Compression-Based Compressed Sensing

Farideh Ebrahim Rezagah, Shirin Jalali, Elza Erkip, H. Vincent Poor

Abstract

Modern compression algorithms exploit complex structures that are present in signals to describe them very efficiently. On the other hand, the field of compressed sensing is built upon the observation that "structured" signals can be recovered from their under-determined set of linear projections. Currently, there is a large gap between the complexity of the structures studied in the area of compressed sensing and those employed by the state-of-the-art compression codes. Recent results in the literature on deterministic signals aim at bridging this gap through devising compressed sensing decoders that employ compression codes. This paper focuses on structured stochastic processes and studies the application of rate-distortion codes to compressed sensing of such signals. The performance of the formerly-proposed compressible signal pursuit (CSP) algorithm is studied in this stochastic setting. It is proved that in the very low distortion regime, as the blocklength grows to infinity, the CSP algorithm reliably and robustly recovers n instances of a stationary process from random linear projections as long as their count is slightly more than n times the rate-distortion dimension (RDD) of the source. It is also shown that under some regularity conditions, the RDD of a stationary process is equal to its information dimension (ID). This connection establishes the optimality of the CSP algorithm at least for memoryless stationary sources, for which the fundamental limits are known. Finally, it is shown that the CSP algorithm combined by a family of universal variable-length fixed-distortion compression codes yields a family of universal compressed sensing recovery algorithms.

Keywords: Compressed Sensing, Lossy Compression, Universal Compression, Rate-Distortion Dimension, Information Dimension.

I. Introduction

Consider the standard setup of a compressed sensing data acquisition system: a decoder observes a noisy linear projection of the high-dimensional signal $\mathbf{x} \in \mathbb{R}^n$, i.e., $\mathbf{y} = A\mathbf{x} + \mathbf{z}$, where $A \in \mathbb{R}^{m \times n}$, $m < n$, is the measurement matrix, and $\mathbf{z} \in \mathbb{R}^m$ denotes the measurement noise. The signal is assumed to be "structured", which typically means that it is sparse in some transform domain. The decoder is expected to recover the signal $\mathbf{x}$ using a computationally efficient algorithm with as few number of measurements, m, as possible. Such modern data acquisition problems, which can be described as solving under-determined systems of linear equations, arise in many different applications, including magnetic resonance imaging (MRI), high resolution imaging, and radar.

While sparsity of the desired signal is the main focus in the compressed sensing literature, started by the key works of Donoho et al. [1] and Candes et al. [2], [3], more recent compressed sesning recovery algorithms capture structures beyond sparsity, such as group sparsity, low-rankness, etc. [4]–[28]. Although the studied structures in the literature and their extensions are present in many signals of interest, and yield promising results, they are to a great extent confined to basic models, compared to more complex underlying structures known to be present in such signals. Employing such elaborate structures that are usually present in the signals can potentially lead to

much more efficient compressed sensing systems that require significantly smaller numbers of measurements, for achieving the same reconstruction quality.

In addition to compressed sensing, the structure of a signal plays an important role in many other fundamental problems in information theory, such as data compression, data prediction and denoising. Data compression is a well-studied topic in information theory, initiated by the Shannon's seminal work [29]. Compression algorithms employ the patterns in a signal to render efficient digital representation of it. After decades of research, the types of structure employed by the state-of-the-art compression algorithms, especially for coding image, audio and video signals, are quite elaborate, and much more complicated than those studied in compressed sensing.

Given the maturity and the efficiency of existing data compression algorithms, one may wonder whether data compression codes can be directly employed to build compressed sensing recovery algorithms. The motivation for such an approach is that in order for a good compression code to represent some process as efficiently as possible, theoretically, it should employ all the structure that is present in it. Therefore, building a compressed sensing decoder based on an efficient data compression code, potentially might enable the decoder to exploit all the structure present in the data, and, thereby minimize the number of measurements. Also, another advantage of this approach would be devising a generic process for building compressed sensing recovery algorithms based on compression codes to simplify this task. Hence, to find the most efficient compressed sensing data recovery algorithm for a given source of data, instead of studying and learning some specific structure in the source model, the already existing data compression codes can be used to directly design efficient compressed sensing decoders.

The idea of utilizing compression codes in designing compressed sensing recovery algorithms was introduced in [30] and [31]. Consider $x^n \in \mathcal{Q}$, where $\mathcal{Q}$ represents a compact subset of $\mathbb{R}^n$. A compression code of rate r for the set $\mathcal{Q}$ is described by encoder and decoder mappings

$$f_n : \mathcal{Q} \to \{1, \dots, 2^r\}$$

and

$$g_n : \{1, \dots, 2^r\} \to \mathbb{R}^n,$$

respectively. The distortion induced by this code is defined as

$$\delta \triangleq \sup_{x^n \in \mathcal{Q}} \|x^n - g_n(f_n(x^n))\|_2.$$

This code defines a codebook $\mathcal{C}_n$, which contains all possible reconstruction vectors generated by this code. That is,

$$\mathcal{C}_n \triangleq \{g_n(f_n(x^n)) : x^n \in \mathcal{Q}\}.$$

Clearly, $|\mathcal{C}_n| \leq 2^r$. Suppose a decoder desires to recover the signal x^n from noisy underdetermined linear projections $y^m = Ax^n + z^m$, by employing the compression code (f_n, g_n), without explicitly studying the set $\mathcal{Q}$. To achieve

this goal, [30] and [31] propose the compressible signal pursuit (CSP) optimization defined as

$$\tilde{x}^n \triangleq \underset{c^n \in \mathcal{C}_n}{\arg\min} \|y^m - Ac^n\|_2^2.$$

In other words, to recover the original signal from sufficient number of random linear projections, through an exhaustive search over the codebook of the compression code, the CSP seeks the reconstruction vector in $\mathcal{C}_n$ that minimizes the measurement error. It can be shown that the required number of measurements for successful recovery depends on the rate-distortion trade-off of the compression code and the desired accuracy [30], [31].

The results of [30] and [31] on deterministic signals establish the foundations of building compression-based compressed sensing decoders. However, since the studied model only concerns deterministic signals, the results do not illustrate the fundamental connections between the source structure, which is captured by its distribution, its information theoretic rate-distortion function and the number of measurements required by compression-based decoders. In this paper we focus on *stationary analog processes*, and study the performance of the CSP algorithm, as a compression-based compressed sensing recovery algorithm. This shift from deterministic signals to stochastic stationary processes enables us to

1) characterize the performance of the CSP algorithm in terms of the information theoretic rate-distortion function of the source, and illustrate the connection between the asymptotic number of measurements required by the CSP and the rate-distortion dimension of the source process;
2) establish new fundamental connections between the rate-distortion dimension of the source, and its information dimension, which serves as its measure of complexity;
3) employ the established connection and prove asymptotic optimality of the CSP algorithm for cases in which the fundamental limits of compressed sensing is known; and
4) employ universal compression codes, and design a compression-based universal compressed sensing recovery algorithm.

Since the sources of interest in compressed sensing applications are usually analog, compression codes employed in building compression-based decoders has to be lossy codes. As a result, the reconstruction given by the CSP algorithm is also a lossy reconstruction. In other words, the resulting compression-based recovery algorithm is a lossy compressed sensing algorithm, where there is a trade-off between the number of measurements, the quality of the reconstruction, and the rate and the distortion of the compression code. In this paper we mainly focus on the this trade-off and leave the complexity issues for future extensions of this work, where a more algorithmic approach would be necessary to handle or at least approximate the minimization in CSP with reasonable time-complexity.

In a standard compressed sensing setting, the decoder recovers the signal *losslessly* or almost losslessly from an underdetermined set of linear equations. While compression-based recovery algorithms enable us to exploit more complex structures, there is an inherent loss due to the underlying lossy compression codes. By letting the distortion of the compression code become arbitrarily small, we can achieve almost lossless recovery which is of interest in compressed sensing problems. Although arbitrarily small distortion for an analog signal dictates an arbitrarily large

compression code rate, the RDD of the code, which is the quantity that directly relates the compression code's rate-distortion behavior to the number of required measurements, remains bounded.

In this paper we consider a stochastic analog source $\mathbf{X} = \{X_i\}_{i=-\infty}^{\infty}$ and signal X^n generated by this source. Instead of observing X^n directly, a decoder measures $Y^m = AX^n + Z^m$, $m < n$, and aims at estimating X^n from Y^m. Here, similar to the deterministic setup, $A \in \mathbb{R}^{m\times n}$ and Z^m denote the measurement matrix and the stochastic noise in the system, respectively. Assume that the data acquisition decoder has access to a "good" lossy compression code for the source $\mathbf{X}$, and employs it to recover the vector X^n via the CSP algorithm. Our first major contribution in this paper is to derive the trade-off between the performance of the compression code, stated in terms of its rate, distortion and excess distortion probability, and the performance of the CSP algorithm, summarized by the required number of linear measurements and its achieved reconstruction quality. We prove that, asymptotically, for large n and as the distortion of the compression codes goes to zero, the normalized number of random linear measurements required by the CSP algorithm is equal to the RDD [32] of the source. It is known that for a random variable (or vector), the (upper and lower) RDD is equal to the (upper and lower) information dimension (ID) of the random variable [32]. Our second major contribution is to extend this result to analog stationary processes, and to prove that, under some regularity conditions, the RDD of a stationary process is equal to its ID, defined in [33]. This combined with the results of [34] establishes the asymptotic optimalilty of CSP for stationary memoryless sources.

We study piecewise-constant signals to illustrate our results on the connection between RDD and ID. Piecewise-constant signals are used widely to model many natural signals in the signal processing, compression, and denoising literature. We derive upper and lower bounds on the rate-distortion functions of such signals, when they are modeled by a first-order Markov process and use these bounds to evaluate the RDD of such processes.

Given our focus on building compression-based compressed sensing algorithms, we also address two related important questions: Can one derive a universal compressed sensing recovery algorithm based on a given universal compression code? How well will such a scheme perform? In information theory, universal codes refer to algorithms that do not require knowledge of the source distribution and yet achieve the optimal performance. Universal lossy or lossless compression [35]–[42], universal denoising [43], [44] and universal prediction [45], [46] are some examples of universal coding problems that have been well-studied in information theory. The problem of universal compressed sensing and the existence of such algorithms that can recover a signal from its underdetermined set of random linear observations without knowing the source model has recently been studied both for deterministic [47] and probabilistic signal models [33], [48], [49]. Our third major contribution is addressing both of the above questions. We prove that a family of universal fixed-distortion compression codes yields a family of universal compressed sensing recovery algorithms. This connection has important implications both in theory and in practice.

The organization of the paper is as follows. Section II studies the performance the CSP algorithm when applied to compressed sensing of a stationary process. Section III examines the properties of complexity measures for analog stationary processes, and establishes a connection between the ID and the RDD of such processes and also provides bounds on the rate-distortion region of the piecewise constant source modeled by a first-order Markov process to

illustrate this relationship. Section IV provides the performance and optimality of CSP for almost lossless recovery using the established connection between RDD and ID. Universal CSP (UCSP) is introduced in Section V as a universal compressed sensing recovery algorithm, and its performance trade-offs are studied. Section VI presents the proofs of some of the results, and Section VII concludes the paper.

A. Notation

Calligraphic letters such as $\mathcal{X}$ and $\mathcal{Y}$ denote sets. The size of a set $\mathcal{X}$ is denoted by $|\mathcal{X}|$. Capital letters like X and Y represent random variables. For a random variable X, $\mathcal{X}$ denotes its alphabet. For $x \in \mathbb{R}$, $\lceil x \rceil$ ($\lfloor x \rfloor$) represents the smallest (largest) integer larger (smaller) than x. For $b \in \mathbb{N}^+$, $[x]_b$ denotes the b-bit approximation of x, i.e., for $x = \lfloor x \rfloor + \sum_{i=1}^{\infty} (x)_i 2^{-i}$, $(x)_i \in \{0,1\}$,

$$[x]_b = \lfloor x \rfloor + \sum_{i=1}^{b} (x)_i 2^{-i}.$$

Also, let $\langle x \rangle_b$ defined as

$$\langle x \rangle_b = \frac{\lfloor bx \rfloor}{b},$$

denote the discretized version of x. For $x \in \mathbb{R}$, δ_x denotes the Dirac measure with an atom at x. Throughout the paper, $\log$ and $\ln$ refer to the logarithm in base 2 and natural logarithm, respectively. $\{0,1\}^* = \cup_{n=1}^{\infty} \{0,1\}^n$ denotes the set of all binary sequences of finite length. For a binary sequence $b \in \{0,1\}^n$, $|b|$ denotes the length of the sequence.

II. Compressible signal pursuit

This section extends the CSP algorithm proposed in [30] and [31] to stochastic processes. The intuition behind the CSP algorithm is that if a set of signals can be compressed efficiently using a compression code, then the structure employed by the compression code can indirectly, through the application of the compression code, be used in building efficient compressed sensing recovery algorithms. In other words, the CSP algorithm, through the compression code, extracts all the useful structure present in the data to reduce the number of linear measurements.

Consider a random vector X^n, generated by stationary process $\mathbf{X} = \{X_i\}_{i=0}^{\infty}$, where $X_i \in \mathcal{X}$. A compressed sensing decoder observes a linear projection of X^n,

$$Y^m = AX^n,$$

where $A \in \mathbb{R}^{m \times n}$ denotes the measurement matrix with $m < n$, and aims at estimating X^n.

A fixed-length lossy compression code for the source $\mathbf{X}$ operating at rate R and blocklength n is specified as (n, f_n, g_n), where

$$f_n : \mathcal{X}^n \mapsto \{1, 2, \ldots, 2^{nR}\}$$

and

$$g_n : \{1, 2, ..., 2^{nR}\} \mapsto \hat{\mathcal{X}}^n$$

denote the encoding and the decoding functions, respectively and $\hat{\mathcal{X}}$ is the reconstruction alphabet. Throughout the paper, we mainly focus on the case where $\mathcal{X} = \hat{\mathcal{X}} = \mathbb{R}$ with squared error distortion $d(x, \hat{x}) = (x - \hat{x})^2$, where, $d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$ denotes a per-letter distortion measure. Traditionally, the performance of a lossy compression code is measured in terms of its rate, R, and expected average distortion $D \triangleq \mathrm{E}[d_n(X^n, \hat{X}^n)]$, where $\hat{X}^n = g_n(f_n(X^n))$, and

$$d_n(x^n, \hat{x}^n) \triangleq \frac{1}{n} \sum_{i=1}^{n} d(x_i, \hat{x}_i),$$

for any $x^n \in \mathcal{X}^n$ and $\hat{x}^n \in \hat{\mathcal{X}}^n$. Another possible performance metric for a lossy compression code is its *excess distortion probability* [50], which is a stronger notion than expected distortion. The excess distortion probability of a code is defined as the probability that the average per-letter distortion between the source and reconstruction blocks exceeds some predetermined threshold, i.e., $\mathrm{P}(d_n(X^n, \hat{X}^n) > D)$. The distortion D is said to be achievable at rate R if for any $\epsilon > 0$, there is a large enough n_0 such that for any $n > n_0$ the (n, f_n, g_n) code satisfies

$$\mathrm{P}(d_n(X^n, \hat{X}^n) > D) \leq \epsilon,$$

i.e. the excess distortion probability ϵ can be driven to zero as $n \to \infty$.

