Lossy Subset Source Coding

Ebrahim MolavianJazi Aylin Yener
Wireless Communications and Networking Laboratory
Department of Electrical Engineering
The Pennsylvania State University, University Park, PA 16802
Email: ebrahim@psu.edu yener@engr.psu.edu

***Abstract*—This paper studies the lossy version of a problem recently proposed by the authors termed *subset source coding*, where the focus is on the fundamental limits of compression for *subsets* of the possible realizations of a discrete memoryless source. An upper bound is derived on the subset rate-distortion function in terms of the *subset mutual information* optimized over the set of conditional distributions that satisfy the expected distortion constraint with respect to the subset-typical distribution and over the set of certain auxiliary subsets. By proving a strong converse result, this upper bound is shown to be tight for a class of *symmetric* subsets. As illustrated in our numerical examples, more often than not, one achieves a gain in the fundamental limit, in that the optimal lossy compression rate for the subset can be strictly smaller than the rate-distortion function of the source, although exceptions can also be constructed.**

I. Introduction

Source coding addresses compression, with or without fidelity, of an information source. Critical to compression of discrete memoryless sources (DMS) is the set of *(source-) typical* sequences that capture essentially all the probability mass of the source. In particular, in lossy compression of a DMS, one essentially groups source-typical sequences and *covers* each group with a sequence that is within a certain distortion distance of them [1], [2]. The fundamental limit for a DMS $X \sim P(x)$ with a distortion requirement D is given by the rate-distortion function [1], [3]:

$$R(D) = R(P, D) := \min_{P_{Y|X}:\mathbb{E}[d(X,Y)]\leq D} \mathbb{I}(X;Y). \qquad (1)$$

This basic setting has been studied extensively and extended to different scenarios and applications, see for example [4]–[7]. An implicit but pivotal consideration in all of these works is that important realizations of the source only consist of the likely and source-typical sequences.

We have argued in [8] that, for some emerging applications, the likelihood and typicality of a source realization may not be the main concern. In semantic communications, bioinformatics, and data mining, only data elements with certain patterns or structures are considered meaningful, valid, or important, although they may occur with potentially low probability. In [8], we have introduced the problem of *subset source coding*, where the encoder and decoder aim at providing a description of only a *subset* of all possible source realizations as determined by the application.

The subset source coding problem inherently involves both probabilistic and combinatorial aspects. On one hand, it has roots in large deviations theory [9] and relates to the generalized asymptotic equipartition property (AEP) [10]. On the other hand, it has a combinatorial element in terms of the exponential number of sequences that satisfy certain structural constraints, and therefore relates to the capacity of magnetic recording channels with constrained coding [11]; the notion of Markov types in compression of Markovian sources [12]; and entropy definitions in statistical mechanics [13]. In its motivation, subset source coding is related to the problem of task encoding in [14] that guarantees certain important but less likely source events are not ignored in data compression, and also the work on information theory of atypical sequences in [15] with applications in signal processing and Big-Data analytics.

In [8], we have analyzed three broad classes of subsets, namely *likely* subsets with not(-so-fast)-vanishing probabilities, *smooth* subsets with continuous structures, and *fluctuating* subsets that alternate between several subset components. In particular, for smooth subsets, we have shown that lossless compression is closely tied to certain *subset-typical sequences*, and that the fundamental limit of lossless compression is the result of a trade-off between the source statistics and the subset structure and is given by a certain *subset entropy* of the *subset-typical distributions*.

Our key contributions in this paper are as follows. For likely subsets, we prove an achievability and a matching strong converse to show that the rate-distortion function of the subset is equal to that of the original source. For the lossy compression of smooth subsets, we prove an achievability that relates the subset rate-distortion function to a certain *subset mutual information* corresponding again to the subset-typical distributions. For the interesting special case of *smooth symmetric* subsets, we show that our achievability result for the lossy case is tight by proving a strong converse. For fluctuating subsets, we prove an achievability and converse to show that the rate-distortion function of the subset is equal to that of the *worst* subset component.

The remainder of the paper is organized as follows. In Section II, we formally introduce the lossy subset source coding problem and provide a motivating example. In Sections III, IV, and V, we state and prove optimal lossy compression rates for likely, smooth, and fluctuating subsets, respectively. The proofs for smooth subsets are given in Section VI. In Section VII, we revisit our numerical examples in [8] and investigate their rate-distortion functions. It turns out that, when focusing only on a subset instead of the entire source space, there is often a gain in the compression rate, but interestingly this is not always the case. We conclude the paper in Section VIII with some discussions and remarks about possible extensions. One technical proof is relegated to the Appendix.

Notation. In order to carefully keep track of the actual

distributions governing the random variables, we follow the notation of Csiszár and Körner [16] for entropy and mutual information quantities. Consider a random variable X with distribution $P(x)$. The standard Shannon entropy $\mathbb{H}(X)$ is denoted by

$$\mathbb{H}(P) := -\sum_{x\in\mathcal{X}} P(x)\log P(x),$$

where the $\log$ operation is understood as base 2 here and throughout this paper. Next, let Y be a random variable conditionally distributed according to $P_{Y|X}(y|x)$. The conditional entropy $\mathbb{H}(Y|X) = \sum_x P_X(x)\mathbb{H}(Y|X=x)$ is denoted by

$$\mathbb{H}(P_{Y|X}|P_X) := \sum_{x\in\mathcal{X}} P_X(x)\mathbb{H}(P_{Y|X=x}).$$

Moreover, the standard average mutual information $\mathbb{I}(X;Y) = \mathbb{H}(Y) - \mathbb{H}(Y|X)$ is denoted by

$$\mathbb{I}(P_X, P_{Y|X}) := \mathbb{H}(P_Y) - \mathbb{H}(P_{Y|X}|P_X),$$

where $P_Y(y) = \sum_x P_X(x)P_{Y|X}(y|x)$ is the marginal distribution of Y. Since we can also write $\mathbb{I}(X;Y) = \mathbb{H}(X) - \mathbb{H}(X|Y)$, we can also use the notation

$$\mathbb{I}(P_X, P_{Y|X}) := \mathbb{H}(P_X) - \mathbb{H}(P_{X|Y}|P_Y),$$

where $P_{X|Y}(x|y) = P_X(x)P_{Y|X}(y|x)/P_Y(y)$ is the induced conditional distribution of X given Y. Finally, the relative entropy is denoted as usual by

$$D(Q\|P) := \sum_{x\in\mathcal{X}} Q(x)\log\frac{Q(x)}{P(x)}.$$

II. Problem Setting

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

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

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. Furthermore, consider a reconstruction alphabet $\mathcal{Y}$ and an additive distortion measure $d : \mathcal{X}\times\mathcal{Y} \to [0, D_{\max}]$ for some maximal distortion value $D_{\max} < \infty$ and define

$$d(x^n, y^n) := \frac{1}{n}\sum_{t=1}^{n} d(x_t, y_t).$$

An $(n, 2^{nR})$ lossy source code for subset $\mathcal{L}$ consists of an encoder $f : \mathcal{L}_n \to \{1, 2, \cdots, 2^{nR}\}$ and a decoder $\phi : \{1, 2, \cdots, 2^{nR}\} \to \mathcal{Y}^n$. For any distortion values $D \geq 0$, the *probability of excess-distortion*[1] is defined as

$$\Pr[\mathcal{E}_{\mathcal{L}}(D)] := \Pr[d(X^n, Y^n) > D | X^n \in \mathcal{L}_n].$$

A rate-distortion pair (R, D) is called achievable if a sequence of $(n, 2^{nR})$ lossy source codes for subset $\mathcal{L}$ exists with $\Pr[\mathcal{E}_{\mathcal{L}}(D)] \to 0$ as $n \to \infty$. The subset rate-distortion function $R_{\mathcal{L}}(D)$ is the infimum of all rates R for which the rate-distortion pair (R, D) is achievable.

Remark 1. The subset rate-distortion function $R_{\mathcal{L}}(D)$, similar to the standard rate-distortion function, is a non-increasing function of D per definition. The convexity of $R_{\mathcal{L}}(D)$, however, is not a priori obvious. The latter would normally build on a time-sharing argument, which is not trivial for the subset source coding problem. In fact, if a codeword x^n belongs to the subset $\mathcal{L}_n$, it may or may not be true that a portion of it $x^{\alpha n}$ belongs to $\mathcal{L}_{\alpha n}$ for some $0 \leq \alpha \leq 1$.

A comment similar to one in [8] for the lossless case is that, the subset rate-distortion function $R_{\mathcal{L}}(D)$ is in fact equal to the standard rate-distortion function of an equivalent conditional source defined as

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

The rate-distortion function of this equivalent conditional source is given either by average mutual information rate results [1], [17] if stationary and ergodic, or by spectral sup-mutual information rate characterizations if non-stationary or non-ergodic [18], [19]. Note that, even the simplest of subsets such as our Example 1 in Section VII lead to equivalent conditional sources that fall into the second category, cf. [18, Example 1.5.1], so numerical computations may not be very straightforward in general. Moreover, the effect of subset structure and the statistics of the original source on the fundamental limits are not quite explicit in such generic approaches. Our methods in the next sections of this paper are in the same spirit as the two aforementioned approaches and essentially lead to the same results, but they are presented in a simpler and more convenient form and language, and they also explicitly clarify the roles of subset structure and source statistics on the fundamental compression performance.

