Rényi Information Dimension: Fundamental Limits of Almost Lossless Analog Compression

Yihong Wu, *Student Member, IEEE*, and Sergio Verdú, *Fellow, IEEE*

***Abstract*—In Shannon theory, lossless source coding deals with the optimal compression of discrete sources. Compressed sensing is a lossless coding strategy for analog sources by means of multiplication by real-valued matrices. In this paper we study almost lossless analog compression for analog memoryless sources in an information-theoretic framework, in which the compressor or decompressor is constrained by various regularity conditions, in particular linearity of the compressor and Lipschitz continuity of the decompressor. The fundamental limit is shown to be the information dimension proposed by Rényi in 1959.**

***Index Terms*—Analog compression, compressed sensing, information measures, Rényi information dimension, Shannon theory, source coding.**

I. Introduction

A. Motivations From Compressed Sensing

THE "Bit" is the universal currency in lossless source coding theory [1], where Shannon entropy is the fundamental limit of compression rate for discrete memoryless sources (DMS). Sources are modeled by stochastic processes and redundancy is exploited as probability is concentrated on a set of exponentially small cardinality as blocklength grows. Therefore, by encoding this subset, data compression is achieved if we tolerate a positive, though arbitrarily small, block error probability.

Compressed sensing [2], [3] has recently emerged as an approach to lossless encoding of analog sources by real numbers rather than bits. It deals with efficient recovery of an unknown real vector from the information provided by linear measurements. The formulation of the problem is reminiscent of the traditional lossless data compression in the following sense.

- Sources are sparse in the sense that each vector is supported on a set much smaller than the blocklength. This kind of redundancy in terms of sparsity is exploited to achieve effective compression by taking fewer number of linear measurements.
- In contrast to lossy data compression, block error probability, instead of distortion, is the performance benchmark.
- The central problem is to determine how many compressed measurements are sufficient/necessary for recovery with vanishing block error probability as blocklength tends to infinity [2]–[4].
- Random coding is employed to show the existence of "good" linear encoders. In particular, when the random projection matrices follow certain distribution (e.g., standard Gaussian), the restricted isometry property (RIP) is satisfied with overwhelming probability and guarantees exact recovery.

On the other hand, there are also significantly different ingredients in compressed sensing in comparison with information theoretic setups.

- Sources are not modeled probabilistically, and the fundamental limits are on a worst case basis rather than on average. Moreover, block error probability is with respect to the distribution of the encoding random matrices.
- Real-valued sparse vectors are encoded by real numbers instead of bits.
- The encoder is confined to be linear while generally in information-theoretical problems such as lossless source coding we have the freedom to choose the best possible coding scheme.

Departing from the compressed sensing literature, we study fundamental limits of lossless source coding for real-valued memoryless sources within an information theoretic setup.

- Sources are modeled by random processes. This method is more flexible to describe source redundancy which encompasses, but is not limited to, sparsity. For example, a mixed discrete-continuous distribution is suitable for characterizing linearly sparse vectors [5], [6], i.e., those with a number of nonzero components proportional to the blocklength with high probability and whose nonzero components are drawn from a given continuous distribution.
- Block error probability is evaluated by averaging with respect to the source.
- While linear compression plays an important role in our development, our treatment encompasses weaker regularity conditions.

Methodologically, the relationship between our approach and compressed sensing is analogous to the relationship between modern coding theory and classical coding theory: classical coding theory adopts a worst case (Hamming) approach whose goal is to obtain codes with a certain *minimum distance*, while modern coding theory adopts a statistical (Shannon) approach whose goal is to obtain codes with small *probability of failure*. Likewise compressed sensing adopts a worst case model in which compressors work provided that the *number of nonzero*

Manuscript received March 02, 2009; revised April 30, 2010. Date of current version July 14, 2010. This work was supported in part by the National Science Foundation under Grants CCF-0635154 and CCF-0728445. The material in this paper was presented in part at the IEEE International Symposium on Information Theory, Seoul, Korea, July 2009 [55].

The authors are with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: yihongwu@princeton.edu; verdu@princeton.edu).

Communicated by H. Yamamoto, Associate Editor for Shannon Theory.

Digital Object Identifier 10.1109/TIT.2010.2050803

components in the source does not exceed a certain threshold, while we adopt a statistical model in which compressors work for *most source realizations*. In this sense, almost lossless analog compression can be viewed as an information theoretic framework for compressed sensing. Probabilistic modeling provides elegant results in terms of fundamental limit, as well as sheds light on constructive schemes on individual sequences. For example, not only is random coding a proof technique in Shannon theory, but also a guiding principle in modern coding theory as well as in compressed sensing.

Recently there have been considerably new developments in using statistical signal models (e.g., mixed distributions) in compressed sensing (e.g., [5]–[8]), where reconstruction performance is evaluated by computing the asymptotic error probability in the large-blocklength limit. As discussed in Section IV-B, the performance of those practical algorithms still lies far from the fundamental limit.

B. Lossless Source Coding for Analog Sources

Discrete sources have been the sole object in lossless data compression theory. The reason is at least twofold. First, nondiscrete sources have infinite entropy, which implies that representation with arbitrarily small block error probability requires arbitrarily large rate. On the other hand, even if we consider encoding analog sources by real numbers, the result is still trivial, as $\mathbb{R}$ and $\mathbb{R}^n$ have the same cardinality. Therefore, a single real number is capable of representing a real vector losslessly, yielding a universal compression scheme for any analog source with *zero* rate and *zero* error probability.

However, it is worth pointing out that the compression method proposed above is *not* robust because the bijection between $\mathbb{R}$ and $\mathbb{R}^n$ is highly irregular. In fact, neither the encoder nor the decoder can be *continuous* [9, Exercise 6(c), p. 385]. Therefore, such a compression scheme is useless in the presence of any observation noise regardless of the signal-to-noise ratio (SNR). This disadvantage motivates us to study how to compress not only *losslessly* but also *gracefully* in the real field. In fact some authors have also noticed the importance of regularity in data compression. In [10] Montanari and Mossel observed that the optimal data compression scheme often exhibits the following inconvenience: codewords tend to depend chaotically on the data; hence, changing a single source symbol leads to a radical change in the codeword. In [10], a source code is said to be *smooth* (resp., *robust*) if the encoder (resp., decoder) is Lipschitz (see Definition 6) with respect to the Hamming distance. The fundamental limits of smooth lossless compression are analyzed in [10] for binary sources via sparse graph codes. In this paper, we focus on sources in the real field with general distributions. Introducing a *topological structure* makes the nature of the problem quite different from traditional formulations in the discrete world, and calls for machinery from dimension theory and geometric measure theory.

C. Operational Characterization of Rényi Information Dimension

In 1959, Alfréd Rényi proposed an information measure for random vectors in Euclidean space named information dimension [11], through the normalized entropy of a finely quantized version of the random vector. It characterizes the rate of growth of the information given by successively finer discretizations of the space. Although a fundamental information measure, the Rényi dimension is far less well known than either the Shannon entropy or the Rényi entropy. Rényi showed that under certain conditions for an absolutely continuous n-dimensional random vector the information dimension is n. Hence, he remarked in [11] that "*the geometrical (or topological) and information-theoretical concepts of dimension coincide for absolutely continuous probability distributions*." However, the operational role of Rényi information dimension has not been addressed before except in the work of Kawabata and Dembo [12], which relates it to the rate-distortion function. It is shown in [12] that when the single-letter distortion function satisfies certain conditions, the rate-distortion function $R(D)$ of a real-valued source scales proportionally to $\log \frac{1}{D}$ as $D \to 0$, with the proportionality constant being the information dimension of the source. This result serves to drop the assumption of continuity in the asymptotic tightness of Shannon's lower bound in the low distortion regime.

In this paper we give an operational characterization of Rényi information dimension as the fundamental limit of almost lossless data compression for analog sources under various regularity constraints of the encoder/decoder. Moreover, we consider the problem of lossless Minkowski dimension compression, where the Minkowski dimension of a set measures its degree of fractality. In this setup we study the minimum upper Minkowski dimension of high-probability events of source realizations. This can be seen as a counterpart of lossless source coding, which seeks the smallest cardinality of high-probability events. Rényi information dimension turns out to be the fundamental limit for lossless Minkowski dimension compression.

D. Organization of the Paper

Notations frequently used throughout the paper are summarized in Section II. Section III gives an overview of Rényi information dimension, a new interpretation in terms of entropy rate and discusses connections with rate-distortion theory. Section IV states the main definitions and results, as well as their connections with compressed sensing. Section V contains definitions and coding theorems of lossless Minkowski dimension compression, which are important intermediate results for Sections VI and VII. New type of concentration-of-measure type of results are proved for memoryless sources, where it is shown that overwhelmingly large probability is concentrated on subsets of low (Minkowski) dimension. Section VI tackles the case of lossless linear compression, where achievability results are given as well as a converse for mixed discrete-continuous sources. Section VII is devoted to lossless Lipschitz decompression, where we establish a general converse in terms of upper information dimension, and its tightness for mixed discrete-continuous and self-similar sources. Some technical lemmas are proved in Appendixes I–X.

II. Notations

The major notations adopted in this paper are summarized as follows.

- $\mathbb{Z}_m = \{0, \ldots, m-1\}$, for $m \in \mathbb{N}$.

- $X^n = [X_1, \ldots, X_n]$ denotes a random vector. $x^n = [x_1, \ldots, x_n]$ denotes a realization of X^n.
- $\langle \cdot \rangle_m$ denotes the quantization operator, which can be applied to real numbers, vectors or subsets of $\mathbb{R}^n$ as follows:

$$\langle x \rangle_m = \frac{\lfloor mx \rfloor}{m} \tag{1}$$

$$\langle x^n \rangle_m = [\langle x_1 \rangle_m, \ldots, \langle x_n \rangle_m] \tag{2}$$

$$\langle S \rangle_m = \{\langle x^n \rangle_m : x^n \in S\}, \qquad S \subset \mathbb{R}^n. \tag{3}$$

- $[\cdot]_m = \langle \cdot \rangle_{2^m}$.
- For $z^n \in \mathbb{Z}^n$

$$C_m(z^n) = \left\{ x^n \in \mathbb{R}^n : [x^n]_m = \frac{z^n}{2^m} \right\} \tag{4}$$

$$= \prod_{i=1}^{n} \left[\frac{z_i}{2^m}, \frac{z_i + 1}{2^m} \right). \tag{5}$$

  Then, $\{C_m(z^n) : z^n \in \mathbb{Z}^n\}$ is a partition of $\mathbb{R}^n$ with called mesh cubes of size 2^{-m}.
- $(x)_i$ denotes the ith bit in the binary expansion of $0 \le x < 1$, that is

$$(x)_i = 2^i([x]_i - [x]_{i-1}) = \lfloor 2^i x \rfloor - 2\lfloor 2^{i-1} x \rfloor \in \mathbb{Z}_2. \tag{6}$$

  Then

$$x = \sum_{i=1}^{\infty} (x)_i 2^{-i} \tag{7}$$

$$[x]_i = \sum_{j=1}^{i} (x)_j 2^{-j}. \tag{8}$$

  Similarly, $(x^n)_i \in \mathbb{Z}_2^n$ is defined componentwise.
- Let (X, d) be a metric space. Denote the closed ball of radius r centered at x by $B(x, r) = \{z \in X : d(z, x) \le r\}$. In particular, in $(\mathbb{R}^m, \|\cdot\|_p)$ $(p \in [1, \infty])$, denote the ℓ_p ball of radius r centered at z by

$$B_p^m(z, r) = \{u \in \mathbb{R}^m : \|u - z\|_p \le r\} \tag{9}$$

  where the ℓ_p-norm on $\mathbb{R}^m$ is defined as

$$|x^m|_p = \begin{cases} \left(\sum_{i=1}^m |x_i|^p\right)^{\frac{1}{p}}, & p \in [1, \infty). \\ \max\{|x_1|, \ldots, |x_m|\}, & p = \infty. \end{cases} \tag{10}$$

- Define $\|x^m\|_0 = \sum_{i=1}^m \mathbf{1}_{\{|x_i| > 0\}}$, which is not a norm since $\|\lambda x^m\|_0 \ne |\lambda| \|x^m\|_0$ for $\lambda \ne 0$ or ± 1. However, $\|x - y\|_0$ is a valid metric on $\mathbb{R}^n$.
- Let $T \subset \{1, \ldots, n\}$. For a $k \times n$ matrix $\mathbf{A}$, denote by $\mathbf{A}_T$ the $k \times |T|$ submatrix formed by those columns of $\mathbf{A}$ whose indices are in T.
- All logarithms in this paper are with respect to base 2.

III. Rényi Information Dimension

In this section, we give an overview of Rényi information dimension and its properties. Moreover, we give a novel interpretation in terms of the entropy rate of the dyadic expansion of the random variable. We also discuss the connection between information dimension and rate-distortion theory established in [12].

A. Definitions

Definition 1 (Information Dimension [11]): Let X be an arbitrary real-valued random variable. Denote for a positive integer m

$$\langle X \rangle_m = \frac{\lfloor mX \rfloor}{m}. \tag{11}$$

Define

$$\underline{d}(X) = \liminf_{m \to \infty} \frac{H(\langle X \rangle_m)}{\log m} \tag{12}$$

and

$$\overline{d}(X) = \limsup_{m \to \infty} \frac{H(\langle X \rangle_m)}{\log m} \tag{13}$$

where $\underline{d}(X)$ and $\overline{d}(X)$ are called lower and upper information dimensions of X, respectively. If $\underline{d}(X) = \overline{d}(X)$, the common value is called the information dimension of X, denoted by $d(X)$, i.e.,

$$d(X) = \lim_{m \to \infty} \frac{H(\langle X \rangle_m)}{\log m}. \tag{14}$$

Rényi also defined the "entropy of dimension $d(X)$" as

$$\hat{H}(X) = \lim_{m \to \infty} (H(\langle X \rangle_m) - d(X) \log m) \tag{15}$$

provided the limit exists.

Definition 1 can be readily extended to random vectors, where the floor function $\lfloor \cdot \rfloor$ is taken componentwise. Since $d(X)$ only depends on the distribution of X, we also denote $d(P_X) = d(X)$. Similar convention also applies to entropy and other information measures.

Apart from discretization, information dimension can be defined from a more general viewpoint: the mesh cubes of size $\epsilon > 0$ in $\mathbb{R}^k$ are the sets $C_{z,\epsilon} = \prod_{j=1}^k [z_j \epsilon, (z_j + 1)\epsilon)$ for $z^k \in \mathbb{Z}^k$. For any $\epsilon > 0$, the collection $\{C_{z,\epsilon} : z \in \mathbb{Z}^k\}$ partitions $\mathbb{R}^k$. Hence, for any probability measure μ on $\mathbb{R}^k$, this partition generates a discrete probability measure μ_ϵ on $\mathbb{Z}^k$ by assigning $\mu_\epsilon(\{z\}) = \mu(C_{z,\epsilon})$. Then, the information dimension of μ can be expressed as

$$d(X) = \lim_{\epsilon \downarrow 0} \frac{H(\mu_\epsilon)}{\log \frac{1}{\epsilon}}. \tag{16}$$

It should be noted that there exist alternative definitions of information dimension in the literature. For example, in [13], the lower and upper information dimensions are defined by replacing $H(\mu_\epsilon)$ with the ϵ-entropy $H_\epsilon(\mu)$ with respect to the ℓ_∞ distance. This definition essentially allows unequal partition of the whole space and lowers the value of information dimension, since $H_\epsilon(\mu) \le H(\mu_\epsilon)$. However, the resulting definition is equivalent (see Theorem 23). As an another example, the following definition is adopted in [14, Def. 4.2]:

$$d(X) = \lim_{\epsilon \downarrow 0} \frac{\mathbb{E} \log \mu(B_2(X, \epsilon))}{\log \epsilon} \tag{17}$$

where μ denotes the distribution of X and $B_2(x, \epsilon)$ is the ℓ_2-ball of radius ϵ centered at x. This definition is equivalent to Defini-

tion 1, as shown in Appendix I. Note that (17) can be generalized to random variables on an arbitrary metric space.

B. Characterizations and Properties

The lower and upper information dimension of a random variable might not always be finite, because $H(\langle X\rangle_n)$ can be infinity for all n. However, as pointed out in [11], if the mild condition $H(\lfloor X\rfloor) < \infty$ is satisfied, we have

$$0 \le \underline{d}(X) \le \bar{d}(X) \le 1. \tag{18}$$

The necessity of this condition is shown in Proposition 1. One sufficient condition for finite information dimension is $\mathbb{E}[\log(1+|X|)] < \infty$. Consequently, if $\mathbb{E}[|X|^\epsilon] < \infty$ for some $\epsilon > 0$, then $\bar{d}(X) < \infty$.

Proposition 1:

$$\mathbb{E}[\log(1+|X|)] < \infty. \tag{19}$$
$$\Rightarrow \quad 0 \le \underline{d}(X) \le \bar{d}(X) \le 1. \tag{20}$$
$$\Leftrightarrow \quad H(\lfloor X\rfloor) < \infty. \tag{21}$$

If $H(\lfloor X\rfloor) = \infty$, then

$$\underline{d}(X) = \bar{d}(X) = \infty. \tag{22}$$

Proof: See Appendix II. □

For $\mathbb{R}^n$-valued X, (20) can be generalized to $0 \le \underline{d}(X) \le \bar{d}(X) \le n$.

To calculate the information dimension in (12) and (13), it is sufficient to restrict to the exponential subsequence $m = 2^l$, as a result of the following proposition.

Proposition 2:

$$\underline{d}(X) = \liminf_{l\to\infty} \frac{H([X]_l)}{l} \tag{23}$$
$$\bar{d}(X) = \limsup_{l\to\infty} \frac{H([X]_l)}{l}. \tag{24}$$

Proof: See Appendix II. □

Similarly to the approach in Proposition 1, we have the following.

Proposition 3: $\underline{d}(X)$ and $\bar{d}(X)$ are unchanged if rounding or ceiling functions are used in Definition 1.

Proof: See Appendix II. □

C. Evaluation of Information Dimension

By the Lebesgue decomposition theorem [15], a probability distribution can be uniquely represented as the mixture

$$\nu = p\nu_{\mathrm{d}} + q\nu_{\mathrm{c}} + r\nu_{\mathrm{s}} \tag{25}$$

where $p+q+r=1$, $p,q,r \ge 0$; ν_{d} is a purely atomic probability measure (discrete part); ν_{c} is a probability measure absolutely continuous with respect to Lebesgue measure, i.e., having a probability density function (continuous part[1]); and ν_{s} is a probability measure singular with respect to Lebesgue measure but with no atoms (singular part).

As shown in [11], the information dimension for the mixture of discrete and absolutely continuous distribution can be determined as follows.

Theorem 1 [11]: Let X be a random variable such that $H(\lfloor X\rfloor)$ is finite. Assume the distribution of X can be represented as

$$\nu = (1-\rho)\nu_{\mathrm{d}} + \rho\nu_{\mathrm{c}} \tag{26}$$

where ν_{d} is a discrete measure, ν_{c} is an absolutely continuous measure, and $0 \le \rho \le 1$. Then

$$d(X) = \rho. \tag{27}$$

Furthermore, given the finiteness of $H(\nu_{\mathrm{d}})$ and $h(\nu_{\mathrm{c}})$, $\hat{H}(X)$ admits a simple formula

$$\hat{H}(X) = (1-\rho)H(\nu_{\mathrm{d}}) + \rho h(\nu_{\mathrm{c}}) + h_2(\rho) \tag{28}$$

where $H(\nu_{\mathrm{d}})$ is the Shannon entropy of ν_{d}, $h(\nu_{\mathrm{c}})$ is the differential entropy of ν_{c}, and $h_2(\rho) = \rho\log\frac{1}{\rho} + (1-\rho)\log\frac{1}{1-\rho}$ is the binary entropy function.

Proof: See [11, Th. 1 and 3] or [16, Th. 1, pp. 588–592]. □

Some consequences of Theorem 1 are as follows. As long as $H(\lfloor X\rfloor) < \infty$:

1) X is discrete: $d(X) = 0$, and $\hat{H}(X)$ coincides with the Shannon entropy of X.
2) X is continuous: $d(X) = 1$, and $\hat{H}(X)$ is equal to the differential entropy of X.
3) X is discrete-continuous-mixed: $d(X) = \rho$, and $\hat{H}(X)$ is the weighted sum of the entropy of discrete and continuous parts plus a term of $h_2(\rho)$.

For mixtures of countably many distributions, we have the following theorem.

Theorem 2: Let Y be a discrete random variable with $H(Y) < \infty$. If $d(P_{X|Y=i})$ exists for all i, then $d(X)$ exists and is given by $d(X) = \sum_{i=1}^{\infty} P_Y(i) d(P_{X|Y=i})$. More generally

$$\bar{d}(X) \le \sum_{i=1}^{\infty} P_Y(i)\bar{d}(P_{X|Y=i}) \tag{29}$$
$$\underline{d}(X) \ge \sum_{i=1}^{\infty} P_Y(i)\underline{d}(P_{X|Y=i}). \tag{30}$$

Proof: For any m, the conditional distribution of $\langle X\rangle_m$ given $Y=i$ is the same as $\langle X^{(i)}\rangle_m$. Then

$$H(\langle X\rangle_m|Y) \le H(\langle X\rangle_m) \le H(\langle X\rangle_m|Y) + H(Y) \tag{31}$$

where

$$H(\langle X\rangle_m|Y) = \sum_{i=1}^{\infty} P_Y(i) H(\langle X\rangle_m|Y=i). \tag{32}$$

Since $H(Y) < \infty$, dividing both sides of (31) by $\log m$ and sending $m \to \infty$ yields (29) and (30). □

[1] In measure theory, sometimes a measure is called continuous if it does not have any atoms, and a singular measure is called singularly continuous. Here we say a measure is continuous if and only if it is absolutely continuous.

To summarize, when X has a discrete-continuous-mixed distribution, the information dimension of X is given by the weight of the continuous part. When the distribution of X has a singular component, its information dimension does not admit a simple formula in general. For instance, it is possible that $\underline{d}(X) < \bar{d}(X)$ [11]. However, for the important class of *self-similar* singular distributions, the information dimension can be explicitly determined. See Section III-E.

D. Interpretation of Rényi Dimension as Entropy Rate

Let $X \in [0,1]$ a.s. Observe that $H(\langle X\rangle_m) \le \log m$, since the range of $\langle X\rangle_m$ contains at most m values. Then, $0 \le \underline{d}(X) \le \bar{d}(X) \le 1$. The dyadic expansion of X can be written as

$$X(\omega) = \sum_{j=1}^{\infty} (X)_j(\omega) 2^{-j} \tag{33}$$

where each $(X)_j$ is a binary random variable. Therefore, there is a one-to-one correspondence between X and the binary random process $\{(X)_j : j \in \mathbb{N}\}$. Note that the partial sum in (33) is

$$[X]_i(\omega) = \sum_{j=1}^{i} (X)_j(\omega) 2^{-j} \tag{34}$$

and $[X]_i$ and $((X)_1, \ldots, (X)_i)$ are in one-to-one correspondence, therefore

$$H([X]_i) = H((X)_1, \ldots, (X)_i). \tag{35}$$

By Proposition 2, we have

$$\underline{d}(X) = \liminf_{i\to\infty} \frac{H((X)_1, \ldots, (X)_i)}{i} \tag{36}$$

$$\bar{d}(X) = \limsup_{i\to\infty} \frac{H((X)_1, \ldots, (X)_i)}{i}. \tag{37}$$

Thus, its information dimension is the *entropy rate* of its dyadic expansion, or the entropy rate of any M-ary expansion of X, divided by $\log M$.

