On the Information Dimension
of Stochastic Processes

Bernhard C. Geiger, *Senior Member, IEEE,* and Tobias Koch, *Senior Member, IEEE*

Abstract—In 1959, Rényi proposed the information dimension and the d-dimensional entropy to measure the information content of general random variables. This paper proposes a generalization of information dimension to stochastic processes by defining the information dimension rate as the entropy rate of the uniformly-quantized stochastic process divided by minus the logarithm of the quantizer step size $1/m$ in the limit as $m \to \infty$. It is demonstrated that the information dimension rate coincides with the rate-distortion dimension, defined as twice the rate-distortion function $R(D)$ of the stochastic process divided by $-\log(D)$ in the limit as $D \downarrow 0$. It is further shown that, among all multivariate stationary processes with a given (matrix-valued) spectral distribution function (SDF), the Gaussian process has the largest information dimension rate, and that the information dimension rate of multivariate stationary Gaussian processes is given by the average rank of the derivative of the SDF. The presented results reveal that the fundamental limits of almost zero-distortion recovery via compressible signal pursuit and almost lossless analog compression are different in general.

Index Terms—Entropy, Gaussian process, information dimension, rate-distortion dimension

I. Introduction

IN 1959, Rényi [1] proposed the *information dimension* and the *d-dimensional entropy* to measure the information content of general random variables (RVs). His idea was to quantize the RV X by a uniform quantizer of step size $1/m$, and to then analyze the entropy of the quantized RV $[X]_m$

in the limit as m tends to infinity. Assuming that the entropy $H([X]_m)$ exists and the asymptotic expansion

$$H([X]_m) = d(X) \log m + H_d(X) + o(1) \qquad (1)$$

holds for $m \to \infty$ (where $o(1)$ refers to remainder terms that vanish as $m \to \infty$), Rényi referred to $d(X)$ as the information dimension and to $H_d(X)$ as the d-dimensional entropy.

In recent years, it was shown that the information dimension is of relevance in various areas of information theory, including rate-distortion theory, almost lossless analog compression, or the analysis of interference channels. For example, Kawabata and Dembo [2] showed that the information dimension of a RV is equal to its *rate-distortion dimension*, defined as twice the rate-distortion function $R(D)$ divided by $-\log(D)$ in the limit as $D \downarrow 0$. Koch [3] demonstrated that the rate-distortion function of a source with infinite information dimension is infinite, and that for any source with finite information dimension and finite differential entropy the Shannon lower bound on the rate-distortion function is asymptotically tight. Wu and Verdú [4] analyzed linear encoding and Lipschitz decoding of discrete-time, independent and identically distributed (i.i.d.), stochastic processes and showed that the information dimension plays a fundamental role in achievability and converse results. Wu *et al.* [5] showed that the degrees of freedom of the K-user Gaussian interference channel can be characterized through the sum of information dimensions. Stotz and Bölcskei [6] generalized this result to vector interference channels.

Jalali and Poor [7] proposed a generalization of information dimension to stationary, discrete-time, stochastic processes by defining the information dimension $d'(\{X_t\})$ of the stochastic process $\{X_t\}$ as the information dimension of $(X_1, \ldots, X_k)$ divided by k in the limit as $k \to \infty$.[1] They showed that, for ψ^*-mixing processes, the information dimension is an achievable rate for universal compressed sensing with linear encoding and decoding via Lagrangian minimum entropy pursuit [7, Th. 8]. Rezagah *et al.* [8] showed that $d'(\{X_t\})$ coincides, under certain conditions, with the rate-distortion dimension $\dim_R(\{X_t\})$, thus generalizing the result by Kawabata and Dembo [2] to stochastic processes. Other notions of information dimensions for stochastic processes are discussed in [9].

In this paper, we propose a different definition for the information dimension of stationary, discrete-time, stochastic processes. Specifically, let $\{[X_t]_m\}$ denote the stochastic

[1]More precisely, Jalali and Poor define the information dimension of a stochastic process via a conditional entropy of the uniformly-quantized process. For stationary processes, their definition coincides with the above-mentioned definition [7, Lemma 3].

process $\{X_t\}$ uniformly quantized with step size $1/m$. We define the *information dimension rate* $d(\{X_t\})$ of $\{X_t\}$ as the entropy rate of $\{[X_t]_m\}$ divided by $\log m$ in the limit as $m \to \infty$. For i.i.d. processes, our definition coincides with that of Jalali and Poor (and, in fact, evaluates to Rényi's information dimension of the marginal RV X_t). More generally, we show that these definitions are equivalent for ψ^*-mixing processes. Nevertheless, there are stochastic processes for which the two definitions disagree. In particular, we derive a closed-form expression for the information dimension rate of stationary, multivariate, Gaussian processes with *power spectral density* (PSD) S_X, which specialized to the univariate case yields that $d(\{X_t\})$ is equal to the Lebesgue measure of the set of harmonics on $[-1/2, 1/2]$ where S_X is positive. For Gaussian processes with a bandlimited PSD, this implies that the information dimension rate $d(\{X_t\})$ is equal to twice the PSD's bandwidth. This is consistent with the intuition that for such processes not all samples contain information. For example, if the bandwidth of the PSD is $1/4$, then we expect that half of the samples in $\{X_t\}$ can be expressed as linear combinations of the other samples and, hence, do not contain information. In contrast, we show that the information dimension $d'(\{X_t\})$ is 1 if S_X is positive on any set with positive Lebesgue measure. In other words, $d'(\{X_t\})$ does not capture the dependence of the information dimension on the support size of S_X.

By emulating the proof of [2, Lemma 3.2], we further show that, for any stochastic process $\{X_t\}$, the information dimension rate $d(\{X_t\})$ coincides with the rate-distortion dimension $\dim_R(\{X_t\})$. This implies that $d'(\{X_t\})$ coincides with $\dim_R(\{X_t\})$ only for those stochastic processes for which $d'(\{X_t\}) = d(\{X_t\})$.

The rest of this paper is organized as follows. In Section II, we introduce the notation used in this paper. In Section III, we present preliminary results on the Rényi information dimension of RVs and random vectors. In Section IV, we present our definition of the information dimension rate of a stochastic process, discuss its connection to the rate-distortion dimension, and compute the information dimension rate of stationary Gaussian processes. In Section V, we review the information dimension proposed by Jalali and Poor and discuss its relation to $d(\{X_t\})$. In Section VI, we briefly discuss the operational meanings of information dimension in compressed sensing and zero-distortion recovery. Section VII concludes the paper with a discussion of the obtained results. Some of the proofs are deferred to the appendices.

II. Notation and Preliminaries

We denote by $\mathbb{R}$, $\mathbb{C}$, and $\mathbb{Z}$ the set of real numbers, the set of complex numbers, and the set of integers, respectively. We further denote by $\mathbb{R}^+$ and $\mathbb{N}$ the set of nonnegative real numbers and the set of positive integers, respectively. We use a calligraphic font, such as $\mathcal{F}$, to denote other sets, and we denote complements as $\mathcal{F}^c$. The set difference between two sets $\mathcal{F}$ and $\mathcal{G}$ is written as $\mathcal{F} \setminus \mathcal{G}$.

The real and imaginary parts of a complex number z are denoted as $\mathfrak{Re}(z)$ and $\mathfrak{Im}(z)$, respectively, i.e., $z = \mathfrak{Re}(z) +$

$\imath \mathfrak{Im}(z)$ where $\imath \triangleq \sqrt{-1}$. The complex conjugate of z is denoted as z^*.

We use uppercase letters to denote deterministic matrices and boldface lowercase letters to denote deterministic vectors. The transpose of a vector or matrix is denoted by $(\cdot)^\mathsf{T}$, the Hermitian transpose by $(\cdot)^\mathsf{H}$. The determinant and rank of a matrix A are $\det A$ and $\mathrm{rank}(A)$, respectively. We denote by I_L the $L \times L$ identity matrix.

We denote RVs by uppercase letters, e.g., X. For a finite or countably infinite collection of RVs we abbreviate $X_\ell^k \triangleq (X_\ell, \ldots, X_{k-1}, X_k)$, $X_\ell^\infty \triangleq (X_\ell, X_{\ell+1}, \ldots)$, and $X_{-\infty}^k \triangleq (\ldots, X_{k-1}, X_k)$.[2] Random vectors are denoted by boldface uppercase letters, e.g., $\mathbf{X} \triangleq (X_1, \ldots, X_L)^\mathsf{T}$. Univariate discrete-time stochastic processes are denoted as $\{X_t, t \in \mathbb{Z}\}$ or, in short, as $\{X_t\}$. For L-variate stochastic processes we use the same notation but with X_t replaced by $\mathbf{X}_t \triangleq (X_{1,t}, \ldots, X_{L,t})^\mathsf{T}$. We call $\{X_{i,t}, t \in \mathbb{Z}\}$ a *component process*.

We denote the probability measure of the RV X by P_X. If P_X is absolutely continuous with respect to (w.r.t.) the Lebesgue measure, then we denote its probability density function (PDF) as f_X. We denote by X_G a Gaussian RV with the same mean and variance as X, and we denote the corresponding Gaussian density as g_X.

We define the quantization of a real-valued RV X with precision m as

$$[X]_m \triangleq \frac{\lfloor mX \rfloor}{m} \tag{2}$$

where $\lfloor a \rfloor$ is the largest integer less than or equal to a. Likewise, $\lceil a \rceil$ denotes the smallest integer greater than or equal to a. We denote by $[X_\ell^k]_m = ([X_\ell]_m, \ldots, [X_k]_m)$ the component-wise quantization of X_ℓ^k (and similarly for other finite or countably infinite collections of RVs and random vectors). For complex RVs Z with real part R and imaginary part I, the quantization $[Z]_m$ is equal to $[R]_m + \imath[I]_m$. We define $\mathcal{C}(z_1^k, a) \triangleq [z_1, z_1 + a) \times \cdots \times [z_k, z_k + a)$ as the k-dimensional hypercube in $\mathbb{R}^k$, with its bottom-left corner at z_1^k and with sidelength a. For example, we have that $[X_1^k]_m = z_1^k$ if $X_1^k \in \mathcal{C}(z_1^k, 1/m)$.

Let $H(\cdot)$, $h(\cdot)$, and $D(\cdot\|\cdot)$ denote entropy, differential entropy, and relative entropy, respectively, and let $I(\cdot;\cdot)$ denote mutual information [10]. We take logarithms to base $e \approx 2.718$, so mutual informations and entropies have dimension *nats*. The entropy rate of a discrete-valued, stationary, L-variate process $\{\mathbf{X}_t\}$ is [10, Sec. 4.2]

$$H'(\{\mathbf{X}_t\}) \triangleq \lim_{k \to \infty} \frac{H(\mathbf{X}_1^k)}{k}. \tag{3}$$

Note that the stationarity of $\{\mathbf{X}_t\}$ guarantees that the limit in (3) exists and is equal to [10, Th. 4.2.1]

$$\lim_{k \to \infty} \frac{H(\mathbf{X}_1^k)}{k} = \lim_{k \to \infty} H(\mathbf{X}_1|\mathbf{X}_{-k}^0). \tag{4}$$

We say that a stationary process $\{\mathbf{X}_t\}$ is ψ^*-*mixing* if

$$\limsup_{k \to \infty} \sup_{A,B} \frac{P_{\mathbf{X}_{-\infty}^0, \mathbf{x}_k^\infty}(A \cap B)}{P_{\mathbf{X}_{-\infty}^0}(A) P_{\mathbf{X}_k^\infty}(B)} = 1 \tag{5}$$

---

[2]If $k < \ell$, then X_ℓ^k is the empty set.

where the supremum is over all $A \in \mathcal{F}_{-\infty}^0$ and $B \in \mathcal{F}_k^\infty$ satisfying $P_{\mathbf{X}_{-\infty}^0}(A)P_{\mathbf{X}_k^\infty}(B) > 0$, and where $\mathcal{F}_{-\infty}^0$ and $\mathcal{F}_k^\infty$ are the σ-fields generated by $\mathbf{X}_{-\infty}^0$ and $\mathbf{X}_k^\infty$, respectively. The ψ^*-mixing property implies that $\{\mathbf{X}_t\}$ is *information regular*, i.e., [11, pp. 111-112]

$$\lim_{k\to\infty} I(\mathbf{X}_k^\infty; \mathbf{X}_{-\infty}^0) = 0. \tag{6}$$

III. RÉNYI INFORMATION DIMENSION

The *Rényi information dimension* of a collection of RVs X_1^k is defined as [1]

$$d(X_1^k) \triangleq \lim_{m\to\infty} \frac{H([X_1^k]_m)}{\log m}, \quad \text{if the limit exists.} \tag{7}$$

When the limit does not exist, we say that the information dimension does not exist. In this case, one may replace the limit either by the *limit superior* or by the *limit inferior* (denoted as $\overline{\lim}$ and $\underline{\lim}$, respectively)

$$\bar{d}(X_1^k) \triangleq \overline{\lim_{m\to\infty}} \frac{H([X_1^k]_m)}{\log m} \tag{8a}$$

$$\underline{d}(X_1^k) \triangleq \underline{\lim_{m\to\infty}} \frac{H([X_1^k]_m)}{\log m} \tag{8b}$$

and call $\bar{d}(X_1^k)$ and $\underline{d}(X_1^k)$ the upper and lower information dimension of X_1^k, respectively. Clearly,

$$\bar{d}(X_1^k) = \underline{d}(X_1^k) = d(X_1^k) \tag{9}$$

if the limit in (7) exists.

We shall follow this notation throughout the document. Specifically, when reporting results in connection with limits, an overline $\overline{(\cdot)}$ indicates that the quantity in the brackets has been computed using the limit superior, an underline $\underline{(\cdot)}$ indicates that it has been computed using the limit inferior, both an overline and an underline $\overline{\underline{(\cdot)}}$ indicates that a result holds irrespective of whether the limit superior or limit inferior is taken. We write no lines if the limit exists.

Definition 1: For two RVs X and W with joint probability measure $P_{X,W}$, the *conditional information dimension* is defined as

$$d(X|W) \triangleq \lim_{m\to\infty} \frac{H([X]_m|W)}{\log m} \tag{10}$$

provided the limit exists. If the limit does not exist, then we define the upper and lower conditional information dimension $\bar{d}(X|W)$ and $\underline{d}(X|W)$ by replacing the limit with the limit superior and the limit inferior, respectively.

A. Properties of Information Dimension

The information dimension of a collection X_1^k is bounded by the number of RVs in the collection, given the integer part of this collection has finite entropy.

Lemma 1 ([1, eq. (7)], [4, Prop. 1]): Let X_1^k be a collection of real-valued RVs. If $H([X_1^k]_1) < \infty$, then

$$0 \le \underline{d}(X_1^k) \le \bar{d}(X_1^k) \le k. \tag{11}$$

If $H([X_1^k]_1) = \infty$, then $\overline{\underline{d}}(X_1^k) = \infty$.

Trivially, if X_1^k is a collection of discrete RVs satisfying $H([X_1^k]_1) < \infty$, then $d(X_1^k) = 0$. Moreover, if the joint distribution of X_1^k is absolutely continuous w.r.t. the Lebesgue measure on $\mathbb{R}^k$ and if $H([X_1^k]_1) < \infty$, then $d(X_1^k) = k$ [1, Th. 4]. More generally, Rényi claims that the information dimension of X_1^k equals $n < k$ if the joint distribution of X_1^k is absolutely continuous on some sufficiently smooth n-dimensional manifold in $\mathbb{R}^k$ [1, p. 209]. Furthermore, if X is a real-valued RV satisfying $H([X]_1) < \infty$ and with probability measure

$$P_X = (1-\rho)P_d + \rho P_c \tag{12}$$

where P_d is a discrete measure, P_c is an absolutely-continuous measure, and $0 \le \rho \le 1$, then [1, Th. 3]

$$d(X) = \rho. \tag{13}$$

Two well-known properties of entropy are that it is reduced by conditioning [10, Th. 2.6.5] and that it obeys a chain rule. Furthermore, the conditional entropy of X given Y can be computed by first calculating the entropy conditioned on the event that $Y = y$, and by then averaging over Y. The corresponding results for information dimension are presented in the following three lemmas.

Lemma 2: Suppose that $H([X]_1) < \infty$. Then, we have for any two RVs X and Y

$$\int \underline{d}(X|Y = y)\mathrm{d}P_Y(y) \le \underline{d}(X|Y)$$

$$\le \bar{d}(X|Y) \le \int \bar{d}(X|Y = y)\mathrm{d}P_Y(y). \tag{14}$$

Consequently, if $d(X|Y = y)$ exists P_Y-almost surely, then the limit in (10) exists and

$$d(X|Y) = \int d(X|Y = y)\mathrm{d}P_Y(y). \tag{15}$$

Proof: See Appendix A-A. ■

Lemma 3: For any two RVs X and Y, we have

$$\overline{\underline{d}}(X|Y) \le \overline{\underline{d}}(X) \tag{16}$$

with equality if X and Y are independent.

Proof: Since conditioning reduces entropy, we have $H([X]_m|Y) \le H([X]_m)$, with equality if X and Y are independent. The lemma follows by dividing both sides of the inequality by $\log m$ and taking limits as $m \to \infty$. ■

Lemma 4: For the collection of RVs X_1^k, we have

$$\sum_{t=1}^k \bar{d}(X_t) \ge \bar{d}(X_1^k) \ge \underline{d}(X_1^k) \ge \sum_{t=1}^k \underline{d}(X_t|X_1^{t-1}). \tag{17}$$

Proof: See Appendix A-B. ■

The left-most inequality in (17) holds with equality if all information dimensions exist and the RVs X_1^k are independent. There are examples where the right-most inequality is strict.

Example 1: Let (X_1, X_2) be uniformly distributed on $[0,1]^2$ and let $Y = g(X_1, X_2)$, where $g: [0,1]^2 \to [0,1]$ is bijective. Such a function can be constructed (see also the discussion in [4, Section IV.B]). Since g is bijective, we have $d(Y|X_1, X_2) = d(X_1, X_2|Y) = 0$. Moreover, since (X_1, X_2) is uniformly distributed on $[0,1]^2$, we have $d(X_1, X_2) = 2$.

Finally, we have $d(Y) \leq 1$ by Lemma 1. From Lemma 4, we get

$$d(X_1, X_2, Y) \geq d(X_1, X_2) + d(Y|X_1, X_2) = 2. \quad (18)$$

However, we also have

$$d(Y) + d(X_1, X_2|Y) \leq 1. \quad (19)$$

It follows that

$$d(X_1, X_2, Y) > d(Y) + d(X_1, X_2|Y) \quad (20)$$

so the chain rule $d(X_1, X_2, Y) \geq d(Y) + d(X_1, X_2|Y)$ holds with strict inequality.

The above example not only demonstrates that the chain rule for information dimension may hold with strict inequality, it also shows that the order in which the chain rule is expanded can be crucial.

B. Information Dimension of Finite-Variance RVs

For RVs X_1^k that have a finite variance, the upper bound on $\bar{d}(X_1^k)$ presented in Lemma 1 can be tightened. To this end, we introduce further notation. We denote the covariance matrix of the vector $\mathbf{X} = (X_1, \ldots, X_k)^\mathsf{T}$ by $C_{X_1^k}$. Furthermore, the cross-covariance matrix between $\mathbf{X} = (X_1, \ldots, X_k)^\mathsf{T}$ and $\mathbf{Y} = (Y_1, \ldots, Y_k)^\mathsf{T}$ is denoted by $C_{X_1^k Y_1^k}$, and the covariance matrix of the vector $(X_1, \ldots, X_k, Y_1, \ldots, Y_k)^\mathsf{T}$ is denoted by $C_{X_1^k, Y_1^k}$. Clearly,

$$C_{X_1^k, Y_1^k} = \begin{bmatrix} C_{X_1^k} & C_{X_1^k Y_1^k} \\ C_{X_1^k Y_1^k}^\mathsf{T} & C_{Y_1^k} \end{bmatrix}. \quad (21)$$

One can show that the information dimension of a collection of real-valued RVs X_1^k cannot exceed the rank of its covariance matrix, i.e.,

$$\bar{\underline{d}}(X_1^k) \leq \operatorname{rank}(C_{X_1^k}). \quad (22)$$

This agrees with the intuition that linearly-dependent components of X_1^k do not contribute to the information dimension. One can further show that collections of Gaussian RVs achieve this upper bound with equality. Thus, among all RVs with a given covariance structure, the Gaussian RV maximizes information dimension. These results follow directly from the more general results for stochastic processes (Theorem 10) in Section IV.

The next theorem evaluates the conditional information dimension of X_1^k given Y_1^ℓ for jointly Gaussian RVs (X_1^k, Y_1^ℓ).

Theorem 5: Let (X_1^k, Y_1^ℓ) be a collection of real-valued, jointly Gaussian RVs. The conditional information dimension of X_1^k given Y_1^ℓ is equal to

$$d(X_1^k|Y_1^\ell) = \operatorname{rank}(C_{X_1^k|Y_1^\ell}) \quad (23)$$

where $C_{X_1^k|Y_1^\ell}$ is the generalized Schur complement of $C_{Y_1^\ell}$ in $C_{X_1^k, Y_1^\ell}$.

Proof: See Appendix A-C. ∎

Theorem 5 implies that the chain rule in Lemma 4 holds with equality for Gaussian RVs. Indeed, if X_1^k is a collection of real-valued, jointly Gaussian RVs, then we have $d(X_1^k) = \operatorname{rank}(C_{X_1^k})$ and $d(X_1^\ell) = \operatorname{rank}(C_{X_1^\ell})$. Moreover, by Theorem 5, $d(X_{\ell+1}^k|X_1^\ell)$ equals the rank of the generalized

Schur complement of $C_{X_1^\ell}$ in $C_{X_1^k}$, denoted by $C_{X_{\ell+1}^k|X_1^\ell}$. Since the rank of $C_{X_1^k}$ can be written as the sum of the ranks of $C_{X_1^\ell}$ and $C_{X_{\ell+1}^k|X_1^\ell}$ [12, 7.1.P28], the claim follows.