A. Motivating Example

Consider a binary DMS, $\mathcal{X} = \{0, 1\}$, with a Bernoulli distribution with parameter $p = 0.11$, so that the Shannon entropy of the source is simply the binary entropy $H_b(p) = 0.5$, and its rate-distortion function with respect to the Hamming distance is $R(D) = 0.5 - H_b(D)$ for $0 \leq D \leq 0.11$ and $R(D) = 0$ otherwise. Now, consider the subset $\mathcal{L} = \{\mathcal{L}_n\}_{n=1}^{\infty}$ with

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

The size of this subset satisfies $\log|\mathcal{L}_n|/n \to 0.69$ as $n \to \infty$. We will show in Example 2 of Section VII that

$$R_{\mathcal{L}}(D) \leq \text{l.c.e.}\Big(\min\Big\{0.44 - H_b(D), (0.91+D)H_b\Big(\frac{0.09-D}{0.91+D}\Big) - 0.09H_b\Big(\frac{D}{0.09}\Big)\Big\}\Big),$$

if $0 \leq D \leq 0.09$, where l.c.e. stands for the lower convex envelope operation, and $R_{\mathcal{L}}(D) = 0$ if $D > 0.09$. In particular, the optimal lossless compression rate is $R_{\mathcal{L}}(D = 0) = 0.43$ [8]. As can be seen, there is no immediate connection between these results, the size of the subset, and the entropy or rate-distortion function of the source.

[1] While the expected distortion $\mathbb{E}[d(X^n, Y^n)]$ is more preferred as the evaluation metric for a lossy source code [1], we adopt the more stringent requirement of vanishing excess-distortion probability as in [16].

III. Compression of Likely Subsets

In this section, we establish our first general result asserting that for *likely* subsets, ones with not so small probability, the optimal lossy compression rate for the subset is identical to that of the original source.

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

$$\liminf_{n\to\infty} \frac{1}{n} \log P_{X^n}[X^n \in \mathcal{L}_n] = 0,$$

the subset rate-distortion function is $R_{\mathcal{L}}(D) = R(D)$.

Theorem 1 is specially interesting and subtle for subsets with slowly vanishing probability. The main idea of the proof is to construct subset codes from appropriately selected standard source codes and vice versa.

Proof: (Achievability) Fix an arbitrary $\epsilon > 0$. Choose an error-exponent-optimal lossy source code in the conventional setting for source P with rate $R(D)+\epsilon$ and $\Pr[d(X^n, Y^n) > D] \to 0$ exponentially fast as $n \to \infty$, so that [16]:

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

Noting that

$$\Pr[d(X^n, Y^n) > D] \geq \Pr[X^n \in \mathcal{L}_n] \cdot \Pr[d(X^n, Y^n) > D | X^n \in \mathcal{L}_n],$$

and that by assumption $\Pr[X^n \in \mathcal{L}_n] \to c > 0$ or $\Pr[X^n \in \mathcal{L}_n] \to 0$ sub-exponentially, we conclude that the same lossy source code, when constrained to only sequences within $\mathcal{L}_n$, achieves $\Pr[d(X^n, Y^n) > D | X^n \in \mathcal{L}_n] \to 0$ as $n \to \infty$. This implies $R_{\mathcal{L}}(D) \leq R(D)$, as the choice of ϵ is arbitrary.

(Converse) Fix an arbitrary lossy subset code achieving some rate R with excess-distortion probability $\Pr[d(X^n, Y^n) > D | X^n \in \mathcal{L}_n] = \epsilon_n \to 0$ as $n \to \infty$. We can consider this code as a conventional lossy source code for the entire space $\mathcal{X}^n$ which maps all sequences in $(\mathcal{X}^n - \mathcal{L}_n)$ to an arbitrary sequence $y_0^n \in \mathcal{Y}^n$ with a distortion not exceeding $D_{\max}$ by assumption. We can analyze the excess-distortion probability as follows.

$$\begin{aligned}
&\Pr[d(X^n, Y^n) > D] \\
&= \Pr[X^n \in \mathcal{L}_n] \cdot \Pr[d(X^n, Y^n) > D | X^n \in \mathcal{L}_n] \\
&\quad + \Pr[X^n \notin \mathcal{L}_n] \cdot \Pr[d(X^n, Y^n) > D | X^n \notin \mathcal{L}_n] \\
&\leq \epsilon_n \cdot \Pr[X^n \in \mathcal{L}_n] + \Pr[X^n \notin \mathcal{L}_n] \\
&= 1 - (1 - \epsilon_n) \cdot \Pr[X^n \in \mathcal{L}_n]. \qquad (2)
\end{aligned}$$

Since $\Pr[X^n \in \mathcal{L}_n] \to c > 0$ or $\Pr[X^n \in \mathcal{L}_n] \to 0$ sub-exponentially with n, the excess-distortion probability of this lossy source code is at least sub-exponentially away from 1. We know, however, that strong converse holds for the lossy compression of a DMS, so that the excess-distortion probability of any lossy source code with rate *below* the rate-distortion function, $R < R(D)$, approaches one [16]:

$$\limsup_{n\to\infty} \frac{1}{n} \log(1 - \Pr[\mathcal{E}(D)]) \leq - \min_{Q} \left[D(Q\|P) + |R(Q, D) - R|^+ \right].$$

Therefore, (2) implies that the rate R is *above* the rate-distortion function $R(D)$. Since the choice of the lossy subset code is arbitrary, this proves that $R_{\mathcal{L}}(D) \geq R(D)$. ∎

IV. Compression of Smooth Subsets

In the following, we state optimal compression rate results for *smooth* subsets as already defined in [8]. Note that, these subsets include ones with exponentially small probability, which are not likely per Section III. Before stating the results, let us recall some definitions for smooth subsets from [8] and introduce new ones needed here for lossy compression analysis. In the following definitions, let $T^n[Q]_{\delta_n}$ denote the set of Q-*typical* sequences, for any given distribution $Q(x)$ and any positive δ_n satisfying the Delta Convention [16].

Definition 1. *[8] We say the subset $\mathcal{L} = \{\mathcal{L}_n \subseteq \mathcal{X}^n\}_{n=1}^{\infty}$ intersects a distribution $Q(x)$ and write $\mathcal{L} \cap T[Q] \neq \emptyset$ if*

$$\limsup_{n\to\infty} |\mathcal{L}_n \cap T^n[Q]_{\delta_n}| \neq 0.$$

In such a case, if it holds for some $H_{\mathcal{L}}(Q)$ and ξ_n that

$$\left| \frac{1}{n} \log |\mathcal{L}_n \cap T^n[Q]_{\delta_n}| - H_{\mathcal{L}}(Q) \right| \leq \xi_n,$$

with $\xi_n \to 0$ as $n \to \infty$, then $H_{\mathcal{L}}(Q)$ is called the subset-$\mathcal{L}$ entropy *of the distribution $Q(x)$.*

Definition 2. *[8] We say the subset $\mathcal{L} = \{\mathcal{L}_n\}_{n=1}^{\infty}$ is smooth if the subset entropy $H_{\mathcal{L}}(Q)$ exists and is continuous in all distributions Q intersecting the subset, $\mathcal{L} \cap T[Q] \neq \emptyset$.*

We have shown in [8] that, the optimal lossless compression of a smooth subset is closely tied to the notion of *subset-typical distributions* Q_X^* that minimize the function

$$g_P(Q) := \mathbb{H}(Q) - H_{\mathcal{L}}(Q) + D(Q\|P), \qquad (3)$$

so that the set of all such subset-typical distributions Q_X^* are

$$\mathcal{Q}_X^* = \operatorname*{arg\,min}_{Q: \mathcal{L} \cap T[Q] \neq \emptyset} [\mathbb{H}(Q) - H_{\mathcal{L}}(Q) + D(Q\|P)]. \qquad (4)$$

Such distributions correspond to typical sets (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 the residual part outside the subset, captured by the term $(\mathbb{H}(Q) - H_{\mathcal{L}}(Q))$, is also small.

In the following, we introduce the conditional version of the aforementioned definitions using the notion of conditional typical sequences. We heavily make use of the notation $T^n[P_{Y|X}|x^n]_{\delta_n}$ as the set of *conditional $P_{Y|X}$-typical* sequences given $x^n \in \mathcal{X}^n$, for any conditional distribution $P_{Y|X}(y|x)$ and any positive δ_n satisfying the Delta Convention [16]. Recall that the size of the conditional typical set for *all* $x^n \in T^n[Q_X]_{\delta_n}$ satisfies [16]

$$\left| \frac{1}{n} \log |T^n[P_{Y|X}|x^n]_{\delta_n}| - \mathbb{H}(P_{Y|X}|Q_X) \right| \leq \epsilon_n, \qquad (5)$$

where $\epsilon_n \to 0$ as $n \to \infty$.

We can now define the notion of conditional subset entropy.