Remark 1. Let $R_m(\mathbf{X}, D)$ and $R_a(\mathbf{X}, D)$ denote the rate-distortion functions of a source $\mathbf{X}$ under vanishing excess distortion probability and expected average distortion, respectively. While $R_m(\mathbf{X}, D)$ and $R_a(\mathbf{X}, D)$ are not equal in general, for stationary ergodic processes $R_m(\mathbf{X}, D) = R_a(\mathbf{X}, D)$ [51]–[53]. Throughout this paper we focus only on such processes; therefore, we drop the subscript m or a, and let $R(\mathbf{X}, D)$ denote the rate-distortion function of the source.

Let

$$\mathcal{C}_n \triangleq \{g_n(f_n(x^n)) : \ x^n \in \mathcal{X}^n\}$$

denote the codebook of this compression code. Clearly, $|\mathcal{C}_n| \leq 2^{nR}$. Given the source output X^n and the observation vector $Y^m = AX^n$, let $\tilde{X}^n$ denote the solution of the CSP algorithm employing the (n, f_n, g_n) code. In other words,

$$\tilde{X}^n = \underset{x^n \in \mathcal{C}_n}{\arg\min} \|Y^m - Ax^n\|_2^2. \tag{1}$$

The following theorem derives an upper bound on the loss incurred by the CSP in recovering X^n. The bound on reconstruction distortion holds with high probability and depends on the parameters of the compression code n, R, D and ϵ, and the number of measurements m. It is important to note that the compression code used by the CSP algorithm is not required to be an optimal code, and the theorem also holds even if the CSP algorithm is based on an off-the-shelf compression code.

Theorem 1. *Consider $Y^m = AX^n$, a system of random linear observations with measurement matrix $A \in \mathbb{R}^{m \times n}$,*

where $A_{i,j}$ are independently and identically distributed (i.i.d.) as $\mathcal{N}(0,1)$. Let $\mathcal{C}_n$ be a lossy compression code for X^n operating at rate R that achieves distortion D with excess distortion probability ϵ. Without any loss of generality assume that the source is normalized such that $D < 1$. For arbitrary $\alpha > 0$ and $\eta > 1$, let $\delta = \frac{\eta}{\log \frac{1}{D}} + \alpha$, and

$$\frac{m}{n} = \frac{2\eta R}{\log \frac{1}{D}},$$

be the normalized number of observations. Let $\tilde{X}^n$ denote to the solution of the CSP algorithm given in (1)*. Then,*

$$\begin{aligned} &\mathrm{P}\Big(\frac{1}{\sqrt{n}}\| X^n - \tilde{X}^n\|_2 \geq (2+\sqrt{n/m})D^{\frac{1}{2}(1-\frac{1+\delta}{\eta})}\Big) \\ &\quad \leq \epsilon + 2^{-\frac{1}{2}nR\alpha} + \mathrm{e}^{-\frac{m}{2}}. \end{aligned}$$

Proof: The proof is provided in Section VI. ■

Theorem 1 states that using a class of compression codes $\mathcal{C}_n$ operating at rate R and distortion D, with $m = \frac{2\eta R n}{\log(1/D)}$ random linear measurements ($\eta > 1$) of n samples of a stochastic process, the distortion incurred by CSP in recovering X^n can be upper-bounded with probability approaching one as n grows without bound. In the limit when D approaches zero, the normalized number of measurements required by the CSP algorithm, for almost lossless recovery of the source, depends on the limit of $\frac{2R}{\log(1/D)}$. If the compression code used by the CSP operates close to the fundamental rate-distortion tradeoff of the source, this limit approaches the rate-distortion dimension of the source [32]. To better understand the performance of the CSP algorithm, in the following section, we focus on this quantity and explore its connections with other known measures of complexity for stationary processes.

As stated in Theorem 1, $\eta > 1$ is a free parameter that affects the performance of the CSP algorithm. Choosing a small η, arbitrarily close to 1, minimizes the number of random linear measurements, m, required by the CSP. On the other hand, since the reconstruction distortion scales as $D^{\frac{1}{2}(1-\frac{1+\delta}{\eta})}$, where $\delta > 0$, for optimal scaling of the distortion ($\sqrt{D}$), η needs to be large. In other words, the closer $\frac{1}{2}(1-\frac{1+\delta}{\eta})$ gets to 1, the better performance we get from CSP in terms of reconstruction distortion. Therefore, as η varies, there is a trade-off between the number of measurements on one hand and the scaling of the reconstruction distortion on the other hand.

Theorem 1 characterizes the performance of the CSP algorithm in recovering a random process, when there is no noise in the measurement process. In reality, there is always some noise in the system. The following theorem proves the robustness of the performance of the CSP algorithm to measurement noise. Specifically assume that instead of $Y^m = AX^n$, the decoder observes $Y^m = AX^n + Z^m$, where Z^m denotes some random measurement noise. Further assume that the decoder employs the CSP algorithm as before to recover X^n from measurements Y^m. That is, $\tilde{X}^n$ is still given by (1). The following theorem states that if the noise power is not very large and the compression code's distortion D stays away from zero, then the performance of the CSP algorithm essentially stays the same.

Theorem 2. *Consider $Y^m = AX^n + Z^m$, a noisy system of random linear observations where Z^m is the additive noise and $A \in \mathbb{R}^{m\times n}$ is the measurement matrix where $A_{i,j}$ are i.i.d. as $\mathcal{N}(0,1)$. Assume that the average power*

of the noise can be bounded by σ_m^2 with probability $1-\epsilon_m$, i.e.

$$\mathrm{P}\left(\frac{1}{\sqrt{m}}\|Z^m\|_2 > \sigma_m\right) < \epsilon_m.$$

Let $\mathcal{C}_n$ be a lossy compression code for X^n operating at rate R that achieves distortion D with excess distortion probability ϵ. Without any loss of generality assume that the source is normalized such that $D<1$. For arbitrary $\alpha>0$ and $\eta>1$, let $\delta=\frac{\eta}{\log\frac{1}{D}}+\alpha$, and $m=\frac{2\eta nR}{\log(1/D)}$ be the normalized number of observations, and let $\tilde{X}^n$ be the solution of the CSP algorithm, as given by (1)*. Then,*

$$\begin{aligned} P\Bigg(&\frac{1}{\sqrt{n}}\| X^n-\tilde{X}^n\|_2 \geq \\ &(2+\sqrt{n/m})D^{\frac{1}{2}(1-\frac{1+\delta}{\eta})}+\frac{2\sigma_m}{\sqrt{D^{\frac{1+\delta}{\eta}}n}}\Bigg) \\ &\leq \epsilon_m+\epsilon+2^{-\frac{1}{2}nR\alpha}+\mathrm{e}^{-\frac{m}{2}}. \end{aligned}$$

Proof: The proof is provided in Section VI. ■

The effect of the noise on the error is captured by the term, $\frac{2\sigma_m}{\sqrt{D^{\frac{1+\delta}{\eta}}n}}$, which disappears as $n\to\infty$. This is due to the fact that by drawing the entries of the measurement matrix based on an i.i.d. $\mathcal{N}(0,1)$ distribution, the signal to noise ratio (SNR) of each measurement goes to infinity as $n\to\infty$. If instead the entries of A are drawn as $\mathcal{N}(0,\frac{1}{n})$, then the term dependent on noise becomes $2\sigma_m/\sqrt{D^{\frac{1+\delta}{\eta}}}$, which does not disappear as n grows to infinity.

III. Information and complexity measures

To develop a unified approach to the problem of structured signal recovery, and also to fundamentally understand the connections between the problems of data compression and compressed sensing, a universal notion of complexity for analog signals is required. Such a notion of complexity is expected to effectively measure all the information contained in the structure of an analog signal.

For discrete signals, there are well-known measures of complexity in the information theory literature. The entropy $H(X)$ and the entropy rate $\bar{H}(\mathbf{X})=\lim_{n\to\infty}H(X_n|X^{n-1})$ measure the complexity of random variable X and stationary process $\mathbf{X}=\{X_i\}$, respectively. Both of these measures are closely connected to the minimum number of bits per symbol required for representing stochastic sources [54]. However, when we shift from discrete alphabet to analog, both the entropy, and the entropy rate become infinite. Therefore, such measures cannot be used for capturing the structure of such signals.

To illustrate what is meant for an analog process to be structured, consider a stationary memoryless (i.e., i.i.d.) process $\mathbf{X}=\{X_i\}_{i=0}^{\infty}$ such that $X_i\sim(1-p)\delta_0+pf_c$, where f_c denotes the probability density function (pdf) of an absolutely continuous distribution. In other words, for each i, with probability p, X_i is exactly equal to zero, otherwise, it is drawn from f_c. From this definition, a block X^n generated by this source contains around $n(1-p)$ entries equal to zero, and the rest of the entries are real numbers in the domain of f_c. To describe X^n with a certain

precision, for zero entries, it suffices to describe their locations. The number of bits required for this description does not depend on the reconstruction quality. However, for the remaining approximately np elements of X^n, it can be proved that the required number of bits grows proportionally to the desired reconstruction quality. This intuitively suggests that the probability p, which controls the number of non-zero elements in X^n, is a fundamental quantity related to the complexity of X^n. This intuition is nicely captured by the notion of ID introduced by Rényi [55].

Defintion 1 (Rényi information dimension [55])**.** *The Rényi upper and lower IDs of an analog random variable X are defined as*

$$\bar{d}(X) = \limsup_{b\to\infty} \frac{H(\langle X\rangle_b)}{\log b},$$

and

$$\underline{d}(X) = \liminf_{b\to\infty} \frac{H(\langle X\rangle_b)}{\log b},$$

respectively. If the two limits coincide, $d(X) = \bar{d}(X) = \underline{d}(X)$ is defined as the Rényi ID of X.

Note that while the above definition of the Rényi IDs is in terms of the entropy of the b-level quantized version of X normalized by the number of bits required for binary representation of it, $\log b$, it is easy to see that we can equivalently find them in terms of the entropy of the b-bit approximation of X, $[X]_b$, normalized by b, the number of bits i.e. $\bar{d}(X) = \limsup_{b\to\infty} \frac{H([X]_b)}{b}$, and $\underline{d}(X) = \liminf_{b\to\infty} \frac{H([X]_b)}{b}$.

The Rényi ID of a random variable serves as a measure of complexity for analog random variables. To shed some light on this measure, consider the i.i.d. sparse source $\mathbf{X}$ described earlier. It can be proved that the Rényi ID of each X_i is equal to p, which is the probability that X_i is non-zero [55]. Decreasing the parameter p increases the sparsity level of the output of such a source, and hence intuitively decreases its complexity. This phenomenon is captured by the Rényi ID of X. In fact, δ_0 can be changed to any discrete probability distribution and the result will not change since the Rényi ID of a discrete source is 0. The notion of Rényi ID for random variables or vectors was extended in [33] to define the ID of analog stationary processes.

Defintion 2 (ID of a stationary process [33])**.** *The k-th order upper and lower IDs of stationary process $\mathbf{X} = \{X_i\}_{i=-\infty}^{\infty}$ are defined as*

$$\bar{d}_k(\mathbf{X}) = \limsup_{b\to\infty} \frac{1}{b} H([X_{k+1}]_b|[X^k]_b),$$

and

$$\underline{d}_k(\mathbf{X}) = \liminf_{b\to\infty} \frac{1}{b} H([X_{k+1}]_b|[X^k]_b),$$

respectively. The upper and lower ID of process $\mathbf{X}$ are defined as

$$\bar{d}_o(\mathbf{X}) = \lim_{k\to\infty} \bar{d}_k(\mathbf{X})$$

and

$$\underline{d}_o(\mathbf{X}) = \lim_{k\to\infty} \underline{d}_k(X),$$

respectively, when the limits exist. If $\bar{d}_o(\mathbf{X}) = \underline{d}_o(\mathbf{X})$*, the ID of process* $\mathbf{X}$*,* $d_o(\mathbf{X})$*, is defined as* $d_o(\mathbf{X}) = \bar{d}_o(\mathbf{X}) = \underline{d}_o(\mathbf{X})$.

For a stationary memoryless i.i.d. process $\mathbf{X} = \{X_i\}_{i=-\infty}^{\infty}$, this definition coincides with that of Rényi's ID of the first order marginal distribution of the process $\mathbf{X}$. That is $\bar{d}_o(\mathbf{X}) = \bar{d}(X_1)$ and $\underline{d}_o(\mathbf{X}) = \underline{d}(X_1)$. For sources with memory, taking the limit as the memory parameter k grows to infinity allows $d_o(\mathbf{X})$ to capture the overall structure that is present in an analog stationary process. It can be proved that $d_o(\mathbf{X}) \leq 1$, for all stationary processes, and if the stationary process $\mathbf{X}$ is structured, $d_o(\mathbf{X})$ is strictly smaller than one [33]. As an example of a structured stationary analog process with memory, consider a piecewise constant signal modeled by a first order Markov process $\mathbf{X} = \{X_i\}_{i=1}^{\infty}$, such that conditioned on $X_{i-1} = x_{i-1}$, X_i is distributed according to $(1-p)\delta_{x_{i-1}} + pf_c$ where f_c denotes the pdf of an absolutely continuous distribution with bounded support, defined over an interval (l, u). In other words, at each time i, the process either makes a jump and takes a value drawn from distribution f_c, or it stays at X_{i-1}. The decision is made based on the outcome of an i.i.d. $\mathrm{Bern}\,(p)$ random variable independent of all past values of $\mathbf{X}$. While the output of this source is not sparse, it is clearly a structured process. This intuition is indeed captured by the ID of the process; it can be proved that $d_o(\mathbf{X}) = p$, i.e., the probability that the process makes a jump determines the complexity of this process [33].

For a stationary memoryless process, under some mild conditions on the distribution, [34] proves that the Rényi ID of the first order marginal distribution of the source characterizes the fundamental limits of compressed sensing. In other words, given a process $\mathbf{X}$, asymptotically, as the blocklength grows to infinity, the minimum number of linear projections, m, normalized by the ambient dimension, n, that is required for recovering source X^n from its linear projections is shown to be equal to $d(X_1)$, which is the Rényi ID of X_1. In [33], it is shown that asymptotically slightly more than $n\bar{d}_o(\mathbf{X})$ random linear projections suffice for *universal* recovery of X^n generated by any Markov process of any order, without knowing the source model, where $\bar{d}_o(\mathbf{X})$ denotes the upper ID of the process $\mathbf{X}$. These results provide an operational interpretation to the Rényi ID of a random variable and its generalization to stationary processes.

The focus of this paper is on the application of compression codes in building compressed sensing recovery algorithms. The rate-distortion function of a stationary source measures the minimum number of bits per source symbol required for achieving a given reconstruction quality. It turns out that for an analog process as the reconstruction becomes finer, the behavior the rate-distortion function is connected to the level of structuredness of the source process and ID notions mentioned earlier. In the rest of this section, we first review the known results on this connection, and then prove our main result of this section, which, under some mild conditions, establishes this connection for general stationary processes.

