

Optimality of Operator-Like Wavelets for Representing Sparse AR(1) Processes

Pedram Pad and Michael Unser

Biomedical Imaging Group, EPFL, Switzerland

Abstract—It is known that the Karhunen-Loève transform (KLT) of Gaussian first-order auto-regressive (AR(1)) processes results in sinusoidal basis functions. The same sinusoidal bases come out of the independent-component analysis (ICA) and actually correspond to processes with completely independent samples. In this paper, we relax the Gaussian hypothesis and study how orthogonal transforms decouple symmetric- α -stable (S α S) AR(1) processes. The Gaussian case is not sparse and corresponds to $\alpha = 2$, while $0 < \alpha < 2$ yields processes with sparse linear-prediction error. In the presence of sparsity, we show that operator-like wavelet bases do outperform the sinusoidal ones. Also, we observe that, for processes with very sparse increments ($0 < \alpha \leq 1$), the operator-like wavelet basis is indistinguishable from the ICA solution obtained through numerical optimization. We consider two criteria for independence. The first is the Kullback-Leibler divergence between the joint probability density function (pdf) of the original signal and the product of the marginals in the transformed domain. The second is a divergence between the joint pdf of the original signal and the product of the marginals in the transformed domain, which is based on Stein's formula for the mean-square estimation error in additive Gaussian noise. Our framework then offers a unified view that encompasses the discrete cosine transform (known to be asymptotically optimal for $\alpha = 2$) and Haar-like wavelets (for which we achieve optimality for $0 < \alpha \leq 1$).

Index Terms—Operator-like wavelets, independent-component analysis, auto-regressive processes, stable distributions.

I. INTRODUCTION

Transform-domain processing is a classical approach to compress signals, model data, and extract features. The underlying idea is that the transform-domain coefficients exhibit a loosened interdependence so that a simple point-wise processing can be applied. For instance, in the discrete domain, the Karhunen-Loève transform (KLT) is famous for yielding optimal transform coefficients that are uncorrelated and therefore also independent, provided the process is Gaussian. Also, if the process is stationary with finite variance and infinite length, then the KLT is a Fourier-like transform (FT-like). Moreover, it is known that FT-like transforms such as the discrete cosine transform are asymptotically equivalent to the KLT for AR(1) processes [1], [2]; thus, for a Gaussian input, all these transforms result in a fully decoupled (independent) representation. However, this favorable independence-related property is extinguished for non-Gaussian processes. In this case, the coefficients are only partially decoupled and the representation of the signal afforded by the KLT is suboptimal.

In recent years, wavelets have emerged as an alternative representation of signals and images. Typical examples of successful applications are JPEG2000 for image compression [3] and shrinkage methods for attenuating noise [4], [5]. Their wide success for transform-domain processing recommends them as good candidates for decoupling practical processes. This empirical observation was established by early studies that include [6], where many natural images were subjected to an independent-component analysis (ICA). It was found that the resulting components have properties that are reminiscent of 2D wavelets and/or Gabor functions. Additional ICA experiments were performed in [7] on realizations of the stationary sawtooth process and of Meyer's ramp process [8]; for both processes, the basis vectors of ICA exhibit a wavelet-like multiresolution structure.

Unfortunately, despite their empirical usefulness, the optimality of wavelets for the representation of non-Gaussian stochastic processes remains poorly understood from a theoretical point of view. An early study can be traced back to [9], where the decomposition of fractional Brownian motions over a wavelet basis was shown to result in almost uncorrelated coefficients, under some conditions. (Ultimately however, the truly optimal domain to represent fractional Brownian motions is known to be simply the FT domain.) Meanwhile, in a deterministic framework, it was shown in [10] that wavelets are optimal (up to some constant) for the N -term approximation of functions in Besov spaces; the extension of this result to a statistical framework could be achieved only experimentally.

Recently, a new tool for the study of generalized innovation models has been proposed in [11], [12]. It is particularly well suited to the investigation of symmetric- α -stable (S α S) white noises, which can be used to drive first-order stochastic differential equations (SDE) to synthesize AR(1) processes. As it turns out, AR(1) systems and α -stable distributions are at the core of signal modeling and probability theory. The classical Gaussian processes correspond to $\alpha = 2$, while $0 < \alpha < 2$ yields stable processes that have heavy-tailed statistics and that are prototypical representatives for sparse signals [13].

In this paper, we take advantage of this tool to establish the optimality of a certain class of wavelets in a stochastic sense. We start by characterizing the amount of dependency between the coefficients of stochastic processes represented in an arbitrary transform domain. We consider two measures of dependency. The first is based on the Kullback-Leibler divergence and the second is based on Stein's formula for the variance of estimation of a signal distorted by additive white Gaussian noise (AWGN). Then, we seek the orthogonal trans-

formation that minimizes these measures of dependency. We confirm the extinction of the optimality of FT-like transforms for $0 < \alpha < 2$ and validate the superiority of the operator-like wavelet transforms proposed in [14]. Also, by finding the optimal transform for different values of α , we demonstrate that, for a positive α less than some threshold, operator-like wavelets are optimal.

This paper is organized as follows: We start by exposing three preliminary concepts like i) measures of divergence between distributions that would be suitable for either noise attenuation or compression applications (Section II); ii) the signal model fundamental to this paper (Section III-A); and iii) operator-like wavelets, (Section III-B). In Section IV, we describe our performance criteria in the context of transform-domain compression and noise attenuation. In addition, we provide an iterative algorithm to find the optimal basis. Results for different AR(1) processes and different transform domains are discussed in Section V. The last section is dedicated to the recapitulation of the main results, the relation to prior works, and topics for future studies.

II. PERFORMANCE MEASURES

In statistical modeling, one interesting problem is the best-achievable performance when the model does not match the reality. In the following we address this issue for the two problems of compression and denoising when the assumed distribution and the real one may differ.

1) *Compression Based on Non-Exact Distribution:* It is well-known that, if we have a source \mathbf{s} of random vectors with pdf $p_{\mathbf{s}}$, then the minimum coding set log-measure (MCSM) of these vectors is

$$\text{MCSM} = \mathbb{H}(p_{\mathbf{s}}) = - \int p_{\mathbf{s}}(\mathbf{s}) \log p_{\mathbf{s}}(\mathbf{s}) d\mathbf{s} \quad (1)$$

which is the entropy of the source. However, if we compress \mathbf{s} assuming $q_{\mathbf{s}}$ as its distribution, then

$$\begin{aligned} \text{CSM}(q_{\mathbf{s}}) &= \text{MCSM} + \mathbb{D}(p_{\mathbf{s}} \| q_{\mathbf{s}}) \\ &= \text{MCSM} + \int p_{\mathbf{s}}(\mathbf{s}) \log \frac{p_{\mathbf{s}}(\mathbf{s})}{q_{\mathbf{s}}(\mathbf{s})} d\mathbf{s} \end{aligned} \quad (2)$$

in which $\mathbb{D}(\cdot \| \cdot)$ is the Kullback-Leibler divergence.

Typically, when there is a statistical dependency between the entries of \mathbf{s} , compressing the vector based on the exact distribution is often intractable. Thus, the common strategy is to expand the vector in some other basis and to then do the compression entry-wise (neglecting the dependency between entries of the transformed vector). This is equivalent to do the compression assuming that the signal distribution is the product of the marginal distributions. Thus, if the transformed vector is $\mathbf{y} = \mathbf{H}\mathbf{s}$, then the normalized redundant information remaining in the compressed signal is

$$\begin{aligned} R(\mathbf{H}) &= \frac{1}{N} (\text{CSM}(p_{y_1}(y_1) \cdots p_{y_N}(y_N)) - \text{MCSM}) \\ &= \frac{1}{N} \mathbb{D}(p_{\mathbf{y}}(\mathbf{y}) \| p_{y_1}(y_1) \cdots p_{y_N}(y_N)), \end{aligned} \quad (3)$$

where N is the number of entries in \mathbf{s} . This is the first measure of performance of the transform \mathbf{H} that we are going to use in this paper. Also, this criterion is commonly used in ICA to find the “most-independent” representation [15].