Definition 3. *Consider any arbitrary subset $\mathcal{L} = \{\mathcal{L}_n \subseteq \mathcal{X}^n\}_{n=1}^{\infty}$, any arbitrary distribution $Q_X(x)$, and any arbitrary*

conditional distribution $P_{Y|X}(y|x)$. Let $Q_Y(y)$ be the induced distribution $Q_Y(y) = \sum_x Q_X(x)P_{Y|X}(y|x)$. We say the auxiliary subset $\bar{\mathcal{L}} = \{\bar{\mathcal{L}}_n \subseteq \mathcal{Y}^n\}_{n=1}^{\infty}$ is $(Q_X(x), P_{Y|X}(y|x), \mathcal{L})$-smooth if it holds for some $H_{\bar{\mathcal{L}}}(Q_Y)$, some $H_{\bar{\mathcal{L}}|\mathcal{L}}(P_{Y|X}|Q_X)$ and some ξ_n that

$$\left|\frac{1}{n}\log\left|T^n[Q_Y]_{\delta_n} \cap \bar{\mathcal{L}}_n\right| - H_{\bar{\mathcal{L}}}(Q_Y)\right| \leq \xi_n,$$

and

$$\Big|\min_{x^n \in \mathcal{L}_n \cap T^n[Q_X]_{\delta_n}} \frac{1}{n}\log\left|\bar{\mathcal{L}}_n \cap T^n[P_{Y|X}|x^n]_{\delta_n}\right| - H_{\bar{\mathcal{L}}|\mathcal{L}}(P_{Y|X}|Q_X)\Big| \leq \xi_n, \quad (6)$$

such that $\xi_n \to 0$ as $n \to \infty$. In such a case, $H_{\bar{\mathcal{L}}|\mathcal{L}}(P_{Y|X}|Q_X)$ is called the conditional subset entropy of $P_{Y|X}$ given Q_X and the subsets $\mathcal{L}$ and $\bar{\mathcal{L}}$.

The quantity $H_{\bar{\mathcal{L}}}(Q_Y)$ is a subset entropy with respect to the auxiliary subset $\bar{\mathcal{L}}$ on the $\mathcal{Y}$ domain. Therefore, it satisfies the property $0 \leq H_{\bar{\mathcal{L}}}(Q_Y) \leq \mathbb{H}(Q_Y)$ [8].

Comparing expressions (5) and (6) suggests that, the conditional subset entropy $H_{\bar{\mathcal{L}}|\mathcal{L}}(P_{Y|X}|Q_X)$ is a dual of the conventional conditional entropy $\mathbb{H}(P_{Y|X}|Q_X)$. In fact, we can readily observe the appealing property that $0 \leq H_{\bar{\mathcal{L}}|\mathcal{L}}(P_{Y|X}|Q_X) \leq \mathbb{H}(P_{Y|X}|Q_X)$ for any pair of distributions $P_{Y|X}$ and Q_X. In particular, for the extreme case of $\mathcal{L}_n = \mathcal{X}^n$ and $\bar{\mathcal{L}}_n = \mathcal{Y}^n$, we have $H_{\bar{\mathcal{L}}|\mathcal{L}}(P_{Y|X}|Q_X) = \mathbb{H}(P_{Y|X}|Q_X)$ for all pairs of distributions $P_{Y|X}$ and Q_X.

We also need a further continuity condition on the conditional subset entropy as introduced in the following definition.

Definition 4. *Consider any arbitrary subset $\mathcal{L} = \{\mathcal{L}_n \subseteq \mathcal{X}^n\}_{n=1}^{\infty}$ and any arbitrary conditional distribution $P_{Y|X}(y|x)$. We say the auxiliary subset $\bar{\mathcal{L}} = \{\bar{\mathcal{L}}_n \subseteq \mathcal{Y}^n\}_{n=1}^{\infty}$ is $(P_{Y|X}(y|x), \mathcal{L})$-smooth if (i) the subset $\bar{\mathcal{L}}$ is $(Q_X(x), P_{Y|X}(y|x), \mathcal{L})$-smooth in the sense of Definition 3 for all distributions $Q_X(x)$ in a δ_n-neighborhood of all subset-$\mathcal{L}$-typical distributions $Q_X^*(x)$ as defined in* (4) *for some δ_n satisfying the Delta-Convention, and (ii) the corresponding subset entropy $H_{\bar{\mathcal{L}}}(Q_Y)$ and the conditional subset entropy $H_{\bar{\mathcal{L}}|\mathcal{L}}(P_{Y|X}|Q_X)$ are continuous in all those $Q_X(x)$ distributions.*

We now state our lossy compression results for smooth subsets.

Theorem 2. *For a discrete memoryless source $P(x)$, the rate-distortion function of any smooth subset $\mathcal{L} = \{\mathcal{L}_n \subseteq \mathcal{X}^n\}_{n=1}^{\infty}$ satisfies*

$$R_{\mathcal{L}}(D) \leq \max_{Q_X^* \in \mathcal{Q}_X^*} \min_{P_{Y|X}:\mathbb{E}[d(X^*,Y^*)]\leq D} \min_{\bar{\mathcal{L}}:(P_{Y|X},\mathcal{L})\text{-smooth}} I_{\mathcal{L},\bar{\mathcal{L}}}(Q_X^*, P_{Y|X}), \quad (7)$$

where: $\mathcal{Q}_X^*$ is the set of all subset-typical distributions as defined in (4)*; $I_{\mathcal{L},\bar{\mathcal{L}}}(Q_X^*, P_{Y|X})$ is the subset mutual information*

$$I_{\mathcal{L},\bar{\mathcal{L}}}(Q_X^*, P_{Y|X}) := H_{\bar{\mathcal{L}}}(Q_Y^*) - H_{\bar{\mathcal{L}}|\mathcal{L}}(P_{Y|X}|Q_X^*), \quad (8)$$

with the notations as in Definitions 3 and 4; and the pair of random variables (X^*, Y^*) are distributed according to $Q_X^*(x)P_{Y|X}(y|x)$ so that

$$\mathbb{E}[d(X^*, Y^*)] = \sum_{x,y} Q_X^*(x)P_{Y|X}(y|x)d(x,y).$$

Proof: Proof is provided in Section VI. ∎

Theorem 2 presents a dual of the classical rate-distortion result (1) for a DMS. The result mainly states that a certain *subset mutual information* $I_{\mathcal{L},\bar{\mathcal{L}}}(Q_X^*, P_{Y|X})$ is critical to this achievability result for lossy compression of smooth subsets. As in the standard rate-distortion result (1), a key part is minimization of this mutual information over conditional distributions $P_{Y|X}(y|x)$ satisfying the expected distortion constraint $\mathbb{E}[d(X^*, Y^*)] \leq D$. The fact that subset-typical distributions $Q_X^*(x)$ play a role in this subset rate-distortion result has an intuition similar to that for the lossless case, so that the balance between the source statistics P and the subset structure $\mathcal{L}$ determines the subset-typical sequences that must be covered by the lossy subset code, and if multiple subset-typical distributions $Q_X^*(x)$ exist, one must code for the worst case, thereby the $\max_{Q_X^* \in \mathcal{Q}_X^*}$ term in (7).

The last key element of our lossy compression result in Theorem 2 is the choice of an auxiliary subset $\bar{\mathcal{L}} = \{\bar{\mathcal{L}}_n \subseteq \mathcal{Y}^n\}_{n=1}^{\infty}$ which is $(P_{Y|X}(y|x), \mathcal{L})$-smooth and minimizes the subset mutual information $I_{\mathcal{L},\bar{\mathcal{L}}}(Q_X^*, P_{Y|X})$. Since the original subset $\mathcal{L} = \{\mathcal{L}_n \subseteq \mathcal{X}^n\}_{n=1}^{\infty}$ is considered to be smooth, the $(P_{Y|X}(y|x), \mathcal{L})$-smoothness condition essentially requires the auxiliary subset $\bar{\mathcal{L}} = \{\bar{\mathcal{L}}_n \subseteq \mathcal{Y}^n\}_{n=1}^{\infty}$ to *preserve* the structure of $\mathcal{L}$ under the stochastic transformation $P_{Y|X}(y|x)$. On the other hand, since we aim at minimizing the subset mutual information $I_{\mathcal{L},\bar{\mathcal{L}}}(Q_X^*, P_{Y|X})$, we require the auxiliary subset $\bar{\mathcal{L}}$ to be large enough to prevent an empty intersection $\bar{\mathcal{L}}_n \cap T^n[P_{Y|X}|x^n]_{\delta_n}$ for all $x^n \in T_{\mathcal{L}}^n[Q_X^*]_{\delta_n}$ and therefore an infinite conditional subset entropy $H_{\bar{\mathcal{L}}|\mathcal{L}}(P_{Y|X}|Q_X^*)$, but also small enough to achieve a small intersection size $\left|T^n[Q_Y^*]_{\delta_n} \cap \bar{\mathcal{L}}_n\right|$ and thus a small subset entropy $H_{\bar{\mathcal{L}}}(Q_Y^*)$. Hence, the optimal auxiliary subset $\bar{\mathcal{L}} = \{\bar{\mathcal{L}}_n \subseteq \mathcal{Y}^n\}_{n=1}^{\infty}$ should be a *good image* of the original subset $\mathcal{L} = \{\mathcal{L}_n \subseteq \mathcal{X}^n\}_{n=1}^{\infty}$ in terms of the scaling of the size of $\mathcal{L}$ under the stochastic transformation $P_{Y|X}(y|x)$.

An immediate but possibly suboptimal selection for the auxiliary subset $\bar{\mathcal{L}}$ is $\bar{\mathcal{L}}_n = \mathcal{Y}^n$ for all n. In this case, the subset entropy and conditional subset entropy reduce to standard (Shannon) entropy and conditional entropies, respectively, which readily give the following achievable rate-distortion result.