Consider a metric space $(\mathbb{R}^k, \rho)$, and random vector X^k. The rate-distortion function of X^k under expected

distortion constraint

$$d(x^k, \hat{x}^k) = \rho(x^k, \hat{x}^k)^r$$

is defined as

$$R_r(X^k, D) = \inf_{\mathrm{E}[d(X^k, \hat{X}^k)] \leq D} I(X^k; \hat{X}^k).$$

Defintion 3 (Rate-distortion dimension (RDD) of a random vector [32])**.** *The upper and lower RDDs of X^k are defined as*

$$\overline{\dim}_R(X^k) = r \limsup_{D \to 0} \frac{R_r(X^k, D)}{\log \frac{1}{D}},$$

and

$$\underline{\dim}_R(X^k) = r \liminf_{D \to 0} \frac{R_r(X^k, D)}{\log \frac{1}{D}},$$

respectively. If $\overline{\dim}_R(X^k) = \underline{\dim}_R(X^k)$, the RDD of X^n is defined as $\dim_R(X^k) = r \lim_{D \to 0} \frac{R_r(X^k, D)}{\log \frac{1}{D}}$.

The following theorem from [32] establishes the connection between the Rényi ID of a random vector X^k and its RDD, for any general distribution on X^k.

Theorem 3 (Proposition 3.3 in [32])**.** *Consider the metric space $(\mathbb{R}^k, \rho)$, such that there exists $0 < a_1 \leq a_2 < \infty$ for which $a_1 \max_{i=1}^k |x_i - \hat{x}_i| \leq \rho(x^k, \hat{x}^k) \leq a_2 \max_{i=1}^k |x_i - \hat{x}_i|$, for all $x^k, \hat{x}^k \in \mathbb{R}^k$. Then, for any distribution of X^k,*

$$\overline{\dim}_R(X^k) = \bar{d}(X^k),$$

and

$$\underline{\dim}_R(X^k) = \underline{d}(X^k),$$

where $\overline{\dim}_R(X^k)$, and $\underline{\dim}_R(X^k)$ denote the upper and lower RDD of X^k under fidelity constraint $d(x^k, \hat{x}^k) = \rho(x^k, \hat{x}^k)^r$.

Consider an analog stationary process $\mathbf{X} = \{X_i\}_{i=-\infty}^{\infty}$. The rate-distortion function $R(\mathbf{X}, D)$ of the source $\mathbf{X}$ under squared error distortion can be computed as [56], [57]

$$R(\mathbf{X}, D) = \lim_{m \to \infty} R^{(m)}(\mathbf{X}, D),$$

where

$$R^{(m)}(\mathbf{X}, D) = \inf_{\mathrm{E}[d_m(X^m, \hat{X}^m)] \leq D} \frac{1}{m} I(X^m; \hat{X}^m).$$

and

$$d_m(x^m, \hat{x}^m) = \frac{1}{m} \|x^m - \hat{x}^m\|_2^2. \tag{2}$$

Note that with this distortion metric, we have $r = 2$ and $R^{(m)}(\mathbf{X}, D) = \frac{1}{m} R_2(X^m, D)$. It can also be shown

that $\inf_m R^{(m)}(\mathbf{X}, D) = R(\mathbf{X}, D)$ [57].

Defintion 4 (RDD of a stationary process)**.** *The upper and lower RDDs of this stationary process* $\mathbf{X}$ *can be defined as*

$$\overline{\dim}_R(\mathbf{X}) = 2 \limsup_{D\to 0} \frac{R(\mathbf{X}, D)}{\log \frac{1}{D}}$$

and

$$\underline{\dim}_R(\mathbf{X}) = 2 \liminf_{D\to 0} \frac{R(\mathbf{X}, D)}{\log \frac{1}{D}}.$$

If $\overline{\dim}_R(\mathbf{X}) = \underline{\dim}_R(\mathbf{X})$*, then* $\dim_R(\mathbf{X}) = \overline{\dim}_R(\mathbf{X}) = \underline{\dim}_R(\mathbf{X})$ *is the RDD of* $\mathbf{X}$*.*

The main result of this section is the following theorem which extends the equivalence of Rényi ID and RDD shown in [32] for i.i.d. random vectors to stationary processes.

Theorem 4. *For a stationary process* $\mathbf{X} = \{X_i\}_{i=-\infty}^{\infty}$*, assume that* $\lim_{D\to 0} \frac{R^{(m)}(\mathbf{X},D)}{\log \frac{1}{D}}$ *exists for all* m*. Then,*

$$\dim_R(\mathbf{X}) = \bar{d}_o(\mathbf{X}).$$

The main ingredients of the proof of Theorem 4 are the following two lemmas.

Lemma 1. *For any stationary process* $\mathbf{X}$*, we have*

$$\overline{\dim}_R(\mathbf{X}) \le \bar{d}_o(\mathbf{X}) \le \inf_m 2\Big(\limsup_{D\to 0} \frac{R^{(m)}(\mathbf{X}, D)}{\log \frac{1}{D}} \Big).$$

Lemma 2. *Assume that* $\lim_{D\to 0} \frac{R^{(m)}(\mathbf{X},D)}{\log \frac{1}{D}}$ *exists for all* m*, and also there exists* $\sigma^2_{\max} > 0$*, such that* $R^{(m)}(\mathbf{X}, D)$ *uniformly converges to* $R(\mathbf{X}, D)$*, for* $D \in (0, \sigma^2_{\max})$*, as* m *grows to infinity. Then,* $\dim_R(\mathbf{X}) = \bar{d}_o(\mathbf{X})$*.*

Proofs of Theorem 4 and Lemmas 1 and 2 are provided in Section VI.

To illustrate the relationship between RDD and ID, as an example, consider the piecewise-constant signal described earlier. To directly evaluate the RDD of this process, its rate-distortion characterization is required. However, deriving the rate-distortion function of sources with memory is in general very challenging. For instance, even for the binary symmetric Markov chain, the rate-distortion function is not known, except in a low-distortion region [58], and we have to resort to upper and lower bounds in general [59], [60]. The following theorem provides upper and lower bounds on the $R(\mathbf{X}, D)$ of the piecewise-constant source. While there is a gap between the bounds on the $R(\mathbf{X}, D)$, since the gap does not depend on D, as shown in the following corollary, they can be used to evaluate the RDD of the source exactly.

Theorem 5. *Consider a first-order stationary Markov process* $\mathbf{X} = \{X_i\}_{i=0}^{\infty}$*, such that conditioned on* $X_{i-1} = x_{i-1}$*,* X_i *is distributed according to* $(1-p)\delta_{x_{i-1}} + pf_c$*, where* f_c *denotes the pdf of an absolutely continuous distribution*

with bounded support, (l,u)*. If* $d_{\max} \triangleq \sup_{x,\hat{x}\in(l,u)} d(x,\hat{x}) < \infty$*, then*

$$pR_{f_c}(D) \leq R(\mathbf{X},D) \leq H(p) + pR_{f_c}(D),$$

where $R_{f_c}(D)$ *and* $H(p)$ *denote the rate distortion function of an i.i.d. process distributed according to pdf* f_c*, and the binary entropy function (*$-p\log_2 p-(1-p)\log_2(1-p)$*), respectively.*

Proof: A detailed proof of Theorem 5 is presented in Section VI. To prove the upper bound (achievability), we consider a code that describes the positions of the jumps losslessly at rate $H(p)$. Since the source is piecewise constant, after describing the positions of the jumps, the encoder removes the repeated values and applies a lossy compression code of blocklength length close to np. Therefore, to describe the values at distortion D the encoder roughly needs to spend $npR_{f_c}(D)$ bits. For the lower bound (converse), we consider a genie-aided decoder that has access to the positions of the jumps. Then intuitively, to describe the values at distortion D, it still needs a rate of at least $pR_{f_c}(D)$. The proof in Section VI makes these steps formal by properly analyzing the reduced block length which is a random number. ■

Corollary 1. *For the piecewise constant source in Theorem 5, we have*

$$\dim_R(\mathbf{X}) = \bar{d}_o(\mathbf{X}) = p.$$

In other words, the RDD is equal to p *which is in turn equal to the ID of this source.*

Proof: Given the bound on the rate-distortion process derived in Theorem 5, it is easy to directly derive the RDD of such a source. More precisely, given the upper bound, it follows that

$$\begin{aligned}
\overline{\dim}_R(\mathbf{X}) &= 2\limsup_{D\to 0}\frac{R(\mathbf{X},D)}{\log\frac{1}{D}} \\
&\leq 2\limsup_{D\to 0}\frac{H(p)+pR_{f_c}(D)}{\log\frac{1}{D}} \\
&= p(\limsup_{D\to 0}\frac{R_{f_c}(D)}{\log\frac{1}{D}}) \\
&= p.
\end{aligned}$$

Similarly, given the lower bound, we have

$$\begin{aligned}
\underline{\dim}_R(\mathbf{X}) &= 2\liminf_{D\to 0}\frac{R(\mathbf{X},D)}{\log\frac{1}{D}} \\
&\geq 2\liminf_{D\to 0}\frac{pR_{f_c}(D)}{\log\frac{1}{D}} \\
&= p(\liminf_{D\to 0}\frac{R_{f_c}(D)}{\log\frac{1}{D}}) \\
&= p,
\end{aligned}$$

where the last lines in both the upper and the lower RDDs follow from [32] and [55]. Therefore, $p \leq \underline{\dim}_R(\mathbf{X}) \leq \overline{\dim}_R(\mathbf{X}) \leq p$. In other words, for this source RDD exists and is equal to $\dim_R(\mathbf{X}) = p$. Hence, the condition of Theorem 4 holds and we have

$$\dim_R(\mathbf{X}) = \bar{d}_o(\mathbf{X}).$$

This agrees with the ID of this source found in Theorem 2 in [33],

$$\bar{d}_o(\mathbf{X}) = \underline{d}_o(\mathbf{X}) = p.$$

■

Remark 2. Corollary 1 states that the RDD of the piecewise constant source described in Theorem 5 is equal to p, which is also the ID of this process [33]. While [33] directly computes the ID of such processes, Theorem 4, by proving the equivalence of ID and RDD, provides a potentially easier alternative path to computing the ID of stochastic processes. Note that to be able to calculate the RDD of a process, the exact characterization of the rate-distortion function is not required. In fact, it is easy to see that it would be enough to have upper and lower bounds on the rate-distortion function of the source, $R(\mathbf{X}, D)$, that are within a reasonable gap. More precisely, as long as the gap between the bounds grows as $o(\log \frac{1}{D})$, they can be used to evaluate the RDD. Moreover, since the RDD only depends on the low-distortion behavior of the rate-distortion function, studying its asymptotic small distortion performance is sufficient for computing the RDD and as a result the ID of a source, without knowing the rate-distortion function explicitly. For instance, [61] studies the asymptotic behavior of the rate-distortion function of some stochastic sources and employs those results to evaluate the RDD of some i.i.d. processes.

IV. Almost lossless recovery

Section II formulated the performance of the CSP algorithm which employs a lossy compression code to recover the output of a stationary process from random linear projections. Specifically, Theorem 1 and Theorem 2 characterize the performance of the CSP algorithm, for noiseless and noisy measurements, respectively. In this section, we focus on the special case in which the lossy compression code is a high-resolution one, and therefore, D is very small. As a result, with high probability, the CSP algorithm generates a high-fidelity or almost lossless reconstruction of the input vector. While the CSP algorithm is inherently a lossy CS recovery algorithm due to the utilized lossy compression code, the almost lossless recovery performance can be achieved by letting the distortion of the compression code become arbitrarily small. To build the analytical tools and insights required to evaluate the CSP performance when the distortion approaches zero, in Section III, we focused on measures of complexity for stationary analog processes, and established a connection with the RDD of a stationary process and its ID.

In a noiseless setting, Theorem 1 asserted that given a compression code operating at rate R and distortion D, the CSP algorithm is able to recover signal X^n from $\frac{2\eta R}{\log(\frac{1}{D})}$ randomized linear measurements. Note that the RDD of the source was defined in Section III as $\lim_{D\to 0} \frac{2R(\mathbf{X},D)}{\log(\frac{1}{D})}$. Therefore, considering a family of optimal compression codes that operate at a very low distortion level, Theorem 1 predicts that, asymptotically, if the normalized number

of measurements is slightly higher than the RDD of the source, then the CSP algorithm generates an almost lossless reconstruction. This result is formalized in the following corollary, which studies the performance of the CSP algorithm in the extreme case, where D approaches zero. It proves that as long as the normalized number of measurements is larger than $\overline{\dim}_R(\mathbf{X})$, CSP recovers the source vector almost losslessly.

Corollary 2. *Consider a stationary process* $\mathbf{X}$ *and a system of random linear observations,* $Y^m = AX^n$*, with measurement matrix* $A \in \mathbb{R}^{m\times n}$*, where* $A_{i,j}$ *are i.i.d. as* $\mathcal{N}(0,1)$*. For any observation error* $\Delta > 0$*, if the number of measurements* $m = m_n$ *satisfies*

$$\liminf_{n\to\infty} \frac{m_n}{n} > \overline{\dim}_R(\mathbf{X}),$$

then there exists a family of compression codes which, when used by the CSP algorithm, yields

$$\lim_{n\to\infty} \mathrm{P}(\frac{1}{\sqrt{n}}\| X^n - \tilde{X}^n\|_2 \geq \Delta) \to 0,$$

where $\tilde{X}^n$ *refers to the solution of the CSP algorithm as in* (1)*.*

Proof: The proof is provided in Section VI. ■

Applying Corollary 2 to the piecewise constant source described in Theorem 5, implies that the almost lossless performance of the CSP algorithm for such a source depends on the RDD of this source. Corollary 1 shows that the for the piecewise constant source we have $\dim_R(\mathbf{X}) = \bar{d}_o(\mathbf{X}) = p$. Combining this together with Corollary 2 implies that there exists a family of compression codes that when employed by the CSP algorithm with a number of measurements satisfying $\liminf_{n\to\infty} \frac{m_n}{n} > p$ yields an asymptotic almost lossless recovery of this source.

Remark 3. Corollary 2 states that the CSP algorithm can achieve almost lossless recovery, using slightly more than $n\overline{\dim}_R(\mathbf{X})$ random linear measurements. On the other hand, for i.i.d. sources, under some mild conditions, $nd_o(\mathbf{X})$ characterizes the minimum required number of measurements for almost lossless recovery [34]. Note that if the rate-distortion function of the source satisfies the condition of Theorem 4, then $\dim_R(\mathbf{X}) = \bar{d}_o(\mathbf{X})$. Even without such assumption on the rate-distortion function, we can employ Lemma 1 to upper bound $\overline{\dim}_R(\mathbf{X})$ by $\bar{d}_o(\mathbf{X})$ and get the same result. Therefore, at least for memoryless i.i.d. sources, the CSP algorithm achieves the optimal performance, in terms of achieving the minimum number of measurements.

For general stationary sources, if the decoder is restricted to be Lipschitz-continuous, which is formally defined below, then asymptotically the normalized number of measurements should be larger than $\limsup_{n\to\infty} \bar{d}(X^n)/n$ [62], which is equal to $\bar{d}_o(\mathbf{X})$ [33]. While the CSP decoder is not Lipschitz continuous, we conjecture that the lower bound also holds for less-restricted decoders. Proving or disproving this is an interesting topic for future research.

Defintion 5 (Lipschitz continuity)**.** *Consider a set* $\mathcal{A} \subset \mathbb{R}^k$*. A function* $f : \mathcal{A} \to \mathbb{R}^n$ *is called Lipschitz continuous*

if there exists constant $c \in \mathbb{R}$, such that

$$\|f(x) - f(y)\| \leq c\|x - y\|,$$

for all $x, y \in \mathcal{A}$.

