Approximation of Functions over Redundant Dictionaries Using Coherence

Anna C. Gilbert* S. Muthukrishnan* Martin J. Strauss*

October 7, 2002

Abstract

One of the central problems of modern mathematical approximation theory is to approximate functions, or *signals*, concisely, with elements from a large candidate set called a *dictionary*. Formally, we are given a signal $\mathbf{A} \in \mathbb{R}^N$ and a dictionary $\mathcal{D} = \{\phi_i\}_{i \in I}$ of unit vectors that span $\mathbb{R}^N$. A representation $\mathbf{R}$ of B terms for input $\mathbf{A} \in \mathbb{R}^N$ is a linear combination of dictionary elements, $\mathbf{R} = \sum_{i \in \Lambda} \alpha_i \phi_i$, for $\phi_i \in \mathcal{D}$ and some Λ, $|\Lambda| \leq B$. Typically, $B \ll N$, so that $\mathbf{R}$ is a concise approximation to signal $\mathbf{A}$. The error of the representation indicates by how well it approximates $\mathbf{A}$, and is given by $\|\mathbf{A} - \mathbf{R}\|_2 = \sqrt{\sum_t |\mathbf{A}[t] - \mathbf{R}[t]|^2}$. The problem is to find the best B-term representation, *i.e.*, find a $\mathbf{R}$ that minimizes $\|\mathbf{A} - \mathbf{R}\|_2$. A dictionary may be redundant in the sense that there is more than one possible *exact* representation for $\mathbf{A}$, *i.e.*, $|\mathcal{D}| > N = \dim(\mathbb{R}^N)$. Redundant dictionaries are used because, both theoretically and in practice, for important classes of signals, as the size of a dictionary increases, the error and the conciseness of the approximations improve.

We present the first known efficient algorithm for finding a provably approximate representation for an input signal over redundant dictionaries. We identify and focus on redundant dictionaries with small coherence (ie., vectors are nearly orthogonal). We present an algorithm that preprocesses any such dictionary in time and space polynomial in $|\mathcal{D}|$, and obtains an $1 + \epsilon$ approximate representation of the given signal in time nearly linear in signal size N and polylogarithmic in $|\mathcal{D}|$; by contrast, most algorithms in the literature require $\Omega(|\mathcal{D}|)$ time, and, yet, provide no provable bounds. The technical crux of our result is our proof that two commonly used local search techniques, when combined appropriately, gives a provably near-optimal signal representation over redundant dictionaries with small coherence. Our result immediately applies to several specific redundant dictionaries considered by the domain experts thus far. In addition, we present new redundant dictionaries which have small coherence (and therefore are amenable to our algorithms) and yet have significantly large sizes, thereby adding to the redundant dictionary construction literature.

Work with redundant dictionaries forms the emerging field of *highly nonlinear approximation theory*. We have presented algorithmic results for some of the most basic problems in this area, but other mathematical and algorithmic questions remain to be explored.

*AT&T Labs—Research, 180 Park Avenue, Florham Park, NJ 07932 USA, {agilbert, muthu, mstrauss}@research.att.com

1 Introduction

1.1 Background on Mathematical Approximation Theory The main problem of *mathematical approximation theory* is to approximate functions compactly, *i.e.*, in small space, or using a "small number of terms." Formally, we are given a signal $\mathbf{A} \in \mathbb{R}^N$ and a dictionary $\mathcal{D} = \{\phi_i\}_{i \in I}$ of unit vectors that span $\mathbb{R}^N$. A representation $\mathbf{R}$ of B terms for input $\mathbf{A} \in \mathbb{R}^N$ is a linear combination of dictionary elements, $\mathbf{R} = \sum_{i \in \Lambda} \alpha_i \phi_i$, for $\phi_i \in \mathcal{D}$ and some Λ, $|\Lambda| \leq B$. Typically, $B \ll N$. The error of the representation indicates how well it approximates $\mathbf{A}$. Following the most common practice, we use ℓ^2 norm: the error of approximation is $\|\mathbf{A} - \mathbf{R}\|_2 = \sqrt{\sum_t |\mathbf{A}[t] - \mathbf{R}[t]|^2}$, henceforth written $\|\mathbf{A} - \mathbf{R}\|$, or (equivalently) its square.

Mathematical approximation theory has applications in numerical computations, *e.g.*, adaptive PDE solvers, audio signal compression, image compression and statistical estimation with its applications to classification. It is a rich area with significant mathematical achievements and successful applications, both classical and novel; see [8], [9] for good surveys, and *The Journal of Approximation Theory* [16] and *Constructive Approximation* [2] for current mathematical developments.

There are two approaches to mathematical approximation theory:

- *Linear Approximation Theory.* We approximate functions using a fixed linear subspace of the dictionary independent of the signal. For example, if the dictionary is the Fourier basis, a B-term approximation of $\mathbf{A}$ is given by the *first* (lowest) B frequencies of its Fourier expansion.
- *Nonlinear Approximation Theory.* We seek the

best or optimal B-term approximation, *i.e.*, $\mathbf{R}$ with $|\Lambda| = B$ such that $\|\mathbf{A} - \mathbf{R}\|$ is minimized. In this setting, the terms used depend on the signal and do not come from a fixed subspace. There are two further variations. In *standard* nonlinear approximation theory, the dictionary is an *orthonormal basis* of size N and each function has a unique exact representation $\mathbf{R}$. It is easy to prove by Parseval's equality that the best B term representation comprises the B largest $|\langle \mathbf{A}, \phi \rangle|$ over $\phi \in \mathcal{D}$. In *highly* nonlinear approximation theory, the dictionary is *redundant* and is larger than N. Signals have more than one exact representation over $\mathcal{D}$ and the best collection of B terms need not be the B largest.

Generally, one wants to relate the quality of a representation to other parameters, such as B, the number of terms in the representation (a central mathematical question) or computational cost (in terms of N, B and $|\mathcal{D}|$, a central algorithmic question). There is considerable mathematical analysis of linear and standard nonlinear approximation theory and the algorithmic issues in both are fairly straightforward. Generalizing from linear to standard non-linear to highly-non-linear theory yields much better representations, but pose much harder mathematical and algorithmic challenges. "*Redundancy on the one hand offers much promise for greater efficiency in terms of approximation rate, but on the other hand, gives rise to highly nontrivial theoretical and practical problems*," (quoted from [23]). In fact, the mathematics of the theory of highly nonlinear approximation is only now emerging, and algorithmic complexity of these problems is wide open. The subject of this paper is the central algorithmic issue in highly non-linear approximation theory: for a given function $\mathbf{A}$, parameter B and a prespecified dictionary $\mathcal{D}$, finding the optimal representation of B terms.

1.2 State of the Art The general highly-non-linear approximation problem of our interest is NP-hard [7]. In fact, the proof there implicitly shows that the problem is NP-hard even to determine whether the optimal error is zero for the given B; hence, unless P=NP, no polynomial time algorithm exists that approximates the best B-term representation over an arbitrary dictionary even if we wish only to approximate the optimal error by a factor, however large.

As a consequence, research in nonlinear approximation theory has mostly focused on specific dictionaries. Mathematicians have studied specific dictionaries—spikes and sinusoids [4], wavelet packets [5, 25, 24], frames [6], time/frequency tilings with algebraic hierarchical structure [19]—and presented individual algorithms (some provable, some heuristic) for constructing the best B-term approximation for any given function. Also, certain "mathematical recipes" such as algorithms for infinite dimensional Hilbert and Banach spaces have been proposed [22], [23]. Save these exceptions, no provable, algorithmic results—running time for finding provably approximate B-term representations for inputs over general redundant dictionaries—are currently known.

1.3 Local Search Heuristics Most of the heuristic algorithms for general redundant dictionaries are local search methods. We briefly summarize them because they are relevant to our results. There are two common approaches.

- Find the j such that $\phi = \phi_j$ maximizes $\langle \mathbf{A} - \mathbf{R}, \phi \rangle$ over $\phi \in \mathcal{D}$. Update $\mathbf{R}$ by $\mathbf{R} \leftarrow \mathbf{R} + \langle \mathbf{A} - \mathbf{R}, \phi \rangle \phi$, and repeat a total of B times. More generally, find the set Λ of $B' \leq B$ of j's with the largest dot products, and put $\mathbf{R} \leftarrow \mathbf{R} + \sum_{j \in \Lambda} \langle \mathbf{A}, \phi_j \rangle \phi_j$. We call this technique B'-fold *matching pursuit*, B'-MP, for $B' \geq 1$.
- Maintain a small subset $\Lambda \subseteq \mathcal{D}$. Find the j such that $\phi = \phi_j$ maximizes $\langle \mathbf{A} - \mathbf{R}, \phi \rangle$ over $\phi \in \mathcal{D}$, and update $\Lambda \leftarrow \Lambda \cup \{\phi_j\}$. Update $\mathbf{R}$ to be the optimal representation of the signal over the subspace spanned by $\{\phi : \phi \in \Lambda\}$, and repeat this process a total of B times. This technique is called *orthogonal matching pursuit*, OMP.