Corollary 1. *For a discrete memoryless source $P(x)$, the rate-distortion function for any smooth subset $\mathcal{L} = \{\mathcal{L}_n \subseteq \mathcal{X}^n\}_{n=1}^{\infty}$ satisfies*

$$R_{\mathcal{L}}(D) \leq \max_{Q_X^* \in \mathcal{Q}_X^*} R(Q_X^*, D)$$

where $\mathcal{Q}_X^*$ is the set of all subset-typical distributions as defined in (4)*, and $R(Q_X^*, D)$ is the standard rate-distortion function* (1) *for distribution $Q_X^*(x)$.*

An interesting special case, for which the achievability in Corollary 1 is tight, is one in which the subset *fully* intersects

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 permutation groups, formally called *type classes* [16], this motivates the following definition.

Definition 5. *A subset $\mathcal{L}=\{\mathcal{L}_n \subseteq \mathcal{X}^n\}_{n=1}^{\infty}$ 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 $n = 1, 2, \cdots$.*

For such symmetric subsets, the role of subset structure vanishes, and a standard rate-distortion code is intuitively sufficient for the lossy compression of the subset. By stating a proof of converse, we show that the achievable rate-distortion in Corollary 1 is optimal for the case of *smooth symmetric* subsets for which Q_X^* is unique. Hence, we find the following simpler characterization for such subsets.

Theorem 3. *Consider a discrete memoryless source $P(x)$ and any smooth symmetric subset $\mathcal{L} = \{\mathcal{L}_n \subseteq \mathcal{X}^n\}_{n=1}^{\infty}$ as in Definition 5 for which the solution to*

$$Q_X^* = \underset{Q:\mathcal{L}\cap T[Q]\neq\emptyset}{\arg\min} D(Q\|P)$$

is unique. Then, the rate-distortion function for the subset $\mathcal{L}$ is

$$R_{\mathcal{L}}(D) = R(Q_X^*, D), \tag{9}$$

where $R(Q_X^*, D)$ is the standard rate-distortion function (1) *for distribution $Q_X^*(x)$.*

Proof: Proof is provided in Section VI. ∎

Remark 2. As a sanity check, we observe that for the case $\mathcal{L}_n = \mathcal{X}^n$, which is smooth and symmetric, the subset-typical distribution is uniquely given by $Q_X^* \equiv P$. Therefore, the characterization (9) and the more general bound (7) on the subset rate-distortion function reduce to the standard rate-distortion function (1). It is worth mentioning that, one could also arrive at the same result via Theorem 1 for likely subsets, since for this case $P_{X^n}[X^n \in \mathcal{L}_n] = 1$ for all n.

V. Fluctuating Subsets

In this section, we consider *fluctuating* subsets which are constructed by superimposing several subsets, so that the resulting subset takes the structure of each component for certain time indices. In such cases, one should code for the *worst* subset component as described below. Before stating our result, let us formally define these subsets.

Definition 6. *Consider a finite collection of subsets $\mathcal{L}_j = \{\mathcal{L}_{j,n}\}_{n=1}^{\infty}$ with $1 \leq j \leq J$ as well as a finite collection of index subsequences $n_j = \{n_{j,k}\}_{k=1}^{\infty}$ with $1 \leq j \leq J$ such that for each $n = 1, 2, \cdots$ we have $n = n_{j,k}$ for a unique pair (j, k). We say $\mathcal{L} = \{\mathcal{L}_n\}_{n=1}^{\infty}$ is an $(\mathcal{L}_j, n_j)_{j=1}^{J}$-fluctuating subset when $\mathcal{L}_n = \mathcal{L}_{j,n}$ if $n \in \{n_{j,k}\}_{k=1}^{\infty}$.*

We are now ready to state our result for fluctuating subsets.

Theorem 4. *Consider a discrete memoryless source $P(x)$ and an $(\mathcal{L}_j, n_j)_{j=1}^{J}$-fluctuating subset. Then, the rate-distortion function for the subset $\mathcal{L}$ is*

$$R_{\mathcal{L}}(D) = \max_{1\leq j\leq J} R_{\mathcal{L}_j}(D).$$

Proof: (Achievability) Fix an arbitrary $\epsilon > 0$. For each $1 \leq j \leq J$, let $\{(f_{j,n}, \phi_{j,n})\}_{n=1}^{\infty}$ be the optimal encoder and decoder sequence for lossy compression of the subset $\mathcal{L}_j$ with distortion D, achieving a rate $R_{\mathcal{L}_j}(D) + \epsilon$ with vanishing excess-distortion probability $\Pr[d(X^n, Y_j^n) > D | X^n \in \mathcal{L}_{j,n}] \to 0$ as $n \to \infty$, where $Y_j^n = \phi_{j,n}(f_{j,n}(X^n))$. We consider the following code for the fluctuating subset: let $f_n \equiv f_{j,n}$ and $\phi_n \equiv \phi_{j,n}$ if $n \in \{n_{j,k}\}_{k=1}^{\infty}$. Then, we have

$$\begin{aligned}&\limsup_{n\to\infty} \Pr[d(X^n, Y^n) > D | X^n \in \mathcal{L}_n] \\ &= \max_{1\leq j\leq J} \limsup_{n\to\infty} \Pr[d(X^n, Y_j^n) > D | X^n \in \mathcal{L}_{j,n}] = 0.\end{aligned}$$

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

(Converse) Assume $R < \max_{1\leq j\leq J} R_{\mathcal{L}_j}(D)$, then there exists at least one $1 \leq \bar{j} \leq J$ such that $R < R_{\mathcal{L}_{\bar{j}}}(D)$. By the definition of $R_{\mathcal{L}_{\bar{j}}}(D)$, any arbitrary compression code for subset $\mathcal{L}_{\bar{j}}$ will satisfy $\limsup_{n\to\infty} \Pr[d(X^n, Y_j^n) > D | X^n \in \mathcal{L}_{\bar{j},n}] > 0$. Hence,

$$\begin{aligned}&\limsup_{n\to\infty} \Pr[d(X^n, Y^n) > D | X^n \in \mathcal{L}_n] \\ &\quad \geq \limsup_{n\to\infty} \Pr[d(X^n, Y_j^n) > D | X^n \in \mathcal{L}_{\bar{j},n}] > 0,\end{aligned}$$

which proves the converse for the fluctuating subset $\mathcal{L}$ and completes the proof of Theorem 4. ∎

In particular, if all components of a fluctuating subset are likely, smooth, or symmetric, we can build on Theorems 1, 2, and 3 to readily get explicit bounds.

VI. Proofs for Smooth Subsets

In this section, we state the achievability proof of our lossy compression result for smooth subsets, Theorem 2, and then the (strong converse) proof of the lossy result for symmetric smooth subsets, Theorem 3. Before starting with the proofs, we make a quick hint at the notion of type classes which we use a few times in our statements and proofs below.

Definition 7. *[16] 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*

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

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 $\hat{P}$.*

One recalls from the *method of types* [16] that, the number of the distinct types in $\mathcal{X}^n$ is only polynomial in n and does not exceed $(n+1)^{|\mathcal{X}|}$, a result referred to as the Type Counting Lemma. In the following, we frequently use the notations

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

for the intersection of subset $\mathcal{L}_n \subseteq \mathcal{X}^n$ with type class $T^n(\hat{P})$ and typical set $T^n[Q]_{\delta_n}$, respectively.

The proof of Theorem 2 on lossy compression of smooth subsets builds upon the following lemma, which is a dual of the *Type Covering Lemma* [16, Lemma 9.1] and states the rate

sufficient for the lossy compression of the intersection of the subset of interest with a single type class.

Lemma 1. *(The Subset-Type Covering Lemma) For any type $\hat{P}_X(x)$ of sequences in $\mathcal{X}^n$, any smooth subset $\mathcal{L} = \{\mathcal{L}_n \subseteq \mathcal{X}^n\}_{n=1}^{\infty}$, any distortion measure $d(x,y)$, any target distortion level $D \geq 0$, and any arbitrary constant δ, there exists a set $B(\hat{P}_X, \mathcal{L}) \subseteq \mathcal{Y}^n$ that satisfies*

$$d(x^n, B(\hat{P}_X, \mathcal{L})) := \min_{y^n \in B(\hat{P}_X, \mathcal{L})} d(x^n, y^n) \leq D, \ \forall x^n \in T_{\mathcal{L}}^n(\hat{P}_X),$$

for sufficiently large n, and whose size is bounded as

$$\frac{1}{n}\log|B(\hat{P}_X, \mathcal{L})| \leq \min_{P_{Y|X}:\mathbb{E}[d(X,Y)]\leq D} \ \min_{\bar{\mathcal{L}}:(\hat{P}_X, P_{Y|X}, \mathcal{L})\text{-smooth}} I_{\mathcal{L},\bar{\mathcal{L}}}(\hat{P}_X, P_{Y|X}) + 3\xi_n, \tag{10}$$

where $\xi_n \to 0$, and $\bar{\mathcal{L}} := \{\bar{\mathcal{L}}_n \subseteq \mathcal{Y}^n\}_{n=1}^{\infty}$ is an $(\hat{P}_X, P_{Y|X}, \mathcal{L})$-smooth auxiliary subset per Definition 3; $I_{\mathcal{L},\bar{\mathcal{L}}}(\hat{P}_X, P_{Y|X})$ is defined in (8)*; and the expected distortion is calculated with respect to the type distribution,*

$$\mathbb{E}[d(X,Y)] = \sum_{x,y} \hat{P}_X(x) P_{Y|X}(y|x) d(x,y).$$

Proof: See the Appendix. ∎

We are now ready to prove Theorem 2 with elements similar to the proof of error exponents for the standard rate-distortion problem [16, Theorem 9.5] and our proof of the lossless subset compression [8].

