March 1989 LIDS-P-1857

RECURSIVE AND ITERATIVE ESTIMATION ALGORITHMS FOR MULTI-RESOLUTION STOCHASTIC PROCESSES

K. C. Chou[1]
A. S. Willsky[1]
A. Benveniste[2]
M. Basseville[2]

Abstract

A current topic of great interest is the multi-resolution analysis of signals and the development of multi-scale or multigrid algorithms. In this paper we describe part of a research effort aimed at developing a corresponding theory for stochastic processes described at multiple scales and for their efficient estimation or reconstruction given partial and/or noisy measurements which may also be at several scales. The theories of multi-scale signal representations and wavelet transforms lead naturally to models of signals(in one or several dimensions) on trees and lattices. In this paper we focus on one particular class of processes defined on dyadic trees. The central results of the paper are three algorithms for optimal estimation/reconstruction for such processes: one reminiscent of the Laplacian pyramid and the efficient use of Haar transforms, a second that is iterative in nature and can be viewed as a multigrid relaxation algorithm, and a third that represents an extension of the Rauch-Tung-Striebel algorithm to processes on dyadic trees and involves a new discrete Riccati equation, which in this case has *three* steps: predict, *merge*, and measurement update. Related work and extensions are also briefly discussed.

---

[1]Dept. of Electrical Eng. and Computer Science and Lab. for Information and Decision Systems, MIT, Cambridge, MA 02139. The work of these authors was supported in part by the Air Force Office of Scientific Research under Grant AFOSR-88-0032, in part by the National Science Foundation under Grant ECS-8700903, and in part by the US Army Research Office under Contract DAAL03-86-K-0171. In addition some of this research was performed while these authors were visitors at Institut de Recherche en Informatique et Systemes Aleatoires(IRISA), Rennes, France during which time A.S.W. received partial support from Institut National de Recherche en Informatique et en Automatique(INRIA).

[2]IRISA, Campus Universitaire de Beaulieu, 35042 Rennes Cedex, France. A.B. is also with INRIA, and M.B. is also with Centre National de la Recherche Scientifique(CNRS). The research of these authors was also supported in part by Grant CNRS GO134.

Contents

List of Figures

1 Introduction

The investigation of multi-scale representations of signals and the development of multi-scale algorithms has been and remains a topic of much interest in many applications for a variety of reasons. In some cases the motivation has directly been the fact that the phenomenon of interest exhibits patterns of importance at multiple scales. One well-known example is the use of fractal models for images[4][25] but many others exist, such as in the modeling of layering structures in the earth using for example self-similar processes as the basis for geophysical signal processing algorithms[9]. A second motivation has been primarily computational: many problems, especially those in several spatial dimensions, are of enormous computational complexity, and thus extremely efficient, possibly highly parallel and iterative algorithms are essential. Multigrid methods for solving partial differential equations[7,20,28] or for performing Monte Carlo experiments[18] are a good example.

A third motivation stems from so-called "sensor fusion" problems in which one is interested in combining together measurements with very different spatial resolutions. Geophysical problems often have this character as do problems of combining multi-spectral measurements(IF, radar, millimeter wave, ...) or of combining gravity-related measurements from different sources(inertial system vertical deflections, gravimeters, gradiometers, ...). Finally, renormalization group ideas, developed originally to assist in the difficult study of near-critical phenomena in statistical mechanical systems, also now find application in methods for improving convergence in large-scale simulated annealing algorithms for Markov random field estimation[17].

One of the more recent areas of investigation in multi-scale analysis has been the development of a theory of multi-scale representations of signals[22,23] and the closely related topic of wavelet transforms[27,14]. These methods have drawn considerable attention in several disciplines including signal processing because they appear to be a natural way to perform a time-scale decomposition of signals and because examples that have been given of such transforms seem to indicate that it should be possible to develop efficient optimal processing algorithms based on these representations. The

development of such optimal algorithms - e.g. for the reconstruction of noise-degraded signals or for the detection and localization of transient signals of different durations - requires, of course, the development of a corresponding theory of stochastic processes and their estimation. The research presented in this and several other papers and reports[5,6] has the development of this theory as its objective.

In the next section we introduce multi-scale representations of signals and wavelet transforms and from these we motivate the investigation of stochastic processes on trees and lattices, the former of which has been the focus of all of our work to date. In that section we also introduce the processes studied in this paper. In Section III we formulate a multi-scale estimation problem and present three algorithms for its solution. The first, which is a fine-to-coarse algorithm is reminiscent of the so-called Laplacian pyramid[10] for image coding and can be implemented efficiently via the Haar transform. The second is an iterative, relaxation algorithm resembling multigrid methods. The third is a fine-to-coarse-to-fine algorithm that represents a generalization of the Rauch-Tung-Striebel smoothing algorithm. The Riccati equation in this case has a fine-to-coarse *prediction* step, a *merging* step in which predicted estimates from neighboring regions of fine data are merged, and a *measurement update* step. Finally, in Section IV we discuss extensions of this work and briefly describe some related research.

2 Multiscale Representations and Stochastic Processes on Trees

2.1 Multiscale, Wavelets, and Trees

As developed in [23,24], the multi-scale representation of a continuous-time signal[1] $x(t)$ consists of a sequence of approximations of that signal on finer and finer subspaces of functions. The entire representation is completely specified by a single function $\phi(t)$, where the approximation of $x(t)$ at the mth scale is given by

$$x_m(t) = \sum_{n=-\infty}^{+\infty} x(m,n)\phi(2^m t - n) \tag{2.1}$$

Thus as $m \to \infty$ the approximation consists of a sum of many highly compressed, weighted, and shifted versions of ϕ. This function is far from arbitrary in its choice. In particular $\phi(t)$ must be orthogonal[2] to its integer translates $\phi(t-n)$, $n = \overset{+}{-} 1, \overset{+}{-} 2, ...$, and also, in order that the $(m+1)$st approximation is indeed a *refinement* of the mth, we require that $\phi(t)$ be exactly representable at the next scale:

$$\phi(t) = \sum_n h(n)\phi(2t - n) \tag{2.2}$$

As developed in [14], the sequence $h(n)$ must satisfy several conditions for the desired properties of $\phi(t)$ to hold and for $x_m(t)$ to converge to $x(t)$ as $m \to \infty$. One of these is that $h(n)$ must be the impulse response of a so-called *quadrature mirror filter*[14,29]. The simplest example of such a ϕ, h pair is the Haar approximation in which

$$\phi(t) = \begin{cases} 1 & 0 \le t < 1 \\ 0 & otherwise \end{cases} \tag{2.3}$$

and

$$h(n) = \begin{cases} 1 & n = 0,1 \\ 0 & otherwise \end{cases} \tag{2.4}$$

[1] This entire theory, and ours as well, extends easily to signals in several spatial dimensions. For simplicity of discussion and notation we focus here on the 1D case.

[2] Actually, it is possible to relax this by requiring only the so-called condition of *quasi-orthogonality*[16].

As shown in [14] there exists a family of FIR $h(n)$'s and corresponding compactly supported $\phi(t)$'s, where the degree of smoothness of $\phi(t)$ increases(albeit slowly) with the length of $h(n)$.

The representation just described is closely connected to the *wavelet transform*, which is based on a single function $\psi(t)$ that has the property that the full set of its scaled translates $\left\{2^{m/2}\psi(2^m t - n)\right\}$ form a complete orthonormal basis for L^2. In [14] it is shown that if ϕ and ψ are related via an equation of the form

$$\psi(t) = \sum_n g(n)\phi(2t - n) \tag{2.5}$$

where $g(n)$ and $h(n)$ form a *conjugate mirror filter* pair[29], then

$$x_{m+1}(t) = x_m(t) + \sum_n d(m,n)\psi(2^m t - n) \tag{2.6}$$

and indeed $x_m(t)$ is simply the partial orthonormal expansion of $x(t)$, up to scale m, with respect to the basis defined by ψ. For example if ϕ and h are as in eq.(2.3), eq.(2.4), then

$$\psi(t) = \begin{cases} 1 & 0 \leq t < 1/2 \\ -1 & 1/2 \leq t < 1 \\ 0 & otherwise \end{cases} \tag{2.7}$$

$$g(n) = \begin{cases} 1 & n = 0 \\ -1 & n = 1 \\ 0 & otherwise \end{cases} \tag{2.8}$$

and $\left\{2^{m/2}\psi(2^m t - n)\right\}$ is the *Haar basis.*

Using eq.(2.1), eq.(2.2), eq.(2.5), and eq.(2.6) we see that we have a *dynamical* relationship between the coefficients $x(m,n)$ at one scale and those at the next. Indeed this relationship defines a lattice on the points (m,n), where $(m+1,k)$ is connected to (m,n) if $x(m,n)$ influences $x(m+1,k)$. For example the Haar representation naturally defines a dyadic tree structure on the points (m,n) in which each point has two equally-weighted(i.e. $h(0) = h(1)$) descendents corresponding to the two subdivisions of the support interval of $\phi(2^m t - n)$, namely those of $\phi(2^{(m+1)}t - 2n)$ and $\phi(2^{(m+1)}t - 2n - 1)$.

