Estimation via Sparse Approximation:
Error Bounds and Random Frame Analysis

by

Alyson Kerry Fletcher

A thesis submitted in partial satisfaction
of the requirements for the degree of

Master of Arts

in

Mathematics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor F. Alberto Grünbaum, Chair
Professor David J. Aldous
Professor Bin Yu

Spring 2005

Estimation via Sparse Approximation:

Error Bounds and Random Frame Analysis

by

Alyson Kerry Fletcher

Abstract

Estimation via Sparse Approximation:
Error Bounds and Random Frame Analysis

by

Alyson Kerry Fletcher

Master of Arts in Mathematics

University of California, Berkeley

Professor F. Alberto Grünbaum, Chair

If a signal x can be represented as a linear combination of K elements from a set of vectors Φ, then x is said to have a K-sparse representation with respect to Φ. Sparseness can be used to remove noise in that x can be estimated from a noise-corrupted observation y by finding the best K-sparse approximation to y. Sparse approximation-based estimation has proven to be an effective method in many areas including wavelet image processing and pattern recognition. However, exactly quantifying the performance of a sparse approximation estimator is, in general, difficult due to the discrete, nonlinear nature of the estimation process.

This work considers two approaches for quantifying the performance of sparse approximation-based estimation. The first approach provides a lower bound (Theorem 1) on the ability to represent a Gaussian random vector sparsely with respect to a given frame Φ. The bound applies to arbitrary frames and depends only on the signal dimension, frame size, and sparsity. The bound shows that, for moderate-sized frames, Gaussian noise is *not* well represented sparsely and thus suggests that sparse approximation will reject such noise well. The bound is derived using rate–distortion theory and may be of independent interest in the study of lossy source coding.

The second approach considers frames generated randomly according to a spherically-symmetric distribution and signals expressible with single dictionary elements. Easily-computed estimates for the probability of selecting the correct dictionary element and the mean-squared error are given (Theorems 3 and 4). Monte Carlo simulations demonstrate the accuracies of these estimates. In the limit as the dimension of the space grows without bound, the estimates reduce to very simple forms. The large-dimension asymptotics (Theorems 5 and 6) reveal a critical signal-to-noise ratio threshold above which the probability of error approaches zero and below which the probability of error approaches one.

Professor F. Alberto Grünbaum
Thesis Committee Chair

To Lola and Simone[2]

Contents

List of Figures

Acknowledgements

This work would not have been possible without an advisor who encouraged me to explore questions outside the scope of my electrical engineering dissertation research. For this freedom and his unwavering support through many ordeals, I thank Professor Kannan Ramchandran.

This research was a pleasure because of interactions with Dr. Sundeep Rangan and Professor Vivek Goyal. I thank them for extensive feedback. I gratefully acknowledge Professor Martin Vetterli's early encouragement of this work. I also would like to thank my thesis committee, Professors F. Alberto Grünbaum, David Aldous, and Bin Yu, for providing their valuable insights.

Ruth Gjerde and Mary Byrnes were absolutely essential in navigating the confusing waters of UC-Berkeley bureaucracy. At the same time, Adriana Schoenberg, Jeff Nelson, Gerald Keane, and most importantly, the incomparable John King, were keeping me healthy enough to complete this thesis.

I couldn't have done this without a few of my friends who kept me sane, or at least tried: Mareike Claassen, Leon Abrams, Ron & Sally Goldstein, Alan, Jane, & Anna Schoenfeld, Dave Nguyen, Kristie Korneluk, James Yeh, Gabe Moy, Mark Johnson, Abhik Majumdar, June Wang, and DeLynn Bettencourt. Lola made getting out of bed every morning and coming home at night sheer joy. Finally, thank you to Sundeep Rangan and Vivek Goyal. The help, encouragement, support, and friendship that Sundeep has given me is more than one could ever ask for or expect. Without the laughter and kindness that he brings to both my worst and best days, my world would not be the same. Most of all, I couldn't have written this thesis or even be here without the unwavering belief and support of Vivek, my oldest dearest friend and the most generous person on earth.

I gratefully acknowledge the financial support of the National Science Foundation through a Graduate Fellowship, Sigma Xi for a Grant-In-Aid of Research, the Soroptimist International Founder's Region through a Dissertation Year Fellowship, and the Henry Luce Foundation through a Clare Boothe Luce Scholarship.

Chapter 1

Introduction

Estimating a signal from a noise-corrupted observation of the signal is a recurring task in science and engineering. This thesis explores the limits of estimation performance in the case where the only *a priori* structure on the signal $x \in \mathbb{R}^N$ is that it has known sparsity K with respect to a given set of vectors $\Phi = \{\varphi_i\}_{i=1}^M \subset \mathbb{R}^N$. The set Φ is called a *dictionary* and is generally a *frame* [22, 14]. The sparsity of K with respect to Φ means that the signal x lies in the set

$$\Phi_K \;=\; \left\{ v \in \mathbb{R}^N \;\middle|\; v = \sum_{i=1}^{M} \alpha_i \varphi_i \quad \text{with at most } K \text{ nonzero } \alpha_i\text{'s} \right\}. \tag{1.1}$$

In many areas of computation, exploiting sparsity is motivated by reduction in complexity [16]; if $K \ll N$ then certain computations may be more efficiently made on α than on x. In compression, representing a signal exactly or approximately by a member of Φ_K is a common first step in efficiently representing the signal, though much more is known when Φ is a basis or union of wavelet bases than is known in the general case [21]. Of more direct interest here is that sparsity models are becoming prevalent in estimation problems; see, *e.g.*, [31, 41].

The parameters of dimension N, dictionary size M, and sparsity K determine the importance of the sparsity model. Representative illustrations of Φ_K are given in Figure 1.1. With dimension $N = 2$, sparsity of $K = 1$ with respect to a dictionary of size $M = 3$

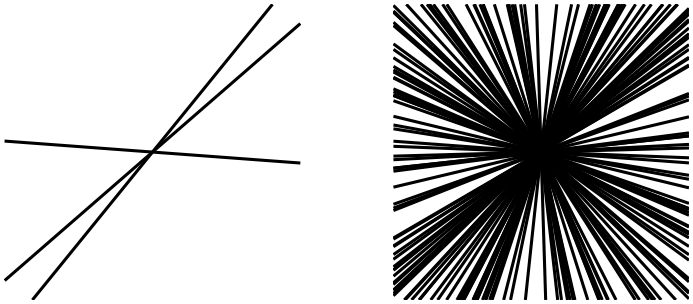


Figure 1.1. Two sparsity models in dimension $N = 2$. Left: Having sparsity $K = 1$ with respect to a dictionary with $M = 3$ elements restricts the possible signals greatly. Right: With the dictionary size increased to $M = 100$, the possible signals still occupy a set of measure zero, but a much larger fraction of signals are approximately sparse.

indicates that x lies on one of three lines, as shown in the left panel. This is a restrictive model, even if there is some approximation error in (1.1). When M is increased, the model stops seeming restrictive, even though the set of possible values for x has measure zero in $\mathbb{R}^2$. The reason is that, unless the dictionary has gaps, all of $\mathbb{R}^2$ is nearly covered. This thesis presents progress in explaining the value of a sparsity model for signal denoising as a function of (N, M, K).

1.1 Denoising by Sparse Approximation with a Frame

Consider the problem of estimating a signal $x \in \mathbb{R}^N$ from the noisy observation $y = x+d$ where $d \in \mathbb{R}^N$ has the i.i.d. Gaussian $\mathcal{N}(0, \sigma^2 I_N)$ distribution. Suppose we know that x lies in given K-dimensional subspace of $\mathbb{R}^N$. Then projecting y to the given subspace would remove a fraction of the noise without affecting the signal component. Denoting the projection operator by P, we would have

$$\hat{x} \;=\; Py \;=\; P(x+d) \;=\; Px + Pd \;=\; x + Pd,$$

and Pd has only K/N fraction of the power of d.

In this thesis we consider the more general signal model $x \in \Phi_K$. The set Φ_K defined in (1.1) is the union of at most $J = \binom{M}{K}$ subspaces of dimension K. We henceforth assume $M > K$ (thus $J > 1$); if not, the model reduces to the classical case of knowing a single

subspace that contains x. The distribution of x, if available, could also be exploited to remove noise. However, in this thesis the denoising operation is based only on the geometry of the signal model Φ_K and the distribution of d.

With the addition of the noise d, the observed vector y will (almost surely) not be represented sparsely, *i.e.*, not be in Φ_K. Intuitively, a good estimate for x is the point from Φ_K that is closest to y in Euclidean distance. Formally, because the probability density function of d is a strictly decreasing function of $\|d\|_2$, this is the maximum likelihood estimate of x given y. The estimate is obtained by applying an optimal sparse approximation procedure to y. We will write

$$\hat{x}_{\mathrm{SA}} = \operatorname*{argmin}_{x \in \Phi_K} \|y - x\|_2 \tag{1.2}$$

for this estimate and call it the optimal K-term approximation of y. Henceforth we omit the subscript 2 indicating the Euclidean norm.

The main results of this thesis are bounds on the per-component mean-squared estimation error $\frac{1}{N}\mathbf{E}\left[\|x - \hat{x}_{\mathrm{SA}}\|^2\right]$ for denoising via sparse approximation.[1] These bounds depend on (N, M, K) but avoid further dependence on the dictionary Φ (such as the coherence of Φ); some results hold for all Φ and others are for randomly generated Φ. To the best of our knowledge, the results differ from any in the literature in several ways:

(a) We study mean-squared estimation error for additive Gaussian noise, which is a standard approach to performance analysis in signal processing. In contrast, analyses such as [20] impose a deterministic bound on the norm of the noise.

(b) We concentrate on having dependence solely on dictionary size rather than more fine-grained properties of the dictionary. In particular, most signal recovery results in the literature are based on noise being bounded above by a function of the *coherence* of the dictionary [23, 18, 35, 27, 50, 19, 51].

(c) Some of our results are for spherically-symmetric random dictionaries. The series of papers [5, 7, 6] is superficially related because of randomness, but in these papers

[1] The expectation is always over the noise d and is over the dictionary Φ and signal x in some cases. However, the estimator does not use the distribution of x.

the signals of interest are sparse with respect to a single known, orthogonal basis and the observations are random inner products. The natural questions include a consideration of the number of measurements needed to robustly recover the signal.

(d) We use source coding thought experiments in bounding estimation performance. This technique may be useful in answering other related questions, especially in sparse approximation source coding.

Our preliminary results were first presented in [26]. Probability of error results in a rather different framework for basis pursuit appear in a manuscript currently under review [24].

1.2 Connections to Approximation

A signal with an exact K-term representation might arise because it was generated synthetically, for example, by a compression system. A more likely situation in practice is that there is an underlying true signal x that has a good K-term *approximation* rather than an exact K-term *representation*. At very least, this is the goal in designing the dictionary Φ for a signal class of interest. It is then still reasonable to compute (1.2) to estimate x from y, but there are trade-offs in the selections of K and M.

Let $f_{M,K}$ denote the squared Euclidean approximation error of the optimal K-term approximation using an M-element dictionary. It is obvious that $f_{M,K}$ decreases with increasing K, and with suitably designed dictionaries it also decreases with increasing M. One concern of approximation theory is to study the decay of $f_{M,K}$ precisely. (For this we should consider N very large or infinite.) For piecewise smooth signals, for example, wavelet frames give exponential decay with K [10, 17, 21].

When one uses sparse approximation to denoise, the performance depends on both the ability to approximate x and the ability to reject the noise. Approximation is improved by increasing M and K, but noise rejection is diminished. The dependence on K is clear, as the fraction of the original noise that remains on average is at least K/N. For the dependence on M, note that increasing M increases the number of subspaces and thus increases the

chance that the selected subspace is not the best one for approximating x. Loosely, when M is very large and the dictionary elements are not too unevenly spread, there is some subspace very close to y and thus $\hat{x}_{\mathrm{SA}} \approx y$. This was illustrated in Figure 1.1.

Fortunately, there are many classes of signals for which M need not grow too quickly as a function of N to get good sparse approximations. Examples of dictionaries with good computational properties that efficiently represent audio signals were given by Goodwin [30]. For iterative design procedures, see papers by Engan *et al.* [25] and Tropp *et al.* [52].

One initial motivation for this work was to give guidance for the selection of M. This requires the combination of approximation results (*e.g.*, bounds on $f_{M,K}$) with results such as ours. The results presented here do not address approximation quality.

1.3 Related Work

Computing optimal K-term approximations is generally a difficult problem. Given $\epsilon \in \mathbb{R}^+$ and $K \in \mathbb{Z}^+$, to determine if there exists a K-term approximation $\hat{x}$ such that $\|x - \hat{x}\| \leq \epsilon$ is an NP-complete problem [15, 45]. This computational intractability of optimal sparse approximation has prompted study of heuristics. A greedy heuristic that is standard for finding sparse approximate solutions to linear equations [29] has been known as *matching pursuit* in the signal processing literature since the work of Mallat and Zhang [42]. Also, Chen, Donoho and Saunders [9] proposed a convex relaxation of the approximation problem (1.2) called *basis pursuit*.

