Adaptive Greedy Approximations[1]

Geoffrey Davis
Mathematics Department, Dartmouth College
Hanover, NH 03755
gdavis@cs.dartmouth.edu

Stéphane Mallat, Marco Avellaneda
Courant Institute, New York University
251 Mercer Street, New York, NY 10012
mallat@cs.nyu.edu, avellane@nyu.edu

Abstract

The problem of optimally approximating a function with a linear expansion over a redundant dictionary of waveforms is NP-hard. The greedy matching pursuit algorithm and its orthogonalized variant produce sub-optimal function expansions by iteratively choosing dictionary waveforms that best match the function's structures. A matching pursuit provides a means of quickly computing compact, adaptive function approximations.

Numerical experiments show that the approximation errors from matching pursuits initially decrease rapidly, but the asymptotic decay rate of the errors is slow. We explain this behavior by showing that matching pursuits are chaotic, ergodic maps. The statistical properties of the approximation errors of a pursuit can be obtained from the invariant measure of the pursuit. We characterize these measures using group symmetries of dictionaries and by constructing a stochastic differential equation model.

We derive a notion of the coherence of a signal with respect to a dictionary from our characterization of the approximation errors of a pursuit. The dictionary elements selected during the initial iterations of a pursuit correspond to a function's coherent structures. The tail of the expansion, on the other hand, corresponds to a noise which is characterized by the invariant measure of the pursuit map.

When using a suitable dictionary, the expansion of a function into its coherent structures yields a compact approximation. We demonstrate a denoising algorithm based on coherent function expansions.

1 Introduction

For data compression applications and fast numerical methods it is important to accurately approximate functions from a Hilbert space $\mathcal{H}$ using a small number of vectors from a given

[1] This work was supported by an ONR graduate fellowship, the AFOSR grant F49620-93-1-0102, ONR grant N00014-91-J-1967, and the Alfred Sloan Foundation

family $\{g_\gamma\}_{\gamma \in \mathbf{\Gamma}}$. For any $M > 0$, we want to minimize the approximation error

$$\epsilon(M) = \|f - \sum_{\gamma \in I_M} \beta_\gamma g_\gamma\|$$

where $I_M \subset \mathbf{\Gamma}$ is an index set of cardinality M. If the family $\{g_\gamma\}_{\gamma \in \mathbf{\Gamma}}$ is an orthonormal basis, then because

$$\epsilon(M) = \sum_{\gamma \in \mathbf{\Gamma} - I_M} | < g_\gamma, f > |^2,$$

the error is minimized by taking I_M to correspond to the M vectors which have the largest inner products $(| < f, g_\gamma > |)_{\gamma \in I_M}$.

Depending upon the basis and the space $\mathcal{H}$, it is possible to estimate the decay rate of the minimal approximation error $\epsilon_0(M) = \inf_{I_M} \epsilon(M)$ as M increases. For example, when $\{g_\gamma\}_{\gamma \in \mathbf{\Gamma}}$ is a wavelet basis the rate of decay of $\epsilon_0(M)$ can be estimated for functions that belong to a particular class of Besov spaces[9]. Conversely, the rate of decay of $\epsilon_0(M)$ can be used to determine to which Besov space in this class f belongs.

We can greatly improve these approximations to f by enlarging the collection $\{g_\gamma\}_{\gamma \in \mathbf{\Gamma}}$ beyond a basis. This enlarged, redundant family of vectors we call a dictionary. The advantage of redundancy in obtaining compact representations can be seen by considering the problem of representing a two-dimensional surface given by $f(x, y)$ on a subset of the plane, $I \times I$. An adaptive square mesh representation of f in the Besov space $B_q^\alpha(L^q(I))$, where $\frac{1}{q} = \frac{\alpha+1}{2}$, can be obtained using a wavelet basis. This wavelet representation can be shown to be asymptotically near optimal in the sense that the decay rate of the error $\epsilon(M)$ is equal to the largest decay attainable by a general class of non-linear transform-based approximation schemes [10].

Even these near-optimal representations are constrained by the fact that the decompositions are over a basis. The regular grid structure of the wavelet basis prevents the compact representation of many functions. For example, when f is equal to a basis wavelet at the largest scale, it can be represented exactly by a expansion consisting of a single element. However, if we translate this f by a small amount, then an accurate approximation can require many elements. One way to improve our approximations is to add to the set $\{g_\gamma\}_{\gamma \in \mathbf{\Gamma}}$ the collection of all translates of the wavelets. The class of functions which can be compactly represented will then be translation invariant. We can obtain even better compact approximations by expanding the dictionary to contain the extremely redundant set of all piecewise polynomial functions on arbitrary triangles.

When the dictionary is redundant, finding a family of M vectors that approximates f with an error close to the minimum $\epsilon_0(M)$ is clearly not achieved by selecting the vectors that have maximal inner products with f. In section 2 we prove that for general dictionaries the problem of finding M-element optimal approximations belongs to a class of computationally intractable problems, the set of NP-hard problems. It is widely believed (but unproven) that the number of operations required to solve an NP-hard problem grows faster than any polynomial in the input size [13].

Because of the difficulty of computing optimal expansions, we turn to suboptimal algorithms. In section 3 we review the performance of greedy algorithms, called matching pursuits, that were introduced in [24] [7]. We describe a fast implementation of these algorithms, and we give numerical examples for a dictionary composed of waveforms that are well-localized in time and frequency. Such dictionaries are particularly important for audio signal processing.

In our numerical experiments we find that the rate of decay of the greedy approximation error $\epsilon(M)$ decreases as M becomes large. These observations are explained by showing that a matching pursuit is a chaotic map which has an ergodic invariant measure. The proof of chaos in section 5 is given for a particular dictionary in a low-dimensional space, and we show numerical results which indicate that higher dimensional matching pursuits are also ergodic maps. Although the chaotic properties prevent prediction of the exact values of the approximation errors, the invariant measure provides a statistical description of these errors after a sufficient number of iterations of the pursuit. In section 6 we characterize these invariant measures using dictionary group invariances and by constructing a stochastic differential equation model for the distribution of asymptotic approximation errors for a dictionary consisting of a Dirac and a Fourier basis.

Our analysis of the asymptotic behavior of matching pursuits leads us to a notion of signal coherence with respect to a dictionary. Matching pursuit approximations yield efficient approximations when the number of terms is small, giving expansions of functions into what we call "coherent" structures. The error incurred by truncating function expansions when the convergence rate $\epsilon(M)$ becomes small corresponds to the realization of a process which is characterized by the invariant measure of the pursuit. We call these realizations "dictionary noise," and we describe an method for denoising signals based on our analysis of the convergence properties of pursuits. In section 7 we compare the numerical performance and complexity of orthogonal and non-orthogonal matching pursuits.

2 Complexity of Optimal Approximation

Let $\mathcal{H}$ be a Hilbert space. A dictionary for $\mathcal{H}$ is a family $\mathcal{D} = \{g_\gamma\}_{\gamma\in\Gamma}$ of unit vectors in $\mathcal{H}$ such that finite linear combinations of the g_γ are dense in $\mathcal{H}$. The smallest possible dictionary is a basis of $\mathcal{H}$; general dictionaries are redundant families of vectors. Vectors in $\mathcal{H}$ do not have unique representations as linear combinations of redundant dictionary elements. We prove below that for a redundant dictionary we must pay a high computational price to find an expansion with M dictionary vectors that yields the minimum approximation error.

Definition 2.1 *Let $\mathcal{D}$ be a dictionary of vectors in an N-dimensional Hilbert space $\mathcal{H}$. Let $\epsilon > 0$ and $M \leq N$. For a given $f \in \mathbf{R}^N$ an (ϵ, M)-***approximation** *is an expansion*

$$\tilde{f} = \sum_{i=1}^{M} \beta_i g_{\gamma_i}, \tag{1}$$

where $\beta_i \in \mathbf{C}$ and $g_{\gamma_i} \in \mathcal{D}$, for which

$$||\tilde{f} - f|| < \epsilon.$$

An **M-optimal approximation** *is an expansion that minimizes $||\tilde{f} - f||$.*

If our dictionary is an orthogonal basis, we can obtain an M-optimal approximation for any $f \in \mathcal{H}$ by computing the inner products $\{< f, g_\gamma >\}_{\gamma \in \mathbf{\Gamma}}$ and sorting the dictionary elements so that $|< f, g_{\gamma_i} >| \geq |< f, g_{\gamma_{i+1}} >|$. The signal $\tilde{f} = \sum_{i=1}^{M} < f, g_{\gamma_i} > g_{\gamma_i}$ is then an M-optimal approximation to f. In an N dimensional space, computing the inner products requires $O(N^2)$ operations and sorting $O(N \log N)$ so the overall algorithm is $O(N^2)$.

Finding M-optimal approximations for general dictionaries is a much more difficult problem. We briefly introduce some basic concepts from complexity theory in order to characterize the difficulty of the M-optimal approximation problem. Further details may be found in [13] and [4].

2.1 NP-completeness

A *concrete problem* is a map from a set of problem instances to a set of solutions. An instance of a concrete problem is a binary string which specifies all parameters and input data for the problem. We first consider the set of *concrete decision problems*, concrete problems for which the solution set is $\{0, 1\}$, i.e. problems with a "yes" or "no" answer.

A very important class of decision problems is the complexity class NP. The class NP consists of the set of problems for which solutions can be verified (but not necessarily solved) in polynomial time. This is a less standard definition from [4] but is simpler than and equivalent to the more commonly used definition of [13]. The verification process makes use of outside information, called a *certificate* which is not in general available to the algorithm computing the solution.

A well-known problem in NP is the Travelling Salesman Problem (TSP). The salesman seeks a path which visits each of n cities exactly once and which ends in the city in which he started. The cost of travelling from city i to city j is an integer $c(i, j)$. An instance of the TSP consists of a binary encoding of n, a threshold C, and the set of costs $c(i, j)$. The problem is to determine whether there exists a path of total cost less than C. The TSP is in NP because the list of cities in any path of total cost less than C that passes through each city exactly once serves as a certificate. That is, we can verify the existence of a tour of the requisite cost in polynomial time if we are given the list of cities which comprise the tour.

We can compare the relative difficulty of problems in NP through the use of *polynomial-time reducibility.* We say that a problem Q_1 is polynomial-time reducible to the problem Q_2, and write $Q_1 \leq_P Q_2$, if we can solve Q_1 by first mapping each instance of Q_1 to an instance of Q_2 in polynomial time, and then solving Q_2. The problem Q_2 is at least as hard as Q_1 (up to a polynomial amount of work) since any algorithm which solves Q_2 can also be used to solve Q_1.

A particularly important subset of NP is the set of *NP-complete* problems, the set $\{Q : Q \in \text{NP and } Q' \leq_P Q \text{ for all } Q' \in \text{NP}\}$. The NP-complete problems are the most difficult problems in NP with respect to our polynomial-time reducibility relation. The Travelling Salesman Problem is an NP-complete problem. It is widely believed, though unproven, that there exist problems in NP which cannot be solved by polynomial time algorithms. If indeed such a set of intractable problems exists, the set of NP-complete problems is contained within it.

We prove below that deciding whether an (ϵ, M)-approximation exists is an NP-complete problem. We further show that the problem of finding an M-optimal approximation is an *NP-hard* problem. NP-hard problems extend the notion of NP-completeness beyond the class of decision problems and consist of the problems $\{Q : Q' \leq_P Q$ for all $Q' \in \mathrm{NP}\}$.

2.2 Complexity of Optimal Approximation

Theorem 2.1 *Let $\mathcal{H}$ be an N-dimensional Hilbert space. Let $\mathcal{D}_N$ be the set of all dictionaries for $\mathcal{H}$ that contain $O(N^k)$ vectors, where $k \geq 1$. Let $0 < \alpha_1 < \alpha_2 < 1$ and $M \leq N$ such that $\alpha_1 N \leq M \leq \alpha_2 N$. The finite-input* **($\epsilon$, M)-approximation problem**, *determining for any given $\epsilon > 0, \mathcal{D} \in \mathcal{D}_N$, and $f \in \mathcal{H}$, whether an (ϵ, M)-approximation exists, is NP-complete. The finite-input* **M-optimal approximation problem**, *finding the optimal M-approximation, is NP-hard.*

We note that we must alter the approximation problems slightly in order to make use of the theory of NP-completeness as described. By the *finite-input* versions of the problems, we mean that f, the elements of the dictionaries, and their coefficients are restricted to binary representations of $\Theta(N^m)$ bits, for some m. This restriction does not affect the proof, because the problems that must be solved for the proof are discrete and unaffected by small perturbations. The theory of NP-completeness over the reals is described in [22].

We emphasize that the theorem does not imply that the M-approximation problem is intractible for *specific* dictionaries $\mathcal{D} \in \mathcal{D}_N$. Indeed, we saw above that for orthonormal dictionaries, the problem can be solved in polynomial time. Rather, the result is that if we have an algorithm which finds the optimal approximation to any given $f \in \mathbf{R}^N$ for *any* dictionary $\mathcal{D} \in \mathcal{D}_N$, the algorithm solves an NP-hard problem.

Proof: For any ϵ we can solve the (ϵ, M)-approximation problem by first solving the M-optimal approximation problem, computing $\epsilon_{min} = \|\tilde{f} - f\|$, and then checking whether $\epsilon_{min} < \epsilon$. Hence the M-optimal approximation problem must be at least as hard as the (ϵ, M)-approximation problem. Proving that the (ϵ, M)-approximation problem is NP-complete thus implies that the M-optimal approximation problem is NP-hard.

The (ϵ, M)-approximation problem is in NP, because we can verify in polynomial time that $\|\tilde{f} - f\| < \epsilon$ once we are given the certificate consisting of the set of M elements and the coefficients used to construct $\tilde{f}$. To prove that the problem is NP-complete we prove that a known NP-complete problem, the exact cover by 3-sets problem, is polynomial-time reducible to our approximation problem. In other words, we prove that our approximation problem is at least as hard as a problem which is known to be NP-complete.

Definition 2.2 *Let X be a set containing $N = 3M$ elements, and let $\mathcal{C}$ be a collection of 3-element subsets of X. The* **exact cover by 3-sets** *problem is to decide whether $\mathcal{C}$ contains an exact cover for X, i.e. to determine whether $\mathcal{C}$ contains a subcollection $\mathcal{C}'$ such that every member of X occurs in exactly one member of $\mathcal{C}'$ [13].*

Lemma 2.1 *We can transform in polynomial time any instance $(X, \mathcal{C})$ of the exact cover by 3-sets problem of size $|X| = 3M$ into an equivalent instance of the (ϵ, M)-approximation problem with a dictionary of size $O(N^3)$ in an N-dimensional Hilbert space.*

This lemma implies that if we can solve the (ϵ, M)-approximation problem for $M = N/3$, we can also solve an NP-complete problem. Therefore the approximation problem must be NP-complete as well. We thus obtain a proof of the theorem for the case $M = \frac{N}{3}$, i.e. $\alpha_1 = \alpha_2 = \frac{1}{3}$, and $k = 3$.

Proof: Let $\mathcal{H}$ be an N dimensional space with an orthonormal basis $\{e_i\}_{1 \leq i \leq N}$. For notational convenience we take the set X to be the set of $N = 3M$ integers between 1 and N. Let $\mathcal{C}$ be a collection of 3-element subsets of X. To each subset $S \subseteq X$ we associate a unit vector in $\mathcal{H}$ given by

$$T(S) = \frac{1}{\sqrt{|S|}} \sum_{i \in S} e_i, \tag{2}$$

where $|S|$ is the cardinality of the set S. Let $\mathcal{D}$ be the dictionary for $\mathcal{H}$ defined by

$$\mathcal{D} = \{T(S_i) : S_i \in \mathcal{C}\}, \tag{3}$$

where the S_i's are the three-element subsets of X contained in $\mathcal{C}$. Since $\mathcal{C}$ contains at most $\begin{pmatrix} N \\ 3 \end{pmatrix} = O(N^3)$ three-element subsets of X, this transformation can be done in polynomial time.

We now show that solving the (ϵ, M)-approximation problem for

$$f = T(X) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} e_i \tag{4}$$

and $\epsilon < \frac{1}{\sqrt{N}}$ is equivalent to solving the exact cover by 3-sets problem $(X, \mathcal{C})$. Suppose $\mathcal{C}$ contains an exact cover $\mathcal{C}'$ for X. Then

$$\|f - \sqrt{\frac{3}{N}} \sum_{S_i \in \mathcal{C}'} T(S_i)\| = 0. \tag{5}$$

Since there are $M = \frac{1}{3}N$ such S_i's, the approximation problem has a solution. Thus, a solution to the exact cover problem implies a solution to the approximation problem.