Both these approaches have been studied extensively in the literature (MP appears in [18] and OMP appears in [21]). However, there are two drawbacks in the state of art, as follows:

The first issue is of proving bounds on the error of the solution.[1] If the dictionary is an orthonormal basis, then the local search techniques above are equivalent and provably find the global optimum. In general, however, these methods do not provide any useful approximation results for finding the best representation for input functions (these are polynomial time heuristics for the NP-hard problem). Such an approximation result is not known even for special redundant dictionaries studied in the literature.

The second issue is of providing an efficient implementation, in particular, for performing each step of

[1] In the mathematical literature, the usual guarantee proved—if any bound at all is given—is that the error of the representation drops off at a prescribed rate as a function of B, for all signals in a given input class. From an algorithmic viewpoint such as ours, we wish to compare the error of the algorithms' output to the error of the optimal representation, whatever the optimal error may be, for a given signal and B. This viewpoint appears to be novel in mathematical approximation theory literature.

the local search. All known algorithms require at least $\Omega(|\mathcal{D}|)$ time. Indeed, the general problem of finding the ϕ from the dictionary $\mathcal{D}$ that has the *largest* inner product with the signal is equivalent to the Farthest Neighbors Problem, which faces the same "dimensionality bottleneck" as the Nearest Neighbors Problem [14]–[15], [17]. Designing efficient algorithms for this problem is still an open challenge, in particular, since the local search heuristics rely on finding the optimal such ϕ's. This is a serious drawback because the community of researchers in highly nonlinear approximation theory are applied mathematicians who care about practical, implementable algorithms.

1.4 Our Contributions One might be tempted to study the problem assuming the dictionary is arbitrary so as to be general. However, an arbitrary dictionary is not a well-motivated choice. We would like to make the dictionaries as large as possible to achieve as high a rate of compression as possible; however, for specific classes of applications (*e.g.*, encoding low-quality speech signals, compressing high motion video), the salient features of these signals are well-described by specific inherently suitable dictionaries. These dictionaries exhibit certain a naturalness for these applications such as smoothness, oscillatory behavior, etc. Redundant dictionaries must, therefore, balance the two needs—naturalness and succinctness of vectors on one hand and size on the other—and designing them is sometimes an art. We would like our algorithmic results to apply to largest number of redundant dictionaries proposed by domain experts in this area.

Further, it is natural to consider the scenario when the dictionary is provided ahead of time for preprocessing. The dictionary is typically large, say, of size $\Omega(N)$. Our goal is to process the input signal and determine a (near) optimal representation very efficiently. In particular, we would prefer to avoid scanning the dictionary entirely while processing the input signal.

We identify and focus on redundant dictionaries of small coherence. A μ-coherent dictionary $\mathcal{D}$ has *coherence* μ, for $0 \leq \mu \leq 1$, if $|\langle \phi_1, \phi_2 \rangle| \leq \mu$ for all distinct $\phi_1, \phi_2 \in \mathcal{D}$. A typical way to generate redundant dictionaries in practice is to take several orthonormal bases and combine them, and one gets the most out of this combination if the bases were as orthogonal to each other as possible. (Using sinusoids and spikes, as in [11], is an example.) Dictionaries generated in this manner have small coherence, thus motivating our work here. Coherence has been discussed as an important notion in the context of local search heuristics in the literature [18], but we appear to the first to formalize the concept of coherent dictionaries and study the algorithmic problem of representing function near-optimally over μ-coherent dictionaries.

We make two contributions. Recall that the problem is, given input signal $\mathbf{A}$ and parameter B, determine the optimal B-term representation $\mathbf{R}$ over dictionary $\mathcal{D}$ of coherence μ, *i.e.*, such that $\|\mathbf{A} - \mathbf{R}\|$ is minimized.

- We present an algorithm that for any ϵ, $\mu B^2 = O(\epsilon)$, finds an $(1+\epsilon)$-factor approximation to the optimal representation in time[2] $N(B\log(|\mathcal{D}|)/\epsilon)^{O(1)}$; the dictionary preprocessing takes time and space polynomial in $|\mathcal{D}|$.

  This is the first known provably approximate result for approximating a function under *any* redundant dictionary with small coherence $O(\epsilon/B^2)$. In addition, it is very fast, taking time nearly linear in the input size modulo the polylogarithmic factor. This is an exponential speedup over previous heuristics that take time $\Omega(|\mathcal{D}|)$ (recall that for redundant dictionaries, $|\mathcal{D}|$ dominates N). For the dictionary which consists of spikes and sinusoids, our approach specializes to give an algorithm that finds a near-optimal representation in $O(N^2)$ time, improving the previous best [4] of $O(N^{3.5})$ time via Linear Programming.

- We explore the concept of μ-coherent redundant dictionaries further, adding to the growing knowledge of redundant dictionary construction [3], [10]–[11].

  - Known redundant dictionaries are of size only $O(N)$ or $O(N\log(N))$; it is desirable to construct larger dictionaries if "natural" properties of a specific application can still be captured. We consider image processing applications and propose a dictionary of significantly larger size: $N^{3/2}/B^6$, that we call *segmentlets*. This generalizes a number of known natural constructions in that area [3], [10]–[11], such as beamlets, wedgelets, and ridgelets. By design, segmentlets are μ-coherent dictionaries for small μ and hence our main algorithmic result applies to them too.
  - Without focusing on any application, we focus on constructing large redundant dictionaries of small coherence. Using the Nisan-Wigderson [20] combinatorial design, we obtain such a dictionary of size exponential in

[2] We assume that entries in $\mathbf{A}$ are bounded by $N^{O(1)}$. The general result requires an additional factor in time and space polynomial in the number of bits in entries of $\mathbf{A}$.

N. We note that there are other combinatorial design that provide such large dictionaries with small coherence. However, this one can be constructed using small-space which may prove valuable for some application. Our motivation for including the Nisan-Wigderson combinatorial design was primarily to show that even though we focus on small conherence, nontrivial and exponentially large dictionaries can still be constructed.

The first category of result above is our main result, presented in §2. All our results on specific redundant dictionaries can be found in §3.

1.5 Technical Overview Our algorithm is two-phased local search, OMP followed by MP. The crux of our technical contribution is our proof that in $O(B/\epsilon)$ iterations, our algorithm converges to within $1+\epsilon$ of the optimal representation. At each iteration, this requires one to determine the dictionary element ϕ_j whose dot product has magnitude at least $(1-\eta)$ times the largest dot product magnitude, and it suffices to estimate $\langle \mathbf{A}, \phi_j \rangle$ to within $\eta \|\mathbf{A}\|$, additively, for some appropriate η. We are able to implement each such iteration using the approximate nearest neighbors data structures that have been designed recently. That gives the overall result.

There are some similarities between this work and our earlier work on finding near-optimal representations of functions over Fourier [13] and Haar wavelet bases [12], but also some crucial differences. Both Fourier and wavelet bases are *non*-redundant dictionaries, and so efficient algorithms already existed for finding even optimal representations over them. The emphasis in the previous work was on working under additional constraints: using only small (polylogarithmic) space and time, a la sampling or streaming. In our case here with redundant dictionaries, even given polynomial time and space, no previous result was known for optimal (or near-optimal) representation of functions. We do not focus on polylogarthmic space/time models, and leave that additional complication open. From a technical point of view, either OMP or MP by itself will suffice if we wanted to specialize our result for non-redundant basis such as Fourier or Haar wavelet; furthermore, one would not need approximate nearest neighbors. Thus, both the proof of approximation as well as the algorithms are more sophisticated for redundant dictionaries.

2 Representation over Redundant Dictionaries

We consider vectors over $\mathbb{R}^N$ for simplicity; everything in this paper can be extended to $\mathbb{C}^N$ in an obvious way.

2.1 Small Coherence Dictionary

DEFINITION 2.1. *A set $\mathcal{D} = \{\phi_i\}$ of elements from $\mathbb{R}^N$ is a* dictionary with coherence μ *if* $\text{span}(\mathcal{D}) = \mathbb{R}^N$, $\|\phi_i\| = 1$ *for all i, and, for all distinct i and j,* $|\langle \phi_i, \phi_j \rangle| \leq \mu$.

The definition of dictionary gives a condition only on pairs of dictionary elements. The following lemma gives an implication to larger sets of vectors.