Two related discoveries have touched off a flurry of recent research:[2]

(a) *Stability of sparsity*—Under certain conditions, the positions of the nonzero entries in a sparse representation of a signal are stable: applying optimal sparse approximation to a noisy observation of the signal will give a coefficient vector with the original support. Typical results are upper bounds (functions of the norm of the signal and

[2]The intensity of activity in this area is reflected by the number of manuscripts currently in review that we have cited.

the coherence of the dictionary) on the norm of the noise that allows a guarantee of stability [23, 35, 34, 18, 20].

(b) *Effectiveness of heuristics*—Both basis pursuit and matching pursuit are able to find optimal sparse approximations, under certain conditions on the dictionary and the sparsity of signal [18, 20, 50, 51, 37, 36].

To contrast: in this thesis we consider noise with unbounded support and thus a positive probability of failing to satisfy a sufficient condition for stability as in (a) above; and we do not address algorithmic issues in finding sparse approximations. It bears repeating that finding optimal sparse approximations is presumably computationally intractable except in the cases where a greedy algorithm or convex relaxation happens to succeed. Our results are thus bounds on the performance of the algorithms that one would probably use in practice.

Denoising by finding a sparse approximation is similar to the concept of denoising by compression popularized by Saito [47] and Natarajan [44]. More recent works in this area include those by Krim *et al.* [39], Chang *et al.* [8] and Liu and Moulin [40]. All of these works use bases rather than frames. To put the present work into a similar framework would require a "rate" penalty for redundancy. Instead, the only penalty for redundancy comes from choosing a subspace that does not contain the true signal ("overfitting" or "fitting the noise"). The literature on compression with frames notably includes [3, 46, 32, 1, 43].

This thesis uses quantization and rate–distortion theory only as a proof technique; there are no encoding rates because the problem is purely one of estimation. However, the "negative" results on representing white Gaussian signals with frames presented here should be contrasted with the "positive" encoding results of Goyal *et al.* [32]. The positive results of [32] are limited to low rates (and hence signal-to-noise ratios that are usually uninteresting). A natural extension of the present work is to derive negative results for encoding. This would support the assertion that frames in compression are useful not universally, but only when they can be designed to yield very good sparseness for the signal class of interest.

1.4 Preview of Results and Outline

To motivate the thesis, we present a set of numerical results from Monte Carlo simulations that qualitatively reflect our main results. In these experiments, N, M, and K are small because of the high complexity of computing optimal approximations and because a large number of independent trials is needed to get adequate precision. Each data point shown is the average of 100 000 trials.

Consider a true signal $x \in \mathbb{R}^4$ ($N = 4$) that has an exact 1-term representation ($K = 1$) with respect to M-element dictionary Φ. We observe $y = x + d$ with $d \sim \mathcal{N}(0, \sigma^2 I_4)$ and compute estimate $\hat{x}_{\mathrm{SA}}$ from (1.2). The signal is generated with unit norm so that the signal-to-noise ratio (SNR) is $1/\sigma^2$ or $-10 \log_{10} \sigma^2$ dB. Throughout we use the following definition for mean-squared error:

$$\mathrm{MSE} = \frac{1}{N} \mathbf{E} \left[\|x - \hat{x}_{\mathrm{SA}}\|^2 \right].$$

To have tunable M, we used dictionaries that are M maximally separated unit vectors in $\mathbb{R}^N$, where separation is measured by the minimum pairwise angle among the vectors and their negations. These are cases of Grassmannian packings [11, 49] in the simplest case of packing one-dimensional subspaces (lines). We used packings tabulated by Sloane with Hardin, Smith and others [48].

Figure 1.2 shows the MSE as a function of σ for several values of M. Note that for visual clarity, MSE / σ^2 is plotted, and all of the same properties are illustrated for $K = 2$ in Figure 1.3. For small values of σ, the MSE is $(1/4)\sigma^2$. This is an example of the general statement that

$$\mathrm{MSE} = \frac{K}{N} \sigma^2 \qquad \text{for small } \sigma,$$

as described in detail in Chapter 2. For large values of σ, the scaled MSE approaches a constant value:

$$\lim_{\sigma \to \infty} \frac{\mathrm{MSE}}{\sigma^2} = g_{K,M},$$

where $g_{K,M}$ is a slowly increasing function of M and $\lim_{M \to \infty} g_{K,M} = 1$. This limiting value makes sense because in the limit $\hat{x}_{\mathrm{SA}} \approx y = x + d$ and each component of d has variance σ^2;

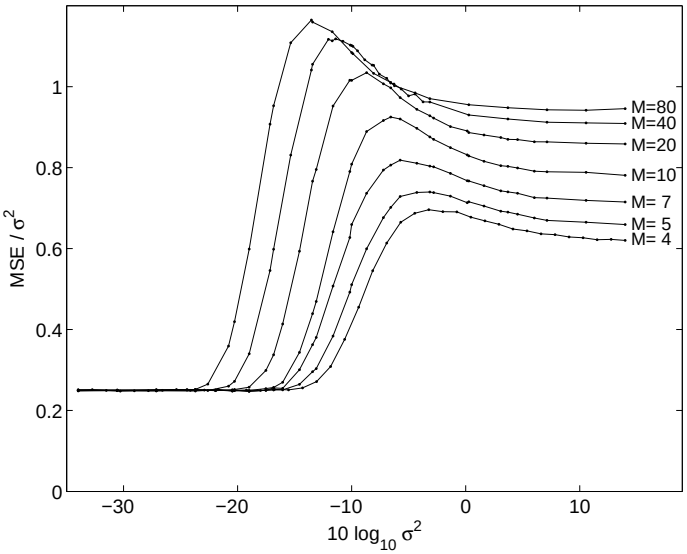


Figure 1.2. Performance of denoising by sparse approximation when the true signal $x \in \mathbb{R}^4$ has an exact 1-term representation with respect to a dictionary that is an optimal M-element Grassmannian packing.

the denoising does not do anything. The characterization of the dependence of $g_{K,M}$ on K and M is the main contribution of Chapter 3.

Another apparent pattern in Figure 1.2 that we would like to explain is the transition between low and high SNR behavior. The transition occurs at smaller values of σ for larger values of M. Also, MSE / σ^2 can exceed 1, so in fact the sparse approximation procedure can *increase* the noise. We are not able to characterize the transition well for general frames. However, in Chapter 4 we obtain results for large frames that are generated by choosing vectors uniformly at random from the unit sphere in $\mathbb{R}^N$. There we get a sharp transition between low and high SNR behavior.

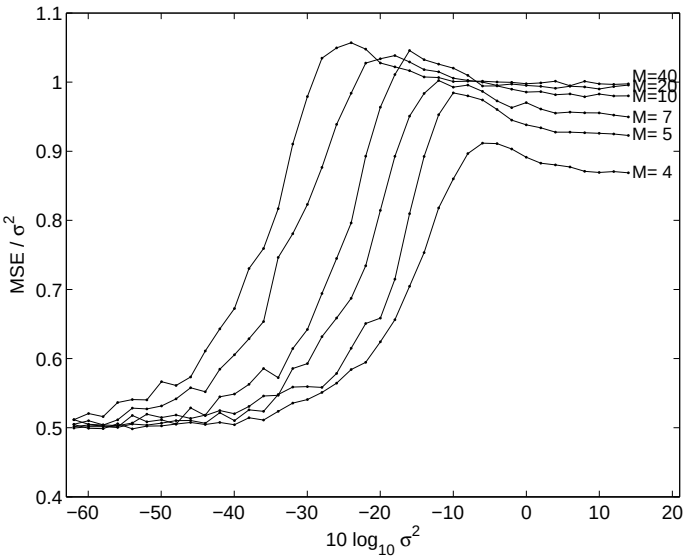


Figure 1.3. Performance of denoising by sparse approximation when the true signal $x \in \mathbb{R}^4$ has an exact 2-term representation with respect to a dictionary that is an optimal M-element Grassmannian packing.

Chapter 2

Preliminary Computations

Recall from the introduction that we are estimating a signal $x \in \Phi_K \subset \mathbb{R}^N$ from an observation $y = x + d$ where $d \sim \mathcal{N}(0, \sigma^2 I_N)$. Φ_K was defined in (1.1) as the set of vectors that can be represented as a linear combination of K vectors from $\Phi = \{\varphi_m\}_{m=1}^M$. We are studying the performance of the estimator

$$\hat{x}_{\mathrm{SA}} = \operatorname*{argmin}_{x \in \Phi_K} \|y - x\|.$$

This estimator is the maximum likelihood estimator of x in this scenario in which d has a Gaussian density and the estimator has no probabilistic prior information on x. The subscript SA denotes "sparse approximation" because the estimate is obtained by finding the optimal sparse approximation of y. There are values of y such that $\hat{x}_{\mathrm{SA}}$ is not uniquely defined. These collectively have probability zero and we ignore them.

Finding $\hat{x}_{\mathrm{SA}}$ can be viewed as a two-step procedure: first, find the subspace spanned by K elements of Φ that contains $\hat{x}_{\mathrm{SA}}$; then, project y to that subspace. The identification of a subspace and the orthogonality of $y - \hat{x}_{\mathrm{SA}}$ to that subspace will be used in our analyses. Let $\mathcal{P}_K = \{P_i\}_i$ be the set of the projections onto subspaces spanned by K of the M vectors in Φ. Then $\mathcal{P}_K$ has at most $J = \binom{M}{K}$ elements,[1] and the estimate of interest is given by

$$\hat{x}_{\mathrm{SA}} \;=\; P_T y, \qquad \text{where} \qquad T \;=\; \operatorname*{argmax}_i \|P_i y\|. \tag{2.1}$$

[1] It is possible for distinct subsets of Φ to span the same subspace.

The distribution of the error $x - \hat{x}_{\mathrm{SA}}$ and the average performance of the estimator both depend on the true signal x. Where there is no distribution on x, the performance measure analyzed here is the conditional MSE

$$e(x) \;=\; \frac{1}{N}\mathbf{E}\left[\|x - \hat{x}_{\mathrm{SA}}\|^2 \;\mid\; x\right]; \tag{2.2}$$

one could say that showing conditioning in (2.2) is merely for emphasis.

In the case that T is independent of d, the projection in (2.1) is to a fixed K-dimensional subspace, so

$$e(x) \;=\; \frac{K}{N}\,\sigma^2. \tag{2.3}$$

This occurs when $M = K$ (there is just one element in $\mathcal{P}_K$) or in the limit of high SNR (small σ^2). In the latter case, the subspace selection is determined by x, unperturbed by d.

Chapter 3

Rate-Distortion Analysis and Low SNR Bound

In this section, we establish bounds on the performance of sparse approximation denoising that apply for any dictionary Φ. One such bound qualitatively explains the low-SNR performance shown in Figures 1.2 and 1.3, *i.e.*, the right-hand side asymptotes in these plots.

The denoising bound depends on a performance bound for sparse approximation *signal representation* developed in Section 3.1. The signal representation bound is empirically evaluated in Section 3.2 and then related to low-SNR denoising in Section 3.3. We will also discuss the difficulties in extending this bound for moderate SNR. To obtain interesting results for moderate SNR, we consider randomly generated Φ in Chapter 4.

3.1 Sparse Approximation of a Gaussian Source

Before addressing the denoising performance of sparse approximation, we give an *approximation* result for Gaussian signals. This result is a *lower bound* on the MSE when sparsely approximating a Gaussian signal; it is the basis for an *upper bound* on the MSE for

denoising when the SNR is low. These bounds are in terms of the problem size parameters (M, N, K).

Theorem 1 *Let Φ be an M-element dictionary, let $J = \binom{M}{K}$, and let $v \in \mathbb{R}^N$ have the distribution $\mathcal{N}(\bar{v}, \sigma^2 I_N)$. If $\hat{v}$ is the optimal K-sparse approximation of v with respect to Φ, then*

$$\frac{1}{N}\boldsymbol{E}\left[\|v - \hat{v}\|^2\right] \;\geq\; \sigma^2 c_1 \left(1 - \frac{K}{N}\right) \tag{3.1}$$

where

$$c_1 = J^{-2/(N-K)} \left(\frac{K}{N}\right)^{K/(N-K)}.$$

For $\bar{v} = 0$, the stronger bound

$$\frac{1}{N}\boldsymbol{E}\left[\|v - \hat{v}\|^2\right] \;\geq\; \sigma^2 \cdot \frac{c_1}{1 - c_1} \cdot \left(1 - \frac{K}{N}\right) \tag{3.2}$$

also holds.

Proof: This follows from Theorem 2. See Section 6.1.

Remarks:

(i) Theorem 1 shows that for any Φ, there is an approximation error lower bound that depends only on the frame size M, the dimension of the signal N, and the dimension of the signal model K.

(ii) As $M \to \infty$ with K and N fixed, $c_1 \to 0$. This is consistent with the fact that it is possible to drive the approximation error to zero by letting the dictionary grow.

(iii) The decay of c_1 as M increases is slow. To see this, define a sparsity measure $\alpha = K/N$ and a redundancy factor $\rho = M/N$. Now using the approximation (see, *e.g.*, [28, p. 530])

$$\binom{\rho N}{\alpha N} \;\approx\; \left(\frac{\rho}{\alpha}\right)^{\alpha N} \left(\frac{\rho}{\rho - \alpha}\right)^{(\rho-\alpha)N},$$

we can compute the limit

$$\lim_{N\to\infty} c_1 \;=\; \left[\left(\frac{\alpha}{\rho}\right)^{2\alpha} \left(1 - \frac{\alpha}{\rho}\right)^{2(\rho-\alpha)} \alpha^{\alpha}\right]^{1/(1-\alpha)}.$$

Thus the decay of the lower bound in (3.1) as ρ is increased behaves as $\rho^{-2\alpha/(1-\alpha)}$. This is slow when α is small.