V. From universal compression to universal compressed sensing

In compressed sensing, the decoder tries to find the signal that matches the measurements and also has the same structure as the unknown input signal. In many applications, the structure of the input signal is not known by the decoder or is known only partially. For instance, for an image, the decoder might know that the wavelet coefficients of the image are sparse, but the image might have much more structure that is not known by the decoder. Moreover, in many application, it is desired to have decoders that work well for sources with different statistics. In summary, from a practical viewpoint, it appealing to have decoders that take advantage of all the information contained in the structure of the signal, without having any prior knowledge about the source distribution, Such decoders, potentially, lead to very efficient compressed sensing algorithms that work for various source models.

A universal compressed sensing decoder aims at recovering an input signal from its under-determined linear measurements, without having access to the source distribution or the source model. The existence of such universal recovery algorithms is known for both deterministic [47] and stochastic [33], [48], [49] settings. In this section, we prove that a family of universal compression codes combined by the CSP algorithm leads to a family of universal compressed sensing recovery algorithms.

Consider a family of variable-length point-wise universal lossy compression codes (n, f_n, g_n) for analog stationary ergodic processes with alphabet $\mathcal{X} \subset \mathbb{R}$. Assume that the family of codes (n, f_n, g_n) operates at fixed distortion D. That is, for any stationary ergodic process $\mathbf{X} = \{X_i\}_i$, with $X_i \in \mathcal{X}$,

i) $\lim_{n\to\infty} \frac{1}{n}|f_n(X^n)| = R(\mathbf{X}, D)$, almost surely,

ii) $\lim_{n\to\infty} \mathrm{P}(\frac{1}{n}\|X^n - \hat{X}^n\|_2^2 \geq D + \epsilon) = 0$,

for any $\epsilon > 0$.

Consider X^n generated by a stationary ergodic process $\mathbf{X}$ with rate-distortion function $R(\mathbf{X}, D)$. A universal compressed sensing decoder observes $Y^m = AX^n$, and aims at estimating X^n from Y^m, employing the code (f_n, g_n), without having access to the distribution of the source. To achieve this goal, consider the following slightly modified version of the CSP algorithm, which we refer to as universal CSP (UCSP):

$$\begin{aligned} \min \quad & \|Au^n - Y^m\|_2 \\ \text{s.t.} \quad & u^n = g_n(b), \\ & b \in \{0,1\}^*, |b| \leq n(R(\mathbf{X}, D) + \epsilon). \end{aligned} \tag{3}$$

In other words, among all binary sequences of length smaller than $n(R(\mathbf{X}, D) + \epsilon)$, UCSP searches for the one whose decompressed version via the universal decoder g_n yields the smallest measurement error.

For $A_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, let $\tilde{X}^n$ denote the minimizer of the UCSP algorithm that employs a point-wise universal compression code operating at distortion D. The following theorem characterizes the performance of the UCSP algorithm and proves that a universal compression code leads to a universal compressed sensing algorithm.

Theorem 6. *Consider $Y^m = AX^n$, a system of random linear observations with measurement matrix $A \in \mathbb{R}^{m\times n}$, where $A_{i,j} \sim$ i.i.d $\mathcal{N}(0,1)$. Let $\mathcal{C}_n$ be variable-length point-wise universal lossy compression code operating at rate R that achieves distortion D with excess distortion probability ϵ. For $\alpha > 0$, $\epsilon > 0$, and $\eta > 1$, such that $\frac{\eta}{\log\frac{1}{D}} + \alpha > \epsilon$, let $\delta = \frac{\eta}{\log\frac{1}{D}} + \alpha - \epsilon$. Suppose $\tilde{X}^n$ refers to the solution of the UCSP algorithm as in* (3)*. Then, for n large enough, and*

$$\frac{m}{n} = \frac{2\eta R(\mathbf{X}, D)}{\log\frac{1}{D}},$$

we have

$$\begin{aligned}\mathrm{P}\Big(\frac{1}{\sqrt{n}}\| X^n - \tilde{X}^n\|_2 \geq (2+\sqrt{n/m})D^{\frac{1}{2}(1-\frac{1+\delta}{\eta})}\Big)\\ \leq \epsilon + 2^{-\frac{1}{2}nR\alpha} + \mathrm{e}^{-\frac{m}{2}}.\end{aligned}$$

Proof: The proof is provided in Section VI. ■

Comparing Theorem 6 with Theorem 1, it can be observed that the performance trade-offs for the CSP and the UCSP are exactly the same in terms of the rate-distortion behavior of the underlying compression code. The difference between the two is in the fact that the CSP optimization employs a compression code that is designed for input source distribution, but the UCSP optimization requires a universal compression code. This might suggest that since UCSP has then same asymptotic performance as CSP, and in addition works for any input distribution, it is always a better choice than the CSP. However, note that UCSP is build upon point-wise universal compression codes for analog sources. While such codes theoretically exist, practical instances of such codes are yet to be found. Moreover, another potential disadvantage of the UCSP compared to the CSP optimization is that while universal codes usually achieve the same asymptotic performance as non-universal codes, their finite blocklength performance is worse than non-universal codes.

Similar to Corollary 2, the following corollary considers the special case where the distortion approaches zero, and proves that, as long as the normalized number of measurements is larger than $\overline{\dim}_R(\mathbf{X})$, there exist universal compression codes that yield universal compressed sensing algorithms that can estimate the source almost losslessly. Note that $\overline{\dim}_R(\mathbf{X})$ is the RDD of the source $\mathbf{X}$, which depends on the source model and captures all the structure within the signal.

Corollary 3. *Consider a stationary process $\mathbf{X}$ and a system of random linear observations, $Y^m = AX^n$, with measurement matrix $A \in \mathbb{R}^{m\times n}$, where $A_{i,j}$ are i.i.d. as $\mathcal{N}(0,1)$ and $m = m_n$ is the number of observations. For any observation error $\Delta > 0$, if the sequence m_n satisfies*

$$\liminf_{n\to\infty} \frac{m_n}{n} > \overline{\dim}_R(\mathbf{X}),$$

then there exists a family of variable-length point-wise universal lossy compression codes which, when used by the UCSP algorithm, yields

$$\lim_{n\to\infty} \mathrm{P}\left(\frac{1}{\sqrt{n}}\| X^n - \tilde{X}^n\|_2 \geq \Delta\right) \to 0.$$

where $\tilde{X}^n$ refers to the solution of the UCSP algorithm as in (3)*.*

Proof: The proof is very similar to the proof of Corollary 2 and is omitted. ■

VI. Proofs

The following lemma from [30] is used in some of the proofs.

Lemma 3 (χ^2-construction)**.** *Fix $\tau > 0$ and let $Z_i \sim \mathcal{N}(0,1)$, $i = 1, 2, \ldots, m$. Then,*

$$\mathrm{P}\left(\sum_{i=1}^{m} Z_i^2 < m(1-\tau)\right) \leq \mathrm{e}^{\frac{m}{2}(\tau+\ln(1-\tau))}$$

and

$$\mathrm{P}\left(\sum_{i=1}^{m} Z_i^2 > m(1+\tau)\right) \leq \mathrm{e}^{-\frac{m}{2}(\tau-\ln(1+\tau))}.$$

A. Proof of Theorem 1

Let $\hat{X}^n = g_n(f_n(X^n))$. Since $\tilde{X}^n = \arg\min_{x^n \in \mathcal{C}_n} \|Y^m - Ax^n\|_2^2$, and $\hat{X}^n \in \mathcal{C}_n$,

$$\|Y^m - A\tilde{X}^n\|_2 \leq \|Y^m - A\hat{X}^n\|_2.$$

Substituting AX^n for Y^m, it follows that

$$\|A(X^n - \tilde{X}^n)\|_2 \leq \|A(X^n - \hat{X}^n)\|_2. \tag{4}$$

Define the event $\mathcal{E}_0$ as

$$\mathcal{E}_0 \triangleq \{\|X^n - \hat{X}^n\|_2^2 \leq nD\}.$$

By assumption, $\mathrm{P}(\mathcal{E}_0^c) \leq \epsilon$. Conditioned on $\mathcal{E}_0$, from (4), we have

$$\|A(X^n - \tilde{X}^n)\|_2 \leq \sigma_{\max}(A)\sqrt{nD} \tag{5}$$

where $\sigma_{\max}(A)$ is the maximum singular value of A. Define events $\mathcal{E}_1$ and $\mathcal{E}_2$ as

$$\mathcal{E}_1 \triangleq \big\{\forall \tilde{x}^n \in \mathcal{C} : \|A(X^n - \tilde{x}^n)\|_2 \geq \sqrt{(1-\tau)m}\|X^n - \tilde{x}^n\|_2\big\},$$

where $\tau \in (0,1)$, and

$$\mathcal{E}_2 \triangleq \{\sigma_{\max}(A) - \sqrt{m} - \sqrt{n} < \sqrt{m}\}.$$

Then, conditioned on $\mathcal{E}_0 \cap \mathcal{E}_1 \cap \mathcal{E}_2$, it follows from (5) that

$$\begin{aligned}\sqrt{m(1-\tau)}\|X^n - \tilde{X}^n\|_2 &\leq \|A(X^n - \tilde{X}^n)\|_2 \\ &\leq \sigma_{\max}(A)\sqrt{nD} \\ &\leq (\sqrt{n} + 2\sqrt{m})\sqrt{nD}. \end{aligned} \tag{6}$$

Rearranging the terms and setting $m = \frac{2\eta n R}{\log(1/D)}$ and $\tau = 1 - D^{(1+\delta)/\eta}$ in (6) yields

$$\frac{1}{\sqrt{n}}\|X^n - \tilde{X}^n\|_2 \leq \sqrt{\frac{D}{1-\tau}}\left(\sqrt{\frac{n}{m}} + 2\right) \tag{7}$$

$$=\sqrt{\frac{D}{D^{(1+\delta)/\eta}}}\left(\sqrt{\frac{n}{m}} + 2\right)$$

$$=D^{0.5(1-(1+\delta)/\eta)}\left(\sqrt{\frac{n}{m}} + 2\right). \tag{8}$$

The inequality in (8) holds with probability $\mathrm{P}(\mathcal{E}_0 \cap \mathcal{E}_1 \cap \mathcal{E}_2)$. In the last step of the proof, a lower bound on this probability or equivalently an upper bound on $\mathrm{P}(\mathcal{E}_0^c \cup \mathcal{E}_1^c \cup \mathcal{E}_2^c)$ is derived.

Fixing $X^n = x^n$ and $\tilde{x}^n$, $A(x^n - \tilde{x}^n)/\|x^n - \tilde{x}^n\|_2$ is a vector of i.i.d. $\mathcal{N}(0,1)$ random variables. Therefore, by Lemma 3,

$$\begin{aligned}&\mathrm{P}_A\left(\|A(x^n - \tilde{x}^n)\|_2 \leq \sqrt{(1-\tau)m}\|x^n - \tilde{x}^n\|_2\right) \\ &\quad\leq \mathrm{e}^{\frac{m}{2}(\tau + \ln(1-\tau))},\end{aligned}$$

and by the union bound, for a fixed $X^n = x^n$,

$$\begin{aligned}&\mathrm{P}_A(\exists \tilde{x}^n \in \mathcal{C} :\ \|A(x^n - \tilde{x}^n)\|_2 \leq \sqrt{(1-\tau)m}\|x^n - \tilde{x}^n\|_2) \\ &\quad\leq 2^{nR}\mathrm{e}^{\frac{m}{2}(\tau + \ln(1-\tau))} \\ &\quad= 2^{nR + \frac{m}{2}(\tau \log \mathrm{e} + \log(1-\tau))}.\end{aligned} \tag{9}$$

Taking the expected value of the both sides of (9) with respect to X^n, and noting that the right hand side of (9) is not random, it follows that

$$\begin{aligned}&\mathrm{E}_{X^n}\Big[\mathrm{P}_A(\exists \tilde{x}^n \in \mathcal{C} : \\ &\qquad\|A(X^n - \tilde{x}^n)\|_2 \leq \sqrt{(1-\tau)m}\|X^n - \tilde{x}^n\|_2)\Big] \\ &\quad\leq 2^{nR + \frac{m}{2}(\tau \log \mathrm{e} + \log(1-\tau))}.\end{aligned} \tag{10}$$

Rewriting $\mathrm{P}_A(\mathcal{E})$ as $\mathrm{E}_A[\mathbb{1}_{\mathcal{E}}]$ and employing Fubini's theorem to exchange the order of integration, it follows

from (10) that

$$\begin{aligned}
&\mathrm{E}_A\Big[\,\mathrm{P}_{X^n}(\exists \tilde{x}^n \in \mathcal{C}: \\
&\qquad \|A(X^n-\tilde{x}^n)\|_2 \le \sqrt{(1-\tau)m}\|X^n-\tilde{x}^n\|_2)\Big] \\
&\le 2^{nR+\frac{m}{2}(\tau\log \mathrm{e}+\log(1-\tau))}. \qquad (11)
\end{aligned}$$

Substituting for m, the exponent in (11) can be upper-bounded as follows:

$$\begin{aligned}
&nR+\frac{m}{2}(\tau\log \mathrm{e}+\log(1-\tau)) \\
&\quad = nR\Big(1+\frac{\eta}{-\log D}(1-D^{(1+\delta)/\eta}+\frac{(1+\delta)}{\eta}\log D)\Big) \\
&\quad = nR\Big(1-(1+\delta)-\frac{\eta}{\log D}(1-D^{(1+\delta)/\eta})\Big) \\
&\quad \le nR(-\delta-\frac{\eta}{\log D}) \\
&\quad = -nR\alpha.
\end{aligned}$$

Define the function υ_n, where $\upsilon_n : \mathbb{R}^{m\times n} \to [0,1]$, as $\upsilon_n(A) \triangleq \mathrm{P}_{X^n}(\exists \tilde{x}^n \in \mathcal{C} : \ \|A(X^n-\tilde{x}^n)\|_2 \le \sqrt{(1-\tau)m}\|X^n-\tilde{x}^n\|_2)$. Note that in our model, m is also a function of n. We prove that $\upsilon(n)$ converges to zero, almost surely. By Markov's inequality, from (11) and (12), it follows that

$$\mathrm{P}(\upsilon_n(A) > 2^{-\frac{1}{2}nR\alpha}) \le \frac{\mathrm{E}[\upsilon_n(A)]}{2^{-\frac{1}{2}nR\alpha}} \le 2^{-\frac{1}{2}nR\alpha}. \qquad (12)$$

Therefore, by the Borel Cantelli Lemma, $\upsilon_n(A) < 2^{-\frac{1}{2}nR\alpha}$, eventually almost surely, and hence $\upsilon_n(A)$ converges to zero, almost surely. This results implies that with probability one $\mathrm{P}_{X^n}(\mathcal{E}_1^c)$ converges to zero.