The preceding development provides the motivation for the study of stochastic processes $x(m,n)$ defined on the types of lattices just described. In our work to date we have focused attention on the case of the dyadic tree. Let us make several comments about this case. First, as illustrated in Figure 1, with this and any of the other lattices, the scale index m is time-like. For example it defines a natural direction of recursion for our representation: from coarse-to-fine in the synthesis of a signal and from fine to coarse in the analysis(e.g. in the Haar case $x(m,n)$ is directly obtainable from $x(m+1,2n)$, $x(m+1,2n+1)$. In the case of our tree, with increasing m - i.e. the direction of synthesis - denoting the forward direction, we then can define a unique backward shift γ^{-1} and two forward shifts α and β(see Figure 1). Also, for notational convenience we denote each node of the tree by a single abstract index t and let T denote the set of all nodes. Thus if $t=(m,n)$ then $\alpha t=(m+1,2n)$, $\beta t=(m+1,2n+1)$, and $\gamma^{-1}t=(m-1,[\frac{n}{2}])$ where $[x]$ =integer part of x. Also we use the notation $m(t)$ to denote the scale(i.e. the m-component of t). Finally, it is worth noting that while we have described multi-scale representations for continuous-time signals on $(-\infty,\infty)$, they can also be used for signals on compact intervals or in discrete-time. For example a signal defined for $t=0,1,...,2^{M-1}$ can be represented by M scales, each of which represents in essence an averaged, decimated version of the finer scale immediately below it. In this case the tree of Figure 1 has a bottom level, representing the samples of the signal itself, and a single root node, denoted by 0, at the top. Such a root node also exists in the representation of continuous-time signals defined on a compact interval.

2.2 Dynamic Stochastic Models on Trees

As in the synthesis description of multi-scale representations, the stochastic models we consider are naturally described as evolving from coarse-to-fine scales. Specifically, we consider the following class of state-space models on trees:

$$x(t)=A(m(t))x(\gamma^{-1}t)+B(m(t))w(t) \tag{2.9}$$

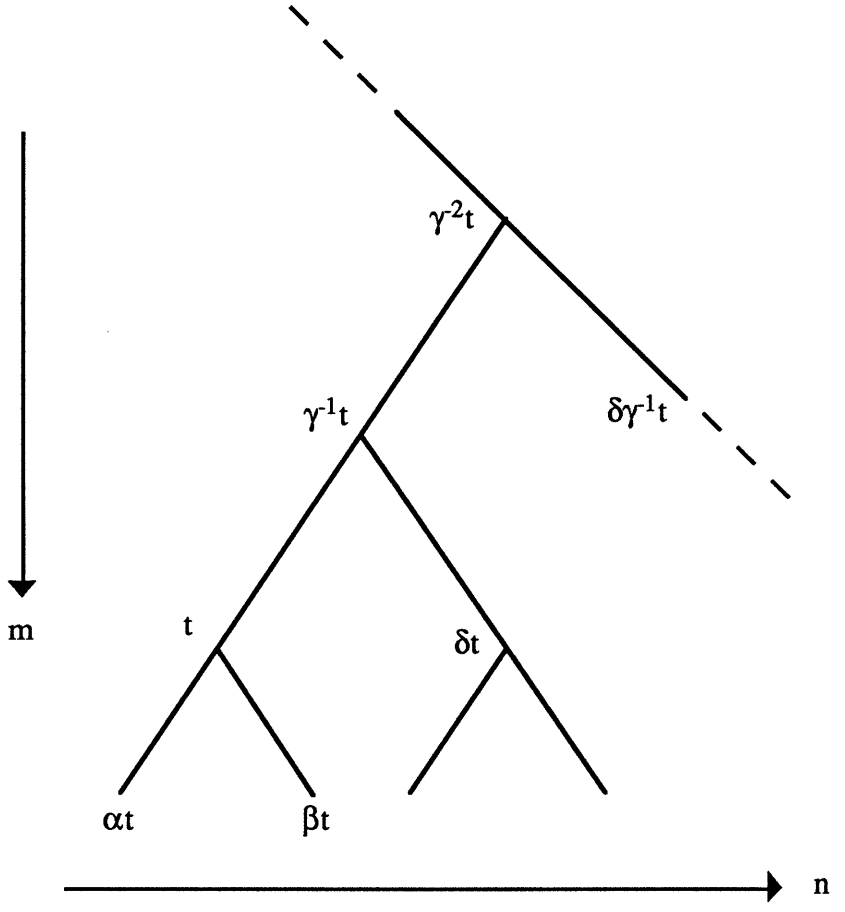

Figure 1: Dyadic Tree Representation

where $\{w(t), t \in T\}$ is a set of independent, zero-mean Gaussian random variables. If we are dealing with a tree with unique root node, 0, we require $w(t)$ to be independent of $x(0)$, the zero-mean initial condition. The covariance of $w(t)$ is I and that of $x(0)$ is $P_x(0)$. If we wish the model eq.(2.9) to define a process over the entire infinite tree, we simply require that $w(t)$ is independent of the "past" of x, i.e. $\{x(\tau)|m(\tau) < m(t)\}$. If $A(m)$ is invertible for all m, this is equivalent to requiring $w(t)$ to be independent of *some* $x(\tau)$ with $\tau \neq t$, $m(\tau) < m(t)$.

Let us make several comments about this model. Note first that the model *does* evolve along the tree, as both $x(\alpha t)$ and $x(\beta t)$ evolve from $x(t)$. Secondly, we note that this process has a Markovian property: given x at scale m, x at scale $m+1$ is independent of x at scales less than or equal to $m-1$. Indeed for this to hold all we need is for w to be independent from scale to scale and not necessarily at each individual node. Also while the analysis we perform is easily extended to the case in which A and B are arbitrary functions of t, we have chosen to focus here on a translation-invariant model: we allow these quantities to depend only on scale. As we will see this leads to significant computational efficiencies and also, when this dependence is chosen appropriately, these models lead to processes possessing self-similar properties from scale to scale.

Note that the second-order statistics of $x(t)$ are easily computed. In particular the covariance $P_x(t) = E[x(t)x^T(t)]$ evolves according to a Lyapunov equation on the tree:

$$P_x(t) = A(m(t))P_x(\gamma^{-1}t)A^T(m(t)) + B(m(t))B^T(m(t)) \tag{2.10}$$

Note in particular that if $P_x(\tau)$ depends only on $m(\tau)$ for $m(\tau) \leq m(t)-1$, then $P_x(t)$ depends only on $m(t)$. We will assume that this is the case and therefore will write $P_x(t) = P_x(m(t))$. Note that this is always true if we are considering the subtree with single root node 0. Also if $A(m)$ is invertible for all m, and if $P_x(t) = P_x(m(t))$ at *some* scale(i.e. at all t for which $m(t)$ equals m for some m), then $P_x(t) = P_x(m(t))$ for *all* t. Let $K_{xx}(t,s) = E[x(t)x^T(s)]$. Let $s \wedge t$ denote the least upper bound of s and t, i.e. the first node that is a predecessor of both t and s. Then

$$K_{xx}(t,s) = \Phi(m(t), m(s \wedge t))P_x(m(s \wedge t))\Phi^T(m(s), m(s \wedge t)) \tag{2.11}$$

where for $m_1 \geq m_2$

$$\Phi(m_1, m_2) = \begin{cases} I & m_1 = m_2 \\ A(m_1)\Phi(m_1 - 1, m_2) & m_1 > m_2 \end{cases} \tag{2.12}$$

Also, let $d(s,t)$ denote the distance from s to t, i.e. the number of branches on the shortest path from s to t. Then $d(s,t) = d(s, s \wedge t) + d(t, s \wedge t) = d(t,s)$ and if $A(m(t)) = A$, then

$$K_{xx}(t,s) = A^{d(t,s\wedge t)} P_x(m(s \wedge t))(A^T)^{d(s,s\wedge t)} \tag{2.13}$$

Furthermore, if A is stable and if $B(m(t)) = B$, let P_x be the solution to the algebraic Lyapunov equation

$$P_x = AP_xA^T + BB^T \tag{2.14}$$

In this case if $P_x(0) = P_x$(if we have a root node), or if we assume that $P_x(\tau) = P_x$ for $m(\tau)$ sufficiently negative[3], then $P_x(t) = P_x$ for all t, and we have the stationary model

$$\begin{aligned} K_{xx}(t,s) &= A^{d(t,s\wedge t)} P_x (A^T)^{d(s,s\wedge t)} \\ &= K_{xx}(d(t, s \wedge t), d(s, s \wedge t)) \end{aligned} \tag{2.15}$$

As a final note, we point out that there is one class of scalar stochastic processes on trees that has been the subject of substantial analysis. In [3] these are referred to as stationary processes but we prefer to use that terminology for the larger class of processes for which $K_{xx}(t,s)$ depends only on $d(t, s \wedge t)$ and $d(s, s \wedge t)$. The class of processes considered in [3] is characterized by the condition that $K_{xx}(t,s)$ depends only on $d(s,t)$ and we refer to these as *isotropic processes.* Note that eq.(2.15) represents an isotropic covariance if $AP_x = P_xA^T$, which shows the connection to the class of reversible stochastic processes[1]. For example in the scalar case

$$K_{xx}(t,s) = \left\{ \frac{B^2}{1 - A^2} \right\} A^{d(s,t)} \tag{2.16}$$

Some of our other research has examined the modeling of isotropic processes on trees, and we briefly describe this in Section IV.

