December 1996

LIDS-P-2376

Research Supported By:

Army Research Office (DAAL-03-92-G-115)
Air Force Office of Scientific Research
(F49620-95-1-0083 and BU GC12391NGD)

ON THE DISTRIBUTIONS OF OPTIMIZED MULTISCALE REPRESENTATIONS

Hamid Krim

On the Distributions of Optimized Multiscale Representations*

Hamid Krim
Stochastic Systems Group, LIDS
Room 35-431, MIT, Cambridge, MA 02139,
tel: 617-253-3370, fax: 617-258-8553
e-mail: ahk@mit.edu

Abstract

Adapted wavelet analysis of signals is achieved by optimizing a selected criterion. We recently introduced a majorization framework for constructing selection functionals, which can be as well suited to compression as *entropy* or others. We show how these functionals operate on the basis selection and their effect on the statistics of the resulting representation.

1 Introduction

Multiscale analysis has permeated most applied science and engineering applications largely on account of its simple and efficient implementation. In addition it provides a highly flexible adaptive framework using Wavelet Packet (WP) and local trigonometric dictionaries [1, 2, 3]. The remarkable impact it has had on signal processing applications is reflected by the vibrant interest from the basic/applied research communities in its apparently naturally suited framework for signal compression [4]. Adapted wavelet representations have further raised enthusiasm in providing a perhaps *optimal* and yet efficiently achievable transform domain for compression (merely via a selection criterion).

Various criteria for optimizing adapted representations, have been proposed in the literature [5, 6, 7], the first and perhaps the best known being the *entropy* criterion. This was proposed on the basis that the most preferable representation for a given signal is that which is the most parsimonious, i.e. that which compresses the energy into the fewest number of basis function coefficients. We have recently recast the search for an optimized wavelet basis into a *majorization* theoretic framework and briefly described later [8]. This framework not only makes the construction of new criteria simple, but raises questions about their physical interpretation and their impact on the statistics of the resulting representation as well. While the first question was addressed and answered quite satisfactorily [8], the second, to the best of our knowledge remains open. To address this issue, we view the basis search as an optimization of a functional over a family of probability density functions which result from the various possible representations of the WP dictionary. We show that for an appropriately selected optimization (or cost) criterion, the resulting Probability Density Function (PDF) of the coefficients for the optimized representation will decrease rapidly (at least as fast as linearly).

In the next section, we present some relevant background as well as the problem formulation. In Section 3 we present the analysis of the optimization leading to an adapted wavelet basis of a given signal $y(t)$. In Section 4 we provide some illustrative examples.

2 Background and Formulation

2.1 Best Basis Representations

The determination of the "best representation" or Best Basis (BB) of a signal in a wavelet packet or Malvar's wavelet basis generally relies on the minimization of an additive criterion. The entropy is usually retained as a cost function but, as will be shown later, other criteria may be constructed to introduce an alternative viewpoint. To obtain an efficient search of the BB, the dictionary $\mathcal{D}$ of possible bases is structured according to a binary tree. Each node (j,m) (with $j \in \{0,\ldots,J\}$ and $m \in \{0,\ldots,2^j-1\}$) of the tree then corresponds to a given orthonormal basis $\mathcal{B}_{j,m}$ of a vector subspace of $\ell^2(\{1,\ldots,K\})$. An orthonormal basis of $\ell^2(\{1,\ldots,K\})$ is then $\mathcal{B}_{\mathcal{P}} = \cup_{(j,m)/I_{j,m}\in\mathcal{P}}\mathcal{B}_{j,m}$ where $\mathcal{P}$ is a partition of $[0,1[$ in intervals $I_{j,m} = [2^{-j}m, 2^{-j}(m+1)[$. By taking advantage of the property

$$\text{Span}\{\mathcal{B}_{j,m}\} = \text{Span}\{\mathcal{B}_{j+1,2m}\} \overset{\perp}{\oplus} \text{Span}\{\mathcal{B}_{j+1,2m+1}\},$$

a fast bottom-up tree search algorithm was developed in [1] to optimize the partition $\mathcal{P}$. The coefficients

*The work of the author was supported in part by the Army Research Office (DAAL-03-92-G-115), Air Force Office of Scientific Research (F49620-95-1-0083 and BU GC12391NGD).

of an observed signal $y(t)$ are henceforth denoted by $\{x_i\}$.

2.2 Majorization Theoretic Approach

We have recently recast this BB search problem [8] into the context of *majorization* theory developed in mathematical analysis in the 1930's [9]. Evaluating two candidate representations for an observed process $y(t)$ in a dictionary of bases, entails a comparison of two corresponding quantitative measures. These can in theory be defined to reflect any desired specific property of the process [8], and thereby afford us to generalize the class of possible criteria mentioned in the previous section. This was in fact inspired by an effective mechanism first proposed in econometry [10] and later formalized and further generalized in [9].