Proof: (of Theorem 2) Fix an arbitrary $\epsilon > 0$ and consider the following code for the subset $\mathcal{L} = \{\mathcal{L}_n \subseteq \mathcal{X}^n\}_{n=1}^{\infty}$. Using the Subset-Type Covering lemma above, we aim the lossy compression of the following set of x^n sequences.

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

where

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

Our lossy subset code consists of the following reconstructions sequences:

$$B(\mathcal{L}) := \bigcup_{\hat{P}_X:n\text{-type}, \hat{P}_X\in\Omega(3\epsilon)} B(\hat{P}_X, \mathcal{L}),$$

where $B(\hat{P}_X, \mathcal{L})$ is the cover set for $T_{\mathcal{L}}^n(\hat{P}_X)$ as defined in the Subset-Type Covering lemma above, whose size is bounded as in (10), and satisfies $d(x^n, B(\hat{P}_X, \mathcal{L})) \leq D$ for all sequences $x^n \in T_{\mathcal{L}}^n(\hat{P}_X)$. Therefore, we get for all sequences $x^n \in \mathcal{A}_n$ that

$$d(x^n, B(\mathcal{L})) \leq d(x^n, B(\hat{P}_{x^n}, \mathcal{L})) \leq D,$$

where $\hat{P}_{x^n}$ denotes the type of the sequence x^n. We can therefore bound the excess-distortion probability as follows.

$$\begin{aligned}
\Pr[\mathcal{E}_{\mathcal{L}}(D)] &= \Pr[d(X^n, Y^n) > D | X^n \in \mathcal{L}_n] \\
&\leq \Pr[X^n \notin \mathcal{A}_n | X^n \in \mathcal{L}_n] \\
&= \frac{\Pr[X^n \in (\mathcal{A}_n^c \cap \mathcal{L}_n)]}{\Pr[X^n \in \mathcal{L}_n]} \\
&\leq \frac{(n+1)^{|\mathcal{X}|} 2^{-n\left[\min_{Q\notin\Omega(3\epsilon), \mathcal{L}\cap T[Q]\neq\emptyset} g_P(Q) - \epsilon_n\right]}}{2^{-n\left[\min_{Q:\mathcal{L}\cap T[Q]\neq\emptyset} g_P(Q) + \epsilon_n\right]}} \qquad (11) \\
&\leq (n+1)^{|\mathcal{X}|} 2^{-n\epsilon},
\end{aligned}$$

where (11) is a result of the following calculation via the Type Counting Lemma and [8, Lm. 1].

$$\begin{aligned}
\Pr[X^n \in (\mathcal{A}_n^c \cap \mathcal{L}_n)] &= \sum_{\hat{P}:n\text{-type}, \hat{P}\notin\Omega(3\epsilon)} P_{X^n}[X^n \in T_{\mathcal{L}}^n(\hat{P})] \\
&\leq (n+1)^{|\mathcal{X}|} \max_{Q\notin\Omega(3\epsilon), \mathcal{L}\cap T[Q]\neq\emptyset} P_{X^n}[X^n \in T_{\mathcal{L}}^n[Q]_{\delta_n}] \\
&\leq (n+1)^{|\mathcal{X}|} 2^{-n\left[\min_{Q\notin\Omega(3\epsilon), \mathcal{L}\cap T[Q]\neq\emptyset} g_P(Q) - \epsilon_n\right]}.
\end{aligned}$$

Hence, it only remains to determine the compression rate. From (10) and the Type Counting Lemma, we have

$$\begin{aligned}
\frac{1}{n}\log|B(\mathcal{L})| &= \frac{1}{n}\log \sum_{\hat{P}_X:n\text{-type}, \hat{P}_X\in\Omega(3\epsilon)} |B(\hat{P}_X, \mathcal{L})| \\
&\leq \frac{1}{n}\log\left((n+1)^{|\mathcal{X}|} \max_{\hat{P}_X:n\text{-type}, \hat{P}_X\in\Omega(3\epsilon)} |B(\hat{P}_X, \mathcal{L})|\right) \\
&\leq \max_{\hat{P}_X:n\text{-type}, \hat{P}_X\in\Omega(3\epsilon)} \ \min_{P_{Y|X}:\mathbb{E}[d(X,Y)]\leq D} \\
&\qquad \min_{\bar{\mathcal{L}}:(\hat{P}_X, P_{Y|X}, \mathcal{L})\text{-smooth}} I_{\mathcal{L},\bar{\mathcal{L}}}(\hat{P}_X, P_{Y|X}) \\
&\qquad + 3\xi_n + \frac{|\mathcal{X}|\log(n+1)}{n} \\
&\leq \max_{Q\in\Omega(3\epsilon)} \min_{P_{Y|X}:\mathbb{E}[d(X,Y)]\leq D} \ \min_{\bar{\mathcal{L}}:(P_{Y|X}, \mathcal{L})\text{-smooth}} I_{\mathcal{L},\bar{\mathcal{L}}}(Q, P_{Y|X}) + 5\xi_n,
\end{aligned}$$

where the last line follows from the continuity of the subset mutual information $I_{\mathcal{L},\bar{\mathcal{L}}}(Q, P_{Y|X})$ for all distributions in a neighborhood of the subset-typical distributions. Since $n \to \infty$ and the choice of $\epsilon > 0$ is arbitrary, this completes the proof of Theorem 2. ∎

In the following, we prove Theorem 3 on lossy compression of smooth symmetric subsets. The achievability immediately follows from Corollary 1, and the converse given below is analogous to that for the standard rate-distortion theorem [16, Theorem 7.3] and uses the following two technical lemmas, whose proofs are omitted due to space limitations.

The first technical lemma is a generalized asymptotic equipartition property (AEP) and a dual of [16, Lm. 2.12] which asserts that essentially all of the probability mass of a smooth subset is concentrated only in the subset-typical sets.

Lemma 2. *Consider a discrete memoryless source $P(x)$ and a smooth subset $\mathcal{L} = \{\mathcal{L}_n \subseteq \mathcal{X}^n\}_{n=1}^{\infty}$ with subset-typical distributions $\mathcal{Q}_X^*$ as given in* (4)*. Then, there exists*

a sequence $\epsilon_n \to 0$ as $n \to \infty$ such that

$$\Pr\left[X^n \in \bigcup_{Q_X^* \in \mathcal{Q}_X^*} T_{\mathcal{L}}^n[Q_X^*]_{\delta_n} \middle| X^n \in \mathcal{L}_n\right] \geq 1 - \epsilon_n.$$

The second technical lemma is a dual of [16, Lm. 2.14] and states that, when constrained to only a subset of the source realizations, any set with high probability has a size essentially no smaller than the subset-typical set.

Lemma 3. *Consider a discrete memoryless source $P(x)$ and a smooth subset $\mathcal{L} = \{\mathcal{L}_n \subseteq \mathcal{X}^n\}_{n=1}^{\infty}$ for which the subset-typical distribution $Q_X^*(x)$ per* (4) *is unique. Given any $0 < \eta < 1$, there exists a sequence $\epsilon_n \to 0$ as $n \to \infty$ such that, if a set $\mathcal{A} \subseteq \mathcal{X}^n$ satisfies*

$$\Pr\left[X^n \in \mathcal{A} \,|\, X^n \in \mathcal{L}_n\right] \geq \eta,$$

then

$$|\mathcal{A}| \geq 2^{n[H_{\mathcal{L}}(Q_X^*) - \epsilon_n]}.$$

We are now ready to prove the result for smooth symmetric subsets.

Proof: (of Theorem 3) We only state the proof of (strong) converse, since the achievability readily follows from Corollary 1 as well as the fact that for smooth symmetric subsets the function $g_P(Q)$ defined in (3) reduces to $D(Q\|P)$.

