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

[19] F. C. Hoppensteadt, *Analysis and Simulation of Chaotic Systems*, Springer-Verlag, New York, 1993.

[20] 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.

[21] A. Lasota and J. A. Yorke, "On the existence of invariant measures for piecewise monotonic transformations," *Transactions of the American Mathematical Society*, 186 (1983), 481-488.

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

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

[24] 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.

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

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

[27] 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.

[28] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, 27:379-423, 623-656, 1948.

[29] 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.

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

[31] R. M. Young, *An Introduction to Nonharmonic Fourier Series,* Academic Press, New York, 1980.

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

[2] S. A. Billings, M. J. Korenberg, and S. Chen, "Identification of non-linear optput-affine systems using an orthogonal least-squares algorithm," *International Journal of Systems Science*, Vol. 19, 1559-1568.

[3] 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.

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

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

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

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

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

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

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

[11] 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.

[12] R. A. DeVore, B. Jawerth, B. J. Lucier, "Surface compression," *Computer Aided Geometric Design*, Vol. 9, 1992, 219-239.

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

[14] 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.

[15] R. J. Duffin and A. C. Schaeffer, "A class of nonharmonic Fourier series," *Trans. Amer. Math. Soc.*, 72, pp. 341-366.

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

[17] 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.

as k increases. Hence, as the number of iterations increases beyond the number of coherent structures, the set grows more and more singular. For the orthogonal pursuit, on the other hand, we have

$$\frac{|< R^{n+1}f, g_\gamma >|^2}{\|R^{n+1}f\|^2} \leq \|(I - P_n)g_\gamma\|^2, \tag{103}$$

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. The table below summarizes the results.

Pursuit	mean($\log_{10}(\kappa(N))$)	mean($\log_{10}(\kappa(N_{coherent}))$)
Non-orthogonal	12.2	1.53
Orthogonal	2.09	0.621

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 quickly computing compact approximations. 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 the signal.

Renormalized matching pursuits possess local topological properties like those of chaotic maps, including local separation of points, and local mixing of the domain. 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 residues.

For dictionaries which are invariant under the action of a group operator, we can construct a choice function which preserves this 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 residues 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] J. Banks, J. Brooks, G. Cairns, G. Davis, and P. Stacey, "On Devaney's Definition of Chaos," *American Mathematical Monthly*, Vol. 99, No. 4, 332-334. April 1992.

of the $\|R^n f\|$'s is within 2 percent of λ_∞. We denote by $N_{coherent}(f)$ the number of coherent structures in f.

For the 274 speech segments tested, the average number of coherent structures was 72.7. For the coherent portion of the signal, 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 $R^n f$ denote the non-orthogonal pursuit residue, and let $R_o^n f$ denote the orthogonal pursuit residue. For the speech segments tested, the ratio $\frac{\|R^{N_{coherent}} f\|}{\|R_o^{N_{coherent}} f\|}$ 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. However, because of the better convergence properties of the orthogonal pursuit, we need not perform as many iterations to obtain the same accuracy as a non-orthogonal pursuit. For the coherent portions of the tested speech segments, the orthogonal pursuit required an average of $N_{coherent} - 4$ iterations to obtain an error equivalent to that of the non-orthogonal pursuit with $N_{coherent}$ iterations. 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.

7.2 Stability

Orthogonal pursuits yield expansions of the form

$$f = \sum_{k=0}^{n} \beta_k u_k + R^n f \tag{102}$$

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 [13] 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 reason is that for the non-orthogonal pursuit, as components of previously selected vectors are reintroduced into the residuals, the penalty (100) against selecting a $g_{\gamma_{n+k}}$ that correlates with g_{γ_n} decreases

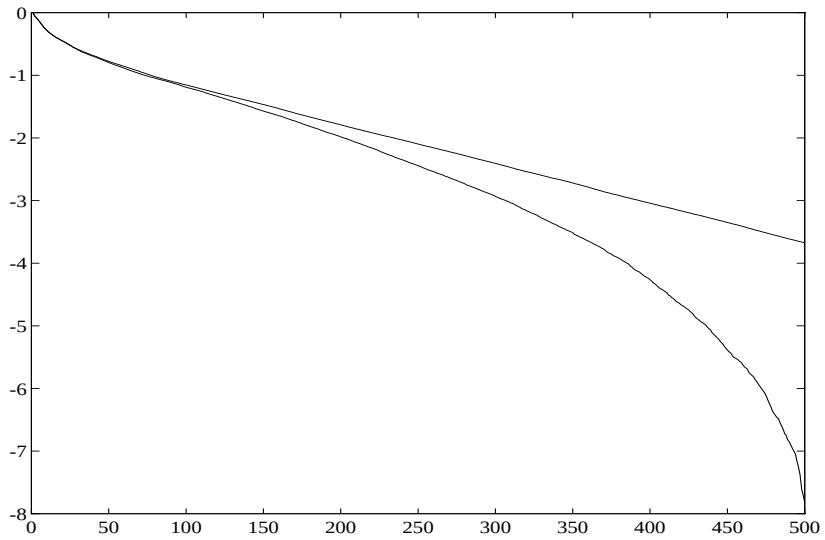

Figure 17: $\log \|R^n f\|$ as a function of the n. The top curve shows the decay of $\|R^n f\|$ for a non-orthogonal pursuit and the bottom for an orthogonal pursuit.

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

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

The vector $g_{\gamma_{n+1}}$ is chosen by finding the $\gamma \in \Gamma$ for which the left hand side of (100) 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

$$| < R^{n+2} f, g_\gamma > |^2 \leq \|R^{n+2} f\|^2 (1 - | < g_\gamma, g_{\gamma_n} > |^2)(1 - | < g_\gamma, g_{\gamma_{n+1}}), \tag{101}$$

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.

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 of f. Although for large expansions, the orthogonal pursuit produces a much smaller error, for the coherent portion of the expansion, the difference between the two algorithms is not great. 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

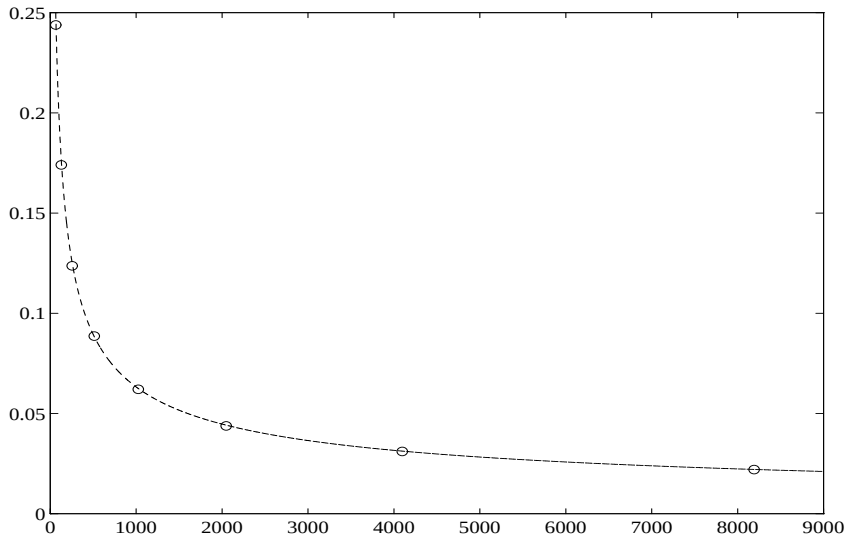

Figure 16: Measured versus predicted values of λ_∞ for the Dirac-Fourier dictionary as a function of the dimension N of the space $\mathcal{H}$. The circles correspond to empirically determined values of λ_∞.

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 orthogonal pursuit algorithm converges much faster than the non-orthogonal pursuit when the number of terms in the expansion is large. However, for expansions over a small number of terms, the performance of the two algorithms is comparable. Our earlier analysis showed that potential stability problems exist with the orthogonal pursuit. We find in our experiments that instabilities do not occur.

