle rs\rangle}, \cdots, X_{\langle rs\rangle+\ell-1})$ are independent with marginal distributions $X_{\langle r's'\rangle} \sim Q_{x|u}(\,\cdot\,|u_{r'})$ (note that independence of the coordinates holds because $\ell < m/2$ and therefore $\boldsymbol{X}_{\langle rs\rangle}^{\langle rs\rangle+\ell-1}$ does not include two entries in the same column). Note that the collection of marginal distributions $Q_{x|u}(\,\cdot\,|u)$, $u \in \mathcal{L}$ satisfies the conditions of Lemma F.5 by assumption. Further, the vector $\boldsymbol{X}_{\langle rs\rangle}^{\langle rs\rangle+\ell-1}$ can have at most one of $K = |\mathcal{L}|^2(\ell+1)$ distributions (depending on the latents value and the occurrence of a line break in the block.)

Applying Lemma F.5, we obtain:

$$B(t) \le e^{-\varepsilon_0 \ell} \le C\, N^{-\varepsilon} \tag{F.44}$$

Summing over $t \in \{0, \dots, \ell-1\}$ and adjusting the constants yields the claim (F.8).

Next consider the bound (F.9). Fix $\boldsymbol{u} \in \mathcal{L}^m$, $i \le m$,$j \le n'$, and write $\ell = \ell(-\delta, u_i)$ for brevity below

$$\mathbb{P}\big(\mathcal{E}_{i,j}^c(\ell)|\boldsymbol{u}\big) \le \mathbb{P}\Big(\forall (i',j') \in S_{ij}(t) \text{ s.t. } u_{i'} = u_i, j' < n' : \boldsymbol{X}_{\langle i'j'\rangle}^{\langle i'j'\rangle+\ell-1} \neq \boldsymbol{X}_{\langle ij\rangle}^{\langle ij\rangle+\ell-1}\Big|\boldsymbol{u}\Big)\,. \tag{F.45}$$

Here $t \in \{0, \dots, \ell-1\}$ can be chosen arbitrarily. Let $S_{ij}(t; \boldsymbol{u}) := \{(i', j') \in S_{ij}(t) \text{ s.t. } u_{i'} = u_i, j' < n'\}$. Conditional on $\boldsymbol{u}$, the vectors $(\boldsymbol{X}_{\langle i'j'\rangle}^{\langle i'j'\rangle+\ell-1})_{(i',j')\in S_{ij}(t;\boldsymbol{u})}$ are i.i.d. and independent of $\boldsymbol{X}_{\langle ij\rangle}^{\langle ij\rangle+\ell-1}$. Further, they are distributed as $\boldsymbol{X}_{\langle ij\rangle}^{\langle ij\rangle+\ell-1}$. Finally,

$$\begin{aligned}
N_{ij}(\boldsymbol{u}) := \big|S_{ij}(t;\boldsymbol{u})\big| &\ge \frac{n\big(m_i(u) - C\log N\big)}{\ell} \\
&\ge \frac{m_i(u) n}{C \log N} - C' n\,.
\end{aligned}$$

where $m_i(u)$ is the number rows $i' < i$ such that $u_{i'} = u$. Since $i \ge i$ and $\mathbb{P}(u_{i'} = u) \ge \min_{u'} q_{\mathrm{r}}(u') > 0$, by Chernoff bound there exist constants C, c_0 such that, for all m, n large enough (since $i \ge m_0$)

$$\mathbb{P}\Big(N_{ij}(\boldsymbol{u}) \ge \frac{c_0 m_0 n}{\log N}\Big) \ge 1 - C e^{-m_0/C}\,. \tag{F.46}$$

Further, for any $\delta > 0$ we can choose positive constants $\varepsilon_0, \varepsilon_1 > 0$ such that the following holds for all m, n, large enough

$$\frac{c_0 m_0 n}{\log N} \ge N^{1-\varepsilon_1} \ge e^{\ell[H(X|U=u_i)+\varepsilon_0]} \tag{F.47}$$

Let T_{ij} be the rank of the first (i',j') (moving backward)in the set defined above such that $\boldsymbol{X}_{\langle i'j'\rangle}^{\langle i'j'\rangle+\ell-1} = \boldsymbol{X}_{\langle ij\rangle}^{\langle ij\rangle+\ell-1}$, and $T_{ij}=\infty$ if no such vector exists. We can continue from Eq. (F.45) to get

$$\begin{aligned}\mathbb{P}\big(\mathcal{E}_{i,j}^c(\ell)\big) &\le \mathbb{P}(T_{ij}\ge N_{ij}(\boldsymbol{u}))\\ &\le \mathbb{P}\big(T_{ij}\ge N_{ij}(\boldsymbol{u});\ N_{ij}(\boldsymbol{u})\ge e^{\ell[H(X|U=u_i)+\varepsilon_0]}\big) + Ce^{-m_0/C}\\ &\overset{(a)}{\le} \exp\big\{-\delta_0 \min_{u\in\mathcal{L}}\big(\ell(-\delta;u)\big)\big\} + Ce^{-m_0/C} \le CN^{-\varepsilon}\,,\end{aligned}$$

where in (a) we used Lemma F.6. This completes the proof of Eq. (F.9). □

F.3. Proof of Lemma F.3

We begin by considering the bound (F.10).