IV. THE INFORMATION DIMENSION RATE

We next propose the *information dimension rate* as a generalization of information dimension to stochastic processes. We define the information dimension rate for general (possibly non-stationary) processes. However, for the sake of simplicity, most of our results will then be presented for stationary processes.

Definition 2: The *information dimension rate* of the L-variate stochastic process $\{\mathbf{X}_t\}$ is defined as

$$d(\{\mathbf{X}_t\}) \triangleq \lim_{m \to \infty} \lim_{k \to \infty} \frac{H([\mathbf{X}_1^k]_m)}{k \log m} \quad (24)$$

provided the limits exist. If the limits do not exist, then we define the upper and lower information dimension rate $\bar{d}(\{\mathbf{X}_t\})$ and $\underline{d}(\{\mathbf{X}_t\})$ by replacing the limits with the limits superior and limits inferior, respectively.

A. Properties of the Information Dimension Rate

The information dimension rate satisfies properties similar to those presented in Lemma 1 for the information dimension. We summarize them in the following lemma.

Lemma 6: Let $\{\mathbf{X}_t\}$ be a stationary, L-variate, real-valued process. If $H([\mathbf{X}_1]_1) < \infty$, then

$$0 \leq \underline{d}(\{\mathbf{X}_t\}) \leq \bar{d}(\mathbf{X}_1) \leq L. \quad (25)$$

If $H([\mathbf{X}_1]_1) = \infty$, then $\bar{d}(\{\mathbf{X}_t\}) = \infty$.

Proof: Suppose first that $H([\mathbf{X}_1]_1) < \infty$. Then, the rightmost inequality in (25) follows from (11). The left-most inequality follows from the nonnegativity of entropy. Finally, the center inequality follows since conditioning reduces entropy, hence $H'(\{[\mathbf{X}_t]_m\}) \leq H([\mathbf{X}_1]_m)$.

Now suppose that $H([\mathbf{X}_1]_1) = \infty$. Since $[\mathbf{X}_1]_1$ is a function of $[\mathbf{X}_1^k]_m$ for every $m, k \in \mathbb{N}$, we have

$$H([\mathbf{X}_1]_1) \leq H([\mathbf{X}_1^k]_m). \quad (26)$$

This implies that $H'(\{[\mathbf{X}_t]_m\}) = \infty$, and the claim that $\underline{d}(\{\mathbf{X}_t\}) = \infty$ follows from Definition 2. ∎

The next result discusses how Lipschitz transformations affect the information dimension rate.

Lemma 7: Let $\{\mathbf{X}_t\}$ be a stationary, L-variate, real-valued process, and let $\{f_t, t \in \mathbb{Z}\}$ be a sequence of Lipschitz functions from $\mathbb{R}^L$ to $\mathbb{R}^M$ with Lipschitz constants K_t satisfying

$$\sup_{t \in \mathbb{Z}} \mathsf{K}_t < \infty. \quad (27)$$

Then,

$$\bar{d}(\{f_t(\mathbf{X}_t)\}) \leq \bar{d}(\{\mathbf{X}_t\}) = \bar{d}(\{\mathbf{X}_t, f_t(\mathbf{X}_t)\}). \quad (28)$$

Proof: See Appendix B. ∎

If $\{f_t\}$ is a sequence of bi-Lipschitz functions with uniformly-bounded Lipschitz constants, then Lemma 7 implies that $\bar{d}(\{f_t(\mathbf{X}_t)\}) = \bar{d}(\{\mathbf{X}_t\})$. As a corollary, we thus obtain that the information dimension rate is invariant under

scaling and translation. More generally, it follows that, if $\{\mathbf{c}_t\}$ and $\{W_t\}$ are sequences of L-variate vectors and $(L \times L)$-dimensional matrices, the latter satisfying $\sup_{t\in\mathbb{Z}} \|W_t\| < \infty$ and $\sup_{t\in\mathbb{Z}} \|W_t^{-1}\| < \infty$ for some induced matrix norm $\|\cdot\|$, then

$$\overline{d}(\{W_t\mathbf{X}_t + \mathbf{c}_t\}) = \overline{d}(\{\mathbf{X}_t\}). \tag{29}$$

Since the information dimension rate of an i.i.d. process equals the information dimension of its marginal RVs, we further recover the well-known result that the information dimension of collections of RVs is invariant under scaling and translation [13, Lemma 3].

The next lemma shows that the information dimension rate of a collection of stochastic processes is unaffected by those that have zero information dimension rate.

Lemma 8: Let $\{\mathbf{X}_t\}$ and $\{\mathbf{Z}_t\}$ be two jointly stationary, L-variate, real-valued processes, and assume that $d(\{\mathbf{Z}_t\}) = 0$. Then,

$$\overline{d}(\{\mathbf{X}_t\}, \{\mathbf{Z}_t\}) = \overline{d}(\{\mathbf{X}_t\}) \tag{30}$$

$$\overline{d}(\{\mathbf{X}_t + \mathbf{Z}_t\}) = \overline{d}(\{\mathbf{X}_t\}). \tag{31}$$

Moreover, if Z is discrete with $H(Z) < \infty$, then we further have

$$\overline{d}(\{\mathbf{X}_t\}|Z) \triangleq \lim_{m\to\infty} \lim_{k\to\infty} \frac{H([\mathbf{X}_1^k]_m|Z)}{k\log m} = \overline{d}(\{\mathbf{X}_t\}). \tag{32}$$

Proof: See Appendix C. ∎

Inter alia, Lemma 8 can be used to compute the information dimension rate of a countable mixture of stochastic processes. For example, specialized to i.i.d. processes, (32) together with Lemma 2 recovers (13) by choosing $X_1 \sim P_d$, $X_2 \sim P_c$, and $P_Z(1) = 1 - P_Z(2) = 1 - \rho$.

B. Information Dimension Rate vs. Rate-Distortion Dimension

Let $R(\mathbf{X}_1^k, D)$ denote the rate-distortion function of the source $\mathbf{X}_1^k$, i.e.,

$$R(\mathbf{X}_1^k, D) \triangleq \inf_{\substack{P_{\hat{\mathbf{X}}_1^k|\mathbf{X}_1^k}: \\ \mathsf{E}[\|\hat{\mathbf{X}}_1^k - \mathbf{X}_1^k\|_2^2] \leq D}} I(\mathbf{X}_1^k; \hat{\mathbf{X}}_1^k) \tag{33}$$

where the infimum is over all conditional distributions of $\hat{\mathbf{X}}_1^k$ given $\mathbf{X}_1^k$ such that

$$\mathsf{E}[\|\hat{\mathbf{X}}_1^k - \mathbf{X}_1^k\|_2^2] \leq D \tag{34}$$

and where $\|\cdot\|_2$ denotes the Euclidean norm. We have the following definition.

Definition 3: The *rate-distortion dimension* of the L-variate stochastic process $\{\mathbf{X}_t\}$ is defined as

$$\mathsf{dim}_R(\{\mathbf{X}_t\}) \triangleq 2 \lim_{D\downarrow 0} \lim_{k\to\infty} \frac{R(\mathbf{X}_1^k, kD)}{-k\log D} \tag{35}$$

provided the limits over D and k exist. (When the process $\{\mathbf{X}_t\}$ is stationary, the limit over k always exists [14, Th. 9.8.1].) If the limits do not exist, then we define the upper and lower rate-distortion dimension $\overline{\mathsf{dim}}_R(\{\mathbf{X}_t\})$ and $\underline{\mathsf{dim}}_R(\{\mathbf{X}_t\})$ by replacing the limits with the limits superior and limits inferior, respectively.

Intuitively, the rate-distortion function

$$R(D) \triangleq \lim_{k\to\infty} \frac{R(\mathbf{X}_1^k, kD)}{k} \tag{36}$$

corresponds to the minimum number of nats per source symbol required to compress a stationary and ergodic source $\{\mathbf{X}_t\}$ with a vector quantizer of average per-symbol distortion not exceeding D [14, Sec. 9.8]. The rate-distortion dimension characterizes the growth of $R(D)$ as D vanishes. For example, for an i.i.d. Gaussian source with variance σ^2, we have [10, Th. 13.3.2]

$$R(D) = \frac{1}{2}\log\left(\frac{\sigma^2}{D}\right)\mathbf{1}\{0 \leq D \leq \sigma^2\} \tag{37}$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function. Observe that in this case $R(D)$ grows like $1/2\log(1/D)$ as $D \to 0$. The rate-distortion dimension corresponds to twice the pre-log factor of the rate-distortion function $R(D)$, which in this case is 1.

In contrast, the information dimension rate characterizes the growth of the entropy rate $H'(\{[\mathbf{X}_t]_m\})$ as m increases. This entropy rate, in turn, corresponds essentially to the number of nats per source symbol required to compress each symbol $\mathbf{X}_t$ of a stationary and ergodic source $\{\mathbf{X}_t\}$ with a uniform quantizer of step size $1/m$. Since a symbol-wise, uniform quantizer cannot outperform the best vector quantizer, it follows that the information dimension rate is lower-bounded by the rate-distortion dimension.

For RVs, Kawabata and Dembo showed that the rate-distortion dimension is actually equal to its information dimension [2, Prop. 3.3]. Thus, a symbol-wise, uniform quantizer achieves the same information dimension as the best vector quantizer. The following theorem generalizes this result to stochastic processes.

Theorem 9: For any L-variate, real-valued process $\{\mathbf{X}_t\}$,

$$\overline{\mathsf{dim}}_R(\{\mathbf{X}_t\}) = \overline{d}(\{\mathbf{X}_t\}). \tag{38}$$

Proof: See Appendix D. ∎

Note that Theorem 9 also holds for non-stationary processes.

C. Information Dimension Rate of Finite-Variance Processes

Let $\{\mathbf{X}_t\}$ be a stationary, L-variate, real-valued process with mean vector $\boldsymbol{\mu}$ and (matrix-valued) spectral distribution function (SDF) $\theta \mapsto \mathsf{F_X}(\theta)$. Thus, $\mathsf{F_X}$ is a bounded, non-decreasing, and right-continuous function on $[-1/2, 1/2]$ such that the autocovariance function

$$K_\mathbf{X}(\tau) \triangleq \mathsf{E}\big[(\mathbf{X}_{t+\tau} - \boldsymbol{\mu})(\mathbf{X}_t - \boldsymbol{\mu})^\intercal\big], \quad \tau \in \mathbb{Z} \tag{39}$$

is given by the Lebesgue-Stieltjes integral [15, (7.3), p. 141]

$$K_\mathbf{X}(\tau) = \int_{-1/2}^{1/2} e^{-\imath 2\pi\tau\theta}\mathsf{dF_X}(\theta), \quad \tau \in \mathbb{Z}. \tag{40}$$

It follows that the (i, j)-th element of $\mathsf{F_X}$ is the cross SDF $\theta \mapsto \mathsf{F}_{X_iX_j}(\theta)$ of the component processes $\{X_{i,t}\}$ and $\{X_{j,t}\}$, i.e.,

$$K_{X_iX_j}(\tau) = \int_{-1/2}^{1/2} e^{-\imath 2\pi\tau\theta}\mathsf{dF}_{X_iX_j}(\theta), \quad \tau \in \mathbb{Z} \tag{41}$$

where

$$K_{X_i X_j}(\tau) \triangleq \mathsf{E}\left[(X_{i,t+\tau} - \mu_i)(X_{j,t} - \mu_j)\right], \quad \tau \in \mathbb{Z} \quad (42)$$

denotes the cross-covariance function. It further follows that the diagonal elements of $\mathsf{F}_\mathbf{X}$ are real and non-decreasing, and they satisfy $\mathsf{F}_{X_i}(1/2) - \mathsf{F}_{X_i}(-1/2) = \sigma_i^2$, where σ_i denotes the standard deviation of $X_{i,t}$. It can be shown that $\theta \mapsto \mathsf{F}_\mathbf{X}(\theta)$ has a derivative almost everywhere, which has positive semi-definite, Hermitian values [15, (7.4), p. 141]. We shall denote the derivative of $\mathsf{F}_\mathbf{X}$ by $\mathsf{F}'_\mathbf{X}$. When $\mathsf{F}_\mathbf{X}$ is absolutely continuous w.r.t. the Lebesgue measure, its derivative $\mathsf{F}'_\mathbf{X}$ coincides with the PSD $\mathsf{S}_\mathbf{X}$ of $\{\mathbf{X}_t\}$.

The following theorem shows that, among all processes of a given SDF, the Gaussian process maximizes the information dimension rate. It further characterizes the information dimension rate of such processes in terms of the SDF.

Theorem 10: Let $\{\mathbf{X}_t\}$ be a stationary, L-variate, real-valued process with SDF $\mathsf{F}_\mathbf{X}$. Then,

$$\overline{d}(\{\mathbf{X}_t\}) \leq \int_{-1/2}^{1/2} \mathrm{rank}(\mathsf{F}'_\mathbf{X}(\theta)) \mathrm{d}\theta \quad (43)$$

with equality if $\{\mathbf{X}_t\}$ is Gaussian.

Proof: See Appendix E. ∎

In order to prove Theorem 10, we invoke Bussgang's theorem to obtain an expression for the SDF of a quantized Gaussian process $\{[\mathbf{X}_t]_m\}$ as a function of the SDF of the original process $\{\mathbf{X}_t\}$. Since we believe that this result is interesting on its own, we present it below.

Lemma 11: Let $\{\mathbf{X}_t\}$ be a stationary, L-variate, real-valued, Gaussian process with mean vector $\boldsymbol{\mu}$ and SDF $\mathsf{F}_\mathbf{X}$. Then, the (i,j)-th entry of the SDF $\theta \mapsto \mathsf{F}_{[\mathbf{X}_t]_m}(\theta)$ of $\{[\mathbf{X}_t]_m\}$ satisfies

$$\mathsf{F}_{[X_i]_m [X_j]_m}(\theta) = (a_i + a_j - 1)\mathsf{F}_{X_i X_j}(\theta) + \mathsf{F}_{N_i N_j}(\theta) \quad (44)$$

where $N_{i,t} \triangleq X_{i,t} - [X_{i,t}]_m$ and

$$a_i \triangleq \frac{1}{\sigma_i^2} \mathsf{E}\left[(X_{i,t} - \mu_i)([X_{i,t}]_m - \mathsf{E}\left[[X_{i,t}]_m\right])\right]. \quad (45)$$

(In (45), μ_i and σ_i denote the mean and standard deviation of $X_{i,t}$.) For every $i = 1, \ldots, L$, we have

$$|1 - a_i| \leq \frac{1}{m}\sqrt{\frac{2}{\pi \sigma_i^2}} \quad (46)$$

and

$$\int_{-1/2}^{1/2} \mathrm{d}\mathsf{F}_{N_i}(\theta) \leq \frac{1}{m^2}. \quad (47)$$

Moreover, if all component processes have zero mean and unit variance, then $a_1 = \ldots = a_L$ and

$$\mathsf{F}_{[\mathbf{X}]_m}(\theta) = (2a_1 - 1)\mathsf{F}_\mathbf{X}(\theta) + \mathsf{F}_\mathbf{N}(\theta). \quad (48)$$

Proof: See Appendix F. ∎

As a corollary to Theorem 10, we obtain that for univariate, stationary, Gaussian processes with PSD S_X, the information dimension rate is equal to the Lebesgue measure of the set of harmonics on $[-1/2, 1/2]$ where S_X is positive, i.e.,

$$d(\{X_t\}) = \lambda(\{\theta \colon \mathsf{S}_X(\theta) > 0\}) \quad (49)$$

where $\lambda(\cdot)$ denotes the Lebesgue measure. As pointed out by one of the reviewers, (49) can also be obtained directly by using the equivalence of information dimension rate and rate-distortion dimension (Theorem 9) together with the parametric representation of the rate-distortion function [14, eqs. (9.7.42) & (9.7.43)]

$$R(D_\beta) = \frac{1}{2} \int_{\mathcal{B}_\beta} \log\left(\frac{\mathsf{S}_X(\theta)}{\beta}\right) \mathrm{d}\theta \quad (50)$$

$$D_\beta = \int_{-1/2}^{1/2} \min\{\mathsf{S}_X(\theta), \beta\} \mathrm{d}\theta \quad (51)$$

for $\beta > 0$, where $\mathcal{B}_\beta \triangleq \{\theta \in [-1/2, 1/2] \colon \mathsf{S}_X(\theta) > \beta\}$. Indeed, when $\lambda(\mathcal{B}_0)$ is zero, we have $d(\{X_t\}) = 0$ since in this case the process $\{X_t\}$ has zero variance and, hence, the entropy rate of the quantized process $\{[X_t]_m\}$ is zero, too. When $\lambda(\mathcal{B}_0)$ is strictly positive, the distortion D_β can be bounded as

$$\beta \lambda(\mathcal{B}_\beta) \leq D_\beta \leq \beta \lambda(\mathcal{B}_0). \quad (52)$$

It follows by the continuity of the Lebesgue measure that $D_\beta / \beta \to \lambda(\mathcal{B}_0)$ as $\beta \to 0$. Consequently, $D_\beta \to 0$ if, and only if, $\beta \to 0$ and the rate-distortion dimension can be written as

$$\begin{aligned}
\dim_R(\{X_t\}) &= \lim_{\beta \downarrow 0} \frac{\int_{\mathcal{B}_\beta} \log\left(\frac{\mathsf{S}_X(\theta)}{\beta}\right) \mathrm{d}\theta}{-\log D_\beta} \\
&= \lim_{\beta \downarrow 0} \left\{ \lambda(\mathcal{B}_\beta) + \frac{\int_{\mathcal{B}_\beta} \log \mathsf{S}_X(\theta) \mathrm{d}\theta}{-\log \beta} \right\} \\
&= \lambda(\mathcal{B}_0) + \lim_{\beta \downarrow 0} \frac{\int_{\mathcal{B}_\beta} \log \mathsf{S}_X(\theta) \mathrm{d}\theta}{-\log \beta}. \quad (53)
\end{aligned}$$

By the continuity of the Lebesgue measure, for every $\varepsilon > 0$ there exists a $\beta' \in (0, 1)$ such that $\lambda(\mathcal{B}_{\beta'}) \geq \lambda(\mathcal{B}_0) - \varepsilon$. Since $\mathcal{B}_\beta \subseteq \mathcal{B}_0$, it follows that

$$\lambda(\mathcal{B}_\beta \setminus \mathcal{B}_{\beta'}) \leq \lambda(\mathcal{B}_0) - \lambda(\mathcal{B}_{\beta'}) \leq \varepsilon, \quad \beta < \beta'. \quad (54)$$

Thus, for every $0 < \beta < \beta' < 1$,

$$\begin{aligned}
\int_{\mathcal{B}_\beta} \log \mathsf{S}_X(\theta) \mathrm{d}\theta &= \int_{\mathcal{B}_\beta \setminus \mathcal{B}_{\beta'}} \log \mathsf{S}_X(\theta) \mathrm{d}\theta + \int_{\mathcal{B}_{\beta'}} \log \mathsf{S}_X(\theta) \mathrm{d}\theta \\
&\geq \lambda(\mathcal{B}_\beta \setminus \mathcal{B}_{\beta'}) \log \beta + \lambda(\mathcal{B}_{\beta'}) \log \beta' \\
&\geq \varepsilon \log \beta + \lambda(\mathcal{B}_{\beta'}) \log \beta'. \quad (55)
\end{aligned}$$

Dividing both sides of (55) by $-\log \beta$, and letting first β and then ε tend to zero, we obtain that the second term on the RHS of (53) is nonnegative. However, by assumption the process $\{X_t\}$ has finite variance, so its PSD S_X is integrable over $[-1/2, 1/2]$. Consequently, using the inequality $\log x \leq x - 1$ and the nonnegativity of S_X, we obtain that

$$\int_{\mathcal{B}_\beta} \log \mathsf{S}_X(\theta) \mathrm{d}\theta \leq \int_{-1/2}^{1/2} \mathsf{S}_X(\theta) \mathrm{d}\theta - 1. \quad (56)$$

Dividing both sides of (56) by $-\log \beta$, and letting β tend to zero, we obtain that the second term on the RHS of (53) is also nonpositive. We conclude that this term is zero, so (49) follows from (53) and Theorem 9.

We observe from Theorem 10 that the information dimension rate of a Gaussian process $\{\mathbf{X}_t\}$ depends only on the

derivative of its SDF $F_\mathbf{X}$, which coincides almost everywhere with the derivative of the absolutely-continuous part of $F_\mathbf{X}$. Indeed, any SDF $F_\mathbf{X}$ can be decomposed as [15, (4.3), p. 124]

$$F_\mathbf{X}(\theta) = F_{\mathbf{X},a}(\theta) + F_{\mathbf{X},d}(\theta) + F_{\mathbf{X},s}(\theta) \qquad (57)$$

where $F_{\mathbf{X},a}$ is *absolutely continuous* w.r.t. the Lebesgue measure, $F_{\mathbf{X},d}$ is *discrete*, and $F_{\mathbf{X},s}$ is *singular*. Furthermore, $F'_\mathbf{X}(\theta) = F'_{\mathbf{X},a}(\theta)$ almost everywhere [15, Sec. 4]. Consequently, the information dimension rate of a Gaussian process depends only on the absolutely-continuous part of its SDF. By combining (57) with Theorem 10 and Lemma 8, we can show that the same is true for non-Gaussian processes.

