On Maximum Rényi Entropy Rate

Christoph Bunte and Amos Lapidoth
ETH Zurich
Switzerland
Email: {bunte,lapidoth}@isi.ee.ethz.ch

Abstract—**We compute the supremum of the Rényi entropy rate over the class of stationary stochastic processes having autocovariance sequences that begin with $p+1$ given values. Our results are closely related to Burg's maximum entropy theorem on the supremum over the same class but of the Shannon entropy rate.**

I. Introduction

Motivated by spectral estimation, Burg found the maximum of the differential Shannon entropy rate over the class of stationary stochastic processes whose autocovariance sequences begin with $p+1$ given values [1], [2, Theorem 12.6.1]. Here we consider the same class, but we maximize a different objective function: the Rényi entropy rate.

To recall the definition of the Rényi rate of a stochastic process, we begin with the Rényi entropy of a random vector or of a (joint) density. The order-α Rényi entropy of a probability density function (PDF) f is defined as

$$h_\alpha(f) = \frac{1}{1-\alpha} \log \int_{-\infty}^{\infty} f(x)^\alpha \,\mathrm{d}x, \tag{1}$$

where α can be any positive number other than one. The integral on the RHS of (1) always exists, possibly taking on the value $+\infty$, in which case we define $h_\alpha(f) = +\infty$ if $0 < \alpha < 1$ and $h_\alpha(f) = -\infty$ if $\alpha > 1$. When a random variable (or random vector) X is of density f_X we sometimes write $h_\alpha(X)$ instead of $h_\alpha(f_X)$.

The order-α Rényi entropy rate (or "Rényi rate" for short) of a stochastic process (SP) $\{X_k\}$ is defined as

$$h_\alpha(\{X_k\}) = \lim_{n\to\infty} \frac{1}{n} h_\alpha\bigl(X_1^n\bigr),$$

whenever the limit exists. Here we use the notation X_i^j to denote the tuple $(X_i, \dots, X_j)$.

The Rényi entropy rate of finite-state Markov chains was computed by Rached, Alajaji, and Campbell [3] with extensions to countable state space in [4].[1] The Rényi entropy rate of stationary Gaussian processes was found by Golshani and Pasha in [5]. Extensions to other types of rate are explored in [6].

The Rényi entropy is closely related to the differential Shannon entropy:

$$h(f) = -\int_{-\infty}^{\infty} f(x) \log f(x) \,\mathrm{d}x. \tag{2}$$

(The integral on the RHS of (2) need not exist. If it does not, then we say that $h(f)$ does not exist.) Under some mild technical conditions [7],

$$h_\alpha(f) \le h(f), \qquad \text{for } \alpha > 1; \tag{3}$$

$$h_\alpha(f) \ge h(f), \qquad \text{for } 0 < \alpha < 1; \tag{4}$$

and

$$\lim_{\alpha\to 1} h_\alpha(f) = h(f). \tag{5}$$

The entropy of a pair of independent random variables is the sum of the individual entropies. This is true for both differential Shannon entropy and Rényi entropy. But the two entropies behave differently when the random variables are dependent. While the differential Shannon entropy of a pair is always upper-bounded by the sum of the individual entropies, this need not hold for Rényi entropy: the Rényi entropy of a random vector can exceed the sum of the Rényi entropies of its components. Consequently, the random vector of highest Rényi entropy among all those whose components have some prespecified distribution need not have independent components. This is, of course, also true if the distributions of the components are not specified but only constrained.[2] Likewise, the supremum of the Rényi *rate* subject to constraints on the marginal distribution is not achieved by memoryless processes [8].

Here we focus on the supremum of the Rényi rate subject to autocovariance constraints. We show that the solution exhibits a dichotomy: when the order α is smaller than one, the supremum is infinite; and when it is greater than one the supremum is the same as if we were maximizing the Shannon rate (with the supremum thus being computable using Burg's theorem). Note, however, that the supremum—unlike the supremum in Burg's theorem—is not achieved by a Gauss-Markov process. It is, however, approachable by stochastic processes having the same autocovariance sequence as the Gauss-Markov process.