7.1 Accuracy

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 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 n is close to 512, the dimension of the signal space, the orthogonal pursuit residue converges very rapidly to 0. The non-orthogonal pursuit, on the other hand, converges exponentially with a slow rate when n is large. We see, then, that orthogonal pursuits yield much better approximations when n is large.

The performance of the two algorithms is similar in the early part of the expansion, however.

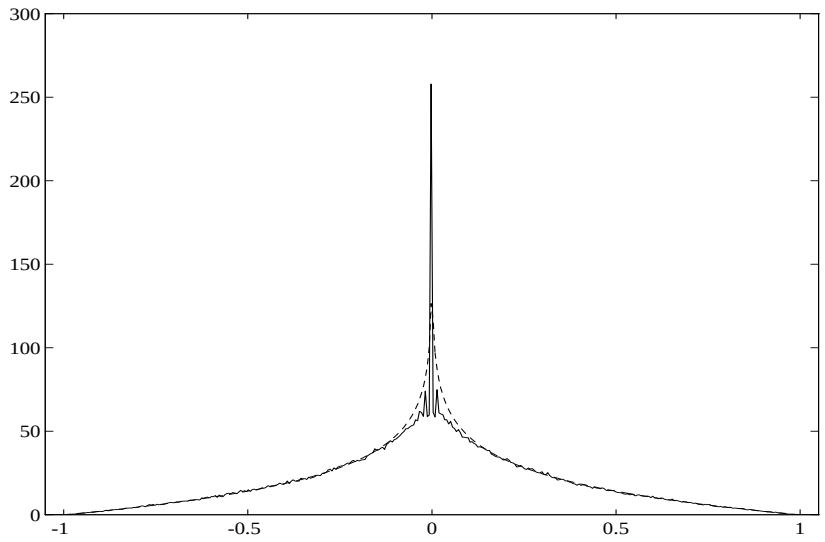

Figure 15: A cross section of the function $p(r, \theta)$ which describes the distribution of the inner products $< P^n, g_\gamma >$ along the $\theta = 0$ axis. The solid curve is determined empirically by computing the Cesaro sums. The dashed curve is a graph of the predicted density from our model.

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

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

The $2N$ inner products $< P^t, g_\gamma >$ with dictionary elements correspond to particles of $2N$ different ages (where a particle's age is the time since it was last set to zero). 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

$$\int_{|z|<\lambda_\infty} z^2 p(z) dz = \frac{1}{N}. \tag{98}$$

Inserting conditions (97) and (98) into (96) yields

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

Hence

$$p(r) = \frac{2}{\pi {\lambda_\infty}^2} \ln(\frac{\lambda_\infty}{r}). \tag{99}$$

Figure 15 compares the graph of (99) 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 higher resolution. The invariant density function is invariant by 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. Figure 16 compares predicted values of λ_∞ with empirically determined values. These results justify *a posteriori* the validity our approximation hypotheses. The discrepancy near the origin is due to the fact that the approximation of the of the complex exponential term in (88) with a Gaussian is not valid for the first few iterations after $< P^n, g_\gamma >$ is set to 0.

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

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

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 truly well spread over the dictionary elements.

$P^t, g_\gamma >$. Since the real and complex parts of $\eta(t)$ have same variance, the solution can be written $p(z,t) = p(r,t)$ where $r = |z|$ and

$$\frac{\partial p(r,t)}{\partial t} = \frac{{\lambda_\infty}^2}{8N} \triangle p(r,t) \tag{90}$$

which reduces to

$$\triangle p(r) = 0 \tag{91}$$

at equilibrium. The general solution to (91) 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 (89) describes the evolution of the inner products before selection; the selection process is modeled by the boundary conditions.

We can write (90) in the form of a local conservation equation,

$$\frac{\partial p(r,t)}{\partial t} + \frac{\partial J(r,t)}{\partial r} = 0, \tag{92}$$

where J, the probability current, is given by

$$J = \frac{-{\lambda_\infty}^2}{8N} \nabla p. \tag{93}$$

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 $\frac{1}{2N}$ and thus implies that

$$\lim_{r \to 0} \oint_{|z|=r} J \cdot \hat{n} \; d\ell = \frac{1}{2N} \tag{94}$$

$$\oint_{|z|=\lambda_\infty} J \cdot \hat{n} \; d\ell = \frac{1}{2N} \tag{95}$$

Integrating (91) we find that $r p_r(r) = C$. By performing the line integrals in (94) and (95), we find that $C = \frac{-2}{\pi {\lambda_\infty}^2}$. Thus, we have

$$p(r) = \frac{-2}{\pi {\lambda_\infty}^2} \ln(r) + D. \tag{96}$$

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

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

The Dirac-Fourier dictionary is an example of dictionary for which all these 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

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

So one of the approximations of (87) 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 (87) thus apply with equal frequency. We derive an average updating equation which incorporates both equations for $g_{P^n} \neq g_\gamma$,

$$< P^{n+2}, g_\gamma > - < P^n, g_\gamma >= \frac{\lambda_\infty e^{i\phi_\gamma^n}}{\sqrt{N}}. \tag{88}$$

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 any fixed γ, we also suppose that at equilibrium the phase random variables ϕ_γ^n are independent as a function of n. The difference $< P^{n+2K} f, g_\gamma > - < P^n, g_\gamma >$ is thus 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 (87) 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 process, we approximate the difference equation with a continuous time Langevin differential equation

$$\frac{d}{dt} < P^t, g_\gamma >= \frac{\lambda_\infty}{2\sqrt{N}} \eta(t), \tag{89}$$

where $\eta(t)$ is a complex Weiner process with mean 0 and variance 1. The corresponding Fokker-Planck equation [16] describes the evolution of the probability distribution $p(z,t)$ of $z =<$

The behavior 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

$$< P^{n+1}, g_\gamma > = \frac{< P^n, g_\gamma >}{\sqrt{1-\lambda_\infty^2}}. \tag{81}$$

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

$$< g_{P^n}, g_\gamma >=< g_{P^n}, P^n >< P^n, g_\gamma > + < g_{P^n}, Q^n >< Q^n, g_\gamma > . \tag{82}$$

Inserting this equation into (79) yields

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

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 (82) that

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

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

$$|A_\gamma^n| = \frac{\lambda_\infty |< g_{P^n}, g_\gamma >|}{\sqrt{1-\lambda_\infty^2}}. \tag{85}$$

Equation (79) is then reduced to

$$< P^{n+1}, g_\gamma > = < P^n, g_\gamma > \sqrt{1-\lambda_\infty^2} + \frac{\lambda_\infty |< g_{P^n}, g_\gamma >| e^{i\phi_\gamma^n}}{\sqrt{1-\lambda_\infty^2}}, \tag{86}$$

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

$$< P^{n+1}, g_\gamma >= \begin{cases} \frac{< P^n, g_\gamma >}{\sqrt{1-{\lambda_\infty}^2}}, & \text{if } < g_\gamma, g_{P^n} >= 0, \\ \sqrt{1-{\lambda_\infty}^2} < P^n, g_\gamma > + \frac{\lambda_\infty |< g_\gamma, g_{P^n} >| e^{i\phi_\gamma^n}}{\sqrt{1-{\lambda_\infty}^2}} & \\ & \text{if } \lambda_\infty^2 \ll |< g_\gamma, g_{P^n} >| \\ 0, & \text{if } g_\gamma = g_{P^n}. \end{cases} \tag{87}$$

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 of the matching pursuit invariant measure obtained with this dictionary.

6.4 An Invariant Measure Model

This section describes an approximate invariant measure model that we apply to the discrete Dirac-Fourier dictionary. The model is verified 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 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} has the same distribution as P^n. 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

$$\lambda(P^n) = |< P^n, g_{P^n} >|. \tag{80}$$

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, and this supposition is verified numerically with several large dimensional dictionaries.

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 residues satisfy (74)

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

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 $\{g_1, g_2, g_3\}$ have equal angles of 60 degrees between themselves, 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} where the index i is taken modulo 3. This implies that the invariant measure of the normalized matching pursuit is invariant with respect to G. It thus admits the same invariant measure over the unit circle in each plane P_i.