Corollary 12: Let $\{\mathbf{X}_t\}$ be a stationary, L-variate, real-valued process with SDF $F_\mathbf{X}$, and let $\{\mathbf{X}_{t,a}\}$ be a stationary, L-variate, real-valued process with SDF $F_{\mathbf{X},a}$, where $F_{\mathbf{X},a}$ is the absolutely-continuous part of $F_\mathbf{X}$, cf. (57). Then

$$\overline{d}(\{\mathbf{X}_t\}) = \overline{d}(\{\mathbf{X}_{t,a}\}). \qquad (58)$$

Proof: Combining the decomposition (57) with the spectral representation of stationary processes [16, Sec. 4.11], it can be shown that every stationary process can be written as

$$\mathbf{X}_t = \mathbf{X}_{t,a} + \mathbf{X}_{t,d} + \mathbf{X}_{t,s}, \quad t \in \mathbb{Z} \qquad (59)$$

where $\{\mathbf{X}_{t,a}\}$, $\{\mathbf{X}_{t,d}\}$, and $\{\mathbf{X}_{t,s}\}$ are stationary, mutually uncorrelated, stochastic processes with the respective SDFs $F_{\mathbf{X},a}$, $F_{\mathbf{X},d}$, and $F_{\mathbf{X},s}$; see [16, p. 758] and references therein. Since $F'_{\mathbf{X},d}$ and $F'_{\mathbf{X},s}$ are zero almost everywhere [15, Sec. 4], we obtain from Theorem 10 and the nonnegativity of the information dimension rate (Lemma 6) that

$$d(\{\mathbf{X}_{t,d}\}) = d(\{\mathbf{X}_{t,s}\}) = 0. \qquad (60)$$

Corollary 12 follows by applying Lemma 8 first together with (60) to show that

$$d(\{\mathbf{X}_{t,d} + \mathbf{X}_{t,s}\}) = 0 \qquad (61)$$

and then together with (61) to show that

$$\overline{d}(\{\mathbf{X}_{t,a} + \mathbf{X}_{t,d} + \mathbf{X}_{t,s}\}) = \overline{d}(\{\mathbf{X}_{t,a}\}). \qquad (62)$$

∎

D. Information Dimension Rate of Complex-Valued Processes

So far, we have considered real-valued stochastic processes. However, every complex-valued RV can be written as a two-dimensional, real-valued, random vector, so the previous results directly generalize to the complex case. In particular, one can define the information dimension rate of the L-variate, complex-valued process $\{\mathbf{Z}_t\}$ as the information dimension rate of the $(2L)$-variate, real-valued process $\{\hat{\mathbf{X}}_t\}$ that follows by stacking the real part of $\mathbf{Z}_t$ on top of the imaginary part of $\mathbf{Z}_t$.

Let $\{\mathbf{Z}_t\}$ be a stationary, L-variate, complex-valued process with mean vector $\boldsymbol{\mu}$ and matrix-valued SDF $F_\mathbf{Z}$, i.e.,

$$K_\mathbf{Z}(\tau) = \int_{-1/2}^{1/2} e^{-\imath 2\pi\tau\theta} \mathrm{d}F_\mathbf{Z}(\theta), \quad \tau \in \mathbb{Z} \qquad (63)$$

where

$$K_\mathbf{Z}(\tau) \triangleq \mathsf{E}\left[(\mathbf{Z}_{t+\tau} - \boldsymbol{\mu})(\mathbf{Z}_t - \boldsymbol{\mu})^\mathsf{H}\right], \quad \tau \in \mathbb{Z} \qquad (64)$$

is the autocovariance function. We say that a stationary, L-variate, complex-valued process $\{\mathbf{Z}_t\}$ is *proper* if it has finite variance, its mean vector is the zero vector, and its pseudo-autocovariance function satisfies

$$\overline{K}_\mathbf{Z}(\tau) \triangleq \mathsf{E}\left[\mathbf{Z}_{t+\tau}\mathbf{Z}_t^\mathsf{T}\right] = 0, \quad \tau \in \mathbb{Z}. \qquad (65)$$

The following result generalizes Theorem 10 to complex-valued stochastic processes.

Theorem 13: Let $\{\mathbf{Z}_t\}$ be a stationary, L-variate, complex-valued process with matrix-valued SDF $F_\mathbf{Z}$. Then,

$$\underline{d}(\{\mathbf{Z}_t\}) \leq 2 \int_{-1/2}^{1/2} \mathrm{rank}(F'_\mathbf{Z}(\theta))\mathrm{d}\theta \qquad (66)$$

with equality if $\{\mathbf{Z}_t\}$ is Gaussian and proper.

Proof: See Appendix G. ∎

Note that neither Gaussianity nor properness is sufficient for equality in Theorem 13. Conversely, Gaussianity and properness are not necessary for equality. For example, any univariate stationary Gaussian process achieves (66) with equality if its real and imaginary components are independent and if the derivatives of their SDFs have matching support.

V. ANOTHER DEFINITION OF INFORMATION DIMENSION

Jalali and Poor [7] proposed a different definition for the information dimension of a univariate stochastic process. We shall refer to this information dimension as the *block-average information dimension* and denote it by $d'(\{X_t\})$. In this section, we discuss scenarios in which the information dimension rate (Definition 2) coincides with and differs from the block-average information dimension. For ease of exposition, in this section we follow [7] and restrict our attention to univariate real-valued processes.

The following definition for the information dimension of stochastic processes was proposed in [7].

Definition 4: The *block-average information dimension* of the stochastic process $\{X_t\}$ is defined as

$$d'(\{X_t\}) \triangleq \lim_{k\to\infty} \lim_{m\to\infty} \frac{H([X_k]_m|[X_1^{k-1}]_m)}{\log m} \qquad (67)$$

provided the limits exist. If the limits do not exist, then one can define the upper and lower block-average information dimension $\overline{d}'(\{X_t\})$ and $\underline{d}'(\{X_t\})$ by replacing the limits by limits superior and limits inferior, respectively.

In the following, we restrict ourselves to stationary processes, in which case the limit over k in (67) is guaranteed to exist. We refer to $d'(\{X_t\})$ as the block-average information dimension because it was shown in [7, Lemma 3] that, if $\{X_t\}$ is stationary and the information dimension $d(X_1^k)$ exists for every k, then

$$d'(\{X_t\}) = \lim_{k\to\infty} \frac{d(X_1^k)}{k}. \qquad (68)$$

If $d(X_1^k)$ does not exist, then the proof of [7, Lemma 3] reveals that

$$\underline{d}'(\{X_t\}) \leq \lim_{k\to\infty} \frac{\underline{d}(X_1^k)}{k} \leq \lim_{k\to\infty} \frac{\overline{d}(X_1^k)}{k} = \overline{d}'(\{X_t\}). \qquad (69)$$

Since conditioning reduces entropy, it follows immediately that

$$\underline{d}'(\{X_t\}) \leq \overline{d}(X_t). \tag{70}$$

Thus, like the information dimension rate, the block-average information dimension of the stochastic process $\{X_t\}$ cannot exceed the information dimension of the marginal RV X_t.

While the entropy rate $H'(\{X_t\})$ of a stationary process $\{X_t\}$ can alternatively be written as the conditional entropy of X_1 given $X_{-\infty}^0$, cf. (4), the block-average information dimension $d'(\{X_t\})$ does, in general, not permit a similar expression. In fact, let

$$d(X_1|X_{-\infty}^0) \triangleq \lim_{m \to \infty} \lim_{k \to \infty} \frac{H([X_1]_m|X_{-k}^0)}{\log m} \tag{71}$$

provided the limit over m exists. (Since conditioning reduces entropy, the limit over k always exists.) The upper and lower information dimensions $\overline{d}(X_1|X_{-\infty}^0)$ and $\underline{d}(X_1|X_{-\infty}^0)$ are defined analogously by replacing the limit over m by the limit superior and limit inferior, respectively. Then, we have that

$$\overline{d}(X_1|X_{-\infty}^0) \leq \underline{d}'(\{X_t\}) \tag{72}$$

where the inequality can be strict; see Theorem 14 and Example 4 below.

A. Block-Average Information Dimension vs. Information Dimension Rate

We next demonstrate that, for ψ^*-mixing processes, the information dimension rate $d(\{X_t\})$ coincides with the block-average information dimension $d'(\{X_t\})$. However, in general the two definitions do not coincide, but there exists an ordering between them.

Theorem 14: Let $\{X_t\}$ be a stationary process. Then,

$$\overline{d}(X_1|X_{-\infty}^0) \leq \overline{d}(\{X_t\}) \leq \overline{d}'(\{X_t\}). \tag{73}$$

Moreover,

$$\lim_{k \to \infty} \overline{\lim_{m \to \infty}} \frac{I([X_k]_m; [X_{-\infty}^0]_m | [X_1^{k-1}]_m)}{\log m}$$
$$\leq \overline{d}'(\{X_t\}) - \overline{d}(\{X_t\})$$
$$\leq \lim_{k \to \infty} \overline{\lim_{m \to \infty}} \frac{I([X_k]_m; [X_{-\infty}^0]_m | [X_1^{k-1}]_m)}{\log m} \tag{74}$$

where the limits over k exist because, by the stationarity of $\{X_t\}$, the mutual information $I([X_k]_m; [X_{-\infty}^0]_m | [X_1^{k-1}]_m)$ is monotonically decreasing in k.

Proof: See Appendix H-A. ∎

The inequalities in (74) imply that, if the limits over m exist, then

$$\lim_{k \to \infty} \lim_{m \to \infty} \frac{I([X_k]_m; [X_{-\infty}^0]_m | [X_1^{k-1}]_m)}{\log m} = 0 \tag{75}$$

is a necessary and sufficient condition for the equality of $\overline{d}(\{X_t\})$ and $\overline{d}'(\{X_t\})$. Note that, for every $m = 2, 3, \ldots$, we have [17, eq. (8.9)]

$$\lim_{k \to \infty} I([X_k]_m; [X_{-\infty}^0]_m | [X_1^{k-1}]_m) = 0. \tag{76}$$

Thus, (75) is satisfied for processes $\{X_t\}$ that allow us to change the order of taking limits as k and m tend to infinity. However, in general (75) is difficult to check. We next present a sufficient condition that is easier to verify.

Corollary 15: Let $\{X_t\}$ be a stationary process. Assume that there exists a nonnegative integer n such that

$$I(X_1^k; X_{-\infty}^{-n}) < \infty, \quad k = 1, 2, \ldots \tag{77}$$

Then, $\overline{d}(\{X_t\}) = \overline{d}'(\{X_t\})$.

Proof: See Appendix H-B. ∎

Condition (77) holds for ψ^*-mixing processes. Indeed, since every ψ^*-mixing process satisfies (6), it follows that one can find an n such that $I(X_1^\infty; X_{-\infty}^{-n}) < \infty$. The condition (77) holds then by the data processing inequality.

If (77) holds for $n = 0$, then we even have that

$$\overline{d}(X_1|X_{-\infty}^0) = \overline{d}(\{X_t\}) = \overline{d}'(\{X_t\}) = \overline{d}(X_t). \tag{78}$$

Thus, in this case all presented generalizations of information dimension to stochastic processes coincide with the information dimension of the marginal RV. To prove (78), we note that (77) with $n = 0$ gives

$$I(X_1; X_{-\infty}^0) < \infty. \tag{79}$$

It then follows by the data processing inequality that

$$I([X_1]_m; X_{-\infty}^0) \leq I(X_1; X_{-\infty}^0) < \infty. \tag{80}$$

Consequently,

$$d(X_1|X_{-\infty}^0) = \lim_{m \to \infty} \frac{H([X_1]_m) - I([X_1]_m; X_{-\infty}^0)}{\log m}$$
$$= d(X_1) \tag{81}$$

if the limit exists. In general, we have $\overline{d}(X_1|X_{-\infty}^0) = \overline{d}(X_t)$. The claim (78) follows then by (73) and because, by (70), $\overline{d}'(\{X_t\}) \leq \overline{d}(X_t)$.

Condition (77) with $n = 0$ is satisfied, for example, if $\{X_t\}$ is a sequence of i.i.d. RVs, if it is a discrete-valued stochastic process with finite marginal entropy, or if it is a continuous-valued stochastic process with finite marginal differential entropy and finite differential entropy rate.

In the following, we present two examples of processes $\{X_t\}$ for which $\overline{d}(X_1|X_{-\infty}^0) = \overline{d}(\{X_t\}) = \overline{d}'(\{X_t\})$. As we shall argue, neither of these examples satisfies (77), hence (77) is sufficient but not necessary.

Example 2: Let $\{B_t\}$ be a sequence of i.i.d. Bernoulli-ρ RVs, i.e., $P_{B_t}(0) = 1 - P_{B_t}(1) = \rho$, and let $\{Y_t\}$ be a sequence of i.i.d. RVs with PDF f_Y supported on $[0, 1]$ and finite differential entropy. By (13), we thus have that $d(Y_t) = 1$ for every t. We define the stochastic process $\{X_t\}$ as

$$X_t = B_t X_{t-1} + (1 - B_t) Y_t, \quad t \in \mathbb{Z} \tag{82}$$

and assume that X_t has the same marginal distribution as Y_t. Note that $\{X_t\}$ is first-order Markov, so

$$d(X_1|X_{-\infty}^0) = d(X_1|X_0) = \rho d(Y) = \rho. \tag{83}$$

Furthermore, [7, Th. 3] demonstrates that $d'(\{X_t\}) = \rho$. Thus, together with (73), this yields that

$$d(X_1|X_{-\infty}^0) = d(\{X_t\}) = d'(\{X_t\}) = \rho. \tag{84}$$

The stochastic process $\{X_t\}$, as defined by (82), satisfies (75) but not (77). Indeed, for every nonnegative integer n, we have $I(X_1; X_{-n}) = \infty$, since X_1 has finite differential entropy and the event $X_1 = X_{-n}$ has positive probability. It follows that $I(X_1^k; X_{-\infty}^{-n}) = \infty$ for every k and n, so (77) is violated. In contrast, we have

$$\varlimsup_{m \to \infty} \frac{I([X_k]_m; [X_{-\infty}^0]_m | [X_1^{k-1}]_m)}{\log m}$$
$$= \varlimsup_{m \to \infty} \frac{I([X_k]_m; [X_{-\infty}^0]_m | [X_1^{k-1}]_m, B_k)}{\log m} \quad (85)$$

since conditioning on the binary random variable B_k changes mutual information by at most one bit. If $B_k = 1$, then $[X_k]_m = [X_{k-1}]_m$; if $B_k = 0$, then $[X_k]_m = [Y_k]_m$, which is independent of $[X_{-\infty}^0]_m$. In both cases, the conditional mutual information between $[X_k]_m$ and $[X_{-\infty}^0]_m$ given $[X_1^{k-1}]_m$ is zero, so (75) is satisfied.

Example 3: Let the process $\{\tilde{X}_t\}$ be periodic with period $P \in \mathbb{N}$ and have finite marginal differential entropy. Further let Δ be uniformly distributed on $\{0, \dots, P-1\}$. Then, the shifted process $\{X_t\}$, defined by

$$X_t = \tilde{X}_{t+\Delta}, \quad t \in \mathbb{Z} \quad (86)$$

is stationary [18, Th. 10-5] and has finite marginal differential entropy. For every $k = P, P+1, \dots$ and $m = 2, 3, \dots$, we have that $H([X_1]_m | X_{-k+1}^0) = 0$ and $H([X_1^k]_m) = H([X_1^P]_m)$, hence

$$d(X_1 | X_{-\infty}^0) = d(\{X_t\}) = d'(\{X_t\}) = 0. \quad (87)$$

As in the previous example, the stochastic process $\{X_t\}$ satisfies (75) but not (77). Indeed, for every nonnegative integer n, we have $I(X_1^k; X_{-\infty}^{-n}) = \infty$ since X_1 has finite differential entropy and the process is periodic. In contrast, $[X_k]_m = [X_{k-P}]_m$, so the conditional mutual information between $[X_k]_m$ and $[X_{-\infty}^0]_m$ given $[X_1^{k-1}]_m$ is zero when $k = P+1, P+2 \dots$

In many cases, the inequalities in Theorem 14 can be strict. The following example shows such a strict inequality for the class of stationary Gaussian processes $\{X_t\}$ with PSD supported on a set of positive Lebesgue measure.[3]

Example 4: Let $\{X_t\}$ be a stationary Gaussian process with zero mean, variance σ^2, and PSD S_X having support $\mathcal{B}_0$. It follows from Theorem 10 that

$$d(\{X_t\}) = \lambda(\mathcal{B}_0). \quad (88)$$

We next argue that if $0 < \lambda(\mathcal{B}_0) < 1$ then $d'(\{X_t\}) = 1$ and $d(X_1 | X_{-\infty}^0) = 0$. Consequently,

$$d(X_1 | X_{-\infty}^0) < d(\{X_t\}) < d'(\{X_t\}). \quad (89)$$

To show that $d'(\{X_t\}) = 1$, we note that

$$\underline{d}'(\{X_t\}) = \lim_{k \to \infty} \varliminf_{m \to \infty} \frac{H([X_k]_m | [X_1^{k-1}]_m)}{k \log m}$$
$$\geq \lim_{k \to \infty} \varliminf_{m \to \infty} \frac{H([X_0]_m | X_{-k}^{-1})}{\log m} \quad (90)$$

[3]The assumption that $\{X_t\}$ has a PSD is made for notational convenience and is not essential. All steps in Example 4 continue to hold if we replace S_X by the derivative of the SDF F_X.

where the inequality follows by the stationarity of $\{X_t\}$; because conditioning reduces entropy; and because, conditioned on X_{-k}^{-1}, $[X_0]_m$ is independent of $[X_{-k}^{-1}]_m$. Since $\{X_t\}$ is Gaussian, it follows that, conditioned on X_{-k}^{-1}, the RV X_0 is Gaussian with mean $\mathsf{E}[X_0 | X_{-k}^{-1}]$ and variance σ_k^2, which is independent of X_{-k}^{-1}. It can be further shown that if $\lambda(\mathcal{B}_0) > 0$, then $\sigma_k^2 > 0$ for every finite k (see Lemma 16 below). It follows that, conditioned on any $X_{-k}^{-1} = x_{-k}^{-1}$, the RV X_0 has a PDF, so by (13)

$$\lim_{m \to \infty} \frac{H([X_0]_m \mid X_{-k}^{-1} = x_{-k}^{-1})}{\log m} = 1, \quad k = 1, 2, \dots \quad (91)$$

Together with Fatou's lemma, this shows that the RHS of (90) is 1, hence $d'(\{X_t\}) = 1$.

To demonstrate that $d(X_1 | X_{-\infty}^0) = 0$, we note that $\lambda(\mathcal{B}_0) < 1$ implies that

$$\int_{-1/2}^{1/2} \log \mathsf{S}_X(\theta) \mathrm{d}\theta = -\infty. \quad (92)$$

This is a necessary and sufficient condition for $\sigma_k^2 \to 0$ as $k \to \infty$; see, e.g., [19, Sec. 10.6]. Intuitively, the fact that $\sigma_k^2 \to 0$ implies that the conditional distribution of X_1 given $X_{-\infty}^0$ is almost surely degenerate, hence $d(X_1 | X_{-\infty}^0) = 0$. To prove this rigorously, we apply [13, Lemma 30] together with the fact that conditioning reduces entropy to upper-bound

$$H([X_1]_m | X_{-k}^0) \leq H([X_1 - \mathsf{E}[X_1 | X_{-k}^0]]_m) + \log 2. \quad (93)$$

Expressing $X_1 - \mathsf{E}[X_1 | X_{-k}^0]$ as $\sigma_{k+1} Z$, where Z is zero-mean, unit-variance Gaussian, the RHS of (93) can be written as $H(\lfloor m\sigma_{k+1} Z \rfloor) + \log 2$. Since $\sigma_k^2 \to 0$ as $k \to \infty$, we obtain from [3, Lemma 1] that

$$\lim_{k \to \infty} H([X_1]_m | X_{-k}^0) \leq \log 2, \quad m = 2, 3, \dots \quad (94)$$

Consequently, the claim follows from the definition of $d(X_1 | X_{-\infty}^0)$.

Lemma 16: Let $\{X_t\}$ be a stationary, univariate, real-valued, Gaussian process with zero mean, variance σ^2, and SDF F_X. Suppose that $\sigma_k^2 = 0$ for some finite k. Then,

$$\lambda(\{\theta : \mathsf{F}_X'(\theta) > 0\}) = 0. \quad (95)$$

Proof: See Appendix H-C. ∎

B. Block-Average Information Dimension vs. Rate-Distortion Dimension

The connection between the block-average information dimension and the rate-distortion dimension of a stochastic process was studied in [8]. The equivalence between the rate-distortion dimension and the information dimension [2, Prop. 3.3] directly implies that

$$\overline{d}'(\{X_t\}) = 2 \lim_{k \to \infty} \varlimsup_{D \to 0} \frac{R(X_1^k, kD)}{-k \log D} \quad (96)$$

Rezagah *et al.* [8] demonstrated that the order of the limits on the RHS of (96) can be exchanged. More precisely, [8, Th. 2] states that if $\lim_{D \to 0} \frac{R(X_1^k, kD)}{-k \log D}$ exists for all k, then

$$\mathsf{dim}_R(\{X_t\}) = \overline{d}'(\{X_t\}). \quad (97)$$

This may appear as a contradiction to our results, since we demonstrate in Theorem 9 that $\overline{\dim}_R(\{X_t\}) = \overline{d}(\{X_t\})$, and Example 4 demonstrates that there are stochastic processes for which $\overline{d}(\{X_t\}) < \overline{d}'(\{X_t\})$. However, the proof of (97) relies on the fact that [8, Sec. VI-E]

$$\left| \frac{1}{k} R(X_1^k, kD) - \lim_{\kappa \to \infty} \frac{1}{\kappa} R(X_1^\kappa, \kappa D) \right| \le \frac{1}{k} I(X_1^k; X_{-\infty}^0) \quad (98)$$

and that the RHS of (98) vanishes as $k \to \infty$. If (79) holds, then this is indeed the case; see [17, eqs. (8.6)–(8.10)]. As shown in Corollary 15, in this case we also have that $\overline{d}'(\{X_t\}) = \overline{d}(\{X_t\})$. In fact, as discussed after Corollary 15, in this case all presented generalizations of information dimension to stochastic processes coincide with the information dimension of the marginal RV. In contrast, if (79) does not hold then, by the data processing inequality, the RHS of (98) is infinite. This is, for example, the case if $\{X_t\}$ is a stationary process with positive variance and a PSD that is zero on a set of positive Lebesgue measure, since for such processes the differential entropy $h(X_1|X_{-\infty}^0)$ is $-\infty$. Our proof of Theorem 9 does not rely on (98). We thus conclude that $\overline{\dim}_R(\{X_t\}) = \overline{d}(\{X_t\})$ for every stochastic process $\{X_t\}$, but that $\overline{\dim}_R(\{X_t\}) = \overline{d}'(\{X_t\})$ only for those processes for which $\overline{d}'(\{X_t\}) = \overline{d}(\{X_t\})$.