Finally, to upper bound $\mathrm{P}(\mathcal{E}_2^c)$, from [63], by the concentration of Lipschitz functions of a Gaussian vector,

$$\mathrm{P}(\sigma_{\max}(A)-\sqrt{m}-\sqrt{n} \ge t\sqrt{m}) \le \mathrm{e}^{-mt^2/2}. \qquad (13)$$

Letting $t=1$ in (13), it follows that

$$\begin{aligned}
\mathrm{P}(\mathcal{E}_2^c) &= \mathrm{P}(\sigma_{\max}(A)-\sqrt{m}-\sqrt{n} \ge \sqrt{m}) \\
&\le \mathrm{e}^{-m/2}. \qquad (14)
\end{aligned}$$

B. Proof of Theorem 2

Let $\hat{X}^n = g_n(f_n(X^n))$. Similar to the proof of Theorem 1, since $\tilde{X}^n = \arg\min_{x^n\in\mathcal{C}_n}\|Y^m - Ax^n\|_2^2$, and $\hat{X}^n \in \mathcal{C}_n$, $\|Y^m - A\tilde{X}^n\|_2 \le \|Y^m - A\hat{X}^n\|_2$. Substituting $Y^m = AX^n + Z^m$, we have

$$\begin{aligned}
&\|AX^n - A\tilde{X}^n\|_2 - \|Z^m\|_2 \\
&\le \|AX^n - A\hat{X}^n\|_2 + \|Z^m\|_2,
\end{aligned}$$

or

$$\|AX^n - A\tilde{X}^n\|_2 \leq \|AX^n - A\hat{X}^n\|_2 + 2\|Z^m\|_2. \tag{15}$$

Define events $\mathcal{E}_0$, $\mathcal{E}_1$ and $\mathcal{E}_2$ as in the proof of Theorem 1 and $\mathcal{E}_3 \triangleq \{\frac{1}{\sqrt{m}}\|Z^m\|_2 \leq \sigma_m\}$. Following similar steps as before, conditioned on $\mathcal{E}_0 \cap \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$, we have

$$\begin{aligned}\sqrt{m(1-\tau)}\|X^n - \tilde{X}^n\|_2 &\leq (\sqrt{n} + 2\sqrt{m})\sqrt{nD} + 2\|Z^m\|_2 \\ &\leq (\sqrt{n} + 2\sqrt{m})\sqrt{nD} + 2\sigma_m\sqrt{m}. \end{aligned}\tag{16}$$

The rest of the proof follows by setting $\tau = 1 - D^{(1+\delta)/\eta}$, and substituting the values of the parameters in (16).

C. Proof of Lemma 1

Given k, define distance measure ρ_k such that for $x^k, \hat{x}^k \in \mathbb{R}^k$, $\rho_k(x^k, \hat{x}^k) \triangleq \sqrt{kd_k(x^k, \hat{x}^k)}$ where $d_k(\cdot,\cdot)$ is defined in (2). Note that $(\mathbb{R}^k, \rho_k)$ is a metric space. Furthermore, since $\max_{i=1}^k |x_i - \hat{x}_i| \leq \rho_k(x^k, \hat{x}^k) \leq \sqrt{k}\max_{i=1}^k |x_i - \hat{x}_i|$, from Theorem 3,

$$2\limsup_{D\to 0} \frac{kR^{(k)}(\mathbf{X}, \frac{D}{k})}{\log\frac{1}{D}} = \bar{d}(X^k).$$

By a change of variable, $2\limsup_{D\to 0} \frac{kR^{(k)}(\mathbf{X},D)}{\log\frac{1}{D}+\log\frac{1}{k}} = \bar{d}(X^k)$, or

$$2\limsup_{D\to 0} \frac{R^{(k)}(\mathbf{X}, D)}{\log\frac{1}{D}} = \frac{1}{k}\bar{d}(X^k).$$

Taking the limit of both sides as k grows to infinity, and employing Lemma 2 from [33], which shows that the upper ID of a process $\mathbf{X}$ can be alternatively be represented as

$$\bar{d}_o(\mathbf{X}) = \lim_{k\to\infty}\frac{1}{k}\left(\limsup_{b\to\infty}\frac{H([X^k]_b)}{b}\right),$$

yields

$$\begin{aligned}\lim_{k\to\infty}\left(2\limsup_{D\to 0}\frac{R^{(k)}(\mathbf{X},D)}{\log\frac{1}{D}}\right) &= \lim_{k\to\infty}\frac{1}{k}\bar{d}(X^k) \\ &= \bar{d}_o(\mathbf{X}).\end{aligned}\tag{17}$$

Since $R^{(k)}(\mathbf{X}, D) \geq \inf_m R^{(m)}(\mathbf{X}, D)$, from (17),

$$\begin{aligned}\bar{d}_o(\mathbf{X}) &\geq \lim_{k\to\infty}\left(2\limsup_{D\to 0}\frac{\inf_m R^{(m)}(\mathbf{X},D)}{\log\frac{1}{D}}\right) \\ &\overset{(a)}{=} \lim_{k\to\infty}\left(2\limsup_{D\to 0}\frac{R(\mathbf{X},D)}{\log\frac{1}{D}}\right) = \overline{\dim}_R(\mathbf{X}),\end{aligned}$$

where (a) follows from the fact that $R(\mathbf{X}, D) = \inf_m R^{(m)}(\mathbf{X}, D)$ [57]. This proves the lower bound in the desired result.

To prove the upper bound, fix a positive integer $m \in \mathbb{N}$. Any integer k can be written as $k = sm + r$, where $r \in \{0, \ldots, m-1\}$. Since $kR^{(k)}(\mathbf{X}, D)$ is a sub-additive sequence [57], $kR^{(k)}(\mathbf{X}, D) \leq smR^{(m)}(\mathbf{X}, D) + rR^{(r)}(\mathbf{X}, D)$, or

$$R^{(k)}(\mathbf{X}, D) \leq \frac{sm}{k} R^{(m)}(\mathbf{X}, D) + \frac{r}{k} R^{(r)}(\mathbf{X}, D). \tag{18}$$

Combining (17) and (18), it follows that

$$\begin{aligned}
\bar{d}_o(\mathbf{X}) &\leq 2 \lim_{k\to\infty} \left(\limsup_{D\to 0} \frac{sm}{k} \frac{R^{(m)}(\mathbf{X}, D)}{\log \frac{1}{D}} \right) \\
&\quad + 2 \lim_{k\to\infty} \left(\limsup_{D\to 0} \frac{r}{k} \frac{R^{(r)}(\mathbf{X}, D)}{\log \frac{1}{D}} \right) \\
&= 2 \lim_{k\to\infty} \left(\frac{sm}{k} \right) \left(\limsup_{D\to 0} \frac{R^{(m)}(\mathbf{X}, D)}{\log \frac{1}{D}} \right) \\
&\quad + 2 \lim_{k\to\infty} \left(\frac{r}{k} \right) \left(\limsup_{D\to 0} \frac{R^{(r)}(\mathbf{X}, D)}{\log \frac{1}{D}} \right) \\
&= 2 \left(\limsup_{D\to 0} \frac{R^{(m)}(\mathbf{X}, D)}{\log \frac{1}{D}} \right).
\end{aligned} \tag{19}$$

Since m is selected arbitrarily, we can take infimum of the right hand side of (19) and derive the desired result.

D. Proof or Lemma 2

By the lemma's assumption, $\overline{\dim}_R(\mathbf{X}) = \dim_R(\mathbf{X})$; therefore, from Lemma 1,

$$\dim_R(\mathbf{X}) \leq \bar{d}_o(\mathbf{X}) \leq 2 \Big(\lim_{D\to 0} \frac{R^{(m)}(\mathbf{X}, D)}{\log \frac{1}{D}} \Big), \tag{20}$$

for all m. Given the uniform convergence assumption, for any $\epsilon > 0$, there exists $m_\epsilon \in \mathbb{N}$, such that for all $m > m_\epsilon$,

$$\left| \frac{R^{(m)}(\mathbf{X}, D)}{\log \frac{1}{D}} - \frac{R(\mathbf{X}, D)}{\log \frac{1}{D}} \right| < \epsilon, \tag{21}$$

for all $D \in (0, \sigma^2_{\max})$.

On the other hand, for any $\epsilon' > 0$ and m, there exists $\delta_{\epsilon',m} > 0$, such that for all $D \in (0, \delta_{\epsilon',m})$,

$$\lim_{D\to 0} \frac{R^{(m)}(\mathbf{X}, D)}{\log \frac{1}{D}} \leq \frac{R^{(m)}(\mathbf{X}, D)}{\log \frac{1}{D}} + \epsilon'. \tag{22}$$

Also, for any $\epsilon'' > 0$, there exists $\delta_{\epsilon''} > 0$, such that for all $D \in (0, \delta_{\epsilon''})$,

$$\frac{R(\mathbf{X}, D)}{\log \frac{1}{D}} \leq \frac{1}{2} \left(\dim_R(\mathbf{X}) + \epsilon'' \right). \tag{23}$$

Therefore, for any ϵ, ϵ' and ϵ'', choosing $m > m_\epsilon$, and $D \in (0, \min(\delta_{\epsilon',m}, \delta_{\epsilon''}))$, and combining (21), (22) and (23) yields

$$\bar{d}_o(\mathbf{X}) \leq \dim_R(\mathbf{X}) + \epsilon + \epsilon' + \epsilon''. \tag{24}$$

Since ϵ, ϵ' and ϵ'' are selected arbitrarily, combining (20) and (24) proves that $\dim_R(\mathbf{X}) = \bar{d}_o(\mathbf{X})$.

E. Proof of Theorem 4

It is shown in [64] that for any stationary process $\mathbf{X}$

$$|R^{(m)}(\mathbf{X}, D) - R(\mathbf{X}, D)| \leq \frac{1}{m} I(X^m; X_{-\infty}^0). \tag{25}$$

Note that while some of the results in [64] only hold for sources that are either absolutely continuous or discrete, as shown in Appendix A, this bound holds for general sources. Since the right hand side of (25) does not depend on D, it shows that $R^{(m)}(\mathbf{X}, D)$ uniformly converges to $R(\mathbf{X}, D)$ for all $D > 0$. On the other hand, for any $0 < \sigma_{\max} < 1$, and any $D \in (0, \sigma_{\max}^2)$, $0 < 1/\log\frac{1}{D} < 1/\log\frac{1}{\sigma_{\max}^2}$. Therefore, $\frac{R^{(m)}(\mathbf{X},D)}{\log\frac{1}{D}}$ uniformly converges to $\frac{R(\mathbf{X},D)}{\log\frac{1}{D}}$, for $D \in (0, \sigma_{\max}^2)$, and by Lemma 2, $\dim_R(\mathbf{X}) = \bar{d}_o(\mathbf{X})$.

F. Proof of Theorem 5

Let X^n denote the output of the source. Given the source model, X^n can be written as

$$X^n = \underbrace{S_1, \ldots, S_1}_{T_1}, \underbrace{S_2, \ldots, S_2}_{T_2}, \ldots, \underbrace{S_N, \ldots, S_N}_{T_N},$$

where $S_1, S_2, \ldots, S_N$ are i.i.d. distributed according to f_c, and $\sum_{i=1}^N T_i = n$. Moreover, $T_1, \ldots, T_{N-1}$ are i.i.d. distributed geometric random variables with parameter p. That is, for $i = 1, \ldots, N-1$ and $m \geq 1$, $\mathrm{P}(T_i = m) = (1-p)^{m-1}p$.

1) Converse: Assume that the pair (R, D) is achievable for the coding source X. Then for any $\epsilon > 0$, there exists a code of blocklength n sufficiently large, which operates at rate R and achieves distortion $D + \epsilon$. We prove

that $R \geq pR_{f_c}(D)$:

$$\begin{aligned}
nR &\geq H(M) \geq I(M;\hat{X}^n) \\
&\geq I(X^n;\hat{X}^n) = I(S^N,T^N,N;\hat{X}^n) \\
&\geq I(S^N;\hat{X}^n \mid T^N,N) \\
&= h(S^N \mid T^N,N) - h(S^N \mid \hat{X}^n,T^N,N) \\
&= \sum_{k=1}^{n} p_N(k)\Big(h(S^k \mid T^k, N=k) \\
&\qquad\qquad - h(S^k|\hat{X}^n,T^k,N=k)\Big) \\
&= \sum_{k=1}^{n} p_N(k)\left(h(S^k) - h(S^k \mid \hat{X}^n,T^k)\right) \\
&= \sum_{k=1}^{n} p_N(k)\left(\sum_{i=1}^{k}\left(h(S_i) - h(S_i \mid S^{i-1},\hat{X}^n,T^k)\right)\right) \\
&\geq \sum_{k=1}^{n} p_N(k)\left(\sum_{i=1}^{k}\left(h(S_i) - h(S_i \mid \hat{X}_{L_i}^{L_{(i+1)}-1},T^k)\right)\right) \qquad (26) \\
&= \sum_{k=1}^{n} p_N(k)\left(\sum_{i=1}^{k}\left(I(S_i;\hat{X}_{L_i}^{L_{(i+1)}-1} \mid T^k)\right)\right) \qquad (27)
\end{aligned}$$

where in (26) $L_i = 1 + \sum_{j=1}^{i-1} T_j$ and (27) holds because S and T are independent. Given T^k define $\hat{S}_i$ as follows:

$$\hat{S}_i = \underset{x \in \{\hat{X}_j : j = L_i,\ldots,L_{(i+1)-1}\}}{\arg\min} d(S_i,x)$$

Hence,

$$
\begin{aligned}
nR &\geq \sum_{k=1}^{n} p_N(k) \left(\sum_{i=1}^{k} \left(I(S_i; \hat{X}_{L_i}^{L_{(i+1)}-1} \mid T^k) \right) \right) \\
&\geq \sum_{k=1}^{n} p_N(k) \left(\sum_{i=1}^{k} \left(I(S_i; \hat{S}_i \mid T^k) \right) \right) \\
&= \sum_{k=1}^{n} p_N(k) \left(\sum_{i=1}^{k} \left(I(S_i; \hat{S}_i T^k) \right) \right) \qquad (28) \\
&\geq \sum_{k=1}^{n} p_N(k) \left(\sum_{i=1}^{k} \left(I(S_i; \hat{S}_i) \right) \right) \\
&\geq \sum_{k=1}^{n} p_N(k) \sum_{i=1}^{k} R_{f_c}(\mathrm{E}[d(S_i, \hat{S}_i)]) \qquad (29) \\
&= \sum_{k=1}^{n} k p_N(k) \frac{1}{k} \sum_{i=1}^{k} R_{f_c}(\mathrm{E}[d(S_i, \hat{S}_i)]) \\
&\geq \sum_{k=1}^{n} k p_N(k) R_{f_c}(\frac{1}{k} \sum_{i=1}^{k} \mathrm{E}[d(S_i, \hat{S}_i)]) \qquad (30) \\
&= \sum_{k=1}^{n} k p_N(k) R_{f_c}(\mathrm{E}[d_k(S^N, \hat{S}^N) | N = k]) \\
&= \mathrm{E}[N R_{f_c}(\mathrm{E}[d_N(S^N, \hat{S}^N)])], \qquad (31)
\end{aligned}
$$

where step (28) follows from the independence of S_i and T^k for all i, step (29) uses the definition of the rate-distortion function for source S, and step (30) follows from the convexity of $R_{f_c}(D)$ and Jensen's inequality. On the other hand, given that $N = k$,

$$
\begin{aligned}
\frac{1}{n} d(X^n; \hat{X}^n) &= \frac{1}{n} \sum_{i=1}^{k} \sum_{j=L_i}^{L_{i+1}-1} d(X_j, \hat{X}_j) \\
&\geq \frac{1}{n} \sum_{i=1}^{k} \sum_{j=L_i}^{L_{i+1}-1} d(S_i, \hat{S}_i) \\
&= \frac{1}{n} \sum_{i=1}^{k} T_i d(S_i, \hat{S}_i). \qquad (32)
\end{aligned}
$$