[3]Once again if A is invertible, if $P_x(t) = P_x$ at *any* single node, $P_x(t) = P_x$ at *all* nodes.

3 Optimal Estimation on Trees

In this section we consider the estimation of the stochastic process described by eq.(2.9). For simplicity we assume that there is a root node 0 and M scales on which we have data and wish to focus[4]. The measurements on which our estimates are based are of the form

$$y(t) = C(m(t))x(t) + v(t) \tag{3.1}$$

where $\{v(t), t \in T\}$ is a set of independent zero-mean Gaussian random variables independent of $x(0)$ and $\{w(t), t \in T\}$. The covariance of $v(t)$ is $R(m(t))$. The model eq.(3.1) allows us to consider multiple resolution measurements of our process. The single resolution problem, i.e. when $C(m) = 0$ unless $m = M$(the finest level), is also of interest as it corresponds to the problem of restoring a noise corrupted version of a stochastic process possessing a multi-scale description.

In the following three subsections we describe three different algorithmic structures. The first of these deals with single scale measurements and a batch algorithm from scale-to-scale. The latter two allow multi-scale measurements and yield iterative multigrid and recursive fine-to-coarse-to-fine algorithms, respectively.

3.1 Noisy Interpolation and the Laplacian Pyramid

Consider the model eq.(2.9) with a single scale of measurements:

$$y(n) = Cx(M,n) + v(n) \quad n = 0, 1, ...2^M - 1 \tag{3.2}$$

where without loss of generality we assume that the covariance of $v(n)$ is I. Let us look first at the batch estimation of x at this finest scale. To do this we define the stacked vectors and block matrices

$$\begin{aligned} Y^T &= [y^T(0), ..., y^T(2^M - 1)] \\ X_M^T &= [x^T(M,0), ..., x^T(M, 2^M - 1)] \\ V^T &= [v^T(0), ..., v^T(2^M - 1)] \end{aligned} \tag{3.3}$$

[4]Steady-state properties as $M \to \infty$ will be reported in a subsequent paper.

$$\mathcal{C} = diag(C, ..., C) \tag{3.4}$$

$$\mathcal{P}_M = E[X_M X_M^T] \tag{3.5}$$

The optimal estimate is given by

$$\hat{X}_M = \mathcal{P}_M \mathcal{C}^T [\mathcal{C} \mathcal{P}_M \mathcal{C}^T + I]^{-1} Y \tag{3.6}$$

Furthermore, suppose that we consider estimates at coarser scales, i.e. interpolation up to higher levels in the tree. For example from eq.(2.9) we see that

$$X_{k+1} = \mathcal{A}_{k+1} X_k + \mathcal{B}_{k+1} W_{k+1} \tag{3.7}$$

where X_{k+1}, X_k, and W_{k+1} are defined in a similar manner with X_{k+1} and W_{k+1} of dimension 2^{k+1} and

$$\mathcal{A}_{k+1} = \begin{bmatrix} A(k+1) & 0 & 0 & \cdots & 0 \\ A(k+1) & 0 & 0 & \cdots & 0 \\ 0 & A(k+1) & 0 & \cdots & 0 \\ 0 & A(k+1) & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & & \vdots \\ 0 & 0 & 0 & \cdots & A(k+1) \\ 0 & 0 & 0 & \cdots & A(k+1) \end{bmatrix} \tag{3.8}$$

$$\mathcal{B}_{k+1} = diag(B_{k+1}, ..., B_{k+1}) \tag{3.9}$$

$$\mathcal{P}_{k+1} = E[X_{k+1} X_{k+1}^T] \tag{3.10}$$

An iterated expectation computation applied to eq.(3.7) then yields a recursive procedure for computing $\hat{X}_k$ from fine to coarse scales, starting with $\hat{X}_M$ in eq.(3.6).

$$\hat{X}_k = \mathcal{P}_k \mathcal{A}_{k+1}^T \mathcal{P}_{k+1}^{-1} \hat{X}_{k+1} \tag{3.11}$$

The computation of these coarse-scale estimates is of importance if one wishes to consider efficient coding of data possessing multiple-scale descriptions. Indeed the algorithm, eq.(3.6) and eq.(3.11), possesses structure reminiscent of the Laplacian

pyramid approach[10] to multiscale coding. To see this and to obtain insight into efficient implementations of eq.(3.6) and eq.(3.11) requires a careful examination of the structure of $\mathcal{P}_k$. It turns out that the eigenstructure of $\mathcal{P}_k$ is directly related to the Haar basis, which should come as no surprise considering the correspondance between the dyadic tree and the Haar wavelet basis.

The covariance matrix at any given level, $\mathcal{P}_k$, can be described as follows. For all pairs $k > l$, let $S(k,l)$ denote the block matrix with $2^{k-l-1} \times 2^{k-l-1}$ blocks each of which equals

$$T(k,l) = \Phi(k,l)P_x(l)\Phi^T(k,l) \tag{3.12}$$

Note that these matrices can be computed recursively.

$$T(k+1,l) = A(k+1)T(k,l)A^T(k+1) \tag{3.13}$$

The $2^k \times 2^k$ block matrix $\mathcal{P}_k$ can then be constructed recursively.

$$U(k,k) = P_x(k) \tag{3.14}$$

$$U(k,l) = \begin{bmatrix} U(k,l+1) & S(k,l) \\ S(k,l) & U(k,l+1) \end{bmatrix} \tag{3.15}$$

$$\mathcal{P}_k = U(k,0) \tag{3.16}$$

For example,

$$\mathcal{P}_2 = \left[\begin{array}{cc:cc} P_x(2) & T(2,1) & T(2,0) & T(2,0) \\ T(2,1) & P_x(2) & T(2,0) & T(2,0) \\ \cdots & \cdots & \cdots & \cdots \\ T(2,0) & T(2,0) & P_x(2) & T(2,1) \\ T(2,0) & T(2,0) & T(2,1) & P_x(2) \end{array}\right] \tag{3.17}$$

so that off the main diagonal $\mathcal{P}_k$ has block submatrices of geometrically growing size each of which has its blocks equal to $T(k,l)$.

The $2^k \times 2^{k+1}$ block matrix $\tilde{\mathcal{P}}_{k,k+1} \triangleq \mathcal{P}_k \mathcal{A}_{k+1}^T = E[X_k|X_{k+1}]$, has a very similar structure. For $k > l$, let $\tilde{S}(k,l)$ denote the block matrix with $2^{k-l-1} \times 2^{k-l}$ blocks each of which equals

$$T(k,l)A^T(k+1) = \Phi(k,l)P_x(l)\Phi^T(k+1,l) \tag{3.18}$$

The $2^k \times 2^{k+1}$ block matrix $\tilde{\mathcal{P}}_k$ is then constructed recursively as follows.

$$\tilde{U}(k,k) = [P_x(k)A^T(k+1)\ P_x(k)A^T(k+1)] \tag{3.19}$$

$$\tilde{U}(k,l) = \begin{bmatrix} \tilde{U}(k,l+1) & \tilde{S}(k,l) \\ \tilde{S}(k,l) & \tilde{U}(k,l+1) \end{bmatrix} \tag{3.20}$$

$$\tilde{\mathcal{P}}_{k,k+1} = \tilde{U}(k,0) \tag{3.21}$$

For example,

$$\tilde{\mathcal{P}}_{2,3} = \left[\begin{array}{cccc:cccc} M_1 & M_1 & M_2 & M_2 & M_3 & M_3 & M_3 & M_3 \\ M_2 & M_2 & M_1 & M_1 & M_3 & M_3 & M_3 & M_3 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ M_3 & M_3 & M_3 & M_3 & M_1 & M_1 & M_2 & M_2 \\ M_3 & M_3 & M_3 & M_3 & M_2 & M_2 & M_1 & M_1 \end{array}\right] \tag{3.22}$$

where

$$M_1 = P_x(2),\ M_2 = T(2,1)A^T(3),\ M_3 = T(2,0)A^T(3) \tag{3.23}$$

The structure of these matrices directly provides us with insight into the form of the estimation equations. In particular let us examine eq.(3.11) component by component. Then from the structure of the matrices and the tree, we can deduce that the contribution of $\hat{x}(t)$ with $m(t) = k+1$ to $\hat{x}(s)$ with $m(s) = k$ depends only on $d(s, s \wedge t)$(or equivalently on $d(s,t)$ which since s and t are on adjacent levels equals $2d(s, s \wedge t) + 1$). Furthermore, there are exactly 2 nodes at scale $k+1$(namely αs and βs) satisfying $d(s, s \wedge t) = 0$ and for any other values of $d(s, s \wedge t)$ there are exactly $2^{d(s, s \wedge t)}$ nodes at level $k+1$. Thus, eq.(3.11) has the following form for each node s with $m(s) = k$.

$$\hat{x}(s) = \sum_{i=0}^{k} H(k,i) \sum_{t \in \Theta_x(k,i)} \hat{x}(t) \tag{3.24}$$

where

$$\Theta_x(k,i) = \{t' | m(t') = k+1, d(s, s \wedge t') = i\} \tag{3.25}$$

This computation from level to level, as we successively decimate our estimated signal and in which processing from scale to scale involves averaging of values bears some

resemblance to the Laplacian pyramid, although in this case the weighting function $H(k,i)$ is of full extent and in general varies from scale to scale. Note that if $A(m) = A$, $B(m) = B$ and $P_x(m) = P$, $H(k,i) = H(i)$.