Fix $i\le m$, $j\le n_c'$, $\boldsymbol{u}\in\mathcal{L}^m$, $\boldsymbol{v}\in\mathcal{L}^n$, $\delta>0$, and write $\ell=\ell_c(\delta,u_i)$, $n'=n_c'$. By union bound:

$$\begin{aligned}\mathbb{P}\big(\mathcal{F}_{i,j}(\ell)\big|\boldsymbol{u},\boldsymbol{v}\big) &= \mathbb{P}\Big(\cup_{s\in[n],|j-s|<\ell}\boldsymbol{B}(s)\Big|\boldsymbol{u},\boldsymbol{v}\Big)\,,\\ \mathcal{B}(s) &:= \Big\{\exists r<i:\ \boldsymbol{X}_{\langle rs\rangle}^{\langle rs\rangle+\ell-1} = \boldsymbol{X}_{\langle ij\rangle}^{\langle ij\rangle+\ell-1}\Big\}\,.\end{aligned}$$

Note that for a fixed s, and conditional on $\boldsymbol{u},\boldsymbol{v}$, the vectors $(\boldsymbol{X}_{\langle rs\rangle}^{\langle rs\rangle+\ell-1})_{1\le s\le i-1}$ are mutually independent and independent of $\boldsymbol{X}_{<ij>}^{<ij>+\ell-1}$. Further, $\boldsymbol{X}_{\langle rs\rangle}^{\langle rs\rangle+\ell-1}$ has independent coordinates with marginals $X_{\langle r's'\rangle}\sim Q_{x|u}(\,\cdot\,|u_{r'},v_{s'})$ (recall that we are conditioning both on $\boldsymbol{u}$ and $\boldsymbol{v}$). In particular, the marginal distributions satisfy the assumption of Lemma F.5 and the law of $\boldsymbol{X}_{\langle rs\rangle}^{\langle rs\rangle+\ell-1}$ can take one of $K=|\mathcal{L}|^2(\ell+1)$ possible values. Letting $i-T(s)$ the last row at which $\boldsymbol{X}_{\langle rs\rangle}^{\langle rs\rangle+\ell-1} = \boldsymbol{X}_{\langle ij\rangle}^{\langle ij\rangle+\ell-1}$ (with $T(s)\ge i$ if no such row exists), we have, for some constants $C,c_0>0$,

$$\begin{aligned}\mathbb{P}\big(\mathcal{F}_{i,j}(\ell)\big|\boldsymbol{u},\boldsymbol{v}\big) &\le \mathbb{P}\Big(\cup_{s\in[n],|j-s|<\ell}\{T(s)\le i-1\}\cap\{i-1\le e^{\ell[\overline{H}-\varepsilon]}\}\Big|\boldsymbol{u},\boldsymbol{v}\Big) + \mathbf{1}(i-1>e^{\ell[\overline{H}-\varepsilon]})\\ &\overset{(a)}{\le} 2\ell\, e^{-\ell\varepsilon} + \mathbf{1}(m>e^{\ell[\overline{H}-\varepsilon]})\\ &\le Cm^{-c_0\varepsilon} + \mathbf{1}(m>e^{\ell[\overline{H}-\varepsilon]})\,,\end{aligned}$$

where in (a) we used Lemma F.5, and we defined $\overline{H}:=\ell^{-1}\sum_{k=j}^{j+\ell-1}H(X|U=u_i,V=v_k)$.

Taking expectation with respect to $\boldsymbol{v}$, we get

$$\begin{aligned}\mathbb{P}\big(\mathcal{F}_{i,j}(\ell)\big|\boldsymbol{u}\big) &\le Cm^{-c_0\varepsilon} + \mathbb{P}\Big(\frac{1}{\ell}\sum_{k=j}^{j+\ell-1}H(X|U=u_i,V=v_k) < \frac{1}{1+\delta}\big(H(X|U=u_i,V)+\varepsilon\big)\Big)\\ &\overset{(a)}{\le} Cm^{-c_0\varepsilon} + e^{-\ell\varepsilon} \le C'm^{-c_0\varepsilon}\,,\end{aligned}$$

where in (a) we used Chernoff bound. This completes the proof of Eq. (F.10).

Finally, the proof Eq. (F.11) is similar to the one of Eq. (F.9). We fix $\boldsymbol{u}\in\mathcal{L}^m$, $i\le m$, $j\le n_c'$, and write $\ell=\ell_c(-\delta,u_i)$.

$$\mathbb{P}\big(\mathcal{F}_{i,j}^c(\ell)|\boldsymbol{u}\big) \le \mathbb{P}\Big(\forall i'<i u_{i'}=u_i:\boldsymbol{X}_{\langle i'j\rangle}^{\langle i'j\rangle+\ell-1}\ne\boldsymbol{X}_{\langle ij\rangle}^{\langle ij\rangle+\ell-1}\Big|\boldsymbol{u}\Big)\,. \tag{F.48}$$

Let $S_{ij}^c(\boldsymbol{u}):=\{(i',j)\ \text{ s.t. } u_{i'}=u_i, i'<i\}$. Conditional on $\boldsymbol{u},\boldsymbol{v}$, the vectors $(\boldsymbol{X}_{\langle i'j'\rangle}^{\langle i'j'\rangle+\ell-1})_{(i',j')\in S_{ij}^c(u)}$ are i.i.d. and independent copies of $\boldsymbol{X}_{\langle ij\rangle}^{\langle ij\rangle+\ell-1}$. Finally, $N_i^c(\boldsymbol{u}):=\big|S_{ij}^c(\boldsymbol{u})\big|$ is the number rows $i'<i$ such that $u_{i'}=u$. By Chernoff bound there exist constants C,c_0 such that, for all m,n large enough (recalling that we need only to consider $i\ge m_0$)

$$\mathbb{P}\Big(N_i^c(\boldsymbol{u})\ge c_0m_0\Big)\ge 1-Ce^{-m_0/C}\,. \tag{F.49}$$

Since $m_0\ge m^{1-o_n(1)}$, for any $\delta>0$ we can choose constants $\varepsilon_0,\varepsilon_1>0$ so that

$$c_0m_0\ge m^{1-\varepsilon_1}\ge e^{\ell[H(X|U=u_i,V)+2\varepsilon_0]}\,. \tag{F.50}$$

