

Subset Source Coding

Ebrahim MolavianJazi and Aylin Yener

Wireless Communications and Networking Laboratory (WCAN)
 Department of Electrical Engineering
 The Pennsylvania State University, University Park, PA 16802
eum20@psu.edu yener@enr.psu.edu

Abstract—Emerging applications including semantic information processing impose priorities on the possible realizations of information sources, so that not all source sequences are important. This paper proposes an initial framework for optimal lossless compression of subsets of the output of a discrete memoryless source (DMS). It turns out that, the optimal source code may not index the conventional source-typical sequences, but rather index certain *subset-typical* sequences determined by the source statistics as well as the subset structure. Building upon an achievability and a strong converse, an analytic expression is given, based on the Shannon entropy, relative entropy, and *subset entropy*, which identifies such subset-typical sequences for a broad class of subsets of a DMS. Interestingly, one often achieves a gain in the fundamental limit, in that the optimal compression rate for the subset can be strictly smaller than the source entropy, although this is not always the case.

I. INTRODUCTION

Source coding addresses the compression, with or without fidelity, of an information source. In particular, in lossless compression of a discrete memoryless source (DMS), one identifies and indexes the set of *typical* sequences that capture essentially all of the probability mass of the entire source. For a DMS with probability distribution $P(x)$ over alphabet \mathcal{X} , the number of such typical sequences is approximately $2^{n\mathbb{H}(P)}$, so that a lossless compression of the DMS can be achieved with a rate $\mathbb{H}(P)$ bits per source symbol [1], where $\mathbb{H}(P) = -\sum_{x \in \mathcal{X}} P(x) \log P(x)$ denotes the (Shannon) entropy of the source.¹ That is, lossless source codes exist with rates above the source entropy, $R > \mathbb{H}(P)$, for which the error probability vanishes exponentially fast [2]:

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \Pr[\mathcal{E}^{(n)}] \leq - \min_{Q: \mathbb{H}(Q) \geq R} D(Q||P), \quad (1)$$

where $\Pr[\mathcal{E}^{(n)}] = \Pr[\hat{X}^n \neq X^n]$ is the error probability that the decoder's estimate is different from the source realization, and $D(Q||P) := \sum_{x \in \mathcal{X}} Q(x) \log(Q(x)/P(x))$ denotes the relative entropy. Further, strong converse holds for the lossless compression of a DMS, so that the error probability of any lossless source code with rate below the entropy, $R < \mathbb{H}(P)$, has an error floor [2]

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log(1 - \Pr[\mathcal{E}^{(n)}]) \leq - \min_{Q: \mathbb{H}(Q) \leq R} D(Q||P). \quad (2)$$

These basic settings have been extensively extended to scenarios with unknown statistics [3] and find applications

in database management [4]. A key underlying consideration in all of these works is that important realizations of the source only consist of likely sequences of the source, i.e., source-typical sequences.

In some emerging applications in data analytics and information processing including semantic communications, database management and bioinformatics, however, the likelihood and typicality of a source realization may not be the main factor to determine the importance of that sequence. In particular, in semantic communications [5], only certain pieces of information might be meaningful according to semantic and logic rules. In such scenarios, therefore, one is interested in processing and conveying only certain source outputs with potentially low probability, rather than capturing the collective probability mass of the source embodied in the source-typical sequences. Posed as a compression problem, the encoder and decoder aim at providing a lossless or lossy description of only a *subset* of all possible source realizations as determined by the application.

Our goal in this paper is to provide an initial treatment of such a *subset source coding* problem. In terms of motivation, our work is related to the problem of task encoding in [6] that guarantees certain important but less likely source events are not neglected in data compression, as well as information theory of atypical sequences in [7] with applications in signal processing and big-data analytics. Our investigation also has roots in large deviations theory and relates to generalized asymptotic equipartition property (AEP) [8] for lossless and lossy compression of subsets.

In Section II, we formally introduce the subset source coding problem. A key point in the problem formulation is selecting an appropriate definition of the error probability. Here, we adopt a conditional error probability, with respect to the total subset probability, as the metric. This definition, particularly the normalization involved in this conditional probability, plays a key role in the analysis and introduces technical subtleties. This renders a rigorous analysis necessary and the problem itself non-trivial, although the results are intuitively pleasing. Having this definition at hand, one might first imagine that subset source coding is simply a compression problem for an *equivalent conditional* source, whose distribution is given by the conditional distribution of the original source conditioned on the subset of interest. In Section III, we show that such an analysis, although valid, is not generally tractable using available tools, such as the

¹Throughout this paper, all log operations are understood as base 2.

information-spectrum approach [9]. Therefore, we resort to the original DMS and build upon the large deviations theory [1] and elements of combinatorics for the analysis.

In Section IV, we present three optimality results that apply to broad classes of potential subsets of a DMS: one for *likely* subsets based on the error exponent results for conventional source coding; one for *smooth* subsets using the method of types [2]; and finally one for *fluctuating* subsets via superimposing several smooth subsets. The proof of the second result is given in Section V. We next present in Section VI several numerical examples of the subset source coding problem and provide observations on the compression rate gain/loss that is achieved/incurred by focusing only on subsets, instead of the entire source space. An interesting observation is that, the subset-compression rate is the result of a tension between the source statistics and the subset structure; in some extreme cases, the compression rate can be even totally independent of the source statistics. Section VII concludes the paper with some remarks about the extension of our treatment to the case of lossy compression.

II. PROBLEM SETTING

Consider a discrete memoryless source with distribution $P_X(x)$ over the discrete finite alphabet \mathcal{X} , such that the n -fold distribution of the source, for all $n = 1, 2, \dots$, satisfies

$$P_{X^n}(x^n) = \prod_{t=1}^n P_X(x_t). \quad (3)$$

For simplicity of notation, we will sometimes write P_X as P . Let $\mathcal{L} = \{\mathcal{L}_n\}_{n=1}^\infty$ be a sequence of subsets of the source realizations such that $\mathcal{L}_n \subseteq \mathcal{X}^n$ and $\Pr[X^n \in \mathcal{L}_n] \neq 0$ for all n . The problem is to find the minimum lossless compression rate for the subset \mathcal{L} .

More formally, an $(n, 2^{nR})$ source code for subset \mathcal{L} consists of an encoder $m : \mathcal{L}_n \rightarrow \{1, 2, \dots, 2^{nR}\}$ and a decoder $\hat{x}^n : \{1, 2, \dots, 2^{nR}\} \rightarrow \mathcal{L}_n \cup \{\mathcal{E}\}$ that assigns to an index $1 \leq m \leq 2^{nR}$ either an estimate $\hat{x}^n(m) \in \mathcal{L}_n$ or an error \mathcal{E} . The error probability of the code is defined as

$$\Pr[\mathcal{E}^{(n)}] := \Pr[\hat{X}^n \neq X^n | X^n \in \mathcal{L}_n]. \quad (4)$$

A rate R is called achievable if an $(n, 2^{nR})$ source code for subset \mathcal{L} exists with $\Pr[\mathcal{E}^{(n)}] \rightarrow 0$ as $n \rightarrow \infty$. The optimal compression rate $R_{\mathcal{L}}^*$ is the infimum of all achievable rates.