The theorem below strengthens Theorem 1 by having a dependence on the entropy of the subspace selection random variable T in addition to the problem size parameters (M, N, K). The entropy of T is defined as

$$H(T) \;=\; -\sum_{i=1}^{|\mathcal{P}_K|} p_T(i) \log_2 p_T(i) \quad \text{bits}$$

where $p_T(i)$ is the probability mass function of T.

Theorem 2 *Let Φ be an M-element dictionary, and let $v \in \mathbb{R}^N$ have the distribution $\mathcal{N}(\bar{v}, \sigma^2 I_N)$. If $\hat{v}$ is the optimal K-sparse approximation of v with respect to Φ and T is the index of the subspace that contains $\hat{v}$, then*

$$\frac{1}{N}\boldsymbol{E}\left[\|v - \hat{v}\|^2\right] \;\geq\; \sigma^2 c_2 \left(1 - \frac{K}{N}\right) \tag{3.3}$$

where

$$c_2 = 2^{-2H(T)/(N-K)} \left(\frac{K}{N}\right)^{K/(N-K)}.$$

For $\bar{v} = 0$, the stronger bound

$$\frac{1}{N}\boldsymbol{E}\left[\|v - \hat{v}\|^2\right] \;\geq\; \sigma^2 \cdot \frac{c_2}{1 - c_2} \cdot \left(1 - \frac{K}{N}\right) \tag{3.4}$$

also holds.

Proof: See Section 6.1.

3.2 Empirical Evaluation of Approximation Error Bounds

The bound in Theorem 1 does not depend on any characteristics of the dictionary other than M and N. Thus it will be nearest to tight when the dictionary is well-suited to representing the Gaussian signal v. That the expression (3.1) is not just a bound but also a useful approximation is supported by the Monte Carlo simulations described in this section.

To empirically evaluate the tightness of the bound, we compare it to the MSE obtained with Grassmannian frames and certain random frames. The Grassmannian frames are from the same tabulation described in Section 1.4 [48]. The random frames are generated by choosing M vectors uniformly at random from the surface of a unit sphere. One such vector can be generated, for example, by drawing an i.i.d. Gaussian vector and normalizing.

Figure 3.1 shows comparisons between the bound in Theorem 1 and the simulated approximation errors as a function of M for several values of N and K. For all the simulations, $\bar{v} = 0$; it is for $\bar{v} = 0$ that T is closest to uniformly distributed and hence the bound is tightest. Parts (a)–(c) each cover a single value of N and combine $K = 1$ and $K = 2$. Part (d) shows results for $N = 10$ and $N = 100$ for $K = 1$. In all cases, the bound holds and gives a qualitative match in the dependence of the approximation error on K and M. In particular, the slopes on these log-log plots correspond to the decay as a function of ρ discussed in Remark (iii) above. We also find that the difference in approximation error between using a Grassmannian frame or a random frame is small.

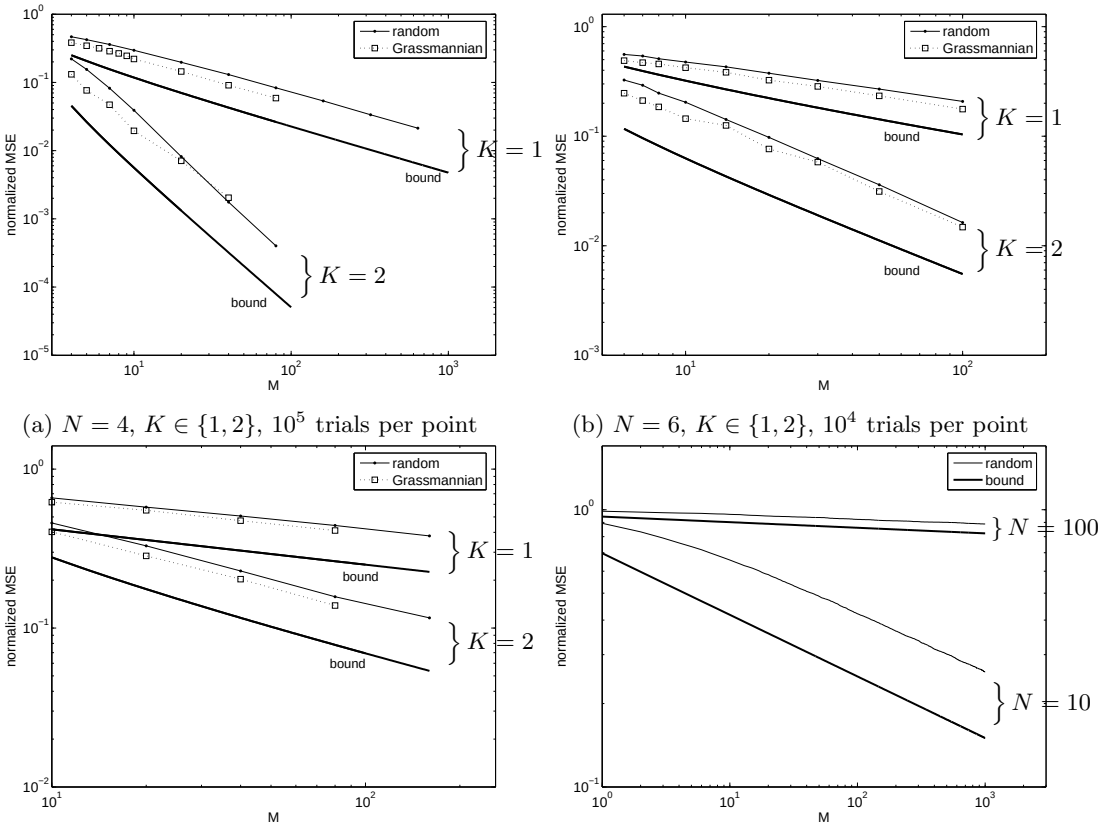
3.3 Bounds on Denoising MSE

We now return to the analysis of the performance of sparse approximation denoising as defined in Chapter 2. We wish to bound the estimation error $e(x)$ for a given signal x and frame Φ.

To create an analogy between the approximation problem considered in Section 3.1 and the denoising problem, let $\bar{v} = x$, $v - \bar{v} = d$, and $v = y$. These correspondences fit perfectly, since $d \sim \mathcal{N}(0, \sigma^2 I_N)$ and we apply sparse approximation to y to get $\hat{x}_{\mathrm{SA}}$. Theorem 2 gives the bound

$$\frac{1}{N}\mathbf{E}\left[\|y - \hat{x}_{\mathrm{SA}}\|^2 \mid x\right] \;\geq\; \sigma^2 c_2 \left(1 - \frac{K}{N}\right)$$

where c_2 is defined as before. As illustrated in Figure 3.2, it is as if we are attempting to represent d by sparse approximation and we obtain $\hat{d} = \hat{x}_{\mathrm{SA}} - x$. The quantity we are interested in is $e(x) = \frac{1}{N}\mathbf{E}\left[\|\hat{d}\|^2 \mid x\right]$.



(a) $N = 4$, $K \in \{1, 2\}$, 10^5 trials per point

(b) $N = 6$, $K \in \{1, 2\}$, 10^4 trials per point

(c) $N = 10$, $K \in \{1, 2\}$, 10^4 trials per point

(d) $N \in \{10, 100\}$, $K = 1$, 10^2 trials per point

Figure 3.1. Comparison betweeen the bound in Theorem 1 and the approximation errors obtained with Grassmannian and spherically-symmetric random frames. The horizontal axis in all plots is M.

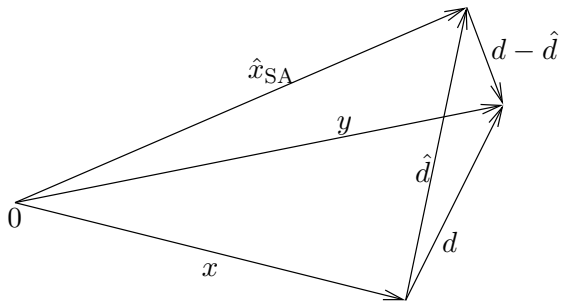


Figure 3.2. Illustration of variables to relate approximation and denoising problems. (An undesirable case in which $\hat{x}_{\mathrm{SA}}$ is not in the same subspace as x.)

In the case that x and $\hat{x}_{\mathrm{SA}}$ are in the same subspace, $d - \hat{d}$ is orthogonal to $\hat{d}$ so $\|d\|^2 = \|\hat{d}\|^2 + \|d - \hat{d}\|^2$. Thus knowing $\mathbf{E}\left[\|d\|^2 \mid x\right] = N\sigma^2$ and having a lower bound on $\mathbf{E}\left[\|\hat{d}\|^2 \mid x\right]$ immediately gives an upper bound on $e(x)$.

The interesting case is when x and $\hat{x}_{\mathrm{SA}}$ are not necessarily in the same subspace. Recalling that T is the index of the subspace selected in sparse approximation, orthogonally decompose d as $d = d_T \oplus d_{T^\perp}$ with d_T in the selected subspace and similarly decompose $\hat{d}$. Then $\hat{d}_T = d_T$ and the expected squared norm of this component can be bounded above as in the previous paragraph. Unfortunately, $\|\hat{d}_{T^\perp}\|$ can be larger than $\|d_{T^\perp}\|$ in proportion to $\|x\|$, as illustrated in Figure 3.2. The worst case is for $\|\hat{d}_{T^\perp}\| = 2\|d_{T^\perp}\|$, when y lies equidistant from the subspace of x and the subspace of $\hat{x}_{\mathrm{SA}}$.

From this analysis we obtain the weak bound

$$e(x) \;=\; \frac{1}{N}\mathbf{E}\left[\|x - \hat{x}_{\mathrm{SA}}\|^2 \mid x\right] \;\leq\; 4\sigma^2 \tag{3.5}$$

and the limiting low SNR bound

$$e(0) \;=\; \frac{1}{N}\mathbf{E}\left[\|x - \hat{x}_{\mathrm{SA}}\|^2 \mid x\right]|_{x=0} \;\leq\; \sigma^2\left(1 - c_2\left(1 - \frac{K}{N}\right)\right). \tag{3.6}$$

Chapter 4

Analysis for Isotropic Random Frames

In general, the performance of sparse approximation denoising is given by

$$e(x) = \frac{1}{N}\mathbf{E}\left[\|x - \hat{x}_{\mathrm{SA}}\|^2\right] = \frac{1}{N}\int_{R^{\mathbb{N}}} \left\| x - \left(\underset{\hat{x}\in\Phi_K}{\operatorname{argmin}} \|x + \eta - \hat{x}\|_2 \right) \right\|^2 f(\eta)\, d\eta$$

where $f(\cdot)$ is the density of the noise d. While this expression does not give any fresh insight, it does remind us that the performance depends on every element of Φ. In this section, we improve greatly upon (3.5) with an analysis that depends on each dictionary element being an independent random vector and on the dictionary being large. The results are expectations over both the noise d and the dictionary itself. In addition to analyzing the MSE, we also analyze the probability of error in the subspace selection, *i.e.*, the probability that x and $\hat{x}_{\mathrm{SA}}$ lie in different subspaces. In light of the simulations in Section 3.2, we expect these analyses to qualitatively match the performance of a variety of dictionaries.

Section 4.1 delineates the additional assumptions made in this section. The probability of error and MSE analyses are then given in Section 4.2. Estimates of the probability of error and MSE are numerically validated in Section 4.3, and finally limits as $N \to \infty$ are studied in Section 4.4

4.1 Modeling Assumptions

This section specifies the precise modeling assumptions in analyzing denoising performance with large, isotropic random frames. Though the results are limited to the case of $K = 1$, the model is described for general K. Difficulties in extending the results to general K are described in the concluding comments of the thesis. While many practical problems involve $K > 1$, the analysis of the $K = 1$ case presented here illustrates a number of unexpected qualitative phenomena, some of which have been observed for higher values of K.

The model is unchanged from earlier in the thesis except that the dictionary Φ and signal x are random:

(a) *Dictionary generation:* The dictionary Φ consists of M i.i.d. random vectors uniformly distributed on the unit sphere in $\mathbb{R}^N$.

(b) *Signal generation:* The true signal x is a linear combination of the first K dictionary elements so that

$$x = \sum_{i=1}^{K} \alpha_i \varphi_i,$$

for some random coefficients $\{\alpha_i\}$. The coefficients $\{\alpha_i\}$ are independent of the dictionary except in that x is normalized to have $\|x\|^2 = N$ for all realizations of the dictionary and coefficients.

(c) *Noise:* The noisy signal y is given by $y = x + d$ where, as before, $d \sim \mathcal{N}(0, \sigma^2 I_N)$. d is independent of Φ and x. We will let

$$\gamma = 1/\sigma^2,$$

which is the input SNR because of the scaling of x.