LEMMA 2.1. *Let $\phi_0, \phi_1, \ldots, \phi_B$ be an arbitrary set of $B+1$ vectors from a dictionary of coherence μ. We have*

1. *If $\mu B < 1$, then $\{\phi_1, \ldots, \phi_B\}$ is independent.*
2. *If $\mu B < 1/2$, then the projection of ϕ_0 onto the span of the other B vectors has norm at most $\sqrt{2\mu^2 B}$.*
3. *If $\mu B < 1/2$, then there exists a set $\{\psi_i : i = 1, \ldots, B\}$ of vectors such that:*
   (a) *The ψ's form an orthonormal system.*
   (b) $\text{span}(\psi_1, \ldots, \psi_B) = \text{span}(\phi_1, \ldots, \phi_B)$.
   (c) $\|\psi_i - \phi_i\|^2 \leq 8\mu^2 B$.

Proof. First consider taking any linear combination of $\phi_1, \ldots, \phi_B$. We have $\left\|\sum_{i=1}^B \alpha_i \phi_i\right\|^2 \geq \sum_{i=1}^B \alpha_i^2 - \sum_{i \neq j} |\alpha_i \alpha_j \langle \phi_i, \phi_j \rangle|$, and $\sum_{i \neq j} |\alpha_i \alpha_j \langle \phi_i, \phi_j \rangle| \leq \mu \sum_{i \neq j} |\alpha_i \alpha_j| \leq \mu \left(\sum_{i=1}^B |\alpha_i|\right)^2 \leq \mu B \sum_{i=1}^B \alpha_i^2$ by the Cauchy-Schwarz inequality, so

$$\left\|\sum_{i=1}^B \alpha_i \phi_i\right\|^2 \geq (1-\mu B) \sum_{i=1}^B \alpha_i^2. \tag{2.1}$$

This gives the first statement.

The length of the projection π_0 of ϕ_0 onto $\text{span}(\phi_1, \ldots, \phi_B)$ is equal to the dot product $\langle \pi_0, u_0 \rangle$ of π_0 with the unit vector u_0 along π_0, which also equals $\langle \phi_0, u_0 \rangle$. By the Cauchy-Schwarz inequality, this is the maximum, over all unit vectors $u \in \text{span}(\phi_1, \ldots, \phi_B)$, of $\langle \phi_0, u \rangle$. Write $u = \frac{\sum_i \alpha_i \phi_i}{\|\sum_i \alpha_i \phi_i\|}$ where $\sum_i \alpha_i^2 = 1$; note that the denominator is at least $\sqrt{1-\mu B}$, by the above. Thus $\|\pi_0\|^2$ is at most $\max_{\|u\|^2=1} \langle u, \phi_0 \rangle^2$, which is at most $\max_{\sum_i \alpha_i^2 = 1} \frac{\langle \sum_i \alpha_i \phi_i, \phi_0 \rangle^2}{\|\sum_i \alpha_i \phi_i\|^2} \leq \max_{\sum_i \alpha_i^2 = 1} \frac{\langle \sum_i \alpha_i \phi_i, \phi_0 \rangle^2}{1-\mu B} \leq \max_{\sum_i \alpha_i^2 = 1} \frac{(\sum_i \mu \alpha_i)^2}{1-\mu B} \leq \frac{\mu^2 B}{1-\mu B} \leq 2\mu^2 B$, using the Cauchy-Schwarz inequality. This gives the second statement.

As for the third statement, let ψ_1' be $\phi_1 - \pi_1$, where π_1 is the projection of ϕ_1 onto the span of $\{\phi_2, \ldots, \phi_B\}$. Thus ψ_1' is orthogonal to $\{\phi_2, \ldots, \phi_B\}$ and

$$\text{span}(\psi_1', \phi_2, \phi_3, \ldots, \phi_B) = \text{span}(\phi_1, \phi_2, \phi_3, \ldots, \phi_B).$$

By the above, $\|\psi_1' - \phi_1\|^2 = \|\pi_1\|^2 \leq 2\mu^2 B$. Let $\psi_1 = \frac{\psi_1'}{\|\psi_1'\|}$. Since $\|\psi_1'\|^2 + \|\pi_1\|^2 = \|\phi_1\|^2 = 1$, it follows that $1 - 2\mu^2 B \leq \|\psi_1'\|^2 \leq 1$, so $\sqrt{1-2\mu^2 B} \leq \|\psi_1'\| \leq 1$, and $\|\psi_1 - \psi_1'\| \leq 1 - \sqrt{1-2\mu^2 B} \leq 2\mu^2 B$. Since $\|\psi_1' - \phi_1\| \leq \sqrt{2\mu^2 B}$, it follows that $\|\psi_1 - \phi_1\| \leq \sqrt{2\mu^2 B} + 2\mu^2 B \leq 2\sqrt{2\mu^2 B}$, using the fact that $x^2 \leq x$ for $0 \leq x \leq 1$.

Recursively find $\psi_2, \ldots, \psi_B$ that are orthogonal, have the same span as $\phi_2, \ldots, \phi_B$, and such that $\|\psi_i - \phi_i\|^2 \leq 8\mu^2(B-1) \leq 8\mu^2 B$.

Given $\{\phi_i\} \subseteq \mathcal{D}$ as above, we say that $\{\psi_i\}$ is an *orthogonalization* of $\{\phi_i\}$.

2.2 Overall Algorithm Our overall algorithm is as follows. Starting with the zero representation $\mathbf{R}_1 = 0$, we perform OMP (the "error phase") until $\|\mathbf{A} - \mathbf{R}_1\|^2 \leq 64B \|\mathbf{A} - \mathbf{R}_{\text{opt}}\|^2$. Suppose that $\mathbf{R}_1$ consists of $B' \leq B$ terms. We then perform a single round of MP (the "energy phase") to find a near-best $(B-B')$-term representation $\mathbf{R}_2$ to $\mathbf{A} - \mathbf{R}_1$. We then output $\mathbf{R}_1 + \mathbf{R}_2$.

More quantitatively, we proceed as follows. Below, we first consider the energy phase. That is, given a signal $\mathbf{A}$, we show how to find a representation $\mathbf{R}$ for $\mathbf{A}$ with square error $\|\mathbf{A} - \mathbf{R}\|^2 \leq \|\mathbf{A} - \mathbf{R}_{\text{opt}}\|^2 + \epsilon' \|\mathbf{A}\|^2$; *i.e.*, worse than optimal by a small multiple of the *energy* of the signal. Our ultimate goal is the stronger statement, $\|\mathbf{A} - \mathbf{R}\|^2 \leq (1+\epsilon) \|\mathbf{A} - \mathbf{R}_{\text{opt}}\|^2 = \|\mathbf{A} - \mathbf{R}_{\text{opt}}\|^2 + \epsilon \|\mathbf{A} - \mathbf{R}_{\text{opt}}\|^2$, where ϵ and ϵ' depend on B and μ. To do that, we show a modest *error* result (that is used first in the overall algorithm): given a signal $\mathbf{A}$, we can find a representation $\mathbf{R}$ to $\mathbf{A}$ with square error $\|\mathbf{A} - \mathbf{R}\|^2 \leq 64B \|\mathbf{A} - \mathbf{R}_{\text{opt}}\|^2$. That is, the square error of $\mathbf{R}$ is at most a moderate multiple of the *optimal square error* to the signal. Finally, combining these results, letting $\delta = \|\mathbf{A} - \mathbf{R}_{\text{opt}}\|$ denote the optimal square error, we first reduce the square error to $64B\delta^2$, then, by representing the residual, we reduce the square error to be additively suboptimal by $\epsilon'(64B\delta^2)$. The result will have square error $(1 + 64B\epsilon')\delta^2 = (1+\epsilon)\delta^2$, by definition of ϵ.

2.3 Algorithm Details

2.3.1 Energy Phase We first show that we can roughly compare the error of two candidate representations for $\mathbf{A}$, $\sum_{i \in \Lambda_1} \alpha_i \phi_i$ and $\sum_{i \in \Lambda_2} \beta_i \phi_i$ by comparing $\sum_{i \in \Lambda_1} \langle \mathbf{A}, \phi_i \rangle$ with $\sum_{i \in \Lambda_2} \langle \mathbf{A}, \phi_i \rangle$.