Eq.(3.24) provides one efficient algorithm for the recursion eq.(3.11). A second also comes from the exploitation of the structure of the matrices $\mathcal{P}_k$ and $\tilde{\mathcal{P}}_{k,k+1}$, and the fact that the *discrete Haar transform* block diagonalizes both of them. For simplicity let us first describe this for the case in which x and y are scalar processes.

Definition 3.1.1 *The discrete Haar basis is an orthonormal basis for $\mathcal{R}^N$ where $N = 2^k$. The matrix V_k whose columns form this basis consists of vectors representing "dilated, translated, and scaled" versions of the vector $[1,-1]^T$. For example for $k = 3$,*

$$V_3 = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} & 0 & 0 & -\frac{1}{2} & 0 & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \\ 0 & -\frac{1}{\sqrt{2}} & 0 & 0 & -\frac{1}{2} & 0 & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \\ 0 & 0 & \frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{2} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \\ 0 & 0 & -\frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{2} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \\ 0 & 0 & 0 & \frac{1}{\sqrt{2}} & 0 & -\frac{1}{2} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \\ 0 & 0 & 0 & -\frac{1}{\sqrt{2}} & 0 & -\frac{1}{2} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \end{bmatrix} \tag{3.26}$$

We now state the following two lemmas without proof[5], providing a link between the discrete Haar basis and $\mathcal{P}_k$.

Lemma 3.1.1 *Consider the case in which $x(t)$ is a scalar process. The discrete Haar matrix V_k provides a complete orthonormal set of eigenvectors for the matrix $\mathcal{P}_k$; i.e.*

$$\mathcal{P}_k = V_k \Lambda_k V_k^T \tag{3.27}$$

where Λ_k is a diagonal matrix.

Lemma 3.1.2 *Given $\tilde{\mathcal{P}}_{k,k+1}$ and V_{k+1},*

$$\tilde{\mathcal{P}}_{k,k+1} V_{k+1} = \left[0 \mid V_k \tilde{\Lambda}_k\right] \tag{3.28}$$

[5]See [13]

where $\tilde{\Lambda}_k$ is a diagonal matrix of dimension 2^k.

Note that the matrix $\tilde{\mathcal{P}}_{k,k+1}$ is $2^k \times 2^{k+1}$. Lemma 3.1.2 simply says that the first 2^k columns of V_{k+1}, i.e. the part that lives on scale $k+1$, are orthogonal to $\tilde{\mathcal{P}}_{k,k+1}$. Meanwhile, the remaining columns of V_{k+1} are quasi-eigenvectors of $\tilde{\mathcal{P}}_{k,k+1}$. Note also that the previous two lemmas are easily extended to the case of *vector* processes $x(t)$. In this case we must consider the block version of the discrete Haar matrix, defined as in Definition 3.1.1 except we now consider "dilated, translated, and scaled" versions of the block matrix $[I \ -I]^T$ instead of the vector $[1, -1]^T$, where each block is of size equal to the dimension of x. It is important to note that the discrete Haar transform, i.e. the computation of $V_k z$ can be performed in an extremely efficient manner(in the block case as well), by successive additions and subtractions of pairs of elements.

Returning to eq.(3.11) we see that we can obtain an extremely efficient transform version of the recursion. Specifically, let

$$\hat{z}_k = V_k^T \hat{X}_k \tag{3.29}$$

Then

$$\hat{z}_k = [0 \mid \tilde{\Lambda}_k]\Lambda_{k+1}^{-1}\hat{z}_{k+1} \tag{3.30}$$

Thus, we see that the fine scale components of $\hat{X}_k$ are unneeded at the coarser scale; i.e. only the lower half of $\hat{z}_{k+1}$, which depends only on pairwise sums of the elements of $\hat{X}_k$, enters in the computation. So, if we let

$$\Lambda_{k+1} = diag(M_{k+1}, D_{k+1}) \tag{3.31}$$

$$\hat{z}_{k+1} = \begin{bmatrix} \hat{z}_{k+1}^1 \\ \hat{z}_{k+1}^2 \end{bmatrix} \tag{3.32}$$

where M_{k+1} and D_{k+1} each have $2^k \times 2^k$ blocks, we see that

$$\hat{z}_k = \tilde{\Lambda}_k D_{k+1}^{-1}\hat{z}_{k+1}^2 \tag{3.33}$$

Finally, while we have focused on the structure of eq.(3.11) , it should be clear that analogous algorithmic structures - i.e. the summation form as in eq.(3.24) or

the transform form in eq.(3.33) - exist for the initial data incorporation step eq.(3.6). Thus, once we perform a single Haar transform on the original data Y, we can compute the transformed optimal estimates $\hat{z}_M, \hat{z}_{M-1}, ...$ in a block-diagonalized manner as in eq.(3.33), where the work required to compute eq.(3.33) is only $O(2^k \times$ dim. of state). Also, it is certainly possible to consider multi-scale measurements in this context, developing filtering(fine-to-coarse) and full smoothing(multigrid or fine-to-coarse-to-fine) algorithms in the transform domain. These will be described in detail in [13].

3.2 A Multigrid Relaxation Algorithm

In this section we use the Markov structure of eq.(2.9) to define an iterative algorithm for the computation of the optimal estimates at all scales given measurements at all scales. As in the multigrid solution of partial differential equations, this approach may have significant computational advantages even if only the finest level estimates are actually desired and if only fine level measurements are available.

Let Y denote the full set of measurements at all scales. Then, thanks to Markovianity we have the following: For $m(t) = M$, the finest scale

$$\begin{aligned} E[x(t)|Y] &= E\left\{E[x(t)|x(\gamma^{-1}t), Y]|Y\right\} \\ &= E\left\{E[x(t)|x(\gamma^{-1}t), y(t)]|Y\right\} \end{aligned} \tag{3.34}$$

For $m(t) < M$

$$\begin{aligned} E[x(t)|Y] &= E\left\{E[x(t)|x(\gamma^{-1}t), x(\alpha t), x(\beta t), Y]|Y\right\} \\ &= E\left\{E[x(t)|x(\gamma^{-1}t), x(\alpha t), x(\beta t), y(t)]|Y\right\} \end{aligned} \tag{3.35}$$

The key now is to compute the inner expectations in eq.(3.34) and eq.(3.35), and to do this we need to view $x(\gamma^{-1}t)$, $x(\alpha t)$, and $x(\beta t)$ as *measurements* of $x(t)$. For the latter two, this comes directly from eq.(2.9). For $x(\gamma^{-1}t)$, however, we need the reverse-time version of eq.(2.9). Assuming that $A(m)$ is invertible for all m we can directly apply the results of [30]:

$$x(\gamma^{-1}t) = F(m(t))x(t) - A^{-1}(m(t))B(m(t))\tilde{w}(t) \tag{3.36}$$

with

$$F(m(t)) = A^{-1}(m(t))[I - B(m(t))B^T(m(t))P_x^{-1}(m(t))] \tag{3.37}$$

and where $\tilde{w}(t)$ is a white noise process with covariance

$$\begin{aligned} E[\tilde{w}(t)\tilde{w}^T(t)] &= I - B^T(m(t))P_x^{-1}(m(t))B(m(t)) \\ &\triangleq \tilde{Q}(m(t)) \end{aligned} \tag{3.38}$$

Let us now focus on the computation of the inner expectation of eq.(3.35). We can write the following equations for $x(\gamma^{-1}t)$, $x(\alpha t)$ $x(\beta t)$, $y(t)$.