A more interesting application of this result concerns dictionaries that are invariant by translation and frequency modulation groups. Let $\mathcal{H} = \mathbf{R}^N$ and $\{\delta_n\}_{0 \leq n < N}$ be the canonical (or Dirac) basis. The translation group is composed of $\{T^k\}_{0 \leq k < N}$ where T is the elementary 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 4.1 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.

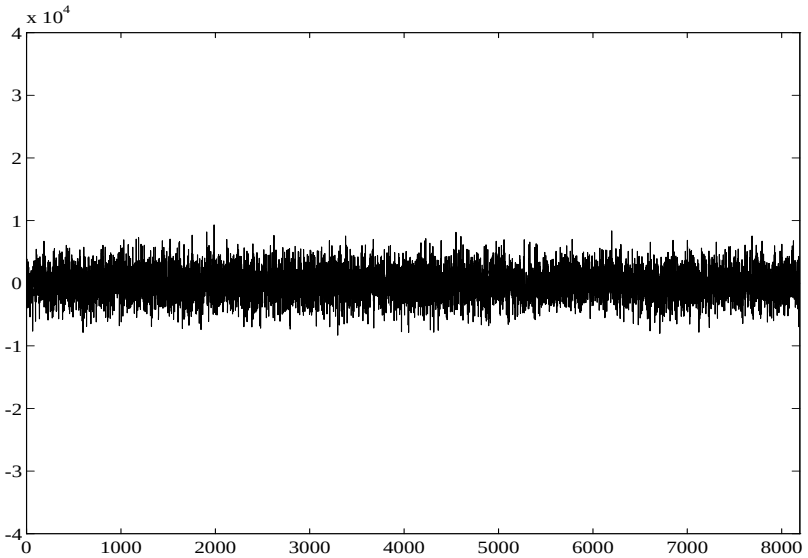

Figure 14: The residual $R^{200}f$ of the "wavelets" signal shown in Fig. 9.

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 an ergodic matching pursuit map with an invariant measure μ defined on the unit sphere $\mathcal{S}$ with $\mu(\mathcal{S}) < +\infty$. If there exists a set of $f \in \mathcal{S}$ 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 $U \in \mathcal{S}$*

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

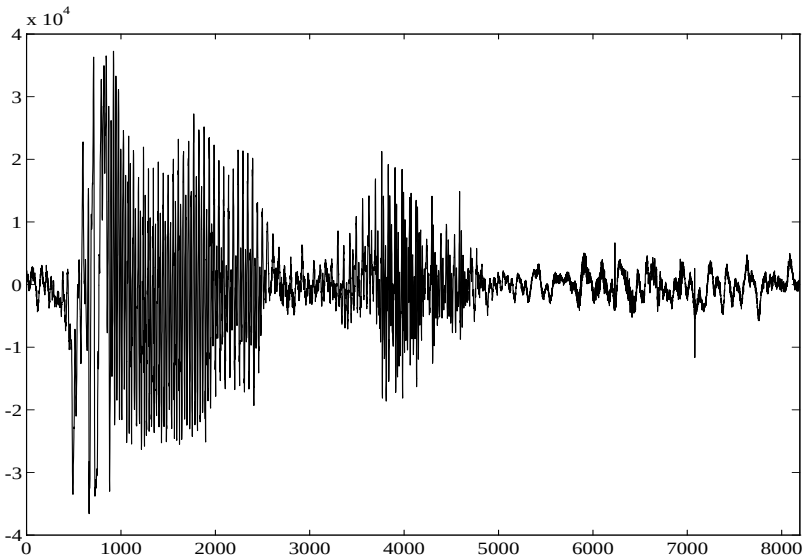

Figure 12: The "wavelets" signal reconstructed from the 200 coherent structures. The number of coherent structures was determined by setting $d = 5$ and $\epsilon = 0.02$.

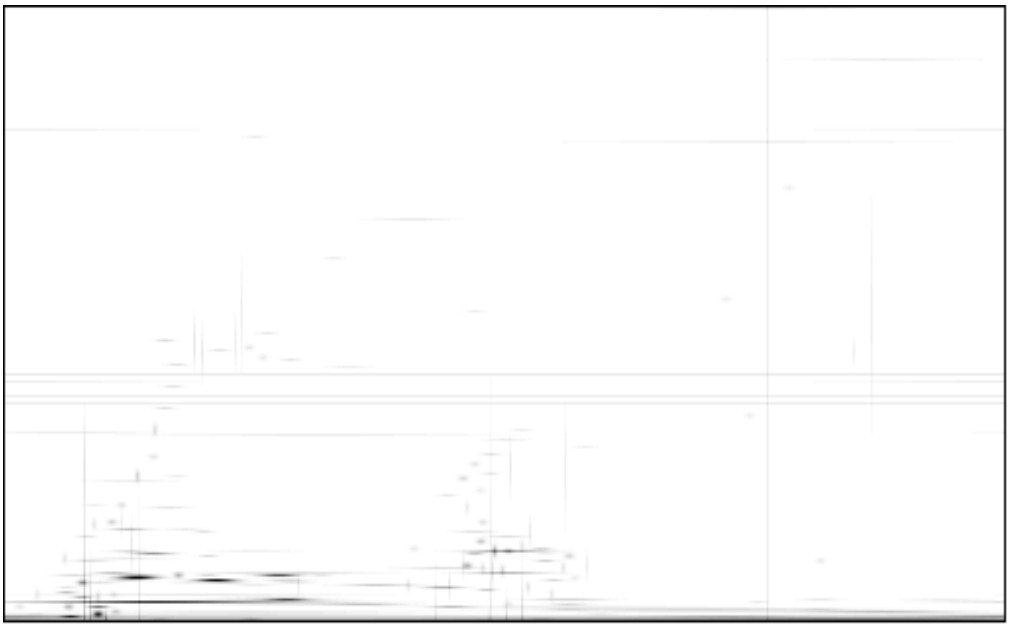

Figure 13: The time-frequency energy distribution of the 200 coherent structures of the speech recording shown in Fig. 9.

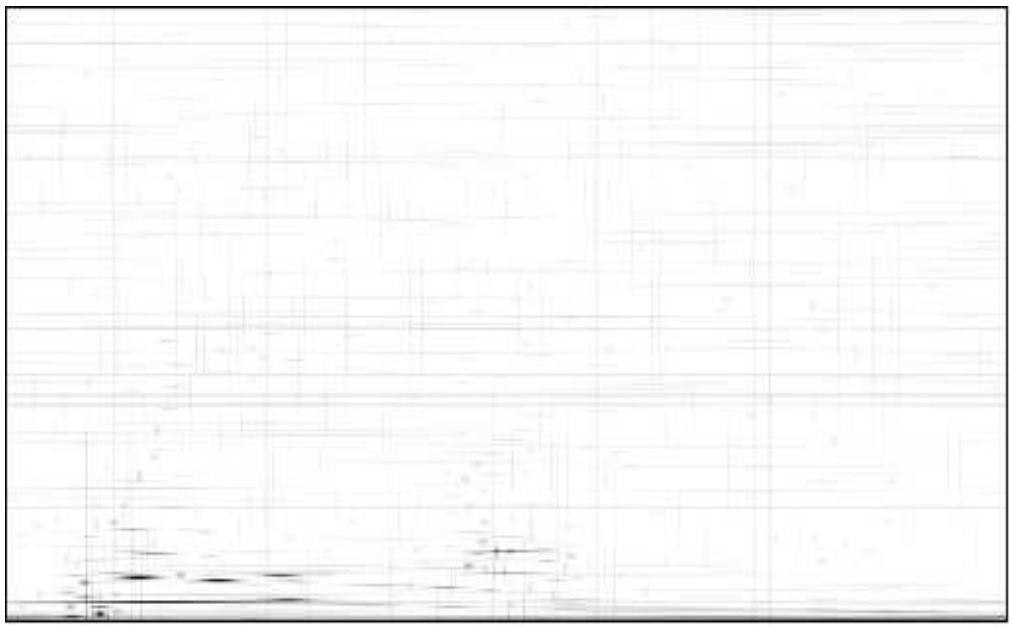

Figure 11: The time-frequency energy distribution of the speech recording shown in Fig. 9. The initial cluster which contains the low-frequency "w" and the harmonics of the long "a". The second cluster is the "le". The final portion of the signal is the "s", which resembles a band-limited noise. The scattered horizontal and vertical bars are components of the noise.

