# The maximum entropy ansatz in the absence of a time arrow: fractional-pole models

Tryphon T. Georgiou , *IEEE Fellow*

**Abstract**

The maximum entropy ansatz, as it is often invoked in the context of time-series analysis, suggests the selection of a power spectrum which is consistent with autocorrelation data and corresponds to a random process least predictable from past observations. We introduce and compare a class of spectra with the property that the underlying random process is least predictable at any given point from the complete set of past and future observations. In this context, randomness is quantified by the size of the corresponding *smoothing error* and deterministic processes are characterized by integrability of the inverse of their power spectral densities—as opposed to the log-integrability in the classical setting. The power spectrum which is consistent with a partial autocorrelation sequence and corresponds to the *most random process* in this new sense, is no longer rational but generated by finitely many fractional-poles.

**Index Terms**

Entropy rate, randomness, time-arrow, predictability, smoothing.

## I. Introduction

THERE is a special place reserved in the spectral analysis literature for the maximum entropy ansatz, and rightly so, due to the multitude of analytic, computational, and practical qualities of maximum entropy spectra. The relevant theory is firmly rooted in analytic interpolation, the moment problem, and the Hilbert space geometry of random processes. The *maximum entropy* (*ME*) *ansatz*, in its basic form, calls for selecting the unique power spectrum which is consistent with a known finite set of autocorrelation moments and is the maximizer of a convex logarithmic functional which represents the entropy rate of the underlying random process. A closely related alternative justification relies on the fact that this *maximum entropy process* (*ME process*) is the least predictable from past observations and hence, it represents a worst-case situation.

The entropy rate of a random process is an inherently time-dependent concept. This fact becomes apparent in multivariable prediction theory where the variance of optimal least-variance predictors depends on the choice of the time-arrow [6, Remark 3]. It is our contention that often there is no natural direction of time. This is the case when statistics are obtained from an array of sensors and the index of the autocorrelation moments represents spacial separation. It is the case when we consider sparse records with both, past and future data available but with possible gaps. It is also the case, when we want to estimate the power spectrum and have no plans to use it for prediction in one way or another. In all such cases the rationale of the ME ansatz may be called into question. Hence, the purpose of this work is to study a time-arrow independent counterpart. In this, a power spectrum is selected so that the underlying random process evaluated at any point in time is the least predictable from the complete set of all other past and future values. In other words, it is the variance of the optimal smoothing filter which is sought to be maximal, as opposed to the variance of the optimal (time-arrow dependent) predictor.

Power spectra which are consistent with a finite set of (contiguous) autocorrelation statistics and correspond to a worst-case smoothing error for the relevant random process, turn out to have an all-pole representation, very much like the ones that result in from the ME ansatz but with one important

Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455; tryphon@ece.umn.edu

difference. These spectra are inverses of the *square root* of positive trigonometric polynomials, and hence, their poles are fractional. They also share a similar property with ME spectra in that they are extrema of a corresponding convex functional —which, however, is no longer logarithmic. Computation of their respective parameters is slightly more involved than having to solve linear (Yule-Walker-Levinson) equations. They can be computed as fixed points of suitable differential equations originating from a homotopy-based method in determining functional extrema. For convenience, and lacking a better terminology, we refer to this new class of spectra and the respective processes as *most random* (*MR*).

The maximum entropy ansatz has a fifty year history or more. We will not attempt to overview significant milestones but refer to [13] for textbook exposition of relevant material, to [9], [11] for an overview of relevant research in signal processing, to Burg [2] who is credited with introducing the maximum entropy ansatz in time series analysis, and to Jaynes [10] and Csiszar [3] for systematic analyses of the ansatz and its relevance in scientific modeling.

## II. Development and main results

As explained in the introduction, we consider the problem of spectral analysis based on partial autocorrelation statistics. Thus, we begin with a finite set of autocorrelation samples $R_k := \mathcal{E}\{u_\ell u^*_{\ell-k}\}$, for $k = 0, 1, 2, \ldots, n$, of a zero-mean, stationary scalar random process $\{u_\ell \;:\; \ell \in \mathbb{Z}\}$, where "*" denotes complex conjugation (together with transposition when applied to vectorial quantities). The discrete "time index" may represent a spatial coordinate when the $u_\ell$'s are readings at, say, a number of uniformly and linearly spaced sensor locations.

Without loss of generality we assume that

$$\mathbf{R}_n := \begin{bmatrix} R_0 & R_1^* & \ldots & R_n^* \\ R_1 & R_0 & \ldots & R_{n-1}^* \\ \vdots & & \ddots & \vdots \\ R_n & R_{n-1} & \ldots & R_0 \end{bmatrix} > 0,$$

i.e., that it is *positive definite*, for otherwise there is a unique power spectrum $d\mu(\theta)$ for which

$$R_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-jk\theta} d\mu(\theta), \text{ for } k = 0, 1, \ldots, n \tag{1}$$

see e.g., [13], [8]. The following theorem summarizes known facts about the maximum entropy power spectrum which is consistent with $\mathbf{R}_n$.

*Theorem 1:* Provided $\mathbf{R}_n > 0$ there exists a unique power spectrum $d\mu_{\rm ME}$ (i.e., a nonnegative measure on $[-\pi,\, \pi)$) which satisfies (1) and is a maximizer of the following convex functional

$$\mathbb{I}(d\mu/d\theta) := \frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left(\left(\frac{d\mu(\theta)}{d\theta}\right)^{-1}\right) d\theta. \tag{2}$$

Further, $d\mu_{\rm ME}$ is absolutely continuous (with respect to the Lebesgue measure) and of the form

$$d\mu_{\rm ME}(\theta) = f_{\rm ME}(\theta) d\theta$$

where the spectral density $f_{\rm ME}(\theta)$ is the inverse of a positive trigonometric polynomial of degree at most $n$, i.e.,

$$f_{\rm ME}(\theta) = \frac{k^2_{\rm ME}}{|a(e^{j\theta})|^2}$$

with $k^2_{\rm ME} > 0$, and $a(z) = 1 + a_1 z + \ldots + a_n z^n$. The polynomial $a(z)$ can be selected to have all of its roots in the complement $\mathbb{D}^c := \{z \;:\; |z| > 1\}$ of the unit disc $\mathbb{D}$ ($\mathbb{D} := \{z \;:\; |z| \leq 1\}$) of the complex plane, in which case

$$\alpha_k = \begin{cases} -a_k & \text{for } k \leq n \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

is the (unique) minimizer of the variance

$$\mathcal{E}_{d\mu_{\rm ME}}\{|u_0 - \hat{u}_{0|{\rm past}}|^2\}$$

of the (one-step-ahead) prediction error when the *predictor*

$$\hat{u}_{0|{\rm past}} := \sum_{k>0} \alpha_k u_{-k} \tag{4}$$

is sought to depend linearly on past observations. In general, the minimal variance of the prediction error depends on the choice of $d\mu$ (which is subject to (1)). This variance is maximal when $d\mu_{\rm ME}$ is selected, i.e., the maximum entropy power spectrum solves the min-max problem:

$$\max_{d\mu} \min_{\alpha_k,\, k>0} \left\{ \mathcal{E}_{d\mu}\{|u_0 - \sum_{k>0} \alpha_k u_{-k}|^2\} \;:\; \text{(1) holds} \right\}. \square$$

In the theorem and throughout, $d\mu/d\theta = f$ denotes the power spectral density function which is independent of any possible singular part of the spectral measure $d\mu$. The theorem is well known and has its roots in the classical theory of moments and the theory of orthogonal polynomials. For a proof see [7], [8], cf. [13]. More specifically, the extremal properties of $a(z)$ are established in e.g., [8, page 38], see also [7, Chapter VIII]. The fact that $f_{\rm ME}$ is consistent with the autocorrelation moment constraints inherited by $\mathbf{R}_n$ follows from [7, Equations (1.15), (1.18)]. On the other hand, the entropy functional $\mathbb{I}(\cdot)$ is clearly convex and a variational argument easily shows that the minimizer is of the form indicated. The last statement follows from the fact that (see [8, page 38, section 2.2])

$$\min_{\alpha_k,\, k>0} \mathcal{E}_{d\mu_{\rm ME}}\{|u_0 - \sum_{k>0} \alpha_k u_{-k}|^2\} = \frac{\det \mathbf{R}_n}{\det \mathbf{R}_{n-1}}$$

is achieved for the choice $\alpha_k = a_k$, while

$$\mathcal{E}_{d\mu}\{|u_0 - \sum_{k>0} \alpha_k u_{-k}|^2\} = \frac{\det \mathbf{R}_n}{\det \mathbf{R}_{n-1}}$$

is independent of $d\mu$ as long as (1) holds. An alternative derivation of all the claims in the theorem can be contructed in a way analogous to the steps used in the proof of Theorem 2 below, which we provide in Section VI.