This interpretation of information dimension enables us to gain more intuition about the result in Theorem 1. When X has a discrete distribution, its dyadic expansion has zero entropy rate. When X is uniform on $[0,1]$, its dyadic expansion is independent identically distributed (i.i.d.) equiprobable, and therefore it has unit entropy rate in bits. If X is continuous, but nonuniform, its dyadic expansion still has unit entropy rate. Moreover, from (36), (37), and Theorem 1, we have

$$\lim_{i\to\infty} D((X)_1, \ldots, (X)_i \| \text{equiprobable}) = -h(X) \tag{38}$$

where $D(\cdot\|\cdot)$ denotes the relative entropy and the differential entropy is $h(X) \le 0$ since $X \in [0,1]$ a.s. The information dimension of a discrete-continuous mixture is also easily understood from this point of view, because the entropy rate of a mixed process is the weighted sum of entropy rates of each component. Moreover, random variables whose lower and upper information dimensions differ can be easily constructed from processes with different lower and upper entropy rates.

E. Self-Similar Distribution

An *iterated function system* (IFS) is a family of contractions $\{F_1, \ldots, F_m\}$ on $\mathbb{R}^n$, where $2 \le m < \infty$, and $F_j : \mathbb{R}^n \to \mathbb{R}^n$ satisfies $\|F_j(x) - F_j(y)\|_2 \le r_j \|x - y\|_2$ for all $x, y \in \mathbb{R}^n$ with $0 < r_j < 1$. By [17, Theorem 2.6], given an IFS, there is a unique nonempty compact set E, called the *invariant set* of the IFS, such that $E = \bigcup_{j=1}^m F_j(E)$. We say that the IFS satisfies the *strong separation condition*, if $\{F_j(E) : j = 1, \ldots, m\}$ are disjoint. The corresponding invariant set is called *self-similar*, if the IFS consists of *similarity transformations*, that is, $F_j(x) = r_j \boldsymbol{O}_j x + t_j$ with $\boldsymbol{O}_j$ an orthogonal matrix and $t_j \in \mathbb{R}^n$, in which case

$$\frac{\|F_j(x) - F_j(y)\|_2}{\|x - y\|_2} = r_j \tag{39}$$

is called the *similarity ratio* of F_j. Self-similar sets are usually fractal. For example, consider the IFS on $\mathbb{R}$ with

$$F_1(x) = x/3, \quad F_2(x) = (x+2)/3. \tag{40}$$

The resulting invariant set is the middle-third Cantor set.

Now we define measures supported on a self-similar set E associated with the IFS $\{F_1, \ldots, F_m\}$. A continuous mapping π from the space $\{1, \ldots, m\}^{\mathbb{N}}$ (equipped with the product topology) onto E is defined as follows:

$$\pi(\{b_1, b_2, \ldots\}) = \bigcap_{j=1}^{\infty} F_{b_j} \circ F_{b_{j-1}} \circ \cdots \circ F_{b_1}(E), \tag{41}$$

where the right-hand side is a singleton [12]. Therefore, every measure μ^* on $\{1, \ldots, m\}^{\mathbb{N}}$ induces a measure μ on E as the image measure of μ^* under π, that is, $\mu(A) = \mu^*(\pi^{-1}(A))$. If μ^* is stationary and ergodic, μ is called a *self-similar* measure. In the special case when μ^* corresponds to a memoryless process with common distribution $P = \{p_1, \ldots, p_m\}$, μ satisfies [17, Th. 2.8]

$$\mu(A) = \sum_{j=1}^{m} p_j \mu(F_j^{-1}(A)) \tag{42}$$

and $\mu(F_j(E)) = p_j$ for each j. The usual Cantor distribution [15] can be defined through the IFS in (40) and $P = \{1/2, 1/2\}$.

The next result gives the information dimension of a self-similar measure μ with IFS satisfying the *open set condition*[2] [18, p. 129], that is, there exists a nonempty bounded open set $U \subset \mathbb{R}^n$, such that $\bigcup_j F_j(U) \subset U$ and $F_i(U) \cap F_j(U) = \emptyset$ for $i \ne j$.

Theorem 3 [17], [12]: Let the distribution of X be a self-similar measure generated from the stationary ergodic measure μ^* on $\{1, \ldots, m\}^{\mathbb{N}}$ and the IFS $\{F_1, \ldots, F_m\}$ with similarity ratios $r_1, \ldots, r_m$ and invariant set E. Then

$$d(X) = \frac{H(\mu^*)}{\sum_{j=1}^m \mu[F_j(E)] \log \frac{1}{r_j}}. \tag{43}$$

When μ^* is the distribution of a memoryless process with common distribution $P = \{p_1, \ldots, p_m\}$, (43) is reduced to

$$d(X) = \frac{H(P)}{\sum_{j=1}^m p_j \log \frac{1}{r_j}}. \tag{44}$$

[2]The open set condition is weaker than the previous strong separation condition.

Note that the open set condition implies that $\sum_{j=1}^m \mathrm{Leb}(F_j(U)) = \mathrm{Leb}(U)\sum_{j=1}^m r_j^n \le \mathrm{Leb}(U)$. Since $0 < \mathrm{Leb}(U) < \infty$, it follows that

$$\sum_{j=1}^m r_j^n \le 1. \tag{45}$$

In view of (44), we have

$$d(X) = \frac{nH(P)}{H(P) + D(P\,\|\,R)} \tag{46}$$

where $R = (r_1^n, \ldots, r_m^n)$ is a subprobability measure. Since $D(P\,\|\,R) \ge 0$, we have $0 \le d(X) \le n$, which agrees with Proposition 1.

F. Connections With Rate-Distortion Theory

The asymptotic behavior of the rate-distortion function, in particular, the asymptotic tightness of the Shannon lower bound in the high-rate regime, has been addressed in [19] and [20] for continuous sources. In [12], Kawabata and Dembo generalized it to real-valued sources that do not necessarily possess a density, and showed that the information dimension plays a central role. For completeness, we summarize the main results from [12] in Appendix III.

G. Rényi Dimension of Order α

With Shannon entropy replaced by Rényi entropy in (12)–(13), the generalized notion of dimension of order α is defined similarly.

Definition 2 (Information Dimension of Order α): Let $\alpha \in [0, \infty]$. Define

$$\underline{d}_\alpha(X) = \liminf_{m\to\infty} \frac{H_\alpha(\langle X\rangle_m)}{\log m} \tag{47}$$

and

$$\bar{d}_\alpha(X) = \limsup_{m\to\infty} \frac{H_\alpha(\langle X\rangle_m)}{\log m} \tag{48}$$

where $H_\alpha(Y)$ denotes the Rényi entropy of order α of a discrete random variable Y with probability mass function $\{p_y : y \in \mathcal{Y}\}$, defined as

$$H_\alpha(Y) = \begin{cases} \log\left(\dfrac{1}{\max_{y\in\mathcal{Y}} p_y}\right), & \alpha = \infty \\ \dfrac{1}{1-\alpha}\log\left(\sum_{y\in\mathcal{Y}} p_y^\alpha\right), & \alpha \ne 1, \infty. \end{cases} \tag{49}$$

$\underline{d}_\alpha(X)$ and $\bar{d}_\alpha(X)$ are called lower and upper dimensions of X of order α, respectively. If $\underline{d}_\alpha(X) = \bar{d}_\alpha(X)$, the common value is called the information dimension of X of order α, denoted by $d_\alpha(X)$. Rényi also defined in [16] "the entropy of X of order α and dimension $d_\alpha(X)$" as

$$\hat{H}_\alpha(X) = \lim_{m\to\infty} (H_\alpha(\langle X\rangle_m) - d_\alpha(X)\log m) \tag{50}$$

provided the limit exists.

As a consequence of the monotonicity of Rényi entropy, information dimensions of different orders satisfy the following result.

Lemma 1: For $\alpha \in [0, \infty]$, $\underline{d}_\alpha$ and $\bar{d}_\alpha$ both decrease with α. Define

$$\hat{d}(X) = \lim_{\alpha\uparrow 1} \bar{d}_\alpha(X). \tag{51}$$

Then

$$0 \le \underline{d}_\infty \le \lim_{\alpha\downarrow 1} \underline{d}_\alpha \le \underline{d} \le \bar{d} \le \hat{d} \le \bar{d}_0. \tag{52}$$

Proof: All inequalities follow from the fact that for a fixed random variable Y, $H_\alpha(Y)$ decreases with α in $[0, \infty]$. □

For dimension of order α, we highlight the following result from [21].

Theorem 4 [21, Th. 3]: Let X be a random variable whose distribution has Lebesgue decomposition as in (25). Then, we have the following.

1) $\alpha > 1$: if $p > 0$, that is, X has a discrete component, we have $d_\alpha(X) = 0$.
2) $\alpha < 1$: if $q > 0$, that is, X has a continuous component, and $H_\alpha(\lfloor X\rfloor) = h_\alpha(\nu_{\mathrm{c}}) < \infty$, we have $d_\alpha(X) = 1$. The differential Rényi entropy $h_\alpha(\nu_{\mathrm{c}})$ is defined using its density f_{c} as

$$h_\alpha(\nu_{\mathrm{c}}) = \frac{1}{1-\alpha}\log\int f^\alpha(x)\mathrm{d}x. \tag{53}$$

In general, $d_\alpha(X)$ is discontinuous in α. For discrete-continuous-mixed distributions, $d_\alpha(X) = \hat{d}(X) = 1$ for all $\alpha < 1$, while $d(X)$ equals to the weight of the continuous part. However, for Cantor distribution, $d_\alpha(X) = \hat{d}(X) = d(X) = \log_3 2$ for all α.

IV. Definitions and Main Results

This section presents a unified framework for lossless data compression and our main results in the form of coding theorems under various regularity conditions. Proofs are relegated to Sections V–VII.

A. Lossless Data Compression

Let the source $\{X_i : i \in \mathbb{N}\}$ be a stochastic process on $(\mathcal{X}^{\mathbb{N}}, \mathcal{F}^{\otimes\mathbb{N}})$, with $\mathcal{X}$ denoting the source alphabet and $\mathcal{F}$ a σ-algebra over $\mathcal{X}$. Let $(\mathcal{Y}, \mathcal{G})$ be a measurable space, where $\mathcal{Y}$ is called the *code alphabet*. The main objective of lossless data compression is to find efficient representations for source realizations $x^n \in \mathcal{X}^n$ by $y^k \in \mathcal{Y}^k$.

Definition 3: A (n, k)-code for $\{X_i : i \in \mathbb{N}\}$ over the code space $(\mathcal{Y}, \mathcal{G})$ is a pair of mappings:

1) encoder: $f_n : \mathcal{X}^n \to \mathcal{Y}^k$ that is measurable relative to $\mathcal{F}^n$ and $\mathcal{G}^k$;
2) decoder: $g_n : \mathcal{Y}^k \to \mathcal{X}^n$ that is measurable relative to $\mathcal{G}^k$ and $\mathcal{F}^n$.

The block error probability is $\mathbb{P}\{g_n(f_n(X^n)) \ne X^n\}$.

The fundamental limit in lossless source coding is as follows.

Definition 4 (Lossless Data Compression): Let $\{X_i : i \in \mathbb{N}\}$ be a stochastic process on $(\mathcal{X}^{\mathbb{N}}, \mathcal{F}^{\otimes \mathbb{N}})$. Define $r(\epsilon)$ to be the infimum of $r > 0$ such that there exists a sequence of $(n, \lfloor rn \rfloor)$-codes over the code space $(\mathcal{Y}, \mathcal{G})$, such that

$$\mathbb{P}\{g_n(f_n(X^n)) \neq X^n\} \leq \epsilon \tag{54}$$

for all sufficiently large n.

According to the classical discrete almost-lossless source coding theorem, if $\mathcal{X}$ is countable and $\mathcal{Y}$ is finite, the minimum achievable rate for any i.i.d. process with distribution P is

$$r(\epsilon) = \begin{cases} \frac{\log |\mathcal{X}|}{\log |\mathcal{Y}|}, & \epsilon = 0 \\ \frac{H(P)}{\log |\mathcal{Y}|}, & 0 < \epsilon < 1 \\ 0, & \epsilon = 1. \end{cases} \tag{55}$$

Using codes over an infinite alphabet, any discrete source can be compressed with zero rate and zero block error probability. In other words, if both $\mathcal{X}$ and $\mathcal{Y}$ are countably infinite, then for all $0 \leq \epsilon \leq 1$

$$r(\epsilon) = 0 \tag{56}$$

for any random process.

B. Lossless Analog Compression With Regularity Conditions

In this section, we consider the problem of encoding analog sources with analog symbols, that is, $(\mathcal{X}, \mathcal{F}) = (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ and $(\mathcal{Y}, \mathcal{G}) = (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ or $([0,1], \mathcal{B}_{[0,1]})$ if *bounded* encoders are required, where $\mathcal{B}_{[0,1]}$ denotes the Borel σ-algebra. As in the countably infinite case, zero rate is achievable even for zero block error probability, because the cardinality of $\mathbb{R}^n$ is the same for any n [22]. This conclusion holds even if we require the encoder/decoder to be Borel measurable, because according to Kuratowski's theorem [23, Remark (i), p. 451] every uncountable standard Borel space is isomorphic[3] to $([0,1], \mathcal{B}_{[0,1]})$. Therefore, a single real number has the capability of encoding a real vector, or even a real sequence, with a coding scheme that is both *universal* and *deterministic*.

However, the rich structure of $\mathbb{R}$ equipped with a metric topology (e.g., that induced by Euclidean distance) enables us to probe the problem further. If we seek the fundamental limits of not only *lossless* coding but *"graceful" lossless* coding, the result is not trivial anymore. In this spirit, our various definitions share the basic information-theoretic setup where a random vector is encoded with a function $f_n : \mathbb{R}^n \to \mathbb{R}^{\lfloor Rn \rfloor}$ and decoded with $g_n : \mathbb{R}^{\lfloor Rn \rfloor} \to \mathbb{R}^n$ with $R \leq 1$ such that f_n and g_n satisfy certain regularity conditions and the probability of incorrect reproduction vanishes as $n \to \infty$.

Regularity in encoder and decoder is imposed for the sake of both less complexity and more robustness. For example, although a surjection g from $[0,1]$ to $\mathbb{R}^n$ is capable of lossless encoding, its irregularity requires specifying *uncountably* many real numbers to determine this mapping. Moreover, regularity in encoder/decoder is crucial to guarantee noise resilience of the coding scheme.

[3]Two measurable spaces are isomorphic if there exists a measurable bijection whose inverse is also measurable.

TABLE I
REGULARITY CONDITIONS OF ENCODER/DECODERS AND CORRESPONDING MINIMUM ϵ-ACHIEVABLE RATES

Encoder	Decoder	Minimum ϵ-achievable rate
Linear	Borel	$\mathsf{R}^*(\epsilon)$
Continuous	Continuous	$\mathsf{R}_0(\epsilon)$
Borel	Lipschitz	$\mathsf{R}(\epsilon)$

Definition 5: Let $\{X_i : i \in \mathbb{N}\}$ be a stochastic process on $(\mathbb{R}^{\mathbb{N}}, \mathcal{B}_{\mathbb{R}}^{\otimes \mathbb{N}})$. Define the minimum ϵ-achievable rate to be the infimum of $R > 0$ such that there exists a sequence of $(n, \lfloor Rn \rfloor)$-codes (f_n, g_n), such that

$$\mathbb{P}\{g_n(f_n(X^n)) \neq X^n\} \leq \epsilon \tag{57}$$

for all sufficiently large n, and the encoder f_n and decoder g_n are constrained according to Table I. Except for linear encoding where $\mathcal{Y} = \mathbb{R}$, it is assumed that $\mathcal{Y} = [0,1]$.

In Definition 5, we have used the following definitions.

Definition 6 (Hölder and Lipschitz Continuity): Let (U, d_U) and (V, d_V) be metric spaces. A function $g : U \to V$ is called (L, γ)-*Hölder continuous* if there exists $L, \gamma \geq 0$ such that for any $x, y \in U$

$$d_V(g(x), g(y)) \leq L d_U(x, y)^{\gamma}. \tag{58}$$

g is called L-*Lipschitz* if g is $(L, 1)$-Hölder continuous. g is simply called *Lipschitz* (resp., β-*Hölder continuous*) if g is L-Lipschitz (resp., (L, β)-Hölder continuous) for some $L \geq 0$.

We proceed to give results for each of the minimum ϵ-achievable rates introduced in Definition 5. Motivated by compressed sensing theory, it is interesting to consider the case where the encoder is restricted to be linear.

Theorem 5 (Linear Encoding: General Achievability): Suppose that the source is memoryless. Then

$$\mathsf{R}^*(\epsilon) \leq \hat{d}(X) \tag{59}$$

for all $0 < \epsilon < 1$, where $\hat{d}(X)$ is defined in (51). Moreover, we have the following.

1) For all linear encoders (except possibly those in a set of zero Lebesgue measure on the space of real matrices), block error probability ϵ is achievable.
2) The decoder can be chosen to be β-Hölder continuous for all $0 < \beta < \frac{R - \hat{d}(X)}{R}$, where $R > \hat{d}(X)$ is the compression rate.

Proof: See Section VI-C. □

Theorem 6 (Linear Encoding: Discrete-Continuous Mixture): Suppose that the source is memoryless with a discrete-continuous mixed distribution. Then

$$\mathsf{R}^*(\epsilon) = d(X) \tag{60}$$

for all $0 < \epsilon < 1$.

Proof: See Section VI-C. □

Theorem 7 (Linear Encoding: Achievability for Self-Similar Sources): Suppose that the source is memoryless with a self-similar distribution that satisfies the open set condition. Then

$$\mathsf{R}^*(\epsilon) \le d(X) \tag{61}$$

for all $0 < \epsilon < 1$.

Proof: See Section VI-C. □

In Theorems 5 and 6, it has been shown that block error probability ϵ is achievable for Lebesgue-a.e. linear encoder. Therefore, choosing any random matrix with i.i.d. entries distributed according to some absolutely continuous distribution on $\mathbb{R}$ (e.g., a Gaussian random matrix) satisfies block error probability ϵ almost surely.

Now, we drop the restriction that the encoder is linear, allowing very general encoding rules. Let us first consider the case where both the encoder and decoder are constrained to be continuous. It turns out that zero rate is achievable in this case.

Theorem 8 (Continuous Encoder and Decoder): For general sources

$$\mathsf{R}_0(\epsilon) = 0 \tag{62}$$

for all $0 < \epsilon \le 1$.

Proof: Since $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$ and $([0,1], \mathcal{B}_{[0,1]})$ are Borel isomorphic, there exist Borel measurable $f : \mathbb{R}^n \to [0,1]$ and $g : [0,1] \to \mathbb{R}^n$, such that $g = f^{-1}$. By Lusin's theorem [24, Th. 7.10], there exists a compact set $K \subset \mathbb{R}^n$ such that f restricted on K is continuous and $\mathbb{P}\{X^n \notin K\} < \epsilon$. Since K is compact and f is injective on K, f is a homeomorphism from K to $T \triangleq f(K)$. Hence, $g : T \to \mathbb{R}^n$ is continuous. Since both K and T are closed, by Tietze extension theorem [9], f and g can be extended to continuous $f' : \mathbb{R}^n \to [0,1]$ and $g' : [0,1] \to \mathbb{R}^n$, respectively. Using f' and g' as the new encoder and decoder, the error probability satisfies $\mathbb{P}\{g'(f'(X^n)) \ne X^n\} \le \mathbb{P}\{X^n \notin K\} < \epsilon$. □

Employing similar arguments as in the proof of Theorem 8, we see that imposing additional continuity constraints on the encoder (resp., decoder) has almost no impact on the fundamental limit $\mathsf{R}(\epsilon)$ (resp., $\mathsf{R}^*(\epsilon)$). This is because a continuous encoder (resp., decoder) can be obtained at the price of an arbitrarily small increase of error probability, which can be chosen to vanish as n grows.

Theorems 9–11 deal with Lipschitz decoding in Euclidean spaces.

Theorem 9 (Lipschitz Decoding: General Converse): Suppose that the source is memoryless. If $\bar{d}(X) < \infty$, then

$$\mathsf{R}(\epsilon) \ge \bar{d}(X) \tag{63}$$

for all $0 < \epsilon < 1$.

Proof: See Section VII-B. □

Theorem 10 (Lipschitz Decoding: Achievability for Discrete/Continuous Mixture): Suppose that the source is memoryless with a discrete-continuous mixed distribution. Then

$$\mathsf{R}(\epsilon) = d(X) \tag{64}$$

for all $0 < \epsilon < 1$.

Proof: See Section VII-C. □

For sources with a singular distribution, in general there is no simple answer due to their fractal nature. For an important class of singular measures, namely self-similar measures generated from i.i.d. digits (e.g., generalized Cantor distribution), the information dimension turns out to be the fundamental limit for lossless compression with Lipschitz decoder.

Theorem 11 (Lipschitz Decoding: Achievability for Self-Similar Measures): Suppose that the source is memoryless and bounded, and its M-ary expansion consists of independent identically distributed digits. Then

$$\mathsf{R}(\epsilon) = d(X) \tag{65}$$

for all $0 < \epsilon < 1$. Moreover, if the distribution of each bit is equiprobable on its support, then (65) holds for $\epsilon = 0$

Proof: See Section VII-D. □

Example 1: As an example, we consider the setup in Theorem 11 with $M = 3$ and $P = \{p, 0, q\}$, where $p + q = 1$. The associated invariant set is the middle third Cantor set C [18] and X is supported on C. The distribution of X, denoted by μ, is called the generalized Cantor distribution [25]. In the ternary expansion of X, each digit is independent and takes value 0 and 2 with probability p and q respectively. Then, by Theorem 11, for any $0 < \epsilon < 1$, $\mathsf{R}(\epsilon) = \frac{h_2(p)}{\log 3}$. Furthermore, when $p = 1/2$, μ coincides with the "uniform" distribution on C, i.e., the standard Cantor distribution. Hence, we have a stronger result that $\mathsf{R}(0) = \log_3 2 \approx 0.63$, i.e., exact lossless compression can be achieved with a Lipschitz continuous decompressor at the rate of the information dimension.

Let $S_n = \{x^n : g_n(f_n(x^n)) = x^n\}$. Then, f_n is the inverse of g_n on S_n. Due to the L-Lipschitz continuity of g_n, f_n is an expansive mapping, that is

$$\|f_n(x) - f_n(y)\| \ge \frac{1}{L}\|x - y\|. \tag{66}$$

Note that (66) implies the injectivity of f_n, a necessary condition for decodability. Moreover, not only does f_n assign different codewords to different source symbols, but also it keeps them sufficiently separated proportionally to their distance. Therefore, the encoder f_n respects the metric structure of the source alphabet.

We conclude this section by introducing *stable decoding*, a weaker condition than Lipschitz continuity.

Definition 7 ((L,Δ)-Stable): Let (U, d_U) and (V, d_V) be metric spaces and $T \subset U$. $g : U \to V$ is called (L,Δ)-stable on T if for all $x, y \in T$

$$d_U(x,y) \le \Delta \Rightarrow d_V(g(x), g(y)) \le L\Delta. \tag{67}$$

We say g is Δ-stable if g is $(1,\Delta)$-stable.