To compare, say, two vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma} \in \mathbb{R}_+^n$ (i.e. positive real), we could evaluate the spreads of their components to establish a property of majorization of one vector by the other. Let these vectors be rank ordered in a decreasing manner and subsequently denoted by $\{\alpha_{[i]}\}$ (i.e. $\alpha_{[i]} \geq \alpha_{[i+1]}, i = 1, \cdots n$), we then have,

Definition 1. *For* $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma} \in \mathbb{R}_+^n$, *we say that* $\boldsymbol{\alpha} \prec \boldsymbol{\gamma}$, *or* $\boldsymbol{\alpha}$ *is majorized by* $\boldsymbol{\gamma}$ *if*

$$\begin{cases} \sum_{i=1}^{k} \alpha_{[i]} \leq \sum_{i=1}^{k} \gamma_{[i]}, & k = 1, \cdots, n-1 \\ \sum_{i=1}^{n} \alpha_{[i]} = \sum_{i=1}^{n} \gamma_{[i]}. \end{cases}$$

Note that in the case of an entropy-based BB search, the comparison carried out on the wavelet packet coefficients is similar to the majorization procedure described above. This theory has also spawned a variety of questions in regards to the choice of functionals (or criteria) acting upon these vectors and preserving the majorization. Many properties have been established [9] and one which is of central importance herein is that *any optimization functional* $g(\cdot)$ *we select, must be order preserving, i.e.*

$$\text{If } \boldsymbol{\alpha} \prec \boldsymbol{\gamma} \Rightarrow g(\boldsymbol{\alpha}) \leq g(\boldsymbol{\gamma}).$$

This not only brings insight into the problem, but provides the impetus as well to further study the various convex/concave criteria typically invoked in the optimization.

2.3 Formulation

The criteria used in majorization are based on using isotonic or order-preserving functionals $\mathcal{I}(\cdot)$ which can be shown to satisfy Schur convexity/concavity [1] [9]. In its general form, a BB search aims at then minimizing a functional $\mathcal{J}(f(x), x)$, where $f(x)$ represents the common PDF of the wavelet coefficients, which are also subject to constraints. Formally, we may state the problem as

$$\min_{f(x)} \mathcal{J}(x, f(x)) = \min_{f(x)} \int [\mathcal{I}(f(x)) + \lambda \mathcal{C}(f(x), x)] dx \quad (1)$$

where $\mathcal{C}(\cdot)$ specifies some implicit or explicit constraints. Our focus in this paper is, for a given $\mathcal{I}(\cdot)$, to determine the statistical properties of the coefficients in the *optimized* or more precisely the class of " $f(x)$" which leads to the minimization of a given functional.

3 Statistical Analysis

The majorization approach may be viewed as a unifying framework which provides the necessary theoretical justifications for all previously proposed BB criteria (e.g. the *entropy* criterion), and which equips one with the theoretical underpinnings and insight for other extensions. This indeed paves the way for a plethora of other possible search principles aimed at reflecting characteristics other than parsimony for instance[8].

Recall, however, that the parsimony of representation, lies at the heart of the originally proposed criteria [1], and various heuristic/justifying statements about the distributions of wavelet coefficients were presented.

Proposition 1. *Any order preserving continuous functional* $\mathcal{I}(\cdot)$ *satisfying the above (convexity/concavity) properties, and which when optimized leads to a BB selection of a signal* $y(t)$, *results in an overall density function* $f(x)$ *of the coefficients which is at least* $o(x^\alpha)$ *as* $x \to \infty$ *(i.e. decreases at least at a linear rate).*

Proof: Concentrating on a general and to be specified functional $\mathcal{I}(\cdot)$ in Eq. 1, with the constraints on $f(x)$ to be a valid PDF and on the coefficients to have finite moment, we may (e.g.) write the following,

$$\mathcal{J}_{min}(x, f(x)) = \min_{f(x)} \left\{ \int_{-\infty}^{\infty} \mathcal{I}(f(x)) + \right. \quad (2)$$
$$\left. \lambda_1 \left(\int_{-\infty}^{\infty} x^\alpha f(x) dx - \mu \right) + \lambda_2 \left(\int_{-\infty}^{\infty} f(x) dx - 1 \right) \right\}.$$

Using standard variational techniques of optimization [11] to find the stationary point of $\mathcal{J}(\cdot, \cdot)$ the following results,

$$\delta \mathcal{J} = \mathcal{I}_{f(x)}(f(x)) + \lambda_1 x^\alpha + \lambda_2 = 0, \quad (3)$$

where $\mathcal{I}_{f(x)}$ denotes a differentiation with respect to $f(\cdot)$. The functional $\mathcal{I}(\cdot)$ being concave/convexe, leads to a decreasing/increasing $\mathcal{I}_{f(x)}(\cdot)$. Using the following standard theorem on monotone increasing/decreasing functions,

[1] Schur convexity/concavity is tied to convexity/concavity and isotonicity (or order-preservation).