Conversely, suppose the (ϵ, M)-approximation problem has a solution for $\epsilon < \frac{1}{\sqrt{N}}$. There exist M 3-element sets $S_i \in \mathcal{C}$ and M coefficients β_n such that

$$\|f - \sum_{n=1}^{M} \beta_n T(S_n)\| < \frac{1}{\sqrt{N}}.$$

The inner product of each basis vector $\{e_i\}_{1 \leq i \leq N}$ with $\sum_{n=1}^{M} \beta_n T(S_n)$ must be non-zero, for otherwise we would have $\| \sum_{i=1}^{n} \beta_i T(S_i) - f\| \geq \frac{1}{\sqrt{N}}$ (recall that all components of f are equal to $\frac{1}{\sqrt{N}}$). Since each $T(S_i)$ has non-zero inner products with exactly three basis vectors and $N = 3M$, the M sets $\{S_i\}_{1\leq i\leq M}$ do not intersect and thus define an exact 3-set cover of X. This proves that a solution to the approximation problem implies a solution to the exact cover problem, which finishes the proof of the lemma.

□

We have proved that the (ϵ, M)-approximation problem is NP-complete for $\alpha_1 = \alpha_2 = \frac{1}{3}$ and dictionaries of size $O(N^3)$. We can extend the result to arbitrary $0 < \alpha_1 < \alpha_2 < 1$ and dictionaries of size $O(N^k)$ for $k > 1$. Let $(X, \mathcal{C})$ be an instance of the exact cover by 3-sets problem where X is a set of n elements. Following lemma 2.1, we can construct an equivalent (ϵ, M)-approximation problem on a $3n$-dimensional space $\mathcal{H}_1$. We then embed this approximation problem in a larger Hilbert space $\mathcal{H}$ in order to satisfy the dictionary size and expansion length constraints. In $\mathcal{H}_2$, the orthogonal complement of $\mathcal{H}_1$ in $\mathcal{H}$, we construct a $(0, \alpha_2 N - n)$-approximation problem which has a unique solution. The combined approximation problem in $\mathcal{H}$ will be equivalent to the exact cover problem and will have the requisite M and dictionary size [7].

The optimal approximation criterion of definition 2.1 has number of undesirable properties which are responsible for its NP-completeness. The elements contained in the expansions are unstable in that functions which are only slightly different can have optimal expansions containing completely different dictionary elements. The expansions also lack an optimal substructure property, i.e. the expansion in M elements with minimal error does not necessarily contain an expansion in $M-1$ elements with minimal error. The expansions therefore cannot be progressively refined. Finally, depending upon the dictionary, the coefficients of optimal approximations can exhibit instability in that the expansion coefficients β_i of the M-optimal approximation (1) to a vector f can have

$$\sum_{i=1}^{M} |\beta_i|^2 >> ||f||^2.$$

Consider the case when $\mathcal{H} = \mathbf{R}^3$, $f = (1,1,1)$, and $\mathcal{D} = \{e_1, e_2, e_3, v\}$, where the e_i's are the Euclidean basis of $\mathbf{R}^3$ and $v = \frac{e_1+\epsilon f}{||e_1+\epsilon f||}$. The M-optimal approximation to f for $M = 2$ is

$$\tilde{f} = \frac{||e_1 + \epsilon f||}{\epsilon} v - \frac{1}{\epsilon} e_1, \tag{6}$$

so we see that $\sum_{i=1}^{M} |\beta_i|^2$ can be made arbitrarily large. In the next section we describe an approximation algorithm, based on a greedy refinement of the vector approximation, which maintains an energy conservation relation that guarantees stability.

3 Matching Pursuits

A matching pursuit is a greedy algorithm which, rather than solving the optimal approximation problem, instead progressively refines the signal approximation with an iterative procedure

[24]. We describe this algorithm in the next section and present an orthogonalized version in section 3.2. Section 3.3 describes a fast numerical implementation, and section 3.4 describes an application to computing adaptive time-frequency decompositions.

3.1 Non-Orthogonal Matching Pursuits

Let $\mathcal{D} = \{g_\gamma\}_{\gamma \in \mathbf{\Gamma}}$ be a dictionary of vectors with unit norm in a Hilbert space $\mathcal{H}$. Let $f \in \mathcal{H}$. The first step of a matching pursuit is to approximate f by projecting it on a vector $g_{\gamma_0} \in \mathcal{D}$

$$f = < f, g_{\gamma_0} > g_{\gamma_0} + Rf. \tag{7}$$

Since the residual Rf is orthogonal to g_{γ_0},

$$\|f\|^2 = |< f, g_{\gamma_0} >|^2 + \|Rf\|^2. \tag{8}$$

We minimize the norm of the residual by choosing the $g_{\gamma_0} \in \mathcal{D}$ that maximizes $|< f, g_\gamma >|$. In infinite dimensions, the supremum of $|< f, g_\gamma >|$ may not be attained, so we relax our selection criterion. We choose g_{γ_0} such that

$$|< f, g_{\gamma_0} >| \geq \alpha \sup_{\gamma \in \Gamma} |< f, g_\gamma >|, \tag{9}$$

where $\alpha \in (0, 1]$ is an optimality factor. The vector g_{γ_0} is chosen from the set of dictionary vectors that satisfy (9), with a choice function whose properties vary depending upon the application. The use of the optimality factor α is discussed further in [17].

The pursuit iterates this procedure by subdecomposing the residual. Let $R^0 f = f$. Suppose that we have already computed the residual $R^k f$. We choose $g_{\gamma_k} \in \mathcal{D}$ such that

$$|< R^k f, g_{\gamma_k} >| \geq \alpha \sup_{\gamma \in \mathbf{\Gamma}} |< R^k f, g_\gamma >| \tag{10}$$

and project $R^k f$ on g_{γ_k}

$$R^{k+1} f = R^k f - < R^k f, g_{\gamma_k} > g_{\gamma_k}. \tag{11}$$

The orthogonality of $R^{k+1} f$ and g_{γ_k} implies that

$$\|R^{k+1} f\|^2 = \|R^k f\|^2 - |< R^k f, g_{\gamma_k} >|^2. \tag{12}$$

By summing (11) for k between 0 and $n-1$ we obtain

$$f = \sum_{k=0}^{n-1} < R^k f, g_{\gamma_k} > g_{\gamma_k} + R^n f. \tag{13}$$

Similary, summing (12) for k between 0 and $n-1$ yields

$$\|f\|^2 = \sum_{k=0}^{n-1} |< R^k f, g_{\gamma_k} >|^2 + \|R^n f\|^2. \tag{14}$$

The residual $R^n f$ is the approximation error of f after choosing n vectors in the dictionary and the energy of this error is given by (14). For any $f \in \mathcal{H}$, the convergence of the error to zero is shown in [24] to be a consequence of a theorem proved by Jones [15], i.e.

$$\lim_{n\to\infty} ||R^n f|| = 0. \tag{15}$$

Hence

$$f = \sum_{k=0}^{\infty} < R^k f, g_{\gamma_k} > g_{\gamma_k}, \tag{16}$$

and

$$||f||^2 = \sum_{k=0}^{\infty} |< R^k f, g_{\gamma_k} >|^2. \tag{17}$$

In infinite dimensions, the convergence rate of this error can be extremely slow. In finite dimensions, we now prove that the convergence is exponential. For any vector $e \in \mathcal{H}$, we define

$$\lambda(e) = \sup_{\gamma\in\mathbf{\Gamma}} |< \frac{e}{||e||}, g_\gamma >|.$$

We will take the optimality factor α to be 1 for finite dimensional spaces unless otherwise specified. Hence, the chosen vector g_{γ_k} satisfies

$$\frac{|< R^k f, g_{\gamma_k} >|}{||R^k f||} = \lambda(R^k f).$$

Equation (12) thus implies that

$$||R^{k+1} f||^2 = ||R^k f||^2 (1 - \lambda^2(R^k f)). \tag{18}$$

Hence norm of the residual decays exponentially with a rate equal to $-\frac{1}{2}\log(1 - \lambda^2(R^k f))$. Since $\mathcal{D}$ contains at least a basis of $\mathcal{H}$ and the unit sphere of $\mathcal{H}$ is compact in finite dimensions, we can derive [24] that there exists $\lambda_{min} > 0$ such that for any $e \in \mathcal{H}$

$$\lambda(e) \geq \lambda_{min}. \tag{19}$$

Equation (18) thus proves that the energy of the residual $R^k f$ decreases exponentially with a minimum decay rate equal to $-\frac{1}{2}\log(1 - \lambda^2_{min})$.

3.2 Orthogonal Matching Pursuits

The approximations derived from a matching pursuit can be refined by orthogonalizing the directions of projection. The resulting orthogonal pursuit converges with a finite number of iterations in finite dimensional spaces, which is not the case for a non-orthogonal pursuit. Equivalent algorithms have been introduced independently in [6], [18], and [1].

The vector g_{γ_k} selected at each iteration by the matching pursuit algorithm is not in general orthogonal to the previously selected vectors $\{g_{\gamma_p}\}_{0\leq p<k}$. In subtracting the projection of $R^k f$

onto g_{γ_k} the algorithm reintroduces new components in the directions of the $\{g_{\gamma_p}\}_{0\leq p<k}$. This can be avoided by orthogonalizing the $\{g_{\gamma_p}\}_{0\leq p<k}$ with a Gram-Schmidt procedure. Let $u_0 = g_{\gamma_0}$. As in a matching pursuit, we choose g_{γ_k} that satisfies (10). This vector is orthogonalized with respect to the previously selected vectors by computing

$$u_k = g_{\gamma_k} - \sum_{p=0}^{k-1} \frac{< g_{\gamma_k}, u_p >}{||u_p||^2} u_p. \tag{20}$$

The residual is then defined by

$$R^{k+1}f = R^k f - \frac{< R^k f, u_k >}{||u_k||^2} u_k. \tag{21}$$

The vector $R^k f$ is the orthogonal projection of f onto the orthogonal complement to the space generated by the vectors $\{g_{\gamma_p}\}_{0\leq p<k}$. Equation (20) implies that $< R^k f, u_k > = < R^k f, g_{\gamma_k} >$ and thus

$$R^{k+1}f = R^k f - \frac{< R^k f, g_{\gamma_k} >}{||u_k||^2} u_k. \tag{22}$$

Since $R^{k+1}f$ and u_k are orthogonal,

$$||R^{k+1}f||^2 = ||R^k f||^2 - \frac{| < R^k f, g_{\gamma_k} > |^2}{||u_k||^2}. \tag{23}$$

If $R^k f \neq 0$, $< R^k f, g_{\gamma_k} > \neq 0$, and since $R^k f$ is orthogonal to all previously selected vectors the selected vectors $\{g_{\gamma_p}\}_{0\leq p<k}$ are linearly independent. Since $R^0 f = f$, from equations (22) and (23), we derive in a manner similar to that used for (13) and (14), that for any $n > 0$

$$f = \sum_{k=0}^{n-1} \frac{< R^k f, g_{\gamma_k} >}{||u_k||^2} u_k + R^n f, \tag{24}$$

and

$$\|f\|^2 = \sum_{k=0}^{n-1} \frac{| < R^k f, g_{\gamma_k} > |^2}{||u_k||^2} + \|R^n f\|^2. \tag{25}$$

The theorem below shows that the residuals of an orthogonal pursuit converge strongly to zero and that the number of iterations required for convergence is less than or equal to the dimension of the space $\mathcal{H}$. Thus in finite dimensional spaces, orthogonal matching pursuits are guaranteed to converge in a finite number of steps, unlike non-orthogonal pursuits.

Theorem 3.1 *Let $\mathcal{H}$ be an N-dimensional Hilbert space (N may be infinite), and let $f \in \mathcal{H}$. An orthogonal pursuit converges in less than or equal to N iterations. The residual $R^n f$ defined in (22) satisfies $||R^N f|| = 0$ when N is finite, or*

$$\lim_{n\to\infty} ||R^n f|| = 0 \tag{26}$$

when N is infinite. Hence

$$f = \sum_{n=0}^{N-1} \frac{< R^n f, g_{\gamma_n} >}{||u_k||^2} u_n \tag{27}$$

and

$$||f||^2 = \sum_{n=0}^{N-1} \frac{|< R^n f, g_{\gamma_n} >|^2}{||u_k||^2}. \tag{28}$$

Proof: We prove the result for the case $N = \infty$; the finite case is trivial. We have from the Bessel inequality that

$$\sum_{k=0}^{\infty} \frac{|< f, u_k >|^2}{||u_k||^2} \leq ||f||^2, \tag{29}$$

so we must have

$$\lim_{k \to \infty} |< f, u_k >| = \lim_{k \to \infty} |< R^n f, g_{\gamma_n} >| = 0. \tag{30}$$

By (9), we must have

$$\lim_{k \to \infty} \sup_{\gamma \in \mathbf{\Gamma}} |< R^k f, g_\gamma >| = 0, \tag{31}$$

so $R^k f$ converges weakly to 0. To show strong convergence, we compute for $n < m$ the difference

$$\begin{aligned} ||R^n f - R^m f||^2 &= \sum_{k=n+1}^{m} \frac{|< f, u_k >|^2}{||u_k||^2} \\ &\leq \sum_{k=n+1}^{\infty} \frac{|< f, u_k >|^2}{||u_k||^2}, \end{aligned} \tag{32}$$

which goes to zero as n goes to infinity since the sum is bounded. The Cauchy criterion is satisfied, so $R^n f$ converges strongly to its weak limit of 0, thus proving the result.

□

The orthogonal pursuit yields a function expansion over an orthogonal family of vectors $\{u_k\}_{0 \leq p < n}$. To obtain an expansion of f over $\{g_{\gamma_n}\}_{0 \leq k < N}$ we must make a change of basis. The Gram-Schmidt vector u_k can be expanded within $\{g_{\gamma_p}\}_{0 \leq p \leq k}$

$$u_k = \sum_{p=0}^{k} b_{p,k} g_{\gamma_p}. \tag{33}$$

Inserting this expression into (27) yields

$$f = \sum_{n=0}^{M} \frac{< R^n f, g_{\gamma_n} >}{||u_n||^2} \sum_{p=0}^{n} b_{p,n} g_{\gamma_p}. \tag{34}$$

In the infinite dimensional case, without absolute convergence of the above infinite series, we cannot rearrange the terms of this double summation to obtain

$$f = \sum_{0 \leq p < M} g_{\gamma_p} \sum_{p \leq n < M} b_{p,n} \frac{< R^n f, g_{\gamma_n} >}{||u_n||^2}. \tag{35}$$

The second summation that defines the expansion coefficients over the family $\{g_{\gamma_p}\}_{0 \leq p < M}$ can indeed diverge. This happens when the family of selected elements is not a Riesz basis of the closed space it generates. In [7] it is proved that it is indeed possible for the set $\{g_{\gamma_k}\}$ of vectors selected by an orthogonal pursuit to be degenerate, even when the dictionary contains an orthonormal basis.

The residuals of orthogonal matching pursuits in general decrease faster than the non-orthogonal matching pursuits. However, this orthogonal procedure can yield unstable expansions by selecting ill-conditioned family of vectors. Orthogonal pursuits also require many more operations to compute than non-orthogonal pursuits because of the Gram-Schmidt orthogonalization procedure. In the next section we compare the complexity of these two algorithms.

3.3 Implementation of Matching Pursuits

We consider the case for which $\mathcal{H}$ is a finite dimensional space and $\mathcal{D}$ is a dictionary with a finite number of vectors. The optimality factor α is set to 1.

The matching pursuit is initialized by computing the inner products $\{< f, g_\gamma >\}_{\gamma \in \mathbf{\Gamma}}$. These inner products are stored in an open hash table [4], where they are partially sorted. The algorithm is defined by induction as follows. Suppose that we have already computed $\{< R^n f, g_\gamma >\}_{\gamma \in \mathbf{\Gamma}}$, for $n \geq 0$. We must first find g_{γ_n} such that

$$|< R^n f, g_{\gamma_n} >| = \sup_{\gamma \in \mathbf{\Gamma}} |< R^n f, g_\gamma >|. \tag{36}$$

Since all inner products are stored in an open hash table, this requires $O(1)$ operations on average. Once g_{γ_n} is selected, we compute the inner product of the new residual $R^{n+1} f$ with all $g_\gamma \in \mathcal{D}$ using an updating formula derived from equation (11)

$$< R^{n+1} f, g_\gamma > = < R^n f, g_\gamma > - < R^n f, g_{\gamma_n} > < g_{\gamma_n}, g_\gamma > . \tag{37}$$

Since we have already computed $< R^n f, g_\gamma >$ and $< R^n f, g_{\gamma_n} >$, this update requires only that we compute $< g_{\gamma_n}, g_\gamma >$. Dictionaries are generally built so that few such inner products are non-zero, and non-zero inner products are either precomputed and stored or computed with a small number of operations. Suppose that the inner product of any two dictionary elements can be obtained with $O(I)$ operations and that there are $O(Z)$ non-zero inner products. Computing the products $\{< R^{n+1} f, g_\gamma >\}_{\gamma \in \mathbf{\Gamma}}$ and storing them in the hash table thus requires $O(IZ)$ operations. The total complexity of P matching pursuit iterations is thus $O(PIZ)$.

