ARMA Processes with Discrete-Continuous Excitation: Compressibility Beyond Sparsity

Mohammad-Amin Charusaie, Arash Amini, *Senior, IEEE*, and Stefano Rini, *Senior, IEEE*

***Abstract*—The Rényi Information Dimension (RID) is a fundamental measure for quantifying the compressibility of random variables with singularities in their distributions, extending beyond classical notions of sparsity. At a high level, RID represents the average number of bits required to encode i.i.d. samples of a random variable with high precision. For stochastic processes, two main extensions of RID exist: the information dimension rate (IDR) and the block information dimension (BID). A more recent approach to characterizing the compressibility of stochastic processes is through ϵ-achievable compression rates, which treat a random process as the limit of finite-dimensional random vectors and leverage tools from compressed sensing. However, the interplay between BID, IDR, and ϵ-achievable compression rates remains poorly understood. Furthermore, explicit values of IDR and BID are known only for a limited class of processes, such as i.i.d. sequences (i.e., discrete-domain white noise) and moving-average (MA) processes. This paper investigates the IDR and BID of discrete-time Auto-Regressive Moving-Average (ARMA) processes and their relationship with ϵ-achievable compression rates when the excitation noise follows a discrete-continuous distribution. Specifically, we show that the RID and ϵ-achievable compression rates of such ARMA processes are equal to those of their excitation noise. In other words, despite the fact that ARMA process samples are not sparse, their compressibility matches that of their sparse excitation noise. To establish this result, we demonstrate that the singular components of the sample distribution are supported on affine sets, with relative dimensions that concentrate around the BID. Leveraging a known result on typical affinely singular sources, we further prove that in this setting, the RID coincides with ϵ-achievable compression rates. The findings of this paper provide new insights into the compressibility of locally correlated data with finite- or infinite-memory, which are commonly modeled using ARMA processes.**

***Index Terms*—ARMA processes, discrete-continuous random variables, Rényi Information Dimension.**

I. Introduction

Discrete-domain auto-regressive moving-average (ARMA) processes are popular stochastic models for explaining data with long-range dependencies; these models are used for estimation and classification purposes [1]. Data with long-range dependencies occur in phenomena such as network traffic [2], fading channels [3], fluid velocity [4], solar irradiance [5], automotive traffic [6], and housing investments [7].

These processes consist of a shaping filter that acts on an innovation process, also referred to as an excitation noise. The distribution of the ARMA model is determined by the innovation process, whereas the dependency patterns among samples of the ARMA process is controlled by the shaping filter, which is in turn determined by a set of parameters (AR/MA parameters). The main advantage of ARMA models is that they allow for deriving optimal estimators in certain settings. However, the distribution of a statistical model and more specifically, its compressibility properties, is a key factor in forming an effective model for realistic applications [8].

Despite the important role of ARMA processes, their compressibility is investigated only in special cases, e.g., auto-regressive processes with Gaussian innovations [9]. While Gaussian models often lend themselves to analytical solutions and closed-form expressions, they do not adequately capture a significant portion of natural signals, including spectral, seismic, and biological data. These data are better described by sparsity inducing laws such as Bernoulli-Gaussian [10]. However, conventional approaches fail to measure the compressibility of an ARMA processes with such sparsity inducing laws, and alternative approaches have yet to emerge in the literature.

In this paper, we evaluate the compressibility of ARMA processes in a rather broad setting. Our main result is to show that the information-theoretic compressibility of an ARMA process is equal to that of its innovation process, and independent of its AR/MA parameters. Furthermore, this value coincides with the notion of smooth and robust compressibility in [11] which extends some compressed sensing concepts to stochastic processes. Our approach results are derived using a technique that is recently developed in [12] and quantifies the compressibility of random vectors with affinely singular distributions. In simple words, the singular part (e.g. mass probabilities) of the distribution of these vectors are supported on affine subsets. By applying this technique, we show that finite dimensional samples of ARMA processes with discrete-continuous excitation noise are affinely singular random vectors. In turn, this result enables us to quantify the smooth and robust compressibility of such ARMA processes.

Relevant Literature

Information-theoretic compressibility of discrete-domain stochastic processes is studied mainly in special cases. In [13] it is shown that ARMA processes have the highest entropy among all processes with a given auto-correlation matrix. In [9], it is proved that the rate-distortion function (RDF) of an autoregressive process with Gaussian excitation is equal to that of its excitation noise. A similar result is shown for first-order autoregressive processes, i.e., known as AR(1) processes,

M. Charusaie was with Sharif university of technology, Tehran, Iran, at the time of writing this manuscript and is now with Max Planck Institute for Intelligent Systems, Tübingen, Germany (email: mcharusaie@tuebingen.mpg.de).

A. Amini is with the department of Electrical Engineering, Sharif University of Technology, Tehran, Iran (email: aamini@sharif.edu).

S. Rini is with the Electrical and Computer Engineering department, National Yang-Ming Chiao-Tung University (NYCU), Hsinchu, Taiwan. (email: stefano.rini@nycu.edu.tw)

with discrete excitation. The results of [9] further implies that information dimension rates of such processes are equal to that of their excitation noise. The RDF of non-stationary Gaussian autoregressive processes and vector Gaussian autoregressive processes are studied in [14]–[15] and [16], respectively. The RDF of the process obtained by applying an FIR filter to a general wide-sense stationary Gaussian process is covered in [17]. For this class of processes, the authors of [18] derive the differential entropy rate. A learning-based compression approach is introduced in [19] which achieves the block information dimension on a class of stationary processes; we note that this technique is applicable to some MA processes studied in this paper.

Almost lossless compression rates, also referred to as ϵ-achievable compression rates, are introduced in [11] and studied for i.i.d. processes. In [20], fundamental bounds are derived for worst ϵ-achievable compression-rates of bounded processes. In [12], almost lossless compression rates of MA processes are investigated; the considered processes are examples of processes with affinely singular sample distributions.

Several other measures of compressibility for stochastic processes are proposed in the literature. An energy-based measure of compressibility is introduced and studied for continuous i.i.d. random variables in [21] and for ergodic sequences in [22]. The non-asymptotic lossless compression rates of a random vector has been studied in [23] via modifying the Minkowski dimension of the support set of the vector. For n-dimensional AR(1) processes with Gaussian excitation, the compression rate is evaluated in terms of the Hausdorff dimension in [24].

Continuous-domain ARMA processes in which the excitation noise is a Lévy process are studied in [25]. The compressibility of continuous-domain innovation processes and their comparisons are provided in [26]; the compressibility measure is based on the entropy rate of finely quantized samples. The study of differential entropy of finely sampled Lévy process with continuously distributed excitation noise is achieved in [27].

Contributions

In this paper, we study the compressibility of ARMA processes, in which the excitation noise is not limited to the Gaussian or absolutely continuous probability measures, but a larger class of discrete-continuous measures. Accordingly, we study the information-theoretic measures of compressibility, such as block-average information dimension (BID) and the information dimension rate (IDR), originally introduced in [28] and [29], respectively, and almost lossless compressibility measures, which is introduced in compressed sensing literature [11].

As we discuss in Section II-B, BID and IDR are similar in definition, however, they induce two different approaches towards compressibility. On one hand, the IDR, $d_I\big(\{\mathbf{X}_t\}\big)$, of a process $\{\mathbf{X}_t\}$ quantifies the ratio of minimum number of bits needed to encode the high-resolution (i.e., $m \to \infty$) quantized version $\big\{[\mathbf{X}_t]_m\big\}$ of the process $\{\mathbf{X}_t\}$, to the minimum number of bits needed to encode the quantized version $\big\{[\mathbf{Y}_t]_m\big\}$ of all encode-able processes $\{\mathbf{Y}_t\}$. On the other hand, the BID, $d_B\big(\{\mathbf{X}_t\}\big)$ calculates the average information dimension of truncated samples of a process, for large number of samples.

Following above definitions, one can argue that the IDR encodes the definition of compressibility in a more sensible way. However, finding its value is theoretically more complex, as calculating this measure incolves finding the entropy rate of a quantized version of the stochastic process in a non-asymptotic setting for quantization step-size.

To address this issue, there are some works that show the equality of IDR and BID —where the latter is generally easier to calculate —under some conditions on the process [30], [29]. One of the conditions that generally simplify the evaluation of IDR is finiteness of mutual information among samples of the process. As we will show by an example in Section III-A, this is not the case for ARMA processes with discrete-continuous excitation noise. Although this condition is violated for the mentioned ARMA case, as a first contribution of this work, we show that IDR and BID are still equal.

The classical information-theoretic measures of compressibility (such as entropy and RID) search over all possible functions that can encode the data. Oftentimes, the optimal encoding function demonstrates irregular behavior in response to the input data, making the compression scheme sensitive to noise and non-idealities. This is in contrast with the compressed sensing scenario [31] where by restricting the encoder to linear functions, the effect of noise is kept under control. Besides, under certain conditions (the encoder satisfies the restricted isometry property with a suitable constant) the standard decoder acts as a Lipschitz operator [32] (noise and non-idealities could be linearly bounded in the output).

Drawing inspiration from the compressed sensing scenario, in this paper we mainly study the latent dimension of the optimal linear/Borel and Borel/Lipschitz encoder/decoder pairs. The minimum compression rate among all encoder/decoder pairs such that the decoding error probability is at most $\epsilon \in (0, 1)$ is called the minimum ϵ-achievable compression rate (linear/Borel minimum ϵ-achievable and Borel/Lipschitz minimum ϵ-achievable compression rates). As the second contribution of this paper, we show that both minimum ϵ-achievable rates for ARMA processes coincide with BID and IDR (as BID and IDR are equal here). To elaborate, in the settings similar to compressed sensing where the encoder is a linear function and the decoder ensures the robustness with respect to noise, the optimal rate of compression is equal to the Rényi information dimension rate of the process. The equality of ϵ-achievable compression rates and the IDR was previously shown for an i.i.d. sequence of random variables in [11], and for moving-average processes in [12]. The result in this paper for ARMA processes extends both results.

Notations: In this study, we adopt to the following notation: capitalized letters (e.g., X) represents random variables (RVs), while lowercase letters (e.g., x) denote fixed scalars. Bold capitalized letters (e.g., $\mathbf{X}$) correspond to random vectors, and bold lowercase letters (e.g., $\mathbf{x}$) indicate fixed vectors. The uniform quantization of a random vector with precision $1/m$ is denoted as $[\mathbf{X}^n]_m$. A comprehensive summary of the notation used throughout this work is provided in Appendix A.

II. Preliminaries

In this section, we present useful definitions of probability measures, ARMA processes, affinely singular RVs, and information-theoretic compressibility measures.

A. Types of Measures

The cornerstones of our work are discrete-continuous probability measures. To properly introduce such measures, we start with two simpler definitions and review the well-known Lebesgue-Radon-Nikodym theorem for characterizing generic measures (see [33, p. 121]).

In the following, let Σ be a σ-field of $\mathbb{R}^n$.

Definition 1 (Absolutely continuous measures): We refer to the probability measure $\mu(\cdot)$ on Σ as *absolutely continuous*, if for every set $S \in \Sigma$ with zero Lebesgue measure, we have that $\mu(S) = 0$.

Definition 2 (Singular measures): A measure $\mu(\cdot)$ on Σ is called *singular*, if there exists a subset $S \subset \mathbb{R}^n$ with zero Lebesgue measure such that

$$\mu(\mathbb{R}^n \setminus S) = 0. \tag{1}$$

If S is further countable, then, $\mu(\cdot)$ is called *discrete*.

Theorem 1 (Lebesgue-Radon-Nikodym): Every probability measure μ on $\mathbb{R}^n$ is associated with a unique singular measure $\mu_s(\cdot)$ and an absolutely continuous measure $\mu_c(\cdot)$, such that for a $\alpha \in [0, 1]$ we have

$$\mu = \alpha\, \mu_c + (1 - \alpha)\, \mu_s. \tag{2}$$

We refer to α as the *continuity chance* of the measure $\mu(\cdot)$. If $\mu_s(\cdot)$ is discrete and $\alpha \in (0, 1)$, i.e., α does not take the endpoints of the interval, then, we call μ a α-discrete-continuous probability measure.

B. Compressibility Measures

The classical notion of entropy is well-defined for discrete-valued RVs. This notion is generalized for continuous-valued RVs via the limiting entropy of the quantized RV. The uniform quantization of a RV $\mathbf{X}^n$ with precision m is defined in (28) as $[\mathbf{X}^n]_m \in (\mathbb{N}/m)^n$.

Definition 3 ([34]): The *Rényi information dimension* (RID) for a RV $\mathbf{X}^n$ is

$$d(\mathbf{X}^n) = \lim_{m\to\infty} \tfrac{\mathsf{H}([\mathbf{X}^n]_m)}{\log m}, \tag{3}$$

if the limit exists, where $\mathsf{H}(\cdot)$ is the Shannon entropy function.

Theorem 2 ([34]): For a random variable X with p-discrete-continuous measure, we have $d(X) = p$.

Definition 4 ([28]): For a discrete-domain stationary stochastic process $\{\mathbf{X}_t\}$, the block-average information dimension (BID) is defined as

$$d_B\big(\{\mathbf{X}_t\}\big) = \lim_{n\to\infty} \lim_{m\to\infty} \frac{\mathsf{H}\big([X_n]_m|[\mathbf{X}^{n-1}]_m\big)}{\log m}, \tag{4}$$

if the limit exists. If $\lim_{m\to\infty}$ in (4) does not exist, $\overline{d}_B\big(\{\mathbf{X}_t\}\big)$ and $\underline{d}_B\big(\{\mathbf{X}_t\}\big)$ represent (4) when $\lim_{m\to\infty}$ is replaced with $\limsup_{m\to\infty}$ and $\liminf_{m\to\infty}$, respectively.

Theorem 3: [28, Lem. 3] For a discrete-domain stationary stochastic process $\{\mathbf{X}_t\}$ the BID equals

$$\begin{aligned} d_B\big(\{\mathbf{X}_t\}\big) &= \lim_{n\to\infty} \lim_{m\to\infty} \frac{\mathsf{H}\big([\mathbf{X}^n]_m\big)}{n \log m} \\ &= \lim_{n\to\infty} \frac{1}{n} d(\mathbf{X}^n). \end{aligned} \tag{5}$$

Definition 5 ([29]): For a discrete-time stochastic process $\{\mathbf{X}_t\}$, *information dimension rate* (IDR) is defined as

$$d_I\big(\{\mathbf{X}_t\}\big) = \lim_{m\to\infty} \lim_{n\to\infty} \frac{\mathsf{H}\big([\mathbf{X}^n]_m\big)}{n \log m}, \tag{6}$$

if $\lim_{m\to\infty}$ exists; otherwise, by replacing $\lim_{m\to\infty}$ with $\liminf_{m\to\infty}$ and $\limsup_{m\to\infty}$, we can define $\underline{d}_I\big(\{\mathbf{X}_t\}\big)$ and $\overline{d}_I\big(\{\mathbf{X}_t\}\big)$, respectively.

Theorem 4 (Thm. 14,[29]): For a discrete-domain stationary stochastic process $\{\mathbf{X}_t\}$, we have that

$$d_I\big(\{\mathbf{X}_t\}\big) \le d_B\big(\{\mathbf{X}_t\}\big). \tag{7}$$

The smooth and robust compression/decompression methods are studied in [35] and [11]. For the following compressibility measures, we use the notations introduced in [11] which are inspired by similar notions in the field of compressed sensing [36].