Theorem 1. *Let $G : D \rightarrow \mathbb{R}$ be strictly increasing (or decreasing) on D. Then there exists a unique inverse function G^{-1} which is strictly monotone increasing (or decreasing) on f(D),*

we conclude that we have an increasing/decreasing inverse function everywhere, except possibly at a finite set of points, or

$$f(x) = \mathcal{I}'^{-1}(-\lambda_1 - \lambda_2 x^{\alpha}),$$

with the λ_i's ensuring the properties of $f(x)$. ■

3.1 Criteria

3.1.1 Entropy:

The *entropy* criterion first proposed in [1] is $(\mathcal{I}(x) = -x \log x)$.

Property 1. *$\mathcal{I}(f(x)) = f(x) \log f(x)$ is a convex functional of $f(x)$.*

Proof: This can easily be seen by taking the second derivative w.r.t. $f(x)$ and noting that $f(x)$ is nonegative. ■

Using the approach described above, one can simply derive the maximizing density as

$$f(x) = \exp\{\lambda_1 + \lambda_2 \mid x \mid +1\}, \tag{4}$$

which when using Definition 1 for the BB search, also leads to the minimization of the entropy of the resulting representation.

3.1.2 Lorenz Criterion:

In studying the spread of components of a vector, one might consider looking at the center of mass and at its variation as a function of x. Let us define

$$F(x) = \int_{-\infty}^{x} f(u)du \tag{5}$$

$$\Phi(x) = \frac{1}{\mu}\int_{-\infty}^{x} uf(u)du, \tag{6}$$

where we recognize in $\Phi(x)$ the "local center" of gravity (or local mean) and in $F(x)$ the cumulative population or the probability at a point x. The graph of the former versus the latter coincides precisely with the Lorenz curve [10] shown in Fig. 1 which also forms the basis of Gini's concentration criterion[9]. The lower curve "B" is more concentrated than curve "A" which clearly represents a more uniform distribution of the coefficients. In this case, the goal is to maximize the distance (or the area enclosed) between

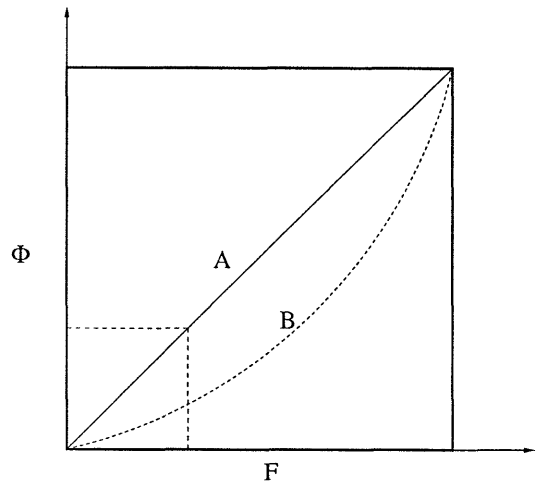

Figure 1: Continuous Lorenz Curve

the "uniform-indicating" curve and that indicating more concentration, or,

$$\mathcal{I}_b(f(x)) = \int F(x)d\Phi(x) - \int \Phi(x)dF(x), \tag{7}$$

leading once again to the following optimization problem,

$$\min_{f(x)} \mathcal{J}(x, f(x)) = \min_{f(x)} \{\mathcal{I}_b(f(x)) + \lambda_1\left(\int_{-\infty}^{\infty} f(x)dx - 1\right) + \lambda_2\left(\int_{-\infty}^{\infty} xf(x)dx - \mu\right)\}$$

Using techniques from calculus of variations[11], this criterion may be "extremized" (maximize $\mathcal{I}_b(\cdot)$) to solve for the class of $f(x)$, which can be solved after much algebra. Instead we can use the method of the Legendre transform which is precisely constructed using the distance between "A" and "B"[11],

$$L(p, F) = pF - \Phi(F) \tag{8}$$

which will achieve an extremum for $\partial L/\partial F = 0$ or $d\Phi/dF = p$ which can be rewritten as,

$$\frac{d\Phi}{dx}/\frac{dF}{dx} = x/\mu$$

or for $p = 1$,

$$x = \mu = \int_{-\infty}^{\infty} xf(x)dx, \tag{9}$$

leading to the fact that $f(x)$ must necessarily be decreasing much more rapidly than x. ■

Our analysis results in a rigorous solution stating that the *class* of distributions which lead to the extrema of the criteria, is of *polynomial/exponential* decay. This is a significant result in its own right, since, to the best of our knowledge, it is the first rigorous proof whose result, not surprisingly corroborates with the appealing and heuristic notion of energy concentration, and which has been the basis of all previously proposed algorithms.