VI. Operational Characterizations

Information dimension was recently given an operational characterization in almost lossless data compression [4]. Specifically, Wu and Verdú defined the *minimum ϵ-achievable rate* $R(\epsilon)$ to be the minimum of $R > 0$ such that there exists a sequence of encoders $f_k \colon \mathbb{R}^k \to \mathbb{R}^{\lfloor Rk \rfloor}$ and decoders $g_k \colon \mathbb{R}^{\lfloor Rk \rfloor} \to \mathbb{R}^k$ satisfying [4, Def. 4]

$$\mathbb{P}[g_k(f_k(X_1^k)) \ne X_1^k] \le \epsilon \quad (99)$$

for all k sufficiently large. As argued in [4, Sec. IV-B], if we impose no restrictions on f_k and g_k, then zero rate is achievable even for $\epsilon = 0$, since the cardinality of $\mathbb{R}^k$ is the same for any k. However, if we restrict ourselves either to encoders f_k that are linear or to decoders g_k that are Lipschitz continuous, then the minimum ϵ-achievable rate for collections of i.i.d. RVs X_1^k with a discrete-continuous mixed distribution, i.e., a distribution of the form (12), is given by

$$R(\epsilon) = d(X). \quad (100)$$

Thus, for such RVs, information dimension has an operational characterization.

For stochastic processes $\{X_t\}$, Wu and Verdú further demonstrated that the minimum ϵ-achievable rate, achievable with Lipschitz-continuous decoders g_k, can be lower-bounded as [20, Remark 4]

$$R(\epsilon) \ge \overline{d}'(\{X_t\}) - \epsilon. \quad (101)$$

To the best of our knowledge, for non-i.i.d. processes $\{X_t\}$, no matching achievability result exists for almost lossless data compression.

In contrast, for universal compressed sensing with linear encoding and decoding via Lagrangian minimum entropy

pursuit, it was shown by Jalali and Poor that $\overline{d}'(\{X_t\})$ is an achievable rate when $\{X_t\}$ is ψ^*-mixing:

Theorem 17 ([7, Th. 8]): Consider a ψ^*-mixing stationary process $\{X_t\}$ taking value in $[0,1]$ with upper block-average information dimension $\overline{d}'(\{X_t\})$. For each k, let the entries of the measurement matrix $A \in \mathbb{R}^{\ell \times k}$ be drawn i.i.d. according to a zero-mean, unit-variance, Gaussian distribution. Given X_1^k generated by $\{X_t\}$ and $(Y_1, \ldots, Y_\ell)^\intercal = A(X_1, \ldots, X_k)^\intercal$, let

$$\hat{\mathbf{X}} \triangleq \operatorname*{arg\,min}_{\mathbf{u} \in \mathcal{X}_m^k} \left\{ \hat{H}_j(\mathbf{u}) + \frac{\gamma}{k^2} \|A\mathbf{u} - (Y_1, \ldots, Y_\ell)^\intercal\|_2^2 \right\} \quad (102)$$

where $\mathcal{X}_m \triangleq \{[x]_{2^m} \colon x \in [0,1]\}$, $\hat{H}_j(\cdot)$ is the conditional empirical entropy [7, Def. 1], $m = \lceil r \log \log k \rceil$ (for $r > 1$), $j = o(\frac{\log k}{\log \log k})$, and $\gamma = (\log k)^{2r}$. If the number of measurements $\ell = \ell_k$ satisfies

$$\frac{\ell_k}{k} \ge (1 + \delta)\overline{d}'(\{X_t\}), \quad \text{for some arbitrary } \delta > 0 \quad (103)$$

then

$$\frac{1}{\sqrt{k}} \|\hat{\mathbf{X}} - (X_1, \ldots, X_k)^\intercal\|_2 \to 0 \quad \text{in probability} \quad (104)$$

as $k \to \infty$.

In words, Theorem 17 states that if the rate of random linear measurements of X_1^k is slightly larger than the block-average information dimension, then the Lagrangian relaxation of minimum entropy pursuit provides an asymptotically distortion-free estimate of X_1^k in terms of the Euclidean norm. Thus, for ψ^*-mixing processes, the block-average information dimension is an achievable rate for almost zero-distortion recovery.

We next discuss an operational characterization of the rate-distortion dimension. By Theorem 9, this is also an operational characterization of the information dimension rate. In [8], Rezagah *et al.* considered the almost zero-distortion recovery of stationary processes when the decoder employs *compressible signal pursuit (CSP)* optimization:

Theorem 18 ([8, Cor. 2]): Consider a stationary, real-valued process $\{X_t\}$ and a system of random linear observations $(Y_1, \ldots, Y_\ell)^\intercal = A(X_1, \ldots, X_k)^\intercal$ with measurement matrix $A \in \mathbb{R}^{\ell \times k}$ composed of i.i.d. zero-mean, unit-variance, Gaussian RVs. If the number of measurements $\ell = \ell_k$ satisfies

$$\lim_{k \to \infty} \frac{\ell_k}{k} > \overline{\dim}_R(\{X_t\}) \quad (105)$$

then there exists a family of compression codes such that

$$\frac{1}{\sqrt{k}} \|\hat{\mathbf{X}} - (X_1, \ldots, X_k)^\intercal\|_2 \to 0 \quad \text{in probability} \quad (106)$$

as $k \to \infty$, where $\hat{\mathbf{X}}$ is the solution of the CSP optimization

$$\hat{\mathbf{X}} = \operatorname*{arg\,min}_{\mathbf{u} \in \mathcal{C}_k} \|(Y_1, \ldots, Y_\ell)^\intercal - A\mathbf{u}\|_2 \quad (107)$$

and $\mathcal{C}_k$ denotes the codebook of the compression code.

In words, if the rate of random linear measurements of X_1^k is slightly larger than the rate-distortion dimension, then there exists a family of compression codes for which CSP optimization yields an asymptotically distortion-free estimate of X_1^k in terms of the Euclidean norm. Thus, the rate-distortion

dimension is an achievable rate for almost zero-distortion recovery.

To summarize, (101) demonstrates that $\bar{d}'(\{X_t\})$ yields a lower bound on the sampling rate required for almost lossless recovery with Lipschitz-continuous decoders. In contrast, Theorem 18 demonstrates that $\overline{\dim}_R(\{X_t\})$ (and hence also $\bar{d}(\{X_t\})$) is an achievable rate for almost zero-distortion recovery. Furthermore, as illustrated by Example 4, there are processes $\{X_t\}$ for which

$$d(\{X_t\}) = \dim_R(\{X_t\}) < d'(\{X_t\}). \tag{108}$$

Our results thus demonstrate that there exist stationary processes for which the sampling rate required for almost zero-distortion recovery is strictly smaller than the sampling rate required for almost lossless recovery with Lipschitz-continuous decoders. In other words, the fundamental limits of almost zero-distortion recovery and almost lossless recovery are different in general.

Comparing the lower bound (101) for almost lossless recovery with Theorem 18 for almost zero-distortion recovery, we observe that there are two main differences in the setup:

i) (101) is obtained for a Lipschitz-continuous decoder g_k, whereas Theorem 18 is based on CSP optimization;

ii) for almost lossless recovery, $\hat{\mathbf{X}} = g_k(f_k(X_1^k))$ is required to be exactly equal to X_1^k with high probability (cf. (99)), whereas for almost zero-distortion recovery it suffices that $\frac{1}{\sqrt{k}}\|\hat{\mathbf{X}} - (X_1, \ldots, X_k)^\intercal\|_2$ be small.

The following example presents a class of stationary processes for which almost zero-distortion recovery at rate $\bar{d}(\{X_t\})$ may also be achieved with linear encoders and decoders. This suggests that the second difference has greater impact.

Example 5: Let $\{X_t\}$ be a stationary, univariate, real-valued, Gaussian process possessing a PSD S_X with support $[-1/4, 1/4]$. By Theorem 10, we have that $d(\{X_t\}) = 1/2$. We next invoke the sampling theorem to demonstrate that there exist linear encoders $f_k\colon \mathbb{R}^k \to \mathbb{R}^{\ell_k}$ and decoders $g_k\colon \mathbb{R}^{\ell_k} \to \mathbb{R}^k$ such that

$$\lim_{k\to\infty} \frac{\ell_k}{k} = \frac{1}{2} \tag{109}$$

and

$$\frac{1}{\sqrt{k}}\|\hat{\mathbf{X}} - (X_1, \ldots, X_k)^\intercal\|_2 \to 0 \quad \text{in probability} \tag{110}$$

as $k \to \infty$, where $\hat{\mathbf{X}} = g_k(f_k(X_1^k))$.

To describe f_k and g_k, we divide the indices $t = 1, \ldots, k$ into three groups:

$$\mathcal{I}_1 \triangleq \{1, \ldots, \Delta_k\} \cup \{k - \Delta_k + 1, \ldots, k\} \tag{111}$$

$$\mathcal{I}_2 \triangleq \{2i\colon i \in \mathbb{Z}\} \cap \{\Delta_k + 1, \ldots, k - \Delta_k\} \tag{112}$$

$$\mathcal{I}_3 \triangleq \{2i + 1\colon i \in \mathbb{Z}\} \cap \{\Delta_k + 1, \ldots, k - \Delta_k\} \tag{113}$$

where $\{\Delta_k\}$ is an arbitrary sequence of even integers that tends to infinity sublinearly in k. The encoder f_k only reproduces the values of X_t with indices $t \in \mathcal{I}_1 \cup \mathcal{I}_2$, i.e., $f_k(X_1^k) = \{X_t, t \in \mathcal{I}_1 \cup \mathcal{I}_2\}$. Consequently,

$$\ell_k = 2\Delta_k + \left\lfloor \frac{k - 2\Delta_k}{2} \right\rfloor \tag{114}$$

and the rate ℓ_k/k converges to $\frac{1}{2}$ as $k \to \infty$.

We next show that we can find a decoder g_k for which (110) holds. Clearly, the values $\{X_t, t \in \mathcal{I}_1 \cup \mathcal{I}_2\}$ are directly observed. It therefore remains to estimate the missing values of X_1^k, which is done via the interpolation formula

$$\hat{X}_t = \sum_{i=-\Delta_k/2}^{\Delta_k/2} X_{2i+t-1} \frac{\sin\left(\pi\left[\frac{1}{2} - i\right]\right)}{\pi\left(\frac{1}{2} - i\right)}, \quad t \in \mathcal{I}_3. \tag{115}$$

It follows that

$$\begin{aligned}
&\mathsf{E}\left[\|(\hat{X}_1, \ldots, \hat{X}_k)^\intercal - (X_1, \ldots, X_k)^\intercal\|_2^2\right] \\
&= \sum_{t\in\mathcal{I}_3} \mathsf{E}\left[(\hat{X}_t - X_t)^2\right] \\
&= \left\lceil \frac{k - 2\Delta_k}{2} \right\rceil \mathsf{E}\left[(\hat{X}_{\Delta_k+1} - X_{\Delta_k+1})^2\right]
\end{aligned} \tag{116}$$

where the last step is due to stationarity. By the sampling theorem for stochastic processes, the expected value on the RHS of (116) vanishes as $\Delta_k \to \infty$ [21, Th. 1]. Thus, dividing both sides of (116) by k and letting $k \to \infty$ gives

$$\lim_{k\to\infty} \frac{1}{k}\mathsf{E}\left[\|(\hat{X}_1, \ldots, \hat{X}_k)^\intercal - (X_1, \ldots, X_k)^\intercal\|_2^2\right] = 0 \tag{117}$$

which together with Chebyshev's inequality [22, Th. 4.10.7] yields (110).

VII. Conclusions

Rényi [1] proposed the information dimension and the d-dimensional entropy to measure the information content of general RVs. His idea was to quantize the real-valued RV X by a uniform quantizer of step size $1/m$, and to then analyze the entropy of the quantized RV $[X]_m$ in the limit as m tends to infinity. His results demonstrate that any RV with positive information dimension has infinite information content. This is, e.g., the case for RVs whose probability measure has an absolutely-continuous part. The problem becomes even more interesting for stochastic processes $\{X_t\}$, since their information content is not only determined by the distribution of the marginals X_t, but also by their temporal dependence. For example, consider a stationary Gaussian process $\{X_t\}$ with bandlimited PSD. On the one hand, Gaussian processes have absolutely-continuous marginals, so one would expect that their information content is infinite. On the other hand, for processes with a bandlimited PSD, the present sample X_0 can be perfectly predicted from its infinite past $X_{-1}, X_{-2}, \ldots$ (see Example 4), which suggests that the information content of $\{X_t\}$ is zero.

To shed some light on such questions, we proposed a generalization of information dimension to stochastic processes by defining the information dimension rate $d(\{\mathbf{X}_t\})$ as the entropy rate $H'(\{[\mathbf{X}_t]_m\})$ divided by $\log m$ in the limit as $m \to \infty$. We demonstrated that the information dimension rate coincides with the rate-distortion dimension, defined as twice the pre-log factor of the rate-distortion function $R(D)$. We further showed that among all stationary process with PSD $S_\mathbf{X}$, the Gaussian process has the largest information dimension rate. This is consistent with the observation that Gaussian

processes are the hardest to predict, hence they are expected to have the largest information content. We then showed that the information dimension rate of stationary Gaussian processes is given by the average rank of $S_\mathbf{X}$, i.e.,

$$d(\{\mathbf{X}_t\}) = \int_{-1/2}^{1/2} \operatorname{rank}(S_\mathbf{X}(\theta)) d\theta. \qquad (118)$$

Specialized to the univariate case, this yields that the information dimension rate is given by the Lebesgue measure of the support of S_X, i.e.,

$$d(\{X_t\}) = \lambda(\{\theta : S_X(\theta) > 0\}). \qquad (119)$$

This agrees with the intuition that if the PSD of $\{X_t\}$ is zero on a set of positive Lebesgue measure, then some samples can be expressed in terms of the remaining samples and have therefore no information content. It further answers the above question whether stationary Gaussian processes with a bandlimited PSD have infinite information content in the positive, unless the PSD is zero almost everywhere.

An alternative definition for the information dimension of a stochastic process was proposed by Jalali and Poor [7] as the information dimension of X_1^k divided by k in the limit as $k \to \infty$. We referred to this quantity as the block-average information dimension $d'(\{X_t\})$. While $d(\{X_t\})$ and $d'(\{X_t\})$ coincide for ψ^*-mixing processes, in general we have that $d(\{X_t\}) \le d'(\{X_t\})$, where the inequality can be strict. In particular, as illustrated by Example 4, if the support of S_X of the Gaussian process $\{X_t\}$ has positive Lebesgue measure, then $d'(\{X_t\}) = 1$. Thus, in contrast to the information dimension rate, the block-average information dimension does not capture the dependence of the information dimension on the support size of S_X.

The essential difference between the definitions of $d(\{X_t\})$ and $d'(\{X_t\})$ is the order in which the limits over the quantization bin size $1/m$ and the block size k are taken. Rezagah *et al.* [8] showed that these limits can be exchanged if the process satisfies

$$I(X_1; X_{-\infty}^0) < \infty \qquad (120)$$

in which case $\dim_R(\{X_t\}) = d'(\{X_t\})$. However, in this case the information dimension of the stochastic process $\{X_t\}$ coincides with the information dimension of the marginal RV X_t. In other words, for such processes a generalization of information dimension to stochastic processes is redundant. In contrast, we showed in Theorem 9 that, for any stochastic process $\{X_t\}$, the information dimension rate $d(\{X_t\})$ coincides with the rate-distortion dimension $\dim_R(\{X_t\})$. This implies that $d'(\{X_t\})$ coincides with $\dim_R(\{X_t\})$ only for those stochastic processes for which $d'(\{X_t\}) = d(\{X_t\})$.

The equivalence between the information dimension rate $d(\{X_t\})$ and the rate-distortion dimension $\dim_R(\{X_t\})$ implies that $d(\{X_t\})$ inherits the operational characterizations of $\dim_R(\{X_t\})$. For example, it was demonstrated in [8] that $\dim_R(\{X_t\})$ is an achievable rate for almost zero-distortion recovery. In contrast, [20] shows that $d'(\{X_t\})$ is a lower bound on the minimum ϵ-achievable rate, achievable with Lipschitz-continuous decoders. By demonstrating that there are processes for which

$$d(\{X_t\}) = \dim_R(\{X_t\}) < d'(\{X_t\}) \qquad (121)$$

our results show that the fundamental limits of almost zero-distortion recovery and almost lossless recovery are different in general. Jalali and Poor [7] further showed that $d'(\{X_t\})$ is an achievable rate for universal lossless compressed sensing with linear encoding and decoding via Lagrangian minimum entropy pursuit when $\{X_t\}$ is ψ^*-mixing. Since for ψ^*-mixing processes we have $d(\{X_t\}) = d'(\{X_t\})$, our definition also inherits this operational characterization.

APPENDIX A
APPENDIX TO SECTION III

A. Proof of Lemma 2

The first inequality in (14), namely,

$$\int \varliminf_{m \to \infty} \frac{H([X]_m | Y = y)}{\log m} dP_Y(y)$$
$$\le \varliminf_{m \to \infty} \int \frac{H([X]_m | Y = y)}{\log m} dP_Y(y) \quad (122)$$

follows directly from Fatou's lemma [22, Th. 1.6.8, p. 50]. The second inequality follows because the limit inferior is upper-bounded by the limit superior. For the third inequality, note that for every $m = 2, 3, \dots$ and $Y = y$ [1, eq. (11)]

$$\frac{H([X]_m | Y = y)}{\log m} \le \frac{H([X]_1 | Y = y)}{\log 2} + 1. \qquad (123)$$

Furthermore, since conditioning reduces entropy, we have

$$\int \left(\frac{H([X]_1 | Y = y)}{\log 2} + 1 \right) dP_Y(y) = \frac{H([X]_1 | Y)}{\log 2} + 1$$
$$\le \frac{H([X]_1)}{\log 2} + 1$$
$$< \infty \qquad (124)$$

for every $m = 2, 3, \dots$ Hence, the RHS of (123) is integrable, and the third inequality in (14) follows again by Fatou's lemma.

B. Proof of Lemma 4

If $H([X_1^k]_1) = \infty$, then we have $d(X_1^k) = \infty$ and the rightmost inequality in (17) holds trivially. Moreover, in this case $H([X_t]_1) = \infty$ for at least one t, so for this t we also have $\bar{d}(X_t) = \infty$. Thus, also the left-most inequality holds.

If $H([X_1^k]_1) < \infty$, then we have

$$H([X_t]_1 | X_1^{t-1}) \le H([X_t]_1) < \infty, \quad t = 1, \dots, k \qquad (125)$$

hence the upper information dimensions are finite. It follows by the chain rule of entropy and because conditioning reduces entropy that

$$
\begin{aligned}
\bar{d}(X_1^k) &= \varlimsup_{m\to\infty} \sum_{t=1}^{k} \frac{H([X_t]_m | [X_1^{t-1}]_m)}{\log m} \\
&\leq \sum_{t=1}^{k} \varlimsup_{m\to\infty} \frac{H([X_t]_m)}{\log m} \\
&= \sum_{i=t}^{k} \bar{d}(X_t).
\end{aligned} \tag{126}
$$

Likewise, we have

$$
\begin{aligned}
\underline{d}(X_1^k) &= \varliminf_{m\to\infty} \sum_{t=1}^{k} \frac{H([X_t]_m | [X_1^{t-1}]_m)}{\log m} \\
&\geq \sum_{t=1}^{k} \varliminf_{m\to\infty} \frac{H([X_t]_m | X_1^{t-1})}{\log m} \\
&= \sum_{t=1}^{k} \underline{d}(X_t | X_1^{t-1})
\end{aligned} \tag{127}
$$

where the inequality follows because conditioning reduces entropy and because, conditioned on X_1^{t-1}, $[X_t]_m$ is independent of $[X_1^{t-1}]_m$.

C. Proof of Theorem 5

To simplify notation, we shall write collections of RVs as vectors, namely, $\mathbf{X} = (X_1, \ldots, X_k)^\mathsf{T}$ and $\mathbf{Y} = (Y_1, \ldots, Y_\ell)^\mathsf{T}$. The proof of Theorem 5 is based on the following lemma.

Lemma 19: Let $\mathbf{X}$ and $\mathbf{Y}$ be k- and ℓ-dimensional, jointly Gaussian vectors with mean vectors $\boldsymbol{\mu}_{\mathbf{X}}$ and $\boldsymbol{\mu}_{\mathbf{Y}}$ and joint covariance matrix $C_{\mathbf{X},\mathbf{Y}}$. Then, there exists a $k \times \ell$ matrix T and a length-k vector $\boldsymbol{\mu}$ such that $\mathsf{E}[\mathbf{X}|\mathbf{Y}] = \boldsymbol{\mu} + T\mathbf{Y}$. Moreover, $\mathbf{E} \triangleq \mathbf{X} - \boldsymbol{\mu} - T\mathbf{Y}$ has zero mean, is uncorrelated with $\mathbf{Y}$, and satisfies $C_{\mathbf{E}} = C_{\mathbf{X}} - TC_{\mathbf{Y}}T^\mathsf{T}$.