Taking expectations on both sides, it follows that

$$\begin{aligned}\mathrm{E}[d_n(X^n;\hat{X}^n)] &\geq \mathrm{E}[\frac{1}{n}\sum_{i=1}^{N} T_i d(S_i,\hat{S}_i)] \\ &\geq \sum_{k=1}^{n} \mathrm{E}[\frac{1}{n}\sum_{i=1}^{k} T_i d(S_i,\hat{S}_i)|N=k]p_N(k) \\ &= \frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{k}\Big(\mathrm{E}[T_i|N=k] \\ &\qquad \mathrm{E}[d(S_i,\hat{S}_i)|N=k]p_N(k)\Big). \end{aligned} \tag{33}$$

Note that $T_1, T_2, \ldots, T_{N-1}$ are i.i.d. and there exists $\tilde{T}_N$ such that $T_1, T_2, \ldots, T_{N-1}, \tilde{T}_N$ are i.i.d. and $\sum_{i=1}^{N-1} T_i + \tilde{T}_N \geq n$. Given $N=k$, $\mathrm{E}[T_1|N=k] = \ldots = \mathrm{E}[T_{k-1}|N=k] = \mathrm{E}[\tilde{T}_k|N=k]$, and therefore

$$\mathrm{E}[T_i|N=k] \geq \frac{n}{k}. \tag{34}$$

Combining (33) and (34), since $\mathrm{E}[T_k d(S_k,\hat{S}_k)|N=k] \geq 0$, it follows that

$$\begin{aligned}\mathrm{E}[d_n(X^n;\hat{X}^n)] &\geq \frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{k-1}\frac{n}{k}\,\mathrm{E}[d(S_i,\hat{S}_i)|N=k]p_N(k) \\ &= \mathrm{E}\left[\frac{N-1}{N} d_{N-1}(S^{N-1},\hat{S}^{N-1})\right]. \end{aligned} \tag{35}$$

But, $N d_N(S^N,\hat{S}^N) = (N-1)d_{N-1}(S^{N-1},\hat{S}^{N-1}) + d(S_N,\hat{S}_N)$. Hence,

$$\begin{aligned}\mathrm{E}[d_N(S^N,\hat{S}^N)] &\leq \mathrm{E}\left[\frac{N-1}{N} d_{N-1}(S^{N-1},\hat{S}^{N-1})\right] \\ &\quad + E\left[\frac{d_{\max}}{N}\right]. \end{aligned} \tag{36}$$

Combining (35) and (36) yields

$$\mathrm{E}[d_n(X^n;\hat{X}^n)] \geq \mathrm{E}[d_N(S^N,\hat{S}^N)] - E\left[\frac{d_{\max}}{N}\right]. \tag{37}$$

Since N counts the number of jumps in X^n, it can be written as $\sum_{i=1}^{n} \mathbb{1}_{X_i \neq X_{i-1}}$. Let $U_i = \mathbb{1}_{X_i \neq X_{i-1}}$. By construction, $\{U_i\}_{i=1}^n$ is a sequence of i.i.d. $\mathrm{Bern}(p)$ random variables. Therefore, by Hoeffding's inequality [65],

$$\mathrm{P}(|\frac{1}{n}\sum_{i=1}^{n} U_i - p| > \epsilon_1) \leq 2\mathrm{e}^{-2n\epsilon_1^2}. \tag{38}$$

Now let $\epsilon_n = \frac{p}{n^{1/4}}$, and define the event $\mathcal{E}_1$ as

$$\mathcal{E}_1 = \{|\frac{N}{n} - p| < \epsilon_n\}. \tag{39}$$

Conditioning on $\mathcal{E}_1$ we can rewrite (37) as

$$\begin{aligned}
\mathrm{E}[d_n(X^n;\hat{X}^n)] &\geq \mathrm{E}[d_N(S^N,\hat{S}^N)] \\
&\quad - P(\mathcal{E}_1)\frac{d_{\max}}{n(p-\epsilon_n)} - P(\mathcal{E}_1^c)d_{\max} \\
&\geq \mathrm{E}[d_N(S^N,\hat{S}^N)] \\
&\quad - \frac{d_{\max}}{n(p-\epsilon_n)} - 2\mathrm{e}^{-2n\epsilon_n^2}d_{\max} \\
&= \mathrm{E}[d_N(S^N,\hat{S}^N)] - \delta_n, \qquad (40)
\end{aligned}$$

where $\delta_n \to 0$ as $n \to \infty$. Combining (31) and (40) yields

$$\begin{aligned}
R &\geq \mathrm{E}\left[\frac{N}{n}R_{f_c}(\mathrm{E}[d_N(S^N,\hat{S}^N)|N])\right] \\
&= \mathrm{E}\left[\frac{N}{n}R_{f_c}(\mathrm{E}[d_N(S^N,\hat{S}^N)|N])|\mathcal{E}_1\right]\mathrm{P}(\mathcal{E}_1) \\
&\quad + \mathrm{E}\left[\frac{N}{n}R_{f_c}(\mathrm{E}[d_N(S^N,\hat{S}^N)|N])|\mathcal{E}_1^c\right]\mathrm{P}(\mathcal{E}_1^c) \\
&\geq \mathrm{E}\left[\frac{N}{n}R_{f_c}(\mathrm{E}[d_N(S^N,\hat{S}^N)|N])|\mathcal{E}_1\right]\mathrm{P}(\mathcal{E}_1) \\
&\geq (p-\epsilon_n)\,\mathrm{E}[R_{f_c}(\mathrm{E}[d_N(S^N,\hat{S}^N)|N])|\mathcal{E}_1]\,\mathrm{P}(\mathcal{E}_1) \\
&= (p-\epsilon_n)\sum_{k=n(p-\epsilon_n)}^{n(p+\epsilon_n)} p_N(k)R_{f_c}(\mathrm{E}[d_k(S^k,\hat{S}^k)|N=k]) \\
&\geq (p-\epsilon_n)\,\mathrm{P}(\mathcal{E}_1)R_{f_c}(\mathrm{E}[d_N(S^N,\hat{S}^N)|\mathcal{E}_1]), \qquad (41)
\end{aligned}$$

where the last step follows from Jensen's inequality. Now we already know that $\mathrm{P}(\mathcal{E}_1)$ is very close to one. Also, from (40),

$$\begin{aligned}
\mathrm{E}[d_n(X^n;\hat{X}^n)] + \delta_n &\geq \mathrm{E}[d_N(S^N,\hat{S}^N)] \\
&\geq \mathrm{E}[d_N(S^N,\hat{S}^N)|\mathcal{E}_1]\,\mathrm{P}(\mathcal{E}_1). \qquad (42)
\end{aligned}$$

Therefore,

$$\mathrm{E}[d_N(S^N,\hat{S}^N)|\mathcal{E}_1] \leq \frac{\mathrm{E}[d_n(X^n;\hat{X}^n)] + \delta_n}{\mathrm{P}(\mathcal{E}_1)},$$

which again since $\mathrm{P}(\mathcal{E}_1)$ is close to one yields the desired result.

2) Achievability: Consider the following encoder: to encode X^n, first describe $T_1,\ldots,T_N$ losslessly and then lossy encode $S_1,\ldots,S_N$. Assuming that the decoder already knows the blocklength n, to convey $T_1,\ldots,T_N$ to the decoder, it suffices to code $T_1,\ldots,T_{N-1}$, because $T_N = n - \sum_{i=1}^{N-1} T_i$. To losslessly describe $T_1,\ldots,T_{N-1}$, the encoder first encodes N using the Elias gamma code [66]. Since $N \in \{1,\ldots,n\}$, this requires at most $2\lfloor \log n \rfloor + 1$ bits. Also, as showed earlier in (38), $\mathrm{P}(|\frac{1}{n}N - p| > \epsilon_1) \leq 2\mathrm{e}^{-2n\epsilon_1^2}$.

Define ϵ_n and $\mathcal{E}_1$ as in (39) in the converse part. Consider a family of lossless compression codes $(n_1, \mathcal{E}_{n_1}^{(T)}, \mathcal{D}_{n_1}^{(T)})$ for the i.i.d. source $T = \{T_i\}_{i=1}^{\infty}$, operating at rate $H(T)+\epsilon_{n_1}^{(T)}$, $\epsilon_{n_1}^{(T)} > 0$, such that $\mathrm{P}(T^{n_1} \neq \hat{T}^{n_1}) \to 0$, as $n_1 \to \infty$, where $\hat{T}^{n_1} = \mathcal{D}_{n_1}^{(T)}(\mathcal{E}_{n_1}^{(T)}(T^{n_1}))$ and $\lim_{n_1\to\infty} \epsilon_{n_1}^{(T)} = 0$. By Shannon's lossless compression theorem, there exists such a family of codes satisfying these conditions [54]. Note that $H(T) = \sum_{m=1}^{\infty}(1-p)^{m-1}p\log((1-p)^{m-1}p) = \frac{H(p)}{p}$. After describing N to the decoder, if $\mathcal{E}_1$ holds, the encoder employs the $(N-1, \mathcal{E}_{N-1}^{(T)}, \mathcal{D}_{N-1}^{(T)})$ code to losslessly convey $T_1, \ldots, T_{N-1}$ to the decoder. This requires $(N-1)(H(T)+\epsilon_{N-1}^{(T)})$, bits. If $\mathcal{E}_1$ does not hold, it sends nothing else. Since the decoder knows N, it can determine whether $\mathcal{E}_1$ holds or not. Define the event $\mathcal{E}_2$ as $\mathcal{E}_2 = \{T^N = \hat{T}^N\}$.

The last encoding step is, conditioned on $\mathcal{E}_1$ holding, to describe $S_1, \ldots, S_N$. Let $(n_2, \mathcal{E}_{n_2}^{(S)}, \mathcal{D}_{n_2}^{(S)})$ be a family of lossy compression codes for the i.i.d. source $S = \{S_i\}_{i=0}^{\infty}$ operating at rate $R_{f_c}(D)$, and expected distortion not exceeding $D + \epsilon_{n_2}^{(S)}$, such that $\epsilon_{n_2}^{(S)} > 0$ and $\lim_{n_2\to\infty} \epsilon_{n_2}^{(S)} = 0$.

Overall the number of transmitted bits is either equal to $2\lfloor\log n\rfloor + 1$ if $\mathcal{E}_1$ does not hold, or $2\lfloor\log n\rfloor + 1 + (N-1)(H(T)+\epsilon_{N-1}^{(T)}) + NR_{f_c}(D)$, otherwise. In the latter case, the rate of the code can be upper bounded as

$$\begin{aligned}\frac{2\lfloor\log n\rfloor + 1}{n} + (p + \frac{p}{n^{1/4}})(R_{f_c}(D) + H(T) + \epsilon_*^{(T)}) \\ = pR_{f_c}(D) + H(p) + \epsilon_X,\end{aligned} \tag{43}$$

where $\epsilon_*^{(T)} = \max_{|n_1-p|\leq\epsilon_n} \epsilon_{n_1}^{(T)}$. Hence, ϵ_X can be made arbitrarily small by choosing n large enough.

After receiving all encoded bits, if only N is transmitted to the decoder, it reconstructs the all-zero sequence. Otherwise, it outputs $\hat{X}^n = \underbrace{\hat{S}_1, \ldots, \hat{S}_1}_{\hat{T}_1}, \underbrace{\hat{S}_2, \ldots, \hat{S}_2}_{\hat{T}_2}, \ldots, \underbrace{\hat{S}_N, \ldots, \hat{S}_N}_{\hat{T}_N}$. Note that by construction $\hat{N} = N$, with probability one.

By the tower property,

$$\begin{aligned}\mathrm{E}[d_n(X^n, \hat{X}^n)] &= \sum_{n_2=1}^{\infty} \mathrm{E}[d_n(X^n, \hat{X}^n)|N = n_2]\,\mathrm{P}(N = n_2) \\ &\leq \sum_{n_2=n(p-\epsilon_n)}^{n(p+\epsilon_n)} \mathrm{E}[d_n(X^n, \hat{X}^n)|N = n_2]\,\mathrm{P}(N = n_2) \\ &\quad + d_{\max}\,\mathrm{P}(\mathcal{E}_1^c) \\ &\leq \sum_{n_2=n(p-\epsilon_n)}^{n(p+\epsilon_n)} \mathrm{E}[d_n(X^n, \hat{X}^n)|N = n_2, \mathcal{E}_2]\,\mathrm{P}(N = n_2, \mathcal{E}_2) \\ &\quad + d_{\max}\,\mathrm{P}(\mathcal{E}_1^c \cup \mathcal{E}_2^c).\end{aligned}$$

Conditioned on $\mathcal{E}_1 \cap \mathcal{E}_2$, the distortion between the source block X^n, and its reconstruction $\hat{X}_n$ can be written as

$d_n(X^n,\hat{X}^n)=\frac{1}{n}\sum_{i=1}^{n}d(X_i,\hat{X}_i)=\frac{1}{n}\sum_{k=1}^{N}T_k d(S_k,\hat{S}_k)$. Therefore,

$$\begin{aligned}&\mathrm{E}[d_n(X^n,\hat{X}^n)]\\&\quad\le\frac{1}{n}\sum_{n_2=n(p-\epsilon_n)}^{n(p+\epsilon_n)}\Bigg(\mathrm{E}[\sum_{k=1}^{n_2}T_k d(S_k,\hat{S}_k)|N=n_2]\\&\qquad\qquad\mathrm{P}(N=n_2,\mathcal{E}_2)\Bigg)+d_{\max}\mathrm{P}(\mathcal{E}_1^c\cup\mathcal{E}_2^c).\end{aligned}$$

Conditioned on N, T_k and $d(S_k,\hat{S}_k)$ are independent, and $T_1,\ldots,T_{N-1}$ are i.i.d. Also, there exists $\tilde{T}_N$ such that $T_N\le\tilde{T}_N$, and $T_1,\ldots,T_{N-1},\tilde{T}_N$ are all i.i.d. Therefore,

$$\begin{aligned}&\mathrm{E}[d_n(X^n,\hat{X}^n)]\\&\le\frac{1}{n}\sum_{n_2=n(p-\epsilon_n)}^{n(p+\epsilon_n)}\Bigg((n_2-1)\,\mathrm{E}[d_{n_2-1}(S^{n_2},\hat{S}^{n_2})]\,\mathrm{E}[T_1|N=n_2]\\&\quad+\mathrm{E}[d(S_{n_2},\hat{S}_{n_2})]\,\mathrm{E}[\tilde{T}_N|N=n_2]\Bigg)\mathrm{P}(N=n_2,\mathcal{E}_2)\\&\quad+d_{\max}\mathrm{P}(\mathcal{E}_1^c\cup\mathcal{E}_2^c)\\&\le\frac{1}{n}\sum_{n_2=n(p-\epsilon_n)}^{n(p+\epsilon_n)}\Bigg((n_2-1)(D+\epsilon_{n_2-1}^{(S)})\,\mathrm{E}[T_1|N=n_2]\\&\quad+\mathrm{E}[d(S_{n_2},\hat{S}_{n_2})]\,\mathrm{E}[\tilde{T}_N|N=n_2]\Bigg)\mathrm{P}(N=n_2,\mathcal{E}_2)\\&\quad+d_{\max}\mathrm{P}(\mathcal{E}_1^c\cup\mathcal{E}_2^c).\end{aligned}\tag{44}$$

On the other hand, since $T_1,\ldots,T_{N-1}$ are i.i.d., we have $\mathrm{E}[T_1|N=n_2]=\ldots=\mathrm{E}[T_{N-1}|N=n_2]$. But $\sum_{i=1}^{N-1}T_i\le n$. Therefore, $\mathrm{E}[\sum_{i=1}^{N-1}T_i|N=n_2]=(n_2-1)\,\mathrm{E}[T_1|N=n_2]\le n$, and

$$\mathrm{E}[T_1|N=n_2]\le\frac{n}{n_2-1}.\tag{45}$$

Also,

$$\begin{aligned}\mathrm{E}[\tilde{T}_N|N=n_2]\,\mathrm{P}(N=n_2,\mathcal{E}_2)&\le\mathrm{E}[\tilde{T}_N|N=n_2]\,\mathrm{P}(N=n_2)\\&\le\mathrm{E}[\tilde{T}_N]=\frac{1}{p}.\end{aligned}\tag{46}$$

Hence, combining (44), (45) and (46) yields $\mathrm{E}[d_n(X^n,\hat{X}^n)]\le\max_{n_2=n(p-\epsilon_n)}^{n(p+\epsilon_n)}(D+\epsilon_{n_2-1}^{(S)})+d_{\max}(\frac{2\epsilon_n}{p}+\mathrm{P}(\mathcal{E}_1^c)+\mathrm{P}(\mathcal{E}_2^c\cap\mathcal{E}_1))\le D+\delta_n$, where $\delta_n\to 0$, as n grows to infinity.