The functional $\mathbb{I}(\cdot)$ in Theorem 1 can be interpreted to represent *entropy rate* (see [9]) and has been introduced into time-series modeling by Burg [2]. It is also interesting to note that the maximum entropy solution $d\mu_{\rm ME}$ together with $\alpha_k$'s in (3) represent a saddle point of $\mathcal{E}_{d\mu}\{|u_0 - \sum_{k>0} \alpha_k u_{-k}|^2\}$ seen as function of two variables, $d\mu$ and the infinite coefficient vector $(\alpha_1, \alpha_2, \ldots)$.

An alternative choice for a solution to (1) corresponding to the least predictable process (MR-process) from combined past and future values can be also obtained via convex optimization of a suitable functional. The following proposition presents this MR-solution and highlights its justification as the worst-case senario with regard to a corresponding smoothing problem. The development mirrors the case of the ME-solution.

*Theorem 2:* Provided $\mathbf{R}_n > 0$ there exists a unique power spectrum $d\mu_{\rm MR}$ (nonnegative measure on $[-\pi,\, \pi]$) which satisfies (1) and is a minimizer of the following concave functional

$$\mathbb{J}(d\mu/d\theta) := \frac{1}{2\pi} \int_{-\pi}^{\pi} \left( \frac{d\mu(\theta)}{d\theta} \right)^{-1} d\theta. \tag{5}$$

Further, $d\mu_{\rm MR}$ is absolutely continuous (with respect to the Lebesgue measure) and of the form

$$d\mu_{\rm MR}(\theta) = f_{\rm MR}(\theta) d\theta$$

where the spectral density $f_{\mathrm{MR}}(\theta)$ is the square root of the inverse of a positive trigonometric polynomial of degree at most $n$, i.e.,

$$f_{\mathrm{MR}}(\theta)=\frac{k_{\mathrm{MR}}^2}{\sqrt{b(e^{j\theta})}}$$

with $k_{\mathrm{MR}}^2>0$ and

$$b(e^{j\theta})=b_{-n}e^{-nj\theta}+\ldots+b_0+\ldots+b_ne^{nj\theta}>0$$

for $\theta\in[-\pi,\,\pi]$ (and $b_{-k}:=b_k^*$). The constant $k_{\mathrm{MR}}^2$ can be selected so that the trigonometric polynomial $b(e^{j\theta})$ satisfies

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\sqrt{b(e^{j\theta})}d\theta=1,$$

in which case,

$$\beta_k=\begin{cases}-\rho_k & \text{when } 1<|k|\\ 0 & 0 \text{ when } k=0\end{cases}\tag{6}$$

with $\rho_\ell$ the coefficients of the Fourier series of

$$\sqrt{b(e^{j\theta})}=\ldots+\rho_{-2}e^{-2j\theta}+\rho_{-1}e^{-j\theta}+1+\rho_1e^{j\theta}+\rho_2e^{2j\theta}+\ldots$$

is the (unique) minimizer of the variance

$$\mathcal{E}_{d\mu_{\mathrm{MR}}}\{|u_0-\hat{u}_{0|\text{past \& future}}|^2\}$$

of the smoothing error when the *smoothing filter*

$$\hat{u}_{0|\text{past \& future}}:=\sum_{k\neq 0}\beta_k u_{-k}\tag{7}$$

is sought to depend linearly on past and future observations. In general, the minimal variance of the error depends on the choice of $d\mu$ (which is subject to (1)). This variance is maximal when $d\mu_{\mathrm{MR}}$ is selected, i.e., the most random power spectrum solves the min-max problem:

$$\max_{d\mu}\min_{\beta_k,\,k\neq 0}\left\{\mathcal{E}_{d\mu}\{|u_0-\sum_{k\neq 0}\beta_k u_{-k}|^2\}\ :\ \text{(1) holds}\right\}.$$

The last statement of the theorem echoes the analogous property of the maximum entropy solution. In fact, it can be seen that in the present case $d\mu_{\mathrm{MR}}$, together with the coefficients $\beta_k$'s in (6) for the smoothing filter, represent a saddle point of $\mathcal{E}_{d\mu}\{|u_0-\sum_{k\neq 0}\beta_k u_{-k}|^2\}$.

The ME-power spectrum is rational and its coefficients can be obtained by solving a system of linear equations (the Yule-Walker-Levinson equations) which give rise to the following expression for

$$a(z)=\frac{1}{\det(\mathbf{R}_{n-1})}\det\begin{pmatrix}R_0 & R_{-1} & \ldots & R_{-n}\\ R_1 & R_0 & \ldots & R_{-n+1}\\ \vdots & & & \vdots\\ R_{n-1} & R_{n-1} & \ldots & R_{-1}\\ z^n & z^{n-1} & \ldots & 1\end{pmatrix},\tag{8}$$

while $k_{\mathrm{ME}}^2=\det(\mathbf{R}_n)/\det(\mathbf{R}_{n-1})$ e.g., see [13] and also [7, page 156]. The corresponding random process can then be simulated via a Markovian realization—in fact via an autoregressive model with transfer function $k_{\mathrm{ME}}/a(z)$ driven by a unit-variance, white-noise input, cf. [13].

The case of the MR-power spectrum differs substantially in this respect. The power spectral density function is not rational. However, its coefficients can be readily obtained from the data $\mathbf{R}_n$ using the formalism in [4], [5]. This is explained in the following statement.

*Theorem 3:* Let $\mathbf{R}_n > 0$, define the column vectors

$$\begin{aligned} \mathrm{R}_1 &:= [\ R_n^* \ \ \ldots \ \ R_1^* \ \ R_0 \ \ R_1 \ \ \ldots \ \ R_n\ ]', \text{ and} \\ G(e^{j\theta}) &:= [\ e^{jn\theta} \ \ \ldots \ \ e^{j\theta} \ \ 1 \ \ e^{-j\theta} \ \ \ldots \ \ e^{-jn\theta}\ ]', \end{aligned}$$

of size $2n+1$, where $'$ denotes transposition (without complex conjugation), and let the $\lambda(t) \in \mathbb{C}^{(2n+1)\times 1}$ represent the solution of the differential equation

$$\frac{d\lambda(t)}{dt} = M(\lambda(t))^{-1}\left(\mathrm{R}_1 - \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{G(e^{j\theta})}{\sqrt{\lambda(t)G(e^{j\theta})}}d\theta\right) \tag{9}$$

on $[0, \infty)$, where

$$M(\lambda(t)) := -\frac{1}{2\pi}\int_{-\pi}^{\pi} G(e^{j\theta})\frac{1/2}{(\lambda(t)G(e^{j\theta}))^{3/2}}G(e^{j\theta})^* d\theta \tag{10}$$

and

$$\lambda(0) = \lambda_0 := [\ \underbrace{0\ \ \ldots\ \ 0}_{n}\ 1\ \underbrace{0\ \ \ldots\ \ 0}_{n}\ ].$$

Then the following hold:

(i) $\lambda(t)$ tends to a limit $\lambda_{\text{MR}} \in \mathbb{C}^{(n+1)\times 1}$ as $t \to \infty$,

(ii) $\lambda_{\text{MR}}G(e^{j\theta}) > 0$ for all $\theta \in [-\pi,\, \pi)$ and

$$d\mu(\theta) = \frac{1}{\sqrt{\lambda_{\text{MR}}G(e^{j\theta})}}d\theta \text{ satisfies } (1), \tag{11}$$

(iii) $\lambda_{\text{MR}}$ is the unique value in $\mathbb{C}^{2n+1}$ for which (ii) holds.