A function is L-Lipschitz if and only if it is (L,Δ)-stable for every $\Delta > 0$. We denote by $\bar{\mathsf{R}}(\epsilon,\Delta)$ the minimum ϵ-achievable rate such that there exists a sequence of Borel encoders and Δ-stable decoders that achieve block error probability ϵ. The fundamental limit of stable decoding is given by the following tight result, whose proof is omitted for conciseness.

Theorem 12 (Δ-Stable Decoding): Let the underlying metric be the ℓ_∞ distance. Suppose that the source is memoryless. Then, for all $0 < \epsilon < 1$

$$\limsup_{\Delta \downarrow 0} \bar{\mathsf{R}}(\epsilon, \Delta) = \bar{d}(X) \tag{68}$$

that is, the minimum ϵ-achievable rate such that for all sufficiently small Δ there exists a Δ-stable coding strategy is given by $\bar{d}(X)$.

C. Connections With Compressed Sensing

As an application of Theorem 6, we consider the following source distribution:

$$\mu = (1-d)\delta_0 + d\,\nu \tag{69}$$

where $0 \le d \le 1$, δ_0 is the Dirac measure with atom at 0 and ν is an absolutely continuous distribution. This is the model for linearly sparse signals used in [5] and [6], where a universal[4] iterative thresholding decoding algorithm is proposed. Under certain assumptions on ν, the asymptotic error probability turns out to exhibit a "phase transition" [5], [6]: there is a sparsity-dependent threshold $\bar{R}(d)$ on the measurement rate above which the error probability vanishes and below which the error probability goes to one. This behavior is predicted by Theorem 6, which shows the optimal threshold is d, irrespective of the prior ν. Moreover, the decoding algorithm in the achievability proof of Section VI-C is universal and robust (Hölder continuous), although it has exponential complexity. The threshold $\bar{R}(d)$ is not given in closed form (in [5, eq. (5) and Fig. 1]), but its numerical evaluation shows that it lies far from the optimal threshold except in the nonsparse regime (d close to 1). Moreover, it can be shown that $\bar{R}(d) = \mathrm{o}(d)$ as $d \to 0$. The performance of several other suboptimal factor-graph-based reconstruction algorithms is analyzed in [7]. Practical robust algorithms that approach the fundamental limit of compressed sensing given by Theorem 6 are not yet known.

Robust reconstruction is of great importance in the theory of compressed sensing [26]–[28], since *noise resilience* is an indispensable property for decompressing sparse signals from real-valued measurements. For example, consider the following robustness result.

Theorem 13 [26]: Suppose we wish to recover a vector $x_0 \in \mathbb{R}^n$ from k noisy compressed linear measurements $y = \mathbf{A}x_0 + e$, where $\mathbf{A} \in \mathbb{R}^{k\times n}$, $e, y \in \mathbb{R}^k$ and $||e||_2 \le \epsilon$. Let x^* be a solution of the following ℓ_1-regularization problem:

$$\begin{aligned} \min \quad & ||x||_1 \\ \text{s.t.} \quad & ||\mathbf{A}x - y||_2 \le \epsilon. \end{aligned} \tag{70}$$

Let $||x_0||_0 = S$ satisfy $\delta_{3S} + 3\delta_{4S} \le 2$, where δ_S is the S-restricted isometry constant δ_S of matrix $\mathbf{A}$, defined as the smallest positive number such that

$$(1-\delta_S)||u||_2^2 \le ||\mathbf{A}u||_2^2 \le (1+\delta_S)||u||_2^2 \tag{71}$$

for all $T \subset \{1, \ldots, n\}$ with $|T| \le S$ and for all u in $\mathbb{R}^n$ supported on T. Then

$$||x^* - x_0||_2 \le C(\delta_{4S})\epsilon. \tag{72}$$

By Theorem 13, using (70) as the decoder, the ℓ_2-norm of the decoding error is upper bounded *proportionally* to the ℓ_2-norm of the noise.

In our framework, a stable or Lipschitz continuous coding scheme also implies *robustness* with respect to noise added at the input of the decompressor, which could result from quantization, finite wordlength or other inaccuracies. For example, suppose that the encoder output $y^k = f_n(x^n)$ is quantized by a q-bit uniform quantizer, resulting in $\tilde{y}^k$. With a 2^{-q}-stable coding strategy (f_n, g_n), we can use the following decoder. Denote the following nonempty set:

$$\mathcal{D}(\tilde{y}^k) = \{z^k \in C_n : ||z^k - \tilde{y}^k||_\infty \le 2^{-q}\} \tag{73}$$

where $C_n = \{f_n(x^n) : x^n \in \mathbb{R}^n, g_n(f_n(x^n)) = x^n\}$. Pick any z^k in $\mathcal{D}(\tilde{y}^k)$ and output $\hat{x}^n = g_n(z^k)$. Then, by the stability of g_n, we have

$$||\hat{x}^n - x^n||_\infty \le 2^{-(q-1)} \tag{74}$$

i.e., each component in the decoder output will suffer at most twice the inaccuracy of the decoder input. Similarly, an L-Lipschitz coding scheme with respect to ℓ_∞ distance incurs an error no more than $L2^{-q}$.

V. Lossless Minkowski-Dimension Compression

As a counterpart to lossless data compression, in this section, we investigate the problem of lossless Minkowski dimension[5] compression for general sources, where the minimum ϵ-achievable rate is defined as $\mathsf{R}_\mathsf{B}(\epsilon)$. This is an important intermediate tool for studying fundamental limits of lossless linear encoding and Lipschitz decoding. Bridging the three compression frameworks, in Sections VI-C and VII-B, we prove the following inequality:

$$\mathsf{R}^*(\epsilon) \le \mathsf{R}_\mathsf{B}(\epsilon) \le \mathsf{R}(\epsilon). \tag{75}$$

Hence, studying $\mathsf{R}_\mathsf{B}(\epsilon)$ provides an *achievability* bound for lossless linear encoding and a *converse* bound for Lipschitz decoding. We present bounds for $\mathsf{R}_\mathsf{B}(\epsilon)$ for general sources, as well as tight results for discrete-continuous mixed and self-similar sources.

A. Minkowski Dimension of Sets in Metric Spaces

In fractal geometry, the Minkowski dimension is a way of determining the fractality of a subset in metric spaces.

Definition 8 (Covering Number): Let A be a nonempty bounded subset of the metric space (X, d). For $\epsilon > 0$, define

[4]The decoding algorithm is universal if it requires no knowledge of the prior distribution ν of nonzero entries.

[5]Also known as Minkowski–Bouligand dimension, fractal dimension or box-counting dimension.

$N_A(\epsilon)$, the ϵ-covering number of A, to be the smallest number of ϵ-balls needed to cover A, that is

$$N_A(\epsilon) = \min\left\{k : A \subset \bigcup_{i=1}^{k} B(x_i, \epsilon), x_i \in X\right\}. \tag{76}$$

Definition 9 (Minkowski Dimensions): Let A be a nonempty bounded subset of metric space (X, d). Define the lower and upper Minkowski dimensions of A as

$$\underline{\dim}_{\mathrm{B}} A = \liminf_{\epsilon \to 0} \frac{\log N_A(\epsilon)}{\log \frac{1}{\epsilon}} \tag{77}$$

$$\overline{\dim}_{\mathrm{B}} A = \limsup_{\epsilon \to 0} \frac{\log N_A(\epsilon)}{\log \frac{1}{\epsilon}} \tag{78}$$

respectively. If $\underline{\dim}_{\mathrm{B}} A = \overline{\dim}_{\mathrm{B}} A$, the common value is called the Minkowski dimension of A, denoted by $\dim_{\mathrm{B}} A$.

It should be pointed out that the Minkowski dimension depends on the underlying metric. Nevertheless, equivalent metrics result in the same dimension. A few examples are as follows.

- $\dim_{\mathrm{B}} A = 0$ for any finite set A.
- $\dim_{\mathrm{B}} A = n$ for any bounded set A of *nonempty interior* in Euclidean space $\mathbb{R}^n$.
- Let C be the middle-third Cantor set in the unit interval. Then, $\dim_{\mathrm{B}} C = \log_3 2$ [18, Example 3.3].
- $\dim_{\mathrm{B}}\{\frac{1}{i} : i \in \mathbb{N}\} = \frac{1}{2} > 0$ [18, Example 3.5]. From this example, wee see that Minkowski dimension lacks certain stability properties one would expect of a dimension, since it is often desirable that adding a countable set would have no effect on dimension. This property fails for Minkowski dimension. On the contrary, we observe that Rényi information dimension exhibits stability with respect to adding a discrete component as long as the entropy is finite. However, mixing any distribution with a discrete measure with unbounded support and infinite entropy will necessarily result in infinite information dimension.

The upper Minkowski dimension satisfies the following properties (see [18, p. 48 (iii) and p. 102 (7.9)]), which will be used in the proof of Theorem 14.

Lemma 2: For bounded sets $A_1, \ldots, A_k$

$$\overline{\dim}_{\mathrm{B}} \bigcup_{i=1}^{k} A_i = \max_i \overline{\dim}_{\mathrm{B}} A_i \tag{79}$$

$$\overline{\dim}_{\mathrm{B}} A_1 \times \cdots \times A_k \le \sum_i \overline{\dim}_{\mathrm{B}} A_i. \tag{80}$$

The following lemma shows that in Euclidean spaces, without loss of generality we can restrict attention to covering A with *mesh cubes* defined in (5). Since all the mesh cubes partition the whole space, to calculate lower or upper Minkowski dimension of a set, it is sufficient to count the number of mesh cubes it intersects, hence justifying the name of box-counting dimension. Denote by $\tilde{N}_A(2^{-m})$ the smallest number of mesh cubes of size 2^{-m} that covers A, that is

$$\tilde{N}_A(2^{-m}) = \min\left\{k : A \subset \bigcup_{i=1}^{k} C_m(z_i), z_i \in \mathbb{Z}^n\right\} \tag{81}$$

$$= |\{z \in \mathbb{Z}^n : A \cap C_m(z) \neq \emptyset\}| \tag{82}$$

$$= |[A]_m|. \tag{83}$$

Lemma 3: Let A be a bounded subset in $(\mathbb{R}^n, \|\cdot\|_p)$, $1 \le p \le \infty$. The Minkowski dimensions satisfy

$$\underline{\dim}_{\mathrm{B}} A = \liminf_{m \to \infty} \frac{\log |[A]_m|}{m} \tag{84}$$

$$\overline{\dim}_{\mathrm{B}} A = \limsup_{m \to \infty} \frac{\log |[A]_m|}{m}. \tag{85}$$

Proof: See Appendix V. □

B. Definitions and Coding Theorems

Consider a source X^n in $\mathbb{R}^n$ equipped with an ℓ_p-norm. We define the minimum ϵ-achievable rate for Minkowski-dimension compression as follows.

Definition 10 (Minkowski-Dimension Compression Rate): Let $\{X_i : i \in \mathbb{N}\}$ be a stochastic process on $(\mathbb{R}^{\mathbb{N}}, \mathcal{B}^{\otimes \mathbb{N}})$. Define

$$\mathsf{R}_{\mathsf{B}}(\epsilon) = \limsup_{n \to \infty} \inf\left\{\frac{1}{n}\overline{\dim}_{\mathrm{B}} S : S \subset \mathbb{R}^n, \mathbb{P}\{X^n \in S\} \ge 1 - \epsilon\right\}. \tag{86}$$

Note that the conventional minimum source coding rate $r(\epsilon)$ in Definition 4 is defined like in (86) replacing $\overline{\dim}_{\mathrm{B}} S$ by $\log |S|$.

In general $0 \le \mathsf{R}_{\mathsf{B}}(\epsilon) \le 1$ for any $0 < \epsilon \le 1$. This is because for any n, there exists a compact subset $S \subset \mathbb{R}^n$, such that $\mathbb{P}\{X^n \in S\} \ge 1 - \epsilon$, and $\overline{\dim}_{\mathrm{B}} S \le n$ by definition. Several coding theorems for $\mathsf{R}_{\mathsf{B}}(\epsilon)$ are given as follows.

Theorem 14: Suppose that the source is memoryless with distribution such that $\bar{d}(X) < \infty$. Then

$$\mathsf{R}_{\mathsf{B}}(\epsilon) \le \lim_{\alpha \uparrow 1} \bar{d}_\alpha(X) \tag{87}$$

and

$$\mathsf{R}_{\mathsf{B}}(\epsilon) \ge \bar{d}(X) \tag{88}$$

for $0 < \epsilon < 1$.

Proof: See Section V-C. □

For the special cases of discrete-continuous-mixed and self-similar sources, we have the following tight results.

Theorem 15: Suppose that the source is memoryless with a discrete-continuous mixed distribution. Then

$$\mathsf{R}_{\mathsf{B}}(\epsilon) = \rho \tag{89}$$

for all $0 < \epsilon < 1$, where ρ is the weight of the continuous part of the distribution. If $d(X)$ is finite, then

$$\mathsf{R}_\mathsf{B}(\epsilon) = d(X). \tag{90}$$

Theorem 16: Suppose that the source is memoryless with a self-similar distribution that satisfies the strong separation condition. Then

$$\mathsf{R}_\mathsf{B}(\epsilon) = d(X) \tag{91}$$

for all $0 < \epsilon < 1$.

Theorem 17: Suppose that the source is memoryless and bounded, and its M-ary expansion consists of independent digits. Then

$$\mathsf{R}_\mathsf{B}(\epsilon) = \bar{d}(X) \tag{92}$$

for all $0 < \epsilon < 1$.

When $\bar{d}(X) = \infty$, $\mathsf{R}_\mathsf{B}(\epsilon)$ can take any value in $[0, 1]$ for all $0 < \epsilon < 1$. Such a source can be constructed using Theorem 15 as follows: Let the distribution of X be a mixture of a continuous and a discrete distribution with weights d and $1-d$ respectively, where the discrete part is supported on $\mathbb{N}$ and has infinite entropy. Then, $\bar{d}(X) = \infty$ by Proposition 1 and Theorem 1, but $\mathsf{R}_\mathsf{B}(\epsilon) = d$ by Theorem 15.

C. Proofs

Before showing the converse part of Theorem 14, we state two lemmas which are of independent interest in conventional lossless source coding theory.

Lemma 4: Assume that $\{X_i : i \in \mathbb{N}\}$ is a discrete memoryless source with common distribution P on the alphabet $\mathcal{X}$. $H(P) < \infty$. Let $\delta \geq -H(P)$. Denote by $p_n^*(P, \delta)$ the block error probability of the optimal $(n, \lfloor (H(X) + \delta)n \rfloor)$-code. Then, for any $n \in \mathbb{N}$

$$p_n^*(P, \delta) \leq \exp[-nE_0(P, \delta)] \tag{93}$$
$$p_n^*(P, \delta) \geq 1 - \exp[-nE_1(P, \delta)] \tag{94}$$

where the exponents are given by

$$E_0(P, \delta) = \inf_{Q: H(Q) > H(P) + \delta} D(Q\|P) \tag{95}$$
$$= \max_{\lambda \geq 0} \lambda \left[H(P) + \delta - H_{\frac{1}{1+\lambda}}(P) \right] \tag{96}$$
$$E_1(P, \delta) = \inf_{Q: H(Q) < H(P) + \delta} D(Q\|P) \tag{97}$$
$$= \max_{\lambda \geq 0} \lambda \left[H_{\frac{1}{1-\lambda}}(P) - H(P) - \delta \right]. \tag{98}$$

Lemma 5: For $\delta \geq 0$

$$\min \{E_0(P, \delta), E_1(P, -\delta)\} \geq \min \left\{ \frac{1}{8}, \frac{1}{2} \left[\frac{(\delta - \mathrm{e}^{-1} \log \mathrm{e})^+}{\log |\mathcal{X}|} \right]^2 \right\} \log \mathrm{e}. \tag{99}$$

Proof: See Appendix VII. □

Lemma 4 shows that the error exponents for lossless source coding are not only asymptotically tight, but also apply to *every* block length. This has been shown for rates above entropy in [29, Exercise 1.2.7, pp. 41–42] via a combinatorial method. A unified proof can be given through the method of Rényi entropy, which we omit for conciseness. The idea of using Rényi entropy to study lossless source coding error exponents was previously introduced by [30]–[32].

Lemma 5 deals with universal lower bounds on the source coding error exponents, in the sense that these bounds are independent of the source distribution. A better bound on $E_0(P, \delta)$ has been shown in [33]: for $\delta > 0$

$$E_0(P, \delta) \geq \frac{1}{2} \left(\frac{\delta}{\log |\mathcal{X}|} \right)^2 \log \mathrm{e}. \tag{100}$$

However, the proof of (100) was based on the dual expression (96) and a similar lower bound of random channel coding error exponent due to Gallager [34, Exercise 5.23], which cannot be applied to $E_1(P, \delta)$. Here we give a common lower bound on both exponents, which is a consequence of Pinsker's inequality [35] combined with the lower bound on entropy difference by variational distance [29].

Proof of Theorem 14: (Converse) Let $0 < \epsilon < 1$ and abbreviate $\bar{d}(X)$ as $\bar{d}$. Suppose $\mathsf{R}_\mathsf{B}(\epsilon) < \bar{d} - 4\delta$ for some $\delta > 0$. Then, for sufficiently large n there exists $S^n \subset \mathbb{R}^n$, such that $\mathbb{P}\{X^n \in S^n\} \geq 1 - \epsilon$ and $\overline{\dim}_\mathrm{B} S^n \leq (\bar{d} - 3\delta)n$.

First we assume that the source has *bounded* support, that is, $|X| \leq K$ a.s. for some $K > 0$. By Proposition 2, choose M such that for all $m > M$,

$$H([X]_m) \geq (\bar{d} - \delta)m. \tag{101}$$

By Lemma 3, for all n, there exists $M_n > M$, such that for all $m > M_n$

$$|[S^n]_m| \leq 2^{m(\overline{\dim}_\mathrm{B}(S^n) + \delta n)} \leq 2^{mn(\bar{d} - 2\delta)} \tag{102}$$
$$\mathbb{P}\{[X^n]_m \in [S^n]_m\} \geq \mathbb{P}\{X^n \in S^n\} \geq 1 - \epsilon. \tag{103}$$

Then, by (101), we have

$$|[S^n]_m| \leq 2^{(H([X]_m) - m\delta)n}. \tag{104}$$

Note that $[X^n]_m$ is a memoryless source with alphabet size at most $\log K + m$. By Lemmas 4 and 5, for all m, n, for all A_m^n such that $|A_m^n| \leq 2^{(H([X]_m) - m\delta)n}$, we have

$$\mathbb{P}\left\{[X^n]_m \in A_m^n\right\} \leq 2^{-nE_1(m\delta)} \tag{105}$$
$$\leq 2^{-\frac{n}{2}\left(\frac{(m\delta - \mathrm{e}^{-1}\log \mathrm{e})^+}{m + \log K}\right)^2}. \tag{106}$$

Choose m and n so large that the right-hand side of (106) is less than $1 - \epsilon$. In the special case of $A_m^n = [S^n]_m$, (106) contradicts (103) in view of (104).

Next we drop the restriction that X is almost surely bounded. Denote by μ the distribution of X. Let K be so large that

$$p = \mu(B(0, K)) > 1 - \delta. \tag{107}$$

Denote the normalized restriction of μ on $B(0, K)$ by μ_1, that is

$$\mu_1(A) = \frac{\mu(A \cap B(0, K))}{\mu(B(0, K))} \tag{108}$$

and the normalized restriction of μ on $B(0,K)^{\mathrm{c}}$ by μ_2. Then

$$\mu = p\mu_1 + (1-p)\mu_2. \tag{109}$$

By Theorem 2

$$\bar{d} \le p\bar{d}(\mu_1) + (1-p)\bar{d}(\mu_2) \tag{110}$$
$$\le p\bar{d}(\mu_1) + (1-p) \tag{111}$$

where $\bar{d}(\mu_2) \le 1$ because of the following: since $\bar{d}$ is finite, we have $\mathbb{E}[\log(1+|X|)] < \infty$ in view of Proposition 1. Consequently, $\mathbb{E}[\log(1+|X|)\mathbf{1}_{\{|X|>K\}}] < \infty$, hence $\bar{d}(\mu_2) \le 1$.

The distribution of $\tilde{X} \triangleq X\mathbf{1}_{\{X\in B(0,K)\}}$ is given by

$$\tilde{\mu} = p\mu_1 + (1-p)\delta_0 \tag{112}$$

where δ_0 denotes the Dirac measure with atom at 0. Then

$$\bar{d}(\tilde{X}) = p\bar{d}(\mu_1) \tag{113}$$
$$> \bar{d} - \delta \tag{114}$$

where:

- (113): by (112) and (31), since Theorem 1 implies $d(\delta_0) = 0$;
- (114): by (111) and (107).

For $T \subset \{1,\dots,n\}$, let $S_T^n = \{x_T^n : x^n \in S^n\}$. Define

$$\tilde{S}^n = \bigcup_{T\subset\{1,\dots,n\}} \{x^n \in \mathbb{R}^n : x_T^n \in S_T^n, x_{T^{\mathrm{c}}}^n = 0\}. \tag{115}$$

Then, for all n

$$\mathbb{P}\{\tilde{X}^n \in \tilde{S}^n\} \ge \mathbb{P}\{X^n \in S^n\} \ge 1-\epsilon \tag{116}$$

and

$$\overline{\dim}_{\mathrm{B}}\tilde{S}^n$$
$$= \max_{T\subset\{1,\dots,n\}} \overline{\dim}_{\mathrm{B}}\{x^n \in \mathbb{R}^n : x_T^n \in S_T^n, x_{T^{\mathrm{c}}}^n = 0\} \tag{117}$$
$$\le \overline{\dim}_{\mathrm{B}} S^n \tag{118}$$
$$\le (\bar{d} - 3\delta)n \tag{119}$$
$$< (\bar{d}(\tilde{X}) - 2\delta)n \tag{120}$$

where (120), (117), and (118) follow from (114), (79), and (80), respectively. But now (120) and (116) contradict the converse part of Theorem 14 for the bounded source $\tilde{X}$, which we have already proved.

(Achievability) Recall that

$$\hat{d} = \lim_{\alpha\uparrow 1} \bar{d}_\alpha. \tag{121}$$

We show that for all $0 < \epsilon < 1$, for all $\delta > 0$, there exists N such that for all $n > N$, there exists $S_n \subset \mathbb{R}^n$ with $\mathbb{P}\{X^n \in S^n\} \ge 1-\epsilon$ and $\overline{\dim}_{\mathrm{B}} S^n \le n(\hat{d}+\delta)$. Therefore, $\mathsf{R}_{\mathsf{B}}(\epsilon) \le \hat{d}$ readily follows.