LEMMA 2.2. *Suppose $\mu B \leq 1$. Let $\mathbf{R}_1 = \sum_{i \in \Lambda_1} \alpha_i \phi_i$ and $\mathbf{R}_2 = \sum_{i \in \Lambda_2} \beta_i \phi_i$ be two B-term representations for $\mathbf{A}$, such that $\mathbf{R}_j$ is the optimal representation in $\text{span}\left(\bigcup_{i \in \Lambda_j} \phi_i\right)$, $j = 1, 2$. If $\sum_{i \in \Lambda_1} \langle \mathbf{A}, \phi_i \rangle^2 \geq \sum_{i \in \Lambda_2} \langle \mathbf{A}, \phi_i \rangle^2$, then*

$$\|\mathbf{A} - \mathbf{R}_1\|^2 \leq \|\mathbf{A} - \mathbf{R}_2\|^2 + 32\mu B \|\mathbf{A}\|^2.$$

Proof. Let $\{\psi_i : i \in \Lambda_1\}$ be an orthogonalization of $\{\phi_i : i \in \Lambda_1\}$. Observe that $\mathbf{R}_j = \sum_{i \in \Lambda_j} \langle \mathbf{A}, \psi_i \rangle \psi_i$, since each side is the unique best representation over $\text{span}(\{\phi_i : i \in \Lambda_j\}) = \text{span}(\{\psi_i : i \in \Lambda_j\})$. Since $\|\mathbf{A} - \mathbf{R}_j\|^2 = \|\mathbf{A}\|^2 - \sum_{i \in \Lambda_j} \langle \mathbf{A}, \psi_i \rangle^2$, it suffices to show that $\left| \sum_{i \in \Lambda_j} \langle \mathbf{A}, \psi_i \rangle^2 - \sum_{i \in \Lambda_j} \langle \mathbf{A}, \phi_i \rangle^2 \right| \leq 16\mu B \|\mathbf{A}\|^2$.

Proceeding, we have

$$\begin{aligned}
&\sum_{i \in \Lambda_j} |\langle \mathbf{A}, \psi_i \rangle|^2 \\
&\quad = \sum_{i \in \Lambda_j} |\langle \mathbf{A}, \phi_i \rangle + \langle \mathbf{A}, \psi_i - \phi_i \rangle|^2 \\
&\quad \geq \sum_{i \in \Lambda_j} \left(|\langle \mathbf{A}, \phi_i \rangle| - |\langle \mathbf{A}, \psi_i - \phi_i \rangle| \right)^2 \\
&\quad \geq \sum_{i \in \Lambda_j} |\langle \mathbf{A}, \phi_i \rangle|^2 - 2 \sum_{i \in \Lambda_j} |\langle \mathbf{A}, \phi_i \rangle| \, |\langle \mathbf{A}, \psi_i - \phi_i \rangle|.
\end{aligned}$$

A bound for the last term will be reused below, so we isolate it. By Lemma 2.1, it is at most $2\sqrt{8\mu^2 B} \|\mathbf{A}\| \sum_{i \in \Lambda_j} |\langle \mathbf{A}, \phi_i \rangle|$, which, by the Cauchy-Schwarz inequality, is at most $2\sqrt{8\mu^2 B} \|\mathbf{A}\| \sqrt{B \sum_{i \in \Lambda_j} |\langle \mathbf{A}, \phi_i \rangle|^2}$. Continuing,

$$\begin{aligned}
&\sqrt{B \sum_{i \in \Lambda_j} |\langle \mathbf{A}, \phi_i \rangle|^2} \\
&\quad \leq \sqrt{B \sum_{i \in \Lambda_j} \left(|\langle \mathbf{A}, \psi_i \rangle| + |\langle \mathbf{A}, \phi_i - \psi_i \rangle| \right)^2} \\
&\quad \leq \sqrt{2B \sum_{i \in \Lambda_j} \left(|\langle \mathbf{A}, \psi_i \rangle|^2 + |\langle \mathbf{A}, \phi_i - \psi_i \rangle|^2 \right)} \\
&\quad \leq \sqrt{2B \|\mathbf{A}\|^2 + 8\mu^2 B^3 \|\mathbf{A}\|^2},
\end{aligned}$$

so that $2 \sum_{i \in \Lambda_j} |\langle \mathbf{A}, \phi_i \rangle| \, |\langle \mathbf{A}, \psi_i - \phi_i \rangle| \leq 8\mu B \|\mathbf{A}\|^2 \sqrt{1 + 4\mu^2 B^2} \leq 8\mu B \sqrt{2} \|\mathbf{A}\|^2$, and $\sum_{i \in \Lambda_j} \langle \mathbf{A}, \psi_i \rangle^2 \geq \sum_{i \in \Lambda_j} \langle \mathbf{A}, \phi_i \rangle^2 - 16\mu B \|\mathbf{A}\|^2$.

Similarly, $\sum_{i \in \Lambda_j} |\langle \mathbf{A}, \psi_i \rangle|^2 \leq \sum_{i \in \Lambda_j} |\langle \mathbf{A}, \phi_i \rangle|^2 + 2 \sum_{i \in \Lambda_j} |\langle \mathbf{A}, \phi_i \rangle| \cdot |\langle \mathbf{A}, \psi_i - \phi_i \rangle| + \sum_{i \in \Lambda_j} |\langle \mathbf{A}, \psi_i - \phi_i \rangle|^2$,

which is at most $\sum_{i\in\Lambda_j} |\langle \mathbf{A}, \phi_i\rangle|^2 + 8\mu B\sqrt{2}\,\|\mathbf{A}\|^2 + 8\mu^2 B^2\,\|\mathbf{A}\|^2 \le \sum_{i\in\Lambda_j} |\langle \mathbf{A}, \phi_i\rangle|^2 + 16\mu B\,\|\mathbf{A}\|^2$, since $\mu B \le 1/2$. The result follows.

Algorithmically, one finds the B vectors $\phi \in \mathcal{D}$ with largest dot products to $\mathbf{A}$, then finds the best representation $\mathbf{R}$ to $\mathbf{A}$ over the span of those vectors (for example, by orthogonalizing the vectors first). The resulting representation $\mathbf{R}$ will have square error $\|\mathbf{A}-\mathbf{R}\|^2 \le \|\mathbf{A}-\mathbf{R}_{\text{opt}}\|^2 + \epsilon\,\|\mathbf{A}\|^2$, for $\epsilon = 32\mu B$.

COROLLARY 2.1. *For a signal $\mathbf{A}$, a single iteration of B-fold Matching Pursuit over a μ-coherent dictionary $\mathcal{D}$ returns a representation $\mathbf{R}$ with $\|\mathbf{A}-\mathbf{R}\|^2 \le \|\mathbf{A}-\mathbf{R}_{\text{opt}}\|^2 + 32\mu B\,\|\mathbf{A}\|^2$.*

2.3.2 Error Phase Lemma 2.4 below says that if we have a representation $\mathbf{R}$ whose error is significantly worse than optimal (roughly, error $\sqrt{B}\delta$ compared with optimal value of δ), then the ϕ with largest dot product represents so much of the signal that we are forced to take it in any optimal or near-optimal representation. This principle will be used, later, to show correctness of a greedy algorithm.

First we prove a lemma that will be useful.

LEMMA 2.3. *Fix $B > 0$. Let $\mathbf{A}$ be a signal. Let Λ be a set of size less than B, and let $\mathbf{R} = \sum_{i\in\Lambda} \alpha_i\phi_i$ be the best representation for $\mathbf{A}$ over $\{\phi_i : i \in \Lambda\}$. Let $\mathbf{R}_{\text{opt}} = \sum_{i\in\Lambda_{\text{opt}}} \alpha_i^{\text{opt}}\phi_i$ be the best B-term representation for $\mathbf{A}$ over all $\mathcal{D}$, subject to $\Lambda \subseteq \Lambda_{\text{opt}}$. (Note that α_i does not necessarily equal α_i^{opt}.) If $\|\mathbf{A}-\mathbf{R}\|^2 \ge (1+\epsilon)\,\|\mathbf{A}-\mathbf{R}_{\text{opt}}\|^2$ for $64\mu^2B^2 \le \epsilon < 1$, then there exists $i \in \Lambda_{\text{opt}} \setminus \Lambda$ such that*

$$\langle \mathbf{A}-\mathbf{R}, \phi_i\rangle^2 \ge \frac{\epsilon}{8B}\,\|\mathbf{A}-\mathbf{R}\|^2 .$$

Proof. Note that $(1+\epsilon) > (1-\epsilon/2)^{-1}$, so the hypothesis $\|\mathbf{A}-\mathbf{R}\|^2 \ge (1+\epsilon)\,\|\mathbf{A}-\mathbf{R}_{\text{opt}}\|^2$ implies $(1-\epsilon/2)\,\|\mathbf{A}-\mathbf{R}\|^2 \ge \|\mathbf{A}-\mathbf{R}_{\text{opt}}\|^2$, so that $\|\mathbf{A}-\mathbf{R}\|^2 - \frac{\epsilon}{2}\,\|\mathbf{A}-\mathbf{R}\|^2 \ge \|\mathbf{A}-\mathbf{R}_{\text{opt}}\|^2$ and $\|\mathbf{A}-\mathbf{R}\|^2 - \|\mathbf{A}-\mathbf{R}_{\text{opt}}\|^2 \ge \frac{\epsilon}{2}\,\|\mathbf{A}-\mathbf{R}\|^2$.