(d) *Estimator:* The estimator $\hat{x}_{\mathrm{SA}}$ is defined as before to be the optimal K-sparse approximation of y with respect to Φ. Specifically, we enumerate the $J = \binom{M}{K}$ K-element subsets of Φ. The jth subset spans a subspace denoted V_j and P_j denotes the projec-

tion operator onto V_j. Then

$$\hat{x}_{\mathrm{SA}} = P_T y \qquad \text{where} \qquad T = \underset{j \in \{1,2,\ldots,J\}}{\operatorname{argmin}} \|y - P_j y\|^2. \tag{4.1}$$

For the special case when M and N are large and $K = 1$, we will estimate two quantities:

Definition 1 *The* subspace selection error probability p_{err} *is defined as*

$$p_{\mathrm{err}} = \Pr\left(T \neq j_{\mathrm{true}}\right), \tag{4.2}$$

where T is the subspace selection index and j_{true} is the index of the subspace containing the true signal x, i.e.*, j_{true} is the index of the subset $\{1, 2, \ldots, K\}$.*

Definition 2 *The* normalized expected MSE *is defined as*

$$E_{\mathrm{MSE}} = \frac{1}{N\sigma^2} \boldsymbol{E}\left[\|x - \hat{x}_{\mathrm{SA}}\|^2\right] = \frac{\gamma}{N} \boldsymbol{E}\left[\|x - \hat{x}_{\mathrm{SA}}\|^2\right]. \tag{4.3}$$

Normalized expected MSE is the per-component MSE normalized by the per-component noise variance $\frac{1}{N}\mathbf{E}\left[\|d\|^2\right] = \sigma^2$. The term "expected MSE" emphasizes that the expectation in (4.3) is over not just the noise d, but also the dictionary Φ and signal x.

We will give tractable computations to estimate both p_{err} and E_{MSE}. Specifically, p_{err} can be approximated from a simple line integral and E_{MSE} can be computed from a double integral.

4.2 Analyses of Subspace Selection Error and MSE

The first result shows that the subspace selection error probability can be bounded by a double integral and approximately computed as a single integral. The integrands are simple functions of the problem parameters M, N, K and γ. While the result is only proven for the case of $K = 1$, K is left in the expressions to indicate the precise role of this parameter.

Theorem 3 *Consider the model described in Section 4.1. When $K = 1$ and M and N are large, the subspace selection error probability defined in (4.2) is bounded above by*

$$p_{\mathrm{err}} \;<\; 1 - \int_0^\infty \int_0^\infty f_r(u) f_s(v) \exp\left(-\frac{(CG(u,v))^r}{1 - G(u,v)}\right) 1_{\{G(u,v) \leq G_{\max}\}}\, dv\, du, \tag{4.4}$$

and p_{err} *is approximated well by*

$$\begin{aligned}\widehat{p}_{\text{err}}(N,M,K,\gamma) &= 1-\int_0^\infty f_r(u)\exp\left(-\left(\frac{C(N-K)\sigma^2u}{N+(N-K)\sigma^2u}\right)^r\right)du \\ &= 1-\int_0^\infty f_r(u)\exp\left(-\left(\frac{Cau}{1+au}\right)^r\right)du, \end{aligned} \tag{4.5}$$

where

$$G(u,v) = \frac{au}{au+\left(1-\sigma\sqrt{\frac{Kv}{N}}\right)^2} \tag{4.6}$$

$$G_{\max} = (r\beta(r,s))^{1/(r-1)},$$

$$C = \left(\frac{J-1}{r\beta(r,s)}\right)^{1/r}, \quad J=\binom{M}{K} \tag{4.7}$$

$$r = \frac{N-K}{2}, \quad s=\frac{K}{2}, \tag{4.8}$$

$$a = \frac{(N-K)\sigma^2}{N} = \frac{N-K}{N\gamma}, \tag{4.9}$$

$f_r(u)$ *is the probability distribution*

$$f_r(u) = r^r\Gamma(r)u^{r-1}e^{-ru}, \quad u\in[0,\infty), \tag{4.10}$$

$\beta(r,s)$ *is the beta function, and* $\Gamma(r)$ *is the Gamma function [2].*

Proof: See Section 6.2.

It is interesting to evaluate $\widehat{p}_{\text{err}}$ in two limiting cases. First, suppose that $J=1$. This corresponds to the situation where there is only one subspace. In this case, $C=0$ and (4.5) gives $\widehat{p}_{\text{err}}=0$. This is expected, since with one subspace there is no chance of a subspace selection error.

At the other extreme, suppose that N, K, and γ are fixed and $M\to\infty$. Then $C\to\infty$ and $\widehat{p}_{\text{err}}\to 1$. Again, this is expected since as the size of the frame increases, the number of possible subspaces increases and the probability of error increases.

The next result approximates the normalized expected MSE with a double integral. The integrand is relatively simple to evaluate and decays quickly as $\rho\to\infty$ and $u\to\infty$ so numerically approximating the double integral is not difficult.

Theorem 4 *Consider the model described in Section 4.1. When $K = 1$ and M and N are large, the normalized expected MSE defined in (4.3) is given approximately by*

$$\widehat{E}_{\mathrm{MSE}}(N, M, K, \gamma) = \frac{K}{N} + \int_0^\infty \int_0^\infty f_r(u) g_r(\rho) F(\rho, u)\, d\rho\, du, \tag{4.11}$$

where $f_r(u)$ is given in (4.10), $g_r(\rho)$ is the probability distribution

$$g_r(\rho) = rC^r r^{r-1} \exp\left(-(C\rho)^r\right), \tag{4.12}$$

$$F(\rho, u) = \begin{cases} \gamma(au(1-\rho)+\rho), & \text{if } \rho(1+au) < au; \\ 0, & \text{otherwise,} \end{cases} \tag{4.13}$$

and C, r, and a are defined in (4.7)–(4.9).

Proof: See Section 6.3.

4.3 Numerical Examples

We now present simulation results to examine the accuracy of the approximations in Theorems 3 and 4. Three pairs of (N, M) values were used: (5,1000), (10,100), and (10,1000). For each integer SNR from -10 dB to 35 dB, the subspace selection and normalized MSE were measured for 5×10^5 independent experiments. The resulting empirical probabilities of subspace selection error and normalized expected MSEs are shown in Figure 4.1. Plotted alongside the empirical results are the estimates $\widehat{p}_{\mathrm{err}}$ and $\widehat{E}_{\mathrm{MSE}}$ from (4.5) and (4.11).

Comparing the theoretical and measured values in Figure 4.1, we see that the theoretical values match the simulation closely over the entire SNR range. Also note that the bottom panel of Figure 4.1 shows qualitatively the same behavior as Figures 1.2 and 1.3 (the direction of the horizontal axis is reversed). In particular, $E_{\mathrm{MSE}} \approx \frac{K}{N}$ for high SNR and the low SNR behavior depends on M and N as described by (3.6).

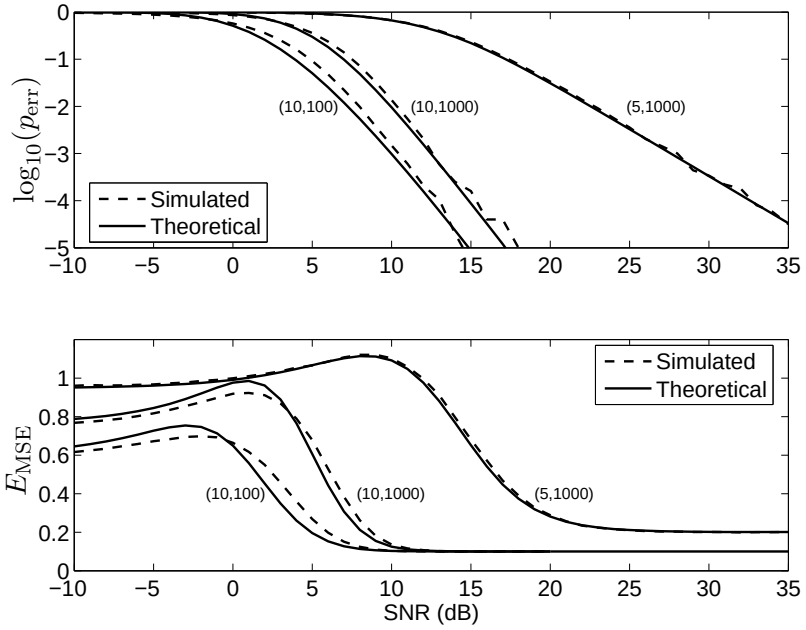


Figure 4.1. Simulation of subspace selection error probability and normalized expected MSE for isotropic random dictionaries. Calculations were made for integer SNRs (in dB), with 5×10^5 independent simulations per data point. In all cases $K = 1$. The curve pairs are labeled by (N, M). Simulation results are compared to the estimates from Theorems 3 and 4.

4.4 Asymptotic Analysis

The estimates $\widehat{p}_{\rm err}$ and $\widehat{E}_{\rm MSE}$ are not difficult to compute numerically, but the expressions (4.5) and (4.11) provide little direct insight. It is thus interesting to examine the asymptotic behavior of $\widehat{p}_{\rm err}$ and $\widehat{E}_{\rm MSE}$ as N and M grow. The following theorem gives an asymptotic expression for the limiting value of the error probability function.

Theorem 5 *Consider the function $\widehat{p}_{\rm err}(N,M,K,\gamma)$ defined in (4.5). Define the* critical SNR *as a function of M, N, and K as*

$$\gamma_{\rm crit} \;=\; C - 1 \;=\; \left(\frac{J-1}{r\beta(r,s)}\right)^{1/r} - 1. \tag{4.14}$$

where C, r, s and J are defined in (4.7) and (4.8). For $K = 1$ and any fixed γ and $\gamma_{\rm crit}$,

$$\lim_{\substack{N,M\to\infty \\ \gamma_{\rm crit}\ constant}} \widehat{p}_{\rm err}(N,M,K,\gamma) \;=\; \begin{cases} 1, & \text{if } \gamma < \gamma_{\rm crit}; \\ 0, & \text{if } \gamma > \gamma_{\rm crit}, \end{cases} \tag{4.15}$$

where the limit is on any sequence of M and N with $\gamma_{\rm crit}$ constant.

Proof: See Section 6.4.

The theorem shows that, asymptotically, there is a critical SNR $\gamma_{\rm crit}$ above which the error probability goes to one and below which the probability is zero. Thus, even though the frame is random, the error event asymptotically becomes deterministic.

A similar result holds for the asymptotic MSE.

Theorem 6 *Consider the function $\widehat{E}_{\rm MSE}(M,N,K,\gamma)$ defined in (4.11) and the critical SNR $\gamma_{\rm crit}$ defined in (4.14). For $K = 1$ and any fixed γ and $\gamma_{\rm crit}$,*

$$\lim_{\substack{N,M\to\infty \\ \gamma_{\rm crit}\ constant}} \widehat{E}_{\rm MSE}(M,N,K,\gamma) \;=\; \begin{cases} \widehat{E}_{\rm lim}(\gamma), & \text{if } \gamma < \gamma_{\rm crit}; \\ 0, & \text{if } \gamma > \gamma_{\rm crit}, \end{cases} \tag{4.16}$$

where the limit is on any sequence of M and N with $\gamma_{\rm crit}$ constant, and

$$\widehat{E}_{\rm lim}(\gamma) = \frac{\gamma + \gamma_{\rm crit}}{1 + \gamma_{\rm crit}}.$$

Proof: See Section 6.5.

Remarks:

(i) Theorems 5 and 6 hold for any values of K. They are stated for $K = 1$ because the significance of $\widehat{p}_{\text{err}}(N, M, K, \gamma)$ and $\widehat{E}_{\text{MSE}}(M, N, K, \gamma)$ is proven only for $K = 1$.

(ii) Both Theorems 5 and 6 involve limits with γ_{crit} constant. It is useful to examine how M, N and K must be related asymptotically for this condition to hold. One can use the definition of the beta function, $\beta(r, s) = \Gamma(r)\Gamma(s)/\Gamma(r + s)$, along with Stirling's approximation, to show that when $K \ll N$,

$$(r\beta(r, s))^{1/r} \approx 1. \tag{4.17}$$

Substituting (4.17) into (4.14), we see that $\gamma_{\text{crit}} \approx J^{1/r} - 1$. Also, for $K \ll N$ and $K \ll M$,

$$J^{1/r} = \binom{M}{K}^{2/(N-K)} \approx (M/K)^{2K/N},$$

so that

$$\gamma_{\text{crit}} \approx (M/K)^{2K/N} - 1$$

for small K and large M and N. Therefore, for γ_{crit} to be constant, $(M/K)^{2K/N}$ must be constant. Equivalently, the dictionary size M must grow as $K(1 + \gamma_{\text{crit}})^{N/(2K)}$, which is exponential in the inverse sparsity N/K.

The asymptotic normalized MSE is plotted in Figure 4.2 for various values of the critical SNR γ_{crit}. When $\gamma > \gamma_{\text{crit}}$, the normalized MSE is zero. This is expected: from Theorem 5, when $\gamma > \gamma_{\text{crit}}$, the estimator will always pick the correct subspace. We know that for a fixed subspace estimator, the normalized MSE is K/N. Thus, as $N \to \infty$, the normalized MSE approaches zero.