Definition 6: [[11]] Let $\{\mathbf{X}_t\}$ be a discrete-domain stochastic process. For an integer n, we refer to $f_n : \mathbb{R}^n \to \mathbb{R}^{\lfloor nR_n \rfloor}$ and $g_n : \mathbb{R}^{\lfloor nR_n \rfloor} \to \mathbb{R}^n$ as an ϵ-encode/decode pair with rate R_n (for the process), once we have

$$\mathbb{P}\Big(g_n\big(f_n(\mathbf{X}^n)\big) \neq \mathbf{X}^n\Big) \le \epsilon. \tag{8}$$

For a given $\epsilon > 0$, we define the minimum ϵ-achievable rate as $\liminf_{n\to\infty} R_n$. Further, if we restrict the encoder f_n to be linear, we obtain the linear-encode ϵ-achievable rate represented by $R^*(\epsilon)$. Alternatively, if we restrict the decoder to be Lipschitz, then, we obtain the minimum Lipschitz-decode ϵ-achievable rate denoted by $R(\epsilon)$.

C. Stochastic Processes

A stationary process is a stochastic process whose unconditional joint probability distribution does not change when shifted in time.

Definition 7 (Stationary process): A stochastic process $\{\mathbf{X}_t\}$ is "strictly stationary" if

$$\mu_{X_{i_1+l},\ldots,X_{i_n+l}}(\cdot) = \mu_{X_{i_1},\ldots,X_{i_n}}(\cdot),$$

for every $n \in \mathbb{N}$ and $l \in \mathbb{Z}$.

The class of *Auto-Regressive Moving-Average* (ARMA) processes is defined as following:

Definition 8 (ARMA Process [37]): A (p, q)-ARMA process with $p, q \in \mathbb{N}$ is defined by the recursive expression

$$X_t = \xi_t + \sum_{i\in[q]} \theta_i \xi_{t-i} - \sum_{i\in[p]} \phi_i X_{t-i}, \tag{9}$$

for $t \in \mathbb{Z}$ and where the constants $\phi_i i \in [p]$ and $\theta_i i \in [q]$ represent the autoregressive (AR) and moving-average (MA) parameters, respectively. The sequence $\{\xi_i\}_{i\in\mathbb{Z}}$ is defined as the excitation process and is a sequence of i.i.d. RVs.

Remark 1: A stationary ARMA process is an ARMA(p,q) process that is strictly stationary. The sufficient and necessary condition for an ARMA(p,q) process to be stationary is provided in Appendix C.

When the excitation white noise in an ARMA process has a discrete-continuous distribution, the vectors of ARMA samples might also have singular components in their distribution. Due to the linear mixture of the white noise samples, these components form affine subsets. The class of vectors with singularities over affine sets was the topic of our previous publication [38]. We refer to this class of vectors as Affinely Singular Random Vectors (*Affinely Singular Random Vectors* (ASRV)).

Definition 9 ([12]): A random vector $\mathbf{Z}^n$ is an ASRV if there exists a finite or countably infinite number of affine subsets $\mathcal{A}_i \subset \mathbb{R}^n$ of dimension $0 \leq d_i \leq n$ and absolutely continuous [1] d_i-dimensional measures μ_i over $\mathcal{A}_i$ such that

$$\forall\, \mathcal{B} \subseteq \mathbb{R}^n :\ \mathbb{P}\big(\mathbf{Z}^n \in \mathcal{B}\big) = \sum_i \mu_i\big(\mathcal{B} \cap \mathcal{A}_i\big). \tag{10}$$

We finally come to a definition of *Discrete–Continuous Excitation Autoregressive–Moving–Average* (DCE-ARMA) processes; a stationary ARMA process in which the excitation noise is discrete-continuous, so that samples of the process are ASRV.

Definition 10: A DCE-ARMA process is defined as the (p,q)-ARMA processes in which the excitation process $\{\xi_t\}$ has a discrete-continuous distribution. Such a process is indicated as (p,q)-DCE-ARMA process.

In the next section, we analyze the compressibility of DCE-ARMA processes using the measures introduced in Sec. II-B.

III. Main Results

We begin our analysis by deriving information-theoretic measures of compressibility for a DCE-ARMA process.

A. BID and IDR of a DCE-ARMA process

In this section, we use the calculus of information dimension to study the BID and IDR of a DCE-ARMA process. In particular, in the next theorem, we show that the BID of all stationary ARMA processes, whether the excitation noise is discrete-continuous or not, is equal to the information dimension of their excitation noise.

Under mild conditions, we further establish that IDR, BID, and RID of the excitation noise are identical.

Theorem 5: If $d(\xi_1)$ is well-defined for the i.i.d. excitation process $\{\xi_t\}$, then, for the resulting stationary ARMA process $\{\mathbf{X}_t\}$ we have

$$d_B\big(\{\mathbf{X}_t\}\big) = d(\xi_1). \tag{11}$$

[1]Here, we abused the notion of absolute continuity. In fact, if f_i is the affine transformation with which we can generate the set $\mathcal{A}_i$, then, by the absolute continuity over $\mathcal{A}_i$ we mean the absolute continuity with respect to the pushforward measure $\ell_i(f_i^{-1}(\cdot))$, where ℓ_i is a d_i-dimensional Lebesgue measure.

Further, if $H\big([\xi_1]_1\big) < \infty$, then,

$$d_I\big(\{\mathbf{X}_t\}\big) = d_B\big(\{\mathbf{X}_t\}\big) = d(\xi_1). \tag{12}$$

While the full proof is provided in Appendix F, we present a proof sketch of (11), which establishes the equality of the BID and the RID of the excitation noise, to highlight the key techniques used. Through this sketch, we use several properties of information dimension, namely, Properties I through IV, that were previously introduced in [39]. Firstly, due to the recurrence relation of the ARMA process $\{\mathbf{X}_t\}$, we prove that a truncation $\mathbf{X}^{m+p}$ of this process, where m is an arbitrary integer, and p is the number of poles of the process, can be linearly obtained from the excitation noise $\boldsymbol{\xi}_{1-q+p}^{m+p}$ and a smaller truncation $\mathbf{X}_1^p$. Next, we use the following property to show that in this case, the information dimension can only increase $d(\mathbf{X}^{m+p}) \leq d(\boldsymbol{\xi}_{1-q+p}^{m+p}, \mathbf{X}_1^p)$.

- **Property I:** A Lipschitz function increases the information dimension of a variable.

Next, using the following properties, we can further upper-bound the right-hand side of the inequality by $d(\boldsymbol{\xi}_{1-q+p}^{m+p}) + d(\mathbf{X}_1^p) \leq (m+q)d(\xi_1) + p$.

- **Property II:** The information dimension of a random vector $\mathbf{X}^m$ is bounded above by m.
- **Property III:** For two independent random vectors $\mathbf{X}^m$ and $\mathbf{Y}^n$, the joint information dimension is equal to the sum of the information dimension of each, i.e., $d(\mathbf{X}^m, \mathbf{Y}^n) = d(\mathbf{X}^m) + d(\mathbf{Y}^n)$.

Furthermore, once more due to the linear dependence of process and the excitation noise, and as a result of the following property, we can lower-bound the information dimension as $d(\mathbf{X}^{m+p}) \geq d(\boldsymbol{\xi}_{1+p}^{m+p}) = m d(\xi_1)$.

- **Property IV:** For two independent random vectors $\mathbf{X}^m$ and $\mathbf{Y}^m$, the information dimension of the sum of random vectors is bounded below by the information dimension of each, i.e., $d(\mathbf{X}^m + \mathbf{Y}^m) \geq \max\{d(\mathbf{X}^m), d(\mathbf{Y}^m)\}$.

As a result, the limit $\lim_{m\to\infty} \frac{d(\mathbf{X}^{m+p})}{m+p}$ that quantifies the BID is equal to $d(\xi_1)$. The upper- and lower-bound of this proof is illustrated for an AR(2) process in Figure 1.

Here, we should mention that the technique that is used in proving the equality of $d_I\big(\{X_t\}\big) = d_B\big(\{X_t\}\big)$ and its consequences vary from the ones investigated in [29] and [30]. This literature establishes sufficient conditions on mutual information among process samples for ensuring this equality; however, these conditions do not necessarily hold in our setting. In particular, [29] shows that the equality holds when there exists a nonnegative integer n for which

$$I(\mathbf{X}_1^k; \mathbf{X}_{-\infty}^{-n}) < \infty, \quad k = 1, 2, \infty. \tag{13}$$

However, this condition is not necessarily satisfied in the case of DCE-ARMA processes. To see this, let $\{\mathbf{X}_t\}$ be an $AR(1)$ process; using the Markovity of the process, we have that

$$I(\mathbf{X}_1^k; \mathbf{X}_{-\infty}^{-n}) = I(X_1; X_{-n}). \tag{14}$$

Moreover, we know

$$X_1 = \sum_{k=-n+1}^{1} (-\phi_1)^{1-k} \xi_k + (-\phi_1)^{n+1} X_{-n}$$

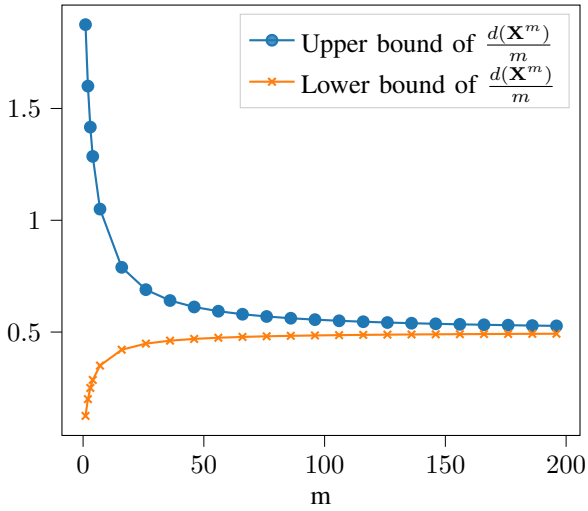


Fig. 1: Lower-bound and upper-bound of the average RID of an AR(2) process with $d(\xi_1) = 0.5$ that is derived in Theorem 5, and in terms of the number of samples of the process.

$$= U + (-\phi_1)^{n+1} X_{-n}, \tag{15}$$

where U is a discrete-continuous RV with $d(U) = 1 - \big(1 - d(\xi_1)\big)^n$ using [26, Lem. 11]. Equivalently, with probability $\big(1 - d(\xi_1)\big)^n$, U takes a value from a countable set. Therefore, if we choose a member x_d from that set[2], then $\mathbb{P}(U = x_d) := p > 0$, and the joint probability measure $\mu_{X_1 X_{-n}}(\cdot)$ will have a nonzero probability p on the line $X_1 - (-\phi_1)^{n+1} X_{-n} = x_d$. As an instance, if ξ_1 takes the value 0 with probability $1 - d(\xi_1)$, then U takes the value 0 with probability $p = \big(1 - d(\xi_1)\big)^n$, and therefore, the line $X_1 - (-\phi_1)^{n+1} X_{-n} = 0$ will have a nonzero probability p. As we show in Lemma 5, we know that X_1 and X_{-n} both have absolutely continuous measures. Hence, the product measure $\mu_{X_1} \times \mu_{X_{-n}}(\cdot)$ is absolutely continuous. As a result, $\mu_{X_1 X_{-n}}(\cdot)$ has singularity w.r.t the measure $\mu_{X_1} \times \mu_{X_{-n}}(\cdot)$ which prevents the Radon-Nikodym derivative and the mutual information from being well-defined (see Figure 2 for illustration). Following from the discussion above, we conclude that Theorem 5 does not directly follow from the results in [29] and [30].

B. Smooth and robust compressibility of a DCE-ARMA process

In the previous section, we showed that the extent to which an ARMA process can be compressed is directly connected to the RID of each instance of its excitation noise. This compressibility rate is determined by the Information Dimension Rate (IDR), an extension of Shannon's entropy. Similar to entropy, IDR evaluates all potential compression functions for the process. As an alternative to that approach, the compressed sensing field focuses on such compression and decompression pairs of functions. It assumes linearity in the former and smoothness in the latter to make the system simple and robust to noise.

[2]The subscript d indicates the discrete set of which x_d is a member.

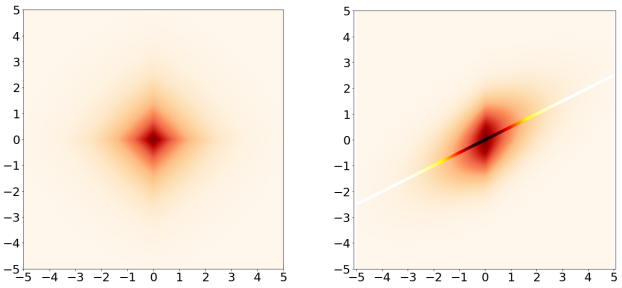


Fig. 2: (a) Product measure of two samples of an AR(1) process generated by Bernoulli Gaussian excitation noise with the RID 0.5 and $\phi_1 = 0.5$, and (b) their joint measure.

In this section, we extend our results to the cases of linear compression and smooth decompression and quantify minimum ϵ-achievable rate $R(\epsilon)$ as in Definition 6 for DCE-ARMA processes. To achieve that, we first need to generalize Theorem 5 to further specify the probability distribution of each truncation of an ARMA process. In particular, Theorem 5 showed that the BID of the process is $d(\xi_1)$. In case each truncation of the process has an affinely singular probability distribution, then as shown in [38, Lemma 2] such result is equivalent to the average dimension of the singular components of probability measure converge to $d(\xi_1)$. In the following theorem, we show that the dimensions of such components further do not only have the average equal to $d(\xi_1)$ but further is concentrated around that value.

Theorem 6: For a stationary (p, q)-DCE-ARMA process in Definition 10, it holds that:

(i) if the excitation noise is absolutely continuous, then, the vector of truncated samples $\mathbf{X}^n$ is also absolutely continuous,

(ii) if the excitation noise $\{\xi_i\}$ is α-discrete-continuous, then, the vector of truncated samples $\mathbf{X}^n$ is affinely singular when $n \geq p+1$. In addition, the affinely singular distribution of $\mathbf{X}^n$ can be expressed as

$$\mathbf{X}^n \overset{d}{=} S_V[\mathbf{Y}_V; 0_{n-d_V}] + \mathbf{e}_V, \tag{16}$$

where S_is are unitary $n \times n$ matrices, V is a discrete RV on $\mathbb{N}$, and $\mathbf{e}_i$s are fixed n-dimensional vectors. Furthermore, $\mathbf{Y}_i$s are d_i-dimensional absolutely continuous RV and

$$\mathbb{P}(d_V > k) \geq 1 - \exp\Big(- (n+q-p) D\big(\tfrac{k+q-p}{n+q-p} \| \alpha\big)\Big), \tag{17}$$

for $\frac{k+q-p}{n+q-p} < \alpha$ and

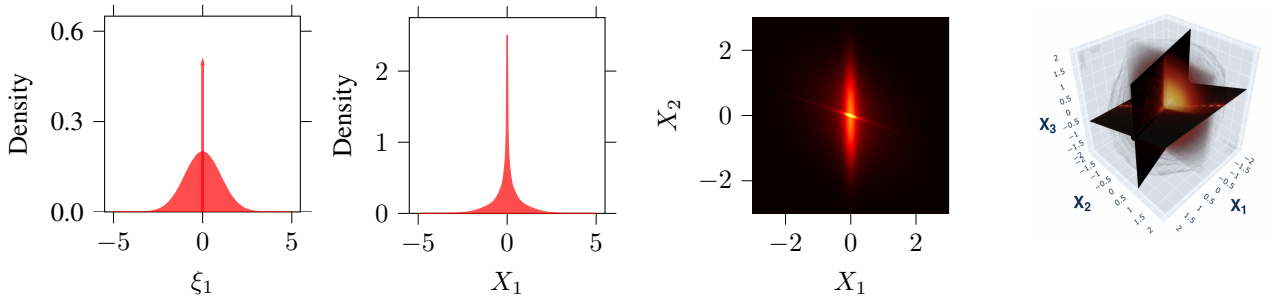
$$\mathbb{P}(d_V < k) \geq 1 - \exp\Big(- (n+q-p) D\big(\tfrac{k-p}{n-p} \| \alpha\big)\Big), \tag{18}$$

for $\frac{k-p}{n-p} > \alpha$.