2) *Denoising Based on Non-Exact Distribution:* Now, consider the problem of estimating \mathbf{s} from the noisy measurement

$$\mathbf{z} = \mathbf{s} + \mathbf{n} \quad (4)$$

where \mathbf{n} is an N -dimensional white Gaussian noise independent from \mathbf{s} . Our prior knowledge is the N th order pdf $p_{\mathbf{s}}(\cdot)$ of the signal. Under these assumptions and according to Stein [16], the estimator of minimum mean square error (MMSE) can be represented as

$$\mathbb{E}\{\mathbf{s}|\mathbf{z}\} = \mathbf{z} + \sigma^2 \nabla \log p_{\mathbf{z}}(\mathbf{z}). \quad (5)$$

where $p_{\mathbf{z}}(\mathbf{z}) = (p_{\mathbf{s}} * p_{\mathbf{n}})(\mathbf{z})$ is the N -dimensional pdf of the noisy measurements. Thus, the MSE given \mathbf{z} is

$$\begin{aligned} &\mathbb{E}_{\mathbf{s}|\mathbf{z}} \left\{ (\mathbf{s} - \mathbb{E}\{\mathbf{s}|\mathbf{z}\})^2 \right\} \\ &= \int \|\mathbf{s} - \mathbf{z}\|^2 p(\mathbf{s}|\mathbf{z}) d\mathbf{s} - \sigma^4 \|\nabla \log p_{\mathbf{z}}(\mathbf{z})\|^2 \\ &= N\sigma^2 + \sigma^4 \Delta \log p_{\mathbf{z}}(\mathbf{z}). \end{aligned} \quad (6)$$

Averaging over \mathbf{z} , we have

$$\begin{aligned} \text{MMSE} &= N\sigma^2 - \sigma^4 \int p_{\mathbf{z}}(\mathbf{z}) \|\nabla \log p_{\mathbf{z}}(\mathbf{z})\|^2 d\mathbf{z} \\ &= N\sigma^2 + \sigma^4 \int p_{\mathbf{z}}(\mathbf{z}) \Delta \log p_{\mathbf{z}}(\mathbf{z}) d\mathbf{z}, \end{aligned} \quad (7)$$

However, if we perform the MMSE estimator assuming $q_{\mathbf{s}}$ as the distribution of \mathbf{s} , then by using (5)-(7), the MSE of estimation becomes

$$\text{MSE}(q_{\mathbf{s}}) = \text{MMSE} + \sigma^4 \int p_{\mathbf{z}}(\mathbf{z}) \left\| \nabla \log \frac{p_{\mathbf{z}}(\mathbf{z})}{q_{\mathbf{z}}(\mathbf{z})} \right\|^2 d\mathbf{z} \quad (8)$$

where $q_{\mathbf{z}}(\mathbf{z})$ is the distribution induced on \mathbf{z} in (4) when the distribution on \mathbf{s} is $q_{\mathbf{s}}(\mathbf{s})$. Here, notice the pleasing similarity between (1)-(2) and (7)-(8).

If the entries of \mathbf{s} are dependent, then the entries of \mathbf{z} are dependent, too. Then, performing the exact MMSE estimator is once again often infeasible. The common scheme is then to take \mathbf{z} into a transform domain, perform an entry-wise denoising (regardless of the dependency between coefficients), and map the result back into the original domain. If the transformation \mathbf{H} is unitary, the performance of this scheme would be $\text{MSE}(p_{y_1}(y_1) \cdots p_{y_N}(y_N))$ where $\mathbf{y} = \mathbf{H}\mathbf{z}$ due to Parseval's relation. We write this as a function of \mathbf{H} normalized by the dimensionality of \mathbf{s} , with

$$\text{MSE}(\mathbf{H}) = \frac{1}{N} \text{MSE}(p_{y_1}(y_1) \cdots p_{y_N}(y_N)) \quad (9)$$

which is the second measure of performance that we are going to consider in this paper.

III. MODELING AND WAVELET ANALYSIS OF $S_{\alpha}S$ AR(1) PROCESSES

In this section, we first give the definition of continuous-domain $S_{\alpha}S$ AR(1) processes and their discrete-domain counterparts. Then, we discuss about the operator-like wavelets that are tuned to this kind of processes.

A. S α S AR(1) Processes

In [11], the authors model the stochastic signal s as a purely innovative process (i.e., *white noise*), having undergone a linear operation. Thus,

$$s = L^{-1}w \quad (10)$$

where w is a continuous-domain white noise and L^{-1} (the inverse of the whitening operator L) is a linear operator.

A general white noise is a probability measure on the dual space of a set of test functions that has the following properties [17]:

- For a given test function φ , the statistics of the random variable $\langle w, \varphi \rangle$ do not change upon shifting φ , where w (the realization of the noise) denotes a generic random element in the dual space of test functions (typically, Schwartz space of tempered distributions).
- If the test functions in the collection $\{\varphi_\beta\}_{\beta \in B}$ (B is an index set) have disjoint supports, then the random variables in $\{\langle w, \varphi_\beta \rangle\}_{\beta \in B}$ are independent.

Under some mild regularity conditions, there is a one-to-one correspondence between the infinitely divisible (id) random variables and the white noises specified above. Thus, specifying a white noise is equivalent to having the random variable $\langle w, \varphi \rangle$ for any test function φ .

Correspondingly, if L^* denotes the adjoint operator of L , then we have

$$\langle s, \varphi \rangle = \langle w, L^{-1*} \varphi \rangle \quad (11)$$

which means that one can readily deduce the statistical distribution of $\langle s, \varphi \rangle$ from the characterization of the process w .

Now, if w is S α S white noise, then the random variable $\langle w, \varphi \rangle$ has an S α S distribution whose characteristic function is given by

$$\hat{p}_{\langle w, \varphi \rangle}(\omega) = \mathbb{E}\{e^{j\omega \langle w, \varphi \rangle}\} = e^{-|\|\varphi\|_\alpha \omega|^\alpha}. \quad (12)$$

In the case of an AR(1) process, we have that

$$L = D + \kappa I \quad (13)$$

where D and I are respectively the differentiator and the identity operator; then, s in (10) is a continuous-domain S α S AR(1) process. The impulse response of L^{-1} is the causal exponential

$$\rho_\kappa(t) = e^{-\kappa t} \mathbf{1}_+(t) \quad (14)$$

where $\mathbf{1}_+(t)$ is the unit step. Thus, as a function of t , we can write

$$s(t) = (\rho_\kappa * w)(t). \quad (15)$$

The AR(1) process is well-defined for $\kappa > 0$. The limit case $\kappa = 0$ can also be handled by setting the boundary condition $s(0) = 0$, which results in a Lévy process that is non-stationary. Realizations of AR(1) processes for $\kappa = 0.05$ and for different values of α are depicted in Figure 1. When α decreases, the process becomes sparser in the sense that its innovation becomes more and more heavy-tailed.

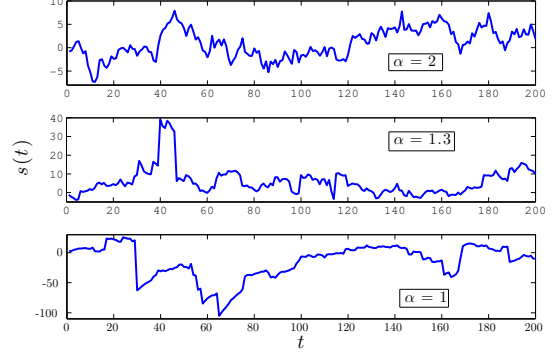


Fig. 1: Examples of AR(1) processes for different α .