What is perhaps surprising is the behavior for $\gamma < \gamma_{\text{crit}}$. In this regime, the normalized MSE actually *increases* with increasing SNR. At the critical level, $\gamma = \gamma_{\text{crit}}$, the normalized MSE approaches its maximum value

$$\max \widehat{E}_{\text{lim}} = \frac{2\gamma_{\text{crit}}}{1 + \gamma_{\text{crit}}}.$$

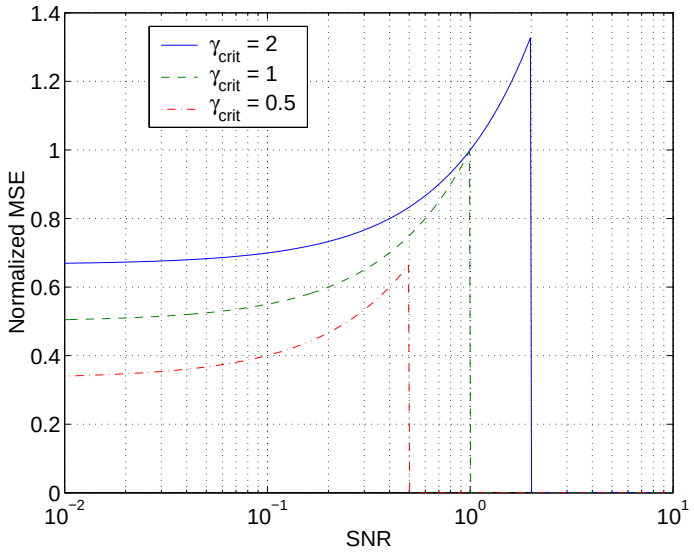


Figure 4.2. Asymptotic normalized MSE as $N \to \infty$ (from Theorem 6) for various critical SNRs γ_{crit}.

When $\gamma_{\text{crit}} > 1$, the limit of the normalized MSE $\widehat{E}_{\text{lim}}(\gamma)$ satisfies $\widehat{E}_{\text{lim}}(\gamma) > 1$. Consequently, the sparse approximation results in noise *amplification* instead of noise reduction. In the worst case, as $\gamma_{\text{crit}} \to \infty$, $\widehat{E}_{\text{lim}}(\gamma) \to 2$. Thus, sparse approximation can result in a noise amplification by a factor as large as 2. Contrast this with the factor of 4 in (3.5), which seems to be a very weak bound.

Chapter 5

Comments and Conclusions

This thesis has addressed properties of denoising by sparse approximation that are geometric in that the signal model is membership in a specified union of subspaces, without a probability density on that set. The denoised estimate is the feasible signal closest to the noisy observed signal.

The first main result (Theorems 1 and 2) is a bound on the performance of sparse approximation applied to a Gaussian signal. This lower bound on mean-squared approximation error is used to determine an upper bound on denoising MSE in the limit of low input SNR.

The remaining results apply to the expected performance when the dictionary itself is random with i.i.d. entries selected according to an isotropic distribution. Easy-to-compute estimates for the probability that the subspace containing the true signal is not selected and for the MSE are given (Theorems 3 and 4). The accuracy of these estimates is verified through simulations. Unfortunately, these results are proven only for the case of $K = 1$. The main technical difficulty in extending these results to general K is that the distances to the various subspaces are not mutually independent. (Though Lemma 2 does not extend to $K > 1$, we expect that a relation similar to (6.11) holds.)

Asymptotic analysis ($N \to \infty$) of the situation with a random dictionary reveals a critical value of the SNR (Theorems 5 and 6). Below the critical SNR, the probability of

selecting the subspace containing the true signal approaches zero and the expected MSE approaches a constant with a simple, closed form; above the critical SNR, the probability of selecting the subspace containing the true signal approaches one and the expected MSE approaches zero.

Sparsity with respect to a randomly generated dictionary is a strange model for naturally-occurring signals. However, most indications are that a variety of dictionaries lead to performance that is qualitatively similar to that of random dictionaries. Also, sparsity with respect to randomly generated dictionaries occurs when the dictionary elements are produced as the random instantiation of a communication channel. Both of these observations require further investigation.

Chapter 6

Proofs

6.1 Proof of Theorems 1 and 2

We begin with a proof of Theorem 2; Theorem 1 will follow easily. The proof is based on analyzing an idealized encoder for v. Note that despite the idealization and use of source coding theory, the bounds hold for any values of (N, M, K)—the results are *not* merely asymptotic. Readers unfamiliar with the basics of source coding theory are referred to any standard text, such as [4, 13, 33], though the necessary facts are summarized below.

Consider the encoder for v shown in Figure 6.1. The encoder operates by first finding the optimal sparse approximation of v, which is denoted by $\hat{v}$. The subspaces in Φ_K are assumed to be numbered, and the index of the subspace containing $\hat{v}$ is denoted by T. $\hat{v}$ is then quantized with a K-dimensional, b-bit quantizer represented by the box "Q" to produce the encoded version of v, which is denoted by $\hat{v}_{\mathrm{Q}}$.

The subspace selection T is a discrete random variable that depends on v. The average number of bits needed to communicate T to a receiver that knows the probability mass function of T is given by the entropy of T, which is denoted $H(T)$ [13]. In analyzing the encoder for v, we assume that a large number of independent realizations of v are encoded at once. This allows b to be an arbitrary real number (rather than an integer) and allows the average number of bits used to represent T to be arbitrarily close to $H(T)$. The encoder

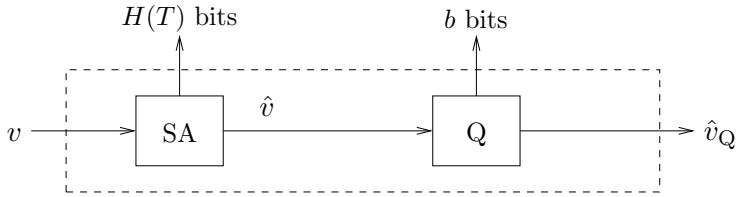


Figure 6.1. The proof of Theorem 2 is based on the analysis of a hypothetical encoder for v. The sparse approximation box "SA" finds the optimal K-sparse approximation of v, denoted $\hat{v}$, by computing $\hat{v} = P_T v$. The subspace selection T can be represented with $H(T)$ bits. The quantizer box "Q" quantizes $\hat{v}$ with b bits, with knowledge of T. The overall output of the encoder is denoted $\hat{v}_{\mathrm{Q}}$.

of Figure 6.1 can thus be considered to use $H(T) + b$ bits to represent v approximately as $\hat{v}_{\mathrm{Q}}$.

The crux of the proof is to represent the squared error that we are interested in, $\|v - \hat{v}\|^2$, in terms of squared errors of the overall encoder $v \mapsto \hat{v}_{\mathrm{Q}}$ and the quantizer $\hat{v} \mapsto \hat{v}_{\mathrm{Q}}$. We will show the orthogonality relationship below and bound both terms:

$$\mathbf{E}\left[\|v - \hat{v}\|^2\right] = \underbrace{\mathbf{E}\left[\|v - \hat{v}_{\mathrm{Q}}\|^2\right]}_{\text{bounded below using fact (a)}} - \underbrace{\mathbf{E}\left[\|\hat{v} - \hat{v}_{\mathrm{Q}}\|^2\right]}_{\text{bounded above using fact (b)}}.$$

The two facts we need from rate–distortion theory are [4, 13, 33]:

(a) The lowest possible per-component MSE for encoding an i.i.d. Gaussian source with per-component variance σ^2 with R bits per component is $\sigma^2 2^{-2R}$.

(b) Any source with per-component variance σ^2 can be encoded with R bits per component to achieve per-component MSE $\sigma^2 2^{-2R}$.

(The combination of facts (a) and (b) tells us that Gaussian sources are the hardest to represent when distortion is measured by MSE.)

Applying fact (a) to the $v \mapsto \hat{v}_{\mathrm{Q}}$ encoding, we get

$$\frac{1}{N}\mathbf{E}\left[\|v - \hat{v}_{\mathrm{Q}}\|^2\right] \geq \sigma^2 2^{-2(H(T)+b)/N}. \tag{6.1}$$

Now we would like to define the quantizer "Q" in Figure 6.1 to get the smallest possible upper bound on $\mathbf{E}\left[\|\hat{v} - \hat{v}_{\mathrm{Q}}\|^2\right]$.

Since the distribution of $\hat{v}$ does not have a simple form (*e.g.*, it is not Gaussian), we have no better tool than fact (b), which requires us only to find (or upper bound) the variance of the input to a quantizer. Consider a two-stage quantization process for $\hat{v}$. The first stage (with access to T) applies an affine, length-preserving transformation to $\hat{v}$ such that the result has mean zero and lies in a K-dimensional space. The output of the first stage is passed to an optimal b-bit quantizer. Using fact (b), the performance of such a quantizer must satisfy

$$\frac{1}{K}\mathbf{E}\left[\|\hat{v} - \hat{v}_{\mathrm{Q}}\|^2\right] \leq \sigma^2_{\hat{v}|T} 2^{-2b/K}, \tag{6.2}$$

where $\sigma^2_{\hat{v}|T}$ is the per-component conditional variance of $\hat{v}$, in the K-dimensional space, conditioned on T.

From here on we have slightly different reasoning for the $\bar{v} = 0$ and $\bar{v} \neq 0$ cases. For $\bar{v} = 0$, we get an exact expression for the desired conditional variance; for $\bar{v} \neq 0$, we use an upper bound.

When $\bar{v} = 0$, symmetry dictates that $\mathbf{E}\left[\hat{v} \mid T\right] = 0$ for all T and $\mathbf{E}\left[\hat{v}\right] = 0$. Thus the conditional variance $\sigma^2_{\hat{v}|T}$ and unconditional variance $\sigma^2_{\hat{v}}$ are equal. Taking the expectation of

$$\|v\|^2 = \|\hat{v}\|^2 + \|v - \hat{v}\|^2$$

gives

$$N\sigma^2 = K\sigma^2_{\hat{v}} + \mathbf{E}\left[\|v - \hat{v}\|^2\right].$$

Thus

$$\sigma^2_{\hat{v}|T} = \sigma^2_{\hat{v}} = \frac{1}{K}\left(N\sigma^2 - \mathbf{E}\left[\|v - \hat{v}\|^2\right]\right) = \frac{N}{K}(\sigma^2 - D_{\mathrm{SA}}), \tag{6.3}$$

where we have used D_{SA} to denote $\frac{1}{N}\mathbf{E}\left[\|v - \hat{v}\|^2\right]$—which is the quantity we are bounding in the theorem. Substituting (6.3) into (6.2) now gives

$$\frac{1}{K}\mathbf{E}\left[\|\hat{v} - \hat{v}_{\mathrm{Q}}\|^2\right] \leq \frac{N(\sigma^2 - D_{\mathrm{SA}})}{K} 2^{-2b/K}. \tag{6.4}$$

To usefully combine (6.1) and (6.4), we need one more orthogonality fact. Since the quantizer Q operates in subspace T, its quantization error is also in subspace T. On the

other hand, because $\hat{v}$ is produced by orthogonal projection to subspace T, $v-\hat{v}$ is orthogonal to subspace T. So

$$\|v - \hat{v}_{\mathrm{Q}}\|^2 = \|\hat{v} - \hat{v}_{\mathrm{Q}}\|^2 + \|v - \hat{v}\|^2.$$

Taking expectations, rearranging, and substituting (6.1) and (6.4) gives

$$\begin{aligned} \mathbf{E}\left[\|v - \hat{v}\|^2\right] &= \mathbf{E}\left[\|v - \hat{v}_{\mathrm{Q}}\|^2\right] - \mathbf{E}\left[\|\hat{v} - \hat{v}_{\mathrm{Q}}\|^2\right] \\ &\geq N\sigma^2 2^{-2(H(T)+b)/N} - N(\sigma^2 - D_{\mathrm{SA}})2^{-2b/K}. \end{aligned} \tag{6.5}$$

Recalling that the left-hand side of (6.5) is ND_{SA} and rearranging gives

$$D_{\mathrm{SA}} \;\geq\; \sigma^2 \left(\frac{2^{-2(H(T)+b)/N} - 2^{-2b/K}}{1 - 2^{-2b/K}} \right). \tag{6.6}$$

Since this bound must be true for all $b \geq 0$, one can maximize with respect to b to obtain the strongest bound. This maximization is messy; however, maximizing the numerator is easier and gives almost as strong a bound. The numerator is maximized when

$$b \;=\; \frac{K}{N-K}\left(H(T) + \frac{N}{2}\log_2 \frac{N}{K} \right),$$

and substituting this value of b in (6.6) gives

$$D_{\mathrm{SA}} \;\geq\; \sigma^2 \cdot \frac{2^{-2H(T)/(N-K)}\left(1 - \frac{K}{N}\right)\left(\frac{K}{N}\right)^{K/(N-K)}}{1 - 2^{-2H(T)/(N-K)}\left(\frac{K}{N}\right)^{N/(N-K)}}.$$

We have now completed the proof of Theorem 2 for $\bar{v} = 0$.

For $\bar{v} \neq 0$, there is no simple expression for $\sigma^2_{\hat{v}|T}$ that does not depend on the geometry of the dictionary, such as (6.3), to use in (6.2). Instead, use

$$\sigma^2_{\hat{v}|T} \;\leq\; \sigma^2_{\hat{v}} \;\leq\; \frac{N}{K}\sigma^2,$$

where the first inequality holds because conditioning cannot increase variance and the second follows from fact that the orthogonal projection of v cannot increase its variance, even if the choice of projection depends on v. Now following the same steps as for the $\bar{v} = 0$ case yields

$$D_{\mathrm{SA}} \;\geq\; \sigma^2 \left(2^{-2(H(T)+b)/N} - 2^{-2b/K} \right)$$

in place of (6.6). The bound is optimized over b to obtain

$$D_{\mathrm{SA}} \;\geq\; \sigma^2 \cdot 2^{-2H(T)/(N-K)} \left(1 - \tfrac{K}{N}\right) \left(\tfrac{K}{N}\right)^{K/(N-K)}.$$

The proof of Theorem 1 now follows directly: since T is a discrete random variable that can take at most J values, $H(T) \leq \log_2 J$.