II. Preliminaries

Key to our results is the following proposition [8, Corollary 4]:

Proposition 1 (Rényi Rate under a Variance Constraint)**.**

1) For every $\alpha > 1$, every $\sigma > 0$, and every $\varepsilon > 0$ there exists a centered stationary SP $\{Y_k\}$ whose Rényi entropy

[1] In the discrete case the density in (1) is replaced by the probability mass function, and the integral is replaced by a sum.

[2] Nevertheless, the maximization of Rényi entropy subject to linear constraints does typically have a simple solution [8].

rate exceeds $\frac{1}{2}\log(2\pi e\sigma^2)-\varepsilon$ *and which satisfies*

$$\mathrm{E}[Y_k Y_{k'}] = \sigma^2\,\mathrm{I}\{k=k'\}, \tag{6}$$

where $\mathrm{I}\{\text{statement}\}$ *is* 1 *when* statement *is true and* 0 *when it is not.*

2) *For every* $0<\alpha<1$, *every* $\sigma>0$, *and every* $\mathsf{M}>0$ *there exists a centered stationary SP* $\{Y_k\}$ *whose Rényi entropy rate exceeds* M *and which satisfies* (6).

To address some technical boundary issues we shall also need the following lemma.

Lemma 2. *Let* $f_1,\ldots,f_p$ *be probability density functions on* $\mathbb{R}^n$, *and let* $q_1,\ldots,q_p\geq 0$ *be nonnegative numbers that sum to one. Let* f *be the mixture density*

$$f(\mathbf{x}) = \sum_{\ell=1}^{p} q_\ell f_\ell(\mathbf{x}), \quad \mathbf{x}\in\mathbb{R}^n.$$

Then

$$h_\alpha(f) \geq \min_{1\leq\ell\leq p} h_\alpha(f_\ell).$$

Proof. For $0<\alpha<1$ this follows by the concavity of Rényi entropy. Consider now $\alpha>1$:

$$\begin{aligned}\log\int f^\alpha(\mathbf{x})\,\mathrm{d}\mathbf{x} &= \log\int\left(\sum_{\ell=1}^{p} q_\ell f_\ell(\mathbf{x})\right)^\alpha \mathrm{d}\mathbf{x}\\ &\leq \log\int\sum_{\ell=1}^{p} q_\ell f_\ell^\alpha(\mathbf{x})\,\mathrm{d}\mathbf{x}\\ &= \log\left(\sum_{\ell=1}^{p} q_\ell\int f_\ell^\alpha(\mathbf{x})\,\mathrm{d}\mathbf{x}\right)\\ &\leq \log\max_{1\leq\ell\leq p}\int f_\ell^\alpha(\mathbf{x})\,\mathrm{d}\mathbf{x}\\ &= \max_{1\leq\ell\leq p}\log\int f_\ell^\alpha(\mathbf{x})\,\mathrm{d}\mathbf{x},\end{aligned}$$

from which the claim follows because $1/(1-\alpha)$ is negative. □

III. Results

Given $\alpha_0,\ldots,\alpha_p\in\mathbb{R}$, we consider the family of all stochastic processes $X_1,X_2,\ldots$ for which

$$\mathrm{E}[X_i X_{i+k}] = \alpha_k, \quad \Big(i\in\mathbb{N},\ k\in\{0,\ldots,p\}\Big). \tag{7}$$

We assume that the $(p+1)\times(p+1)$ matrix whose Row-ℓ Column-m element is $\alpha_{|\ell-m|}$ is positive definite. This implies [9] that there exist constants $a_1,\ldots,a_p,\sigma^2$ and a $p\times p$ positive definite matrix K_p such that the following holds:[3] if the random p-vector $(W_{1-p},\ldots,W_0)$ is of second-moment matrix K_p (not necessarily centered) and if $\{Z_i\}_{i=1}^\infty$ are independent of $(W_{1-p},\ldots,W_0)$ with

$$\mathrm{E}[Z_i] = 0, \qquad i\in\mathbb{N}, \tag{8a}$$

$$\mathrm{E}[Z_i Z_j] = \sigma^2\,\mathrm{I}\{i=j\}, \qquad i,j\in\mathbb{N}, \tag{8b}$$

then the process defined inductively via

$$X_i = \sum_{k=1}^{p} a_i X_{i-k} + Z_i, \quad i\in\mathbb{N} \tag{9}$$

with the initialization

$$(X_{1-p},\ldots,X_0) = (W_{1-p},\ldots,W_0) \tag{10}$$

satisfies the constraints (7).