Proof: See Appendix E. ∎

In the following example, we show that how such affinely singular random vector is formed.

Example 1 (Affine Singularity): Assume that the ARMA



(a) μ_{ξ_1} (b) μ_{X_1} (c) $\mu_{X_1X_2}$ (d) $\mu_{X_1X_2X_3}$

Fig. 3: The distribution of (a) excitation noise, (b) a single sample, (c) two sequential samples, and (d) three sequential samples from the AR(1) process that is discussed in Example 1

process defined recursively as

$$X_t = \xi_t + \sum_{i=1}^{\infty} h_i X_{t-i}, \tag{19}$$

where $h_1 = 1/3$ and $h_i = 0$ for all other values $i \neq 1$. Assume that the excitation noise is standard Bernoulli-Gaussian (i.e., $\xi_t = b_t n_t$ where $b_t \sim Bern(1/2)$ and $n_t \sim \mathcal{N}(0,1)$). For this excitation, as we show in Lemma 5, the probability distribution of the sample X_1 is absolutely continuous. To be more precise, with a simple calculation, one can show that the distribution of X_1 is a uniform mixture of Gaussian distributions with variance $\sigma^2 = \frac{t}{9^k}$ for limiting case of $k \to \infty$ for $t \in T_k$ where T_k is a set of either powers of 9 or a sum of distinct powers of 9 that are less than 9^k. Furthermore, due to (19), $\mathbf{X}_2^3$ can be obtained as a function of the tuple (X_1, ξ_2, ξ_3) as

$$X_2 = \frac{X_1}{3} + \xi_2 \tag{20a}$$

$$X_3 = \frac{X_2}{3} + \xi_3 = \frac{X_1}{9} + \frac{\xi_2}{3} + \xi_3. \tag{20b}$$

These identities together with the definition of ξ_t illustrate the underlying distribution of $\mathbf{X}_1^3$. In fact, in case of $\xi_2 = 0$ and when ξ_3 takes values according to the Gaussian distribution, that is an event with probability $\mathbb{P}(b_2 = 0, b_3 = 1) = 1/4$, we have $X_2 = X_1/3$. This event induces a plane on which $\mathbf{X}_1^3$ is distributed. Similarly, the event $\xi_3 = 0$ concludes in plane $X_3 = X_2/3$ on which the random vector is distributed. If we further assume the event $\xi_2 = \xi_3 = 0$, then we conclude that the random vector is distributed along the line $X_3 = X_2/3 = X_1/9$. The manifestation of singularity in such AR(1) processes is demonstrated in Figure 3. As we see, while the marginal distribution of X_1 is absolutely continuous, the atomic singularity of ξ_1 leads to the emergence of singularities on lower-dimensional sub-spaces in joint probability measures of $\mu_{X_1X_2}$ and $\mu_{X_1X_2X_3}$.

Remark 2: To explain the intuition behind the inequalities (17) and (18), we recall the recurrence relation of the samples of an ARMA process as stated in Definition 8. Using this relation, we observe that the value of each truncation $\boldsymbol{X}^n$ of the process can be obtained using (i) the value of the first p samples $\boldsymbol{X}^p$ and (ii) the value of the excitation noise ξ_{p-q+1}^n. More concretely, the value of $\boldsymbol{X}^n$ is obtained by a matrix multiplication on the concatenation of $\boldsymbol{X}^p$ and $\boldsymbol{\xi}_{p-q+1}^n$.

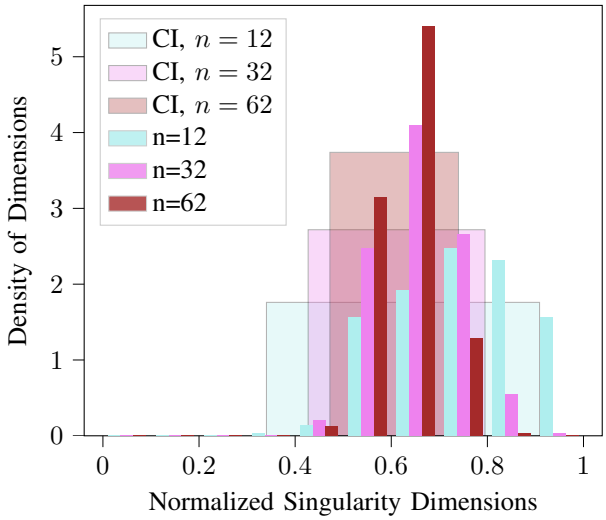


Fig. 4: The concentration of normalized singularity dimensions for a (2, 3)-DCE-ARMA process with $d(\xi_1) = 0.6$ compared with the concentration inequality (CI) of Corollary 1 with 80% confidence

The applied matrix is block-diagonal and is composed of the identity matrix I_p and a Hankel matrix H. In the proof of Theorem 6, we show that if a binary vector $\boldsymbol{s} \in \{0,1\}^{n-p+q}$ controls whether each component of the excitation variables $\boldsymbol{\xi}_{p-q+1}^n$ takes value from the continuous part of its distribution, then, the result of the matrix multiplication lies on an affine set with dimension $p + \text{rank}(H^{[\boldsymbol{s}]})$. Here, $H^{[\boldsymbol{s}]}$ is the matrix formed by columns $\mathcal{I} \subseteq \mathbb{Z}$ of the Hankel matrix where $s_i = 1$ for each $i \in \mathcal{I}$. Finally, the probability distribution of $\text{rank}(H^{[\boldsymbol{s}]})$ where each component of $\boldsymbol{s}$ takes value from a Bernoulli distribution $Bern(\alpha)$ is calculated in [38, Lemma 8], and is shown to take the form of (17) and (18).

The result of Theorem 6 can further be rephrased as an (ϵ, δ) convergence criteria as in the following, showing that if the width of the truncation is large enough, then with probability at least $1 - \epsilon$, the dimensions of singular components are concentrated within the interval $\big[n(\alpha - \delta), n(\alpha + \delta)\big]$.

Corollary 1: For a (p, q)-DCE-ARMA process and given

$\epsilon, \delta > 0$, set

$$n \geq \max\Bigg\{ \frac{2\big[q(1+\delta/2-\alpha)-p\big]}{\delta}, \frac{-\log\frac{\epsilon}{2}}{\mathsf{D}(\alpha-\delta/2\|\alpha)} - q, \frac{2p}{\delta}, \frac{-\log\frac{\epsilon}{2}}{\mathsf{D}(\alpha+\delta/2\|\alpha)} - q \Bigg\}, \tag{21}$$

and define $k_- = n(\alpha - \delta)$ and $k_+ = n(\alpha + \delta)$. Theorem 6 implies that $\mathbb{P}(d_i > k_-) \geq 1 - \frac{\epsilon}{2}$ and $\mathbb{P}(d_i < k_+) \geq 1 - \frac{\epsilon}{2}$, which in turn shows that

$$\mathbb{P}\big(\big|\tfrac{d_i}{n} - \alpha\big| < \delta\big) \geq 1 - \epsilon. \tag{22}$$

Figure 4 illustrates the concentration of measure described in the above corollary.

To leverage such concentration property in our compressibility analysis, we now review a result that connects the BID of certain sources to their minimum ϵ-achievable and Minkowski dimension compression rates.

Theorem 7 ([12]): Let $\{\mathbf{Z}_t\}$ be a discrete-domain stochastic process such that each sequence of length n from its samples has an affinely singular distribution (see Definition 9) with finitely many affine subsets $\{\mathcal{A}_i\}_{i=1}^{k_n}$ with dimensions $\{d_i\}_{i=1}^{k_n}$ and corresponding measures $\{\mu_i\}_{i=1}^{k_n}$. Let V_n be a discrete random variable that takes the values $1 \leq i \leq k_n$ with probability $\mu_i(\mathcal{A}_i)$. If for all $\epsilon, \delta \in \mathbb{R}^+$, there exists a large enough n such that

$$\mathbb{P}\left(\left|\tfrac{d_{V_n}}{n} - d_B\big(\{\mathbf{Z}_t\}\big)\right| < \delta\right) > 1 - \epsilon, \tag{23}$$

then, for the process $\{\mathbf{Z}_t\}$ (the typical source) we know that

$$R^*(\epsilon) = R_B(\epsilon) = R(\epsilon) = d_B\big(\{\mathbf{Z}_t\}\big), \tag{24}$$

where $R_B(\epsilon)$ is the Minkowski-dimension compression rate defined in [11, Definition 10].

Based on Theorem 6 and Corollary 1, we know that (23) holds for DCE-ARMA processes with finite discrete space. Therefore, Theorem 7 implies that (24) is valid for such processes:

Theorem 8: Let the DCE-ARMA process $\{\mathbf{Z}_t\}$ as in Definition 10 be such that

$$\xi_i = \nu_i X_{c,i} + (1 - \nu_i) X_{D,i}, \tag{25}$$

where ν_is are i.i.d Bernoulli RVs with $\mathbb{P}(\nu_i = 1) = \alpha$, $X_{c,i}$s are absolutely continuous i.i.d RVs, and $X_{D,i}$s are discrete i.i.d RVs with $X_{D,i} \in \mathcal{D}$ for a finite subset $\mathcal{D}$ of $\mathbb{R}$. For the process $\{\mathbf{Z}_t\}$ we have

$$R^*(\epsilon) = R_B(\epsilon) = R(\epsilon) = d(\xi_1) = \alpha. \tag{26}$$

Proof: See Appendix G. ∎

It is important to note that when the excitation noise is purely discrete ($\alpha = 0$), the truncated process samples are no longer affinely singular. However, Theorem 8 remains valid.

In the following, we provide an example of why such a process does not generate affinely singular random vectors.

Example 2 (Self-similar Singularity): A widely-studied instance of the singularity in ARMA processes occurs when the process is AR(1) and the excitation noise is a Rademacher process, i.e.,

$$\xi_k = \begin{cases} -1 & 1/2 \\ +1 & 1/2 \end{cases}. \tag{27}$$

As an instance, if the process is according to (19) and where $h_1 = 1/3$ and $h_i = 0$ for all other values $i \neq 1$, then, X_N corresponds to a Bernoulli convolution of order N, and [40], [41] show that such convolution converges in limit $N \to \infty$ to a scaled version of a random variable with Cantor's distribution (see Fig. 5). Since Cantor's set has 0 Lebesgue measure [42, Chapter 2, Section 5], and because the probability on such set is 1, then the distribution is not absolutely continuous as defined in Definition 2. Furthermore, since such set is not countable [42, Corollary 62], then such distribution is further not discrete. Therefore, such distribution is not in form of the distribution of affinely singular random vectors.

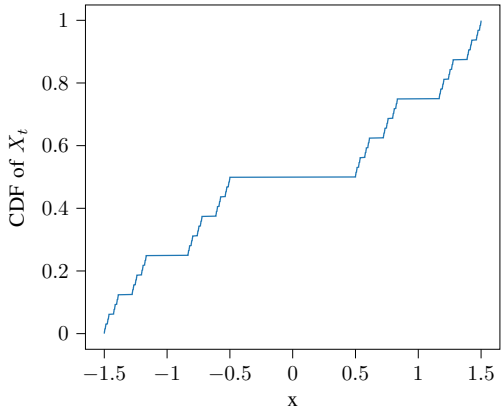


Fig. 5: The CDF of a Bernoulli convolution (a realization of the AR(1) process $X_t = \xi_t + aX_{t-1}$ with Rademacher excitation noise) for $a = 1/3$ which coincides with a scaled Cantor function.

IV. Discussion

The results in Section III-A and III-B support the claim that an ARMA filter preserves the compressibility of the underlying excitation noise, regardless of how compressible is each component of the resulting process.

Firstly, Theorem 5 and 8 show that if the excitation noise is purely discrete, then the distribution of truncated samples of the resulting DCE-ARMA process contains a singularity such that its compressibility measure of Section III-A and III-B behave similarly to that of a purely discrete measure. This means that asymptotically the necessary information for reconstructing the discretized version of the process is concentrated within a diminishing number of bits per sample, as we increase the number of samples. At the same time, the more samples of the process we want to linearly compress and robustly recover, the less the ratio of the hidden variables that we need.

Secondly, when the excitation process is purely continuous, Theorem 6 shows that the truncated DCE-ARMA process has

a continuous distribution. This means that theoretically, it is impossible to compress this process with a non-trivial ratio.

Finally, the second part of Theorem 6 shows that when the excitation process is discrete-continuous then, the truncated DCE-ARMA process has affinely singular distribution where the dimension of the underlying affine subsets concentrate around the continuity chance of the excitation noise. This concludes in Theorem 8 which expresses that the optimal compressibility rates of the ARMA process are equal to the continuity chance of the excitation noise.

V. Conclusion

In this paper, we studied the compressibility of a class of ARMA processes from an information-theoretical perspective. More specifically, we considered the discrete-continuous DCE-ARMA processes, the excitation process of which have discrete-continuous distributions. We showed that the discrete part of the excitation process induces certain types of singularity in the distribution of the sample path of the ARMA process, which greatly affects the overall compressibility. Besides evaluating various compressibility measures for these processes such as sample RID of the excitation process, minimum ϵ-achievable compression rates of the ARMA process, BID of the ARMA process, and the IDR of the ARMA process, we proved their equality in this special case.

References

[1] J. A. Cadzow, "ARMA modeling of time series," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-4, no. 2, pp. 124–128, 1982.

[2] M. Laner, P. Svoboda, and M. Rupp, "Parsimonious fitting of long-range dependent network traffic using ARMA models," *IEEE Communications Letters*, vol. 17, no. 12, pp. 2368–2371, 2013.

[3] H. Mehrpouyan and S. D. Blostein, "Arma synthesis of fading channels," *IEEE Transactions on Wireless Communications*, vol. 7, no. 8, pp. 2846–2850, 2008.

[4] S. de Waele and P. M. Broersen, "Reliable LDA-spectra by resampling and arma-modeling," *IEEE Transactions on Instrumentation and Measurement*, vol. 48, no. 6, pp. 1117–1121, 1999.

[5] M. David, F. Ramahatana, P.-J. Trombe, and P. Lauret, "Probabilistic forecasting of the solar irradiance with recursive ARMA and garch models," *Solar Energy*, vol. 133, pp. 55–72, 2016.

[6] J. Klepsch, C. Klüppelberg, and T. Wei, "Prediction of functional ARMA processes with an application to traffic data," *Econometrics and Statistics*, vol. 1, pp. 128–149, 2017.

[7] W. Miles, "Irreversibility, uncertainty and housing investment," *The Journal of Real Estate Finance and Economics*, vol. 38, no. 2, pp. 173–182, 2009.

[8] E. Bostan, U. S. Kamilov, M. Nilchian, and M. Unser, "Sparse stochastic processes and discretization of linear inverse problems," *IEEE Transactions on Image Processing*, vol. 22, no. 7, pp. 2699–2710, 2013.

[9] R. Gray, "Information rates of autoregressive processes," *IEEE Transactions On Information Theory*, vol. 16, no. 4, pp. 412–421, 1970.

[10] C. Soussen, J. Idier, D. Brie, and J. Duan, "From bernoulli–gaussian deconvolution to sparse signal restoration," *IEEE Transactions on Signal Processing*, vol. 59, no. 10, pp. 4572–4584, 2011.

[11] Y. Wu and S. Verdú, "Rényi information dimension: Fundamental limits of almost lossless analog compression," *IEEE Transactions on Information Theory*, vol. 56, no. 8, pp. 3721–3748, 2010.