III. AN INITIAL ATTEMPT VIA THE EQUIVALENT CONDITIONAL SOURCE

The way our problem of subset compression is posed, specifically the error probability definition (4), may spur the idea that a conditional setting readily captures this problem. In particular, one may define an equivalent conditional source \tilde{X}^n as

$$P_{\tilde{X}^n}(x^n) := \frac{P_{X^n}(x^n)}{P_{X^n}[X^n \in \mathcal{L}_n]} 1\{x^n \in \mathcal{L}_n\}, \quad (5)$$

and claim the fundamental lossless compression rate of this conditional source to be equivalent to our $R_{\mathcal{L}}^*$ of interest. This claim is indeed valid, since the error probability for both cases is readily shown to be the same.

The fundamental compression limits of this conditional source, however, are not in general very straightforward to analyze. One possibility is to use the information-spectrum approach [9] to characterize the fundamental compression limits of this potentially non-stationary and non-ergodic source. In particular, we have from the result of Han and Verdú [10] that the fundamental limit is given by the *spectral sup-entropy rate* of the equivalent conditional source:

$$R_{\mathcal{L}}^* = \bar{H}(\tilde{\mathbf{X}}) := \mathfrak{p}\text{-lim sup}_{n \rightarrow \infty} \frac{1}{n} \log \frac{1}{P_{\tilde{X}^n}(\tilde{X}^n)}, \quad (6)$$

where $\tilde{\mathbf{X}} = \{\tilde{X}^n\}_{n=1}^\infty$ is the equivalent conditional source process, and the $\mathfrak{p}\text{-lim sup}$ operation, *limit superior in probability*, is basically defined as the supremum of the support set of the limiting distribution [9]. For those subsets for which the equivalent conditional source is stationary and ergodic, the results (slightly) simplify to average entropy rate results [1], [9]:

$$R_{\mathcal{L}}^* = \lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{H}(\tilde{X}^n). \quad (7)$$

Although the above information-spectrum approach and limiting analysis yield a complete characterization of the fundamental compression limits, its numerical evaluation for arbitrary subsets is cumbersome, if at all possible, and may require tedious manipulations. Moreover, the effect of subset structure and the statistics of the original source on the fundamental limits are not quite explicit. In the remainder of this paper, we give three more tractable optimality results that apply to broad classes of general subsets.

IV. COMPRESSION RATE FOR GENERAL SUBSETS

In this section, we present three results that apply to general subsets.

A. Likely Subsets

The first general result asserts that, for subsets with not so small probability, the optimal compression rate is identical to that of the original source.

Theorem 1. *For a discrete memoryless source P and any subset $\mathcal{L} = \{\mathcal{L}_n\}_{n=1}^\infty$ whose probability $P_{X^n}[X^n \in \mathcal{L}_n]$ as $n \rightarrow \infty$ either is a constant or decays sub-exponentially to zero, the optimal subset compression rate is $R_{\mathcal{L}}^* = \mathbb{H}(P)$.*

Theorem 1 is more intuitive for subsets \mathcal{L} with an asymptotically constant probability, $P_{X^n}[X^n \in \mathcal{L}_n] \rightarrow c$ where $0 < c \leq 1$, since excluding any constant fraction of sequences in \mathcal{X}^n does not reduce the required compression rate. The case of subsets with slowly vanishing probability is somewhat more subtle, as explained in the following proof of Theorem 1.

Proof: (Achievability) Fix an arbitrary $\epsilon > 0$ and choose an error-exponent optimal source code in the conventional setting with rate $\mathbb{H}(P) + \epsilon$ and $\Pr[\hat{X}^n \neq X^n] \rightarrow 0$ exponentially fast as $n \rightarrow \infty$; cf. (1). Noting that

$\Pr[\hat{X}^n \neq X^n] \geq \Pr[X^n \in \mathcal{L}_n] \cdot \Pr[\hat{X}^n \neq X^n | X^n \in \mathcal{L}_n]$, and that by assumption $\Pr[X^n \in \mathcal{L}_n] \rightarrow c > 0$ or $\Pr[X^n \in \mathcal{L}_n] \rightarrow 0$ sub-exponentially, we conclude that the

same code, when constrained to only sequences within \mathcal{L}_n , achieves $\Pr[\hat{X}^n \neq X^n | X^n \in \mathcal{L}_n] \rightarrow 0$ as $n \rightarrow \infty$. This implies $R_{\mathcal{L}}^* \leq \mathbb{H}(P)$, as the choice of ϵ is arbitrary.

(Converse) Fix an arbitrary source code for the subset $\mathcal{L} = \{\mathcal{L}_n\}_{n=1}^{\infty}$ achieving the rate R with error probability $\Pr[\hat{X}^n \neq X^n | X^n \in \mathcal{L}_n] = \epsilon_n \rightarrow 0$ as $n \rightarrow \infty$. We can consider this code as a conventional source code for the entire space \mathcal{X}^n which maps all sequences in $(\mathcal{X}^n - \mathcal{L}_n)$ to an error \mathcal{E} . We can analyze the error probability as follows.

$$\begin{aligned} \Pr[\hat{X}^n \neq X^n] &= \Pr[X^n \in \mathcal{L}_n] \cdot \Pr[\hat{X}^n \neq X^n | X^n \in \mathcal{L}_n] \\ &\quad + \Pr[X^n \notin \mathcal{L}_n] \cdot \Pr[\hat{X}^n \neq X^n | X^n \notin \mathcal{L}_n] \\ &\leq \epsilon_n \cdot \Pr[X^n \in \mathcal{L}_n] + \Pr[X^n \notin \mathcal{L}_n] \\ &= 1 - (1 - \epsilon_n) \cdot \Pr[X^n \in \mathcal{L}_n]. \end{aligned} \quad (8)$$

Since $\Pr[X^n \in \mathcal{L}_n] \rightarrow c > 0$ or $\Pr[X^n \in \mathcal{L}_n] \rightarrow 0$ sub-exponentially with n , the error probability of this code is at least sub-exponentially away from 1. By the error exponent result (2) for conventional source coding, this implies that the rate R is above the source entropy $\mathbb{H}(P)$. Since the choice of the code is arbitrary, we have $R_{\mathcal{L}}^* \geq \mathbb{H}(P)$.² ■

Theorem 1 immediately captures a large class of subsets by asserting that only subsets with exponentially small probability need further study. In fact, one might imagine that all important subsets with non-negligible probability are already addressed by this theorem and therefore the remaining possible subsets with exponentially small probability are so rare that their analysis may seem irrelevant. In particular, one may assume that such subsets with negligible probability only contain the *atypical* sequences of the source [1], which are anyways ignored in conventional compression. However, as will be further clarified in the remainder of the paper including the numerical examples of Section VI, one can find subsets containing many source-typical sequences, which yet have an exponentially small probability. Moreover, as discussed in the Introduction, even the atypical sequences of the source may be important for certain applications.