Recall the definition $\overline{H} := \ell^{-1}\sum_{k=j}^{j+\ell-1} H(X|U=u_i, V=v_k)$. By an an application of Chernoff bound

$$\mathbb{P}\Big(N_i^c(\boldsymbol{u}) \geq e^{\ell[H(X|U=u_i,V)+\varepsilon_0]}\Big) \geq 1 - Cm^{-\varepsilon} - Ce^{-m_0/C}\,. \tag{F.51}$$

Let T_i be the rank of the first i' (moving backward)in the set defined above such that $\boldsymbol{X}_{\langle i'j\rangle}^{\langle i'j\rangle+\ell-1} = \boldsymbol{X}_{\langle ij\rangle}^{\langle ij\rangle+\ell-1}$, and $T_i = \infty$ if no such vector exists. From Eq. (F.48) we get

$$\begin{aligned}
\mathbb{P}\big(\mathcal{F}_{i,j}^c(\ell)\big) &\leq \mathbb{P}(T_i \geq N_i(\boldsymbol{u})) \\
&\leq \mathbb{P}\big(T_i \geq N_i(\boldsymbol{u});\ N_i(\boldsymbol{u}) \geq e^{\ell[H(X|U=u_i,V)+\varepsilon_0]}\big) + Cm^{-\varepsilon} \\
&\overset{(a)}{\leq} \exp\big\{-\delta_0 \min_{u\in\mathcal{L}}\big(\ell_c(-\delta;u)\big)\big\} + Cm^{-\varepsilon} \leq 2Cm^{-\varepsilon}\,,
\end{aligned}$$

where in (a) we used Lemma F.6.

G. Proofs for latent-based encoders

G.1. Proof of Lemma 5.7

G.1.1. General bound (5.8)

From Eq. (1.3), we get

$$\mathsf{R}_{\text{lat}}(\boldsymbol{X}) = \frac{1}{mn\log_2|\mathcal{X}|}\Big\{\mathsf{len}(\mathsf{header}) + \mathsf{len}(\mathsf{Z}_{\mathcal{L}}(\widehat{\boldsymbol{u}})) + \mathsf{len}(\mathsf{Z}_{\mathcal{L}}(\widehat{\boldsymbol{v}})) + \sum_{u,v\in\mathcal{L}}\mathsf{len}(\mathsf{Z}_{\mathcal{X}}(\hat{\boldsymbol{X}}(u,v)))\Big\} \wedge 1\,, \tag{G.1}$$

where $\hat{\boldsymbol{X}}(u,v) := \mathsf{vec}\big(X_{ij}:\ \hat{u}_i(\boldsymbol{X}) = u, \hat{v}_j(\boldsymbol{X}) = v\big)$ are the estimated blocks of $\boldsymbol{X}$. Note that this rate depends on the base compressors $\mathsf{Z}_{\mathcal{L}}$, $\mathsf{Z}_{\mathcal{X}}$ but we will omit these from our notations.

Define the 'ideal' expected compression rate (i.e. the rate achieved by a compressor that is given the latents):

$$\mathsf{R}_{\#} := \frac{1}{mn\log_2|\mathcal{X}|}\Big\{\mathbb{E}[\mathsf{len}(\mathsf{header})] + \mathbb{E}[\mathsf{len}(\mathsf{Z}(\boldsymbol{u}))] + \mathbb{E}[\mathsf{len}(\mathsf{Z}(\boldsymbol{v}))] + \sum_{u,v\in\mathcal{L}}\mathbb{E}[\mathsf{len}(\mathsf{Z}(\boldsymbol{X}(u,v)))]\Big\}\,.$$

Since $\mathsf{R}_{\text{lat}}(\boldsymbol{X}) \leq 1$ by construction, we have

$$\begin{aligned}
\mathbb{E}\,\mathsf{R}_{\text{lat}}(\boldsymbol{X}) &\leq \mathbb{E}\left\{\mathsf{R}_{\text{lat}}(\boldsymbol{X})\,\mathbf{1}_{\mathsf{Err}_U(\boldsymbol{X};\widehat{\boldsymbol{u}})=1}\mathbf{1}_{\mathsf{Err}_V(\boldsymbol{X};\widehat{\boldsymbol{v}})=1}\right\} + \mathbb{P}\big(\mathsf{Err}_U(\boldsymbol{X};\widehat{\boldsymbol{u}})<1\big) + \mathbb{P}\big(\mathsf{Err}_V(\boldsymbol{X};\widehat{\boldsymbol{v}})<1\big) \\
&\overset{(*)}{\leq} \mathsf{R}_{\#} + \mathbb{P}\big(\mathsf{Err}_U(\boldsymbol{X};\widehat{\boldsymbol{u}})<1\big) + \mathbb{P}\big(\mathsf{Err}_V(\boldsymbol{X};\widehat{\boldsymbol{v}})<1\big)\,,
\end{aligned}$$

where in step $(*)$ we bounded $\mathbb{E}[\mathsf{len}(\mathsf{Z}(\widehat{\boldsymbol{u}}))\mathbf{1}_{\mathsf{Err}_U(\boldsymbol{X};\widehat{\boldsymbol{u}})=1}] = \mathbb{E}[\mathsf{len}(\mathsf{Z}(\boldsymbol{u}))\mathbf{1}_{\mathsf{Err}_U(\boldsymbol{X};\widehat{\boldsymbol{u}})=1}] \leq \mathbb{E}[\mathsf{len}(\mathsf{Z}(\boldsymbol{u}))]$, because, on the event $\{\mathsf{Err}_U(\boldsymbol{X};\widehat{\boldsymbol{u}})=1\}$, $\widehat{\boldsymbol{u}}$ coincides with $\boldsymbol{u}$ up to relabelings, and the compressed length is invariant under relabelings. Similar arguments were applied to $\mathsf{len}(\mathsf{Z}(\boldsymbol{v}))$ and $\mathsf{len}(\mathsf{Z}(\boldsymbol{X}(u,v)))$.