4 Applications

The appeal of this result is twofold:

1. It provides a strong theoretical argument/justification for previously proposed BB search criteria
2. It provides insight for further improving BB searches, particularly in noisy environments

In particular, these results can be turned around to specify one of the properties of an exponential distribution which is known to be "optimal", as the criterion of optimization. More specifically, we may use the "shape factor" of the density $f(x)$ which can be viewed as a robust global measure, less prone to variability in the presence of noise. The shape factor can be evaluated in the Maximum Likelihood sense for the WP tableau for instance, and used to efficiently prune the binary tree to result in a BB. In contrast to recently proposed algorithms, we avoid to explicitly use the (perhaps) strong a priori assumption of normality of the noise, and our criterion here is obtained by proceeding "in reverse" (i.e. in light of the distribution properties of the "optimal" representation, we optimize the intermediate distributions in order to achieve it). Similarly, the second criterion analyzed above is used as a measure of the distribution of the coefficients on the tree and optimized to achieve a BB.

In Fig. 2, we show for illustration the histograms of a typical signal (ramp signal) in noise and that of resulting BB coefficients.

Acknowledgement: Thanks are due to Dr. J-C Pesquet for comments.

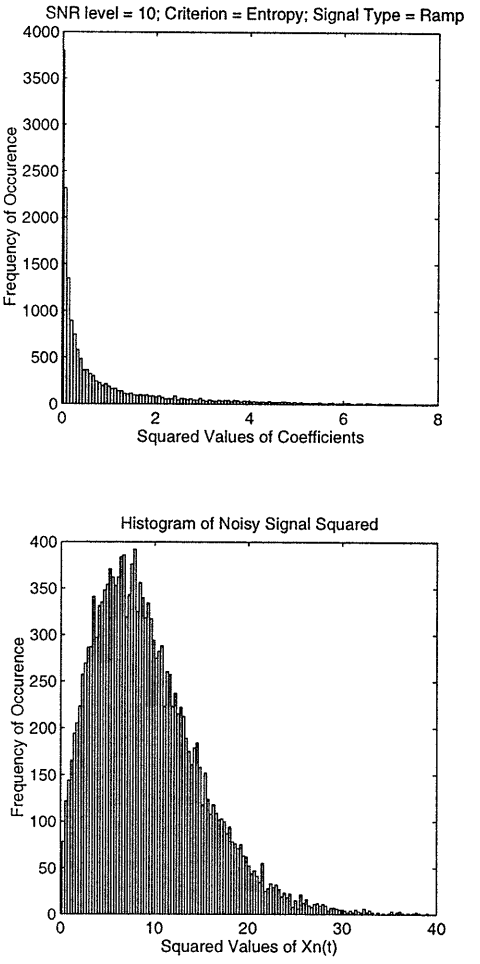

Figure 2: Histograms of Signal + Noise and of its MS representation .

References

[1] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Trans. Inform. Theory*, vol. IT-38, pp. 713–718, Mar. 1992.

[2] Y. Meyer, *Wavelets and Applications.* Philadelphia: SIAM, first ed., 1992.

[3] F. Meyer and R. Coifman, "Brushlets: a tool for directional image analysis and image compression," *preprint.*

[4] J. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Trans on Sig. Proc.*, vol. 41, no. 12, pp. 3445–3462, 1993.

[5] D. Donoho and I. Johnstone, "Ideal denoising in an orthogonal basis chosen from a library of bases," Oct. 1994. To appear in C. R. Acad. Sci. Paris, 1994.

[6] H. Krim, S. Mallat, D. Donoho, and A. Willsky, "Best basis algorithm for signal enhancement," in *ICASSP'95*, (Detroit, MI), IEEE, May 1995.

[7] H. Krim and J.-C. Pesquet, *On the Statistics of Best Bases Criteria*, vol. Wavelets in Statistics of *Lecture Notes in Statistics.* Springer-Verlag, July 1995.

[8] H. Krim and D. Brooks, "Feature-based best basis segmentation of ecg signals," in *IEEE Symposium on Time-Freq./Time Scale Analysis*, (Paris, France), June 1996.

[9] G. Hardy, J. Littlewood, and G. Pòlya, *Inequalities.* Cambridge Press, second edition ed., 1934.

[10] M. O. Lorentz, "Methods of measuring concentration of wealth," *Jour. Amer. Statist. Assoc.*, vol. 9, pp. 209–219, 1905.

[11] F. B. Hildebrand, *Methods of Applied Mathematics.* Prentice-Hall, sec. edition ed., 1965.