$$\begin{aligned} y(t) &= C(m(t))x(t) + v(t) && (3.39)\\ x(\gamma^{-1}t) &= F(m(t))x(t) - A^{-1}(m(t))B(m(t))\tilde{w}(t) && (3.40)\\ x(\alpha t) &= A(m(\alpha t))x(t) + B(m(\alpha t))w(\alpha t) && (3.41)\\ x(\beta t) &= A(m(\beta t))x(t) + B(m(\beta t))w(\beta t) && (3.42) \end{aligned}$$

which can be rewritten as

$$\mathcal{Y} = Hx(t) + \xi \tag{3.43}$$

where

$$H \triangleq \begin{bmatrix} C(m(t)) \\ F(m(t)) \\ A(m(\alpha t)) \\ A(m(\beta t)) \end{bmatrix}, \xi \triangleq \begin{bmatrix} v(t) \\ A^{-1}(m(t))B(m(t))\tilde{w}(t) \\ B(m(\alpha t))w(\alpha t) \\ B(m(\beta t))w(\beta t) \end{bmatrix} \tag{3.44}$$

and[6]

$$x(t) \perp \xi \tag{3.45}$$

Note the covariance of ξ has the following structure.

$$\begin{aligned} E[\xi\xi^T] &= \begin{bmatrix} R(m(t)) & 0 & 0 & 0 \\ 0 & R_1(m(t)) & 0 & 0 \\ 0 & 0 & R_2(m(\alpha t)) & 0 \\ 0 & 0 & 0 & R_2(m(\alpha t)) \end{bmatrix} \\ &\triangleq \mathcal{R} \end{aligned} \tag{3.46}$$

where

$$\begin{aligned} R_1(m(t)) &\triangleq A^{-1}(m(t))B(m(t))\tilde{Q}(t)B^T(m(t))A^{-T}(m(t)) && (3.47)\\ R_2(m(\alpha t)) &\triangleq B(m(\alpha t))B^T(m(\alpha t)) && (3.48)\\ &= B(m(\beta t))B^T(m(\beta t)) && (3.49) \end{aligned}$$

[6]We denote $\perp$ to denote orthogonal in the sense that $a \perp b$ if $E[ab^T] = 0$.

The inner expectation in eq.(3.35) can now be computed as follows.

$$\begin{aligned} E[x(t)|\mathcal{Y}] &= (P_x^{-1}(t) + H^T\mathcal{R}^{-1}H)^{-1}H^T\mathcal{R}^{-1}\mathcal{Y} \\ &= \mathcal{P}^{-1}\left\{K_1 y(t) + K_2 x(\gamma^{-1}t) + K_3 x(\alpha t) + K_4 x(\beta t)\right\} \end{aligned} \tag{3.50}$$

where

$$K_1 = C^T(m(t))R^{-1}(m(t)) \tag{3.51}$$

$$K_2 = F^T(m(t))R_1^{-1}(m(t)) \tag{3.52}$$

$$K_3 = A^T(m(\alpha t))R_2^{-1}(m(\alpha t)) \tag{3.53}$$

$$K_4 = A^T(m(\beta t))R_2^{-1}(m(\alpha t)) \tag{3.54}$$

$$\mathcal{P} = P_x^{-1}(t) + K_1C(m(t)) + K_2F(m(t)) + K_3A(m(\alpha t)) + K_4A(m(\beta t)) \tag{3.55}$$

We can use a similar procedure for computing $E[x(t)|x(\gamma^{-1}t), y(t)]$ so that we can now carry out the outer expectations in eq.(3.34) and eq.(3.35) to yield the following formulas for $\hat{x}(t) \triangleq E[x(t)|Y]$.
For $m(t) = M$

$$\hat{x}(t) = (\mathcal{P}')^{-1}\left\{C^T(m(t))R^{-1}(m(t))y(t) + F^T(m(t))R_1^{-1}(m(t))\hat{x}(\gamma^{-1}t)\right\} \tag{3.56}$$

For $m(t) < M$

$$\hat{x}(t) = \mathcal{P}^{-1}\left\{K_1 y(t) + K_2\hat{x}(\gamma^{-1}t) + K_3\hat{x}(\alpha t) + K_4\hat{x}(\beta t)\right\} \tag{3.57}$$

where

$$\mathcal{P}' = P_x^{-1}(t) + C^T(m(t))R^{-1}(m(t))C(m(t)) + F^T(m(t))R_1^{-1}(m(t))F(m(t)) \tag{3.58}$$

Thus, eq.(3.56) and eq.(3.57) are an implicit set of equations for $\{\hat{x}(t)|t \in T\}$. Note that the computation involved at each point on the tree involves only its three nearesest neighbors and the measurement at that point. This suggests the use of a Gauss-Seidel relaxation algorithm for solving this set of equations. Note that the computations of all the points along a particular scale are independent of each other,

allowing these computations to be performed in parallel. We could then arrange the computations of the relaxation algorithm so that we do all the computations at a particular scale in parallel, i.e. a Jacobi sweep at this scale, and the sweeps can be performed consecutively moving up and down the tree. The possibilities for parallelization are plentiful; the fact that the computations can now be counted in terms of scales rather that in terms of individual points already reduces the size of the problem from $O(2^{M+1})$, which is the number of nodes on the tree, to $O(M)$. The following is one possible algorithm(recall our previous notation where X_k denotes the vector of points along the kth level of the tree; $\hat{X}_k$ now denotes the smoothed estimate of X_k).

Algorithm 3.2.1 *Multigrid Relaxation Algorithm:*

1. *Initialize* $\hat{X}_0, ..., \hat{X}_M$ *to* 0.
2. *Do Until Desired Convergence is Attained:*
   (a) *Compute in parallel eq.(3.56) for each entry of* $\hat{X}_M$
   (b) *For* $k = M-1$ *to* 0
   Compute in parallel eq.(3.57) for each entry of $\hat{X}_k$
   (c) *For* $k = 1$ *to* $M-1$
   Compute in parallel eq.(3.57) for each entry of $\hat{X}_k$

Essentially, Algorithm 3.2.1 starts at the finest scale, moves sequentially up the tree to the coarsest scale, moves sequentially back down to the finest scale, then cycles through this procedure until convergence is attained. In multigrid terminology[8] this is a *V-cycle*. The issue of convergence is normally studied via the analysis of the global matrix formed from the set of implicit equations, eq.(3.56)-(3.57). However, our problem has a particular structure that allows us to give the following relatively simple argument for the convergence of a Gauss-seidel relaxation algorithm under any ordering of the local computations. We can think of the computation of $E[x(t)|Y]$ for all $t \in T$ as performing the minimization of a convex quadratic cost function with respect to $\{x(t) : t \in T\}$ and each Gauss-Seidel step is the local minimization with

respect to a particular $x(t)$ with the remaining x's held constant. Since each local minimization results in a reduction of the overall cost function and this function is convex, then the limit of the sequence of local minimizations results in the global minimization of the cost function.

3.3 Two-Sweep, Rauch-Tung-Striebel Algorithm

In this section we consider the same problem as in Section 3.2, but the algorithm structure we define is recursive, rather than iterative, and in fact is a generalization of the well-known Rauch-Tung-Striebel(RTS) smoothing algorithm for causal state models. The algorithm once again involves a pyramidal set of steps and a considerable level of parallelism.

To begin, let us recall the structure of the RTS algorithm for a state model with state $\hat{x}(t)$. The first step of the process consists of a Kalman filter for computing $\hat{x}(t|t)$, predicting to obtain $\hat{x}(t+1|t)$ and updating with the new measurement $y(t)$. The second step propagates backward combining the smoothed estimate $\hat{x}_s(t+1)$ with the filtered estimate at the previous point in time $\hat{x}(t|t)$(or equivalently $\hat{x}(t+1|t)$) to compute $\hat{x}_s(t)$. In the case of estimation on trees, we have a very similar structure; indeed the backward sweep and measurement update are identical in form to the RTS algorithm. The prediction step is, however, somewhat more complex, and while it can be written as a single step, we prefer to think of it as two parallel prediction steps, each as in RTS, followed by a *merge* step that has no counterpart for state models evolving in *time*. One other difference is that the forward sweep of our algorithm is from fine-to-coarse and thus involves the backward version eq.(3.36) of our original model eq.(2.9).

To begin let us define some notation:

$$\begin{aligned} Y_t &= \{y(s)|s = t \text{ or } s \text{ is a descendent of } t\} \\ &= \{y(s)|s \in (\alpha,\beta)^* t \ , \ m(s) \leq M\} \qquad (3.59) \\ Y_t^+ &= \{y(s)|s \in (\alpha,\beta)^* t \ , \ t < m(s) \leq M\} \qquad (3.60) \\ \hat{x}(\cdot|t) &= E[x(\cdot)|Y_t] \qquad (3.61) \\ \hat{x}(\cdot|t+) &= E[x(\cdot)|Y_t^+] \qquad (3.62) \end{aligned}$$

The interpretation of these estimates is provided in Figure 2.