We now have, by the definition of $\Delta_{\mathsf{Z}}(N;k)$ in Eq. (5.12),

$$\frac{\mathbb{E}[\mathsf{len}(\mathsf{Z}(\boldsymbol{u}))]}{mn\log_2|\mathcal{X}|} \leq \frac{H(U)}{n\log_2|\mathcal{X}|} + +\frac{1}{n}\Delta_{\mathsf{Z}}(m\wedge n;\{r,c\})\,, \tag{G.2}$$

$$\frac{\mathbb{E}[\mathsf{len}(\mathsf{Z}(\boldsymbol{v}))]}{mn\log_2|\mathcal{X}|} \leq \frac{H(V)}{m\log_2|\mathcal{X}|} + +\frac{1}{m}\Delta_{\mathsf{Z}}(m\wedge n;\{r,c\})\,, \tag{G.3}$$

$$\frac{\mathbb{E}[\mathsf{len}(\mathsf{Z}(\boldsymbol{X}(u,v)))|\boldsymbol{u},\boldsymbol{v}]}{mn\log_2|\mathcal{X}|} \leq \hat{q}_{\mathsf{r}}(u)\hat{q}_{\mathsf{c}}(v)\frac{H(X|U=u,V=v)}{\log_2|\mathcal{X}|} + \Delta_{\mathsf{Z}}(c\cdot mn;\{Q(\,\cdot\,|u,v)\}_{i,v\in\mathcal{L}})\,, \tag{G.4}$$

where in the last line $\hat{r}$ is the empirical distribution of the row latents and $\hat{c}$ is the empirical distribution of the column latents. By taking expectation in the last expression, we get

$$\sum_{u,v\in\mathcal{L}}\frac{\mathbb{E}[\mathsf{len}(\mathsf{Z}(\boldsymbol{X}(u,v)))|\boldsymbol{u},\boldsymbol{v}]}{mn\log_2|\mathcal{X}|} \leq \frac{H(X|U,V)}{\log_2|\mathcal{X}|} + |\mathcal{L}|^2\Delta_{\mathsf{Z}}(c\cdot mn;\{Q(\,\cdot\,|u,v)\}_{u,v\in\mathcal{L}})\,. \tag{G.5}$$

Finally, the header contains $|\mathcal{L}|^2+2$ integers of maximum size mn, whence $\mathsf{len}(\mathsf{header}) \le 4\log_2(mn)$. We conclude that

$$\begin{aligned}\mathsf{R}_{\#} \le & \frac{1}{\log_2|\mathcal{X}|}\Big\{H(X|U,V)+\frac{1}{n}H(U)+\frac{1}{n}H(V)\Big\}+\frac{2\log_2(mn)}{mn}\\ & +|\mathcal{L}|^2\Delta_{\mathsf{Z}}(c\cdot mn;\{Q(\cdot\,|u,v)\}_{u,v\in\mathcal{L}})+2\Delta_{\mathsf{Z}}(m\wedge n;\{r,c\})\,.\end{aligned}$$

The claim (5.8) follows from the first bound in Eq. (5.2) noticing that, under the stated assumptions on m,n,

$$\frac{1}{n}\big[\mathsf{h}(\varepsilon_U)+\varepsilon_u\log(|\mathcal{L}|-1)\big] \le \varepsilon_U \le \mathbb{P}\big(\mathsf{Err}_U(\boldsymbol{X}^{m,n};\widehat{\boldsymbol{u}})<1\big)\,. \tag{G.6}$$

G.1.2. Redundancy bounds for specific encoders: Eqs. (5.9)–(5.11)

LZ coding. Let $\boldsymbol{X}^N=(X_1,\dots,X_N)$ be a vector with i.i.d. symbols $X_i\sim q$ with q a probability distribution over $\mathcal{X}$. The analysis is similar to the one in Appendix F, and we will adopt the same notations here. There are two important differences: data are i.i.d. (not matrix-structured) and we want to derive a sharper estimate (not just the entropy term, but bounding the overhead as well).

We define $L_k(\boldsymbol{X}^N)$, $T_k(\boldsymbol{X}^N)$ as per Eqs. (F.1), (F.2). We let $(k(1),\dots,k(M_N))$ be the values taken by k in the while loop of the Lempel-Ziv pseudocode of Section 5.2.2. In particular

$$k(1)=1\,, \tag{G.7}$$
$$k(\ell+1)=k(\ell)+L_{k(\ell)}(\boldsymbol{X}^N)\,, \tag{G.8}$$
$$k(M_N)=N\,. \tag{G.9}$$

(We set $k(0)=0$ by convention.) Therefore the total length of the code is

$$\begin{aligned}\mathsf{len}(\mathsf{LZ}(\boldsymbol{X}^N)) &= M_N\lceil\log_2(N+|\mathcal{X}|)\rceil+\sum_{\ell=1}^{M_N}\mathsf{len}(\mathsf{elias}(L_{k(\ell)}))\\ &\le M_N\lceil\log_2(N+|\mathcal{X}|)\rceil+2\sum_{\ell=1}^{M_N}\log_2(L_{k(\ell)})\\ &\le M_N\lceil\log_2(N+|\mathcal{X}|)\rceil+2M_N\log_2(N/M_N)\,,\end{aligned}$$

where the last step follows by Jensen's inequality. By one more application of Jensen, we obtain

$$\mathbb{E}\,\mathsf{R}_{\mathsf{LZ}}(\boldsymbol{X}^N) \le \frac{1}{\log_2|\mathcal{X}|}\cdot\frac{\mathbb{E}M_N}{N}\cdot\big\{\lceil\log_2(N+|\mathcal{X}|)\rceil+2\log_2(N/\mathbb{E}M_N)\big\}\,. \tag{G.10}$$

Define the set of break points and bad positions as

$$S_N:=\big\{k(1),k(2),\dots,k(M_N)\big\}\,, \tag{G.11}$$
$$B_N(\ell):=\big\{k\in[N/2,N]:\ L_k(\boldsymbol{X}^N)<\ell\big\}\,. \tag{G.12}$$