## III. Notation and preliminaries

We consider the scalar zero-mean stationary random process $\{u_k,\ k \in \mathbb{Z}\}$ and, as before, we let $R_0,\, R_1,\, R_2, \ldots$ represent its sequence of autocorrelation samples and $d\mu(\theta)$ its power spectrum. We study quadratic optimization problems with respect to the usual inner product

$$\begin{aligned} \langle\sum_k a_k u_k, \sum_\ell b_\ell u_\ell\rangle_{d\mu} &:= \mathcal{E}_{d\mu}\{(\sum_k a_k u_k)(\sum_\ell b_\ell u_\ell)^*\} \\ &= \sum_{k,\ell} a_k R_{k-\ell} b_\ell^* \end{aligned} \tag{12}$$

where $R_{-m} := R_m^*$. As usual [12], the closure of $\mathrm{span}\{u_k\ :\ k \in \mathbb{Z}\}$, which we denote by $\mathcal{U}$, can be identified with the space $L_{2,d\mu}$ of functions which are square integrable with respect to $d\mu(\theta)$ on the unit circle with inner product

$$\langle a, b\rangle_{d\mu} := \frac{1}{2\pi}\int_{-\pi}^{\pi} a(\theta)(b(\theta))^* d\mu(\theta)$$

where $a(\theta) = \sum_k a_k e^{jk\theta}$ and $b(\theta) = \sum_\ell b_\ell e^{j\ell\theta}$. Then it is quite standard that the correspondence

$$\mathcal{U} \to L_{2,d\mu}\ \ :\ \ \sum_k a_k u_k \mapsto \sum_k a_k e^{jk\theta}$$

is a Hilbert space isomorphism.

Least-variance approximation problems can equivalently be expressed in $L_{2,d\mu}$. In particular, the variance $\mathcal{E}_{d\mu}\{|u_0 - \hat{u}_{0|\text{past}}|^2\}$ of the one-step-ahead prediction error

$$u_0 - \hat{u}_{0|\text{past}}$$

with $\hat{u}_{0|\mathrm{past}} = \sum_{k>0} \alpha_k u_{-k}$ as in (4), can equivalently be expressed in the form

$$\|1 - \sum_{k>0} \alpha_k e^{jk\theta}\|_{d\mu}^2, \tag{13}$$

and similarly, the variance of the smoothing error $\mathcal{E}_{d\mu}\{|u_0 - \hat{u}_{0|\mathrm{past\ \&\ future}}|^2\}$ is simply

$$\|1 - \sum_{k\neq 0} \beta_k e^{jk\theta}\|_{d\mu}^2 \tag{14}$$

in view of $\hat{u}_{0|\mathrm{past\ \&\ future}}$ given in (7).

The power spectrum $d\mu$ is a bounded nonnegative measure on $[-\pi,\,\pi)$ and admits a decomposition $d\mu = d\mu_{\mathrm{s}} + f d\theta$ with $d\mu_{\mathrm{s}}$ a singular measure and $f d\theta$ the absolutely continuous part of $d\mu$ (with respect to the Lebesgue measure). Then, the variance of the optimal one-step-ahead prediction error is given in terms of the power spectral density function $f$ by the celebrated Szegö-Kolmogorov formula given below.

*Theorem 4:* [14] With $d\mu = d\mu_{\mathrm{s}} + f d\theta$ as above

$$\inf_{\alpha} \|1 - \sum_{k>0} \alpha_k e^{jk\theta}\|_{d\mu}^2 = \exp\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi} \log f(\theta) d\theta\right\}$$

when $\log f \in L_1$, and zero otherwise.

For a proof see [8, page 183], and also [15, Chapter 6]. In the next section we derive an analogous formula for the variance of the optimal smoothing error when using both past and future values of $u_\ell$.

## IV. Least-variance smoothing

Given the power spectrum $d\mu$ of a random process we seek the optimal linear smoothing filter using both past and future observations. It turns out that the variance of the smoothing error is the harmonic mean of the spectral density of the random process, i.e., it relates to the $0$th Fourier coefficient of the inverse of the spectral density of the process. This result will be used in the next section for the purpose of identifying the MR-spectra which are consistent with a finite set of autocorrelation samples.

*Theorem 5:* Let $d\mu$ be a bounded nonnegative measure on $[-\pi,\pi)$, let $d\mu = d\mu_{\mathrm{s}} + f d\theta$ be the decomposition of $d\mu$ into its singular and absolutely continuous parts. Then, the infimum of

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} |\alpha(\theta)|^2 d\mu(\theta) \tag{15}$$

subject to the constraints

$$\alpha(\theta) \in L_1, \tag{16}$$

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \alpha(\theta) d\theta = 1 \tag{17}$$

is equal to

$$\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} f(\theta)^{-1} d\theta\right)^{-1} \tag{18}$$

when $f^{-1} \in L_1$, and zero otherwise.

An important step in the proof of the theorem is provided by the following lemma.

*Lemma 6:* [8] Let $d\mu_{\mathrm{s}}$ be a bounded singular measure on $\theta \in I := [-\pi,\pi)$ (i.e., the absolutely continuous part of $d\mu_{\mathrm{s}}$ is identically zero) and let $\epsilon_1$ be an arbitrary positive number. Then, it is always possible to decompose the interval $I$ into a finite number of intervals such that for a certain class $I_1$ of these intervals (i.e., their union) and for the complementary class $I_2 = I \backslash I_1$, the following inequalities hold:

$$\begin{aligned} \frac{1}{2\pi}\int_{I_1} d\mu_{\rm s}(\theta) &< \epsilon_1, \\ \frac{1}{2\pi}\int_{I_2} d\theta &< \epsilon_1. \end{aligned}$$

For a proof of Lemma 6 see [8, page 7]. We now proceed with the proof of Theorem 5.

*Proof: of Theorem 5:* Assume first that $d\mu$ is absolutely continuous with no singular part. Given any positive number $\epsilon$ define

$$\alpha_\epsilon(\theta) := \frac{(f(\theta)+\epsilon)^{-1}}{\frac{1}{2\pi}\int_{-\pi}^{\pi}(f(\theta)+\epsilon)^{-1}d\theta}.$$

We note that $\alpha_\epsilon \in L_1$ (also in $L_2$ and in fact, it is even bounded and positive),

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\alpha_\epsilon(\theta)d\theta = 1,$$

and we observe that

$$\begin{aligned} &\frac{1}{2\pi}\int_{-\pi}^{\pi}|\alpha_\epsilon(\theta)|^2 f(\theta)d\theta \\ =\ & \frac{\frac{1}{2\pi}\int_{-\pi}^{\pi}(f(\theta)+\epsilon)^{-2}f(\theta)d\theta}{\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}(f(\theta)+\epsilon)^{-1}d\theta\right)^2} \\ <\ & \left(\frac{1}{2\pi}\int_{-\pi}^{\pi}(f(\theta)+\epsilon)^{-1}d\theta\right)^{-1} \end{aligned}$$

because $f(\theta)/(f(\theta)+\epsilon) < 1$. If $f^{-1} \notin L_1$ then

$$\lim_{\epsilon\to 0}\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}(f(\theta)+\epsilon)^{-1}d\theta\right)^{-1} = 0,$$

whereas if $f^{-1} \in L_1$ the limit equals the expression given in (18). To prove our claims for the case where $d\mu$ is absolutely continuous, it remains to show that when $f^{-1} \in L_1$ the infimal value for (15) is never strictly less than (18).

Continuing on, we assume that $f^{-1} \in L_1$. We normalize $f^{-1}$ to have the identity as its $0$th Fourier coefficient

$$\alpha_0 = \frac{f^{-1}}{\frac{1}{2\pi}\int_{-\pi}^{\pi}f^{-1}(\theta)d\theta}$$

and consider the perturbation

$$\alpha = \alpha_0 + \delta$$

for an arbitrary $\delta \in L_1$ with vanishing $0$th Fourier coefficient (i.e., a $\delta \in L_1$ satisfying $\int_{-\pi}^{\pi}\delta(\theta)d\theta = 0$ so that $\alpha$ satisfies (17)). It readily follows that

$$\begin{aligned} &\frac{1}{2\pi}\int_{-\pi}^{\pi}|\alpha(\theta)|^2 d\mu(\theta) \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(|\alpha_0(\theta)|^2 + 2\delta(\theta)\alpha_0(\theta) + |\delta(\theta)|^2\right)f(\theta)d\theta \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi}|\alpha_0(\theta)|^2 f(\theta)d\theta + \frac{1}{2\pi}\int_{-\pi}^{\pi}|\delta(\theta)|^2 f(\theta)d\theta, \end{aligned} \tag{19}$$

where for the last step we note that

$$\begin{aligned}\int_{-\pi}^{\pi}\delta(\theta)\alpha_0(\theta)f(\theta)d\theta &= \int_{-\pi}^{\pi}\delta(\theta)\frac{f^{-1}(\theta)}{\frac{1}{2\pi}\int_{-\pi}^{\pi}f^{-1}(\theta)d\theta}f(\theta)d\theta\\ &= \frac{1}{\frac{1}{2\pi}\int_{-\pi}^{\pi}f^{-1}(\theta)d\theta}\int_{-\pi}^{\pi}\delta(\theta)d\theta\\ &= 0.\end{aligned}$$

The first term in (19) is precisely the claimed infimal value in (18) and the second term is clearly nonnegative. This proves our claim in the case where $d\mu$ is absolutely continuous.