6.2 Proof of Theorem 3

Using the notation of Section 4.1, let V_j, $j = 1, 2, \ldots, , J$, be the subspaces spanned by the J possible K-element subsets of the dictionary Φ. Let P_j be the projection operator onto V_j, and let T be index of the subspace closest to y. Let j_{true} be the index of the subspace containing the true signal x, so that the probability of error is

$$p_{\mathrm{err}} = \Pr(T \neq j_{\mathrm{true}}).$$

For each j, let $\hat{x}_j = P_j y$, so that the estimator, $\hat{x}_{\mathrm{SA}}$ in (4.1) can be rewritten as $\hat{x}_{\mathrm{SA}} = \hat{x}_T$. Also, define random variables

$$\rho_j = \|y - \hat{x}_j\|^2 / \|y\|^2, \quad j = 1, 2, \ldots, J$$

to represent the normalized distances between y and the V_j's. Henceforth, the ρ_j's will be called *angles*, since $\rho_j = \sin^2 \theta_j$ where θ_j is the angle between y and V_j. The angles are well defined since $\|y\|^2 > 0$ with probability one.

Lemma 1 *For all $j \neq j_{\mathrm{true}}$, the angle ρ_j is independent of x and d.*

Proof: Given a subspace V and vector y, define the function

$$R(y, V) = \|y - P_V y\|^2 / \|y\|^2, \tag{6.7}$$

where P_V is the projection operator onto the subspace y. Thus, $R(y, V)$ is the angle between y and V. With this notation, $\rho_j = R(y, V_j)$. Since ρ_j is a deterministic function of y and V_j and $y = x + d$, to show ρ_j is independent of x and d, it suffices to prove that ρ_j is

independent of y. Equivalently, we need to show that for any function $G(\rho)$ and vectors y_0 and y_1,

$$\mathbf{E}\left[G(\rho_j) \mid y = y_0\right] = \mathbf{E}\left[G(\rho_j) \mid y = y_1\right].$$

This property can be proven with the following symmetry argument: Let U be any orthogonal transformation. Since U is orthogonal, $P_{UV}(Uy) = UP_V y$ for all subspaces V and vectors y. Combining this with the fact that $\|Uv\| = \|v\|$ for all v, we see that

$$\begin{aligned} R(Uy, UV) &= \|Uy - P_{UV}(Uy)\|^2/\|Uy\|^2 \;=\; \|U(y - P_V(y))\|^2/\|Uy\|^2 \\ &= \|y - P_V(y)\|^2/\|y\|^2 \;=\; R(y, V). \end{aligned} \tag{6.8}$$

Also, for any scalar $\alpha > 0$, it can be verified that $R(\alpha y, V) = R(y, V)$.

Now, let y_0 and y_1 be any two possible non-zero values for the vector y. Then, there exists an orthogonal transformation U and scalar $\alpha > 0$ such that $y_1 = \alpha U y_0$. Since $j \neq j_{\text{true}}$ and $K = 1$, the subspace V_j is spanned by vectors φ_i, independent of the vector y. Therefore

$$\begin{aligned} \mathbf{E}\left[G(\rho_j) \mid y = y_1\right] &= \mathbf{E}\left[G(R(y_1, V_j))\right] \;=\; \mathbf{E}\left[G(R(\alpha U y_0, V_j))\right] \\ &= \mathbf{E}\left[G(R(U y_0, V_j))\right]. \end{aligned} \tag{6.9}$$

Now since the elements of Φ are distributed uniformly on the unit sphere, the subspace UV_j is identically distributed to V_j. Combining this with (6.8) and (6.9),

$$\begin{aligned} \mathbf{E}\left[G(\rho_j) \mid y = y_1\right] &= \mathbf{E}\left[G(R(U y_0, V_j))\right] \;=\; \mathbf{E}\left[G(R(U y_0, UV_j))\right] \\ &= \mathbf{E}\left[G(R(y_0, V_j))\right] \;=\; \mathbf{E}\left[G(\rho_j) \mid y = y_0\right] \end{aligned}$$

and this completes the proof. □

Lemma 2 *The random angles ρ_j, $j \neq j_{\text{true}}$ are i.i.d., each with a probability density function given by the beta distribution,*

$$p_\rho(\rho) = \frac{1}{\beta(r, s)} \rho^{r-1}(1 - \rho)^{s-1}, \qquad 0 \leq \rho \leq 1, \tag{6.10}$$

where $r = (N - K)/2$ and $s = K/2$ as defined in (4.8).

Proof: Since $K = 1$, each of the subspaces V_j for $j \neq j_{\text{true}}$ is spanned by a single, unique vector in Φ. Since the vectors in Φ are independent and the random variables ρ_j are the angles between y and the spaces V_j, the angles are independent.

Now consider a single angle ρ_j for $j \neq j_{\text{true}}$. The angle ρ_j is the angle between y and a random subspace V_j. Since the distribution of the random vectors defining V_j is spherically symmetric and ρ_j is independent of y, ρ_j is identically distributed to the angle between any fixed subspace V and a random vector z uniformly distributed on the unit sphere. One way to create such a random vector z is to take $z = w/\|w\|$, where $w \sim \mathcal{N}(0, I_N)$. Let $w_1, w_2, \ldots, w_K$ be the components of w in V, and $w_{K+1}, w_{K+2}, \ldots, w_N$ be the components in the orthogonal complement to V. If we define

$$X = \sum_{i=1}^{K} w_i^2 \qquad \text{and} \qquad Y = \sum_{i=K+1}^{N} w_i^2,$$

then the angle between z and V is $\rho = Y/(X+Y)$. Since X and Y are the sums of K and $N-K$ i.i.d. squared Gaussian random variables, they are Chi-squared random variables with K and $N-K$ degrees of freedom, respectively [38]. Now, a well-known property of Chi-squared random variables is that if X and Y are Chi-squared random variables with m and n degrees of freedom, $Y/(X+Y)$ will have the beta distribution with parameters $m/2$ and $n/2$. Thus, $\rho = Y/(X+Y)$ has the beta distribution, with parameters r and s defined in (4.8). The probability density function for the beta distribution is given in (6.10). □

Lemma 3 *Let $\rho_{\min} = \min_{j \neq j_{\text{true}}} \rho_j$. Then $\rho_{\min}$ is independent of x and d and has the approximate distribution*

$$\Pr(\rho_{\min} > \epsilon) \approx \exp\left(-(C\epsilon)^r\right) \tag{6.11}$$

for small ϵ, where C is given in (4.7). More precisely,

$$\Pr(\rho_{\min} > \epsilon) \;<\; \exp(-(C\epsilon)^r(1-\epsilon)^{s-1}) \quad \textit{for all } \epsilon \in (0,1) \tag{6.12}$$

and

$$\Pr(\rho_{\min} > \epsilon) \;>\; \exp(-(C\epsilon)^r/(1-\epsilon)) \quad \textit{for } 0 < \epsilon < (r\beta(r,s))^{1/(r-1)}. \tag{6.13}$$

Proof: Since Lemma 1 shows that each ρ_j is independent of x and d, it follows that $\rho_{\min}$ is independent of x and d as well. Also, for any $j \neq j_{\text{true}}$, by bounding the integrand of

$$\Pr(\rho_j < \epsilon) \;=\; \frac{1}{\beta(r,s)} \int_0^\epsilon \rho^{r-1}(1-\rho)^{s-1}\, d\rho$$

from above and below we obtain the bounds

$$\frac{(1-\epsilon)^{s-1}}{\beta(r,s)} \int_0^\epsilon \rho^{r-1}\, d\rho \;<\; \Pr(\rho_j < \epsilon) \;<\; \frac{1}{\beta(r,s)} \int_0^\epsilon \rho^{r-1}\, d\rho,$$

which simplifies to

$$\frac{(1-\epsilon)^{s-1}\epsilon^r}{r\beta(r,s)} \;<\; \Pr(\rho_j < \epsilon) \;<\; \frac{\epsilon^r}{r\beta(r,s)}. \tag{6.14}$$

Now, there are $J-1$ subspaces V_j where $j \neq j_{\text{true}}$, and, by Lemma 2, the ρ_j's are mutually independent. Consequently, if we apply the upper bound of (6.14) and $1-\delta > \exp(-\delta/(1-\delta))$ for $\delta \in (0,1)$, with $\delta = \epsilon^r/(r\beta(r,s))$, we obtain

$$\begin{aligned}
\Pr(\rho_{\min} > \epsilon) \;&=\; \prod_{j \neq j_{\text{true}}} \Pr(\rho_j > \epsilon) \;>\; \left(1 - \frac{\epsilon^r}{r\beta(r,s)}\right)^{J-1} \\
&>\; \exp\left(-\frac{\epsilon^r(J-1)}{r\beta(r,s)(1-\delta)}\right) \quad \text{for } 0 < \epsilon < (r\beta(r,s))^{1/r} \\
&>\; \exp\left(-\frac{\epsilon^r(J-1)}{r\beta(r,s)(1-\epsilon)}\right) \quad \text{for } 0 < \epsilon < (r\beta(r,s))^{1/(r-1)}.
\end{aligned}$$

Similarly, using the lower bound of (6.14), we obtain

$$\begin{aligned}
\Pr(\rho_{\min} > \epsilon) \;&=\; \prod_{j \neq j_{\text{true}}} \Pr(\rho_j > \epsilon) \;<\; \left(1 - \frac{(1-\epsilon)^{s-1}\epsilon^r}{r\beta(r,s)}\right)^{J-1} \\
&<\; \exp\left(-\frac{(1-\epsilon)^{s-1}\epsilon^r(J-1)}{r\beta(r,s)}\right).
\end{aligned}$$

□

Proof of Theorem 3: Let V_{true} be the "correct" subspace, *i.e.*, $V_{\text{true}} = V_j$ for $j = j_{\text{true}}$. Let D_{true} be the squared distance from y to V_{true}, and let $D_{\min}$ be the minimum of the squared distances from y to the "incorrect" subspaces, V_j, $j \neq j_{\text{true}}$. Since the estimator selects the closest subspace, there is an error if and only if $D_{\min} \leq D_{\text{true}}$. Thus,

$$p_{\text{err}} = \Pr(D_{\min} \leq D_{\text{true}}). \tag{6.15}$$

To estimate this quantity, we will approximate the probability distributions of $D_{\min}$ and D_{true}.

First consider D_{true}. Write the noise vector d as $d = d_0 + d_1$, where d_0 is the component in V_{true} and d_1 is in $V_{\text{true}}^{\perp}$. Let $D_0 = \|d_0\|^2$ and $D_1 = \|d_1\|^2$. Since $y = x + d$ and $x \in V_{\text{true}}$, the squared distance from y to V_{true} is D_1. Thus,

$$D_{\text{true}} = D_1. \tag{6.16}$$

Now consider $D_{\min}$. For any j, $\hat{x}_j$ is the projection of y onto V_j. Thus, the squared distance from y to any space V_j is $\|y - \hat{x}_j\|^2 = \rho_j \|y\|^2$. Hence, the minimum of the squared distances from y to the spaces V_j, $j \neq j_{\text{true}}$ is

$$D_{\min} = \rho_{\min} \|y\|^2. \tag{6.17}$$

We will bound and approximate $\|y\|^2$ to obtain the bound and approximation of the theorem. Notice that $y = x + d = x + d_0 + d_1$ where $x + d_0 \in V_{\text{true}}$ and $d_1 \in V_{\text{true}}^{\perp}$. Using this orthogonality and the triangle inequality we obtain the bound

$$\begin{aligned} \|y\|^2 &= \|x + d_0\|^2 + \|d_1\|^2 \;\geq\; (\|x\| - \|d_0\|)^2 + \|d_1\|^2 \\ &= (\sqrt{N} - \sqrt{D_0})^2 + D_1. \end{aligned} \tag{6.18}$$

For an accurate approximation, note that since d_0 is the component of d in the K-dimensional space V_{true}, we have $D_0 \ll N$ unless the SNR is very low. Thus

$$\|y\|^2 \;\approx\; N + D_1. \tag{6.19}$$

Combining (6.15), (6.16), and (6.17) gives

$$p_{\text{err}} \;=\; \Pr\left(D_{\min} \leq D_{\text{true}}\right) \;=\; \Pr\left(\rho_{\min}\|y\|^2 \leq D_1\right) \;=\; \Pr\left(\rho_{\min} \leq D_1/\|y\|^2\right). \tag{6.20}$$

Note that by Lemma 3, $\rho_{\min}$ is independent of x and d. Therefore, $\rho_{\min}$ is independent of D_0 and D_1. We can now obtain a bound and an approximation from (6.20) by taking expectations over D_0 and D_1.