B. Smooth Subsets

Our second result provides an analytic formula for the optimal compression rate of a broad class of *smooth* subsets, including even those with exponentially small probability. To state this result, we first recall the definition of type classes and typical sequences of a DMS and introduce a new quantity called the *subset entropy*.

Definition 1. [2] Let $N(x; x^n)$ be the number of occurrences of the symbol $x \in \mathcal{X}$ in the sequence x^n . The type of a sequence x^n is the empirical distribution $\hat{P}_{x^n}(x)$ defined as

$$\hat{P}_{x^n}(x) := \frac{1}{n}N(x; x^n), \quad \forall x \in \mathcal{X}. \quad (9)$$

Accordingly, the set of all sequences in \mathcal{X}^n with type \hat{P} is denoted by $T^n(\hat{P})$ and called the type class of \hat{P} . Furthermore, given any general distribution Q and any positive sequence δ_n , the set $T^n[Q]_{\delta_n}$ of Q -typical sequences is defined as the union of type classes $T^n(\hat{P})$ for those types \hat{P} in \mathcal{X}^n which

satisfy $|\hat{P}(x) - Q(x)| \leq \delta_n$ for all $x \in \mathcal{X}$ with $Q(x) > 0$ and $\hat{P}(x) = 0$ otherwise. Here and throughout, the sequence δ_n is assumed to satisfy the Delta-Convention i.e., as $n \rightarrow \infty$, we have $\delta_n \rightarrow 0$ and $\sqrt{n}\delta_n \rightarrow \infty$.

One recalls from the *method of types* [2] that, the number of the distinct types in \mathcal{X}^n does not exceed $(n+1)^{|\mathcal{X}|}$, a result referred to as the Type Counting Lemma. Furthermore, for every distribution Q , the size of the Q -typical set satisfies

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log |T^n[Q]_{\delta_n}| = \mathbb{H}(Q). \quad (10)$$

We can now define the notion of subset entropy.

Definition 2. We say the subset $\mathcal{L} = \{\mathcal{L}_n\}_{n=1}^{\infty}$ intersects a distribution $Q(x)$ and write $\mathcal{L} \cap T[Q] \neq \emptyset$ if

$$\limsup_{n \rightarrow \infty} |\mathcal{L}_n \cap T^n[Q]_{\delta_n}| \neq 0. \quad (11)$$

In such a case, we define the subset- \mathcal{L} entropy of distribution $Q(x)$ as

$$H_{\mathcal{L}}(Q) := \lim_{n \rightarrow \infty} \frac{1}{n} \log |\mathcal{L}_n \cap T^n[Q]_{\delta_n}|, \quad (12)$$

if the above limit exists.

Comparing expressions (10) and (12) suggests that, the subset entropy $H_{\mathcal{L}}(Q)$ is a dual of the conventional entropy $\mathbb{H}(Q)$. In fact, for any distribution Q with $\mathcal{L} \cap T[Q] \neq \emptyset$ for which the subset entropy is well-defined, we can readily observe the appealing property that $0 \leq H_{\mathcal{L}}(Q) \leq \mathbb{H}(Q)$. In particular, for the extreme case of $\mathcal{L}_n = \mathcal{X}^n$, we have $H_{\mathcal{L}}(Q) = \mathbb{H}(Q)$ for all distributions Q .

We are now ready to express our second result whose proof is relegated to Section V.

Theorem 2. For a discrete memoryless source P and any subset $\mathcal{L} = \{\mathcal{L}_n\}_{n=1}^{\infty}$, if the subset entropy $H_{\mathcal{L}}(Q)$ exists and is continuous in all distributions Q intersecting the subset, $\mathcal{L} \cap T[Q] \neq \emptyset$, the optimal compression rate for the subset \mathcal{L} is

$$\begin{aligned} R_{\mathcal{L}}^* &= \max \left\{ H_{\mathcal{L}}(Q^*) : \mathcal{L} \cap T[Q^*] \neq \emptyset, \right. \\ &\quad \left. Q^* \in \arg \min_{Q: \mathcal{L} \cap T[Q] \neq \emptyset} [\mathbb{H}(Q) - H_{\mathcal{L}}(Q) + D(Q||P)] \right\}. \end{aligned}$$

Theorem 2 has an interesting interpretation in terms of a tension between the source statistics and the subset structure. It suggests that, for a given subset, the most likely sequences of the source which should be indexed by a lossless source code belong, not necessarily to the source-typical set with distribution P , but to a typical set (i) whose distribution Q is potentially close to the source statistics in the sense of relative entropy so that the term $D(Q||P)$ is relatively small and (ii) with potentially large intersection with the subset so that the size of its residual part outside the subset, captured by the term $(\mathbb{H}(Q) - H_{\mathcal{L}}(Q))$, is also rather small. The *subset-typical* distributions Q^* optimize the trade-off between these two elements by minimizing the function

$$g_P(Q) = \mathbb{H}(Q) - H_{\mathcal{L}}(Q) + D(Q||P), \quad (13)$$

²An alternative proof of converse follows from [2, Lemma 2.14 and Problem 2.11].

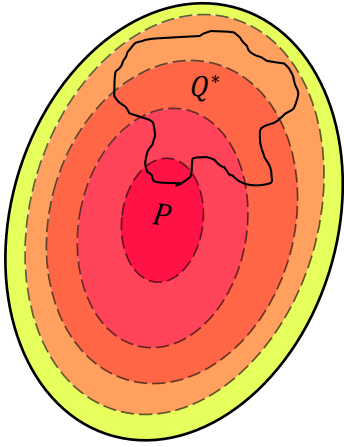


Fig. 1. Schematic description of Theorem 2 for smooth subsets. The subset is depicted with a bizarre curly shape. The dashed rings denote the typical sets, which are shown in the order of closeness to the source statistic P in the sense of relative entropy. The subset-typical distribution Q^* corresponds to the typical set that highly intersects the subset but is also close to the source statistic P .

and the size of the corresponding subset-typical sets determines $\max H_{\mathcal{L}}(Q^*)$ to be the rate of the lossless source code for this subset. This interpretation is schematically depicted in Figure 1.

The conditions mentioned in Theorem 2 are to guarantee that the subset sequence \mathcal{L} is smooth. We will further clarify this through several examples in Section VI. For now, as a sanity check, note for the extreme case of $\mathcal{L}_n = \mathcal{X}^n$ that, since the subset \mathcal{L} intersects all distributions Q and $H_{\mathcal{L}}(Q) = \mathbb{H}(Q)$, our objective function of interest reduces to $g_P(Q) = D(Q||P)$ which is minimized by $Q^* = P$ for which $H_{\mathcal{L}}(Q^*) = \mathbb{H}(Q^*) = \mathbb{H}(P)$, which is consistent with our first impression. Of course, one could arrive at the same result via Theorem 1, since for this case $P_{X^n}[X^n \in \mathcal{L}_n] = 1$ for all n .