We now consider the case where

$$d\mu(\theta) = d\mu_{\rm s}(\theta) + f(\theta)d\theta$$

with $d\mu_{\rm s}$ a singular measure (always with respect to the Lebesgue measure). For an arbitrary $\epsilon > 0$ we consider a decomposition of

$$[-\pi,\,\pi) = I_1 \cup I_2$$

where

$$\int_{I_1} d\mu_{\rm s}(\theta) \;<\; \epsilon^3, \tag{20}$$

$$\int_{I_2} d\theta \;<\; \epsilon^3. \tag{21}$$

That such a decomposition exists follows from Lemma 6 taking $\epsilon_1 = \epsilon^3$ in the statement of the lemma. Now let $\chi_{I_1}$ denote the characteristic function of $I_1$ which takes the value $1$ when $\theta \in I_1$ and zero otherwise, and set

$$\alpha_\epsilon = \frac{(f+\epsilon)^{-1}\chi_{I_1}}{\frac{1}{2\pi}\int_{-\pi}^{\pi}(f(\theta)+\epsilon)^{-1}\chi_{I_1}(\theta)d\theta} \tag{22}$$

which is in $L_1$ and has the identity as its $0$th Fourier coefficient. Then

$$\begin{aligned}\frac{1}{2\pi}\int_{-\pi}^{\pi}|\alpha_\epsilon(\theta)|^2 d\mu(\theta) &= \frac{1}{2\pi}\int_{-\pi}^{\pi}|\alpha_\epsilon(\theta)|^2 d\mu_{\rm s}(\theta)\\ &+\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{(f+\epsilon)^{-2}\chi_{I_1}}{\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}(f(\theta)+\epsilon)^{-1}\chi_{I_1}(\theta)d\theta\right)^2}f(\theta)d\theta.\end{aligned}$$

The first term on the right hand side is bounded above by

$$\frac{1/\epsilon^2}{\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}(f(\theta)+\epsilon)^{-1}\chi_{I_1}(\theta)d\theta\right)^2}\epsilon^3$$

which decays to $0$ with $\epsilon$, whereas the second term is bounded above by

$$\frac{1}{\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}(f(\theta)+\epsilon)^{-1}\chi_{I_1}(\theta)d\theta\right)}$$

which in the limit recovers the claimed bound (18). The earlier argument for the case of absolutely continuous $d\mu$ applies and shows that this bound is in fact the correct value for the infimum and that no lower value is possible. ■

*Remark 7:* It is clear from the proof that if $f^{-1} \in L_1$ and $d\mu = f d\theta$ is absolutely continuous, then

$$\alpha_0 = \frac{f^{-1}}{\frac{1}{2\pi}\int_{-\pi}^{\pi}f^{-1}(\theta)d\theta}$$

is the unique optimal solution which achieves the minimal value

$$\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} f(\theta)^{-1}d\theta\right)^{-1}$$

for

$$\|\alpha\|_{d\mu}^2 := \frac{1}{2\pi}\int_{-\pi}^{\pi} |\alpha(\theta)|^2 d\mu(\theta)$$

subject to $\alpha \in L_1$ with $0$th Fourier coefficient the identity. Thus, if

$$\alpha_0(\theta) \sim \ldots + \rho_{-1}e^{-j\theta} + 1 + \rho_1 e^{j\theta} + \ldots$$

and $\rho_\ell$, $\ell = \pm 1, \pm 2, \ldots$ the corresponding Fourier coefficients, then

$$\hat{u}_0 = -\sum_{\ell \neq 0} \rho_\ell u_{-\ell}$$

is the optimal in the least variance sense estimate for $u_0$, and $u_\ell$ is a random process with $d\mu$ as its power spectrum. In this case the infimum is achieved, and hence it represents the minimum variance of the error. When the power spectrum has either singular part or $f^{-1} \notin L_1$, then $\alpha_\epsilon$ as in (22) provides suboptimal solutions. This is completely analogous to the Szegö-Kolmogorov setting where optimal one-step ahead predictors (which use only past observations) exist when $\log f \in L_1$ otherwise the least variance is not attained but can be gotten arbitrarily closely [7, Chapter II].

*Remark 8:* It is interesting to observe that while the minimal variance of a smoothing error for a random process having $f$ as spectral density is the *harmonic mean*

$$m_{-1,f} := \left(\frac{1}{2\pi}\int_{-\pi}^{\pi} f(\theta)^{-1}d\theta\right)^{-1}$$

of the values of $f$ on the $[-\pi, \pi)$, the minimal variance of the optimal one-step-ahead predictor using only past observations is the *geometric mean* (see [8, page 183], [15, Chapter 6])

$$m_{0,f} := \exp\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left(f(\theta)\right)d\theta\right).$$

The former is the inverse of $\mathbb{J}(d\mu/d\theta)$ whereas the latter is exponential of $-\mathbb{I}(d\mu/d\theta)$. Naturally, $m_{-1,f} \leq m_{0,f}$ (see also [1, page 23]). This ordering is clear from the interpretation of the two quantities as variances of best predictors which use "past+future" and "only past" observations, respectively.

## V. On deterministic processes: an example

It may be rather surprising, at first glance, that the value of a random process with power spectral density

$$f_o(\theta) = |1 - e^{j\theta}|^2 = 2 - 2\cos(\theta) \tag{23}$$

can be predicted at any given point with arbitrarily small variance, when both past and future observations are available. Yet this is the case, and this is due to the fact that $f_o^{-1} \notin L_1$ (equivalently $m_{-1,f_o} = 0$). This example highlights the difference between "deterministic processes" in the sense of $m_{-1,f} = 0$ and those in the sense of Szegö-Kolmogorov which are characterized by $m_{0,f} = 0$ or, equivalently, by $\log f \notin L_1$ instead.

For our particular example, the fact that $f_o^{-1} \notin L_1$ follows from the divergence of

$$\int_{\epsilon}^{\pi} \frac{1}{1-\cos(\theta)}d\theta > \int_{\epsilon}^{\pi} \frac{1}{\theta^2}d\theta$$

as $\epsilon \to 0$. On the other hand, the fact that $\log f_o \in L_1$ can be seen as follows. Since $g(z) := 1 - z$ is analytic and does not vanish in $\mathbb{D}$, $\log |g|$ is harmonic and

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \log(|g(re^{j\theta})|)d\theta = \log(|g(0)|) = 0$$

for any value of $r \in [0, 1)$. Therefore, the integral of the logarithm of $\lim_{r\to 1} |g(re^{j\theta})|$ also vanishes, and the same applies to $f_o(\theta) = \lim_{r\to 1} |g(re^{j\theta})|^2$.

In the rest of this section we explain how a random process corresponding to $f_o$ can be predicted with vanishingly small variance from the combined past and future record. We do so, for didactic purposes, by sketching a specialized and more direct construction than that of Section IV.