To obtain a *bound*, combine the lower bound on $\Pr(\rho_{\min} > \epsilon)$ from (6.13) with (6.18):

$$\begin{aligned} p_{\text{err}} &< \Pr\left(\rho_{\min} \leq \frac{D_1}{D_1 + (\sqrt{N} - \sqrt{D_0})^2}\right) \\ &= \Pr\left(\rho_{\min} \leq \frac{\sigma^2(N-K)U}{\sigma^2(N-K)U + (\sqrt{N} - \sigma\sqrt{KV})^2}\right) \\ &= \Pr\left(\rho_{\min} \leq \frac{aU}{aU + \left(1 - \sigma\sqrt{\frac{KV}{N}}\right)^2}\right) \\ &= \Pr\left(\rho_{\min} \leq G(U,V)\right) \\ &\leq \mathbf{E}\left[1 - \exp\left(-\frac{(CG(U,V))^r}{1-G(U,V)} 1_{\{G(U,V) \leq G_{\max}\}}\right)\right], \end{aligned}$$

where we have started with (6.18) substituted in (6.20); the first equality uses $U = D_1/((N-K)\sigma^2)$, which is a normalized Chi-squared random variable with $N - K = 2r$ degrees of freedom and $V = D_0/(K\sigma^2)$, which is a normalized Chi-squared random variable with $K = 2s$ degrees of freedom [38]; the last equality uses the definition of G from the statement of the theorem; and the final inequality is an application of Lemma 3. This yields (4.4).

To obtain an *approximation*, combine the approximation of $\Pr(\rho_{\min} > \epsilon)$ from (6.11) with (6.19):

$$\begin{aligned} p_{\text{err}} &\approx \Pr\left(\rho_{\min} \leq \frac{D_1}{N + D_1}\right) \\ &= \Pr\left(\rho_{\min} \leq \frac{\sigma^2(N-K)U}{N + \sigma^2(N-K)U}\right) \\ &= \Pr\left(\rho_{\min} \leq \frac{aU}{1+aU}\right) \\ &\approx \mathbf{E}\left[1 - \exp\left(-\left(C\frac{aU}{1+aU}\right)^r\right)\right], \end{aligned}$$

which yields (4.5). This completes the proof.

6.3 Proof of Theorem 4

We will continue with the notation of the proof of Theorem 3. To approximate the MSE, we will need yet another property of the random angles ρ_j.

Lemma 4 *For any subspace $j \neq j_{\text{true}}$, $\boldsymbol{E}[\hat{x}_j \mid \rho_j, y] = (1-\rho_j)y$.*

Proof: Define the random variable $w_j = \hat{x}_j - (1-\rho_j)y$, and let $\mu_j = \mathbf{E}[w_j \mid \rho_j, y]$. Then

$$\mathbf{E}[\hat{x}_j \mid \rho_j, y] = (1-\rho_j)y + \mu_j.$$

So the lemma will be proven if we can show $\mu_j = 0$. To this end, first observe that since $\hat{x}_j$ is the projection of y onto the space V_j, $\hat{x}_j - y$ is orthogonal to $\hat{x}_j$. Using this fact along with the definition of ρ_j,

$$\begin{aligned} w_j'y &= (\hat{x}_j - (1-\rho_j)y)'y \;=\; \hat{x}_j'y - \|y\|^2 + \rho_j\|y\|^2 \\ &= \hat{x}_j'y - \|y\|^2 + \|\hat{x}_j - y\|^2 \;=\; \hat{x}_j'(\hat{x}_j - y) \;=\; 0. \end{aligned}$$

That is, w_j is orthogonal to y. Consequently, $\mu_j = \mathbf{E}[w_j \mid \rho_j, y]$ is orthogonal to y as well.

We can now show $\mu_j = 0$ from a symmetry argument similar to that used in the proof of Lemma 1. For any vector y and subspace V, define the function

$$W(y, V) \;=\; P_V y - (1 - R(y,V))y,$$

where, as in the proof of Lemma 1, P_V is the projection operator onto V, and $R(y,V)$ is given in (6.7). Since $\rho_j = R(y, V_j)$, we can rewrite w_j as

$$w_j \;=\; \hat{x}_j - \rho_j y \;=\; P_{V_j} y - (1 - R(y,V_j))y \;=\; W(y, V_j).$$

The proof of Lemma 1 showed that, for any orthogonal transformation U, $P_{UV}(Uy) = UP_V y$ and $R(Uy, UV) = R(y,V)$. Therefore,

$$\begin{aligned} W(Uy, UV) &= P_{UV}(Uy) - (1 - R(Uy, UV))(Uy) \\ &= UP_V y - U(1 - R(y,V))y \;=\; U(P_V y - (1 - R(y,V))y) \;=\; UW(y, V). \end{aligned} \tag{6.21}$$

Now, fix y and let U be any fixed orthogonal transformation of $\mathbb{R}^N$ with the property that $Uy = y$. Since U is orthogonal and the space V_j is generated by random vectors with a spherically symmetric distribution, UV_j is identically distributed to V_j. Combining this

with (6.21) and the fact that $Uy = y$ gives

$$\begin{aligned}
\mu_j &= \mathbf{E}\left[w_j \mid \rho_j, y\right] = \mathbf{E}\left[W(y, V_j) \mid \rho_j, y\right] \\
&= \mathbf{E}\left[W(Uy, V_j) \mid \rho_j, y\right] \quad (\text{since } Uy = y) \\
&= \mathbf{E}\left[W(Uy, UV_j) \mid \rho_j, y\right] \quad (\text{since } UV_j \text{ is distributed identically to } V_j) \\
&= \mathbf{E}\left[UW(y, V_j) \mid \rho_j, y\right] = U\mu_j.
\end{aligned}$$

Therefore, $\mu_j = U\mu_j$ for all orthogonal transformations U such that $Uy = y$. Hence, μ_j must be spanned by y. But, we showed above that μ_j is orthogonal to y. Thus $\mu_j = 0$, and this proves the lemma. □

Proof of Theorem 4: As in the proof of Theorem 3, let D_0 and D_1 be the squared norms of the components of d on the spaces V_{true} and $V_{\text{true}}^{\perp}$ respectively. Also, let $U = D_1/((N-K)\sigma^2)$. Define the random variable

$$E_0 = \frac{1}{N\sigma^2}(\|x - \hat{x}_{\text{SA}}\|^2 - D_0)$$

and its conditional expectation

$$F_0(\rho, u) = \mathbf{E}\left[E_0 \mid \rho_{min} = \rho, U = u\right].$$

Differentiating the approximate cumulative distribution function of $\rho_{\min}$ in Lemma 3, we see that $\rho_{\min}$ has an approximate probability density function of $f_r(\rho)$. Also, as argued in the proof of Theorem 3, U has the probability density function given by $g_r(u)$. Therefore

$$\begin{aligned}
E_{\text{MSE}} &= \frac{1}{N\sigma^2}\mathbf{E}\left[\|x - \hat{x}_{\text{SA}}\|^2\right] \\
&\approx \frac{1}{N\sigma^2}\int_0^\infty \int_0^\infty f_r(\rho) g_r(u) \mathbf{E}\left[\|x - \hat{x}_{\text{SA}}\|^2 \mid \rho_{\min} = \rho, U = u\right] du\, d\rho \\
&= \int_0^\infty \int_0^\infty f_r(\rho) g_r(u) \left(F_0(\rho, u) + \frac{1}{N\sigma^2}\mathbf{E}\left[D_0 \mid \rho_{min} = \rho, U = u\right]\right) du\, d\rho \\
&= \frac{1}{N\sigma^2}\mathbf{E}\left[D_0\right] + \int_0^\infty \int_0^\infty f_r(\rho) g_r(u) F_0(\rho, u)\, du\, d\rho \\
&= \frac{K}{N} + \int_0^\infty \int_0^\infty f_r(\rho) g_r(u) F_0(\rho, u)\, du\, d\rho. \qquad (6.22)
\end{aligned}$$

In the last step, we have used the fact that $D_0 = \|d_0\|^2$, where d_0 is the projection of d onto the K-dimensional subspace V_{true}. Since d has variance σ^2 per dimension, $\mathbf{E}\left[D_0\right] = K\sigma^2$.

Comparing (6.22) with (4.11), the theorem will be proven if we can show that

$$F_0(\rho, u) \approx F(\rho, u),$$

where $F(\rho, u)$ is given in (4.13).

We consider two cases: when $T = j_{\text{true}}$ and $T \neq j_{\text{true}}$. First, consider the case $T = j_{\text{true}}$. In this case, $\hat{x}_{\text{SA}}$ is the projection of y onto the true subspace V_{true}. The error $x - \hat{x}_{\text{SA}}$ will be precisely d_0, the component of the noise d on V_{true}. Thus,

$$\|x - \hat{x}_{\text{SA}}\|^2 = \|d_0\|^2 = D_0.$$

Consequently, when $T = j_{\text{true}}$,

$$E_0 = \frac{1}{N\sigma^2}(\|x - \hat{x}_{\text{SA}}\|^2 - D_0) = 0.$$

Taking the conditional expectation with respect to $\rho_{\min}$, U and the event that $T = j_{\text{true}}$,

$$\mathbf{E}\left[E_0 \mid T = j_{\text{true}}, \rho_{\min}, U\right] = 0. \tag{6.23}$$

Next consider the case $T \neq j_{\text{true}}$. In this case, we divide the approximation error into three terms:

$$\|x - \hat{x}_{\text{SA}}\|^2 = \|y - x\|^2 + \|y - \hat{x}_{\text{SA}}\|^2 - 2(y - x)'(y - \hat{x}_{\text{SA}}). \tag{6.24}$$

We take the conditional expectation of the three terms in (6.24) given $T \neq j_{\text{true}}$, D_0, D_1 and ρ_{min}.

For the first term in (6.24), observe that since $y - x = d$, and $\|d\|^2 = D_0 + D_1$,

$$\|y - x\|^2 = \|d\|^2 = D_0 + D_1.$$

Therefore, since $\rho_{\min}$ is independent of d,

$$\mathbf{E}\left[\|y - x\|^2 \mid T \neq j_{\text{true}}, D_0, D_1, \rho_{\min}\right] = D_0 + D_1. \tag{6.25}$$

For the second term in (6.24), let $\hat{x}_j$ be the projection of y onto the jth subspace V_j. By the definition of ρ_j,

$$\|y - \hat{x}_j\|^2 = \rho_j \|y\|^2.$$

Therefore, when $T \neq j_{\text{true}}$,

$$\|y - \hat{x}_{\text{SA}}\|^2 = \rho_{\min}\|y\|^2.$$

Using the approximation in the proof of Theorem 3 that $\|y\|^2 \approx N + D_1$,

$$\|y - \hat{x}_{\text{SA}}\|^2 \approx \rho_{\min}(N + D_1).$$

Hence,

$$\mathbf{E}\left[\|y - \hat{x}_{\text{SA}}\|^2 \mid T \neq j_{\text{true}}, \rho_{\min}, D_0, D_1\right] \approx \rho_{\min}(N + D_1). \tag{6.26}$$

Evaluating the last term in (6.24) with Lemma 4 we obtain

$$\begin{aligned}\mathbf{E}\left[(y-x)'(y-\hat{x}_j) \mid x, d, \rho_j\right] &= \mathbf{E}\left[d'(y-\hat{x}_j) \mid x, d, \rho_j\right] \\ &= d'y - d'\mathbf{E}\left[\hat{x}_j \mid x, d, \rho_j\right] = d'y - (1-\rho_j)d'y = \rho_j d'y = \rho_j d'(x+d).\end{aligned}$$

Therefore,

$$\mathbf{E}\left[(y-x)'(y-\hat{x}_{\text{SA}}) \mid T \neq j_{\text{true}}, x, d, \rho_j\right] = \rho_{\min} d'(x+d).$$

Since d is independent of x and $d'd = \|d\|^2 = D_0 + D_1$,

$$\mathbf{E}\left[(y-x)'(y-\hat{x}_{\text{SA}}) \mid T \neq j_{\text{true}}, \rho_{\min}, D_0, D_1\right] = \rho_{\min}(D_0 + D_1) \approx \rho_{\min} D_1 \tag{6.27}$$

since $D_1 \gg D_0$. Substituting (6.25), (6.26) and (6.27) into (6.24),

$$\begin{aligned}\mathbf{E}\left[\|x - \hat{x}_{\text{SA}}\|^2 \mid T \neq j_{\text{true}}, D_0, D_1, \rho_{\min}\right] &\approx D_0 + D_1 + \rho_{\min}(N + D_1) - 2\rho_{\min} D_1 \\ &= D_0 + D_1(1 - \rho_{\min}) + \rho_{\min} N.\end{aligned}$$

Combining this with the definitions $U = D_1/\sigma^2(N-K)$, $a = (N-K)/N\gamma$, and $\gamma = 1/\sigma^2$,

$$\begin{aligned}\mathbf{E}\left[E_0 \mid T \neq j_{\text{true}}, D_0, D_1, \rho_{\min}\right] &= \frac{1}{N\sigma^2}\mathbf{E}\left[\|x - \hat{x}_{\text{SA}}\|^2 - D_0 \mid T \neq j_{\text{true}}, D_0, D_1, \rho_{\min}\right] \\ &= \frac{1}{N\sigma^2}(D_1(1-\rho_{\min}) + N\rho_{\min}) \\ &= \gamma\left(aU(1-\rho_{\min}) + \rho_{\min}\right).\end{aligned}$$

Hence,

$$\mathbf{E}\left[E_0 \mid T \neq j_{\text{true}}, U, \rho_{\min}\right] \approx \gamma\left(aU(1-\rho_{\min}) + \rho_{\min}\right). \tag{6.28}$$

Now, from the proof of Theorem 3, we saw that $T \neq j_{\text{true}}$ is approximately equivalent to the condition that

$$\rho_{\min} < D_1/(N + D_1) = aU/(1 + aU).$$

Combining this with (6.23) and (6.28),

$$\begin{aligned} F_0(\rho, u) &= \mathbf{E}\left[E_0 \mid \rho_{\min} = \rho, U = u\right] \\ &\approx \begin{cases} \gamma(au(1-\rho)+\rho), & \text{if } \rho < au/(1+au); \\ 0, & \text{otherwise} \end{cases} \\ &= F(\rho, u) \end{aligned}$$

This shows that $F_0(\rho, u) \approx F(\rho, u)$ and completes the proof.