Proof: If $\mathbf{X}$ and $\mathbf{Y}$ are jointly Gaussian, then $\mathbf{X}$ can be written as a linear transformation of $\mathbf{Y}$ and an uncorrelated error. This follows from the fact that, for jointly Gaussian $\mathbf{X}$ and $\mathbf{Y}$, the linear minimum mean-square error (LMMSE) estimator of $\mathbf{X}$ given $\mathbf{Y}$ always exists and is given by $\mathsf{E}[\mathbf{X}|\mathbf{Y}] = \boldsymbol{\mu} + T\mathbf{Y}$. The result that $\mathbf{E}$ has zero mean, is uncorrelated with $\mathbf{Y}$, and satisfies $C_{\mathbf{E}} = C_{\mathbf{X}} - TC_{\mathbf{Y}}T^\mathsf{T}$ follows by direct calculation. ∎

Since information dimension is translation invariant, it follows that

$$
\begin{aligned}
d(\mathbf{X}|\mathbf{Y}=\mathbf{y}) &= d(\boldsymbol{\mu} + T\mathbf{Y} + \mathbf{E}|\mathbf{Y}=\mathbf{y}) \\
&= d(\mathbf{E}|\mathbf{Y}=\mathbf{y}).
\end{aligned} \tag{128}
$$

Furthermore, since $\mathbf{X}$ and $\mathbf{Y}$ are jointly Gaussian, so are $\mathbf{Y}$ and $\mathbf{E}$, and from the fact that they are uncorrelated follows that they are independent. Thus,

$$
d(\mathbf{E}|\mathbf{Y}=\mathbf{y}) = d(\mathbf{E}) = \operatorname{rank}(C_{\mathbf{E}}) \tag{129}
$$

where $C_{\mathbf{E}}$ is the covariance matrix of $\mathbf{E}$. The identities (128) and (129) hold for every $\mathbf{y}$, so it follows from Lemma 2 that

$d(\mathbf{X}|\mathbf{Y}) = \operatorname{rank}(C_{\mathbf{E}})$. It remains to show that $C_{\mathbf{E}}$ is the generalized Schur complement of $C_{\mathbf{Y}}$ in $C_{\mathbf{X},\mathbf{Y}}$. Indeed, by [12, 7.1.P28] there exists a matrix W such that $C_{\mathbf{YX}} = C_{\mathbf{Y}}W$. The generalized Schur complement of $C_{\mathbf{Y}}$ in $C_{\mathbf{X},\mathbf{Y}}$ is then given by

$$
C_{\mathbf{X}|\mathbf{Y}} = C_{\mathbf{X}} - W^\mathsf{H} C_{\mathbf{Y}} W. \tag{130}
$$

Comparing (130) with the expression of $C_{\mathbf{E}}$ given in Lemma 19, we observe that $C_{\mathbf{E}} = C_{\mathbf{X}|\mathbf{Y}}$ if the matrix T in Lemma 19 satisfies $C_{\mathbf{YX}} = C_{\mathbf{Y}}T^\mathsf{T}$. This is indeed the case: since $\mathbf{X} = \boldsymbol{\mu} + T\mathbf{Y} + \mathbf{E}$, and since $\mathbf{Y}$ and $\mathbf{E}$ are uncorrelated, we have that

$$
\begin{aligned}
C_{\mathbf{YX}} &= \mathsf{E}[\mathbf{YX}^\mathsf{T}] - \mathsf{E}[\mathbf{Y}]\mathsf{E}[\mathbf{X}^\mathsf{T}] \\
&= \mathsf{E}[\mathbf{Y}(\boldsymbol{\mu} + T\mathbf{Y} + \mathbf{E})^\mathsf{T}] - \mathsf{E}[\mathbf{Y}]\mathsf{E}[(\boldsymbol{\mu} + T\mathbf{Y} + \mathbf{E})^\mathsf{T}] \\
&= \mathsf{E}[\mathbf{Y}]\boldsymbol{\mu}^\mathsf{T} + \mathsf{E}[\mathbf{YY}^\mathsf{T}]T^\mathsf{T} - \mathsf{E}[\mathbf{Y}]\boldsymbol{\mu}^\mathsf{T} - \mathsf{E}[\mathbf{Y}]\mathsf{E}[\mathbf{Y}^\mathsf{T}]T^\mathsf{T} \\
&= C_{\mathbf{Y}}T^\mathsf{T}.
\end{aligned} \tag{131}
$$

This proves Theorem 5.

APPENDIX B
PROOF OF LEMMA 7

To prove Lemma 7, we shall need the following auxiliary result.

Lemma 20: Let X_1^k be a collection of real-valued RVs, and let $f\colon \mathbb{R}^k \to \mathbb{R}^\ell$ be Lipschitz continuous with Lipschitz constant K. Then,

$$
H([f(X_1^k)]_m | [X_1^k]_m) \leq \ell \log\lceil \mathsf{K}\sqrt{k} + 1 \rceil. \tag{132}
$$

Proof: Note that if $[X_1^k]_m = z_1^k/m$ for some $z_1^k \in \mathbb{Z}^k$, then $X_1^k \in \mathcal{C}(z_1^k/m, 1/m)$, a cube with diameter $\sqrt{k}/m$. The image of this cube under the Lipschitz function f has a diameter not greater than $\mathsf{K}\sqrt{k}/m$. Computing $[f(X_1^k)]_m$ induces a partition of $\mathbb{R}^\ell$ into ℓ-dimensional cubes. Of this partition, at most $\lceil \mathsf{K}\sqrt{k} + 1 \rceil^\ell$ elements have a nonempty intersection with the image of $\mathcal{C}(z_1^k/m, 1/m)$ under f. Therefore,

$$
H([f(X_1^k)]_m | [X_1^k]_m = z_1^k/m) \leq \ell \log\lceil \mathsf{K}\sqrt{k} + 1 \rceil \tag{133}
$$

for every $z_1^k \in \mathbb{Z}^k$, so Lemma 20 follows by averaging over $[X_1^k]_m$. ∎

We next prove Lemma 7. Let $\mathbf{Y}_t = f_t(\mathbf{X}_t)$. To prove the right-most relation in (28), we use that for every k and m

$$
\begin{aligned}
H([\mathbf{X}_1^k]_m) &\leq H([\mathbf{X}_1^k]_m, [\mathbf{Y}_1^k]_m) \\
&= H([\mathbf{X}_1^k]_m) + H([\mathbf{Y}_1^k]_m | [\mathbf{X}_1^k]_m).
\end{aligned} \tag{134}
$$

The second summand can be further upper-bounded by

$$
H([\mathbf{Y}_1^k]_m | [\mathbf{X}_1^k]_m) \leq \sum_{t=1}^{k} H([f_t(\mathbf{X}_t)]_m | [\mathbf{X}_t]_m). \tag{135}
$$

Since every function f_t is Lipschitz with a Lipschitz constant at most $\mathsf{K} \triangleq \sup_{t\in\mathbb{Z}} \mathsf{K}_t$, we can use Lemma 20 to bound the RHS of (135) by $kM \log\lceil \mathsf{K}\sqrt{L} + 1 \rceil$. Since this term is independent of m, the contribution of the second summand on the RHS of (134) vanishes as $m \to \infty$. We thus obtain $\bar{d}(\{\mathbf{X}_t, \mathbf{Y}_t\}) = \bar{d}(\{\mathbf{X}_t\})$ by dividing both sides of (134) by $k\log m$ and letting k and m tend to infinity.

To prove the left-most relation in (28), we use that for every k and m

$$H([\mathbf{Y}_1^k]_m) \leq H([\mathbf{Y}_1^k]_m, [\mathbf{X}_1^k]_m). \tag{136}$$

The claim follows then by dividing both sides of (136) by $k \log m$ and letting k and m tend to infinity.

APPENDIX C
PROOF OF LEMMA 8

For every m and k, we have

$$
\begin{aligned}
H([\mathbf{X}_1^k]_m) &\leq H([\mathbf{X}_1^k]_m, [\mathbf{Z}_1^k]_m) \\
&\leq H([\mathbf{X}_1^k]_m) + H([\mathbf{Z}_1^k]_m). \tag{137}
\end{aligned}
$$

Dividing by $k \log m$ and letting first k and then m tend to infinity yields (30).

To prove (31), we note that Lemma 7 and (30) yield $\overline{d}(\{\mathbf{X}_t + \mathbf{Z}_t\}) \leq \overline{d}(\{\mathbf{X}_t\})$. For the reverse inequality, we use [13, Lemma 30] and the fact that conditioning reduces entropy to obtain

$$
\begin{aligned}
H([\mathbf{X}_1^k &+ \mathbf{Z}_1^k]_m) \\
&\geq H([\mathbf{X}_1^k + \mathbf{Z}_1^k]_m | [\mathbf{Z}_1^k]_m) \\
&\geq H([\mathbf{X}_1^k]_m + [\mathbf{Z}_1^k]_m | [\mathbf{Z}_1^k]_m) - kL \log(2) \\
&\geq H([\mathbf{X}_1^k]_m) - H([\mathbf{Z}_1^k]_m) - kL \log(2). \tag{138}
\end{aligned}
$$

Dividing both sides of (138) by $k \log m$, and letting first k and then m tend to infinity, yields $\overline{d}(\{\mathbf{X}_t + \mathbf{Z}_t\}) \geq \overline{d}(\{\mathbf{X}_t\})$ and proves (31).

Finally, if Z is discrete and $H([\mathbf{X}_1]_1) = \infty$, then $H([\mathbf{X}_1^k]_m | Z) = \infty$, since

$$H([\mathbf{X}_1^k]_m | Z) \geq H([\mathbf{X}_1^k]_m) - H(Z) \tag{139}$$

where the second entropy is finite by assumption and the first entropy satisfies $H([\mathbf{X}_1]_m) \geq H([\mathbf{X}_1]_1) = \infty$. Conversely, if Z is discrete and $H([\mathbf{X}_1]_1) < \infty$, then

$$
\begin{aligned}
H([\mathbf{X}_1^k]_m | Z) &\leq H([\mathbf{X}_1^k]_m) \\
&= H([\mathbf{X}_1^k]_m | Z) + I([\mathbf{X}_1^k]_m; Z). \tag{140}
\end{aligned}
$$

Dividing all terms by k and letting k tend to infinity thus yields

$$
\begin{aligned}
\lim_{k \to \infty} &\frac{H([\mathbf{X}_1^k]_m | Z)}{k} \\
&\leq H'(\{[\mathbf{X}_t]_m\}) \\
&\leq \lim_{k \to \infty} \frac{H([\mathbf{X}_1^k]_m | Z)}{k} + \overline{\lim_{k \to \infty}} \frac{1}{k} I([\mathbf{X}_1^k]_m; Z). \tag{141}
\end{aligned}
$$

Since $I([\mathbf{X}_1^k]_m; Z) \leq H(Z) < \infty$, the second term on the RHS of (141) tends to zero as k tends to infinity. Thus, dividing (141) by $\log m$, and letting m tend to infinity, yields (32).

APPENDIX D
PROOF OF THEOREM 9

The proof of Theorem 9 is essentially identical to the proof of [2, Lemma 3.2]. For the sake of completeness, we reproduce the full proof here. Indeed, choosing in (33)

$$\hat{\mathbf{X}}_1^k = [\mathbf{X}_1^k]_m, \quad m = \sqrt{\frac{L}{D}} \tag{142}$$

yields

$$R(\mathbf{X}_1^k, kD) \leq H([\mathbf{X}_1^k]_m) \tag{143}$$

since for the choice (142) we have $\|\mathbf{X}_1^k - \hat{\mathbf{X}}_1^k\|_2^2 \leq \frac{kL}{m^2} = kD$, hence it satisfies (34). Consequently, dividing by $-k \log D$, and taking limits as $k \to \infty$ and $D \downarrow 0$, we obtain

$$
\begin{aligned}
2 \lim_{D \downarrow 0} \lim_{k \to \infty} \frac{R(\mathbf{X}_1^k, kD)}{-k \log D} &\leq 2 \lim_{D \downarrow 0} \lim_{k \to \infty} \frac{H([\mathbf{X}_1^k]_m)}{-k \log D} \\
&= \lim_{m \to \infty} \lim_{k \to \infty} \frac{H([\mathbf{X}_1^k]_m)}{k \log m} \tag{144}
\end{aligned}
$$

if the limits exist. If the limits do not exist, then we obtain the same upper bound for the limits replaced by limits superior and limits inferior.[4]

We next derive a lower bound on the rate-distortion dimension. To simplify notation, we treat the collection $\mathbf{X}_1^k$ of k L-variate random vectors as a collection of $k' = kL$ RVs. To show that the upper bound (144) holds with equality, we use the following lower bound on $R(\mathbf{X}_1^k, D)$ given in [23], [2, eq. (A.1)]:

$$R(X_1^{k'}, D) \geq \sup_{s \leq 0, \lambda_s} \left\{ sD + \mathsf{E}\left[\log \lambda_s \left(X_1^{k'} \right) \right] \right\} \tag{145}$$

where $\lambda_s \colon \mathbb{R}^{k'} \to [0, \infty)$ is an arbitrary nonnegative measurable function satisfying

$$\sup_{y_1^{k'} \in \mathbb{R}^{k'}} \mathsf{E}\left[\lambda_s \left(X_1^{k'} \right) e^{s \|y_1^{k'} - X_1^{k'}\|^2} \right] \leq 1. \tag{146}$$

Following the proof of [2, Lemma 3.2], we apply (145) with

$$s = -m^2 \tag{147}$$

$$\lambda_s(x_1^{k'}) = \frac{1}{N^{k'}} \sum_{i_1^{k'} \in \mathbb{Z}^{k'}} \frac{\mathbf{1}\left\{ [x_1^{k'}]_m = \frac{i_1^{k'}}{m} \right\}}{\mathbb{P}([X_1^{k'}]_m = \frac{i_1^{k'}}{m})} \tag{148}$$

$$N = 1 + 2 \sum_{i=0}^{\infty} e^{-i^2}. \tag{149}$$

---

[4]Since $m^2 = L/D$, taking the limit as $D \downarrow 0$ is tantamount to taking the limit as $m \to \infty$.

We first show that this choice of λ_s satisfies (146). Indeed,

$$\sup_{y_1^{k'} \in \mathbb{R}^{k'}} \mathsf{E}\left[\lambda_s\left(X_1^{k'}\right) e^{s\|y_1^{k'} - X_1^{k'}\|^2}\right]$$

$$\leq \sup_{y_1^{k'} \in \mathbb{R}^{k'}} \frac{1}{N^{k'}} \sum_{i_1^{k'} \in \mathbb{Z}^{k'}} \sup_{x_1^{k'}:[x_1^{k'}]_m = \frac{i_1^{k'}}{m}} e^{-m^2\|y_1^{k'} - x_1^{k'}\|^2}$$

$$= \frac{1}{N^{k'}} \sup_{j_1^{k'} \in \mathbb{Z}^{k'}} \sup_{\tilde{y}_1^{k'} \in [0,1)^{k'}} \sum_{i_1^{k'} \in \mathbb{Z}^{k'}} \prod_{\ell=1}^{k'} \sup_{0 \leq \tilde{x}_\ell < 1} e^{-(\tilde{y}_\ell + j_\ell - \tilde{x}_\ell - i_\ell)^2}$$

$$= \frac{1}{N^{k'}} \prod_{\ell=1}^{k'} \sup_{j_\ell \in \mathbb{Z}} \sup_{0 \leq \tilde{y}_\ell < 1} \sum_{i_\ell \in \mathbb{Z}} \sup_{0 \leq \tilde{x}_\ell < 1} e^{-(\tilde{y}_\ell + j_\ell - \tilde{x}_\ell - i_\ell)^2} \quad (150)$$

where the second step follows by substituting $\tilde{x}_\ell = mx_\ell - i_\ell$ and $\tilde{y}_\ell = my_\ell - j_\ell$. Since the sum over i_ℓ does not depend on j_ℓ, it follows that

$$\sup_{j_\ell \in \mathbb{Z}} \sup_{0 \leq \tilde{y}_\ell < 1} \sum_{i_\ell \in \mathbb{Z}} \sup_{0 \leq \tilde{x}_\ell < 1} e^{-(\tilde{y}_\ell + j_\ell - \tilde{x}_\ell - i_\ell)^2}$$
$$= \sup_{0 \leq \tilde{y}_\ell < 1} \sum_{i_\ell \in \mathbb{Z}} \sup_{0 \leq \tilde{x}_\ell < 1} e^{-(\tilde{y}_\ell - \tilde{x}_\ell - i_\ell)^2} \quad (151)$$

which can be upper-bounded as

$$\sup_{0 \leq \tilde{y}_\ell < 1} \sum_{i_\ell \in \mathbb{Z}} \sup_{0 \leq \tilde{x}_\ell < 1} e^{-(\tilde{y}_\ell - \tilde{x}_\ell - i_\ell)^2} \leq 1 + 2\sum_{i=0}^{\infty} e^{-i^2}. \quad (152)$$

Hence,

$$\sup_{y_1^{k'} \in \mathbb{R}^{k'}} \mathsf{E}\left[\lambda_s\left(X_1^{k'}\right) e^{s\|y_1^{k'} - X_1^{k'}\|^2}\right]$$
$$\leq \frac{1}{N^{k'}}\left(1 + 2\sum_{i=0}^{\infty} e^{-i^2}\right)^{k'} \quad (153)$$

which, by (149), is equal to 1. It follows that s and λ_s, as chosen in (147) and (148), satisfy (146).

We next evaluate (145) for this choice of s and λ_s and for distortion kD. This yields

$$R(X_1^{k'}, kD) \geq -km^2 D - k' \log N$$
$$+ \mathsf{E}\left[\log\left(\sum_{i_1^{k'} \in \mathbb{Z}^{k'}} \frac{\mathbf{1}\left\{[X_1^{k'}]_m = \frac{i_1^{k'}}{m}\right\}}{\mathbb{P}\left([X_1^{k'}]_m = \frac{i_1^{k'}}{m}\right)}\right)\right]$$
$$= H([X_1^{k'}]_m) - k(m^2 D + L \log N). \quad (154)$$

For $m^2 = L/D$, this becomes

$$R(X_1^{k'}, kD) \geq H([X_1^{k'}]_m) - k'(1 + \log N). \quad (155)$$

We next replace again the collection $X_1^{k'}$ of RVs by the equivalent collection $\mathbf{X}_1^k$ of random vectors. Dividing both sides of (155) by $-k \log D$, and taking the limits as $k \to \infty$ and $D \downarrow 0$, yields

$$2 \lim_{D \downarrow 0} \lim_{k \to \infty} \frac{R(\mathbf{X}_1^k, kD)}{-k \log D}$$
$$\geq 2 \lim_{D \downarrow 0} \lim_{k \to \infty} \frac{H([\mathbf{X}_1^k]_m) - kL(1 + \log N)}{-k \log D}$$
$$= \lim_{m \to \infty} \lim_{k \to \infty} \frac{H([\mathbf{X}_1^k]_m)}{k \log m} \quad (156)$$

if the limits over k and D exist. If the limits do not exist, then we obtain the same lower bound for the limits replaced by limits superior and limits inferior. Combining (156) with (144) proves Theorem 9.

APPENDIX E
PROOF OF THEOREM 10

The proof consists of two parts. In the first part, we show that of all processes $\{\mathbf{X}_t\}$ with a given SDF $\mathsf{F}_\mathbf{X}$, the Gaussian process has the largest information dimension rate (Section E-A). In the second part, we demonstrate that the information dimension rate of Gaussian processes is given by the average rank of the derivative of the SDF (Section E-B).

A. Gaussian Processes Maximize the Information Dimension

By Theorem 9, the upper information dimension rate is given by

$$\bar{d}(\{\mathbf{X}_t\}) = 2 \overline{\lim_{D \downarrow 0}} \lim_{k \to \infty} \frac{R(\mathbf{X}_1^k, kD)}{-k \log D}. \quad (157)$$

The claim that the information dimension is maximized by a Gaussian process then follows by the well-known fact that of all random vectors $\mathbf{X}_1^k$ with a given covariance matrix $C_{\mathbf{X}_1^k}$, the Gaussian random vector has the largest rate-distortion function $R(\mathbf{X}_1^k, kD)$.

To prove this claim for multivariate sources, we shall write the collection of L-variate vectors $\mathbf{X}_1^k$ as a collection of k' RVs $X_1^{k'}$, where $k' = kL$. Since the information dimension rate is translation invariant (Lemma 7), we can assume without loss of optimality that the RVs $X_1^{k'}$ have zero mean. Furthermore, by the eigenvalue decomposition, there exists an orthogonal matrix W such that the random variables $Y_1^{k'}$ given by $(Y_1, \ldots, Y_{k'})^\mathsf{T} = W^\mathsf{T}(X_1, \ldots, X_{k'})^\mathsf{T}$ are uncorrelated and their variances are the eigenvalues of $C_{X_1^{k'}}$, which we shall denote by $\lambda_1, \ldots, \lambda_{k'}$. Since mutual information is invariant under bijections, and the Euclidean norm is invariant under multiplications by orthogonal matrices, it follows that $R(Y_1^{k'}, kD) = R(X_1^{k'}, kD)$.

For the case where $Y_1^{k'}$ are independent, zero-mean, Gaussian random variables with variances $\lambda_1, \ldots, \lambda_{k'}$, the rate-distortion function is given by [10, Th. 13.3.3]

$$R(Y_1^{k'}, kD) = \sum_{t=1}^{k'} \frac{1}{2} \log \frac{\lambda_t}{D_t} \quad (158)$$

where

$$D_t = \begin{cases} \xi, & \text{if } \xi < \lambda_t \\ \lambda_t, & \text{if } \xi \geq \lambda_t \end{cases} \quad (159)$$

and ξ is chosen so that $D_1 + \ldots + D_{k'} = kD$.

The RHS of (158) can also be achieved for non-Gaussian RVs by choosing the following (possibly suboptimal) distribution of the reconstruction values $\hat{Y}_1^{k'}$:

$$\hat{Y}_t = \begin{cases} (Y_t + Z_t)\frac{\lambda_t - D_t}{\lambda_t}, & \text{if } \xi < \lambda_t \\ 0, & \text{if } \xi \geq \lambda_t \end{cases} \quad (160)$$

where $Z_1^{k'}$ are independent, zero-mean, Gaussian RVs with variances $\frac{D_t \lambda_t}{\lambda_t - D_t}$, and ξ is as in (159). Indeed, it is easy

to check that (160) satisfies the distortion constraint (34). Furthermore, by using that conditioning reduces entropy and that Gaussian RVs maximize differential entropy, it can be shown that

$$I(\hat{Y}_1^{k'}; Y_1^{k'}) \leq \sum_{t=1}^{k'} \frac{1}{2} \log \frac{\lambda_t}{D_t}. \tag{161}$$

Comparing (161) with (158), we conclude that, for uncorrelated RVs $Y_1^{k'}$,

$$R(Y_1^{k'}, kD) \leq R((Y_1^{k'})_G, kD) \tag{162}$$

where $(Y_1^{k'})_G$ are jointly Gaussian with the same covariance matrix as $Y_1^{k'}$. Since $R(Y_1^{k'}, kD) = R(X_1^{k'}, kD)$ and $R((Y_1^{k'})_G, kD) = R((X_1^{k'})_G, kD)$, the same is also true for general RVs. Together with (157), this proves that of all processes $\{\mathbf{X}_t\}$ with a given SDF $\mathsf{F_X}$, the Gaussian process has the largest information dimension rate.