Suppose for now that we can construct a sequence of subsets $V_m^n \subset 2^{-m}\mathbb{Z}^n$, such that for any n the following holds:

$$|V_m^n| \le 2^{mn(\hat{d}+\delta)} \tag{122}$$
$$\sum_{m\ge 1} \mathbb{P}\{[X^n]_m \notin V_m^n\} < \infty. \tag{123}$$

Denote

$$U_m^n = \sum_{z^n \in V_m^n} C_m(2^m z^n) \tag{124}$$
$$T_m^n = \bigcap_{j\ge m} U_j^n \tag{125}$$
$$T^n = \bigcup_{m\ge 1} T_m^n = \liminf_{m\to\infty} U_m^n. \tag{126}$$

Then, $[U_m^n]_m = V_m^n$. Now we claim that for each m, $\overline{\dim}_{\mathrm{B}} T_m^n \le (\hat{d}+\delta)n$ and $\mathbb{P}\{X^n \in T^n\} = 1$. First observe that for each $j \ge m$, $T_m^n \subset U_j^n$, therefore $\{C_j(2^j z^n) : z^n \in V_j^n\}$ covers T_m^n. Hence, $|[T_m^n]_j| \le |V_j^n| \le 2^{jn(\hat{d}+\delta)}$, by (122). Therefore, by Lemma 3

$$\overline{\dim}_{\mathrm{B}} T_m^n = \limsup_{j\to\infty} \frac{\log|[T_m^n]_j|}{j} \le n(\hat{d}+\delta). \tag{127}$$

By the Borel–Cantelli lemma, (123) implies that $\mathbb{P}\{X^n \in T^n\} = 1$. Let $S^n = \bigcup_{m=1}^M T_m^n$ where M is so large that $\mathbb{P}\{X^n \in S^n\} \ge 1-\epsilon$. By the finite subadditivity[6] of upper Minkowski dimension in Lemma 2, $\overline{\dim}_{\mathrm{B}} S^n = \max_{m=1,\dots,M} \overline{\dim}_{\mathrm{B}} T_m^n \le n(\hat{d}+\delta)$. By the arbitrariness of δ, the ϵ-achievability of rate $\hat{d}$ is proved.

Now let us proceed to the construction of the required $\{V_m^n\}$. To that end, denote by μ_m the probability mass function of $[X]_m$. Let

$$V_m^n = \left\{ z^n \in 2^{-m}\mathbb{Z}^n : \frac{1}{mn}\log\frac{1}{\mu_m^n(\{z^n\})} \le \hat{d}+\delta \right\} \tag{128}$$
$$= \left\{ z^n \in 2^{-m}\mathbb{Z}^n : \mu_m^n(\{z^n\}) \ge 2^{-mn(\hat{d}+\delta)} \right\}. \tag{129}$$

Then, immediately (122) follows from $\mathbb{P}\{[X^n]_m \in V_m^n\} \le 1$. Also, for $\alpha < 1$

$$\mathbb{P}\left\{[X^n]_m \notin V_m^n\right\}$$
$$= \sum_{z^n \in 2^{-m}\mathbb{Z}^n} \mu_m^n(\{z^n\})\mathbf{1}_{\{z^n\notin V_m^n\}} \tag{130}$$
$$\le \sum_{z^n \in 2^{-m}\mathbb{Z}^n} \mu_m^n(\{z^n\}) \left(\frac{2^{-mn(\hat{d}+\delta)}}{\mu_m^n(\{z^n\})}\right)^{1-\alpha} \tag{131}$$
$$= 2^{-nm(1-\alpha)(\hat{d}+\delta)} \sum_{z^n \in 2^{-m}\mathbb{Z}^n} [\mu_m^n(\{z^n\})]^\alpha \tag{132}$$
$$= 2^{(1-\alpha)(H_\alpha([X^n]_m) - mn(\hat{d}+\delta))} \tag{133}$$
$$= 2^{n(1-\alpha)(H_\alpha([X]_m) - m\hat{d} - m\delta)} \tag{134}$$

[6]Countable subadditivity fails for upper Minkowski dimension. Had it been satisfied, we could have picked $S^n = T^n$ to achieve $\epsilon = 0$.

where (134) follows from the fact that $[X^n]_m = [[X_1]_m, \ldots, [X_n]_m]^{\mathrm{T}}$ are i.i.d. and the joint Rényi entropy is the sum of individual Rényi entropies. Hence

$$\frac{\log \mathbb{P}\left\{[X^n]_m \notin V_m^n\right\}}{m} \le (1-\alpha)\left(\frac{H_\alpha([X]_m)}{m} - \hat{d} - \delta\right) n. \tag{135}$$

Choose $0 < \alpha < 1$ such that $\bar{d}_\alpha < \hat{d} + \delta/2$, which is guaranteed to exist in view of (121) and the fact that $\bar{d}_\alpha$ is nonincreasing in α according to Lemma 1. Then

$$\begin{aligned}
&\limsup_{m\to\infty} \frac{\log \mathbb{P}\{[X^n]_m \notin V_m^n\}}{m} \\
&\quad \le (1-\alpha)\left(\limsup_{m\to\infty} \frac{H_\alpha([X]_m)}{m} - \hat{d} - \delta\right) n && (136)\\
&\quad = (1-\alpha)(\bar{d}_\alpha - \hat{d} - \delta) n && (137)\\
&\quad \le -(1-\alpha) n\delta/2 < 0. && (138)
\end{aligned}$$

Hence, $\mathbb{P}\{[X^n]_m \notin V_m^n\}$ decays at least exponentially with m. Accordingly, (123) holds, and the proof of $\mathsf{R}_{\mathsf{B}}(\epsilon) \le \hat{d}$ is complete. □

Next we prove $\mathsf{R}_{\mathsf{B}}(\epsilon) = d(X)$ for the special case of discrete-continuous mixed sources. The following lemma is needed in the converse proof.

Lemma 6 [36, Th. 4.16]: Any Borel set $A \subset \mathbb{R}^k$ whose upper Minkowski dimension is strictly less than k has zero Lebesgue measure.

Proof of Theorem 15: (Achievability) Let the distribution of X be

$$\mu = (1-\rho)\mu_{\mathrm{d}} + \rho\mu_{\mathrm{c}} \tag{139}$$

where $0 \le \rho \le 1$, μ_{c} is a probability measure on $(\mathbb{R}, \mathcal{B})$ absolutely continuous with respect to Lebesgue measure and μ_{d} is a discrete probability measure. Let $\mathcal{A}$ be the collection of all the atoms of μ_{d}, which is, by definition, a countable subset of $\mathbb{R}$.

Let $W_i = \mathbf{1}_{\{X_i \notin \mathcal{A}\}}$. Then, $\{W_i\}$ is a sequence of i.i.d. binary random variables with expectation

$$\mathbb{E}W_i = \mathbb{P}\{X_i \notin \mathcal{A}\} = (1-\rho)\mu_{\mathrm{d}}(\mathcal{A}^{\mathrm{c}}) + \rho\mu_{\mathrm{c}}(\mathcal{A}^{\mathrm{c}}) = \rho. \tag{140}$$

By the weak law of large numbers (WLLN)

$$\begin{aligned}
\frac{1}{n}|\mathrm{spt}(X^n)| &= \frac{1}{n}\sum_{i=1}^n W_i && (141)\\
&\xrightarrow{\mathbb{P}} \rho && (142)
\end{aligned}$$

where the *generalized support* of vector x^n is defined as

$$\mathrm{spt}(x^n) = \{i = 1, \ldots, n : x_i \notin \mathcal{A}\}. \tag{143}$$

Fix an arbitrary $\delta > 0$ and let

$$C_n = \{x^n \in \mathbb{R}^n : |\mathrm{spt}(x^n)| < (\rho + \delta)\, n\}. \tag{144}$$

By (142), for sufficiently large n, $\mathbb{P}\{X^n \in C_n\} \ge 1 - \epsilon/2$. Decompose C_n as

$$C_n = \bigcup_{\substack{T\subset\{1,\ldots,n\}\\ |T|<(\rho+\delta)n}} \bigcup_{z\in\mathcal{A}^{n-|T|}} U_{T,z} \tag{145}$$

where

$$U_{T,z} = \{x^n \in \mathbb{R}^n : \mathrm{spt}(x^n) = T, x^n_{T^{\mathrm{c}}} = z\}. \tag{146}$$

Note that the collection of all $U_{T,z}$ is *countable*, and thus we may relabel them as $\{U_j : j \in \mathbb{N}\}$. Then, $C_n = \bigcup_j U_j$. Hence, there exists $J \in \mathbb{N}$ and $R > 0$, such that $\mathbb{P}\{X^n \in S^n\} \ge 1 - \epsilon$, where

$$S_n = B(0,R) \cap \bigcup_{j=1}^J U_j. \tag{147}$$

Then

$$\begin{aligned}
\overline{\dim}_{\mathrm{B}} S_n &= \max_{j=1,\ldots,J} \overline{\dim}_{\mathrm{B}} U_j \cap B(0,R) && (148)\\
&\le (\rho+\delta)n && (149)
\end{aligned}$$

where (148) is by Lemma 2, and (149) follows from $\overline{\dim}_{\mathrm{B}} U_j \cap B(0,R) \le |T| < (\rho+\delta)n$ for each j. This proves the ϵ-achievability of $\rho + \delta$. By the arbitrariness of δ, $\mathsf{R}_{\mathsf{B}}(\epsilon) \le \rho$.

(Converse) Since $\mathsf{R}_{\mathsf{B}}(\epsilon) \ge 0$, we can assume $\rho > 0$. Let $S_n \in \mathbb{R}^n$ be such that $\mathbb{P}\{X^n \in S_n\} \ge 1 - \epsilon$. Define

$$T_n = \{x^n \in \mathbb{R}^n : |\mathrm{spt}(x^n)| > (\rho - \delta)n\}. \tag{150}$$

By (142), $\mathbb{P}\{X^n \in T_n\} \ge 1 - \epsilon$ for sufficiently large n. Let $G_n = S_n \cap T_n$, then $\mathbb{P}\{X^n \in G_n\} \ge 1 - 2\epsilon$. Write G_n as the disjoint union

$$G_n = \bigcup_{\substack{T\subset\{1,\ldots,n\}\\ |T|>(\rho-\delta)n}} \bigcup_{z\in\mathcal{A}^{n-|T|}} V_{T,z} \tag{151}$$

where

$$V_{T,z} = U_{T,z} \cap G_n. \tag{152}$$

Also let $E_{T,z} = \{x^n_T : x^n \in V_{T,z}\}$. Since $\mathbb{P}\{X^n \in G_n\} \ge 1 - 2\epsilon > 0$, there exists $T \subset \{1, \ldots, n\}$ and $z \in \mathcal{A}^{n-|T|}$, such that $\mathbb{P}\{X^n \in V_{T,z}\} > 0$. Note that

$$\begin{aligned}
\mathbb{P}\{X^n \in V_{T,z}\} &= \rho^{|T|}(1-\rho)^{n-|T|}\mu_{\mathrm{c}}^{|T|}(E_{T,z})\\
&\quad \times \prod_{i=1}^{n-|T|} \mu_{\mathrm{d}}(\{z_i\}) > 0. && (153)
\end{aligned}$$

If $\rho < 1$, $\mu_{\mathrm{c}}^{|T|}(E_{T,z}) > 0$, which implies $E_{T,z} \subset \mathbb{R}^{|T|}$ has positive Lebesgue measure. By Lemma 6, $\overline{\dim}_{\mathrm{B}} E_{T,z} = |T| > (\rho - \delta)n$, hence

$$\overline{\dim}_{\mathrm{B}} S_n \ge \overline{\dim}_{\mathrm{B}} V_{T,z} = \overline{\dim}_{\mathrm{B}} E_{T,z} > (\rho - \delta)n. \tag{154}$$

If $\rho = 1$, $\mu = \mu_c$ implies that $S_n \subset \mathbb{R}^n$ has positive Lebesgue measure. Thus, $\overline{\dim}_B S_n = n$. By the arbitrariness of $\delta > 0$, the proof of $\mathsf{R}_B(\epsilon) \geq \rho$ is complete. □

Next we prove $\mathsf{R}_B(\epsilon) = d(X)$ for self-similar sources under the same assumption of Theorem 3. This result is due to the stationarity and ergodicity of the underlying discrete process that generates the analog source distribution.

Proof of Theorem 16: By Theorem 3, $d(X)$ is finite. Therefore, the converse follows from Theorem 14. To show achievability, we invoke the following definition. Define the *local dimension* of a Borel measure ν on $\mathbb{R}^n$ as the function (if the limit exists) [17, p. 169]

$$\dim_{\mathrm{loc}} \nu(x) = \lim_{r \downarrow 0} \frac{\log \nu(B(x,r))}{\log r}. \tag{155}$$

Denote the distribution of X^n by the product measure μ^n, which is also self-similar and satisfies the strong separation theorem. By [17, Lemma 6.4(b) and Prop. 10.6]

$$\dim_{\mathrm{loc}} \mu^n(x) = d(\mu^n) = nd(X) \tag{156}$$

holds for μ^n-almost every x. Define the sequence of random variables

$$D_m(X^n) = \frac{1}{m} \log \frac{1}{\mu^n(B(X^n, 2^{-m}))}. \tag{157}$$

Then, (156) implies that $D_m \xrightarrow{\text{a.s.}} nd(X)$ as $m \to \infty$. Therefore, for all $0 < \epsilon < 1$ and $\delta > 0$, there exists M, such that

$$\mathbb{P}\{ \bigcap_{m=M}^{\infty} \{D_m(X^n) \leq nd(X) + \delta\}\} \geq 1 - \epsilon. \tag{158}$$

Let

$$S_n = \bigcap_{m=M}^{\infty} \{x^n : D_m(x^n) \leq nd(X) + \delta\} \tag{159}$$

$$= \bigcap_{m=M}^{\infty} \left\{x^n : \mu^n(B(x^n, 2^{-m})) \geq 2^{-m(nd(X)+\delta)}\right\}. \tag{160}$$

Then, $\mathbb{P}\{X^n \in S_n\} \geq 1 - \epsilon$ in view of (158), and

$$\overline{\dim}_B S_n = \limsup_{m \to \infty} \frac{\log N_{S_n}(2^{-m})}{m} \tag{161}$$

$$\leq nd(X) + \delta \tag{162}$$

where (162) follows from (160) and $\mu^n(S_n) \leq 1$. Hence, $\mathsf{R}_B(\epsilon) \leq d(X)$ is proved. □

Finally, we prove $\mathsf{R}_B(\epsilon) = \bar{d}(X)$ for memoryless sources whose M-ary expansion consisting of independent digits.

Proof of Theorem 17: Without loss of generality, assume $M = 2$, that is, the binary expansion of X consists of independent bits. We follow the same steps as in the achievability proof of Theorem 14. Suppose for now that we can construct a sequence of subsets $\{V_m^n\}$, such that (123) holds and their cardinality does not exceed

$$|V_m^n| \leq 2^{n(H([X]_m) + m\delta)}. \tag{163}$$

By the same arguments that lead to (127), the sets defined in (125) satisfy

$$\overline{\dim}_B T_m^n = \limsup_{j \to \infty} \frac{\log |[T_m^n]_j|}{j} \tag{164}$$

$$= n \limsup_{j \to \infty} \frac{H([X]_j) + j\delta}{j} \tag{165}$$

$$= n(\bar{d} + \delta). \tag{166}$$

Since $\lim_{m \to \infty} \mathbb{P}\{X^n \in T_m^n\} = 1$, this shows the ϵ-achievability of $\bar{d}$.

Next we proceed to construct the required $\{V_m^n\}$. Applying Lemma 4 to the DMS $[X]_m$ and blocklength n yields

$$p_n^*([X]_m, m\delta) \leq 2^{-nE_0(m\delta)}. \tag{167}$$

By the assumption that X is bounded, without loss of generality, we shall assume $|X| \leq 1$ a.s. Therefore, the alphabet size of $[X]_m$ is at most 2^m. Simply applying (100) to $[X]_m$ yields

$$E_0([X]_m, m\delta) \geq \frac{\delta^2}{2} \tag{168}$$

which does not grow with m and cannot suffice for our purpose of constructing V_m^n. Exploiting the structure that $[X]_m$ consists of independent bits, in Appendix VIII, we show a much better bound

$$E_0([X]_m, m\delta) \geq \frac{m\delta^2}{2}. \tag{169}$$

Then, by (167) and (169), there exists V_m^n, such that (163) holds and

$$\mathbb{P}\{[X^n]_m \notin V_m^n\} \leq 2^{-\frac{nm\delta^2}{2}} \tag{170}$$

which implies (123). This concludes the proof of $\mathsf{R}_B(\epsilon) \leq \bar{d}$. □

VI. Lossless Linear Compression

In this section, we analyze lossless compression with linear encoders, which are the basic elements in compressed sensing. Capitalizing on the approach of Minkowski-dimension compression developed in Section V, we obtain achievability results for linear compression. For memoryless sources with a discrete-continuous mixed distribution, we also establish a converse in Theorem 6 which shows that the information dimension is the fundamental limit of lossless linear encoding.

A. Minkowski-Dimension Compression and Linear Compression

The following theorem establishes the relationship between Minkowski-dimension compression and linear data compression.

Theorem 18 (Linear Encoding: General Achievability): For general sources

$$\mathsf{R}^*(\epsilon) \leq \mathsf{R}_B(\epsilon). \tag{171}$$

Moreover, we have the following.

1) For all linear encoders (except possibly those in a set of zero Lebesgue measure on the space of real matrices), block error probability ϵ is achievable.
2) For all $\epsilon' > \epsilon$ and

$$0 < \beta < \frac{R - \mathsf{R}_\mathsf{B}(\epsilon)}{R} \tag{172}$$

where $R > \mathsf{R}_\mathsf{B}(\epsilon)$ is the compression rate, there exists a β-Hölder continuous decoder that achieves block error probability ϵ'.

Consequently, in view of Theorem 18, the results on $\mathsf{R}_\mathsf{B}(\epsilon)$ for memoryless sources in Theorems 14–16 yield the achievability results in Theorems 5–7, respectively. Hölder exponents of the decoder can be found by replacing $\mathsf{R}_\mathsf{B}(\epsilon)$ in (172) by its respective upper bound.

For discrete-continuous sources, the achievability in Theorem 6 can be shown directly without invoking the general result in Theorem 18. See Remark 3. From the converse proof of Theorem 6, we see that effective compression can be achieved with linear encoders, i.e., $\mathsf{R}^*(\epsilon) < 1$, only if the source distribution is *not* absolutely continuous with respect to Lebesgue measure.

Remark 1: Linear embedding of low-dimensional subsets in Banach spaces was previously studied in [37]–[39], etc., in a nonprobabilistic setting. For example, [38, Th 1.1] showed that: for a subset S of a Banach space with $\overline{\dim}_\mathrm{B} S < k/2$, there exists a bounded linear function that embeds S into $\mathbb{R}^k$. Here in a probabilistic setup, the embedding dimension can be improved by a factor of two.

Following the idea in the proof of Theorem 18, we obtain a nonasymptotic result of lossless linear compression for k-sparse vectors, which is relevant to compressed sensing.

Corollary 1: Denote the collection of all k-sparse vectors in $\mathbb{R}^n$ by

$$\Sigma_k = \{x^n \in \mathbb{R}^n : \|x^n\|_0 \le k\}. \tag{173}$$

Let μ be a σ-finite Borel measure on $\mathbb{R}^n$. Then, given any $l \ge k+1$, for Lebesgue-a.e. $l \times n$ real matrix $\mathbf{H}$, there exists a Borel function $g : \mathbb{R}^l \to \mathbb{R}^n$, such that $g(\mathbf{H}x^n) = x^n$ for μ-a.e. $x^n \in \Sigma_k$. Moreover, when μ is finite, for any $\epsilon > 0$ and $0 < \beta < 1 - \frac{k}{l}$, there exists a $l \times n$ matrix $\mathbf{H}^*$ and $g^* : \mathbb{R}^l \to \mathbb{R}^n$, such that $\mu\{x^n \in \Sigma_k : g^*(\mathbf{H}^* x^n) \ne x^n\} \le \epsilon$ and g^* is β-Hölder continuous.

Remark 2: The assumption that the measure μ is σ-finite is essential, because the validity of Corollary 1 hinges upon Fubini's theorem, where σ-finiteness is an indispensable requirement. Consequently, if μ is the distribution of a k-sparse random vector with uniformly chosen support and Gaussian distributed nonzero entries, we conclude that all k-sparse vectors can be linearly compressed except for a subset of zero measure under μ. On the other hand, if μ is the counting measure on Σ_k, Corollary 1 no longer applies because that μ is not σ-finite. In fact, if $l < 2k$, no linear encoder from $\mathbb{R}^n$ to $\mathbb{R}^l$ works for *every* k-sparse vector, even if no regularity constraint is imposed on the decoder. This is because no $l \times n$ matrix acts injectively on Σ_k. To see this, introduce the notation

$$\mathcal{C} \ominus \mathcal{C} = \{x - y : x, y \in \mathcal{C}\} \tag{174}$$

where $\mathcal{C} \subset \mathbb{R}^n$. Then, for any $l \times n$ matrix $\mathbf{H}$

$$(\Sigma_k \ominus \Sigma_k) \cap \mathrm{Ker}(\mathbf{H}) = \Sigma_{2k} \cap \mathrm{Ker}(\mathbf{H}) \ne \{0\}. \tag{175}$$

Hence, there exist two k-sparse vectors that have the same image under $\mathbf{H}$.

On the other hand, $l = 2k$ is sufficient to linearly compress all k-sparse vectors, because (175) holds for Lebesgue-a.e. $2k \times n$ matrix $\mathbf{H}$. To see this, choose $\mathbf{H}$ to be a random matrix with i.i.d. entries according to some continuous distribution (e.g., Gaussian). Then, (190) holds if and only if all $2k \times 2k$ submatrices formed by $2k$ columns of $\mathbf{H}$ are invertible. This is an almost sure event, because the determinant of each of the $\binom{n}{2k}$ submatrices is an absolutely continuous random variable. The sufficiency of $l = 2k$ is a bit stronger than the result in Remark 1, which gives $l = 2k + 1$. For an explicit construction of such a $2k \times n$ matrix, we can choose $\mathbf{H}$ to be the matrix $H_{ij} = \cos(\frac{(i-1)j\pi}{n})$ (see Appendix IV).

B. Auxiliary Results

Let $0 < k < n$. Denote by $G(n,k)$ the Grassmannian manifold [25] consisting of all k-dimensional subspaces of $\mathbb{R}^n$. For $V \in G(n,k)$, the orthogonal projection from $\mathbb{R}^n$ to V defines a linear mapping $\mathrm{proj}_V : \mathbb{R}^n \to \mathbb{R}^n$ of rank k. The technique we use in the achievability proof of linear analog compression is to use the random orthogonal projection proj_V as the encoder, where V is distributed according to the invariant probability measure on $G(n,k)$, denoted by $\gamma_{n,k}$ [25]. The relationship between $\gamma_{n,k}$ and the Lebesgue measure on $\mathbb{R}^{k\times n}$ is shown in the following lemma.

Lemma 7 [25, Exercise 3.6]: Denote the rows of a $k \times n$ matrix $\mathbf{H}$ by $H_1, \ldots, H_k$, the row span of $\mathbf{H}$ by $\mathrm{Im}(\mathbf{H}^\mathrm{T})$, and the volume of the unit ℓ_2-ball in $\mathbb{R}^n$ by $\alpha(n)$. Set

$$\mathcal{B}_L = \{\mathbf{H} \in \mathbb{R}^{k\times n} : \|H_i\|_2 \le L, i = 1, \ldots, k\}. \tag{176}$$

Then, for $\mathcal{S} \subset G(n,k)$ measurable, i.e., a collection of k-dimensional subspaces of $\mathbb{R}^n$

$$\gamma_{n,k}(\mathcal{S}) = \alpha(n)^{-k} \mathrm{Leb}\{\mathbf{H} \in \mathcal{B}_1 : \mathrm{Im}(\mathbf{H}^\mathrm{T}) \in \mathcal{S}\}. \tag{177}$$

The following result states that a random projection of a given vector is not too small with high probability. It plays a central role in estimating the probability of "bad" linear encoders.