[12] M. A. Charusaie, S. Rini, and A. Amini, "On the compressibility of affinely singular random vectors," in *2020 IEEE International Symposium on Information Theory (ISIT)*, 2020, pp. 2240–2245.

[13] J. Franke, "ARMA processes have maximal entropy among time series with prescribed autocovariances and impulse responses," *Advances in applied probability*, vol. 17, no. 4, pp. 810–840, 1985.

[14] R. M. Gray and T. Hashimoto, "A note on rate-distortion functions for nonstationary gaussian autoregressive processes," *IEEE Transactions On Information Theory*, vol. 54, no. 3, pp. 1319–1322, 2008.

[15] T. Hashimoto and S. Arimoto, "On the rate-distortion function for the nonstationary gaussian autoregressive process (corresp.)," *IEEE Transactions on Information Theory*, vol. 26, no. 4, pp. 478–480, 1980.

[16] V. Kafedziski, "Rate distortion of stationary and nonstationary vector Gaussian sources," in *IEEE/SP 13th Workshop on Statistical Signal Processing, 2005*. IEEE, 2005, pp. 1054–1059.

[17] J. Gutiérrez-Gutiérrez and P. M. Crespo, "Asymptotically equivalent sequences of matrices and Hermitian block Toeplitz matrices with continuous symbols: Applications to MIMO systems," *IEEE Transactions on Information Theory*, vol. 54, no. 12, pp. 5671–5680, 2008.

[18] J. Gutierrez-Gutierrez and P. M. Crespo, "Asymptotically equivalent sequences of matrices and multivariate ARMA processes," *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 5444–5454, 2011.

[19] S. Jalali, "Toward theoretically founded learning-based compressed sensing," *IEEE Transactions on Information Theory*, vol. 66, no. 1, pp. 387–400, 2020.

[20] Y. Gutman and A. Spiewak, "Metric mean dimension and analog compression," *IEEE Transactions on Information Theory*, vol. 66, no. 11, pp. 6977–6998, 2020.

[21] A. Amini, M. Unser, and F. Marvasti, "Compressibility of deterministic and random infinite sequences," *IEEE Transactions on Signal Processing*, vol. 59, no. 11, pp. 5193–5201, 2011.

[22] J. F. Silva and M. S. Derpich, "On the characterization of ℓ_p-compressible ergodic sequences," *IEEE Transactions on Signal Processing*, vol. 63, no. 11, pp. 2915–2928, 2015.

[23] G. Alberti, H. Bölcskei, C. De Lellis, G. Koliander, and E. Riegler, "Lossless analog compression," *IEEE Transactions on Information Theory*, vol. 65, no. 11, pp. 7480–7513, 2019.

[24] G. Lawler *et al.*, "Hausdorff dimension of cut points for Brownian motion," *Electronic Journal of Probability*, vol. 1, 1996.

[25] P. J. Brockwell and T. Marquardt, "Lévy-driven and fractionally integrated ARMA processes with continuous time parameter," *Statistica Sinica*, pp. 477–494, 2005.

[26] H. Ghourchian, A. Amini, and A. Gohari, "How compressible are innovation processes?" *IEEE Transactions on Information Theory*, vol. 64, no. 7, pp. 4843–4871, July 2018.

[27] J. Fageot, A. Fallah, and T. Horel, "Entropic compressibility of Lévy processes," 2020.

[28] S. Jalali and H. V. Poor, "Universal compressed sensing for almost lossless recovery," *IEEE Transactions on Information Theory*, vol. 63, no. 5, pp. 2933–2953, 2017.

[29] B. C. Geiger and T. Koch, "On the information dimension of stochastic processes," *EEE Transactions on Information Theory*, 2019.

[30] F. E. Rezagah, S. Jalali, E. Erkip, and H. V. Poor, "Compression-based compressed sensing," *IEEE Transactions on Information Theory*, vol. 63, no. 10, pp. 6735–6752, 2017.

[31] D. L. Donoho, "Compressed sensing," *IEEE Transactions on information theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[32] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE transactions on information theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[33] W. Rudin, *Real and complex analysis*. Tata McGraw-hill education, 2006.

[34] A. Rényi, "On the dimension and entropy of probability distributions," *Acta Mathematica Hungarica*, vol. 10, no. 1-2, pp. 193–215, 1959.

[35] A. Montanari and E. Mossel, "Smooth compression, gallager bound and nonlinear sparse-graph codes," in *2008 IEEE International Symposium on Information Theory*, 2008, pp. 2474–2478.

[36] E. J. Candés, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.20124

[37] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[38] M.-A. Charusaie, A. Amini, and S. Rini, "Compressibility measures for affinely singular random vectors," *IEEE Transactions on Information Theory*, vol. 68, no. 9, pp. 6245–6275, 2022.

[39] Y. Wu, "Shannon theory for compressed sensing," Ph.D. dissertation, Princeton University, 2011.

[40] R. Kershner and A. Wintner, "On symmetric bernoulli convolutions," *American Journal of Mathematics*, vol. 57, no. 3, pp. 541–548, 1935.

[41] B. Picinbono and J. . Tourneret, "Singular random signals," *IEEE Transactions on Signal Processing*, vol. 53, no. 2, pp. 499–504, 2005.

[42] C. C. Pugh, *Real mathematical analysis*. Springer, 2015.

[43] F. R. Gantmakher, *The theory of matrices*. American Mathematical Soc., 1959, vol. 131.

[44] P. J. Brockwell and A. Lindner, "Strictly stationary solutions of autoregressive moving average equations," *Biometrika*, vol. 97, no. 3, pp. 765–772, 2010.

[45] W. Feller, *An introduction to probability theory and its application*. John Wiley and Sons, 1971, vol. II.

[46] O. Nielson *et al.*, *An introduction to integration and measure theory*. John Wiley & Sons, Inc., New York., 1997.

[47] P. Billingsley, *Probability and measure*. John Wiley & Sons, 2008.

[48] K. B. Athreya and S. N. Lahiri, *Measure theory and probability theory*. Springer Science & Business Media, 2006.

[49] A. Gyorgy, T. Linder, and K. Zeger, "On the rate-distortion function of random vectors and stationary sources with mixed distributions," *IEEE Transactions on Information Theory*, vol. 45, no. 6, pp. 2110–2115, 1999.

APPENDIX A
NOTATIONS

The following notations are adopted in this work.

• *Random variables (RV) and distributions:* the set of RVs $\{X_m, \dots X_n\}$ is abbreviated as $\{X_i\}_{i=m}^n$. For brevity, we define $\{X_i\} = \{X_i\}_{i=-\infty}^{\infty}$. When this set of random variables is used to construct a random vector, we employ the notation $\mathbf{X}_m^n = [X_m, X_{m+1}, \dots, X_n]$ with $n \geq m$. Again, when $m = 1$, the subscript is omitted, i.e. $\mathbf{X}_1^n = \mathbf{X}^n$. By abuse of notation, for a binary vector $\mathbf{s} \in \{0,1\}^n$, $\mathbf{X}^{\mathbf{s}}$ denotes a random vector formed by the elements X_i of $\mathbf{X}^n$, where $s_i = 1$. Equality in distribution is indicated as $\stackrel{\mathrm{d}}{=}$. The discrete/continuous part of the RV X is indicated as $X_{\mathrm{d}}/X_{\mathrm{c}}$, respectively. The Bernoulli RV with success probability p is indicated as $\mathrm{Bern}(p)$.

Shannon entropy function is shown by $\mathsf{H}(\cdot)$, the differential entropy by $\mathsf{H}(\cdot)$, the mutual information by $\mathsf{I}(\cdot;\cdot)$ and the Kullback-Leibler divergence by $\mathsf{D}(\cdot||\cdot)$. For the sake of simplicity in expressing our results, we extend the notion of random vectors to 0-dimensional random vector $\mathbf{X}$ or *null*, and with an abuse of notation, we assume that in such case $\mathsf{H}(\mathbf{X}) = 0$.

• *Set theory*: Set subtraction is shown as $\mathcal{A} \setminus \mathcal{B} = \mathcal{A} \cap \mathcal{B}^c$. For an affine set $\mathcal{A}$, $\dim(\mathcal{A})$ stands for its Euclidean dimension. The sets $\{i, i+1, \dots, j\} \subseteq \mathbb{N}$ and $\{1, \dots, j\}$ are abbreviated as $[i:j]$ and $[j]$, respectively.

• *Vectors and matrices:* Given an $m \times n$ matrix A, we denote the i-th column of A by $A^{[i]}$, for $i \in [n]$. In addition, for a binary vector $\mathbf{s} \in \{0,1\}^n$, $A^{[\mathbf{s}]}$ denotes the sub-matrix of A formed by columns $A^{[i]}$ of A for $i \in [n]$, where $s_i = 1$. The rank of a matrix A are represented by $\mathrm{rank}(A)$. The conjugate transpose of the matrix A is indicated as $A^\dagger$. For $\mathbf{v}^n \in \{0,1\}^n$, $\overline{\mathbf{v}}^n$ is the vector obtained by obtaining logical negation of all the elements in $\mathbf{v}$. The n-dimensional column vector of all zeros/ones is indicated as $\mathbf{0}_n/\mathbf{1}_n$. Similarly, the $n \times m$ all zeros/ones matrix is indicated as $\mathbf{0}_{n\times m}/\mathbf{1}_{n\times m}$.

• *Other notations:* For $\alpha \in [0,1]$, define $\overline{\alpha} = 1 - \alpha$. The uniform quantization of the RV $\mathbf{X}^n$ with precision $1/m$ is defined as

$$[\mathbf{X}^n]_m \triangleq \left[\frac{\lfloor mX_1 \rfloor}{m}, \dots, \frac{\lfloor mX_n \rfloor}{m}\right], \tag{28}$$

with $[\mathbf{X}^n]_m \in (\mathbb{N}/m)^n$ and where $\lfloor x \rfloor$ is the floor of x.

APPENDIX B
RATIONAL FUNCTIONS AND HANKEL MATRICES

Definition 11: The function $R : \mathbb{C} \to \mathbb{C}$ is rational, if there exists $h, g \in \mathbb{C}[x]$ (the set of all polynomials with complex-valued coefficients) with $g \neq 0$ such that

$$R(z) = \frac{h(z)}{g(z)}. \tag{29}$$

The function is further called "proper" if $\deg(g) \geq \deg(h)$. The roots of g and h are commonly referred to as the poles and zeros of R with their multiplicities. If p is simultaneously a root of both g and h with multiplicities k_g and k_h where $k_h \geq k_g$, then p is called a removable pole.

Definition 12 (Zeros and Poles): We say $+\infty$ (or $-\infty$) is a zero of $R(\cdot)$ if $\lim_{z\to+\infty} g(z) = 0$ (or $\lim_{z\to-\infty} h(z) = 0$). A similar statement holds for the poles at infinity.

Remark 3: A rational function is proper, if and only if it has no $\pm\infty$ poles.

Remark 4: The set of rational functions are closed under multiplication and linear combination.

Lemma 1 ([43] Chapter V, Theorem 8 & Corollary in p.245): Let $H(z)$ be a proper rational function with p non-removable poles (including multiplicities). If $h[n]$ is such that $\sum_{n=0}^{\infty} h[n]z^{-n} = H(z)$ (i.e., $h[n]$s are the Laurent series of $H(z)$ around $z = 0$, or the causal inverse z-transform of $H(z)$), then, the Hankel matrix $A = \big[h[i+j]\big]_{i,j=0}^{p-1}$ has non-zero determinant.

APPENDIX C
STATIONARY ARMA PROCESSES

For any $n \geq p+1$, the process in (9) can be expressed through the vector equality

$$\Phi \cdot \mathbf{X}^n = \Theta \cdot \boldsymbol{\xi}_{p-q+1}^n, \tag{30}$$

where $\Phi \in \mathbb{R}^{(n-p)\times n}$ and $\Theta \in \mathbb{R}^{n-p\times(n+q-p)}$ are Toeplitz matrices defined as

$$\Phi = \begin{bmatrix} \phi_p & \phi_{p-1} & \dots & \phi_1 & 1 & \dots & 0 \\ \dots & \ddots & \ddots & \ddots & \ddots & \ddots & \dots \\ \dots & 0 & \phi_p & \phi_{p-1} & \dots & \phi_1 & 1 \end{bmatrix}, \tag{31}$$

and

$$\Theta = \begin{bmatrix} \theta_q & \theta_{q-1} & \dots & \theta_1 & 1 & \dots & 0 \\ \dots & \ddots & \ddots & \ddots & \ddots & \ddots & \dots \\ \dots & 0 & \theta_q & \theta_{q-1} & \dots & \theta_1 & 1 \end{bmatrix}. \tag{32}$$

An alternative representation of the ARMA process in Definition 8 through the $\mathcal{Z}$-transform as the filtered version of the excitation process, through the filter

$$H(z) = \frac{1 + \sum_{i\in[q]} \theta_i z^{-i}}{1 + \sum_{i\in[p]} \phi_i z^{-i}}, \tag{33}$$

where $H(z)$ stands for the $\mathcal{Z}$-transform of the filter's impulse response.

One can also rewrite this rational function in the canonical form as

$$H(z) = \frac{\Pi_{i=1}^{n_n}(r_i z^{-1} - 1)^{s_i}}{\Pi_{i=1}^{n_d}(a_i z^{-1} - 1)^{p_i}}, \tag{34}$$

where the pairs a_i/r_i describe the filter poles/zeros while p_i/s_i are their multiplicity.

Remark 5: It is shown in [44] that an ARMA(p,q) process with the excitation noise $\{\xi_i\}$ is stationary if and only if one of the below conditions hold:

1) all poles of $H(z)$ on $\mathbb{C}$ is removable (see Appendix B for the definition of a removable pole),
2) all poles of $H(z)$ on the unit circle $|z|=1$ are removable, and

$$\mathbb{E}\bigg[\max\Big\{0,\log\big(|\xi_1|\big)\Big\}\bigg]<\infty. \tag{35}$$

Appendix D
Continuity of ARMA samples

We first state the following lemma due to the clarity of our proofs.

Lemma 2: For an absolutely continuous RV $\mathbf{X}^n$ and an arbitrary RV $\mathbf{Y}^n$, the sum $\mathbf{Z}^n=\mathbf{X}^n+\mathbf{Y}^n$ is an absolutely continuous RV.

Proof: See [45, Theorem 2 and 4, Chapter V.4]. ∎

Corollary 2: Let $\mathbf{X}_1$ and $\mathbf{X}_2$ be independent n-dimensional RVs with continuity chances (defined in Theorem 1) α_1, α_2, respectively. The continuity chance β of $\mathbf{Z}=\mathbf{X}_1+\mathbf{X}_2$ satisfies

$$\beta\geq 1-(1-\alpha_1)(1-\alpha_2)\geq\max\{\alpha_1,\alpha_2\}. \tag{36}$$

Proof: According to Lemma 2, we have that

$$\begin{aligned}\mathbb{P}\big(\mathbf{Z}\in{}^{\text{continuous}}_{\text{part}}\big)&\geq\mathbb{P}\Big(\big(\mathbf{X}_1\in{}^{\text{continuous}}_{\text{part}}\big)\text{ or }\big(\mathbf{X}_2\in{}^{\text{continuous}}_{\text{part}}\big)\Big)\\&\geq 1-\underbrace{\mathbb{P}\big(\mathbf{X}_1\notin{}^{\text{continuous}}_{\text{part}}\big)}_{1-\alpha_1}\underbrace{\mathbb{P}\big(\mathbf{X}_2\notin{}^{\text{continuous}}_{\text{part}}\big)}_{1-\alpha_2}\\&=1-(1-\alpha_1)(1-\alpha_2)\geq\max\{\alpha_1,\alpha_2\},\end{aligned} \tag{37}$$

where we used the independence of $\mathbf{X}_1$ and $\mathbf{X}_2$ to obtain the product of the probabilities. ∎

Lemma 3: Let $\mathbf{X}^{k_1}$ and $\mathbf{Z}^{k_2}$ be absolutely continuous random vectors and $\mathbf{Y}^{k_1}$ be a given random vector, and not necessarily absolutely continuous. Further, let V be a random variable over $\mathbb{N}$ such that given $V=i$, $\mathbf{X}^{k_1}$ is independent of $\mathbf{Z}^{k_2}$ and $\mathbf{Y}^{k_1}$. Then, for $\mathcal{S}\subseteq\mathbb{R}^{k_1+k_2}$, the probability measure

$$\mu(\mathcal{S})=\mathbb{P}\big([\mathbf{X}^{k_1}+\mathbf{Y}^{k_1};\mathbf{Z}^{k_2}]\in\mathcal{S}\,\big|\,V=i\big)$$

is absolutely continuous for every $i\in\mathbb{N}$.