An interesting special case is one in which the subset fully intersect a continuous spectrum of distributions. In this case, the subset must contain all sequences of a certain range of typical sets. Since all sequences within a typical set can be decomposed into a few type classes, and each type class is a permutation group, this motivates the following definition.

Definition 3. A subset $\mathcal{L} = \{\mathcal{L}_n\}_{n=1}^{\infty}$ is called symmetric if it has the property that, for any sequence $x^n \in \mathcal{L}_n$, all permutations of x^n also belong to \mathcal{L}_n , for all $n = 1, 2, \dots$

One readily observes that, for symmetric subsets over a continuous range of distributions Q intersecting the subset, we have the property $H_{\mathcal{L}}(Q) = \mathbb{H}(Q)$ and the objective function thus reduces to $g_P(Q) = D(Q||P)$. Hence, we arrive at the following simpler expression.

Corollary 1. For a discrete memoryless source P and any symmetric subset $\mathcal{L} = \{\mathcal{L}_n\}_{n=1}^{\infty}$ for which $\mathbb{H}(Q)$ is continuous in all distributions Q intersecting the subset, i.e. $\mathcal{L} \cap T[Q] \neq \emptyset$, the optimal lossless compression rate for the

subset \mathcal{L} is

$$R_{\mathcal{L}}^* = \max \left\{ \mathbb{H}(Q^*) : Q^* \in \arg \min_{Q: \mathcal{L} \cap T[Q] \neq \emptyset} D(Q||P) \right\}.$$

C. Fluctuating Subsets with Smooth Components

In this subsection, we consider subsets that are not smooth so that the limit in (12) does not exist and the subset-entropy is not well-defined, but are fluctuating among a finite number of smooth components. In such cases, one can code for the *worst* subset component as described in the following theorem.

Theorem 3. Consider a discrete memoryless source P ; a finite collection of subsets $\mathcal{L}_j = \{\mathcal{L}_{j,n}\}_{n=1}^{\infty}$ with $1 \leq j \leq J$ for which the subset entropy $H_{\mathcal{L}_j}(Q)$ exists and is continuous in all distributions Q intersecting each subset, $\mathcal{L}_j \cap T[Q] \neq \emptyset$; and a finite collection of index subsequences $\{n_{j,k}\}_{k=1}^{\infty}$ with $1 \leq j \leq J$ such that for each $n = 1, 2, \dots$ we have $n = n_{j,k}$ for a unique pair (j, k) . Define the fluctuating subset $\mathcal{L} = \{\mathcal{L}_n\}_{n=1}^{\infty}$ as $\mathcal{L}_n = \mathcal{L}_{j,n}$ if $n \in \{n_{j,k}\}_{k=1}^{\infty}$. Then, the optimal compression rate for the subset \mathcal{L} is

$$R_{\mathcal{L}}^* = \max_{1 \leq j \leq J} \max \left\{ H_{\mathcal{L}_j}(Q_j^*) : Q_j^* \in \arg \min_{Q: \mathcal{L}_j \cap T[Q] \neq \emptyset} [\mathbb{H}(Q) - H_{\mathcal{L}_j}(Q) + D(Q||P)] \right\}.$$

Proof: (Achievability) We build upon the achievability statement of Theorem 2. Fix an arbitrary $\epsilon > 0$. For each $1 \leq j \leq J$, let $\{(m_{j,n}, \hat{x}_j^n)\}_{n=1}^{\infty}$ be the optimal encoder and decoder sequence for compression of the subset \mathcal{L}_j , achieving $\Pr[\hat{X}_j^n \neq X^n | X^n \in \mathcal{L}_{j,n}] \rightarrow 0$ as $n \rightarrow \infty$ with a rate $R_{\mathcal{L}_j}^* + \epsilon = \max_{Q_j^*} H_{\mathcal{L}_j}(Q_j^*) + \epsilon$. We consider the following code for the fluctuating subset: let $m_n \equiv m_{j,n}$ and $\hat{x}^n \equiv \hat{x}_j^n$ if $n \in \{n_{j,k}\}_{k=1}^{\infty}$. Then, we have

$$\begin{aligned} \limsup_{n \rightarrow \infty} \Pr[\hat{X}^n \neq X^n | X^n \in \mathcal{L}_n] \\ = \max_{1 \leq j \leq J} \limsup_{n \rightarrow \infty} \Pr[\hat{X}_j^n \neq X^n | X^n \in \mathcal{L}_{j,n}] = 0. \end{aligned} \quad (14)$$

The rate of this code is $\max_{1 \leq j \leq J} R_{\mathcal{L}_j}^* + \epsilon$. Since ϵ is arbitrary, this completes the achievability proof.

(Converse) We build upon the strong converse statement of Theorem 2. Assume $R < \max_{1 \leq j \leq J} \max_{Q_j^*} H_{\mathcal{L}_j}(Q_j^*)$, then there exists at least one $1 \leq \bar{j} \leq J$ such that $R < R_{\mathcal{L}_{\bar{j}}}^* = \max_{Q_{\bar{j}}^*} H_{\mathcal{L}_{\bar{j}}}(Q_{\bar{j}}^*)$. By the strong converse result of Theorem 2, we have $\limsup_{n \rightarrow \infty} \Pr[\hat{X}_{\bar{j}}^n \neq X^n | X^n \in \mathcal{L}_{\bar{j},n}] = 1$ for any arbitrary compression code for subset $\mathcal{L}_{\bar{j}}$. Hence,

$$\begin{aligned} \limsup_{n \rightarrow \infty} \Pr[\hat{X}^n \neq X^n | X^n \in \mathcal{L}_n] \\ \geq \limsup_{n \rightarrow \infty} \Pr[\hat{X}_{\bar{j}}^n \neq X^n | X^n \in \mathcal{L}_{\bar{j},n}] = 1, \end{aligned} \quad (15)$$

which proves the strong converse property for the fluctuating subset \mathcal{L} and completes the proof of Theorem 3. ■

V. PROOF OF THEOREM 2

In this section, we provide the proof of Theorem 2. The proof builds upon the following lemma, which is a dual of Sanov's theorem [2, Pr. 2.12] and summarizes the properties of typical sequences intersecting a subset of the source. In particular, we frequently use the notations

$$T_{\mathcal{L}}^n(\hat{P}) := \mathcal{L}_n \cap T^n(\hat{P}), \quad T_{\mathcal{L}}^n[Q]_{\delta_n} := \mathcal{L}_n \cap T^n[Q]_{\delta_n}. \quad (16)$$