Let $\{\psi_i\}$ be an orthogonalization of $\{\phi_i : \phi_i \in \Lambda_{\text{opt}}\}$ extending an orthogonalization of $\{\phi_i : \phi_i \in \Lambda\}$. Then $\mathbf{R}_{\text{opt}}$ is the best representation over $\{\psi_i : i \in \Lambda_{\text{opt}}\}$ and $\mathbf{R}$ is the best representation over $\{\psi_i : i \in \Lambda\}$. Furthermore, by orthogonality, $\mathbf{R}_{\text{opt}} = \sum_{i\in\Lambda_{\text{opt}}} \langle \mathbf{A}, \psi_i\rangle\,\psi_i$ and $\mathbf{R} = \sum_{i\in\Lambda} \langle \mathbf{A}, \psi_i\rangle\,\psi_i$, using the same coefficients as $\mathbf{R}_{\text{opt}}$.

By Parseval's equality and orthogonality,

$$\|\mathbf{A}-\mathbf{R}\|^2 - \|\mathbf{A}-\mathbf{R}_{\text{opt}}\|^2 = \|\mathbf{R}_{\text{opt}}-\mathbf{R}\|^2$$

which equals $\sum_{i\in\Lambda_{\text{opt}}\setminus\Lambda} \langle \mathbf{A}, \psi_i\rangle^2$, so it follows that, for some $i \in \Lambda_{\text{opt}} \setminus \Lambda$, we have $\langle \mathbf{A}, \psi_i\rangle^2 \ge \frac{\|\mathbf{A}-\mathbf{R}\|^2 - \|\mathbf{A}-\mathbf{R}_{\text{opt}}\|^2}{B} \ge \frac{\epsilon}{2B}\,\|\mathbf{A}-\mathbf{R}\|^2$. Since $\mathbf{R}$ is orthogonal to ψ_i, it follows that $\langle \mathbf{A}-\mathbf{R}, \psi_i\rangle^2 = \langle \mathbf{A}, \psi_i\rangle^2 \ge \frac{\epsilon}{2B}\,\|\mathbf{A}-\mathbf{R}\|^2$.

Now, $|\langle \phi_i - \psi_i, \mathbf{A}-\mathbf{R}\rangle| \le \|\phi_i-\psi_i\| \cdot \|\mathbf{A}-\mathbf{R}\| \le \left(\sqrt{8\mu^2 B}\right)\|\mathbf{A}-\mathbf{R}\| \le \left(\sqrt{\frac{\epsilon}{8B}}\right)\|\mathbf{A}-\mathbf{R}\|$, so $|\langle \mathbf{A}-\mathbf{R}, \phi_i\rangle| \ge |\langle \mathbf{A}-\mathbf{R}, \psi_i\rangle| - |\langle \mathbf{A}-\mathbf{R}, \psi_i-\phi_i\rangle| \ge \left(\sqrt{\frac{\epsilon}{2B}} - \sqrt{\frac{\epsilon}{8B}}\right)\|\mathbf{A}-\mathbf{R}\| = \sqrt{\frac{\epsilon}{8B}}\,\|\mathbf{A}-\mathbf{R}\|$, and the result follows.

LEMMA 2.4. *Let $\mathbf{A}$ be a signal and let $\mathbf{R} = \sum_{i\in\Lambda} \alpha_i\phi_i$ be the best representation for $\mathbf{A}$ over Λ of size less than B. Suppose there's a set $\Lambda_{\text{opt}} \supseteq \Lambda$ with $|\Lambda_{\text{opt}}| = B$, and a representation $\mathbf{R}_{\text{opt}} = \sum_{i\in\Lambda_{\text{opt}}} \beta_i\phi_i$ for $\mathbf{A}$ such that $\|\mathbf{A}-\mathbf{R}\|^2 > 64B\,\|\mathbf{A}-\mathbf{R}_{\text{opt}}\|^2$. Finally, assume $8\mu^2(B+1) < \frac{1}{64B}$. If $\phi = \phi_i$ maximizes $\langle \mathbf{A}-\mathbf{R}, \phi\rangle$ over $\phi \in \mathcal{D}$, then $i \in \Lambda_{\text{opt}}$.*

Proof. Suppose not. By Lemma 2.3 with $\epsilon = 1/2$, since ϕ_i is maximal, $\langle \mathbf{A}-\mathbf{R}, \phi_i\rangle^2 \ge \frac{1}{16B}\,\|\mathbf{A}-\mathbf{R}\|^2$. Let $\{\psi_j\}$ denote an orthogonalization of $\{\phi_j : j \in \Lambda_{\text{opt}} \cup \{i\}\}$. Then

$$\begin{aligned}
\langle \mathbf{A}-\mathbf{R}, \psi_i\rangle &\ge \langle \mathbf{A}-\mathbf{R}, \phi_i\rangle - \|\mathbf{A}-\mathbf{R}\|\,\|\phi_i-\psi_i\| \\
&\ge \left(\sqrt{\frac{1}{16B}} - \sqrt{8\mu^2(B+1)}\right)\|\mathbf{A}-\mathbf{R}\| \\
&\ge \left(\sqrt{\frac{1}{16B}} - \sqrt{\frac{1}{64B}}\right)\|\mathbf{A}-\mathbf{R}\| ,
\end{aligned}$$

so that $\langle \mathbf{A}-\mathbf{R}, \psi_i\rangle^2 \ge \frac{1}{64B}\,\|\mathbf{A}-\mathbf{R}\|^2 > \|\mathbf{A}-\mathbf{R}_{\text{opt}}\|^2$. On the other hand, since ψ_i is orthogonal to $\Lambda_{\text{opt}} \ni \text{span}(\Lambda_{\text{opt}} \setminus \Lambda)$, it follows that $\langle \mathbf{A}-\mathbf{R}, \psi_i\rangle = \langle \mathbf{A}-\mathbf{R}_{\text{opt}}, \psi_i\rangle + \langle \mathbf{R}_{\text{opt}}-\mathbf{R}, \psi_i\rangle = \langle \mathbf{A}-\mathbf{R}_{\text{opt}}, \psi_i\rangle$. Also, by optimality of $\mathbf{R}_{\text{opt}}$, $\mathbf{A}-\mathbf{R}_{\text{opt}}$ is orthogonal to $\mathbf{R}_{\text{opt}}$, so that $\|\mathbf{A}-\mathbf{R}_{\text{opt}}\|^2 = \|\mathbf{A}\|^2 - \|\mathbf{R}_{\text{opt}}\|$. Finally, since ψ_i is orthogonal to $\Lambda_{\text{opt}} \supseteq \Lambda$, $\Lambda \cup \{i\}$ can be extended to a basis, and we have $\|\mathbf{A}\|^2 \ge \|\mathbf{R}_{\text{opt}}\|^2 + \langle \mathbf{A}-\mathbf{R}_{\text{opt}}, \psi_i\rangle^2$, *i.e.*, $\langle \mathbf{A}-\mathbf{R}, \psi_i\rangle^2 = \langle \mathbf{A}-\mathbf{R}_{\text{opt}}, \psi_i\rangle^2 \le \|\mathbf{A}\|^2 - \|\mathbf{R}_{\text{opt}}\|^2 = \|\mathbf{A}-\mathbf{R}_{\text{opt}}\|^2$, a contradiction.

Algorithmically, we can start with a signal $\mathbf{A}$ and representation $\mathbf{R} = 0$ over subspace $\Lambda = \emptyset$. As long as $\|\mathbf{A}-\mathbf{R}\|^2 > 64B\,\|\mathbf{A}-\mathbf{R}_{\text{opt}}\|^2$, we can identify a vector $\phi \in \mathcal{D}$ that is in every optimal representation. We add ϕ to Λ, let $\mathbf{R}$ be the best representation over Λ, and continue looping. We terminate after at most B iterations, and, when we terminate, $\|\mathbf{A}-\mathbf{R}\| \le 64B\,\|\mathbf{A}-\mathbf{R}_{\text{opt}}\|$.

