Subset Source Coding

Ebrahim MolavianJazi, *Member, IEEE*, and Aylin Yener, *Fellow, IEEE*

***Abstract*— This paper studies the fundamental limits of storage for structured data, where statistics and structure are both critical to the application. Accordingly, a framework is proposed for optimal lossless and lossy compression of subsets of the possible realizations of a discrete memoryless source (DMS). For the lossless subset-compression problem, it turns out that the optimal source code may not index the conventional source-typical sequences, but rather index certain subset-typical sequences consistent with the subset of interest. 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. Intuitively, subset-typical sequences belong to those typical sets which highly intersect the subset of interest but are still closest to the source distribution in the sense of relative entropy. For the lossy subset-compression problem, 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 a 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 shown in our numerical examples, more often than not, one achieves a gain in the fundamental limits, in that the optimal compression rate for the subset in both the lossless and lossy settings can be strictly smaller than the source entropy and the source rate-distortion function, respectively, although exceptions are also possible.**

***Index Terms*— Subset-typical sequences, subset entropy, subset mutual information, subset-type covering lemma, method of types, semantic information processing.**

I. Introduction

SOURCE coding addresses compression, with or without fidelity, of an information source. In particular, in (near-) lossless compression of a discrete memoryless source (DMS), one identifies and indexes *source-typical* sequences that capture essentially all the probability mass of the source. For a DMS X with probability distribution $P(x)$ over alphabet $\mathcal{X}$, the number of such typical sequences is approximately $2^{n\mathbb{H}(X)}$, so the fundamental limit of lossless compression is given by the Shannon entropy [1], [2]:

$$R^* = \mathbb{H}(X). \tag{1}$$

In lossy compression of a DMS, on the other hand, one essentially groups source-typical sequences and *covers* each group with a sequence that is within a certain distortion distance of them [1], [2]. For a DMS $X \sim P(x)$, the number of such cover sequences is approximately $2^{nR(D)}$, so the fundamental limit of lossy compression with a distortion requirement D is given by the rate-distortion function, defined as the average mutual information optimized over the set of conditional distributions that satisfy the expected distortion constraint [1], [3]:

$$R(D) = R(P, D) := \min_{P_{Y|X}:\mathbb{E}[d(X,Y)]\leq D} \mathbb{I}(X; Y). \tag{2}$$

These basic settings have been studied extensively and extended to scenarios with unknown statistics [4] and to network and distributed settings [5], [6] and find applications in database management [7], [8]. An implicit but pivotal consideration in all of these works is that important realizations of interest for the source to reconstruct consist only of the likely and source-typical sequences.

In some emerging applications in information processing including 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 [9], [10], only information with certain patterns and structures 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.

The goal of this paper is to provide a treatment of a *subset source coding* problem, where the encoder and decoder aim at providing a (near-)lossless or lossy description of only a *subset* of all possible source realizations as determined by the application. To explain the subset source coding problem more concretely, we provide in the following a motivating example with a toy setup to showcase the kind of results we obtain, and the connections with and distinctions from the standard source coding problem.

A. Motivating Example

Consider a binary DMS, $\mathcal{X} = \{0, 1\}$, with a Bernoulli distribution with parameter $\Pr[X = 1] = p = 0.11$, so that the Shannon entropy of the source is simply the binary entropy $H_b(p) = -p\log p - (1-p)\log(1-p) = 0.5$, and its

Manuscript received June 6, 2016; revised September 9, 2017; accepted June 5, 2018. Date of publication July 9, 2018; date of current version August 16, 2018. This work was supported by the U.S. Army Research Laboratory through the Network Science Collaborative Technology Alliance under Grant W911NF-09-2-0053. This paper was presented in part at the 2015 Allerton Conference on Communications, Controls, and Computing and in part at the 2016 Information Theory and Applications Workshop.

E. MolavianJazi was with the Department of Electrical Engineering, The Pennsylvania State University, University Park, PA 16802 USA. He is now with Motorola Mobility LLC, Chicago, IL 60654 USA (e-mail: ebrahim.molavian@gmail.com).

A. Yener is with the Wireless Communications and Networking Laboratory, Department of Electrical Engineering, The Pennsylvania State University, University Park, PA 16802 USA (e-mail: yener@engr.psu.edu).

Communicated by S. S. Pradhan, Associate Editor for Shannon Theory.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2018.2854601

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

$$\mathcal{L}_n := \{x^n \in \mathcal{X}^n : x^n \text{ has no consecutive 1s}\}. \tag{3}$$

The size of this subset satisfies $(1/n)\log|\mathcal{L}_n| \to 0.69$ as $n \to \infty$. We will show in Example 4 of Section VII that the optimal (near-)lossless compression rate of this subset is $R_{\mathcal{L}}^* = 0.43$, and that the rate-distortion function of this subset satisfies

$$R_{\mathcal{L}}(D) \leq \text{l.c.e.}\Bigg(\min\Bigg\{0.44 - H_b(D), (0.91 + D)H_b \times \left(\frac{0.09 - D}{0.91 + D}\right) 0.09 H_b\left(\frac{D}{0.09}\right)\Bigg\}\Bigg), \tag{4}$$

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 $D > 0.09$. It is clear that, entropy and rate-distortion for this subset is completely different from that of the original source, due to the specific structure imposed by the subset.

To shed more light on the connections between these results and the structure of this subset, we consider the following. The binary sequences in this subset can also be generated by a certain asymmetric two-state binary Markov chain [11], [12], namely, with $\alpha := \Pr[X_{n+1} = 1|X_n = 0]$ and $\beta := \Pr[X_{n+1} = 0|X_n = 1] = 1$. Accordingly, we can consider an infinite continuous set of *Markovian types* (i.e., a type with a with a Markov structure) [13] with $0 \leq \alpha \leq 1$, and so with different cardinalities, that satisfy the constraint in $\mathcal{L}_n$. The entropy-rate for such a Markovian type can be calculated as [1]

$$\mathcal{H}(X) = \frac{1}{\alpha + 1} H_b(\alpha), \tag{5}$$

If one tries to choose an α that maximizes the entropy rate $\mathcal{H}(X)$, one gets $\alpha = 0.38$, which leads to $\mathcal{H}(X) = 0.69$, which is indeed the exponent of the subset size we derived above. However, if we evaluate (4) at $D = 0$, we get $R(D = 0) = 0.43$ coming from the second term in (4).

The reason why $R(D = 0)$ is strictly (and by far) smaller than the entropy-rate $\mathcal{H}(X) = \lim_{n\to\infty}(1/n)\log|\mathcal{L}_n|$ calculated above is that, the Markovian type with $\alpha = 0.38$ is not "the most likely type class within the subset" due to the bias introduced by the prior of Bernoulli($p = 0.11$). In other words, the generating distribution is closer, in the sense of KL divergence, to a Markovian type class that is of smaller cardinality than the Markov type class with the most elements. This phenomenon biases which elements of $\mathcal{L}_n$ need to be encoded; cf., Figure 1 and its discussion for more details. It turns out that the parameter for the optimal Markovian type class is $\alpha^* = 0.0995$, for which the corresponding stationary distribution satisfies $\Pr[X = 1] = 0.09$ and the corresponding entropy rate is $1/(1 + \alpha^*)H_b(\alpha^*) = 0.09/0.91$.

One notes that, there is a second $\alpha = 0.78$ that generates the same entropy-rate value, for which the stationary distribution has $\Pr[X = 1] = 0.44$. Compared to the optimal solution with $\Pr[X = 1] = 0.09$, this type class is further from the generating distribution Bernoulli($p = 0.11$) in a KL divergence sense.

Finally, for further insight into the issues and knobs in this example, consider for a moment the situation with a Bernoulli $p = \Pr[X = 1] = 0.18$ that satisfies $H_b(0.18) = 0.68$. In such a situation, one can show that the optimal Markovian type indeed satisfies $\alpha^* = 0.38$, namely the largest Markovian type that satisfies the constraint of subset $\mathcal{L}_n$. Therefore, in the original setting with $p = 0.11$, the prior distribution naturally pulls the most likely Markov type away from the one with the largest cardinality.

B. Background

Previous efforts sharing similar motivations as in this work include task encoding in [14] that guarantees certain important but less likely source events are not ignored in data compression, and information theory of atypical sequences in [15] with applications in signal processing and big-data analytics.

The subset source coding problem inherently involves both probabilistic and combinatorial aspects. On one hand, it has roots in large deviations theory [16] and relates to the generalized asymptotic equipartition property (AEP) [17] and Sanov's theorem [1], and particularly to the conditional limit theorem (CoLT) [1], [18], maximum entropy distribution [19], Gibbs conditioning principle [20], and conditional law of large numbers [21]. We discuss the latter relations in more details in Section III-B.

On the other hand, the subset source coding problem has a combinatorial element in terms of the exponential number of information sequences that satisfy certain structural constraints, and therefore relates to the capacity for magnetic recording channels with constrained coding [22]–[24]; the notion of Markov types in compression of Markovian sources [25], [26]; and entropy definitions in statistical mechanics models [27].

C. Outline and Contributions

In Section II, we formally introduce the subset source coding problem in both lossless and lossy versions. In Section III, we discuss two possible alternative approaches for this problem via (i) the Verdú-Han information spectrum approach [28] and (ii) the conditional limit theorem (CoLT) [1] with the quasi-independence feature [18]. In this paper, we instead blueprovide a rather elementary analysis from first principles of the method of types [29] and large deviations theory [1], [16] along with elements of combinatorics for the analysis. In Section IV, we use error exponent results for conventional source coding to state our first general result for *likely subsets*, those with not(-so-fast)-vanishing probabilities. In Section V, we extend the notion of typical sequences and present optimality results for a broad class of *smooth subsets*, those satisfying certain regularity conditions and continuous structures. In Section VI, we prove optimal compression rates for *fluctuating subsets* that alternate between several structures. Our key contributions in these three main sections are as follows.

- For likely subsets, we prove an achievability and a matching strong converse to show that the fundamental

limits of lossy and lossless compression of the subset are equal to those of the original source.
- For smooth subsets, we prove an achievability and a strong converse for the lossless case which shows that the fundamental limit 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, both defined in this paper. For 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 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 a converse to show that the fundamental limits of lossy and lossless compression of the subset are equal to those of the *worst* structure, i.e., the one which requires the highest compression rate.

We next present in Section VII several numerical examples of the subset source coding problem which suggest, when focusing only on a subset instead of the entire source, there is often a gain in the compression rate, although there are exceptions. We devote Section VIII to a generalization of our framework to the case of subsets with weighted priorities, which has relations to the problem of unequal error protection in channel coding [30], [31]. We conclude the paper in Section IX with a recap of the results, some discussions regarding the computability of our results, and a few remarks about possible extensions. The proofs of the main results are presented in the main text, while those of the technical underlying lemmas are relegated to the Appendices.

In the following, we would like to briefly highlight the main novel features of our work.
- One key novel aspect of our work is posing a new basic setup in the source coding literature, including the problem formulation and the performance metrics.
- We have provided a set of rather elementary proofs from first principles, which are quite readable for a broad audience in information theory, but at the same time does not appear to sacrifice the extent to which performance results can be developed, e.g., compared to potential results that one can obtain using CoLT.
- We have treated, among other cases, fluctuating subsets and subsets with weighted priorities, that do not appear to be (directly) handled by existing versions of CoLT.
- The lossy compression aspects of subset source coding and the techniques we have proposed, e.g., the subset-type covering lemma using "auxiliary subsets" as structure-preserving images of the original subsets (cf. Lemma 2), are novel concepts and contributions that are interesting on their own.

Notation: We use capital letters X and Y to denote random variables, and lower case letters x and y to denote their realizations. We use calligraphic letters $\mathcal{X}$ and $\mathcal{Y}$ to denote sets or alphabets. We use P_X and Q_Y to denote marginal distributions, and $P_{Y|X}$ to denote conditional distributions. We use $\hat{P}_X$, $\hat{P}_{Y|X}$, and $\hat{P}_{XY}$ to denote types, conditional types, and joint types, respectively. Throughout this paper, all log operations are understood as base 2. We follow the notation of Csiszár and Körner [29] for denoting entropy and mutual information. Consider a random variable X with distribution $P(x)$. The Shannon entropy $\mathbb{H}(X)$ is denoted by

$$\mathbb{H}(P) := -\sum_{x\in\mathcal{X}} P(x)\log P(x). \tag{6}$$

Analogously, consider a random variable X with marginal distribution $P_X(x)$, and let Y be a random variable conditionally distributed according to $P_{Y|X}(y|x)$. The conditional Shannon entropy $\mathbb{H}(Y|X)=\sum_x P_X(x)\mathbb{H}(Y|X=x)$ is denoted by

$$\mathbb{H}(P_{Y|X}|P_X) := \sum_{x\in\mathcal{X}} P_X(x)\mathbb{H}(P_{Y|X=x}). \tag{7}$$

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

$$\begin{aligned}\mathbb{I}(P_X, P_{Y|X}) &:= \mathbb{H}(P_Y)-\mathbb{H}(P_{Y|X}|P_X)\\ &= \mathbb{H}(P_X)-\mathbb{H}(P_{X|Y}|P_Y),\end{aligned} \tag{8}$$

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

$$D(Q\|P) := \sum_{x\in\mathcal{X}} Q(x)\log\frac{Q(x)}{P(x)}. \tag{9}$$

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

$$P_{X^n}(x^n)=\prod_{t=1}^{n} P_X(x_t). \tag{10}$$

For simplicity, 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. We wish to find the minimum (near-)lossless and lossy compression rate for the subset sequence $\mathcal{L}$.

More formally, an $(n,2^{nR})$ near-lossless (or simply, lossless) code for subset $\mathcal{L}$ consists of an encoder $m:\mathcal{L}_n\to\{1,2,\cdots,2^{nR}\}$ and a decoder $\hat{x}^n:\{1,2,\cdots,2^{nR}\}\to\mathcal{L}_n\cup\{\mathcal{E}\}$ that assigns to an index $1\le m\le 2^{nR}$ 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

$$\Pr[\mathcal{E}_{\mathcal{L}}] := \Pr[\hat{X}^n\neq X^n|X^n\in\mathcal{L}_n]. \tag{11}$$

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

For the lossy subset compression problem, we 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

$$d(x^n, y^n) := \frac{1}{n} \sum_{t=1}^{n} d(x_t, y_t). \tag{12}$$

An $(n, 2^{nR})$ lossy 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

$$\Pr[\mathcal{E}_{\mathcal{L}}(D)] := \Pr[d(X^n, Y^n) > D | X^n \in \mathcal{L}_n]. \tag{13}$$

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 interpretation of the conditioning used in the problem formulation above is that the encoder never sees or cares about the source realizations outside the subset $\mathcal{L}$. Note that, the prior distribution $P_{X^n}(x^n)$ induces a prior on the subset $\mathcal{L}_n$; cf., Section III-A for mode details.

Remark 2: The subset rate-distortion function $R_{\mathcal{L}}(D)$, similar to the standard rate-distortion function, is a non-increasing function of D, by definition. The convexity of $R_{\mathcal{L}}(D)$ in D, however, is not a priori obvious. The latter would normally build on a time-sharing argument, namely, to combine shorter codes achieving distortions D_1 and D_2 with appropriate rates R_1 and R_2, respectively, to form a longer code that achieves the convex combination of those rate-distortion pairs. However, such an argument 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 $0 < \alpha < 1$.

III. Alternative Approaches

In this section, we discuss two alternative analysis and proof approaches for the subset source coding problem, namely, the information spectrum approach and the conditional limit theorem approach, and argue that these two approaches although interesting, are quite challenging and their application to the subset source coding might not be straightforward.

A. The Information Spectrum Approach

At the outset, one may think that a conditional source formulation can readily capture the subset compression problem. In particular, one can define an equivalent conditional source $\tilde{X}^n$ as

$$P_{\tilde{X}^n}(x^n) := \frac{P_{X^n}(x^n)}{P_{X^n}[X^n \in \mathcal{L}_n]} 1\{x^n \in \mathcal{L}_n\}, \tag{14}$$

and claim the fundamental lossless and lossy compression rates of this conditional source to be equivalent to our $R^*_{\mathcal{L}}$ and $R_{\mathcal{L}}(D)$ of interest, respectively. This claim is indeed valid, since the error probability and the excess-distortion probability for both cases are the same, as readily shown in Appendix A.

The fundamental compression limits of this equivalent conditional source, however, are not in general very straightforward to analyze. For those subsets for which the equivalent conditional source is stationary and ergodic, the fundamental lossless and lossy compression limits are given by average entropy rate and average mutual information rate, respectively [1], [28], [32], [33]:

$$\begin{aligned} R^*_{\mathcal{L}} &= \lim_{n\to\infty} \frac{1}{n} \mathbb{H}(\tilde{X}^n), \\ R_{\mathcal{L}}(D) &= \lim_{n\to\infty} \inf_{Y^n : \frac{1}{n}\mathbb{E}[d(\tilde{X}^n, Y^n)] \leq D} \frac{1}{n} \mathbb{I}(\tilde{X}^n; Y^n). \end{aligned} \tag{15}$$

However, the stationarity and ergodicity assumptions do not hold for most subsets, even the simplest ones such as our Example 1 in Section VII; cf. [28, Example 1.5.1]. Therefore, one would need to utilize the more advanced information-spectrum approach [28] to characterize the fundamental compression limits of this potentially non-stationary and non-ergodic equivalent source. In particular, for the lossless case, the fundamental limit is given by the *spectral sup-entropy rate* of the conditional source [34]:

$$R^*_{\mathcal{L}} = \bar{H}(\tilde{\mathbf{X}}) := \text{p-}\limsup_{n\to\infty} \frac{1}{n} \log \frac{1}{P_{\tilde{X}^n}(\tilde{X}^n)}, \tag{16}$$

where $\tilde{\mathbf{X}} = \{\tilde{X}^n\}_{n=1}^{\infty}$ is the equivalent conditional source process, and the p-lim sup operation, limit superior in probability, is defined as the supremum of the support set of the limiting distribution [28]. Analogously, for the lossy case, the fundamental limit is given by the *spectral sup-mutual information rate* of the conditional source [35]:

$$R_{\mathcal{L}}(D) = \inf_{\mathbf{Y} : \bar{d}(\tilde{\mathbf{X}}, \mathbf{Y}) \leq D} \bar{I}(\tilde{\mathbf{X}}; \mathbf{Y}), \tag{17}$$

where $\mathbf{Y} = \{Y^n\}_{n=1}^{\infty}$ is a reconstruction process and

$$\begin{aligned} \bar{I}(\tilde{\mathbf{X}}; \mathbf{Y}) &:= \text{p-}\limsup_{n\to\infty} \frac{1}{n} \log \frac{P_{Y^n|\tilde{X}^n}(Y^n|\tilde{X}^n)}{P_{Y^n}(Y^n)}, \\ \bar{d}(\tilde{\mathbf{X}}, \mathbf{Y}) &:= \text{p-}\limsup_{n\to\infty} \frac{1}{n} d(\tilde{X}^n, Y^n). \end{aligned} \tag{18}$$

Although the above limiting analysis and information-spectrum approach yield a complete characterization of the fundamental compression limits, its numerical evaluation for arbitrary subsets is cumbersome and may require tedious manipulations. Moreover, the general form of the fundamental limit results in (15), (16), (17) are not explicit about the effect of subset structure and the statistics of the original source on the compression rate, and each example needs to be individually analyzed. In Sections IV, V, and VI, we present a more accessible form of treatment and give three tractable optimality results that apply to broad classes of subsets.