Consider a realization $\{u_k,\, k \in \mathbb{Z}\}$ of a random process corresponding to $f_o$ as follows:

$$u_k = w_k - w_{k-1}$$

where $\{w_k,\, k \in \mathbb{Z}\}$ is a sequence of independent, identically distributed, random variables with zero mean and unit variance (i.e., a white-noise process). We assume that "past" ($\{u_k,\, k < 0\}$) as well as "future" ($\{u_k,\, k > 0\}$) observations are available, and that we wish to estimate the "present" $u_0 = w_0 + w_{-1}$ based on this two-sided observation record. Then,

$$\mathbf{u}_{<0} := \begin{bmatrix} u_{-1} \\ u_{-2} \\ u_{-3} \\ \vdots \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 & \dots \\ 0 & 1 & -1 & \ddots \\ 0 & 0 & 1 & \ddots \\ \vdots & \vdots & \ddots & \ddots \end{bmatrix} \begin{bmatrix} w_{-1} \\ w_{-2} \\ w_{-3} \\ \vdots \end{bmatrix}$$

and

$$\mathbf{u}_{>0} := \begin{bmatrix} u_{1} \\ u_{2} \\ u_{3} \\ \vdots \end{bmatrix} = \begin{bmatrix} -1 & 1 & 0 & \dots \\ 0 & -1 & 1 & \ddots \\ 0 & 0 & -1 & \ddots \\ \vdots & \vdots & \ddots & \ddots \end{bmatrix} \begin{bmatrix} w_{0} \\ w_{1} \\ w_{2} \\ \vdots \end{bmatrix}$$

In both cases the mapping is Toeplitz, and identical except for a sign change. Let now

$$v := \begin{bmatrix} 1 & (1-\epsilon) & (1-\epsilon^2) & \dots \end{bmatrix},$$

and for $1 > \epsilon > 0$ and define

$$\begin{aligned} \hat{w}_{-1} &:= v\mathbf{u}_{<0} = w_{-1} + \epsilon(1-\epsilon)\sum_{k=-2}^{-\infty} \epsilon^{-k+2} w_k \\ \hat{w}_0 &:= -v\mathbf{u}_{>0} = w_0 + \epsilon(1-\epsilon)\sum_{k=1}^{\infty} \epsilon^{k-1} w_k \\ \hat{u}_0 &:= \hat{w}_0 - \hat{w}_{-1}. \end{aligned}$$

Each of the above can be taken as an estimator for the corresponding un-hatted variable. The variance of estimation in all cases can be made arbitrarily small with appropriately small choice for $\epsilon$. This justifies our claim.

## VI. Proofs of Theorems 2 and 3

Due to the strict concavity of the inversion map $x \mapsto 1/x$ on $\mathbb{R}_+$, $\mathbb{J}(\cdot)$ is also a strictly concave functional on (non-negative) density functions. We first show that a spectral density $f_{\rm ME}$ of the form claimed in Theorem 2 is indeed a minimizer of $\mathbb{J}(\cdot)$ subject to the moment constraints

$$R_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-jk\theta} f(\theta)d\theta, \text{ for } k = 0, 1, \ldots, n. \tag{24}$$

Existence of suitable values for the corresponding parameters requires proving Theorem 3 next, which claims that these values correspond to an attractive equilibrium of a certain differential equation. The form of $f_{\rm ME}$ ensures stationarity and hence, due to the strict concavity of $\mathbb{J}(\cdot)$, it ensures that this is indeed the unique extremal point. Finally, we revisit the optimizaton problem and consider measures with possible singular part. The singular part does not affect the value of $\mathbb{J}(\cdot)$, but the fact that a singular part is allowed, relaxes the constraint (24) to (1). Yet, as we will see, $f_{\rm ME}$ is still the minimizer and, hence, the extremal spectral measure $d\mu$ cannot have a singular part. In the end, we return to the remaining claims in Theorem 2 regarding properties of the minimizer.

### A. Functional form of minimizer

Consider first the problem of minimizing $\mathbb{J}(f)$ with $f$ constrained to satisfy (24). If

$$\lambda := \left[\begin{array}{ccccc} \lambda_{-n} & \ldots & \lambda_0 & \ldots & \lambda_n \end{array}\right]$$

denotes a vector of Lagrange multipliers, the corresponding Lagrangian is

$$\mathcal{L}(f,\lambda) \;\; := \;\; \mathbb{J}(f) - \lambda(\mathrm{R}_1 - \frac{1}{2\pi}\int_{-\pi}^{\pi} G(e^{j\theta})f(\theta)d\theta)$$

where $\mathrm{R}_1, G$ are defined in the statement of Theorem 3. If we set the variation

$$\delta\mathcal{L}(f,\lambda;\delta f) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\frac{-1}{f(\theta)^2} + \lambda G(e^{j\theta})\right)\delta f(\theta)d\theta$$

identically equal to zero for all perturbations $\delta f$ (assuming that $f > 0$ and hence $\delta f$ unconstrained), then we conclude that

$$f(\theta) = \frac{1}{\sqrt{\lambda G(e^{j\theta})}}, \tag{25}$$

which is the form claimed in Theorem 2 for $f_{\rm ME}$. Our next step is to prove that, provided $\mathbf{R}_n > 0$, there always exists such a density function which satisfies (24) and that the trigonometric polynomial $\lambda G(e^{j\theta})$ is in fact strictly positive.

### B. Proof of Theorem 3

We follow the formalism in [5] for solving moment problems. We denote by $\mathfrak{R}$ the positive cone

$$\mathfrak{R} := \{\mathrm{R} \;:\; \mathrm{R} = \frac{1}{2\pi}\int_{-\pi}^{\pi} G(e^{j\theta})d\mu(\theta), \text{ where } d\mu \geq 0\}$$

and by $\mathfrak{K}$ the dual cone

$$\mathfrak{K} := \{\lambda \;:\; \lambda G(e^{j\theta}) \geq 0 \text{ for } \theta \in [-\pi,\pi]\}.$$

Both are subsets of $\mathbb{R}\times\mathbb{C}^{2n}$ since their "0th" entries $R_0, \lambda_0 \in \mathbb{R}_+$ while the remaining entries $R_\ell, \lambda_\ell \in \mathbb{C}$ ($\ell = \pm 1, \pm 2, \ldots, \pm n$). Also, both are convex. The interior of $\mathfrak{R}$ is denoted by $\mathrm{int}(\mathfrak{R})$ and the interior of the dual cone, which consists of all vectors $\lambda$ such that the trigonometric polynomial $\lambda G$ is strictly positive on the unit circle, is denoted by $\mathfrak{K}_+$. The Jacobian $\frac{\partial H}{\partial \lambda}$ of the mapping

$$H\ :\ \mathfrak{K}_+ \to \mathrm{int}(\mathfrak{R})\ :\ \lambda \mapsto \frac{1}{2\pi}\int_{-\pi}^{\pi} G(e^{j\theta})\frac{1}{\sqrt{\lambda G(e^{j\theta})}}\,d\theta$$

between Lagrange vectors and moments is given in (10) and is denoted by $M(\lambda)$. As long as $\lambda \in \mathfrak{K}_+$ the Jacobian is an invertible matrix. Our goal is to find a value for $\lambda$ so that condition (ii) of Theorem 3 holds. We do this as follows.

We begin with $\lambda_0$ as in Theorem 3 for which we readily observe that $\lambda_0 G \equiv 1 > 0$. It follows that

$$\mathrm{R}_0 := \frac{1}{2\pi}\int_{-\pi}^{\pi} G(e^{j\theta})\frac{1}{\sqrt{\lambda_0 G(e^{j\theta})}}\,d\theta \in \mathrm{int}(\mathfrak{R}).$$

Since $\mathbf{R}_n > 0$, we also know that $\mathrm{R}_1 \in \mathrm{int}(\mathfrak{R})$. Since $\mathrm{int}(\mathfrak{R})$ is convex and $\mathrm{R}_1, \mathrm{R}_0 \in \mathrm{int}(\mathfrak{R})$, the interval $[\mathrm{R}_0,\, \mathrm{R}_1] \subset \mathrm{int}(\mathfrak{R})$, i.e.,

$$\mathrm{R}_\tau := \tau \mathrm{R}_1 + (1-\tau)\mathrm{R}_0 \tag{26}$$

belongs to $\mathrm{int}(\mathfrak{R})$ for all $\tau \in [0,1]$. The key idea is now to trace $\mathrm{R}_\tau$ by following corresponding values for $\lambda_\tau$ in the dual cone. This is not always possible. It depends on the functional form for the sought spectral density function $f$. The critical issue that may prevent such path-following in the dual space is whether any $\lambda$ in the boundary of $\mathfrak{K}_+$ maps onto a point in the interior of $\mathfrak{R}$. When this happens, there are interior points in $\mathfrak{R}$ which do not admit the assumed representation. We will see below that this does not happen for the functional form $1/\sqrt{\lambda G}$ and hence, that the plan we have outlined applies. We discuss these key steps/facts next.