Note that $S_N=S_N^{\le}\cup S_N^{>}$ where:

$$S_N^{\le}:=\Big\{k(j):\ j\le M_N, k(j-1)\le\lfloor N/2\rfloor\Big\}\,,\qquad S_N^{>}\qquad :=\Big\{k(j):\ j\le M_N, k(j-1)>\lfloor N/2\rfloor\Big\}\,. \tag{G.13}$$

Further $|S_N^{\le}|\stackrel{\mathrm{d}}{=}M_{\lfloor N/2\rfloor}$ and, for any $\ell\in\mathbb{N}$,

$$\frac{N}{2}\ge\sum_{k\in S_N^{>}}L_k\ge\big(|S_N^{>}|-|B_N(\ell)|\big)\ell\,. \tag{G.14}$$

Therefore,

$$\begin{aligned}\mathbb{E}\,|S_N^{>}| &\le \frac{N}{2\ell} + \mathbb{E}\,|B_N(\ell)| \\ &\le \frac{N}{2\ell} + \sum_{k=\lceil N/2\rceil}^{N} \mathbb{P}\big(L_k(\boldsymbol{X}^N) < \ell\big)\,. \end{aligned} \tag{G.15}$$

We claim that this implies, for $C_0 = 20c_*\log|\mathcal{X}|$ and $\log N \ge (2\log(2/H(q)))^2$,

$$\frac{1}{N}\mathbb{E}\,|S_N^{>}| \le \frac{H(q)}{2\log_2 N} + C_0\frac{(\log\log_2 N)^{1/2}}{(\log_2 N)^{3/2}} =: \psi(\log_2 N)\,. \tag{G.16}$$

Before proving this claim, let us show that it implies the thesis. Recall that $M_N = |S_N|$ and $S_N = S_N^{\le} \cup S_N^{>}$ where $|S_N^{\le}| \stackrel{\mathrm{d}}{=} M_{\lfloor N/2\rfloor}$. Therefore, we have proven

$$\begin{aligned}\mathbb{E}M_N &\le N\psi(\log_2 N) + \mathbb{E}M_{\lfloor N/2\rfloor} \\ &\le \sum_{\ell=0}^{K-1} N_\ell\psi(\log_2 N_\ell) + \mathbb{E}M_{N_K}\,, \end{aligned} \tag{G.17}$$

where we defined recursively $N_0 = N$, $N_{\ell+1} = \lfloor N/2\rfloor$, and $K := \min\{\ell : \log_2 N_\ell < (2\log(2/H(q)))^2\}$. Of course, $M_{N_K} \le N_K \le \exp((2\log 2/H(q))^2)$. Further $\underline{N}_\ell \le N_\ell \le \overline{N}_\ell$, where $\underline{N}_0 = \overline{N}_0 = N$ and $\overline{N}_{\ell+1} = \overline{N}_\ell/2$, $\underline{N}_{\ell+1} = (\underline{N}_\ell - 1)/2$ for $\ell \ge 0$. We thus get $\underline{N}_\ell = (N+1)2^{-\ell} - 1$, $\overline{N}_\ell = N\,2^{-\ell}$ and therefore

$$\begin{aligned}\frac{1}{N}\sum_{\ell=0}^{K-1} N_\ell\psi(\log_2 N_\ell) &\le \frac{1}{N}\sum_{\ell=0}^{\infty}\overline{N}_\ell\psi(\log_2\underline{N}_\ell) \\ &\le \frac{H(q)}{2}\sum_{\ell=0}^{\infty}2^{-\ell}\frac{1}{\log_2(N2^{-\ell}-1)} + C_0\sum_{\ell=0}^{\infty}2^{-\ell}\frac{(\log\log_2 N)^{1/2}}{(\log_2(N2^{-\ell}-1))^{3/2}} \\ &\le \frac{H(q)}{\log_2 N} + 2C_0\frac{(\log\log_2 N)^{1/2}}{(\log_2 N)^{3/2}}\,. \end{aligned}$$

Substituting in Eq. (G.17), we get

$$\begin{aligned}\frac{1}{N}\mathbb{E}M_N &\le \frac{H(q)}{\log_2 N} + 2C_0\frac{(\log\log_2 N)^{1/2}}{(\log_2 N)^{3/2}} + \frac{1}{N}\exp\left\{\big(2\log(2/H(q))\big)^2\right\} \\ &\le \frac{H(q)}{\log_2 N} + 3C_0\frac{(\log\log_2 N)^{1/2}}{(\log_2 N)^{3/2}}\,, \end{aligned}$$

where the last inequality follows for $N \ge \exp\left\{\big(4\log(2/H(q))\big)^2\right\}$ (noting that $C_0 > 1$). Finally, the desired bound (5.9) follows by substituting the last estimate in Eq. (G.10).

We are left with the task of proving claim (G.16). Fix any k, $\lceil N/2\rceil \le k \le N$ and write q^ℓ for the product distribution $q \times \cdots \times q$ (ℓ times). Setting $H = H_{\text{nats}}(q)$ (measuring here entropy in nats), for any $\delta > 0$,

$$\begin{aligned}\mathbb{P}\big(L_k(\boldsymbol{X}^N) < \ell\big) &= \mathbb{P}\Big(\boldsymbol{X}_i^{i+\ell-1} \ne \boldsymbol{X}_k^{k+\ell-1}\ \forall i < k\Big) \qquad \text{(G.18)} \\ &\le \sum_{\boldsymbol{x}^\ell\in\mathcal{X}^\ell}\mathbb{P}\big(Z_k(\boldsymbol{x}^\ell) = 0\big)\cdot\mathbb{P}\big(\boldsymbol{X}_k^{k+\ell-1} = \boldsymbol{x}^\ell\big) \\ &\le \sum_{\boldsymbol{x}^\ell\in\mathcal{X}^\ell} q^\ell(\boldsymbol{x}^\ell)\big(1 - q^\ell(\boldsymbol{x}^\ell)\big)^{N/2\ell} \\ &\le \sum_{\boldsymbol{x}^\ell\in\mathcal{X}^\ell} q^\ell(\boldsymbol{x}^\ell)\,\mathbf{1}\big(q^\ell(\boldsymbol{x}^\ell) \le e^{-\ell[H+\delta]}\big) + \exp\Big\{-\frac{N}{2\ell}\cdot e^{-\ell[H+\delta]}\Big\} \\ &=: P_{\le}(\ell;\delta) + P_{>}(\ell,N;\delta)\,. \qquad \text{(G.19)} \end{aligned}$$