To begin, consider the measurement update. Specifically, suppose that we have computed $\hat{x}(t|t+)$ and the corresponding error covariance, $P(m(t)|m(t)+)$; the fact

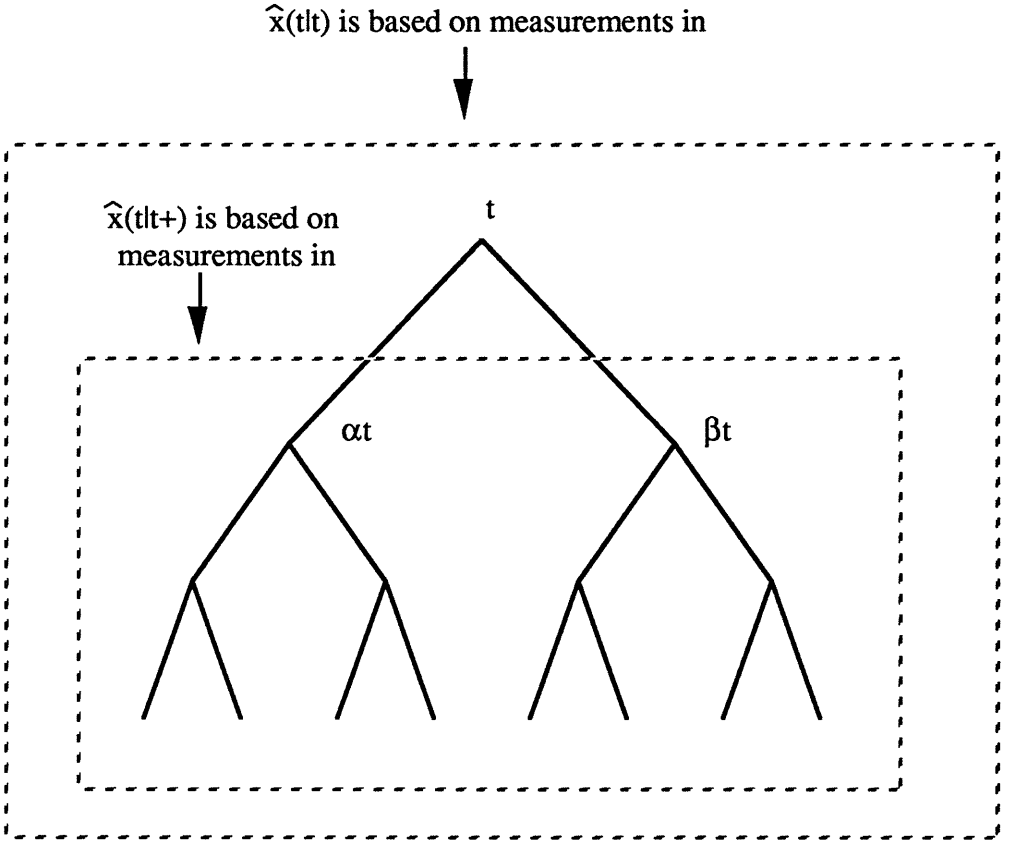

Figure 2: Representation of Meaurement Update and Merged Estimates

that this depends only on scale should be evident from the structure of the problem. Then, standard estimation results yield

$$\hat{x}(t|t) = \hat{x}(t|t+) + K(m(t))[y(t) - C(m(t))\hat{x}(t|t+)] \quad (3.63)$$

$$K(m(t)) = P(m(t)|m(t)+)C^T(m(t))V^{-1}(m(t)) \quad (3.64)$$

$$V(m(t)) = C(m(t))P(m(t)|m(t)+)C^T(m(t)) + R(m(t)) \quad (3.65)$$

and the resulting error covariance is given by

$$P(m(t)|m(t)) = [I - K(m(t))C(m(t))]P(m(t)|m(t)+) \quad (3.66)$$

Note that the computations begin on the finest level(m(t)=M) with $\hat{x}(t|t+) = 0$, $P(M|M+) = P_x(M)$.

Suppose now that we have computed $\hat{x}(\alpha t|\alpha t)$ and $\hat{x}(\beta t|\beta t)$. Note that $Y_{\alpha t}$ and $Y_{\beta t}$ are disjoint and these estimates can be calculated in parallel. Furthermore, once again they have equal error covariances, denoted by $P(m(\alpha t)|m(\alpha t)) = P(m(t)+1|m(t)+1)$. We then compute $\hat{x}(t|\alpha t)$ and $\hat{x}(t|\beta t)$ which are given by

$$\hat{x}(t|\alpha t) = F(m(t)+1)\hat{x}(\alpha t|\alpha t) \quad (3.67)$$

$$\hat{x}(t|\beta t) = F(m(t)+1)\hat{x}(\beta t|\beta t) \quad (3.68)$$

with corresponding identical error covariances $P(m(t)|m(t)+1)$ given by

$$P(m(t)|m(t)+1) = F(m(t)+1)P(m(t)+1|m(t)+1)F^T(m(t)+1) + \mathcal{Q}(m(t)+1) \quad (3.69)$$

$$\mathcal{Q}(m(t)+1) = A^{-1}(m(t)+1)B(m(t)+1)\tilde{Q}(m(t)+1)B^T(m(t)+1)A^{-T}(m(t)+1) \quad (3.70)$$

Eq.(3.67) and eq.(3.68) follow from projecting both sides of our backward model eq.(3.36) onto $Y_{\alpha t}$ and $Y_{\beta t}$, respectively. By noting that the dynamics of the one-step prediction error are identical to the dynamics of our backward model eq.(3.36), we arrive at eq.(3.69) by squaring both sides of the equation and taking expectations.

These estimates must then be merged to form $\hat{x}(t|t+)$. The derivation of this computation can be given as follows. By definition

$$\hat{x}(t|t+) = E[x(t)|Y_{\alpha t}, Y_{\beta t}] \tag{3.71}$$

But from our model, eq.(2.9), we can decompose $Y_{\alpha t}$ and $Y_{\beta t}$ in the following way.

$$Y_{\alpha t} = M_{\alpha t}x(t) + \xi_1 \tag{3.72}$$

$$Y_{\beta t} = M_{\beta t}x(t) + \xi_2 \tag{3.73}$$

where the matrices $M_{\alpha t}$ and $M_{\beta t}$ contain products of $A(m(s))$, $m(s) > m(t)$, and the vectors ξ_1 and ξ_2 are functions of the driving noises $w(s)$ and the measurement noises $v(s)$ for s in the subtree strictly below αt and s in the subtree strictly βt, respectively, the latter fact implying $\xi_1 \perp \xi_2$. We also let

$$R_{\alpha t} = E[\xi_1 \xi_1^T] \tag{3.74}$$

$$R_{\beta t} = E[\xi_2 \xi_2^T] \tag{3.75}$$

We then write eq.(3.72) and eq.(3.73) as a single equation in the following way.

$$\mathcal{Y} = \mathcal{H}x(t) + \Xi \tag{3.76}$$

where

$$\mathcal{H} = \begin{bmatrix} M_{\alpha t} \\ M_{\beta t} \end{bmatrix}, \Xi = \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix}, \mathcal{R} = E[\Xi\Xi^T] \tag{3.77}$$

and $x(t) \perp \Xi$. As we did for the computations producing the implicit equations for the multigrid algorithm we can write the optimal estimate of $x(t)$ given $\mathcal{Y}$ in the following way.

$$\begin{aligned} \hat{x}(t|t+) &= [P_x^{-1}(t) + \mathcal{H}^T\mathcal{R}^{-1}\mathcal{H}]^{-1}\mathcal{H}^T\mathcal{R}^{-1}\mathcal{Y} \\ &= [P_x^{-1}(t) + M_{\alpha t}^T R_{\alpha t}^{-1} M_{\alpha t} + M_{\beta t}^T R_{\beta t}^{-1} M_{\beta t}]^{-1}[M_{\alpha t}^T R_{\alpha t}^{-1} Y_{\alpha t} + M_{\beta t}^T R_{\beta t}^{-1} Y_{\beta t}] \end{aligned} \tag{3.78}$$

But since

$$P(t|\alpha t) = [P_x^{-1}(t) + M_{\alpha t}^T R_{\alpha t}^{-1} M_{\alpha t}]^{-1} \tag{3.79}$$

$$P(t|\beta t) = [P_x^{-1}(t) + M_{\beta t}^T R_{\beta t}^{-1} M_{\beta t}]^{-1} \tag{3.80}$$

we can rewrite eq.(3.78) as

$$\hat{x}(t|t+) = P(m(t)|m(t)+)P^{-1}(m(t)|m(t)+1)[\hat{x}(t|\alpha t) + \hat{x}(t|\beta t)] \tag{3.81}$$

$$P(m(t)|m(t)+) = [2P^{-1}(m(t)|m(t)+1) - P_x^{-1}(t)]^{-1} \tag{3.82}$$

where we have used the fact that $P(t|\alpha t) = P(t|\beta t) = P(m(t)|m(t)+1)$.

We now derive the formulas for the second part of the RTS algorithm involving the propogation downward along the tree combining the smoothed estimate $\hat{x}_s(\gamma^{-1}t)$ with the filtered estimate $\hat{x}(t|t)$ to produce $\hat{x}_s(t)$. Our derivation relies essentially on the following orthogonal decomposition of Y_0(the measurements at every node on the tree).