Remark 6: If $\mathbf{X}^{k_1}$ is independent of $\mathbf{Y}^{k_1}$, $\mathbf{Z}^{k_2}$ and $V=0$ with probability 1, Lemma 3 implies that $[\mathbf{X}^{k_1}+\mathbf{Y}^{k_1};\mathbf{Z}^{k_2}]$, and therefore $\mathbf{X}^{k_1}+\mathbf{Y}^{k_1}$, are absolutely continuous. This shows that Lemma 2 is a special case of Lemma 3. Another special case is when $\mathbf{X}^{k_1}$, $\mathbf{Z}^{k_2}$ are independent, $\mathbf{Y}^{k_1}\equiv\mathbf{0}$ and $V=0$ with probability 1; in this case, Lemma 3 shows that concatenation of two independent absolutely continuous random vectors is again absolutely continuous.

Proof: We prove this lemma in two steps: (i) we show that the joint probability on a $(k_1\times k_2)$-dimensional box (cartesian product of closed intervals) can be rewritten in an integration form of a function $g(\mathbf{a},\mathbf{b}):\mathbb{R}^{k_1}\times\mathbb{R}^{k_2}\to\mathbb{R}$, and (ii) we show that for any zero Lebesgue-measure set $\mathcal{S}$, the probability $\mathbb{P}\big([\mathbf{X}^{k_1}+\mathbf{Y}^{k_1};\mathbf{Z}^{k_2}]\in\mathcal{S}\big|V=i\big)$ is zero.

• **Step (i)**: Since for two sets $\mathcal{R}^{k_1}$ and $\mathcal{R}^{k_2}$ in $\mathbb{R}^{k_1}$ and $\mathbb{R}^{k_2}$, we have

$$\begin{aligned}&\mathbb{P}([\mathbf{X}^{k_1}+\mathbf{Y}^{k_1};\mathbf{Z}^{k_2}]\in\mathcal{R}^{k_1}\times\mathcal{R}^{k_2}|V=i)\\&\qquad\leq\mathbb{P}(\mathbf{Z}^{k_2}\in\mathcal{R}^{k_2}|V=i),\end{aligned} \tag{38}$$

we can see that the LHS in (38) is absolutely continuous w.r.t. the RHS (See [45, Chapter V, 10.3]). Hence, using Radon-Nikodym theorem, we can rewrite the LHS as

$$\begin{aligned}&\mathbb{P}([\mathbf{X}^{k_1}+\mathbf{Y}^{k_1};\mathbf{Z}^{k_2}]\in\mathcal{R}^{k_1}\times\mathcal{R}^{k_2}|V=i)\\&\qquad=\int_{\mathbf{z}\in\mathcal{R}^{k_2}}g(\mathcal{R}^{k_1},\mathbf{z})\mathbb{P}(\mathrm{d}\mathbf{z}|V=i),\end{aligned} \tag{39}$$

where $g(\mathcal{R}^{k_1},\mathbf{z})$ is defined as

$$g(\mathcal{R}^{k_1},\mathbf{z})=\lim_{h\to 0}\mathbb{P}(\mathbf{X}^{k_1}+\mathbf{Y}^{k_1}\in\mathcal{R}^{k_1}|\mathbf{Z}^{k_2}\in\mathcal{I}_h^{\mathbf{z}},V=i), \tag{40}$$

in which $\mathcal{I}_h^{\mathbf{z}}=[z_1,z_1+h]\times\ldots\times[z_{k_2},z_{k_2}+h]$. This probability measure always exists (See [45, Chapter V, 9.4]).

Next, we rewrite the RHS of (40) as follows

$$\begin{aligned}&\mathbb{P}(\mathbf{X}^{k_1}+\mathbf{Y}^{k_1}\in\mathcal{R}^{k_1}|\mathbf{Z}^{k_2}\in\mathcal{I}_h^{\mathbf{z}},V=i)\\&=\int_{\mathbf{y}\in\mathbb{R}^{k_1}}\mathbb{P}(\mathbf{X}^{k_1}\in\mathcal{R}^{k_1}-\mathbf{y}|\mathbf{Z}^{k_2}\in\mathcal{I}_h^{\mathbf{z}},V=i)\\&\qquad\times\mathbb{P}(\mathbf{Y}^{k_1}\in\mathrm{d}\mathbf{y}|\mathbf{Z}^{k_2}\in\mathcal{I}_h^{\mathbf{z}},V=i)\\&\overset{(a)}{=}\int_{\mathbf{y}\in\mathbb{R}^{k_1}}\mathbb{P}(\mathbf{X}^{k_1}\in\mathcal{R}^{k_1}-\mathbf{y}|V=i)\\&\qquad\times\mathbb{P}(\mathbf{Y}^{k_1}\in\mathrm{d}\mathbf{y}|\mathbf{Z}^{k_2}\in\mathcal{I}_h^{\mathbf{z}},V=i),\end{aligned} \tag{41}$$

where (a) is due to conditional independence of $\mathbf{X}^{k_1}$ and $\mathbf{Z}^{k_2}$. Let $m(\mathcal{A})$ stand for the Lebesgue measure of the Borel set $\mathcal{A}$. If $\mathbb{P}(X\in\cdot|V=i)$ is an absolutely continuous measure, [46, Prop. 15.5] implies that for every $\epsilon>0$ there exists $\delta_\epsilon>0$ such that $\mathbb{P}(X\in\mathcal{A}|V=i)\leq\epsilon$ for every Borel set $\mathcal{A}\subset\mathbb{R}^{k_1}$ with $m(\mathcal{A})<\delta_\epsilon$. Further, using [47, Theorem 12.1] we know that $m(\mathcal{A})=m(\mathcal{A}-\mathbf{y})$ for every $\mathbf{y}\in\mathbb{R}^{k_1}$. Hence, for every choice of $\mathbf{y}$ we have $\mathbb{P}(X\in\mathcal{A}-\mathbf{y}|V=i)\leq\epsilon$. Thus, using (41) we have that

$$\begin{aligned}&\mathbb{P}(\mathbf{X}^{k_1}+\mathbf{Y}^{k_1}\in\mathcal{A}|\mathbf{Z}^{k_2}\in\mathcal{I}_h^{\mathbf{z}},V=i)\\&\leq\epsilon\int_{\mathbf{y}\in\mathbb{R}^{k_1}}\mathbb{P}(\mathbf{Y}^{k_1}\in\mathrm{d}\mathbf{y}|\mathbf{Z}^{k_2}\in\mathcal{I}_h^{\mathbf{z}},V=i)\leq\epsilon.\end{aligned} \tag{42}$$

This means that for every $\epsilon>0$, there exists $\delta_\epsilon>0$ such that if $m(\mathcal{A})<\delta_\epsilon$ we can bound $g(\mathcal{A},\mathbf{z})$ as

$$g(\mathcal{A},\mathbf{z})=\lim_{h\to 0}\mathbb{P}(\mathbf{X}^{k_1}+\mathbf{Y}^{k_1}\in\mathcal{A}|\mathbf{Z}^{k_2}\in\mathcal{I}_h^{\mathbf{z}},V=i)\leq\epsilon. \tag{43}$$

Recall [46, Prop. 15.5] , (43) shows that $g(\mathcal{A},\mathbf{z})$ is an absolutely continuous measure.

Now, the Radon-Nikdoym theorem in conjunction with (39) reveals the existence of functions $q(\mathbf{u},\mathbf{z})$ and $q'(\mathbf{z})$ such that

$$\begin{aligned}&\mathbb{P}([\mathbf{X}^{k_1}+\mathbf{Y}^{k_1};\mathbf{Z}^{k_2}]\in\mathcal{R}^{k_1}\times\mathcal{R}^{k_2}|V=i)\\&\qquad\overset{(a)}{=}\int_{\mathbf{z}\in\mathcal{R}^{k_2}}\int_{\mathbf{u}\in\mathcal{R}^{k_1}}q(\mathbf{u},\mathbf{z})\mathrm{d}\mathbf{u}\mathbb{P}(\mathrm{d}\mathbf{z}|V=i)\end{aligned}$$

$$\overset{(b)}{=} \int_{\mathbf{z}\in\mathcal{R}^{k_2}} \int_{\mathbf{u}\in\mathcal{R}^{k_1}} q(\mathbf{u},\mathbf{z})q'(\mathbf{z})\mathrm{d}\mathbf{u}\mathrm{d}\mathbf{z},$$

where (a) is because of the absolute continuity of $g(\cdot,\mathbf{z})$ and (b) holds due to the absolute continuity of $\mathbb{P}(\mathbf{Z}^{k_2}\in\cdot|V=i)$.

• **Step (ii)**: Let $\{\mathcal{R}_k\}_{k=1}^{\infty}$ be an arbitrary countable set of (possibly intersecting) boxes. If we define $\mathcal{F}_i=\mathcal{R}_i\backslash\bigcup_{k=1}^{i-1}\mathcal{R}_i$, we know that $\mathcal{F}_i$s are disjoint:

$$m\big(\cup_i\mathcal{R}_i\big)=m\big(\cup_i\mathcal{F}_i\big)=\sum_i m(\mathcal{F}_i). \tag{44}$$

Besides, each $\mathcal{F}_i$ can be decomposed into finitely-many almost-disjoint boxes $\{\widetilde{\mathcal{R}}_{i,j}\}_j$:

$$\exists\{\widetilde{\mathcal{R}}_{i,j}\}_{j=1}^{n_i}:\ \begin{cases} m\big(\widetilde{\mathcal{R}}_{i,j_1}\cap\widetilde{\mathcal{R}}_{i,j_2}\big)=0, & j_1\neq j_2,\\ \mathcal{F}_i\subseteq\cup_{j=1}^{n_i}\widetilde{\mathcal{R}}_{i,j}, & \\ m\Big(\big(\cup_{j=1}^{n_i}\widetilde{\mathcal{R}}_{i,j}\big)\setminus\mathcal{F}_i\Big)=0. & \end{cases} \tag{45}$$

Note that $\widetilde{\mathcal{R}}_{i,j}$s are closed boxes and can overlap only at their boundaries, i.e., $m\big(\widetilde{\mathcal{R}}_{i,j_1}\cap\widetilde{\mathcal{R}}_{i,j_2}\big)=0$. Similarly, $\mathcal{F}_i$ and $\cup_{j=1}^{n_i}\widetilde{\mathcal{R}}_{i,j}$ might differ only at parts of the borders of some of $\widetilde{\mathcal{R}}_{i,j}$s. Overall, we have that

$$\big(\cup_i\mathcal{R}_i\big)=\big(\cup_i\mathcal{F}_i\big)\subseteq\big(\cup_{i,j}\widetilde{\mathcal{R}}_{i,j}\big), \tag{46}$$

and

$$m\Big(\big(\cup_{i,j}\widetilde{\mathcal{R}}_{i,j}\big)\setminus\big(\cup_i\mathcal{R}_i\big)\Big)=0. \tag{47}$$

In simple words, we showed that the union of an arbitrary countable set of boxes can be effectively decomposed into a union of countable set of almost-disjoint boxes.

As a result, for every set of boxes $\{\mathcal{R}_i\}_{i=1}^{\infty}$ we rewrite

$$\begin{aligned}
&\mathbb{P}([\mathbf{X}^{k_1}+\mathbf{Y}^{k_1};\mathbf{Z}^{k_2}]\in\cup_{k=1}^{\infty}\mathcal{R}_k|V=i)\\
&\overset{(a)}{\leq}\mathbb{P}([\mathbf{X}^{k_1}+\mathbf{Y}^{k_1};\mathbf{Z}^{k_2}]\in\cup_{i,j}\widetilde{\mathcal{R}}_{i,j}|V=i)\\
&\overset{(b)}{=}\sum_{i,j}\int_{(\mathbf{u},\mathbf{z})\in\widetilde{\mathcal{R}}_{i,j}} q(\mathbf{u},\mathbf{z})q'(\mathbf{z})\mathrm{d}\mathbf{u}\mathrm{d}\mathbf{z}\\
&\overset{(c)}{=}\int_{(\mathbf{u},\mathbf{z})\in\cup_{i,j}\widetilde{\mathcal{R}}_{i,j}} q(\mathbf{u},\mathbf{z})q'(\mathbf{z})\mathrm{d}\mathbf{u}\mathrm{d}\mathbf{z}\\
&\overset{(d)}{=}\int_{(\mathbf{u},\mathbf{z})\in\cup_{k=1}^{\infty}\mathcal{R}_k} q(\mathbf{u},\mathbf{z})q'(\mathbf{z})\mathrm{d}\mathbf{u}\mathrm{d}\mathbf{z}.
\end{aligned} \tag{48}$$

Here, (a) and (d) are due to (46) and (47), respectively. Further, (b) is followed by Step (i), and (c) is correct because $\widetilde{\mathcal{R}}_j$s are almost-disjoint sets as (45) specifies.

Now, we recall a result from [48, Prop. 2.5.8] that for every L_1-measurable function $f(\cdot)$ and constant $\epsilon>0$, there exists $\delta_\epsilon>0$ such that for every measurable set $\mathcal{A}$ with $m(\mathcal{A})<\delta_\epsilon$, we have $\int_{\mathcal{A}}|f(\mathbf{x})|\mathrm{d}\mathbf{x}<\epsilon$. If we let $f(\mathbf{u},\mathbf{z})=q(\mathbf{u},\mathbf{z})q'(\mathbf{z})$, one asserts that f is an L_1 measurable function. The reason is that $\int_{\mathbb{R}^{k_1+k_2}}f(\mathbf{u},\mathbf{z})\mathrm{d}\mathbf{u}\mathrm{d}\mathbf{z}=\mathbb{P}\big([\mathbf{X}^{k_1}+\mathbf{Y}^{k_1};\mathbf{Z}^{k_2}]\in\mathbb{R}^{k_1+k_2}|V=i\big)=1<\infty$.