[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 [29].

B. The Conditional Limit Theorem Approach

Large deviations (LD) is indeed a closely related area to the subset source coding problem we propose in this paper. In particular, Sanov's theorem and conditional limit theorem (CoLT) are two key results in LD that have strong connections with our problem. In this subsection, we make these connections clearer: we briefly state these two results, argue how one could potentially use CoLT to tackle the problem of subset source coding, and explain the challenges for using such a CoLT-based approach.

As was hinted in the Introduction and will be also discussed in Section IV, for most cases of interest, the subset $\mathcal{L}$ is an unlikely or *rare event*, in that $\Pr[X^n \in \mathcal{L}_n]$ is vanishing exponentially fast with n. Treating rare events is the essence of large deviations. One of the key results in LD is Sanov's theorem [1] that considers a set E of probability distributions, possibly with additional regularity conditions (e.g., that E is the closure of its interior). Then for i.i.d. random variables $X^n \sim P$, one gets $-(1/n)\log \Pr[X^n \in E] \to D(Q^*\|P)$ as $n \to \infty$, where $Q^* = \arg\min_{Q\in E} D(Q\|P)$ is called the (generalized) I-projection of P on the set E [1], [18]. An important class of such sets E is defined by the sample mean or empirical block average constraint $(1/n)\sum_{i=1}^{n} g(X_i) > \alpha$ for a given function $g(x)$ and a given constant α, usually satisfying $\alpha > \mathbb{E}_P[g(X)]$. Also, an extension is possible to handle the intersection of multiple such constraints.

A second key result in LD is to study conditional distributions given a rare event, which is referred to as the conditional limit theorem (CoLT). The standard version of CoLT, e.g., per [1], states that for an i.i.d. $X^n \sim P$ and for a closed convex set E of probability measures, such that $P \notin E$, we have [1]

$$\Pr[X_1 = a_1, X_2 = a_2, \ldots, X_m = a_m | P_{X^n} \in E] \to \prod_{i=1}^{m} Q^*(a_i) \text{ in probability}$$

for *fixed* m as $n \to \infty$, where P_{X^n} denotes the empirical block average or type of X^n, and Q^* is the (generalized) I-projection of P on the set E (defined above). This basically means that, conditioned on the event that type of X^n belongs to the set E, the first few elements of X^n are *asymptotically conditionally independent* with common distribution Q^*.

Recall from our problem formulation in Section II that, the kind of conditional probabilities that we are interested in are of the form $\Pr[X^n \in \mathcal{A}_n | X^n \in \mathcal{L}_n]$, for which the conditioning part, in some cases, might reduce to a type (or Markovian type) constraint $\Pr[X^n \in \mathcal{A}_n | P_{X^n} \in E_n]$. This is in fact very similar to what the CoLT result addresses, except for the fact that here we have $m = n$.

However, an important observation is that, the standard version of CoLT does not guarantee independence for long sequences: "[asymptotic conditional independence] *is not true* for $m = n$, since there are end effects; given that the type of the sequence is in E, the last elements of the sequence can be determined from the remaining elements, and the elements are no longer independent." [1, pp. 374–375]. In fact, the asymptotic conditional independence property presented in the standard version of CoLT is much more limited than just the case $m = n$ mentioned by Cover and Thomas [1]. Dembo and Zeitouni [20] discuss that, to get such an asymptotic conditional independence property, one could go beyond a fixed m and also extend to the case of $m = m_n$ being a function of the blocklength n, but the speed of growth should satisfy $m_n \frac{\log n}{n} \to 0$ as $n \to \infty$ or sometimes $m_n = o(n)$, but there is no hope for conditional independence beyond that growth speed. These results suggest that the standard form of CoLT cannot be applied to the subset source coding problem.

Nonetheless, a more relaxed result is possible if one is satisfied with *almost independence*. Csiszár [18] proves that, for certain (almost completely convex) sets E of probability measures, i.i.d. random variables $(X_1, \ldots, X_n) \sim P$ under the condition $P_{X^n} \in E$ are *asymptotically quasi-independent* with limiting distribution Q^* (defined above), namely, $\lim_{n\to\infty}(1/n)D(P_{X^n|E}\|(Q^*)^n) = 0$. It follows that, in the words of Csiszar (with slight changes in notation), "whatever probabilistic statement holds, except for an event of exponentially small probability, for i.i.d. RV's with common distribution Q^*, it holds …with conditional probability tending to 1 for $(X_1, \ldots, X_n)$ given that $P_{X^n} \in E$" [18].

The above-mentioned advanced form of CoLT (with the quasi-independence feature) is indeed an incredible result that can potentially serve as an alternative approach to tackle the subset source coding problem. We believe such a solution would consist of the following steps: (i) Identify and prove an appropriately general form of advanced CoLT that can handle potentially general/arbitrary subsets $\mathcal{L}_n$; (ii) Specialize the CoLT result to the problem at hand (i.e., the error events or the excess-distortion events) including appropriate change of measures to Q^*; and finally (iii) Incorporate any additional steps needed for the subset compression problem, e.g., rate calculation via counting arguments, and distortion analysis via covering lemmas. Note that, in our view, the third step is rather independent of the first two steps so, regardless of the approach taken for the first two steps (CoLT-based or not), appropriate techniques and solutions need to be devised (such as the subset type covering lemma we have developed in Section V-C2) which are inherent to the subset source coding problem, and therefore novel and interesting on their own.

We suspect such an alternative CoLT-based approach, although very interesting and valuable, can be quite challenging. Firstly, the mere development of advanced CoLT results with the quasi-independence feature for (rather) generic and arbitrary sets E (that directly capture, e.g., fluctuating subsets or other n-dependent structures), instead of an almost completely convex set E of probability measures, appears to be difficult. In fact, based on a rather detailed check of the citations of Csiszár [18] within the literature of probability and information theory, very few works have focused on the quasi-independence property and its extension for more general E sets [36]–[38]. Secondly, it is not obvious to us that, wherever the appropriate form of advanced CoLT is already known or is newly characterized, such a CoLT-based approach would lead to results far beyond what we have developed here with a rather elementary analysis. In fact, in some sense, one

could conceive our proof methods as special cases of (the existing or a potential) CoLT for the problem at hand.

IV. Compression of Likely Subsets

In this section, we establish our first result asserting that for *likely* subsets, ones with not so small probability, not so unexpectedly, the optimal lossless and lossy compression rates for the subset turn out identical to those of the original source.

Theorem 1: For a discrete memoryless source $P(x)$ and any subset $\mathcal{L} = \{\mathcal{L}_n \subseteq \mathcal{X}^n\}_{n=1}^{\infty}$ that satisfies

$$\liminf_{n\to\infty} \frac{1}{n} \log P_{X^n}[X^n \in \mathcal{L}_n] = 0, \tag{19}$$

the optimal lossless subset-compression rate is $R_{\mathcal{L}}^* = \mathbb{H}(P)$ and the subset rate-distortion function is $R_{\mathcal{L}}(D) = R(D)$. In particular, the result holds if $P_{X^n}[X^n \in \mathcal{L}_n]$ as $n \to \infty$ either converges to a constant or decays sub-exponentially to zero.

Theorem 1 is more intuitive for subsets $\mathcal{L}$ with an asymptotically constant probability so that $P_{X^n}[X^n \in \mathcal{L}_n] \to c$ 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 and the case that subset probability does not converge at all but an asymptotic lower bound to the subset probability is constant or decaying at most sub-exponentially to zero are somewhat more subtle, as explained in the following proof of Theorem 1. The main idea is to construct subset codes from appropriately selected source codes and vice versa.

Proof: Below, we first provide the proof for the lossless case. Then, we describe the few changes needed to make the proof work for the lossy case. First note that, from the definition of liminf, the assumption in (19) implies that, for any $\epsilon > 0$, we have $\Pr[X^n \in \mathcal{L}_n] > \exp(-n\epsilon)$ for large enough n.

(Achievability) Fix an arbitrary $\epsilon > 0$. Choose an error-exponent-optimal lossless source code in the conventional setting for source P with rate $\mathbb{H}(P)+\epsilon$ and $\Pr[\hat{X}^n \neq X^n] \to 0$ exponentially fast as $n \to \infty$, so that [29]

$$\lim_{n\to\infty} \frac{1}{n} \log \Pr[\hat{X}^n \neq X^n] \leq - \min_{Q:\mathbb{H}(Q)\geq R} D(Q\|P). \tag{20}$$

Noting that

$$\Pr[\hat{X}^n \neq X^n] \geq \Pr[X^n \in \mathcal{L}_n] \cdot \Pr[\hat{X}^n \neq X^n | X^n \in \mathcal{L}_n], \tag{21}$$

and that by assumption $\Pr[X^n \in \mathcal{L}_n] > \exp(-n\epsilon)$, we conclude that the same lossless source code, when constrained to only sequences within $\mathcal{L}_n$, achieves $\Pr[\hat{X}^n \neq X^n | X^n \in \mathcal{L}_n] \to 0$ as $n \to \infty$. This implies $R_{\mathcal{L}}^* \leq \mathbb{H}(P)$, as the choice of ϵ is arbitrary.

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

$$\begin{aligned}
&\Pr[\hat{X}^n \neq X^n] \\
&\quad = \Pr[X^n \in \mathcal{L}_n] \cdot \Pr[\hat{X}^n \neq X^n | X^n \in \mathcal{L}_n] \\
&\qquad + \Pr[X^n \notin \mathcal{L}_n] \cdot \Pr[\hat{X}^n \neq X^n | X^n \notin \mathcal{L}_n] && (22)\\
&\quad \leq \epsilon_n \cdot \Pr[X^n \in \mathcal{L}_n] + \Pr[X^n \notin \mathcal{L}_n] && (23)\\
&\quad = 1 - (1-\epsilon_n) \cdot \Pr[X^n \in \mathcal{L}_n]. && (24)
\end{aligned}$$

Since $\Pr[X^n \in \mathcal{L}_n] > \exp(-n\epsilon)$, the error probability of this code is at least sub-exponentially away from 1. We know, however, that 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, $R < \mathbb{H}(P)$, approaches one [29]

$$\limsup_{n\to\infty} \frac{1}{n} \log(1 - \Pr[\hat{X}^n \neq X^n]) \leq - \min_{Q:\mathbb{H}(Q)\leq R} D(Q\|P). \tag{25}$$

Therefore, (24) implies that the rate R is above the source entropy $\mathbb{H}(P)$. Since the choice of the lossless code is arbitrary, this proves that $R_{\mathcal{L}}^* \geq \mathbb{H}(P)$. [2]

The proof for the lossy case is identical after making the following changes: use $R(Q, D)$ instead of $\mathbb{H}(Q)$; $R(D)$ instead of $\mathbb{H}(P)$; $R_{\mathcal{L}}(D)$ instead of $R^*(D)$; and the excess distortion event $d(X^n, Y^n) > D$ instead of the error event $\hat{X}^n \neq X^n$. Also note [29], [39] for error exponent results for standard lossy compression, including the following result about the excess-distortion probability of any lossy source code with rate below the rate-distortion function, $R < R(D)$:

$$\begin{aligned}
&\limsup_{n\to\infty} \frac{1}{n} \log(1 - \Pr[d(X^n, Y^n) > D]) \\
&\qquad \leq - \min_{Q} \left[D(Q\|P) + |R(Q, D) - R|^+ \right]. && (26)
\end{aligned}$$

□

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 be tempted to think that subsets with non-negligible probability are already addressed by this theorem and, since the remaining possible subsets with exponentially small probability are so rare, their analysis is not very relevant. In particular, one may be tempted to think that such subsets with *negligible* probability only contain the *atypical* sequences of the source, which are ignored in conventional compression anyway. However, as will be clarified in the remainder of the paper, one can find subsets containing source-typical sequences, which yet have an exponentially small probability; see Section VII. Moreover, as discussed in the Introduction, even the atypical sequences of the source may be important for certain applications.

V. Compression of Smooth Subsets

In this section, we state optimal compression rate results for a broad class of *smooth* subsets, ones with continuous

[2] An alternative proof of converse for the lossless case follows from [29, Lemma 2.14 and Problem 2.11].

structures, including subsets with exponentially small probability. We present the results for the lossless and lossy cases in Subsections V-A and V-B, respectively, and provide the proofs in Subsection V-C.

A. Lossless Compression of Smooth Subsets

In this subsection, we state our lossless compression result for smooth subsets. Our result relies on a new quantity termed the *subset entropy*, to introduce which we first recall the definition and properties of standard (source-) typical sequences of a DMS.

In the following definitions, let $N(x;x^n)$ be the number of occurrences of the symbol $x \in \mathcal{X}$ in the sequence x^n.

Definition 1 [29]: Given any distribution $Q(x)$ and any positive δ_n, the set $T^n[Q]_{\delta_n}$ of Q-typical *sequences is defined as the set of all sequences $x^n \in \mathcal{X}^n$ that satisfy*

$$\left|\frac{1}{n}N(x;x^n) - Q(x)\right| \leq \delta_n, \tag{27}$$

for all $x \in \mathcal{X}$ with $Q(x) > 0$ and $N(x;x^n) = 0$ otherwise.

Remark 3: In the definition above and throughout the paper, the sequence δ_n is assumed to satisfy the Delta-Convention, i.e., as $n \to \infty$, we have $\delta_n \to 0$ and $\sqrt{n}\delta_n \to \infty$ [29]. One recalls from the properties of the typical sequences that, for every distribution $Q(x)$, the size of the Q-typical set satisfies [29]

$$\lim_{n\to\infty} \frac{1}{n} \log |T^n[Q]_{\delta_n}| = \mathbb{H}(Q). \tag{28}$$

We can now define the notion of subset entropy.

Definition 2: 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

$$\limsup_{n\to\infty} |\mathcal{L}_n \cap T^n[Q]_{\delta_n}| \neq 0. \tag{29}$$

Remark 4: In condition (29) of Definition 2, for any fixed distribution $Q(x)$, we may have empty intersection for several (or even many) values of n, but this of course would not violate the original assumption that $\Pr[X^n \in \mathcal{L}_n] \neq 0$ for all n.

Definition 3: Consider a subset $\mathcal{L} = \{\mathcal{L}_n \subseteq \mathcal{X}^n\}_{n=1}^{\infty}$ that intersects a distribution $Q(x)$, i.e., $\mathcal{L} \cap T[Q] \neq \emptyset$. A constant $H_{\mathcal{L}}(Q)$ is called the subset-$\mathcal{L}$ entropy *of distribution $Q(x)$ if*

$$\lim_{n\to\infty} \frac{1}{n} \log \left|\mathcal{L}_n \cap T^n[Q]_{\delta_n}\right| = H_{\mathcal{L}}(Q), \tag{30}$$

provided that the limit exists.

Comparing expressions (28) and (30) suggests that, subset entropy $H_{\mathcal{L}}(Q)$ is an analog of the standard entropy $\mathbb{H}(Q)$. In fact, we readily observe the appealing property $0 \leq H_{\mathcal{L}}(Q) \leq \mathbb{H}(Q)$ for any distribution Q with $\mathcal{L} \cap T[Q] \neq \emptyset$. In particular, for $\mathcal{L}_n = \mathcal{X}^n$, we have $H_{\mathcal{L}}(Q) = \mathbb{H}(Q)$ for all distributions Q.

Our focus in this section is on *smooth* subsets, ones for which the subset entropy is a continuous function, essentially suggesting that the subset intersects only a *continuous spectrum* of distributions and nothing outside of it.

Definition 4: We say the subset $\mathcal{L} = \{\mathcal{L}_n \subseteq \mathcal{X}^n\}_{n=1}^{\infty}$ is smooth if the subset entropy $H_{\mathcal{L}}(Q)$ exists and is continuous in all distributions Q intersecting the subset, $\mathcal{L} \cap T[Q] \neq \emptyset$.

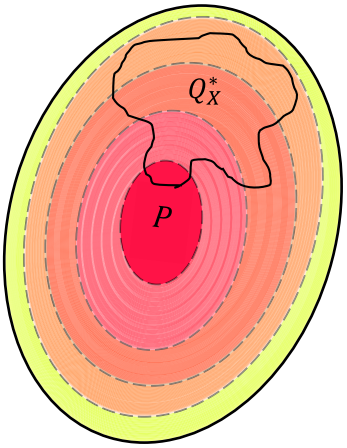

Fig. 1. Schematic description of Theorem 2 for lossless compression of smooth subsets. The subset is depicted with a curved shape. The dashed rings denote the typical sets, which are shown in the order of closeness to the source statistic P in the sense of relative entropy. The subset-typical distribution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

In the following, we state our lossless compression result for smooth subsets.

Theorem 2: For a discrete memoryless source $P(x)$, the optimal lossless compression rate for any smooth subset $\mathcal{L} = \{\mathcal{L}_n \subseteq \mathcal{X}^n\}_{n=1}^{\infty}$ is

$$R_{\mathcal{L}}^* = \max_{Q_X^* \in \mathcal{Q}_X^*} H_{\mathcal{L}}(Q_X^*), \tag{31}$$

where the set $\mathcal{Q}_X^*$ is defined as

$$\mathcal{Q}_X^* = \arg\min_{Q:\mathcal{L}\cap T[Q]\neq\emptyset} [\mathbb{H}(Q) - H_{\mathcal{L}}(Q) + D(Q\|P)], \tag{32}$$

$$\mathcal{Q}_X^* = \arg\min_{Q:\mathcal{L}\cap T[Q]\neq\emptyset} g_P(Q), \tag{33}$$

with the function $g_P(Q)$ given by

$$g_P(Q) = \mathbb{H}(Q) - H_{\mathcal{L}}(Q) + D(Q\|P). \tag{34}$$

Proof: Proof is provided in Section V-C1. □

Theorem 2 has an interesting interpretation in terms of a tension between the source statistics and the subset structure. It suggest that, within a given subset, the most likely sequences of the source which should be indexed by a lossless subset code do not necessarily belong to the *source-typical* set with distribution P. Rather, they belong 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 relatively small. The *subset-typical* distributions Q_X^* optimize the trade-off between these two elements by minimizing the function $g_P(Q)$ introduced in (34), and the size of the corresponding subset-typical set dictates $H_{\mathcal{L}}(Q_X^*)$ to be the rate of the lossless compression code for this subset. In most cases, there is only a single minimizing distribution Q_X^*, so the set of subset-typical distributions $\mathcal{Q}_X^*$ has a single element, but in case there are multiple minimizers, one should code for the worst case, thereby the $\max_{Q_X^* \in \mathcal{Q}_X^*}$ term in (31). This interpretation is schematically depicted in Figure 1.