Consider any arbitrary lossy source code for subset $\mathcal{L}$ that uses M codewords and satisfies

$$\Pr\left[d\left(X^n, \phi(f(X^n))\right) \leq D \,|\, X^n \in \mathcal{L}_n\right] \geq 1 - \epsilon,$$

for a potentially non-vanishing $0 < \epsilon < 1$. Define the set $\mathcal{A}$ as follows:

$$\mathcal{A} := \left\{x^n \in T_{\mathcal{L}}^n[Q_X^*]_{\delta_n} : d(x^n, \phi(f(x^n))) \leq D\right\}.$$

Then, Lemma 2 and the simple inequality $\Pr[A \cap B] \geq \Pr[A] - \Pr[B^c]$ imply

$$\Pr\left[X^n \in \mathcal{A} \,|\, X^n \in \mathcal{L}_n\right] \geq 1 - \epsilon - \tau_n,$$

for some $\tau_n \to 0$, which, on account of Lemma 3, yields

$$|\mathcal{A}| \geq \exp\left(n\left[\mathbb{H}(Q_X^*) - \epsilon_n\right]\right), \tag{12}$$

since $H_{\mathcal{L}}(Q_X) = \mathbb{H}(Q_X)$ for all distributions $Q_X(x)$ intersecting the smooth symmetric subset $\mathcal{L}$. On the other hand, define the set of reconstruction codewords corresponding to the set $\mathcal{A}$ as

$$\mathcal{C} := \left\{y^n \in \mathcal{Y}^n : y^n = \phi(f(x^n)) \text{ for some } x^n \in \mathcal{A}\right\},$$

and accordingly decompose the set $\mathcal{A}$ as follows.

$$\mathcal{A} := \bigcup_{y^n \in \mathcal{C}} \mathcal{A}(y^n),$$

where for any fixed $y^n \in \mathcal{C}$ we have defined

$$\mathcal{A}(y^n) := \left\{x^n \in \mathcal{A} : \phi(f(x^n)) = y^n\right\}.$$

We can further decompose the sequences x^n in $\mathcal{A}(y^n)$ according to their joint type $\tilde{P}_{XY}(x,y)$ with y^n, so that

$$\mathcal{A}(y^n) = \bigcup_{\substack{\tilde{P}_{XY}(x,y):\ n\text{-joint type} \\ \mathbb{E}[d(\tilde{X},\tilde{Y})] \leq D \\ |\tilde{P}_X(x) - Q_X^*(x)| \leq \delta_n}} \left(\mathcal{A}(y^n) \cap T_{\mathcal{L}}^n(\tilde{P}_{X|Y}|y^n)\right),$$

where the constraints hold (i) since $d(x^n, \phi(f(x^n))) \leq D$ for all $x^n \in \mathcal{A}$ implies $\mathbb{E}[d(\tilde{X},\tilde{Y})] \leq D$, and (ii) since $x^n \in \mathcal{A} \subseteq T^n[Q_X^*]_{\delta_n}$ implies $|\tilde{P}_X(x) - Q_X^*(x)| \leq \delta_n$ for all $x \in \mathcal{X}$. Recalling that the size of the conditional type $T^n(\tilde{P}_{X|Y}|y^n)$ for all $y^n \in T^n(\tilde{P}_Y)$ satisfies

$$\left|T^n(\tilde{P}_{X|Y}|y^n)\right| \leq 2^{n\mathbb{H}(\tilde{P}_{X|Y}|\tilde{P}_Y)},$$

we get

$$\begin{aligned}|\mathcal{A}| &= \sum_{y^n \in \mathcal{C}} |\mathcal{A}(y^n)| \\ &\leq |\mathcal{C}| \cdot (n+1)^{|\mathcal{X}||\mathcal{Y}|} \max_{\substack{\tilde{P}_{XY}(x,y):\ n\text{-joint type} \\ \mathbb{E}[d(\tilde{X},\tilde{Y})] \leq D \\ |\tilde{P}_X(x) - Q_X^*(x)| \leq \delta_n}} 2^{n\mathbb{H}(\tilde{P}_{X|Y}|\tilde{P}_Y)}. \end{aligned} \tag{13}$$

Combining (12) and (13), we have proved that the size of any lossy source code for the symmetric smooth subset $\mathcal{L}$ satisfies

$$M \geq |\mathcal{C}| \geq (n+1)^{-|\mathcal{X}||\mathcal{Y}|} \times \exp\left(n \min_{\substack{\tilde{P}_{XY}(x,y):\ n\text{-joint type} \\ \mathbb{E}[d(\tilde{X},\tilde{Y})] \leq D \\ |\tilde{P}_X(x) - Q_X^*(x)| \leq \delta_n}} \left[\mathbb{H}(Q_X^*) - \mathbb{H}(\tilde{P}_{X|Y}|\tilde{P}_Y) - \epsilon_n\right]\right).$$

Due to the continuity of the conditional Shannon entropy, we have proved that

$$R_{\mathcal{L}}(D) \geq \min_{P_{Y|X}: \mathbb{E}[d(X^*,Y^*)] \leq D} \mathbb{I}(Q_X^*, P_{Y|X}) - 3\epsilon_n.$$

This concludes the proof of the strong converse and that of Theorem 3. ∎

VII. Numerical Examples

In this section, we revisit some of the numerical examples in [8] and investigate their lossy compression performance. In all of these examples, we consider a binary DMS, $\mathcal{X} = \{0,1\}$, with a Bernoulli distribution $B(p)$ with parameter $0 \leq p \leq 1/2$. The lossy compression is considered with respect to $\mathcal{Y} = \{0,1\}$ and the Hamming distortion. The rate-distortion function of the source is $R(D) = H_b(p) - H_b(D)$ if $0 \leq D \leq p$ and $R(D) = 0$ if $D > p$ [1]. 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 $D_b(q\|p) := q\log(q/p) + (1-q)\log\left((1-q)/(1-p)\right)$.

We first consider a symmetric example.

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

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

This subset is smooth and symmetric, and $B(q)$ is the only distribution that intersects the subset $\mathcal{L}$, so $Q_X^* = B(q)$. We obtain from Theorem 3 that

$$R_{\mathcal{L}}(D) = R(Q_X^*, D) = \begin{cases} H_b(q) - H_b(D), & 0 \leq D \leq \min\{q, \bar{q}\}, \\ 0, & D > \min\{q, \bar{q}\}, \end{cases}$$

which follows from the calculation of the standard rate-distortion function of the binary source [1]. It is evident that the subset rate-distortion function in this example can be below or

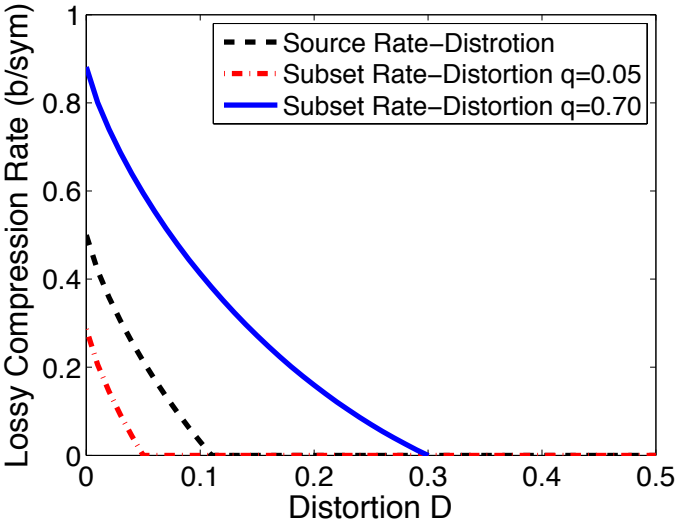

Fig. 1. Comparison of the subset rate-distortion function of Example 1 with the rate-distortion function of the source for a Bernoulli DMS with parameter $p = 0.11$.

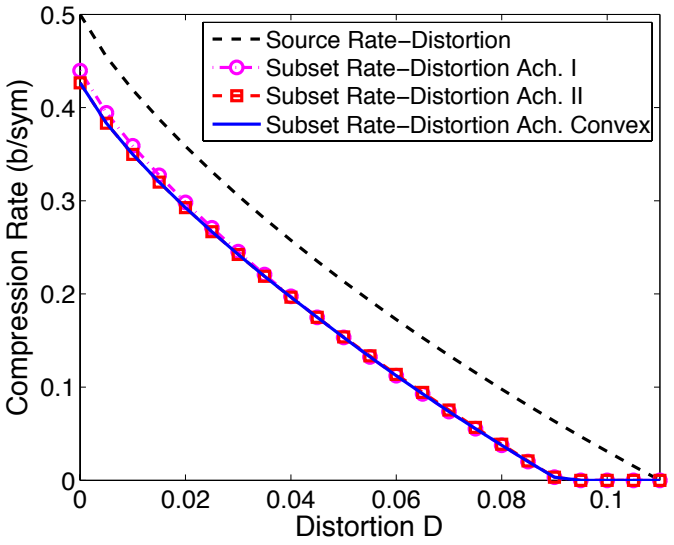

Fig. 2. Comparison of the achievable rate-distortion pair (16) and its components for the subset in Example 2 with the rate-distortion function of the source for a Bernoulli DMS with parameter $p = 0.11$.

above the rate-distortion function of the source. We illustrate this comparison in Figure 1.

Next, we focus on a smooth but non-symmetric example.