The initialization and selection portions of the orthogonal matching pursuit algorithm are implemented in the same way as they are for the non-orthogonal algorithm. The difference between the two algorithms is in the updating of the inner products $< R^n f, g_\gamma >$ after a vector has been selected. Once the vector g_{γ_n} is selected, we must compute the expansion coefficients of the orthogonal vector u_n

$$u_n = \sum_{p=0}^{n} b_{p,n} g_{\gamma_p} \tag{38}$$

with the Gram-Schmidt formula (20). For $p < n$, we suppose that we already computed the expansion coefficients u_p over the family $\{g_{\gamma_k}\}_{0 \leq k \leq p}$. If the inner products of any two elements in $\mathcal{D}$ is calculated in $O(I)$ operations the expansion coefficient $\{b_{p,n}\}_{0 \leq p \leq n}$ are obtained in $O(nI + n^2)$ operations. We use the Gram-Schmidt formula for computational simplicity, although it has poor numerical properties,

We compute the inner product of the new residual $R^{n+1}f$ with all $g_\gamma \in \mathcal{D}$ using the orthogonal updating formula (22)

$$< R^{n+1} f, g_\gamma > = < R^n f, g_\gamma > - \frac{< R^n f, g_{\gamma_n} > < u_n, g_\gamma >}{||u_n||^2}. \tag{39}$$

Since

$$< u_n, g_\gamma > = \sum_{p=0}^{n} b_{p,n} < g_{\gamma_p}, g_\gamma >, \tag{40}$$

computing $\{< R^{n+1} f, g_\gamma >\}_{\gamma \in \mathbf{\Gamma}}$ requires $O(nIZ)$ operations. The total number of operations to compute P orthogonal matching pursuit iterations is therefore $O(P^3 + P^2 IZ)$. For P iterations, the non-orthogonal pursuit algorithm is P times faster than the orthogonal one. When P is large, which is the case in many signal processing applications, the orthogonal pursuits give much better approximations, but the amount of computation required is prohibitive. For small P, non-orthogonal and orthogonal pursuits give roughly equivalent approximations. We make a detailed comparison of the performance of these algorithms in section 7.

3.4 Application to Dictionaries of Time-Frequency Atoms

Signals such as sound recordings contain structures that are well localized both in time and frequency. This localization varies according to the sound source, which makes it difficult to find a basis that is *a priori* well adapted to all components of a particular sound signal. Dictionaries of time-frequency atoms include waveforms with a wide range of time-frequency localizations and are thus much larger than a single basis. Such dictionaries are generated by translating, modulating, and scaling a single real window function $g(t) \in L^2(\mathbf{R})$. We suppose that that $g(t)$ is even, $||g|| = 1$, $\int g(t)dt \neq 0$, and $g(0) \neq 0$. We denote $\gamma = (s, u, \xi)$ and

$$g_\gamma(t) = \frac{1}{\sqrt{s}} g(\frac{t-u}{s}) e^{i\xi t}. \tag{41}$$

The time-frequency atom $g_\gamma(t)$ is centered at $t = u$ with a support proportional to s. Its Fourier transform is

$$\hat{g}_\gamma(\omega) = \sqrt{s}\hat{g}(s(\omega - \xi)) e^{-i(\omega - \xi)u}, \tag{42}$$

and $\hat{g}_\gamma(\omega)$ is centered at $\omega = \xi$ and concentrated over a domain proportional to $\frac{1}{s}$. For small values of s the atoms are well localized in time but poorly localized in frequency; for large values of s the atoms are well localized in space but poorly localized in time.

The dictionary of time-frequency atoms $\mathcal{D} = \{g_\gamma(t)\}_{\gamma \in \Gamma}$ is a very redundant set of functions that includes both window Fourier frames and wavelet frames [5]. When the window function g is the Gaussian $g(t) = 2^{1/4}e^{-\pi t^2}$, the resulting time-frequency atoms are Gabor functions, and have optimal localization in time and in frequency. A matching pursuit decomposes any $f \in \mathbf{L}^2(\mathbf{R})$ into the sum

$$f = \sum_{n=0}^{+\infty} < R^n f, g_{\gamma_n} > g_{\gamma_n}, \tag{43}$$

where the scales, position and frequency $\gamma_n = (s_n, u_n, \xi_n)$ of each atom

$$g_{\gamma_n}(t) = \frac{1}{\sqrt{s}_n} g(\frac{t - u_n}{s_n}) e^{i\xi_n t} \tag{44}$$

are chosen to best match the structures of f. This procedure efficiently approximates any signal structure that is well-localized in the time-frequency plane, regardless of whether its localization is in time or in frequency.

To any matching pursuit expansion[17][19], we can associate a time-frequency energy distribution defined by

$$Ef(t,\omega) = \sum_{n=0}^{\infty} | < R^n f, g_{\gamma_n} > |^2 W g_{\gamma_n}(t,\omega), \tag{45}$$

where

$$W g_{\gamma_n}(t,\omega) = 2 \exp\left[-2\pi\left(\frac{(t-u)^2}{s^2} + s^2(\omega - \xi)^2\right)\right],$$

is the Wigner distribution [3] of the Gabor atom g_{γ_n}. Its energy is concentrated in the time and frequency domains where g_{γ_n} is localized. Figure 3.4 shows $Ef(t,\omega)$ for the signal f of 512 samples displayed in Fig. 3.4. This signal is built by adding waveforms of different time-frequency localizations. It is the sum of $\cos((1 - \cos(ax))bx)$, two truncated sinusoids, two Dirac functions, and $\cos(cx)$.

The signal is decomposed into a dictionary consisting of a discretized version of the Gabor dictionary described above. The translation and modulation parameters are given in units proportional to the sample spacing, and the scaling is restricted to powers of 2. The size of this discretized dictionary is $N \log_2(N)$ for an N-sample dictionary.

Each Gabor time-frequency atom selected by the matching pursuit corresponds to a dark elongated Gaussian blob in the time-frequency plane. The arch of the the $\cos((1 - cos(ax))bx)$ is decomposed into a sum of atoms that covers its time-frequency support. The truncated sinusoids are in the center and upper left-hand corner of the plane. The middle horizontal dark line is an atom well localized frequency that corresponds to the component $cos(cx)$ of the signal. The two vertical dark lines are atoms very well localized in time that correspond to the two Diracs.

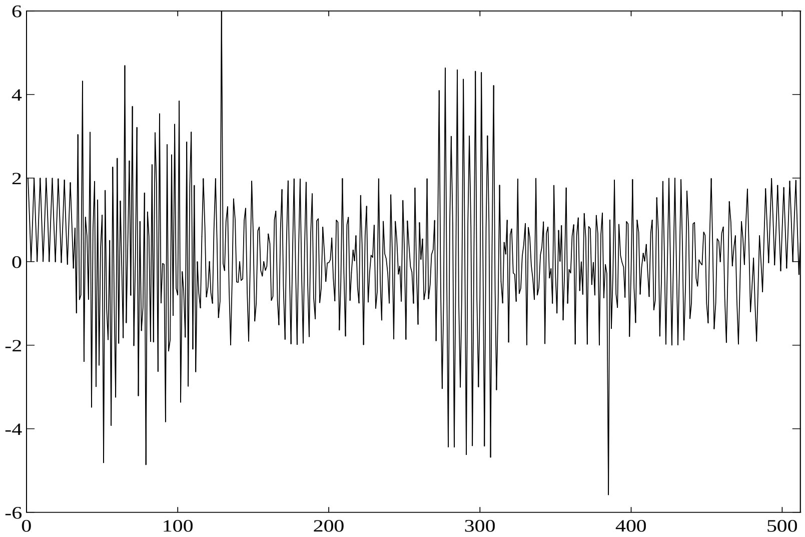

Figure 1: Synthetic signal of 512 samples built by adding $\cos((1-\cos(ax))bx)$, two truncated sinusoids, two Dirac functions, and $\cos(cx)$.

Software implementing matching pursuits for time-frequency dictionaries is available through anonymous ftp at the address cs.nyu.edu . Instructions are in the file README of the directory /pub/wave/software.

4 Group Invariant Dictionaries

The translation, dilation, and frequency modulation of any vector from the Gabor dictionary is also contained in the Gabor dictionary. This invariance under the action of any of the operators that belong to the group of translations, dilations or frequency modulations implies important invariance properties of matching pursuits with the Gabor dictionary. In this section we examine the properties of pursuits when the dictionary is left invariant by a given group of unitary linear operators $\mathcal{G} = \{G_\tau\}_{\tau\in\Omega}$ that is a representation of the group Ω. Since each operator G_τ is unitary, its inverse and adjoint is $G_{\tau^{-1}}$, where τ^{-1} is the inverse of τ in Ω. For example, the unitary groups of translation, frequency modulation and dilation over $\mathcal{H} = \mathbf{L}^2(\mathbf{R})$ are defined respectively by $G_\tau f(t) = f(t-\tau)$, $G_\tau f(t) = e^{i\tau t} f(t)$ and $G_\tau f(t) = \frac{1}{\sqrt{s^\tau}} f(\frac{t}{s^\tau})$.

Definition 4.1 *A dictionary $\mathcal{D} = \{g_\gamma\}_{\gamma\in\mathbf{\Gamma}}$ is invariant with respect to the group of unitary operators $\mathcal{G} = \{G_\tau\}_{\tau\in\Omega}$ if and only if for every $g_\gamma \in \mathcal{D}$ and $\tau \in \Omega$ we have $e^{i\phi} G_\tau g_\gamma \in \mathcal{D}$ for a unique $\phi \in [0, 2\pi)$.*

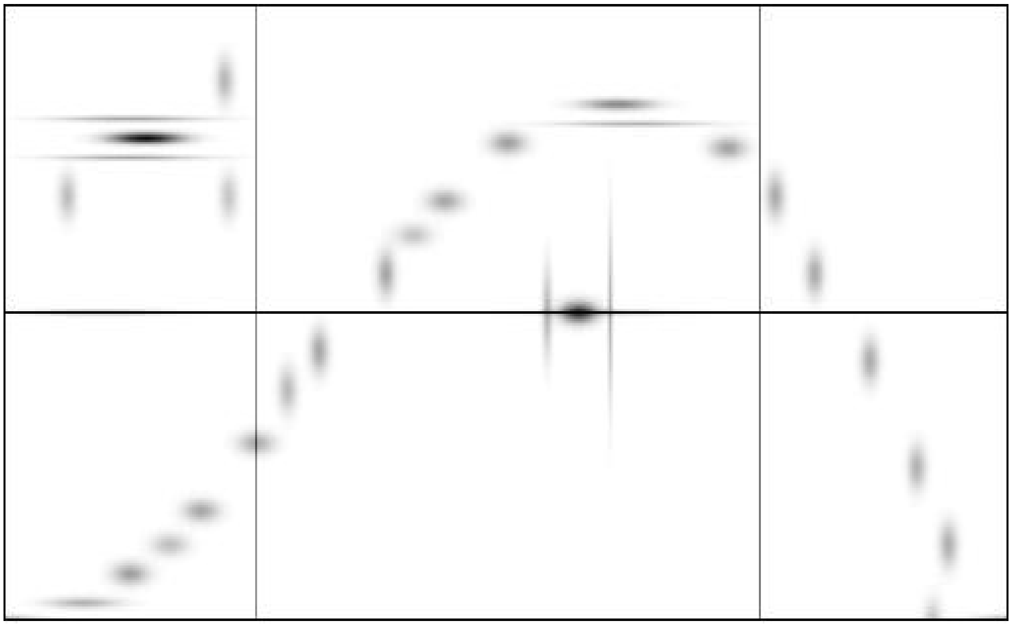

Figure 2: Time frequency energy distribution $Ef(t, \omega)$ of the signal in the figure above. The horizontal axis is time and the vertical axis is frequency. The darkness of the image increases with $Ef(t, \omega)$.

The Gabor dictionary is invariant under the group generated by translations, modulations and dilations. The properties of the corresponding matching pursuit depends upon the choice function C that chooses for any $f \in \mathcal{H}$ an element $g_{\gamma_0} = C(\mathbf{E}[f])$ from the set

$$\mathbf{E}[f] = \{g \in \mathcal{D} : |<f, g>| \geq \alpha \sup_{g_\gamma \in \mathcal{D}} |<f, g_\gamma>|\}$$

onto which f is then projected. The following proposition imposes a commutatitivity condition on the choice function C so that the matching pursuit commutes with the group operators.

Proposition 4.1 *Let $\mathcal{D}$ be invariant with respect to the group of unitary operators $\mathcal{G} = \{G_\tau\}_{\tau \in \Omega}$. Let $f \in \mathcal{H}$ and*

$$f = \sum_{k=0}^{n-1} a_n g_{\gamma_n} + R^n f$$

be its matching pursuit computed with the choice function C. If for any $n \in \mathbf{N}$ we have

$$CG_\tau \mathbf{E}[R^n f] = e^{i\phi} G_\tau C \mathbf{E}[R^n f], \tag{46}$$

then the matching pursuit decomposition of $G_\tau f$ is

$$G_\tau f = \sum_{k=0}^{n-1} a_n e^{i\phi_n} G_\tau g_{\gamma_n} + G_\tau R^n f, \tag{47}$$

where the phases ϕ_n are determined so that $e^{i\phi_n} G_\tau g_{\gamma_n} \in \mathcal{D}$.

The condition (46) means that the element chosen from the $\mathbf{E}[R^n f]$ transformed by G_τ is the transformation by G_τ of the element chosen from $\mathbf{E}[R^n f]$ up to a complex phase. Equation (47) implies that the vectors selected by the matching pursuit for $G_\tau f$ are, up to a complex phase, the vectors selected for the matching pursuit of f transformed by G_τ, and the residuals of $G_\tau f$ are equal to the residuals of f transformed by G_τ.

Proof: Since the group is unitary, for any $g_\gamma \in \mathcal{D}$

$$< G_\tau f, g_\gamma > = < f, G_{\tau^{-1}} g_\gamma > .$$

Hence $g_\gamma \in \mathbf{E}[G_\tau f]$ if and only if $e^{i\phi} G_{\tau^{-1}} g_\gamma \in \mathbf{E}[f]$ for some ϕ, which proves that $\mathbf{E}[G_\tau f] = G_\tau \mathbf{E}[f]$ up to a complex phase. By using the commutativity (46) of the choice function with respect to G_τ we then easily prove (47) by induction.

□

We must now prove that choice functions exist that satisfy the commutativity relation (46) for all $f \in \mathcal{H}$ or at least for almost all $f \in \mathcal{H}$. The following proposition gives a necessary condition for constructing such choice functions.

Proposition 4.2 *Let $\mathbf{K}$ be set of functions $f \in \mathcal{H}$ such that there exists a $G_\tau \neq I$ for which*

$$\mathbf{E}[f] = \mathbf{E}[G_\tau f] \tag{48}$$

up to a complex phase. There exists a choice function C such that for any $f \in \mathcal{H} - \mathbf{K}$ and $G_\tau \in \mathcal{G}$

$$C G_\tau \mathbf{E}[f] = e^{i\phi} G_\tau C \mathbf{E}[f]. \tag{49}$$

Proof: We construct this choice function from the equivalence classes of an equivalence relation on the sets $E[f]$ for $f \in \mathcal{H} - \mathbf{K}$. We define R on $\mathcal{H} - \mathbf{K}$ such that $E[f]\ R\ E[h]$ if and only if there exists a $G_\tau \in \mathcal{G}$ such that $v \in E[h]$ if and only if $e^{i\phi} G_\tau v \in E[f]$ for some ϕ. Note that because we have excluded the set $\mathbf{K}$ from the domain of R this G_τ is unique.

For all $f \in \mathcal{H} - \mathbf{K}$ the set $E[f]$ belongs to exactly one R equivalence class. By the axiom of choice, we can select a representative set $E[h]$ from each equivalence class. Let S be the set of these representatives. Again, by the axiom of choice, we can define a choice function C on S by selecting from each set $E[h] \in S$ a single element $C(E[h])$.

We now extend C to all $\mathcal{H} - \mathbf{K}$. For any $f \in \mathcal{H} - \mathbf{K}$, $E[f]$ is contained in some R equivalence class. Hence there exists an $h \in S$ and a unique G_τ such that $v \in E[h]$ if and only if $e^{i\phi} G_\tau v \in E[f]$. We set $C(E[f]) = e^{i\phi} G_\tau C(E[h])$ where ϕ is determined so that $C(E[f]) \in \mathcal{D}$.