Lemma 1. Consider a discrete memoryless source P , a subset $\mathcal{L} = \{\mathcal{L}_n\}_{n=1}^{\infty}$, and a distribution Q intersecting the subset, $\mathcal{L} \cap T[Q] \neq \emptyset$. If the subset entropy $H_{\mathcal{L}}(Q)$ exists, then there exists some $\epsilon_n \rightarrow 0$ as $n \rightarrow \infty$ such that

$$2^{-n[g_P(Q)+\epsilon_n]} \leq P_{X^n}[X^n \in T_{\mathcal{L}}^n[Q]_{\delta_n}] \leq 2^{-n[g_P(Q)-\epsilon_n]}, \quad (17)$$

where function $g_P(Q)$ is defined in (13). Moreover, if $H_{\mathcal{L}}(Q)$ exists and is continuous for all Q satisfying $\mathcal{L} \cap T[Q] \neq \emptyset$, then

$$P_{X^n}[X^n \in \mathcal{L}_n] \geq 2^{-n[\min_{Q:\mathcal{L} \cap T[Q] \neq \emptyset} g_P(Q) + \epsilon_n]}. \quad (18)$$

Proof: Recall from the properties of type classes [2] that, all sequences $x^n \in T^n(\hat{P})$ satisfy

$$P_{X^n}(x^n) = 2^{-n[\mathbb{H}(\hat{P}) + D(\hat{P} \| P)]}. \quad (19)$$

On the other hand, the existence of the subset entropy $H_{\mathcal{L}}(Q)$ as defined by the limit (12) implies that, there exists some $\epsilon'_n \rightarrow 0$ as $n \rightarrow \infty$ such that

$$H_{\mathcal{L}}(Q) - \epsilon'_n \leq \frac{1}{n} \log |T_{\mathcal{L}}^n[Q]_{\delta_n}| \leq H_{\mathcal{L}}(Q) + \epsilon'_n. \quad (20)$$

Now, note that

$$\begin{aligned} |T_{\mathcal{L}}^n[Q]_{\delta_n}| \min_{x^n \in T^n[Q]_{\delta_n}} P_{X^n}(x^n) &\leq P_{X^n}[X^n \in T_{\mathcal{L}}^n[Q]_{\delta_n}] \leq \\ &\leq |T_{\mathcal{L}}^n[Q]_{\delta_n}| \max_{x^n \in T^n[Q]_{\delta_n}} P_{X^n}(x^n). \end{aligned} \quad (21)$$

But, (19) and the continuity of the Shannon entropy and relative entropy implies the existence of $\epsilon''_n \rightarrow 0$ such that

$$\begin{aligned} \min_{x^n \in T^n[Q]_{\delta_n}} P_{X^n}(x^n) &\geq 2^{-n[\mathbb{H}(Q) + D(Q \| P) + \epsilon''_n]}, \\ \max_{x^n \in T^n[Q]_{\delta_n}} P_{X^n}(x^n) &\leq 2^{-n[\mathbb{H}(Q) + D(Q \| P) - \epsilon''_n]}. \end{aligned} \quad (22)$$

Combining (21) with (20), (22), and recalling the definition (13) of the function $g_P(Q)$ completes the proof of the first part of the lemma in (17) with $\epsilon_n := \epsilon'_n + \epsilon''_n$. The proof of the second part then immediately follows, since the subset entropy $H_{\mathcal{L}}(Q)$ and therefore the function $g_P(Q)$ is assumed to be continuous:

$$P_{X^n}[X^n \in \mathcal{L}_n] = P_{X^n}[X^n \in \bigcup_{\hat{P}: n\text{-type}} T_{\mathcal{L}}^n(\hat{P})] \quad (23)$$

$$\geq \max_{Q:\mathcal{L} \cap T[Q] \neq \emptyset} P_{X^n}[X^n \in T_{\mathcal{L}}^n[Q]_{\delta_n}] \quad (24)$$

which implies (18) and completes the proof of Lemma 1. ■

We are now ready to prove Theorem 2, which is inspired by [2, Th. 2.15 and Pr. 2.6].

Proof: (of Theorem 2) To prove the achievability side, we consider the following source code for the subset $\mathcal{L} =$

$\{\mathcal{L}_n\}_{n=1}^{\infty}$. Fix an arbitrary $\epsilon > 0$. The encoder indexes all sequences x^n belonging to the set \mathcal{A}_n defined as

$$\mathcal{A}_n := \bigcup_{\hat{P}: n\text{-type}, \hat{P} \in \Omega(3\epsilon)} T_{\mathcal{L}}^n(\hat{P}), \quad (25)$$

where

$$\Omega(\epsilon) := \left\{ Q: \mathcal{L} \cap T[Q] \neq \emptyset, g_P(Q) < \min_{Q:\mathcal{L} \cap T[Q] \neq \emptyset} g_P(Q) + \epsilon \right\}. \quad (26)$$

All other sequences in $(\mathcal{L}_n - \mathcal{A}_n)$ lead to an error $\mathcal{E}_{\mathcal{L}}$. Note that, the use of \min for $g_P(Q)$ in the definition (26) is justified by the continuity of the subset entropy $H_{\mathcal{L}}(Q)$ and thus the function $g_P(Q)$. We can write

$$\begin{aligned} \Pr[X^n \in (\mathcal{A}_n^c \cap \mathcal{L}_n)] &= \sum_{\hat{P}: n\text{-type}, \hat{P} \notin \Omega(3\epsilon)} P_{X^n}[X^n \in T_{\mathcal{L}}^n(\hat{P})] \\ &\leq (n+1)^{|\mathcal{X}|} \max_{Q \notin \Omega(3\epsilon), \mathcal{L} \cap T[Q] \neq \emptyset} P_{X^n}[X^n \in T_{\mathcal{L}}^n[Q]_{\delta_n}]. \end{aligned} \quad (27)$$

Combining (27) and Lemma 1, the error probability is bounded as

$$\begin{aligned} \Pr[\mathcal{E}_{\mathcal{L}}] &= \Pr[X^n \notin \mathcal{A}_n | X^n \in \mathcal{L}_n] \quad (28) \\ &\leq \frac{(n+1)^{|\mathcal{X}|} \max_{Q \notin \Omega(3\epsilon), \mathcal{L} \cap T[Q] \neq \emptyset} P_{X^n}[X^n \in T_{\mathcal{L}}^n[Q]_{\delta_n}]}{P_{X^n}[X^n \in \mathcal{L}_n]} \\ &\leq \frac{(n+1)^{|\mathcal{X}|} 2^{-n[\min_{Q \notin \Omega(3\epsilon), \mathcal{L} \cap T[Q] \neq \emptyset} g_P(Q) - \epsilon_n]}}{2^{-n[\min_{Q:\mathcal{L} \cap T[Q] \neq \emptyset} g_P(Q) + \epsilon_n]}}. \end{aligned} \quad (29)$$