Now, for a given integer k and time period T , set

$$\varphi_k(t) = \delta(t - kT) - e^{-\kappa T} \delta(t - (k-1)T) \quad (16)$$

and define w_k as

$$w_k = \langle s, \varphi_k(t) \rangle = s(kT) - e^{-\kappa T} s((k-1)T). \quad (17)$$

This means that the sampled version $\{s_k = s((k-1)T)\}_{k \in \mathbb{Z}}$ of $s(t)$ satisfies the first-order difference equation

$$s_k = e^{-\kappa T} s_{k-1} + w_k. \quad (18)$$

Also, we have that

$$w_k = \langle s, \varphi_k(t) \rangle = \langle w, (\check{\rho}_\kappa * \varphi_k)(t) \rangle \quad (19)$$

where $\check{\rho}_\kappa(t) = \rho_\kappa(-t)$ is the impulse response of L^{-1*} , the inverse of the adjoint operator of L . Also,

$$(\check{\rho}_\kappa * \varphi_k)(t) = \beta_{\kappa, T}(t - kT) = \mathbf{1}_{[kT, (k+1)T)} e^{-\kappa(t-kT)} \quad (20)$$

is the exponential B-spline with parameters κ and T [12]. The fundamental property here is that the kernels $\{\beta_{\kappa, T}(\cdot - kT)\}_{k \in \mathbb{Z}}$ are shifted replicates of each other and have compact and disjoint supports. Thus, according to the definition of a white noise, $\{w_k\}_{k \in \mathbb{Z}}$ is an iid sequence of S α S random variables with the common characteristic function

$$\hat{p}_w(\omega) = \mathbb{E}\{e^{j\omega \langle w, \beta_{\kappa, T} \rangle}\} = e^{-|\|\beta_{\kappa, T}\|_\alpha \omega|^\alpha}. \quad (21)$$

The conclusion is that a continuous-domain AR(1) process maps into the discrete AR(1) process $\{s_k\}_{k \in \mathbb{Z}}$ that is uniquely specified by (18) and (21).

We now consider N consecutive samples of the process and define the random vectors $\mathbf{s} = [s_1 \cdots s_N]^\top$ and $\mathbf{w} = [w_1 \cdots w_N]^\top$. This allows us to rewrite (18) as

$$\mathbf{s} = \mathbf{L}^{-1} \mathbf{w} \quad (22)$$

where $\mathbf{L}^{-1} = [\bar{l}_{ij}]_{N \times N}$ and

$$\bar{l}_{ij} = e^{-\kappa T(j-i)} \cdot \mathbf{1}_{\{j \geq i\}} \quad (23)$$

which is the discrete-domain counterpart of (14).

In the next sections, we are going to study linear transforms applied to the signal s (or \mathbf{s}). Here, we recall a fundamental property of stable distributions that we shall use in our derivations.

Property 1 (Linear combination of S α S random variables): Let $\bar{r} = \sum_{m=1}^M a_m r_m$ where r_m are iid S α S random variables with dispersion parameter c . Then, \bar{r} is an S α S as well with dispersion parameter $\| [a_1, \dots, a_M] \|_\alpha c$.

For more explanation, assume that r_1, \dots, r_M are M iid S α S random variables with common characteristic function $e^{-|c\omega|^\alpha}$, and a_1, \dots, a_M are M arbitrary real numbers. Then, for the characteristic function of the random variable $r^* = \sum_{m=1}^M a_m r_m$, we have that

$$\hat{p}_{r^*}(\omega) = \prod_{m=1}^M e^{-|a_m c \omega|^\alpha} = e^{-|(\sum_{m=1}^M |a_m|^\alpha)^{1/\alpha} c \omega|^\alpha}. \quad (24)$$

Thus, r^* , which is a linear combination of iid S α S random variables, is an S α S random variable with the same distribution as one of them multiplied by the factor $(\sum_{m=1}^M |a_m|^\alpha)^{1/\alpha}$; i.e.

$$r^* \stackrel{d}{=} \left(\sum_{m=1}^M |a_m|^\alpha \right)^{1/\alpha} r_1. \quad (25)$$

B. Operator-Like Wavelets

Conventional wavelet bases act as smoothed versions of the derivative operator. To decouple the AR(1) process in (15) by a wavelet-like transform, we need to choose basis functions that essentially behave like the whitening operator L in (13). Such wavelet-like basis functions are called operator-like wavelets and can be tailored to any given differential operator L [14]. The operator-like wavelet at scale i and location k is given by

$$\psi_{i,k} = L^* \phi_i(\cdot - 2^i k T), \quad (26)$$

where ϕ_i is a scale-dependent smoothing kernel. Since $\{\psi_{i,k}\}$ is an orthonormal basis and $s = L^{-1}w$, the wavelet coefficients of the signal s are

$$\begin{aligned} v_{i,k} &= \langle s, \psi_{i,k} \rangle = \langle L^{-1}w, \psi_{i,k} \rangle \\ &= \langle w, L^{-1*} L^* \phi_i(\cdot - 2^i k T) \rangle = \langle w, \phi_i(\cdot - 2^i k T) \rangle. \end{aligned} \quad (27)$$

Based on this equality, we understand that, for any given i and for all k , the $v_{i,k}$ follows an S α S distribution with width parameter $\|\phi_i\|_\alpha$ [11]. Also, since w is independent at every point, intuitively, the level of decoupling has a direct relation to the overlap of the smoothing kernels $\phi_i(\cdot - 2^i k T)$. The operator-like wavelets proposed in [14] are very similar to Haar wavelets, except that they are piecewise exponential instead of piecewise constant (for $\kappa = 0$). Then satisfy

$$\psi_{i,k}(t) \propto \quad (28)$$

$$\begin{aligned} & e^{-2^{-i}\kappa T} \beta_{\kappa, 2^{-i}T}(t - k2^{-i}T) - \beta_{\kappa, 2^{-i}T}(t - (k+1)2^{-i}T) \\ &= \begin{cases} 0 & t < k2^{-i}T \\ e^{-\kappa(t-(k-1)2^{-i}T)} & k2^{-i}T \leq t < (k+1)2^{-i}T \\ -e^{-\kappa(t-(k+1)2^{-i}T)} & (k+1)2^{-i}T \leq t < (k+2)2^{-i}T \\ 0 & (k+2)2^{-i}T \leq t \end{cases} \end{aligned}$$

For these wavelets, the supports of $\phi_{i,k}$ do not overlap within the given scale i . Thus, the wavelet coefficients at scale i are independent and identically distributed. This property

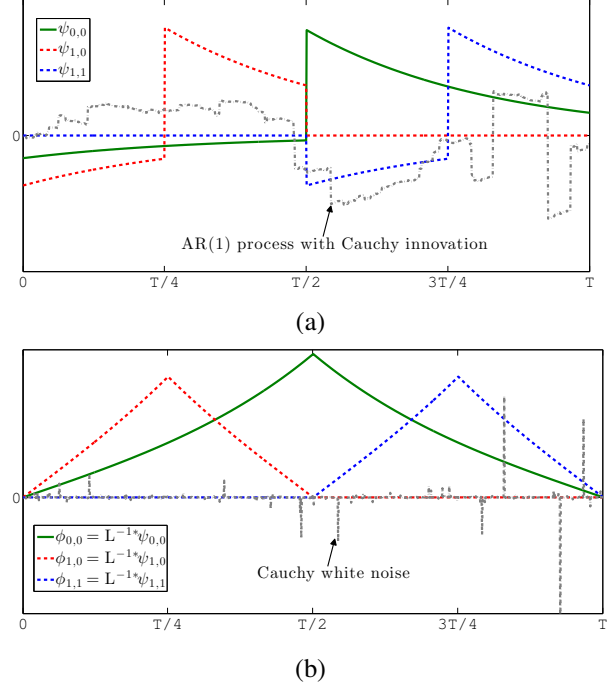


Fig. 2: (a) Operator-like wavelets in two consecutive scales acting on an AR(1) process with Cauchy excitation. (b) The equivalent windows (smoothing kernels) acting on the underlying Cauchy white noise. Note that $\psi_{1,0}$ and $\psi_{1,1}$ ($\phi_{1,0}$ and $\phi_{1,1}$, respectively) are non-overlapping.

suggests that this type of transform is an excellent candidate for decoupling AR(1) processes. The illustration of plugging these wavelets into (27) is given in Figure 2.