The moments $\mathrm{R}_\tau$, $\tau \in [0,1]$, satisfy the differential equation

$$\frac{d\mathrm{R}_\tau}{d\tau} = \mathrm{R}_1 - \mathrm{R}_0 \tag{27}$$

as follows readily from (26). Then the dual parameters $\lambda(\tau)$ satisfy

$$\frac{d\lambda(\tau)}{d\tau} = M(\lambda)^{-1}(\mathrm{R}_1 - \mathrm{R}_0), \tag{28}$$

as long as $\lambda(\tau)$ remains in the interior of $\mathfrak{K}_+$ —in which case $M(\lambda)$ is invertible being the (inverse of the) autocorrelation matrix of a positive spectral density function. We claim that this is always the case. To prove it, assume that the contrary is true and that $[0,\tau_0)$ is a maximal subinterval of $[0,1]$ for which $\lambda(\tau) \in \mathfrak{K}_+$ for $0 \le \tau < \tau_0$. Thus, the family of positive trigonometric polynomials

$$\{\lambda(\tau)G(e^{j\theta})\ :\ \tau \in [0,\tau_0)\}$$

has either a limit point on the boundary of $\mathfrak{K}_+$ or it grows unbounded. In either case we will draw a contradiction.

In the first case, there must exist an accumulation point $\hat{\lambda}$ for which $\hat{\lambda}G(e^{j\theta})$ vanishes on the unit circle. But then $\hat{\lambda}G(e^{j\theta})$, which is a nonnegative trigonometric polynomial, must have a double root at some point $e^{j\theta_0}$. Therefore

$$\frac{1}{\sqrt{\hat{\lambda}G(e^{j\theta})}}, \tag{29}$$

which has at least a single pole at $e^{j\theta_0}$, is not integrable. The assertion that the inverse of the square root of a nonnegative trigonometric polynomial which vanishes on the circle is not integrable is elementary. It suffices to consider a typical case, such as $1-\cos(\theta)$, where $\frac{1}{\sqrt{1-\cos(\theta)}} = \frac{1}{\sqrt{2}|\sin(\theta/2)|} > \frac{\sqrt{2}}{|\theta|}$ is clearly not integrable—the general case is similar. The nonintegrability of (29) implies that the family of vectors

$$\{\frac{1}{2\pi}\int_{-\pi}^{\pi} G(e^{j\theta})\frac{1}{\sqrt{\lambda(\tau)G(e^{j\theta})}}\,d\theta \;:\; 0\leq\tau<\tau_0\}$$

is unbounded, in contradiction to the assumption that the image of $\{\lambda(\tau)\;:\;0\leq\tau<\tau_0\}$ under $H$ is the subset

$$\{\mathrm{R}_\tau \;:\; 0\leq\tau<\tau_0\}$$

of the bounded interval $[\mathrm{R}_0,\mathrm{R}_1]$.

We now draw a contradiction for the second case. We assume that $\lambda(\tau)$ grows unbounded as $\tau\to\tau_0$. It follows that there is sequence $\tau_i\in[0,\tau_0)$, $i=1,2,\ldots$ such that $\tau_i\to\tau_0$ and $\|\lambda(\tau_i)\|\to\infty$ while the unit-length vectors

$$\hat{\lambda}_i := \frac{\lambda(\tau_i)}{\|\lambda(\tau_i)\|}\to\hat{\lambda}\in\mathfrak{K}$$

converge as $i\to\infty$, with $\|\cdot\|$ being the Euclidean norm. At the same time, the sequence $\mathrm{R}_{\tau_i}=H(\lambda(\tau_i))$, $i=1,2,\ldots$, converges to $R_{\tau_0}\in\mathrm{int}(\mathfrak{R})$. But any interior point $\mathrm{R}\in\mathfrak{R}$ is characterized by the property that the functional

$$\mathfrak{C}_{\mathrm{R}} \;:\; \mathfrak{K}\to\mathbb{R}_+ \;:\; \lambda\mapsto\lambda\mathrm{R}$$

is strictly positive (e.g., see [5, Proposition 3]). (This is due to the fact any such $\mathrm{R}$ assumes a representation $\frac{1}{2\pi}\int_{-\pi}^{\pi}G(e^{j\theta})f(\theta)d\theta$ for some strictly positive density function $f(\theta)$.) On the other hand, returning to the sequence $R_{\tau_i}$ $i=1,2,\ldots$, we observe that

$$\begin{aligned}\mathfrak{C}_{\mathrm{R}_{\tau_i}} \;:\; \hat{\lambda}_i\mapsto\hat{\lambda}_iR_{\tau_i} &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\hat{\lambda}_iG(e^{j\theta})}{\sqrt{\lambda(\tau_i)G(e^{j\theta})}}\,d\theta\\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\sqrt{\frac{\hat{\lambda}_iG(e^{j\theta})}{\|\lambda(\tau_i)\|}}\,d\theta\end{aligned}$$

tends to $0$ as $\|\lambda(\tau_i)\|$ grows unbounded. Therefore, the functionals $\mathfrak{C}_{\mathrm{R}_{\tau_i}}$, $i\to\infty$, are not uniformly bounded away from zero. Yet, their limit $\mathfrak{C}_{\mathrm{R}_{\tau_0}}$ is, due to the fact that $\mathrm{R}_{\tau_0}\in\mathrm{int}(\mathfrak{R})$. This is a contradiction. Therefore (28) can be integrated over the complete interval $[0,1]$ and $\lambda(\tau)$ remains bounded and in the interior of the dual cone (i.e., the trajectory lies in $\mathfrak{K}_+$). We identify $\lambda(1)=\lambda_{\mathrm{MR}}$.

We now re-scale the independent variable in (27-28) by replacing $\tau$ with $t=-\log(1-\tau)$. We simplify notation and denote $\mathrm{R}_{\tau(t)}$ by $\mathrm{R}_t$ and $\lambda(t(\tau))$ by $\lambda(t)$. Using $\frac{\partial\tau}{\partial t}=1-\tau$ and $\mathrm{R}_1-\mathrm{R}_0=\frac{1}{1-\tau}(\mathrm{R}_1-\mathrm{R}_\tau)$, we rewrite (27) as

$$\frac{d\mathrm{R}_t}{dt}=\mathrm{R}_1-\mathrm{R}_t,\;\text{ for } t\in[0,\infty),$$

and (28) as

$$\frac{d\lambda(t)}{dt}=M(\lambda(t))^{-1}(\mathrm{R}_1-\mathrm{R}_t), \tag{30}$$

where, as usual, $\mathrm{R}_t=\frac{1}{2\pi}\int_{-\pi}^{\pi}G\frac{1}{\lambda(t)G}\,d\theta$. We have now established claims (i) and (ii) of Theorem 3. I.e., we have shown that as $t\to\infty$ in (30) the trajectory $\lambda(t)$ converges in $\mathfrak{K}_+$, and that the limit point $\lambda_{\mathrm{MR}}$ is such that (1) holds. Claim (iii) of the theorem follows from the concavity of $\mathbb{J}(\cdot)$. More specifically, the functional form of $f_{\mathrm{MR}}$ guarantees that it is a minimizer of $\mathbb{J}(\cdot)$. There can only be one such minimizer since $\mathbb{J}(\cdot)$ is strictly concave.

### C. Proof of Theorem 2

Define first the column vector

$$g(e^{j\theta}) := \left[\begin{array}{cccc} 1 & e^{-j\theta} & \dots & e^{-jn\theta} \end{array}\right]'.$$

Assuming that $d\mu = d\mu_{\rm s} + fd\theta$ with $d\mu_s$ a singular measure and $fd\theta$ the absolutely continuous part of $d\mu$, the minimization of $\mathbb{J}(f)$ subject to (1) is equivalent to minimization of $\mathbb{J}(f)$ subject to

$$\mathbf{R}_n \geq \frac{1}{2\pi}\int_{-\pi}^{\pi} g(e^{j\theta})f(\theta)g(e^{j\theta})^* d\theta. \tag{31}$$

The corresponding Lagrangian is now

$$\begin{aligned} \mathcal{L}_o(f,\Lambda) &:= \mathbb{J}(f(\theta)) + \qquad (32)\\ &+\mathrm{trace}\left(\Lambda\left(\mathbf{R}_n - \frac{1}{2\pi}\int_{-\pi}^{\pi} g(e^{j\theta})f(\theta)g(e^{j\theta})^* d\theta\right)\right)\\ &= \mathbb{J}(f(\theta)) + \mathrm{trace}(\Lambda\mathbf{R}_n)\\ &-\frac{1}{2\pi}\int_{-\pi}^{\pi}\left(g(e^{j\theta})^*\Lambda g(e^{j\theta})\right) f(\theta)d\theta \qquad (33)\end{aligned}$$