By Burg's maximum entropy theorem [2, Theorem 12.6.1], of all stochastic processes satisfying (7) the one of highest (differential) Shannon entropy rate is the p-th order Gauss-Markov process. It is obtained when $(W_{1-p},\ldots,W_0)$ is a centered Gaussian and $\{Z_i\}$ are IID $\sim\mathcal{N}\big(0,\sigma^2\big)$. Its Shannon entropy rate is

$$\lim_{n\to\infty}\frac{1}{n}h(X_1,\ldots,X_n) = \frac{1}{2}\log(2\pi e\sigma^2).$$

Our interest is in the maximum Rényi entropy rate.

Theorem 3. *The supremum of the order-*α *Rényi entropy rate over all stochastic processes satisfying* (7) *is* $+\infty$ *for* $0<\alpha<1$ *and is equal to the Shannon entropy rate of the* p*-th order Gauss-Markov process for* $\alpha>1$.

Proof. We first consider the case where $\alpha>1$. Let $a_1,\ldots,a_p,\sigma^2$ and K_p be as above, and let $\varepsilon>0$ be arbitrarily small. By Proposition 1 there exists a stochastic process $\{Z_i\}$ such that (8) holds and such that

$$\lim_{n\to\infty}\frac{1}{n}h_\alpha(Z_1,\ldots,Z_n) \geq \frac{1}{2}\log(2\pi e\sigma^2)-\varepsilon. \tag{11}$$

The matrix K_p is positive definite, so by the spectral representation theorem we can find vectors $\mathbf{w}_1,\ldots,\mathbf{w}_p\in\mathbb{R}^p$ and constants $q_1,\ldots,q_p>0$ with $q_1+\cdots+q_p=1$ such that

$$\mathsf{K}_p = \sum_{\ell=1}^{p} q_\ell \mathbf{w}_\ell \mathbf{w}_\ell^\mathsf{T}. \tag{12}$$

(The vectors are eigenvectors of K_p, and the constants $q_1,\ldots,q_p$ are the scaled eigenvalues of K_p.) Draw the random vector $\mathbf{W}$ independently of $\{Z_i\}$ with

$$\Pr[\mathbf{W}=\mathbf{w}_\ell] = q_\ell,$$

so that, by (12),

$$\mathrm{E}[\mathbf{W}\mathbf{W}^\mathsf{T}] = \mathsf{K}_p.$$

Construct now the stochastic process $\{X_i\}$ using (9) initialized with $(X_{1-p},\ldots,X_0)^\mathsf{T}$ being set to $\mathbf{W}$.

The resulting stochastic process thus satisfies the constraints (7). We next study its Rényi entropy rate. To that end, we study the Rényi entropy of the vector X_1^n. Let $f_\mathbf{X}$ denote its density, and let $f_{\mathbf{X}|\mathbf{w}_\ell}$ denote its conditional density given $\mathbf{W}=\mathbf{w}_\ell$, so

$$f_\mathbf{X}(\mathbf{x}) = \sum_{\ell=1}^{p} q_\ell f_{\mathbf{X}|\mathbf{w}_\ell}(\mathbf{x}), \quad \mathbf{x}\in\mathbb{R}^n.$$

[3] The Row-ℓ Column-m element element of the matrix K_p is $\alpha_{|\ell-m|}$.

Consequently, by Lemma 2,

$$h_\alpha(f_{\mathbf{X}}) \geq \min_{1\leq \ell \leq p} h_\alpha(f_{\mathbf{X}|\mathbf{w}_\ell}). \tag{13}$$