IV. SEARCH FOR THE OPTIMAL TRANSFORMATION

For now on, we assume that the signal vector $\mathbf{s} = [s_1 \dots s_N]^\top$ with $s_k = s((k-1)T)$ is obtained from the samples of an S α S AR(1) process and satisfies the discrete innovation model (18). The representation of the signal \mathbf{s} in (22) in the transform domain is denoted by $\mathbf{y} = [y_1 \dots y_N]^\top = \mathbf{H}\mathbf{s}$, where $\mathbf{H} = [h_{ij}]_{N \times N}$ is the underlying orthogonal transformation matrix (e.g., DCT, wavelet transform). Let us now use (3) to characterize the performance of a given transformation matrix \mathbf{H} . First, we simplify (3) to

$$\begin{aligned} \mathbf{R}(\mathbf{H}) &= \frac{1}{N} \sum_{n=1}^N \mathbb{H}(y_n) - \frac{1}{N} \mathbb{H}(\mathbf{y}) \\ &= \frac{1}{N} \sum_{n=1}^N \mathbb{H}(y_n) - \mathbb{H}(w_1) - \frac{1}{N} \log \det \mathbf{H} \mathbf{L}^{-1}, \end{aligned} \quad (29)$$

where we also observe that $\log \det \mathbf{H} \mathbf{L}^{-1} = 0$. In addition, since the w_m is α -stable, we can write $y_n \stackrel{d}{=} \bar{h}_n w_1$, where \bar{h}_n is the α -(pseudo)norm of the n th row of $\mathbf{H} \mathbf{L}^{-1}$ (see Property 1) given by

$$\bar{h}_n = \left(\sum_{r=1}^N \left| \sum_{m=1}^N h_{nm} \bar{l}_{mr} \right|^\alpha \right)^{\frac{1}{\alpha}}. \quad (30)$$

It follows that

$$\mathbf{R}(\mathbf{H}) = \frac{1}{N} \sum_{n=1}^N \log \bar{h}_n, \quad (31)$$

which can be readily calculated for any given \mathbf{H} .

Note 1: This criterion is reminiscent of the sum-of-dispersion criterion $\sum_{n=1}^N \bar{h}_n$ which is frequently used in the study of α -stable stochastic processes [18], [19]. Unlike (31), the dispersion criterion does not have a direct information-theoretic interpretation.

As second option, we use the criterion (9) to measure the performance of a given transform matrix \mathbf{H} . Again, it can be simplified to

$$\text{MSE}(\mathbf{H}) = \sigma^2 - \frac{\sigma^4}{N} \sum_{n=1}^N \int \frac{(p'_{y_n}(y_n))^2}{p_{y_n}(y_n)} dy_n, \quad (32)$$

in which y_n is the n th entry of $\tilde{\mathbf{y}} = \mathbf{H}\mathbf{z}$. According to Property 1 of α -stable random variables, we write $\tilde{y}_n \stackrel{d}{=} \bar{h}_n w_1 + n_1$ where n_1 is a standard Gaussian random variable. This allows us to deduce the pdf expression

$$p_{\tilde{y}_n}(y) = \frac{1}{\bar{h}_n} p_{w_1}\left(\frac{y}{\bar{h}_n}\right) * p_{n_1}(y). \quad (33)$$

Thus, (32) is calculable through one-dimensional integrals.

For the sake of the optimization process, we also need to derive the gradient of the cost functions \mathbf{R} and MSE with respect to \mathbf{H} . Specifically, according to (30) and (31), the partial derivative of $\mathbf{R}(\mathbf{H})$ is

$$\frac{\partial \mathbf{R}}{\partial h_{ij}} = \frac{1}{N \alpha \bar{h}_i^\alpha} \frac{\partial \bar{h}_i^\alpha}{\partial h_{ij}} \quad (34)$$

where

$$\frac{\partial \bar{h}_i^\alpha}{\partial h_{ij}} = \alpha \sum_{r=1}^N l_{jr} \text{sgn} \left(\sum_{n=1}^N h_{ik} l_{kr} \right) \left| \sum_{n=1}^N h_{ik} l_{kr} \right|^{\alpha-1}. \quad (35)$$

Also, the partial derivative of $\text{MSE}(\mathbf{H})$ in (32) is

$$\begin{aligned} \frac{\partial \text{MSE}}{\partial h_{ij}} &= -\frac{\sigma^4}{N} \frac{\partial}{\partial \bar{h}_i} \int \frac{(p_{\tilde{y}_i}^{(1)}(u))^2}{p_{\tilde{y}_i}(u)} du \times \frac{\bar{h}_i^{1-\alpha}}{\alpha} \frac{\partial \bar{h}_i^\alpha}{\partial h_{ij}} \\ &= -\frac{\sigma^4}{N} \left(2 \int \frac{\partial}{\partial \bar{h}_i} p_{\tilde{y}_i}^{(1)}(u) \frac{p_{\tilde{y}_i}^{(1)}(u)}{p_{\tilde{y}_i}(u)} du \right. \\ &\quad \left. - \int \frac{\partial}{\partial \bar{h}_i} p_{\tilde{y}_i}(u) \left(\frac{p_{\tilde{y}_i}^{(1)}(u)}{p_{\tilde{y}_i}(u)} \right)^2 du \right) \\ &\quad \times \frac{\bar{h}_i^{1-\alpha}}{\alpha} \frac{\partial \bar{h}_i^\alpha}{\partial h_{ij}} \end{aligned} \quad (36)$$

in which $p_{\tilde{y}_i}^{(k)}(\cdot)$ is the k th derivative of $p_{\tilde{y}_i}(\cdot)$ which, according to (33), can be written as

$$p_{\tilde{y}_i}^{(k)}(\cdot) = p_{y_i}(s) * \frac{d^k}{ds^k} \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{s^2}{2\sigma^2}} \right). \quad (37)$$

Also, we have that

$$\frac{\partial}{\partial h_i} p_{\tilde{y}_i}(y) = -\frac{1}{h_i} p_{\tilde{y}_i}(y) - \frac{y}{h_i} p_{\tilde{y}_i}^{(1)}(y) - \frac{1}{h_i} p_{\tilde{y}_i}^{(2)}(y) \quad (38)$$

and

$$\frac{\partial}{\partial h_i} p_{\tilde{y}_i}^{(1)}(y) = -\frac{2}{h_i} p_{\tilde{y}_i}^{(1)}(y) - \frac{y}{h_i} p_{\tilde{y}_i}^{(2)}(y) - \frac{1}{h_i} p_{\tilde{y}_i}^{(3)}(y). \quad (39)$$

Now, since the y_i have nice characteristic functions, we can calculate (37) efficiently through the inverse Fourier transform

$$p_{y_i}^{(k)}(\cdot) = \mathcal{F}_\omega^{-1} \left\{ (j\omega)^k e^{-|\bar{h}_i \omega|^\alpha - \frac{\sigma^2}{2} \omega^2} \right\} \quad (40)$$

using the FFT algorithm.

Thus, we can use gradient-based optimization to obtain the optimal transformations for different values of κ , α , and N . For our experiments, we implemented a gradient-descent algorithm with adaptive step size to efficiently find the optimal transform matrix. Since the transform matrix may deviate from the space of unitary matrices, after each step, we project it on that space using the method explained in Appendix I. The algorithm is as follows where C is the chosen measure of independence (i.e., \mathbf{R} or MSE):

Algorithm 1: ICA for S α S AR(1) Processes

- 1: **input:** N, α, κ
 - 2: **initialize:** $\mathbf{H}_{\text{old}}, \mu, a \in [1, +\infty)$ and $b \in [0, 1]$
 - 3: **repeat**
 - 4: $\tilde{\mathbf{H}}_{\text{new}} = \mathbf{H}_{\text{old}} - \mu \nabla C|_{\mathbf{H}_{\text{old}}}$
 - 5: Set \mathbf{H}_{new} to the projection of $\tilde{\mathbf{H}}_{\text{new}}$ onto the space of unitary matrices
 - 6: **if** $C(\mathbf{H}_{\text{new}}) < C(\mathbf{H}_{\text{old}})$ **then**
 - 7: $\mathbf{H}_{\text{old}} \leftarrow \mathbf{H}_{\text{new}}$
 - 8: $\mu \leftarrow a \cdot \mu$
 - 9: **else**
 - 10: $\mathbf{H}_{\text{new}} \leftarrow \mathbf{H}_{\text{old}}$
 - 11: $\mu \leftarrow b \cdot \mu$
 - 12: **end if**
 - 13: **until** convergence
 - 14: **return** \mathbf{H}_{new}
-

Algorithm 1 can be viewed as a model-based version of ICA. We take advantage of the underlying stochastic model to derive an optimal solution based on the minimization of (31) and (32), which involves the computation of ℓ_α -norms of the transformation matrix. By contrast, the classical version of ICA is usually determined empirically based on the observations of a process, but the ultimate aim is similar; namely, the decoupling of the data vector.