From proposition 4.1, $E[f]$ and $E[G_\tau f]$ belong to the same R equivalence class for all $f \in \mathcal{H} - \mathbf{K}$, and $E[G_\tau f] = G_\tau E[f]$ up to a complex phase. Hence there exists an $h \in S$ and a unique G_ρ and G_σ such that up to a complex phase $E[f] = E[G_\rho h] = G_\rho E[h]$ and $E[G_\tau f] = E[G_\sigma h] = G_\sigma E[h]$. This implies that $E[h] = E[G_{\sigma^{-1}} G_\tau G_\rho h]$, and because h is not in $\mathbf{K}$, we must have $G_\sigma = G_\tau G_\rho$. By our construction of C we have $C(E[f]) = e^{i\omega} G_\rho C(E[h])$ and $C(E[G_\tau f]) = e^{i\psi} G_\sigma C(E[h]) = e^{i\psi} G_\tau G_\rho C(E[h]) = e^{i(\psi - \omega)} G_\tau C(E[f])$. Thus, the choice function we have constructed satisfies the commutativity condition (46)

□

Proposition 4.3 *Let $\mathcal{H}$ be an infinite dimensional space. If for any $g_\gamma \in \mathcal{D}$ and $G_\tau \neq I$ there exists A such that for any $h \in \mathcal{H}$*

$$||h||^2 \geq A \sum_{n \in \mathbf{N}} | < h, G_{\tau^n} g_\gamma > |^2, \tag{50}$$

then $\mathbf{K} = \{0\}$.

Proof: If there exists $f \in \mathcal{H}$ and G_τ such that $\mathbf{E}[f] = \mathbf{E}[G_\tau f]$, then for any $n \in \mathbf{N}$, $\mathbf{E}[f] = \mathbf{E}[G_{\tau^n} f]$, where $G_{\tau^n} f$ is the n^{th} power of G_τ. Hence, for any $g_\gamma \in E[f]$ and $n \in \mathbf{N}$

$$| < f, G_{\tau^n} g_\gamma > | \geq \alpha \lambda(f).$$

If we set $h = f$ in (50), this property implies that $\lambda(f) = 0$, for otherwise f would have an infinite norm. Since linear combinations of elements in $\mathcal{D}$ are dense in $\mathcal{H}$, if $\lambda(f) = 0$ then $f = 0$.

□

Property (50) is satisfied for the Gabor dictionary and the group $\mathcal{G}$ composed of dilations, translations, and modulations for $\mathcal{H} = \mathbf{L}^2(\mathbf{R})$. This comes from our ability to construct frames of $\mathbf{L}^2(\mathbf{R})$ through translations and dilations or frequency modulations of Gaussian functions [5]. Proposition 4.2 implies that there exists a choice function such that matching pursuits with a Gabor dictionary commute with dilations, translations, and frequency modulations.

For dictionaries corresponding to a finite cyclic group, such as the group of unit translations modulo N, we can obtain group invariant decompositions for all functions in a complex space $\mathcal{H}$ if and only if the dictionary contains all eigenvectors of the group operators G_τ [7]. This result cannot be extended to groups generated by two non-commuting elements, such as the set of all unit translations and modulations, because non-commuting operators have different eigenvectors.

For general groups, when $\mathcal{H}$ has a finite dimension and $\mathcal{D}$ is a finite dictionary, we set the optimality factor $\alpha = 1$. Then $f \in \mathbf{K}$ if and only if there exists $g_\gamma \in \mathcal{D}$ and G_τ such that

$$\lambda(f) = | < f, g_\gamma > | = | < f, G_\tau g_\gamma > |.$$

This set $\mathbf{K}$ is of measure 0 in $\mathcal{H}$. If $R^n f$ is not in $\mathbf{K}$ for all $n \in \mathbf{N}$, the proof of proposition 4.1 shows that the commutativity relation (47) remains valid for f. If the set of functions f for which $R^n f \in \mathbf{K}$ for some n is of measure 0 in $\mathcal{H}$ we say that the matching pursuit commutes with operators in $\mathcal{G}$ almost everywhere in $\mathcal{H}$.

5 Chaos in Matching Pursuits

Each iteration of a matching pursuit is a solution of an M-optimal approximation problem where $M = 1$. Hence the pursuit exhibits some of the same instabilities in its choice of dictionary vectors as solutions to the M-optimal approximation problem. In this section we study these instabilities and prove that for a particular dictionary the pursuit is chaotic.

5.1 Renormalized Matching Pursuits

We renormalize the residuals $R^n f$ to prevent the convergence of residuals to zero so we can study their asymptotic properties. Let $R^n f$ be the residual after step n of a matching pursuit. The renormalized residual $\tilde{R}^n f$ is

$$\tilde{R}^n f = \frac{R^n f}{\|R^n f\|}. \tag{51}$$

The *renormalized matching pursuit map* is defined by

$$M(\tilde{R}^n f) = \tilde{R}^{n+1} f. \tag{52}$$

Since $R^{n+1} f = R^n f - < R^n f, g_{\gamma_n} > g_{\gamma_n}$ and

$$\|R^{n+1} f\|^2 = \|R^n f\|^2 - | < R^n f, g_{\gamma_n} > |^2,$$

we derive that if $| < \tilde{R}^n f, g_{\gamma_n} > | \neq 1$

$$M(\tilde{R}^n f) = \tilde{R}^{n+1} f = \frac{\tilde{R}^n f - < \tilde{R}^n f, g_{\gamma_n} > g_{\gamma_n}}{\sqrt{1 - | < \tilde{R}^n f, g_{\gamma_n} > |^2}}. \tag{53}$$

We set $M(\tilde{R}^n f) = 0$ if $| < \tilde{R}^n f, g_{\gamma_n} > | = 1$.

At each iteration the renormalized matching pursuit map removes the largest dictionary component of the residual and renormalizes the new residual. This action is much like that of a binary shift operator acting on a binary decimal: the shift operator removes the most significant digit of the expansion and then multiplies the decimal by 2, which is analogous to a renormalization.

Definition 5.1 *Let $s \in [0,1]$ be expanded in binary form $0.s_1 s_2 s_3 \ldots$, where $s_i \in \{0,1\}$. The binary left-shift map $L : [0,1] \rightarrow [0,1]$ is defined by*

$$L(0.s_1 s_2 s_3 \ldots) = 0.s_2 s_3 s_4 \ldots. \tag{54}$$

The binary shift map is well-known to be chaotic with respect to the Lebesgue measure on $[0,1]$. We recall the three conditions that characterize a chaotic map $T : \Sigma \rightarrow \Sigma$ [8] [2].

1. T must have a *sensitive dependence on initial conditions*. Let $T^{(k)} = T \circ T \circ \ldots \circ T$, k times. There exists an $\epsilon > 0$ such that in every neighborhood of $x \in \Sigma$ we can find a point y such that $|T^{(k)}(x) - T^{(k)}(y)| > \epsilon$ for some $k \geq 0$.

2. Successive iterations of T must mix the domain. T is said to be *topologically transitive* if for every pair of open sets $U, V \subseteq \Sigma$, there is a $k > 0$ for which $T^{(k)}(U) \cap V \neq \emptyset$.

3. The *periodic points* of T must be *dense* in Σ.

The topological properties of the renormalized matching pursuit map are similar to those of the left shift map which suggests the possibility of chaotic behavior. The renormalized matching pursuit map has "sensitive dependence" on the initial signal f when f is near a dictionary element or at the midpoint of a line joining two different dictionary elements. Let $f \in \mathcal{H}$ and g_{γ_1} and g_{γ_2} be two dictionary elements such that

$$|<f, g_{\gamma_1}>| = |<f, g_{\gamma_2}>| > |<f, g_{\gamma}>| \quad \text{for } \gamma_1, \gamma_2 \neq \gamma \in \mathbf{\Gamma}.$$

We can change the residual Rf completely by moving f an arbitrarily small distance towards either g_{γ_1} or g_{γ_2}. The map thus separates points in particular regions of the space. Alternatively, consider two signals f_1 and f_2 defined by

$$f_1 = (1-\epsilon) g_{\gamma_0} + \epsilon h_1 \tag{55}$$

and

$$f_2 = (1-\epsilon) g_{\gamma_0} + \epsilon h_2 \tag{56}$$

where g_{γ_0} is the closest dictionary element to f_1 and f_2, $\|h_1 - h_2\| = 1$, and $<h_1, g_{\gamma_0}>=<h_2, g_{\gamma_0}>= 0$. Then $\|f_1 - f_2\| = \epsilon \|h_1 - h_2\|$ can be made arbitrarily small, while $\|\tilde{R} f_1 - \tilde{R} f_2\| = \|h_1 - h_2\| = 1$. The open ball around g_{γ_0} is mapped to the entire orthogonal complement of g_{γ_0} in the function space, which shows that in some regions of the space, the renormalized matching pursuit map also shares the domain-mixing properties of chaotic maps.

5.2 Chaotic Three-Dimensional Matching Pursuit

Proving that a non-linear map is topologically transitive can be extremely difficult. We therefore consider first a simple dictionary for $\mathcal{H} = \mathbf{R}^3$, for which we prove that the renormalized matching pursuit is topologically equivalent to a shift map. The dictionary $\mathcal{D}$ consists of three unit vectors g_0, g_1, and g_2 in $\mathbf{R}^3$ oriented such that $<g_i, g_j>= \frac{1}{2}$ for $i \neq j$. The vectors form the edges of a regular tetrahedron emerging from a common vertex; each vector is separated by a 60 degree angle from the other two.

To prove the topological equivalence, we first use symmetries to reduce the normalized matching pursuit to a one-dimensional map. The residual $R^n f$ is formed by projecting $R^{n-1} f$ onto the plane perpendicular to the selected $g_{\gamma_{n-1}}$. Hence, the residuals $R^n f$ are all contained in one of the three planes P_i orthogonal to the vectors g_i. We can expand the residual $R^n f \in P_i$ over the orthonormal basis $(e_{i,1}, e_{i,2})$ of P_i given by

$$\begin{aligned} e_{i,1} &= g_{i+1} - g_{i-1} && (57) \\ e_{i,2} &= \frac{g_{i+1} + g_{i-1} - g_i}{\sqrt{2}}. && (58) \end{aligned}$$

All subscripts above and for the remainder of this section will be taken modulo 3.

Let (x_n, y_n) be the coordinates of $R^n f$ with respect to the basis $(e_{i,1}, e_{i,2})$. Since $R^n f$ is orthogonal to g_i, the next dictionary vector that is selected will be either g_{i-1} or g_{i+1}. One can

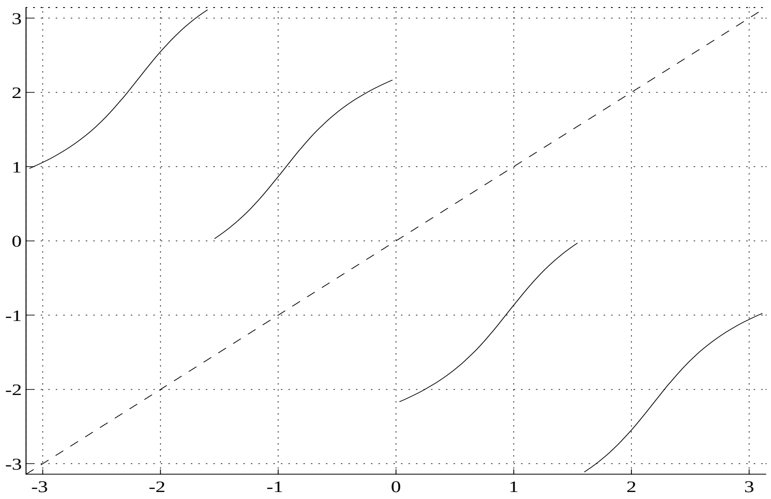

Figure 3: $F(\theta)$ on $[-\pi, \pi)$. The discontinuities occur between quadrants and correspond to the points at which the dictionary element selected by the pursuit changes. The first and third pieces are mapped to P_{i+1} and the second and fourth are mapped to P_{i-1}. The line $y = \theta$ is plotted for reference.

verify that the residual $R^n f$ is mapped to a point in P_{i-1} if $x_n y_n < 0$ and to a point in P_{i+1} if $x_n y_n > 0$. The coordinates of the residual $R^{n+1} f$ in either of these planes are

$$F_{xy}\begin{pmatrix} x_n \\ y_n \end{pmatrix} = \begin{cases} \begin{bmatrix} -\frac{1}{2} & \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & 0 \end{bmatrix} \begin{pmatrix} x_n \\ y_n \end{pmatrix} & \begin{array}{r} x_n > 0, y_n \geq 0 \\ \text{or } x_n < 0, y_n \leq 0 \end{array} \\ \begin{bmatrix} -\frac{1}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & 0 \end{bmatrix} \begin{pmatrix} x_n \\ y_n \end{pmatrix} & \begin{array}{r} x_n \geq 0, y_n < 0 \\ \text{or } x_n \leq 0, y_n > 0 \end{array} \end{cases} \tag{59}$$

The normalized residual $\tilde{R}^n f$ has a unit norm and hence lies on a unit circle in one of the planes P_i. We can thus parameterize this residual by an angle $\theta \in [-\pi, \pi)$ where the angle 0 corresponds to the orthogonal basis vector $e_{i,1}$. The angle of the next renormalized residual $\tilde{R}^{n+1} f$ in P_{i+1} or P_{i-1} is $F(\theta) = \mathrm{Arg}(F_{xy}(\cos\theta, \sin\theta))$. The graph of $F(\theta)$ is shown in Figure 5.2. To simplify the analysis, we identify the three unit circles on the planes P_i to a single circle so that the map $F(\theta)$ becomes a map from the unit circle onto itself. The index of the plane in which a residual vector $R^n f$ lies can be obtained from the index of the plane P_i in which Rf lies and the sequence of the angles in the planes of the residuals $Rf, R^2 f, R^3 f, \ldots$, so the map encodes the plane P_i containing $R^n f$.

F is piecewise strictly monotonically increasing with discontinuities at integer multiples of

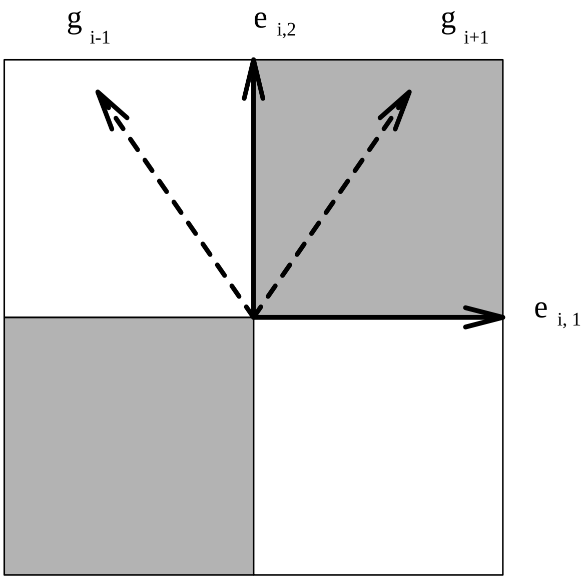

Figure 4: The plane P_i. The projections of g_{i+1} and g_{i-1} are shown as the dotted vectors in the first and second quadrants. The shaded and white areas are mapped to P_{i+1} and P_{i-1}, respectively, by the next iteration of the pursuit.

$\frac{\pi}{2}$. Moreover, F possesses the following symmetries which we make use of below.

$$F(\theta) = \begin{cases} \pi + F(\theta+\pi) & -\pi \leq \theta < -\frac{\pi}{2} \\ -F(-\theta) & -\frac{\pi}{2} \leq \theta < 0 \\ F(\theta) & 0 \leq \theta < \frac{\pi}{2} \\ \pi - F(\pi-\theta) & \frac{\pi}{2} \leq \theta < \pi \end{cases} \tag{60}$$

To analyze the chaotic behavior of this map, we examine $F^{(2)}$, the graph of which is shown in Figure 5. The map $F^{(2)}$ partitions $[-\pi,\pi)$ into four invariant sets $I_+ = [p_1,p_2)$, $I_- = [-p_2,-p_1)$, $J_+ = [0,p_1) \cup [p_2,\pi)$, and $J_- = [-\pi,-p_2) \cup [-p_1,0)$. Here $\pm p_1$ and $\pm p_2$ are the four fixed points of the map given by $p_1 = \tan^{-1}(\sqrt{2})$ and $p_2 = \pi - \tan^{-1}(\sqrt{2})$.