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

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

This subset is smooth, and all distributions $B(q)$ with $0 \le q \le 1/2$ intersect it. Hence, the subset-typical distribution is $Q_X^* = B(q^*)$ where [8]

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

Therefore, we can use Corollary 1 to find an achievable rate-distortion pair as follows.

$$R_{\mathcal{L}}^{(1)}(D) = \begin{cases} H_b(q^*) - H_b(D), & 0 \le D \le q^* \\ 0, & D > q^*. \end{cases} \tag{14}$$

We can also build on Theorem 2 to obtain another achievable rate-distortion pair. Let $0 \le D \le q^*$, and consider the following conditional distribution:

$$P_{Y|X}(0|0) = 1, \quad P_{Y|X}(0|1) = D/q^*,$$

so that $Q_Y^* = B(q^* - D)$. Note that, under this conditional distribution, $P_{Y|X}(1|0) = 0$ thus no 0 in x^n will flip to a 1 in y^n, hence the no-consecutive-1 structure will be preserved. Also, note that $\mathbb{E}[d(X^*, Y^*)] = \Pr[X^* \ne Y^*] = D$. Now, consider the auxiliary subset $\bar{\mathcal{L}} = \{\bar{\mathcal{L}}_n \subseteq \mathcal{Y}^n\}_{n=1}^{\infty}$ with

$$\bar{\mathcal{L}}_n := \{y^n \in \mathcal{Y}^n : y^n \text{ has no consecutive 1s}\}.$$

In this case, we get

$$H_{\bar{\mathcal{L}}}(Q_Y^*) = (1 - q^* + D)H_b\left(\frac{q^* - D}{1 - q^* + D}\right),$$

$$H_{\bar{\mathcal{L}}|\mathcal{L}}(P_{Y|X}|Q_X^*) = \lim_{n\to\infty} \frac{1}{n} \log \binom{nq^*}{nD} = q^* H_b\left(\frac{D}{q^*}\right),$$

and obtain the following achievable rate-distortion pair:

$$R_{\mathcal{L}}^{(2)}(D) = \begin{cases} (1-q^*+D)H_b\left(\frac{q^*-D}{1-q^*+D}\right) - q^* H_b\left(\frac{D}{q^*}\right), & 0 \le D \le q^* \\ 0, & D > q^*. \end{cases} \tag{15}$$

Note that, we can use time-sharing for this subset, since any portion of a sequence belonging to this subset will also have no consecutive ones. Hence, we arrive at the following result:

$$R_{\mathcal{L}}(D) \le \text{l.c.e.}\left(\min\{R_{\mathcal{L}}^{(1)}(D), R_{\mathcal{L}}^{(2)}(D)\}\right), \tag{16}$$

where l.c.e. stands for the lower convex envelope operation. This immediately implies that $R_{\mathcal{L}}(D) = 0$ for $D > q^*$, but since no converse for our Theorem 2 is currently known, we cannot guarantee that (16) is optimal for $0 \le D \le q^*$. However, Figure 2 shows that even the achievable subset rate-distortion in (16) can sometimes outperform the rate-distortion function of the original source and already provide lossy compression gains.

Finally, we present a fluctuating example with smooth components.

Example 3. Consider a subset $\mathcal{L}_1 = \{\mathcal{L}_{1,n}\}_{n=1}^{\infty}$ with

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

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

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

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

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

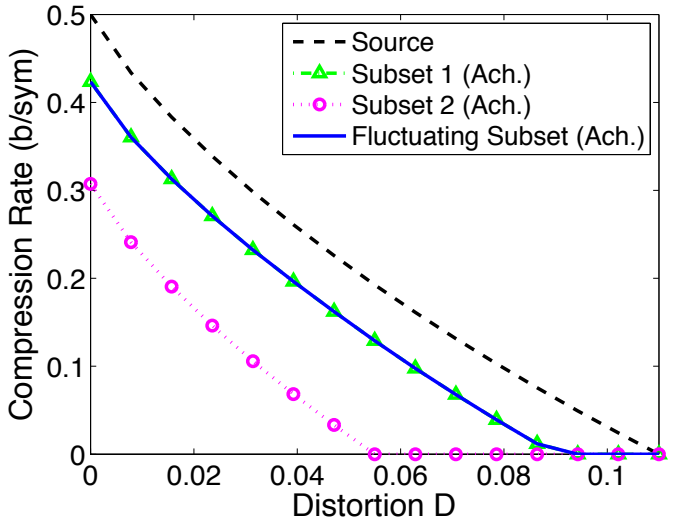

Fig. 3. The achievable rate-distortion pair via (18) and (19) for the fluctuating subset of Example 3 with parameters $q_1 = 0$, $q_2 = 0.09$ and $w_1 = 0$, $w_2 = 0.18$ and a binary DMS with parameter $p = 0.11$.

The first component $\mathcal{L}_1$ is smooth and intersects all distributions $B(q)$ with $q_1 \leq q \leq q_2$, so that its subset-typical distribution is $Q_X^{*(1)} = B(q^*)$ where [8]

$$q^* = \underset{q_1 \leq q \leq q_2}{\arg\min} \left[H_b(q) - (1-q) H_b\left(\frac{q}{1-q}\right) + D_b(q\|p) \right]. \quad (17)$$

An analysis similar to that in Example 2 implies

$$R_{\mathcal{L}_1}(D) \leq \min\{R_{\mathcal{L}_1}^{(1)}(D), R_{\mathcal{L}_2}^{(2)}(D)\}, \quad (18)$$

where $R_{\mathcal{L}_1}^{(1)}(D)$ and $R_{\mathcal{L}_2}^{(2)}(D)$ are as in (14) and (15), respectively, with q^* as given in (17). Note that, (18) is not a convex function in D, but it is also unclear whether a time-sharing argument can be applied to this subset to convexify the result, since a portion of a sequence belonging to this subset may not retain the same structure as that in the original sequence.

The second component $\mathcal{L}_2$ is also smooth and intersects all distributions $B(w)$ with $w_1 \leq w \leq w_2$, so that $Q_X^{*(2)} = B(w^*)$ where

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

From Corollary 1, we get

$$R_{\mathcal{L}_2}(D) \leq \begin{cases} H_b(w^*) - H_b(D), & 0 \leq D \leq w^* \\ 0, & D > w^*. \end{cases} \quad (19)$$

Substituting (18) and (19) in Theorem 4 yields

$$R_{\mathcal{L}}(D) \leq \max\left\{ \min\{R_{\mathcal{L}_1}^{(1)}(D), R_{\mathcal{L}_1}^{(2)}(D)\}, R_{\mathcal{L}_2}(D) \right\}. \quad (20)$$

This readily implies $R_{\mathcal{L}}(D) = 0$ for $D > \max\{q^*, w^*\}$, but the optimality of (20) for $0 \leq D \leq \max\{q^*, w^*\}$ is unknown in the absence of a converse for our Theorem 2. However, Figure 3 shows that even the achievable subset rate-distortion in (20) can sometimes outperform the rate-distortion function of the original source and already provide lossy compression gains. Furthermore, depending on the parameter selection and the distortion value, the performance of the fluctuating subset may be dominated by that of one subset component or the other.

VIII. Concluding Remarks

We have provided a framework as well as optimality results for lossy compression of likely, smooth, and fluctuating subsets of discrete memoryless sources. One of our key findings is that, lossy compression of smooth subsets involves covering the subset-typical sequences with those conditionally typical sequences of the reconstruction alphabet that belong to an auxiliary subset, which is smooth by selection. In our proposed achievability, the number of cover sequences is then related to the size of the smallest intersection of the conditional typical sets with the selected auxiliary subset. Therefore, achieving lower compression rates requires a smart selection of the auxiliary subset that is a good image of the original subset and preserves its structure.

We envision two immediate directions for future research on this topic. One is to close the open parts of our current results, namely: (i) to prove a lower bound or converse for the lossy compression of smooth subsets to complement the result in Theorem 2; (ii) to extend the complete characterization of Theorem 3 for the lossy compression of smooth symmetric subsets to the case in which the optimal subset-typical distribution $Q_X^*(x)$ is not unique. Another future direction is to analyze the fundamental compression limits of other subsets that are not covered by our current analysis.

Appendix
Proof of Lemma 1

Fix an arbitrarily small constant $\eta > 0$, and consider a pair of random variables (X, Y) such that $\mathbb{E}[d(X,Y)] \leq |D - \eta|^+$ and the X-marginal is distributed according to $\hat{P}(x)$. Denote the Y-marginal distribution by $P_Y(y)$. Fix an auxiliary subset $\bar{\mathcal{L}} = \{\bar{\mathcal{L}}_n \subseteq \mathcal{Y}^n\}_{n=1}^{\infty}$ which is $(\hat{P}_X, P_{Y|X}, \mathcal{L})$-smooth per Definition 3. We use the following random coding argument to prove the existence of the set $B(\hat{P}_X, \mathcal{L})$ as described in the lemma. Generate M independently and identically distributed (i.i.d.) sequences $\{Y^n(m)\}_{m=1}^M$ at random according to the uniform distribution over the set $\bar{\mathcal{L}}_n \cap T^n[P_Y]_{\delta_n}$; the exact value of M will be specified later in the proof. We define the set of *uncovered* x^n sequences in $T_{\mathcal{L}}^n(\hat{P})$ by the Y^n sequences as follows.

$$U\left(\{Y^n(m)\}_{m=1}^M\right) := \left\{ x^n \in T_{\mathcal{L}}^n(\hat{P}) : d\left(x^n, \{Y^n(m)\}_{m=1}^M\right) > D \right\}.$$

Our goal is to prove that, if M is chosen appropriately, then we obtain $\mathbb{E}\left[\left|U\left(\{Y^n(m)\}_{m=1}^M\right)\right|\right] < 1$ for sufficiently large n, which implies that a deterministic set $B(\hat{P}, \mathcal{L}) := \{y^n(m)\}_{m=1}^M$ with elements belonging to $\bar{\mathcal{L}}_n \cap T^n[P_Y]_{\delta_n}$ exists such that all sequences $x^n \in T_{\mathcal{L}}^n(\hat{P}_X)$ are covered in the sense that $d(x^n, B(\hat{P}_X, \mathcal{L})) \leq D$.