COROLLARY 2.2. *For a signal* $\mathbf{A}$*, Orthogonal Matching Pursuit, in at most* B *iterations, over a* μ*-coherent dictionary* $\mathcal{D}$*, returns a representation* $\mathbf{R}$ *with* $\|\mathbf{A}-\mathbf{R}\|^2 \leq 64B\|\mathbf{A}-\mathbf{R}_{\text{opt}}\|^2$*. OMP stopped when* $\|\mathbf{A}-\mathbf{R}\|^2 < 64B\|\mathbf{A}-\mathbf{R}_{\text{opt}}\|^2$ *returns a representation over a subspace of dimension* $B' \leq B$ *that can be extended to a* B*-dimensional subspace containing an optimal representation.*

2.3.3 Putting it All Together

THEOREM 2.1. *Fix a dictionary* $\mathcal{D}$ *with coherence* μ*. Let* $\mathbf{A}$ *be a signal and suppose it has a* B*-term representation over* $\mathcal{D}$ *with error* $\|\mathbf{A}-\mathbf{R}_{\text{opt}}\| = \delta$*, where* $B < 1/(32\mu)$*. Then, in iterations polynomial in* B*, we can find a representation with error at most* $\sqrt{(1+2064\mu B^2)}\delta$*.*

Proof. The algorithm is as follows. Assume we know $\delta = \|\mathbf{A}-\mathbf{R}_{\text{opt}}\|$ in advance; we will remove this assumption below. Use Corollary 2.2 to find a B'-term representation $\mathbf{R}_1$ over $\Lambda_1 \subseteq \mathcal{D}$ with $\|\mathbf{A}-\mathbf{R}\|^2 \leq 64B\delta^2$ and such that Λ_1 is a subset of a space containing an optimal representation. Then use Corollary 2.1 to find a representation $\mathbf{R}_2$ with square error at most $32\mu B\|\mathbf{A}-\mathbf{R}\|^2 \leq 32\cdot 64\mu B^2\|\mathbf{A}-\mathbf{R}_{\text{opt}}\|^2$ worse than the best representation for $\mathbf{A}-\mathbf{R}_1$. Output $\mathbf{R}_1+\mathbf{R}_2$.

We note that the hypotheses on μ and B in Corollaries 2.2 and 2.1 are implied by our hypothesis $B < 1/(32\mu)$. Since, as we show below, the best representation for $\mathbf{A}-\mathbf{R}_1$ has square error at most $\delta^2(1+512\mu^2B^3)$, the overall representation has square error at most $\delta^2(1+512\mu^2B^3+2048\mu B^2) \leq (1+2064\mu B^2)\delta^2$, since $\mu B < 1/32$. It remains only to show that $\mathbf{A}-\mathbf{R}_1$ has a representation with square error at most $\delta^2(1+512\mu^2B^3)$.

To see this, fix an optimal representation $\mathbf{R}_{\text{opt}} = \sum_{i\in\Lambda}\beta_i\phi_i$ consisting of vectors in some $\Lambda \supseteq \Lambda_1$. Let $\{\psi_i\}$ denote an orthogonalization of $\{\phi_i : \phi_i \in \Lambda\}$ that extends an orthogonalization of $\{\phi_i : \phi_i \in \Lambda_1\}$. Then $\mathbf{R}_{\text{opt}} = \sum_{i\in\Lambda}\langle\mathbf{A},\psi_i\rangle\psi_i$ and $\mathbf{R}_1 = \sum_{i\in\Lambda_1}\langle\mathbf{A},\psi_i\rangle\psi_i$.

Consider $\mathbf{R}_{\text{opt}}-\mathbf{R}_1$, which is orthogonal to $\mathbf{R}_1$ and to $\mathbf{A}-\mathbf{R}_{\text{opt}}$. We first claim that $\mathbf{R}_{\text{opt}}-\mathbf{R}$ has some good $(B-|\Lambda_1|)$-term representation over $\mathcal{D}$. Specifically, we will approximate $\mathbf{R}_{\text{opt}}-\mathbf{R}_1 = \sum_{i\in\Lambda\setminus\Lambda_1}\langle\mathbf{R}_{\text{opt}}-\mathbf{R}_1,\psi_i\rangle\psi_i$ by $\sum_{i\in\Lambda\setminus\Lambda_1}\langle\mathbf{R}_{\text{opt}}-\mathbf{R}_1,\psi_i\rangle\phi_i$. Since each $\|\psi_i-\phi_i\|$ is small, we expect that substituting ϕ_i for ψ_i in the expansion for $\mathbf{R}_{\text{opt}}-\mathbf{R}$ to be a small perturbation. We have $\left\|(\mathbf{R}_{\text{opt}}-\mathbf{R}_1)-\sum_{i\in\Lambda\setminus\Lambda_1}\langle\mathbf{R}_{\text{opt}}-\mathbf{R}_1,\psi_i\rangle\phi_i\right\|^2 = \left\|\sum_{i\in\Lambda\setminus\Lambda_1}\langle\mathbf{R}_{\text{opt}}-\mathbf{R}_1,\psi_i\rangle(\psi_i-\phi_i)\right\|^2$ is at most $\sum_{i\in\Lambda\setminus\Lambda_1}\langle\mathbf{R}_{\text{opt}}-\mathbf{R}_1,\psi_i\rangle^2\sum_{i\in\Lambda\setminus\Lambda_1}\|\psi_i-\phi_i\|^2$, which is at most

$$\begin{aligned}\|\mathbf{R}_{\text{opt}}-\mathbf{R}_1\|^2\cdot B(8\mu^2B) &\leq \|\mathbf{A}-\mathbf{R}_1\|^2\cdot B(8\mu^2B)\\ &\leq 64B\delta^2\cdot B(8\mu^2B) = 512\mu^2B^3\delta^2.\end{aligned}$$

Since $\mathbf{A}-\mathbf{R}_{\text{opt}}$ is orthogonal to $\text{span}(\Lambda)$, it follows that the representation $\sum_{i\in\Lambda\setminus\Lambda_1}\langle\mathbf{R}_{\text{opt}}-\mathbf{R}_1,\psi_i\rangle\phi_i$ gives a representation for $\mathbf{A}-\mathbf{R}_1$ with corresponding error, namely

$$\begin{aligned}&\left\|(\mathbf{A}-\mathbf{R}_1)-\sum_{i\in\Lambda\setminus\Lambda_1}\langle\mathbf{R}_{\text{opt}}-\mathbf{R}_1,\psi_i\rangle\phi_i\right\|^2\\ &= \|\mathbf{A}-\mathbf{R}_{\text{opt}}\|^2\\ &\quad+\left\|(\mathbf{R}_{\text{opt}}-\mathbf{R}_1)-\sum_{i\in\Lambda\setminus\Lambda_1}\langle\mathbf{R}_{\text{opt}}-\mathbf{R}_1,\psi_i\rangle\phi_i\right\|^2\\ &= \delta^2(1+512\mu^2B^3).\end{aligned}$$

Using Lemma 2.2, one can find a $(B-|\Lambda_1|)$-term approximation $\mathbf{R}_2$ to $\mathbf{A}-\mathbf{R}_1$ whose error is at most $32\mu B\|\mathbf{A}-\mathbf{R}_1\|^2 \leq 32\mu B(64B\delta^2)$ worse, additively, then the best such representation. It follows that

$$\begin{aligned}&\|\mathbf{A}-(\mathbf{R}_1+\mathbf{R}_2)\|^2\\ &\leq \|(\mathbf{A}-\mathbf{R}_1)-\mathbf{R}_2\|^2\\ &\leq \delta^2(1+512\mu^2B^3)+32\mu B(64B\delta^2)\\ &\leq \delta^2(1+16\mu B^2+2048\mu B^2), \text{ since } \mu B<1/32\\ &\leq \delta^2(1+2064\mu B^2).\end{aligned}$$

Now suppose we do not know δ in advance. Then we would not know $B' = |\Lambda_1|$, *i.e.*, we would not know when to switch from the error phase to the energy phase. We simply try all possibilities for $B' \leq B$.

So far, we have assumed that one can find the $\phi \in \mathcal{D}$ with largest $|\langle\mathbf{A},\phi\rangle|$ and estimate that dot product in unit time. We can, in fact, make this algorithm feasible and obtain a result similar to Theorem 2.1 using Nearest Neighbor Data Structures [14], [15]–[17]. For some c, fix ϵ, $c\mu B^2 \leq \epsilon \leq 1$; our goal will be to find a representation with error at most $(1+\epsilon)$ worse than optimal.