Proposition 5.1 *The restriction of $F^{(2)}$ to each of the invariant regions $I_\pm$ and $J_\pm$ is topologically conjugate to the binary shift map L. The restrictions are therefore chaotic. F is chaotic on the inverse images of $I_\pm$ and $J_\pm$.*

Proof: To prove that $F^{(2)}$ is *topologically conjugate* to L, we must construct an homeomorphism h such that

$$h \circ F^{(2)} = L \circ h. \tag{61}$$

This homeomorphism guarantees that $F^{(2)}$ shares the shift map's topological transitivity, sensitive dependence on initial conditions, and dense periodic points.

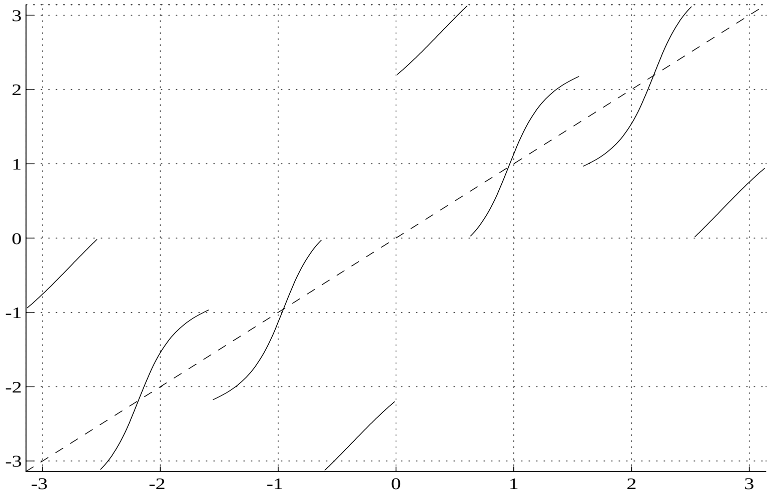

Figure 5: $F^{(2)}(\theta)$ on $[-\pi, \pi)$. The discontinuities correspond to the different selected elements in two iterations of the pursuit. From left to right in $[-\pi, 0)$, the pieces correspond to selecting (1) g_{i+1} followed by g_{i+2}, (2) g_{i+1} followed by g_i, (3) g_{i-1} followed by g_i, (4) g_{i-1} followed by g_{i-2}. The cycle is repeated in $[0, \pi)$. The fixed points correspond to the projections of $\pm g_{i\pm 1}$ onto P_i.

We first focus on the region I_+. Due to symmetry, the construction is identical for I_-, so we drop the subscript of I below. The map $F^{(2)}$ is differentiable over $I_0 = [p_1, \frac{\pi}{2})$ and $I_1 = [\frac{\pi}{2}, p_2)$. For x in I, we define the *index* of x by

$$i(x) = \begin{cases} 0, & x \in I_0 \\ 1, & x \in I_1 \end{cases} \tag{62}$$

The *itinerary* of a point $x \in I$ is the sequence of indices of the images of x under successive applications of $F^{(2)}$. Following a standard technique [8], we construct a homeomorphism h_I that satisfies (61) by assigning to each point $x \in I$ a binary decimal in $[0,1]$ with digits corresponding to the itinerary of x

$$h_I(x) = 0\ .\ i(x)\ i(F^{(2)}(x))\ i(F^{(4)}(x))\ i(F^{(6)}(x)) \ldots \tag{63}$$

The itinerary of $F^{(2)}(x)$ is just the itinerary of x shifted left, so we have

$$\begin{aligned} h_I \circ F^{(2)}(x) &= 0\ .\ i(F^{(2)}(x))\ i(F^{(4)}(x))\ i(F^{(6)}(x)) \ldots \\ &= L \circ h_I(x). \end{aligned} \tag{64}$$

Thus, (61) is satisfied. The details of the proof that $h_I(x)$ is a homeomorphism are similar to those in [8], §1.7, with one technical difference. The method of proving that the map h_I is one-to-one from [8] requires that $(F^{(2)})'$ be bounded above one. This is not the case here. However, we can show that for $\theta \in [-\pi, \pi)$ $(F^{(4)})'(\theta) \geq \frac{11}{6} > 1$. The injectivity of h_I is then obtained with minor modifications of the proof in [8].

The proof that $F^{(2)} : J_\pm \to J_\pm$ is a chaotic map is similar to the proof for $F^{(2)} : I_\pm \to I_\pm$. We consider J_+. We first modify the metric over our domain so that the points p_1 and p_2 have a zero distance as well as the points 0 and π. This metric over our domain is equivalent to a uniform metric over a circle. With this modification, we obtain a map which is differentiable over $[0, p_1)$ and $[p_2, \pi)$ and which maps each of these intervals to the entire domain. The proof now proceeds exactly as above. We note that in the proof that $F^{(2)}$ is chaotic on J we define the index function $i(x)$ so that

$$i(x) = \begin{cases} 0, & x \in F(I_0) \\ 1, & x \in F(I_1) \end{cases} \tag{65}$$

With this construction we obtain a conjugacy between F and the shift map with a homeomorphism $h_J(x) = h_I(F(x))$.

□

The similarities between $F^{(2)}$ and L become much clearer when we compare the graph of $F^{(2)}$ on I in Figure 6 with the graph of the binary shift L on $[0,1)$, given by $y = 2x \bmod 1$. Both maps are piecewise differentiable and monotonically increasing, and both map each continuous piece onto the entire domain. The slope of the graph of L is strictly greater than 1, and although the slope of the pieces of $F^{(2)}$ is not everywhere greater than 1, the slope of the pieces of $F^{(4)}$ is. The itinerary for a point in $[0,1)$ under L is just its binary decimal expansion, so we see that the homeomorphism we have constructed is a natural one.

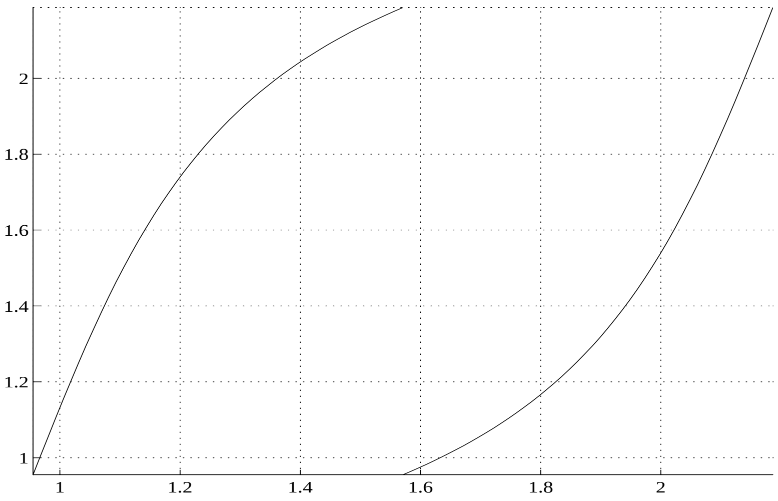

Figure 6: $F^{(2)}(\theta)$ on I.

6 Invariant Measure

The chaotic properties of matching pursuits make it impossible to predict the exact evolution of the residuals, but we can can obtain a statistical description of their properties. For an ergodic map, asymptotic statistics can be obtained from the invariant measure. The residuals $R^n f$ for n large can be interpreted as realizations of an equilibrium process whose distribution is characterized by the invariant measure of the map. The next section describes the basic properties of these invariant measures and analyzes the particular case of the three-dimensional dictionary.

In higher dimensional spaces, numerical experiments show that the norm of the residuals $||R^n f||$ decreases quickly for the first few iterations of a pursuit, but afterwards the decay rate slows down and remains approximately constant. The average decay rate can be computed from the invariant measure, and the measurement of this decay rate has applications in the approximation of signals using a small number of "coherent structures".

Families such as the Gabor dictionary that are invariant under the action of group operators yield invariant measures with invariant properties that are studied in section 6.3. To refine our understanding of the invariant measures, we construct an approximate stochastic model of the equilibrium process and provide numerical verifications for a dictionary composed of discrete Dirac and Fourier bases.

6.1 Ergodicity

We first summarize some results of ergodic theory [14] [20]. Let μ be a measure and let Σ be a measurable set with $\mu(\Sigma) > 0$. Let T be a map from Σ onto Σ. T is said to be *measure-preserving* if for any measurable set $S \subset \Sigma$ we have

$$\mu(S) = \mu(T^{-1}(S)), \tag{66}$$

where $T^{-1}(S)$ is the inverse image of S under T. The measure μ is said to be an *invariant measure* under T. A set E is said to be an *invariant set* under T if $T^{-1}E = E$. The measure preserving map T is *ergodic* with respect to a measure μ if for all invariant sets $E \subset \Sigma$ we have either $\mu(E) = 0$ or $\mu(\Sigma - E) = 0$.

Ergodicity is a measure-theoretical notion that is related to the topological transitivity property of chaos [23]. It implies that the map T moves around the points in its domain. For example, if T is ergodic with respect to a non-atomic measure μ, then only for x in a set of μ-measure 0 do the iterates $Tx, T^2x, T^3x, \ldots$ converge to a cycle of finite length. Hence, for almost all $x \in \Sigma$, T^nx neither tends to a fixed point nor a limit cycle. For most of Σ the asymptotic behavior of T^nx is complicated.

The binary left shift map on [0,1] is ergodic with respect the Lebesgue measure ν [16]. We can use the topological conjugacy relation (61) we derived in section 5 to prove that the renormalized matching pursuit map $F^{(2)}$ is also ergodic with respect to the measure $\mu(S) = \nu(h(S))$, where h is the conjugacy relation from (61), restricted to one of invariant sets $I_\pm, J_\pm$. If the set S is invariant with respect to $F^{(2)}$, i.e. $F^{(2)}(S) = S$, then from (61), the set $h(S)$ must be invariant under L. Because L is ergodic, either $\nu(h(S))$ or $\nu([0,1] - h(S))$ must be zero. From our definition of μ, we have either $\mu(S) = 0$ or $\mu(\Sigma - S) = 0$, proving the result. Using the Birkhoff ergodic theorem, we can use this result to prove that the renormalized matching pursuit map F is ergodic when restricted to one of its two invariant sets.

The ergodicity of a map T allows us to numerically estimate the invariant measure by counting for points $x \in \Sigma$ how often the iterates $Tx, T^2x, T^3x, \ldots$ lie in a particular subset S of Σ. The Birkhoff ergodic theorem [14] states that when $\mu(\Sigma) < \infty$,

$$\mu(S) = \mu(\Sigma) \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \chi_S(T^k x) \tag{67}$$

except possibly for x in a set of μ-measure 0.

When an invariant measure μ is absolutely continuous with respect to the Lebesgue measure, by the Radon-Nikodym theorem there exists a function p such that

$$\mu(S) = \int_S p(x)dx \tag{68}$$

The function p is called an *invariant density*. For the invariant measure of F, this density is given by

$$p(x) = |h'(x)|$$

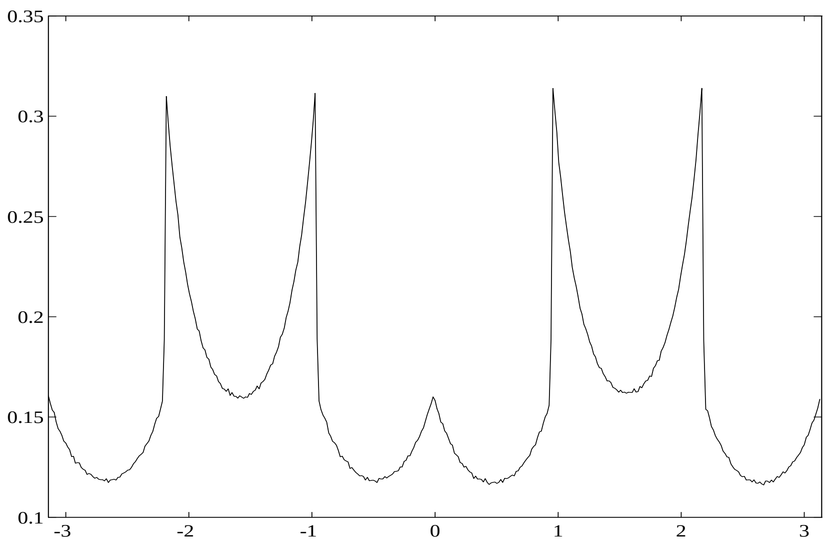

Figure 7: The invariant densities of F for the two invariant sets $I_{\pm} \cup J_{\mp}$ superimposed on the interval $[-\pi, \pi)$. The densities have been obtained by computing the Cesaro sums.

provided that $h(x)$ is absolutely continuous. This invariant density measure can be computed numerically by estimating the limit (67) when the density exists. Fig. 7 is the result of numerically computing the Cesaro sums (67) for a large set of random values of x with sets S of the form $[a, a+\delta)$. In this case, the support of the invariant measure of the normalized matching pursuit is the three unit circles of the planes P_i. On each of these circles, the invariant density measures are the same and equal to $p(\theta)$.

6.2 Coherent Structures

The Birkhoff ergodic theorem implies that an ergodic invariant measure reflects the distributions of successive iterates of the map T. The average number of times the map takes its value in a set is proportional to the invariant measure of this set. The invariant measure provides a statistical description of the behavior of the residuals after a large number of iterations, although the residuals may initially display transient behavior. For example, for the three-dimensional dictionary of section 5, there is one chance in three that the residual is on the unit circle of any particular plane P_i, and over this plane the probability that it is located between the angles θ_1 and θ_2 is $\frac{\mu([\theta_1,\theta_2])}{\mu[0,2\pi]}$.

In higher dimensional spaces the invariant measure μ can be viewed as the distribution of a stochastic process over the unit sphere $\mathcal{S}$ of the space $\mathcal{H}$. After a sufficient number of iterations, the renormalized residuals of the map can be considered to be realizations of this process. We call "dictionary noise" the process P corresponding to the invariant ergodic measure of the renormalized matching pursuit (if it exists). If the dictionary is invariant under translations and frequency modulations, we prove in the next section that the dictionary noise is a stationary white noise. Realizations of a dictionary noise have inner products that are as small and as uniformly spread across the dictionary vectors as possible. Indeed, the measure is invariant under the action of the normalized matching pursuit which sets to zero the largest inner product and makes the appropriate renormalization with (76). Since the statistical properties of a realization x of P are not modified by setting to zero the largest inner product, the value $\lambda(x)$ of this largest inner product cannot be much larger the next larger ones. The average value of this maximum inner product for realizations of this process is by definition

$$\lambda_\infty = \int_{\mathcal{S}} \lambda(x) d\mu(x) = E[\lambda(P)].$$

The ergodicity of the invariant measure implies that

$$\lambda_\infty = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \lambda(\mathbf{R}^k f) \tag{69}$$

for almost all f. We recall from (18) that if the optimality factor $\alpha = 1$ we have

$$||R^{n+1} f|| = ||R^n f|| \sqrt{1 - \lambda^2(R^n f)}. \tag{70}$$

The average decay rate is thus

$$d_\infty = \lim_{n\to\infty} \frac{\log ||f|| - \log ||R^{n-1} f||}{n} = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \frac{-1}{2} \log(1 - \lambda^2(R^k f)). \tag{71}$$

The ergodicity of the renormalized map implies that this average decay rate is

$$d_\infty = -\frac{1}{2} \int_{\mathcal{S}} \log(1 - \lambda^2(x)) d\mu(x) = -\frac{1}{2} E[\log(1 - \lambda^2(P))]. \tag{72}$$

Since $\lambda(x) \geq \lambda_{min}$,

$$d_\infty \geq -\frac{1}{2} \log(1 - \lambda_{min}^2),$$

but numerical experiments show that there is often not a large factor between these two values.

The decay rate of the norms of the residuals $R^n f$ was studied numerically in [24]. The numerical experiments show that when the original vector f is well-correlated with a few dictionary vectors, the first iterations of the matching pursuit remove these highly correlated components, called *coherent structures*. Afterwards, the average decay rate decreases quickly to d_∞.

The chaotic behavior of the matching pursuit map that we have demonstrated provides a theoretical explanation for this behavior of the decay rate. As the coherent structures are removed, the energy of the residuals becomes spread out over many dictionary vectors, as it is for realizations of the dictionary noise P, and the decay rate of the residuals becomes small and on average equal to d_∞. The convergence of the average decay rate to d_∞ can be interpreted as the the residuals of an ergodic map converging to the support of the invariant measure.

We emphasize that our notion of coherence here is entirely dependent upon the dictionary in question. A residual which is considered dictionary noise with respect to one dictionary may contain many coherent structures with respect to another dictionary. For example, a sinusoidal wave has no coherent components in a dictionary composed of Diracs but is clearly very coherent in a dictionary of complex exponentials.

For many signal processing applications, the dictionary defines a set of structures which we wish to isolate. We truncate signal expansions after most of the coherent structures have been removed because the dictionary noise which remains does not resemble the features we are looking for, and because the convergence of the approximations is slow for the dictionary noise. Expansions into coherent structures allow us to compress much of the signal energy into a few elements.