Lemma 8 [25, Lemma 3.11]: For any $x^n \in \mathbb{R}^n \backslash \{0\}$

$$\gamma_{n,k}\{V : \|\mathrm{proj}_V x^n\|_2 \le \delta\} \le \frac{2^n \delta^k}{\alpha(n)\,\|x^n\|_2^k}. \tag{178}$$

To show the converse part of Theorem 6, we will invoke the Steinhaus theorem as an auxiliary result.

Lemma 9 (Steinhaus [40]): For any measurable set $\mathcal{C} \subset \mathbb{R}^n$ with positive Lebesgue measure, there exists an open ball centered at 0 contained in $\mathcal{C} \ominus \mathcal{C}$.

Lastly, with the notation in (174), we give a characterization of the fundamental limit of lossless linear encoding as follows. The proof is omitted for conciseness.

Lemma 10: $\mathsf{R}^*(\epsilon)$ is the infimum of $R > 0$ such that for sufficiently large n, there exists a Borel set $S^n \subset \mathbb{R}^n$ and a linear subspace $H^n \subset \mathbb{R}^n$ of dimension at least $\lceil (1-R)n \rceil$, such that

$$(S^n \ominus S^n) \cap H^n = \{0\}. \tag{179}$$

C. Proofs

Proof of Theorem 18: We first show (171). Fix $0 < \delta' < \delta$ arbitrarily. Let $R = \mathsf{R}_{\mathsf{B}}(\epsilon) + \delta$, $k = \lfloor Rn \rfloor$ and $k' = \lfloor (\mathsf{R}_{\mathsf{B}}(\epsilon) + \delta')n \rfloor$. We show that there exists a matrix $\mathbf{H}_n \in \mathbb{R}^{n \times n}$ of rank k and $g_n : \mathrm{Im}(\mathbf{H}_n) \to \mathbb{R}^n$ Borel measurable such that

$$\mathbb{P}\{g_n(\mathbf{H}_n X^n) = X^n\} \geq 1 - \epsilon \tag{180}$$

for sufficiently large n.

By definition of $\mathsf{R}_{\mathsf{B}}(\epsilon)$, there exists $N_0 \in \mathbb{N}$ such that for all $n \geq N_0$, there exists a compact $U_n \subset \mathbb{R}^n$ such that $\mathbb{P}\{X^n \in U_n\} \geq 1 - \epsilon$ and $\overline{\dim}_{\mathrm{B}} U_n \leq k'$. Given an encoding matrix $\mathbf{H}$, define the decoder $g_{\mathbf{H}} : \mathrm{Im}(\mathbf{H}) \to \mathbb{R}^n$ as

$$g_{\mathbf{H}}(y^k) = \min \mathbf{H}^{-1}(\{y^k\}) \cap U_n \tag{181}$$

where the $\min$ is taken componentwise[7]. Since $\mathbf{H}^{-1}(\{y^k\})$ is closed and U_n is compact, $\mathbf{H}^{-1}(\{y^k\}) \cap U_n$ is compact. Hence, $g_{\mathbf{H}}$ is well defined.

Next consider a random orthogonal projection matrix $\mathbf{\Xi} = \mathrm{proj}_{\boldsymbol{V}}$ independent of X^n, where $\boldsymbol{V} \in G(n,k)$ is a random k-dimensional subspace distributed according to the invariant measure $\gamma_{n,k}$ on $G(n,k)$. We show that for all $n \geq N_0$

$$\mathbb{P}\{g_{\mathbf{\Xi}}(\mathbf{\Xi} X^n) \neq X^n\} \leq \epsilon \tag{182}$$

which implies that there exists at least one realization of $\mathbf{\Xi}$ that satisfies (180). To that end, we define

$$p_{\mathrm{e}}(\mathbf{H}) = \mathbb{P}\{X^n \in U_n, g_{\mathbf{H}}(\mathbf{H}X^n) \neq X^n\} \tag{183}$$

and use the union bound

$$\mathbb{P}\{g_{\mathbf{\Xi}}(\mathbf{\Xi} X^n) \neq X^n\} \leq \mathbb{P}\{X^n \notin U_n\} + \mathbb{E}[p_{\mathrm{e}}(\mathbf{\Xi})] \tag{184}$$

where the first term $\mathbb{P}\{X^n \notin U_n\} \leq \epsilon$. Next we show that the second term is *zero*. Let $\mathcal{U}(x^n) = U_n - x^n$. Then

$$\begin{aligned}
&\mathbb{E}[p_{\mathrm{e}}(\mathbf{\Xi})] \\
&\quad = \int_{U_n} \mathbb{P}\{g_{\mathbf{\Xi}}(\mathbf{\Xi} x^n) \neq x^n\} \mu^n(\mathrm{d}x^n) && (185)\\
&\quad \leq \int_{U_n} \mathbb{P}\{\exists y^n \in \mathcal{U}(x^n) : \mathbf{\Xi} y^n = 0\} \mu^n(\mathrm{d}x^n) && (186)\\
&\quad = \int_{U_n} \mathbb{P}\{\mathrm{Ker}(\mathbf{\Xi}) \cap \mathcal{U}(x^n) \neq \{0\}\} \mu^n(\mathrm{d}x^n). && (187)
\end{aligned}$$

[7]Alternatively, we can use any other tie-breaking strategy as long as Borel measurability is satisfied.

We show that for all $x^n \in U_n$, $\mathbb{P}\{\mathrm{Ker}(\mathbf{\Xi}) \cap \mathcal{U}(x^n) \neq \{0\}\} = 0$. To this end, let

$$0 < \beta < \frac{\delta - \delta'}{R} = \frac{k - k'}{k}. \tag{188}$$

Define

$$T_j(x^n) = \{y^n \in \mathcal{U}(x^n) : \|y^n\|_2 \geq 2^{-j\beta}\} \tag{189}$$

$$Q_j = \{\mathbf{H} : \forall y^n \in T_j(x^n), \|\mathbf{H}y^n\|_2 \geq 2^{-j}\}. \tag{190}$$

Then

$$\bigcup_{j \geq 1} T_j(x^n) = \mathcal{U}(x^n) \backslash \{0\}. \tag{191}$$

Observe that $\mathbf{H} \in \bigcap_{j \geq J} Q_j$ implies that

$$\|\mathbf{H}y^n\|_2 \geq 2^{-(J+1)} \|y^n\|_2^{\frac{1}{\beta}} \qquad \forall y^n \in \mathcal{U}(x^n). \tag{192}$$

Therefore, $\mathbf{H} \in Q_j$ for all but a finite number of j's if and only if

$$\|\mathbf{H}y^n\|_2 \geq C(\mathbf{H}, x^n) \|y^n\|_2^{\frac{1}{\beta}} \qquad \forall y^n \in \mathcal{U}(x^n) \tag{193}$$

for some $C(\mathbf{H}, x^n) > 0$.

Next we show that $\mathbf{\Xi} \in Q_j$ for all but a finite number of j's with probability one. Cover $\mathcal{U}(x^n)$ with $2^{-(j+1)}$-balls. The centers of those balls that intersect $T_j(x^n)$ are denoted by $\{w_{i,j}, i = 1, \ldots, M_j\}$. Pick $y_{i,j} \in T_j(x^n) \cap B(w_{i,j}, 2^{-(j+1)})$. Then, $B(w_{i,j}, 2^{-(j+1)}) \subset B(y_{i,j}, 2^{-j})$, hence $\{B(y_{i,j}, 2^{-j}) : i = 1, \ldots, M_j\}$ cover $T_j(x^n)$. Suppose $\|\mathbf{H}y_{i,j}\|_2 \geq 2^{-(j-1)}$, then for any $z \in B(y_{i,j}, 2^{-j})$

$$\begin{aligned}
\|\mathbf{H}z\|_2 &\geq \|\mathbf{H}y_{i,j}\|_2 - \|\mathbf{H}(y_{i,j} - z)\|_2 && (194)\\
&\geq \|\mathbf{H}y_{i,j}\|_2 - \|y_{i,j} - z\|_2 && (195)\\
&\geq 2^{-j} && (196)
\end{aligned}$$

where (195) follows because $\mathbf{H}$ is an orthogonal projection. Thus, $\|\mathbf{H}y_{i,j}\|_2 \geq 2^{-(j-1)}$ for all $i = 1, \ldots, M_j$ implies that $\mathbf{H} \in Q_j$. Therefore, by the union bound

$$\sum_{j \geq 1} \mathbb{P}\{\mathbf{\Xi} \in {Q_j}^{\mathrm{c}}\} \leq \sum_{j \geq 1} \sum_{i=1}^{M_j} \mathbb{P}\{\|\mathbf{\Xi} y_{i,j}\|_2 \leq 2^{-(j-1)}\}. \tag{197}$$

By Lemma 8

$$\begin{aligned}
&\mathbb{P}\{\|\mathbf{\Xi} y_{i,j}\|_2 \leq 2^{-(j-1)}\} \\
&\quad = \gamma_{n,k}\left(\left\{V : \|\mathrm{proj}_V y_{i,j}\|_2 \leq 2^{-(j-1)}\right\}\right) && (198)\\
&\quad \leq \frac{2^{n-(j-1)k}}{\alpha(n) \|y_{i,j}\|_2^k} && (199)\\
&\quad \leq \frac{2^{n+k(1-j+j\beta)}}{\alpha(n)} && (200)
\end{aligned}$$

where (200) is due to $\|y_{i,j}\|_2 \geq 2^{-j\beta}$, because $y_{i,j} \in T_j(x^n)$. Since $\overline{\dim}_{\mathrm{B}} U_n < k'$, there is a constant C_1 such that $N_{U_n}(2^{-j-1}) \leq C_1 2^{jk'}$. Since $\mathcal{U}(x^n)$ is a translation of U_n, it follows that

$$M_j \leq C_1 2^{jk'}. \tag{201}$$

Thus

$$\begin{aligned}
&\sum_{j \geq 1} \mathbb{P}\{\mathbf{\Xi} \in {Q_j}^{\mathrm{c}}\} \\
&\quad \leq C_1 \alpha(n)^{-1} 2^{n+k} \sum_{j \geq 1} 2^{j(k' - (1-\beta)k)} && (202)\\
&\quad < \infty && (203)
\end{aligned}$$

where:
- (202): by substituting (200) and (201) into (197);
- (203): by (188).

Therefore, by the Borel–Cantelli lemma, $\mathbf{\Xi} \in Q_j$ for all but a finite number of j's with probability one. Hence

$$\mathbb{P}\{||\mathbf{\Xi}y^n||_2 \geq C(\mathbf{\Xi}, x^n)||y^n||_2^{\frac{1}{\beta}}, \forall y^n \in \mathcal{U}(x^n)\} = 1 \quad (204)$$

which implies that for any $x^n \in U_n$

$$\mathbb{P}\{\mathrm{Ker}(\mathbf{\Xi}) \cap \mathcal{U}(x^n) = \{0\}\} = 1. \quad (205)$$

In view of (187)

$$\mathbb{E}[p_{\mathrm{e}}(\mathbf{\Xi})] = \mathbb{P}\{X^n \in U_n, g_{\mathbf{\Xi}}(\mathbf{\Xi}X^n) \neq X^n\} = 0 \quad (206)$$

whence (182) follows. This shows the ϵ-achievability of R. By the arbitrariness of $\delta > 0$, (171) is proved.

Now we show that

$$\mathbb{P}\{g_{\mathbf{A}}(\mathbf{A}X^n) \neq X^n\} \leq \epsilon \quad (207)$$

holds for all $\mathbf{A} \in \mathbb{R}^{k\times n}$ except possibly on a set of zero Lebesgue measure, where $g_{\mathbf{A}}$ is the corresponding decoder for $\mathbf{A}$ defined in (181). Note that

$$\begin{aligned}
&\{\mathbf{A} : \mathbb{P}\{g_{\mathbf{A}}(\mathbf{A}X^n) \neq X^n\} > \epsilon\} \\
&\quad \subset \{\mathbf{A} : p_{\mathrm{e}}(\mathbf{A}) > 0\} && (208)\\
&\quad = \left\{\mathbf{A} : p_{\mathrm{e}}(\mathrm{proj}_{\mathrm{Im}(\mathbf{A}^{\mathrm{T}})}) > 0\right\} && (209)
\end{aligned}$$

where:
- (208): by (184);
- (209): by (187) and $\mathrm{Ker}(\mathbf{A}) = \mathrm{Ker}(\mathrm{proj}_{\mathrm{Im}(\mathbf{A}^{\mathrm{T}})})$.

Define

$$\mathcal{S} \triangleq \{V \in G(n,k) : p_{\mathrm{e}}(\mathrm{proj}_V) > 0\}. \quad (210)$$

Recalling $\mathcal{B}_L$ defined in (176), we have

$$\begin{aligned}
&\mathrm{Leb}\left\{\mathbf{A} \in \mathcal{B}_1 : p_{\mathrm{e}}(\mathrm{proj}_{\mathrm{Im}(\mathbf{A}^{\mathrm{T}})}) > 0\right\} \\
&\quad = \mathrm{Leb}\{\mathbf{A} \in \mathcal{B}_1 : \mathrm{Im}(\mathbf{A}^{\mathrm{T}}) \in \mathcal{S}\} && (211)\\
&\quad = \alpha(n)^k \gamma_{n,k}(\mathcal{S}) && (212)\\
&\quad = 0 && (213)
\end{aligned}$$

where:
- (211): by (209) and since $\mathrm{Im}(\mathbf{A}^{\mathrm{T}}) \in G(n,k)$ holds Lebesgue-a.e.;
- (212): by Lemma 7;
- (213): by (206).

Observe that (213) implies that for any L

$$\mathrm{Leb}\left\{\mathbf{A} \in \mathcal{B}_L : p_{\mathrm{e}}(\mathrm{proj}_{\mathrm{Im}(\mathbf{A}^{\mathrm{T}})}) > 0\right\} = 0. \quad (214)$$

Since $\mathbb{R}^{n\times k} = \bigcup_{L\geq 1} \mathcal{B}_L$, in view of (209) and (214), we conclude that (207) holds Lebesgue-a.e.

Finally, we show that for any $\epsilon' > \epsilon$, there exists a sequence of matrices and β-Hölder continuous decoders that achieves compression rate R and block error probability ϵ'. Since $\mathbf{\Xi} \in Q_j$ for all but a finite number of j's a.s., there exists a J_n (independent of x^n), such that

$$\mathbb{P}\left\{\mathbf{\Xi} \in \bigcap_{j\geq J_n} Q_j\right\} \geq 1 - \epsilon' + \epsilon. \quad (215)$$

Thus, by (192), for any $x^n \in U_n$

$$\mathbb{P}\left\{||\mathbf{\Xi}y^n||_2 \geq 2^{-(J_n+1)}\,||y^n||_2^{\frac{1}{\beta}}\,,\ \forall y^n \in \mathcal{U}(x^n)\right\} \geq 1 - \epsilon' + \epsilon. \quad (216)$$

Integrating (216) with respect to $\mu^n(\mathrm{d}x^n)$ on U_n and by Fubini's theorem, we have

$$\begin{aligned}
&\mathbb{P}\Big\{X^n \in U_n, ||\mathbf{\Xi}(y^n - X^n)||_2 \\
&\quad \geq 2^{-J_n+1}\,||y^n - X^n||_2^{\frac{1}{\beta}}\ \forall y^n \in U_n\Big\} \\
&\quad = \int \mathbb{P}\Big\{X^n \in U_n, ||\mathbf{H}(y^n - X^n)||_2 \\
&\qquad\qquad \geq 2^{-J_n+1}\,||y^n - X^n||_2^{\frac{1}{\beta}}\ \forall y^n \in U_n\Big\} \\
&\qquad \times P_{\mathbf{\Xi}}(\mathrm{d}\mathbf{H}) && (217)\\
&\quad \geq 1 - \epsilon'. && (218)
\end{aligned}$$

Hence, there exists $S_n \subset U_n$ and an orthogonal projection matrix $\mathbf{H}_n$ of rank k, such that $\mathbb{P}\{X^n \in S_n\} \geq 1 - \epsilon'$ and for all $x^n, y^n \in S_n$

$$||\mathbf{H}_n(y^n - x^n)||_2 \geq 2^{-(J_n+1)}\,||y^n - x^n||_2^{\frac{1}{\beta}} \quad (219)$$

for all $x^n, y^n \in S_n$. Therefore[8] $\mathbf{H}_n^{-1}|_{\mathbf{H}_n(S_n)}$ is $(2^{\beta(J_n+1)}, \beta)$-Hölder continuous. By the extension theorem of Hölder continuous mappings [41], $\mathbf{H}_n^{-1}$ can be extended to $g_n : \mathbb{R}^k \to \mathbb{R}^n$ that is β-Hölder continuous. Then

$$\mathbb{P}\{g_n(\mathbf{H}_n X^n) \neq X^n\} \leq \epsilon'. \quad (220)$$

Recall from (188) that $0 < \beta < \frac{R - \mathsf{R}_{\mathsf{B}}(\epsilon) - \delta'}{R}$. By the arbitrariness of δ', (172) holds. □

Remark 3: Without recourse to the general result in Theorem 18, the achievability for discrete-continuous sources in Theorem 6 can be proved directly as follows. In (184), choose

$$U_n = \{x^n \in \mathbb{R}^n : |\mathrm{spt}(x^n)| \leq (\rho + \delta/2)n\} \quad (221)$$

and consider $\mathbf{\Xi}$ whose entries are i.i.d. standard Gaussian (or any other absolutely continuous distribution on $\mathbb{R}$). Using linear algebra, it is straightforward to show that the second term in (184) is zero. Thus, the block error probability vanishes since $\mathbb{P}\{X^n \notin U_n\} \to 0$.

Finally, we complete the proof of Theorem 6 by proving the converse.

Converse Proof of Theorem 6: Let the distribution of X be defined as in (139). We show that for any $\epsilon < 1$, $\mathsf{R}^*(\epsilon) \geq \rho$. Since $\mathsf{R}^*(\epsilon) \geq 0$, assume $\rho > 0$. Fix an arbitrary $0 < \delta < \rho$. Suppose $R = \rho - \delta$ is ϵ-achievable. Let $k = \lfloor(\rho - \delta)n\rfloor$ and $k' = \lfloor(\rho - \delta/2)n\rfloor$. By Lemma 10, for sufficiently large n, there exist a Borel set S^n and a linear subspace $H^n \subset \mathbb{R}^n$,

[8] $f|_A$ denotes the restriction of f on the subset A.

such that $\mathbb{P}\{X^n \in S^n\} \geq 1-\epsilon$, $S^n \ominus S^n \cap H^n = \{0\}$ and $\dim H^n \geq n-k$.

If $\rho = 1$, $\mu = \mu_{\mathrm{c}}$ is absolutely continuous with respect to Lebesgue measure. Therefore, S^n has positive Lebesgue measure. By Lemma 9, $S^n \ominus S^n$ contains an open ball in $\mathbb{R}^n$. Hence, $S^n \ominus S^n \cap H^n = \{0\}$ cannot hold for any subspace H^n with positive dimension. This proves $\mathsf{R}^*(\epsilon) \geq 1$. Next we assume that $0 < \rho < 1$.

Let

$$T_n = \{x^n \in \mathbb{R}^n : |\mathrm{spt}(x^n)| \geq k'\} \tag{222}$$

and $G_n = S^n \cap T_n$. By (142), for sufficiently large n, $\mathbb{P}\{X^n \notin T_n\} \leq (1-\epsilon)/2$, hence $\mathbb{P}\{X^n \in G_n\} \geq (1-\epsilon)/2$.

Next we decompose G_n according to the generalized support of x^n

$$G_n = \bigcup_{\substack{U \subset \{1,\ldots,n\} \\ |U| \geq k'}} C_U \tag{223}$$

where we have denoted the disjoint subsets

$$C_U = \{x^n \in S^n : \mathrm{spt}(x^n) = U\}. \tag{224}$$

Then

$$\sum_{\substack{U \subset \{1,\ldots,n\} \\ |U| \leq k'}} \mathbb{P}\{X^n \in C_U\} = \mathbb{P}\{X^n \in G_n\} \tag{225}$$

$$\geq (1-\epsilon)/2 > 0. \tag{226}$$

So there exists $U \subset \{1,\ldots,n\}$ such that $|U| \leq k'$ and $\mathbb{P}\{X^n \in C_U\} > 0$.

Next we decompose each C_U according to $x^n_{U^{\mathrm{c}}}$ which can only take countably many values. Let $j = |U|$. For $y \in \mathcal{A}^{n-j}$, let

$$B_y = \{x^n : x^n \in C_U, x^n_{U^{\mathrm{c}}} = y\} \tag{227}$$

$$D_y = \{x^n_U : x^n \in B_y\}. \tag{228}$$

Then, C_U can be written as a disjoint union of B_y

$$C_U = \bigcup_{y \in \mathcal{A}^{n-j}} B_y. \tag{229}$$

Since $\mathbb{P}\{X^n \in C_U\} > 0$, there exists $y \in \mathcal{A}^{n-j}$ such that $\mathbb{P}\{X^n \in B_y\} > 0$.

Note that

$$\mathbb{P}\{X^n \in B_y\} = \mu^n(B_y) \tag{230}$$

$$= \rho^j (1-\rho)^{n-j} \mu_{\mathrm{c}}^j(D_y) \prod_{i=1}^{n-j} \mu_{\mathrm{d}}(\{y_i\}) \tag{231}$$

$$> 0. \tag{232}$$

Therefore, $\mu_{\mathrm{c}}^j(D_y) > 0$. Since μ_{c}^j is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^j$, D_y has positive Lebesgue measure. By Lemma 9, $D_y \ominus D_y$ contains an open ball $B_2^j(0,\delta_0)$ for some $\delta_0 > 0$. Therefore, we have

$$K \subset B_y \ominus B_y \subset S^n \ominus S^n \tag{233}$$

where $K = \{x^n \in \mathbb{R}^n : x^n_{U^{\mathrm{c}}} = 0, x^n_U \in B_2^j(0,\delta_0)\}$. Hence, K contains j linear independent vectors, denoted by $\{a_1,\ldots,a_j\}$. Let $\{b_1,\ldots,b_m\}$ be a basis for H^n, where $m \geq n-k$ by assumption. Since $j = |U| \geq k' > k$, we conclude that $\{a_1,\ldots,a_j,b_1,\ldots,b_m\}$ are linearly *dependent*. Therefore

$$\sum_{l=1}^{m} \beta_l b_l = \sum_{l=1}^{j} \alpha_l a_l \neq 0 \tag{234}$$

where $\alpha_i \neq 0$ and $\beta_l \neq 0$ for some i and l. If we choose those nonzero coefficients sufficiently small, then $\sum_{l=1}^{j} \alpha_l a_l \in K$ and $\sum_{l=1}^{m} \beta_l b_l \in H^n$ since H^n is a linear subspace. This contradicts $S^n \ominus S^n \cap H^n = \{0\}$. Thus, $\mathsf{R}^*(\epsilon) \geq \rho - \delta$, and $\mathsf{R}^*(\epsilon) \geq \rho$ follows from the arbitrariness of δ. □

VII. Lossless Lipschitz Decompression

In this section, we study the fundamental limit of lossless compression with Lipschitz decoders. To facilitate the discussion, we first introduce several important concepts from geometric measure theory. Then, we proceed to give proofs of Theorems 9–11.

A. Geometry Measure Theory

Geometric measure theory [42], [25] is an area of analysis studying the geometric properties of sets (typically in Euclidean spaces) through measure theoretic methods. One of the core concepts in this theory is *rectifiability*, a notion of smoothness or regularity of sets and measures. Basically a set is rectifiable if it is the image of a subset of a Euclidean space under some Lipschitz function. Rectifiable sets admit a smooth analog coding strategy. Therefore, lossless compression with Lipschitz decoders boils down to finding a subset of source realizations that is rectifiable and has high probability. In contrast, the goal of conventional almost-lossless data compression is to show concentration of probability on sets of small cardinality. This characterization enables us to use results from geometric measure theory to study Lipschitz coding schemes.