As a sanity check for Theorem 2, 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 (34) reduces to $g_P(Q) = D(Q\|P)$ which is minimized by $Q_X^* = P$ for which $H_{\mathcal{L}}(Q_X^*) = \mathbb{H}(Q_X^*) = \mathbb{H}(P)$, which is consistent with first impressions. 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 the case of symmetric subsets, as defined below, where the subset either fully intersects with a *type class* [29] or does not intersect at all.

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 $n = 1, 2, \cdots$.*

One readily observes that, a symmetric subset $\mathcal{L}$ satisfies the property $H_{\mathcal{L}}(Q) = \mathbb{H}(Q)$ for any distributions Q intersecting the subset, i.e., $\mathcal{L} \cap T[Q] \neq \emptyset$. This is because the subset is fully intersecting a type class, and from the properties proved in the Method of Types [13], [29], one knows that the size of a type class $T_n(Q)$ is (on an exponential scale) equal to $\exp(n\mathbb{H}(Q))$. Therefore, a symmetric subset $\mathcal{L}$ is smooth if $\mathbb{H}(Q)$ is continuous in all distributions Q intersecting the subset, $\mathcal{L} \cap T[Q] \neq \emptyset$. In such a case, the objective function (34) reduces to $g_P(Q) = D(Q\|P)$. Hence, we arrive at the following simpler expression.

Corollary 1: For a discrete memoryless source $P(x)$, the optimal lossless compression rate for any smooth symmetric subset $\mathcal{L} = \{\mathcal{L}_n \subseteq \mathcal{X}^n\}_{n=1}^{\infty}$ is

$$R_{\mathcal{L}}^* = \max_{Q_X^* \in \mathcal{Q}_X^{symm}} \mathbb{H}(Q_X^*), \tag{35}$$

where the set $\mathcal{Q}_X^{symm}$ is defined as

$$\mathcal{Q}_X^{symm} = \operatorname*{arg\,min}_{Q:\mathcal{L}\cap T[Q]\neq\emptyset} D(Q\|P). \tag{36}$$

Remark 5: It turns out that, in the context of subset source coding, we may face initially counterintuitive situations with $R_L^* > R^* = \mathbb{H}(X)$. That is, we may need more than $\mathbb{H}(X) = \mathbb{H}(P_X)$ bits for lossless compression of certain subsets; see, e.g., Examples 1 and 3 in Section VII. The reason for this phenomenon is that, within the subset compression framework, the typical sequences that we have to code empirically follow the *subset-typical distribution* $Q_X^*(x)$ rather than the *source-typical distribution* $P_X(x)$. Since all such Q_X^*-typical sequences are statistically (almost) similar, we need to index all of them for lossless compression. Now, if $Q_X^*(x)$ is a *more uniform* distribution than $P_X(x)$, then $\mathbb{H}(Q_X^*) > \mathbb{H}(X)$, so that the total number of Q_X^*-typical sequences, $2^{n\mathbb{H}(Q_X^*)}$, is larger than the number of $P_X(x)$-typical sequences, $2^{n\mathbb{H}(X)}$. This reasoning is sufficient for symmetric subsets; cf. Definition 5. For non-symmetric subsets, the subset structure must also be taken into account, so if $Q_X^*(x)$ is more uniform than $P_X(x)$ *and* the majority of Q_X^*-typical sequences belong to the subset, there is still a possibility for exceeding the source entropy $\mathbb{H}(X)$. Similar arguments can be stated for exceeding $R(D) = R(P_X, D)$ bits in lossy compression of certain subsets.

B. Lossy Compression for Smooth Subsets

In this subsection, we state our lossy compression result for smooth subsets. Our result relies on a quantity we term the *subset mutual information*, to introduce which we first recall some definitions and introduce a few notations.

Definition 6 [29]: Given any conditional distribution $P_{Y|X}(y|x)$ and any positive δ_n, the set $T^n[P_{Y|X}|x^n]_{\delta_n}$ of conditional $P_{Y|X}$-typical *sequences given $x^n \in \mathcal{X}^n$ is defined as the set of all sequences $y^n \in \mathcal{Y}^n$ that satisfy*

$$\left|\frac{1}{n}N((x,y);(x^n,y^n)) - \frac{1}{n}N(x;x^n)P_{Y|X}(y|x)\right| \leq \delta_n, \tag{37}$$

for all $x \in \mathcal{X}, y \in \mathcal{Y}$ with $P_{Y|X}(y|x) > 0$ and $N((x,y);(x^n,y^n)) = 0$ otherwise.

Remark 6: In the definition above and throughout, the sequence δ_n is assumed to satisfy an extension of the Delta-Convention mentioned in Remark 3, i.e., (i) as $n \to \infty$, we have $\delta_n \to 0$ and $\sqrt{n}\delta_n \to \infty$ and (ii) when going from conditional to nonconditional typical sets, the δ_n sequence, with some abuse of notation, also stands for sums and constant multiples like $\delta_n'' := |\mathcal{Y}|(\delta_n + \delta_n')$ and so on [29].

One recalls from the properties of the typical sequences that, for every conditional distribution $P_{Y|X}(y|x)$ and any arbitrary distribution $Q_X(x)$, the size of the conditional $P_{Y|X}$-typical set satisfies [29]

$$\begin{aligned}&\lim_{n\to\infty} \min_{x^n \in T^n[Q_X]_{\delta_n}} \frac{1}{n}\log|T^n[P_{Y|X}|x^n]_{\delta_n}| \\ &\quad = \lim_{n\to\infty} \max_{x^n \in T^n[Q_X]_{\delta_n}} \frac{1}{n}\log|T^n[P_{Y|X}|x^n]_{\delta_n}| = \mathbb{H}(P_{Y|X}|Q_X).\end{aligned} \tag{38}$$

We can now define the notions of conditional subset entropy and subset mutual information.

Definition 7: Consider a subset $\mathcal{L} = \{\mathcal{L}_n \subseteq \mathcal{X}^n\}_{n=1}^{\infty}$, a distribution $Q_X(x)$ and a conditional distribution $P_{Y|X}(y|x)$. Let $Q_Y(y)$ be the induced distribution $Q_Y(y) = \sum_x Q_X(x)P_{Y|X}(y|x)$. Consider an auxiliary subset $\bar{\mathcal{L}} = \{\bar{\mathcal{L}}_n \subseteq \mathcal{Y}^n\}_{n=1}^{\infty}$ for which the subset entropy

$$H_{\bar{\mathcal{L}}}(Q_Y) = \lim_{n\to\infty} \frac{1}{n}\log\left|T^n[Q_Y]_{\delta_n} \cap \bar{\mathcal{L}}_n\right| \tag{39}$$

exists. A constant $H_{\bar{\mathcal{L}}|\mathcal{L}}(P_{Y|X}|Q_X)$ is called the conditional subset entropy of $P_{Y|X}$ given Q_X and the subsets $\mathcal{L}$ and $\bar{\mathcal{L}}$ if

$$\begin{aligned}&\lim_{n\to\infty} \min_{x^n \in \mathcal{L}_n \cap T^n[Q_X]_{\delta_n}} \frac{1}{n}\log\left|\bar{\mathcal{L}}_n \cap T^n[P_{Y|X}|x^n]_{\delta_n}\right| \\ &\qquad = H_{\bar{\mathcal{L}}|\mathcal{L}}(P_{Y|X}|Q_X),\end{aligned} \tag{40}$$

provided that the limit exists. Accordingly, the subset mutual information is defined as

$$I_{\mathcal{L},\bar{\mathcal{L}}}(Q_X, P_{Y|X}) := H_{\bar{\mathcal{L}}}(Q_Y) - H_{\bar{\mathcal{L}}|\mathcal{L}}(P_{Y|X}|Q_X). \tag{41}$$

Definition 8: Consider a subset $\mathcal{L} = \{\mathcal{L}_n \subseteq \mathcal{X}^n\}_{n=1}^{\infty}$, a distribution $Q_X(x)$, a conditional distribution $P_{Y|X}(y|x)$, and an induced marginal distribution $Q_Y(y)$. We say an 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 both the subset entropy $H_{\bar{\mathcal{L}}}(Q_Y)$ and the conditional subset entropy $H_{\bar{\mathcal{L}}|\mathcal{L}}(P_{Y|X}|Q_X)$ exist.

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, as discussed before, it satisfies the property $0 \leq H_{\bar{\mathcal{L}}}(Q_Y) \leq \mathbb{H}(Q_Y)$.

Comparing expressions (38) and (40) we see that, the conditional subset entropy $H_{\bar{\mathcal{L}}|\mathcal{L}}(P_{Y|X}|Q_X)$ is an analog of the conventional conditional entropy $\mathbb{H}(P_{Y|X}|Q_X)$. 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 $\mathcal{L}_n = \mathcal{X}^n$ 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 and subset mutual information as introduced in the following definition.

Definition 9: Consider a subset $\mathcal{L} = \{\mathcal{L}_n \subseteq \mathcal{X}^n\}_{n=1}^{\infty}$ and a 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 8 for all distributions $Q_X(x)$ in a δ_n-neighborhood of all subset-$\mathcal{L}$-typical distributions $Q_X^*(x) \in \mathcal{Q}_X^*$ as defined in (33) 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 $H_{\bar{\mathcal{L}}|\mathcal{L}}(P_{Y|X}|Q_X)$ and hence the subset mutual information $I_{\mathcal{L},\bar{\mathcal{L}}}(Q_X, P_{Y|X})$ are continuous in all those $Q_X(x)$ distributions.

We now state our lossy compression results for smooth subsets.

Theorem 3: 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

$$R_{\mathcal{L}}(D) \leq \max_{Q_X^* \in \mathcal{Q}_X^*} \inf_{P_{Y|X}: \mathbb{E}[d(X^*,Y^*)] \leq D} \inf_{\bar{\mathcal{L}}:(P_{Y|X},\mathcal{L})\text{-smooth}} I_{\mathcal{L},\bar{\mathcal{L}}}(Q_X^*, P_{Y|X}), \tag{42}$$

where: $\mathcal{Q}_X^*$ is the set of all subset-typical distributions as defined in (33); $I_{\mathcal{L},\bar{\mathcal{L}}}(Q_X^*, P_{Y|X})$ is the subset mutual information as in Definition 7; $\bar{\mathcal{L}}$ is the smooth auxiliary subset as in Definition 9; and the pair of random variables (X^*, Y^*) are distributed according to $Q_X^*(x)P_{Y|X}(y|x)$ so that

$$\mathbb{E}[d(X^*, Y^*)] = \sum_{x,y} Q_X^*(x) P_{Y|X}(y|x) d(x, y). \tag{43}$$

Proof: Proof is provided in Section V-C2. □

Theorem 3 presents a result that is analogous to the classical rate-distortion result (2) for a DMS. This theorem 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 classical rate-distortion result (2), a key is minimization of this mutual information over the conditional distributions $P_{Y|X}(y|x)$ that satisfy the expected distortion constraint $\mathbb{E}[d(X^*, Y^*)] \leq D$. The fact that the collection of subset-typical distributions $\mathcal{Q}_X^*$ plays 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 encoded via the lossy subset compression code, see the discussion below Theorem 2. Further, if multiple subset-typical distributions $Q_X^*(x)$ exist, one must code for the worst case, hence the $\max_{Q_X^* \in \mathcal{Q}_X^*}$ term in (42).

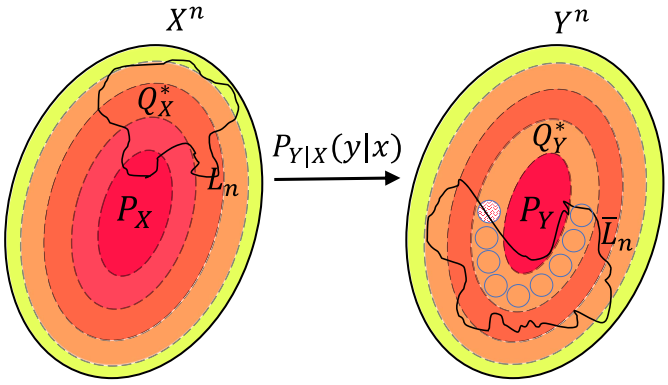

Fig. 2. Schematic description of Theorem 3 for lossy compression of smooth subsets. The transformation of objects in the original $\mathcal{X}^n$ domain to the reconstruction $\mathcal{Y}^n$ domain via the conditional distribution $P_{Y|X}(y|x)$ is illustrated. The source distribution is denoted by P_X and its induced distribution on the Y domain is denoted by P_Y. The subset-typical distribution is denoted by Q_X^* and its induced distribution on the Y domain is denoted by Q_Y^*. The dashed rings on both sides denote the typical sets corresponding to different distributions. The original subset $\mathcal{L}_n \subseteq \mathcal{X}^n$ and the auxiliary subset $\bar{\mathcal{L}}_n \subseteq \mathcal{Y}^n$ are depicted with the curved shapes. The circles on the right side depict the conditional typical sets $T^n[P_{Y|X}|x^n]_{\delta_n}$ for several x^n sequences belonging to $T^n[Q_X^*]_{\delta_n} \cap \mathcal{L}_n$. One observes that, the size of the intersection of the auxiliary subset $\bar{\mathcal{L}}_n$ with different conditional typical sets $T^n[P_{Y|X}|x^n]_{\delta_n}$ varies with the choice of x^n, and the one with the least intersection size, shown as a hatched circle, dictates the compression rate.

The last key element of our lossy compression result in Theorem 3 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 $I_{\mathcal{L},\bar{\mathcal{L}}}(Q_X^*, P_{Y|X})$ as defined in (41), 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 $P_{Y|X}(y|x)$. This interpretation is schematically depicted in Figure 2.

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 mutual information reduces to the average mutual information, which readily gives the following achievable rate-distortion result.

Corollary 2: 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

$$R_{\mathcal{L}}(D) \leq \max_{Q_X^* \in \mathcal{Q}_X^*} R(Q_X^*, D), \tag{44}$$

where $\mathcal{Q}_X^*$ is the set of all subset-typical distributions as defined in (33)*, and $R(Q_X^*, D)$ is the standard rate-distortion function* (2) *for distribution $Q_X^*(x)$.*

For the special case of symmetric subsets, the subset $\mathcal{L}$ fully intersects the subset-typical distributions $Q_X^*(x)$, therefore the role of subset structure vanishes and a standard rate-distortion code for this distribution is sufficient for the lossy compression of the subset. By stating a proof of converse, we show that the achievable rate-distortion in Corollary 2 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 4: 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

$$Q_X^* = \operatorname*{arg\,min}_{Q:\mathcal{L}\cap T[Q]\neq\emptyset} D(Q\|P) \tag{45}$$

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

$$R_{\mathcal{L}}(D) = R(Q_X^*, D), \tag{46}$$

where $R(Q_X^*, D)$ is the standard rate-distortion function (2) *for distribution $Q_X^*(x)$.*

Proof: Proof is provided in Section V-C3. □

As a sanity check, we can again observe that for the standard case of $\mathcal{L}_n = \mathcal{X}^n$, the subset-typical distribution is uniquely given by $Q_X^* \equiv P$. Therefore, the specific characterization (46) and in turn the more general bound (42) on the subset rate-distortion formula reduce to the standard rate-distortion function (2). 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.

C. Proofs for Smooth Subsets

In the remainder of this section, we state the proof of our compression results for smooth subsets. We provide the proofs for achievability and strong converse of our lossless result, Theorem 2, in Subsection V-C1; achievability of our lossy result, Theorem 3, in Subsection V-C2; and strong converse of the lossy result for smooth symmetric subsets, Theorem 4, in Subsection V-C3.

Before starting with the proofs, we recall the notion of type classes used frequently herein.

Definition 10 [29]: The type *of a sequence x^n is the empirical distribution $\hat{P}_{x^n}(x)$ defined as*

$$\hat{P}_{x^n}(x) := \frac{1}{n} N(x; x^n), \quad \forall x \in \mathcal{X}. \tag{47}$$

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 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 [29]. In the following, we frequently use the notations

$$T_{\mathcal{L}}^n(\hat{P}) := \mathcal{L}_n \cap T^n(\hat{P}), \quad T_{\mathcal{L}}^n[Q]_{\delta_n} := \mathcal{L}_n \cap T^n[Q]_{\delta_n}, \tag{48}$$

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.

1) Proof of the Lossless Result: In this part, we provide the proof of Theorem 2 on lossless compression of smooth subsets. The strong converse proof is inspired by [29, Pr. 2.6], while the achievability proof readily builds upon the following lemma, which is related to Sanov's theorem [29, Pr. 2.12] and summarizes the properties of typical sequences intersecting a subset of the source.

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

$$2^{-n[g_P(Q)+\epsilon_n]} \leq P_{X^n}[X^n \in T_{\mathcal{L}}^n[Q]_{\delta_n}] \leq 2^{-n[g_P(Q)-\epsilon_n]}, \tag{49}$$

where function $g_P(Q)$ is defined in (34)*. Moreover, if $\mathcal{L}$ is a smooth subset, then*

$$2^{-n\left[\min_{Q:\mathcal{L}\cap T[Q]\neq\emptyset} g_P(Q)+\epsilon_n\right]} \leq P_{X^n}[X^n \in \mathcal{L}_n] \leq (n+1)^{|\mathcal{X}|} 2^{-n\left[\min_{Q:\mathcal{L}\cap T[Q]\neq\emptyset} g_P(Q)-\epsilon_n\right]}. \tag{50}$$

Proof: Proof is provided in Appendix B. □

We are now ready to prove Theorem 2, which is inspired by [29, Th. 2.15 and Pr. 2.6].