By Chernoff bound

$$P_{\le}(\ell;\delta) \le e^{-\ell \max_{\lambda>0}\psi_\delta(\lambda)}\,,$$
$$\psi_\delta(\lambda) := \lambda[H+\delta] - \log\Big\{\sum_{x\in\mathcal{X}} q(x)^{1-\lambda}\Big\}\,.$$

Note that $\lambda \mapsto \psi_\delta(\lambda)$ is continuous, concave, with $\psi_\delta'(0)=\delta$, $\psi_\delta(0)=0$, $\psi_\delta(1)=\delta+H-\log|\mathcal{X}|$. Hence (assuming $H<\log|\mathcal{X}|$ because otherwise there is nothing to prove) for all δ small enough ψ is maximized for $\lambda\in(0,1)$. Further, defining the random variable $Q=q(x)$ for $x\sim\mathsf{Unif}(\mathcal{X})$,

$$\psi_\delta''(\lambda) = -\frac{\mathbb{E}[Q^{1-\lambda}(\log Q)]}{\mathbb{E}[Q^{1-\lambda}]} + \Big[\frac{\mathbb{E}[Q^{1-\lambda}(\log Q)^2]}{\mathbb{E}[Q^{1-\lambda}]}\Big]^2 \tag{G.20}$$
$$\ge -\frac{\mathbb{E}[Q^{1-\lambda}(\log Q)^2]}{\mathbb{E}[Q^{1-\lambda}]} \tag{G.21}$$
$$\ge -\mathbb{E}[(\log Q)^2] =: -\underline{c}_*(q)\,. \tag{G.22}$$

Here the last inequality holds because $Q\mapsto Q^{1-\lambda}$ is monotone increasing (for $\lambda\in[0,1]$) and $Q\mapsto(\log Q)^2$ is monotone decreasing over $Q\in[0,1]$, and therefore $\mathbb{E}[Q^{1-\lambda}(\log Q)^2]\le\mathbb{E}[Q^{1-\lambda}]\,\mathbb{E}[(\log Q)^2]$. In what follows, we set $c_*:=\underline{c}_*(q)\wedge 1$. Hence $\psi_\delta(\lambda)\ge\delta\lambda - c_*\lambda^2/2$ for $\lambda\in[0,1]$ and therefore using Eq. (G.20),

$$P_{\le}(\ell;\delta) \le \exp\Big\{-\ell\min\Big(\frac{\delta^2}{2c_*};\delta-\frac{c_*}{2}\Big)\Big\}\,.$$

Substituting in Eq. (G.19), and using this in Eq. (G.15), we get, for $\delta\in[0,c_*]$:

$$\frac{1}{N}\mathbb{E}|S_N^{>}| \le \frac{1}{2\ell} + \exp\Big\{-\frac{\ell\delta^2}{2c_*}\Big\} + \exp\Big\{-\frac{N}{2\ell}\cdot e^{-\ell[H+\delta]}\Big\}$$

We set

$$\ell = \frac{\log N}{H}(1-\varepsilon)\,,\qquad \delta = \frac{1}{2}H\varepsilon\,,$$

for $\varepsilon\le(2c_*/H)\wedge(1/2)$. Substituting in the previous bound, we get

$$\frac{1}{N}\mathbb{E}|S_N^{>}| \le \frac{H}{2\log N}(1+2\varepsilon) + \exp\Big\{-\frac{H\varepsilon^2\log N}{16c_*}\Big\} + \exp\Big\{-\frac{HN^{(\varepsilon+\varepsilon^2)/2}}{2\log N}\Big\}\,.$$

We finally select $\varepsilon = c_0(c_*\log|\mathcal{X}|/H)(\log\log N/\log N)^{1/2}$, with c_0 a sufficiently small absolute constant. Substituting above,

$$\begin{aligned}\frac{1}{N}\mathbb{E}|S_N^{>}| - \frac{H}{2\log N} &\le c_0\,c_*\,\log|\mathcal{X}|\frac{(\log\log N)^{1/2}}{(\log N)^{3/2}} + \exp\Big\{-\frac{c_0^2(\log|\mathcal{X}|)^2c_*}{16H}\log\log N\Big\}\\ &\quad + \exp\Big\{-\frac{H}{2\log N}e^{(c_0c_*\log|\mathcal{X}|/2H)\sqrt{\log N}}\Big\}\\ &\le c_0\,c_*\,\log|\mathcal{X}|\frac{(\log\log N)^{1/2}}{(\log N)^{3/2}} + (\log N)^{-c_0^2/16} + \exp\Big\{-\frac{H}{2\log N}e^{(c_0/2)\sqrt{\log N}}\Big\}\,.\end{aligned}$$

Setting $c_0=5$, we get

$$\frac{1}{N}\mathbb{E}|S_N^{>}| - \frac{H}{2\log N} \le 6\,c_*\,\log|\mathcal{X}|\frac{(\log\log N)^{1/2}}{(\log N)^{3/2}} + \exp\Big\{-\frac{H}{2\log N}e^{2\sqrt{\log N}}\Big\}\,,$$

whence, the claim (G.16) follows for $\log N\ge(\log(2/H))^2$.