As long as a signal f contains coherent structures, the sequence $\lambda(R^n f)$ has different properties than realizations of the random variable $\lambda(P)$, where P is the dictionary noise process. A simple procedure to decide when the coherent structures have mostly disappeared by iteration n is to test whether a running average of the $\lambda(R^k f)$'s satisfy

$$\frac{1}{d}\sum_{k=n}^{n+d} \lambda(R^k f) \leq \lambda_\infty(1+\epsilon), \tag{73}$$

where d and ϵ are smoothing and confidence parameters, respectively, which are adjusted according to the variance of $\lambda(P)$.

Numerical experiments suggest that the normalized matching pursuit with a Gabor dictionary does have an ergodic invariant measure. After a number of iterations, the residuals behave like realizations of a stationary white noise. The next section shows why this occurs. In our discrete implementation of this dictionary, where the scale is discretized in powers of 2 and $\mathcal{H} = \mathbf{R}^N$ where $N = 8192$, we measured numerically that $\lambda_\infty \approx 0.043$. Fig. 9 displays $\lambda(R^n f)$ as a function of the number of iterations n for a noisy recording of the word "wavelets" shown in Fig. 8. We see that the Cesaro sums of $\lambda(R^n f)$ converge to λ_∞. The time-frequency energy distribution $Ef(t,\omega)$ of the first $n = 200$ coherent structures is shown in Fig. 12. Fig. 11 is the signal reconstructed from these coherent structures, and Fig. 13 shows the approximation error $R^n f$. The signal recovered from the coherent structures has a very good sound quality despite the fact that it was approximated by far fewer elements than the number of samples.

When we use the Gabor dictionary, the coherent structures of a signal are those portions of a signal that are well-localized in the time-frequency plane. Gaussian white noise is not efficiently represented in this dictionary because it is stationary and white, and its energy is spread uniformly over the entire dictionary, much like the realizations of the dicionary noise. Speech contains many structures that are well-localized in the time-frequency plane, especially voiced segments of speech, so these signals are efficiently represented by the Gabor dictionary. As a result, the coherent portion of a noisy speech signal is a much better approximation to the speech than to the noise. The coherent reconstruction of the "wavelet" signal has a 14.9

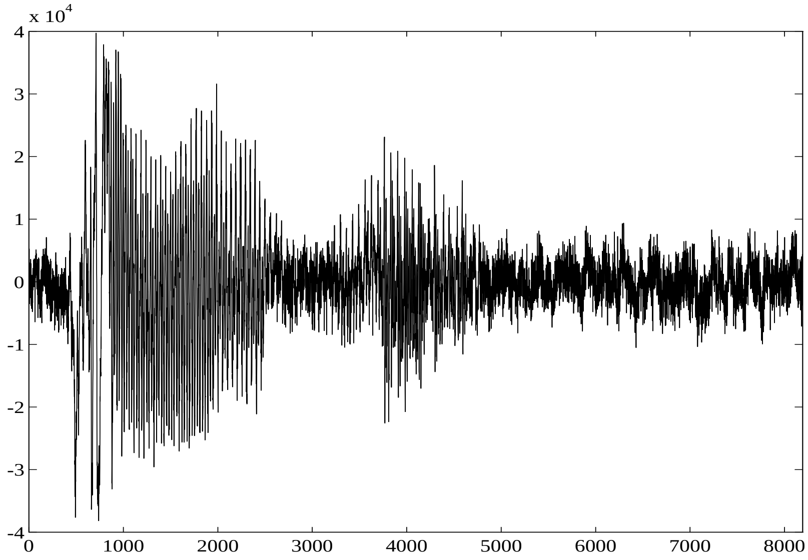

Figure 8: Digitized recording of a female speaker pronouncing the word "wavelets" to which white noise has been added. Sampling is at 11 KHz, and the signal to noise ratio is 10dB.

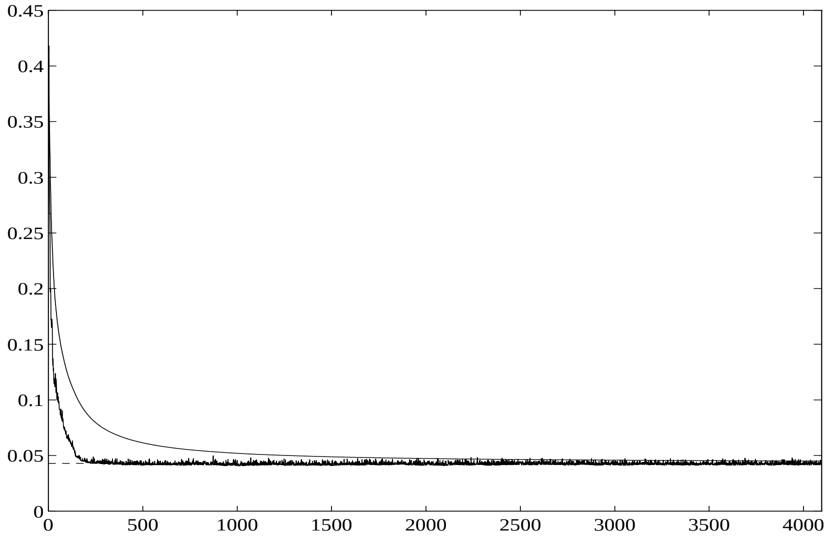

Figure 9: $\lambda(R^n f)$ and the Cesaro sum $\frac{1}{n}\sum_{k=1}^{n}\lambda(R^k f)$ as a function of n for the "wavelets" signal with a dictionary of discrete Gabor functions. The top curve is the Cesaro sum, the middle curve is $\lambda(R^n f)$, and the dashed line is λ_∞.

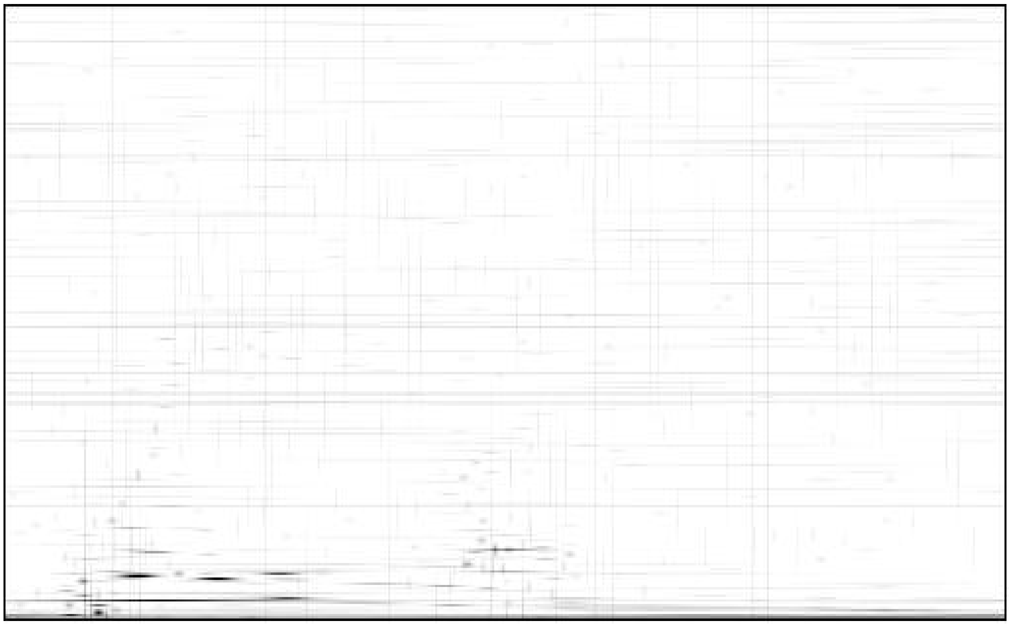

Figure 10: The time-frequency energy distribution of the speech recording shown in Fig. 8. The initial cluster which contains the low-frequency “w” and the harmonics of the long “a”. The second cluster is the “le”. The final portion of the signal is the “s”, which resembles a band-limited noise. The scattered horizontal and vertical bars are components of the noise.

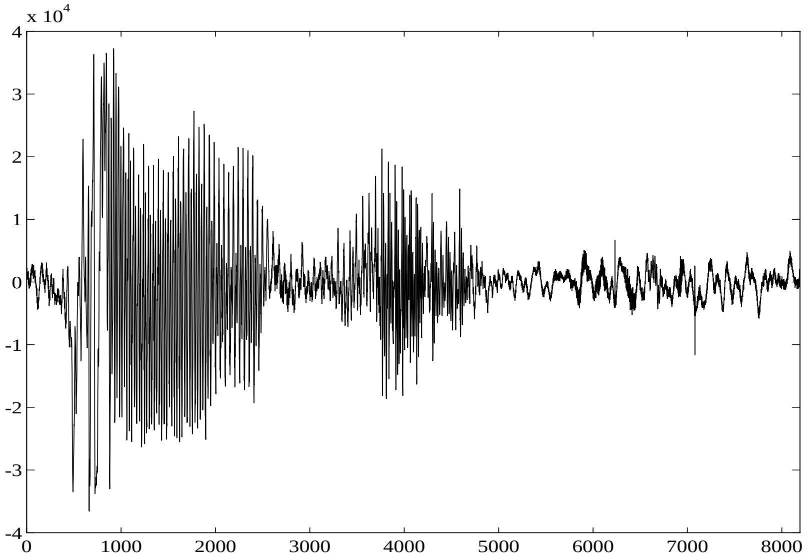

Figure 11: The “wavelets” signal reconstructed from the 200 coherent structures. The number of coherent structures was determined by setting $d = 5$ and $\epsilon = 0.02$.

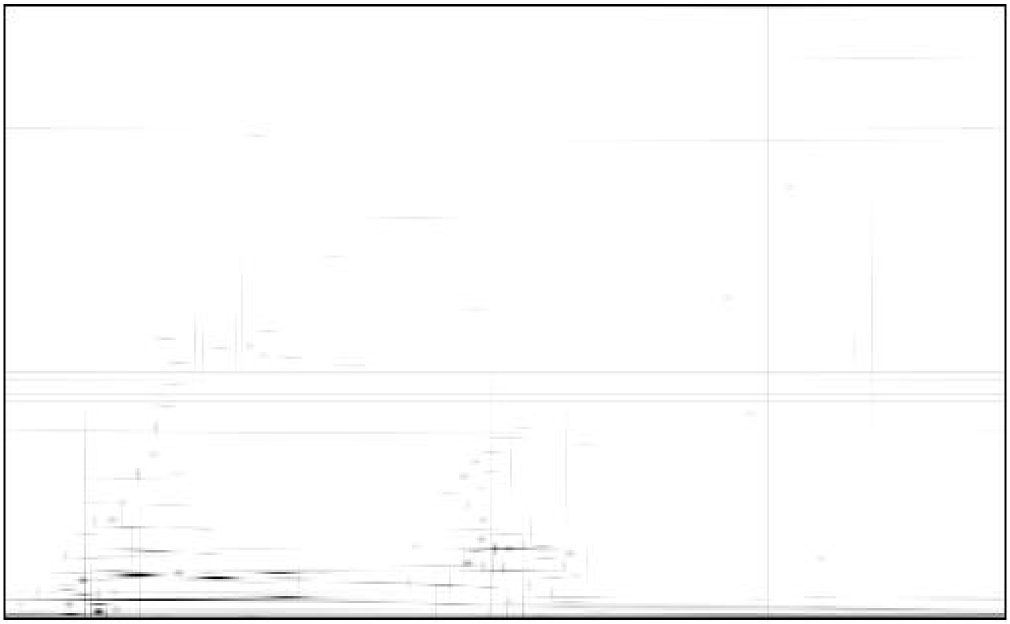

Figure 12: The time-frequency energy distribution of the 200 coherent structures of the speech recording shown in Fig. 8. Note that the phonetic features described in Fig. 8 are all clearly visible in the coherent portion of the signal.

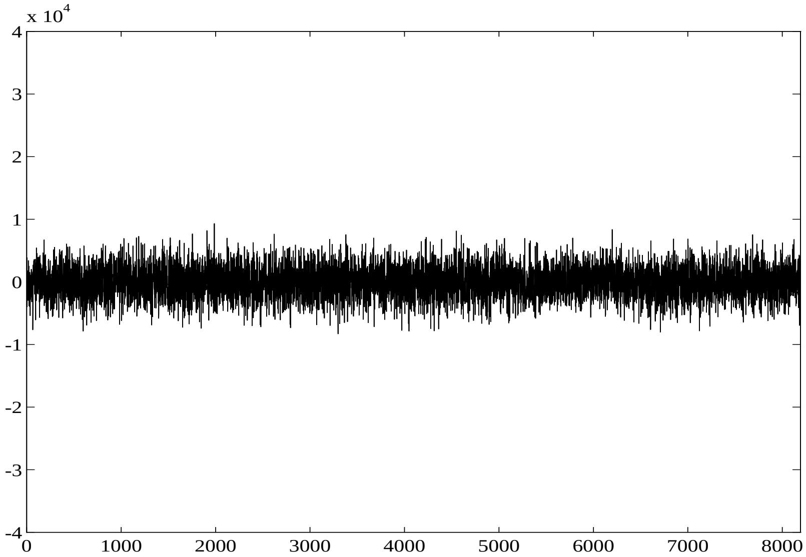

Figure 13: The residual $R^{200}f$ of the "wavelets" signal shown in Fig. 8.

dB signal to noise ratio whereas the original signal had only a 10.0 dB SNR. Moreover, the coherent reconstruction is audibly less noisy than the original signal.

A denoising procedure proposed in [24] is based upon the fact that white noise is poorly represented in the Gabor dictionary, and was inspired by numerical experiments with the decay of the residuals. Similar ideas have been described by [21] [11]. Namely, to separate "noise" from a signal, we approximate the noisy signal using a scheme which efficiently approximates the portion of interest but inefficiently approximates the noise. In order to implement a denoising scheme with a matching pursuit, it is essential that the dictionary be well-adapted to decomposing the portion of signals we wish to retain and poorly-adapted to decomposing that portion we wish to discard. In [7] an algorithm is described for optimizing a dictionary so that the coherence of signals of interest can be maximized. The analysis in the remainder of this section can be used to characterize the types of signals that a given dictionary is inefficient for representing, the realizations of a dictionary noise, so that we can determine what types of "noise" we can remove from signals.

6.3 Invariant Measure of Group Invariant Dictionaries

The Gabor dictionary is a particular example of a dictionary that is invariant under the action of group operators $\mathcal{G} = \{G_\tau\}_{\tau \in \Omega}$. We proved in section 4 that with an appropriate choice function the resulting matching pursuit commutes with the corresponding group operators. If the matching pursuit commutes with G_τ, the renormalized matching pursuit map also satisfies the commutativity property

$$M(G_\tau \tilde{R}^n f) = G_\tau M(\tilde{R}^n f). \tag{74}$$

The following proposition studies a consequence of this commutativity for the invariant measure in a finite dimensional space.

Proposition 6.1 *Let M be a matching pursuit map which is ergodic with respect to an invariant measure μ defined on the unit sphere $\mathcal{S}$ with $\mu(\mathcal{S}) < +\infty$. If there exists a subset of $\mathcal{S}$ of non-zero μ-measure such that (74) is satisfied for all $n \in \mathbf{N}$, then for any $G_\tau \in \mathcal{G}$ and $U \subset \mathcal{S}$*

$$\mu(G_\tau U) = \mu(U).$$

Proof: This result is a simple consequence of the Birkhoff ergodic theorem. Indeed for any $U \subset \Sigma$ and almost any $f \in \mathcal{S}$ whose residuals satisfy (74) we have

$$\mu(U) = \mu(\mathcal{S}) \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \chi_U(M^k f). \tag{75}$$

Hence

$$\mu(G_\tau U) = \mu(\mathcal{S}) \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \chi_{G_\tau U}(M^k f).$$

Since $M^k G_{\tau^{-1}} f = G_{\tau^{-1}} M^k f$

$$\chi_{G_\tau U}(M^k f) = \chi_U(M^k G_{\tau^{-1}} f).$$

But since the limit (75) is independent of f for almost all f, we derive that $\mu(G_\tau U) = \mu(U)$.

□

This result trivially applies to the invariant measure of the three dimensional dictionary studied in section 5. Since the three vectors $\{g_0, g_1, g_2\}$ are all separated by 60 degree angles, this dictionary is invariant under the action of the rotation group composed of $\{I, G, G^2\}$ where I is the identity and G the rotation operator that maps g_i to g_{i+1}. This implies that the invariant measure of the normalized matching pursuit is invariant with respect to G. We thus have the same invariant measure over the unit circle in each plane P_i.