G. Proof of Corollary 2

Since $\liminf_{n\to\infty}\frac{m_n}{n}>2\overline{\dim}_R(\mathbf{X})$, there exists $\eta>1$, such that $\liminf_{n\to\infty}\frac{m_n}{n}>2\eta\overline{\dim}_R(\mathbf{X})$. Therefore, there exists $n_\eta>0$, such that for all $n>n_\eta$, $\frac{m_n}{n}\ge 2\eta\overline{\dim}_R(\mathbf{X})$. On the other hand, for any $\gamma>0$, there exists

$D_\gamma > 0$, such that for all $D \leq D_\gamma$,

$$2\frac{R(\mathbf{X}, D)}{\log \frac{1}{D}} \leq \overline{\dim}_R(\mathbf{X}) + \gamma.$$

Hence, there exists $\eta' \in (1, \eta)$, such that choosing γ small enough, we have

$$\frac{m}{n} \geq \frac{4\eta' R(\mathbf{X}, D)}{\log \frac{1}{D}},$$

for all $n > n_\eta$ and $D < D_\gamma$.

Since $\lim_{D\to 0}(2D^{\frac{1}{2}(1-\frac{1+\delta}{\eta})}(\sqrt{\frac{\log \frac{1}{D}}{4\eta R}} + 2) + \sqrt{D}) = 0$, there exists $D_\Delta < D_\gamma$, such that

$$2D_\Delta^{\frac{1}{2}(1-\frac{1+\delta}{\eta})}(\sqrt{\frac{\log \frac{1}{D_\Delta}}{4\eta R}} + 2) + \sqrt{D_\Delta} < \Delta.$$

Considering a family of lossy compression codes achieving $(R(D_\Delta), D_\Delta)$ and the CSP algorithm that employs this family of codes, Theorem 1 proves the desired result.

H. Proof of Theorem 6

Let $\hat{X}^n = g_n(f_n(X^n))$. Since (n, f_n, g_n) denotes a family of point-wise universal lossy compression codes operating at distortion level D, for any $\epsilon > 0$, for all n large enough,

$$\mathrm{P}(\frac{1}{n}|f_n(X^n)| > R(\mathbf{X}, D) + \epsilon) < \frac{\epsilon}{2},$$

and

$$\mathrm{P}(\frac{1}{\sqrt{n}}\|X^n - \hat{X}^n\|_2 > \sqrt{D + \epsilon}) \leq \frac{\epsilon}{2}.$$

Let $\mathcal{E}_1 \triangleq \{\frac{1}{n}|f_n(X^n)| \leq R(\mathbf{X}, D) + \epsilon\} \cup \{\frac{1}{\sqrt{n}}\|X^n - \hat{X}^n\|_2 \leq \sqrt{D + \epsilon}\}$. Then, $\mathrm{P}(\mathcal{E}_1) \leq \epsilon$, and conditioned on $\mathcal{E}_1^c$, $f_n(X^n)$ satisfies the condition of the UCSP optimization. Therefore, conditioned on $\mathcal{E}_1^c$,

$$\begin{aligned}\|Y^m - A\tilde{X}^n\|_2 &\leq \|Y^m - Ag_n(f_n(X^n))\|_2 \\ &\leq \sigma_{\max}(A)\|X^n - \hat{X}^n\|_2 \\ &\leq \sigma_{\max}(A)\sqrt{n(D + \epsilon)}. \end{aligned} \tag{47}$$

The rest of the proof is very similar to the proof of Theorem 1. The only difference is that in this case, instead of the size of the codebook, we need to bound the size of the set $\mathcal{B} = \{b : b \in \{0, 1\}^*, |b| \leq n(R(\mathbf{X}, D) + \epsilon)\}$. But $|\mathcal{B}| = \sum_{i=1}^{n(R(\mathbf{X},D)+\epsilon)} 2^i = 2^{n(R(\mathbf{X},D)+\epsilon)+1} - 1$. The rest of the proof follows similar to the proof of Theorem 1.

VII. Conclusions

In this paper, we have studied the application of rate-distortion codes in building compressed sensing recovery algorithms for stochastic processes. Establishing such connections between rate-distortion coding and compressed sensing potentially enables application of well-studied state-of-the-art lossy compression codes in building highly efficient compressed sensing recovery algorithms.

We have focused on the CSP algorithm proposed in [30] as a compression-based compressed sensing recovery algorithm for deterministic signals. For the CSP algorithm that employs a rate-distortion code with a certain rate R and distortion D, we have derived an upper bound on the normalized distance between the original vector and its reconstruction that holds with high probability.

To analyze the asymptotic performance of the CSP algorithm when the distortion D approaches zero, we have defined the RDD of stationary processes, as a generalization of the RDD of stochastic vectors introduced in [32]. We have proved that under some mild conditions the RDD of a stationary process is equal to its ID introduced in [33]. Our results have demonstrated that in the limit, as $D \to 0$, for sufficiently large blocklengths n, CSP renders a reliable reconstruction of the source vector with almost zero-distortion, with slightly more than n times the RDD of the source. This is equal to the fundamental limit of compressed sensing in memoryless stationary sources shown in [34], which proves the optimality of CSP at least in cases where the lower bounds are known.

There are two major directions that remain open for future study: the first is to design algorithms to solve the minimization problem in CSP with manageable complexity; and the second is to find the fundamental limits on compressed sensing for general stochastic stationary sources, which would enable us to see whether CSP is always optimal, and if not how far from optimal it is.

APPENDIX A
RATE OF APPROACH OF $R^{(m)}(\mathbf{X}, D)$ TO $R(\mathbf{X}, D)$

Consider a pair of random variables $(X, \hat{X}) \in \mathcal{X} \times \hat{\mathcal{X}}$, with alphabet sets $\mathcal{X}, \hat{\mathcal{X}} \subset \mathbb{R}$, distributed as $p_{X,\hat{X}}$, where $p_{X,\hat{X}}$ denotes a general measure. For sets $\mathcal{E} \in \mathcal{X}$ and $\mathcal{F} \in \hat{\mathcal{X}}$, the probability of the set $\mathcal{E} \times \mathcal{F}$ under $p_{X,\hat{X}}$ is computed as

$$\mathrm{P}_{X,\hat{X}}(\mathcal{E} \times \mathcal{F}) = \int_{u \in \mathcal{E} \times \mathcal{F}} p_{X,\hat{X}}(du).$$

The marginal distributions under X and $\hat{X}$ are defined as

$$\mathrm{P}_X(\mathcal{E}) = \int_{u \in \mathcal{E} \times \hat{\mathcal{X}}} p_{X,\hat{X}}(du),$$

and

$$\mathrm{P}_{\hat{X}}(\mathcal{F}) = \int_{u \in \mathcal{X} \times \mathcal{F}} p_{X,\hat{X}}(du),$$

respectively. Let $\mathcal{P}$ denote a partition of $\mathcal{X} \times \mathcal{F}$ into finitely many rectangles, $\{\mathcal{E}_i, \mathcal{F}_j\}_{i,j}$. Dobrushin [32], [67] established that for random variables $(X, \hat{X})$ with a general distribution, the mutual information can be generalized as

$$I(X; \hat{X}) = \sup_{\mathcal{P}} \sum_{i,j} \mathrm{P}_{X,\hat{X}}(\mathcal{E}_i \times \mathcal{F}_j) \log \frac{\mathrm{P}_{X,\hat{X}}(\mathcal{E}_i \times \mathcal{F}_j)}{\mathrm{P}_X(\mathcal{E}_i)\, \mathrm{P}_{\hat{X}}(\mathcal{F}_j)}.$$

Wyner and Ziv in [64] proved that with sources with either discrete or absolutely continuous distributions we

have

$$\sum_{k=1}^{N} I(X_k;\hat{X}_k) - N\Delta_N - I(X^N;\hat{X}^N) \le 0,$$

where

$$\Delta_N = \frac{1}{N}\sup_{\mathcal{P}} \sum_{i_1,\ldots,i_N} \mathrm{P}_{X^N}\Big(\prod_{k=1}^{N}\mathcal{E}_{i_k}\Big)\log\frac{\mathrm{P}_{X^N}\big(\prod_{k=1}^{N}\mathcal{E}_{i_k}\big)}{\prod_{k=1}^{N}\mathrm{P}_{X_k}(\mathcal{E}_{i_k})}. \tag{48}$$

In the following, we prove that this inequality also holds for sources with general distributions. Given $(X_k,\hat{X}_k)$ and $\epsilon>0$, let $\mathcal{P}_k=\{\mathcal{E}_{i_k}\times\mathcal{F}_{j_k}\}_{i_k,j_k}$ denote the partitioning of $\mathcal{X}\times\hat{\mathcal{X}}$ that ensures

$$\begin{aligned} &I(X_k;\hat{X}_k)\\ &\quad-\sum_{i_k,j_k}\mathrm{P}_{X_k,\hat{X}_k}(\mathcal{E}_{i_k}\times\mathcal{F}_{j_k})\log\frac{\mathrm{P}_{X_k,\hat{X}_k}(\mathcal{E}_{i_k}\times\mathcal{F}_{j_k})}{\mathrm{P}_{X_k}(\mathcal{E}_i)\,\mathrm{P}_{\hat{X}_k}(\mathcal{F}_{j_k})}\le\frac{\epsilon}{N}. \end{aligned}$$

Since $I(X_k;\hat{X}_k)$ is defined as the supremum of the objective function over all partitions, such a partition always exists. Combining these partitions yields a natural partitioning of $\mathcal{X}^N\times\hat{\mathcal{X}}^N$. Since to evaluate $I(X^N;\hat{X}^N)$ and Δ_N involves taking suprema of the corresponding objective functions, we have

$$\begin{aligned} &\sum_{k=1}^{N} I(X_k;\hat{X}_k) - N\Delta_N - I(X^N;\hat{X}^N)\\ &\le\sum_{k=1}^{N}\Big(\sum_{i_k,j_k}\mathrm{P}_{X_k,\hat{X}_k}(\mathcal{E}_{i_k}\times\mathcal{F}_{j_k})\\ &\qquad\qquad\log\frac{\mathrm{P}_{X_k,\hat{X}_k}(\mathcal{E}_{i_k}\times\mathcal{F}_{j_k})}{\mathrm{P}_{X_k}(\mathcal{E}_{i_k})\,\mathrm{P}_{\hat{X}_k}(\mathcal{F}_{j_k})}+\frac{\epsilon}{N}\Big)\\ &\quad-\sum_{i_1,\ldots,i_N}\mathrm{P}_{X^N}\Big(\prod_{k=1}^{N}\mathcal{E}_{i_k}\Big)\log\frac{\mathrm{P}_{X^N}\big(\prod_{k=1}^{N}\mathcal{E}_{i_k}\big)}{\prod_{k=1}^{N}\mathrm{P}_{X_k}(\mathcal{E}_{i_k})}\\ &\quad-\sum_{\substack{i_1,\ldots,i_k\\ j_1,\ldots,j_k}}\mathrm{P}_{X^N,\hat{X}^N}\Big(\prod_{k=1}^{N}\mathcal{E}_{i_k}\times\prod_{k=1}^{N}\mathcal{F}_{j_k}\Big)\\ &\qquad\qquad\log\frac{\mathrm{P}_{X^N,\hat{X}^N}\big(\prod_{k=1}^{N}\mathcal{E}_{i_k}\times\prod_{k=1}^{N}\mathcal{F}_{j_k}\big)}{\mathrm{P}_{X^N}\big(\prod_{k=1}^{N}\mathcal{E}_{i_k}\big)\,\mathrm{P}_{\hat{X}^N}\big(\prod_{k=1}^{N}\mathcal{F}_{j_k}\big)}\\ &=\epsilon+\sum_{i^N,j^N}\mathrm{P}_{X^N,\hat{X}^N}\Big(\prod_{k=1}^{N}\mathcal{E}_{i_k}\times\prod_{k=1}^{N}\mathcal{F}_{j_k}\Big)\\ &\qquad\qquad\log\Big[\prod_{k=1}^{N}\frac{\mathrm{P}_{X_k,\hat{X}_k}(\mathcal{E}_{i_k}\times\mathcal{F}_{j_k})}{\mathrm{P}_{X_k}(\mathcal{E}_{i_k})\times\mathrm{P}_{\hat{X}_k}(\mathcal{F}_{j_k})}\\ &\qquad\qquad\qquad\times\frac{\prod_{k=1}^{N}\mathrm{P}_{X_k}(\mathcal{E}_{i_k})}{\mathrm{P}_{X^N}\big(\prod_{k=1}^{N}\mathcal{E}_{i_k}\big)}\\ &\qquad\qquad\qquad\times\frac{\mathrm{P}_{X^N}\big(\prod_{k=1}^{N}\mathcal{E}_{i_k}\big)\,\mathrm{P}_{\hat{X}^N}\big(\prod_{k=1}^{N}\mathcal{F}_{j_k}\big)}{\mathrm{P}_{X^N,\hat{X}^N}\big(\prod_{k=1}^{N}\mathcal{E}_{i_k}\times\prod_{k=1}^{N}\mathcal{F}_{j_k}\big)}\Big]. \end{aligned} \tag{49}$$