Let $\mathcal{S}$ be a zero Lebesgue-measure set. We choose $\epsilon>0$ arbitrarily small; according to [48, Prop. 2.5.8] and (48), there exists δ_ϵ such that if $m\big(\cup_{k=1}^{\infty}\mathcal{R}_k\big)<\delta_\epsilon$, then, $\mathbb{P}([\mathbf{X}^{k_1}+\mathbf{Y}^{k_1};\mathbf{Z}^{k_2}]\in\cup_{k=1}^{\infty}\mathcal{R}_k|V=i)<\epsilon$. Since (see [42, pp. 385, 389])

$$\underbrace{m(\mathcal{S})}_{=0}\triangleq\inf\Big\{\sum_k m(\mathcal{O}_k):\ \mathcal{O}_k=\overset{\text{open}}{\text{box}},\ \mathcal{S}\subseteq\cup_k\mathcal{R}_k\Big\}, \tag{49}$$

we shall have a set of open boxes $\{\mathcal{O}_k\}_k$ with $\sum_k m(\mathcal{O}_k)<\delta_\epsilon$ such that $\mathcal{S}\subseteq\cup_k\mathcal{O}_k$. If we set $\mathcal{R}_k=\overline{\mathcal{O}}_k$ (the closure of $\mathcal{O}_k$), we have

$$m\big(\cup_k\mathcal{R}_k\big)\leq\sum_k m(\mathcal{R}_k)=\sum_k m(\mathcal{O}_k)<\delta_\epsilon. \tag{50}$$

In addition,

$$\mathcal{S}\subseteq\cup_k\mathcal{O}_k\subseteq\cup_k\mathcal{R}_k. \tag{51}$$

If we summarize the above results, we achieve

$$\begin{aligned}
&\mathbb{P}([\mathbf{X}^{k_1}+\mathbf{Y}^{k_1};\mathbf{Z}^{k_2}]\in\mathcal{S}|V=i)\\
&\quad\leq\mathbb{P}([\mathbf{X}^{k_1}+\mathbf{Y}^{k_1};\mathbf{Z}^{k_2}]\in\cup_{k=1}^{\infty}\mathcal{R}_k|V=i)<\epsilon.
\end{aligned} \tag{52}$$

Since the choice of $\epsilon>0$ is arbitrary, we shall have $\mathbb{P}([\mathbf{X}^{k_1}+\mathbf{Y}^{k_1};\mathbf{Z}^{k_2}]\in\mathcal{S}|V=i)=0$.

■

Lemma 4: Let $h[\cdot]$ be the causal inverse z-transform of a stationary ARMA filter $\mathsf{H}(z)$ with $p+p'$ non-removable poles (see (34) for the definition of non-removable) from which p are non-zero poles. If for $i\geq\lceil\frac{p'+1}{p}\rceil+1$ we define

$$H^{(i)}=\Big[h[ip-p-1+j+k]\Big]_{j,k=0}^{p-1}, \tag{53}$$

then, $H^{(i)}$ is full-rank.

Proof: Let us define

$$\tilde{h}_i[n]=\begin{cases} h[ip-p-1+n], & n\geq 0,\\ 0, & n<0. \end{cases} \tag{54}$$

We first show that $\tilde{h}_i[\cdot]$ is the impulse response of a causal ARMA process with p non-zero and non-removable poles:

$$\begin{aligned}
\tilde{H}_i(z)&=\sum_{n=0}^{\infty}\tilde{h}_i[n]z^{-n}=z^{ip-p-1}\sum_{n=ip-p-1}^{\infty}h[n]z^{-n}\\
&=z^{ip-p-1}\mathsf{H}(z)-\sum_{n=0}^{ip-p-2}h[n]z^{ip-p-1-n}.
\end{aligned} \tag{55}$$

Note that the last term of (55) is polynomial in terms of z. Since $i\geq\lceil\frac{p'+1}{p}\rceil+1$, we know that $ip-p-1\geq p'$; therefore, $z^{ip-p-1}\mathsf{H}(z)$ has no zero poles. This confirms that $\tilde{H}_i(z)$ has exactly p non-removable poles all of which are non-zero. Thus, we can conclude the claim by recalling Lemma 1.

We should highlight that since $\tilde{H}_i(z)=\sum_{n=0}^{\infty}\tilde{h}_i[n]z^{-n}$, we know that $\tilde{H}_i(z)$ is bounded as $z\to\infty$. Therefore, $\tilde{H}_i(z)$ is a *proper* rational function.

■

Lemma 5: Let $\{\mathbf{X}_t\}$ be a D/C-ARMA process (see Definition 10) with excitation noise $\{\boldsymbol{\xi}_t\}$ and the corresponding RID of $d(\xi_1)\neq 0$. We know that a truncation $\boldsymbol{\xi}_{\mathcal{S}}=\{\xi_i:\ i\in\mathcal{S}\}$ of the excitation noise is affinely singular with a set of absolutely continuous RVs $\{\tilde{\boldsymbol{\xi}}_i\}$ laid on affine sets and a selection variable $V_{\mathcal{S}}$. The measure $\mathbb{P}([\mathbf{X}_1^p;\tilde{\boldsymbol{\xi}}_i]\in\cdot|V_{\mathcal{S}}=i)$ is absolutely continuous for all $i\in\mathbb{Z}$.

Proof: One could write $\mathbf{X}_1^p$ as

$$\mathbf{X}_1^p = \sum_{i=0}^{\infty} \underbrace{H^{(i)} \boldsymbol{\xi}_{1-ip}^{p-ip}}_{\mathbf{Y}_i}, \tag{56}$$

where $H^{(i)} = \Big[h[ip-p-1+j+k]\Big]_{j,k=1}^{p}$, and $h[\cdot]$ is the causal inverse Z-transform of transfer function $\mathsf{H}(z)$ of the D/C-ARMA process. Next, we rewrite $\mathbf{X}_1^p$ as

$$\mathbf{X}_1^p = \underbrace{\sum_{i=0}^{i_o-1} \mathbf{Y}_i}_{\mathbf{R}} + \underbrace{\sum_{i=i_0}^{\infty} \mathbf{Y}_i}_{\mathbf{U}}, \tag{57}$$

where $i_0 = \max\big\{1 + \lfloor \frac{1-\min \mathcal{S}}{p} \rfloor, 1 + \lfloor \frac{(q-p)_+ + 1}{p} \rfloor\big\}$.

To prove the theorem, we follow three steps: (i) we prove that $\mathbf{U}$ is independent of $\mathbf{R}$ and $\tilde{\boldsymbol{\xi}}_i$ given $V_{\mathcal{S}} = i$, (ii) we show that $\mathbf{U}$ is an absolutely continuous RV, and (iii) we use Lemma 3 and the properties that are shown in the two previous steps to prove absolute continuity of $\mathbb{P}\big([\mathbf{R}+\mathbf{U}; \tilde{\boldsymbol{\xi}}_i] \in \cdot | V_{\mathcal{S}} = i\big)$.

• **Step (i)**: Firstly, using [49, Lemma 3], we know that

$$\mathsf{I}(V_{\mathcal{S}}; \boldsymbol{\xi}_{\mathcal{S}}) = \mathsf{H}(V_{\mathcal{S}}), \tag{58}$$

or equivalently,

$$\mathsf{H}(V_{\mathcal{S}} | \boldsymbol{\xi}_{\mathcal{S}}) = 0. \tag{59}$$

As a result, for an arbitrary random variable T, and all sets $\mathcal{S}'$ that contain $\mathcal{S}$ we have

$$\mathsf{I}(V_{\mathcal{S}}; T | \boldsymbol{\xi}_{\mathcal{S}'}) \leq \mathsf{H}(V_{\mathcal{S}} | \boldsymbol{\xi}_{\mathcal{S}'}) \leq \mathsf{H}(V_{\mathcal{S}} | \boldsymbol{\xi}_{\mathcal{S}}) = 0. \tag{60}$$

As a result, by assuming $\mathcal{S}''$ being mutually exclusive with $\mathcal{S}'$, we have

$$\begin{aligned} \mathsf{I}(\boldsymbol{\xi}_{\mathcal{S}''}; \boldsymbol{\xi}_{\mathcal{S}'} | V_{\mathcal{S}}) &\leq \mathsf{I}(\boldsymbol{\xi}_{\mathcal{S}''}; \boldsymbol{\xi}_{\mathcal{S}'}, V_{\mathcal{S}}) && (61) \\ &= \mathsf{I}(\boldsymbol{\xi}_{\mathcal{S}''}; \boldsymbol{\xi}_{\mathcal{S}'}) + \mathsf{I}(\boldsymbol{\xi}_{\mathcal{S}''}; V_{\mathcal{S}} | \boldsymbol{\xi}_{\mathcal{S}'}) = 0, && (62) \end{aligned}$$

where the last equality holds because of (60), and mutually exclusiveness $\mathcal{S}'$ and $\mathcal{S}''$, and that $\{\tilde{\boldsymbol{\xi}}_t\}$ are drawn independently.

Moreover, because of chain rule we have

$$\mathsf{I}(\boldsymbol{\xi}_{\mathcal{S}''}; \boldsymbol{\xi}_{\mathcal{S}'} | V_{\mathcal{S}}) = \sum_i \mathbb{P}(V_{\mathcal{S}} = i) \mathsf{I}(\boldsymbol{\xi}_{\mathcal{S}''}; \boldsymbol{\xi}_{\mathcal{S}'} | V_{\mathcal{S}} = i) = 0, \tag{63}$$

and since all the terms in RHS are non-negative, then for all i such that $\mathbb{P}(V_{\mathcal{S}} = i)$ is non-zero, we have

$$\mathsf{I}(\boldsymbol{\xi}_{\mathcal{S}''}; \boldsymbol{\xi}_{\mathcal{S}'} | V_{\mathcal{S}} = i) = 0. \tag{64}$$

Moreover, we know that given $V_{\mathcal{S}} = i$, $\tilde{\boldsymbol{\xi}}_i$ is merely a projection of $\boldsymbol{\xi}_{\mathcal{S}}$ on an e_i-dimensional space where $e_i \leq |\mathcal{S}|$. As a result, given $V_{\mathcal{S}} = i$, $\tilde{\boldsymbol{\xi}}_i$ is a function of $\boldsymbol{\xi}_{\mathcal{S}}$. Hence, using data processing inequality we have

$$\begin{aligned} 0 \leq \mathsf{I}(U; R, \tilde{\boldsymbol{\xi}}_i | V_{\mathcal{S}} = i) &\leq \mathsf{I}(U; R, \boldsymbol{\xi}_{\mathcal{S}} | V_{\mathcal{S}} = i) \\ &\overset{(a)}{\leq} \mathsf{I}(\boldsymbol{\xi}_{-\infty}^{p-i_0 p}; R, \boldsymbol{\xi}_{\mathcal{S}} | V_{\mathcal{S}} = i) \\ &\overset{(b)}{\leq} \mathsf{I}(\boldsymbol{\xi}_{-\infty}^{p-i_0 p}; \boldsymbol{\xi}_{p-i_0 p+1}^{p}, \boldsymbol{\xi}_{\mathcal{S}} | V_{\mathcal{S}} = i) \\ &\overset{(c)}{\leq} \mathsf{I}(\boldsymbol{\xi}_{-\infty}^{p-i_0 p}; \boldsymbol{\xi}_{p-i_0 p+1}^{\max|\mathcal{S}|} | V_{\mathcal{S}} = i) \\ &\overset{(d)}{=} 0, \end{aligned}$$

where (a) holds because U is a function of $\boldsymbol{\xi}_{-\infty}^{p-i_0 p}$ and because of data processing inequality, and (b) holds because R is a function of $\boldsymbol{\xi}_{p-i_0 p+1}^{p}$ and as a result of data processing inequality. Moreover, (c) is followed by $i_0 \geq 1 + \lfloor \frac{1-\min \mathcal{S}}{p} \rfloor$ and as a result $\min \mathcal{S} \geq p - i_0 p + 1$. Finally (d) is followed by (64). This proves that U and R are independent given $V_{\mathcal{S}} = i$.

• **Step (ii)**: Firstly, since $d(\xi_1) = d > 0$ and as a result of Lemma 5, the continuity chance of $\boldsymbol{\xi}_{1-ip}^{p-ip}$ is d^p. On the other hand, using Lemma 4, $H^{(i)}$ is a $p \times p$ full-rank matrix for $i \geq 3$. Further, it is shown in [12, Lemma 6] that the any full column-rank linear transformation of absolutely continuous random vectors is absolutely continuous. Hence, the continuity chance of $\mathbf{Y}_i$ is at least d^p for every $i \geq 3$. Next, using the independence of $\mathbf{Y}_i$s, and Corollary 2, the continuity chance of $\mathbf{U}_t := \sum_{i=i_0}^{t} \mathbf{Y}_i$ and also $\mathbf{U}$ is lower bounded by $1 - (1-d^p)^{t-i_0}$. Finally, by arbitrariness of t, one can prove this claim. • **Step (iii)**: We know by definition that $\tilde{\boldsymbol{\xi}}_i$ is an absolutely continuous random vector, and the absolute continuity of $\mathbf{U}$ is a result of Step (ii). Therefore, by making use of independence of $\mathbf{U}$ and $\mathbf{R}, \tilde{\boldsymbol{\xi}}_i$ given $V_{\mathcal{S}} = i$, and using Lemma 3 we conclude that $\mathbb{P}\big([\mathbf{R}+\mathbf{U}; \tilde{\boldsymbol{\xi}}_i] \in \cdot | V_{\mathcal{S}} = i\big)$ is an absolutely continuous measure that completes the proof. ■

Appendix E
Proof of Theorem 6

Proof: We prove this theorem using the following steps: (i) we show that a truncation ξ_{p-q+1}^n of the excitation noise has an affinely singular probability distribution, (ii) using step (i) we show that the joint random vector $[\mathbf{X}^p; \xi_{p-q+1}^n]$ of a truncation of the ARMA process and its excitation noise is affinely singular, (iii) using the previous steps and linear recursive relation in ARMA processes we show that each truncation $\mathbf{X}^n$ of the ARMA process has affinely singular distribution, (iv) we show that if the excitation noise is absolutely continuous, then the truncation $\mathbf{X}^n$ is absolutely continuous, and (v) we use Lemma [38, Lemma 7] and Hankel property of the linear recursion of the ARMA process to show that the singularity dimensions of the truncation $\mathbf{X}^n$ concentrates around $n\mathrm{d}(\xi_i)$ for large n.

• **Step (i)**: We first should note that a truncation of discrete-continuous excitation noise $\{\xi_t\}$ of the DCE-ARMA process has *orthogonally singular* probability measure that is defined in [38, Definition 2], which is known to be an affinely singular measure with singular components along the axes of the Euclidean space. To formalize this notion, assume that ν_t is a Bernoulli random variable that denotes whether ξ_t takes its discrete ($\nu_t = 0$) or continuous values ($\nu_t = 1$). Using this choice of notation, [38, Lemma 3] shows that the truncation ξ_{p-q+1}^n of the excitation noise has a probability distribution equivalent to

$$\boldsymbol{\xi}_{p-q+1}^n \overset{d}{=} U_V[\tilde{\boldsymbol{\xi}}_V; 0_{d_O}] + \mathbf{b}_V, \tag{65}$$

where V denotes the singular component on which the drawn instance of the random vector $\boldsymbol{\xi}_{p-q+1}^n$ is laid. Further, U_V denotes a $(n+q-p)\times(n+q-p)$ permutation matrix that maps the first d_V components of a vector to the indices of the components of $\boldsymbol{\xi}_{p-q+1}^n$ that take continuous values, i.e., $U_V = \big[I_{n+q-p}^{[\boldsymbol{\nu}]}, I_{n+q-p}^{[\overline{\boldsymbol{\nu}}]}\big]$, where $\boldsymbol{\nu}$ is, with an abuse of notation, defined as

$$\boldsymbol{\nu} := \boldsymbol{\nu}_{p-q+1}^n. \tag{66}$$

Moreover, $\tilde{\xi}_V$ quantifies the distribution of the random vector $\boldsymbol{\xi}_{p-q+1}^n$ on its singular component with index V, d_V is the Euclidean dimension of that singular component, and d_O is the difference between the Euclidean dimension of that component and the encompassing space $d_O = n+q-p-d_V$. Furthermore, $\mathbf{b}_V$ is a fixed vector that determines the bias of the singular component from the origin.