Proof (of Theorem 2): To prove the achievability side, we consider the following 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

$$\mathcal{A}_n := \bigcup_{\hat{P}:n\text{-type}, \hat{P}\in\Omega(3\epsilon)} T_{\mathcal{L}}^n(\hat{P}), \tag{51}$$

where

$$\Omega(\epsilon) := \left\{Q : \mathcal{L} \cap T[Q] \neq \emptyset, g_P(Q) < \min_{Q:\mathcal{L}\cap T[Q]\neq\emptyset} g_P(Q) + \epsilon\right\}. \tag{52}$$

All other sequences in $(\mathcal{L}_n - \mathcal{A}_n)$ lead to an error. Note that, the use of min for $g_P(Q)$ in the definition (52) is justified by the continuity of the subset entropy $H_{\mathcal{L}}(Q)$ and thus the function $g_P(Q)$. We can write

$$\begin{aligned}
&\Pr[X^n \in (\mathcal{A}_n^c \cap \mathcal{L}_n)] \\
&\quad = \sum_{\hat{P}:n\text{-type}, \hat{P}\notin\Omega(3\epsilon)} P_{X^n}[X^n \in T_{\mathcal{L}}^n(\hat{P})] && (53)\\
&\quad \leq (n+1)^{|\mathcal{X}|} \max_{\substack{\hat{P}:n\text{-type} \\ \hat{P}\notin\Omega(3\epsilon), T_{\mathcal{L}}^n(\hat{P})\neq\emptyset}} P_{X^n}[X^n \in T_{\mathcal{L}}^n(\hat{P})] && (54)\\
&\quad \leq (n+1)^{|\mathcal{X}|} \max_{Q\notin\Omega(3\epsilon), \mathcal{L}\cap T[Q]\neq\emptyset} P_{X^n}[X^n \in T_{\mathcal{L}}^n[Q]_{\delta_n}]. && (55)
\end{aligned}$$

Combining (55) and Lemma 1, the error probability is bounded as

$$\Pr[\mathcal{E}_{\mathcal{L}}] = \Pr[X^n \notin \mathcal{A}_n | X^n \in \mathcal{L}_n] \tag{56}$$

$$= \frac{\Pr[X^n \in (\mathcal{A}_n^c \cap \mathcal{L}_n)]}{\Pr[X^n \in \mathcal{L}_n]} \tag{57}$$

$$\leq \frac{(n+1)^{|\mathcal{X}|} 2^{-n\left[\min\limits_{Q \notin \Omega(3\epsilon), \mathcal{L} \cap T[Q] \neq \emptyset} g_P(Q) - \epsilon_n\right]}}{2^{-n\left[\min\limits_{Q: \mathcal{L} \cap T[Q] \neq \emptyset} g_P(Q) + \epsilon_n\right]}}. \tag{58}$$

Therefore, from definition (52) 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

$$\frac{1}{n} \log |\mathcal{A}_n| = \frac{1}{n} \log \sum_{\hat{P}: n\text{-type}, \hat{P} \in \Omega(3\epsilon)} |T_{\mathcal{L}}^n(\hat{P})| \tag{59}$$

$$\leq \frac{1}{n} \log \left((n+1)^{|\mathcal{X}|} \max_{\hat{P}: n\text{-type}, \hat{P} \in \Omega(3\epsilon)} |T_{\mathcal{L}}^n(\hat{P})| \right) \tag{60}$$

$$\leq \frac{1}{n} \log \left((n+1)^{|\mathcal{X}|} \max_{Q \in \Omega(3\epsilon)} |T_{\mathcal{L}}^n[Q]_{\delta_n}| \right) \tag{61}$$

$$\leq \max_{Q \in \Omega(3\epsilon)} H_{\mathcal{L}}(Q) + \xi_n + \frac{|\mathcal{X}| \log(n+1)}{n}, \tag{62}$$

where (60) follows from the Type Counting Lemma, and (62) from (30) and the continuity of the subset entropy $H_{\mathcal{L}}(Q)$. This completes the achievability proof for Theorem 2 since $n \to \infty$ and the choice of $\epsilon > 0$ is arbitrary.

In the following, we prove a strong converse for Theorem 2, that is, we prove any arbitrary lossless code for the 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

$$\Pr[X^n \in (\mathcal{A}_n \cap \mathcal{L}_n)]$$

$$= \sum_{\hat{P}: n\text{-type}} P_{X^n}[X^n \in (\mathcal{A}_n \cap T_{\mathcal{L}}^n(\hat{P}))] \tag{63}$$

$$\leq (n+1)^{|\mathcal{X}|} \max_{\substack{\hat{P}: n\text{-type} \\ T_{\mathcal{L}}^n(\hat{P}) \neq \emptyset}} P_{X^n}[X^n \in (\mathcal{A}_n \cap T_{\mathcal{L}}^n(\hat{P}))] \tag{64}$$

$$\leq (n+1)^{|\mathcal{X}|} \max_{Q: \mathcal{L} \cap T[Q] \neq \emptyset} P_{X^n}[X^n \in (\mathcal{A}_n \cap T_{\mathcal{L}}^n[Q]_{\delta_n})]. \tag{65}$$

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

$$P_{X^n}[X^n \in (\mathcal{A}_n \cap T_{\mathcal{L}}^n[Q]_{\delta_n})]$$

$$\leq |\mathcal{A}_n \cap T_{\mathcal{L}}^n[Q]_{\delta_n}| \max_{x^n \in T^n[Q]_{\delta_n}} P_{X^n}(x^n) \tag{66}$$

$$\leq \min\{2^{nR}, 2^{n[H_{\mathcal{L}}(Q) + \xi_n]}\} \times 2^{-n[\mathbb{H}(Q) + D(Q\|P) - \xi_n']} \tag{67}$$

$$= 2^{-n[g_P(Q) + |H_{\mathcal{L}}(Q) - R + \xi_n|^+ - \epsilon_n]}, \tag{68}$$

where (67) follows form (30) and that $P_{X^n}(x^n) = 2^{-n[\mathbb{H}(\hat{P}_{x^n}) + D(\hat{P}_{x^n}\|P)]}$ with some $\xi_n \to 0$ and $\xi_n' \to 0$ as $n \to \infty$ [1], [29], and (68) follows from the definition of $g_P(Q)$ and $\epsilon_n := \xi_n + \xi_n'$. Combining (65), (68) and Lemma 1, the correct decoding probability is bounded as

$$1 - \Pr[\mathcal{E}_{\mathcal{L}}]$$

$$= \Pr[X^n \in \mathcal{A}_n | X^n \in \mathcal{L}_n] \tag{69}$$

$$= \frac{P_{X^n}[X^n \in (\mathcal{A}_n \cap \mathcal{L}_n)]}{P_{X^n}[X^n \in \mathcal{L}_n]} \tag{70}$$

$$\leq \frac{(n+1)^{|\mathcal{X}|} 2^{-n\left[\min\limits_{Q: \mathcal{L} \cap T[Q] \neq \emptyset} g_P(Q) + |H_{\mathcal{L}}(Q) - R + \xi_n|^+ - \epsilon_n\right]}}{2^{-n\left[\min\limits_{Q: \mathcal{L} \cap T[Q] \neq \emptyset} g_P(Q) + \epsilon_n\right]}}. \tag{71}$$

Inspecting the lower bound (71) on error probability suggests that, if $R < H_{\mathcal{L}}(Q_X^*) - 2\epsilon_n$ for any distribution Q_X^* satisfying $g_P(Q_X^*) \leq \min\limits_{Q: \mathcal{L} \cap T[Q] \neq \emptyset} g_P(Q) + 3\epsilon_n$, then the error probability is bounded at least as $\Pr[\mathcal{E}_{\mathcal{L}}] \geq 1 - (n+1)^{|\mathcal{X}|} 2^{-n\xi_n}$. Since ϵ_n and $(n+1)^{|\mathcal{X}|} 2^{-n\xi_n}$ are both vanishing[3] as $n \to \infty$, this proves the strong converse and completes the proof of Theorem 2. □

2) Proof of the Lossy Result: In this part, we provide the proof of Theorem 3 on lossy compression of smooth subsets. The proof of this achievability result builds upon the following lemma, which is an analog of the *Type Covering Lemma* [29, 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 2 (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 exist a set $B(\hat{P}_X, \mathcal{L}) \subseteq \mathcal{Y}^n$ that satisfies

$$d(x^n, B(\hat{P}_X, \mathcal{L})) := \min_{y^n \in B(\hat{P}_X, \mathcal{L})} d(x^n, y^n) \leq D, \quad \forall x^n \in T_{\mathcal{L}}^n(\hat{P}_X), \tag{72}$$

for sufficiently large n, and whose size is bounded as

$$\frac{1}{n} \log |B(\hat{P}_X, \mathcal{L})| \leq \inf_{P_{Y|X}: \mathbb{E}[d(X,Y)] \leq D} \inf_{\bar{\mathcal{L}}: (\hat{P}_X, P_{Y|X}, \mathcal{L})\text{-smooth}} I_{\mathcal{L}, \bar{\mathcal{L}}}(\hat{P}_X, P_{Y|X}) + 3\xi_n, \tag{73}$$

where $\xi_n \to 0$ as $n \to \infty$, and $\bar{\mathcal{L}} := \{\bar{\mathcal{L}}_n \subseteq \mathcal{Y}^n\}_{n=1}^{\infty}$ is a $(\hat{P}_X, P_{Y|X}, \mathcal{L})$-smooth auxiliary subset, and $I_{\mathcal{L}, \bar{\mathcal{L}}}(\hat{P}_X, P_{Y|X})$ is the subset mutual information, both as introduced in Definitions 7 and 8; and the expected distortion is calculated with respect to the type distribution,

$$\mathbb{E}[d(X, Y)] = \sum_{x,y} \hat{P}_X(x) P_{Y|X}(y|x) d(x, y). \tag{74}$$

Proof: Proof is provided in Appendix C. □

We are now ready to prove Theorem 3 with elements similar to the proof of error exponents for the classical rate-distortion problem [29, Th. 9.5] and our proof of the lossless subset compression in Theorem 2.

[3] A sequence $\tilde{\xi}_n \geq \xi_n$ can be found such that, as $n \to \infty$, not only $\tilde{\xi}_n \to 0$ but also $n\tilde{\xi}_n \to \infty$ and furthermore $n^{|\mathcal{X}||\mathcal{Y}|} \exp(-n\tilde{\xi}_n) \to 0$. Therefore, we always assume that ξ_n satisfies the latter conditions.

Proof (of Theorem 3): As in the achievability proof of Theorem 2, we fix an arbitrary $\epsilon > 0$ and consider the following lossy 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.

$$\mathcal{A}_n := \bigcup_{\hat{P}:n\text{-type}, \hat{P}\in\Omega(3\epsilon)} T_{\mathcal{L}}^n(\hat{P}), \tag{75}$$

where

$$\Omega(\epsilon):=\left\{Q:\mathcal{L}\cap T[Q]\neq\emptyset, g_P(Q) < \min_{Q:\mathcal{L}\cap T[Q]\neq\emptyset} g_P(Q)+\epsilon\right\}. \tag{76}$$

Our lossy source code consists of the following reconstructions sequences:

$$B(\mathcal{L}) := \bigcup_{\hat{P}_X:n\text{-type}, \hat{P}_X\in\Omega(3\epsilon)} B(\hat{P}_X, \mathcal{L}), \tag{77}$$

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, so it has the size

$$\frac{1}{n}\log|B(\hat{P}_X, \mathcal{L})| \leq \inf_{P_{Y|X}:\mathbb{E}[d(X,Y)]\leq D} \inf_{\bar{\mathcal{L}}:(\hat{P}_X, P_{Y|X}, \mathcal{L})\text{-smooth}} I_{\mathcal{L},\bar{\mathcal{L}}}(\hat{P}_X, P_{Y|X}) + 3\xi_n, \tag{78}$$

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

$$d(x^n, B(\mathcal{L})) \leq d(x^n, B(\hat{P}_{x^n}, \mathcal{L})) \leq D, \tag{79}$$

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

$$\begin{aligned}
\Pr[\mathcal{E}_{\mathcal{L}}(D)] &= \Pr[d(X^n, Y^n) > D | X^n \in \mathcal{L}_n] && (80)\\
&\leq \Pr[X^n \notin \mathcal{A}_n | X^n \in \mathcal{L}_n] && (81)\\
&\leq \frac{(n+1)^{|\mathcal{X}|} 2^{-n\left[\min_{Q\notin\Omega(3\epsilon), \mathcal{L}\cap T[Q]\neq\emptyset} g_P(Q)-\epsilon_n\right]}}{2^{-n\left[\min_{Q:\mathcal{L}\cap T[Q]\neq\emptyset} g_P(Q)+\epsilon_n\right]}} \\
&\leq (n+1)^{|\mathcal{X}|} 2^{-n\epsilon}, && (82)
\end{aligned}$$

where the last line follows from our calculations in the lossless case; cf. (53)-(58). Hence, it only remains to determine the compression rate. From the Subset-Type Covering Lemma above and the Type Counting Lemma, we have

$$\begin{aligned}
&\frac{1}{n}\log|B(\mathcal{L})| \\
&= \frac{1}{n}\log \sum_{\hat{P}_X:n\text{-type}, \hat{P}_X\in\Omega(3\epsilon)} |B(\hat{P}_X, \mathcal{L})| && (83)\\
&\leq \frac{1}{n}\log\left((n+1)^{|\mathcal{X}|} \max_{\hat{P}_X:n\text{-type}, \hat{P}_X\in\Omega(3\epsilon)} |B(\hat{P}_X, \mathcal{L})|\right) && (84)\\
&\leq \max_{\hat{P}_X:n\text{-type}, \hat{P}_X\in\Omega(3\epsilon)} \inf_{P_{Y|X}:\mathbb{E}[d(X,Y)]\leq D} \\
&\quad \times \inf_{\bar{\mathcal{L}}:(\hat{P}_X, P_{Y|X}, \mathcal{L})\text{-smooth}} I_{\mathcal{L},\bar{\mathcal{L}}}(\hat{P}_X, P_{Y|X}) \\
&\quad + 3\xi_n + \frac{|\mathcal{X}|\log(n+1)}{n} && (85)\\
&\leq \max_{Q\in\Omega(3\epsilon)} \inf_{P_{Y|X}:\mathbb{E}[d(X,Y)]\leq D} \inf_{\bar{\mathcal{L}}:(P_{Y|X}, \mathcal{L})\text{-smooth}} I_{\mathcal{L},\bar{\mathcal{L}}}(Q, P_{Y|X}) \\
&\quad + 5\xi_n, && (86)
\end{aligned}$$

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 3. □

3) Proof of the Lossy Result for Symmetric Subsets: In this part, we prove Theorem 4 on lossy compression of smooth symmetric subsets. The achievability immediately follows from Corollary 2. The converse is analogous to that for the standard rate-distortion theorem [29, Th. 7.3] and uses the following two technical lemmas.

The first technical lemma is a generalized asymptotic equipartition property (AEP) and an analog of [29, Lemma 2.12] which asserts that essentially all of the probability mass of a smooth subset, symmetric or not, is concentrated only in the subset-typical sequences.

Lemma 3: Consider a discrete memoryless source $P(x)$ and a smooth subset $\mathcal{L} = \{\mathcal{L}_n \subseteq \mathcal{X}^n\}_{n=1}^{\infty}$ with the set of subset-typical distributions $\mathcal{Q}_X^*$ as defined in (33)*. Then, there exists a sequence $\epsilon_n \to 0$ as $n \to \infty$ such that*

$$\Pr\left[X^n \in \bigcup_{Q_X^*\in\mathcal{Q}_X^*} T_{\mathcal{L}}^n[Q_X^*]_{\delta_n} \middle| X^n \in \mathcal{L}_n\right] \geq 1-\epsilon_n. \tag{87}$$

Proof: Proof is provided in Appendix D. □

The second technical lemma is an analog of [29, Lemma 2.14] and states that, when constrained to only a smooth subset of the source realizations, symmetric or not, any set with high probability has a size essentially no smaller than the subset-typical set.

Lemma 4: 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 (33) *is unique. Given $0 < \eta < 1$, there exists a sequence $\epsilon_n \to 0$ as $n \to \infty$ such that, if a set $\mathcal{A} \subseteq \mathcal{X}^n$ satisfies*

$$\Pr\left[X^n \in \mathcal{A} \mid X^n \in \mathcal{L}_n\right] \geq \eta, \tag{88}$$

then

$$|\mathcal{A}| \geq 2^{n[H_{\mathcal{L}}(Q_X^*)-\epsilon_n]}. \tag{89}$$

Proof: Proof is provided in Appendix E. □

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

Proof of Theorem 4): We only state the proof of (strong) converse, since the achievability readily follows from Corollary 2 as well as the fact that for smooth symmetric subsets the function $g_P(Q)$ defined in (34) reduces to $D(Q\|P)$.