Definition 11 (Hausdorff Measure and Dimension): Let $s > 0$ and $A \subset \mathbb{R}^n$. Define

$$\mathcal{H}^s_\delta(A) = \inf\left\{\sum_i \mathrm{diam}(E_i)^s : A \subset \bigcup_i E_i, \mathrm{diam}(E_i) \leq \delta\right\} \tag{235}$$

where $\mathrm{diam}(E_i) = \sup\{\|x-y\|_2 : x,y \in E_i\}$. Define the *$s$-dimensional Hausdorff measure* on $\mathbb{R}^n$ by

$$\mathcal{H}^s(A) = \lim_{\delta \downarrow 0} \mathcal{H}^s_\delta(A). \tag{236}$$

The Hausdorff dimension of A is defined by

$$\dim_{\mathrm{H}}(A) = \inf\{s : \mathcal{H}^s(A) < \infty\}. \tag{237}$$

Hausdorff measure generalizes both the counting measure and Lebesgue measure and provides a nontrivial way to measure low-dimensional sets in a high-dimensional space. When $s = n$, $\mathcal{H}^n$ is just a rescaled version of the usual n-dimensional Lebesgue measure [25, 4.3]; when $s = 0$, $\mathcal{H}^0$ reduces

to the counting measure. For $0 < s < n$, $\mathcal{H}^s$ gives a nontrivial measure for sets of Hausdorff dimension s in $\mathbb{R}^n$, because if $\dim_{\mathrm{H}} A < s$, $\mathcal{H}^s(A) = 0$; if $\dim_{\mathrm{H}} A > s$, $\mathcal{H}^s(A) = \infty$. As an example, consider $n = 1$ and $s = \log_3 2$. Let C be the middle-third Cantor set in the unit interval, which has zero Lebesgue measure. Then, $\dim_{\mathrm{H}} C = s$ and $\mathcal{H}^s(C) = 1$ [18, 2.3].

Definition 12 (Rectifiable Sets [42, 3.2.14]): $E \subset \mathbb{R}^n$ is called *m-rectifiable* if there exists a Lipschitz mapping from some bounded set in $\mathbb{R}^m$ onto E.

Definition 13 (Rectifiable Measures [25, Definition 16.6]): Let μ be a measure on $\mathbb{R}^n$. μ is called *m-rectifiable* if $\mu \ll \mathcal{H}^m$ and there exists a μ-a.s. set $E \subset \mathbb{R}^n$ that is m-rectifiable.

Several useful facts about rectifiability are presented as follows.

Lemma 11 [42]:

1) An l-rectifiable set is also m-rectifiable for $m \geq l$.
2) The Cartesian product of an m-rectifiable set and an l-rectifiable set is $(m + l)$-rectifiable.
3) The finite union of m-rectifiable sets is m-rectifiable.
4) Countable sets are 0-rectifiable.

Using the notion of rectifiability, we give a sufficient condition for the ϵ-achievability of Lipschitz decompression by the following lemma.

Lemma 12: $\mathsf{R}(\epsilon) \leq R$ if there exists a sequence of $\lfloor Rn \rfloor$-rectifiable sets $S^n \subset \mathbb{R}^n$ with

$$\mathbb{P}\{X^n \in S^n\} \geq 1 - \epsilon \tag{238}$$

for all sufficiently large n.

Proof: See Appendix V. □

Definition 14 (k-Dimensional Density [25, Def. 6.8]): Let μ be a measure on $\mathbb{R}^n$. The k-dimensional *upper and lower densities* of μ at x are defined as

$$\overline{D}_k(\mu, x) = \limsup_{r \downarrow 0} \frac{\mu(B(x,r))}{r^k} \tag{239}$$

$$\underline{D}_k(\mu, x) = \liminf_{r \downarrow 0} \frac{\mu(B(x,r))}{r^k}. \tag{240}$$

If $\overline{D}_k(\mu, x) = \underline{D}_k(\mu, x)$, the common value is called the k-dimensional density of μ at x, denoted by $D_k(\mu, x)$.

The following important result in geometric measure theory gives a density characterization of rectifiability for Borel measures.

Theorem 19 (Preiss Theorem [43, Th. 5.6]): A σ-finite Borel measure on $\mathbb{R}^n$ is m-rectifiable if and only if $0 < D_m(\mu, x) < \infty$ for μ-a.e. $x \in \mathbb{R}^n$.

Recalling the expression for information dimension $d(P_X)$ in (17), we see that for the information dimension of a measure to be equal to m it requires that the exponent of the *average* measure of ϵ-balls equals m, whereas m-rectifiability of a measure requires that the measure of *almost every* ϵ-ball scales as $\mathrm{O}(\epsilon^m)$, a much stronger condition than the existence of information dimension. Obviously, if a probability measure μ is m-rectifiable, then $d(\mu) = m$.

B. Converse

In view of the lossless Minkowski dimension compression results developed in Section V, the general converse in Theorem 9 is rather straightforward. We need the following lemma to complete the proof.

Lemma 13: Let $S \subset \mathbb{R}^n$ be k-rectifiable. Then

$$k \geq \overline{\dim}_{\mathrm{B}} S. \tag{241}$$

Proof: See Appendix IX. □

Proof of Theorem 9: Lemma 13 implies the following general inequality:

$$\mathsf{R}(\epsilon) \geq \mathsf{R}_{\mathsf{B}}(\epsilon). \tag{242}$$

If the source is memoryless and $\bar{d}(X) < \infty$, then it follows from Theorem 14 that $\mathsf{R}(\epsilon) \geq \bar{d}(X)$. □

C. Achievability for Finite Mixture

We first prove a general achievability result for finite mixtures, a corollary of which applies to discrete-continuous mixed distributions in Theorem 10.

Theorem 20 (Achievability of Finite Mixtures): Let the distribution μ of X be a mixture of finitely many Borel probability measures on $\mathbb{R}$, i.e.,

$$\mu = \sum_{i=1}^{N} \rho_i \mu_i \tag{243}$$

where $\{\rho_1, \ldots, \rho_N\}$ is a probability mass function. If R_i is ϵ_i-achievable with Lipschitz decoders for μ_i, $i = 1, \ldots, N$, then R is ϵ-achievable for μ with Lipschitz decoders, where

$$\epsilon = \sum_{i=1}^{N} \epsilon_i \tag{244}$$

$$R = \sum_{i=1}^{N} \rho_i R_i. \tag{245}$$

Proof: By induction, it is sufficient to show the result for $N = 2$. Denote $\rho_1 = \rho = 1 - \rho_2$. Let $\{W_i\}$ be a sequence of i.i.d. binary random variables with $\mathbb{P}\{W_i = 1\} = \rho$. Let $\{X_i\}$ be a i.i.d. sequence of real-valued random variables, such that the distribution of each X_i conditioned on the events $\{W_i = 1\}$ and $\{W_i = 0\}$ are μ_1 and μ_2 respectively. Then, $\{X_i : i \in \mathbb{N}\}$ is a memoryless process with common distribution μ. Since the claim of the theorem depends only on the *probability law* of the source, we base our calculation of block error probability on this specific construction.

Fix $\delta > 0$. Since R_1 and R_2 are achievable for μ_1 and μ_2 respectively, by Lemma 12, there exists N_1 such that for all

$n > N_1$, there exists $S_1^n, S_2^n \subset \mathbb{R}^n$, with $\mu_1^n(S_1^n) \geq 1-\epsilon_1$, $\mu_2^n(S_2^n) \geq 1-\epsilon_2$ and S_1^n is $\lfloor R_1 n \rfloor$-rectifiable and S_2^n is $\lfloor R_2 n \rfloor$-rectifiable. Let

$$\mathcal{W}_n = \{w^n \in \mathbb{Z}_2^n : (\rho-\delta)n \leq \|w^n\|_0 \leq (\rho+\delta)n\}. \quad (246)$$

By WLLN, $\mathbb{P}\{W^n \in \mathcal{W}_n\} \to 1$. Hence, for any $\epsilon' > 0$, there exists N_2, such that for all $n > N_2$

$$\mathbb{P}\{W^n \in \mathcal{W}_n\} \geq 1-\epsilon'. \quad (247)$$

Let

$$n > \max\left\{\frac{N_1}{\rho-\delta}, \frac{N_1}{1-\rho-\delta}, N_2\right\}. \quad (248)$$

Define

$$S^n = \{x^n \in \mathbb{R}^n : \exists m, T, (\rho-\delta)n \leq m \leq (\rho+\delta)n, \\ T \subset \{1,\ldots,n\}, |T| = m, x_T \in S_1^m, x_{T^c} \in S_2^{n-m}\}. \quad (249)$$

Next we show that S^n is $\lfloor (R+2\delta)n \rfloor$-rectifiable. For all $(\rho-\delta)n \leq m \leq (\rho+\delta)n$, it follows from (248) that

$$N_1 \leq (\rho-\delta)n \leq m \quad (250)$$

$$N_1 \leq (1-\rho-\delta)n \leq n-m. \quad (251)$$

and

$$R_1 m + R_2(n-m) \leq R_1(\rho+\delta)n + R_2(1-\rho+\delta)n \quad (252)$$

$$= ((R_1+R_2)\delta + R)n \quad (253)$$

$$\leq (2\delta+R)n \quad (254)$$

where $R = \rho R_1 + (1-\rho)R_2$ according to (245). Observe that S^n is a finite union of subsets, each of which is a Cartesian product of a $\lfloor R_1 m \rfloor$-rectifiable set in $\mathbb{R}^m$ and a $\lfloor R_2(n-m) \rfloor$-rectifiable set in $\mathbb{R}^{n-m}$. Recalling Lemma 11, S^n is $\lfloor (R+2\delta)n \rfloor$-rectifiable, in view of (254).

Now we calculate the measure of S^n under μ^n

$$\mathbb{P}\{X^n \in S^n\}$$

$$\geq \mathbb{P}\{X^n \in S^n, W^n \in \mathcal{W}_n\} \quad (255)$$

$$= \sum_{m=\lceil(\rho-\delta)n\rceil}^{\lfloor(\rho+\delta)n\rfloor} \sum_{\|w^n\|_0=m} \mathbb{P}\{W^n = w^n\} \times \mathbb{P}\{X^n \in S^n \mid W^n = w^n\} \quad (256)$$

$$\geq \sum_{m=\lceil(\rho-\delta)n\rceil}^{\lfloor(\rho+\delta)n\rfloor} \sum_{|T|=m} \sum_{\mathrm{supp}(w^n)=T} \mathbb{P}\{W^n = w^n\} \times \mathbb{P}\{X_T \in S_1^m, X_{T^c} \in S_2^{n-m} \mid W^n = w^n\} \quad (257)$$

$$= \sum_{m=\lceil(\rho-\delta)n\rceil}^{\lfloor(\rho+\delta)n\rfloor} \sum_{\|w^n\|_0=m} \mathbb{P}\{W^n = w^n\} \times \mu_1^m(S_1^m)\mu_2^{n-m}(S_2^{n-m}) \quad (258)$$

$$\geq \sum_{m=\lceil(\rho-\delta)n\rceil}^{\lfloor(\rho+\delta)n\rfloor} \sum_{\|w^n\|_0=m} (1-\epsilon_1)(1-\epsilon_2)\mathbb{P}\{W^n = w^n\} \quad (259)$$

$$\geq (1-\epsilon)\,\mathbb{P}\{W^n \in \mathcal{W}_n\} \quad (260)$$

$$\geq 1-\epsilon-\epsilon' \quad (261)$$

where:

- (257): by construction of S_n in (249);
- (259): by (250) and (251);
- (260): $\epsilon = \epsilon_1 + \epsilon_2$ according to (244);
- (261): by (247).

In view of Lemma 12, $R+2\delta$ is $\epsilon+\epsilon'$-achievable for X. By the arbitrariness of δ and ϵ', R is ϵ-achievable for X. □

Proof of Theorem 10: Let the distribution of X be $\mu = (1-\rho)\mu_{\mathrm{d}} + \rho\mu_{\mathrm{c}}$ as defined in (139), where μ_{d} is discrete and μ_{c} is absolutely continuous. By Lemma 11, countable sets are 0-rectifiable. For any $0 < \epsilon < 1$, there exists $A > 0$ such that $\mu_{\mathrm{c}}([-A, A]) > 1-\epsilon$. By definition, $[-A, A]$ is 1-rectifiable. Therefore, by Lemma 12 and Theorem 20, ρ is an ϵ-achievable rate for X. The converse follows from (242) and Theorem 15. □

D. Achievability for Singular Distributions

In this section, we prove Theorem 11 for memoryless sources, using isomorphism results in ergodic theory. The proof outline is as follows: a classical result in ergodic theory states that Bernoulli shifts are isomorphic if they have the same entropy. Moreover, the homomorphism can be chosen to be *finitary*, that is, each coordinate only depends on finitely many coordinates. This finitary homomorphism naturally induces a Lipschitz decoder in our setup; however, the caveat is that the Lipschitz continuity is with respect to an *ultrametric* (Definition 15) that is *not* equivalent to the usual Euclidean distance. Nonetheless, by an arbitrarily small increase in the compression rate, the decoder can be modified to be Lipschitz with respect to the Euclidean distance. Before proceeding to the proof, we first present some necessary results of ultrametric spaces and finitary coding in ergodic theory.

Definition 15: Let (X, d) be a metric space. d is called an *ultrametric* if

$$d(x,z) \leq \max\{d(x,y), d(y,z)\} \quad (262)$$

for all $x, y, z \in X$.

A canonical class of ultrametric spaces is the ultrametric Cantor space [44]: let $\mathcal{X} = \mathbb{Z}_M^{\mathbb{Z}_+}$ denote the set of all one-sided M-ary sequences $x = (x_0, \ldots)$. To endow $\mathcal{X}$ with an ultrametric, define

$$d_\alpha(x,y) = \begin{cases} 0, & x = y \\ \alpha^{-\min\{n \in \mathbb{N} : x_n \neq y_n\}}, & x \neq y. \end{cases} \quad (263)$$

Then, for every $\alpha > 1$, d_α is an ultrametric on $\mathcal{X}$. In a similar fashion, we define an ultrametric on $[0,1]^k$ by considering the M-ary expansion of real vectors. Similar to the binary expansion defined in (6), for $x^k \in [0,1]^k$, $i \in \mathbb{Z}_+$, $M \in \mathbb{N}$ and $M \geq 2$, define

$$(x^k)_{M,i} = \lfloor M^i x^k \rfloor - M\lfloor M^{i-1} x^k \rfloor \in \mathbb{Z}_M^k \quad (264)$$

then

$$x^k = \sum_{i \in \mathbb{Z}_+} (x^k)_{M,i} M^{-i}. \quad (265)$$

Denoting for brevity

$$(x^k) = ((x^k)_{M,0}, (x^k)_{M,1}, \ldots) \quad (266)$$

(263) induces an ultrametric on $[0,1]^k$

$$\hat{d}(x^k, y^k) = d_M((x^k), (y^k)). \tag{267}$$

It is important to note that $\hat{d}$ is *not* equivalent to the ℓ_∞ distance (or any ℓ_p distance), since we only have

$$\hat{d}(x^k, y^k) \geq \frac{1}{M}\|x^k - y^k\|_\infty. \tag{268}$$

To see the impossibility of the other direction of (268), consider $x = 1/M$ and $y = \sum_{i=2}^{l} M^{-i}$. As $l \to \infty$, $|x - y| \to 0$ but $\hat{d}(x, y)$ remains $1/M$. Therefore, a Lipschitz function with respect to $\hat{d}$ is *not* necessarily Lipschitz under $\|\cdot\|_\infty$. However, the following lemma bridges the gap if the dimension of the domain and the Lipschitz constant are allowed to increase.

Lemma 14: Let $\hat{d}$ be the ultrametric on $[0,1]^k$ defined in (267). $W \subset [0,1]^k$ and $g : (W, \hat{d}) \to (\mathbb{R}^n, \|\cdot\|_\infty)$ is Lipschitz. Then, there exists $W' \subset [0,1]^{k+1}$ and $g' : (W', \|\cdot\|_\infty) \to (\mathbb{R}^n, \|\cdot\|_\infty)$ such that $g(W) = g'(W')$ and g' is Lipschitz.

Proof: See Appendix X. □

Next we recall several results on finitary coding of Bernoulli shifts. Kolmogorov–Ornstein theory studies whether two processes with the same entropy rate are isomorphic. Kean and Smorodinsky [45] showed that two *double-sided* Bernoulli shifts of the same entropy are *finitarily isomorphic*. For the single-sided case, Del Junco [46] showed that there is a *finitary homomorphism* between two *single-sided* Bernoulli shifts of the same entropy, which is a finitary improvement of Sinai's theorem [47], [55]. We will see how a finitary homomorphism of the digits is related to a real-valued Lipschitz function, and how to apply Del Junco's ergodic-theoretic result to our problem.

Definition 16 (Finitary Homomorphisms): Let $\mathcal{C}$ and $\mathcal{D}$ be finite sets. Let σ and τ denote the left shift operators on the product spaces $\mathcal{X} = \mathcal{C}^{\mathbb{Z}_+}$ and $\mathcal{Y} = \mathcal{D}^{\mathbb{Z}_+}$ respectively. Let μ and ν be measures on $\mathcal{X}$ and $\mathcal{Y}$ (with product σ-algebras). A *homomorphism* $\phi : (\mathcal{X}, \mu, \sigma) \to (\mathcal{Y}, \nu, \tau)$ is a measure preserving mapping that commutes with the shift operator, i.e., $\nu = \mu \circ \phi^{-1}$ and $\phi \circ \sigma = \tau \circ \phi$ μ-a.e. ϕ is said to be *finitary* if there exist sets of zero measure $A \subset \mathcal{X}$ and $B \subset \mathcal{Y}$ such that $\phi : \mathcal{X} \backslash A \to \mathcal{Y} \backslash B$ is continuous (with respect to the product topology).

Informally, finitariness means that for almost every $x \in \mathcal{X}$, $\phi(x)_0$ is determined by finitely many coordinates in x. The following lemma characterizes this intuition in precise terms.

Lemma 15 [48, Conditions 5.1, p. 281]: Let $\mathcal{X} = \mathcal{C}^{\mathbb{Z}_+}$ and $\mathcal{Y} = \mathcal{D}^{\mathbb{Z}_+}$. Let $\phi : (\mathcal{X}, \mu, \sigma) \to (\mathcal{Y}, \nu, \tau)$ be a homomorphism. Then, the following statements are equivalent.

1) ϕ is finitary.
2) For μ-a.e. $x \in \mathcal{X}$, there exists $j(x) \in \mathbb{N}$, such that for any $x' \in \mathcal{X}$, $(x')_0^{j(x)} = (x)_0^{j(x)}$ implies that $(\phi(x))_0 = (\phi(x'))_0$.
3) For each $j \in \mathcal{D}$, the inverse image $\phi^{-1}\{y \in \mathcal{Y} : y_0 = j\}$ of each time-0 cylinder set in $\mathcal{Y}$ is, up to a set of measure 0, a countable union of cylinder sets in $\mathcal{X}$.

Theorem 21 [46, Th. 1]: Let P and Q be probability distributions on finite sets $\mathcal{C}$ and $\mathcal{D}$. Let $(\mathcal{X}, \mu) = (\mathcal{C}^{\mathbb{Z}_+}, P^{\mathbb{Z}_+})$ and $(\mathcal{Y}, \nu) = (\mathcal{D}^{\mathbb{Z}_+}, Q^{\mathbb{Z}_+})$. If P and Q each have at least three non-zero components and $H(P) = H(Q)$, then there is a finitary homomorphism $\phi : (\mathcal{X}, \mu) \to (\mathcal{Y}, \nu)$.

We now use Lemmas 14–15 and Theorem 21 to prove Theorem 11.

Proof of Theorem 11: Without loss of generality, assume that the random variable satisfies $0 \leq X \leq 1$. Denote by P the common distribution of the M-ary digits of X. By Proposition 2

$$d(X) = \frac{H(P)}{\log M}. \tag{269}$$

Fix n. Let $d = d(X)$, $k = \lceil dn \rceil$ and $\mathcal{C} = \mathbb{Z}_M^n$. Let Q be a probability measure on $\mathcal{D} = \mathbb{Z}_M^k$ such that $H(P^n) = H(Q)$. Such a Q always exists because $\log|\mathcal{D}| = k \log M \geq nH(P)$. Let $\mu = (P^n)^{\mathbb{Z}_+}$ and $\nu = Q^{\mathbb{Z}_+}$ denote the product measure on $\mathcal{C}^{\mathbb{Z}_+}$ and $\mathcal{D}^{\mathbb{Z}_+}$, respectively. Since μ and ν has the same entropy rate, by Theorem 21, there exists a finitary homomorphism $\phi : (\mathcal{D}^{\mathbb{Z}_+}, \nu, \sigma) \to (\mathcal{C}^{\mathbb{Z}_+}, \mu, \tau)$. By the characterization of finitariness in Lemma 15, for any $u \in \mathcal{D}^{\mathbb{Z}_+}$, there exists $j(u) \in \mathbb{N}$ such that $\phi(u)_0$ is determined only by $u_0, \ldots, u_{j(u)}$. Denote the closed ultrametric ball

$$B_u = \{v \in \mathcal{D}^{\mathbb{Z}_+} : \hat{d}(u, v) \leq M^{-(j(u)+1)}\} \tag{270}$$

where $\hat{d}$ is defined in (267). Then, for any $v \in B_u$, $\phi(u)_0 = \phi(v)_0$. Note that $\{B_u : u \in \mathcal{D}^{\mathbb{Z}_+}\}$ forms a *countable* cover of $\mathcal{D}^{\mathbb{Z}_+}$. This is because B_u is just a cylinder set in $\mathcal{D}^{\mathbb{Z}_+}$ with base $(u)_0^{j(u)}$, and the total number of cylinders is countable. Furthermore, since intersecting ultrametric balls are contained in each other [49], there exists a sequence $\{u^{(i)}\}$ in $\mathcal{D}^{\mathbb{Z}_+}$, such that $\bigcup_{i \in \mathbb{N}} B_{u^{(i)}}$ *partitions* $\mathcal{D}^{\mathbb{Z}_+}$. Therefore, for all $0 < \epsilon < 1$, there exists N, such that $\nu(E) \geq 1 - \epsilon$, where

$$E = \bigcup_{i=1}^{N} B_{u^{(i)}}. \tag{271}$$

For $x \in [0,1]^k$, recall the M-ary expansion of x defined in (266), denoted by $(x) \in \mathcal{D}^{\mathbb{Z}_+}$. Let

$$F = \phi(E) \subset \mathcal{C}^{\mathbb{Z}_+} \tag{272}$$

$$W = \{x \in [0,1]^k : (x) \in E\} \tag{273}$$

$$S = \{x \in [0,1]^n : (x) \in F\}. \tag{274}$$

Since ϕ is measure preserving, $\mu = \nu \circ \phi^{-1}$, therefore

$$\mathbb{P}\{X^n \in S\} = \mu(F) = \nu(\phi^{-1}(F)) \geq \nu(E) \geq 1 - \epsilon. \tag{275}$$

Next we use ϕ to construct a real-valued Lipschitz mapping g. Define $g : [0,1]^k \to [0,1]^n$ by

$$g(x) = \sum_{i \in \mathbb{Z}_+} \phi((x))_i M^{-i}. \tag{276}$$

Since ϕ commutes with the shift operator, for all $z \in \mathcal{D}^{\mathbb{Z}_+}$, $\phi(z)_i = (\tau^i \phi(z))_0 = \phi(\tau^i z)_0$. Also, for $x \in [0,1]^k$, $\tau^i(x) = (M^i x)$. Therefore

$$g(x) = \sum_{i \in \mathbb{Z}_+} \phi((M^i x))_0 M^{-i}. \tag{277}$$

Next we proceed to show that $g:(W,\hat{d})\to([0,1]^n,\|\cdot\|_\infty)$ is Lipschitz. In view of (268) and (263), it is sufficient to show that $\phi:(E,d_M)\to(\mathcal{C}^{\mathbb{Z}_+},d_M)$ is Lipschitz. Let

$$J=\max_{1\le i\le N} j(u^{(i)}). \tag{278}$$

First observe that ϕ is M^J-Lipschitz on each ultrametric ball $B_{u^{(i)}}$ in E. To see this, consider distinct points $v,w\in B_{u^{(i)}}$. Let $d_M(v,w)=M^{-m}$. Then, $m\ge j(u^{(i)})+1$. Since $(v)_0^{m-1}=(w)_0^{m-1}$, $\phi(v)$ and $\phi(w)$ coincide on their first $m-j(u^{(i)})-1$ digits. Therefore

$$\begin{aligned} d_M(\phi(v),\phi(w)) &\le M^{-m+j(u^{(i)})} && (279)\\ &\le M^{j(u^{(i)})}d_M(v,w) && (280)\\ &\le M^J d_M(v,w). && (281)\end{aligned}$$

Since every closed ultrametric ball is also open [49, Prop. 18.4], $B_{u^{(1)}},\dots,B_{u^{(N)}}$ are disjoint, therefore ϕ is L-Lipschitz on E for some $L>0$. Then, for any $y,z\in W$

$$\begin{aligned} \|g(y)-g(z)\|_\infty &\le M\hat{d}(g(y),g(z)) && (282)\\ &= Md_M(\phi((y)),\phi((z))) && (283)\\ &\le ML\,d_M((y),(z)) && (284)\\ &= ML\,\hat{d}(y,z) && (285)\end{aligned}$$

where:

- (282): by (268);
- (283) and (285): by (267);
- (284): by (281).