B. The Information Dimension of Gaussian Processes

We now assume that $\{\mathbf{X}_t\}$ is Gaussian. For every i, we define

$$N_{i,t} \triangleq X_{i,t} - [X_{i,t}]_m. \tag{163}$$

Furthermore, let $U_{i,t}$ be i.i.d. (over all i and t) and uniformly distributed on $[0, 1/m)$, and let $W_{i,t} \triangleq [X_{i,t}]_m + U_{i,t}$. We define $\{[\mathbf{X}_t]_m\}$, $\{\mathbf{N}_t\}$, and $\{\mathbf{U}_t\}$ as the corresponding multivariate processes. Since $\{U_{i,t}\}$ is independent of $\{[X_{j,t}]_m\}$ for every i, j, the (matrix-valued) SDFs of $\{\mathbf{W}_t\}$, $\{[\mathbf{X}_t]_m\}$, and $\{\mathbf{U}_t\}$ satisfy

$$\mathsf{F_W}(\theta) = \mathsf{F}_{[\mathbf{X}]_m}(\theta) + \mathsf{F_U}(\theta). \tag{164}$$

Moreover, the (matrix-valued) PSD of $\{\mathbf{U}_t\}$ exists and equals

$$\mathsf{S_U}(\theta) = \frac{1}{12m^2} I_L. \tag{165}$$

Since the information dimension rate is translation invariant (Lemma 7), and since the SDF $\mathsf{F_X}$ does not depend on the mean vector $\boldsymbol{\mu}$, we can assume without loss of generality that $\{\mathbf{X}_t\}$ has zero mean. We further show in Lemma 21 in Appendix E-C that we can assume without loss of generality that every component process of $\{\mathbf{X}_t\}$ has unit variance. By (48) in Lemma 11, it thus follows that

$$\mathsf{F}_{[\mathbf{X}]_m}(\theta) = (2a_1 - 1)\mathsf{F_X}(\theta) + \mathsf{F_N}(\theta). \tag{166}$$

We continue by writing the entropy of $[\mathbf{X}_1^k]_m$ in terms of a differential entropy, i.e.,

$$H([\mathbf{X}_1^k]_m) = h(\mathbf{W}_1^k) + kL \log m. \tag{167}$$

Denoting by $(\mathbf{W}_1^k)_G$ a Gaussian vector with the same mean and covariance matrix as $\mathbf{W}_1^k$, and denoting by $f_{\mathbf{W}_1^k}$ and $g_{\mathbf{W}_1^k}$ the PDFs of $\mathbf{W}_1^k$ and $(\mathbf{W}_1^k)_G$, respectively, this can be expressed as

$$H([\mathbf{X}_1^k]_m) = h((\mathbf{W}_1^k)_G) - D(f_{\mathbf{W}_1^k} \| g_{\mathbf{W}_1^k}) + kL \log m. \tag{168}$$

Dividing by $k \log m$, and letting first k then m tend to infinity, yields the information dimension rate $d(\{\mathbf{X}_t\})$. Lemma 22 in Appendix E-C shows that

$$D(f_{\mathbf{W}_1^k} \| g_{\mathbf{W}_1^k}) \leq k\Xi \tag{169}$$

for some constant Ξ that is independent of (k, m). Moreover, the differential entropy rate of the stationary, L-variate, Gaussian process $(\{\mathbf{W}_t\})_G$ is given by [15, Th. 7.10]

$$\lim_{k \to \infty} \frac{h((\mathbf{W}_1^k)_G)}{k}$$
$$= \frac{L}{2} \log(2\pi e) + \frac{1}{2} \int_{-1/2}^{1/2} \log \det \mathsf{F'_W}(\theta) \mathrm{d}\theta. \tag{170}$$

It thus follows that the information dimension rate of $\{\mathbf{X}_t\}$ equals

$$d(\{\mathbf{X}_t\}) = L + \lim_{m \to \infty} \frac{1}{2 \log m} \int_{-1/2}^{1/2} \log \det \mathsf{F'_W}(\theta) \mathrm{d}\theta. \tag{171}$$

It remains to show that the RHS of (171) is equal to the RHS of (43). To do so, we first show that the integral on the RHS of (171) can be restricted to a subset $\mathcal{F}_\Upsilon^c \subseteq [-1/2, 1/2]$ on which the entries of $\mathsf{F'_N}(\theta)$ are bounded from above by Υ/m^2 for some $\Upsilon > 0$. We then show that, on this set, $\det \mathsf{F'_W}(\theta)$ can be bounded from above and from below by products of affine transforms of the *eigenvalues of* $\mathsf{F'_X}(\theta)$. These bounds are asymptotically tight, i.e., they are equal in the limit as m tends to infinity. We complete the proof by showing that the order of limit and integration can be exchanged.

1) Restriction on $\mathcal{F}_\Upsilon^c \subseteq [-1/2, 1/2]$: Choose $\Upsilon > 0$ and let

$$\mathcal{F}_\Upsilon^{(i)} \triangleq \{\theta : \mathsf{F'}_{N_i}(\theta) > \Upsilon/m^2\} \tag{172}$$

$$\mathcal{F}_\Upsilon \triangleq \{\theta : \max_{i=1,\ldots,L} \mathsf{F'}_{N_i}(\theta) > \Upsilon/m^2\}. \tag{173}$$

By (47) in Lemma 11, we have for every i

$$\lambda\left(\mathcal{F}_\Upsilon^{(i)}\right) \leq \frac{1}{\Upsilon}. \tag{174}$$

Since the set $\mathcal{F}_\Upsilon$ is the union of $\mathcal{F}_\Upsilon^{(i)}$, $i = 1, \ldots, L$, it then follows by the union bound that

$$\lambda(\mathcal{F}_\Upsilon) \leq \frac{L}{\Upsilon}. \tag{175}$$

To prove (174), we note that, by the Lebesgue decomposition theorem [22, Th. 2.2.6] and the fact that the Radon-Nikodym derivative $\mathrm{d}\mathsf{F}_{N_i}(\theta)/\mathrm{d}\lambda$ coincides with $\mathsf{F'}_{N_i}(\theta)$ almost everywhere [22, Sec. 2.3],

$$\int_{-1/2}^{1/2} \mathrm{d}\mathsf{F}_{N_i}(\theta) \geq \int_{-1/2}^{1/2} \mathsf{F'}_{N_i}(\theta) \mathrm{d}\theta$$
$$\geq \int_{\mathcal{F}_\Upsilon^{(i)}} \mathsf{F'}_{N_i}(\theta) \mathrm{d}\theta$$
$$\geq \frac{\Upsilon}{m^2} \lambda\left(\mathcal{F}_\Upsilon^{(i)}\right) \tag{176}$$

where the second inequality follows because $\theta \mapsto \mathsf{F'}_{N_i}(\theta)$ is nonnegative, and the third inequality follows by definition of $\mathcal{F}_\Upsilon^{(i)}$. By (47) in Lemma 11, the integral on the left-hand side (LHS) of (176) is upper-bounded by $1/m^2$, hence (174) follows.

By (164) and (165), we have that

$$\mathsf{F'_W}(\theta) = \mathsf{F'}_{[\mathbf{X}]_m}(\theta) + \frac{1}{12m^2} I_L. \tag{177}$$

Since derivatives of matrix-valued SDFs are positive semidefinite, it follows that

$$\det \mathsf{F}'_{\mathbf{W}}(\theta) \geq \det \mathsf{S}_{\mathbf{U}}(\theta) = 1/(12m^2)^L. \qquad (178)$$

Hence,

$$\varliminf_{m \to \infty} \frac{\int_{\mathcal{F}_\Upsilon} \log \det \mathsf{F}'_{\mathbf{W}}(\theta) \mathrm{d}\theta}{\log m} \geq -\varlimsup_{m \to \infty} \lambda(\mathcal{F}_\Upsilon) \frac{L \log(12m^2)}{\log m}$$

$$\geq -\frac{2L^2}{\Upsilon} \qquad (179)$$

where the last step follows from (175). Applying Hadamard's and Jensen's inequality, we further get

$$\int_{\mathcal{F}_\Upsilon} \log \det \mathsf{F}'_{\mathbf{W}}(\theta) \mathrm{d}\theta$$

$$\leq \sum_{i=1}^{L} \int_{\mathcal{F}_\Upsilon} \log \mathsf{F}'_{W_i}(\theta) \mathrm{d}\theta$$

$$\leq \sum_{i=1}^{L} \lambda(\mathcal{F}_\Upsilon) \log \left(\frac{\int_{\mathcal{F}_\Upsilon} \mathsf{F}'_{W_i}(\theta) \mathrm{d}\theta}{\lambda(\mathcal{F}_\Upsilon)} \right)$$

$$\leq L\lambda(\mathcal{F}_\Upsilon) \log \left(\frac{(2a_1 - 1) + \frac{1}{12m^2} + \frac{1}{m^2}}{\lambda(\mathcal{F}_\Upsilon)} \right) \quad (180)$$

where the last step follows from (47), (164), (166), and the assumption that every component process of $\{\mathbf{X}_t\}$ has zero mean and unit variance. Since, by (46) in Lemma 11, $a_1 \to 1$ as $m \to \infty$, (180) yields

$$\varlimsup_{m \to \infty} \int_{\mathcal{F}_\Upsilon} \log \det \mathsf{F}'_{\mathbf{W}}(\theta) \mathrm{d}\theta \leq -\varlimsup_{m \to \infty} L\lambda(\mathcal{F}_\Upsilon) \log \left(\lambda(\mathcal{F}_\Upsilon) \right)$$

$$\leq \frac{L}{e}. \qquad (181)$$

Consequently,

$$-\frac{2L^2}{\Upsilon} \leq \varliminf_{m \to \infty} \frac{\int_{\mathcal{F}_\Upsilon} \log \det \mathsf{F}'_{\mathbf{W}}(\theta) \mathrm{d}\theta}{\log m}$$

$$\leq \varlimsup_{m \to \infty} \frac{\int_{\mathcal{F}_\Upsilon} \log \det \mathsf{F}'_{\mathbf{W}}(\theta) \mathrm{d}\theta}{\log m} \leq 0 \quad (182)$$

for every Υ. It follows that this integral does not contribute to the information dimension rate if we let Υ tend to infinity. In view of (171), we thus obtain the information dimension rate $d(\{\mathbf{X}_t\})$ by evaluating

$$L + \frac{1}{2 \log m} \int_{\mathcal{F}_\Upsilon^{\mathsf{c}}} \log \det \mathsf{F}'_{\mathbf{W}}(\theta) \mathrm{d}\theta \qquad (183)$$

in the limit as first m and then Υ tends to infinity.

2) Bounding $\det \mathsf{F}'_{\mathbf{W}}(\theta)$ *by the Eigenvalues of* $\mathsf{F}'_{\mathbf{X}}(\theta)$: Lemma 11 and (177) yield

$$\mathsf{F}'_{\mathbf{W}}(\theta) = (2a_1 - 1)\mathsf{F}'_{\mathbf{X}}(\theta) + \mathsf{F}'_{\mathbf{N}}(\theta) + \frac{1}{12m^2} I_L. \qquad (184)$$

Let $\chi_i(\theta)$, $i = 1, \ldots, L$, denote the eigenvalues of $\mathsf{F}'_{\mathbf{X}}(\theta)$. Since $\mathsf{F}'_{\mathbf{N}}(\theta)$ is positive semidefinite, we obtain

$$\det \mathsf{F}'_{\mathbf{W}}(\theta) \geq \det \left((2a_1 - 1)\mathsf{F}'_{\mathbf{X}}(\theta) + \frac{1}{12m^2} I_L \right)$$

$$= \prod_{i=1}^{L} \left((2a_1 - 1)\chi_i(\theta) + \frac{1}{12m^2} \right). \qquad (185)$$

We next derive an upper bound on $\det \mathsf{F}'_{\mathbf{W}}(\theta)$. Let $\|\mathsf{F}'_{\mathbf{N}}(\theta)\|_1 \triangleq \sum_{i,j=1}^{n} |\mathsf{F}'_{N_i N_j}(\theta)|$ denote the ℓ_1-*matrix norm* of $\mathsf{F}'_{\mathbf{N}}(\theta)$. Since $\mathsf{F}'_{\mathbf{N}}(\theta)$ is positive semidefinite, the element with the maximum modulus is on the main diagonal; cf. [12, Problem 7.1.P1]. Furthermore, by assumption, on $\mathcal{F}_\Upsilon^{\mathsf{c}}$ the diagonal elements of $\mathsf{F}_{\mathbf{N}}(\theta)$ are bounded from above by $\frac{\Upsilon}{m^2}$. We hence obtain that

$$\|\mathsf{F}'_{\mathbf{N}}(\theta)\|_1 \leq L^2 \frac{\Upsilon}{m^2}. \qquad (186)$$

It is known that all matrix norms bound the largest eigenvalue of the matrix from above [12, Th. 5.6.9].[5] Thus, the upper bound (186) is also an upper bound on the largest eigenvalue of $\mathsf{F}'_{\mathbf{N}}(\theta)$. Let $\omega_i(\theta)$, $i = 1 \ldots, L$, denote the eigenvalues of $\mathsf{F}'_{\mathbf{W}}(\theta)$. Then, we have for $m^2 \geq 8/\pi$ (such that $2a_1 - 1 \geq 0$) [12, Cor. 4.3.15]

$$\det \mathsf{F}'_{\mathbf{W}}(\theta) = \prod_{i=1}^{L} \omega_i(\theta)$$

$$\leq \prod_{i=1}^{L} \left((2a_1 - 1)\chi_i(\theta) + \frac{L^2\Upsilon}{m^2} + \frac{1}{12m^2} \right). \quad (187)$$

Combining (185) and (187) with (183), we obtain

$$\lim_{m \to \infty} \sum_{i=1}^{L} \frac{\int_{\mathcal{F}_\Upsilon^{\mathsf{c}}} \log \left((2a_1 - 1)\chi_i(\theta) + \frac{1}{12m^2} \right) \mathrm{d}\theta}{\log m}$$

$$\leq \varliminf_{m \to \infty} \frac{\int_{\mathcal{F}_\Upsilon^{\mathsf{c}}} \log \det \mathsf{F}'_{\mathbf{W}}(\theta) \mathrm{d}\theta}{\log m}$$

$$\leq \varlimsup_{m \to \infty} \frac{\int_{\mathcal{F}_\Upsilon^{\mathsf{c}}} \log \det \mathsf{F}'_{\mathbf{W}}(\theta) \mathrm{d}\theta}{\log m}$$

$$\leq \lim_{m \to \infty} \sum_{i=1}^{L} \frac{\int_{\mathcal{F}_\Upsilon^{\mathsf{c}}} \log \left((2a_1 - 1)\chi_i(\theta) + \frac{\frac{1}{12} + L^2\Upsilon}{m^2} \right) \mathrm{d}\theta}{\log m}. \quad (188)$$

To compute the limit of (183) as $m \to \infty$, we thus need to evaluate

$$L - \sum_{i=1}^{L} \lim_{m \to \infty} \int_{\mathcal{F}_\Upsilon^{\mathsf{c}}} \frac{\log \left((2a_1 - 1)\chi_i(\theta) + \frac{\mathsf{K}}{m^2} \right)}{\log(1/m^2)} \mathrm{d}\theta \quad (189)$$

where K is either $1/12$ (left-most inequality in (188)) or $1/12 + L^2\Upsilon$ (right-most inequality in (188)).

3) Exchanging Limit and Integration: To evaluate (189), we continue along the lines of [24, Sec. VIII]. Specifically, for each i, we split the integral on the RHS of (189) into three parts:

$$\mathcal{F}_I \triangleq \{\theta \in \mathcal{F}_\Upsilon^{\mathsf{c}} : \chi_i(\theta) = 0\} \qquad (190)$$

$$\mathcal{F}_{II} \triangleq \{\theta \in \mathcal{F}_\Upsilon^{\mathsf{c}} : \chi_i(\theta) \geq \mathsf{K}/(1 - \varepsilon)\} \qquad (191)$$

$$\mathcal{F}_{III} \triangleq \{\theta \in \mathcal{F}_\Upsilon^{\mathsf{c}} : 0 < \chi_i(\theta) < \mathsf{K}/(1 - \varepsilon)\} \qquad (192)$$

where $0 < \varepsilon < 1$ is arbitrary.

---

[5]This bound holds without a multiplicative constant, since the spectral radius of a matrix is the infimum of all matrix norms [12, Lemma 5.6.10].

For the first part, we obtain

$$\int_{\mathcal{F}_I} \frac{\log\left((2a_1-1)\chi_i(\theta)+\frac{\mathsf{K}}{m^2}\right)}{\log(1/m^2)}\mathrm{d}\theta$$

$$= \int_{\mathcal{F}_I} \frac{\log\mathsf{K} + \log(1/m^2)}{\log(1/m^2)}\mathrm{d}\theta$$

$$= \lambda(\mathcal{F}_I)\left(1 + \frac{\log\mathsf{K}}{\log(1/m^2)}\right) \tag{193}$$

which evaluates to $\lambda(\mathcal{F}_I)$ in the limit as $m \to \infty$.

We next show that the integrals over $\mathcal{F}_{II}$ and $\mathcal{F}_{III}$ do not contribute to (189). To this end, it suffices to consider the integral of the function

$$\frac{\log\left((2a_1-1)\frac{\chi_i(\theta)}{\mathsf{K}}+\frac{1}{m^2}\right)}{\log(1/m^2)} \triangleq \frac{\log\left(A_m(\theta)+\frac{1}{m^2}\right)}{\log(1/m^2)}. \tag{194}$$

In the remainder of the proof, we shall assume without loss of generality that $m^2 > 8/\pi$, in which case $A_m(\theta) > 0$ on $\theta \in \mathcal{F}_{II} \cup \mathcal{F}_{III}$. Clearly, whenever $A_m(\theta) > 0$, the function in (194) converges to zero as $m \to \infty$. Moreover, for $A_m(\theta) \geq 1$, this function is nonpositive.

For all $\theta \in \mathcal{F}_{II}$ we have $A_m(\theta) \geq (2a_1-1)/(1-\varepsilon)$, hence we can find a sufficiently large m_0 such that, by (46) in Lemma 11, we have $A_m(\theta) \geq 1$, $m \geq m_0$. Since by the same result we also have $2a_1 - 1 \leq 2$, $m^2 > 8/\pi$, it follows that, for $m > \max\{m_0, \sqrt{8/\pi}\}$,

$$\frac{\log\left(2\frac{\chi_i(\theta)}{\mathsf{K}}+\frac{1}{m^2}\right)}{\log(1/m^2)} \leq \frac{\log\left(A_m(\theta)+\frac{1}{m^2}\right)}{\log(1/m^2)} \leq 0. \tag{195}$$

The LHS of (195) is nonpositive and monotonically increases to zero as $m \to \infty$. We can thus apply the monotone convergence theorem [22, Th. 1.6.7, p. 49] to get

$$0 \geq \varlimsup_{m\to\infty} \int_{\mathcal{F}_{II}} \frac{\log\left(A_m(\theta)+\frac{1}{m^2}\right)}{\log(1/m^2)}\mathrm{d}\theta$$

$$\geq \varliminf_{m\to\infty} \int_{\mathcal{F}_{II}} \frac{\log\left(A_m(\theta)+\frac{1}{m^2}\right)}{\log(1/m^2)}\mathrm{d}\theta$$

$$\geq \lim_{m\to\infty} \int_{\mathcal{F}_{II}} \frac{\log\left(2\frac{\chi_i(\theta)}{\mathsf{K}}+\frac{1}{m^2}\right)}{\log(1/m^2)}\mathrm{d}\theta$$

$$= \int_{\mathcal{F}_{II}} \lim_{m\to\infty} \frac{\log\left(2\frac{\chi_i(\theta)}{\mathsf{K}}+\frac{1}{m^2}\right)}{\log(1/m^2)}\mathrm{d}\theta$$

$$= 0. \tag{196}$$

We next turn to the case $\theta \in \mathcal{F}_{III}$. It was shown in [24, p. 443] that if $A_m(\theta) < 1$, then the function in (194) is bounded from above by 1. Furthermore, if $A_m(\theta) < 1 - \frac{1}{m^2}$ then it is nonnegative, and if $A_m(\theta) \geq 1 - \frac{1}{m^2}$ then it is nonpositive and monotonically increasing in m. Restricting ourselves to the case $m^2 > 8/\pi$, we thus obtain for $\theta \in \mathcal{F}_{III}$

$$\frac{\log\left(A_m(\theta)+\frac{1}{m^2}\right)}{\log(1/m^2)} \geq \begin{cases} \frac{\log\left(\frac{2}{1-\varepsilon}+\frac{\pi}{8}\right)}{\log(\pi/8)}, & A_m(\theta) \geq 1 - \frac{\pi}{8} \\ 0, & \text{otherwise} \end{cases} \tag{197}$$

where we made use of the fact that $A_m(\theta) < (2a_1-1)/(1-\varepsilon)$, $\theta \in \mathcal{F}_{III}$ and, by (46) in Lemma 11, $2a_1 - 1 \leq 2$, $m^2 > 8/\pi$.