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

$$\Pr\left[d\left(X^n, \phi(f(X^n))\right) \leq D \mid X^n \in \mathcal{L}_n\right] \geq 1-\epsilon, \quad (90)$$

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

$$\mathcal{A} := \left\{x^n \in T^n_{\mathcal{L}}[Q^*_X]_{\delta_n} : d(x^n, \phi(f(x^n))) \leq D\right\}. \quad (91)$$

From Lemma 3, we have

$$\Pr\left[X^n \in T^n_{\mathcal{L}}[Q^*_X]_{\delta_n} \mid X^n \in \mathcal{L}_n\right] \geq 1-\tau_n, \quad (92)$$

for some $\tau_n \to 0$ as $n \to \infty$. Then, the simple inequality $\Pr[A \cap B] \geq \Pr[A] - \Pr[B^c]$ implies

$$\Pr\left[X^n \in \mathcal{A} \mid X^n \in \mathcal{L}_n\right] \geq 1-\epsilon-\tau_n, \quad (93)$$

which, on account of Lemma 4 and since $H_{\mathcal{L}}(Q_X) = \mathbb{H}(Q_X)$ for all distributions $Q_X(x)$ intersecting the subset $\mathcal{L}$, yields

$$|\mathcal{A}| \geq 2^{n[\mathbb{H}(Q^*_X)-\epsilon_n]}. \quad (94)$$

On the other hand, define the set of reconstruction codewords corresponding to the set $\mathcal{A}$ as

$$\mathcal{C} := \left\{y^n \in \mathcal{Y}^n : y^n = \phi(f(x^n)) \text{ for some } x^n \in \mathcal{A}\right\}, \quad (95)$$

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

$$\mathcal{A} := \bigcup_{y^n \in \mathcal{C}} \mathcal{A}(y^n), \quad (96)$$

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

$$\mathcal{A}(y^n) := \left\{x^n \in \mathcal{A} : \phi(f(x^n)) = y^n\right\}. \quad (97)$$

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

$$\mathcal{A}(y^n) = \bigcup_{\substack{\hat{P}_{XY}(x,y):\, n\text{-joint type} \\ \mathbb{E}[d(\hat{X},\hat{Y})] \leq D \\ |\hat{P}_X(x) - Q^*_X(x)| \leq \delta_n}} \left(\mathcal{A}(y^n) \cap T^n_{\mathcal{L}}(\hat{P}_{X|Y}|y^n)\right), \quad (98)$$

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

$$\left|T^n(\hat{P}_{X|Y}|y^n)\right| \leq 2^{n\mathbb{H}(\hat{P}_{X|Y}|\hat{P}_Y)}, \quad (99)$$

we get

$$\begin{aligned} |\mathcal{A}| &\leq \sum_{y^n \in \mathcal{C}} \left|\mathcal{A}(y^n)\right| \\ &\leq |\mathcal{C}| \cdot (n+1)^{|\mathcal{X}||\mathcal{Y}|} \max_{\substack{\hat{P}_{XY}(x,y):\, n\text{-joint type} \\ \mathbb{E}[d(\hat{X},\hat{Y})] \leq D \\ |\hat{P}_X(x) - Q^*_X(x)| \leq \delta_n}} 2^{n\mathbb{H}(\hat{P}_{X|Y}|\hat{P}_Y)}. \end{aligned} \quad (100)$$

Combining (94) and (100), we have proved that the size of any lossy code for the smooth symmetric subset $\mathcal{L}$ satisfies

$$M \geq |\mathcal{C}| \geq (n+1)^{-|\mathcal{X}||\mathcal{Y}|} \times \exp\left(n \min_{\substack{\hat{P}_{XY}(x,y):\, n\text{-joint type} \\ \mathbb{E}[d(\hat{X},\hat{Y})] \leq D \\ |\hat{P}_X(x) - Q^*_X(x)| \leq \delta_n}} \left[\mathbb{H}(Q^*_X) - \mathbb{H}(\hat{P}_{X|Y}|\hat{P}_Y) - \epsilon_n\right]\right). \quad (101)$$

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

$$R_{\mathcal{L}}(D) \geq \min_{P_{Y|X}:\mathbb{E}[d(X^*,Y^*)] \leq D} \mathbb{I}(Q^*_X, P_{Y|X}) - 3\epsilon_n. \quad (102)$$

This concludes the proof of the strong converse and that of Theorem 4. □

VI. 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 particular, we focus on subsets that are not likely or smooth, but are fluctuating among a finite number of such components. 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 11: 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 infinite 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 5: Consider a discrete memoryless source $P(x)$ and an $(\mathcal{L}_j, n_j)_{j=1}^{J}$-fluctuating subset. Then, the optimal lossless compression rate and rate-distortion function for the subset $\mathcal{L}$ respectively satisfy:

$$R^*_{\mathcal{L}} = \max_{1 \leq j \leq J} R^*_{\mathcal{L}_j}, \quad (103)$$

$$R_{\mathcal{L}}(D) = \max_{1 \leq j \leq J} R_{\mathcal{L}_j}(D). \quad (104)$$

Proof: We state only the proof for the lossless case; that for the lossy case is very similar and we skip the details for brevity. (Achievability) Fix an arbitrary $\epsilon > 0$. For each $1 \leq j \leq J$, let $\{(m_{j,n}, \hat{x}^n_j)\}_{n=1}^{\infty}$ be the optimal encoder and decoder sequence for lossless compression of the subset $\mathcal{L}_j$, achieving a rate $R^*_{\mathcal{L}_j} + \epsilon$ with vanishing error probability $\Pr[\hat{X}^n_j \neq X^n | X^n \in \mathcal{L}_{j,n}] \to 0$ as $n \to \infty$. We consider the following code for the fluctuating subset: let $m_n \equiv m_{j,n}$ and $\hat{x}^n \equiv \hat{x}^n_j$ if $n \in \{n_{j,k}\}_{k=1}^{\infty}$. Then, we have

$$\begin{aligned} &\limsup_{n\to\infty} \Pr[\hat{X}^n \neq X^n | X^n \in \mathcal{L}_n] \\ &\qquad = \max_{1 \leq j \leq J} \limsup_{n\to\infty} \Pr[\hat{X}^n_j \neq X^n | X^n \in \mathcal{L}_{j,n}] = 0. \end{aligned} \quad (105)$$

The rate of this code is $\max_{1\le j\le J} R^*_{\mathcal{L}_j}+\epsilon$. Since ϵ is arbitrary, this completes the achievability proof in the lossless case.

(Converse) Assume $R < \max_{1\le j\le J} R^*_{\mathcal{L}_j}$, then at least one $1 \le \bar{j} \le J$ exists such that $R < R^*_{\mathcal{L}_{\bar{j}}}$. By the definition of $R^*_{\mathcal{L}_{\bar{j}}}$, we have $\limsup_{n\to\infty} \Pr[\hat{X}^n_{\bar{j}} \neq X^n | X^n \in \mathcal{L}_{\bar{j},n}] > 0$ for any arbitrary lossless code for the subset $\mathcal{L}_{\bar{j}}$ with rate R. Hence,

$$\limsup_{n\to\infty} \Pr[\hat{X}^n \neq X^n | X^n \in \mathcal{L}_n] \ge \limsup_{n\to\infty} \Pr[\hat{X}^n_{\bar{j}} \neq X^n | X^n \in \mathcal{L}_{\bar{j},n}] > 0, \quad (106)$$

which proves the converse for the fluctuating subset $\mathcal{L}$ for the lossless case. □

If all components of a fluctuating subset are smooth, Theorem 5 readily specializes as follows, using Theorem 2 and Corollary 2.

Corollary 3: Consider a discrete memoryless source $P(x)$; an $(\mathcal{L}_j, n_j)_{j=1}^J$-fluctuating subset whose components are all smooth with subset typical distributions given by

$$\mathcal{Q}^*_j = \arg\min_{Q:\mathcal{L}_j\cap T[Q]\neq\emptyset} [\mathbb{H}(Q) - H_{\mathcal{L}_j}(Q) + D(Q\|P)]. \quad (107)$$

Then, the optimal lossless compression rate for the fluctuating subset $\mathcal{L}$ is

$$R^*_{\mathcal{L}} = \max_{1\le j\le J} \max_{Q^*_j\in\mathcal{Q}^*_j} H_{\mathcal{L}_j}(Q^*_j), \quad (108)$$

and the rate-distortion function for the subset $\mathcal{L}$ satisfies

$$R_{\mathcal{L}}(D) \le \max_{1\le j\le J} \max_{Q^*_j\in\mathcal{Q}^*_j} R(Q^*_j, D), \quad (109)$$

where $R(Q^*_j, D)$ is the standard rate-distortion function (2) for distribution $Q^*_j(x)$.

VII. Examples

In this section, we present several examples to better illustrate our models and results. In all of these examples, we consider a binary DMS, $\mathcal{X} = \{0, 1\}$, with a Bernoulli distribution $B(p)$ with parameter $0 \le p \le 1/2$. The fundamental limit of lossless compression is given by the source entropy, which is $R^* = H_b(p)$ for this source. The lossy compression is considered with respect to the Hamming distance. In particular, the rate-distortion function of the source is $R(D) = H_b(p) - H_b(D)$ if $0 \le D \le p$ and $R(D) = 0$ if $D > p$ [1]. In this section, we frequently use some notations: the Hamming weight $w_H(x^n)$ of a binary sequence x^n, the binary convolution operation $p * q := p\bar{q} + \bar{p}q$, the binary entropy function $H_b(p) := -p\log p - (1-p)\log(1-p)$, and the binary divergence function $D_b(q\|p) := q\log(q/p) + (1-q)\log((1-q)/(1-p))$.

We first focus on two examples with symmetric subsets.

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

$$\mathcal{L}_n := \{x^n \in \mathcal{X}^n : w_H(x^n) = \lfloor nq \rfloor\}, \quad 0 \le q \le 1. \quad (110)$$

This subset is smooth and symmetric, and $B(q)$ is the only distribution that intersects the subset $\mathcal{L}$. One can verify the latter by computing $H_{\mathcal{L}}(B(q))$ as defined in Definition 3 based on the normalized log-size of the intersection of the subset $\mathcal{L}_n$ with the typical set corresponding to $B(q)$. The symmetric property is trivial based on the weight constraint. Finally, the smoothness property comes from the fact that the function $g_P(Q) = D_b(q\|p)$ introduced in (34) is defined only at one q point, so is trivially continuous in q. Therefore, $Q^*_X = B(q)$. We obtain from Corollary 1 for the lossless compression that

$$R^*_{\mathcal{L}} = \mathbb{H}(Q^*_X) = H_b(q), \quad (111)$$

and from Theorem 4 for the lossy compression that

$$R_{\mathcal{L}}(D) = R(Q^*_X, D) = \begin{cases} H_b(q) - H_b(D), & 0 \le D \le \min\{q, \bar{q}\} \\ 0, & D > \min\{q, \bar{q}\}, \end{cases} \quad (112)$$

where the latter follows from the calculations for the standard rate-distortion function of the binary source [1], [2]. It is evident that the optimal lossless compression rate (111) for this subset can be below or above the source entropy. Similarly, the subset rate-distortion function (112) in this example can be below or above the rate-distortion function of the source; cf., Remark 5. We illustrate the latter comparison in Figure 3. The subset in this example, in the limit of large n, converges to an i.i.d. probability distribution, therefore falls in the framework of CoLT as discussed in [1] and [18]. Accordingly, a CoLT-based analysis, as discussed in Section III-B, can be also invoked to derive similar results for this example.

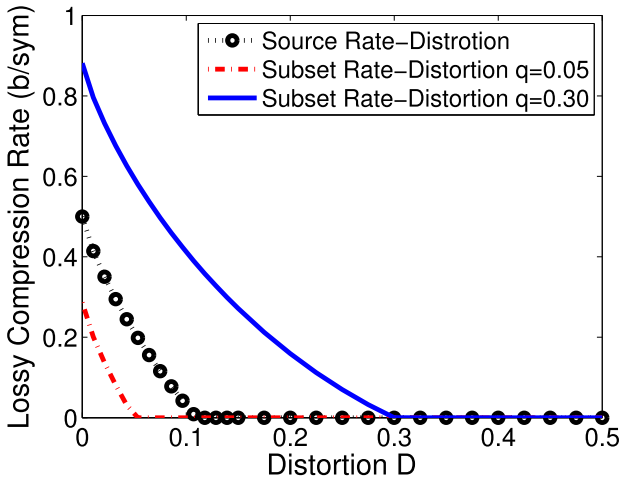

Fig. 3. Comparison of the subset rate-distortion function of Example 1, binary sequences with normalized Hamming weight q, with the rate-distortion function of the source for a Bernoulli DMS with parameter $p = 0.11$.

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

$$\mathcal{L}_n := \{x^n \in \mathcal{X}^n : 0 \le w_H(x^n) \le nq\}, \quad 0 \le q \le 1. \quad (113)$$

For the case $q \ge p$, the subset is likely so Theorem 1 implies that $R^*_{\mathcal{L}} = R^* = H_b(p)$, and $R_{\mathcal{L}}(D) = R(D) = H_b(p) - H_b(D)$ for $0 \le D \le p$ and $R_{\mathcal{L}}(D) = R(D) = 0$ for $D > p$. This subset is again smooth and symmetric. Moreover, $B(\bar{q})$ with $0 \le \bar{q} \le q$ are the only distributions that intersect the subset $\mathcal{L}$ and $H_b(\bar{q})$ is continuous over the interval $0 \le \bar{q} \le q$. One can verify these properties similar to Example 1, with the only exception that smoothness now follows from the function

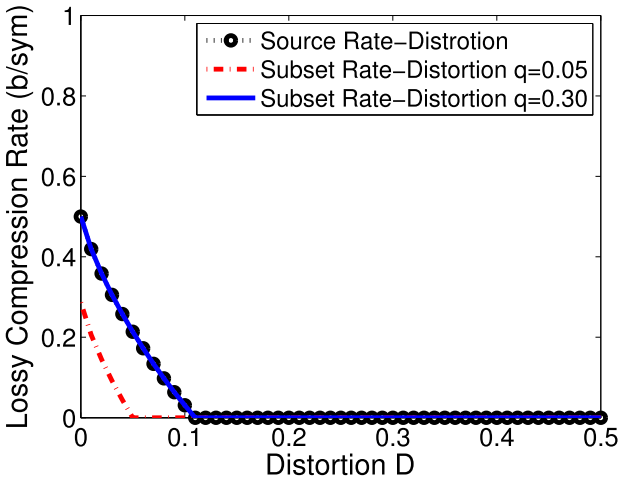

Fig. 4. Comparison of the subset rate-distortion function of Example 2, binary sequences with normalized Hamming weight not exceeding q, with the rate-distortion function of the source for a Bernoulli DMS with parameter $p = 0.11$.

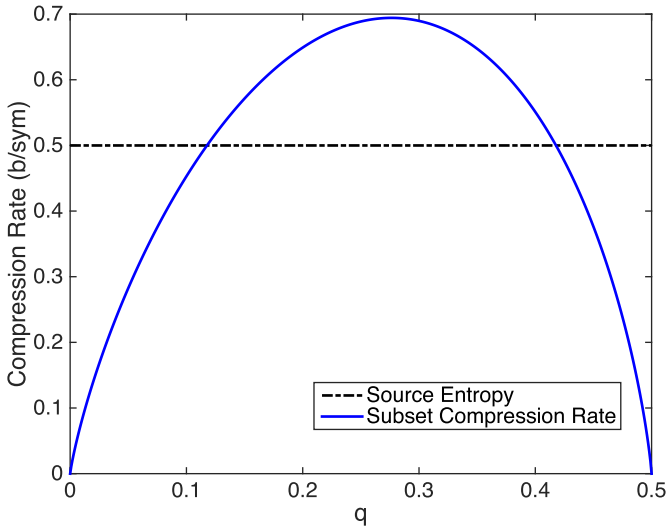

Fig. 5. Comparison of the optimal lossless subset-compression rate of Example 3, binary sequences with normalized Hamming weight q and with no consecutive 1s, with the source entropy for a Bernoulli DMS with parameter $p = 0.11$.

$g_P(Q) = D_b(\bar{q}\|p)$ being continuous in $\bar{q}$ over the interval $0 \le \bar{q} \le q$. Therefore, the subset-typical distribution is given by $Q_X^* = B(q^*)$, where

$$q^* = \arg\min_{\bar{q}:0\le\bar{q}\le q} D_b(\bar{q}\|p) = \min\{q, p\}. \tag{114}$$

Hence, we can use Corollary 1 to obtain for the optimal lossless compression rate that

$$R_{\mathcal{L}}^* = \mathbb{H}(Q_X^*) = H_b\left(\min\{q, p\}\right). \tag{115}$$

Analogously, we can use Theorem 4 to obtain for the subset rate-distortion that

$$\begin{aligned} R_{\mathcal{L}}(D) &= R(Q_X^*, D) \\ &= \begin{cases} H_b\left(\min\{q, p\}\right) - H_b(D), & 0 \le D \le \min\{q, p\} \\ 0, & D > \min\{q, p\}. \end{cases} \end{aligned} \tag{116}$$

It is evident that the optimal lossless compression rate for this subset never exceeds the source entropy. Similarly, the rate-distortion satisfies $R_{\mathcal{L}}(D) \le R(D)$ for all distortion values D. We illustrate the latter comparison in Figure 4. As both the formulas and the figures suggest, if $q < p$, a strictly positive rate gain can be achieved in both the lossless and lossy case by focusing only on the subset.

The subset in this example, in the limit of large n, converges to a convex set of i.i.d. probability distributions, therefore falls in the framework of CoLT as discussed in [1] and [18]. Accordingly, a CoLT-based analysis, as discussed in Section III-B, can be also invoked to derive similar results for this example.

In the following, we consider two smooth but non-symmetric examples to which Corollary 1 and Theorem 4 do not apply and instead require Theorems 2 and 3 and Corollary 2.

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

$$\mathcal{L}_n := \{x^n \in \mathcal{X}^n : w_H(x^n) = \lfloor nq \rfloor,\ x^n \text{ has no consecutive 1s}\}, \tag{117}$$

where $0 \le q \le 1/2$, since clearly $\mathcal{L}_n = \emptyset$ with $q > 1/2$.

This subset is not likely since it has exponentially small probability. However, it is smooth, and the only distribution that intersects the subset $\mathcal{L}$ is $B(q)$ with a subset entropy given by

$$\begin{aligned} H_{\mathcal{L}}(B(q)) &= \lim_{n\to\infty} \frac{1}{n} \log \binom{n - \lfloor nq \rfloor + 1}{\lfloor nq \rfloor} \\ &= (1-q) H_b\left(\frac{q}{1-q}\right). \end{aligned} \tag{118}$$

Therefore, we obtain from Theorem 2 for the lossless compression rate of this subset that

$$R_{\mathcal{L}}^* = H_{\mathcal{L}}(B(q)) = (1-q) H_b\left(\frac{q}{1-q}\right). \tag{119}$$

A plot of this compression rate is illustrated in Figure 5, which shows the subset compression rate (119) can be below or above the source entropy.

For the lossy compression, we can use Corollary 2 to obtain the following achievable rate-distortion pair:

$$R_{\mathcal{L}}^{(1)}(D) = \begin{cases} H_b(q) - H_b(D), & 0 \le D \le q, \\ 0, & D > q. \end{cases} \tag{120}$$

We can also use Theorem 3 to obtain another achievable rate-distortion pair. Let $0 \le D \le q$, and consider the following conditional distribution:

$$P_{Y|X}(0|0) = 1, \quad P_{Y|X}(0|1) = D/q, \tag{121}$$

so that $Q_Y^* = B(q - D)$. Note that, $P_{Y|X}(1|0) = 0$ under this conditional distribution, thus no 0 in x^n will flip to a 1 in y^n, hence the no-consecutive-1 structure will be preserved by the stochastic transformation from x to y. 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

$$\bar{\mathcal{L}}_n := \{y^n \in \mathcal{Y}^n : y^n \text{ has no consecutive 1s}\}. \tag{122}$$

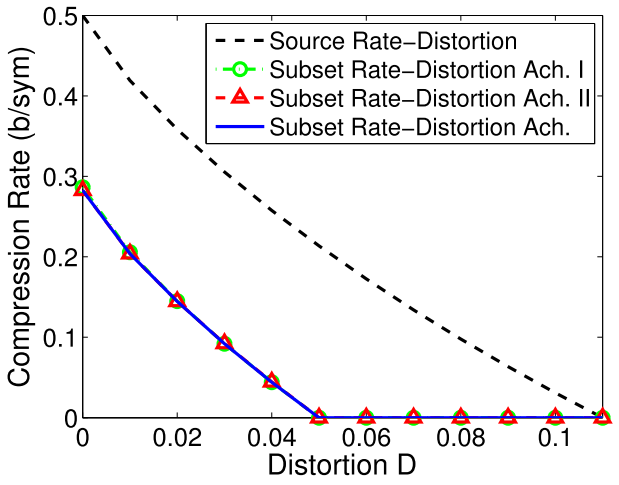

Fig. 6. Comparison of the subset rate-distortion function of Example 3, binary sequences with normalized Hamming weight q and with no consecutive 1s, with the rate-distortion function of the source for a Bernoulli DMS with parameter $p = 0.11$.