Hence, $g:(W,\hat{d})\to([0,1]^n,\|\cdot\|_\infty)$ is Lipschitz.

By Lemma 14, there exists a subset $W'\subset[0,1]^{k+1}$ and a Lipschitz mapping $g':(W',\|\cdot\|_\infty)\to(\mathbb{R}^n,\|\cdot\|_\infty)$ such that $g'(W')=g(W)=S$. By Kirszbraun's theorem [42, 2.10.43], we extend g' to a Lipschitz function $g_n:[0,1]^{k+1}\to\mathbb{R}^n$. Then, $S=g_n(W)$. Since g_n is continuous and $[0,1]^{k+1}$ is compact, by Lemma 18, there exists a Borel function $f_n:\mathbb{R}^n\to[0,1]^{k+1}$, such that $g_n=f_n^{-1}$ on S.

To summarize, we have obtained a Borel function $f_n:\mathbb{R}^n\to[0,1]^{k+1}$ and a Lipschitz function $g_n:[0,1]^{k+1}\to\mathbb{R}^n$, where $k=\lceil dn\rceil$, such that $\mathbb{P}\{g_n(f_n(X^n))=X^n\}\ge\mathbb{P}\{X^n\in S\}\ge 1-\epsilon$. Therefore, we conclude that $\mathsf{R}(\epsilon)\le d$. The converse follows from Theorem 9.

Last, we show that

$$\mathsf{R}(0)=d=\log_M m \tag{286}$$

for the special case when P is equiprobable on the support, where $m=|\mathrm{supp}(P)|$. Recalling the construction of self-similar measures in Section III-C, we first note that the distribution of X is a self-similar measure that is generated by the IFS $(S_0,\dots,S_{M-1})$, where

$$S_i(x)=\frac{x+i}{M},\qquad i=0,\dots,M-1. \tag{287}$$

This IFS satisfies the open set condition, since $(0,1)\supset\bigcup_{i=0}^{M-1}S_i((0,1))$ and the union is disjoint. Denote by $E\subset[0,1]$ the invariant set of the reduced IFS $\{S_i:i\in\mathrm{supp}(P)\}$. By [12, Corollary 4.1], the distribution of X, denoted by μ_X, is in fact the normalized d-dimensional Hausdorff measure on E, i.e., $\mu_X(\cdot)=\mathcal{H}^d(\cdot\cap E)/\mathcal{H}^d(E)$. Therefore, $\mu_{X^n}(\cdot)=\mathcal{H}^{dn}(\cdot\cap E^n)/\mathcal{H}^{dn}(E^n)$. By [18, Exercise 9.11], there exists a constant $c_1>0$, such that for all $x^n\in E^n$

$$\underline{D}_{dn}(\mathcal{H}^{dn},x^n)\ge c_1 \tag{288}$$

that is, E^n has positive lower dn-density everywhere. By [50, Th. 4.1(1)], for any $k>dn$, there exists $F\subset E^n$ such that $\mathcal{H}^{dn}(F)=0$ and $E^n\backslash F$ is k-rectifiable. Therefore, $\mathbb{P}\{X^n\in F\}=0$. By Lemma 12, the rate d is 0-achievable. □

VIII. Concluding Remarks

Compressed sensing, as an analog compression paradigm, imposes two basic requirements: the linearity of the encoder and the robustness of the decoder; the rationale is that low complexity of encoding operations and noise resilience of decoding operations are indispensable in dealing with analog sources. To better understand the fundamental limits imposed by the requirements of low complexity and noise resilience, it is pedagogically sound to study them separately and in a more general paradigm than compressed sensing. Motivated by this observation, in this paper we have proposed an information theoretic framework for lossless analog compression of analog sources under regularity conditions of the coding schemes. Abstractly, the approach boils down to probabilistic dimension reduction with smooth embedding. In this framework, obtaining fundamental limits requires tools quite different from those used in traditional information theory, calling for machineries from dimension theory and geometric measure theory in addition to ergodic theory.

Within this general framework, we analyzed the fundamental limits under different regularity constraints imposed on compressor and decompressor. Perhaps the most surprising result is, as shown in (75)

$$\mathsf{R}^*(\epsilon)\le\mathsf{R}(\epsilon) \tag{289}$$

which holds for any real-valued source. This conclusion implies that a Lipschitz constraint at the decompressor results in less efficient compression than a linearity constraint at the compressor. For memoryless sources, we have also obtained bounds or exact expressions for various ϵ-achievable rates. As seen in Theorems 5–12, Rényi's information dimension plays an important role in the associated coding theorems. These results provide new operational characterizations for Rényi's information dimension in a lossless compression framework.

In the important case of discrete-continuous mixed sources, which is a probabilistic generalization of the linearly sparse source model used in compressed sensing (a fixed fraction of observations are zero), we have shown that the fundamental limit is Rényi information dimension, which coincides with the weight on the continuous part in the source distribution. In the memoryless case, this corresponds to the fraction of analog symbols in the source realization. This might suggest that the mixed discrete-continuous nature of the source is of fundamental importance in the analog compression framework; sparsity is just one manifestation of a mixed distribution.

It should be remarked that random linear coding is not only an important achievability proof technique in Shannon theory, but also an inspiration to obtain efficient schemes in modern coding

theory, compressed sensing as well as our analog compression framework. Moreover, it also provides information about how close practical encoders are to the fundamental limit. For instance, in lossless linear compression, the achievability bound on $\mathsf{R}^*(\epsilon)$ in Theorem 18 can be achieved with Lebesgue-a.e. linear encoders. This implies that generating random linear encoders using any continuous distribution achieves the desired error probability almost surely.

As far as future research directions, there are regularity conditions beyond those in Table I that are worth studying. For example, it is interesting to investigate the fundamental limit of *bilipshitz* coding schemes, i.e., the encoder and decoder both being Lipschitz continuous. This is a probabilistic version of bilipschitz embedding in Euclidean spaces, e.g., Dvoretzky's theorem [51] and Johnson–Lindenstrauss lemma [52]. As a more restricted case, the fundamental limit of linear compression with Lipschitz decompression is the most desirable result.

Appendix I
Proof of Equivalence of (17) and (14)

Proposition 4: The information dimension of a random variable X on $\mathbb{R}^n$ can be calculated as follows:

$$d(X) = \lim_{\epsilon \downarrow 0} \frac{\mathbb{E} \log \mu(B_p(X, \epsilon))}{\log \epsilon} \tag{290}$$

where μ is the distribution of X and $B_p(x, \epsilon)$ is the ℓ_p-ball of radius ϵ centered at x ($1 \le p \le \infty$). The lower (upper) information dimension can be obtained by replacing $\lim$ by $\liminf$ ($\limsup$).

Proof: Let X be a random vector in $\mathbb{R}^n$ and denote its distribution by μ. Due to the equivalence of ℓ_p-norms, it is sufficient to show (290) for $p = \infty$. Recall the notation in (5) and note that

$$H([X]_m) = \mathbb{E} \log \frac{1}{\mu(C_m(2^m [X]_m))}. \tag{291}$$

For any $0 < \epsilon < 1$, there exists $m \in \mathbb{N}$, such that $2^{-m} \le \epsilon < 2^{m-1}$. Then

$$C_m([x]_m 2^m) \subset B_\infty(x, 2^{-m}) \subset B_\infty(x, \epsilon) \tag{292}$$

$$\subset B_\infty(x, 2^{-(m-1)}) \subset B_\infty([x]_m, 2^{-(m-2)}). \tag{293}$$

As a result of (292), we have

$$\mathbb{E} \log \frac{1}{\mu(B_\infty(X, \epsilon))} \le H([X]_m). \tag{294}$$

On the other hand, note that $B_\infty([x]_m, 2^{-(m-2)})$ is a disjoint union of 8^n mesh cubes. By (293), we have

$$\mathbb{E} \log \frac{1}{\mu(B_\infty(X, \epsilon))} \ge \mathbb{E} \log \frac{1}{\mu(B_\infty([X]_m, 2^{-(m-2)}))} \tag{295}$$

$$\ge H([X]_m) - n \log 8. \tag{296}$$

Combining (294) and (296) yields

$$\frac{H([X]_m) - n \log 8}{m} \le \frac{\mathbb{E} \log \frac{1}{\mu(B_\infty(X, \epsilon))}}{\log \frac{1}{\epsilon}} \le \frac{H([X]_m)}{m - 1}. \tag{297}$$

By Proposition 2, sending $\epsilon \downarrow 0$ and $m \to \infty$ yields (290). □

Appendix II
Proofs of Propositions 1–3

Lemma 16: For all $p, q \in \mathbb{N}$

$$H(\langle X \rangle_p) \le H(\langle X \rangle_q) + \log \left(\left\lceil \frac{p}{q} \right\rceil + 1 \right). \tag{298}$$

Proof:

$$H(\langle X \rangle_p) \le H(\langle X \rangle_p, \langle X \rangle_q) \tag{299}$$

$$= H(\langle X \rangle_p | \langle X \rangle_q) + H(\langle X \rangle_q). \tag{300}$$

Note that for any $l \in \mathbb{Z}$

$$H\left(\langle X \rangle_p \middle| \langle X \rangle_q = \frac{l}{q} \right) \tag{301}$$

$$= H\left(\langle X \rangle_p \middle| \frac{l}{q} \le \frac{X}{q} < \frac{l+1}{q} \right) \tag{302}$$

$$= H\left(\lfloor pX \rfloor \middle| \frac{pl}{q} \le pX < \frac{p(l+1)}{q} \right). \tag{303}$$

Given $pX \in [\frac{pl}{q}, \frac{p(l+1)}{q})$, the range of $\lfloor pX \rfloor$ is upper bounded by $\lceil \frac{p}{q} \rceil + 1$. Therefore, for all $l \in \mathbb{Z}$

$$H\left(\langle X \rangle_p \middle| \langle X \rangle_q = \frac{l}{q} \right) \le \log \left(\left\lceil \frac{p}{q} \right\rceil + 1 \right). \tag{304}$$

Hence, $H(\langle X \rangle_p | \langle X \rangle_q)$ admits the same upper bound and (298) holds. □

Lemma 17 [53, p. 2102]: Let W be an $\mathbb{N}$-valued random variable. Then, $H(W) < \infty$ if $\mathbb{E} \log W < \infty$.

Proof of Proposition 1: Using Lemma 16 with $p = n, q = 1$ and $p = 1, q = n$, we have

$$H(\lfloor X \rfloor) - \log 2 \le H(\langle X \rangle_n) \le H(\lfloor X \rfloor) + \log n. \tag{305}$$

Equation (21) ⇒ (20): When $H(\lfloor X \rfloor)$ is finite, dividing both sides of (305) by $\log n$ and letting $n \to \infty$ results in (20).

Equation (20) ⇒ (21): Suppose $H(\lfloor X \rfloor) = \infty$. By (305), $H(\langle X \rangle_n) = \infty$ for every n and (20) fails. This also proves (22).

(19) ⇒ (21)

$$\mathbb{E} \log(|X| + 1) < \infty \tag{306}$$

$$\Rightarrow \mathbb{E} \log(\lfloor |X| \rfloor + 1) < \infty \tag{307}$$

$$\Rightarrow H(\lfloor |X| \rfloor + 1) < \infty \tag{308}$$

$$\Rightarrow H(\lfloor |X| \rfloor) < \infty \tag{309}$$

$$\Rightarrow H(\lfloor X \rfloor) < \infty \tag{310}$$

where:

- (308): by Lemma 17;
- (310): by

$$H(\lfloor X \rfloor) \le H(\lfloor |X| \rfloor) + H(\lfloor X \rfloor | \lfloor |X| \rfloor) \tag{311}$$

$$\le H(\lfloor |X| \rfloor) + \log 2. \tag{312}$$

□

Proof of Proposition 2: Fix any $m \in \mathbb{N}$ and $l \in \mathbb{N}$, such that $2^{l-1} \le m < 2^l$. By Lemma 16, we have

$$H([X]_{l-1}) \le H(\langle X \rangle_m) + \log 3, \tag{313}$$

$$H(\langle X \rangle_m) \le H([X]_l) + \log 3. \tag{314}$$

Therefore

$$\frac{H([X]_{l-1})-\log 3}{l}\le\frac{H(\langle X\rangle_m)}{\log m}\le\frac{H([X]_l)+\log 3}{l-1} \quad (315)$$

and hence (23) and (24) follow. □

Proof of Proposition 3: Note that

$$H(\langle X\rangle_m)\le H\left(\frac{\lceil mX\rceil}{m}\right)+H\left(\langle X\rangle_m\middle|\frac{\lceil mX\rceil}{m}\right) \quad (316)$$

$$\le H\left(\frac{\lceil mX\rceil}{m}\right)+\log 2 \quad (317)$$

$$H\left(\frac{\lceil mX\rceil}{m}\right)\le H(\langle X\rangle_m)+H\left(\frac{\lceil mX\rceil}{m}\middle|\langle X\rangle_m\right) \quad (318)$$

$$\le H(\langle X\rangle_m)+\log 2. \quad (319)$$

The same bound works for rounding. □

Appendix III
Information Dimension and Rate-Distortion Theory

The asymptotic tightness of the Shannon lower bound in the high-rate regime is shown by the following result.

Theorem 22 [20]: Let X be a random variable on the normed space $(\mathbb{R}^k,\|\cdot\|)$ with a density such that $h(X)>-\infty$. Let the distortion function be $\rho(x,y)=\|x-y\|^r$ with $r>0$, and the single-letter rate-distortion function is given by

$$R(D)=\inf_{\mathbb{E}\|Y-X\|^r\le D} I(X;Y). \quad (320)$$

Suppose that there exists an $\alpha>0$ such that $\mathbb{E}\|X\|^\alpha<\infty$. Then, $R(D)\ge R_{\rm L}(D)$ and

$$\lim_{D\to 0}(R(D)-R_{\rm L}(D))=0 \quad (321)$$

where the Shannon lower bound $R_{\rm L}(D)$ takes the following form:

$$R_{\rm L}(D)=\frac{k}{r}\log\frac{1}{D}+h(X)+\frac{k}{r}\log\frac{k}{r\mathrm{e}}-\frac{k}{r}\Gamma\left(\frac{k}{r}\right)\log V_k \quad (322)$$

where V_k denotes the volume of the unit ball $B_k=\{x\in\mathbb{R}^k:\|x\|\le 1\}$.

Special cases of Theorem 22 include the following.

- MSE distortion and scalar source: $\rho(x,y)=(x-y)^2$, $\|x\|=|x|$, $r=2$, $k=1$ and $V_k=2$. Then, the Shannon lower bound takes the familiar form

$$R_{\rm L}(D)=h(X)-\frac{1}{2}\log(2\pi\mathrm{e}D). \quad (323)$$

- Absolute distortion and scalar source: $\rho(x,y)=|x-y|$, $\|x\|=|x|$, $r=1$, $k=1$ and $V_k=2$. Then

$$R_{\rm L}(D)=h(X)-\log(2\mathrm{e}D). \quad (324)$$

- ℓ_∞-norm and vector source: $\rho(x,y)=\|x-y\|_\infty$, $\|x\|=\|x\|_\infty$, $r=1$ and $V_k=2^k$. Then

$$R_{\rm L}(D)=h(X)+\log\left(\frac{1}{k!}\left(\frac{k}{2\mathrm{e}D}\right)^k\right). \quad (325)$$

For general sources, Kawabata and Dembo introduced the concept of *rate-distortion dimension* in [12]. The rate-distortion dimension of a measure μ (or a random variable X with the distribution μ) on the metric space $(\mathbb{R}^k,d)$ is defined as follows:

$$\overline{\dim}_{\rm R}(X)=r\limsup_{D\downarrow 0}\frac{R_r(D)}{\log\frac{1}{D}} \quad (326)$$

$$\underline{\dim}_{\rm R}(X)=r\liminf_{D\downarrow 0}\frac{R_r(D)}{\log\frac{1}{D}} \quad (327)$$

where $R_r(D)$ is the single-letter rate-distortion function of X with distortion function $\rho(x,y)=d(x,y)^r$ $(r>0)$. Then, under the equivalence of the metric and the ℓ_∞-norm as in (328), the rate-distortion dimension coincides with the information dimension of X.

Theorem 23 [12, Prop. 3.3]: Consider the metric space $(\mathbb{R}^k,d)$. If there exists $a_1,a_2>0$, such that for all $x,y\in\mathbb{R}^k$

$$a_2\|x-y\|_\infty\le d(x,y)\le a_1\|x-y\|_\infty \quad (328)$$

then

$$\overline{\dim}_{\rm R}(X)=\bar{d}(X) \quad (329)$$

$$\underline{\dim}_{\rm R}(X)=\underline{d}(X). \quad (330)$$

Moreover, (329) and (330) hold even if the ϵ-entropy $H_\epsilon(X)$ instead of the rate-distortion function $R(\epsilon)$ is used in the definition of $\overline{\dim}_{\rm R}(X)$ and $\underline{\dim}_{\rm R}(X)$.

In particular, consider the special case of scalar source and MSE distortion $d(x,y)=|x-y|^2$. Then, whenever $d(X)$ exists and is finite

$$R_2(D)=\frac{1}{2}d(X)\log\frac{1}{D}+\mathrm{o}(\log D). \quad (331)$$

Therefore, $\frac{1}{2}d(X)$ is the scaling factor of $R_2(D)$ with respect to $\log\frac{1}{D}$ in the high-rate regime, which gives an operational characterization of information dimension in Shannon theory. Note that in the most familiar cases we can sharpen (331) to show the following.

- X is discrete and $H(X)<\infty$: $R_2(D)=H(X)+\mathrm{o}(1)$.
- X is continuous and $h(X)>-\infty$: $R_2(D)=\frac{1}{2}\log\frac{1}{2\pi\mathrm{e}D}+h(X)+\mathrm{o}(1)$.

Appendix IV
Injectivity of the Cosine Matrix

We show that the cosine matrix defined in Remark 2 is injective on Σ_k. We consider a more general case. Let $l=2k\le n$ and $\{\omega_1,\ldots,\omega_n\}\subset\mathbb{R}$. Let $\mathbf{H}$ be an $l\times n$ matrix where

$H_{ij} = \cos((i-1)\omega_j)$. We show that each $l \times l$ submatrix formed by columns of $\mathbf{H}$ is non-singular if and only if $\{\cos(\omega_1), \ldots, \cos(\omega_n)\}$ are distinct.

Let $\mathbf{G} = \mathbf{H}_{\{1,\ldots,l\}}$. Then, $G_{ij} = T_{i-1}(x_j)$ where $x_j = \cos(\omega_j)$ and T_m denotes the mth order Chebyshev polynomial of the first kind [54]. Note that $\det(\mathbf{G})$ is a polynomial in $(x_1, \ldots, x_l)$ of degree $l(l-1)/2$. Also $\det(\mathbf{G}) = 0$ if $x_i = x_j$ for some $i \neq j$. Therefore, $\det(\mathbf{G}) = C\prod_{1\le i<j\le l}(x_i - x_j)$. The constant C is given by the coefficient of the highest order term in the contribution from the main diagonal $\prod_{i=1}^{l} T_{i-1}(x_i)$. Since the leading coefficient of T_j is 2^{j-1}, we have $C = 2^{(l-1)(l-2)/2}$. Therefore

$$\det(\mathbf{G}) = 2^{(l-1)(l-2)/2} \prod_{1\le i<j\le l} [\cos(\omega_i) - \cos(\omega_j)] \neq 0. \tag{332}$$

Appendix V
Proof of Lemma 12

Lemma 18: Let $A \subset \mathbb{R}^k$ be compact, and let $g : A \to \mathbb{R}^n$ be continuous. Then, there exists a Borel measurable function $f : g(A) \to A$ such that $g(f(x)) = x$ for all $x \in g(A)$.