V. RESULTS FOR DIFFERENT TRANSFORMATIONS

Initially, we investigate the effect of the signal length N on the value of \mathbf{R} and MSE . We consider the case of a Lévy process (i.e., $\kappa = 0$) and numerically optimized the criteria for different α and plot it as a function of N . Results are depicted in Figure 3. As we see, the criteria values converge quickly to their asymptotic values. Thus, for the remainder of the experiments, we have chosen $N = 64$. This is a block size that is reasonable computationally and large enough to be representative of the asymptotic regime.

Then, we investigate the performance of different transforms for various processes. First, we focus on the Lévy processes. In

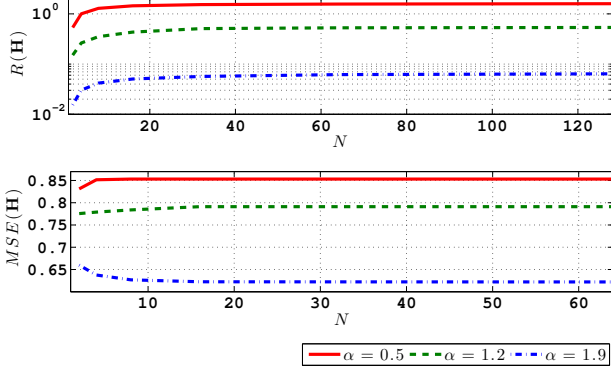


Fig. 3: Minimum value of $R(\mathbf{H})$ and $MSE(\mathbf{H})$ for Lévy processes as a function of N for different values of α . In the second plot $\sigma^2 = 1$.

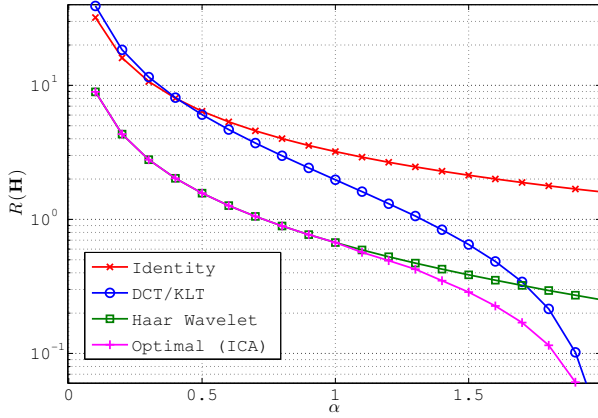


Fig. 4: $R(\mathbf{H})$ of Lévy processes versus α when $n = 64$ for different \mathbf{H} .

this case, the operator-like wavelet transform is the classical Haar wavelet transform (HWT). The performance criteria R and MSE as a function of α for various transforms are plotted in Figures 4 and 5, respectively. The considered transformations are as follows: identity as the baseline, discrete cosine transform (DCT), Haar wavelet transform (HWT), and optimal solution (ICA) provided by the proposed algorithm. In the case of $\alpha = 2$ (Gaussian scenario), the process s is a Brownian motion whose KLT is a sinusoidal transform that is known analytically. In this case, the DCT and the optimal transform converge to the KLT since being decorrelated is equivalent to being independent. We see this coincidence in both Figures 4 and 5. The vanishing of R at $\alpha = 2$ indicates perfect decoupling. By contrast, as α decreases, neither the DCT nor the optimal transform decouples the signal completely. The latter means that there is no unitary transform that completely decouples stable non-Gaussian Lévy processes. However, we see that, based on both criteria R and MSE , and as α decreases, the DCT becomes less favorable while the performance of the HWT gets closer to the optimal one. Moreover, Figures 4 and 5 even suggest that the Haar wavelet transform is equivalent to the ICA solution for $\alpha \leq 1$.

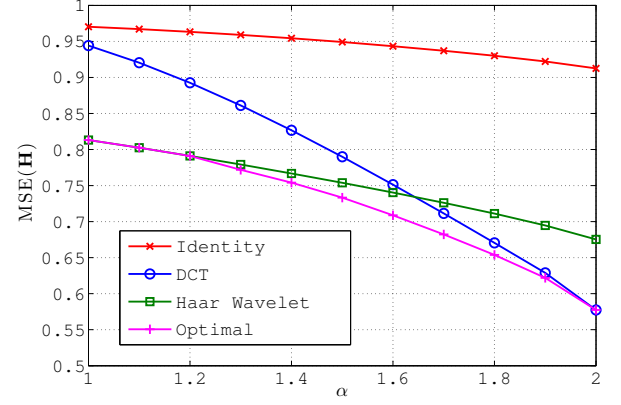


Fig. 5: $MSE(\mathbf{H})$ of Lévy processes versus α when $n = 64$ for different \mathbf{H} .

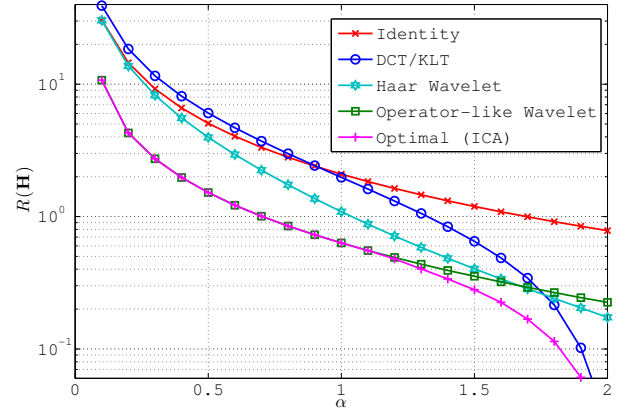


Fig. 7: $R(\mathbf{H})$ versus α when $e^{-\kappa T} = 0.9$ and $n = 64$ for different \mathbf{H} .

Also, to see the transition from sinusoidal bases to Haar wavelet bases, we plot the optimal basis which is obtained by the proposed algorithm at two consequent scales. In Figure 6, we see the progressive evolution of the ICA solution from the sinusoidal basis to the Haar basis while changing the parameter α of the model.

Next, we consider a stationary AR(1) process with $e^{-\kappa T} = 0.9$ and $n = 64$. For $\alpha = 2$, we get the well-known classical Gaussian AR(1) process for which the DCT is known to be asymptotically optimal [1], [2]. The performance criterion R versus α for the DCT, the HWT, the operator-like wavelet matched to the process, and the optimal ICA solution are plotted in Figure 7. Here too we see that, for $\alpha = 2$, ICA is equivalent the DCT. But, as α decreases, the DCT loses its optimality and the matched operator-like wavelet becomes closer to optimum. Again, we observe that, for $\alpha \leq 1$, the ICA solution is the matched operator-like wavelet described in Section III-B. The fact that the matched operator-like wavelet outperforms the HWT shows the benefit of the tuning of the wavelet to the differential characteristics of the process. Also, as shown in Figure 8, experimentally determined ICA basis functions for $\alpha = 1$ are indistinguishable from the wavelets in Figure 2.