We next study $h_\alpha(f_{\mathbf{X}|\mathbf{w}_\ell})$ for any given $\ell \in \{1,\dots,p\}$. Recalling that $\mathbf{W}$ and $\{Z_i\}$ are independent, we conclude that, conditional on $\mathbf{W} = \mathbf{w}_\ell$, the random variables $X_1,\dots,X_n$ are generated inductively via (9) with the initialization

$$(X_{1-p},\dots,X_0)^\mathsf{T} = \mathbf{w}_\ell.$$

Conditionally on $\mathbf{W} = \mathbf{w}_\ell$, the random variables $X_1,\dots,X_n$ are thus an affine transformation of $Z_1,\dots,Z_n$. The transformation is of unit Jacobian, and thus

$$h_\alpha(f_{\mathbf{X}|\mathbf{w}_\ell}) = h_\alpha(Z_1,\dots,Z_n), \quad \ell \in \{1,\dots,p\}. \tag{14}$$

From this and (13) it follows that

$$h_\alpha(f_{\mathbf{X}}) \geq h_\alpha(Z_1,\dots,Z_n).$$

Dividing by n and using (11) establishes the result.

We next turn to the case $0 < \alpha < 1$. For every $\mathsf{M} > 0$ arbitrarily large, we use Proposition 1 to construct $\{Z_i\}$ as above but with

$$\lim_{n\to\infty} \frac{1}{n} h_\alpha(Z_1,\dots,Z_n) \geq \mathsf{M}.$$

The proof continues as for the case where α exceeds one. □

IV. Discussion

Theorem 3 has bearing on the spectral estimation problem, i.e., the problem of extrapolating the values of the autocovariance sequence from its first $p+1$ values. One approach is to choose the extrapolated sequence to be the autocovariance sequence of the stochastic process that—among all stochastic processes that have an autocovariance sequence that starts with these $p+1$ values—maximizes the Shannon rate, namely the p-th order Gauss-Markov process (Burg's theorem).

A different approach might be to choose some $\alpha > 1$ and to replace the maximization of the Shannon rate with that of the order-α Rényi rate. As we next argue, Theorem 3 shows that this would result in the same extrapolated sequence. Indeed, inspecting the proof of the theorem we see that the stochastic process $\{X_i\}$ that we constructed, while not a Gauss-Markov process, has the same autocovariance sequence as the p-th order Gauss-Markov process that satisfies the constraints. And, for $\alpha > 1$ the supremum can only be achieved by a stochastic process of this autocovariance sequence: for any other autocovariance function the Rényi rate is upper bounded by the Shannon rate (because $\alpha > 1$), and the latter is upper bounded by the Shannon rate of the Gaussian process, which, unless the autocovariance sequence is that of the p-th order Gauss-Markov process, is strictly smaller than the supremum (Burg's theorem).

Acknowledgment

Discussions with Stefan M. Moser are gratefully acknowledged.

References

[1] J. P. Burg, "Maximum entropy spectral analysis,," in *Proc. 37th Meet. Society of Exploration Geophysicists, 1967. Reprinted in* Modern Spectrum Analysis, *D. G. Childers, Ed. New York: IEEE Press, 1978 pp. 34–41.*, 1967.

[2] T. Cover and J. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ: John Wiley & Sons, 2006.

[3] Z. Rached, F. Alajaji, and L. Campbell, "Rényi's divergence and entropy rates for finite alphabet Markov sources," *IEEE Trans. Inf. Theory*, vol. 47, no. 4, pp. 1553–1561, May 2001.

[4] L. Golshani, E. Pasha, and G. Yari, "Some properties of Rényi entropy and Rényi entropy rate," *Information Sciences*, vol. 179, no. 14, pp. 2426–2433, 2009.

[5] L. Golshani and E. Pasha, "Rényi entropy rate for Gaussian processes," *Information Sciences*, vol. 180, no. 8, pp. 1486–1491, 2010.

[6] M. Khodabin, "ADK entropy and ADK entropy rate in irreducible-aperiodic Markov chain and Gaussian processes," *Journal of the Iranian Statistical Society*, vol. 9, no. 2, pp. 115–126, 2010.

[7] L. Wang and M. Madiman, "Beyond the entropy power inequality, via rearrangements," July 2013, arXiv:1307.6018.

[8] C. Bunte and A. Lapidoth, in *Proc. of the 2014 IEEE 28-th Convention of Electrical and Electronics Engineers in Israel*, Eilat, Israel, December 3–5 2014.

[9] M. Pourahmadi, *Foundations of Time Series Analysis and Prediction Theory*, ser. Wiley Series in Probability and Statistics. Wiley, 2001.