Hence, on $\mathcal{F}_{III}$ the magnitude of the function in (194) is bounded by

$$\left|\frac{\log\left(A_m(\theta)+\frac{1}{m^2}\right)}{\log(1/m^2)}\right| \leq \max\left\{1, \frac{\log\left(\frac{2}{1-\varepsilon}+\frac{\pi}{8}\right)}{\log(8/\pi)}\right\}. \tag{198}$$

We can thus apply the dominated convergence theorem [22, Th. 1.6.9, p. 50] to get

$$\lim_{m\to\infty} \int_{\mathcal{F}_{III}} \frac{\log\left((2a_1-1)\chi_i(\theta)+\frac{\mathsf{K}}{m^2}\right)}{\log(1/m^2)}\mathrm{d}\theta$$

$$= \int_{\mathcal{F}_{III}} \lim_{m\to\infty} \frac{\log\left((2a_1-1)\frac{\chi_i(\theta)}{\mathsf{K}}+\frac{1}{m^2}\right)}{\log(1/m^2)}\mathrm{d}\theta = 0. \tag{199}$$

Combining (193), (196), and (199), we can evaluate (189) as

$$L - \sum_{i=1}^{L} \lim_{m\to\infty} \int_{\mathcal{F}_\Upsilon^{\mathsf{c}}} \frac{\log\left((2a_1-1)\chi_i(\theta)+\frac{\mathsf{K}}{m^2}\right)}{\log(1/m^2)}\mathrm{d}\theta$$

$$= \sum_{i=1}^{L} \left(1 - \lambda(\{\theta \in \mathcal{F}_\Upsilon^{\mathsf{c}}: \chi_i(\theta) = 0\})\right)$$

$$= \sum_{i=1}^{L} \lambda(\{\theta \in \mathcal{F}_\Upsilon^{\mathsf{c}}: \chi_i(\theta) > 0\}). \tag{200}$$

4) Wrapping Up: To compute the limit of (183) as first m and then Υ tends to infinity, it remains to let $\Upsilon \to \infty$ on the RHS of (200). By the continuity of the Lebesgue measure, this yields

$$\sum_{i=1}^{L} \lambda(\{\theta: \chi_i(\theta) > 0\}) = \int_{-1/2}^{1/2} \mathrm{rank}(\mathsf{F}'_{\mathbf{X}}(\theta))\mathrm{d}\theta. \tag{201}$$

To summarize, combining (171), (182), and (200), we obtain that

$$d(\{\mathbf{X}_t\}) = L + \lim_{\Upsilon\to\infty}\lim_{m\to\infty} \frac{1}{2\log m} \int_{-1/2}^{1/2} \log\det\mathsf{F}'_{\mathbf{W}}(\theta)\mathrm{d}\theta$$

$$= \int_{-1/2}^{1/2} \mathrm{rank}(\mathsf{F}'_{\mathbf{X}}(\theta))\mathrm{d}\theta. \tag{202}$$

This proves Theorem 10.

C. Auxiliary Results

Lemma 21: Suppose that $\{\mathbf{X}_t\}$ is a stationary, L-variate, real-valued, Gaussian process with mean vector $\boldsymbol{\mu}$ and SDF $\mathsf{F}_{\mathbf{X}}$. Suppose that the component processes are ordered by their variances, i.e.,

$$\sigma_1^2 \geq \sigma_2^2 \geq \cdots \sigma_{L'}^2 \geq \sigma_{L'+1}^2 = \cdots \sigma_L^2 = 0. \tag{203}$$

Then,

$$d(\{\mathbf{X}_t\}) = d\left(\{(X_{1,t}/\sigma_1, \ldots, X_{L',t}/\sigma_{L'})\}\right) \tag{204}$$

and, for almost every θ,

$$\mathrm{rank}(\mathsf{F}'_{\mathbf{X}}(\theta)) = \mathrm{rank}(\mathsf{F}'_{(X_1/\sigma_1,\ldots,X_{L'}/\sigma_{L'})}(\theta)). \tag{205}$$

Proof: Normalizing component processes with positive variance to unit variance does not affect the information dimension rate, as follows from Lemma 7. If $\sigma_i^2 = 0$, then

the component process $\{X_{i,t}\}$ is almost surely constant. It follows that $H([X_{i,1}^k]_m) = 0$ for every m and every k, so

$$H([\mathbf{X}_1^k]_m) = H([X_{1,1}^k]_m, \ldots, [X_{L',1}^k]_m). \qquad (206)$$

Dividing by $k \log m$, and letting m and k tend to infinity, shows that $d(\{\mathbf{X}_t\}) = d(\{\frac{1}{\sigma_1} X_{1,t}, \ldots, \frac{1}{\sigma_{L'}} X_{L',t}\})$.

Let Π be an $L' \times L'$ diagonal matrix with values σ_i on the main diagonal. For component processes with zero variance, the corresponding row and column of $\mathsf{F}'_{\mathbf{X}}(\theta)$ is zero almost everywhere. Hence, we have for almost every θ that

$$\mathsf{F}'_{\mathbf{X}}(\theta) = \begin{bmatrix} \Pi \mathsf{F}'_{(X_1/\sigma_1, \ldots, X_{L'}/\sigma_{L'})}(\theta)\Pi & 0 \\ 0 & 0 \end{bmatrix} \qquad (207)$$

where 0 denotes an all-zero matrix of appropriate size. We thus have $\mathrm{rank}(\mathsf{F}'_{\mathbf{X}}(\theta)) = \mathrm{rank}(\mathsf{F}'_{(X_1/\sigma_1, \ldots, X_{L'}/\sigma_{L'})}(\theta))$ for almost every θ. ■

Lemma 22: Let $\mathbf{X}$ be an ℓ-variate, real-valued, Gaussian vector with mean vector $\boldsymbol{\mu}_{\mathbf{X}}$ and covariance matrix $C_{\mathbf{X}}$. Let $\mathbf{W} \triangleq [\mathbf{X}]_m + \mathbf{U}$, where $\mathbf{U}$ is an ℓ-variate vector, independent of $\mathbf{X}$, with components independently and uniformly distributed on $[0, 1/m)$. Then,

$$\frac{D(f_{\mathbf{W}} \| g_{\mathbf{W}})}{\ell} \leq \frac{1}{2} \log \left(2\pi \left(1 + \frac{1}{12} \right) \right) + \frac{75}{2} + \frac{24}{\pi}. \qquad (208)$$

Proof: By [25, Th. 23.6.14], $\mathbf{X} = (X_1, \ldots, X_\ell)^\mathsf{T}$ can be written as

$$\mathbf{X} = A\mathbf{X}' + \boldsymbol{\mu}_{\mathbf{X}} \qquad (209)$$

where $\mathbf{X}'$ is an ℓ'-dimensional, zero-mean, Gaussian vector ($\ell' \leq \ell$) with independent components whose variances are the nonzero eigenvalues of $C_{\mathbf{X}}$ and where A is an $\ell \times \ell'$ matrix satisfying $A^\mathsf{T} A = I_{\ell'}$. We use the data processing inequality, the chain rule for relative entropy, and the fact that $\mathbf{X}'$ is Gaussian, to obtain

$$D(f_{\mathbf{W}} \| g_{\mathbf{W}})$$
$$\leq D(f_{\mathbf{W}, \mathbf{X}'} \| g_{\mathbf{W}, \mathbf{X}'})$$
$$\leq D(f_{\mathbf{X}'} \| g_{\mathbf{X}'}) + \int D(f_{\mathbf{W}|\mathbf{X}'=\mathbf{x}'} \| g_{\mathbf{W}|\mathbf{X}'=\mathbf{x}'}) f_{\mathbf{X}'}(\mathbf{x}') \mathrm{d}\mathbf{x}'$$
$$= \int D(f_{\mathbf{W}|\mathbf{X}'=\mathbf{x}'} \| g_{\mathbf{W}|\mathbf{X}'=\mathbf{x}'}) f_{\mathbf{X}'}(\mathbf{x}') \mathrm{d}\mathbf{x}' \qquad (210)$$

where $g_{\mathbf{W}, \mathbf{X}'}$ denotes the PDF of a Gaussian vector with the same mean vector and covariance matrix as $(\mathbf{W}, \mathbf{X}')$, and

$$f_{\mathbf{W}|\mathbf{X}'=\mathbf{x}'}(\mathbf{w}) \triangleq \frac{f_{\mathbf{W}, \mathbf{X}'}(\mathbf{w}, \mathbf{x}')}{f_{\mathbf{X}'}(\mathbf{x}')} \qquad (211)$$

$$g_{\mathbf{W}|\mathbf{X}'=\mathbf{x}'}(\mathbf{w}) \triangleq \frac{g_{\mathbf{W}, \mathbf{X}'}(\mathbf{w}, \mathbf{x}')}{g_{\mathbf{X}'}(\mathbf{x}')}. \qquad (212)$$

To evaluate the relative entropy on the RHS of (210), we first note that, given $\mathbf{X}$, the random vector $\mathbf{W}$ is uniformly distributed on an ℓ-dimensional cube of length $\frac{1}{m}$. Since $\mathbf{X}$ can be obtained from $\mathbf{X}'$ via (209), the conditional PDF of $\mathbf{W}$ given $\mathbf{X}' = \mathbf{x}'$ is

$$f_{\mathbf{W}|\mathbf{X}'=\mathbf{x}'}(\mathbf{w}) = m^\ell \mathbf{1}\{[\mathbf{w}]_m = [A\mathbf{x}' + \boldsymbol{\mu}_{\mathbf{X}}]_m\}. \qquad (213)$$

Consequently, denoting $\mathbf{z} = [A\mathbf{x}' + \boldsymbol{\mu}_{\mathbf{X}}]_m$,

$$D(f_{\mathbf{W}|\mathbf{X}'=\mathbf{x}'} \| g_{\mathbf{W}|\mathbf{X}'=\mathbf{x}'}) = \log \left(m^\ell \sqrt{(2\pi)^\ell \det C_{\mathbf{W}|\mathbf{X}'}} \right)$$
$$+ \frac{m^\ell}{2} \int_{\mathcal{C}(\mathbf{z}, 1/m)} (\mathbf{w} - \boldsymbol{\mu}_{\mathbf{W}|\mathbf{X}'=\mathbf{x}'})^\mathsf{T} C_{\mathbf{W}|\mathbf{X}'}^{-1} (\mathbf{w} - \boldsymbol{\mu}_{\mathbf{W}|\mathbf{X}'=\mathbf{x}'}) \mathrm{d}\mathbf{w} \qquad (214)$$

where $\boldsymbol{\mu}_{\mathbf{W}|\mathbf{X}'=\mathbf{x}'}$ and $C_{\mathbf{W}|\mathbf{X}'}$ denote the conditional mean and the conditional covariance matrix of $\mathbf{W}$ given $\mathbf{X}' = \mathbf{x}'$. These can be computed as [25, Th. 23.7.4]

$$\boldsymbol{\mu}_{\mathbf{W}|\mathbf{X}'=\mathbf{x}'} = \mathsf{E}[\mathbf{W}] + C_{\mathbf{W}\mathbf{X}'} C_{\mathbf{X}'}^{-1} \mathbf{x}' \qquad (215)$$

$$C_{\mathbf{W}|\mathbf{X}'} = C_{\mathbf{W}} - C_{\mathbf{W}\mathbf{X}'} C_{\mathbf{X}'}^{-1} C_{\mathbf{W}\mathbf{X}'}^\mathsf{T} \qquad (216)$$

where $C_{\mathbf{W}\mathbf{X}'}$ denotes the cross-covariance matrix of $\mathbf{W}$ and $\mathbf{X}'$, and $C_{\mathbf{W}}$ and $C_{\mathbf{X}'}$ denote the covariance matrices of $\mathbf{W}$ and $\mathbf{X}'$, respectively.

Defining $\mathbf{Z} \triangleq [\mathbf{X}]_m$, we have $\mathbf{W} = \mathbf{Z} + \mathbf{U}$. Since $\mathbf{U}$ is independent of $\mathbf{X}$, the cross-covariance matrix of $\mathbf{W}$ and $\mathbf{X}$ is equal to the cross-covariance matrix of $\mathbf{Z}$ and $\mathbf{X}$. Bussgang's theorem [26, eq. (20)] yields $K_{Z_j X_i}(\tau) = a_j K_{X_j X_i}(\tau)$, where a_j is defined in (45). Hence, if $\Lambda_{\mathbf{a}}$ is a diagonal matrix with $\mathbf{a} = (a_1, \ldots, a_\ell)$ on the main diagonal, then $C_{\mathbf{Z}\mathbf{X}} = \Lambda_{\mathbf{a}} C_{\mathbf{X}}$. From (209) we get $C_{\mathbf{X}} = A C_{\mathbf{X}'} A^\mathsf{T}$ and $C_{\mathbf{W}\mathbf{X}'} = C_{\mathbf{W}\mathbf{X}} A$, hence

$$C_{\mathbf{W}\mathbf{X}'} = C_{\mathbf{W}\mathbf{X}} A = C_{\mathbf{Z}\mathbf{X}} A = \Lambda_{\mathbf{a}} C_{\mathbf{X}} A = \Lambda_{\mathbf{a}} A C_{\mathbf{X}'}. \qquad (217)$$

Together with (215) and (216), this yields

$$\boldsymbol{\mu}_{\mathbf{W}|\mathbf{X}'=\mathbf{x}'} = \mathsf{E}[\mathbf{W}] + \Lambda_{\mathbf{a}} A \mathbf{x}' \qquad (218)$$

$$C_{\mathbf{W}|\mathbf{X}'} = C_{\mathbf{W}} - \Lambda_{\mathbf{a}} C_{\mathbf{X}} \Lambda_{\mathbf{a}}. \qquad (219)$$

Combining (218) with (209), and using the triangle inequality, we upper-bound each component of $\mathbf{w} - \boldsymbol{\mu}_{\mathbf{W}|\mathbf{X}'=\mathbf{x}'}$ as

$$|w_j - \mathsf{E}[W_j] - a_j(x_j - \mu_j)| \leq |z_j - x_j| + |u_j - \mathsf{E}[U_j]|$$
$$+ |\mathsf{E}[Z_j] - \mu_j| + |1 - a_j||x_j - \mu_j|. \qquad (220)$$

The first and the third term on the RHS of (220) are both upper-bounded by $\frac{1}{m}$, and the second term is upper-bounded by $\frac{1}{2m}$. From (46) in Lemma 11, we get that the term $|1 - a_j|$ is upper-bounded by $1/m\sqrt{2/\pi\sigma_j^2}$, where σ_j^2 is the variance of X_j. We thus obtain

$$\|\mathbf{w} - \boldsymbol{\mu}_{\mathbf{W}|\mathbf{X}'=\mathbf{x}'}\|_2^2 \leq \frac{1}{m^2} \left(\frac{25\ell}{2} + \frac{4}{\pi} \sum_{j=1}^\ell \frac{(x_j - \mu_j)^2}{\sigma_j^2} \right). \qquad (221)$$

We next note that, since $\mathbf{W} = \mathbf{Z} + \mathbf{U}$, and since $\mathbf{U}$ is independent from $\mathbf{Z}$ and i.i.d. on $[0, 1/m)$,

$$C_{\mathbf{W}|\mathbf{X}'} = C_{\mathbf{Z}} - \Lambda_{\mathbf{a}} C_{\mathbf{X}} \Lambda_{\mathbf{a}} + \frac{1}{12m^2} I_\ell. \qquad (222)$$

It can be shown that $C_{\mathbf{Z}} - \Lambda_{\mathbf{a}} C_{\mathbf{X}} \Lambda_{\mathbf{a}}$ is the conditional covariance matrix of $\mathbf{Z}$ given $\mathbf{X}'$, hence it is positive semidefinite.[6]

---

[6]Indeed, we have $C_{\mathbf{Z}\mathbf{X}} = C_{\mathbf{W}\mathbf{X}}$ and, by (209), $C_{\mathbf{Z}\mathbf{X}'} = C_{\mathbf{Z}\mathbf{X}} A$. Replacing in (216) $\mathbf{W}$ by $\mathbf{Z}$, and repeating the steps leading to (219), we obtain the desired result.

It follows that the smallest eigenvalue of $C_{\mathbf{W}|\mathbf{X}'}$ is lower-bounded by $\frac{1}{12m^2}$. Together with (221), this yields for the second term on the RHS of (214)

$$\frac{m^\ell}{2} \int_{\mathcal{C}(\mathbf{z},1/m)} (\mathbf{w} - \boldsymbol{\mu}_{\mathbf{W}|\mathbf{X}'=\mathbf{x}'})^\mathsf{T} C_{\mathbf{W}|\mathbf{X}'}^{-1} (\mathbf{w} - \boldsymbol{\mu}_{\mathbf{W}|\mathbf{X}'=\mathbf{x}'}) \mathrm{d}\mathbf{w}$$

$$\leq 6m^{\ell+2} \frac{1}{m^\ell} \frac{1}{m^2} \left(\frac{25\ell}{2} + \frac{4}{\pi} \sum_{j=1}^\ell \frac{(x_j - \mu_j)^2}{\sigma_j^2} \right)$$

$$= \frac{75\ell}{2} + \frac{24}{\pi} \sum_{j=1}^\ell \frac{(x_j - \mu_j)^2}{\sigma_j^2}. \tag{223}$$

To upper-bound the first term on the RHS of (214), we use that (222) combined with Lemma 11 implies that every diagonal element of $C_{\mathbf{W}|\mathbf{X}'}$ is given by

$$\mathsf{E}\left[(Z_j - \mathsf{E}[Z_j])^2\right] - a_j^2 \sigma_j^2 + \frac{1}{12m^2} = -(1 - a_j)^2 \sigma_j^2$$
$$+ \mathsf{E}\left[(X_j - \mu_j - Z_j + \mathsf{E}[Z_j])^2\right] + \frac{1}{12m^2}. \tag{224}$$

The first term on the RHS of (224) is negative, and the second term is upper-bounded by $\mathsf{E}\left[(X_j - Z_j)^2\right] \leq 1/m^2$. Hence, every element on the main diagonal of $C_{\mathbf{W}|\mathbf{X}'}$ is upper-bounded by $\frac{1+1/12}{m^2}$. It thus follows from Hadamard's inequality that

$$\log\left(m^\ell \sqrt{(2\pi)^\ell \det C_{\mathbf{W}|\mathbf{X}'}}\right) \leq \frac{\ell}{2}\log\left(2\pi\left(1 + \frac{1}{12}\right)\right). \tag{225}$$

Combining (223) and (225) with (214) and (210) yields

$$D(f_{\mathbf{W}_1^\ell} \| g_{\mathbf{W}_1^\ell})$$
$$\leq \frac{\ell}{2}\log\left(2\pi\left(1 + \frac{1}{12}\right)\right) + \frac{75\ell}{2} + \frac{24}{\pi}\sum_{j=1}^\ell \frac{\mathsf{E}\left[(X_j - \mu_j)^2\right]}{\sigma_j^2}$$
$$= \ell\left(\frac{1}{2}\log\left(2\pi\left(1 + \frac{1}{12}\right)\right) + \frac{75}{2} + \frac{24}{\pi}\right) \tag{226}$$

and completes the proof. ∎

APPENDIX F
SPECTRAL DISTRIBUTION FUNCTION OF $\{[\mathbf{X}_t]_m\}$

Let $\{\mathbf{X}_t\}$ be a stationary, L-variate, Gaussian process with mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_L)^\mathsf{T}$ and SDF $\mathsf{F}_\mathbf{X}$. Let $\{\mathbf{Z}_t\}$ and $\{\mathbf{N}_t\}$ be defined as $Z_{i,t} \triangleq [X_{i,t}]_m$ and $N_{i,t} \triangleq X_{i,t} - [X_{i,t}]_m$, respectively. For every pair $i,j = 1, \ldots, L$, we have

$$K_{N_i N_j}(\tau) = K_{X_i X_j}(\tau) + K_{Z_i Z_j}(\tau)$$
$$- K_{X_i Z_j}(\tau) - K_{Z_i X_j}(\tau). \tag{227}$$

Bussgang's theorem [26, eq. (20)] further yields that $K_{X_i Z_j}(\tau) = K_{Z_j X_i}(-\tau) = a_j K_{X_i X_j}(\tau)$, where a_j is defined in (45). Consequently,

$$K_{N_i N_j}(\tau) = K_{X_i X_j}(\tau) + K_{Z_i Z_j}(\tau)$$
$$- a_j K_{X_i X_j}(\tau) - a_i K_{X_j X_i}(-\tau)$$
$$= (1 - a_j - a_i) K_{X_i X_j}(\tau) + K_{Z_i Z_j}(\tau). \tag{228}$$

Since the SDF is fully determined by the covariance structure of a process [27, Th. 1, p. 206], we obtain (44).

To prove (47), namely,

$$\int_{-1/2}^{1/2} \mathrm{d}\mathsf{F}_{N_i}(\theta) \leq \frac{1}{m^2} \tag{229}$$

we note that

$$\int_{-1/2}^{1/2} \mathrm{d}\mathsf{F}_{N_i}(\theta) = \mathsf{E}\left[(X_{i,t} - Z_{i,t})^2\right] - (\mu_i - \mathsf{E}[Z_{i,t}])^2. \tag{230}$$

Since $|X_{i,t} - Z_{i,t}| \leq \frac{1}{m}$ and $(\mu_i - \mathsf{E}[Z_{i,t}])^2 \geq 0$, the claim (47) follows.