For Lemmas 2.2–2.4 and Theorem 2.1, we can prove that it suffices to find ϕ with *near*-biggest dot product and to *approximate* the dot product; *i.e.*, to find ϕ_j such that, for all i, $|\langle\mathbf{A},\phi_j\rangle|^2 \geq |\langle\mathbf{A},\phi_i\rangle|^2 - \eta\|\mathbf{A}\|^2$ and to estimate $\langle\mathbf{A},\phi_j\rangle$ as $\langle\mathbf{A},\phi_j\rangle_\sim$ with $|\langle\mathbf{A},\phi_j\rangle - \langle\mathbf{A},\phi_j\rangle_\sim|^2 \leq \eta\|\mathbf{A}\|^2$, for some appropriate η, polynomially related to B/ϵ. There is a limited set S of vectors for which we will want dot products; specifically, we will want $\langle\psi_1,\psi_2\rangle$ only if each ψ is a linear combination of

at most B vectors from $\mathcal{D} \cup \{\mathbf{A}\}$, in which the coefficients are, by (2.1), at most $O(\|\mathbf{A}\|^2)$ (here taken to be $N^{O(1)}$) and written to the unit of precision (here taken to be 1). Thus there are at most $(|\mathcal{D}|+1)^{O(B)}$ possible vectors in S. Therefore, we can use the following steps:

- Normalize the signal to $\frac{\mathbf{A}}{\|\mathbf{A}\|}$, find a representation for the normalized signal, and scale back up. For the normalized signal, we can get the approximate dot product operations from approximate ℓ^2 distance operations.
- Using [1], randomly project the signal $\mathbf{A}$ and the dictionary vectors, using a randomly-chosen linear function h from a particular family, into a $\log|S|/\eta^{O(1)}$ dimensional space so the ℓ^2 norm between any pairs of vectors from S is approximated to factor $(1+\eta)$.
- Use the approximate nearest neighbor results in [14], [15]–[17] to return a $(1+\eta)$-approximation to the closest dictionary vector in ℓ^2 norm to a query of the form $\mathbf{A}-\mathbf{R}$, in time $(B\log(|\mathcal{D}|)/\eta)^{O(1)}$. This allows us to compute $\mathbf{R}_1$ of Theorem 2.1. To compute $\mathbf{R}_2$ of Theorem 2.1, we need to find the B largest dot products with $\mathbf{A}-\mathbf{R}_1$. To do that, repeatedly find the ϕ_i with largest dot product to $\mathbf{A}-\mathbf{R}$ and, by properties of the nearest neighbor data structure, remove $h(\phi_i)$ from the dictionary, in time $(B\log|\mathcal{D}|/\eta)^{O(1)}$. Finally, to approximate the best representation of $\mathbf{A}$ over Λ, find the best formal representation of $h(\mathbf{A})$ over $\{h(\phi) : \phi \in \Lambda\}$ and use those coefficients.

Note that, even for $\phi_j \in \Lambda$, it is possible that $\langle \mathbf{A}-\mathbf{R}, \phi_j\rangle \neq 0$. In fact, $\langle \mathbf{A}-\mathbf{R}, \phi_j\rangle$ may be the maximum dot product over $\phi \in \mathcal{D}$, so we may need to choose ϕ_j on multiple iterations. One can show, however, that only a small number of iterations is needed for OMP to find a B'-term representation $\mathbf{R}_1$, $B' \leq B$, with $\|\mathbf{A}-\mathbf{R}_1\|^2 \leq 64B\|\mathbf{A}-\mathbf{R}_{\text{opt}}\|^2$. The energy phase takes just a single iteration of $(B-B')$-fold MP. In summary,

THEOREM 2.2. *Fix a dictionary, $\mathcal{D}$, of coherence μ, over vectors of length N. For certain constants c and c', fix $B < c/\mu$ and fix ϵ, $c'\mu B^2 \leq \epsilon \leq 1$. There is an algorithm that preprocesses $\mathcal{D}$ taking time and space $(B|\mathcal{D}|/\epsilon)^{O(1)}$. For any given signal, the algorithm produces a representation for $\mathbf{A}$ with error $(1+\epsilon)$ times that of the optimal representation taking time $(B\log|\mathcal{D}|/\epsilon)^{O(1)}$ only.*

Note that some cost dependence on N is hidden by dependence on $|\mathcal{D}| \geq N$.

The idea to use nearest neighbor data structures in this context was suggested independently by Piotr Indyk.

3 Special Dictionaries

In this section, we briefly explore the notion of small coherence dictionaries further. We consider several specific dictionaries and analyze their properties.

Three desired properties of redundant dictionaries are their large size, their naturalness, and the possibility to find representations over them quickly. Here "naturalness" encompasses the idea that signals typical of a particular application should have concise, low-error representations, and that the dictionary elements themselves should have semantic significance within the application. These goals are somewhat in conflict with each other; nevertheless, we show that our criterion of low coherence applies to several large and natural dictionaries in the literature (or small variations).

3.1 Spikes and Sinusoids Two fundamental orthonormal bases are the *spikes* and *sinusoids*. A spike is a vector δ_s defined by $\delta_s(s) = 1$ and $\delta_s(t) = 0$ for $t \neq s$. A sinusoid is a complex-valued vector ψ_ω defined by $\psi_\omega(t) = \frac{1}{\sqrt{N}} e^{2\pi i\omega t/N}$. It is immediate that the dot product of any spike with any sinusoid equals $1/\sqrt{N}$ in absolute value and the dictionary $\mathcal{D}$ formed by their union is μ-coherent, for $\mu = 1/\sqrt{N}$. It follows that the algorithm in Theorem 2.1 applies to this dictionary. Note that, in particular, if the signal is exactly represented as a sum of B spikes and sinusoids, so that the optimal error is zero, then the error of our output is zero—we recover the signal exactly.

We note that Theorem 2.1 gives a running time of $O(B^4 + B^2N + BN\log(N))$ to find a representation with error at most $(1 + O(B^2/\sqrt{N}))$ times optimal. For each of B iterations, we have a subset $\Lambda \subseteq \mathcal{D}$ of spikes and sinusoids available for the representation. To find the best representation $\mathbf{R}$ over Λ, we could find an orthonormal basis for span(Λ), which can be done in time B^3. We then need to find the the ϕ_j's with biggest dot products to $\mathbf{A}-\mathbf{R}$. We can find all $\langle \phi_j, \mathbf{A}\rangle$'s explicitly in time $O(N\log(N))$ and all $\langle \phi_j, \mathbf{R}\rangle$'s symbolically in time BN. For the extreme case of $B = \Theta(\sqrt{N})$, our algorithm takes time $O(N^2)$ to recover exactly any signal that is exactly represented as B spikes and sinusoids. By contrast, in [4], the authors give an algorithm with runtime $O(N^{3.5})$ which recovers the signal exactly, assuming it consists of B spikes and/or sinusoids. Even for the limited case of an exact representation, their algorithm is quite expensive.

Our algorithm and [4] apply to the combination of

any two incoherent orthonormal bases. The analysis above for our algorithm assumes that one of the bases is presented in the coordinate system defined by the other (equivalently, that one can take the dot product between any two dictionary elements in constant time.) If this is not the case, then one could compute and store all $O(N^2)$ such dot products in the time to multiply two $N \times N$ matrices, *i.e.*, N^{2+a} for some a, $0 < a < 1$.

Other speedups are possible for our algorithm, using Theorem 2.2 in general or using techniques in [13] for the spikes and sinusoids dictionary in particular. Details will be given in the final version of this paper.

3.2 Combinatorial Designs The dictionary in the previous subsection had small coherence and was a combination of two fundamental orthonormal bases but it was not very large. In fact, we can build a dictionary with small coherence that is considerably larger using a combinatorial design, such as the matrix at the center of the Nisan-Wigderson generator. Specifically, in [20], the authors show how to build a collection $\mathcal{S}$ of subsets of $[1, N]$, for N an even power of a prime, such that each subset has size exactly $\sqrt{N}$, any two subsets intersect in at most d places, and the number of subsets is $\Omega(N^{(d+1)/2})$.

Define $\mathcal{D}$ by $\mathcal{D} = \{N^{-1/4}\chi_S : S \in \mathcal{S}\}$. It follows that $|\mathcal{D}| = |\mathcal{S}| = \Omega(N^{(d+1)/2})$ and, for distinct ϕ_1 and ϕ_2 in $\mathcal{D}$, $|\langle \phi_1, \phi_2 \rangle| \leq d/\sqrt{N}$.

For $d = 2$, we get a superlinear-sized dictionary of size roughly $N^{3/2}$ and coherence roughly $2/\sqrt{N}$. At the other extreme, one can take d almost as large as $\sqrt{N}$, yielding coherence less than 1 and $|\mathcal{D}|$ roughly $2^{\sqrt{N}}$. One can also use values for d between the two extremes—for example, if $d = \log(N)$ or $d = N^{1/4}$, the dictionary has $N^{\Omega(\log(N))} > N^{O(1)}$ or $2^{\Omega(N^{1/4})} \gg N^{O(1)}$ elements and coherence $\log(N)/\sqrt{N}$ or $N^{-1/4}$, not much worse than the expected absolute dot product $1/\sqrt{N}$ of a pair of random unit vectors.

This construction gives us non-trivial dictionaries with extremely small coherence, extremely large size, or both, to which we can apply Theorem 2.1 and obtain efficient algorithms.