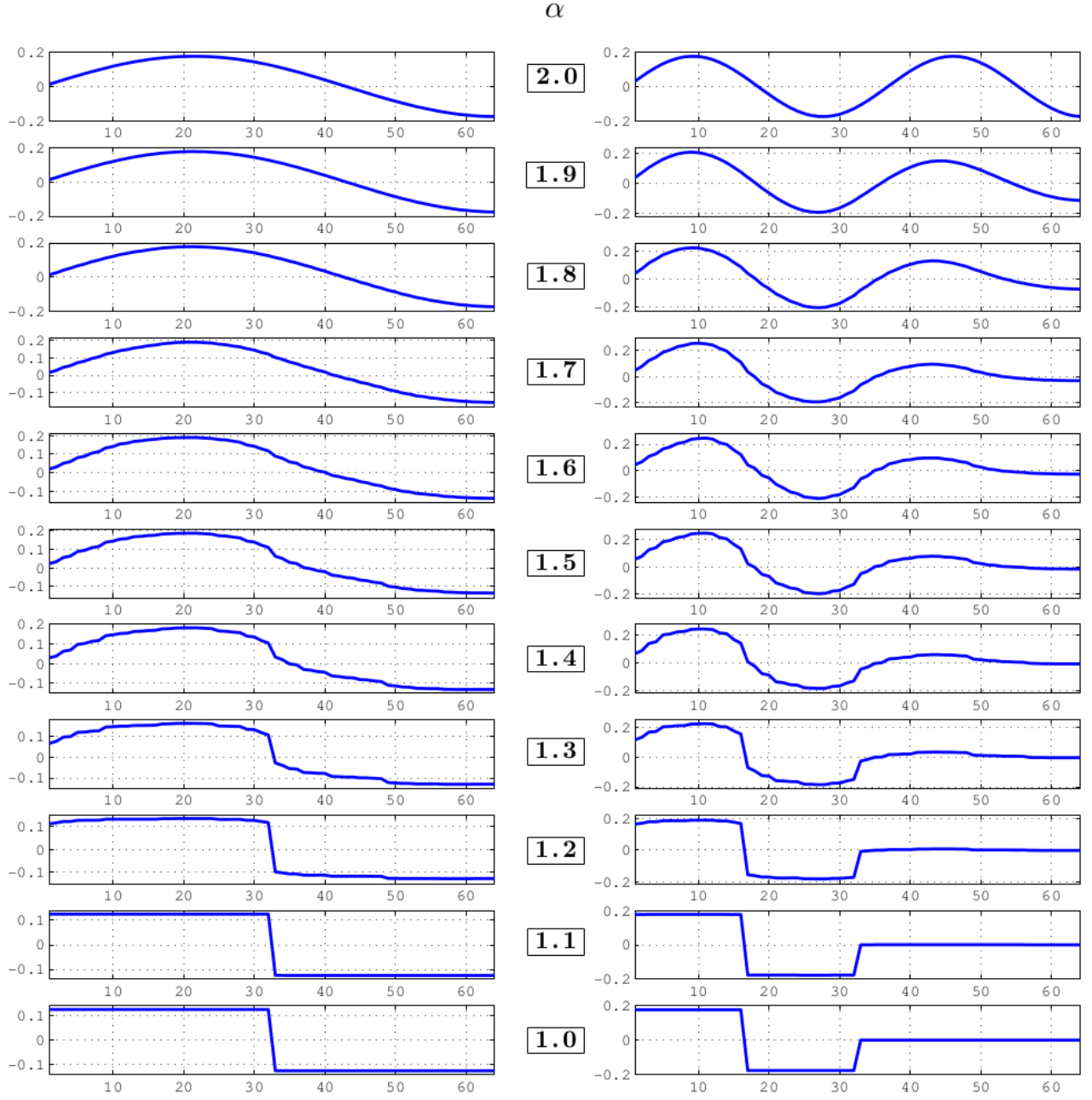


Fig. 6: Two rows of the optimal \mathbf{H} (ICA) for $\alpha = 2$ down to 1 when $N = 64$. In each row, we see the evolution from sinusoidal waves to Haar wavelets by increasing the sparsity of the underlying innovation process.

To substantiate those findings, we present a theorem that states that, based on the above mentioned criteria and for any $\alpha < 2$, the operator-like wavelet transform outperforms the DCT (or, equivalently, the KLT associated with the Gaussian member of the family) as the block-size N tends to infinity.

Theorem 1: If $\alpha < 2$ and $\kappa \geq 0$, we have that

$$\lim_{N \rightarrow \infty} R(\text{OpWT}) < \lim_{N \rightarrow \infty} R(\text{DCT}) = \infty \quad (41)$$

and

$$\lim_{N \rightarrow \infty} \text{MSE}(\text{OpWT}) < \lim_{N \rightarrow \infty} \text{MSE}(\text{DCT}) = \sigma^2, \quad (42)$$

where OpWT stands for the operator-like wavelet transform.

The proof is given in Appendix II.

In addition, this theorem states that, for $\alpha < 2$ and as N tends to ∞ , the performance of the DCT is equivalent to the trivial identity operator. This is surprising because, since the DCT is optimal for the Gaussian case ($\alpha = 2$), one may expect that it has a good result for other AR(1) processes. However, although this theorem does not assert that operator-like wavelets are the optimal basis, it still shows that, by applying them, we obtain a better performance than trivial transformations. Also, through simulations we observed that operator-like wavelets are close to optimal transform, particularly when the underlying white noise becomes very sparse.

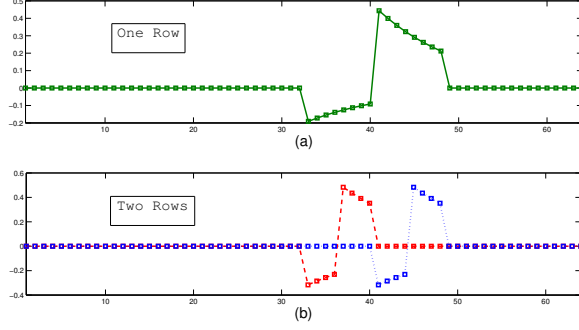


Fig. 8: Three rows of the optimal \mathbf{H} for $\alpha = 1$ and $n = 64$. Parts (a) and (b) show the dyadic structure of the wavelets.

VI. SUMMARY AND FUTURE STUDIES

In this paper, we focused on the simplest version (first-order differential system with an S α S excitation) of the sparse stochastic processes which have been proposed by Unser et al [11], [12]. Because of the underlying innovation model and the properties of S α S random variables, we could obtain a closed-form formula for the performance of different transform-domain representations and characterize the optimal transform. This is a novel model-based point of view for ICA. We proved that operator-like wavelets are better than sinusoidal transforms for decoupling the AR(1) processes with sparse excitations ($\alpha < 2$). This result is remarkable since sinusoidal bases are known to be asymptotically optimal for the classical case of $\alpha = 2$. Moreover, we showed that, for very sparse excitations ($\alpha \lesssim 1$), operator-like wavelets are equivalent to the ICA. As far as we know, this is the first theoretical results on the optimality of wavelet-like bases for a given class of stochastic processes.

Another interesting aspect of this study is that it gives a unified framework for Fourier-type transforms and a class of wavelet transforms. Now, the Fourier transform and the wavelet transforms were based on two different intuitions and philosophies. However, here we have a model in which we obtain both transform families just by changing the underlying parameters.

The next step in this line of research is to investigate the extent to which these findings can be generalized to other white noises or higher-order differential operators. Also, studying the problem in the original continuous domain would be theoretically very valuable.

APPENDIX I

PROJECTION ON THE SPACE OF UNITARY MATRICES

Suppose that \mathbf{A} is an $N \times N$ matrix. Our goal is to find the unitary matrix \mathbf{H}^* that is the closest to \mathbf{A} in Frobenius norm, in the sense that

$$\mathbf{H}^* = \arg \min_{\mathbf{H}} \|\mathbf{A} - \mathbf{H}\|_F. \quad (43)$$

According to singular-value decomposition (SVD), we can write $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top$ where \mathbf{U} and \mathbf{V} are unitary matrices and $\mathbf{\Lambda}$ is a diagonal matrix with nonnegative diagonal entries.

Since the Frobenius norm is unitarily invariant, we have that

$$\|\mathbf{A} - \mathbf{H}\|_F = \|\mathbf{\Lambda} - \mathbf{U}^\top \mathbf{H} \mathbf{V}\|_F \quad (44)$$

in which $\mathbf{U}^\top \mathbf{H} \mathbf{V}$ is a unitary matrix that we call \mathbf{K} . The expansion of the right-hand side of (44) gives

$$\begin{aligned} \|\mathbf{\Lambda} - \mathbf{K}\|_F^2 &= \sum_{1 \leq i, j \leq N} k_{ij}^2 + \sum_{i=1}^N \lambda_{ii}^2 - 2 \sum_{i=1}^N \lambda_{ii} k_{ii} \\ &= N + \sum_{i=1}^N \lambda_{ii}^2 - 2 \sum_{i=1}^N \lambda_{ii} k_{ii}. \end{aligned} \quad (45)$$

Since \mathbf{K} is unitary, $|k_{ii}| \leq 1$ for $i = 1, \dots, N$. Thus, setting $k_{ii} = 1$, which means setting $\mathbf{K} = \mathbf{I}$, minimizes (45). Consequently, the projection of \mathbf{A} on the space of unitary matrices is $\mathbf{H}^* = \mathbf{U}\mathbf{V}^\top$.