For each t, Y_t, as defined in eq.(3.59) is the set of measurements in the subtree beneath t(and including the measurement at t). Let $Y_{\bar{t}}$ denote all the remaining measurements, and viewing this as one large vector, define

$$\nu_{\bar{t}|t} = Y_{\bar{t}} - E[Y_{\bar{t}}|Y_t] \tag{3.83}$$

so that $\nu_{\bar{t}|t} \perp Y_t$ and the linear span of the set of *all* measurements, Y_0, is given by

$$\text{span } Y_0 = \text{ span } \{Y_t, Y_{\bar{t}}\} = \text{ span } \left\{Y_t, \nu_{\bar{t}|t}\right\} \tag{3.84}$$

Then

$$\begin{aligned} \hat{x}_s(t) &= E[x(t)|Y_t, \nu_{\bar{t}|t}] \\ &= \hat{x}(t|t) + E[x(t)|\nu_{\bar{t}|t}] \end{aligned} \tag{3.85}$$

If we write $x(t)$ as

$$x(t) = \tilde{x}(t|t) + \hat{x}(t|t) \tag{3.86}$$

and note that

$$\hat{x}(t|t) \perp \nu_{\bar{t}|t} \tag{3.87}$$

then we can write the following.

$$\hat{x}_s(t) = \hat{x}(t|t) + E[\tilde{x}(t|t)|\nu_{\bar{t}|t}] \tag{3.88}$$

Using the same argument on $\hat{x}_s(\gamma^{-1}t)$ allows us to write

$$\hat{x}_s(\gamma^{-1}t) = \hat{x}(\gamma^{-1}t|t) + E[\tilde{x}(\gamma^{-1}t|t)|\nu_{\bar{t}|t}] \tag{3.89}$$

Suppose the following equality were to hold.

$$E[\tilde{x}(t|t)|\nu_{\bar{t}|t}] = LE[\tilde{x}(\gamma^{-1}t|t)|\nu_{\bar{t}|t}] \tag{3.90}$$

Then eq.(3.88) and eq.(3.89) could be combined to yield the following formula.

$$\hat{x}_s(t) = \hat{x}(t|t) + L\left[\hat{x}_s(\gamma^{-1}t) - \hat{x}(\gamma^{-1}t|t)\right] \tag{3.91}$$

We now proceed to show that eq.(3.90) indeed holds and compute explicitly the matrix L. We begin with the following iterated expectation.

$$E[\tilde{x}(t|t)|\nu_{\bar{t}|t}] = E[E[\tilde{x}(t|t)|\tilde{x}(\gamma^{-1}t|t), \nu_{\bar{t}|t}]|\nu_{\bar{t}|t}] \tag{3.92}$$

We now examine the inner expectation, $E[\tilde{x}(t|t)|\tilde{x}(\gamma^{-1}t|t), \nu_{\bar{t}|t}]$, in detail. In particular the linear span of $\left\{\tilde{x}(\gamma^{-1}t|t), \nu_{\bar{t}|t}\right\}$ has the following structure.

$$\text{span}\left\{\tilde{x}(\gamma^{-1}t|t), \nu_{\bar{t}|t}\right\} = \text{span}\left\{\tilde{x}(\gamma^{-1}t|t), \tilde{w}_{s|x_s}, w_{s'}, v_{s''}\right\} \tag{3.93}$$

$$\tilde{x}(\gamma^{-1}t|t) \perp \tilde{w}_{s|x_s}, w_{s'}, v_{s''} \tag{3.94}$$

where

$$s, s', s'' \notin \text{ subtree under } t \tag{3.95}$$

To show this we note the following decomposition of $Y_{\bar{t}}$.

$$Y_{\bar{t}} = L_1 x(\gamma^{-1}t) + f(\tilde{w}_{s|x_s}, w_{s'}, v_{s''}) \tag{3.96}$$

where f is a linear function of its arguments. Substituting eq.(3.96) into eq.(3.83) yields

$$\nu_{\bar{t}|t} = L_1\tilde{x}(\gamma^{-1}t|t) + f(\tilde{w}_{s|x_s}, w_{s'}, v_{s''}) \tag{3.97}$$

where we have used the fact that $f(\tilde{w}_{s|x_s}, w_{s'}, v_{s''}) \perp Y_t$. The fact that $\tilde{x}(\gamma^{-1}t|t) \perp f(\tilde{w}_{s|x_s}, w_{s'}, v_{s''})$ verifies eq.(3.93). Using eq.(3.93) we have that

$$E[\tilde{x}(t|t)|\tilde{x}(\gamma^{-1}t|t), \nu_{\bar{t}|t}] = E[\tilde{x}(t|t)|\tilde{x}(\gamma^{-1}t|t)] \tag{3.98}$$

where we have also used the fact that $f(\tilde{w}_{s|x_s}, w_{s'}, v_{s''}) \perp \tilde{x}(t|t)$. Substituting eq.(3.98) into eq.(3.92) we get

$$E[\tilde{x}(t|t)|\nu_{\bar{t}|t}] = E[E[\tilde{x}(t|t)|\tilde{x}(\gamma^{-1}t|t)]|\nu_{\bar{t}|t}] \tag{3.99}$$

But by using our backward equations, eq.(3.36), eq.(3.67)(in the latter case with $\alpha t \mapsto t$ and $t \mapsto \gamma^{-1}t$) we find that

$$E[\tilde{x}(t|t)|\tilde{x}(\gamma^{-1}t|t)] = P(m(t)|m(t))F^T(m(t))P^{-1}(m(t)-1|m(t))\tilde{x}(\gamma^{-1}t|t) \tag{3.100}$$

This leads to our desired result.

$$E[\tilde{x}(t|t)|\nu_{\bar{t}|t}] = P(m(t)|m(t))F^T(m(t))P^{-1}(m(t)-1|m(t))E[\tilde{x}(\gamma^{-1}t|t)|\nu_{\bar{t}|t}] \tag{3.101}$$

Finally, eq.(3.90), eq.(3.91), and eq.(3.101) yield the following smoothing formula.

$$\hat{x}_s(t) = \hat{x}(t|t) + P(m(t)|m(t))F^T(m(t))P^{-1}(m(t)-1|m(t))\left[\hat{x}_s(\gamma^{-1}t) - \hat{x}(\gamma^{-1}t|t)\right] \tag{3.102}$$

We now summarize the overall two-sweep algorithm:

Upward Sweep

Measurement Update:

$$\hat{x}(t|t) = \hat{x}(t|t+) + K(m(t))[y(t) - C(m(t))\hat{x}(t|t+)] \tag{3.103}$$

$$K(m(t)) = P(m(t)|m(t)+)C^T(m(t))V^{-1}(m(t)) \tag{3.104}$$

$$V(m(t)) = C(m(t))P(m(t)|m(t)+)C^T(m(t)) + R(m(t)) \tag{3.105}$$

$$P(m(t)|m(t)) = [I - K(m(t))C(m(t))]P(m(t)|m(t)+) \tag{3.106}$$

One-step Prediction:

$$\hat{x}(\gamma^{-1}t|t) = F(m(t))\hat{x}(t|t) \tag{3.107}$$

$$P(m(t)-1|m(t)) = F(m(t))P(m(t)|m(t))F^T(m(t)) + \mathcal{Q}(m(t)) \tag{3.108}$$

$$\mathcal{Q}(m(t)) = A^{-1}(m(t))B(m(t))\tilde{Q}(m(t))B^T(m(t))A^{-T}(m(t)) \tag{3.109}$$

Merge Step:

$$\hat{x}(t|t+) = P(m(t)|m(t)+)P^{-1}(m(t)|m(t)+1)[\hat{x}(t|\alpha t) + \hat{x}(t|\beta t)] \tag{3.110}$$

$$P(m(t)|m(t)+) = [2P^{-1}(m(t)|m(t)+1) - P_x^{-1}(t)]^{-1} \tag{3.111}$$

Downward Sweep

$$\hat{x}_s(t) = \hat{x}(t|t) + P(m(t)|m(t))F(m(t))P^{-1}(m(t)-1|m(t))\left[\hat{x}_s(\gamma^{-1}t) - \hat{x}(\gamma^{-1}t|t)\right] \tag{3.112}$$

4 Discussion

In this paper we have introduced a class of stochastic processes defined on dyadic trees and have described several estimation algorithms for these processes. The consideration of these processes and problems has been motivated by a desire to develop multi-scale descriptions of stochastic processes and in particular by the deterministic theory of multi-scale signal representations and wavelet transforms. The algorithms we have described have connections with Laplacian pyramids, Haar transforms, multigrid relaxation algorithms, and the Rauch-Tung-Striebel form of the optimal smoother for linear state-space models.

In addition to open questions directly related to the models we have considered here there are a number of related research problems under consideration. We limit our comments here to three of these.