It remains to prove (46), namely,

$$|1 - a_i| \leq \frac{1}{m}\sqrt{\frac{2}{\pi\sigma_i^2}}. \tag{231}$$

Set $f(\alpha) \triangleq \frac{\alpha}{\sigma_i} e^{-\alpha^2/2\sigma_i^2}$, $\alpha \in \mathbb{R}$. We have

$$a_i = \frac{\mathsf{E}\left[(X_{i,t} - \mu_i)(Z_{i,t} - \mathsf{E}[Z_{i,t}])\right]}{\sigma_i^2}$$
$$= \frac{\mathsf{E}\left[(X_{i,t} - \mu_i)Z_{i,t}\right]}{\sigma_i^2}$$
$$= \frac{1}{\sqrt{2\pi\sigma_i^2}} \sum_i \frac{i}{m} \int_{\frac{i}{m}}^{\frac{i+1}{m}} f(\alpha)\mathrm{d}\alpha. \tag{232}$$

Furthermore,

$$\frac{1}{\sqrt{2\pi\sigma_i^2}} \int_{-\infty}^\infty \alpha f(\alpha)\mathrm{d}\alpha = 1. \tag{233}$$

It follows that

$$1 - a_i = \frac{1}{\sqrt{2\pi\sigma_i^2}} \sum_i \int_{\frac{i}{m}}^{\frac{i+1}{m}} \left(\alpha - \frac{i}{m}\right) f(\alpha)\mathrm{d}\alpha. \tag{234}$$

Since $|\alpha - i/m| \leq 1/m$ for $\alpha \in [i/m, (i+1)/m]$, this yields

$$|1 - a_i| \leq \frac{1}{\sqrt{2\pi\sigma_i^2}} \sum_i \int_{\frac{i}{m}}^{\frac{i+1}{m}} \left|\alpha - \frac{i}{m}\right| |f(\alpha)|\mathrm{d}\alpha$$
$$\leq \frac{1}{m} \frac{1}{\sqrt{2\pi\sigma_i^2}} \int_{-\infty}^\infty |f(\alpha)|\mathrm{d}\alpha$$
$$= \frac{1}{m}\sqrt{\frac{2}{\pi\sigma_i^2}}. \tag{235}$$

This proves (46) and concludes the proof of Lemma 11.

APPENDIX G
PROOF OF THEOREM 13

Let $\{\mathbf{Z}_t\}$ be a stationary, L-variate, complex-valued process with matrix-valued SDF $\mathsf{F}_\mathbf{Z}$. Let the *real composite process* $\{\hat{\mathbf{X}}_t\}$ be defined as $\hat{\mathbf{X}}_t \triangleq [\mathfrak{Re}(\mathbf{Z}_t^\mathsf{T}), \mathfrak{Im}(\mathbf{Z}_t^\mathsf{T})]^\mathsf{T}$. That is, $\hat{\mathbf{X}}_t$ is obtained by stacking the real part of $\mathbf{Z}_t$ on top of the imaginary part of $\mathbf{Z}_t$. Further let the *augmented process* $\{\hat{\mathbf{Z}}_t\}$ be defined as $\hat{\mathbf{Z}}_t \triangleq [\mathbf{Z}_t^\mathsf{T}, \mathbf{Z}_t^\mathsf{H}]^\mathsf{T}$. Clearly, $\hat{\mathbf{X}}_t$ and $\hat{\mathbf{Z}}_t$ satisfy $\hat{\mathbf{Z}}_t = T\hat{\mathbf{X}}_t$, where

$$T \triangleq \begin{bmatrix} I_L & \imath I_L \\ I_L & -\imath I_L \end{bmatrix} \tag{236}$$

is unitary up to a factor of 2, i.e., $TT^\mathsf{H} = T^\mathsf{H}T = 2I_L$. The matrix-valued autocovariance function of $\{\hat{\mathbf{Z}}_t\}$ reads

$$K_{\hat{\mathbf{Z}}}(\tau) = \begin{bmatrix} K_\mathbf{Z}(\tau) & \overline{K}_\mathbf{Z}(\tau) \\ \overline{K}_\mathbf{Z}^*(\tau) & K_\mathbf{Z}^*(\tau) \end{bmatrix} \tag{237}$$

where $\overline{K}_{\mathbf{Z}}$ denotes the pseudo-autocovariance function of $\{\mathbf{Z}_t\}$. The corresponding matrix-valued SDF is given by

$$\mathsf{F}_{\hat{\mathbf{Z}}}(\theta) = \left[\begin{array}{cc} \mathsf{F}_{\mathbf{Z}}(\theta) & \overline{\mathsf{F}}_{\mathbf{Z}}(\theta) \\ -\overline{\mathsf{F}}_{\mathbf{Z}}^{*}(-\theta) & -\mathsf{F}_{\mathbf{Z}}^{*}(-\theta) \end{array} \right] \qquad (238)$$

where $\overline{\mathsf{F}}_{\mathbf{Z}}$ satisfies

$$\overline{K}_{\mathbf{Z}}(\tau) = \int_{-1/2}^{1/2} e^{-\imath 2\pi\tau\theta} \mathrm{d}\overline{\mathsf{F}}_{\mathbf{Z}}(\theta), \quad \tau \in \mathbb{Z}. \qquad (239)$$

The autocovariance functions and SDFs of $\{\hat{\mathbf{X}}_t\}$ and $\{\hat{\mathbf{Z}}_t\}$ are related via

$$K_{\hat{\mathbf{Z}}}(\tau) = T K_{\hat{\mathbf{X}}}(\tau) T^{\mathsf{H}} \qquad (240)$$
$$\mathsf{F}_{\hat{\mathbf{Z}}}(\theta) = T \mathsf{F}_{\hat{\mathbf{X}}}(\theta) T^{\mathsf{H}}. \qquad (241)$$

By definition, $\overline{d}(\{\mathbf{Z}_t\}) = \overline{d}(\{\hat{\mathbf{X}}_t\})$. It thus follows from Theorem 10 that

$$\overline{d}(\{\mathbf{Z}_t\}) = \overline{d}(\{\hat{\mathbf{X}}_t\}) \leq \int_{-1/2}^{1/2} \mathrm{rank}(\mathsf{F}_{\hat{\mathbf{X}}}'(\theta)) \mathrm{d}\theta. \qquad (242)$$

Since left or right multiplication by a nonsingular matrix leaves the rank unchanged, we obtain from (241) that the rank of $\mathsf{F}_{\hat{\mathbf{X}}}'(\theta)$ is equal to the rank of $\mathsf{F}_{\hat{\mathbf{Z}}}'(\theta)$. Furthermore, by (238), the rank of $\mathsf{F}_{\hat{\mathbf{Z}}}'(\theta)$ is upper-bounded by the rank of $\mathsf{F}_{\mathbf{Z}}'(\theta)$ plus the rank of $(\overline{\mathsf{F}}_{\mathbf{Z}}')^{*}(-\theta)$ [28, Th. 1]. Consequently,

$$\overline{d}(\{\mathbf{Z}_t\}) \leq \int_{-1/2}^{1/2} \mathrm{rank}(\mathsf{F}_{\mathbf{Z}}'(\theta)) + \mathrm{rank}((\overline{\mathsf{F}}_{\mathbf{Z}}')^{*}(-\theta)) \mathrm{d}\theta$$
$$= 2 \int_{-1/2}^{1/2} \mathrm{rank}(\mathsf{F}_{\mathbf{Z}}'(\theta)) \mathrm{d}\theta \qquad (243)$$

where the second step follows because complex conjugation does not affect the rank.

If $\{\mathbf{Z}_t\}$ is Gaussian, then (242) holds with equality by Theorem 10. If $\{\mathbf{Z}_t\}$ is, in addition, proper then $\overline{K}_{\mathbf{Z}}(\tau) = 0$, so the derivative of $\overline{\mathsf{F}}_{\mathbf{Z}}$ is zero almost everywhere. Hence, the derivative of $\mathsf{F}_{\hat{\mathbf{Z}}}$ becomes block diagonal almost everywhere and its rank equals the sum of the ranks of its diagonal elements. We conclude that, if $\{\mathbf{Z}_t\}$ is proper Gaussian, then (243) holds with equality. This proves Theorem 13.

APPENDIX H
APPENDIX TO SECTION V

A. Proof of Theorem 14

For every $m = 2, 3, \dots$ and $k = 1, 2, \dots$ we have

$$H([X_1]_m | X_{-\infty}^0) \leq H([X_k]_m | [X_{-\infty}^{k-1}]_m)$$
$$\leq H([X_k]_m | [X_1^{k-1}]_m) \qquad (244)$$

by stationarity; and because conditioning reduces entropy and, conditioned on $X_{-\infty}^0$, $[X_1]_m$ is independent of $[X_{-\infty}^0]_m$. Note that, by (4) and stationarity,

$$H([X_k]_m | [X_{-\infty}^{k-1}]_m) = H'(\{[X_t]_m\}). \qquad (245)$$

Thus, dividing (244) by $\log m$ and taking first the limit over m and then the limit over k yields

$$\overline{d}(X_1 | X_{-\infty}^0) \leq \overline{d}(\{X_t\}) \leq \overline{d}'(\{X_t\}). \qquad (246)$$

This proves (73).

We next bound the difference $\overline{d}'(\{X_t\}) - \overline{d}(\{X_t\})$. By (245), we have

$$H([X_k]_m | [X_1^{k-1}]_m) - H'(\{[X_t]_m\})$$
$$= I([X_k]_m; [X_{-\infty}^0]_m | [X_1^{k-1}]_m). \qquad (247)$$

Dividing (247) by $\log m$ and taking first the limit over m and then the limit over k yields

$$\lim_{k\to\infty} \varliminf_{m\to\infty} \frac{I([X_k]_m; [X_{-\infty}^0]_m | [X_1^{k-1}]_m)}{\log m}$$
$$\leq \overline{d}'(\{X_t\}) - \overline{d}(\{X_t\})$$
$$\leq \lim_{k\to\infty} \varlimsup_{m\to\infty} \frac{I([X_k]_m; [X_{-\infty}^0]_m | [X_1^{k-1}]_m)}{\log m}. \qquad (248)$$

This concludes the proof of Theorem 14.

B. Proof of Corollary 15

Suppose there exists a nonnegative n such that

$$I(X_1^k; X_{-\infty}^{-n}) < \infty, \quad k = 1, 2, \dots \qquad (249)$$

We first show that

$$I([X_k]_m; [X_{-\infty}^0]_m | [X_1^{k-1}]_m) \leq \frac{I([X_1^k]_m; [X_{-\infty}^0]_m)}{k}. \qquad (250)$$

In a second step, we then show that (249) implies that

$$\lim_{k\to\infty} \varlimsup_{m\to\infty} \frac{I([X_1^k]_m; [X_{-\infty}^0]_m)}{k \log m} = 0 \qquad (251)$$

which together with (250) and (74) demonstrates that $\overline{d}(\{X_t\}) = \overline{d}'(\{X_t\})$, thus proving Corollary 15.

To prove (250), we use the chain rule, stationarity, and the fact that conditioning reduces entropy, to obtain

$$I([X_1^k]_m; [X_{-\infty}^0]_m)$$
$$= \sum_{\ell=1}^{k} \Big[H([X_\ell]_m | [X_1^{\ell-1}]_m) - H([X_\ell]_m | [X_{-\infty}^{\ell-1}]_m) \Big]$$
$$\geq \sum_{\ell=1}^{k} \Big[H([X_k]_m | [X_1^{k-1}]_m) - H([X_k]_m | [X_{-\infty}^{k-1}]_m) \Big]$$
$$= k I([X_k]_m; [X_{-\infty}^0]_m | [X_1^{k-1}]_m). \qquad (252)$$

Having obtained (250), we next show that (249) implies (251). Indeed,

$$\varlimsup_{k\to\infty} \varlimsup_{m\to\infty} \frac{I([X_1^k]_m; [X_{-\infty}^0]_m)}{k \log m}$$
$$\leq \varlimsup_{k\to\infty} \varlimsup_{m\to\infty} \frac{I([X_1^k]_m; [X_{-\infty}^{-n}]_m)}{k \log m}$$
$$+ \varlimsup_{k\to\infty} \varlimsup_{m\to\infty} \frac{I([X_1^k]_m; [X_{-n+1}^0]_m | [X_{-\infty}^{-n}]_m)}{k \log m}$$
$$\leq \varlimsup_{k\to\infty} \varlimsup_{m\to\infty} \frac{I(X_1^k; X_{-\infty}^{-n})}{k \log m}$$
$$+ \varlimsup_{k\to\infty} \varlimsup_{m\to\infty} \frac{H([X_{-n+1}^0]_m)}{k \log m} \qquad (253)$$

where n is a nonnegative integer satisfying (249). Here, the first inequality follows from the chain rule; the second inequality follows from the data processing inequality and by upperbounding the second mutual information by $H([X^0_{-n+1}]_m)$.

The first limit on the RHS of (253) is zero because, by assumption, $I(X^k_1; X^{-n}_{-\infty}) < \infty$. The second limit on the RHS of (253) can be written as $\varlimsup_{k\to\infty} \bar{d}(X^0_{-n+1})/k$, which is zero because, by Lemma 1, $\bar{d}(X^0_{-n+1})$ is bounded in k. This proves (251) and concludes the proof of Corollary 15.

C. Proof of Lemma 16

Since $\{X_t\}$ is Gaussian, the conditional mean of X_k given $X_0, \ldots, X_{k-1}$ can be written as

$$\mathsf{E}[X_k | X_0, \ldots, X_{k-1}] = \sum_{\ell=1}^{k} \alpha_\ell X_{k-\ell} \qquad (254)$$

for some coefficients $\alpha_1, \ldots, \alpha_k$.[7] The conditional variance σ_k^2 is thus given by (see, e.g., [19, Sec. 10.6])

$$\sigma_k^2 = \mathsf{E}\left[\left(X_k - \sum_{\ell=1}^{k} \alpha_\ell X_{k-\ell}\right)^2\right]$$
$$= \int_{-\frac{1}{2}}^{\frac{1}{2}} \left|1 - \sum_{\ell=1}^{k} \alpha_\ell e^{-\imath 2\pi \ell\theta}\right|^2 \mathrm{d}\mathsf{F}_X(\theta). \qquad (255)$$

The function

$$g(\theta) = 1 - \sum_{\ell=1}^{k} \alpha_\ell e^{-\imath 2\pi \ell\theta}, \quad -1/2 \le \theta \le 1/2 \qquad (256)$$

is analytic on the closed interval $[-1/2, 1/2]$, hence it is either constant or it has at most finitely many zeros in $[-1/2, 1/2]$. Moreover, g cannot be the all-zero function, as can be argued by contradiction. Indeed, suppose there exist $\alpha_1, \ldots, \alpha_k$ such that $g(\theta) = 0$ for all θ. Then, by (255), we have $\sigma_k^2 = 0$ irrespective of F_X. In other words, we can find a linear estimator that perfectly predicts X_k from $X_0, \ldots, X_{k-1}$ irrespective of the SDF of $\{X_t\}$. This is clearly a contradiction, since even the best predictor yields $\sigma_k^2 = \sigma^2$ for an i.i.d., zero-mean, variance-σ^2, Gaussian process, i.e., when $\mathsf{F}'_X(\theta) = \sigma^2$. Thus, the set $\mathcal{Z} \triangleq \{\theta : g(\theta) = 0\}$ is finite and has therefore Lebesgue measure zero.

Since $|g(\theta)|^2 = 0$ for $\theta \in \mathcal{Z}$, we have

$$\sigma_k^2 = \int_{\mathcal{Z}^c} |g(\theta)|^2 \mathrm{d}\mathsf{F}_X(\theta). \qquad (257)$$

Since furthermore $|g(\theta)|^2 > 0$ for $\theta \in \mathcal{Z}^c$, we have $\sigma_k^2 = 0$ only if

$$\int_{\mathcal{Z}^c} \mathrm{d}\mathsf{F}_X(\theta) = 0. \qquad (258)$$

This implies that $\mathsf{F}'_X(\theta) = 0$ for all $\theta \in \mathcal{Z}^c$. Hence, the set of harmonics θ for which $\mathsf{F}'_X(\theta) > 0$ is contained in $\mathcal{Z}$. The proof is completed by the monotonicity of measures and the fact that $\mathcal{Z}$ has Lebesgue measure zero.

---

[7]More precisely, the coefficients correspond to the LMMSE estimator for estimating X_k from $X_0, \ldots, X_{k-1}$. The LMMSE estimator always exists, even though it is not necessarily unique.

REFERENCES

[1] A. Rényi, "On the dimension and entropy of probability distributions," *Acta Mathematica Hungarica*, vol. 10, no. 1-2, pp. 193–215, Mar. 1959.
[2] T. Kawabata and A. Dembo, "The rate-distortion dimension of sets and measures," *IEEE Trans. Inf. Theory*, vol. 40, no. 5, pp. 1564–1572, Sep. 1994.
[3] T. Koch, "The Shannon lower bound is asymptotically tight," *IEEE Trans. Inf. Theory*, vol. 62, no. 11, pp. 6155–6161, Nov. 2016.
[4] Y. Wu and S. Verdú, "Rényi information dimension: Fundamental limits of almost lossless analog compression," *IEEE Trans. Inf. Theory*, vol. 56, no. 8, pp. 3721–3748, Aug. 2010.
[5] Y. Wu, S. Shamai (Shitz), and S. Verdú, "Information dimension and the degrees of freedom of the interference channel," *IEEE Trans. Inf. Theory*, vol. 61, no. 1, pp. 256–279, Jan. 2015.
[6] D. Stotz and H. Bölcskei, "Degrees of freedom in vector interference channels," *IEEE Trans. Inf. Theory*, vol. 62, no. 7, pp. 4172–4197, Jul. 2016.
[7] S. Jalali and H. V. Poor, "Universal compressed sensing for almost lossless recovery," *IEEE Trans. Inf. Theory*, vol. 63, no. 5, pp. 2933–2953, May 2017.
[8] F. E. Rezagah, S. Jalali, E. Erkip, and H. V. Poor, "Compression-based compressed sensing," *IEEE Trans. Inf. Theory*, vol. 63, no. 10, pp. 6735–6752, Oct. 2017.
[9] Y. Gutman and A. Śpiewak, "Metric mean dimension and analog compression," `arXiv:1812.00458 [math.DS]`, Dec. 2018.
[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 1st ed. Wiley Interscience, 1991.
[11] R. C. Bradley, "Basic properties of strong mixing conditions. A survey and some open questions," *Probab. Surveys*, vol. 2, pp. 107–144, 2005.
[12] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd ed. Cambridge: Cambridge University Press, 2013.
[13] Y. Wu, "Shannon theory for compressed sensing," Ph.D. dissertation, Princeton University, 2011.
[14] R. G. Gallager, *Information Theory and Reliable Communication*. John Wiley & Sons, 1968.
[15] N. Wiener and P. Masani, "The prediction theory of multivariate stochastic processes," *Acta Mathematica*, vol. 98, no. 1, pp. 111–150, 1957.
[16] M. B. Priestley, *Spectral Analysis and Time Series*. Elsevier Academic Press, 2005.
[17] R. M. Gray, *Entropy and Information Theory*, 2nd ed. New York, NY: Springer, 2011.
[18] A. Papoulis and U. S. Pillai, *Probability, Random Variables and Stochastic Processes*, 4th ed. New York, NY: McGraw Hill, 2002.
[19] U. Grenander and G. Szegö, *Toeplitz Forms and Their Applications*. University of California Press, 1958.
[20] Y. Wu and S. Verdú, "Optimal phase transitions in compressed sensing," *IEEE Trans. Inf. Theory*, vol. 58, no. 10, pp. 6241–6263, Oct. 2012.
[21] A. V. Balakrishnan, "A note on the sampling principle for continuous signals," *IRE Trans. Inf. Theory*, vol. 3, no. 2, pp. 143–146, Jun. 1957.
[22] R. B. Ash and C. A. Doléans-Dade, *Probability and Measure Theory*, 2nd ed. Elsevier/Academic Press, 2000.
[23] I. Csiszár, "On an extremum problem of information theory," *Stud. Sci. Math. Hung.*, vol. 9, pp. 57–71, 1974.
[24] A. Lapidoth, "On the asymptotic capacity of stationary Gaussian fading channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 2, pp. 437–446, Feb. 2005.
[25] ——, *A Foundation in Digital Communication*. Cambridge: Cambridge University Press, 2009.
[26] J. J. Bussgang, "Crosscorrelation functions of amplitude-distorted Gaussian signals," Res. Lab. of Electronics, M.I.T., Cambridge, Mass., Tech. Rep. 216, Mar. 1952.
[27] I. I. Gihman and A. V. Skorohod, *The Theory of Stochastic Processes I*. Springer Verlag, 1980.
[28] M. Lundquist and W. Barrett, "Rank inequalities for positive semidefinite matrices," *Linear Algebra and its Applications*, vol. 248, pp. 91–100, 1996.

Bernhard C. Geiger (S'07–M'14–SM'19) was born in Graz, Austria, in 1984. He received the Dipl.-Ing. degree in electrical engineering (with distinction) and the Dr. techn. degree in electrical and information engineering (with distinction) from Graz University of Technology, Austria, in 2009 and 2014, respectively.

In 2009 he joined the Signal Processing and Speech Communication Laboratory, Graz University of Technology, as a Project Assistant and took a position as a Research and Teaching Associate at the same lab in 2010. He was a Senior Scientist and Erwin Schrödinger Fellow at the Institute for Communications Engineering, Technical University of Munich, Germany from 2014 to 2017 and a postdoctoral researcher at the Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria from 2017 to 2018. He is currently Senior Researcher at Know-Center GmbH, Graz, Austria. His research interests cover information theory for machine learning and information-theoretic model reduction for Markov chains and hidden Markov models.

Tobias Koch (S'02–M'09–SM'16) is a Visiting Professor and Ramón y Cajal Research Fellow with the Signal Theory and Communications Department of Universidad Carlos III de Madrid (UC3M). He received the M.Sc. degree in electrical engineering (with distinction) in 2004 and the Ph.D. degree in electrical engineering in 2009, both from ETH Zurich, Switzerland. From June 2010 until May 2012 he was a Marie Curie Intra-European Research Fellow with the University of Cambridge, UK. He was also a research intern at Bell Labs, Murray Hill, NJ, USA in 2004, and the Universitat Pompeu Fabra (UPF), Barcelona, Spain, in 2007. He joined the Signal Processing Group of UC3M in June 2012. His research interests are in digital communication theory and information theory.

Dr. Koch received a Starting Grant from the European Research Council (ERC), a Ramón y Cajal Research Fellowship, a Marie Curie Intra-European Fellowship, a Marie Curie Career Integration Grant, and a Fellowship for Prospective Researchers from the Swiss National Science Foundation. In 2013–2016 he served as Vice Chair of the Spain Chapter of the IEEE Information Theory Society.