APPENDIX II

PROOF OF THEOREM 1

A. Proof of Part 1 (Equation (41))

According to (31), we have that

$$\begin{aligned} \mathbf{R}(\mathbf{H}) &= \frac{1}{N} \sum_{n=1}^N \log \bar{h}_n = \frac{1}{N} \sum_{n=1}^N \log \left(\frac{1}{\bar{h}_n^{-1}} \right) \\ &= \int_{\mathbb{R}} \log \left(\frac{1}{\gamma} \right) p(\gamma) d\gamma \end{aligned} \quad (46)$$

in which $p(\cdot)$ is the empirical distribution of \bar{h}_n^{-1} .

According to SVD, we can write $\mathbf{L}^{-1} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top$ where $\mathbf{\Lambda}$ is a diagonal matrix with λ_i as diagonal entries. Taking \mathbf{s} in the KLT domain is equivalent to multiplying it by \mathbf{U}^\top . The eigenvalues of the covariance of AR(1) matrices are known in closed form and are given by [20] and [21], for $\kappa \geq 0$, as

$$|\lambda_i|^{-1} = \sqrt{(1 - e^{-\kappa T})^2 + 4e^{-\kappa T} \sin^2 \left(\frac{\omega_i}{2} \right)} \quad (47)$$

and

$$\begin{aligned} v_{ij} &= \sqrt{\frac{2}{N + (1 - e^{-2\kappa T}) \lambda_i^2}} \\ &\quad \times \sin \left(\omega_i \left(j - \frac{N+1}{2} \right) + i \frac{\pi}{2} \right) \end{aligned} \quad (48)$$

in which ω_i , $i = 1, \dots, N$, is the i th positive root of

$$\tan(N\omega) = -\frac{(1 - e^{-2\kappa T}) \sin \omega}{\cos \omega - 2e^{-\kappa T} + e^{-2\kappa T} \cos \omega}. \quad (49)$$

Since $\tan(N\omega)$ is an injective function that sweeps the whole domain of the real numbers while $\omega \in \left[\frac{i-1}{N}\pi, \frac{i}{N}\pi \right]$, for $i = 1, \dots, N$, (49) has a single root in each of such intervals. Thus, as N tends to infinity, the empirical distribution of the ω_i tends to the uniform distribution on $[0, \pi]$. Then, starting from (47), one can obtain the limit empirical distribution of $|\lambda_i|$ as

$$p_\lambda(\lambda) = \frac{2}{\pi} \frac{\lambda}{\sqrt{\lambda^2 - (1 - e^{-\kappa T})^2} \sqrt{(1 + e^{-\kappa T})^2 - \lambda^2}}. \quad (50)$$

Now, $\sum_{j=1}^N v_{ij}^2 = 1$ means that

$$\sum_{j=1}^N \left| \sin \left(\omega_i \left(j - \frac{N+1}{2} \right) + i \frac{\pi}{2} \right) \right|^2 \sim \mathcal{O}(N) \quad (51)$$

as N tends to infinity. But, for $\alpha < 2$, we have that

$$\begin{aligned} & \left(\sum_{j=1}^N \left| \sin \left(\omega_i \left(j - \frac{N+1}{2} \right) + i \frac{\pi}{2} \right) \right|^\alpha \right)^{\frac{1}{\alpha}} \\ & \geq \left(\sum_{j=1}^N \left| \sin \left(\omega_i \left(j - \frac{N+1}{2} \right) + i \frac{\pi}{2} \right) \right|^2 \right)^{\frac{1}{\alpha}} \sim \mathcal{O}(N^{\frac{1}{\alpha}}). \end{aligned} \quad (52)$$

Thus, for $\alpha < 2$, $(\sum_{j=1}^N |v_{ij}|^\alpha)^{\frac{1}{\alpha}}$ grows faster than $\mathcal{O}(N^{\frac{1}{\alpha}-\frac{1}{2}})$ and thus tends to infinity as N tends to infinity. Consequently, the limit empirical distribution of \bar{h}_i^{-1} can be represented as

$$p(\gamma) = \begin{cases} \frac{2}{\pi} \frac{\gamma}{\sqrt{\gamma^2 - (1 - e^{-\kappa T})^2} \sqrt{(1 + e^{-\kappa T})^2 - \gamma^2}} & \alpha = 2 \\ \delta(\gamma) & \alpha \neq 2. \end{cases} \quad (53)$$

By plugging this result into (46), we conclude that, for $\alpha < 2$, $\lim_{N \rightarrow \infty} \mathbf{R}(\text{KLT}) = \infty$. This completes the proof of the right-hand side.

Now, for the proof of the left-hand side, we need to specify the matrix \mathbf{H} for the operator-like wavelet transform. This matrix is given by the recursive construction

$$\begin{aligned} \mathbf{H}_k &= \text{diag} \left(\sqrt{\frac{1 - e^{-2\kappa T}}{1 - e^{-2^{k+1}\kappa T}}}, \sqrt{\frac{1 - e^{-2\kappa T}}{1 - e^{-2^{k+1}\kappa T}}}, \overbrace{1, \dots, 1}^{2^k - 2} \right) \\ &\quad \times \begin{bmatrix} \ell_{k-1} & e^{-2^{k-1}\kappa T} \ell_{k-1} \\ -e^{-2^{k-1}\kappa T} \ell_{k-1} & \ell_{k-1} \\ \mathbf{H}'_{k-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{H}'_{k-1} \end{bmatrix} \end{aligned} \quad (54)$$

in which \mathbf{H}'_{k-1} is the matrix \mathbf{H}_{k-1} omitting the first row and $\ell_{k-1} = [1, e^{-\kappa T}, \dots, e^{-(2^{k-1}-1)\kappa T}]$. Also, $\mathbf{H}_0 = [1]$. Let us denote the empirical distribution of \bar{h}_i^{-1} (the reciprocal of the α -(pseudo) norm of the rows of $\mathbf{H}_k \mathbf{L}_{2^k}$) by $p_k(\gamma) = \sum_{i=1}^k p_i \delta(\gamma - \gamma_i)$. Now, for the sequence of p_i and γ_i , with respect to k , we have the following recursive relation:

- Replace p_{k-1} by $(\frac{p_{k-1}}{2}, \frac{p_{k-1}}{2})$
- Remove γ_{k-1} . Then, if $\kappa > 0$, set

$$\begin{aligned} \gamma_{k-1} &= \sqrt{\frac{1 - e^{-2^{k+1}\kappa T}}{1 - e^{-2\kappa T}}} \\ &\quad \times \left(\sum_{i=-2^{k-1}+1}^{2^k-1} \left(\frac{e^{-|i|\kappa T} - e^{-(2^k-|i|)\kappa T}}{1 - e^{-2\kappa T}} \right)^\alpha \right)^{-\frac{1}{\alpha}} \end{aligned} \quad (55)$$

and

$$\gamma_k = \sqrt{\frac{1 - e^{-2^{k+1}\kappa T}}{1 - e^{-2\kappa T}}} \left(\sum_{i=1}^{2^k} \left(\frac{1 - e^{-2i\kappa T}}{1 - e^{-2\kappa T}} \right)^\alpha \right)^{-\frac{1}{\alpha}} \quad (56)$$

else, if $\kappa = 0$, set

$$\gamma_{k-1} = 2^{\frac{k}{2}} \left(\sum_{i=-2^{k-1}+1}^{2^k-1} (2^{k-1} - |i|)^\alpha \right)^{-\frac{1}{\alpha}} \quad (57)$$

and

$$\gamma_k = 2^{\frac{k}{2}} \left(\sum_{i=1}^{2^k} i^\alpha \right)^{-\frac{1}{\alpha}}. \quad (58)$$