3.3 Segmentlets Next, we present a redundant dictionary that is inspired by beamlets [11], a dictionary used for edge detection in image analysis. Consider the space of functions on the square array of side $\sqrt{N}$. Fix a parameter p, to be determined later. Consider the set of all line segments with endpoints (x_1, y_1) and (x_2, y_2), such that x_1, y_1, and y_2 are multiples of p and $x_2 = x_1 + p$. Then the number of segments is $\Omega(N^{3/2}/p^3)$, much greater than the dimension, N, of the space of functions, and any two segments intersect at most once and have horizontal extent exactly p.

From each line segment, we next construct a set of pixels, each of which intersects the line segment. In general, a line segment may intersect several contiguous pixels in the same column; in that case, put only the middle one or two intersected pixels into the associated set. It follows that any two sets of pixels constructed this way will intersect in at most $O(1)$ pixels and each will consist of $\Theta(p)$ pixels. For the dictionary, take normalized characteristic functions on theses sets of pixels. The coherence is $O(1/\sqrt{p})$ and the size of the dictionary equals the number of segments, $N^{3/2}/p^3$.

Suppose we are interested in B-term representations. To apply the above techniques, we need coherence less than $1/B$, *i.e.*, $p \geq B^2$. It follows that the dictionary can have size $N^{3/2}/B^6$, which is greater than N for sufficiently small B. In particular, the size of the dictionary is superlinear, by more than log factors. Thus the size of the segmentlet dictionary is larger than the beamlet dictionary, and we can apply Theorem 2.1 to obtain efficient algorithms for near-optimal B-term representations over segmentlets. Segmentlets are natural for capturing edges in images. We believe that this dictionary will have exciting applications in image processing.

4 Concluding Remarks

We have presented algorithmic results for the basic problem in highly nonlinear approximation: An efficient algorithm for near-optimal representation of input signals over μ-coherent dictionaries. There are several additional insights we can provide into highly nonlinear approximation problems. We briefly describe them here.

The problem of representing functions using redundant dictionaries has a few variants. As defined above, a representation $\mathbf{R}$ for input $\mathbf{A} \in \mathbb{R}^N$ is a linear combination of dictionary elements, $\mathbf{R} = \sum_{i \in \Lambda} \alpha_i \phi_i$ for $\phi_i \in \mathcal{D}$. In general, there are two measures to assess the goodness of the representation. The first is how well it approximates $\mathbf{A}$ and we measure this error $\|\mathbf{A} - \mathbf{R}\|_p$ in ℓ^p norm for $1 \leq p \leq \infty$; the results in this paper focus on $p = 2$. The second metric attempts to capture varying notions of parsimony and we measure this with the ℓ^q norm of the representation itself, $\|\mathbf{R}\|_q = \sum_{i \in \Lambda} |\alpha_i|^q$ for $q = 0$ or $1 \leq q \leq \infty$. We focused throughout on minimizing the ℓ^2 norm of the error subject to the ℓ^0 norm of the representation equaling zero. In general, other (p, q) combinations may be of interest, and one may want to fix the ℓ^p norm of the error (for example to zero) and minimize the ℓ^q norm of the representation. For example, instead of seeking the best B-term representation, we may seek the representation minimizing

$\sum_i |\alpha_i|$, *i.e.*, $\|\alpha\|_1$. We have a polynomial time solution for this case via linear programming when $p = \infty$. Let Φ be the matrix of vector ϕ_i at component j, $\Phi_{i,j} = \phi_i(j)$. Express each coefficient α_i as a sum of positive and negative parts, $\alpha_i = p_i - n_i$. The linear program is to minimize $\sum_i (p_i + n_i)$ subject to

$$\begin{cases} \begin{pmatrix} \Phi & -\Phi \end{pmatrix} \begin{pmatrix} p \\ n \end{pmatrix} & \leq & \mathbf{A} + \epsilon \\ \begin{pmatrix} \Phi & -\Phi \end{pmatrix} \begin{pmatrix} p \\ n \end{pmatrix} & \geq & \mathbf{A} - \epsilon. \end{cases}$$

Another variation is one in which we wish to minimize the energy of the representation, *i.e.*, minimize $\|\alpha\|_2$. This problem has a polynomial time solution via semidefinite programming. This is because

$$\text{minimize } \alpha^t \alpha \text{ subject to } \|\Phi\alpha - \mathbf{A}\|_p \leq \epsilon$$

is a semidefinite program. Note that $\alpha^t\alpha$ is the squared ℓ^2 norm of the representation and the constraints define a convex region about the point $\mathbf{A} \in \mathbb{R}^N$.

We have initiated the study of the algorithmic complexity of highly nonlinear approximation problems. Approximation theory is a rich area, and it is active in mathematics, signal processing and statistics; theoretical computer scientists can have a significant impact here. Many algorithmic problems remain open. Finally, as remarked earlier, applied mathematicians in approximation theory care about practical algorithms. We believe that our algorithms here are not diffcult to implement, but we leave that study for the future.

Acknowledgement

We thank Joel Tropp for helpful discussions and for pointing out an error in an earlier draft of this paper.

References

[1] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. *JCSS*, 58(1):137–147, 1999.

[2] Constructive Approximation. `http://www.math.vanderbilt.edu/~ca`.

[3] E. Candes. *Ridgelets: Theory and Applications.* PhD thesis, Dept. of Stastistics, Stanford University, 1998.

[4] S. S. B. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Computing*, 20:33–61, 1999.

[5] R. Coifman and M. V. Wickerhauser. Entropy-based algorithms for best basis selection. *IEEE Trans. Inform. Theory*, 38(2), March 1992.

[6] I. Daubechies. *Ten lectures on wavelets.* SIAM, Philadelphia, 1992.

[7] G. Davis, S. Mallat, and M. Avellaneda. Greedy adaptive approximation. *Journal of Constructive Approximation*, 13:57–98, 1997.

[8] R. A. DeVore. Nonlinear approximation. *Acta Numerica*, 7:51–150, 1998.

[9] R. .A. DeVore and G. G. Lorentz. *Constructive Approximation.* Springer-Verlag, New York, 1993.

[10] D. Donoho. Wedgelets: Nearly-minimax estimation of edges. *Annals of Statistics*, 27:859–897, 1999.

[11] D. L. Donoho and X. Huo. Beamlet pyramids: A new form of multiresolution analysis, suited for extracting lines, curves, and objects from very noisy image data. In *Proceedings of SPIE 2000*, volume 4119, 2000.

[12] A. C. Gilbert, S. Guha, P. Indyk, Y. Kotidis, S. Muthukrishnan, and M. Strauss. Fast, small-space algorithms for approximate histogram maintenance. In *Proceedings of ACM STOC 2002*, 2002.

[13] A. C. Gilbert, S. Guha, P. Indyk, S. Muthukrishnan, and M. Strauss. Near-optimal sparse Fourier representations via sampling. In *Proc. of ACM STOC*, 2002.

[14] P. Indyk. *High-Dimensional Computational Geometry.* PhD thesis, Stanford, 2000.

[15] P. Indyk and R. Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of ACM STOC 1998*, pages 604–613, 1998.

[16] The Journal of Approximation Theory. `http://www.math.ohio-state.edu/JAT`.

[17] E. Kushilevitz, R. Ostrovsky, and Y. Rabani. Efficient search for approximate nearest neighbor in high dimensional spaces. In *Proceedings of ACM STOC 1998*, pages 614–623, 1998.

[18] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

[19] C. M.Thiele and L. F. Villemoes. A fast algorithm for adapted time frequency tilings. *Applied and Computational Harmonic Analysis*, 3:91–99, 1996.

[20] N. Nisan and A. Wigderson. Hardness vs. randomness. *J. Comput. System Sci.*, 49:149–167, 1994.

[21] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad. Orthogonal Matching Pursuit: Recursive Function Approximation with Applications to Wavelet Decomposition. In *Proc. of the 27th Annual Asilomar Conference on Signals, Systems, and Computers*, pages 40–44, 1993.

[22] V. N. Temlyakov. The best m-term approximation and greedy algorithms. *Advances in Computational Math.*, 8:249–265, 1998.

[23] V. N. Temlyakov. Greedy algorithms and m-term approximation with regard to redundant dictionaries. *J. Approximation Theory*, to appear.

[24] Lars Villemoes. Best approximation with Walsh atoms. *Constructive Approximation*, 13:329–355, 1997.

[25] Lars Villemoes. Nonlinear approximation with Walsh atoms. In A. Le Méhauté, C. Rabut, and L. L. Schumaker, editors, *Surface Fitting and Multiresolution Methods*, pages 329–336. Vanderbilt University Press, 1997.