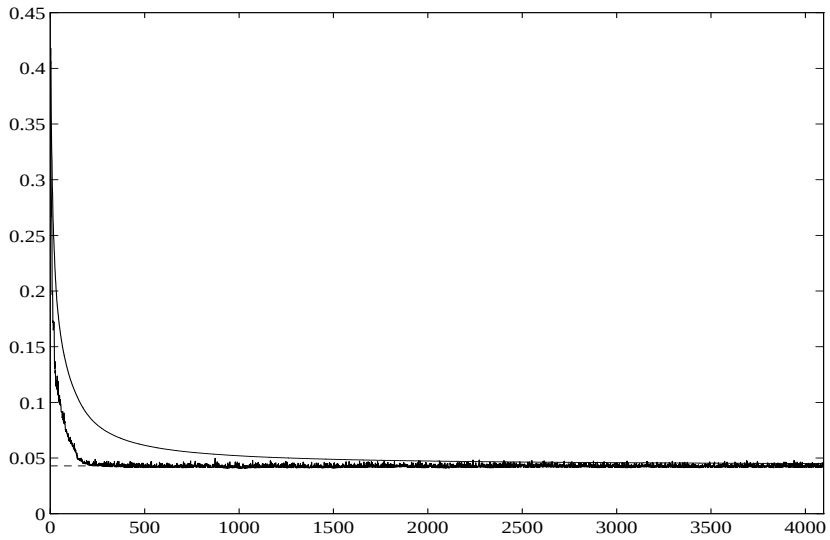

Figure 10: $\lambda(R^n f)$ and the Cesaro sum $\frac{1}{n}\sum_{k=1}^{n}\lambda(R^k f)$ as a function of n for the "wavelets" signal with a dictionary of discrete Gabor functions. The top curve is the Cesaro sum, the middle curve is $\lambda(R^n f)$, and the dashed line is λ_∞.

the number of samples.

When we use the Gabor dictionary, the coherent structures of a signal are those portions of a signal which 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 which are well-localized in the time-frequency plane, especially in voiced segments of speech, so speech signals are efficiently represented by the Gabor dictionary. The coherent portion of a noisy speech signal, therefore, will be a much better approximation to the speech than to the noise. As a result, the coherent reconstruction of the "wavelet" signal has a 14.9 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.

[32] proposes a denoising procedure based upon the fact that white noise is poorly represented in the Gabor dictionary, which was inspired by numerical experiments with the decay of the residues. Similar ideas exist in [27] [14], namely, to separate "noise" from a signal, we approximate a 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 that portion of signals we wish to retain and poorly-adapted to decomposing that portion we wish to discard. In [13] an algorithm is described for optimizing a dictionary so that the coherence of signals of interest can be maximized. Furthermore, the analysis of this section can be used to characterize

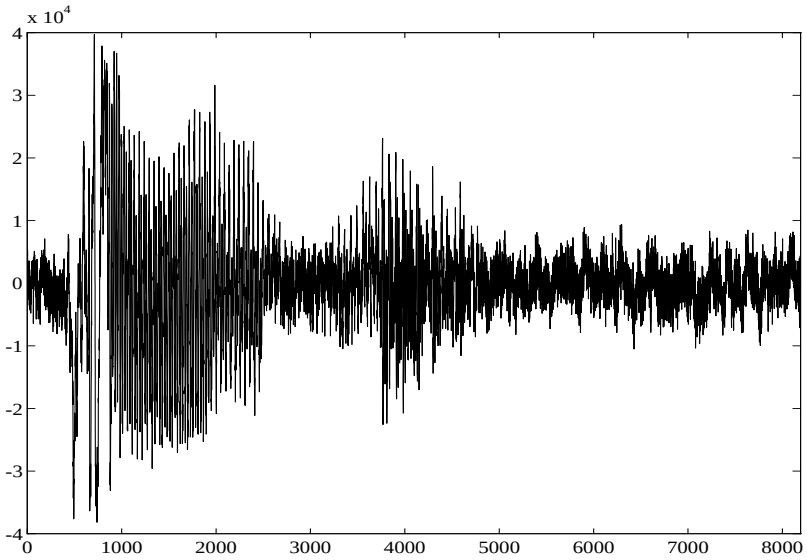

Figure 9: Digitized recording of a female speaker pronouncing the word "wavelets" to which white noise has been added. Sampling is at 11 KHz, and the signal to noise ratio is 10dB.

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 is a smoothing parameter and ϵ is a confidence parameter that are adjusted depending upon the variance of $\lambda(P)$.

Numerical experiments indicate that the normalized matching pursuit with a Gabor dictionary does have an ergodic invariant measure. After a number of iterations, the residues 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 $\lambda_\infty \approx 0.043 = \frac{3.9}{\sqrt{8192}}$. Fig. 10 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. 9. We see that the Cesaro average of $\lambda(R^n f)$ is converging to λ_∞. The time-frequency energy distribution $Ef(t,\omega)$ of the first $n = 200$ coherent structures is shown in Fig. 13. Fig. 12 is the signal reconstructed from these coherent structures whereas Fig. 14 shows the approximation error $R^n f$. The signal recovered from the coherent structures has an excellent sound quality despite the fact that it was approximated by many fewer elements than

The ergodicity of the invariant measure implies that

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

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

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

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

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

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 residues $R^n f$ was studied numerically in [32]. The numerical experiments show that when the original vector f correlates well 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 derived provides a theoretical explanation for this behavior of the decay rate. As the coherent structures are removed, the energy of the residue becomes spread out over many dictionary vectors, as it is for realizations of the dictionary noise P, and the decay rate of the residue becomes small and on average equal to d_∞. The convergence of the average decay rate to d_∞ can be interpreted as the the residues 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 residue 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.

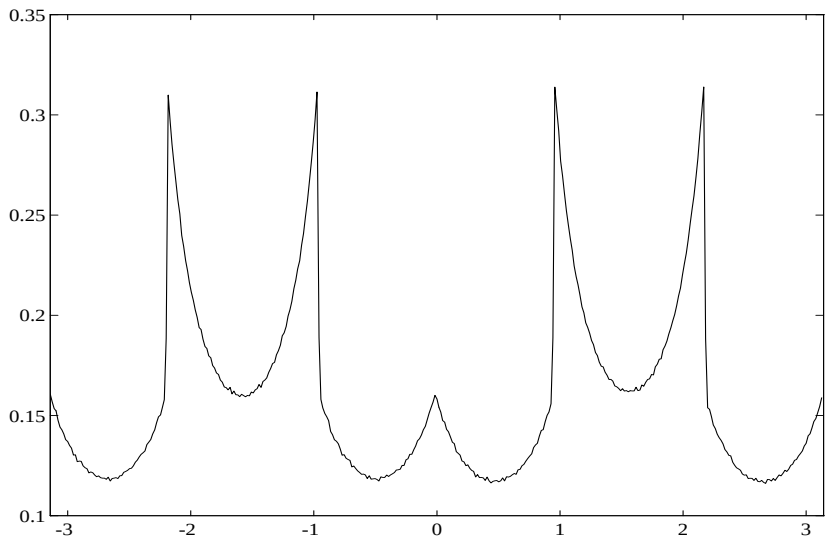

Figure 8: The invariant densities of F for the two invariant sets $I_{\pm} \cup J_{\mp}$ superimposed on the interval $[-\pi, \pi)$. The densities have been obtained by computing the Cesaro sums.

for the three-dimensional dictionary of section 5, there is one chance in three that the residue is on the unit circle of any particular plane P_i, and over this plane the probability that it is located at the angle θ is $p(\theta)$.