Consequently, according to (46), we have that

$$\lim_{n \rightarrow \infty} \mathbf{R}(\text{HWT}) = \sum_{k=1}^{\infty} 2^{-k} \log \gamma_k^{-1}. \quad (59)$$

However, for the case $\kappa > 0$ and $k < N$,

$$\begin{aligned} \gamma_k^{-1} &\leq \frac{2 \left((2^k - 1) (1 - e^{-2^k \kappa T})^\alpha \right)^{\frac{1}{\alpha}}}{\sqrt{(1 - e^{-2\kappa T}) (1 - e^{-2^{k+1}\kappa T})}} \\ &\leq \frac{2}{\sqrt{1 - e^{-2\kappa T}}} \sqrt{\frac{1 - e^{-2^k \kappa T}}{1 + e^{-2^k \kappa T}}} (2^k - 1)^{\frac{1}{\alpha}} \\ &\leq \frac{2^{1+\frac{k}{\alpha}}}{\sqrt{1 - e^{-2\kappa T}}}. \end{aligned} \quad (60)$$

Thus,

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbf{R}(\text{HWT}) &\leq \sum_{k=1}^{\infty} 2^{-k} \log \frac{2^{1+\frac{k}{\alpha}}}{\sqrt{1 - e^{-2\kappa T}}} \\ &= \left(\frac{2}{\alpha} + \frac{1}{2} \log \frac{1}{1 - e^{-2\kappa T}} \right) \log 2. \end{aligned} \quad (61)$$

For the case $\kappa = 0$ and $k < N$,

$$\begin{aligned} \gamma_k^{-1} &\leq 2^{-\frac{k}{2}} \left((2^k - 1) (2^{k-1})^\alpha \right)^{\frac{1}{\alpha}} \\ &\leq 2^{\frac{k}{2} + \frac{k}{\alpha} - 1}. \end{aligned} \quad (62)$$

Thus,

$$\lim_{n \rightarrow \infty} \mathbf{R}(\text{HWT}) \leq \sum_{k=1}^{\infty} 2^{-k} \log 2^{\frac{k}{2} + \frac{k}{\alpha} - 1} = \frac{2}{\alpha} \log 2. \quad (63)$$

Therefore, the proof is complete.

B. Proof of Part 2 (Equation (42))

Proof: We have that

$$\text{MSE}(\mathbf{H}) = \frac{1}{N} \sum_{n=1}^N \nu(\bar{h}_n^{-1}) = \int_{\mathbb{R}} \nu(\gamma^{-1}) p(\gamma) d\gamma \quad (64)$$

in which $\nu(\gamma^{-1})$ is the MMSE of the estimating w from s in the scalar problem

$$s = \gamma^{-1} w + z, \quad (65)$$

where w is a stable random variable with characteristic function $\hat{p}_w(\omega) = \exp(-|\omega|^\alpha)$ and z is a Gaussian random variable with variance σ^2 . We know that $\nu(\cdot)$ is a monotone continuous function that vanishes at zero and tends to σ^2 asymptotically. Also, $p(\cdot)$ is the empirical distribution of the reciprocals of \bar{h}_i in (30). The proof is then essentially the same as the one of Theorem 1 but simpler since the function $\nu(\cdot)$ is bounded.

For \mathbf{H} equal to Fourier transform, the limiting $p(\gamma)$ was given in (53). Thus, for $\alpha < 2$, as n tends to infinity, $\text{MSE}(\mathbf{H})$ tends to σ^2 . This completes the proof of the right-hand side.

For the case that \mathbf{H} is the operator-like wavelet transform, the limit is $p(\gamma) = \sum_{k=1}^{\infty} p_k \delta(\gamma - \gamma_k)$ where $p_k = 2^{-k}$ and γ_k were given in (55) – (58). Thus, we have that

$$\text{MSE}(\text{OpWT}) = \sum_{k=1}^{\infty} 2^{-k} \nu(\gamma_k^{-1}) \leq \frac{1}{2} \nu(\gamma_1^{-1}) + \frac{\sigma^2}{2}. \quad (66)$$

But, obviously, $\gamma_1^{-1} < \infty$; hence, $\nu(\gamma_1^{-1}) < \sigma^2$, which completes the proof.

REFERENCES

- [1] J. Pearl, “On coding and filtering stationary signals by discrete Fourier transforms,” *IEEE Transactions on Information Theory*, vol. 19, no. 2, pp. 229–232, March 1973.
- [2] M. Unser, “On the approximation of the discrete Karhunen-Loève transform for stationary processes,” *Signal Processing*, vol. 7, no. 3, pp. 231–249, December 1984.
- [3] D. S. Taubman and M. W. Marcellin, *JPEG2000: Image Compression Fundamentals, Standards and Practice*. Kluwer Academic Publishers, 2001.
- [4] D. L. Donoho, “Nonlinear wavelet methods for recovery of signals, densities, and spectra from indirect and noisy data,” in *Proceedings of symposia in Applied Mathematics*, vol. 47. American Mathematical Society, 1993, pp. 173–205.
- [5] C. Taswell, “The what, how, and why of wavelet shrinkage denoising,” *Computing in Science Engineering*, vol. 2, no. 3, pp. 12–19, May/June 2000.
- [6] B. A. Olshausen and D. J. Field, “Emergence of simple-cell receptive field properties by learning a sparse code for natural images,” *Nature*, vol. 381, pp. 607–609, 13 June 1996.
- [7] J. F. Cardoso and D. L. Donoho, “Some experiments on independent component analysis of non-Gaussian processes,” in *Proceedings of the IEEE Signal Processing Workshop on Higher-Order Statistics*, Caesarea, 14–16 June 1999, pp. 74–77.
- [8] Y. Meyer, *Wavelets and Applications*, Masson, France, 1992.
- [9] P. Flandrin, “On the spectrum of fractional brownian motions,” *IEEE Transactions on Information Theory*, vol. 35, no. 1, pp. 197–199, January 1989.
- [10] R. A. Devore, “Nonlinear approximation,” *Acta Numerica*, vol. 7, pp. 51–150, January 1998.
- [11] M. Unser, P. D. Tafti, and Q. Sun, “A unified formulation of Gaussian vs. sparse stochastic processes—part I: Continuous-domain theory,” *arXiv:1108.6150v1*.
- [12] M. Unser, P. D. Tafti, A. Amini, and H. Kirshner, “A unified formulation of Gaussian vs. sparse stochastic processes—part II: Discrete-domain theory,” *arXiv:1108.6152v1*.
- [13] A. Amini, M. Unser, and F. Marvasti, “Compressibility of deterministic and random infinite sequences,” *IEEE Transactions on Signal Processing*, vol. 59, no. 11, pp. 5193–5201, November 2011.
- [14] I. Khalidov and M. Unser, “From differential equations to the construction of new wavelet-like bases,” *IEEE Transactions on Signal Processing*, vol. 54, no. 4, pp. 1256–1267, April 2006.
- [15] J. V. Stone, *Independent Component Analysis*. The MIT Press, September 2004.
- [16] C. M. Stein, “Estimation of the mean of a multivariate normal distribution,” *The Annals of Statistics*, vol. 9, pp. 1135–1151, November 1981.
- [17] I. Gelfand and N. Y. Vilenkin, *Generalized Functions*. New York, USA: Academic Press, 1964, vol. 4.
- [18] M. Sahmoudi, K. Abed-Meraim, and M. Benidir, “Blind separation of impulsive alpha-stable sources using minimum dispersion criterion,” *IEEE Signal Processing Letters*, vol. 12, no. 4, pp. 281–284, April 2005.
- [19] A. R. Soltani and R. Moeanaddin, “On dispersion of stable random vectors and its application in the prediction of multivariate stable processes,” *Journal of Applied Probability*, vol. 31, no. 3, pp. 691–699, September 1994.
- [20] W. D. Ray and R. M. Driver, “Further decomposition of the Karhunen-Loève series representation of stationary random process,” *IEEE Transactions on Information Theory*, vol. IT-16, no. 6, pp. 663–668, November 1970.
- [21] U. S. Kamilov, P. Pad, A. Amini, and M. Unser, “MMSE estimation of sparse Lévy processes,” *IEEE Transactions on Signal Processing*, vol. 61, no. 1, pp. 137–147, January 2013.