• **Step (ii):** If we concatenate the truncation $\mathbf{X}^p$ of the process and the truncation $\boldsymbol{\xi}_{p-q+1}^n$ of the excitation noise, then by taking the first d_V columns of U_V and the p-dimensional identity matrix I_p we can generate the tall matrix

$$\tilde{U}_V = \left[\begin{array}{c|c} I_p & 0_{p\times d_V} \\ \hline 0_{(n+q-p)\times p} & I_{n+q-p}^{[\boldsymbol{\nu}]} \end{array}\right], \tag{67}$$

using which the concatenation can be rewritten as

$$[\mathbf{X}^p; \boldsymbol{\xi}_{p-q+1}^n] = \tilde{U}_V[\mathbf{X}^p; \tilde{\boldsymbol{\xi}}_V] + \mathbf{b}_V. \tag{68}$$

Furthermore, Lemma 5 shows that the concatenation $\mathbf{Z}^{p+d_V} := [\mathbf{X}^p; \tilde{\boldsymbol{\xi}}_V]$ in RHS of the above identity is absolutely continuous for each choice of $V = i$. Therefore, if we complete the matrix $\tilde{U}_V$ by adding extra orthogonal columns, we can show that $[\mathbf{X}^p; \boldsymbol{\xi}_{p-q+1}^n]$ takes the form of an affinely singular random vector.

• **Step (iii)**: In this step we use the concatenation $[\mathbf{X}^p; \boldsymbol{\xi}_{p-q+1}^n]$ that is shown to be affinely singular in previous step to specify the distribution of each truncation $\mathbf{X}^n$ of the ARMA process. To that end, we refer the reader to the proof of Theorem 5, and particularly (92) where we show that $\mathbf{X}^n$ for $n \geq p$ can be written in form of

$$\mathbf{X}^n = \widehat{\Phi}^{-1}\widehat{\Theta}\big[\mathbf{X}^p; \xi_{p-q+1}^n\big], \tag{69}$$

where $\widehat{\Phi}$ and $\widehat{\Theta}$ are defined in (91) and (93), respectively. Using (68), the above identity can be rewritten as

$$\mathbf{X}^n = \widehat{\Phi}^{-1}\widehat{\Theta}\tilde{U}_V[\mathbf{X}^p; \tilde{\boldsymbol{\xi}}_V] + \mathbf{b}'_V, \tag{70}$$

where $\mathbf{b}'_V = \widehat{\Phi}^{-1}\widehat{\Theta}$ is a fixed vector for each choice of V. Therefore, in order to show that $\mathbf{X}^n$ has an affinely singular distribution, we need to rewrite the first term of RHS of (70) as a rotation of a low-dimensional absolutely continuous random vector. To that end, we first rewrite the matrix $\widehat{\Phi}^{-1}\widehat{\Theta}\tilde{U}_V$ using the method of singular value decomposition (SVD) as $\widehat{\Phi}^{-1}\widehat{\Theta}\tilde{U}_V = Q_V\Sigma_V P_V^\dagger$, where Q_V and P_V are unitary matrices and Σ_V is a diagonal matrix with $\mathrm{rank}(\widehat{\Phi}^{-1}\widehat{\Theta}\tilde{U}_V)$ non-zero terms. Next, we use [38, Lemma 10] that shows that the result of a multiplication of a full-row rank matrix and an absolutely continuous RV is another absolutely continuous RV. In fact, $\Sigma_V P_V^\dagger$ is a concatenation of a full-row rank matrix and null rows. Therefore, using absolute continuity of $\big[\mathbf{X}^p; \xi_{p-q+1}^n\big]$ in Step (ii), we conclude that $\Sigma_V P_V^\dagger\big[\mathbf{X}^p; \xi_{p-q+1}^n\big]$ is a concatenation of an absolutely continuous RV with dimension $\mathrm{rank}(\widehat{\Phi}^{-1}\widehat{\Theta}\tilde{U}_V)$ and zero components. This, together with the SVD form shows that $\mathbf{X}^n$ is an affinely singular RV in which the singular components have dimension $d_V = \mathrm{rank}(\widehat{\Phi}^{-1}\widehat{\Theta}\tilde{U}_V)$.

By definition of $\widehat{\Phi}$ in (91), we know that it is a full-rank square matrix. Therefore, we have

$$\mathrm{rank}(\widehat{\Phi}^{-1}\widehat{\Theta}\tilde{U}_V) = \mathrm{rank}(\widehat{\Theta}\tilde{U}_V). \tag{71}$$

Moreover, by definition of $\widehat{\theta}$ in (93), we can show that

$$\widehat{\Theta}\tilde{U}_V = \left[\begin{array}{c|c} I_p & 0_{p\times d_V} \\ \hline 0_{n-p\times p} & \Theta^{[\boldsymbol{\nu}]} \end{array}\right], \tag{72}$$

that has the rank of $p+\mathrm{rank}(\Theta^{[\nu]})$. Therefore, using the above two identities, we have

$$d_V = p + \mathrm{rank}(\Theta^{[\nu]}). \tag{73}$$

• **Step (iv)**: In case of absolute continuity of the excitation noise, and the definition of ν in Step (i) assures that ν is an all-one vector. Therefore, d_V in (73) can be obtained as

$$d_V = p + \mathrm{rank}(\Theta), \tag{74}$$

and since Θ is a full-row rank matrix with $n-p$ rows (see the definition of Θ in (32)), therefore we have

$$d_V = n, \tag{75}$$

that is equivalent to say $\mathbf{X}^n$ is a mixture of absolutely continuous RVs, and therefore itself is an absolutely continuous RV.

• **Step (v)**: In case that the excitation noise is discrete-continuous with continuity chance of α, then the random vector ν is distributed as $n-p+q-1$ i.i.d Bernoulli random variables with each components having $\mathbb{P}(\nu_i = 1) = \alpha$. As a result, the matrix $\Theta^{[\nu]}$ is a random choice of columns of a Hankel matrix with probability α. For such random choice, [38, Lemma 8] shows that the rank function is distributed as

$$\begin{aligned}&\mathbb{P}\big(\mathrm{rank}(\Theta^{[\boldsymbol{\nu}]}) > k\big) \\ &\qquad \geq 1 - \exp\Big(-(n+q-p)D\big(\tfrac{k+q}{n+q-p}\|\alpha\big)\Big),\end{aligned} \tag{76}$$

for $\frac{k+q}{n+q-p} < \alpha$ and

$$\begin{aligned}&\mathbb{P}\big(\mathrm{rank}(\Theta^{[\boldsymbol{\nu}]}) < k\big) \\ &\qquad \geq 1 - \exp\Big(-(n+q-p)D\big(\tfrac{k}{n-p}\|\alpha\big)\Big),\end{aligned} \tag{77}$$

for $\frac{k}{m} > \alpha$.

Inequalities (76) and (77) together with (73) conclude in (16), (17), (18), and accordingly complete the proof. ■

Appendix F
Proof of Theorem 5

Before delving into the details of the proof, we introduce the following lemma to bound the entropy of quantized shifted

version of a RV.

Lemma 6: For a RV X, if we have $\mathsf{H}\big([X]_1\big) < m$, then, we can bound $\mathsf{H}\big([X+\epsilon]_1\big)$ as

$$\mathsf{H}\big([X+\epsilon]_1\big) \leq 4m + \mathrm{C}, \tag{78}$$

for every $\epsilon \in (-1,1)$, where C is a fixed value that dependent on the distribution of X, and not ϵ.

Proof: We define $Y = X + \epsilon$, and study the entropy

$$\mathsf{H}\big([Y]_1\big) = -\sum_{i=-\infty}^{\infty} p_i^Y \log p_i^Y, \tag{79}$$

where p_i^Y is defined as

$$p_i^Y = \mathbb{P}(i \leq Y < i+1). \tag{80}$$

Let $c \in \mathbb{N}$ be a large enough integer such that $\mathbb{P}(|X| > c) < 1/4$. We now decompose $\mathsf{H}\big([Y]_1\big)$ as

$$\begin{aligned}\mathsf{H}\big([Y]_1\big) = &-\sum_{i\in[-c-1:c]} p_i^Y \log p_i^Y \\ &-\sum_{i\geq c+1, i\leq -c-2} p_i^Y \log p_i^Y.\end{aligned} \tag{81}$$

By defining

$$\begin{aligned}A = \sum_{i\in[-c-1:c]} p_i^Y &= \mathbb{P}(-c-1 \leq Y < c+1) \\ &\geq \mathbb{P}(|X| < c) \geq \tfrac{3}{4},\end{aligned} \tag{82}$$

we can write

$$-\sum_{i\in[-c-1:c]} p_i^Y \log p_i^Y = -A\sum_{i\in[-c-1:c]} \underbrace{\frac{p_i^Y}{A}}_{\widetilde{p}_i} \log \frac{p_i^Y}{A} A \tag{83}$$

$$= -A\sum_{i\in[-c-1:c]} \widetilde{p}_i \log \widetilde{p}_i - A\log A \tag{84}$$

$$\overset{(a)}{\leq} A(-\sum_{i\in[-c-1:c]} \widetilde{p}_i \log \widetilde{p}_i) + \tfrac{1}{3} \tag{85}$$

$$\overset{(b)}{\leq} \log\big(2(c+1)\big) + \tfrac{1}{3}, \tag{86}$$

where (a) is followed by $A \in [\frac{3}{4}, 1]$, and (b) is because the entropy of any discrete RV with $2(c+1)$ symbols is upper-bounded by $\log\big(2(c+1)\big)$

Let $s_\epsilon = 1$ if $\epsilon \in [0,1)$, and $s_\epsilon = -1$ if $\epsilon \in (-1,0)$. We know that for all $i \geq c+1, i \leq -c-2$

$$p_i^Y \leq p_i^X + p_{i+s_\epsilon}^X \leq \sum_{k\geq c+1, k\leq -c-2} p_k^X \leq \tfrac{1}{4}. \tag{87}$$

Since $-x\log x$ is an increasing function for $x \leq \frac{1}{2}$, we conclude that

$$\begin{aligned}-p_i^Y \log p_i^Y &\leq -(p_i^X + p_{i+s_\epsilon}^X)\log(p_i^X + p_{i+s_\epsilon}^X) \\ &\leq -2\max\{p_i^X, p_{i+s_\epsilon}^X\}\log\big(2\max\{p_i^X, p_{i+s_\epsilon}^X\}\big) \\ &\overset{(a)}{\leq} -2p_i^X \log\big(2p_i^X\big) - 2p_{i+s_\epsilon}^X \log\big(2p_{i+s_\epsilon}^X\big) \\ &\leq -2p_i^X \log p_i^X - 2p_{i+s_\epsilon}^X \log p_{i+s_\epsilon}^X - 2(p_i^X + p_{i+s_\epsilon}^X)\end{aligned} \tag{88}$$

where (a) is followed by positiveness of function $-x\log x$ for $x \leq 1$ coupled with the fact that $p_i^X, p_{i+1}^X \leq \frac{1}{4}$.

Finally, using (81), (86), and (88) we have

$$\mathsf{H}\big([Y]_1\big) \leq \log(c+1) + \frac{4}{3} + 4\mathsf{H}\big([X]_1\big) - 4 \tag{89}$$

$$\leq 4m + \underbrace{\log(c+1) - \tfrac{8}{3}}_{\mathrm{C}}. \tag{90}$$

■

Proof of Theorem 5: Consider the ARMA process at time $t > 0$ and the corresponding Φ and Θ matrices as in (31) and (32). We first modify Φ to make it invertible:

$$\widehat{\Phi} = \left[\begin{array}{c|c} I_p & 0_{p\times m} \\ \hline \Phi_{m\times(m+p)} \end{array}\right]. \tag{91}$$

We can now rewrite (30) as

$$\widehat{\Phi}\cdot\mathbf{X}^{m+p} = \widehat{\Theta}\underbrace{\left[\begin{array}{c} \mathbf{X}^p \\ \hline \boldsymbol{\xi}_{1-q+p}^{m+p} \end{array}\right]}_{\widehat{\mathbf{Q}}} = \widehat{\Theta}\widehat{\mathbf{Q}}, \tag{92}$$

where

$$\widehat{\Theta} = \left[\begin{array}{c|c} I_p & 0_{p\times(m+q)} \\ \hline 0_{m\times p} & \Theta_{m\times(m+q)} \end{array}\right]. \tag{93}$$

From the fact that $\det(\widehat{\Phi}) = 1$ and [39, Thm. 2], it follows that

$$d(\mathbf{X}_1^{m+p}) = d(\widehat{\Theta}\widehat{\mathbf{Q}}). \tag{94}$$

The remainder of the proof relies on the following steps: (i) we show that $[\mathbf{X}^p, \boldsymbol{\xi}_{1-q+p}^p]$ is independent of $\boldsymbol{\xi}_{p+1}^{p+m}$, (ii) we find lower- and upper-bounds for $d(\mathbf{X}^{m+p})$ and determine $d_B\big(\{\mathbf{X}_t\}\big)$, and finally, (iii) we prove that $d_B\big(\{\mathbf{X}_t\}\big)$ is a lower-bound for $d_I\big(\{\mathbf{X}_t\}\big)$. The latter bound together with the result in Theorem 4 completes the proof.

• **Step (i)**: Using the recurrence relationship in (9), we know that $\mathbf{X}^p$ is a function of excitation noise samples $\boldsymbol{\xi}_{-\infty}^p$. Since $\{\xi_t\}$ is an i.i.d. process, it is straightforward to check that $[\mathbf{X}^p, \boldsymbol{\xi}_{1-q+p}^p]$ is again a function of $\boldsymbol{\xi}_{-\infty}^p$ and independent of $\boldsymbol{\xi}_{p+1}^{p+m}$.