A more interesting application of proposition 6.1 concerns dictionaries that are invariant with respect to translation and frequency modulation groups. Let $\mathcal{H} = \mathbf{R}^N$ and let $\{\delta_n\}_{0\leq n<N}$ be the canonical (or Dirac) basis. The translation group is composed of $\{T^k\}_{0\leq k<N}$ where T is translation modulo N, $T\delta_n = \delta_{(n+1) \bmod N}$. The modulation group is composed of $\{F^k\}_{0\leq k<N}$ where F is the frequency modulation operator defined by $F\delta_n = e^{i\frac{2\pi n}{N}}\delta_n$.

Suppose that the matching pursuit is an ergodic map which admits an invariant measure and that it is implemented with a choice function that commutes almost everywhere with the translation and frequency modulation group operators. Proposition 6.1 proves that the invariant measure of M is also invariant with respect to translations T^k and frequency modulations

F^k. The invariance with respect to translations means that the discrete process associated to this measure is stationary (modulo N). The invariance with respect to frequency modulation operators F^k implies that the discrete power spectrum of this process (the discrete Fourier transform of the N point autocorrelation vector) is constant. In other words, the process is a white stationary noise.

A simple example of a translation and frequency modulation invariant dictionary is constructed by aggregating the canonical basis of N discrete Diracs and the discrete Fourier orthonormal basis

$$\mathcal{D} = \{\delta_n, e_n\}_{0 \leq n < N} = \{g_\gamma\}_{\gamma \in \mathbf{\Gamma}},$$

where e_n is the discrete complex exponential

$$e_n = \sum_{k=0}^{N-1} e^{\frac{i2\pi nk}{N}} \delta_k.$$

In the next section we construct a stochastic model to determine the matching pursuit invariant measure for this dictionary.

6.4 An Invariant Measure Model

We now describe an method for determining the invariant measure for the discrete Dirac-Fourier dictionary. We verify our model numerically at the end of the section.

Let g_{γ_n} be the dictionary element selected on iteration n. The normalized matching pursuit map is defined by

$$\tilde{R}^{n+1} f = \frac{\tilde{R}^n f - < \tilde{R}^n f, g_{\gamma_n} > g_{\gamma_n}}{\sqrt{1 - |< \tilde{R}^n f, g_{\gamma_n} >|^2}}. \tag{76}$$

To find the invariant measure we consider the matching pursuit mapping for a realization of a stochastic process P^n

$$P^{n+1} = M(P^n) = \frac{P^n - < P^n, g_{P^n} > g_{P^n}}{\sqrt{1 - |< P^n, g_{P^n} >|^2}}, \tag{77}$$

where g_{P^n} is a random vector that takes its values over the dictionary $\mathcal{D}$ and satisfies

$$|< P^n, g_{P^n} >| = \sup_{\gamma \in \mathbf{\Gamma}} |< P^n, g_\gamma >|. \tag{78}$$

The invariant measure of the map corresponds to an equilibrium state in which $< P^{n+1}, g_\gamma >$ has the same distribution as $< P^n, g_\gamma >$. For any $\gamma \in \mathbf{\Gamma}$,

$$< P^{n+1}, g_\gamma > = \frac{< P^n, g_\gamma >}{\sqrt{1 - |< P, g_{P^n} >|^2}} - \frac{< P^n, g_{P^n} >< g_{P^n}, g_\gamma >}{\sqrt{1 - |< P^n, g_{P^n} >|^2}}. \tag{79}$$

We recall that $\lambda(P^n)$ is defined to be $|< P^n, g_{P^n} >|$. We suppose that in equilibrium the random variable $\lambda(P^n)$ is constant and equal to its mean, λ_∞. This is equivalent to supposing that the standard deviation of $\lambda(P)$ is small when compared to the mean. This assumption has been verified numerically with several large dimensional dictionaries.

The determination of $< P^n, g_{P^n} >< g_{P^n}, g_\gamma >$ can be divided into three cases. If $g_{P^n} = g_\gamma$, then $< P^{n+1}, g_\gamma >= 0$. If $< g_\gamma, g_{P^n} >= 0$ then (79) reduces to

$$< P^{n+1}, g_\gamma > = \frac{< P^n, g_\gamma >}{\sqrt{1-\lambda_\infty^2}}. \tag{80}$$

Otherwise, we decompose

$$g_{P^n} =< g_{P^n}, P^n > P^n + < g_{P^n}, Q^n > Q^n.$$

Since P^n is a process whose realizations are on the unit sphere of $\mathcal{H}$, this is equivalent to an orthogonal projection onto a unit norm vector P^n plus the projection Q^n onto the orthogonal complement of P^n. We thus obtain

$$< g_{P^n}, g_\gamma >=< g_{P^n}, P^n >< P^n, g_\gamma > + < g_{P^n}, Q^n >< Q^n, g_\gamma > . \tag{81}$$

Inserting this equation into (79) yields

$$< P^{n+1}, g_\gamma > = < P^n, g_\gamma > \sqrt{1 - |< g_{P^n}, P^n >|^2} + A_\gamma^n, \tag{82}$$

with

$$A_\gamma^n = -\frac{< P^n, g_{P^n} >< g_{P^n}, Q^n >< Q^n, g_\gamma >}{\sqrt{1 - |< P^n, g_{P^n} >|^2}}.$$

We have from (81) that

$$|A_\gamma^n| = \frac{\lambda_\infty |< g_{P^n}, g_\gamma > - < g_{P^n}, P^n >< P^n, g_\gamma >|}{\sqrt{1-\lambda_\infty^2}}. \tag{83}$$

If ${\lambda_\infty}^2 \ll |< g_\gamma, g_{P^n} >|^2$, then because

$$|< P^n, g_\gamma >| \leq |< P^n, g_{P^n} >| \approx \lambda_\infty,$$

we have to a first approximation that

$$|A_\gamma^n| = \frac{\lambda_\infty |< g_{P^n}, g_\gamma >|}{\sqrt{1-\lambda_\infty^2}}. \tag{84}$$

Equation (79) is then reduced to

$$< P^{n+1}, g_\gamma > = < P^n, g_\gamma > \sqrt{1-\lambda_\infty^2} + \frac{\lambda_\infty |< g_{P^n}, g_\gamma >| e^{i\phi_\gamma^n}}{\sqrt{1-\lambda_\infty^2}}, \tag{85}$$

where ϕ_γ^n is the complex phase of A_γ^n. The three possible new cases for the evolution of $< P^n, g_\gamma >$ are summarized by

$$< P^{n+1}, g_\gamma >= \begin{cases} \frac{<P^n, g_\gamma>}{\sqrt{1-\lambda_\infty{}^2}}, & \text{if } < g_\gamma, g_{P^n} >= 0, \\ \sqrt{1-\lambda_\infty{}^2} < P^n, g_\gamma > + \frac{\lambda_\infty |< g_\gamma, g_{P^n} >| e^{i\phi_\gamma^n}}{\sqrt{1-\lambda_\infty{}^2}} & \\ & \text{if } \lambda_\infty^2 \ll |< g_\gamma, g_{P^n} >| \\ 0, & \text{if } g_\gamma = g_{P^n}. \end{cases} \tag{86}$$

The Dirac-Fourier dictionary is an example of dictionary for which all above simplification assumptions are valid. We observe numerically that in the equilibrium state for a space of dimension N, λ_∞ is of the order of $\frac{1}{\sqrt{N}}$ whereas the standard deviation of $\lambda(P)$ is of the order of $\frac{1}{N}$, which justifies approximating $\lambda(P)$ by its mean λ_∞. Moreover, for any distinct g_γ and g_{P^n} in this dictionary, either both vectors are in the same basis (Dirac or Fourier) and

$$< g_\gamma, g_{P^n} >= 0,$$

or both vectors are in different bases and

$$\lambda_\infty^2 \ll |< g_\gamma, g_{P^n} >| = \frac{1}{\sqrt{N}}.$$

Thus, one of the approximations of (86) always applies. Because of the symmetrical positions of the Dirac and the Fourier dictionary vectors, there is an equal probability that g_{P^n} belongs to the Dirac or Fourier basis. For any fixed g_γ, the first two updating equations of (86) thus apply with equal frequency. We derive an average updating equation which incorporates both equations for $g_{P^n} \neq g_\gamma$,

$$< P^{n+2}, g_\gamma > - < P^n, g_\gamma >= \frac{\lambda_\infty e^{i\phi_\gamma^n}}{\sqrt{N}}. \tag{87}$$

For n and γ fixed, $e^{i\phi_\gamma^n}$ is a complex random variable and the symmetry of the dictionary implies that its real and imaginary parts have the same distributions with a zero mean. For γ fixed, we suppose that at equilibrium the phase random variables ϕ_γ^n are independent as a function of n at equilibrium. The difference $< P^{n+2K} f, g_\gamma > - < P^n, g_\gamma >$ is approximated by the sum of K independent, identically distributed complex random variables of variance 1. By the central limit theorem, the distribution of $\frac{1}{2K}(< P^{n+2K} f, g_\gamma > - < P^n, g_\gamma >)$ tends to a complex Gaussian random variable of variance 1. The inner products $< P^n, g_\gamma >$ thus follow a complex random walk as long as $g_\gamma \neq g_{P^n}$. The last case $g_{P^n} = g_\gamma$ of (86) occurs when $< P^n, g_\gamma >$ is the largest inner product whose amplitude we know to be

$$|< P^n, g_{P^n} >| = \lambda(P) = \lambda_\infty.$$

At equilibrium, the distribution of $< P^n, g_\gamma >$ is that of a random walk with an absorbing boundary at λ_∞.

To find an explicit expression for the distribution of the resulting inner products $< P^n, g_\gamma >$, we approximate the difference equation with a continuous time Langevin differential equation

$$\frac{d}{dt} < P^t, g_\gamma >= \frac{\lambda_\infty}{2\sqrt{N}} \eta(t), \tag{88}$$

where $\eta(t)$ is a complex Weiner process with mean 0 and variance 1. The corresponding Fokker-Planck equation [12] describes the evolution of the probability distribution $p(z,t)$ of $z =< P^t, g_\gamma >$. Since the phase of $\eta(t)$ is uniformly distributed, the solution can be written $p(z,t) = p(r,t)$ where $r = |z|$ and

$$\frac{\partial p(r,t)}{\partial t} = \frac{{\lambda_\infty}^2}{8N} \triangle p(r,t) \tag{89}$$

which reduces to

$$\triangle p(r) = 0 \tag{90}$$

at equilibrium. The general solution to (90) with a singularity at $r = 0$ is

$$p(r) = C \ln(r) + D.$$

The constants C and D are obtained from boundary conditions.

The inner products $< P^n, g_\gamma >$ start at $r = 0$ and diffuse outward until they reach $r = \lambda_\infty$, at which time $g_{P^n} = g_\gamma$, and $< P^n, g_\gamma >$ returns to 0. The Langevin equation (88) describes the evolution of the inner products before selection; the selection process is modeled by the boundary conditions.

We can write (89) in the form of a local conservation equation,

$$\frac{\partial p(r,t)}{\partial t} + \frac{\partial J(r,t)}{\partial r} = 0, \tag{91}$$

where J, the probability current, is given by

$$J = \frac{-{\lambda_\infty}^2}{8N} \nabla p. \tag{92}$$

The aggregate evolution of the inner products is described by a net probability current which flows outward from a source at the origin and which is removed by a sink at $r = \lambda_\infty$. At each time step, exactly one of the $2N$ dictionary elements is selected and set to 0. Thus, the strength of both the sink and the source is $\frac{1}{2N}$ which implies that

$$\lim_{r \to 0} \oint_{|z|=r} J \cdot \hat{n} \, d\ell = \frac{1}{2N} \tag{93}$$

$$\oint_{|z|=\lambda_\infty} J \cdot \hat{n} \, d\ell = \frac{1}{2N} \tag{94}$$

Integrating (90) we find that $rp_r(r) = C$. By performing the line integrals in (93) and (94), we find that $C = \frac{-2}{\pi {\lambda_\infty}^2}$. Thus, we have

$$p(r) = \frac{-2}{\pi {\lambda_\infty}^2} \ln(r) + D. \tag{95}$$

We use additional constraints to find D and λ_∞. Since all inner products lie in $|r| < \lambda_\infty$, we must have

$$\int_{|z|<\lambda_\infty} p(z)dz = 1. \tag{96}$$

Since the dictionary consists of two orthonormal bases and $\|P^t\| = 1$, we have

$$\sum_{\gamma \in \mathbf{\Gamma}} |< P^t, g_\gamma >|^2 = 2.$$

The $2N$ inner products $< P^t, g_\gamma >$ have last been set to zero at $2N$ different times. We thus assume the mean ergodic property

$$E|< P^t, g_\gamma >|^2 = \frac{1}{2N} \sum_{\gamma \in \mathbf{\Gamma}} |< P^t, g_\gamma >|^2 = \frac{1}{N}$$

and hence

$$\int_{|z|<\lambda_\infty} z^2 p(z)dz = \frac{1}{N}. \tag{97}$$

Inserting (95) into conditions (96) and (97) yields

$$\lambda_\infty = \frac{2}{\sqrt{N}} \quad \text{and} \quad D = \frac{2 \ln \lambda_\infty}{\pi {\lambda_\infty}^2}.$$

Hence

$$p(r) = \frac{2}{\pi {\lambda_\infty}^2} \ln(\frac{\lambda_\infty}{r}). \tag{98}$$

Figure 14 compares the graph of (98) for $N = 4096$ with an empirically determined density function. The empirical density function was obtained by computing the Cesaro sums $\frac{1}{n}\sum_{k=0}^{n} < R^k f, g_\gamma >$ where g_γ is a Dirac element and f is a realization of a Gaussian white noise. The first N terms were discarded to eliminate transient behavior and to speed the convergence of the sum. We have aggregated the Cesaro sums for the members of the Dirac basis to obtain a smoother plot. The invariant density function is invariant under translation due to the translation invariance of the decomposition, so this aggregation does not affect our measurements. The figure shows an excellent agreement between the model and measured values. The small discrepancy near the origin is due to the fact that the approximation of the of the complex exponential term in (87) with a Gaussian is not valid for the first few iterations after $< P^n, g_\gamma >$ is set to 0. Figure 15 compares predicted values of λ_∞ with empirically determined values. These results justify *a posteriori* the validity our approximation hypotheses.

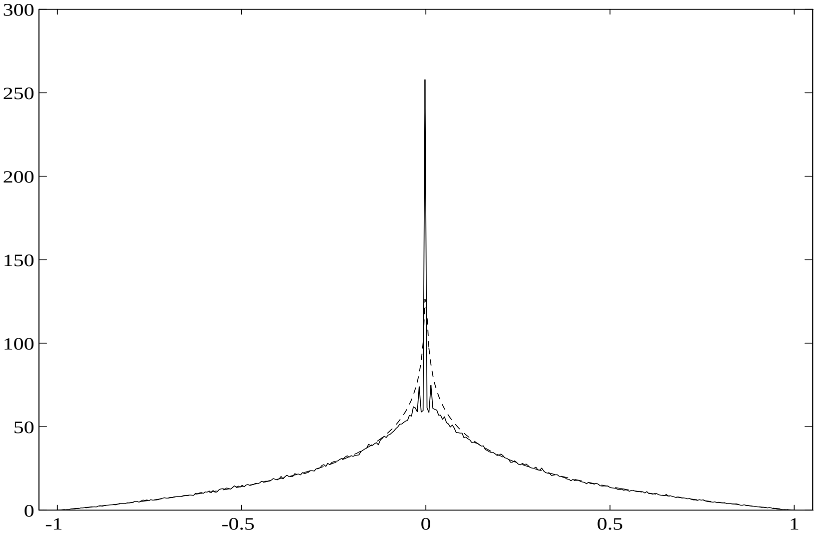

Figure 14: A cross section of the function $p(r,\theta)$ which describes the distribution of the inner products $< P^n, g_\gamma >$ along the $\theta = 0$ axis. The solid curve is determined empirically by computing the Cesaro sums $\frac{1}{n}\sum_{k=0}^{n} < R^k f, g_\gamma >$. The dashed curve is a graph of the predicted density from our model.