The Lagrange multiplier $\Lambda$ is a matrix which has a Toeplitz structure. (To see this note that any possible component of $\Lambda$ which is orthogonal to the subspace of Toeplitz matrices has no effect since it vanishes when taking the inner product $\mathrm{trace}(\Lambda T)$ for any Toeplitz matrix $T$ as done in (32).) The minimizer $f$ would correspond to a measure $d\mu$ with a nontrivial singular part only if the equality constraint in (31) is not active. For this to be the case, the multiplier

$$g(e^{j\theta})^*\Lambda g(e^{j\theta})$$

of $f(\theta)$ in (33) must vanish at least for some values of $\theta$. However, the correspondence

$$\Lambda = \left[\begin{array}{cccc} \frac{1}{n+1}\lambda_0 & \frac{1}{n}\lambda_1 & \dots & \frac{1}{1}\lambda_n\\ \frac{1}{n}\lambda_1 & \frac{1}{n+1}\lambda_0 & \dots & \frac{1}{2}\lambda_{n-1}\\ \vdots & \vdots & \ddots & \vdots\\ \frac{1}{1}\lambda_{-n} & \frac{1}{2}\lambda_{-n+1} & \dots & \frac{1}{n+1}\lambda_0 \end{array}\right]$$

shows that in fact $\mathcal{L}_o(f,\Lambda) = \mathcal{L}(f,\lambda)$, i.e., it is the same Lagrangian as in Section VI-A. The value for the Lagrange multipliers in the latter, as identified in Section VI-B, are such that $\lambda_{\rm MR}G(e^{j\theta})$ is a positive trigonometric polynomial. This polynomial is precisely the multiplier of $f(\theta)$ in (33) and is strictly positive for all $\theta \in [-\pi,\pi]$. Hence, the equality constraint in (31) is active for the extremal $f$ of the relaxed problem corresponding to (32). Then, the analysis in Section VI-A applies. Therefore, the minimizer corresponds to an absolutely continuous power spectral distribution $d\mu_{\rm MR} = f_{\rm MR}(\theta)d\theta$ which is of the form claimed in the theorem.

We now address the remaining claims in the theorem regarding the variance of the smoothing error for the corresponding random process. Given the expression for $f_{\rm MR}$ which is the square root of the inverse of a positive trigonometric polynomial, the form of the optimal smoothing filter for the corresponding random process is provided by Theorem 5. It is a consequence of the same theorem that the variance of the optimal smoothing error $\mathcal{E}_{d\mu_{\rm MR}}\{|u_0 - \hat{u}_{0|\text{past \& future}}|^2\}$ is precisely the inverse of the $\mathbb{J}$-functional evaluated at $f_{\rm MR}$, i.e.,

$$\left(\mathbb{J}(f_{\rm MR})\right)^{-1}.$$

The last part of the theorem is also immediate since

$$\min_{\beta_k,\,k\neq 0}\left\{\mathcal{E}_{d\mu}\{|u_0 - \sum_{k\neq 0}\beta_k u_{-k}|^2\} \;:\; \text{(1) holds}\right\}$$

is $(\mathbb{J}(d\mu/d\theta))^{-1}$ for any spectral measure consistent with (1). But $d\mu_{\rm MR}$ is the unique maximizer of this inverse.

## VII. On spectral analysis: an example

For illustration purposes, we compare the power spectra $f_{\rm ME}$ and $f_{\rm MR}$ given in Theorems 1 and 2 for a basic example. We begin by evaluating the first $4$ autocorrelation moments for the following spectral density:

$$f_{\rm true}(\theta) = 1 + \frac{2}{5}\cos(\theta) + \delta(\theta - \frac{1}{2}) + \delta(\theta + \frac{1}{2}).$$

Here, for convenience, we depart slightly from our earlier notation and incorporate the singular part of the power spectrum into the "spectral density" as a sum of two Dirac functions —the distributions $\delta(\theta-\theta_0)$ for $\theta_0 = \pm\frac{1}{2}$. Thus, the absolutely continuous part of the power spectrum is made up of only the continuous portion $\left(1+\frac{2}{5}\cos(\theta)\right)d\theta$ of $f_{\rm true}(\theta)d\theta$. The corresponding random process consists of a random moving average component generated by

$$u_k^{MA} = w_k + \frac{1}{2}w_{k-1}$$

with $w_k$ a white-noise process with variance $1/(1+1/4)$ (normalized so that $\mathcal{E}\{|u_k^{MA}|^2\} = 1$), and a deterministic sinusoidal component at frequency $\theta_o = 1/2$ [rad/unit of time]. The first $4$ samples of the autocorrelation function of

$$u_k = u_k^{MA} + 2\sin(\frac{k}{2}+\phi)$$

(with $\phi$, say, uniformly distributed on $[-\pi,\,\pi]$) can be readily computed and are as follows:

$$\begin{aligned} &\left[\begin{array}{cccc} R_0 & R_{\pm1} & R_{\pm2} & R_{\pm3}\end{array}\right] \\ &= \left[\begin{array}{cccc} 3.0000 & 2.1552 & 1.0806 & 0.1415\end{array}\right].\end{aligned}$$

The corresponding Toeplitz matrix $\mathbf{R}_3$ is positive definite, and as a result, there is a nontrivial family of power spectra which are consistent with the autocorrelation data –$d\mu(\theta) = f_{\rm true}(\theta)d\theta$ is only of them.

Figure 1 shows the three particular power spectra that concern us here. First, the "moving-average + sinusoids" power spectrum described above is shown with a dashed line ($- - -$). Then, a ME-power spectrum which is consistent with $\mathbf{R}_3$ and obtained following the maximum entropy ansatz is shown with a dash-dotted line ($- \cdot -$). Finally, the MR-power spectrum corresponding to the least smooth process is shown with a continuous line (——-).

*All three power spectra shown are consistent with the covariance data*. Hence, there is no suggestion that one should be preferable. They all describe the same data. A selection can only be based on either prior information or a prejudice —this is where an "ansatz" becomes relevant. Had we known that the "true" spectrum originates from a moving average component plus a minimal number of sinusoids, we could have recovered the exact power spectrum from the covariance data following e.g., [6]. Of course, such knowledge is rarely available and one is called to use other insights. Thence, if the power spectrum and a model for the process is to be used for prediction purposes, the maximum entropy option is quite natural since it represents the relevant "worst-case senario." However, if the model is to be used for filling in gaps in records, then the MR-option is the appropriate "worst-case senario." Then, if our goal is to simply identify features in the power spectrum, either may be appropriate.

Using (8) we determine that

$$f_{\rm ME}(\theta) \;=\; \frac{k_{\rm ME}^2}{|1+a_1e^{j\theta}+a_2e^{2j\theta}+a_3e^{3j\theta}|^2} \tag{34}$$

with $k_{\rm ME} = 1.2732$, and

$$\begin{aligned} &\left[\begin{array}{ccc} a_1 & a_2 & a_3\end{array}\right] \\ &= \left[\begin{array}{ccc} -0.9026 & 0.1829 & 0.1465\end{array}\right].\end{aligned}$$

On the other hand, following Theorem 3 we compute

$$\begin{aligned} f_{\mathrm{MR}}(\theta) &= \frac{1}{\sqrt{\sum_{\ell=-3}^{3} \lambda_\ell e^{\ell j\theta}}} \qquad (35) \\ &= \frac{\kappa^2}{\sqrt{|1+\hat{a}_1 e^{j\theta}+\hat{a}_2 e^{2j\theta}+\hat{a}_3 e^{3j\theta}|^2}} \\ &= \frac{\kappa^2}{|1+\hat{a}_1 e^{j\theta}+\hat{a}_2 e^{2j\theta}+\hat{a}_3 e^{3j\theta}|} \qquad (36) \end{aligned}$$

with

$$\begin{aligned} &\left[\begin{array}{cccc} \lambda_0 & \lambda_{\pm 1} & \lambda_{\pm 2} & \lambda_{\pm 3} \end{array}\right] \\ &= \left[\begin{array}{cccc} 3.4942 & -2.5690 & 0.9598 & -0.1231 \end{array}\right], \end{aligned}$$

or, equivalently, $\kappa = 1.2732$ and

$$\begin{aligned} &\left[\begin{array}{ccc} \hat{a}_1 & \hat{a}_2 & \hat{a}_3 \end{array}\right] \\ &= \left[\begin{array}{ccc} -1.7673 & 1.1795 & -0.1956 \end{array}\right]. \end{aligned}$$

Here, again, we depart slightly from our earlier notation so as to compare the coefficients more directly to the ME-spectral density. The parameters $b_\ell$ and $\rho_\ell$ as in Theorem 2 for $f_{\mathrm{MR}}$ and smoothing filter, respectively, can be readily determined from the above.