Therefore, from definition (26) of the set $\Omega(\epsilon)$, we have proved the existence of a source code for subset \mathcal{L} with vanishing error probability, $\Pr[\mathcal{E}_{\mathcal{L}}] \leq (n+1)^{|\mathcal{X}|} 2^{-n\epsilon}$, and achieving the compression rate

$$\frac{1}{n} \log |\mathcal{A}_n| = \frac{1}{n} \log \sum_{\hat{P}: n\text{-type}, \hat{P} \in \Omega(3\epsilon)} |T_{\mathcal{L}}^n(\hat{P})| \quad (30)$$

$$\leq \frac{1}{n} \log \left((n+1)^{|\mathcal{X}|} \max_{Q \in \Omega(3\epsilon)} |T_{\mathcal{L}}^n[Q]_{\delta_n}| \right) \quad (31)$$

$$\leq \max_{Q \in \Omega(3\epsilon)} H_{\mathcal{L}}(Q) + \epsilon'_n + \frac{|\mathcal{X}| \log(n+1)}{n}, \quad (32)$$

where (31) follows from the Type Counting Lemma, and (32) from (20) and the continuity of the subset entropy $H_{\mathcal{L}}(Q)$. Since $n \rightarrow \infty$ and the choice of $\epsilon > 0$ is arbitrary, this completes the achievability proof for Theorem 2.

In the following, we prove a strong converse for Theorem 2, that is, we prove any arbitrary source code for subset \mathcal{L} with rate $R < R_{\mathcal{L}}^*$ has an error probability approaching one.

To this end, first let $\mathcal{A}_n := \{x^n(j)\}_{j=1}^{2^{nR}}$ be the set of encoded sequences which will be correctly decoded, and note that the Type Counting Lemma implies

$$\begin{aligned} \Pr[X^n \in (\mathcal{A}_n \cap \mathcal{L}_n)] &= \sum_{\hat{P}: n\text{-type}} P_{X^n}[X^n \in (\mathcal{A}_n \cap T_{\mathcal{L}}^n(\hat{P}))] \\ &\leq (n+1)^{|\mathcal{X}|} \max_{Q:\mathcal{L} \cap T[Q] \neq \emptyset} P_{X^n}[X^n \in (\mathcal{A}_n \cap T_{\mathcal{L}}^n[Q]_{\delta_n})]. \end{aligned} \quad (33)$$

However, we have for any distribution $Q(x)$ that

$$P_{X^n}[X^n \in (\mathcal{A}_n \cap T_{\mathcal{L}}^n[Q]_{\delta_n})] \leq |\mathcal{A}_n \cap T_{\mathcal{L}}^n[Q]_{\delta_n}| \max_{x^n \in T^n[Q]_{\delta_n}} P_{X^n}(x^n) \quad (34)$$

$$\leq \min\{2^{nR}, 2^{n[H_{\mathcal{L}}(Q) + \epsilon'_n]}\} \times 2^{-n[\mathbb{H}(Q) + D(Q\|P) - \epsilon''_n]} \quad (35)$$

$$= 2^{-n[g_P(Q) + |H_{\mathcal{L}}(Q) - R + \epsilon'_n|^+ - \epsilon_n]}, \quad (36)$$

where (35) follows from (20) and (22), and (36) from the definition of $g_P(Q)$ and $\epsilon_n = \epsilon'_n + \epsilon''_n$. Combining (33), (36) and Lemma 1, the correct decoding probability is bounded as

$$1 - \Pr[\mathcal{E}_{\mathcal{L}}] = \Pr[X^n \in \mathcal{A}_n | X^n \in \mathcal{L}_n] \leq \frac{(n+1)^{|\mathcal{X}|} \max_{Q: \mathcal{L} \cap T[Q] \neq \emptyset} P_{X^n}[X^n \in (\mathcal{A}_n \cap T_{\mathcal{L}}^n[Q]_{\delta_n})]}{P_{X^n}[X^n \in \mathcal{L}_n]} \quad (37)$$

$$\leq (n+1)^{|\mathcal{X}|} 2^{n[\min_{Q: \mathcal{L} \cap T[Q] \neq \emptyset} g_P(Q) + \epsilon_n]} \times 2^{-n[\min_{Q: \mathcal{L} \cap T[Q] \neq \emptyset} g_P(Q) + |H_{\mathcal{L}}(Q) - R + \epsilon'_n|^+ - \epsilon_n]}. \quad (38)$$

Inspecting the lower bound (38) on error probability suggests that, if $R < H_{\mathcal{L}}(Q^*) - 4\epsilon_n$ for any distribution Q^* satisfying $g(Q^*) \leq \min_{Q: \mathcal{L} \cap T[Q] \neq \emptyset} g_P(Q) + 4\epsilon_n$, then the error probability is bounded at least as $\Pr[\mathcal{E}_{\mathcal{L}}] \geq 1 - 2^{-n\epsilon_n}$. Since ϵ_n is vanishing and $n\epsilon_n \rightarrow \infty$, this proves the strong converse and completes the proof of Theorem 2. ■

VI. NUMERICAL EXAMPLES

In this section, we present several numerical examples to illustrate our models and results. In all of these examples, we consider a binary DMS with a Bernoulli distribution $B(p)$ with parameter p , so that $\mathcal{X} = \{0, 1\}$ and $\Pr[X_t = 1] = p$ for some $0 \leq p \leq 1/2$ and all $t = 1, \dots, n$. We use the Hamming weight $w_H(x^n)$ of a binary sequence x^n , the binary entropy function $H_b(p) := -p \log p - (1-p) \log(1-p)$, and the binary divergence function $D_b(q\|p) := q \log(q/p) + (1-q) \log((1-q)/(1-p))$.

The first two examples are symmetric, thus can be readily handled by Corollary 1.

Example 1. Consider $\mathcal{L} = \{\mathcal{L}_n\}_{n=1}^{\infty}$ with

$$\mathcal{L}_n := \{x^n \in \mathcal{X}^n : w_H(x^n) = \lfloor nq \rfloor\}, \quad 0 \leq q \leq 1.$$

This subset is symmetric and $B(q)$ is the only distribution that intersects the subset \mathcal{L} , so we obtain from Corollary 1 that

$$R_{\mathcal{L}}^* = \mathbb{H}(B(q)) = H_b(q). \quad (39)$$

It is evident that the subset compression rate can be below or beyond the source entropy $H_b(p)$.