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 residue of the map can be considered as a realization 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

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

For example, if T has an ergodic invariant measure μ that is non-atomic (every set of non-zero measure contains a subset of smaller measure), then only for 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 $x \in \Sigma$, $T^n x$ neither goes to a fixed point or a limit cycle, so for most of Σ the asymptotic behavior of $T^n x$ is complicated.

The binary left shift map on [0,1] is ergodic with respect the Lebesgue measure [22]. 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), when 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 [18] states that when $\mu(\Sigma) < \infty$,

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

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

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

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

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

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. 8 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 $[a, a + \delta_a)$. In this case, the support of the invariant measure of the normalized matching pursuit is on 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 Cesaro sum (67) shows that an ergodic invariant measure reflects the distributions of iterates of the map T. The average number of times the map takes its value in a set is proportional to the measure of this set. The invariant measure thus provides a statistical description after a large number of iterations, during which the map may have transient behavior. For example,

□

The similarities of $F^{(2)}$ and L become much clearer when we compare the graph of $F^{(2)}$ on I in Figure 7 with the graph of the binary shift L on $[0,1)$, given by $y = 2x \mod 1$. 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.

6 Invariant Measure

The chaotic properties of matching pursuits make it impossible to predict the exact evolution of the residues, but we can can obtain a statistical description of the properties of the residues. For an ergodic map, asymptotic statistics can be obtained from the invariant measure. The residue can then be interpreted as a realization 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 residues $\|R^n f\|$ decreases quickly for the first few iterations, 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 for 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 [18] [26]. 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 [30]. It implies that the map T mixes around the points in its domain.

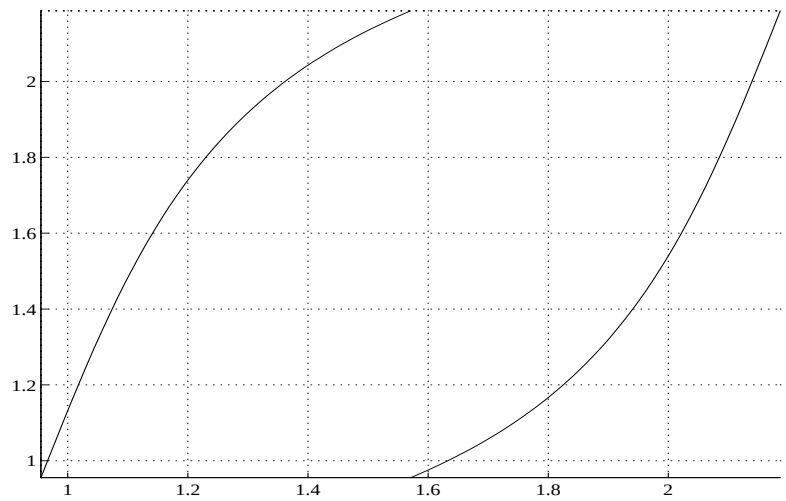

Figure 6: $F^{(2)}(\theta)$ on I.

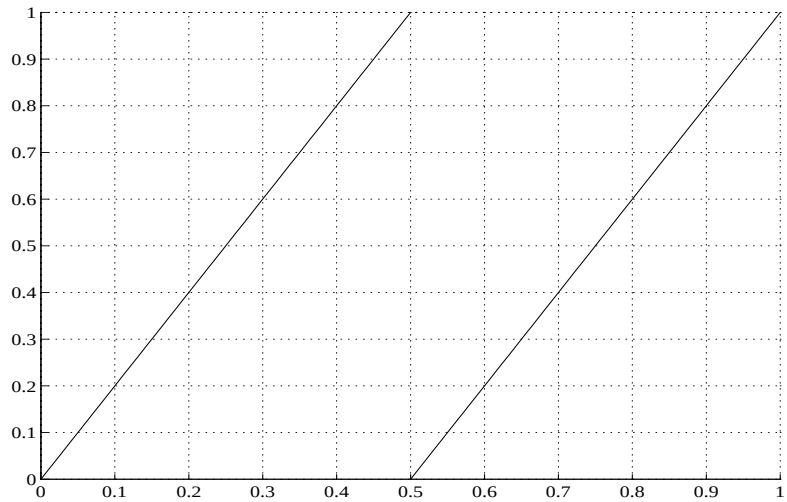

Figure 7: The binary left shift operator L on binary expansions of $[0, 1]$.

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 and are therefore chaotic. Hence 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.

We first focus on the region I_+. Due to the symmetry, the construction is identical for I_-, and we drop the subscript 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 [9], the homeomorphism h is constructed by assigning to each point $x \in I$ a binary decimal in $[0, 1]$ with digits corresponding to the itinerary of x

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

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

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

Thus, (61) is satisfied. The details of the proof that $h(x)$ is a homeomorphism are similar to [9], §1.7, with one minor difference. The fact that h is one-to-one in [9] 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 is then obtained with minor modifications of the original proof.

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 $h'(x) = h(F(x))$.

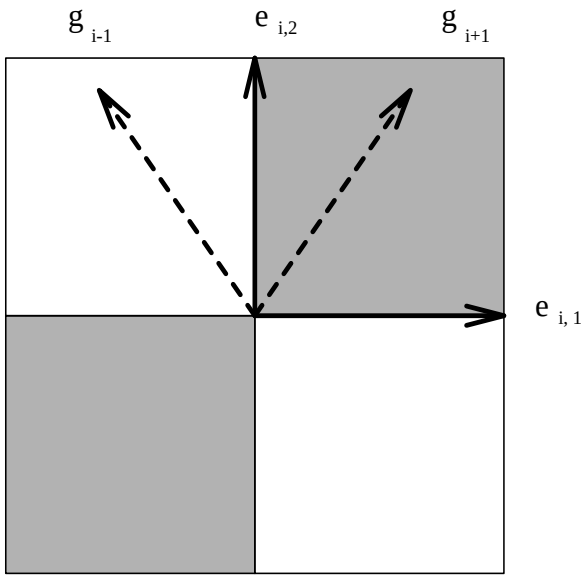

Figure 4: The plane P_i. The projections of g_{i+1} and g_{i-1} are shown as the dotted vectors in the first and second quadrants. The shaded and white areas are mapped to P_{i+1} and P_{i-1}, respectively, by the next iteration of the pursuit.

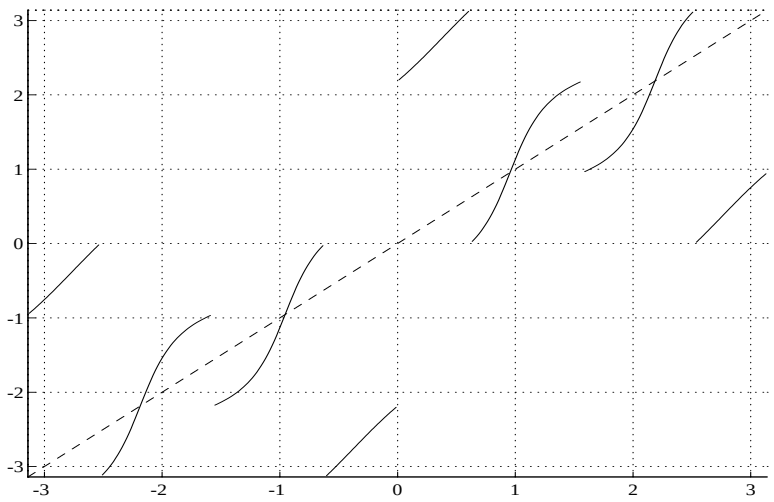

Figure 5: $F^{(2)}(\theta)$ on $[-\pi, \pi)$. The discontinuities again correspond to the different selected elements in the two iterations of the pursuit. From left to right in $[-\pi, 0)$, the pieces correspond to selecting (1) g_{i+1} followed by g_{i+2}, (2) g_{i+1} followed by g_i, (3) g_{i-1} followed by g_i, (4) g_{i-1} followed by g_{i-2}, with the cycle repeated in $[0, \pi)$. The fixed points correspond to the projections of $\pm g_{i\pm1}$ onto P_i.