Canceling the common terms, and rearranging the terms, it follows that

$$\begin{aligned}
&\sum_{k=1}^{N} I(X_k;\hat{X}_k) - N\Delta_N - I(X^N;\hat{X}^N) \\
&\leq \epsilon + \sum_{i^N,j^N} \mathrm{P}_{X^N,\hat{X}^N}\Big(\prod_{k=1}^{N}\mathcal{E}_{i_k} \times \prod_{k=1}^{N}\mathcal{F}_{j_k}\Big) \\
&\quad \log\Big(\frac{\prod_{k=1}^{N}\mathrm{P}_{X_k|\hat{X}_k}(\mathcal{E}_{i_k}|\mathcal{F}_{j_k})}{\mathrm{P}_{X^N|\hat{X}^N}(\prod_{k=1}^{N}\mathcal{E}_{i_k}|\prod_{k=1}^{N}\mathcal{F}_{j_k})}\Big).
\end{aligned} \tag{50}$$

Since $\log x \leq x-1$, the right hand side of (50) can further be upper-bounded as

$$\begin{aligned}
&\sum_{k=1}^{N} I(X_k;\hat{X}_k) - N\Delta_N - I(X^N;\hat{X}^N) \\
&\leq \epsilon + \sum_{i^N,j^N} \mathrm{P}_{X^N,\hat{X}^N}\Big(\prod_{k=1}^{N}\mathcal{E}_{i_k} \times \prod_{k=1}^{N}\mathcal{F}_{j_k}\Big) \\
&\qquad \Big(\frac{\prod_{k=1}^{N}\mathrm{P}_{X_k|\hat{X}_k}(\mathcal{E}_{i_k}|\mathcal{F}_{j_k})}{\mathrm{P}_{X^N|\hat{X}^N}(\prod_{k=1}^{N}\mathcal{E}_{i_k}|\prod_{k=1}^{N}\mathcal{F}_{j_k})} - 1\Big) \\
&= \epsilon + \sum_{i^N,j^N} \mathrm{P}_{\hat{X}^N}\Big(\prod_{k=1}^{N}\mathcal{F}_{j_k}\Big)\prod_{k=1}^{N}\mathrm{P}_{X_k|\hat{X}_k}(\mathcal{E}_{i_k}|\mathcal{F}_{j_k}) \\
&\qquad - \sum_{i^N,j^N} \mathrm{P}_{X^N,\hat{X}^N}\Big(\prod_{k=1}^{N}\mathcal{E}_{i_k} \times \prod_{k=1}^{N}\mathcal{F}_{j_k}\Big) \\
&= \epsilon + 1 - 1 = \epsilon.
\end{aligned} \tag{51}$$

Since $\epsilon > 0$ was selected arbitrarily, this proves the desired inequality, i.e., $\sum_{k=1}^{N} I(X_k;\hat{X}_k) - N\Delta_N - I(X^N;\hat{X}^N) \leq 0$. This result is analogous to Lemma 2 in [64], but holds for sources with general distributions. After this generalization, the next steps required for proving the lower bound established in Section III.B of [64] also hold in this case, with no change. Therefore,

$$R^{(N)}(\mathbf{X},D) \geq R^{(1)}(\mathbf{X},D) - \Delta_N.$$

Using the fact that memory decreases the rate of a source [64] we get an upper bound on $R^{(N)}(\mathbf{X},D)$:

$$R^{(1)}(\mathbf{X},D) - \Delta_N \leq R^{(N)}(\mathbf{X},D) \leq R^{(1)}(\mathbf{X},D). \tag{52}$$

To prove the inequality (25), we first need to review some properties of Δ_N. Following the definition in (48), it can be shown that Δ_N can be represented in terms of mutual information as follows [64]:

$$\Delta_N = \frac{1}{N}\sum_{i=2}^{N} I(X_k;X_1^{k-1}). \tag{53}$$

Note that with this alternative representation it is very easy to see that Δ_N is increasing in N [64]. Putting this

together with (52) we get

$$|R^{(N)}(\mathbf{X}, D) - R(\mathbf{X}, D)| \leq \Delta_N \leq \Delta_\infty, \tag{54}$$

where

$$\Delta_\infty = \lim_{N\to\infty} \Delta_N = I(X_1; X_{-\infty}^0), \tag{55}$$

follows directly from (53). Note that $R(\mathbf{X}, D)$ is the rate-distortion function of the stationary process $\mathbf{X}$.

Let $\mathbf{Y}$ be the supersource whose outputs are successive blocks of m outputs of the source $\mathbf{X}$. Applying (54) to $\mathbf{Y}$ with $N = 1$ we have

$$|R^{(1)}(\mathbf{Y}, D) - R(\mathbf{Y}, D)| \leq \Delta_\infty.$$

Since $\mathbf{Y}$ is defined as a supersource of successive blocks of length m of the source $\mathbf{X}$, it is easy to see that $R^{(1)}(\mathbf{Y}, D) = mR^{(m)}(\mathbf{X}, D)$ and $R(\mathbf{Y}, D) = mR(\mathbf{X}, D)$, and therefore,

$$\begin{aligned} |R^{(m)}(\mathbf{X}, D) - R(\mathbf{X}, D)| &\leq \frac{1}{m}\Delta_\infty \\ &= \frac{1}{m} I(X_1; X_{-\infty}^0), \end{aligned}$$

where the last line follows from (55). Hence, the proof is complete and (25) holds for general stationary sources.

ACKNOWLEDGMENTS

This research was supported in part by the U.S. National Science Foundation grant CCF-1420575.

REFERENCES

[1] D.L. Donoho. Compressed sensing. *IEEE Trans. Inf. Theory*, 52(4):1289–1306, 2006.

[2] E. J Candès and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. Inf. Theory*, 52(12):5406–5425, 2006.

[3] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory*, 52(2):489–509, Feb. 2006.

[4] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde. Model-based compressive sensing. *IEEE Trans. Inf. Theory*, 56(4):1982–2001, Apr. 2010.

[5] M. Vetterli, P. Marziliano, and T. Blu. Sampling signals with finite rate of innovation. *IEEE Trans. Signal Process.*, 50(6):1417–1428, Jun. 2002.

[6] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum rank solutions to linear matrix equations via nuclear norm minimization. *SIAM Rev.*, 52(3):471–501, Apr. 2010.

[7] C. Hegde and R. G. Baraniuk. Sampling and recovery of pulse streams. *IEEE Trans. Signal Process.*, 59(4):1505 –1517, Apr. 2011.

[8] D. L. Donoho, H. Kakavand, and J. Mammen. The simplest solution to an underdetermined system of linear equations. In *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, pages 1924 –1928, Jul. 2006.

[9] S. Bakin. Adaptive regression and model selection in data mining problems. *Ph.D. Thesis, Australian National University*, 1999.

[10] Y. C. Eldar, P. Kuppinger, and H. Bolcskei. Block-sparse signals: Uncertainty relations and efficient recovery. *IEEE Trans. Signal Process.*, 58(6):3042–3054, Jun. 2010.

[11] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. Roy. Statist. Soc. Ser. B*, 68(1):49–67, 2006.

[12] S. Ji, D. Dunson, and L. Carin. Multi-task compressive sensing. *IEEE Trans. Signal Process.*, 57(1):92–106, 2009.

[13] A. Maleki, L. Anitori, Z. Yang, and R. G. Baraniuk. Asymptotic analysis of complex LASSO via complex approximate message passing (CAMP). *arXiv:1108.0477v1*, 2011.

[14] M. Stojnic. Block-length dependent thresholds in block-sparse compressed sensing. *arXiv:0907.3679*, 2009.

[15] M. Stojnic, F. Parvaresh, and B. Hassibi. On the reconstruction of block-sparse signals with an optimal number of measurements. *IEEE Trans. Signal Processing*, 57(8):3075–3085, Aug. 2009.

[16] M. Stojnic. ℓ_2/ℓ_1-optimization in block-sparse compressed sensing and its strong thresholds. *IEEE J. Select. Top. Signal Proc.*, 4(2):350–357, 2010.

[17] L. Meier, S. Van De Geer, and P. Buhlmann. The group LASSO for logistic regression. *J. Roy. Statist. Soc. Ser. B*, 70(1):53–71, 2008.

[18] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. *Found. of Comp. Math.*, 12(6):805–849, 2012.

[19] S. Som and P. Schniter. Compressive imaging using approximate message passing and a Markov-tree prior. *IEEE Trans. Signal Process.*, 60(7):3439–3448, 2012.

[20] D. Donoho and G. Kutyniok. Microlocal analysis of the geometric separation problem. *Comm. Pure App. Math.*, 66(1):1–47, 2013.

[21] E. J. Candés, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *arXiv preprint arXiv:0912.3599*, 2009.

[22] A. E. Waters, A. C. Sankaranarayanan, and R. Baraniuk. Sparcs: Recovering low-rank and sparse matrices from compressive measurements. In *Proc. Adv. Neural Inform. Proc. Sys.*, pages 1089–1097, 2011.

[23] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A.S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM J. Optimization*, 21(2):572–596, 2011.

[24] M. F. Duarte, W. U. Bajwa, and R. Calderbank. The performance of group LASSO for linear regression of grouped variables. Technical report, Tach. Rep. TR-2010-10, Duke University, Dept. Computer Science, Durham, NC, 2011.

[25] T. Blumensath and M. E. Davies. Sampling theorems for signals from the union of finite-dimensional linear subspaces. *IEEE Trans. Inf. Theory*, 55(4):1872–1882, 2009.

[26] M. B. McCoy and J. A. Tropp. Sharp recovery bounds for convex deconvolution, with applications. *arXiv preprint arXiv:1205.1580*, 2012.

[27] C. Studer and R. G. Baraniuk. Stable restoration and separation of approximately sparse signals. *Applied and Comp. Har. Anal.*, 37(1):12–35, 2014.

[28] G. Peyré and J. Fadili. Group sparsity with overlapping partition functions. *Proc. EUSIPCO*, pages 303–307, 2011.

[29] C. E. Shannon. A mathematical theory of communication: Parts I and II. *Bell Syst. Tech. J.*, 27:379–423 and 623–656, 1948.

[30] S. Jalali and A. Maleki. From compression to compressed sensing. In *Proc. IEEE Int. Symp. Inform. Theory*, pages 111–115, 2013.

[31] S. Jalali and A. Maleki. From compression to compressed sensing. *Applied and Comp. Har. Anal.*, 2015.

[32] T. Kawabata and A. Dembo. The rate-distortion dimension of sets and measures. *IEEE Trans. Inf. Theory*, 40(5):1564–1572, Sep. 1994.

[33] S. Jalali and H. V. Poor. Universal compressed sensing of Markov sources. *arXiv preprint arXiv:1406.7807*, 2014.

[34] Y. Wu and S. Verdú. Rényi information dimension: Fundamental limits of almost lossless analog compression. *IEEE Trans. Inf. Theory*, 56(8):3721–3748, Aug. 2010.

[35] J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Trans. Inf. Theory*, 23(3):337–343, 1977.

[36] J. Ziv and A. Lempel. Compression of individual sequences via variable-rate coding. *IEEE Trans. Inf. Theory*, 24(5):530–536, Sep 1978.

[37] D. J. Sakrison. The rate of a class of random processes. *IEEE Trans. Inform. Theory*, 16:10–16, Jan. 1970.

[38] J. Ziv. Coding of sources with unknown statistics part II: Distortion relative to a fidelity criterion. *IEEE Trans. Inf. Theory*, 18:389–394, May 1972.

[39] D. L. Neuhoff, R. M. Gray, and L. D. Davisson. Fixed rate universal block source coding with a fidelity criterion. *IEEE Trans. Inf. Theory*, 21:511–523, May 1972.

[40] D. L. Neuhoff and P. L. Shields. Fixed-rate universal codes for Markov sources. *IEEE Trans. Inf. Theory*, 24:360–367, May 1978.

[41] J. Ziv. Distortion-rate theory for individual sequences. *IEEE Trans. Inf. Theory*, 24:137–143, Jan. 1980.

[42] R. Garcia-Munoz and D. L. Neuhoff. Strong universal source coding subject to a rate-distortion constraint. *IEEE Trans. Inf. Theory*, 28:285295, Mar. 1982.

[43] D. Donoho. The Kolmogorov sampler. Technical Report 2002-04, Stanford University, Jan. 2002.

[44] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú, and M. Weinberger. Universal discrete denoising: Known channel. *IEEE Trans. Inf. Theory*, 51(1):5–28, 2005.

[45] M. Feder, N. Merhav, and M. Gutman. Universal prediction for individual sequences. *IEEE Trans. Inf. Theory*, 38(4):1258–1270, 1992.

[46] N. Merhav and M. Feder. Universal prediction. *IEEE Trans. Inf. Theory*, 44(6):2124–2147, 1998.

[47] S. Jalali, A. Maleki, and R.G. Baraniuk. Minimum complexity pursuit for universal compressed sensing. *IEEE Trans. Inf. Theory*, 60(4):2253–2268, Apr. 2014.

[48] D. Baron and M. F. Duarte. Universal MAP estimation in compressed sensing. In *Proc. 49th Annual Proc. Allerton Conf. Comm., Cont., and Comp.*, Sep. 2011.

[49] D. Baron and M. F. Duarte. Signal recovery in compressed sensing via universal priors. *arXiv:1204.2611*, 2012.

[50] K. Marton. Error exponent for source coding with a fidelity criterion. *IEEE Trans. Inf. Theory*, 20(2):197–199, 1974.

[51] Y. Steinberg and S. Verdu. Simulation of random processes and rate-distortion theory. *IEEE Trans. Inf. Theory*, 42(1):63–86, 1996.

[52] S. Ihara and M. Kubo. Error exponent of coding for stationary memoryless sources with a fidelity criterion. *IEICE Trans. on Fund. of Elec., Comm. and Comp. Sciences*, 88(5):1339–1345, 2005.

[53] K. Iriyama. Probability of error for the fixed-length lossy coding of general sources. *IEEE Trans. Inf. Theory*, 51(4):1498–1507, April 2005.

[54] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, New York, second edition, 2006.

[55] A. Rényi. On the dimension and entropy of probability distributions. *Acta Mathematica Academiae Scientiarum Hungarica*, 10(1-2):193–215, 1959.

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

[57] R. G. Gallager. *Information Theory and Reliable Communication*. NY: John Wiley, 1968.

[58] R. Gray. Rate distortion functions for finite-state finite-alphabet Markov sources. *IEEE Trans. Inf. Theory*, 17(2):127–134, Mar. 1971.

[59] T. Berger. Explicit bounds to r(d) for a binary symmetric Markov source. *IEEE Trans. Inf. Theory*, 23(1):52–59, Jan 1977.

[60] S. Jalali and T. Weissman. New bounds on the rate-distortion function of a binary Markov source. In *Proc. IEEE Int. Symp. Inform. Theory*, pages 571–575. IEEE, 2007.

[61] 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*, 45(6), 1999.

[62] Y. Wu and S. Verdú. Optimal phase transitions in compressed sensing. *IEEE Trans. Inf. Theory*, 58(10):6241–6263, 2012.

[63] E. Candès, J. Romberg, and T. Tao. Decoding by linear programming. *IEEE Trans. Inf. Theory*, 51(12):4203 – 4215, Dec. 2005.

[64] A. D. Wyner and J. Ziv. Bounds on the rate-distortion function for stationary sources with memory. *IEEE Trans. Inf. Theory*, 17(5):508–513, 1971.

[65] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, March 1963.

[66] P. Elias. Universal codeword sets and representations of the integers. *IEEE Trans. Inf. Theory*, 21(2):194–203, 1975.

[67] M. S. Pinsker. *Information and Information Stability of Random Variables and Processes*. translated and edited by A. Feinstein. San Francisco, CA: Holden-Day, Inc., 1964.