6.4 Proof of Theorem 5

The function $g_r(u)$ is the p.d.f. of a normalized Chi-squared random variable with $2r$ degrees of freedom [38]. That is, $g_r(u)$ is the p.d.f. of a variable of the form

$$U_r = \frac{1}{2r} \sum_{i=1}^{2r} X_i^2,$$

where the X_i's are i.i.d. Gaussian random variables with zero mean and unit variance. Therefore, we can rewrite $\widehat{p}_{\text{err}}$ as

$$\widehat{p}_{\text{err}} = 1 - \mathbf{E}\left[\exp\left(-\left(\frac{aCU_r}{1+aU_r}\right)^r\right)\right],$$

where the expectation is over the variable U_r. Now, by the strong law of large numbers,

$$\lim_{r\to\infty} U_r = 1, \text{ a.s.}$$

Also, if $K = 1$ and γ are fixed,

$$\lim_{N\to\infty} a \;=\; \lim_{N\to\infty} \frac{N-K}{\gamma N} \;=\; \gamma^{-1}. \tag{6.29}$$

Taking the limit $N, M \to \infty$, with $K = 1$ and C constant,

$$\begin{aligned}
\lim_{N,M\to\infty} \widehat{p}_{\text{err}} &= \lim_{r\to\infty} \left[1 - \exp\left(-\left(\frac{C}{1+\gamma}\right)^r\right)\right] \\
&= \begin{cases} 1, & \text{if } \gamma + 1 < C; \\ 0, & \text{if } \gamma + 1 > C \end{cases} \\
&= \begin{cases} 1, & \text{if } \gamma < \gamma_{\text{crit}}; \\ 0, & \text{if } \gamma > \gamma_{\text{crit}}. \end{cases}
\end{aligned}$$

6.5 Proof of Theorem 6

As in the proof of Theorem 5, let U_r be a normalized Chi-squared variable with p.d.f. $g_r(u)$. Also let ρ_r be a random variable with p.d.f. $f_r(\rho)$. Then, we can write $\widehat{E}_{\text{MSE}}$ as

$$\widehat{E}_{\text{MSE}} = K/N + \mathbf{E}\left[F(\rho_r, U_r)\right], \tag{6.30}$$

where the expectation is over the random variables U_r and ρ_r. As in the proof of Theorem 5, we saw $U_r \to 1$ almost surely as $r \to \infty$. Integrating the p.d.f. $f_r(\rho)$, we have the c.d.f.

$$\Pr(\rho_r < x) = \exp(-(Cx)^r).$$

Therefore,

$$\lim_{r\to\infty} \Pr(\rho_r < x) = \begin{cases} 1, & \text{if } x < 1/C; \\ 0, & \text{if } x > 1/C. \end{cases}$$

Hence, $\rho_r \to 1/C$ in distribution. Therefore, taking the limit of (6.30) with $K = 1$ and C constant, and $N, M \to \infty$,

$$\begin{aligned}
\lim_{N,M\to\infty} \widehat{E}_{\text{MSE}} &= \lim_{N,M\to\infty} \tfrac{K}{N} + F(\tfrac{1}{C}, 1) \\
&= \lim_{N,M\to\infty} \begin{cases} \gamma(a(1-1/C)+1/C), & \text{if } (1+a)/C < a; \\ 0, & \text{if } (1+a)/C > a. \end{cases}
\end{aligned}$$

Now, using the limit (6.29) and the definition $\gamma_{\text{crit}} = C - 1$,

$$\begin{aligned}
\lim_{N,M\to\infty} \gamma(a(1-1/C)+1/C) &= (1-1/C) + \gamma/C = (C - 1 + \gamma)/C \\
&= (\gamma_{\text{crit}} + \gamma)/(\gamma_{\text{crit}} + 1) = \widehat{E}_{\text{lim}}(\gamma).
\end{aligned}$$

Also, as in the proof of Theorem 5, in the limit as $N \to \infty$, $(1+a)/C < a$ is equivalent to $\gamma < \gamma_{\text{crit}}$. Therefore,

$$\lim_{N,M\to\infty} \widehat{E}_{\text{MSE}} = \begin{cases} \widehat{E}_{\text{lim}}(\gamma), & \text{if } \gamma < \gamma_{\text{crit}}; \\ 0, & \text{if } \gamma > \gamma_{\text{crit}}. \end{cases}$$

Bibliography

[1] O. Al-Shaykh, E. Miloslavsky, T. Nomura, R. Neff, and A. Zakhor. Video compression using matching pursuits. *IEEE Trans. Circuits Syst. Video Technol.*, 9(1):123–143, February 1999.

[2] G. E. Andrews, R. Askey, and R. Roy. *Special Functions*, volume 71 of *Encyclopedia of Mathematics and Its Applications*. Cambridge University Press, Cambridge, 1999.

[3] F. Bergeaud and S. Mallat. Matching pursuit of images. In *Proc. IEEE Int. Conf. Image Proc.*, volume I, pages 53–56, Washington, DC, October 1995.

[4] T. Berger. *Rate Distortion Theory*. Prentice-Hall, Englewood Cliffs, NJ, 1971.

[5] E. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Th.*, June 2004. Submitted.

[6] E. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and innacurate measurements. Comput. & Appl. Math. Report 05-12, UCLA, March 2005.

[7] E. Candès and T. Tao. Near optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. Inform. Th.*, October 2004. Submitted.

[8] S. G. Chang, B. Yu, and M. Vetterli. Adaptive wavelet thresholding for image denoising and compression. *IEEE Trans. Image Proc.*, 9(9):1532–1546, September 2000.

[9] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Rev.*, 43(1):129–159, 2001.

[10] A. Cohen and J.-P. D'Ales. Nonlinear approximation of random functions. *SIAM J. Appl. Math.*, 57(2):518–540, April 1997.

[11] J. H. Conway, R. H. Hardin, and N. J. A. Sloane. Packing lines, planes, etc.: Packings in Grassmannian spaces. *Experimental Mathematics*, 5(2):139–159, 1996. See also [12].

[12] Editors' note on packing lines, planes, etc.: Packings in Grassmannian spaces. *Experimental Mathematics*, 6(2):175, 1997.

[13] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, New York, 1991.

[14] I. Daubechies. *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 1992.

[15] G. Davis. *Adaptive Nonlinear Approximations*. PhD thesis, New York Univ., September 1994.

[16] J. W. Demmel. *Applied Numerical Linear Algebra*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 1997.

[17] R. A. DeVore. Nonlinear approximation. *Acta Numerica*, pages 51–150, 1998.

[18] D. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization. *Proc. Nat. Acad. Sci.*, 100(5):2197–2202, March 2003.

[19] D. L. Donoho and M. Elad. On the stability of basis pursuit in the presence of noise. *EURASIP J. Appl. Sig. Proc.*, October 2004. Submitted.

[20] D. L. Donoho, M. Elad, and V. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inform. Th.*, February 2004. Submitted.

[21] D. L. Donoho, M. Vetterli, R. A. DeVore, and I. Daubechies. Data compression and harmonic analysis. *IEEE Trans. Inform. Th.*, 44(6):2435–2476, October 1998.

[22] R. J. Duffin and A. C. Schaeffer. A class of nonharmonic Fourier series. *Trans. Amer. Math. Soc.*, 72:341–366, 1952.

[23] M. Elad and A. M. Bruckstein. A generalized uncertainty principle and sparse representation in pairs of bases. *IEEE Trans. Inform. Th.*, 48(9):2558–2567, September 2002.

[24] M. Elad and M. Zibulevsky. A probabilistic study of the average performance of basis pursuit. *IEEE Trans. Inform. Th.*, December 2004. Submitted.

[25] K. Engan, S. O. Aase, and J. H. Husøy. Designing frames for matching pursuit algorithms. In *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, volume 3, pages 1817–1820, Seattle, WA, May 1998.

[26] A. K. Fletcher and K. Ramchandran. Estimation error bounds for frame denoising. In *Proc. Wavelets: Appl. in Sig. & Image Proc. X, part of SPIE Int. Symp. on Optical Sci. & Tech.*, volume 5207, pages 40–46, San Diego, CA, August 2003.

[27] J.-J. Fuchs. On sparse representations in arbitrary redundant bases. *IEEE Trans. Inform. Th.*, 50(6):1341–1344, June 2004.

[28] R. G. Gallager. *Information Theory and Reliable Communication*. John Wiley & Sons, New York, 1968.

[29] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins Univ. Press, Baltimore, MD, second edition, 1989.

[30] M. M. Goodwin. *Adaptive Signal Models: Theory, Algorithms and Audio Applications*. Kluwer Acad. Pub., 1998.

[31] I. F. Gorodnitsky and B. D. Rao. Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm. *IEEE Trans. Signal Proc.*, 45(3):600–616, March 1997.

[32] V. K Goyal, M. Vetterli, and N. T. Thao. Quantized overcomplete expansions in $\mathbb{R}^N$: Analysis, synthesis, and algorithms. *IEEE Trans. Inform. Th.*, 44(1):16–31, January 1998.

[33] R. M. Gray. *Source Coding Theory*. Kluwer Acad. Pub., Boston, MA, 1990.

[34] R. Gribonval and M. Nielsen. Highly sparse representations from dictionaries are unique and independent of the sparseness measure. Technical Report R-2003-16, Dept. Mathematical Sciences, Aalborg University, October 2003.

[35] R. Gribonval and M. Nielsen. Sparse representations in unions of bases. *IEEE Trans. Inform. Th.*, 49(12):3320–3325, December 2003.

[36] R. Gribonval and M. Nielsen. Beyond sparsity: Recovering structured representations by ℓ_1 minimization and greedy algorithms—Application to the analysis of sparse underdetermined ICA. Technical Report 1684, IRISA, Rennes, France, January 2005.

[37] R. Gribonval and P. Vandergheynst. On the exponential convergence of matching pursuits in quasi-incoherent dictionaries. Technical Report 1619, IRISA, Rennes, France, April 2004.

[38] G. R. Grimmett and D. R. Stirzaker. *Probability and Random Processes*. Oxford Univ. Press, second edition, 1992.

[39] H. Krim, D. Tucker, S. Mallat, and D. Donoho. On denoising and best signal representation. *IEEE Trans. Inform. Th.*, 45(7):2225–2238, November 1999.

[40] J. Liu and P. Moulin. Complexity-regularized image denoising. *IEEE Trans. Image Proc.*, 10(6):841–851, June 2001.

[41] D. Malioutov, M. Çetin, and A. S. Willsky. Sparse signal reconstruction perspective for source localization with sensor arrays. *IEEE Trans. Signal Proc.*, 53(8):3010–3022, August 2005.

[42] S. G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Proc.*, 41(12):3397–3415, December 1993.

[43] F. Moschetti, L. Granai, P. Vandergheynst, and P. Frossard. New dictionary and fast atom searching method for matching pursuit representation of displaced frame difference. In *Proc. IEEE Int. Conf. Image Proc.*, volume 3, pages 685–688, Rochester, NY, September 2002.

[44] B. K. Natarajan. Filtering random noise from deterministic signals via data compression. *IEEE Trans. Signal Proc.*, 43(11):2595–2605, November 1995.

[45] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Computing*, 24(2):227–234, April 1995.

[46] R. Neff and A. Zakhor. Very low bit-rate video coding based on matching pursuits. *IEEE Trans. Circuits Syst. Video Technol.*, 7(1):158–171, February 1997.

[47] N. Saito. Simultaneous noise suppression and signal compression using a library of orthonormal bases and the minimum description length criterion. In E. Foufoula-Georgiou and P. Kumar, editors, *Wavelets in Geophysics*, pages 299–324. Academic Press, San Diego, CA, 1994.

[48] N. J. A. Sloane, R. H. Hardin, and W. D. Smith. A library of putatively optimal spherical codes, together with other arrangements which may not be optimal but are especially interesting for some reason. URL: `http://www.research.att.com/~njas/packings`.

[49] T. Strohmer and R. W. Heath Jr. Grassmannian frames with applications to coding and communication. *Appl. Comput. Harm. Anal.*, 14(3):257–275, May 2003.

[50] J. A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Inform. Th.*, 50(10):2231–2242, October 2004.

[51] J. A. Tropp. Just relax: Convex programming methods for subset selection and sparse approximation. ICES Report 0404, Univ. of Texas at Austin, February 2004.

[52] J. A. Tropp, I. S. Dhillon, R. W. Heath Jr., and T. Strohmer. Designing structured tight frames via an alternating projection method. *IEEE Trans. Inform. Th.*, 51(1):188–209, January 2005.