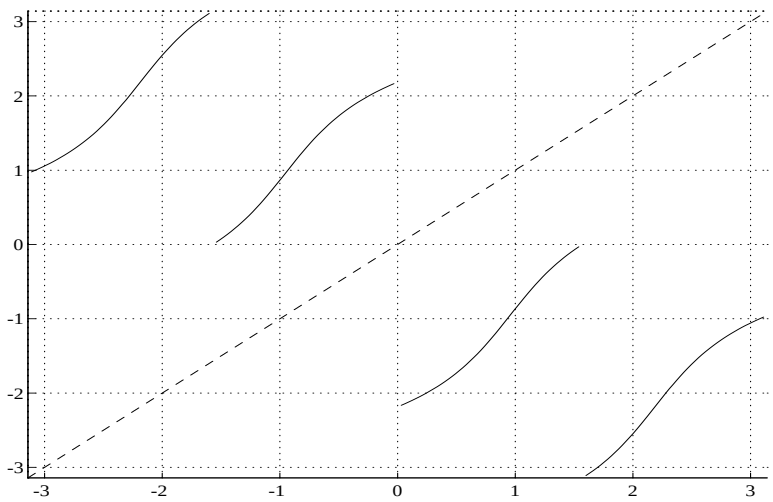

Figure 3: $F(\theta)$ on $[-\pi, \pi)$. The discontinuities occur between quadrants and correspond to the points at which the element selected by the pursuit changes. The first and third pieces are mapped to P_{i+1} and the second and fourth are mapped to P_{i-1}. The line $y = \theta$ is plotted for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

The normalized residue $\tilde{R}^n f$ has a unit norm and hence lies on a unit circle in one of the planes P_i. We can thus parameterize this residue 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 residue $\tilde{R}^{n+1} f$ 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 residues $Rf, R^2 f, R^3 f, \ldots$, so the map encodes the plane P_i containing $R^n f$.

F is piecewise strictly monotonically increasing with discontinuities at integer multiples of $\frac{\pi}{2}$. Moreover, F possesses the following symmetries which we will use later:

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

To analyze the chaotic behavior of this map, we focus on $F^{(2)}$ 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 $p_2 = \pi - \tan^{-1}(\sqrt{2})$.

We can change the residue 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

To prove analytically that a non-linear map is topologically transitive is often extremely difficult. We thus concentrate first on a simple dictionary of $\mathcal{H} = \mathbf{R}^3$, where 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 reduce the normalized matching pursuit to a one-dimensional map. The residue $R^n f$ is formed by projecting $R^{n-1}f$ onto the plane perpendicular to the selected $g_{\gamma_{n-1}}$. Hence, the residues $R^n f$ are all contained in one of the three planes P_i orthogonal to the vectors g_i. We can expand the residue $R^n f \in P_i$ over the orthonormal basis $(e_{i,1}, e_{i,2})$ of P_i given by

$$e_{i,1} = g_{i+1} - g_{i-1} \tag{57}$$

$$e_{i,2} = \frac{g_{i+1} + g_{i-1} - g_i}{\sqrt{2}}. \tag{58}$$

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

Let (x_n, y_n) be the coordinates of $R^n f$ in the basis $(e_{i,1}, e_{i,2})$. Since it is orthogonal to g_i the next dictionary vector that is selected is either g_{i-1} or g_{i+1}. One can verify that the residue $R^n f$ is mapped to a point in P_{i-1} if $x_n y_n \leq 0$ and to a point in P_{i+1} if $x_n y_n \geq 0$. The coordinates of the residue $R^{n+1}f$ is either is these planes are

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

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 residue and renormalizes the new residue. 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 $L : [0,1] \to [0,1]$ 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 $T : \Sigma \to \Sigma$ [9] [4].

1. T must have a *sensitive dependence on initial conditions*. Let $T^{(k)} = T \circ T \circ \ldots \circ T$, k times. There exists $\epsilon > 0$ 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

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

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 $\lambda(f) = 0$ 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 [7]. This result implies that there exists a choice function such that the matching pursuit in a Gabor dictionary commutes 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 translation and invariant decompositions for all functions f in a complex space $\mathcal{H}$ if and only if the dictionary contains all eigenvectors of the group operators G_τ [13]. 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

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

This set $\mathbf{K}$ is of measure 0 in $\mathcal{H}$. If for all $n \in \mathbf{N}$ $R^n f$ is not in $\mathbf{K}$, the proof of proposition 4.1 proves that the commutativity relation (46) remains valid for f. If the set of such functions 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 residues $R^n f$ to prevent the convergence of residues to zero so we can study their asymptotic properties. Let $R^n f$ be the residue after step n of a matching pursuit. The renormalized residue $\tilde{R}^n f$ is

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

Hence $g_\gamma \in \mathbf{E}[G_\tau f]$ if and only if $G_{\tau^{-1}} g_\gamma \in \mathbf{E}[f]$ which proves that $\mathbf{E}[G_\tau f] = G_\tau \mathbf{E}[f]$. By using the commutativity (45) of the choice function with respect to G_τ we then easily prove (46) by induction.

□

The difficulty is now to prove that there exists a choice function that satisfies the commutativity relation (45) for all $f \in \mathcal{H}$ or at least for almost all $f \in \mathcal{H}$. The following proposition gives a necessary condition to construct such choice functions.

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

$$\mathbf{E}[f] = \mathbf{E}[G_\tau f]. \tag{47}$$

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

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

Proof: To define such a choice function we construct the equivalence classes of the equivalence relation R_1 in $\mathcal{H}$ defined by $f \; R_1 \; h$ if and only if there exists $G_\tau \in \mathcal{G}$ such that $f = G_\tau h$. The axiom of choice guarantees that there exists a choice function that chooses an element from each equivalence class. Let $\mathcal{H}_1$ be the set of all class representatives. The axiom of choice also guarantees that for any $f \in \mathcal{H}_1$ there exists a choice function C that associates to any set $\mathbf{E}[f]$ an element within this set. To extend this choice function we define a new equivalence relation R_2 in $\mathcal{H}$ defined by $f \; R_2 \; h$ if and only if there exists $G_\tau \in \mathcal{G}$ such that $f = G_\tau h$ and $\mathbf{E}[f] = \mathbf{E}[h]$. All elements that belong to $\mathcal{H} - \mathbf{K}$ correpond to equivalence classes of 1 vector. For each equivalence class, we choose a representative f. There exists G_τ such that $G_\tau f \in \mathcal{H}_1$. Since G_τ is unitary, for any $g_\gamma \in \mathcal{D}$

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

Hence, $\mathbf{E}[G_\tau f] = G_\tau E[f]$. For any h that is equivalent to f with respect to R_2, we define

$$C(\mathbf{E}[h]) = C(\mathbf{E}[f]) = G_{\tau^{-1}} C(\mathbf{E}[G_\tau f]) = G_{\tau^{-1}} C(G_\tau \mathbf{E}[f]) \in E[f]. \tag{49}$$

This choice function associates a unique element to each different set $E[f]$. If $f \in \mathcal{H} - \mathbf{K}$, property (49) implies that this choice function satisfies (48).

□

Proposition 4.3 *When* $\mathcal{H}$ *is 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 $\mathbf{K} = \{0\}$.