Example 2. Consider $\mathcal{L} = \{\mathcal{L}_n\}_{n=1}^{\infty}$ with

$$\mathcal{L}_n := \{x^n \in \mathcal{X}^n : 0 \leq w_H(x^n) \leq nq\}, \quad 0 \leq q \leq 1.$$

This is again a symmetric subset, so we simply use Corollary 1 to obtain

$$R_{\mathcal{L}}^* = \max\{H_b(q^*) : q^* = \arg \min_{\bar{q}: 0 \leq \bar{q} \leq q} D_b(\bar{q}\|p)\} = H_b(\min\{p, q\}), \quad (40)$$

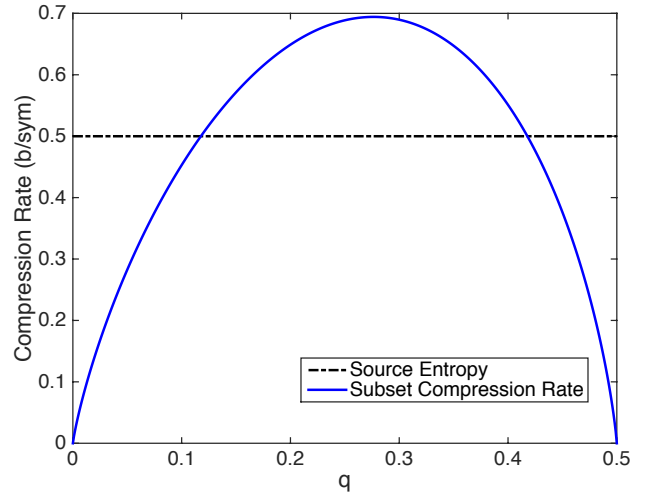


Fig. 2. Comparison of the optimal subset compression rate of Example 3 with the source entropy for a Bernoulli DMS source with parameter $p = 0.11$.

since $B(\bar{q})$ with $0 \leq \bar{q} \leq q$ are the only distributions that intersect the subset \mathcal{L} . The optimal rate $R_{\mathcal{L}}^* = H_b(p)$ for the case $q \geq p$ also follows from Theorem 1 since $\Pr[X^n \in \mathcal{L}_n] \approx 1$. It is evident that the compression rate for this subset never exceeds the source entropy $H_b(p)$, so we achieve a rate gain by focusing only on the subset.

In the following, we consider two non-symmetric examples for which Corollary 1 does not apply and classical entropy quantities are not sufficient for the analysis. Therefore, we need to resort to the original form in Theorem 2 and perform further computations to find the subset-entropy.

Example 3. Consider $\mathcal{L} = \{\mathcal{L}_n\}_{n=1}^{\infty}$ with

$$\mathcal{L}_n := \{x^n \in \mathcal{X}^n : w_H(x^n) = \lfloor nq \rfloor, x^n \text{ has no consecutive 1s}\},$$

where $0 \leq q \leq 1/2$. In this case, Theorem 1 does not apply since the subset has exponentially small probability. In order to employ Theorem 2, we first note that $B(q)$ is the only distribution that intersects the subset \mathcal{L} , and it has a subset entropy given by

$$H_{\mathcal{L}}(B(q)) = \lim_{n \rightarrow \infty} \frac{1}{n} \log \binom{n - \lfloor nq \rfloor + 1}{\lfloor nq \rfloor} = (1-q) H_b\left(\frac{q}{1-q}\right). \quad (41)$$

Therefore, we obtain

$$R_{\mathcal{L}}^* = H_{\mathcal{L}}(B(q)) = (1-q) H_b\left(\frac{q}{1-q}\right). \quad (42)$$

A plot of this compression rate is illustrated in Figure 2, which shows the subset compression rate (42) can be below or beyond the source entropy $H_b(p)$.

Example 4. Consider $\mathcal{L} = \{\mathcal{L}_n\}_{n=1}^{\infty}$ with

$$\mathcal{L}_n := \{x^n \in \mathcal{X}^n : x^n \text{ has no consecutive 1s}\}.$$

Again, Theorem 1 does not apply since the subset is not likely. In order to employ Theorems 2, we first note that all distributions $B(q)$ with $0 \leq q \leq 1/2$ intersect the subset \mathcal{L} ,

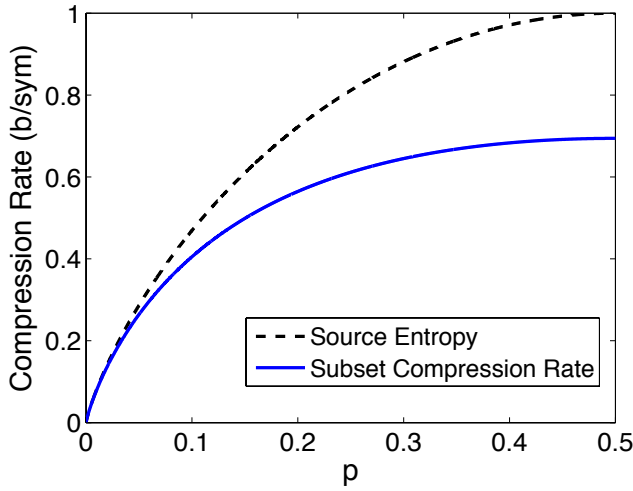


Fig. 3. Comparison of the optimal lossless subset-compression rate of Example 4 with the source entropy for a Bernoulli DMS with parameter p .

and each has a subset entropy given by (41). Therefore, we obtain $R_{\mathcal{L}}^* = (1 - q^*)H_b\left(\frac{q^*}{1 - q^*}\right)$ where

$$q^* = \arg \min_{0 \leq q \leq 1/2} \left[H_b(q) - (1 - q)H_b\left(\frac{q}{1 - q}\right) + D_b(q||p) \right]. \quad (43)$$

A plot of this subset-compression rate is illustrated in Figure 3, which shows the optimal lossless compression rate of this subset is always below the source entropy $H_b(p)$.

Finally, we present a non-smooth example for which Theorem 2 is not directly applicable. However, the characterization of Theorem 3 facilitates the analysis.

Example 5. Consider a subset $\mathcal{L}_1 = \{\mathcal{L}_{1,n}\}_{n=1}^{\infty}$ with

$$\mathcal{L}_{1,n} := \{x^n \in \mathcal{X}^n : nq_1 \leq w_H(x^n) \leq nq_2, x^n \text{ has no consecutive 1s}\},$$

for some $0 \leq q_1 \leq q_2 \leq 1/2$, and another subset $\mathcal{L}_2 = \{\mathcal{L}_{2,n}\}_{n=1}^{\infty}$ with

$$\mathcal{L}_{2,n} := \{x^n \in \mathcal{X}^n : nw_1 \leq w_H(x^n) \leq nw_2, x^n \text{ has 1s only in even positions}\},$$

for some $0 \leq w_1 \leq w_2 \leq 1/2$. Now, consider the superimposed subset $\mathcal{L} = \{\mathcal{L}_n\}_{n=1}^{\infty}$ with

$$\mathcal{L}_n := \begin{cases} \mathcal{L}_{1,n} & \text{if } n \text{ odd} \\ \mathcal{L}_{2,n} & \text{if } n \text{ even} \end{cases}. \quad (44)$$

Note that, Theorem 1 does not apply since the fluctuating subset \mathcal{L} is not likely, and Theorem 2 does not since the subset is not smooth. However, both components are smooth subsets. In particular, the second subset \mathcal{L}_2 intersects all distributions $B(w)$ with $w_1 \leq w \leq w_2$ with a subset entropy given by

$$H_{\mathcal{L}}(B(w)) = \lim_{n \rightarrow \infty} \frac{1}{n} \log \binom{\lfloor n/2 \rfloor}{\lfloor nw \rfloor} = \frac{1}{2} H_b(2w).$$