We first note that

$$\mathbb{E}\left[\left|U\left(\{Y^n(m)\}_{m=1}^M\right)\right|\right] = \sum_{x^n \in T_{\mathcal{L}}^n(\hat{P})} \Pr\left[d\left(x^n, \{Y^n(m)\}_{m=1}^M\right) > D\right]. \quad (21)$$

However, due to the i.i.d. generation of the sequences $\{Y^n(m)\}_{m=1}^M$, we find for all sequences in $T_{\mathcal{L}}^n(\hat{P}_X)$ that,

$$\begin{aligned}&\Pr\left[d\left(x^n,\{Y^n(m)\}_{m=1}^M\right)>D\right]\\&=\left(1-\Pr\left[d\left(x^n,Y^n(1)\right)\le D\right]\right)^M\le 2^{-M\cdot\Pr[d(x^n,Y^n(1))\le D]},\end{aligned}\tag{22}$$

from the inequality $(1-x)^n \le 2^{-nx}$ for all $0 \le x \le 1$, $n > 0$.

We now analyze the probability in (22) using the specific generation of the $\{Y^n(m)\}_{m=1}^M$ sequences. In particular, since these sequences are generated uniformly over the set $\bar{\mathcal{L}}_n \cap T^n[P_Y]_{\delta_n}$, we obtain

$$\begin{aligned}&\Pr\left[d\left(x^n,Y^n(1)\right)\le D\right]\\&\quad=\frac{|\{y^n\in\bar{\mathcal{L}}_n\cap T^n[P_Y]_{\delta_n}:d(x^n,y^n)\le D\}|}{|\bar{\mathcal{L}}_n\cap T^n[P_Y]_{\delta_n}|}\\&\quad\ge\frac{|\bar{\mathcal{L}}_n\cap T^n[P_{Y|X}|x^n]_{\delta_n}|}{|\bar{\mathcal{L}}_n\cap T^n[P_Y]_{\delta_n}|},\end{aligned}\tag{23}$$

where (23) follows from the properties of typical sequences $x^n \in T^n(\hat{P}_X)$ and $y^n \in T^n[P_{Y|X}|x^n]_{\delta_n}$ that, for sufficiently large n, we have

$$\begin{aligned}d(x^n,y^n)&\le\mathbb{E}[d(X,Y)]+|\mathcal{X}||\mathcal{Y}|\delta_n D_{\max}\\&\le(D-\eta)+|\mathcal{X}||\mathcal{Y}|\delta_n D_{\max}\le D,\end{aligned}$$

for the case $D > \eta$. This result also holds for the case $D \le \eta$, since the condition $\mathbb{E}[d(X,Y)] \le |D-\eta|^+ = 0$ implies that for all (x,y) pairs, either $d(x,y) = 0$ or $\hat{P}_X(x)P_{Y|X}(y|x) = 0$, which in turn implies $d(x^n,y^n) = 0 \le D$ for all n and all $x^n \in T^n(\hat{P}_X)$ and $y^n \in T^n[P_{Y|X}|x^n]_{\delta_n}$.

Now, recall from Definition 3 of $(\hat{P}_X, P_{Y|X}, \mathcal{L})$-smooth auxiliary subset that, for any $x^n \in T_{\mathcal{L}}^n[\hat{P}_X]_{\delta_n}$,

$$\begin{aligned}&|\bar{\mathcal{L}}_n\cap T^n[P_{Y|X}|x^n]_{\delta_n}|\\&\quad\ge\min_{x^n\in T_{\mathcal{L}}^n[\hat{P}_X]_{\delta_n}}|\bar{\mathcal{L}}_n\cap T^n[P_{Y|X}|x^n]_{\delta_n}|\\&\quad\ge 2^{n[H_{\bar{\mathcal{L}}|\mathcal{L}}(P_{Y|X}|\hat{P}_X)-\xi_n]}\end{aligned}\tag{24}$$

and that

$$|\bar{\mathcal{L}}_n\cap T^n[P_Y]_{\delta_n}|\le 2^{n[H_{\bar{\mathcal{L}}}(P_Y)+\xi_n]}.\tag{25}$$

Substituting (24) and (25) into (23) yields for all sequences in $T_{\mathcal{L}}^n(\hat{P}_X)$ that,

$$\begin{aligned}\Pr\left[d\left(x^n,Y^n(1)\right)\le D\right]&\ge 2^{-n\left[H_{\bar{\mathcal{L}}}(P_Y)-H_{\bar{\mathcal{L}}|\mathcal{L}}(P_{Y|X}|\hat{P}_X)+2\xi_n\right]}\\&=2^{-n\left[I_{\mathcal{L},\bar{\mathcal{L}}}(\hat{P}_X,P_{Y|X})+2\xi_n\right]}.\end{aligned}\tag{26}$$

Making the selection

$$M=2^{n\left[I_{\mathcal{L},\bar{\mathcal{L}}}(\hat{P}_X,P_{Y|X})+3\xi_n\right]},$$

along with (21), (22), and (26) implies

$$\begin{aligned}\mathbb{E}\left[\left|U\left(\{Y^n(m)\}_{m=1}^M\right)\right|\right]&\le\left|T_{\mathcal{L}}^n(\hat{P}_X)\right|\cdot 2^{-M\cdot 2^{-n\left[I_{\mathcal{L},\bar{\mathcal{L}}}(\hat{P}_X,P_{Y|X})+2\xi_n\right]}}\\&\le 2^{n\log|\mathcal{X}|-M\cdot 2^{-n\left[I_{\mathcal{L},\bar{\mathcal{L}}}(\hat{P}_X,P_{Y|X})+2\xi_n\right]}}\\&\le 2^{n\log|\mathcal{X}|-2^{n\xi_n}}<1,\end{aligned}$$

for sufficiently large n. Since the choice of $\eta > 0$ is arbitrary, and the pair of random variables (X,Y) can be arbitrarily selected subject to distortion and X-marginal distribution constraints, and the auxiliary subset $\bar{\mathcal{L}}$ can be any $(\hat{P}_X, P_{Y|X}, \mathcal{L})$-smooth subset, we have proved the existence of the set $B(\hat{P}_X, \mathcal{L})$ as claimed in the lemma and with size

$$\begin{aligned}&\frac{1}{n}\log B(\hat{P}_X,\mathcal{L})=\frac{1}{n}\log M\\&\le\min_{P_{Y|X}:\mathbb{E}[d(X,Y)]\le D}\ \min_{\bar{\mathcal{L}}:(\hat{P}_X,P_{Y|X},\mathcal{L})\text{-smooth}} I_{\mathcal{L},\bar{\mathcal{L}}}(\hat{P}_X,P_{Y|X})+3\xi_n.\end{aligned}$$

ACKNOWLEDGMENT

This research is sponsored in part by the Army Research Laboratory under the Network Science Collaborative Technology Alliance, Agreement Number W911NF-09-2-0053.

REFERENCES

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

[2] A. E. Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge University Press, UK, 2012.

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

[4] N. Merhav and M. Feder, "Universal prediction," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2124–2147, 1998.

[5] S. S. Pradhan and K. Ramchandran, "Distributed source coding using syndromes (DISCUS): Design and construction," *IEEE Trans. Inf. Theory*, vol. 49, no. 3, pp. 626–643, 2003.

[6] E. Tuncel, P. Koulgi, and K. Rose, "Rate-distortion approach to databases: Storage and content-based retrieval," *IEEE Trans. Inf. Theory*, vol. 50, no. 6, pp. 953–967, 2004.

[7] 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.

[8] E. MolavianJazi and A. Yener, "Subset Source Coding," in *Proc. 53rd Allerton Conference on Communications, Control, and Computing*. Monticello, IL, 2015.

[9] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. Springer Science & Business Media, 2009.

[10] A. Dembo and I. Kontoyiannis, "Source coding, large deviations, and approximate pattern matching," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1590–1615, 2002.

[11] B. Marcus, P. Siegel, and R. Roth, "An introduction to coding for constrained systems," in *Handbook of Coding Theory*, W. C. Huffman and V. Pless, Eds. New York: Elsevier, 1998.

[12] P. Jacquet and W. Szpankowski, "Markov types and minimax redundancy for Markov sources," *IEEE Trans. Inf. Theory*, vol. 50, no. 7, pp. 1393–1402, 2004.

[13] N. Merhav, "Statistical physics and information theory," *Foundations and Trends in Communications and Information Theory*, vol. 6, no. 1–2, pp. 1–212, 2009.

[14] C. Bunte and A. Lapidoth, "Encoding tasks and Rényi entropy," *IEEE Trans. Inf. Theory*, vol. 60, no. 9, pp. 5065–5076, 2014.

[15] A. Høst-Madsen, E. Sabeti, and C. Walton, "Information theory for atypical sequences," in *Proc. IEEE Information Theory Workshop (ITW)*. Honolulu, HI, 2013, pp. 1–5.

[16] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Academic, New York, 1981.

[17] R. M. Gray, *Entropy and Information Theory*. Springer Science & Business Media, 2011.

[18] T.-S. Han, *Information-Spectrum Methods in Information Theory*. Springer-Verlag, Berlin, Germany, 2003.

[19] Y. Steinberg and S. Verdú, "Simulation of random processes and rate-distortion theory," *IEEE Trans. Inf. Theory*, vol. 42, no. 1, pp. 63–86, 1996.