4 Group Invariant Dictionaries

The translation, dilation, and frequency modulation of any vector that belongs to the Gabor dictionary still belongs to this dictionary. The dictionary invariance under the action of any operators that belong to the group of translations, dilations or frequency modulations implies important invariance properties of the matching pursuit. We study such properties when the dictionary is left invariant by any 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 ajoint 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 any $g_\gamma \in \mathcal{D}$ and $\tau \in \Omega$, $G_\tau g_\gamma \in \mathcal{D}$.*

The Gabor dictionary is invariant under the group generated by the groups of 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 $n \in \mathbf{N}$

$$C G_\tau \mathbf{E}[R^n f] = G_\tau C \mathbf{E}[R^n f], \tag{45}$$

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

$$G_\tau f = \sum_{k=0}^{n-1} a_n G_\tau g_{\gamma_n} + G_\tau R^n f. \tag{46}$$

The condition (45) means that an element chosen from the $\mathbf{E}[R^n f]$ transformed by G_τ is the transformation by G_τ of the element chosen among $\mathbf{E}[R^n f]$. Equation (46) proves the vectors selected by the matching pursuit of $G_\tau f$ are the vectors selected for the matching pursuit of f transformed by G_τ and the residues of $G_\tau f$ are equal to the residues of f transformed by G_τ.

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

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

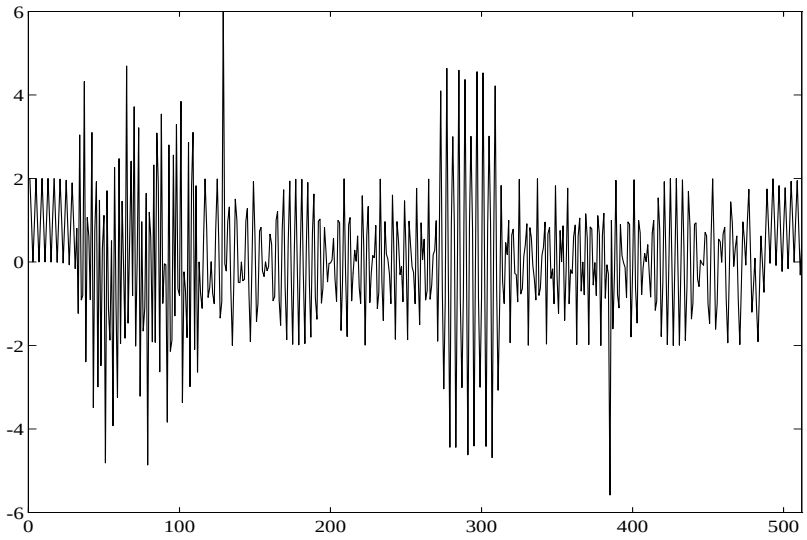

Figure 1: Synthetic signal of 512 samples built by adding $\cos((1-\cos(ax))bx)$, two truncated sinusoids, two Dirac functions, and $\cos(cx)$.

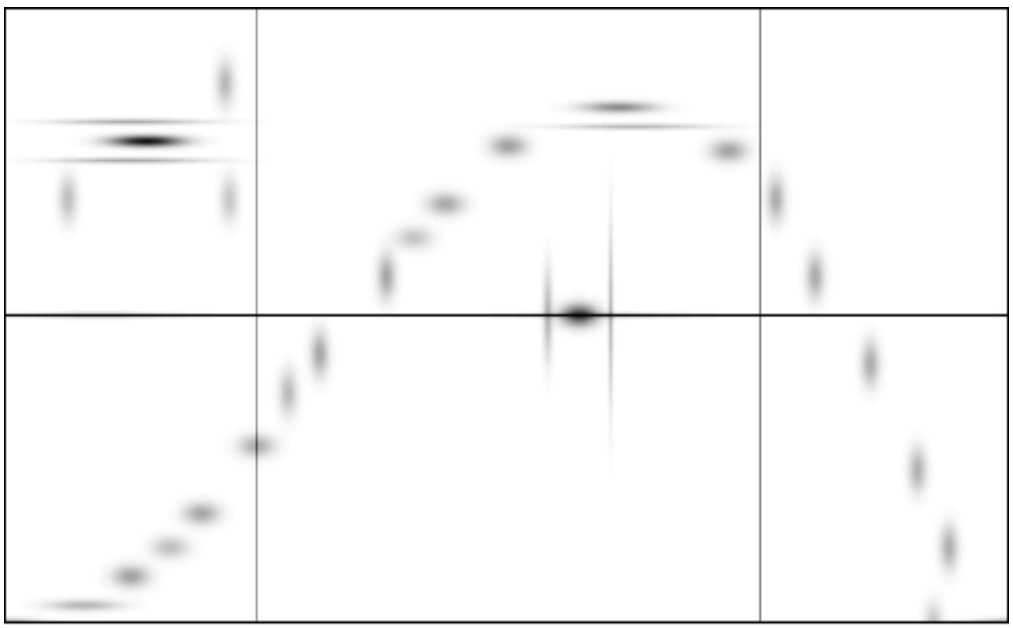

Figure 2: Time frequency energy distribution $Ef(t,\omega)$ of the signal in the figure above. The horizontal axis is time and the vertical axis is frequency. The darkness of the image increases with $Ef(t,\omega)$.

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

$$\hat{g}_\gamma(\omega) = \sqrt{s}\hat{g}(s(\omega - \xi))e^{-i(\omega-\xi)u}, \tag{41}$$

and the Fourier transform 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 vice versa.

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 [7]. 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

$$f = \sum_{n=0}^{+\infty} < R^n f, g_{\gamma_n} > g_{\gamma_n}, \tag{42}$$

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

$$g_{\gamma_n}(t) = \frac{1}{\sqrt{s_n}} g(\frac{t - u_n}{s_n})e^{i\xi_n t} \tag{43}$$

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

To any matching pursuit expansion[23][25], we can associate a time-frequency energy distribution defined by

$$Ef(t,\omega) = \sum_{n=0}^{\infty} |< R^n f, g_{\gamma_n} >|^2 W g_{\gamma_n}(t,\omega), \tag{44}$$

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 [5] 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 $\cos(cx)$. Each Gabor time-frequency atom selected by the matching pursuit is 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.

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.

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

$$u_n = \sum_{p=0}^{n} b_{p,n} g_{\gamma_p} \tag{37}$$

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. Although the Gram-Schmidt formula has poor numerical qualities, we chose to use it for computational simplicity.

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

$$< R^{n+1}f, g_\gamma >=< R^n f, g_\gamma > - < R^n f, g_{\gamma_n} > < u_n, g_\gamma > . \tag{38}$$

Since

$$< u_n, g_\gamma >= \sum_{p=0}^{n} b_{p,n} < g_{\gamma_p}, g_\gamma >, \tag{39}$$

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 $O(P^3 + P^2IZ)$. 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 calculation 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 depending upon the sound, which makes it difficult to find a basis that is *a priori* well adapted to all components of the sound recording. Dictionaries of time-frequency atoms include waveforms with a wide range of time-frequency localization 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

$$g_\gamma(t) = \frac{1}{\sqrt{s}} g(\frac{t-u}{s}) e^{i\xi t}. \tag{40}$$

Inserting this expression into (27) yields

$$f = \sum_{n=0}^{M} \frac{< R^n f, g_{\gamma_n} >}{||u_n||^2} \sum_{p=0}^{n} b_{p,n} g_{\gamma_p}. \tag{33}$$

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

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

The second summation that defines the expansion coefficients over the family $\{g_{\gamma_p}\}_{0 \le p < M}$ can indeed diverge. This happens when the family of selected elements is not a Riesz basis of the closed space it generates. In [13] 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 residues 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. It also requires many more operations to compute because of the Gram-Schmidt orthogonalization procedure. In the next section we compare the complexity of these two algorithms.