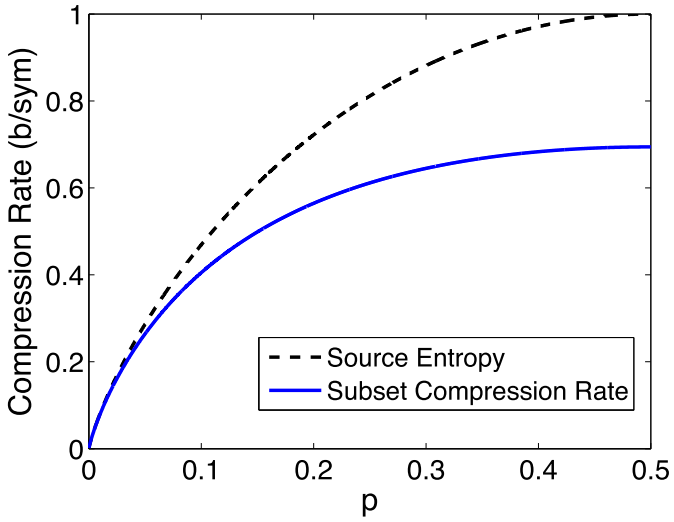

Fig. 7. Comparison of the optimal lossless subset-compression rate of Example 4, binary sequences with no consecutive 1s, with the source entropy for a Bernoulli DMS with parameter p.

In this case, we get

$$H_{\tilde{\mathcal{L}}}(Q_Y^*) = \lim_{n\to\infty} \frac{1}{n} \log \binom{n - n(q-D) + 1}{n(q-D)} = (1 - q + D) H_b\left(\frac{q-D}{1-q+D}\right) \tag{123}$$

and

$$H_{\tilde{\mathcal{L}}|\mathcal{L}}(P_{Y|X}|Q_X^*) = \lim_{n\to\infty} \frac{1}{n} \log \binom{nq}{nD} = q H_b\left(\frac{D}{q}\right), \tag{124}$$

so we obtain the following achievable rate-distortion pair:

$$R_{\mathcal{L}}^{(2)}(D) = \begin{cases} (1-q+D) H_b\left(\frac{q-D}{1-q+D}\right) - q H_b\left(\frac{D}{q}\right), & 0 \le D \le q, \\ 0, & D > q. \end{cases} \tag{125}$$

Hence, we arrive at the following result:

$$R_{\mathcal{L}}(D) \le \min\{R_{\mathcal{L}}^{(1)}(D), R_{\mathcal{L}}^{(2)}(D)\}. \tag{126}$$

Note that, (126) is not a convex function in D, and it is unclear whether time-sharing 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 weight condition as in the original sequence; see Remark 2. In any case, even the achievable rate-distortion (126) already shows gains over the rate-distortion function of the original source for some cases, as shown in Figure 6.

The subset in this example, in the limit of large n, converges to a Markov chain, therefore falls in the framework of CoLT for Markov processes as discussed in [36] and [40]. In principle, a CoLT-based analysis with appropriate extensions, as discussed in Section III-B, can be also invoked to derive similar results for this example.

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

$$\mathcal{L}_n := \{x^n \in \mathcal{X}^n : x^n \text{ has no consecutive 1s}\}. \tag{127}$$

Again, Theorem 1 does not apply since the subset is not likely. In order to employ Theorems 2 and 3, we first note that all distributions $B(q)$ with $0 \le q \le 1/2$ intersect the subset $\mathcal{L}$, and each has a subset entropy given by (118). Therefore, this subset is smooth, and its subset-typical distribution is $Q_X^* = B(q^*)$ where

$$q^* = \arg\min_{0\le q\le 1/2} \left[H_b(q) - (1-q) H_b\left(\frac{q}{1-q}\right) + D_b(q\|p) \right]. \tag{128}$$

Hence, we obtain from Theorem 2 for the optimal lossless compression rate of this subset that

$$R_{\mathcal{L}}^* = (1-q^*) H_b\left(\tfrac{q^*}{1-q^*}\right). \tag{129}$$

A plot of this subset-compression rate is illustrated in Figure 7, which shows the optimal lossless compression rate (129) of this subset is always below the source entropy.

For the lossy compression of this subset, we can use Corollary 2 to find an achievable rate-distortion pair as follows.

$$R_{\mathcal{L}}^{(1)}(D) = \begin{cases} H_b(q^*) - H_b(D), & 0 \le D \le q^* \\ 0, & D > q^*. \end{cases} \tag{130}$$

We can also build on Theorem 3 to obtain another achievable rate-distortion pair. Let $0 \le D \le q^*$, and consider the following conditional distribution:

$$P_{Y|X}(0|0) = 1, \quad P_{Y|X}(0|1) = D/q^*, \tag{131}$$

so that $Q_Y^* = B(q^* - D)$. Note that, $P_{Y|X}(1|0) = 0$ under this conditional distribution, 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

$$\bar{\mathcal{L}}_n := \{y^n \in \mathcal{Y}^n : y^n \text{ has no consecutive 1s}\}. \tag{132}$$

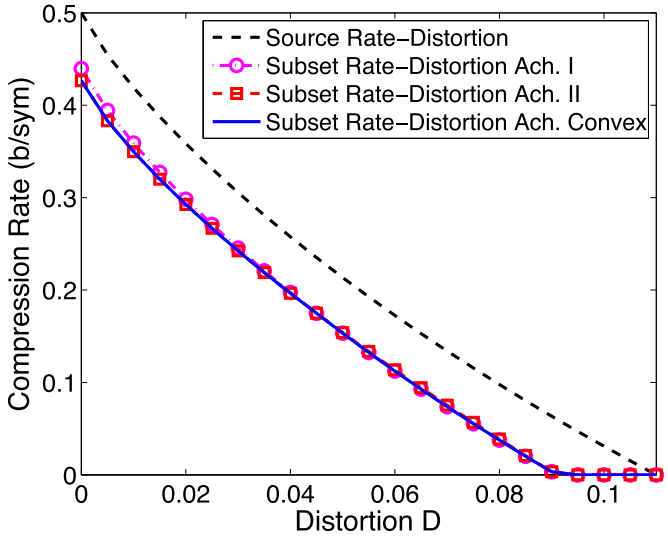

Fig. 8. Comparison of the achievable rate-distortion pair (136) and its components for the subset in Example 4, binary sequences with no consecutive 1s, with the rate-distortion function of the source for a Bernoulli DMS with parameter $p = 0.11$.

In this case, similar to (123) and (124), we get

$$H_{\tilde{\mathcal{L}}}(Q_Y^*) = (1 - q^* + D)H_b\left(\frac{q^* - D}{1 - q^* + D}\right), \quad (133)$$

$$H_{\tilde{\mathcal{L}}|\mathcal{L}}(P_{Y|X}|Q_X^*) = q^* H_b\left(\frac{D}{q^*}\right), \quad (134)$$

and obtain the following achievable rate-distortion pair:

$$R_{\mathcal{L}}^{(2)}(D) = \begin{cases} (1-q^*+D)H_b\left(\frac{q^*-D}{1-q^*+D}\right) - q^* H_b\left(\frac{D}{q^*}\right), & 0 \le D \le q^* \\ 0, & D > q^*. \end{cases} \quad (135)$$

Note that, unlike Example 3, 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:

$$R_{\mathcal{L}}(D) \le \text{l.c.e.}\left(\min\{R_{\mathcal{L}}^{(1)}(D), R_{\mathcal{L}}^{(2)}(D)\}\right), \quad (136)$$

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 3 is currently known, we cannot guarantee that (136) is optimal for $0 \le D \le q^*$. However, Figure 8 shows that even the achievable subset rate-distortion in (136) can sometimes outperform the rate-distortion function of the original source and already provide lossy compression gains.

The subset in this example, in the limit of large n, converges to a set of Markov chains, therefore falls in the framework of CoLT for Markov processes as discussed in [36] and [40]. In principle, a CoLT-based analysis with appropriate extensions, as discussed in Section III-B, can be also invoked to derive similar results for this example.

Finally, we present a fluctuating example for which Theorems 2 and 3 are not directly applicable. However, the characterizations of Theorem 5 and Corollary 3 facilitate the analysis.

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

$$\mathcal{L}_{1,n} := \{x^n \in \mathcal{X}^n : \quad nq_1 \le w_H(x^n) \le nq_2, \quad x^n \text{ has no consecutive 1s}\}, \quad (137)$$

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

$$\mathcal{L}_{2,n} := \{x^n \in \mathcal{X}^n : \quad nw_1 \le w_H(x^n) \le nw_2, \quad x^n \text{ has 1s only in even positions}\}, \quad (138)$$

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

$$\mathcal{L}_n := \begin{cases} \mathcal{L}_{1,n} & \text{if } n \text{ odd} \\ \mathcal{L}_{2,n} & \text{if } n \text{ even} \end{cases}. \quad (139)$$

Note that, Theorem 1 does not apply since the fluctuating subset $\mathcal{L}$ is not likely, and Theorems 2 and 3 do not apply since the subset is not smooth. However, both components are smooth subsets. In particular, the first subset component $\mathcal{L}_1$ is smooth and intersects all distributions $B(q)$ with $q_1 \le q \le q_2$, so that its subset-typical distribution is $Q_X^{*(1)} = B(q^*)$ where

$$q^* = \arg\min_{q_1 \le q \le q_2} \left[H_b(q) - (1-q)H_b\left(\frac{q}{1-q}\right) + D_b(q\|p) \right]. \quad (140)$$

An analysis similar to those in Examples 3 and 4 implies for the lossless compression that

$$R_{\mathcal{L}_1}^* = (1 - q^*)H_b\left(\frac{q^*}{1-q^*}\right), \quad (141)$$

and for lossy compression that

$$R_{\mathcal{L}_1}(D) \le \min\{R_{\mathcal{L}_1}^{(1)}(D), R_{\mathcal{L}_1}^{(2)}(D)\}, \quad (142)$$

where $R_{\mathcal{L}_1}^{(1)}(D)$ and $R_{\mathcal{L}_2}^{(2)}(D)$ are as in (130) and (135), respectively, with q^* as given in (140). Note that, (142) is not a convex function in D, and it is 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 weight structure as that in the original sequence.

The second subset component $\mathcal{L}_2$ is also smooth and intersects all distributions $B(w)$ with $w_1 \le w \le w_2$ with a subset entropy given by

$$H_{\mathcal{L}}(B(w)) = \lim_{n\to\infty} \frac{1}{n} \log \binom{\lfloor n/2 \rfloor}{\lfloor nw \rfloor} = \frac{1}{2} H_b(2w), \quad (143)$$

so that $Q_X^{*(2)} = B(w^*)$ where

$$w^* = \arg\min_{w_1 \le w \le w_2} \left[H_b(w) - \frac{1}{2} H_b(2w) + D_b(w\|p) \right]. \quad (144)$$

We can then use Theorem 2 to find for the optimal lossless compression rate of this subset that

$$R_{\mathcal{L}_2}^* = \frac{1}{2} H_b(2w^*), \quad (145)$$

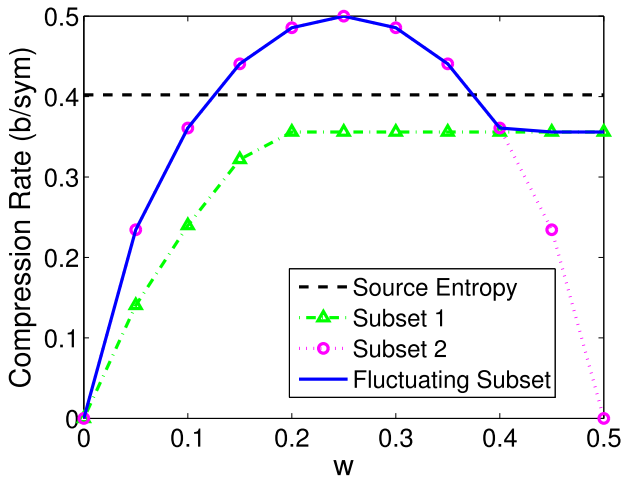

Fig. 9. The optimal lossless compression rate of the fluctuating subset of Example 5 for a binary DMS with fixed parameter $p = 0.08$ and varying subset parameters $q_1 = 0$, $q_2 = 0.4w$, $w_1 = w_2 = w$ where $0 \le w \le 1/2$.

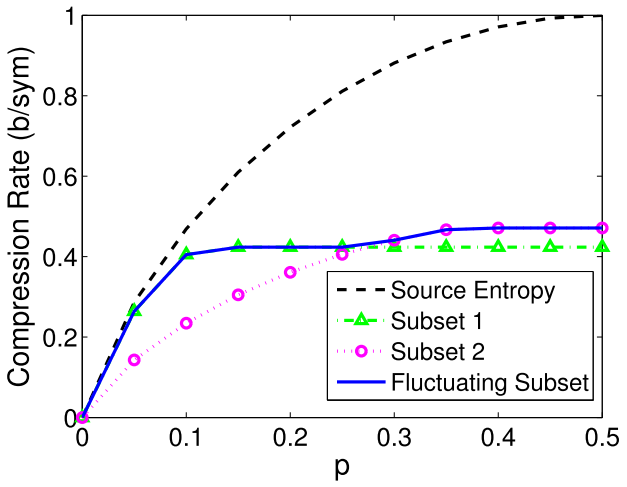

Fig. 10. The optimal lossless compression rate of the fluctuating subset of Example 5 for fixed subset parameters $q_1 = 0$, $q_2 = 0.09$ and $w_1 = 0$, $w_2 = 0.18$ and a binary DMS with varying parameter $0 \le p \le 1/2$.

and apply Corollary 2 to get for the lossy compression that

$$R_{\mathcal{L}_2}(D) \le \begin{cases} H_b(w^*) - H_b(D), & 0 \le D \le w^* \\ 0, & D > w^*. \end{cases} \tag{146}$$

Substituting (141) and (145) in Theorem 5 yields for the lossless compression of the fluctuating that

$$\begin{aligned} R^*_{\mathcal{L}} &= \max\left\{R^*_{\mathcal{L}_1}, R^*_{\mathcal{L}_2}\right\} \\ &= \max\left\{(1-q^*)H_b\left(\frac{q^*}{1-q^*}\right), \frac{1}{2}H_b(2w^*)\right\}. \end{aligned} \tag{147}$$

To show the different aspects of this scenario, we make two comparisons for the lossless case. 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 $0 \le w \le 1/2$. The lossless compression rate for this fluctuating subset is shown in Figure 9. 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 above 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 $0 \le p \le 1/2$. The lossless compression rate for this fluctuating subset is shown in Figure 10. In this case, the compression rate of the fluctuating 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.

Analogously, substituting (142) and (146) in Theorem 5 yields for lossy compression of the fluctuating subset that

$$R_{\mathcal{L}}(D) \le \max\left\{\min\{R^{(1)}_{\mathcal{L}_1}(D), R^{(2)}_{\mathcal{L}_1}(D)\}, R_{\mathcal{L}_2}(D)\right\}. \tag{148}$$

This readily implies $R_{\mathcal{L}}(D) = 0$ for $D > \max\{q^*, w^*\}$, but the optimality of (148) for $0 \le D \le \max\{q^*, w^*\}$ is unknown in the absence of a converse for our Theorem 3. For instance,

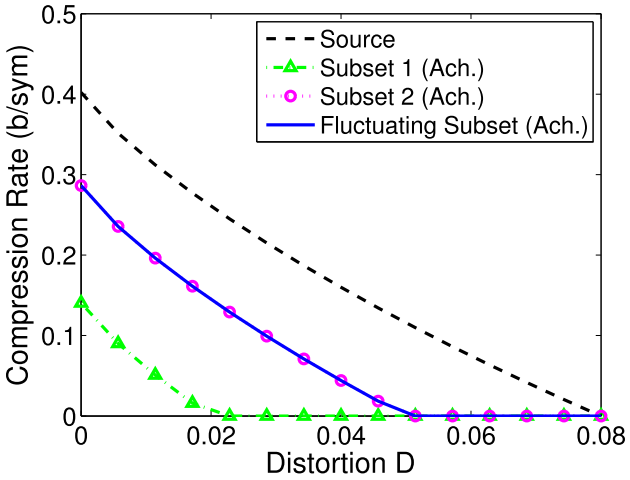

Fig. 11. The achievable rate-distortion pair via (142) and (146) for the fluctuating subset of Example 5 with parameters $q_1 = 0$, $q_2 = 0.4w$, $w_1 = w_2 = w = 0.05$ and a binary DMS with parameter $p = 0.08$.

the compression rate currently achieved by (148) for the zero-distortion case, $D = 0$, is

$$\max\left\{(1-q^*)H_b\left(\frac{q^*}{1-q^*}\right), H_b(w^*)\right\}, \tag{149}$$

which is strictly worse than the anticipated result from the lossless analysis (147). However, Figure 12 shows that even the achievable subset rate-distortion in (148) 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.

The subset components in this example, in the limit of large n, are intersections of a convex set of i.i.d. probability distributions with a Markov chain. Therefore, in principle, one might be able to invoke certain methods (e.g., Bayes' rule) to apply the framework of CoLT for i.i.d. distributions as in [1] and [18] and that for Markov processes as

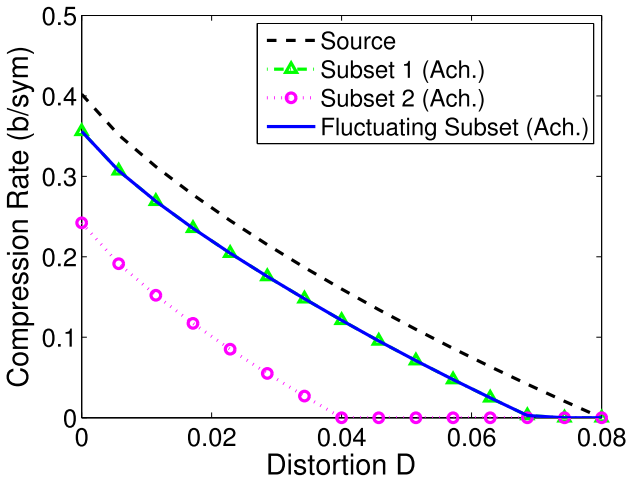

Fig. 12. The achievable rate-distortion pair via (142) and (146) for the fluctuating subset of Example 5 with parameters $q_1 = 0$, $q_2 = 0.09$ and $w_1 = 0$, $w_2 = 0.18$ and a binary DMS with parameter $p = 0.08$.

in [36] and [40], after appropriate extensions as discussed in Section III-B. However, the fluctuating source property is a different aspect, which apparently needs new forms of CoLT (with the quasi-independence feature).

