Sparse Signal Sampling using Noisy Linear Projections

Galen Reeves

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2008-3
http://www.eecs.berkeley.edu/Pubs/TechRpts/2008/EECS-2008-3.html

January 7, 2008

Sparse Signal Sampling using Noisy Linear Projections

Galen Reeves

December 19, 2007

Contents

1 Introduction **9**
- 1.1 Our Contributions . . . 9
- 1.2 General Sampling Model . . . 11
- 1.3 A Sensor network application . . . 12
- 1.4 Overview of Related Work . . . 13

2 Problem Setup **15**
- 2.1 Specific Sampling Model . . . 15
- 2.2 Sparse Signal Classes . . . 16
  - 2.2.1 Non-stochastic Bounded Signals . . . 17
  - 2.2.2 Gaussian Signals . . . 18
- 2.3 Recovery Tasks . . . 18
  - 2.3.1 Support Recovery . . . 18
  - 2.3.2 Signal Estimation . . . 20

3 Support Recovery: Necessary Conditions **21**
- 3.1 Results . . . 21
  - 3.1.1 Perfect Recovery . . . 21
  - 3.1.2 Partial Recovery . . . 22
- 3.2 Discussion . . . 23
- 3.3 Proofs . . . 25
  - 3.3.1 Proof of Theorem 3.1 . . . 25
  - 3.3.2 Proof of Theorems 3.2, 3.3, and 3.4 . . . 26

4 Support Recovery: Sufficient Conditions **29**
- 4.1 ML Support Estimation . . . 29
- 4.2 Results . . . 30
- 4.3 Discussion . . . 32
- 4.4 Proofs . . . 35
  - 4.4.1 Proof of Theorem 4.1 . . . 35
  - 4.4.2 Proof of Lemma 4.4 . . . 36
  - 4.4.3 Proof of Lemma 4.3 . . . 39

5 Signal Estimation: Effects of Noise Prior to Sampling **41**
5.1 Observation Error versus Sampling Error 41
5.2 Optimal Linear Estimator 42
5.3 Results 43
5.4 Discussion 43
5.5 Proofs 44

6 Conclusions and Future Work **47**

A Facts about Chi Squared Variables **49**

B The Asymptotic Spectrum of Random Matrices **53**

List of Figures

1.1 The under-sampled setting 11
1.2 The sparse under-sampled setting when the support K corresponds to the first k elements of x (gray represents zero value) 12
1.3 Sparse signal sampling using a sensor network 13

2.1 Possible distortion metrics 19

3.1 Necessary sampling density ρ as a function of the fractional distortion α for various Ω for any signal class. 24
3.2 Necessary sampling density ρ as a function of the fractional distortion α for various Ω for the class of Gaussian signals. 24

4.1 Sufficient (bold) and necessary (light) sampling densities ρ as a function of the fractional distortion α for various Ω for the class of bounded signals. . . 33
4.2 Sufficient (bold) and necessary (light) sampling densities ρ (Log scale) as a function of the fractional distortion α for various Ω for the class of bounded signals. 33
4.3 Sufficient (bold) and necessary (light) sampling densities ρ as a function of the fractional distortion α for various Ω for the class of Gaussian signals. . . 34
4.4 Sufficient (bold) and necessary (light) sampling densities ρ (log scale) as a function of the fractional distortion α for various Ω for the class of Gaussian signals. 34

5.1 Sparse signal sampling using a sensor network with observation and sampling error 42
5.2 The distortion, $\log_{10} D$, as a function of ρ for various Ω and $\beta = 100$ under sampling noise (solid) and observation noise (dashed). 44
5.3 The distortion, $\log_{10} D$, as a function of ρ for various β and $\Omega = 0.4$ under sampling noise (solid) and observation noise (dashed). 44
5.4 The distortion, $\log_{10} D$, as a function of Ω for various linear scalings of ρ and Ω for $\beta = 100$ under sampling noise (solid) and observation noise (dashed). . 45

Acknowledgements

I thank my advisor Michael Gastpar. His guidance, support, patience, and humor have made this thesis possible. I also thank Martin Wainwright for his helpful discussion throughout the development of this thesis and his comments and suggestions as my second reader.

I appreciate everyone at the Wireless Foundations for giving me an environment where research is fun. Anand Sarwate and Bobak Nazer have been exceptionally helpful with all facets of graduate life, thanks guys. Also, I credit Krish Eswaran, Hari Palaiyanur, and the "sparsity crew" of Dapo Omidiran and Sahand Negahban for many fruitful conversations.

I am grateful to my family for their love, support, and inspiration. In particular I thank David Wilkins for his feedback on this thesis and for reminding me how to pace myself.

Finally, I thank Mary Tai Knox for her feedback on this masters and for giving me the energy to do what I love.

The work in this thesis was supported in part by ARO under the MURI "Heterogeneous Sensor Webs for Automated Target Recognition and Tracking in Urban Terrain," No. W911NF-06-1-0076.

Chapter 1

Introduction

In many engineering applications we choose to view the world (an unknown signal) through a set of samples. Often, a relatively small number of samples tells us all we need to know. For example, the classical Whittaker-Nyquist-Kotelnikov-Shannon sampling theorem states that a continuous-time band-limited signal can be perfectly reconstructed from uniformly spaced discrete samples provided that the sampling rate (number of samples per time) is greater than twice the signal bandwidth. This fact is crucial to the analog-to-digital conversion in signal processing and telecommunications.

For a more general notion of what it means to sample a signal, we may consider a variety of interesting applications where the signals of interest are not band-limited. In fact, even more can be said when we consider that sometimes the information we desire is not an unknown signal per se, but rather some function of it. Examples from the past decade include spectrum blind sampling (Bresler et al. [1, 2, 3]), sampling with a finite rate of innovation (Vetterli et al [4]) and compressed sensing (Donoho [5] and Candes & Tao [6], and many others).

In particular, the field of compressed sensing deals with the digital-to-digital sampling of signals that are somehow compressible. For many such signals, the sampling processes simultaneously *senses* (provides a set of samples sufficient to reconstruct an unknown signal) and *compresses* (the number of samples is far less than the dimension of the original signal).

For a given sampling application, natural questions include: 1) how should the samples be taken? 2) how many samples are required? 3) what fidelity is required to represent each sample? and 4) how do we extract the desired information from the samples?

1.1 Our Contributions

In this thesis, we consider the sampling of discrete-time signals that are sparse, i.e. most of the elements are zero. Such signals contain both discrete information (the support, i.e. the indices of the non-zero components) as well as non-discrete information (the values of the non-zero components). In applications such as signal estimation and compression the goal is to recover the original signal under some mean squared error (MSE) distortion criterion. In

applications such as model selection and regression the main concern is to correctly identify the support.

We focus on a particular form of sampling, namely noisy linear projections, and address questions about the number and quality of samples required. The fidelity of each sample is measured by the per-sample signal-to-noise ratio (SNR), and we assume that the size of the support scales linearly, as opposed to sub-linearly, with the dimension of the unknown signal. We make the following contributions for the under-sampled large system setting where the number of samples is less than the signal dimension, and the signal dimension becomes very large:

- **Perfect support recovery is hard:** If the per-sample SNR does not increase with the dimension of the signal, then exact recovery of the support is not possible. Previous work has shown this to be the case for particular efficient sub-optimal reconstruction algorithms. In Theorem 3.1 of Chapter 3, we show that it is also true for any possible reconstruction algorithm.

- **Fractional support recovery is not as hard:** We introduce a notion of partial support recovery and show that even if the per-sample SNR does not increase with the dimension of the signal, it is still possible to guarantee recovery of some *fraction* of the support. Chapter 2 describes our fractional distortion metric. Theorems 3.2 and 3.4 in Chapter 3 give necessary conditions on the sampling rate and SNR required to recover a given fraction of the support. Equivalently, these results may be seen as upper bound on the fraction that can be recovered using any possible estimator. Theorem 4.1 in Chapter 4 gives a complementary set of sufficient conditions for an ideal estimator that uses exhaustive search.

- **Stochastic versus worst-case analysis:** Previous work on perfect support recovery has used a worst-case analysis that requires a (lower) bound on the smallest non-zero signal component. In this thesis, we consider both stochastic and non-stochastic signal models. An advantage of considering signals with a distribution is that we can give performance guarantees even when the smallest non-zero signal component is arbitrarily small. The nature of these guarantees is made precise in Chapter 2 where we define a notion of the SNR and reliable recovery for both stochastic and non-stocastic signal classes. The results on partial support recovery in Chapters 3 and 4 depend only on certain properties of the assumed signal class.

- **Not all noise is the same:** It is tempting to argue that noise added to the signal prior to sampling can be "pushed through" the sampling process and equivalently considered as noise added after the sampling process. We show that such a consideration is inappropriate. Chapter 5 address the task of signal estimation under two different noise models and Theorem 5.1 gives the corresponding asymptotic MSE distortions as a function of the sampling rate and SNR.

We remark that our results represent fundamental (information-theoretic) limits. For the task of support recovery, our necessary bounds hold for any possible estimator. At the

same time, our achievable results correspond to an ideal estimator that performs exhaustive search. Such an estimator is computationally expensive and thus a interesting extension of this thesis is to find corresponding achievable results for an efficient estimator.

Also, we emphasize that Chapters 3 and 4 are concerned with sampling bonds for support recovery whereas Chapter 5 compares different noise models on the task of signal estimation. Since the tasks of support recovery and signal estimation are related, a natural question not addressed in this thesis is what effect noise added prior to sampling has on the task of support recovery.

The remainder of this introduction is as follows: Section 1.2 describes our general sampling model, Section 1.4 gives a brief overview of previous work in this area, and Section 1.3 provides an example application in sensor networks.

1.2 General Sampling Model

Let $x \in \mathbb{R}^n$ be an unknown sparse signal. The observation model can be generally formulated as a sampling problem where each "sample" consists of an inner product between x and some predetermined measurement vector $\phi_i \in \mathbb{R}^n$ and some random noise w_i

$$y_i = \langle \phi_i, x \rangle + w_i \quad \text{for} \quad i = 1, \cdots, m. \tag{1.1}$$

We will collect our observations into a vector $y = (y_1, y_2, \cdots, y_m)^T$ and let $w = (w_1, w_2, \cdots w_m)^T$. Then, the sampling is given in matrix form as

$$y = \Phi x + w, \tag{1.2}$$

where $\Phi = [\phi_1, \phi_2, \cdots, \phi_n]^T$ is an $m \times n$ sampling matrix. In the under-sampled setting ($m < n$) the matrix Φ is not invertible, and thus general inference problems are challenging.

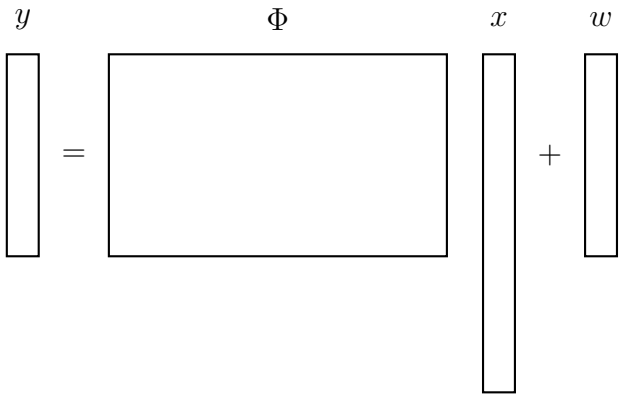

Figure 1.1: The under-sampled setting

The locations of the non-zero elements of x will be referred to as the support of x:

$$K = \{i \in \{1, \cdots n\} \quad \text{s.t.} \quad x_i \neq 0\}, \tag{1.3}$$

where $k = |K|$ is the number of non-zero elements. The signal x is sparse if k is significantly less than n. We use Φ_K to denote the matrix formed by columns of Φ indexed by K and x_K to denote the vector formed by the elements of x indexed by K. Also, $K^\perp$ denotes the complement of the support and thus $|K^\perp| = n - k$. If $k < m$, then the submatrix Φ_K is invertible. This is significant because it means that if we know the support K then our observation model corresponds to an over-constrained set of linear equations. In general, however, the support is considered to be unknown.

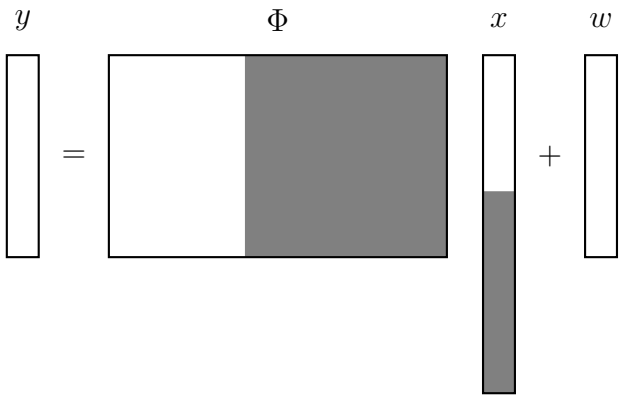

Figure 1.2: The sparse under-sampled setting when the support K corresponds to the first k elements of x (gray represents zero value)

1.3 A Sensor network application

To motivate the problem of sparse support recovery with finite (non-increasing) SNR we provide the following sensor network example (variations can be found in [7, 8, 9]).

Imagine that there is some sparse phenomenon in our environment which may be modeled as a sparse vector $x \in \mathbb{R}^n$ whose indices correspond to specific locations. We desire to locate the phenomenon (i.e. identify the support of x) using a spatially distributed network of n sensors. If the placement of the sensors corresponds to the locations indexed by x, then the vector of sensor observations $\tilde{x} \in \mathbb{R}^n$ will also be sparse. More generally, however, we may assume that there is some known one-to-one linear transformation between x and the sensor observations (represented by $\Psi \in \mathbb{R}^{n \times n}$), and so the observations $\tilde{x} = \Psi x$ are in general non-sparse.

To determine the support of x, it is clearly sufficient to collect the observations from all n sensors. However, since power and bandwidth are likely to be scarce resources, it may be

a much better idea to use one of the procedures outlined in [7, 8, 9] to efficiently compute m linear projections of the data $\tilde{x}$. Then, the data we receive from the network is of the form $y = \Phi\Psi x + w$ where the noise w results from observation noise at each sensor, and computation and communication across the network.

Under assumptions about x, we may pose recovery in terms of the SNR. As the problem size n increases it is important that the size (in bits) of each sample remains constant. Otherwise, the communication will ultimately overwhelm the network even if the ratio m/n stays fixed. Our results are significant because they show that such a network can guarantee recovery of a fixed fraction of the support with $m \ll n$ "network samples".

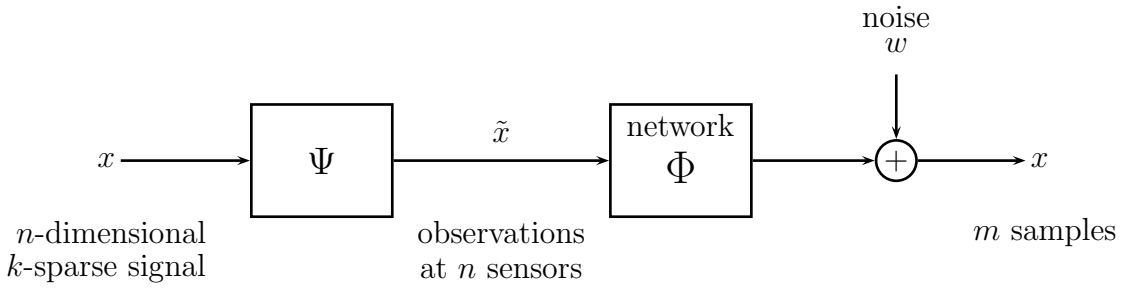

Figure 1.3: Sparse signal sampling using a sensor network

1.4 Overview of Related Work

There is a rich literature on sparse signal reconstruction and support recovery. In this section we present a summary of some of the relevant research with an emphasis on support recovery.

In the noiseless setting ($w = 0$) it has been shown that the support of any signal with exactly k non-zero elements can be recovered from just $m = k + 1$ samples [1]. However, such estimation is a lattice-decoding problem. Compressive sensing [10, 5, 6] has shown that for a small increase in the number of noiseless samples, $m = \mathcal{O}(k \log(n/k))$, perfect recovery can be achieved using an efficient algorithm (linear program) called *Basis Pursuit*.

In the presence of noise, perfect support recovery becomes more difficult. Compressive sensing results [11, 12] show that for $m = \mathcal{O}(k \log(n/k))$ samples there exist efficient algorithms (quadratic programs) which can provide an estimate $\hat{x}$ that is stable, that is $||\hat{x} - x||$ is bounded with respect to $||w||$. Clearly such estimation procedures must supply some information about the support but how much? A partial answer is provided by [13, 14, 15, 11] where it is shown that it is possible to obtain an estimate $\hat{x}$ whose support is contained inside the support of x. However, no guarantees are given on the size of estimated support.

Taking a slightly different approach, the work of [16, 17, 18] has addressed the ability of a particular ℓ_1 constrained quadratic program, the *Lasso*, to guarantee perfect support recovery in the large system setting ($n \to \infty$). Results are formulated in terms of scaling conditions for (n, k, m) and the magnitude of the smallest non-zero component of x denoted

$x_{\min}$. Most relevant to this thesis, is the work of Wainwright [18] which shows that in the linear sparsity regime, if ϕ_i are i.i.d. $\mathcal{N}(0, \frac{1}{n}I)$ then both the sufficient and necessary conditions for perfect recovery (using the Lasso) require that either $m/n \to \infty$ or $x_{\min} \to \infty$.

Another line of research has considered information-theoretic bounds on the performance of the optimal support estimator, also in the large system setting $n \to \infty$. Gastpar and Bresler [3] developed a lower bound on the probability of perfect support recovery in the case where $\{\phi_i\}$ correspond to m rows of the $n \times n$ Fourier matrix and the non-zero elements of x are i.i.d Gaussian. For Gaussian Φ, Sarvotham et al. [19] have given an arbitrary rate-distortion lower bound, and Fletcher et al. [20, 21] have studied the rate-distortion behavior of the signal x. Also, recent work by Aeron et al. [22] has studied the information theoretic properties of sensor networks for sparse signal recovery.

Finally, a major source of inspiration for this thesis comes from Wainwright [23] who provided both lower and upper bounds on the number of samples in terms of the scaling of (n, k, m) and $x_{\min}$. In the linear sparsity setting, the conditions of the sufficient bound are less stringent than in [18] but still require that either $m/n \to \infty$ or that $x_{\min} \to \infty$. As Wainwright points out, there is a significant gap between this achievable bound and the corresponding necessary bound which is finite, namely $m/n < 1$, for a fixed value of $x_{\min}$.

Chapter 2

Problem Setup

This chapter defines three components of our sampling setup. Section 2.1 describes how the samples will be taken. Section 2.2 address what is known, or assumed to be known about the signal of interest prior to sampling and defines two particular signal classes. Section 2.3 addresses the information of interest when decoding the received signal, and describes two related but different recovery tasks.

2.1 Specific Sampling Model

For $x \in \mathbb{R}^n$ we consider a linear observation model (introduced in Section 1.2) in which samples $y \in \mathbb{R}^m$ are taken as

$$y = \Phi x + w, \tag{2.1}$$

where $\Phi \in \mathbb{R}^{m \times n}$ is a sampling matrix and $w \in \mathbb{R}^m$ is noise. We assume that the x has exactly k non-zero elements which are indexed by the support K, and that K is distributed uniformly over the $\binom{n}{k}$ possibilities.

We further assume that the sampling matrix $\Phi \in \mathbb{R}^{m \times n}$ is randomly constructed with i.i.d. rows $\phi_i \sim \mathcal{N}(0, \frac{1}{n} I)$. This matrix construction is a common choice in the compressed sensing literature, and we focus on it for two reasons. First, it also allows us to use powerful asymptotic results for chi squared random variables (Appendix A) and certain well studied random matrices (Appendix B) to develop our bounds.

The second reason for our choice of Φ is that its distribution is invariant to rotation. That is, for any $n \times n$ orthonormal matrix Ψ, the matrix $\Phi\Psi$ is equal in distribution to Φ. The significance of this fact is that all our results also apply to the setting where the observed signal is not sparse per se, but is sparse with respect to some known orthonormal basis Ψ. Thus our sampling model extends to the more general setting

$$y = \Phi\Psi x + w. \tag{2.2}$$

This sampling model is crucial to many applications and is discussed in our examples of sensor network applications in Sections 1.3 and 5.1. However, for the rest of this thesis we find it convenient to describe our results in terms of sampling model (2.1).

In our analysis of partial support recovery in Chapters 3 and 4 we assume that $w \sim \mathcal{N}(0, \sigma_w^2 I)$. In our analysis of signal estimation in Chapter 5 we investigate what happens if noise is added prior to sampling. Thus we compare two noise models: one in which $w \sim \mathcal{N}(0, \sigma_w^2 I)$ and one in which $w \sim \mathcal{N}(0, \sigma_w^2 \Phi\Phi^T)$.

Our our goal is to understand to the asymptotic performance of the sampling model. Hence, we assume that the size of the support $k = k_n$ and number of samples $m = m_n$ depend on ambient dimension n and we see what happens as $n \to \infty$. As mentioned in Section 1.4 there are many interesting choices for the scalings of k_n. We exclusively consider the setting of linear sparsity where k_n is a linear function of n. We are interested in which sampling tasks can (and cannot) be solved in the under-sampled setting where $m_n < n$. We provide the following definitions.

Definition 2.1. The sparsity is $\Omega_n = k_n/n$ and the asymptotic sparsity is $\Omega = \lim_{n\to\infty} k_n/n$. This parameter is a measure of the "bandwidth" of a signal, and linear sparsity corresponds to $0 < \Omega < 1$.

Definition 2.2. The sampling density is $\rho_n = m_n/n$ and the asymptotic sampling density is $\rho = \lim_{n\to\infty} m_n/n$. In the under-sampled setting $\rho < 1$.

We find it convenient to consider a sampling matrix that preserves the magnitude of x. Specifically, we choose to scale the variance of each element of rows ϕ_i^T as $1/n$ so that $\mathbb{E}[\langle \phi_i, x\rangle^2] = ||x||^2/n$, and we will consider signals whose average energy $||x||^2/n$ does not depend on n. We caution the reader that these choices are in contrast to some of the related work [18, 23] where Φ is chosen such that $\mathbb{E}[\langle \phi_i, x\rangle^2] = ||x||^2$ and hence Φ amplifies the signal x.

2.2 Sparse Signal Classes

This thesis is concerned with sampling signals that are sparse. However, to make guarantees on what can or cannot be achieved requires additional information about the class of the unknown signal. Let $\mathcal{X}$ denote a class of sparse signals and let $\mathcal{X}_n$ denote the sub-class of signals with length n. One of the most useful measures we need is the signal-to-noise ratio.

Definition 2.3. For a given signal x, the per-sample signal-to-noise ratio (SNR) is

$$\mathsf{SNR}(x) = \frac{\mathbb{E}\left[||\Phi x||^2\right]}{\mathbb{E}\left[||w||^2\right]} = \frac{1}{n\sigma_w^2}||x||^2. \tag{2.3}$$

Although the above definition also refers to the total signal-to-noise ratio, we call it the per-sample signal-to-noise ratio to emphasize that it is a property of the average signal and noise energies and does not depend on the number of samples.

Performance guarantees will require good bounds on $\mathsf{SNR}(x)$, and the types of bounds we can use depend on the assumed signal class $\mathcal{X}$. In general we may want absolute bounds which hold for all $x \in \mathcal{X}_n$. For stochastic signal classes however, it may be advantageous (and

necessary) to consider bounds which hold with some desired probability. Although a variety of (potentially weaker) bounds could be considered, we will use the following definition in our analysis.

Definition 2.4. For a given signal class $\mathcal{X}$, $\mathsf{SNR}(\mathcal{X})$ is an asymptotic lower bound on $\mathsf{SNR}(x)$. If $\mathcal{X}$ is non-stochastic then the bound must satisfy

$$\mathsf{SNR}(\mathcal{X}) \leq \mathsf{SNR}(x) \qquad \text{for all } x \in \mathcal{X}. \tag{2.4}$$

If $\mathcal{X}$ is stochastic then there must exist some constant $c > 0$ such that

$$\mathbb{P}\left\{\mathsf{SNR}(\mathcal{X}_n) \leq \mathsf{SNR}(x)\right\} > 1 - e^{-n\,c} \tag{2.5}$$

Two other useful measures are the following.

Definition 2.5. For a given signal class $\mathcal{X}$, the relative size of the smallest non-zero element of x is characterized by $\beta_L(\mathcal{X})$. If $\mathcal{X}$ is non-stochastic then

$$\beta_L(\mathcal{X}) = \lim_{n\to\infty} \inf_{x\in\mathcal{X}} \inf_{i\in K} x_i^2/\sigma_w^2 \tag{2.6}$$

If $\mathcal{X}$ is stochastic then $\beta_L(\mathcal{X})$ is a probabilistic upper bound and there must exist some constant $c > 0$ such that

$$\mathbb{P}\Big\{\min_{i\in K} x_i^2/\sigma_w^2 \leq \beta_L(\mathcal{X}_n)\Big\} > 1 - e^{-n\,c} \tag{2.7}$$

Definition 2.6. For a given stochastic signal class $\mathcal{X}$ with $\mathbb{E}(x) = 0$ and $\mathbb{E}[x^2] = \sigma_x^2$, the normalized per-sample signal-to-noise ratio is given by $\beta(\mathcal{X}) = \sigma_x^2/\sigma_w^2$ such that $\mathbb{E}[\mathsf{SNR}(x)] = \Omega\beta(\mathcal{X})$.

In general a wide variety of signal classes may be considered. In the following sections we introduce two example classes, one stochastic and one non-stochastic.

2.2.1 Non-stochastic Bounded Signals

Often it is appealing to have models that do not assume a distribution. Such models may arise naturally when we need a worst-case analysis. Also, resulting claims are robust in that they do not depend on the choice or parameters of an assumed distribution. Previous work on support recovery [16, 17, 18] has focused on the following class.

Definition 2.7. $\mathcal{B}_n$ is the set of all $x \in \mathbb{R}^n$ whose non-zero elements are bounded from below in magnitude, that is $|x_i| \geq x_{\min}$ for all $i \in K$ where $x_{\min}$ is a known constant that does not depend n.

For the class $\mathcal{B}$ it is clear that $\beta_L(\mathcal{B}) = x_{\min}^2/\sigma_w^2$ and $\mathsf{SNR}(\mathcal{B}) = \Omega\beta_L$.

2.2.2 Gaussian Signals

In this thesis, we also consider the class of Gaussian signals which is ubiquitous in information theory, signal processing, and communications.

Definition 2.8. Let $\mathcal{G}_n$ be the set of all $x \in \mathbb{R}^n$ whose non-zero elements are Gaussian, that is x_i are i.i.d. $\mathcal{N}(0, \sigma_x^2)$ for all $i \in K$.

Since any "non-zero" element of x can be arbitrarily close to zero, support recovery becomes more difficult than when there is a fixed bound on $x_{\min}$. We note that $\mathsf{SNR}(x)$ is a random variable that obeys concentration of measure. Using standard large deviation bounds for χ^2 variables (Lemma A.1) we see that for any $\epsilon > 0$ we may choose $\mathsf{SNR}(\mathcal{G}_n) = (1 - \epsilon/n)\Omega_n\beta(\mathcal{G}_n)$ where $\beta(\mathcal{G}_n) = \sigma_x^2/\sigma_w^2$. For simplicity we will use the limit $\mathsf{SNR}(\mathcal{G}) = \mathsf{SNR}(\mathcal{G}_\infty) = \Omega\beta(\mathcal{G})$. Also, we may trivially choose $\beta_L(\mathcal{G}) = \beta(\mathcal{G})$ although much tighter bounds are possible.

2.3 Recovery Tasks

Generally stated, the goal of sampling is to recover some function of the unknown signal to within some desired distortion. In the following sections we look at two different types of tasks, signal estimation and support recovery, and propose corresponding distortion measures.

2.3.1 Support Recovery

Given the true support K and any estimate $\hat{K}$ there are several natural measures for the distortion $d(K, \hat{K})$. One may consider recovery of K as a target recognition problem where for each index $i \in \{1, \cdots, n\}$ we want to determine whether or not i is in the support K.

At one extreme, minimization of

$$d'(K, \hat{K}) = \begin{cases} |K| - |\hat{K}| & \hat{K} \subseteq K \\ \infty & \hat{K} \supset K \end{cases}$$

attempts to find the largest subset $\hat{K}$ that is contained in K. The results of [13, 14, 15, 11] can be interpreted in terms of this metric. Roughly speaking, their results guarantee that $d'(\hat{K}, K) \leq |K|$ but cannot say much more because no guarantees are given on the size $\hat{K}$.

At the other extreme, minimization of

$$d''(K, \hat{K}) = \begin{cases} \infty & \hat{K} \subset K \\ |\hat{K}| - |K| & \hat{K} \supseteq K \end{cases}$$

attempts to find the smallest subset $\hat{K}$ that contains the true support, and in general one may formulate a Neyman-Pearson style tradeoff between the two types of errors.

Our focus is on reconstruction at the point where the number of false positives is equal to the number of false negatives. Since we assume that $|K|$ is known a priori, we can impose this condition by requiring that $|\hat{K}| = |K|$. Accordingly, we use the following metric which is proportional (by a factor of two) to the total number of errors

$$d(K, \hat{K}) = \begin{cases} |K| - |K \cap \hat{K}| & |\hat{K}| = |K| \\ \infty & |\hat{K}| \neq |K| \end{cases}$$

For the remainder of this thesis we consider only candidate supports $\hat{K}$ that are the same size as the true support.

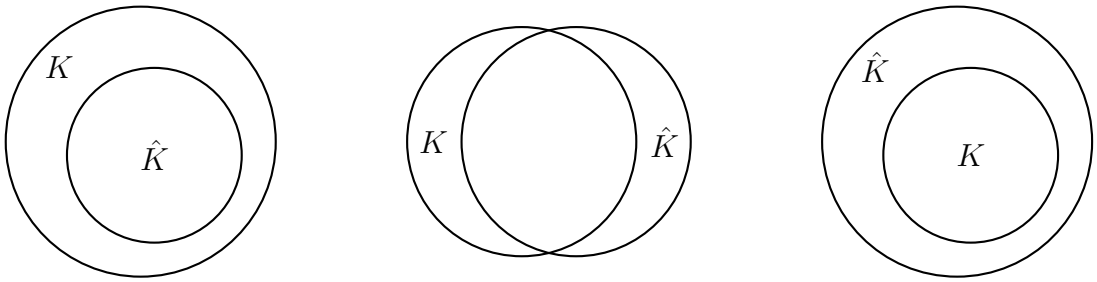

Figure 2.1: Possible distortion metrics

We can define partial recovery as the requirement that $d(K, \hat{K}) \leq a^*$ for some $a^* \geq 0$. If we consider a^* to be a function of n then there are several interesting choices for the scaling of a_n^*. For instance, if recovery is possible with $a_n^* = \mathcal{O}(\log n)$ then as $n \to \infty$ the average distortion $\frac{1}{k_n} d(K, \hat{K}) \to 0$ although the allowable number of errors $d(K, \hat{K}) \to \infty$. Our results, however, pertain to linear scalings between a_n^* and k_n.

Definition 2.9. The fractional distortion is $\alpha \geq 0$, and fractional partial recovery is the requirement that $d(K, \hat{K}) \leq a_n^*$ where $a_n^* = \alpha\, k_n$.

The requirement $\alpha = 0$ corresponds to perfect recovery whereas the requirement $\alpha = 1$ is always satisfied (since we assume $|\hat{K}| = |K|$).

Our analysis will use the following definitions to characterize the performance of an estimator $\hat{K}(y)$ with respect to fractional partial recovery. Recall that y is a function of x and thus the performance depends on $\mathcal{X}$.

Definition 2.10. Let $\hat{K}(y)$ be an estimator of K. If $\mathcal{X}_n$ is non-stochastic, then

$$P_e(\alpha, \mathcal{X}_n) = \max_{x \in \mathcal{X}_n} \mathbb{P}\left\{ d(K, \hat{K}(y)) > \alpha\, k \right\} \tag{2.8}$$

where the probability is over w, and Φ. If $\mathcal{X}_n$ is stochastic, then

$$P_e(\alpha, \mathcal{X}_n) = \mathbb{P}\left\{ d(K, \hat{K}(y)) > \alpha\, k \right\} \tag{2.9}$$

where the probability is over x, w, and Φ.

Definition 2.11. An estimator $\hat{K}(y)$ is said to be *asymptotically reliable* for a class $\mathcal{X}$ if there exists some constant $c > 0$ such that $P_e(\alpha, \mathcal{X}_n) < e^{-n\,c}$.

Definition 2.12. An estimator $\hat{K}(y)$ is said to be *asymptotically unreliable* for a class $\mathcal{X}$ if there exists some constant $c > 0$ and integer N such that $P_e(\alpha, \mathcal{X}_n) > c$ for all $n \geq N$.

We remark that a weaker notation of reliable recovery is to constrain the expected distortion, that is to require that $\mathbb{E}_K[d(K, \hat{K}(y))] \leq \alpha\, k$. Although such a statement means that on average the fractional distortion is less than α, it is still possible that a linear fraction of all possible supports have resulting distortion greater than α. Our notion of asymptotically reliable recovery implies more. It says that although there may be a set of "bad" supports with resulting distortion greater than α, the size of this set very small relative to the total number of possible supports.

For a baseline performance measure, we consider an estimator $\hat{K}_{\mathrm{RG}}$ which randomly guesses a subset of size k independent of the data. It is clear that $\mathbb{E}[d(K, \hat{K}_{\mathrm{RG}})] = (1 - \Omega)k$ which corresponds to a fraction distortion of $\alpha = 1 - \Omega$. Moreover, this value is a sharp threshold as is seen by the following lemma.

Lemma 2.1. *The random guessing estimator $\hat{K}_{RG}$ is asymptotically reliable for any $\alpha > 1 - \Omega$ and asymptotically unreliable for any $\alpha < 1 - \Omega$.*

Proof. This follows from an extension of Hoeffding's Inequality. This particular problem, which corresponds to sampling without replacement, is addressed by Hoeffding [24]. □

The significance of Lemma 2.1 is that no samples are needed to guarantee any fraction distortional $\alpha > 1 - \Omega$. Accordingly, this thesis focuses on support reconstruction for $\alpha \in [0, 1 - \Omega]$.

2.3.2 Signal Estimation

A common distortion measure for signal estimation is the mean squared error (MSE). For simplicity we consider the normalized mean squared error, $||x - \hat{x}||^2/||x||^2$, and define the following metric.

Definition 2.13. Let $\hat{x}(y)$ be an estimator of x. If $\mathcal{X}_n$ is non-stochastic, then

$$D(\mathcal{X}_n) = \max_{x \in \mathcal{X}_n} \frac{\mathbb{E}[||\hat{x}(y) - x||^2]}{||x||^2}, \tag{2.10}$$

where the expectation is over w, and ϕ. If $\mathcal{X}_n$ is stochastic, then

$$D(\mathcal{X}_n) = \frac{\mathbb{E}[||\hat{x}(y) - x||^2]}{\mathbb{E}[||x||^2]}, \tag{2.11}$$

where the probability is over x, w, and ϕ.

Chapter 3

Support Recovery: Necessary Conditions

This chapter deals with fundamental limits on support recovery. For the sampling model described in Chapter 2, we give necessary conditions on the number and quality of samples needed to recover a given fraction of the support. In Section 3.1.1 we show that perfect recovery is not possible unless the SNR increases without bound with dimension n. This means that as n becomes large, either the noise must disappear or the magnitude of each non-zero element of x must increase without bound. In Section 3.1.2 we give an upper bound on fraction of the support that can be recovered for a fixed SNR. Section 3.2 provides discussion, and proofs are given in Section 3.3.

3.1 Results

Our results come in two flavors. The bound on perfect recovery in Section 3.1.1 is a sufficient condition for any estimator to be asymptotically unreliable. The bounds on partial recovery in Section 3.1.2 give necessary conditions any estimator to be asymptotically reliable.

3.1.1 Perfect Recovery

In the paper [23], Wainwright gives sufficient and necessary conditions for perfect support recovery. With respect to our sampling model, the sufficient conditions require $\beta_L(\mathcal{X}) = \infty$ whereas the necessary conditions are satisfied with $\beta_L(\mathcal{X}) < \infty$. In the following theorem we show that $\beta_L(\mathcal{X}) < \infty$ is a sufficient condition for asymptotically unreliable recovery, and thus $\beta_L(\mathcal{X}) = \infty$ is a necessary condition for asymptotically reliable recovery.

Theorem 3.1. *For a given signal class $\mathcal{X}$, sparsity $\Omega \in (0, 1)$, and sampling rate $\rho < 1$, consider the task of perfect support recovery, i.e. the fractional distortion $\alpha = 0$. If $\beta_L(\mathcal{X}) < \infty$ then any estimator $\hat{K}(y)$ is asymptotically unreliable*

We we remark that Theorem 3.1 is very general in that it depends only on the behavior of the smallest non-zero element of x. This means that perfect recovery is not possible unless the per-sample SNR grows without bound with n.

To see why this result makes sense we consider an observation model with $m = n$ and $\tilde{\Phi} = I_n$ such that $y = x + w$. In this case, it is clear that there will always be some positive probability of error. Our result shows that the same is true when we observe $y = \Phi x + w$ for $m \leq n$ and Φ Gaussian.

3.1.2 Partial Recovery

The bounds in this section are information-theortic in nature. The following definitions allow us to characterize the number bits needed to describe the support K to within a desired fractional distortion α.

Definition 3.1. For $p \in [0, 1]$ the function $h(p) = -p\log(p) + (1-p)\log(1-p)$ is the binary entropy function.

Definition 3.2. For $p \in [0, 1]$ and $u \in [0, 1-p]$ we define

$$h(\Omega, \alpha) = \Omega h(\alpha) + (1 - \Omega)h\left(\frac{\alpha}{1/\Omega - 1}\right) \tag{3.1}$$

Let $\mathcal{K}_n(a) = \{U : d(K, U) = a\}$ be the set of supports with distortion equal to a. Using the fact [25] that $\log \binom{n}{k} = n\, h(k/n) + \mathcal{O}(\log n)$ gives

$$\frac{1}{n}\log|\mathcal{K}_n(\alpha\,\Omega\,n)| \quad \rightarrow \quad h(\Omega, \alpha) \quad \text{as} \quad n \rightarrow \infty.$$

Hence, the asymptotic bit rate required to encode the support K to within fractional distortion α is given by $h(\Omega) - h(\Omega, \alpha)$. Note that at $\alpha = 1 - \Omega$ this quantity is zero. This corresponds to Lemma 2.1 which shows that $1 - \Omega$ is a natural upper bound on α.

In the following theorems we essentially plug our rate-distortion function (i.e. the necessary bit rate $h(\Omega) - h(\Omega, \alpha)$) into known bounds. The first result applies to general signal classes.

Theorem 3.2. *For a given signal class $\mathcal{X}$, sparsity $\Omega \in (0, 1)$, sampling rate $\rho < 1$, and fractional distortion $\alpha \in (0, 1-\Omega)$, a necessary condition for asymptotically reliable recovery is*

$$\rho > \frac{h(\Omega) - h(\Omega, \alpha)}{\frac{1}{2}\log\left(1 + \mathsf{SNR}(\mathcal{X})\right)}. \tag{3.2}$$

This bound uses a straightforward application of the data processing inequality and has been previously observed in [19] in terms of some general rate-distortion function. Because the only dependence on the signal class $\mathcal{X}$ is in the SNR, this bound is general and works for both stochastic and non-stochastic classes.

The drawback of Theorem 3.2 is that the bound is very loose. In the paper [3], Gastpar and Bresler analyze a sampling model where the sampling matrix corresponds to m rows of the $n \times n$ Fourier matrix and develop a tighter bound for stochastic signals. This bound can be extended to our sampling matrix and is given below.

Theorem 3.3. *For a given stochastic signal class $\mathcal{X}$, sparsity $\Omega \in (0,1)$, sampling rate $\rho < 1$, and fractional distortion $\alpha \in (0, 1-\Omega)$, a necessary condition for asymptotically reliable recovery is*

$$\rho > \frac{h(\Omega) - h(\Omega,\alpha) + \frac{1}{n} I(x;y|K)}{\frac{1}{2}\log\left(1+\mathsf{SNR}(\mathcal{X})\right)} \tag{3.3}$$

where $I(x;y|K)$ is mutual information between x and y conditioned on K.

Notice that Theorem 3.2 follows immediately from Theorem 3.3 and the non-negativity of mutual information. Hence, the sampling density bound in Theorem 3.3 is greater than or equal to the bound in in Theorem 3.2. However, the conditional mutual information is difficult to compute in general. For the special case of the Gaussian signal class we derive a closed-form expression of Theorem 3.3.

Theorem 3.4. *For the Gaussian signal class $\mathcal{G}$, sparsity $\Omega \in (0,1)$, sampling rate $\rho < 1$, and fractional distortion $\alpha \in (0, 1-\Omega)$, a necessary condition for asymptotically reliable recovery is*

$$\rho > \frac{h(\Omega) - h(\Omega,\alpha) + \rho\, \mathcal{V}_{WS}(\mathsf{SNR}(\mathcal{G}); \rho/\Omega)}{\frac{1}{2}\log\left(1+\mathsf{SNR}(\mathcal{G})\right)} \tag{3.4}$$

where the function $\mathcal{V}_{WS}(\gamma; r)$ is given in Lemma B.3.

3.2 Discussion

The bound in Theorem 3.2 is shown in Figure 3.1 as a function of α for $\mathsf{SNR}(\mathcal{X}) = 100$ and various Ω. The strength of this bound is that it applies generally to all signal classes. However, it is likely to be overly conservative for stochastic signal classes. For the Gaussian signal class, the bound in Theorem 3.4 is shown in Figure 3.2 as a function of α for $\mathsf{SNR}(\mathcal{G}) = 10^5$ and various Ω. Our intuition suggests that support recovery is more difficult for the Gaussian class than for the bounded classes. This difference is supported by the results where we see that the necessary sampling bound in Figure 3.2 is higher than in Figure 3.1 even though the SNR is much larger for the Gaussian bound.

In light of Theorem 3.1, which states that asymptotically reliable recovery cannot be achieved with $\rho < 1$ and $\mathsf{SNR} < \infty$, we see that both Theorems 3.2 and 3.4 are overly conservative at $\alpha = 0$. What is not clear, is how the true performance limit behaves as α becomes very small.

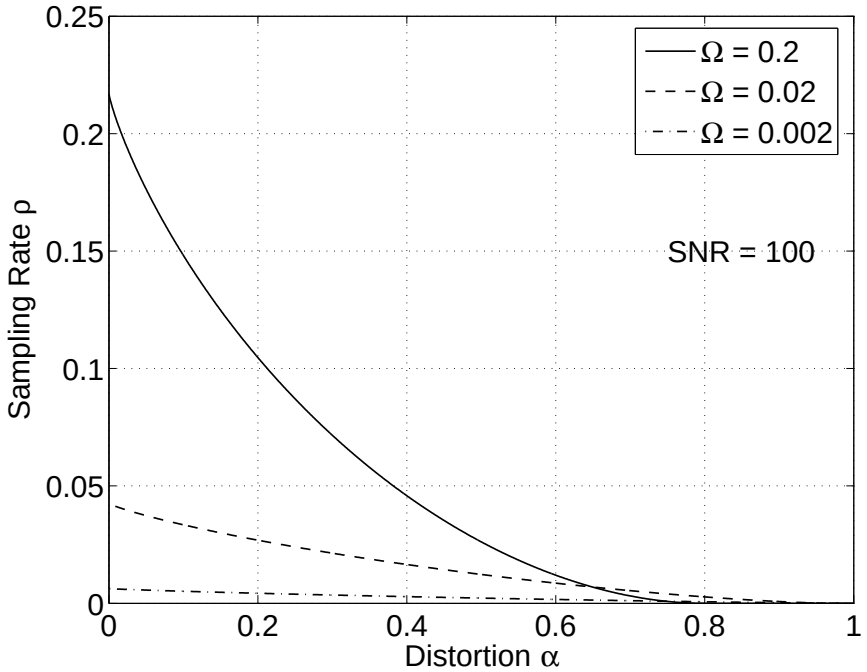

Figure 3.1: Necessary sampling density ρ as a function of the fractional distortion α for various Ω for any signal class.

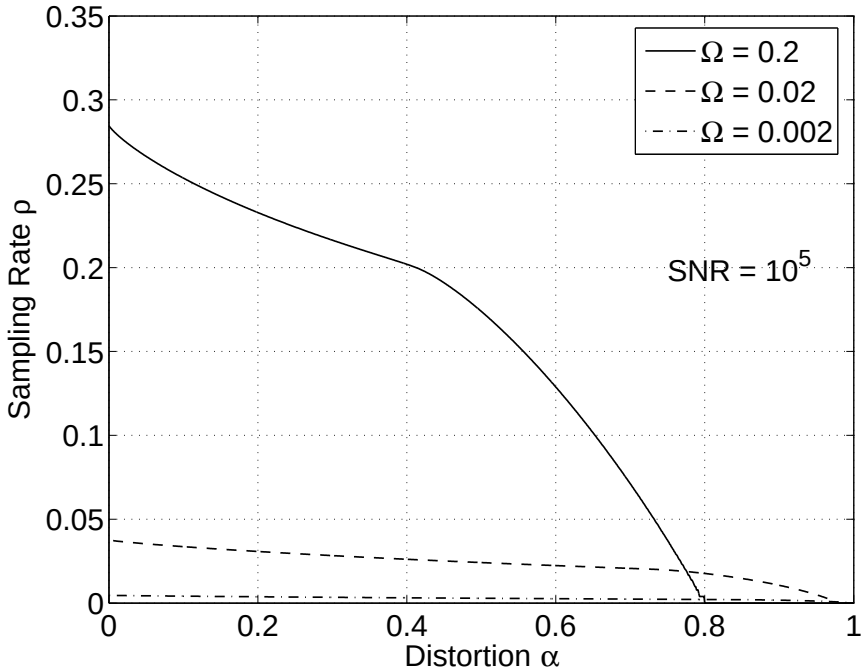

Figure 3.2: Necessary sampling density ρ as a function of the fractional distortion α for various Ω for the class of Gaussian signals.

3.3 Proofs

3.3.1 Proof of Theorem 3.1

In this proof we consider a modified problem in which the estimator has access to additional information about x. In this setting we analyze the optimal estimator $\hat{K}^*$ and show that, for the task of perfect recovery, $\hat{K}^*$ is asymptotically unreliable. Since $\hat{K}^*$ can perform no worse than any estimator in the unmodified problem, this proves our desired result.

For a given signal x, let $i_0 = \arg\min_{i \in K} |x_i|$. Now, imagine that the decoder knows the signal x_K and the set $K_1 = K \backslash i_0$, that is every element of the support except for i_0. Hence, all that remains is to determine which of the remaining $n-k+1$ indices belongs in K. Note that $y - \Phi_{K_0} x_{K_0} \sim \mathcal{N}(x_{i_0} a_i, \sigma_w^2 I)$ and thus the MAP estimate of i_0 is given by

$$\begin{aligned}\hat{i}_0 &= \arg\min_{j \in K_1^\perp} ||y - \Phi_{K_1} x_{K_1} - x_{i_0} a_j||^2 \\ &= \arg\min_{j \in K_1^\perp} ||w + x_{i_0} a_{i_0} - x_{i_0} a_j||^2\end{aligned}$$

For this decoder, an error occurs if there exists $j \in K^\perp$ such that

$$||w + x_{i_0} a_{i_0} - x_{i_0} a_j||^2 < ||w||^2.$$

Let $\beta_L = x_{i_0}^2 / \sigma_w^2$. By multiplying both sides of the above equation by $n/x_{i_0}^2$ we see that the probability of error is equal to the probablity of the following event

$$\min_{1 \le j \le n-k} ||Q + X_0 + X_j||^2 < ||Q||^2 \tag{3.5}$$

where $Q \sim \mathcal{N}(0, \frac{m}{\rho \beta_L} I_m)$ and $X_i \sim \mathcal{N}(0, I_m)$ independently for all i.

To bound this probability we define the following events

$$\begin{aligned}A_1 &= \left\{ \frac{1}{2} m \le ||X_0||^2 \le 2m \right\} \cap \left\{ \frac{1}{2} \frac{m^2}{\beta \rho} \le ||Q||^2 \le 2 \frac{m^2}{\beta \rho} \right\}, \\ A_2 &= \left\{ Q^T X_0 < 0 \right\}.\end{aligned}$$

Using large deviations bound for central χ^2 variables (Lemma A.1) we have $\mathbb{P}\{A_1\} \ge 1 - e^{-c_1 n}$ for some $c_1 > 0$. Further we note that $\mathbb{P}\{A_2 | A_1\} = 1/2$. Hence we can bound the event $A = A_1 \cap A_2$ as

$$\mathbb{P}\{A\} = \mathbb{P}\{A_1\} \mathbb{P}\{A_2 | A_1\} \ge (1/2)(1 - e^{-c_1 n}) \tag{3.6}$$

Note that $||Q + X_0||^2 = ||Q||^2 + ||X_0||^2 + 2 Q^T X_0$ and thus the event A places a bound on the relative difference between $||Q + X_0||^2$ and $||Q||^2$.

Lemma A.3 says that the cumulative distribution function of a non-central χ^2_{NC} variable is decreasing in the non-centrality parameter. Using this fact and conditioning on A gives the following bound for all $j = 1, \cdots, n-k$

$$\begin{aligned}\mathbb{P}\left\{ ||Q + X_0 + X_j||^2 < ||Q||^2 \;\middle|\; A \right\} &\ge \min_{t \in [1/2, 2]} \mathbb{P}\left\{ \chi^2_{NC}\left(m,\ 2m + \frac{m^2}{\beta \rho} t \right) < \frac{m^2}{\beta \rho} t \right\} \\ &= \mathbb{P}\left\{ \chi^2_{NC}\left(m,\ 2m + \gamma_2 m^2 \right) < \gamma_2 m^2 \right\}\end{aligned}$$

where γ_2 correspond to the minimizing value of t.

For a given signal class $\mathcal{X}$ we have $\beta_L \leq \beta_L(\mathcal{X})$ where the bound is tight for non-stochastic classes and holds with exponentially high probability in n for stochastic signal classes. By the assumptions $\beta_L(\mathcal{X}) < \infty$ and $\rho < 1$ we have $\gamma_2 > 0$. Also, since the probability of error only increases as β_L becomes small, it is sufficient to consider $\beta_L > 0$ for all n. In this case, we have $\gamma_2 < \infty$.

Let Z_j be i.i.d. $\chi^2_{NC}\left(m,\, 2m + \gamma_2 m^2\right)$ and let $f_Z(z)$ denote its probability density. We now use Lemma A.4 which gives a lower bound on $f_Z(z)$. Specifically, given some finite constant $\tau > 0$, there exists some constant $c > 0$ and integer $M < \infty$ such that

$$f_Z(z) \geq \frac{c}{m}$$

for all $m \geq M$ and $z \in [\gamma_2 m^2 - \tau\, m, \gamma_2 m^2]$. This means that

$$\begin{aligned}
\mathbb{P}\left\{Z_j \leq \gamma_2 m^2\right\} &= \int_{-\infty}^{\gamma_2 m^2} f_Z(z) dz \\
&\geq \int_{\gamma_2 m^2 - \tau m}^{\gamma_2 m^2} \frac{c}{m} dz \\
&= \tau c > 0
\end{aligned}$$

Hence, the total probability of error is bounded by

$$\begin{aligned}
&\mathbb{P}\left\{\min_{1\leq j\leq n-k} ||Q + X_0 + X_j||^2 < ||Q||^2\right\} \\
&\geq \mathbb{P}\{A\}\mathbb{P}\left\{\min_{1\leq j\leq n-k} ||Q + X_0 + X_j||^2 < ||Q||^2 \;\middle|\; A\right\} \\
&\geq \mathbb{P}\{A\}\mathbb{P}\left\{\min_{1\leq j\leq n-k} Z_j < \gamma_2 m^2\right\} \\
&= \mathbb{P}\{A\}\left[1 - \mathbb{P}\left\{Z_j \geq \gamma_2 m^2\right\}^{n-k}\right] \\
&\geq \mathbb{P}\{A\}\left[1 - \left(1 - \mathbb{P}\left\{Z_j < \gamma_2 m^2\right\}\right)^{n-k}\right] \\
&\geq \mathbb{P}\{A\}\left[1 - (1 - \tau c)^{n-k}\right]
\end{aligned}$$

and we see that the estimator is asymptotically unreliable.

3.3.2 Proof of Theorems 3.2, 3.3, and 3.4

We begin with the proof of Theorem 3.3 which implies Theorem 3.2. This proof is given in the paper [3] for the Gaussian signal class $\mathcal{X}$, the task of perfect recovery, and the case where Φ corresponds to m rows of the $n \times n$ Fourier matrix. We give the proof here for general signal class $\mathcal{X}$, the task of partial recovery, and in terms of our Gaussian sampling matrix Φ.

We define $s = x_K$ and note that the pair (K, s) is equivalent to x. Also, we define $z = \Phi x$ to be the noiseless version of the samples. Finally, we use standard definitions from information theory [25].

The data processing inequality gives

$$I(z; y) \geq I(s, K; y),$$

and the chain rule for mutual information gives

$$I(s, K; y) = I(K; y) + I(s; y|K).$$

Thus we have

$$I(z; y) \geq I(K; y) + I(s; y|K).$$

Since the noise w is i.i.d. Gaussian, and the signal-to-noise ratio between z and $y = z + w$ is given by $\mathsf{SNR}(\mathcal{X})$, we see that an upper bound for the information $I(z; y)$ is given by the channel capacity of an additive white Gaussian noise channel. Thus,

$$I(z; y) \leq m\frac{1}{2}\log(1 + \mathsf{SNR}(\mathcal{X})).$$

Next, we consider how much information is required between K and y. Given that K is uniformly distributed over the $\binom{n}{k}$ possibilities, a simple counting argument and the fact that $\log\binom{n}{k} = n\,h(k/n) + \mathcal{O}(\log n)$ shows that the asymptotic number of bits we need to decode K to within accuracy α is given by $n\,h(\Omega) - n\,h(\Omega, \alpha)$. Using Fano's inequality, we see that $P_e(\alpha) > 0$ unless

$$I(K; y) \geq n\,h(\Omega) - n\,h(\Omega, \alpha)$$

Putting everything together gives the necessary condition of Theorem 3.3

$$m \geq \frac{n\,h(\Omega) - n\,h(\Omega, \alpha) + I(s; y|K)}{\frac{1}{2}\log(1 + \mathsf{SNR}(\mathcal{X}))}$$

Now, the simplified bound in Theorem 3.2 follows from the fact that $I(s; y|K) \geq 0$. For stochastic signal classes, however it may be possible to give a positive lower bound to $I(s; y|K)$. For the Gaussian class $\mathcal{G}$, Gastpar [3] showed that

$$I(s; y|K = U) = \frac{1}{2}\log\det\left(I_k + \mathsf{SNR}(\mathcal{G})\frac{n}{k}\Phi_U^T\Phi_U\right),$$

and thus

$$I(s; y|K) = \mathbb{E}_K I(s; y|K = U) = \frac{1}{2}\mathbb{E}_K\log\det\left(I_k + \mathsf{SNR}(\mathcal{G})\frac{n}{k}\Phi_K^T\Phi_K\right).$$

For a given sampling matrix, such as random rows of the Fourier matrix, this information may be difficult to compute. However, for our sampling matrix Φ we can derive a closed form solution using facts from Appendix B and the following lemma.

Lemma 3.1. *For a nonnegative definite matrix $M \in \mathbb{B}^{n\times n}$ let $\lambda_1(M), \cdots \lambda_n(M)$ denote the eigenvalues of M and let $F_M^n(x)$ denote the empirical eigenvalue distribution (B.1). For $\gamma \geq 0$ we have*

$$\begin{aligned}\frac{1}{n}\log\det(I+\gamma M) &= \sum_{i=1}^{n}\frac{1}{n}\log\det(I+\gamma\lambda_i(M)) \\ &= \int_0^\infty \log(1+\gamma x)d\,F_M^n(x)\end{aligned}$$

Proof. This follows from the properties of the determinant. □

As is shown in Appendix B, the matrix $\Phi_K^T\Phi_K$ is a Wishart Matrix for all K. By the Marcenko-Pastur law [26], the empirical probability distribution of eigenvalues of this matrix converge to a non-random continuous function as $n \to \infty$. This function is given in Lemma B.1. This also means that $I(s;y|K=U)$ converges to a value that does not depend on U. This quantity is given by the so-called Shannon transform $\mathcal{V}_{\mathrm{WS}}(\gamma;r)$ in Lemma B.3. To conclude we have,

$$\frac{1}{n}I(s;y|K=U) \to \rho\,\mathcal{V}_{\mathrm{WS}}(\mathsf{SNR}(\mathcal{G});\rho/\Omega) \tag{3.7}$$

for all U as $n \to \infty$.

Chapter 4

Support Recovery: Sufficient Conditions

This chapter deals with what can be achieved in support recovery. For the sampling model described in Chapter 2, we give sufficient conditions on the number and quality of samples needed to recover a given fraction of the support. These results are complementary to the necessary conditions given in Chapter 3. The bounds correspond to the performance of a maximum likelihood (ML) estimator that performs an exhaustive search over all possible supports. This estimator is described in Section 4.1. The results, which are stated in Section 4.2 and discussed in Section 4.3, show that ML estimator can guarantee some fraction of the support. If this fraction is not too close to one, then only a modest sampling rate is sufficient. However, in accordance with the results of Chapter 3, no guarantees can be made as the desired fraction becomes close to one. Proofs are given in section Section 4.4

4.1 ML Support Estimation

We consider an ideal decoding algorithm which has knowledge of the sparsity parameter k and, given the samples y, performs an exhaustive search over all candidate supports U with size k to determine the most likely estimate of K. Further, our analysis focuses on an estimator that is independent of the signal class $\mathcal{X}$ and assumes that all k-sparse signals $x \in \mathbb{R}^n$ are equally likely.

Definition 4.1. The ML estimator $\hat{K}_{\mathrm{ML}}(y)$ is given by

$$\hat{K}_{\mathrm{ML}}(y) = \arg\max_{|U|=k} \max_{z \in \mathbb{R}^k} \mathbb{P}\{y | K = U, x_K = z\} \tag{4.1}$$

This is the same estimator studied (for the special case of $\alpha = 0$) in Wainwright [23] for the bounded signal class. The following result provides as an equivalent form of the ML estimate.

Lemma 4.1. *The ML estimator is given by*

$$\hat{K}_{ML}(y) = \arg\min_{|U|=k} \min_{z\in\mathbb{R}^k} ||y - \Phi_U z||^2 \tag{4.2}$$

$$= \arg\min_{|U|=k} ||[I_m - \Phi_U(\Phi_U^T\Phi_U)^{-1}\Phi_U^T]y||^2. \tag{4.3}$$

Proof. The follows from our definition of the ML estimate and is shown in the paper [23]. □

The ML estimator is not optimal because it ignores information about $\mathcal{X}$. However we study it for several reasons. First, the ML decoder can be seen as a universal decoder that is appropriate for cases in which we have incomplete or non-existent knowledge of $\mathcal{X}$. For instance, the implementation of ML decoder does not depend on the value of β or β_L. Second, for many signal classes, the difference between the optimal and ML estimates decreases as the SNR becomes large. Finally, as is shown in Lemma 4.1, the ML decoder can be written in terms of orthogonal projections of y which is a formulation that simplifies our analysis.

We remark that ML decoding is computationally hard for any problem of non-trivial size. However, the resulting achievable bound is interesting because it allows us to see where there is potential for improvement in current sub-optimal recovery algorithms. Furthermore, if one is able to lower bound the performance of some efficient estimator with respect to the optimal decoder, then an achievable result is automatically attained.

For completeness we also provide the maximum a posteriori (MAP) estimator for the Gaussian signal class.

Lemma 4.2. *For the signal class $\mathcal{G}$, the MAP estimator is given by*

$$\hat{K}_{MAP}(y,\mathcal{G}) = \arg\min_{|U|=k} \min_{z\in\mathbb{R}^k} ||y - \Phi_U z||^2 + \tfrac{1}{\beta}||z||^2 \tag{4.4}$$

$$= \arg\min_{|U|=k} ||[I_m - \Phi_U(\Phi_U^T\Phi_U + \tfrac{1}{\beta}I)^{-1}\Phi_U^T]\, y||^2. \tag{4.5}$$

Proof. This follows from the definition of a MAP estimate for Gaussian variables. □

4.2 Results

Intuitively, the components of x_K with the smallest magnitudes are the most likely to be left out of the estimated support. Our achievable bound relies on the total magnitude of all the missed components of x_K. Accordingly we introduce the following term which is a function of x.

Definition 4.2. Let $s \in \mathbb{R}^k$ correspond to the non-zero elements of x. Assume that s is indexed such that $|s_1| \leq |s_2| \leq \cdots \leq |s_k|$. Then, for some $\alpha \leq 1$, the normalized magnitude of the smallest $a = \lceil \alpha\, k \rceil$ elements of s is

$$g(\alpha, x) = \frac{1}{\alpha||x||^2} \sum_{i=1}^{a} s_i^2 \tag{4.6}$$

As we saw before with the SNR, we need a bound on $g(\alpha, x)$ that holds for a given signal class $\mathcal{X}$.

Definition 4.3. For a given signal class $\mathcal{X}$, $g(\alpha, \mathcal{X})$ is an asymptotic lower bound on $g(\alpha, x)$. If $\mathcal{X}$ is non-stochastic then the bound must satisfy

$$g(\alpha, \mathcal{X}) \leq g(\alpha, x) \qquad \text{for all } x \in \mathcal{X}. \tag{4.7}$$

If $\mathcal{X}$ is stochastic then there must exist some constant $c > 0$ such that

$$\mathbb{P}\left\{g(\alpha, \mathcal{X}_n) \leq g(\alpha, x)\right\} > 1 - e^{-n\,c} \tag{4.8}$$

For the bounded class, it is clear that we may choose $g(\alpha, \mathcal{B}) = 1$. For the Gaussian class, a suitable choice is provided by the following lemma which is proved in 4.4.3.

Lemma 4.3. *For the Gaussian signal class $\mathcal{G}$ we may choose*

$$g(\alpha, \mathcal{G}) = -W\left(-e^{-(2/\alpha)h(\alpha)-1}\right) > e^{-(2/\alpha)h(\alpha)-1} > \alpha^2/e^3 \tag{4.9}$$

where the Lambert-W function $W(z)$ is the inverse function of $f(z) = z\,e^z$.

Remarkably, the bounds $\mathsf{SNR}(\mathcal{X})$ and $g(\alpha, \mathcal{X})$ tell us all we need to know about the signal class $\mathcal{X}$ for our achievable bound. We now state our main theorem which gives a set of sufficient conditions for asymptotically reliable recovery using the ML estimator.

Theorem 4.1. *For a given signal class $\mathcal{X}$, sparsity $\Omega \in (0,1)$, sampling rate $\rho < 1$, and fractional distortion $\alpha \in (0, 1-\Omega)$, the estimator $\hat{K}_{ML}(y)$ is asymptotically reliable if $\mathsf{SNR}(\mathcal{X}) > 1/(\alpha\, g(\alpha, \mathcal{X}))$ and*

$$\rho > \Omega + \max_{u \in [\alpha, 1-\Omega]} \frac{2\,h(\Omega, u)}{\log\left(\mathsf{SNR}(\mathcal{X})\,u\,g(u, \mathcal{X})\right) + \left(\mathsf{SNR}(\mathcal{X})\,u\,g(u, \mathcal{X})\right)^{-1} - 1} \tag{4.10}$$

where the function $g(u, \mathcal{X})$ satisfies definition 4.3, and $h(\Omega, u)$ is given by (3.1).

In the following corollaries, we provided a simplified, and necessarily weaker set of sufficient conditions for the bounded and Gaussian signal classes. These corollaries make it easier to see the approximate scaling behavior of the bounds.

Corollary 4.1. *For the bounded signal class $\mathcal{B}$, sparsity $\Omega \in (0,1)$, sampling rate $\rho < 1$, and fractional distortion $\alpha \in (0, 1-\Omega)$, the estimator $\hat{K}_{ML}(y)$ is asymptotically reliable if $\mathsf{SNR}(\mathcal{B}) > e/\alpha$ and*

$$\rho > \Omega + \frac{2\,h(\Omega)}{\log\left(\mathsf{SNR}(\mathcal{B})\,\alpha/e\right)}. \tag{4.11}$$

Corollary 4.2. *For the Gaussian signal class $\mathcal{G}$, sparsity $\Omega \in (0,1)$, sampling rate $\rho < 1$, and fractional distortion $\alpha \in (0, 1-\Omega)$, the estimator $\hat{K}_{ML}(y)$ is asymptotically reliable if* $\mathsf{SNR}(\mathcal{G}) > e^4/\alpha^3$ *and*

$$\rho > \Omega + \frac{2\,h(\Omega)}{\log\left(\mathsf{SNR}(\mathcal{G})\,\alpha^3/e^4\right)}. \tag{4.12}$$

In the following corollary we consider a specific relationship between the density and the sparsity and show how the upper bound on the fractional distortion depends on the SNR.

Corollary 4.3 (Achievable Distortion)**.** *Let the sampling density be given by $\rho = \Omega + 2h(\Omega)$. With exponentially high probability in n, the fractional distortion of the estimator $\hat{K}_{ML}(y)$ obeys*

$$\alpha < e^2/\mathsf{SNR}(\mathcal{B}) \tag{4.13}$$

for the bounded signal class and

$$\alpha < (e^2/\mathsf{SNR}(\mathcal{G}))^{1/3} \tag{4.14}$$

for the Gaussian class.

4.3 Discussion

In this section, we discuss the implications of Theorem 4.1 for the bounded and Gaussian signal classes. Figures 4.1 and 4.2 (log scale) show the sufficient sampling bound as a function of α for various Ω for the bounded signal class with $\mathsf{SNR} = 100$. Figures 4.3 and 4.4 (log scale) show the sufficient sampling bound as a function of α for various Ω for the Gaussian signal class with $\mathsf{SNR} = 10^5$. Also shown, are the corresponding lower bounds from Chapter 3.

For both signal classes, we see that recovery in the under-sampled setting with fixed SNR is possible over a range of α. However, as α becomes small, the sampling rate increases without bound. This confirms our results from Chapter 3 which indicate that perfect recovery in the presence of noise is very challenging. At the same time, it shows that if we accept a small fraction of errors, reasonable results can be attained. Also, from Figures 4.2 and 4.4 we see that the upper and lower bond are reasonably tight for values of α that are not near 0 or $1-\Omega$.

We may also consider non-asymptotic results. By paying attention to the exact error exponents we can give a bound on the probability of error for fixed n. In Table 4.1 we show the sufficient conditions to recover 99.9% of the support when $n = 10,000$ and x_K is bounded from below. For example, when $k = 200$ and $m = 522$ the probability that the support is not perfectly recovered is less than 0.0001. In Table 4.2, we show sufficient conditions to recover 90% of the support when $n = 10,000$ and x_K is Gaussian. In this setting we see that a high SNR is required, but that reliable recovery can be performed with very few samples.

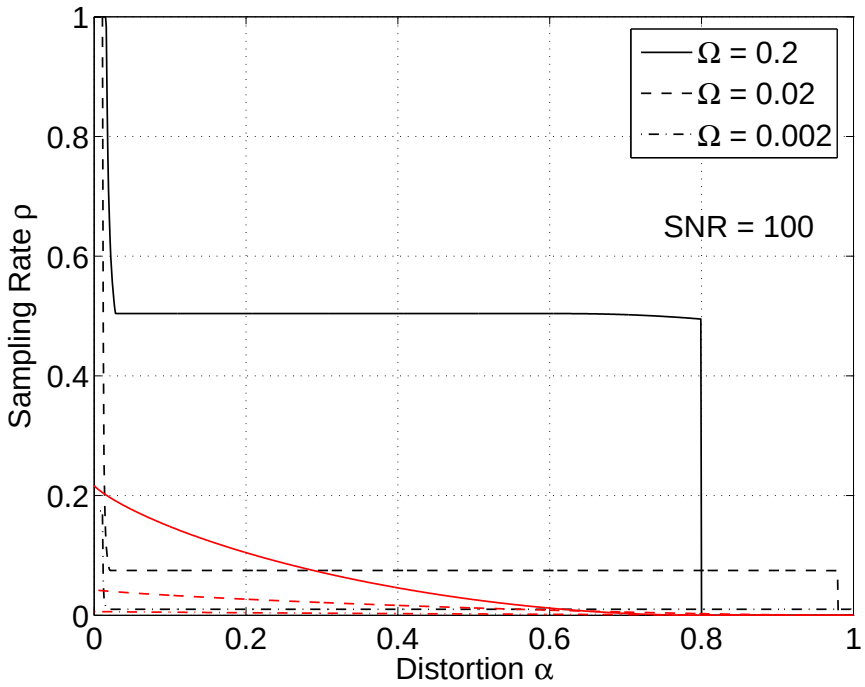

Figure 4.1: Sufficient (bold) and necessary (light) sampling densities ρ as a function of the fractional distortion α for various Ω for the class of bounded signals.

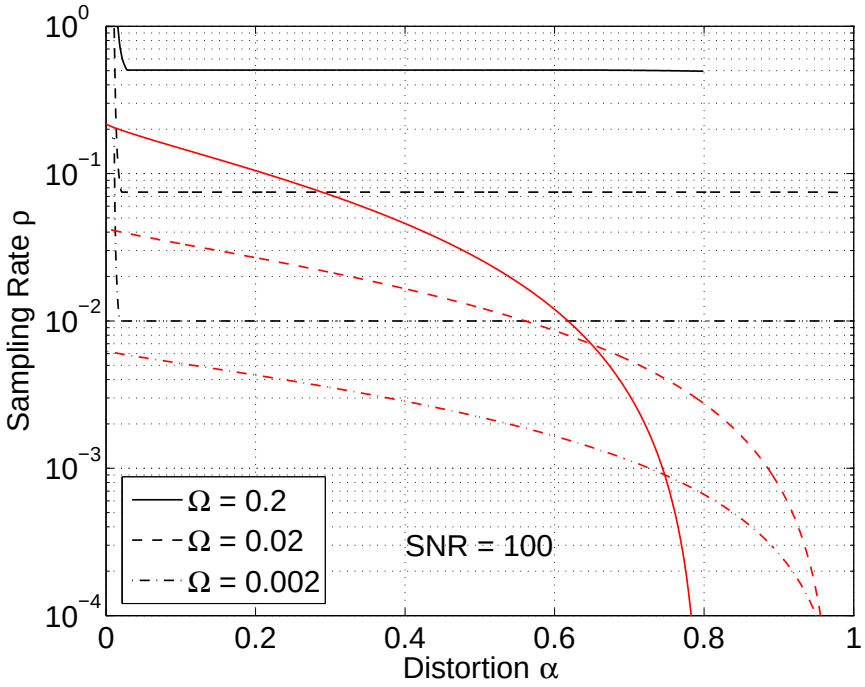

Figure 4.2: Sufficient (bold) and necessary (light) sampling densities ρ (Log scale) as a function of the fractional distortion α for various Ω for the class of bounded signals.

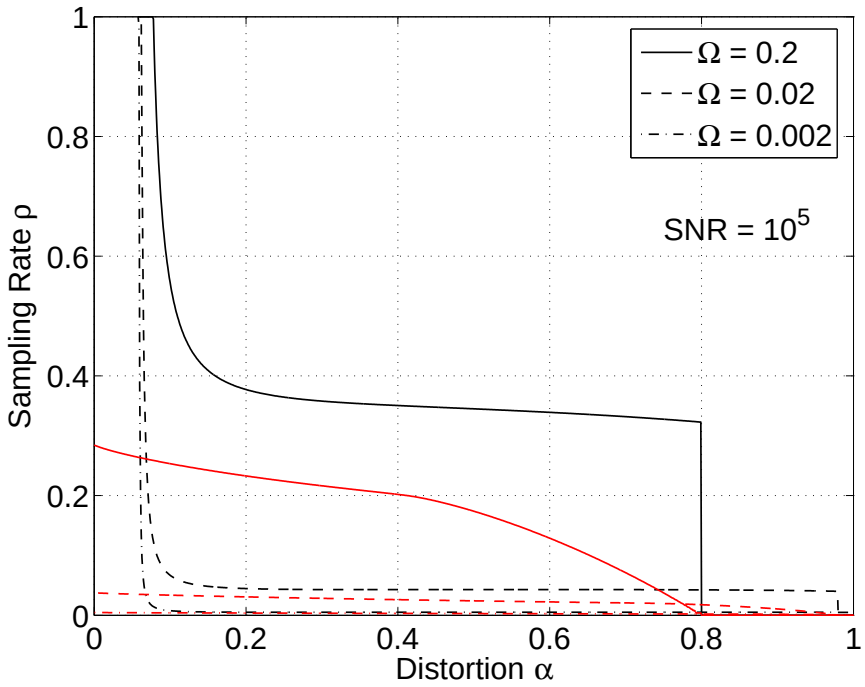

Figure 4.3: Sufficient (bold) and necessary (light) sampling densities ρ as a function of the fractional distortion α for various Ω for the class of Gaussian signals.

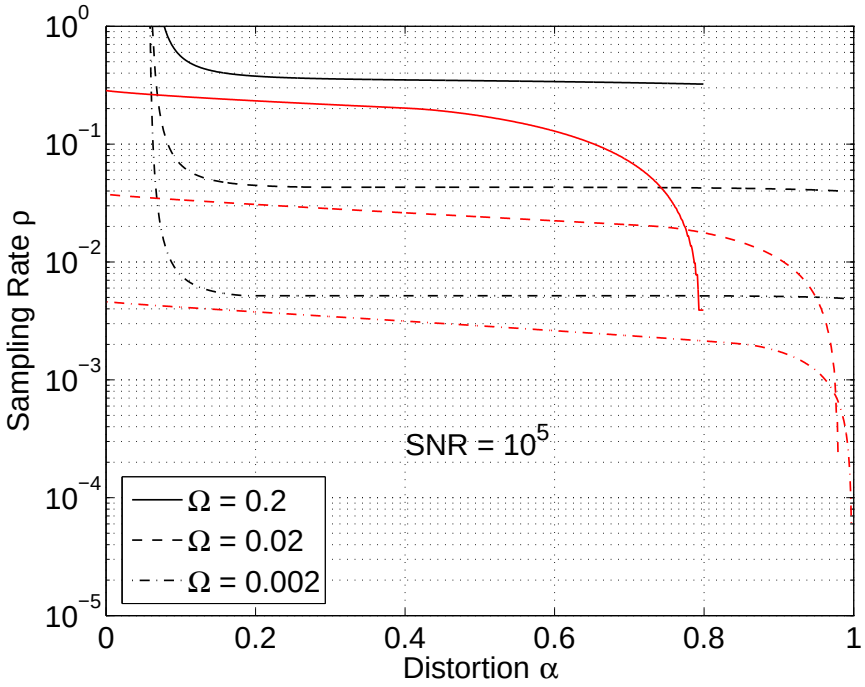

Figure 4.4: Sufficient (bold) and necessary (light) sampling densities ρ (log scale) as a function of the fractional distortion α for various Ω for the class of Gaussian signals.

Table 4.1: Sufficient number of samples to recover 99.9% of the support when n = 10,000 and x is bounded.

k	β_L	$P_e(10^{-3})$	m
200	10^5	1.00%	514
		0.10%	518
		0.01%	522
500	10^5	1.00%	1,050
		0.10%	1,055
		0.01%	1,060

Table 4.2: Sufficient number of samples to recover 90% of the support when n = 10,000 and x is Gaussian

k	β	$P_e(0.1)$	m
200	10^8	1.00%	346
		0.10%	362
		0.01%	381
500	10^7	1.00%	978
		0.10%	1,009
		0.01%	1,042

4.4 Proofs

4.4.1 Proof of Theorem 4.1

The main technical result underlying the proof is the following lemma which relates the desired error probability, $P_e(\alpha)$, to the large deviations behavior of multiple independent chi-squared variables. The proof of this lemma is given in Section 4.4.2.

Lemma 4.4. *For given parameters (n, k, m, α) and signal class $\mathcal{X}_n$, let $g(u, \mathcal{X}_n)$ satisfy definition 4.3 with error exponent $c_0 > 0$ for all $u \in [\alpha, 1 - k/n]$. Then, for any scalar $t > 0$*

$$P_e(\alpha) \leq \mathbb{P}\{\chi^2(m-k) > t\} \tag{4.15}$$
$$+ \sum_{a=\lfloor \alpha k \rfloor}^{\lceil k^2/n \rceil} \left[\binom{k}{a} \binom{n-k}{a} \mathbb{P}\left\{\chi^2(m-k) < \tau(a)\, t\right\} + e^{-n\, c_0} \right]$$

where $\tau(a) = \left[\mathsf{SNR}(\mathcal{X})(a/k) g(a/k, \mathcal{X})\right]^{-1}$.

We proceed by using large deviation bounds and the union bound to lower bound the error exponents of the terms on the right hand side of (4.15). We then determine conditions such that the exponents are positive.

First, for any $\nu > 0$ we may choose $t_\nu = (1+\nu)(m-k)$. By the upper concentration bound in Lemma A.1 we have

$$-\frac{1}{n}\log\left(\mathbb{P}(\chi^2(m-k) > t_\nu)\right) \le E_1$$

where $E_1 = (\rho - \Omega)\nu^2/4$ is positive for all $\rho > \Omega$.

Next, we consider the remaining term in (4.15) evaluated with t_ν for ν arbitrarily close to zero. To use the lower concentration bound in Lemma A.1 requires

$$\mathsf{SNR}(\mathcal{X}) > \max_{u\in[\alpha,1-\Omega]} \frac{1}{u\, g(u,\mathcal{X})} = \frac{1}{\alpha\, g(\alpha,\mathcal{X})}.$$

If this inequality is satisfied, then

$$-\frac{1}{n}\log\left(\mathbb{P}\left\{\chi^2(m-k) < \tau(a)\, t\right\}\right) \le E_2(a)$$

where

$$E_2(a) = (\rho-\Omega)\frac{1}{2}\left[\log\left(\mathsf{SNR}(\mathcal{X})(a/k)g(a/k,\mathcal{X})\right) + \left(\mathsf{SNR}(\mathcal{X})(a/k)g(a/k,\mathcal{X})\right)^{-1} - 1\right]$$

Finally, we note that $\log\binom{k}{a}\binom{n-k}{a} \to n\, h(\Omega, a/k)$ as $n\to\infty$, where $h(\Omega,\alpha)$ is given by (3.1).

Putting everything together gives

$$\begin{aligned} P_e(\alpha) &\le e^{-n\,E_1} + \sum_{a=\lfloor \alpha k\rfloor}^{\lceil k^2/n\rceil} \left[e^{-n\left(E(a)-h(\Omega,a/k)\right)} + e^{-n\,c_0}\right] \\ &< e^{-n\,E_1} + \max_{\alpha k\le a\le (1-\Omega)k} e^{-n\left(E(a)-h(\Omega,a/k)\right)+\log k} + e^{-n\,c_0+\log k}, \end{aligned}$$

and the bound in Theorem 4.1 guarantees that $E(a) - h(\Omega, a/k) > 0$ for $\alpha k \le a \le (1-\Omega)k$.

4.4.2 Proof of Lemma 4.4

For a given set $(n, k, m, \alpha, \mathsf{SNR}(\mathcal{X}_n), g(u,\mathcal{X}_n))$ we derive an upper bound on $P_e(\alpha)$ for the ML decoder. In particular, we analyze $P_e(\alpha|K)$, the error conditioned on the true set K. Because of the symmetry in the sampling procedure, we will see that the conditional probability does not depend on the particular set K. Therefore for any distribution over K we have

$$P_e(\alpha) = \sum_{K:\,|K|=k} \mathbb{P}(K) P_e(\alpha|K) = P_e(\alpha|K).$$

Analysis of ML Decoder

As is shown in [23], the ML decoder in Lemma 4.2 can be written explicitly as a single minimization. For any subset U with $|U| = k < m$ the $m \times k$ matrix Φ_U has full rank with probability one. Thus we may define the following orthogonal projections

$$\Pi_U = \Phi_U[\Phi_U^T\Phi_U]^{-1}\Phi_U^T \tag{4.16}$$

$$\Pi_U^{\perp} = I - \Phi_U[\Phi_U^T\Phi_U]^{-1}\Phi_U^T. \tag{4.17}$$

corresponding to the range (respectively null) space of Φ_U. The ML decoder can now be given as

$$\hat{K}_{ML} = \arg\min_{|U|=k} \text{err}(U). \tag{4.18}$$

where $\text{err}(U) = \frac{1}{\sigma_w^2}||\Pi_U^{\perp} y||^2$.

Analysis of Error

Let $a^* = \lfloor \alpha k \rfloor$ and consider the sets

$$G = \{U \ : \ |U| = k, |U \cap K| > k - a^*\}$$
$$B = \{U \ : \ |U| = k, |U \cap K| \leq k - a^*\}$$

where G represents the "good" set of candidate supports and B represents the "bad" set. An error is declared if $\hat{K} \in B$ and this occurs if and only if

$$\min_{U \in B} \text{err}(U) < \min_{V \in G} \text{err}(V).$$

To develop a bound on the above event, we split it into to sub-events which can be analyzed independently. For any scalar t we define to the the following two "bad" events

$$A_B = \left\{ \min_{U \in B} \text{err}(U) < t \right\}, \quad A_G = \left\{ \min_{V \in G} \text{err}(V) > t \right\}.$$

Note that $A_B^c \cap A_G^c$ is a sufficient condition for success. Accordingly, the probability of error can be bounded as

$$\begin{aligned} P_e(a^*|K) &\leq 1 - \mathbb{P}\left(A_B^c \cap A_G^c\right) \\ &= \mathbb{P}\left(A_B \cup A_G\right) \\ &\leq \mathbb{P}\left(A_B\right) + \mathbb{P}\left(A_G\right) \end{aligned}$$

Our proof proceeds by relating $\mathbb{P}(A_B)$ and $\mathbb{P}(A_B)$ to the the cumulative distribution functions of various χ^2 variables.

Bounding $\mathbb{P}(A_G)$

To tightly characterize the behavior of $\min_{V \in G} \text{err}(V)$ would require significant effort. However, since it is always true that $K \in G$, it is sufficient to consider the weakened bound $\mathbb{P}(A_G) \leq \mathbb{P}(\text{err}(K) > t)$. Since w has zero mean i.i.d. Gaussian elements, and $\Pi_K^{\perp}$ is an orthonormal projection matrix, the random variable

$$\text{err}(K) = \frac{1}{\sigma_w^2}||\Pi_K^{\perp} y||^2 = \frac{1}{\sigma_w^2}||\Pi_K^{\perp} w||^2.$$

has a χ^2 distribution with $m - k$ degrees of freedom (Lemma A.2).

Bounding $\mathbb{P}(A_B)$

To bound the probability that $\min_{U \in B} \text{err}(U) < t$, it is sufficient to bound $\text{err}(U)$ for every possible $U \in B$. At a high level, our approach to this task is to repeatedly partition A_B into smaller sub-events whose behavior we can characterize with respect to χ^2 variables. Then, repeated application of the union bound lead a bound on our desired quantity $P(A_B)$.

We begin by partitioning $U \in B$ based on the size of the overlap $k - a = |U \cap K|$. Let

$$\tilde{B}(a) = \{U \ : \ |U| = k, |U \cap K| = k - a\},$$
$$A_{\tilde{B}(a)} = \left\{ \min_{U \in \tilde{B}(a)} \text{err}(U) < t \right\}.$$

Then, the union bound gives

$$\mathbb{P}(A_B) \leq \sum_{a=a^*}^{a_{\max}} \mathbb{P}(A_{\tilde{B}(a)}). \tag{4.19}$$

where $a_{\max} = \lceil k - k^2/n \rceil$. Accordingly, the next step is to bound the event $\mathbb{P}(A_{\tilde{B}(a)})$ for all $a^* \leq a \leq a_{max}$.

Bounding $\mathbb{P}(A_{\tilde{B}(a)})$

For any $U \in \tilde{B}(a)$ the distribution of $\text{err}(U)$ can be characterized with respect to two quantities. First, conditioned on the set $K \backslash U$, the magnitude of the "missed" components of x is given by $\mathsf{SNR}(x_{K \backslash U})$. Second, for any set U the magnitude of the projected noise is given by

$$\Lambda(U) = \frac{1}{\sigma_w^2}||\Pi_U^{\perp} w||^2.$$

Since w is a Gaussian vector with i.i.d. elements and $\Pi_K^{\perp}$ is an orthonormal projection matrix, $\Lambda(U)$ is a χ^2 random variable with $m - k$ degrees of freedom (Lemma A.2).

Now, conditioned on $\mathsf{SNR}(x_{K\setminus U}) = \theta$ the random vector $(\sigma_w^2\theta)^{-1/2}\,\Phi_{K\setminus U}x_{K\setminus U}$ has i.i.d. zero mean Gaussian elements with variance one. If we also condition on $\Lambda(U) = \lambda$, then we see that

$$\frac{1}{\theta}\,\mathrm{err}(U) = \frac{1}{\sigma_w^2\theta}||\Pi_U^{\perp}\left(\Phi_{K\setminus U}x_{K\setminus U} + w\right)||^2$$

is a non-central χ^2_{NC} variable with non-centrality parameter λ/θ and $m-k$ degrees of freedom due to the orthogonal projection $\Pi_U^{\perp}$ (Lemma A.2). This means that

$$\mathbb{P}\big\{\mathrm{err}(U) < t\big|\mathsf{SNR}(x_{K\setminus U}) = \theta, \Lambda(U) = \lambda\big\} = \mathbb{P}\big\{\chi^2_{NC}(m-k, \lambda/\theta) < t/\theta\big\} \tag{4.20}$$

To reduce amount of conditioning required in (4.20) we use Lemma A.3 which states the the cumulative distribution function of non-central χ^2 variable is decreasing the non-centrality parameter. Noting that $\Lambda(U) \geq 0$ we have

$$\begin{aligned}\mathbb{P}\big\{\mathrm{err}(U) < t\big|\mathsf{SNR}(x_{K\setminus U}) = \theta\big\} &\leq \mathbb{P}\big\{\chi^2_{NC}(m-k, 0) < t/\theta\big\}\\ &= \mathbb{P}\big\{\chi^2(m-k) < t/\theta\big\},\end{aligned}$$

and so

$$\mathbb{P}\big\{\mathrm{err}(U) < t\big|\mathsf{SNR}(x_{K\setminus U}) \geq \theta\big\} \leq \mathbb{P}\big\{\chi^2(m-k) < t/\theta\big\}.$$

By our assumptions we have

$$\mathbb{P}\big\{\min_{U\in\tilde{B}(a)}\mathsf{SNR}(x_{K\setminus U}) < 1/\tau(a)\big\} \leq e^{-n\,c_0}$$

where $\tau(a) = \big[\mathsf{SNR}(\mathcal{X})(a/k)g(a/k,\mathcal{X})\big]^{-1}$. Thus, the union bound gives

$$\begin{aligned}\mathbb{P}\big\{A_{\tilde{B}(a)}\big\} &\leq e^{-n\,c_0} + \sum_{U\in\tilde{B}(a)}\mathbb{P}\big\{\chi^2(m-k) < \tau(a)t\big\}\\ &= e^{-n\,c_0} + \binom{k}{a}\binom{n-k}{a}\mathbb{P}\big\{\chi^2(m-k) < \tau(a)t\big\}.\end{aligned}$$

4.4.3 Proof of Lemma 4.3

We are given $x_K \sim \mathcal{N}(0, \sigma_x^2 I_k)$ and $a_n = \lceil \alpha k_n \rceil$. If we let $a = \lceil \alpha k_n \rceil$ and $\tilde{C}(a) = \{U \subset K \;:\; |U| = a\}$, then by definition we have

$$g(\alpha, x) = \min_{U\in\tilde{C}(a)}\frac{k\,\sigma_x^2}{||x_K||^2}\frac{||x_U||^2}{a\sigma_x^2}$$

We note that the variable $||x_K||^2/\sigma_x^2$ is χ^2 with k degrees of freedom and, for each set U, the variable $||x_U||^2/\sigma_x^2$ is χ^2 with a degrees of freedom. Although these variable are not

independent, we can use the union bound to bound $g(\alpha, x)$ in terms of independent events. For any $\epsilon_1, \epsilon_2 > 0$ we have

$$\begin{aligned}\mathbb{P}\big\{g(\alpha,x) < (1-\epsilon_1)(1-\epsilon_2)\big\} &\leq \mathbb{P}\{||x_K||^2/\sigma_x^2 < (1-\epsilon_1)k\} \\ &\quad + \sum_{U\in\tilde{C}(a)} \mathbb{P}\{||x_U||^2/\sigma_x^2 \leq (1-\epsilon_2)a\} \\ &= \mathbb{P}\{\chi^2(k) < (1-\epsilon_1)k\} \\ &\quad + |\tilde{C}(a)|\mathbb{P}\{\chi^2(a) \leq (1-\epsilon_2)a\}\end{aligned}$$

Using concentration bounds for central χ^2 variables (Lemma A.1) we may set ϵ_1 arbitrarily close to zero. Furthermore, we note that $|\tilde{C}(a)| = \binom{k}{a}$. As $n \to \infty$ this means that

$$-\frac{1}{n}\log\Big(|\tilde{C}(a)|\mathbb{P}\{\chi^2(a) \leq (1-\epsilon_2)a\}\Big) \leq E_0(\Omega,\epsilon,\rho)$$

where

$$E_0(\Omega,\epsilon,\rho) = \Omega\alpha\frac{1}{2}\left[-\log(1-\epsilon_2)+\epsilon_2\right] - \Omega h(\alpha).$$

Solving for the critical value of ϵ_2 such that $\lim_{n\to\infty} E_0(\Omega,\epsilon,\rho) > 0$ leads to the stated bound.

Chapter 5

Signal Estimation: Effects of Noise Prior to Sampling

While Chapters 3 and 4 were concerned with support recovery, the focus of this chapter is signal recovery. Our goal is to understand how the distortion, measured by the mean squared error (MSE), is affected by where the noise enters into the sampling processes. We compare two noise models: the standard model in which the noise is added to each sample, and a variation in which the noise is added to the signal x prior to sampling. For stochastic signal classes we derive closed-form expressions for the distortion of the linear minimal mean squared error (LMMSE) estimator that has a priori knowledge of the support. Section 5.1 describes the noise models, Section 5.2 describes the LMMSE estimator, Section 5.3 gives our results, Section 5.4 provides discussion, and proofs are given in section 5.5.

5.1 Observation Error versus Sampling Error

In this chapter we make a distinction between noise that is added to each sample (sampling error) and noise that is added to the signal prior to sampling (observation error). Accordingly we generalize the sampling model proposed in Chapter 2, to the following

$$y = \Phi(x + w_{\text{obs}}) + w_{\text{smp}} \tag{5.1}$$

where w_{obs} is observation error and w_{smp} is sampling error and all other assumptions about Φ and K are the same.

To motivate these sources of noise, we consider the sensor network application introduced in Section 1.3. Figure 5.1 below shows the same sensor network with the source of observation noise and sampling noise made explicit. In this context, the observation noise arises from the noise and quantization of the sensor reading itself. The sampling noise arises from the computation and communication over the network, and any subsequent quantization.

To understand the relative tradeoff of the two error types we consider samples taken with

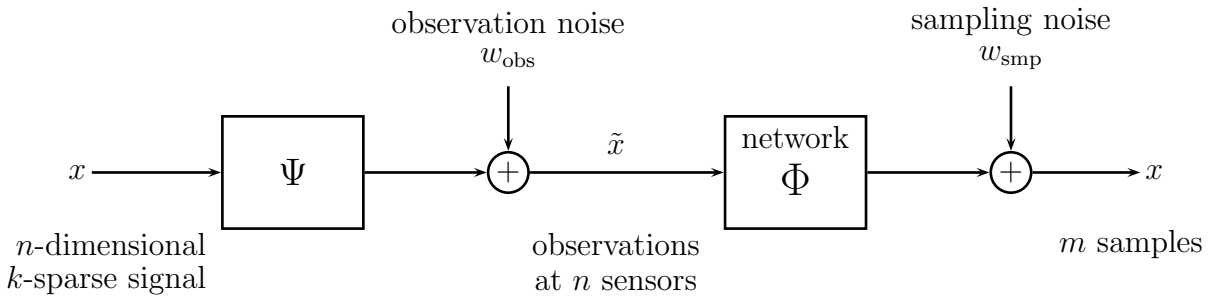

Figure 5.1: Sparse signal sampling using a sensor network with observation and sampling error

only one kind of error, that is

$$y_{\text{obs}} = \Phi(x + w_{\text{obs}}) \tag{5.2}$$

$$y_{\text{smp}} = \Phi x + w_{\text{smp}} \tag{5.3}$$

where both $w_{\text{obs}} \in \mathbb{R}^n$ and $w_{\text{obs}} \in \mathbb{R}^m$ have i.i.d. elements with zero mean and variance σ_w^2. Note that elements of the random vector Φw_{obs} also have variance σ_w^2 but are not independent. We consider any stochastic signal class $\mathcal{X}$ in which the non-zero elements are i.i.d. with zero mean and variance σ_x^2. Thus throughout this chapter $\beta = \beta(\mathcal{X}) = \sigma_x^2/\sigma_w^2$. Finally, we do not constrain our our analysis to the under-sampled setting and consider any $\rho \geq 0$.

5.2 Optimal Linear Estimator

In this section we consider estimation of the signal x when the support K is known. This is interesting when K can be accurately determined or when K is known a priori but there is no control over how the samples are taken. Additionally, the analysis of this case shows the fundamental differences between the types of noise regardless of uncertainty in the support. Finally, we note that for the Gaussian class $\mathcal{G}$ the LMMSE estimator is the optimal MSE estimator. For results concerning a sub-optimal efficient ℓ_1 constrained estimator without knowledge of the true support see [27].

Conditioned on a particular sampling matrix Φ and support K, let Σ_x, Σ_y, and Σ_{xy} denote the covariance and cross correlation matrices of x and y. It is well known that the LLMSE linear estimator is given by

$$\hat{x}(y|K) = \Sigma_{xy}\Sigma_y^{-1} y. \tag{5.4}$$

Moreover, the conditional distortion is given by

$$D(\mathcal{X}|K,\Phi) = \frac{1}{k\sigma_x^2}\mathbb{E}[||x - \hat{x}(y|K)||^2] = \frac{1}{k\sigma_x^2}\text{tr}\left\{\Sigma_x - \Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx}\right\} \tag{5.5}$$

Hence, for a given sampling problem the distortion is given by

$$D(\mathcal{X}) = \mathbb{E}[D(\mathcal{X}|K, \Phi)]. \tag{5.6}$$

5.3 Results

We give a closed form expression for the MSE distortion of LMMSE estimator for both sampling error and observation error. Our results are formulated in terms of the η-transform of the asymptotic spectra of the Wishart and F matrices (see Appendix C for more information).

Theorem 5.1. *Let $\mathcal{X}$ be a stochastic signal whose non-zero elements are i.i.d. with zero mean and variance σ_x. For sparsity $\Omega \in (0, 1)$ and sampling rate ρ consider the distortion of the linear minimum mean squared estimator. Under the influence of sampling noise i.i.d. with zero mean and variance σ_w^2 the distortion is given by*

$$D_{smp}(\mathcal{X}) = \eta_{WS}\left(\rho\beta; \tfrac{\rho}{\Omega}\right), \tag{5.7}$$

where $\beta = \sigma_x^2/\sigma_w^2$ and $\eta_{WS}(x; r)$ is given by Lemma B.4. Under the influence of observation noise i.i.d. with zero mean and variance σ_w^2 the distortion is given by

$$D_{obs}(\mathcal{X}) = \begin{cases} \frac{\beta}{1+\beta}\left[\frac{1}{\beta} + \eta_{FM}\left(\frac{\Omega}{1-\Omega}(\beta+1); \frac{\rho}{\Omega}, \frac{\rho}{1-\Omega}\right)\right] & 0 \le \rho < 1-\Omega \\ \frac{\beta}{1+\beta}\left[\frac{1}{\beta} + \frac{1-\rho}{\Omega}\eta_{FM}\left(\frac{1-\rho}{\rho}(\beta+1); \frac{1-\Omega}{1-\rho}, \frac{1-\Omega}{\rho}\right)\right] & 1-\Omega \le \rho < 1 \\ 1/(1+\beta) & 1 \le \rho \end{cases} \tag{5.8}$$

where $\beta = \sigma_x^2/\sigma_w^2$ and $\eta_{FM}(x; r_1, r_2)$ is given by Lemma B.5.

5.4 Discussion

The distortions in Theorem 5.1 are shown in the following figures. Figure 5.2 shows $\log_{10} D$ as a function of ρ for various Ω, Figure 5.3 shows $\log_{10} D$ as a function of ρ for various β, and Figure 5.4 shows $\log_{10} D$ as a function of Ω for various scalings of $\rho(\Omega)$.

The results are in line with our intuition. For $\rho < 1$ the noise due to observation noise is correlated and is less detrimental to recovery. On the other hand, for $\rho > 1$, D_{obs} remains constant whereas $D_{\text{smp}} \to 0$ as $\rho \to \infty$. Figure 5.3 shows that these results are consistent for a range of SNR. We remark that in all cases, the differences between the types of noise becomes very small as Ω becomes small. This occurs because $\eta_{\text{FM}}(\beta) \to \eta_{\text{WM}}(\beta)$, as the ratio r_2 of the F-Matrix goes to zero.

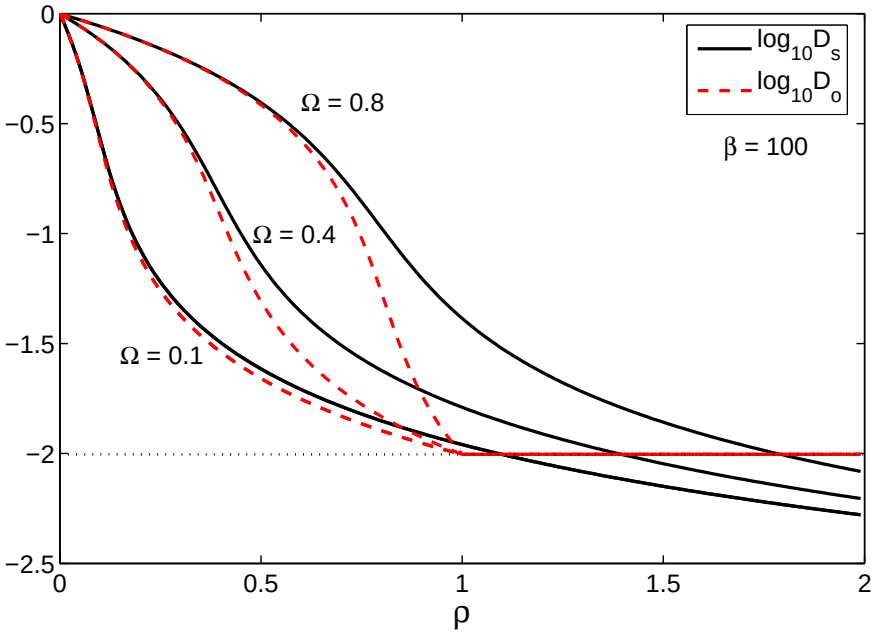

Figure 5.2: The distortion, $\log_{10} D$, as a function of ρ for various Ω and $\beta = 100$ under sampling noise (solid) and observation noise (dashed).

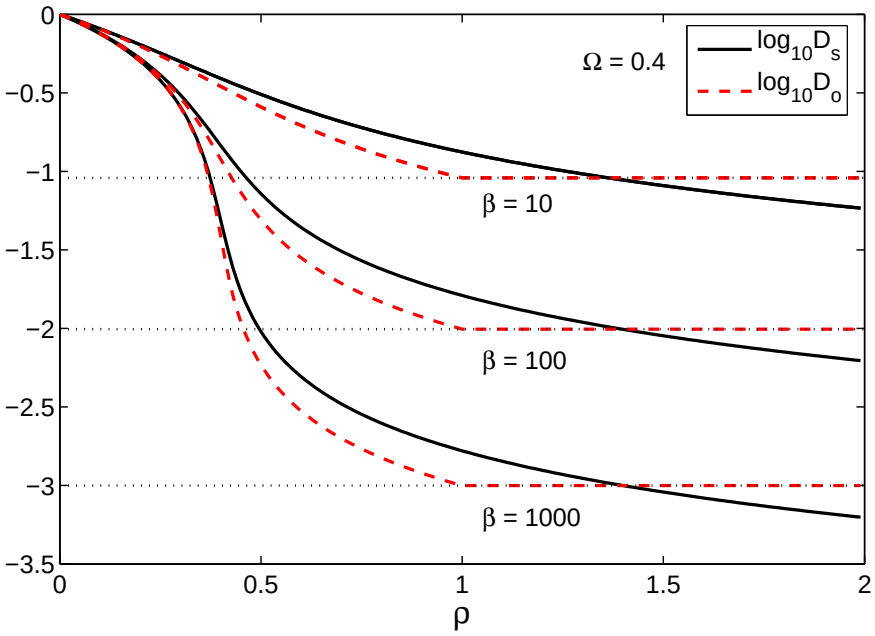

Figure 5.3: The distortion, $\log_{10} D$, as a function of ρ for various β and $\Omega = 0.4$ under sampling noise (solid) and observation noise (dashed).

5.5 Proofs

The first step toward calculating the distortion is to recast the conditional distortion (5.6) in terms of a single random matrix. Under the sampling error model we have

$$\Sigma_x = \sigma_x^2 I_n \qquad \Sigma_y = \sigma_x^2 \Phi_K \Phi_K^T + \sigma_w^2 I_m \qquad \Sigma_{xy} = \Phi_K^T.$$

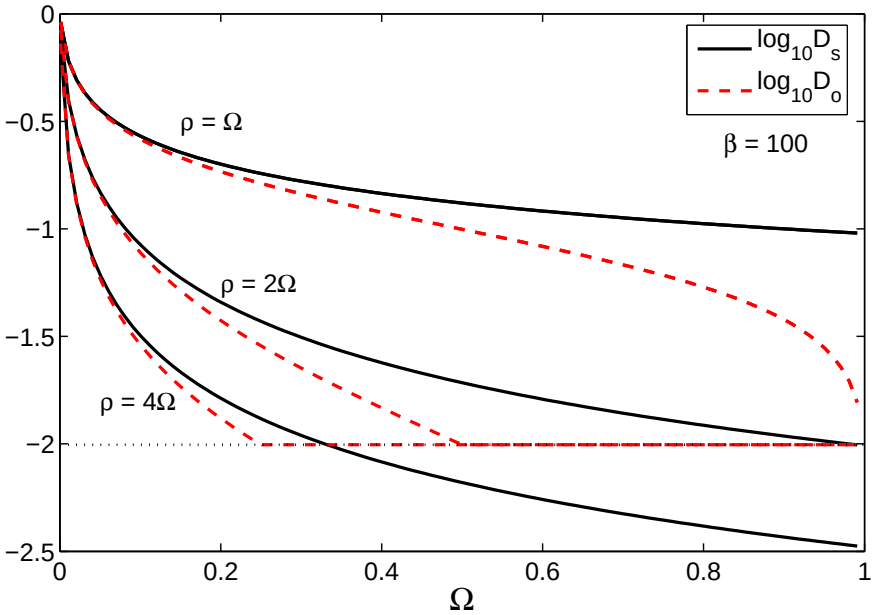

Figure 5.4: The distortion, $\log_{10} D$, as a function of Ω for various linear scalings of ρ and Ω for $\beta = 100$ under sampling noise (solid) and observation noise (dashed).

and thus

$$D_{\text{smp}}(\mathcal{X}|K,\Phi) = \frac{1}{k}\text{tr}\left\{I_n - \Phi_K^T(\Phi_K\Phi_K^T + (1/\beta)I)^{-1}\Phi_K\right\} \tag{5.9}$$

$$= \frac{1}{k}\text{tr}\left\{(I_k + \beta\Phi_K^T\Phi_K)^{-1}\right\}. \tag{5.10}$$

where $(\frac{n}{m})\Phi_K^T\Phi_K$ is a central Wishart matrix.

Under the observation error model we have

$$\Sigma_x = \sigma_x^2 I_n \qquad \Sigma_y = \sigma_x^2\Phi_K\Phi_K^T + \sigma_w^2\Phi_{K^\perp}\Phi_{K^\perp}^T \qquad \Sigma_{xy} = \Phi_K^T.$$

and thus

$$D_{\text{smp}}(\mathcal{X}|K,\Phi) = \frac{1}{k}\text{tr}\left\{I_n - \Phi_K^T(\Phi_K\Phi_K^T + (1\beta)\Phi_{K^\perp}\Phi_{K^\perp}^T)^{-1}\Phi_K\right\} \tag{5.11}$$

To simplify the above equation we must consider whether or not the covariance matrix $\Sigma_{w'} = \Phi_{K^\perp}\Phi_{K^\perp}^T$ corresponding to $\Phi_{K^\perp}w_{\text{obs},K^c}$ is invertible. For $m < n-k$, the matrix $\Sigma_{w'}$ is full rank and the resulting conditional distortion is

$$D_{\text{obs},1}(\mathcal{X}|K,\Phi) = \frac{\beta}{1+\beta}\left[\frac{1}{\beta} + \frac{1}{k}E\,\text{tr}\left\{\left(I_k + (1+\beta)\Phi_K^T\Sigma_{w'}^{-1}\Phi_K\right)^{-1}\right\}\right] \tag{5.12}$$

where $(\frac{n-k}{k})\Phi_K^T\Sigma_{w'}^{-1}\Phi_K$ is a central F matrix.

For $n-k < m < n$, the matrix $\Sigma_{w'}$ has rank $n-k$ and is not invertible. The distortion can be determined using $\tilde{\Sigma}_{w'} = \Phi_{K^\perp}^T\Phi_{K^\perp}$ which has the same non-zero eigenvalues as $\Sigma_{w'}$,

and an $n-k \times n-m$ random matrix G whose i.i.d. elements have the same distribution as the elements of Φ. The conditional distortion is equal in distribution to the following

$$D_{\text{obs},2} \stackrel{d}{=} \frac{\beta}{1+\beta}\left[\frac{1}{\beta} + \frac{1}{k} E \operatorname{tr}\left\{\left(I_{m-n} + (1+\beta) G^* \tilde{T}^{-1} G\right)^{-1}\right\}\right]. \tag{5.13}$$

where $(\frac{m}{m-n}) G^T \tilde{\Sigma}_{w'}^{-1} G$ is a central F matrix. Finally, for $m \geq n$, the matrix Φ is invertible and the distortion is simply $1/(1+\beta)$.

The next step involves evaluating the expectations of the conditional distortions given in (5.10), (5.12), and (5.13). We use the following fact.

Lemma 5.1. *For a nonnegative definite matrix $M \in \mathbb{B}^{n\times n}$ let $\lambda_1(M), \cdots \lambda_n(M)$ denote the eigenvalues of M and let $F_M^n(x)$ denote the empirical eigenvalue distribution (B.1). For $\gamma \geq 0$ we have*

$$\frac{1}{n} tr\left\{(I+\gamma M)^{-1}\right\} = \sum_{i=1}^{n} \frac{1}{1+\gamma\lambda_i(M)} \tag{5.14}$$

$$= \int_0^\infty \frac{1}{1+\gamma x} d\, F_M^n(x) \tag{5.15}$$

Proof. This follows from the properties of the trace. □

We are interested in the asymptotic limits as $n, k, m \to \infty$ with $k/n \to \Omega$ and $m/n \to \rho$. In this setting, it has been shown that for both both the Wishart [26] and F matrices [28] the empirical probability distributions converge to non-random continuous functions with closed form expressions which are given in Lemmas B.1 and B.2. This means that the conditional distributions (5.10), (5.12), and (5.13) converge to some non-random quantity. Moreover, this quantity, which corresponds to the η-transform (defined in appendix B), has a closed from solution for Wishart matrix and the F matrix. To conclude we note that as $n \to \infty$,

$$D_{\text{smp}}(\mathcal{X}|K,\Phi) \quad \to \quad \eta_{\text{WS}}\left(\rho\beta; \tfrac{\rho}{\Omega}\right) \tag{5.16}$$

$$D_{\text{obs},1}(\mathcal{X}|K,\Phi) \quad \to \quad \frac{\beta}{1+\beta}\left[\frac{1}{\beta} + \eta_{\text{FM}}\left(\frac{\Omega}{1-\Omega}(\beta+1); \frac{\rho}{\Omega}, \frac{\rho}{1-\Omega}\right)\right] \tag{5.17}$$

$$D_{\text{obs},2}(\mathcal{X}|K,\Phi) \quad \to \quad \frac{\beta}{1+\beta}\left[\frac{1}{\beta} + \frac{1-\rho}{\Omega}\eta_{\text{FM}}\left(\frac{1-\rho}{\rho}(\beta+1); \frac{1-\Omega}{1-\rho}, \frac{1-\Omega}{\rho}\right).\right]. \tag{5.18}$$

where $\eta_{\text{WS}}(\gamma; r)$ and $\eta_{\text{FM}}(x; r_1, r_2)$ are given by Lemmas B.4 and B.5.

Chapter 6

Conclusions and Future Work

Research over the past decade has changed what it means to "sample" a signal. The field of compressed sensing in particular has focused on recovering sparse signals from a small number of linear projections. In this thesis we have given fundamental limits on what can and cannot be learned about an unknown signal when the samples, which consist of randomly constructed linear projections, are corrupted by noise. Our contributions lie it two areas: support recovery bounds and the effects of different noise models in signal estimation.

For the task of support recovery, Chapter 3 established that perfect recovery cannot be achieved in the setting of linear sparsity unless the SNR grows without bound with the signal dimension. This result is significant because it shows that, in many applications, perfect support recovery is not attainable.

Next, we considered partial support recovery, and showed that it is possible to recover some fraction of the support. Chapters 3 and 4 gave complementary necessary and sufficient conditions on the number of samples and the SNR required to attain a desired accuracy. The results of Chapter 3 gave an upper bound on the fraction of the support that can be recovered by any estimation algorithm. The results of Chapter 4 gave a lower bound on the fraction of the support that can be recovered using a particle maximum likelihood estimator.

Our results on partial support recovery quantify how much information about the support can be extracted from noisy linear projections. Unlike previous bounds which considered only perfect recovery, we are able derive achievable results for support recovery in the under-sampled linear sparsity setting. Also, our results were developed in parallel for both stochastic and non-stochastic signal models.

Our other contribution dealt with the effects of different noise models on the ability to estimate a sparse signal. Chapter 5 compared the traditional sampling model, where noise is added independently to each sample, to a model where noise is added to the signal prior to sampling. We used results on the asymptotic spectrum of certain random matrices to derive closed form expressions of the means squared error distortion for both models. Our results showed that in the under-sampled setting, noise added prior to sampling is less detrimental than noise added to the samples. Furthermore, the difference between the two types of noise decreases as the unknown signal becomes sparser.

We recognize several directions for further work. One potentially useful extension of this

thesis is to derive sufficient conditions, like those given in Chapter 4, that correspond to an efficient estimation algorithm. The main drawback of the ML decoder we analyzed is that it is computationally hard, and is thus not practical in many settings. One way this analysis might be achieved, is to bound the performance of some efficient estimator with respect to the ML estimator. In conjunction with the results of Chapter 4, such an approach would yield an achievable result for the efficient estimator.

Next, we note that our primary motivation for using a gaussian sampling matrix was that it simplified our analysis. In general, however, it would be interesting to see to what degree our results hold for other matrix constructions. This would be particularly interesting for matrices which arise naturally out of the physics of the sampling process, or for matrices that are themselves sparse, and are thus easy to implement.

Finally, our results on signal estimation in Chapter 5 illustrate the impact of where the noise enters into the samples. A natural extension is to see what effect noise added prior to sampling has on the ability to recover some fraction of the support.

Appendix A

Facts about Chi Squared Variables

In this appendix we provide a number of useful facts about χ^2 random variables and prove Lemmas that are used throughout the thesis. We begin with some standard definitions.

Definition A.1. Given d independent variables $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, then

$$\sum_{i=1}^{d} \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2 \tag{A.1}$$

is a central $\chi^2(d)$ variable with d degrees of freedom.

Remark A.1. The variable $Z \sim \chi^2(d)$ is non-negative, has mean d, and variance $2d$. For $z \geq 0$ its probability density function is given by

$$f_Z(z) = \frac{z^{m/2-1} e^{-d/2}}{2^{d/2}\Gamma(d/2)} \tag{A.2}$$

where $\Gamma(s) = \int_0^\infty t^{s-1} e^{-t} dt$ is the Gamma function.

Definition A.2. Given d independent variables $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ then

$$\sum_{i=1}^{d} \left(\frac{X_i}{\sigma_i} \right)^2 \tag{A.3}$$

is a non-central $\chi^2_{NC}(d, \nu)$ variable with d degrees of freedom and non-centrality parameter

$$\nu = \sum_{i=1}^{d} \left(\frac{\mu_i}{\sigma_i} \right)^2 \tag{A.4}$$

Remark A.2. The variable $Z \sim \chi^2_{NC}(d, \nu)$ is non-negative, has mean $d + \nu$, and variance $2(d + 2\nu)$. Let $N \sim \text{Poisson}(\nu/2)$ then Q is equal in distribution to a central χ^2 variable $Y \sim \chi^2(d + 2N)$. Accordingly, its probability density function can be written in terms of the central χ^2 distribution as

$$f_Z(q) = \sum_{i=0}^{\infty} \frac{e^{-\nu/2}(\nu/2)^i}{i!} f_{\chi^2(d+2i)}(z). \tag{A.5}$$

Next we will provide some useful properties. We start with large deviation bounds.

Lemma A.1. *Given $Z \sim \chi^2(d)$ and any $\epsilon > 0$ the following bounds hold*

$$Pr\{Z > (1+\epsilon)d\} \leq \exp\left(-\frac{d}{4}\epsilon^2\right) \tag{A.6}$$

$$Pr\left\{Z < \frac{1}{(1+\epsilon)}d\right\} \leq \exp\left(-\frac{d}{2}\left[\log(1+\epsilon) - \frac{\epsilon}{1+\epsilon}\right]\right) \tag{A.7}$$

Proof. This lemma follows from a Chernoff bound and the proof is outlined in [29]. Let Z be a central χ^2 variable with d degrees of freedom. For any $\mu < 0$

$$Pr(Z \leq z) = Pr(z^{\mu(Z-z)} \geq 1).$$

By Markov's inequality this means

$$\begin{aligned} Pr(Z \leq z) &\leq \inf_{\mu<0} \mathbb{E}\left[e^{\mu(Z-z)}\right] \\ &= \inf_{\mu<0} (1-2\mu)^{-d/2} \exp\left(-x\mu\right) \\ &= \inf_{\mu<0} \exp\left(-d\frac{1}{2}\ln(1-2\mu) - \mu z\right) \end{aligned} \tag{A.8}$$

The infimum occurs at $\mu = (1 - d/z)/2$ and plugging this value into (A.8) results in the lower tail bound (A.6). The well known upper tail bound (A.7) is proved in [29]. □

Lemma A.2. *Given a p-dimensional random vector $X \sim \mathcal{N}(0, I)$ and any $p \times p$ orthonormal projection matrix Π with rank d and independent of X, then $||\Pi X||^2 \sim \chi^2(d)$.*

Proof. We can write $\Pi = U^T \Lambda U$ where Λ is a diagonal matrix with exactly k ones and $p-k$ zeros and $U^T U = I$. Since U is orthonormal, the vector $Y = UX$ is equal in distribution to X. Thus we have

$$||Ux||^2 = \sum_{i=1}^{p} Y^T Y \lambda_i = \sum_{i\,:\,\lambda_i \neq 0} Y_i^2 \tag{A.9}$$

which is χ^2 with d degrees of freedom because Y_i are independently distributed $\mathcal{N}(0,1)$. □

Lemma A.3. *The cumulative distribution function of a non-central χ^2_{NC} variable is decreasing in the non-centrality parameter ν, that is, for any integer $d > 0$ and scalars x, $\nu' \geq \nu \geq 0$ we have*

$$Pr\left\{\chi^2_{NC}(d, \nu') < x\right\} \leq Pr\left\{\chi^2_{NC}(d, \nu) < x\right\} \tag{A.10}$$

Proof. Let $Q(x;d,\nu) = Pr\left\{\chi^2_{NC}(d,\nu') < x\right\}$ and let $F(x;d) = Pr\left\{\chi^2(d) < x\right\}$. Note that $Q(x;d,\nu)$ can be written in terms of $F(x;\cdot)$ as

$$Q(x;d,\nu) = \sum_{j=0}^{\infty} e^{-\nu/2}\frac{(\nu/2)^j}{j!}F(x;d+2j)$$

Taking the partial derivative with respect to ν gives

$$\frac{\partial}{\partial\nu}Q(x;d,\nu) = \frac{1}{2}\sum_{j=0}^{\infty} e^{-\nu/2}\frac{(\nu/2)^j}{j!}\left[F(x;d+2+2j) - F(x;d+2j)\right]$$

Since $F(x;d)$ is strictly decreasing in d the above quantity is strictly negative. Thus, $Q(x;d,\nu)$ is strictly decreasing in ν. □

The following lemma is a counterpart to a large deviation bound. Roughly speaking, it will be used to bound a non-central $\chi^2_{NC}(d,d^2)$ variable a distance of d away from its mean.

Lemma A.4. *Given any $\tau,\gamma_1 < \infty$ and $\gamma_2 > 0$ there exists a constant $c > 0$ and integer $M < \infty$ such that for all $t \in [0,\tau]$ and $m \geq M$ the probability distribution function of the non-central χ^2_{NC} variable $Z \sim \chi^2_{NC}(m,\gamma_1+\gamma_2 m^2)$ is bounded as*

$$f_Z(\mathbb{E}[Z] - t\cdot m) > \frac{c_0}{m} \tag{A.11}$$

Proof. Let $N \sim Poisson((\gamma_1 m + \gamma_2 m^2)/2)$ and let $P = m + 2N$. Then, the pdf of Z can be written as

$$f_Z(z) = \mathbb{E}_P\left[\frac{z^{P/2-1}e^{-z/2}}{2^{P/2}\Gamma(P/2)}\right] = \mathbb{E}_P\left[\frac{1}{2\Gamma(P/2)}\exp\left\{\left(\frac{P}{2}-1\right)\log\left(\frac{z}{2}\right) - \frac{z}{2}\right\}\right]. \tag{A.12}$$

Using Stirling's approximation, $\Gamma(s) = \sqrt{2\pi}e^{-s+\nu/s}s^{s-1/2}$ with $|\nu| < 1/12$, gives

$$f_Z(z) \geq \mathbb{E}_P\left[\frac{1}{\sqrt{4\pi P}}\exp\left\{-\left(\frac{P}{2}-1\right)\log\left(\frac{P}{z}\right) + \frac{P-z}{2} - \frac{1}{6P}\right\}\right].$$

Using the bound $\log(s) \leq s-1$ gives

$$\begin{aligned} f_Z(z) &\geq \mathbb{E}_P\left[\frac{1}{\sqrt{4\pi P}}\exp\left\{-\left(\frac{P}{2}-1\right)\left(\frac{P-z}{z}\right) + \frac{P-z}{2} - \frac{1}{6P}\right\}\right] \\ &> \mathbb{E}_P\left[\frac{1}{\sqrt{4\pi P}}\exp\left\{-\frac{(P-z)^2}{2z} - \frac{1}{6P}\right\}\right]. \end{aligned}$$

Note that $\mathbb{E}[P] = \mathbb{E}[Z] = (1+\gamma_1)m + \gamma_2 m^2$. Thus, with the substitution $z = \mathbb{E}[Z] - t\,m$ we have

$$f_Z(\mathbb{E}[Z] - t\,m) \geq \mathbb{E}_P\left[\frac{1}{\sqrt{4\pi P}}\exp\left\{-\frac{(P-\mathbb{E}[P]+t\,m)^2}{2(\mathbb{E}[P]-t\,m)} - \frac{1}{6P}\right\}\right].$$

Using standard concentration results for Poisson random variable [30], there exists some constant $c_0 > 0$ and integer $M_0 < \infty$ such that, for all $m \geq M_0$ the following bound holds

$$Pr\{|P - \mathbb{E}[P]| < 2\,m\} > c_0. \tag{A.13}$$

Accordingly, for $m \geq M_0$ we have

$$f_Z(\mathbb{E}[Z] - t\,m) \geq C_1 \exp\{-C_2\}$$

where

$$C_1 = \frac{c_0}{\sqrt{4\pi[(3+\gamma_1)m + \gamma_2 m^2]}}$$
$$C_2 = \frac{(2+t)^2 m}{2[(1+\gamma_1 - t)m + \gamma_2 m^2]} + \frac{1}{6[(1+\gamma_1)m + \gamma_2 m^2]}$$

Given τ, we may choose some $M \geq M_1$ and constants $c_1 > 0$ and $c_2 < \infty$ such that for all $m > M$ we have

$$C_1 \geq \frac{c_1}{m},$$
$$C_2 \leq c_2$$

for all $t \in [0, \tau]$. With $c = c_1 e^{-c_2}$ we conclude our proof.

□

Appendix B

The Asymptotic Spectrum of Random Matrices

In this appendix we provide results about the limiting eigenvalue distribution of two matrices that arise from our sampling model. With the exception of Lemma B.5, this is a compilation of known results (see [31] for more information). In general these results hold for complex matrices, but for simplicity we consider only real matrices.

We begin with the definitions of the central Wishart and central F matrices.

Definition B.1. Let $H \in \mathbb{R}^{m\times k}$ have zero mean i.i.d. entries with variance one. The $k \times k$ matrix $(\frac{1}{m})H^T H$ is a central Wishart matrix.

Definition B.2. Let $H \in \mathbb{R}^{m\times k}$ and $M \in \mathbb{R}^{m\times p}$ $(m < p)$ have zero mean i.i.d. Gaussian entries with unit variance. The $k \times k$ matrix $(\frac{1}{k})H^T((\frac{1}{p})MM^T)^{-1}H$ is a central F matrix.

In many applications in signal processing and communications, the relevant performance metrics only depend on the singular values of the matrix M, rather than on its more precise structure. Accordingly, a great deal of research has focused on the behavior of the singular values for certain random matrices. In this work, we need the distribution of the eigenvalues (i.e. the spectrum) for the Wishart and F matrix. For an $n \times n$ matrix M the empirical cumulative distribution of the eigenvalues is given by

$$F_M^n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\lambda_i(M) \leq x\} \tag{B.1}$$

where $\lambda_1(M), \cdots \lambda_n(M)$ are the eigenvalues of M and $\mathbb{1}\{\cdot\}$ is the indicator function. A remarkable fact about the matrices we are interested in, is that their empirical distributions converge to a non-random continuous function as $n \to \infty$. For the Wishart matrix this is the classic Marcenko-Pastur law [26] and is given in Lemma B.1. For the F matrix this result was shown by Silverstein [28] and is given in Lemma B.2.

Lemma B.1. *Let $H \in \mathbb{R}^{m\times k}$ have zero mean i.i.d. entries with variance one. If $k/m \to r$ as $k, m \to \infty$, then the empirical spectral density of the central Wishart matrix $(\frac{1}{m})H^*H$ converges to*

$$f_{WS}(x;r) = \left(1+\frac{1}{r}\right)^+ \delta(x) + \frac{\sqrt{(x-a)^+(b-x)^+}}{2\pi r x} \tag{B.2}$$

with $a = (1-\sqrt{r})^2$, $b = (1+\sqrt{b})^2$ and $(x)^+ = \max\{0, x\}$.

Lemma B.2. *Let $H \in \mathbb{R}^{m\times k}$ and $M \in \mathbb{C}^{m\times p}$ $(m < p)$ have zero mean i.i.d. Gaussian entries with unit variance. If $m/k \to r_1$ and $m/p \to r_2 \in (0,1)$ as $k, m, p \to \infty$, then the empirical spectral density of the central F matrix $(\frac{1}{k})H^*((\frac{1}{p})MM^*)^{-1}H$ converges to*

$$f_{FM}(x;r_1,r_2) = \left(1+\frac{1}{r_1}\right)^+ \delta(x) + \frac{(1-r_2)\sqrt{(x-a)^+(b-x)^+}}{2\pi x(r_1 x + r_2)} \tag{B.3}$$

with

$$a = \left(\frac{1-\sqrt{1-(1-r_1)(1-r_2)}}{1-r_2}\right)^2 \qquad b = \left(\frac{1+\sqrt{1-(1-r_1)(1-r_2)}}{1-r_2}\right)^2$$

In typical applications involving a random matrix M the properties of interest depend on the expected value of some function of M. The following two transforms are taken directly from [31] and are useful for many problems in signal processing and communications.

Definition B.3. Given a nonnegative variable X and $\gamma \geq 0$ the Shannon transform is

$$\mathcal{V}_X(\gamma) = \frac{1}{2}\mathbb{E}\left[\log(1+\gamma X)\right] \tag{B.4}$$

Definition B.4. Given a nonnegative variable X and $\gamma \geq 0$ the η-transform is

$$\eta_X(\gamma) = \mathbb{E}\left[\frac{1}{1+\gamma X}\right] \tag{B.5}$$

The following lemmas provide the transforms necessary for this thesis. Lemmas B.4 and B.4 are taken from [31] whereas Lemma B.4 is, to our knowledge, the first closed form expression of the η-transform for the F-matrix.

Lemma B.3. *The Shannon transform of the distribution $f_{WS}(x;r)$ is given by*

$$\mathcal{V}_{WS}(\gamma;r) = \log\left(1+\gamma-F_1(\gamma,r)\right) + \frac{1}{r}\log\left(1+r\gamma-F_1(\gamma,r)\right) + \frac{1}{r\gamma}F_1(\gamma,r) \tag{B.6}$$

with

$$F_1(\gamma,r) = \frac{1}{4}\left(\sqrt{\gamma(1+\sqrt{r})^2+1} - \sqrt{\gamma(1-\sqrt{r})^2+1}\right)^2. \tag{B.7}$$

Lemma B.4. *The η-transform of the distribution $f_{WS}(x;r)$ is given by*

$$\eta_{WS}(\gamma;r) = 1 - \frac{1}{r\gamma r} F_1(\gamma, r) \tag{B.8}$$

where $F_1(\gamma, r)$ is given by (B.7).

Lemma B.5. *The η-transform of the distribution $f_{FM}(x;r_1,r_2)$ is given by*

$$\eta_{FM}(\gamma, r_1, r_2) = \frac{1}{1-r_2} + \left(\frac{\gamma}{\gamma - r_2/r_1}\right) F_2(\gamma) + \left(\frac{r_2/r_1}{r_2/r_1 - \gamma}\right) F_2(r_2/r_1) \tag{B.9}$$

with

$$a = \left(\frac{1-\sqrt{1-(1-r_1)(1-r_2)}}{1-r_2}\right)^2 \quad b = \left(\frac{1+\sqrt{1-(1-r_1)(1-r_2)}}{1-r_2}\right)^2$$
$$F_2(x) = \frac{1-r1}{2} - \frac{1}{1-r_2} + \frac{1-r_2}{2x}(\sqrt{(1+ax)(1+bx)} - 1).$$

Proof. The key to solving this nontrivial integration problem is to use the substitution $x = 1 + c - 2\sqrt{c}\cos u$ for a properly chosen value of c. □

Bibliography

[1] P. Feng and Y. Bresler, "Spectrum-blind minimum-rate sampling and reconstruction of multiband signals," in *Proc. IEEE Int. Conf. Acoust. Speech Sig. Proc.*, vol. 3, Atlanta, GA, May 1996, pp. 1689–1692.

[2] R. Venkataramani and Y. Bresler, "Sub-Nyquist Sampling of Multiband Signals: Perfect Reconstruction and Bounds on Aliasing Error," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Seattle, WA, Apr 1998, pp. 1633–1636.

[3] M. Gastpar and Y. Bresler, "On the necessary density for spectrum-blind nonuniform sampling subject to quantization," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Istanbul, Turkey, Jun 2000, pp. 248–351.

[4] M. Vetterli, P. Marziliano, and T. Blu, "Sampling signals with finite rate of innovation," *IEEE Transactions on Signal Processing*, vol. 50(6), pp. 1417–1428, 2002.

[5] D. Donoho, "Compressed Sensing," *IEEE Trans. on Information Theory*, vol. 52(4), pp. 1289–1306, Apr. 2006.

[6] E. Candes, J. Romberg, and T. Tao, "Near optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Trans. on Information Theory*, vol. 52(12), pp. 5406–5425, Dec. 2006.

[7] W. Wang, M. Garofalakis, and K. Ramchandran, "Distributed sparse random projections for refinable approximation," in *Proc. Int. Conf. on Info. Processing in Sensor Networks*, Cambridge, MA, Apr. 2007.

[8] W. Bajwa, J. Haupt, A. Sayeed, and R. Nowak, "Compressive wireless sensing," in *Proc. Int. Conf. on Info. Processing in Sensor Networks*, Nashville, TN, Apr. 2006, pp. 134–142.

[9] M. Rabbat, J. Haupt, A. Singh, and R. Nowak, "Decentralized Compression and Predistribution via Randomized Gossiping," in *Proc. Int. Conf. on Info. Processing in Sensor Networks*, Nashville, TN, Apr. 2006, pp. 51–59.

[10] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal of Scientific Computing*, vol. 20(1), pp. 33–61, 1998.

[11] D. Donoho, M. Elad, and V. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. on Information Theory*, vol. 52(1), pp. 6–18, Jan 2006.

[12] E. Candes, J. Romberg, and T. Tao, "Stable Signal Recovery from Incomplete and Inaccurate Measurements," *Comm. on Pure and Applied Math.*, vol. 59(8), pp. 1207–1223, 2006.

[13] J. Tropp, "Just relax: Convex programming methods for identifying sparse signals in noise," *IEEE Trans. on Information Theory*, vol. 52(3), pp. 1030–1051, 2006.

[14] J. Fuchs, "Recovery of exact sparse representations in the presence of noise," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Montreal, QC, Canada, May 2004, pp. 533–536.

[15] ——, "Recovery of exact sparse representations in the presence of bounded noise," *IEEE Trans. on Information Theory*, vol. 51(10), pp. 3601–3608, Oct 2005.

[16] N. Meinshausen and P. Buhlmann, "Consistent neighborhood selection for high-dimensional graphs with Lasso," *Annals of Statistics*, vol. 34(3), 2006.

[17] P. Zhao and B. Yu, "On model selection consistency of lasso," *Journal of Machine Learning Research*, vol. 51(10), pp. 2541–2563, Nov 2006.

[18] M. Wainwright, "Sharp thresholds for high-dimensional and noisy recovery of sparsity," in *Proc. Allerton Conf. on Comm., Control, and Computing*, Monticello, IL, Sep 2006.

[19] S. Sarvotham, D. Baron, and R. Baranuik, "Measurements vs. bits: Compressed sensing meets information theory," in *Proc. Allerton Conference on Communication, Control, and Computing*, Monticello, IL, Sep 2006.

[20] A. Fletcher, S. Rangan, V. Goyal, and K. Ramchandran, "Denoising by sparse approximation:Error bounds based on rate-distortion theory," *Journal on Applied Signal Processing*, vol. 10, pp. 1–19, 2006.

[21] A. Fletcher, S. Rangan, and V. Goyal, "Rate-Distortion bounds for sparse approximation," in *Proc. IEEE Statistical Signal Processing Workshop*, Madison, WI, Aug 2007, pp. 254–258.

[22] S. Aeron, M. Zhao, and V. Saligrama, "Sensing Capacity, Diversity and Sparsity: Fundamental Tradeoffs," in *ITA UCSD Workshop*, La Jolla, CA, Jan 2007.

[23] M. Wainwright, "Information-theoretic bounds on sparsity recovery in the high-dimensional and noisy setting," in *Proc. IEEE Int. Symposium on Information Theory*, Nice, France, Jun 2007.

[24] W. Hoeffding, "Probability Inequalities for Sums of Bounded Random Variables," *Journal of the American Statistical Association*, vol. 58(301), pp. 13–30, 1963.

[25] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[26] V. Marcenko and L. Pastur, "Distribution of eigenvalues for some sets of random matrices," *Math. USSR-Sbornik*, vol. 1, pp. 457–483, 1967.

[27] G. Reeves and M. Gastpar, "Differences between observation and sampling error in sparse signal reconstruction," in *Proc. IEEE Statistical Signal Processing Workshop*, Madison, WI, Aug 2007, pp. 690–694.

[28] J. Silverstein, "The limiting eigenvalue distribution of a multivariate F-matrix," *SIAM J. of Math. Analysis*, vol. 30, pp. 641–646, 1985.

[29] B. Laurent and P. Massart, "Adaptive estimation of a quadratic functional by model selection," *Annals of Statistics*, vol. 28(5), pp. 1302–1338, 2000.

[30] H. Louis, K. Chen, and K. Choi, "Some Asymptotic and Large Deviation Results in Poisson Approximation," *The Annals of Probability*, vol. 20(4), pp. 1867–1876, Oct 1992.

[31] A. Tulino and S. Verdu, *Random Matrix Theory and Wireless Communications*. Hanover, MA: now Publisher Inc., 2004.