3.3 Numerical Implementation of Matching Pursuits

We suppose that $\mathcal{H}$ is a finite dimensional space and $\mathcal{D}$ 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 $\{< f, g_\gamma >\}_{\gamma \in \mathbf{\Gamma}_\alpha}$ and we store these inner products in an open hash table [6], 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 $n \ge 0$. We must first find g_{γ_n} such that

$$| < R^n f, g_{\gamma_n} > | = \sup_{\gamma \in \mathbf{\Gamma}} | < R^n f, g_\gamma > |. \tag{35}$$

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 residue $R^{n+1} f$ with all $g_\gamma \in \mathcal{D}$ using an updating formula derived from equation (11)

$$< R^{n+1} f, g_\gamma > = < R^n f, g_\gamma > - < R^n f, g_{\gamma_n} > < g_{\gamma_n}, g_\gamma > . \tag{36}$$

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

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

$$\lim_{n \to N-1} \|R^n f\| = 0. \tag{26}$$

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 first suppose that $< R^k f, g_{\gamma_k} > \neq 0$ for $k < N$. If not, then we are through, since condition (9) implies that $< R^k f, g_\gamma >= 0$ for all $g_\gamma \in \mathcal{D}$, and because the vectors in $\mathcal{D}$ span $\mathcal{H}$, we must have $R^k f = 0$. Because $R^k f$ is orthogonal to $\{g_{\gamma_p}\}_{0 \leq p < k}$ and $< R^k f, g_{\gamma_k} > \neq 0$, the set $\{g_{\gamma_p}\}_{0 \leq p \leq k}$ must be linearly independent. When N is finite, the N linearly independent vectors $\{g_{\gamma_p}\}_{0 \leq p < N-1}$ form a basis of $\mathcal{H}$, and the orthogonalized vectors $\{u_p\}$ form an orthogonal basis of $\mathcal{H}$. The result follows directly.

When N is infinite, we have from the Bessel inequality that

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

By (9), we must have

$$\lim_{k \to \infty} \sup_{\gamma \in \mathbf{\Gamma}} |< R^k f, g_\gamma >| = 0, \tag{30}$$

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

$$\begin{aligned} \|R^n f - R^m f\|^2 &= \sum_{k=n+1}^{m} \frac{|< f, u_k >|^2}{\|u_k\|^2} \\ &\leq \sum_{k=n+1}^{\infty} \frac{|< f, u_k >|^2}{\|u_k\|^2}, \end{aligned} \tag{31}$$

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 $\{g_{\gamma_p}\}_{0 \leq p \leq k}$

$$u_k = \sum_{p=0}^{n} b_{p,k} g_{\gamma_p}. \tag{32}$$

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 by [8], [24], and [3].

At each iteration, the vector g_{γ_k} selected by the matching algorithm is *a priori* not orthogonal to the previously selected vectors $\{g_{\gamma_p}\}_{0 \leq p < k}$. In subtracting the projection of $R^k f$ over g_{γ_k} the algorithm reintroduces new components in the directions of $\{g_{\gamma_p}\}_{0 \leq p < k}$. This can be avoided by orthogonalizing $\{g_{\gamma_p}\}_{0 \leq p < k}$ 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 residue 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 on 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), similarly to equations (13) and (14), we derive that for any $n > 0$

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

and

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

The theorem below proves that that the residues 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.

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 (11) 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 residue $R^n f$ 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 [32] to be a consequence of a theorem proved by Jones [20]

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

Hence

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

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

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

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

For simplicity, for the remainder of the thesis 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

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

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 residue 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 [32] 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 residue $R^k f$ decreases exponentially with a minimum decay rate equal to $-\frac{1}{2}\log(1-\lambda_{min}^2)$.

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, that maintains an energy conservation relation which guarantees stability.

3 Matching Pursuits

A matching pursuit is a greedy algorithm that progressively refines the signal approximation with an iterative procedure instead of solving the optimal approximation problem [32]. 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 an adaptive time-frequency decomposition.

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 residue Rf is orthogonal to g_{γ_0},

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

We minimize the norm of the residue by choosing g_{γ_0} which maximizes $|< f, g_\gamma >|$. In infinite dimensions, the supremum of $|< f, g_\gamma >|$ may not be attained, so 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 pursuit iterates this procedure by subdecomposing the residue. Let $R^0 f = f$. Suppose that we have already computed the residue $R^k f$. 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

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

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 three-element sets $S_i \in \mathcal{C}$ and M coefficients β_n such that

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

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 bases 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 [13].

The optimal approximation criterion of definition 2.1 has number of undesirable properties which are partly 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. The expansion in M elements with minimal error does not necessarily contain an expansion in $M - 1$ elements with minimal error. The expansions can therefore not 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

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

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 set of M elements and their coefficients. To prove that it is NP-complete we prove that it is as hard as the exact cover by 3-sets problem, 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 $\mathcal{C}$ contain a subcollection $\mathcal{C}'$ such that every member of X occurs in exactly one member of $\mathcal{C}'$? [17]*

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 so the approximation problem must be NP-complete as well. It thus gives a proof of the theorem for $M = \frac{N}{3}$.

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

$$T(S) = \sum_{i \in S} \frac{e_i}{\sqrt{K}}. \tag{2}$$

Let $\mathcal{D}$ be the dictionary of $\mathcal{H}$ 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

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

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

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

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 $\mathcal{D}$. 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 sums 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 functions in an N-dimensional Hilbert space $\mathcal{H}$. Let $\epsilon > 0$ and $M \in \mathbf{N}$. 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 consists of 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 $O(N^2)$.

For general redundant dictionaries, the following theorem proves that finding M-optimal approximations is computationally intractible.

Theorem 2.1 *Let $\mathcal{H}$ be an N dimensional Hilbert space. Let $k \geq 1$ and let $\mathcal{D}_N$ be the set of all dictionaries for $\mathcal{H}$ that contain $O(N^k)$ vectors. Let $0 < \alpha_1 < \alpha_2 < 1$ and $M \in \mathbf{N}$ such that $\alpha_1 N \leq M \leq \alpha_2 N$. The* **$(\epsilon$, 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* **M-optimal approximation problem***, finding the optimal M-approximation, is NP-hard.*

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, we mean 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 decides an NP-hard problem.

Note: in our computations we restrict f, the elements of the dictionaries, and their coefficients to floating point representations of $\Theta(N^m)$ bits for some fixed m [29]. This restriction does not substantially affect the proof of NP-completeness, because the problems that must be solved for the proof are discrete and unaffected by small perturbations.

We can greatly improve these linear 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 [11][12]. 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 one unit, then an accurate approximation can require many elements. One way to improve matters is to add to the set $\{g_\gamma\}_{\gamma\in\mathbf{\Gamma}}$, for example, the collection of all translates of the wavelets. The class of functions which can be compactly represented will then be translation invariant. We can do even better 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 5 we prove that for general dictionaries, the problem of finding M-vector optimal approximations is NP-hard. Because of the numerical intractability of computing optimal expansions, in section 3 we review the performance of greedy algorithms, called matching pursuits, that were introduced in [32] [13]. We also describe a fast numerical 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 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 residues, the invariant measure provides a statistical description of the residues after a sufficient number of iterations of the pursuit. This enables us in section 6 to measure the average asymptotic decay rate of residues, which is often slow. Matching pursuit approximations yield efficient approximations when M is small, though, 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 called "dictionary noise." The properties of invariant measures are studied for the particular case of dictionaries that are invariant under the action of a group of operators, and a stochastic model of the evolution of the residues is developed for a dictionary which is composed of a discrete Dirac basis and discrete Fourier basis.

Adaptive Nonlinear Approximations

Geoffrey Davis, Stéphane Mallat, and Marco Avellaneda

New York University, Courant Institute
251 Mercer Street, New York, NY 10012

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 the dictionary waveforms which best match the function's structures. Matching pursuits provide 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 using a stochastic differential equation model. These invariant measures define a noise with respect to a given dictionary. The dictionary elements selected during the initial iterations of a pursuit correspond to a function's coherent structures. The expansion of a function into its coherent structures provides a compact approximation with a suitable dictionary. 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 H using a small number of vectors from a given 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 $\{g_\gamma\}_{\gamma\in\mathbf{\Gamma}}$ 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, if $\{g_\gamma\}_{\gamma\in\mathbf{\Gamma}}$ is a wavelet basis, the rate of decay of $\epsilon_0(M)$ can be estimated for functions that belong to any particular Besov space. Conversely, the rate of decay of $\epsilon_0(M)$ characterizes the Besov space to which f belongs[10].