VIII. Generalization to Subsets With Weighted Priorities

In this section, we describe a generalization of our framework to subsets with weighted priorities. Our main framework, as explained in Section II, can be thought of as *a 0-1 priority* setting, in which the sequences x^n belonging to the subset $\mathcal{L}_n$ are the only focus and have a *weight* of 1, and all other sequences x^n outside $\mathcal{L}_n$ are not important at all and have a weight of 0. A more general setting can be one in which different weights between 0 and 1 can be assigned to sequences, each capturing the relative importance or *priorities* of the sequences. One can consider this as a source coding dual of the unequal error protection problem in channel coding [30], [31].

In particular, consider the subset $\mathcal{L} = \{\mathcal{L}_n \subseteq \mathcal{X}^n\}_{n=1}^{n}$ and a fixed partition of it such that

$$\mathcal{L}_n = \bigcup_{k=1}^{K} \mathcal{S}_{n,k} \ , \tag{150}$$

with $\mathcal{S}_{n,k} \cap \mathcal{S}_{n,k'} = \emptyset$ for all $n = 1, 2, \cdots$ and any $k \neq k' \in \{1, \cdots, K\}$ for a fixed finite number K. Accordingly, for each $k = 1, \cdots, K$, consider the partition components as individual subsets as defined below.

$$\mathcal{S}_k = \left\{\mathcal{S}_{n,k}\right\}_{n=1}^{\infty}, \tag{151}$$

and denote the overall partition by $\mathcal{S} = (\mathcal{S}_1, \cdots, \mathcal{S}_K)$. Now, consider a priority or weight vector $w = (w_1, \cdots, w_K)$ such that $0 \leq w_k \leq 1$ and $\sum_{k=1}^{K} w_k = 1$. Here, w_k represents the priority of the partition $\mathcal{S}_k$.

We define an $(n, 2^{nR})$ lossless code for the subset $\mathcal{L}$ with partition $\mathcal{S}$ and weight vector w to consist of an encoder $m : \mathcal{L}_n \to \{1, 2, \cdots, 2^{nR}\}$ and a decoder $\hat{x}^n : \{1, 2, \cdots, 2^{nR}\} \to \mathcal{L}_n \cup \{\mathcal{E}\}$ with error probability

$$\Pr[\mathcal{E}_{\mathcal{S},w}] := \sum_{k=1}^{K} w_k \Pr[\hat{X}^n \neq X^n | X^n \in \mathcal{S}_{n,k}]. \tag{152}$$

A rate R is called achievable if a sequence of $(n, 2^{nR})$ lossless codes for the subset $\mathcal{L}$ with partition $\mathcal{S}$ and weight vector w exists such that $\Pr[\mathcal{E}_{\mathcal{S},w}] \to 0$ as $n \to \infty$. The optimal lossless subset compression rate $R^*_{\mathcal{S},w}$ is the infimum of all achievable rates.

Analogously, we define an $(n, 2^{nR})$ lossy code with distortion level D for the subset $\mathcal{L}$ with partition $\mathcal{S}$ and weight vector w to consist of an encoder $f : \mathcal{L}_n \to \{1, 2, \cdots, 2^{nR}\}$ and a decoder $\phi : \{1, 2, \cdots, 2^{nR}\} \to \mathcal{Y}^n$ with the excess-distortion probability

$$\Pr[\mathcal{E}_{\mathcal{S},w}(D)] := \sum_{k=1}^{K} w_k \Pr[d(X^n, Y^n) > D | X^n \in \mathcal{S}_{n,k}]. \tag{153}$$

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

Our main result below states that, the performance is dictated by that of the *worst* partition component. In particular, the priority or weight vector does not affect the performance, so long as the number or weight of the partition does not vary with n. Hence, an immediate consequence of Theorem 6 below is that, even for this generalized model with weighted priorities, we can stay within the original framework of Section II and use the results we have already developed in Sections IV, V, and VI.

Theorem 6: For a discrete memoryless source $P(x)$ and a subset $\mathcal{L} = \{\mathcal{L}_n \subseteq \mathcal{X}^n\}_{n=1}^{\infty}$ with partition $\mathcal{S} = (\mathcal{S}_1, \cdots, \mathcal{S}_K)$ and weight vector $w = (w_1, \cdots, w_K)$, where K is a finite integer, the optimal lossless subset-compression rate is

$$R^*_{\mathcal{S},w} = \max_{1 \leq k \leq K} R^*_{\mathcal{S}_k}, \tag{154}$$

and the subset rate-distortion function is

$$R_{\mathcal{S},w}(D) = \max_{1 \leq k \leq K} R_{\mathcal{S}_k}(D). \tag{155}$$

Proof: We state only the proof for the lossless case; that for the lossy case is very similar and we skip the details for brevity. (Achievability) Fix an arbitrary $\epsilon > 0$. For each $1 \leq k \leq K$, let $\{(m_{n,k}, \hat{x}^n_k)\}_{n=1}^{\infty}$ be the optimal encoder and decoder sequence for lossless compression of the partition component $\mathcal{S}_k$, achieving a rate $R^*_{\mathcal{S}_k} + \epsilon$ with vanishing error probability $\Pr[\hat{X}^n_k \neq X^n | X^n \in \mathcal{S}_{n,k}] \to 0$ as $n \to \infty$. We consider the following code for the weighted subset $\mathcal{L}$: let $m_n(x^n) = (k, m_{n,k}(x^n))$ if $x^n \in \mathcal{S}_{n,k}$, and set $\hat{x}^n(k, m) = \hat{x}^n_k(m)$; basically, we juxtapose the partition index with the

codeword used within that partition. Then, we have

$$\limsup_{n\to\infty} \sum_{k=1}^{K} w_k \Pr[\hat{X}_k^n \neq X^n | X^n \in \mathcal{S}_{n,k}] = \sum_{k=1}^{K} w_k \limsup_{n\to\infty} \Pr[\hat{X}_k^n \neq X^n | X^n \in \mathcal{S}_{n,k}] = 0. \quad (156)$$

The rate of this code is

$$\lim_{n\to\infty} \frac{1}{n} \log \sum_{k=1}^{K} 2^{n(R^*_{\mathcal{S}_k}+\epsilon)} = \max_{1\leq k\leq K} R^*_{\mathcal{S}_k} + \epsilon. \quad (157)$$

Since ϵ is arbitrary, this completes the achievability proof in the lossless case.

(Converse) Consider any arbitrary code of rate R for subset $\mathcal{L}$ with partition $\mathcal{S}$ and weight vector w such that

$$\limsup_{n\to\infty} \sum_{k=1}^{K} w_k \Pr[\hat{X}_k^n \neq X^n | X^n \in \mathcal{S}_{n,k}] = 0. \quad (158)$$

Then, since none of the weights w_k varies with n, we must have for all $1 \leq k \leq K$ that

$$\limsup_{n\to\infty} \Pr[\hat{X}_k^n \neq X^n | X^n \in \mathcal{S}_{n,k}] = 0. \quad (159)$$

The definition of $R^*_{\mathcal{S}_k}$ then implies that $R > R^*_{\mathcal{S}_k}$ for all $1 \leq k \leq K$, which in turn implies $R > \max_{1\leq k\leq K} R^*_{\mathcal{S}_k}$ and proves the converse for the weighted subset $\mathcal{L}$ in the lossless case. □

IX. Concluding Remarks

A. Recap of the Results

We have provided a framework for lossless and lossy compression of subsets of discrete memoryless sources as well as several optimality results for broad classes of subsets including likely subsets, smooth subsets, fluctuating subsets, and subsets with weighted priorities. In particular, for smooth subsets that intersect only a continuous range of distributions, we have demonstrated that the lossless compression performance is mainly dictated by subset-typical distributions that optimize the trade-off between the closeness to the source statistics and the size of intersection with the subset. Moreover, 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 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.

B. Discussion on Computability of the Results

One very valid concern about the results presented in this paper is regarding the extent to which computation of these fundamental results is possible for various subset structures. We believe, the end goal of such studies as ours, at least in part, is to facilitate computable fundamental limits. In this subsection, we would like to clarify this issue and (at least partially) resolve this concern.

In this paper, we have identified several classes of subsets (likely, smooth, fluctuating, weighted priorities) for which we can compute the fundamental limits exactly or via bounds. We have also presented, in Section VII, several examples from more basic structures to more complex ones to showcase some computation scenarios. Of course, our main intention from providing these examples has been to demonstrate the implications of imposing structures on the data compression problem and to show in what situations one might get a gain (or, perhaps surprisingly, loss) by only focusing on a subset of realizations, rather than the entire ensemble.

Our results, similar to many well-known results in probability and information theory, facilitate rather nice, closed-form, and computable solutions provided a "nice structure" is considered in the problem setting. For example, in Sanvo's theorem and CoLT (with exact or quasi independence), one obtains computable results by focusing on closed and (almost completely) convex sets E of probability distributions, such as those resulting from empirical block average or Markovian/sliding empirical block average constraints. Similarly, we focus on smooth subsets that satisfy certain regularity conditions. Our results for fluctuating subsets and subset with weighted priorities are attempts towards more general subsets/structures.

In the problem of subset source coding, our objective functions are the subset entropy $H_{\mathcal{L}}(P_X)$ or the subset mutual information terms $I_{\mathcal{L},\bar{\mathcal{L}}}(P_X, P_{Y|X})$ that are, in general, non-convex objective functions. Moreover, the optimization is over subsets that are, in general, non-convex domains. Therefore, the limits might not be necessarily computable if we consider arbitrary structures. Following the discussion in section III-A, imposing completely arbitrary structures on the subset $\mathcal{L}$ quickly turns the problem into compression of non-stationary and non-ergodic equivalent sources, for which a closed-form, computable solution usually appears far from rich and therefore outside the scope of our work.

As a final remark, we would like to mention that, our generic results on performance limits of rather arbitrary subsets provide at least some insights into the key issues arising in the problem of subset source coding. In that sense, at least to some extent, we may compare our treatment with that of the information spectrum approach of Verdú-Han [28] for non-stationary and non-ergodic sources, whose results are very interesting and insightful but, in general, non-computable. Arguably, the only new scenario for which the information spectrum approach can provide computable results is for the fundamental performance limits of mixed sources and channels with general mixtures [28]. In our work, the category of fluctuating subsets, whose structure switches among a few options as showcased in our Example 5, has a similar spirit and facilitate computable results.

C. Future Directions

We envision at least two immediate directions for future research on this topic. One is to expand the setting of subsets with weighted priorities in Section VIII to partitions whose number K_n or weight vector w_n rapidly varies with blocklength n. Additionally, compression limits of subsets not covered by our current analysis is of future interest.

Appendix A
Analysis of the Equivalent Conditional Source

For the equivalent conditional source $\tilde{X}^n$ defined as

$$P_{\tilde{X}^n}(x^n) := \frac{P_{X^n}(x^n)}{P_{X^n}[X^n \in \mathcal{L}_n]} 1\{x^n \in \mathcal{L}_n\}, \tag{160}$$

the fundamental lossless compression rate is identical to our $R^*_{\mathcal{L}}$ of interest since the error probability for both cases is the same:

$$\begin{aligned}
&\Pr[\hat{X}^n \neq X^n | X^n \in \mathcal{L}_n] \\
&\quad = \frac{P_{X^n}[\hat{X}^n \neq X^n, X^n \in \mathcal{L}_n]}{P_{X^n}[X^n \in \mathcal{L}_n]} && (161)\\
&\quad = \frac{\sum_{x^n} P_{X^n}(x^n) 1\{\hat{x}(m(x^n)) \neq x^n, x^n \in \mathcal{L}_n\}}{P_{X^n}[X^n \in \mathcal{L}_n]} && (162)\\
&\quad = \sum_{x^n} \frac{P_{X^n}(x^n)}{P_{X^n}[X^n \in \mathcal{L}_n]} 1\{x^n \in \mathcal{L}_n\} \cdot 1\{\hat{x}(m(x^n)) \neq x^n\} && (163)\\
&\quad = \sum_{x^n} P_{\tilde{X}^n}(x^n) 1\{\hat{x}(m(x^n)) \neq x^n\} && (164)\\
&\quad = \Pr[\widehat{\tilde{X}^n} \neq \tilde{X}^n]. && (165)
\end{aligned}$$

Analogously, the fundamental lossy compression rate of this equivalent conditional source is identical to our $R_{\mathcal{L}}(D)$ of interest since the probability of excess-distortion for both cases is the same:

$$\begin{aligned}
&\Pr[d(X^n, Y^n) > D | X^n \in \mathcal{L}_n] \\
&\quad = \frac{P_{X^n}[d(X^n, Y^n) > D, X^n \in \mathcal{L}_n]}{P_{X^n}[X^n \in \mathcal{L}_n]} && (166)\\
&\quad = \frac{\sum_{x^n} P_{X^n}(x^n) 1\{d(x^n, \phi(f(x^n))) > D, x^n \in \mathcal{L}_n\}}{P_{X^n}[X^n \in \mathcal{L}_n]} && (167)\\
&\quad = \sum_{x^n} \frac{P_{X^n}(x^n)}{P_{X^n}[X^n \in \mathcal{L}_n]} 1\{x^n \in \mathcal{L}_n\} \cdot 1\{d(x^n, \phi(f(x^n))) > D\} && (168)\\
&\quad = \sum_{x^n} P_{\tilde{X}^n}(x^n) 1\{d(x^n, \phi(f(x^n))) > D\} && (169)\\
&\quad = \Pr[d(\tilde{X}^n, Y^n) > D]. && (170)
\end{aligned}$$

Appendix B
Proof of Lemma 1

Recall from the properties of type classes that, all sequences $x^n \in T^n(\hat{P})$ satisfy [29]

$$P_{X^n}(x^n) = 2^{-n[\mathbb{H}(\hat{P}) + D(\hat{P} \| P)]}. \tag{171}$$

On the other hand, the existence of the subset entropy $H_{\mathcal{L}}(Q)$ as defined in (30) implies that, there exists some $\xi_n \to 0$ as $n \to \infty$ such that

$$H_{\mathcal{L}}(Q) - \xi_n \leq \frac{1}{n} \log \left| T^n_{\mathcal{L}}[Q]_{\delta_n} \right| \leq H_{\mathcal{L}}(Q) + \xi_n. \tag{172}$$

Now, note that

$$\begin{aligned}
|T^n_{\mathcal{L}}[Q]_{\delta_n}| \min_{x^n \in T^n[Q]_{\delta_n}} P_{X^n}(x^n) &\leq P_{X^n}[X^n \in T^n_{\mathcal{L}}[Q]_{\delta_n}] \\
&\leq |T^n_{\mathcal{L}}[Q]_{\delta_n}| \max_{x^n \in T^n[Q]_{\delta_n}} P_{X^n}(x^n). \qquad (173)
\end{aligned}$$

But, (171) and the continuity of the Shannon entropy and relative entropy implies the existence of some $\xi'_n \to 0$ such that

$$\begin{aligned}
\min_{x^n \in T^n[Q]_{\delta_n}} P_{X^n}(x^n) &\geq 2^{-n[\mathbb{H}(Q) + D(Q\|P) + \xi'_n]}, \\
\max_{x^n \in T^n[Q]_{\delta_n}} P_{X^n}(x^n) &\leq 2^{-n[\mathbb{H}(Q) + D(Q\|P) - \xi'_n]}. \qquad (174)
\end{aligned}$$

Combining (173) with (172), (174), and recalling the definition of the function $g_P(Q)$ in (34) completes the proof of the first part of the lemma in (49) with $\epsilon_n := \xi_n + \xi'_n$.
To prove the second part, we note that

$$\begin{aligned}
P_{X^n}[X^n \in \mathcal{L}_n] &= P_{X^n}[X^n \in \bigcup_{\hat{P}: n\text{-type}} T^n_{\mathcal{L}}(\hat{P})] \\
&= \sum_{\hat{P}: n\text{-type}, T^n_{\mathcal{L}}(\hat{P}) \neq \emptyset} P_{X^n}[X^n \in T^n_{\mathcal{L}}(\hat{P})]. \qquad (175)
\end{aligned}$$

We can lower bound the summation in (175) by any group of summands including the ones belonging to the Q-typical set with maximum probability,

$$P_{X^n}[X^n \in \mathcal{L}_n] \geq \max_{Q: \mathcal{L} \cap T[Q] \neq \emptyset} P_{X^n}[X^n \in T^n_{\mathcal{L}}[Q]_{\delta_n}], \tag{176}$$

and we can upper bound (175) by recalling the Type Counting Lemma and upper bounding each term with the one having the maximum probability, and noting that expanding the maximization domain from types to general distributions can not decrease the probability.

$$\begin{aligned}
&P_{X^n}[X^n \in \mathcal{L}_n] \\
&\quad \leq (n+1)^{|\mathcal{X}|} \max_{\hat{P}: n\text{-type}, T^n_{\mathcal{L}}(\hat{P}) \neq \emptyset} P_{X^n}[X^n \in T^n_{\mathcal{L}}(\hat{P})] && (177)\\
&\quad \leq (n+1)^{|\mathcal{X}|} \max_{Q: \mathcal{L} \cap T[Q] \neq \emptyset} P_{X^n}[X^n \in T^n_{\mathcal{L}}[Q]_{\delta_n}]. && (178)
\end{aligned}$$

Combining (176) and (178) with the result in (49) completes the proof of (50) and that of Lemma 1.