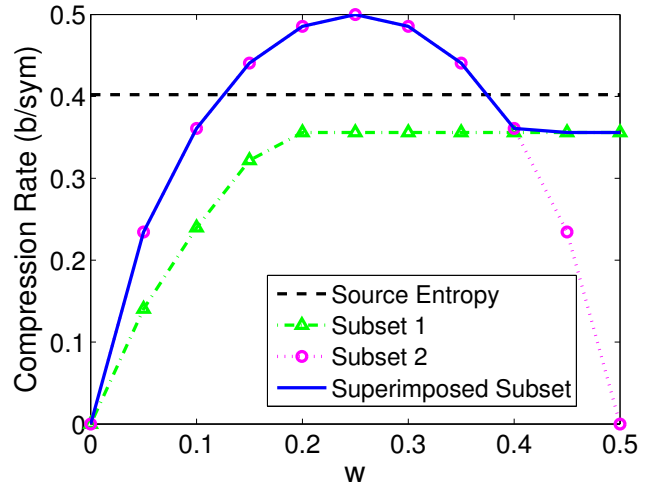


Fig. 4. The optimal lossless compression rate of the superimposed subset of Example 5 for a binary DMS with fixed parameter $p = 0.08$ and varying subset parameters $q_1 = 0, q_2 = 0.4w, w_1 = w_2 = w$ where $0 \leq w \leq 1/2$.

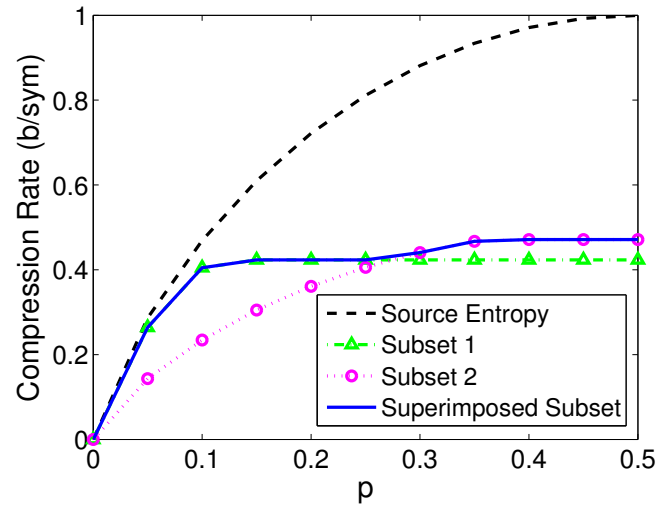


Fig. 5. The optimal lossless compression rate of the superimposed subset of Example 5 for fixed subset parameters $q_1 = 0, q_2 = 0.09$ and $w_1 = 0, w_2 = 0.18$ and a binary DMS with varying parameter $0 \leq p \leq 1/2$.

Hence, we can use Theorem 3 to obtain

$$R_{\mathcal{L}}^* = \max \{ R_{\mathcal{L}_1}^*, R_{\mathcal{L}_2}^* \}, \quad (45)$$

where $R_{\mathcal{L}_1}^* = (1 - q^*)H_b\left(\frac{q^*}{1 - q^*}\right)$ with

$$q^* = \arg \min_{q_1 \leq q \leq q_2} \left[H_b(q) - (1 - q)H_b\left(\frac{q}{1 - q}\right) + D_b(q||p) \right], \quad (46)$$

and $R_{\mathcal{L}_2}^* = \frac{1}{2} H_b(2w^*)$ with

$$w^* = \arg \min_{w_1 \leq w \leq w_2} \left[H_b(w) - \frac{1}{2} H_b(2w) + D_b(w||p) \right]. \quad (47)$$

To show the different aspects of this scenario, we make two comparisons. In one case, we fix the source distribution to $p = 0.08$ and vary the subset parameters as $q_1 = 0$, $q_2 = 0.4w$ and $w_1 = w_2 = w$ where $0 \leq w \leq 1/2$. The compression rate for this superimposed subset is shown in Figure 4. The compression rate is observed to be dominated by that of the second subset for smaller values of w and by that of the first subset for larger values of w . One also notes that the optimal subset-compression rate in this case can be below or beyond the source entropy $H_b(p)$.

In the second case, we fix the subset parameters to $q_1 = 0$, $q_2 = 0.09$ and $w_1 = 0$, $w_2 = 0.18$ and vary the source distribution as $0 \leq p \leq 1/2$. The compression rate for this superimposed subset is shown in Figure 5. In this case, the compression rate of the superimposed subset is observed to be dominated by that of the first subset for smaller values of p and by that of the second subset for larger values of p . In either situations, however, the subset-compression rate always remains below the source entropy $H_b(p)$.

VII. CONCLUDING REMARKS

We have provided a framework, as well as several optimality results, for lossless compression of subsets of discrete memoryless sources. We envision two immediate directions for future research on this topic. One is to analyze the fundamental compression limit of those subsets that are not covered by our current results. Another more important extension is to investigate the *lossy* compression for subsets of discrete memoryless sources. We have explored this direction in [11] and provided results analogous to our lossless theorems given in this paper.

ACKNOWLEDGMENT

Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053 (the ARL Network Science CTA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government.

REFERENCES

- [1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley, New York, 1991.
- [2] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Academic, New York, 1981.
- [3] N. Merhav and M. Feder, "Universal prediction," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2124–2147, 1998.
- [4] A. Ingber and T. Weissman, "Compression for Similarity Identification: Fundamental Limits," in *Proc. IEEE Int. Symp. on Inf. Theory (ISIT)*. Honolulu, HI, 2014, pp. 1–5.
- [5] Y. Bar-Hillel and R. Carnap, "Semantic information," *The British Journal for the Philosophy of Science*, vol. 4, no. 14, pp. 147–157, 1953.
- [6] C. Bunte and A. Lapidoth, "Encoding tasks and Rényi entropy," *IEEE Trans. Inf. Theory*, vol. 60, no. 9, pp. 5065–5076, 2014.
- [7] A. Høst-Madsen, E. Sabeti, and C. Walton, "Information theory for atypical sequences," in *Proc. IEEE Information Theory Workshop (ITW)*. Seville, Spain, 2013, pp. 1–5.
- [8] A. Dembo and I. Kontoyannis, "Source coding, large deviations, and approximate pattern matching," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1590–1615, 2002.
- [9] T.-S. Han, *Information-Spectrum Methods in Information Theory*. Springer-Verlag, Berlin, Germany, 2003.
- [10] T. S. Han and S. Verdú, "Approximation theory of output statistics," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 752–772, 1993.
- [11] E. MolavianJazi and A. Yener, "Subset source coding," *IEEE Trans. Inf. Theory*, in preparation.