Arithmetic Coding. In Arithmetic Coding (AC) we encode the empirical distribution of $\boldsymbol{X}^N$ $\widehat{q}_N(x) := N^{-1}\sum_{i\le N}\mathbf{1}_{X_i=x}$, and then encode $\boldsymbol{X}^N$ in at most $-\log_2\widehat{q}^N(\boldsymbol{X}^N)+1$ bits. The encoding of $\widehat{q}_N$ amounts to encoding the $|\mathcal{X}|-1$ integers $N\widehat{q}_N(x)$, $x\in\mathcal{X}\setminus\{0\}$ (assuming that $0\in\mathcal{X}$, one of the counts can be obtained by difference). We thus have

$$\begin{aligned}\mathsf{len}(\mathsf{AC}(\boldsymbol{X}^N)) &\le -\log_2\widehat{q}_N^{\otimes N}(\boldsymbol{X}^N)+1+\sum_{x\in\mathcal{X}}\mathsf{len}(\mathsf{elias}(N\widehat{q}_N(x)))\\ &\le -\log_2\widehat{q}_N^{\otimes N}(\boldsymbol{X}^N)+2|\mathcal{X}|\log_2 N\\ &=\sum_{i=1}^N -\log_2\widehat{q}_N(X_i)+2|\mathcal{X}|\log_2 N\\ &= N\,H(\widehat{q}_N)+2|\mathcal{X}|\log_2 N\,.\end{aligned}$$

Taking expectations

$$\begin{aligned}\mathbb{E}\mathsf{R}_{\mathsf{AC}}(\boldsymbol{X}^N) &\le \frac{\mathbb{E}\,H(\widehat{q}_N)}{\log_2|\mathcal{X}|}+\frac{2|\mathcal{X}|\log_2 N}{N\log_2|\mathcal{X}|}\\ &\le \frac{H(q)}{\log_2|\mathcal{X}|}+\frac{2|\mathcal{X}|\log_2 N}{N\log_2|\mathcal{X}|}\,.\end{aligned}$$

ANS Coding. The bound (5.11) follows for range ANS coding from the analysis of (Duda, 2009; 2013; Kosolobov, 2022), where encoding of empirical distributions are analyzed as for AC coding.

G.2. Proof of Theorem 5.8

The proof consists in applying Lemma 5.7 and showing that $\mathbb{P}\Big(\mathsf{Err}_U(\boldsymbol{X}^{m,n};\widehat{\boldsymbol{u}})>0\Big)\le\log(mn)/mn$, $\mathbb{P}\Big(\mathsf{Err}_V(\boldsymbol{X}^{m,n};\widehat{\boldsymbol{v}})>0\Big)\le\log(mn)/mn$.

In what follows we will assume without loss of generality $m\le n$, and recall that $|\mathcal{L}|=k$, identifying $\mathcal{L}=\{1,\dots,k\}$. We will assume k fixed. We will use $C,c,c',\dots$ for constants that might depend on k, as well as the constant c_0 in the statement in ways that we do ot track.

We will show that these bounds hold conditional on $\boldsymbol{u}$, $\boldsymbol{v}$, on the events $\min_u\hat{q}_{\mathsf{r}}(u),\min_v\hat{q}_{\mathsf{r}}(v)\ge c/2$ which holds with probability at least $1-\exp(-c'm)\ge 1-\log(mn)/mn$. Hence, hereafter we will treat $\boldsymbol{u},\boldsymbol{v}$ as deterministic. Recall that $\boldsymbol{M}\in\mathbb{R}^{m\times n}$ is the matrix entries with

$$M_{ij}=\psi(X_{i,j})\,,$$

and let $\boldsymbol{M}_*=\mathbb{E}\{\boldsymbol{M}\}$. We collect a few facts about $\boldsymbol{M}$ and its expectation.

Singular values. Note that $\boldsymbol{M}_*$ takes the form

$$\boldsymbol{M}_*=\boldsymbol{L}\boldsymbol{\Psi}\boldsymbol{R}^{\mathsf{T}}\,, \tag{G.23}$$

where $\boldsymbol{\Psi}\in\mathbb{R}^{r\times r}$ is a matrix with entries $\boldsymbol{\Psi}_{u,v}=\overline{\psi}(u,v)$, $\boldsymbol{L}\in\{0,1\}^{m\times r}$, with $L_{ij}=1\Leftrightarrow u_i=j$, and $\boldsymbol{R}\in\{0,1\}^{n\times r}$, with $R_{ij}=1\Leftrightarrow v_i=j$. Define $\boldsymbol{L}=\boldsymbol{L}_0\boldsymbol{D}_L^{1/2}$ where $\boldsymbol{D}_L$ is a diagonal matrix with $(\boldsymbol{D}_L)_{ii}=m\hat{q}_{\mathsf{r}}(i)$, and analogously $\boldsymbol{R}=\boldsymbol{R}_0\boldsymbol{D}_L^{1/2}$, and introduce the singular value decomposition $\boldsymbol{D}_L^{1/2}\boldsymbol{\Psi}\boldsymbol{D}_R^{1/2}=\overline{\boldsymbol{A}}\boldsymbol{\Sigma}\overline{\boldsymbol{B}}^{\mathsf{T}}$. We then have the singular value decomposition

$$\boldsymbol{M}_*=\boldsymbol{A}_*\boldsymbol{\Sigma}\boldsymbol{B}_*^{\mathsf{T}}\,,\quad \boldsymbol{A}_*=\boldsymbol{L}_0\overline{\boldsymbol{A}}\,,\quad \boldsymbol{B}_*=\boldsymbol{R}_0\overline{\boldsymbol{B}}\,. \tag{G.24}$$