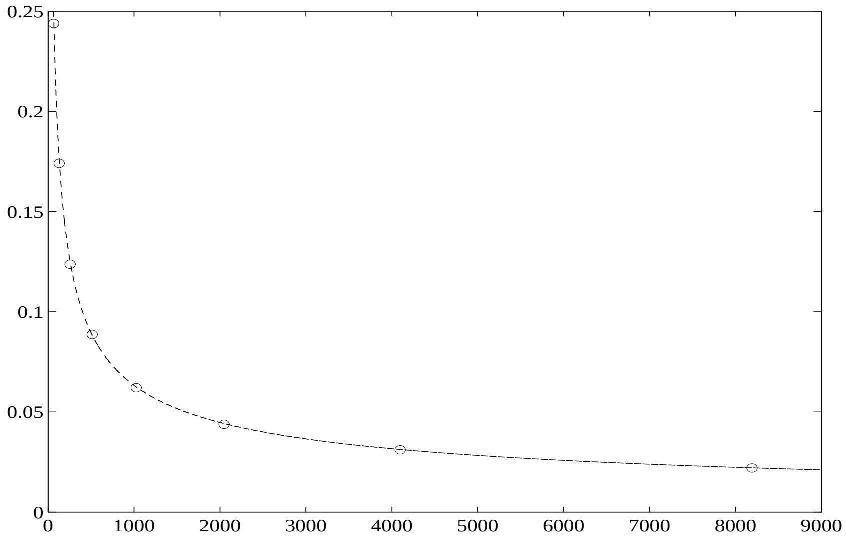

Figure 15: Measured versus predicted values of λ_∞ for the Dirac-Fourier dictionary as a function of the dimension N of the space $\mathcal{H}$. The circles correspond to empirically determined values of λ_∞.

For this dictionary the average value λ_∞ is only twice as large as the minimum λ_{min}. The value λ_{min} is attained for the linear chirp

$$f = \sum_{k=0}^{N-1} e^{\frac{i2\pi k^2}{N}} \delta_k,$$

where

$$\lambda_{min} = \lambda(f) = \frac{1}{\sqrt{N}}.$$

The average value of λ_∞ for this equilibrium process is much smaller than the value $\sqrt{\frac{\log N}{N}}$ which would be obtained from a white stationary Gaussian noise. This shows that the realizations of the dictionary noise have energy that is well spread over the dictionary elements.

7 Comparison of Non-orthogonal and Orthogonal Pursuits

In this section we compare the accuracy and stability of the non-orthogonal and orthogonal pursuit algorithms. Our numerical experiments show that the convergence rates of the algorithms are comparable for the coherent portions of signal expansions. When the number of terms in the expansion is large, however, the orthogonal pursuit algorithm converges much more quickly

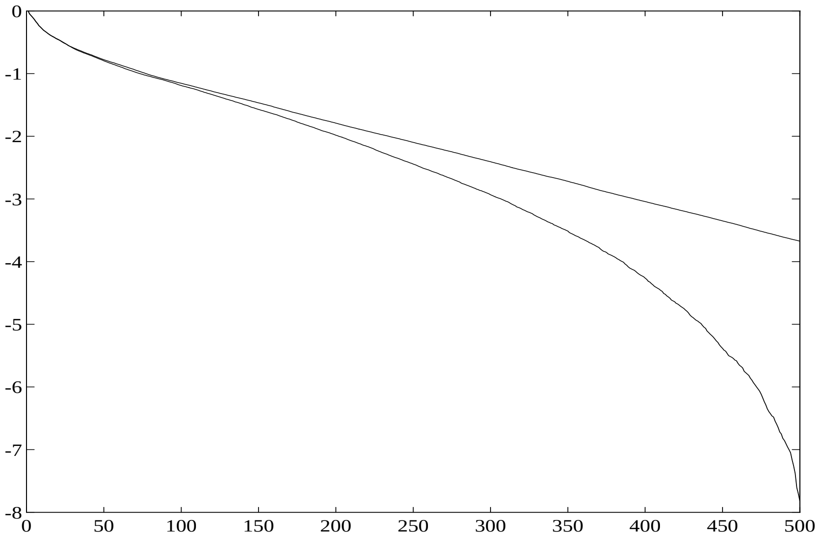

Figure 16: $\log \|R^n f\|$ as a function of n. The top curve shows the decay of $\|R^n f\|$ for a non-orthogonal pursuit and the bottom for an orthogonal pursuit.

than the non-orthogonal pursuit. Our experiments also show that the expansions produced by both pursuits are well-conditioned for the coherent portions of signals. Although [7] showed that it is possible for stability problems to occur with orthogonal pursuits, we do not observe instabilities in our experiments.

7.1 Accuracy of Non-orthogonal and Orthogonal Pursuit Approximations

To compare the performance of the orthogonal and non-orthogonal pursuits, we segmented a digitized speech recording into 512-sample pieces and decomposed the pieces using both algorithms. The dictionary used was the discretized Gabor dictionary described in section 3.4.

Figure 7.1 shows for both algorithms the decay of the residual $\|R^n f\|$ as a function of n for a 512 sample speech segment. When the number of terms in the expansion is close to the dimension of the signal space, we see that the orthogonal pursuit residuals converge very rapidly to 0. The non-orthogonal pursuit residuals, on the other hand, converge exponentially with a slow rate when n is large. We see, then, that orthogonal pursuits yield much better approximations when the number of terms in the expansion is large.

In the initial stages of the expansion, however, the performance of the two algorithms is similar. The reason is that for the early part of the expansion the selected vectors are nearly orthogonal, so the orthogonalization step does not contribute greatly. This near-orthogonality comes from the fact that for both pursuits $< R^{n+1}f, g_{\gamma_n} >= 0$, so

$$|< R^{n+1}f, g_\gamma >|^2 \leq \|R^{n+1}f\|^2(1 - |< g_\gamma, g_{\gamma_n} >|^2). \quad (99)$$

The vector $g_{\gamma_{n+1}}$ is chosen by finding the $\gamma \in \Gamma$ for which the left hand side of (99) is maximized. If the vector g_γ contains a component in the direction of g_{γ_n}, then this portion of g_γ does not contribute to the product $|< R^{n+1}f, g_\gamma >|$. Hence there is a penalty against selecting dictionary elements g_γ for which $|< g_\gamma, g_{\gamma_n} >|$ is large. If $|< g_{\gamma_n}, g_{\gamma_{n+1}} >| = 0$, then we also have

$$|< R^{n+2}f, g_\gamma >|^2 \leq \|R^{n+2}f\|^2(1 - |< g_\gamma, g_{\gamma_n} >|^2)(1 - |< g_\gamma, g_{\gamma_{n+1}} >|), \qquad (100)$$

so we have a similar penalty against selecting a $g_{\gamma_{n+2}}$ which correlates with either g_{γ_n} or $g_{\gamma_{n+1}}$, and so on. Hence, the initially selected vectors tend to be orthogonal. Successive iterations of the pursuit gradually reintroduce a g_{γ_n} component (unless $< g_{\gamma_n}, g_{\gamma_{n+k}} >= 0$), so as k increases, the vectors $g_{\gamma_{n+k}}$ become more correlated with g_{γ_n}.

These nearly-orthogonal elements which comprise the initial terms of the expansion correspond to the signal's coherent structures, the portions of the signal which are well-approximated by dictionary elements. For many applications, we are interested in only the coherent portion of the expansion. Although for large expansions the orthogonal pursuit produces a much smaller error, the difference between the two algorithms is not great for the coherent portion of the expansion. Hence for coherent expansions we can realize a large computational savings by using the faster non-orthogonal pursuits.

To compare the accuracy of approximations generated by the two algorithms, we partitioned a speech recording into 512-sample segments and decomposed the segments using an orthogonal and a non-orthogonal pursuit with the discretized Gabor dictionary. The coherent portions of the non-orthogonal pursuit expansions were determined using by comparing a running average of the $\lambda(R^nf)$'s to λ_∞, as described in section 6.2. For the discretized Gabor dictionary with 512 samples, we find that $\lambda_\infty \approx 0.17$. Selected dictionary elements are deemed to be coherent until a running average of 5 $\|R^nf\|$'s is within 2 percent of λ_∞. We denote by $C(f)$ the number of coherent structures in f.

The average number of coherent structures for the 274 speech segments tested was 72.7. For the coherent portion of the signals, the norm of the residual generated by the orthogonal pursuit was on average only 18.5 percent smaller than the norm of the residual for the matching pursuit. More precisely, let R^nf denote the non-orthogonal pursuit residual, and let R^n_of denote the orthogonal pursuit residual. For the speech segments tested, the ratio $\frac{\|R^{C(f)}f\|}{\|R_o^{C(f)}f\|}$ ranged from 0.864 to 1.771 with an average of 1.185 and a standard deviation of 0.176. We see, then, that for the coherent part of the signal, the benefits of the orthogonalization are not large.

The computational cost of performing a given number of iterations of an orthogonal pursuit is much higher than for the non-orthogonal pursuit, as we showed in section 3. The implementation of the pursuit used requires $I = O(1)$ operations to compute the inner products $< g_\gamma, g_{\gamma'} >$ and on average, $Z = N = 512$ of these inner products are non-zero. The non-orthogonal expansion of the coherent part of the signal thus requires roughly $4 \times 10^4 I$ operations whereas the orthogonal expansion requires roughly $2 \times 10^6 I$ operations. The cost is two orders of magnitude higher for a 20 percent improvement in the error.

The cost of the orthogonal pursuit can be reduced somewhat due to its improved convergence properties. Fewer orthogonal iterations are required to reduce the norm of the residual to a given value. For the coherent portion of the tested speech segments, however, this savings is small. In our experiments, the orthogonal pursuit required an average of $C(f) - 4$ iterations to obtain an error equivalent to that of the non-orthogonal pursuit with $C(f)$ iterations.

7.2 Stability of Non-orthogonal and Orthogonal Pursuit Expansions

Orthogonal pursuits yield expansions of the form

$$f = \sum_{k=0}^{n} \beta_k u_k + R^n f \tag{101}$$

where the u_k's are orthogonalized dictionary elements. When the selected set of dictionary elements is degenerate (when the set does not form a Riesz basis for the space it spans), these expansions cannot be converted into expansions over the dictionary elements g_γ. In [7] it is proved that it is indeed possible for the set $\{g_{\gamma_k}\}$ of vectors selected by an orthogonal pursuit to be degenerate, even when the dictionary contains an orthonormal basis. We now examine numerically the stability of the collection of dictionary elements selected by orthogonal and non-orthogonal pursuits.

To compare the degeneracy of the sets of elements selected by the two algorithms, we computed the 2-norm condition number for the Gram matrix $G_{i,j} =< g_{\gamma_i}, g_{\gamma_j} >$ for twenty 128-sample speech segments. As we discussed above, the initially selected coherent structures are roughly orthogonal and form a well-conditioned set for both pursuits. As the pursuit proceeds, the set selected by the non-orthogonal pursuit grows more and more singular, while the set selected by the orthogonal pursuit stays well-conditioned. The average log of the 2-norm condition numbers for the twenty samples are listed below.

Pursuit	$\log_{10}(\kappa(C))$	$\log_{10}(\kappa(128))$
Non-orthogonal	1.53	12.2
Orthogonal	0.621	2.09

We see that for the non-orthogonal pursuit the condition number of the Gram matrix is small for the coherent portion of the signal (roughly the first 20 vectors in the expansion). The small condition number is the result of the penalty (99). As components of previously selected vectors are reintroduced into the residuals, the penalty (99) against selecting a $g_{\gamma_{n+k}}$ that correlates with g_{γ_n} decreases as k increases. Hence, as the number of iterations becomes close to the dimension of the signal, the set grows more and more singular.

For the orthogonal pursuit, on the other hand, we have

$$\frac{|< R^{n+1} f, g_\gamma >|^2}{||R^{n+1} f||^2} \leq ||(I - P_n) g_\gamma||^2, \tag{102}$$

where P_n is the orthogonal projection onto the space spanned by $g_{\gamma_0} \ldots g_{\gamma_n}$. Hence there is a penalty against selecting for $g_{\gamma_{n+1}}$ a g_γ which correlates strongly with *any* of the previously selected elements. As we see in the table, the condition numbers of the Gram matrices for the orthogonal pursuit are correspondingly smaller.

8 Conclusion

The problem of optimally approximating a function with a linear expansion over a redundant set is a computationally intractable one. The greedy matching pursuit algorithms provide a means of computing compact approximations quickly. The orthogonalized matching pursuit algorithm converges in a finite number of steps in finite dimensional spaces. The much faster non-orthogonal matching pursuits yield comparable expansions for the coherent portion of a signal.

Renormalized matching pursuits possess local topological properties like those of chaotic maps, including local separation of points, and local mixing of the domain. We have shown that for a particular dictionary, the renormalized pursuit is in fact chaotic and ergodic. Ergodic pursuits possess invariant measures from which we obtain a statistical description of the residuals.

For dictionaries which are invariant under the action of a group operator, we can construct a choice function which preserves this group invariance. We can deduce properties of the invariant measure of a pursuit with such a dictionary; in particular, the invariant density function of a translation and modulation invariant pursuit will be stationary and white.

Numerical experiments with the Dirac-Fourier dictionary show that the asymptotic residuals of the pursuit converge to dictionary noise, the realizations of a white, stationary process. The asymptotic convergence rate is slow, and the asymptotic inner products $< R^n f, g_\gamma >$ essentially perform a random walk until they reach a constant λ_∞ and are selected. With an appropriate dictionary, the expansion of a signal into its coherent structures provides a close approximation with a small number of terms.

References

[1] S. Chen, S. A. Billings, and W. Luo, "Orthogonal least squares methods and their application to non-linear system identification," *International Journal of Control*, vol. 50, No. 5, 1873-1896, 1989.

[2] P. Collet and J.P. Eckmann. *Iterated Maps on the Interval as Dynamical Systems*, Birkhauser, Boston, 1980.

[3] L. Cohen, "Time-frequency distributions: a review" Proceedings of the IEEE, Vol. 77, No. 7, 941-979, July 1989.

[4] T. H. Cormen, C. E. Leiserson, and R. L. Rivest. *Introduction to Algorithms*, McGraw-Hill, New York, 1991.

[5] I. Daubechies, *Ten Lectures on Wavelets*, CBMS-NSF Series in Appl. Math., SIAM, 1991.

[6] G. Davis, S. Mallat, and Z. Zhang, "Adaptive time-frequency decompositions," Optical Engineering, July 1994.

[7] G. Davis "Adaptive Nonlinear Approximations," Ph.D. dissertation, Department of Mathematics, New York University, 1994.

[8] R. L. Devaney, *An Introduction to Chaotic Dynamical Systems*, Addison-Wesley Publishing Company, Inc., New York, 1989.

[9] R. A. DeVore, B. Jawerth, V. Popov, "Compression of wavelet decompositions," *American Journal of Mathematics*, 114 (1992): 737-785.

[10] R. A. DeVore, B. Jawerth, B. J. Lucier, "Image compression through wavelet transform coding," *IEEE Trans. Info. Theory*, Vol. 38, No. 2, 719-746, March 1992.

[11] D. L. Donoho, "Wavelet shrinkage and W.V.D.: a 10-minute tour," *Progress in Wavelet Analysis and Applications*, Y. Meyer and S. Roques (eds.), Editions Frontieres, Gif-sur-Yvette, France, 1993, 109-128.

[12] C. W. Gardiner, *Handbook of Stochastic Methods,* Springer-Verlag, New York, 1985.

[13] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W. H. Freeman and Co., New York, 1979.

[14] P. R. Halmos, *Lectures on Ergodic Theory*, The Mathematical Society of Japan, Tokyo, 1956.

[15] L. K. Jones, "On a conjecture of Huber concerning the convergence of projection pursuit regression", *The Annals of Statistics*, vol. 15, No. 2, p. 880-882, 1987.

[16] A. Lasota and M. Mackey, *Probabilistic Properties of Deterministic Systems,* Cambridge University Press, New York, 1985.

[17] S. Mallat and Z. Zhang "Matching Pursuit with Time-Frequency Dictionaries", *IEEE Trans. on Signal Processing*, Dec. 1993.

[18] Y. C. Pati R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal Matching Pursuit: Recursive Function Approximation with Applications to Wavelet Decomposition," *Proceedings of the* 27^{th} *Annual Asilomar Conference on Signals, Systems, and Computers*, Nov. 1993.

[19] S. Qian and D. Chen, "Signal Representation via Adaptive Normalized Gaussian Functions," *IEEE Trans. on Signal Processing*, vol. 36, no. 1, Jan. 1994.

[20] M. Reed and B. Simon, *Methods of Modern Mathematical Statistics, Vol. 1*, Academic Press, New York, 1972.

[21] N. Saito, "Simultaneous Noise Suppression and Signal Compression using a Library of Orthonormal Bases and the Minimum Description Length Criterion," *Wavelets in Geophysics,* to appear.

[22] L. Blum, M. Shub, and S. Smale, "On a theory of computation and complexity over the real numbers: NP-completeness, recursive functions, and universal machines," *Bulletin of the American Mathematical Society*, Vol. 21, No. 1, 1-46, July 1989.

[23] P. Walters, *Ergodic theory–Introductory Lectures*, Springer-Verlag, New York, 1975.

[24] Z. Zhang, "Matching Pursuit," Ph.D. dissertation, New York University, 1993.