Proof: For all $x \in g(A)$, $g^{-1}(\{x\})$ is nonempty and compact since g is continuous. For each $i \in \{1, \ldots, k\}$, let the ith component of f be

$$f_i(x) = \min\{t_i : t^k \in g^{-1}(\{x\})\} \tag{333}$$

where t_i is the ith coordinate of t^k. This defines $f : g(A) \to A$ which satisfies $g(f(x)) = x$ for all $x \in g(A)$. Now we claim that each f_i is *lower semicontinuous*, which implies that f is Borel measurable. To this end, we show that for any $a \in A$, $f^{-1}((a, \infty))$ is open. Assume the opposite, then there exists a sequence $\{y_m\}$ in $g(A)$ that converges to $y \in g(A)$, such that $f(y_m) \le a$ and $f(y) > a$. Due to the compactness of A, there exists a subsequence $f(y_{m_l})$ that converges to some point x in A. Therefore, $x \le a$. But by the continuity of g, we have

$$g(x) = \lim_{l\to\infty} g(f(y_{m_l})) = \lim_{l\to\infty} y_{m_l} = y.$$

Hence, by definition of f and $f(y) > a$, we have $x > a$, which is a contradiction. Therefore, f_i is lower semicontinuous. □

Proof of Lemma 12: Let $S^n \subset \mathbb{R}^n$ be $\lfloor Rn \rfloor$-rectifiable and assume that (238) holds for all $n \ge N$. Then, by definition there exists a bounded subset $T^n \subset \mathbb{R}^{\lfloor Rn \rfloor}$ and a Lipschitz function $g_n : T^n \to \mathbb{R}^n$, such that $S^n = g_n(T^n)$. By continuity, g_n can be extended to the closure $\overline{T^n}$, and $\overline{S^n} = g_n(\overline{T^n})$. Since $\overline{T^n}$ is compact, by Lemma 18, there exists a Borel function $f_n : \mathbb{R}^n \to \mathbb{R}^{\lfloor Rn \rfloor}$, such that $g_n(f_n(x^n)) = x^n$ for all $x^n \in \overline{S^n}$. By Kirszbraun's theorem [42, 2.10.43], g_n can be extended to a Lipschitz function $g_n : \mathbb{R}^{\lfloor Rn \rfloor} \to \mathbb{R}^n$ with the same Lipschitz constant. Then

$$\mathbb{P}\{g_n(f_n(X^n)) = X^n\} \ge \mathbb{P}\{X^n \in \overline{S}^n\} \ge 1 - \epsilon \tag{334}$$

for all $n \ge N$, which proves the ϵ-achievability of R. □

Appendix VI
Proof of Lemma 3

Proof: Since $\|\cdot\|_p$-norms are equivalent, it is sufficient to only consider $p = \infty$. Observe that $\epsilon \mapsto N_A(\epsilon)$ is nonincreasing. Hence, for any $2^{-m} \le \epsilon < 2^{-(m-1)}$, we have

$$\frac{\log N_A(2^{-(m-1)})}{m} \le \frac{\log N_A(\epsilon)}{\log \frac{1}{\epsilon}} \le \frac{\log N_A(2^{-m})}{m-1}. \tag{335}$$

Therefore, it is sufficient to restrict to $\epsilon = 2^{-m}$ and $m \to \infty$ in (77) and (78). To see the equivalence of covering by mesh cubes, first note that $\tilde{N}_A(2^{-m}) \ge N_A(2^{-m})$. On the other hand, any ℓ_∞-ball of radius 2^{-m} is contained in the union of 3^n mesh cubes of size 2^{-m} (by choosing a cube containing some point in the set together with its neighboring cubes). Thus, $\tilde{N}_A(2^{-m}) \le 3^n N_A(2^{-m})$. Hence, the limits in (77) and (78) coincide with those in (84) and (85). □

Appendix VII
Proof of Lemma 5

Proof: By Pinsker's inequality

$$D(P||Q) \ge \frac{1}{2} d(P,Q)^2 \log \mathrm{e} \tag{336}$$

where $d(P,Q)$ is the variational distance between P and Q and $0 \le d(P,Q) \le 2$. In this case, where $\mathcal{X}$ is countable

$$d(P,Q) = \sum_{x\in\mathcal{X}} |P(x) - Q(x)|. \tag{337}$$

By [29, Lemma 2.7, p. 33], when $d(P,Q) \le \frac{1}{2}$

$$|H(P) - H(Q)| \le d(P,Q) \log \frac{|\mathcal{X}|}{d(P,Q)} \tag{338}$$

$$\le d(P,Q) \log |\mathcal{X}| + \mathrm{e}^{-1} \log \mathrm{e}. \tag{339}$$

When $d(P,Q) \ge \frac{1}{2}$, by (336), $D(P||Q) \ge \frac{1}{8}\log \mathrm{e}$; when $0 \le d(P,Q) < 1/2$, by (339)

$$d(P,Q) \ge \frac{(|H(P) - H(Q)| - \mathrm{e}^{-1}\log \mathrm{e})^+}{\log |\mathcal{X}|}. \tag{340}$$

Using (336) again

$$D(P||Q) \ge \frac{1}{2}\left[\frac{(|H(P) - H(Q)| - \mathrm{e}^{-1}\log \mathrm{e})^+}{\log |\mathcal{X}|}\right]^2 \log \mathrm{e}. \tag{341}$$

Since $|H(P) - H(Q)| \ge \delta$ holds in the minimization of (95) and (97), (99) is proved. □

Appendix VIII
Proof of (169)

Proof: By (100), for all $\delta \in \mathbb{R}$

$$E_0(P,\delta) \ge \frac{c(\delta)}{\log^2 |\mathcal{X}|} \tag{342}$$

where

$$c(x) = \frac{\log \mathrm{e}}{2}(x^+)^2 \tag{343}$$

which is a nonnegative nondecreasing convex function.

Since $|X| \leq 1$ a.s., $[X]_m$ is in one-to-one correspondence with $[(X)_1,\ldots,(X)_m]$. Denote the distribution of $[(X)_1,\ldots,(X)_m]$ and $(X)_i$ by P^m and P_i, respectively. By assumption, $(X)_1,\ldots,(X)_m$ are independent, hence

$$P^m = P_1 \times \cdots \times P_m. \tag{344}$$

By (95)

$$E_0(P^m, m\delta) = \min_{Q^m: H(Q^m) \geq H([X]_m)+m\delta} D(Q^m||P^m) \tag{345}$$

where Q^m is a distribution on $\mathbb{Z}_2^m$. Denote the marginals of Q^m by $\{Q_1,\ldots,Q_m\}$. Combining (344) with properties of entropy and relative entropy, we have

$$D(Q^m||P^m) \geq \sum_{i=1}^m D(Q_i||P_i) \tag{346}$$

and

$$H(P^m) = \sum_{i=1}^m H(P_i) \tag{347}$$

$$H(Q^m) \leq \sum_{i=1}^m H(Q_i). \tag{348}$$

Therefore

$$\begin{aligned}
&E_0(P^m, m\delta)\\
&\quad= \min_{H(Q^m) \geq H([X]_m)+m\delta} D(Q^m||P^m) && (349)\\
&\quad\geq \min_{H(Q^m) \geq H([X]_m)+m\delta} \sum_{i=1}^m D(Q_i||P_i) && (350)\\
&\quad\geq \min_{\sum \alpha_i = H([X]_m)+m\delta} \sum_{i=1}^m \min_{H(Q_i)\geq\alpha_i} D(Q_i||P_i) && (351)\\
&\quad\geq \min_{\sum \alpha_i = H([X]_m)+m\delta} \sum_{i=1}^m E_0(P_i, \alpha_i - H(P_i)) && (352)\\
&\quad\geq \min_{\sum \alpha_i = H([X]_m)+m\delta} \sum_{i=1}^m c(\alpha_i - H(P_i)) && (353)\\
&\quad= \min_{\sum \beta_i = m\delta} \sum_{i=1}^m c(\beta_i) && (354)\\
&\quad\geq mc(\delta) && (355)\\
&\quad= \frac{1}{2} m\delta^2 \log \mathrm{e} && (356)
\end{aligned}$$

where:

- (350): by (346);
- (351): by (348), we have

$$\begin{aligned}
&\{Q^m : H(Q^m) \geq H([X]_m) + m\delta\}\\
&\subset \left\{Q^m : H(Q_i) \geq \alpha_i, \sum \alpha_i = H([X]_m) + m\delta\right\}; \qquad (357)
\end{aligned}$$

- (353): by (342);
- (354): let $\beta_i = \alpha_i - H(X_i)$; then, $\sum \beta_i = m\delta$, by (347);
- (355): due to the convexity of $x \mapsto (x^+)^2$. □

Appendix IX
Proof of Lemma 13

Proof: By the k-rectifiability of S, there exists a bounded subset $T \subset \mathbb{R}^k$ and an L-Lipschitz mapping $f : T \to \mathbb{R}^n$ such that $S = f(T)$. Note that

$$\overline{\dim}_{\mathrm{B}} T = \limsup_{\delta \downarrow 0} \frac{\log N_T(\delta)}{\log \frac{1}{\delta}} \leq k. \tag{358}$$

By definition of $N_T(\delta)$, there exists $\{x_i : i = 1,\ldots,N_T(\delta)\}$, such that T is covered by the union of $B(x_i,\delta), i = 1,\ldots,N_T(\delta)$. Then

$$\begin{aligned}
S = f(T) &\subset \bigcup_{i=1}^{N_T(\delta)} f(B(x_i,\delta))\\
&\subset \bigcup_{i=1}^{N_T(\delta)} B(f(x_i), L\delta) && (359)
\end{aligned}$$

which implies that

$$N_S(L\delta) \leq N_T(\delta). \tag{360}$$

Therefore

$$\begin{aligned}
\overline{\dim}_{\mathrm{B}} S &= \limsup_{\delta \downarrow 0} \frac{\log N_S(\delta)}{\log \frac{1}{\delta}} && (361)\\
&\leq \limsup_{\delta \downarrow 0} \frac{\log N_T(\delta/L)}{\log \frac{L}{\delta} + \log \frac{1}{L}} && (362)\\
&\leq k. && (363)
\end{aligned}$$

which the last inequality follows from (358). □

Appendix X
Proof of Lemma 14

Proof: Suppose we can construct a mapping $\tau : [0,1]^k \to [0,1]^{k+1}$ such that

$$\|\tau(x^k) - \tau(y^k)\|_\infty \geq M^{-k} \hat{d}(x^k, y^k) \tag{364}$$

holds for all $x^k, y^k \in [0,1]^k$. By (364), τ is injective. Let $W' = \tau(W)$ and $g' = g \circ \tau^{-1}$. Then, by (364) and the L-Lipschitz continuity of g

$$\begin{aligned}
\|g'(x^k) - g'(y^k)\|_\infty &= \|g(\tau^{-1}(x^k)) - g(\tau^{-1}(y^k))\|_\infty && (365)\\
&\leq L\, \hat{d}(\tau^{-1}(x^k), \tau^{-1}(y^k)) && (366)\\
&\leq LM^k \, \|\tau^{-1}(x^k) - \tau^{-1}(y^k)\|_\infty && (367)
\end{aligned}$$

holds for all $x^k, y^k \in W'$. Hence, $g' : W' \to \mathbb{R}^n$ is Lipschitz with respect to the ℓ_∞ distance, and it satisfies $g'(W') = g(W)$.

To complete the proof of the lemma, we proceed to construct the required τ. The essential idea is to *puncture* the M-ary expansion of x^k such that any component has *at most* k consecu-

$$x^k = \begin{pmatrix} (x_1)_1 & (x_2)_1 & \dots \\ \vdots & \vdots & \dots \\ (x_k)_1 & (x_k)_1 & \dots \end{pmatrix} \xrightarrow{\tau}$$

$$\tau(x^k) = \begin{pmatrix} 0 & (x_1)_2 & \dots & (x_1)_k & 0 & \dots \\ (x_1)_1 & 0 & \dots & (x_2)_k & (x_1)_{k+1} & \dots \\ (x_2)_1 & (x_2)_2 & \dots & (x_3)_k & (x_2)_{k+1} & \dots \\ \vdots & \vdots & \dots & \vdots & \vdots & \dots \\ (x_{k-1})_1 & (x_{k-1})_2 & \dots & (x_k)_k & (x_{k-1})_{k+1} & \dots \\ (x_k)_1 & (x_k)_2 & \dots & 0 & (x_k)_{k+1} & \dots \end{pmatrix}$$

Fig. 1. Schematic illustration of τ in terms of M-ary expansions.

tive nonzero digits. For notational convenience, define $r : \mathbb{N} \to \{0, \dots, k\}$ and $\eta_j : \mathbb{Z}_M^k \to \mathbb{Z}_M^{k+1}$ for $j = 0, \dots, k$ as follows:

$$r(i) = i - \left\lfloor \frac{i}{k+1} \right\rfloor (k+1) \tag{368}$$

$$\eta_j(b^k) = (b_1, \dots, b_j, 0, b_{j+1}, \dots, b_k)^{\mathrm{T}}. \tag{369}$$

Define

$$\tau(x^k) = \sum_{i \in \mathbb{N}} \eta_{r(i)}((x^k)_i) M^{-i}. \tag{370}$$

A schematic illustration of τ in terms of the M-ary expansion is given in Fig. 1.

Next we show that τ satisfies the expansiveness condition in (364). For any $x^k \neq y^k \in [0,1]^k$, let $\hat{d}(x^k, y^k) = M^{-l}$. Then, by definition, for some $m \in \{1, \dots, k\}$, $\hat{d}(x_m, y_m) = M^{-l}$ and $l = \min\{i \in \mathbb{N} : (x_m)_i \neq (y_m)_i\}$. Without loss of generality, assume that $(x_m)_i = 1$ and $(y_m)_i = 0$. Then, by construction of τ, there are no more than k consecutive nonzero digits in $\tau(x)$ or $\tau(y)$. Since the worst case is that $(x_m)_i$ and $(y_m)_i$ are followed by k 0's and k $(M-1)$'s respectively, we have

$$\|\tau(x^k) - \tau(x^k)\|_\infty \geq M^{-(k+l+1)} \tag{371}$$

$$= M^{-(1+k)} \hat{d}(x^k, y^k) \tag{372}$$

which completes the proof of (364). □

ACKNOWLEDGMENT

The authors would like to thank M. Chiang for stimulating discussions and J. Luukkainen of the University of Helsinki for suggesting [50]. They are also grateful for suggestions by an anonymous reviewer.

REFERENCES

[1] C. E. Shannon, "A mathematical theory of communication," in *Bell Syst. Tech. J.*, 1948, vol. 27, pp. 379–423, 623–56.
[2] D. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.
[3] E. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.
[4] M. Wainwright, "Information-theoretic bounds on sparsity recovery in the high-dimensional and noisy setting," in *Proc. IEEE Int. Symp. Inf. Theory*, Nice, France, Jun. 2007.
[5] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proc. Nat. Acad. Sci.*, vol. 106, no. 45, pp. 18914–18919, Nov. 2009.
[6] D. L. Donoho, A. Maleki, and A. Montanari, "Construction of message passing algorithms for compressed sensing," to be submitted to *IEEE Trans Inf. Theory*.
[7] Y. Eftekhari, A. H. Banihashemi, and I. Lambadaris, "An efficient approach toward the asymptotic analysis of node-based recovery algorithms in compressed sensing," 2010 [Online]. Available: http://arxiv.org/abs/1001.2284
[8] P. Schniter, "Turbo reconstruction of structured sparse signals," in *Proc. Conf. Inf. Sci. Syst.*, Princeton, NJ, Mar. 2010.
[9] J. R. Munkres, *Topology*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 2000.
[10] A. Montanari and E. Mossel, "Smooth compression, Gallager bound and nonlinear sparse graph codes," in *Proc. IEEE Int. Symp. Inf. Theory*, Toronto, ON, Canada, Jul. 2008.
[11] A. Rényi, "On the dimension and entropy of probability distributions," *Acta Mathematica Hungarica*, vol. 10, no. 1–2, Mar. 1959.
[12] T. Kawabata and A. Dembo, "The rate-distortion dimension of sets and measures," *IEEE Trans. Inf. Theory*, vol. 40, no. 5, pp. 1564–1572, Sep. 1994.
[13] Y. B. Pesin, *Dimension Theory in Dynamical Systems: Contemporary Views and Applications*. Chicago, IL: Univ. Chicago Press, 1997.
[14] B. R. Hunt and V. Y. Kaloshin, "How projections affect the dimension spectrum of fractal measures," *Nonlinearity*, vol. 10, pp. 1031–1046, 1997.
[15] E. Çinlar, *Probability and Stochastics*. New York: Springer-Verlag, 2010.
[16] A. Rényi, *Probability Theory*. Amsterdam, The Netherlands: North-Holland, 1970.
[17] K. Falconer, *Techniques in Fractal Geometry*. New York: Wiley, 1997.
[18] K. Falconer, *Fractal Geometry: Mathematical Foundations and Applications*, 2nd ed. New York: Wiley, 2003.
[19] A. György, T. Linder, and K. Zeger, "On the rate-distortion function of random vectors and stationary sources with mixed distributions," *IEEE Trans. Inf. Theory*, vol. 45, pp. 2110–2115, 1999.
[20] T. Linder and R. Zamir, "On the asymptotic tightness of the Shannon lower bound," *IEEE Trans. Inf. Theory*, vol. 40, no. 6, pp. 2026–2031, Nov. 1994.
[21] I. Csiszár, "On the dimension and entropy of order α of the mixture of probability distributions," *Acta Mathematica Hungarica*, vol. 13, no. 3–4, pp. 245–255, Sep. 1962.
[22] P. Halmos, *Naive Set Theory*. Princeton, NJ: D. Van Nostrand, 1960.
[23] K. Kuratowski, *Topology*. New York: Academic, 1966, vol. I.
[24] G. Folland, *Real Analysis: Modern Techniques and Their Applications*, 2nd ed. New York: Wiley-Interscience, 1999.
[25] P. Mattila, *Geometry of Sets and Measures in Euclidean Spaces: Fractals and Rectifiability*. Cambridge, U.K.: Cambridge Univ. Press, 1999.
[26] E. Candés, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Commun. Pure Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, Aug. 2006.
[27] R. Calderbank, S. Howard, and S. Jafarpour, "Construction of a large class of deterministic matrices that satisfy a statistical isometry property," *IEEE J. Sel. Topics Signal Process.*, vol. 29, no. 4, 2009.
[28] W. Xu and B. Hassibi, "Compressed sensing over the Grassmann manifold: A unified analytical framework," in *Proc. 46th Annu. Allerton Conf. Commun. Control Comput.*, 2008, pp. 562–567.
[29] I. Csiszár and J. G. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1982.
[30] I. Csiszár, "Generalized cutoff rates and Rényi's information measures," *IEEE Trans. Inf. Theory*, vol. 41, no. 1, pp. 26–34, Jan. 1995.
[31] P. N. Chen and F. Alajaji, "Csiszár's cutoff rates for arbitrary discrete sources," *IEEE Trans. Inf. Theory*, vol. 47, no. 1, pp. 330–338, Jan. 2001.
[32] H. Shimokawa, "Rényi's entropy and error exponent of source coding with countably infinite alphabet," in *Proc. IEEE Int. Symp. Inf. Theory*, Seattle, WA, Jul. 2006.
[33] C. Chang and A. Sahai, "Universal quadratic lower bounds on source coding error exponents," in *Proc. Conf. Inf. Sci. Syst.*, 2007.
[34] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[35] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*. San Francisco, CA: Holden-Day, 1964.
[36] K. Alligood, T. Sauer, and J. A. Yorke, *Chaos: An Introduction to Dynamical Systems*. New York: Springer-Verlag, 1996.
[37] R. Mañé, "On the dimension of the compact invariant sets of certain non-linear maps," in *Dynamical Systems and Turbulence, Warwick 1980*, ser. Lecture Notes in Mathematics. Berlin, Germany: Springer-Verlag, 1981, vol. 898, pp. 230–242.
[38] B. R. Hunt and V. Y. Kaloshin, "Regularity of of embeddings of infinite-dimensional fractal sets into finite-dimensional spaces," *Nonlinearity*, vol. 12, no. 5, pp. 1263–1275, 1999.
[39] A. Ben-Artzi, A. Eden, C. Foias, and B. Nicolaenko, "Hölder continuity for the inverse of Mañé's projection," *J. Math. Anal. Appl.*, vol. 178, pp. 22–29, 1993.
[40] H. Steinhaus, "Sur les distances des points des ensembles de mesure positive," *Fundamenta Mathematicae*, vol. 1, pp. 93–104, 1920.
[41] G. J. Minty, "On the extension of Lipschitz, Lipschitz-Hölder continuous, and monotone functions," *Bull. Amer. Math. Soc.*, vol. 76, no. 2, pp. 334–339, 1970.
[42] H. Federer, *Geometric Measure Theory*. New York: Springer-Verlag, 1969.
[43] D. Preiss, "Geometry of measures in $\mathbb{R}^n$: Distribution, rectifiability, and densities," *Ann. Math.*, vol. 125, no. 3, pp. 537–643, 1987.
[44] G. A. Edgar, *Integral, Probability, and Fractal Measures*. New York: Springer-Verlag, 1997.
[45] M. Keane and M. Smorodinsky, "Bernoulli schemes of the same entropy are finitarily isomorphi," *Ann. Math.*, vol. 109, no. 2, pp. 397–406, 1979.
[46] A. Del Junco, "Finitary codes between one-sided Bernoulli shifts," *Ergodic Theory Dyn. Syst.*, vol. 1, pp. 285–301, 1981.
[47] J. G. Sinaĭ, "A weak isomorphism of transformations with an invariant measure," *Dokl. Akad. Nauk SSSR.*, vol. 147, pp. 797–800, 1962.
[48] K. E. Petersen, *Ergodic Theory*. Cambridge, U.K.: Cambridge Univ. Press, 1990.
[49] W. H. Schikhof, *Ultrametric Calculus: An Introduction to p-Adic Analysis*. New York: Cambridge Univ. Press, 2006.
[50] M. A. Martin and P. Mattila, "k-dimensional regularity classifications for s-fractals," *Trans. Amer. Math. Soc.*, vol. 305, no. 1, pp. 293–315, 1988.
[51] V. Milman, "Dvoretzky's theorem: Thirty years later," *Geom. Funct. Anal.*, vol. 2, no. 4, pp. 455–479, Dec. 1992.
[52] W. Johnson and J. Lindenstrauss, "Extensions of Lipschitz maps into a Hilbert space," *Contemp. Math.*, vol. 26, pp. 189–206, 1984.
[53] E. C. Posner and E. R. Rodemich, "Epsilon entropy and data compression," *Ann. Math. Stat.*, vol. 42, no. 6, pp. 2079–2125, Dec. 1971.
[54] G. Szegö, *Orthogonal Polynomials*. Providence, RI: AMS, 1975.
[55] Y. Wu and S. Verdú, "Fundamental limits of almost lossless analog compression," in *Proc. IEEE Int. Symp. Inf. Theory*, Seoul, Korea, Jun. 2009.

Yihong Wu (S'10) received the B.E. degree in electrical engineering from Tsinghua University, Beijing, China, in 2006 and the M.A. degree in electrical engineering from Princeton University, Princeton, NJ in 2008, where he is currently working towards the Ph.D. degree at the Department of Electrical Engineering.

He is a recipient of the Princeton University Wallace Memorial honorific fellowship in 2010. His research interests are in information theory, signal processing, mathematical statistics, optimization, and distributed algorithms.

Sergio Verdú (S'80–M'84–SM'88–F'93) received the Telecommunications Engineering degree from the Universitat Politècnica de Barcelona, Barcelona, Spain, in 1980 and the Ph.D. degree in electrical engineering from the University of Illinois at Urbana-Champaign, Urbana, in 1984.

Since 1984, he has been a member of the faculty of Princeton University, Princeton, NJ, where he is the Eugene Higgins Professor of Electrical Engineering.

Dr. Verdú is the recipient of the 2007 Claude E. Shannon Award and the 2008 IEEE Richard W. Hamming Medal. He is a member of the National Academy of Engineering and was awarded a Doctorate Honoris Causa from the Universitat Politècnica de Catalunya in 2005. He is a recipient of several paper awards from the IEEE: the 1992 Donald Fink Paper Award, the 1998 Information Theory Outstanding Paper Award, an Information Theory Golden Jubilee Paper Award, the 2002 Leonard Abraham Prize Award, the 2006 Joint Communications/Information Theory Paper Award, and the 2009 Stephen O. Rice Prize from the IEEE Communications Society. He has also received paper awards from the Japanese Telecommunications Advancement Foundation and from Eurasip. He received the 2000 Frederick E. Terman Award from the American Society for Engineering Education for his book *Multiuser Detection* (Cambridge, U.K.: Cambridge Univ. Press, 1998). He served as President of the IEEE Information Theory Society in 1997. He is currently Editor-in-Chief of *Foundations and Trends in Communications and Information Theory*.