Therefore $\sigma_k(\boldsymbol{M}_*)\ge\sigma_{\min}(\boldsymbol{D}_L)^{1/2}\sigma_{\min}(\boldsymbol{\Psi})\sigma_{\min}(\boldsymbol{D}_R)^{1/2}$ (here and below σ_k denotes the k-th largest singular value) and using the assumptions on $\hat{q}_{\mathsf{r}},\hat{q}_{\mathsf{c}}$,

$$\sigma_k(\boldsymbol{M}_*)\ge c\mu\sqrt{mn}\,. \tag{G.25}$$

Concentration. $\boldsymbol{M} - \boldsymbol{M}_*$ is a centered matrix with independent entries with variance bounded by σ^2 and entries bounded by 1 (by the assumption $|\psi(x)| \le 1$). By matrix Bernstein inequality there exists a universal constant C such that the following holds with probability at least $1 - (100n)^{-2}$:

$$\|\boldsymbol{M} - \boldsymbol{M}_*\|_{\text{op}} \le C \max\left(\sigma\sqrt{n \log n}; \log n\right). \tag{G.26}$$

Incoherence. Since all the entries of $\boldsymbol{M}_*$ are bounded by 1, we get

$$\|\boldsymbol{M}_*\|_{2\to\infty} \vee \|\boldsymbol{M}_*^{\mathsf{T}}\|_{2\to\infty} \le \nu\sqrt{n}\,. \tag{G.27}$$

Row concentration. For any $i \le m$, and any $\boldsymbol{W} \in \mathbb{R}^{n\times k}$ fixed, with probability at least $1 - (100n)^{-5}$:

$$\|(\boldsymbol{M} - bM_*)_{i,\cdot}\boldsymbol{W}\|_2 \le C \max\left(\sigma\|\boldsymbol{W}\|_F\sqrt{\log n}; \|\boldsymbol{W}\|_{2\to\infty}\log n\right).$$

Defining $\Delta_* := \sigma_k(\boldsymbol{M}_*) \ge c\mu\sqrt{mn}$, this implies

$$\|(\boldsymbol{M} - \boldsymbol{M}_*)_{i,\cdot}\boldsymbol{W}\|_2 \le \Delta_*\|\boldsymbol{W}\|_{2\to\infty}\varphi\left(\frac{\|\boldsymbol{W}\|_F}{\sqrt{n}\|\boldsymbol{W}\|_{2\to\infty}}\right),$$
$$\varphi(x) := \frac{C}{\mu\sqrt{mn}}\max\left(x\sigma\sqrt{n\log n}; \log n\right).$$

Given these, we apply (Abbe et al., 2020)[Corollary 2.1], with the following estimates of various parameters (see (Abbe et al., 2020) for definitions):

$$\begin{aligned}
\Delta_* &\asymp \mu\sqrt{mn}\,,\\
\gamma &\lesssim \max\left(\frac{\sigma}{\mu}\sqrt{\frac{\log n}{m}}; \frac{\log n}{\mu\sqrt{mn}}\right) \lesssim \frac{\sigma}{\mu}\sqrt{\frac{\log n}{m}}\,,\\
\|\boldsymbol{M}_*\|_{2\to\infty} \vee \|\boldsymbol{M}_*^{\mathsf{T}}\|_{2\to\infty} &\le \nu\sqrt{n} \lesssim \gamma\Delta_*\,,\\
\varphi(1) &\lesssim \frac{\sigma}{\mu}\sqrt{\frac{\log n}{m}}\,,\\
\varphi(\gamma) &\lesssim \max\left(\frac{\sigma^2}{\mu^2}\frac{\log n}{m}; \frac{\log n}{\mu\sqrt{mn}}\right),\\
\kappa &\asymp 1\,.
\end{aligned}$$

Then (Abbe et al., 2020)[Corollary 2.1] implies that there exists a $k \times k$ orthogonal matrix $\tilde{\boldsymbol{Q}}$ such that

$$\frac{\|\boldsymbol{A} - \boldsymbol{A}_*\tilde{\boldsymbol{Q}}\|_{2\to\infty}}{\|\boldsymbol{A}_*\|_{2\to\infty}} \lesssim (1+\varphi(1))(\gamma + \varphi(\gamma)) + \varphi(1) \tag{G.28}$$
$$\lesssim \frac{\sigma}{\mu}\sqrt{\frac{\log m}{n}}\,. \tag{G.29}$$

Recall that $\boldsymbol{A}_* = \boldsymbol{L}_0\overline{\boldsymbol{A}}$ and $\overline{\boldsymbol{A}}$ is an othogonal matrix. Therefore, there exists an orthogonal matrix $\boldsymbol{Q}$ such that (with the desired probability):

$$\frac{\|\boldsymbol{A} - \boldsymbol{L}_0\boldsymbol{Q}\|_{2\to\infty}}{\|\boldsymbol{L}_0\|_{2\to\infty}} \lesssim \frac{\sigma}{\mu}\sqrt{\frac{\log m}{n}}\,. \tag{G.30}$$

Further, the i-th row of $\boldsymbol{L}_0$ is

$$(\boldsymbol{L}_0)_{i,\cdot} = \frac{1}{\sqrt{m\hat{q}_{\mathrm{r}}(u_i)}}\boldsymbol{e}_{u_i}^{\mathsf{T}} =: z(u_i)\boldsymbol{e}_{u_i}^{\mathsf{T}}\,. \tag{G.31}$$

Hence, for any i, $\sqrt{c_0}\|\boldsymbol{L}_0\|_{2\to\infty} \le \|(\boldsymbol{L}_0)_{i,\cdot}\| \le \|\boldsymbol{L}_0\|_{2\to\infty}$. Denoting by $\boldsymbol{q}_j$ the j-th row of $\boldsymbol{Q}$, we thus get for all i,

$$\frac{\|\boldsymbol{a}_i - z(u_i)\boldsymbol{q}_{u_i}\|_2}{\|\boldsymbol{a}_i\|_2} \lesssim \frac{\sigma}{\mu}\sqrt{\frac{\log m}{n}}\,. \tag{G.32}$$

The claim follows immediately using the fact that $\sqrt{c_0}\max_u z(u) \le \min_u z(u) \le \max_u z(u)$.