Figure 2 marks the zero of the moving average component of $f_{\mathrm{true}}$ (inside $\mathbb{D}$) along with the location of the two spectral lines (on the unit circle) with "o". The poles of the ME-spectrum are marked with a "$\diamond$" and the fractional poles of the MR-spectrum with a "$\Box$".

Figure 3 presents realizations of time-series corresponding to $f_{\mathrm{true}}$, $f_{\mathrm{ME}}$, and $f_{\mathrm{MR}}$. The one corresponding to $f_{\mathrm{true}}$ is generated by a Markovian moving-average model plus a sinusoidal component with a random phase. The time-series corresponding to $f_{\mathrm{ME}}$ is generated by a Markovian autoregressive model as usual. Finally, the time-series corresponding to $f_{\mathrm{MR}}$ is generated by a suitable discretization of the standard spectral representation (stochastic integral)

$$u_\ell = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{j\ell\theta} dv(\theta)$$

where $dv(\theta)$ is a zero-mean white noise process for $\theta \in [-\pi, \pi)$ such that $\frac{d}{d\theta}\mathcal{E}\{|v(\theta)|^2\} = f_{\mathrm{MR}}(\theta)$, see e.g., [8, page 183]. There is not apparent observational feature distinguishing these three realizations, at least over the window where they have been drawn, and hence, they are produced here only to satisfy curiosity.

## VIII. Concluding remarks

The present study sought to explore the issue of the time-arrow in the context of the maximum entropy ansatz. When the index of a random process designates a variable other than time, the principle can be called into question. A more abstract version of seeking spectra *maximally noncommittal to unavailable data*, such as gaps in a record, suggests other alternatives, including the one studied herein.

At the moment, the information theoretic significance of $\mathbb{J}(d\mu(\theta)/d\theta)$ is still under consideration. However, it is clear that, in the same way that entropy rates relate to a level of "surprise" when tracking the forward evolution of a random process, similarly $\mathbb{J}$ relates to a situation where we record new values of a random process at widely separated gaps of an earlier record. Regarding the significance of MR-spectra in time-series analysis, examples similar to the one that we presented here suggest similar qualities to the ME-ones (though, admitedly, they are slightly less appealing in terms of their ease of computation).

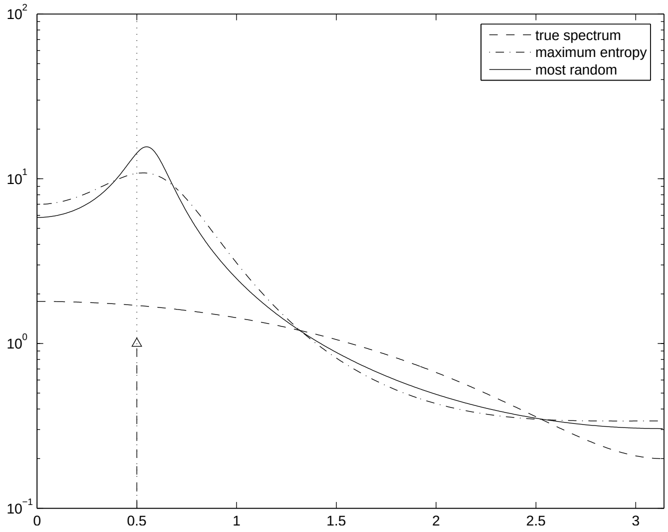

Fig. 1. Power spectra consistent with $R_0$, $R_{\pm1}$, $R_{\pm2}$, $R_{\pm3}$.

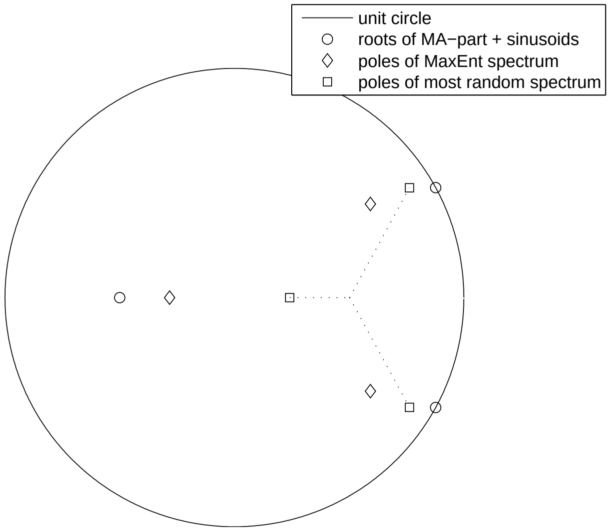

Fig. 2. Poles/zeros of $f_{\rm true}$, and singularities of $f_{\rm ME}$, and $f_{\rm MR}$.

## References

[1] E.F. Beckenbach and R. Bellman, **Inequalities**, Springer-Verlag, Berlin-Heidelberg, 198 pages,1965.

[2] J. Burg, **Maximum entropy spectral analysis**, Ph.D. dissertation, Stanford University, 1975.

[3] I. Csiszár, "Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems," *The Annals of Probability*, **19(4)**: 2032-2066, 1991.

[4] T.T. Georgiou, "Solution of the general moment problem via a one-parameter imbedding," *IEEE Trans. on Automatic Control*, **50(6):** 811-826, June 2005.

[5] T.T. Georgiou, "Relative Entropy and the multi-variable multi-dimensional Moment Problem," *IEEE Trans. on Information Theory*, to appear; preprint available at: arXiv:math.OC/0506124v1.

[6] T.T. Georgiou, "The Carathéodory-Fejér-Pisarenko decomposition and its multivariable counterpart," preprint, 29 pages: arXiv:math.OC/0509225v1.

[7] Ya. L. Geronimus, **Orthogonal Polynomials**, English translation from Russian by Consultants Bureau, New York, 570 pages, 1961.

[8] U. Grenander and G. Szegö, **Toeplitz Forms and their Applications**, Chelsea, 1958.

[9] S. Haykin, **Nonlinear Methods of Spectral Analysis,** Springer-Verlag, New York, 247 pages, 1979.

[10] E.T. Jaynes, "On the rationale of maximum entropy methods," *Proc. IEEE*, **70**: 939-952, 1982.

[11] R.D. Levine and M. Tribus (editors), **The Maximum Entropy Formalism**, MIT Press, Cambridge, 1979.

[12] P. Masani, Recent trends in multivariate prediction theory, in **Multivariate Analysis**, P.R. Krishnaiah, Ed., Academic Press, pp. 351-382, 1966.

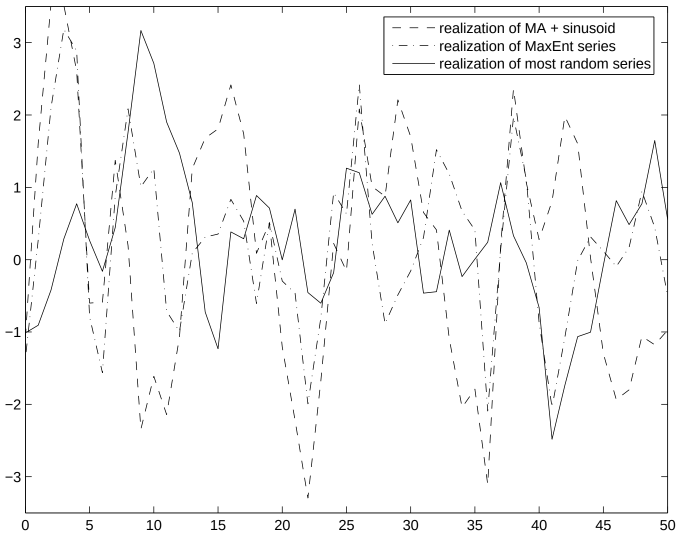

Fig. 3. Realizations of time-series according to $f_{\mathrm{true}}$, $f_{\mathrm{ME}}$, and $f_{\mathrm{MR}}$.

[13] P. Stoica and R. Moses, **Introduction to Spectral Analysis**, Prentice Hall, 2005.
[14] G. Szegö, "Über die randwerten eiher analytischen functionen," *Math. Ann.*, **84:** 232-244, 1921.
[15] S.R.S. Varadhan, **Probability Theory**, AMS, 2000.