• **Step (ii)**: As $\widehat{\Theta}$ in (94) forms a linear, and therefore Lipschitz, transform, we know from [39, Thm. 2] that

$$\begin{aligned}d(\mathbf{X}^{m+p}) = d(\widehat{\Theta}\widehat{\mathbf{Q}}) &\leq d(\mathbf{X}^p, \boldsymbol{\xi}_{1-q+p}^p, \boldsymbol{\xi}_{p+1}^{p+m}) \\ &= d(\mathbf{X}^p, \boldsymbol{\xi}_{1-q+p}^p) + d(\boldsymbol{\xi}_{p+1}^{p+m}) \quad (95a) \\ &\leq p + q + m\alpha, \quad (95b)\end{aligned}$$

where (95a) is due to the independence of $(\mathbf{X}^p, \boldsymbol{\xi}_{1-q+p}^p)$ and $\boldsymbol{\xi}_{p+1}^{p+m}$ followed by [39, Eqn. 2.22], while (95b) is due to [39, Eq. 2.11] because of the independence of the elements of $\{\boldsymbol{\xi}_t\}$. Besides, we can partition $\widehat{\Theta}\widehat{\mathbf{Q}}$ as follows,

$$\widehat{\Theta}\widehat{\mathbf{Q}} = \left[\begin{array}{c|c|c} \widehat{\Theta}_1 & \widehat{\Theta}_2 & \widehat{\Theta}_3 \end{array}\right] \cdot \left[\begin{array}{c} \mathbf{X}^p \\ \boldsymbol{\xi}_{1-q+p}^p \\ \boldsymbol{\xi}_{p+1}^{m+p} \end{array}\right], \tag{96}$$

so that we can define

$$U_1 = \widehat{\Theta}_1\mathbf{X}^p + \widehat{\Theta}_2\boldsymbol{\xi}_{1-q+p}^p, \tag{97a}$$

$$U_2 = \widehat{\Theta}_3 \boldsymbol{\xi}_{p+1}^{m+p}. \tag{97b}$$

Since $\boldsymbol{\xi}_{p+1}^{p+m}$ is independent of $\mathbf{X}^p$ and $\boldsymbol{\xi}_{1-q+p}^{p}$ jointly, we conclude that U_1 and U_2 are also independent. Therefore, using [39, Eqn. 20.20], we have that

$$d(\widehat{\Theta}\widehat{\mathbf{Q}}) = d(U_1 + U_2) \geq d(U_2). \tag{98}$$

We recall that $\widehat{\Theta}_3$ is a lower-triangular square matrix with all the diagonal elements being 1. Hence (based on [39, Thm. 2]),

$$d(U_2) = d\big(\widehat{\Theta}_3 \boldsymbol{\xi}_{p+1}^{m+p}\big) = d\big(\boldsymbol{\xi}_{p+1}^{m+p}\big) = m\alpha. \tag{99}$$

Overall, this implies that

$$d(\mathbf{X}^{p+m}) = d(\widehat{\Theta}\widehat{\mathbf{Q}}) \geq m\alpha. \tag{100}$$

Combining (95) and (100), we can write

$$\frac{m\alpha}{m+p} \leq \frac{d(\mathbf{X}^{p+m})}{m+p} \leq \frac{p+q+m\alpha}{m+p}, \tag{101}$$

which proves that

$$d_B\big(\{\mathbf{X}_m\}\big) = \lim_{m\to\infty} \frac{d(\mathbf{X}^{p+m})}{m+p} = \alpha. \tag{102}$$

- **Step (iii)**: Using [29, Eqn. 16, 73], we write that

$$d_I\big(\{\mathbf{X}_t\}\big) \geq d(X_{p+1}|\mathbf{X}_{-\infty}^{p}, \boldsymbol{\xi}_{1-q+p}^{p}). \tag{103}$$

According to the definition of RID, we know

$$\begin{aligned} d(X_{p+1}|\mathbf{X}_{-\infty}^{p}, \boldsymbol{\xi}_{1-q+p}^{p}) &= \lim_{k\to\infty} \frac{\mathsf{H}\big([X_{p+1}]_k|\mathbf{X}_{-\infty}^{p}, \boldsymbol{\xi}_{1-q+p}^{p}\big)}{\log k} \\ &\overset{(a)}{=} \lim_{k\to\infty} \frac{\mathsf{H}\big([X_{p+1}]_k|\mathbf{X}^{p}, \boldsymbol{\xi}_{1-q+p}^{p}\big)}{\log k} \\ &= \lim_{k\to\infty} \int_{\mathbb{R}^p\times\mathbb{R}^q} \frac{\mathcal{H}_k(\mathbf{x}^p, \boldsymbol{\tau}_{1-q+p}^{p})}{\log k} \mathrm{d}\mu(\mathbf{x}^p, \boldsymbol{\tau}_{1-q+p}^{p}), \end{aligned} \tag{104}$$

where (a) is because of (9), $\mu(\cdot,\cdot)$ is the joint probability measure on $(\mathbf{X}^p, \boldsymbol{\xi}_{1-q+p}^{p})$ and

$$\mathcal{H}_k(\mathbf{x}^p, \boldsymbol{\tau}_{1-q+p}^{p}) = \mathsf{H}\big([X_{p+1}]_k|\mathbf{X}^p = \mathbf{x}^p, \boldsymbol{\xi}_{1-q+p}^{p} = \boldsymbol{\tau}_{1-q+p}^{p}\big). \tag{105}$$

Combining (9) and (104) we have that

$$\mathcal{H}_k(\mathbf{x}^p, \boldsymbol{\tau}_{1-q+p}^{p}) = \mathsf{H}\big([\xi_{p+1} + c]_k\big), \tag{106}$$

where

$$c = \sum_{i=1}^{q} \theta_i \tau_{p+1-i} + \sum_{i=1}^{p} \phi_i x_{p+1-i}. \tag{107}$$

To further simplify (106)

$$\begin{aligned} \mathsf{H}\big([\xi_{p+1} + c]_k\big) &= \mathsf{H}\big(\big[\xi_{p+1} + c - [c]_k\big]_k\big) \\ &\overset{(a)}{\leq} \mathsf{H}\big(\big[\xi_{p+1} + c - [c]_k\big]_1\big) + \log(k), \end{aligned} \tag{108}$$

where (a) is valid because of [34, Eqn. 11]. Consequently, for $k \geq 2$, we have

$$\frac{\mathsf{H}\big([\xi_{p+1} + c]_k\big)}{\log k} \leq \frac{\mathsf{H}\Big(\big[\xi_{p+1} + c - [c]_k\big]_1\Big)}{\log k} + 1. \tag{109}$$

Since $c - [c]_k \in [0,1)$ and $\mathsf{H}\big([\xi_{p+1}]_1\big) \leq m < \infty$, we can apply Lemma 6 to conclude $\mathsf{H}\big([\xi_{p+1} + c - [c]_k]_1\big) < 4m + \mathrm{C}$, where C is a fixed value and merely dependents on the distribution of ξ_{p+1}. As a result, we have that

$$\frac{\mathsf{H}\big([\xi_{p+1} + c]_k\big)}{\log k} \leq \frac{4m + \mathrm{C}}{\log k} + 1. \tag{110}$$

By plugging in the latter result and (106) in (104), we get

$$\begin{aligned} &\int_{\mathbb{R}^p\times\mathbb{R}^q} \frac{\mathcal{H}_k(\mathbf{x}^p, \boldsymbol{\tau}_{1-q+p}^{p})}{\log k} \mathrm{d}\mu(\mathbf{x}^p, \boldsymbol{\tau}_{1-q+p}^{p}) \\ &\leq \Big(\frac{4m + \mathrm{C}}{\log k} + 1\Big) \underbrace{\int_{\mathbb{R}^p\times\mathbb{R}^q} \mathrm{d}\mu(\mathbf{x}^p, \boldsymbol{\tau}_{1-q+p}^{p})}_{1} \leq \frac{4m + \mathrm{C}}{\log 2} + 1. \end{aligned} \tag{111}$$

This allows us to apply the dominant convergence theorem to achieve

$$\begin{aligned} &d(X_{p+1}|\mathbf{X}_{-\infty}^{p}, \boldsymbol{\xi}_{1-q+p}^{p}) \\ &\quad= \int_{\mathbb{R}^p\times\mathbb{R}^q} \Big(\lim_{k\to\infty} \frac{\mathcal{H}_k(\mathbf{x}^p, \boldsymbol{\tau}_{1-q+p}^{p})}{\log k}\Big) \mathrm{d}\mu(\mathbf{x}^p, \boldsymbol{\tau}_{1-q+p}^{p}) \\ &\quad= \int_{\mathbb{R}^p\times\mathbb{R}^q} d\Big(\xi_{p+1} + \sum_{i=1}^{q} \theta_i \tau_{p+1-i} + \sum_{i=1}^{p} \phi_i x_{p+1-i}\Big) \\ &\qquad\qquad \times \mathrm{d}\mu(\mathbf{x}^p, \boldsymbol{\tau}_{1-q+p}^{p}). \end{aligned} \tag{112}$$

[39, Lem. 3] implies that

$$d\Big(\xi_{p+1} + \sum_{i=1}^{q} \theta_i \tau_{p+1-i} + \sum_{i=1}^{p} \phi_i x_{p+1-i}\Big) = d(\xi_{p+1}) = \alpha, \tag{113}$$

which means

$$d(X_{p+1}|\mathbf{X}_{-\infty}^{p}, \boldsymbol{\xi}^{p}) = \alpha. \tag{114}$$

Using this identity and (103) one can see that

$$d_I\big(\{\mathbf{X}_t\}\big) \geq \alpha. \tag{115}$$

We recall from Theorem 4 that $d_I(\{\mathbf{X}_t\}) \leq d_B(\{\mathbf{X}_t\})$. However, (102) states that $d_B(\{\mathbf{X}_t\}) = \alpha$. Now, (115) shows that $d_I(\{\mathbf{X}_t\}) = d_B(\{\mathbf{X}_t\}) = \alpha$. ∎

Appendix G
Proof of Theorem 8

Proof: We consider the three cases of discrete, continuous, and discrete-continuous excitation noise separately.

- **Discrete excitation** ($\mathbb{P}(\nu_i = 1) = 0$): In this case, we are dealing with infinite mixtures of purely discrete RVs; therefore, we no longer have affinely singular random variables (there is no absolutely continuous component even on lower dimensional subspaces). In this case, we show that ϵ-achievable rates are zero. This together with the fact that $d(\xi_i) = 0$ for a discrete random variable ξ_i (see Theorem 2), proves the theorem.

Using the definition of ARMA process in (9), we know that

$$X_t + \sum_{i=1}^{p} \phi_i X_{t-i} = \xi_t + \sum_{i=1}^{q} \theta_i \xi_{t-i}. \tag{116}$$

Since besides the excitation noise, each X_t depends on its p previous samples, the recursive equation is solvable by having p boundary conditions (the the realization of the excitation noise). As a result, $\mathbf{X}^n$ could be expressed as a linear function of $\mathbf{X}^p$ and ξ_{1-q}^n for each n. Since ξ_i takes value from a finite or countably infinite set, then we can find a finite set $\mathcal{A}_i$ such that

$$\mathbb{P}(\xi_i \in \mathcal{A}_i) \geq 1 - \frac{\delta}{2^{i+q}}, \tag{117}$$

for all $i \in \{1-q, \ldots\}$, where $\delta \in (0,1)$. Therefore, the probability of ξ_{1-q}^n taking a value from $\mathcal{B}_n = \mathcal{A}_i \otimes \ldots \otimes \mathcal{A}_{q+n}$ is lower-bounded as

$$\mathbb{P}(\xi_{1-q}^n \in \mathcal{B}_n) \geq \prod_{i=1}^{n+q} \big(1 - \frac{\delta}{2^i}\big) \geq 1 - \delta \sum_{i=1}^{n+q} \frac{1}{2^i} \geq 1 - \delta. \tag{118}$$

Next, given $\xi_{1-q}^n \in \mathcal{B}_n$, since $\mathbf{X}^n$ is obtained via a linear function $\mathbf{X}^n = f(\mathbf{X}^p, \xi_{1-q}^n)$ of the excitation noise and previous samples, and because of finiteness of $\mathcal{B}_n$ we can cover the support of $\mathbf{X}^n$ with a finite set of linear functions of $\mathbf{X}^p$. More formally, if we define a random vector $\mathbf{Y}^n$ that is distributed as

$$\mathbb{P}(\mathbf{Y}^n = \mathbf{y}^n) = \mathbb{P}(\mathbf{X}^n = \mathbf{y}^n | \boldsymbol{\xi}_{1-q}^n \in \mathcal{A}_m^n), \tag{119}$$

then there exists a set of functions $f_i : \mathbb{R}^p \to \mathbb{R}^n$ for $i \in [1 : |\mathcal{B}_n|]$ such that $\mathrm{supp}(\mathbf{Y}^n) \subseteq \cup_{i=1}^n \mathrm{supp}\big(f_i(\mathbf{X}^p)\big)$. A set with such property (i.e., covered by a finite number of Lipschitz images of $\mathbb{R}^p$) is called p-rectifiable. We observe that rectifiability is further satisfied for all higher values of p, i.e. we can cover the support of $\mathbf{Y}^n$ by adding dummy dimensions to $\mathbf{X}^p$. As stated in [11, Lemma 12], if $\mathrm{supp}(\mathbf{Y}^n)$ is $\lfloor Rn \rfloor$-rectifiable for all sufficiently large ns, then the minimum ϵ-achievable rate for $\{\mathbf{Y}_i\}$ is bounded above by R. Hence, this property is met for $\mathbf{Y}^n$, and $R = \frac{p}{m}$ and all $n \in \{m, m+1, \ldots\}$. As a result, $R_{\{\mathbf{Y}_i\}}(\epsilon) \leq \frac{p}{m}$.

Next, to find an upper-bound for minimum ϵ-achievable rate for $\{\mathbf{X}_i\}$ by R', one needs to find a set of $(n, \lfloor R'n \rfloor)$-encoder-decoder pair g_i, h_i such that

$$\mathbb{P}(g_n(h_n(\mathbf{X}^n)) \neq \mathbf{X}^n) \leq \epsilon. \tag{120}$$

Here, if we choose the same pair of encoders and decoders as of $\{\mathbf{Y}_i\}$, we have

$$\begin{aligned} &\mathbb{P}(g_n(h_n(\mathbf{X}^n)) \neq \mathbf{X}^n) \\ &\quad = \mathbb{P}(g_n(h_n(\mathbf{X}^n)) \neq \mathbf{X}^n | \boldsymbol{\xi}_{1-q}^n \in \mathcal{A}_m^n) \mathbb{P}(\boldsymbol{\xi}_{1-q}^n \in \mathcal{A}_m^n) \\ &\quad + \mathbb{P}(g_n(h_n(\mathbf{X}^n)) \neq \mathbf{X}^n | \boldsymbol{\xi}_{1-q}^n \notin \mathcal{A}_m^n) \mathbb{P}(\boldsymbol{\xi}_{1-q}^n \notin \mathcal{A}_m^n), \end{aligned} \tag{121}$$

that coupled with (118) and (119) concludes in

$$\mathbb{P}(g_n(h_n(\mathbf{X}^n)) \neq \mathbf{X}^n) \leq \mathbb{P}(g_n(h_n(\mathbf{Y}^n)) \neq \mathbf{Y}^n) + \frac{1}{m} \tag{122}$$

$$\leq \epsilon' + \delta. \tag{123}$$

By having large enough n such that $\epsilon' \leq \frac{\epsilon}{2}$ and by setting $\delta \in (0, \frac{\epsilon}{2})$, we have $R(\epsilon) \leq \frac{p}{m}$. Since we could fix m arbitrarily large, and by positiveness of $R(\epsilon)$, we can prove that

$$R(\epsilon) = 0. \tag{124}$$

The above identity together with the inequality

$$R^*(\epsilon) \leq R_B(\epsilon) \leq R(\epsilon), \tag{125}$$

in [11, Eq. 75] proves that all above values of $R*(\epsilon), R_B(\epsilon), R(\epsilon)$ are equal to 0.

• **Continuous excitation** ($\mathbb{P}(\nu_i = 1) = 1$): In this case, the first part of Theorem 6 shows that for all n, $\mathbf{X}^n$ is also absolutely continuous. Now, Theorem 2 implies that $d(\mathbf{X}^n) = n$. Besides, Theorem 3 reveals that $d_B\big(\{\mathbf{X}_t\}\big) = \lim_{n\to\infty} \frac{d(\mathbf{X}^n)}{n} = 1$.

Due to absolute continuity of $\mathbf{X}^n$, it can be thought of as an affinely singular random variable $V_n = 1$ and $d_1 = n$. As a result, one can see that $\Big|\frac{d_1}{n} - d_B\big(\{\mathbf{Z}_t\}\big)\Big| = 0$. Equivalently, we have

$$\mathbb{P}_{V_n}\left(i : \Big|\frac{d_i}{n} - d_B\big(\{\mathbf{Z}_t\}\big)\Big| < \delta\right) = 1 > 1 - \epsilon, \tag{126}$$

for every pair of (δ, ϵ). Finally, using Theorem 7, we conclude that

$$R^*(\epsilon) = R_B(\epsilon) = R(\epsilon) = d_B\big(\{\mathbf{Z}_t\}\big) = d(\xi_1) = 1, \tag{127}$$

where the last equality holds because of Theorem 2.

• **Mixed discrete-continuous excitation** ($\mathbb{P}(\nu_i = 1) \in (0,1)$): Recalling Corollary 1 and Theorem 7, we know that a D/C-ARMA process with finite discrete space is a typical source, and

$$R^*(\epsilon) = R_B(\epsilon) = R(\epsilon) = d_B\big(\{\mathbf{Z}_t\}\big) = d(\xi_1) = \alpha, \tag{128}$$

where the last equality holds because of Theorem 2. ■