1. *Modeling of scalar isotropic processes on trees*

   As we mentioned in Section II, isotropic processes have correlation structures that depend only on the distance between points on the tree. A natural extension of a classical 1D time series modeling problem is the construction of dynamic models that match a given isotropic correlation function $K_{xx}(k)$ for a specified number of lags $k = 0, 1, ...N$. This problem is studied in detail in[5,6] and in particular an extension of classical AR modeling is developed and with it a corresponding generalization of the Levinson and Schur recursions for AR models as the order N increases. A few comments about this theory are in order. First, the sequence $K_{xx}(k)$ must satisfy an even more strict set of conditions to be a valid correlation function for an isotropic tree process than it does to be the correlation function of a time series. In particular, since the sequence $x(t), x(\gamma^{-1}t), x(\gamma^{-2}t), ...$ is a standard time series, we see that $K_{xx}(k)$ must be a positive definite function. Moreover, considering the covariance of the three points $x(\alpha t)$, $x(\beta t)$, $x(\gamma^{-1}t)$, we conclude that the following condition must be

satisfied:

$$\begin{bmatrix} K_{xx}(0) & K_{xx}(2) & K_{xx}(0) \\ K_{xx}(2) & K_{xx}(0) & K_{xx}(0) \\ K_{xx}(0) & K_{xx}(2) & K_{xx}(0) \end{bmatrix} \geq 0 \tag{4.1}$$

Such a condition and many others that must be satisfied do not arise in usual time series. In particular an isotropic process $x(t)$ is one whose statistics are invariant to any isometry on the index set T, i.e. any invertible map preserving distance. For time series such isometries are quite limited: translations, $t \mapsto t+n$, and reflections $t \mapsto -t$. For dyadic trees the set of isometries is far richer, placing many more constraints on K_{xx}. Referring to the Levinson algorithm, recall that the validity of $K_{xx}(k)$ as a covariance function manifests itself in a sequence of reflection coefficients that must take values between $\overset{+}{-}$ 1. For trees the situation is more complex: for n odd $|k_n| < 1$ while for n even $-\frac{1}{2} < k_n < 1$, $k(n)$ being the nth reflection coefficient. Furthermore, since dyadic trees are fundamentally infinite dimensional, the Levinson algorithm involves "forward"(with respect to the scale index m) and "backward" prediction filters of dimension that grows with order, as one must predict a window of values at the boundary of the filter domain. Also, the filters are not strictly causal in m. For example, while the first-order AR model is simply the scalar, constant-parameter version of the model eq.(2.9) considered here, the second order model represents a forward prediction of $x(t)$ based on $x(\gamma^{-1}t)$, $x(\gamma^{-2}t)$ *and* $x(\delta t)$, which is at the same scale as $x(t)$(refer to Figure 1). The third-order forward model represents the forward prediction of $x(t)$ *and* $x(\delta t)$ based on $x(\gamma^{-1}t)$, $x(\gamma^{-2}t)$, $x(\gamma^{-2}t)$ *and* $x(\delta\gamma^{-1}t)$. We refer the reader to [5,6] for details.

2. *Models on lattices and more complex correlation structures*

As discussed in Section II, it is the Haar wavelet decomposition that most directly suggests the investigation of processes on trees. While these models can certainly be used to generate stochastic processes using other scaling functions $\phi(t)$, such functions more naturally suggest lattice structures on the set of indices (m,n). The investigation of processes and algorithms analogous to those

considered here but for such lattices are of obvious interest. In this case we expect more complex algorithms since any two points have least upper bounds *and* greatest lower bounds so that the correlation structure of the resulting processes will be more involved. If, however, the filter impulse response $h(n)$ is FIR we expect that finite-dimensional algorithms analogous to those in Section III can be developed, with considerable parallelism, as we have here, but with additional connectivity. An additional extension which should yield similar algorithmic structures is to consider the model as in eq.(2.9) but with $w(t)$ independent from scale to scale but correlated along each scale. Indeed one can argue that the model eq.(2.6) naturally suggests at least a finite length correlation structure for $w(m,n)$ as a function of n. Results related to these ideas will be forthcoming.

3. The problems we have considered in this paper are fundamentally discrete in nature; i.e. there is a finest level scale M at which processes are defined and considered. Given the motivation from and setting of wavelets and multi-resolution representations it is of interest to understand the stochastic version of the limit in eq.(2.1) and the properties and statistics of the limit process. Results along these lines will also be described in future papers as will relations to self-similar processes and fractals.

References

[1] B. Anderson and T. Kailath, "Forwards, backwards, and dynamically reversible Markovian models of second-order processes," IEEE Trans. Circuits and Systems, CAS-26, no. 11, 1978, pp.956-965.

[2] J. Arnaud, "Fonctions spheriques et fonctions definies-positives sur l'arbre homogene," C.R. Acad. Sc., Serie A,1980, pp.99-101.

[3] J. Arnaud and G. Letac, "La formule de representation spectrale d'un processus gaussien stationnaire sur un arbre homogene," Laboratoire de Stat. et. Prob. - U.A.-CNRS 745, Toulouse.

[4] M. Barnsley, *Fractals Everywhere*, Academic Press, San Diego, 1988.

[5] M. Basseville, A. Benveniste, "Traitement statistique du signal multi-echelle," IRISA Report No. 446 1988.

[6] A. Benveniste, M. Basseville, A. Willsky, K. Chou, "Multiresolution Signal Processing and Modeling of Isotropic Processes on Homogeneous Trees," to be presented at Int'l Symp. on Math. Theory of Networks and Systems, Amsterdam, June 1989.

[7] A. Brandt, "Multi-level adaptive solutions to boundary value problems," Math. Comp. Vol. 13, 1977, pp.333-390.

[8] W. Briggs, *A Multigrid Tutorial*, SIAM, Philadelphia, PA, 1987.

[9] C. Bunks, *Random Field Modeling for Interpretation and Analysis of Layered Data*, MIT, Dept. Electrical Engineering, PhD Thesis, 1987.

[10] P. Burt and E. Adelson, "The Laplacian pyramid as a compact image code," IEEE Trans. Comm. 31, 1983, pp.482-540.

[11] P. Cartier, "Harmonic analysis on trees," Proc. Sympos. Pure Math., Vol. 26, Amer. Math. Soc. Providence R.I., 1974, pp.419-424.

[12] P. Cartier, "Geometrie et analyse sur les arbres," Seminaire Bourbaki, 24eme annee, Expose no. 407, 1971/72.

[13] K. Chou, *A Stochastic Modeling Approach to Multiscale Signal Processing*, MIT, Dept. Electrical Engineering, PhD Thesis, 1989(in preparation).

[14] I. Daubechies, "Orthonormal bases of compactly supported wavelets," Comm. on Pure and Applied Math. 91, 1988, pp. 909-996.

[15] I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis," AT&T Bell Laboratories Report.

[16] I. Daubechies, A. Grossman, and Y. Meyer, "Painless non-orthogonal expansions," J. Math. Phys. 27, 1986, pp.1271-1283.

[17] B. Gidas, "A renormalization group approach to image processing problems," IEEE Trans. on Pattern Anal. and Mach. Int. 11, 1989, pp. 164-180.

[18] J. Goodman and A. Sokal, "Multi-grid Monte Carlo I. conceptual foundations," Preprint, Dept. Physics New York University, New York Nov. 1988; to be published.

[19] A. Grossman and J. Morlet, "Decomposition of Hardy functions into square integreable wavelets of constant shape," SIAM J. Math. Anal. 15, 1984, pp. 723-736.

[20] W. Hackbusch and U. Trottenberg, Eds., *Multigrid Methods*, Springer-Verlag, N.Y., N.Y., 1982.

[21] R. Kronland-Martinet, J. Morlet and A. Grossman, "Analysis of sound patterns through wavelet transforms," preprint CPT-87/P, Centre de Physique Theorique, CNRS, Marseille, 1981.

[22] S. G. Mallat, "A compact multiresolution representation: the wavelet model," Dept. of Computer and Info. Science - U. of Penn., MS-CIS-87-69, GRASP LAB 113, Aug. 1987.

[23] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," Dept. of Computer and Info. Science - U. of Penn., MS-CIS-87-22, GRASP LAB 103, May. 1987.

[24] S. G. Mallat, "Multiresolution approximation and wavelets," Dept. of Computer and Info. Science - U. of Penn., MS-CIS-87-87, GRASP LAB 80, Sept. 1987.

[25] B. Mandelbrot, *The Fractal Geometry of Nature*, Freeman, New York, 1982.

[26] S. McCormick, *Multigrid Methods*, Vol. 3 of the SIAM Frontiers Series, SIAM, Philadelphia, 1987.

[27] Y. Meyer, "L'analyse par ondelettes," Pour la Science, Sept. 1987.

[28] D. Paddon and H. Holstein, Eds., *Multigrid Methods for Integral and Differential Equations*, Clarendon Press, Oxford, England, 1985.

[29] M. J. Smith and T. P. Barnwell, "Exact reconstruction techniques for tree-structured subband coders," IEEE Trans. on ASSP 34, 1986, pp.434-441.

[30] T. Verghese and T. Kailath, "A further note on backward Markovian models," IEEE Trans. on Information Theory IT-25, 1979, pp.121-124.