Appendix C
Proof of Lemma 2

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 X is distributed according to $\hat{P}(x)$. Let the Y's distribution be $P_Y(y)$. 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 7. 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 value of M will be specified later in the proof. We define the set of *uncovered* x^n sequences in $T^n_{\mathcal{L}}(\hat{P})$ by these Y^n sequences as follows.

$$U\left(\{Y^n(m)\}_{m=1}^M\right) := \left\{x^n \in T^n_{\mathcal{L}}(\hat{P}) : d\left(x^n, \{Y^n(m)\}_{m=1}^M\right) > D\right\}. \quad (179)$$

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

We first note that

$$\mathbb{E}\left[\left|U\left(\{Y^n(m)\}_{m=1}^M\right)\right|\right] = \mathbb{E}\left[\sum_{x^n \in T^n_{\mathcal{L}}(\hat{P})} 1\left\{d\left(x^n, \{Y^n(m)\}_{m=1}^M\right) > D\right\}\right] \quad (180)$$

$$= \sum_{x^n \in T^n_{\mathcal{L}}(\hat{P})} \Pr\left[d\left(x^n, \{Y^n(m)\}_{m=1}^M\right) > D\right]. \quad (181)$$

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

$$\Pr\left[d\left(x^n, \{Y^n(m)\}_{m=1}^M\right) > D\right] = \Pr\left[\bigcap_{m=1}^M d\left(x^n, Y^n(m)\right) > D\right] \quad (182)$$

$$= \prod_{m=1}^M \Pr\left[d\left(x^n, Y^n(m)\right) > D\right] \quad (183)$$

$$= \left(1 - \Pr\left[d\left(x^n, Y^n(1)\right) \le D\right]\right)^M \quad (184)$$

$$\le 2^{-M \cdot \Pr\left[d(x^n, Y^n(1)) \le D\right]}, \quad (185)$$

where the last line follows from the inequality $(1-x)^n \le 2^{-nx}$ for all $0 \le x \le 1$ and $n > 0$.

We now analyze the probability in (185) 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

$$\Pr\left[d\left(x^n, Y^n(1)\right) \le D\right] = \frac{|\{y^n \in \bar{\mathcal{L}}_n \cap T^n[P_Y]_{\delta_n} : d(x^n, y^n) \le D\}|}{|\bar{\mathcal{L}}_n \cap T^n[P_Y]_{\delta_n}|} \quad (186)$$

$$\ge \frac{|\bar{\mathcal{L}}_n \cap T^n[P_{Y|X}|x^n]_{\delta_n}|}{|\bar{\mathcal{L}}_n \cap T^n[P_Y]_{\delta_n}|}, \quad (187)$$

where (187) 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

$$d(x^n, y^n) = \sum_{x,y} \frac{1}{n} N((x,y); (x^n, y^n)) d(x,y) \quad (188)$$

$$\le \sum_{x,y} (\hat{P}_X(x) P_{Y|X}(y|x) + \delta_n) d(x,y) \quad (189)$$

$$\le \mathbb{E}[d(X,Y)] + |\mathcal{X}||\mathcal{Y}|\delta_n D_{\max} \quad (190)$$

$$\le (D - \eta) + |\mathcal{X}||\mathcal{Y}|\delta_n D_{\max} \quad (191)$$

$$\le D, \quad (192)$$

for the case $D > \eta$. The last result (192) also holds for the case of target distortion level satisfying $D \le \eta$, 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 7 and 8 of $(\hat{P}_X, P_{Y|X}, \mathcal{L})$-smooth auxiliary subset that, there exists some $\xi_n \to 0$ as $n \to \infty$ such that, for any $x^n \in T^n_{\mathcal{L}}[\hat{P}_X]_{\delta_n}$,

$$|\bar{\mathcal{L}}_n \cap T^n[P_{Y|X}|x^n]_{\delta_n}| \ge \min_{x^n \in T^n_{\mathcal{L}}[\hat{P}_X]_{\delta_n}} |\bar{\mathcal{L}}_n \cap T^n[P_{Y|X}|x^n]_{\delta_n}| \ge 2^{n[H_{\bar{\mathcal{L}}|\mathcal{L}}(P_{Y|X}|\hat{P}_X) - \xi_n]}, \quad (193)$$

and that

$$|\bar{\mathcal{L}}_n \cap T^n[P_Y]_{\delta_n}| \le 2^{n[H_{\bar{\mathcal{L}}}(P_Y) + \xi_n]}. \quad (194)$$

Substituting (193) and (194) into (187) yields for all sequences in $T^n_{\mathcal{L}}(\hat{P}_X)$ that,

$$\Pr\left[d\left(x^n, Y^n(1)\right) \le D\right] \ge 2^{-n\left[H_{\bar{\mathcal{L}}}(P_Y) - H_{\bar{\mathcal{L}}|\mathcal{L}}(P_{Y|X}|\hat{P}_X) + 2\xi_n\right]} = 2^{-n\left[I_{\mathcal{L},\bar{\mathcal{L}}}(\hat{P}_X, P_{Y|X}) + 2\xi_n\right]}, \quad (195)$$

which along with (181) and (185) implies

$$\mathbb{E}\left[\left|U\left(\{Y^n(m)\}_{m=1}^M\right)\right|\right] \le \left|T^n_{\mathcal{L}}(\hat{P})\right| \cdot 2^{-M \cdot 2^{-n\left[I_{\mathcal{L},\bar{\mathcal{L}}}(\hat{P}_X, P_{Y|X}) + 2\xi_n\right]}} \quad (196)$$

$$\le 2^{n \log|\mathcal{X}| - M \cdot 2^{-n\left[I_{\mathcal{L},\bar{\mathcal{L}}}(\hat{P}_X, P_{Y|X}) + 2\xi_n\right]}}. \quad (197)$$

If we choose

$$M = 2^{n\left[I_{\mathcal{L},\bar{\mathcal{L}}}(\hat{P}_X, P_{Y|X}) + 3\xi_n\right]}, \quad (198)$$

we get for sufficiently large n that,

$$\mathbb{E}\left[\left|U\left(\{Y^n(m)\}_{m=1}^M\right)\right|\right] \le 2^{n \log|\mathcal{X}| - 2^{n\xi_n}} < 1. \quad (199)$$

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

$$\frac{1}{n} \log B(\hat{P}_X, \mathcal{L}) = \frac{1}{n} \log M \le \inf_{P_{Y|X}: \mathbb{E}[d(X,Y)] \le D} \inf_{\bar{\mathcal{L}}: (\hat{P}_X, P_{Y|X}, \mathcal{L})\text{-smooth}} I_{\mathcal{L},\bar{\mathcal{L}}}(\hat{P}_X, P_{Y|X}) + 3\xi_n. \quad (200)$$

APPENDIX D
PROOF OF LEMMA 3

If δ_n is chosen large enough yet still satisfying the Delta Convention, we can write

$$g_P(\bar{Q}) > g_P(Q_X^*) + 3\epsilon_n \tag{201}$$

for all $\bar{Q}$-typical sets which do not intersect at all with the subset-typical sets $\cup_{Q_X^* \in \mathcal{Q}^*} T_{\mathcal{L}}^n[Q_X^*]_{\delta_n}$. Therefore, we get from Lemma 1 that

$$\begin{aligned}
&\Pr\left[X^n \notin \bigcup_{Q_X^* \in \mathcal{Q}_X^*} T_{\mathcal{L}}^n[Q_X^*]_{\delta_n} \middle| X^n \in \mathcal{L}_n\right] \\
&= \frac{P_{X^n}\left[X^n \in \bigcup_{\bar{Q}: g_P(\bar{Q}) > g_P(Q_X^*)+3\epsilon_n} T_{\mathcal{L}}^n[\bar{Q}]_{\delta_n}\right]}{P_{X^n}[X^n \in \mathcal{L}_n]} && (202)\\
&\le \frac{(n+1)^{|\mathcal{X}|} 2^{-n[g_P(Q_X^*)+3\epsilon_n-\epsilon_n]}}{2^{-n[g_P(Q_X^*)+\epsilon_n]}} && (203)\\
&\le (n+1)^{|\mathcal{X}|} 2^{-n\epsilon_n}, && (204)
\end{aligned}$$

which goes to 0 as $n \to \infty$.

APPENDIX E
PROOF OF LEMMA 4

We build upon Lemma 1, Lemma 3, and inequality (174) to find

$$\begin{aligned}
\frac{1}{2}\eta &\le \Pr\left[X^n \in \left(\mathcal{A} \cap T_{\mathcal{L}}^n[Q_X^*]_{\delta_n}\right) \middle| X^n \in \mathcal{L}_n\right] && (205)\\
&\le \left|\mathcal{A} \cap T_{\mathcal{L}}^n[Q_X^*]_{\delta_n}\right| \max_{x^n \in T^n[Q_X^*]_{\delta_n}} \Pr\left[X^n = x^n \mid X^n \in \mathcal{L}_n\right] && (206)\\
&\le |\mathcal{A}| \frac{2^{-n[g_{P_X}(Q_X^*)+H_{\mathcal{L}}(Q_X^*)-\epsilon_n]}}{2^{-n[g_{P_X}(Q_X^*)+\epsilon_n]}}, && (207)
\end{aligned}$$

which implies

$$|\mathcal{A}| \ge \frac{1}{2}\eta 2^{n[H_{\mathcal{L}}(Q_X^*)-2\epsilon_n]} \ge 2^{n[H_{\mathcal{L}}(Q_X^*)-3\epsilon_n]}. \tag{208}$$

ACKNOWLEDGMENT

The authors greatly appreciate Associate Editor Prof. Sandeep Pradhan and the two anonymous reviewers for their extensive and detailed comments that significantly improved the presentation of this work, particularly regarding the connections of our work with the conditional limit theorem (CoLT). The authors acknowledge the helpful feedback on the conference versions of this work from a number of colleagues, in particular Matthieu Bloch for suggesting the generalization studied in Section VIII, Neri Merhav for pointing out the connections between the combinatorial aspects of our work and constrained coding, Markovian types, and statistical physics, and Or Ordentlich for bringing [11] and [12] to our attention.

REFERENCES

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley, 1991.
[2] A. El Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2012.
[3] T. Berger, *Rate Distortion Theory; A Mathematical Basis for Data Compression*. Englewood, Cliffs, NJ, USA: Prentice-Hall, 1971.
[4] N. Merhav and M. Feder, "Universal prediction," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2124–2147, Oct. 1998.
[5] V. K. Goyal, "Multiple description coding: Compression meets the network," *IEEE Signal Process. Mag.*, vol. 18, no. 5, pp. 74–93, Sep. 2001.
[6] 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, Mar. 2003.
[7] 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, Jun. 2004.
[8] A. Ingber and T. Weissman, "Compression for similarity identification: Fundamental limits," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Honolulu, HI, USA, Jun./Jul. 2014, pp. 1–5.
[9] Y. Bar-Hillel and R. Carnap, "Semantic information," *Brit. J. Philos. Sci.*, vol. 4, no. 14, pp. 147–157, 1953.
[10] R. Carnap and Y. Bar-Hillel, "An outline of a theory of semantic information," Dept. Res. Lab. Electron., Massachusetts Instit. Technol., Cambridge, MA, USA, Tech. Rep. 247, 1952, pp. 221–274.
[11] O. Ordentlich. (2016). "Novel lower bounds on the entropy rate of binary hidden Markov processes." [Online]. Available: https://arxiv.org/abs/1601.06453
[12] E. Ordentlich and T. Weissman, "Bounds on the entropy rate of binary hidden Markov processes," in *Entropy of Hidden Markov Processes and Connections to Dynamical Systems: Papers from the Banff International Research Station Workshop*, B. Marcus, K. Petersen, and T. Weissman, Eds. New York, NY, USA: Cambridge Univ. Press, 2011.
[13] I. Csiszár, "The method of types," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2505–2523, Oct. 1998.
[14] C. Bunte and A. Lapidoth, "Encoding tasks and Rényi entropy," *IEEE Trans. Inf. Theory*, vol. 60, no. 9, pp. 5065–5076, Sep. 2014.
[15] A. Høst-Madsen, E. Sabeti, and C. Walton, "Information theory for atypical sequences," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Honolulu, HI, USA, Sep. 2013, pp. 1–5.
[16] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. Berlin, Germany: Springer, 2009.
[17] A. Dembo and L. Kontoyiannis, "Source coding, large deviations, and approximate pattern matching," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1590–1615, Jun. 2002.
[18] I. Csiszár, "Sanov property, generalized I-projection, and a conditional limit theorem," *Ann. Probability*, vol. 12, no. 3, pp. 768–793, 1984.
[19] J. M. van Campenhout and T. M. Cover, "Maximum entropy and conditional probability," *IEEE Trans. Inf. Theory*, vol. IT-27, no. 4, pp. 483–489, Jul. 1981.
[20] A. Dembo and O. Zeitouni, "Refinements of the Gibbs conditioning principl," *Probability Theory Related Fields*, vol. 104, no. 1, pp. 1–14, 1996.
[21] O. A. Vasicek, "A conditional law of large numbers," *Ann. Probability*, vol. 8, no. 1, pp. 142–147, 1980.
[22] E. Zehavi and J. K. Wolf, "On runlength codes," *IEEE Trans. Inf. Theory*, vol. IT-34, no. 1, pp. 45–54, Jan. 1988.
[23] B. Marcus, P. Siegel, and R. Roth, "An introduction to coding for constrained systems," in *Handbook Coding Theory*, W. C. Huffman and V. Pless, Eds. New York, NY, USA: Elsevier, 1998.
[24] K. A. Immink, *Codes for Mass Data Storage Systems*. Eindhoven, The Netherlands: Shannon Foundation, 2004.
[25] L. Davisson, G. Longo, and A. Sgarro, "The error exponent for the noiseless encoding of finite Ergodic Markov sources," *IEEE Trans. Inf. Theory*, vol. IT-27, no. 4, pp. 431–438, Jul. 1981.
[26] P. Jacquet and W. Szpankowski, "Markov types and minimax redundancy for Markov sources," *IEEE Trans. Inf. Theory*, vol. 50, no. 7, pp. 1393–1402, Jul. 2004.
[27] N. Merhav, "Statistical physics and information theory," *Found. Trends Commun. Inf. Theory*, vol. 6, nos. 1–2, pp. 1–212, 2010.
[28] T.-S. Han, *Information-Spectrum Methods in Information Theory*. Berlin, Germany: Springer-Verlag, 2003.
[29] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York, NY, USA: Academic, 1981.
[30] S. Borade, B. Nakiboğlu, and L. Zheng, "Unequal error protection: An information-theoretic perspective," *IEEE Trans. Inf. Theory*, vol. 55, no. 12, pp. 5511–5539, Dec. 2009.
[31] B. Nazer, Y. Y. Shkel, and S. C. Draper, "The AWGN red alert problem," *IEEE Trans. Inf. Theory*, vol. 59, no. 4, pp. 2188–2200, Apr. 2013.
[32] R. M. Gray, *Entropy and Information Theory*. New York, NY, USA: Springer, 2011.

[33] K. M. Mackenthun and M. B. Pursley, "Strongly and weakly universal source coding," in *Proc. Conf. Inf. Sci. Syst. (CISS)*. Baltimore, MD, USA: John Hopkins Univ., 1977, pp. 286–291.
[34] T. S. Han and S. Verdú, "Approximation theory of output statistics," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 752–772, May 1993.
[35] Y. Steinberg and S. Verdú, "Simulation of random processes and rate-distortion theory," *IEEE Trans. Inf. Theory*, vol. 42, no. 1, pp. 63–86, Jan. 1996.
[36] C. Schroeder, "I-projection and conditional limit theorems for discrete parameter Markov processes," *Ann. Probability*, vol. 21, no. 2, pp. 721–758, 1993.
[37] A. Meda and P. Ney, "The Gibbs conditioning principle for Markov chains," in *Perplexing problems probability. Prog. Probability*, vol. 44, M. Bramson and R. Durrett, Eds. Boston, MA, USA: Birkhäuser, 1999, pp. 385–398.
[38] A. Meda, "Conditional weak laws in Banach spaces," *Amer. Math. Soc.*, vol. 131, no. 8, pp. 2597–2609, 2003.
[39] K. Marton, "Error exponent for source coding with a fidelity criterion," *IEEE Trans. Inf. Theory*, vol. IT-20, no. 2, pp. 197–199, Mar. 1974.
[40] I. Csiszár, T. Cover, and B.-S. Choi, "Conditional limit theorems under Markov conditioning," *IEEE Trans. Inf. Theory*, vol. IT-33, no. 6, pp. 788–801, Nov. 1987.

Ebrahim MolavianJazi (S'08–M'14) received a dual-degree B.S. in Electrical Engineering and Applied Mathematics from Isfahan University of Technology, Isfahan, Iran in 2007, and a M.S. and a Ph.D. in Electrical Engineering from the Wireless Instintute, Department of Electrical Engineering, University of Notre Dame, IN, USA in 2010 and 2014, respectively. He was a postdoctocal scholar with the Wireless Communications and Networking Laboratory (WCAN), Department of Electrical Engineering, The Pennsylvania State University, University Park, PA, USA, from 2014 to 2016. He has been a staff researcher with the Wireless Standards and Research team at Motorola Mobility LLC, Chicago, IL, USA, since 2016. He is currently focused as a company delegate on 3GPP 5G New Radio (NR) standards, and his research interests are broadly in information theory and wireless communications.

Aylin Yener (S'91–M'01–SM'14–F'15) received the B.Sc. degree in electrical and electronics engineering and the B.Sc. degree in physics from Bogazici University, Istanbul, Turkey, and the M.S. and Ph.D. degrees in electrical and computer engineering from the Wireless Information Network Laboratory (WINLAB), Rutgers University, New Brunswick, NJ, USA. She is a Professor of Electrical Engineering at The Pennsylvania State University, University Park, PA, USA, since 2010, where she joined the faculty as an assistant professor in 2002, and was an associate professor 2006-2010. Since 2017, she is a Dean's Fellow in the College of Engineering at The Pennsylvania State University. She was a visiting professor of Electrical Engineering at Stanford University in 2016-2018 and a visiting associate professor in the same department in 2008-2009. Her current research interests are in caching systems, information security, green communications, and more generally in the fields of communication theory, information theory and network science. She received the NSF CAREER Award in 2003, the Best Paper Award in Communication Theory from the IEEE International Conference on Communications in 2010, the Penn State Engineering Alumni Society (PSEAS) Outstanding Research Award in 2010, the IEEE Marconi Prize Paper Award in 2014, the PSEAS Premier Research Award in 2014, and the Leonard A. Doggett Award for Outstanding Writing in Electrical Engineering at Penn State in 2014. She is a distinguished lecturer for the IEEE Communications Society (2018-2020) and the IEEE Vehicular Technology Society (2017-2019).

Dr. Yener is a member of the Board of Governors of the IEEE Information Theory Society (2015-2020), where she was previously the Treasurer from 2012 to 2014. She served as the Student Committee Chair for the IEEE Information Theory Society from 2007 to 2011, and was the co-Founder of the Annual School of Information Theory in North America in 2008. She was a Technical (Co)-Chair for various symposia/tracks at the IEEE ICC, PIMRC, VTC, WCNC, and Asilomar in 2005, 2008-2014 and 2018. She served as an Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS from 2009 to 2012, an Editor and an Editorial Advisory Board Member for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2001 to 2012, and a Guest Editor for the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY in 2011, and the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS in 2015. Currently, she serves on the Editorial Board of the IEEE TRANSACTIONS ON MOBILE COMPUTING and as a Senior Editor for the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS.