UC Berkeley
UC Berkeley Electronic Theses and Dissertations

Title
Sparsity Pattern Recovery in Compressed Sensing

Permalink
https://escholarship.org/uc/item/7775157z

Author
Reeves, Galen

Publication Date
2011

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Sparsity Pattern Recovery in Compressed Sensing

by

Galen Reeves

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Engineering — Electrical Engineering and Computer Sciences
with a Designated Emphasis in Communication, Computation and Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Michael Gastpar, Chair
Professor Martin Wainwright
Professor Peter Bickel

Fall 2011

Sparsity Pattern Recovery in Compressed Sensing

Copyright 2011
by
Galen Reeves

Abstract

Sparsity Pattern Recovery in Compressed Sensing

by

Galen Reeves

Doctor of Philosophy in Engineering — Electrical Engineering and Computer Sciences
with a Designated Emphasis in Communication, Computation and Statistics

University of California, Berkeley

Professor Michael Gastpar, Chair

The problem of recovering sparse signals from a limited number of measurements is now ubiquitous in signal processing, statistics, and machine learning. A natural question of fundamental interest is that of what can and cannot be recovered in the presence of noise. This thesis provides a sharp characterization for the task of sparsity pattern recovery (also known as support recovery). Using tools from information theory, we find a sharp separation into two problem regimes – one in which the problem is fundamentally noise-limited, and a more interesting one in which the problem is limited by the behavior of the sparse components themselves. This analysis allows us to identify settings where existing computationally efficient algorithms, such as the LASSO or approximate message passing, are optimal as well as other settings where these algorithms are highly suboptimal. We compare our results to predictions of phase transitions made via the powerful but heuristic replica method, and find that our rigorous bounds confirm some of these predictions.

The remainder of the thesis explores extensions of our bounds to various scenarios. We consider specially structured sampling matrices and show how such additional structure can make a key difference, analogous to the role of diversity in wireless communications. Finally, we illustrate how the new bounding techniques introduced in this thesis can be used to establish information-theoretic secrecy results for certain communication channel models that involve eavesdroppers.

This thesis is dedicated to my mother.

Acknowledgments

It has been a true pleasure to have Michael Gastpar as an advisor. As a role model, he has inspired me by his integrity and modesty. As a motivator, he has helped me achieve difficult tasks by making sure that I believed in what I was doing at each step along the way. Finally, as a researcher, he has pushed me to always strive for the meaningful research questions, not just the ones that are close at hand.

I would also like to thank Martin Wainwright, Kannan Ramchandran, and Peter Bickel for serving on my committee.

My first taste of research started at Cornell working with Toby Berger (my undergraduate advisor) and Sergio Servetto. During my time at Berkeley, I have benefited greatly from my interactions with Miki Lustig, Jim Pitman, Kannan Ramchandran, Anant Sahai, Shankar Sastry, David Tse, and Martin Wainwright. Beyond Berkeley, my internship with Jie Liu at Microsoft Research introduced me to new and exciting applications, and the many interactions I had as a visitor at TU Delft and EPFL greatly enriched my graduate experience.

My senior group members Bobak Nazer, Anand Sarwate, and Krish Eswaran were particularly helpful to me at the beginning of my graduate career. Thanks also to Salman Avestimehr, Guy Bresler, Alex Dimakis, Alyson Fletcher, Naveen Goela, Pulkit Grover, Nebojsa Milosavljevic, Sahand Negahban, Dapo Omidiran, Hari Palaiyanur, Gireeja Ranade, Changho Suh, Rahul Tandra, I-Hsiang Wang, Jiening Zhan, and all the other members of the Wireless Foundations community for making my time at Berkeley both productive and enjoyable.

I would also like to acknowledge the many friends, both at Berkeley and abroad, who have so greatly enriched my life. A special thanks to my roommates at the Copa Colusa for all of our endless discussions and shared adventures.

I am grateful to my family for their love and inspiration: my mother for teaching me the joy of asking questions, my father for sharing his passion as an academic, and David Wilkins for his deep and wonderful insights into the world. Thanks also to my sister for being my constant friend since the day I was born.

Finally, thanks to Mary Knox for her amazing love and support (and proofreading!). She gave me freedom follow my dreams but also brought me back to reality when I went astray. I owe her more than words can say.

Contents

Chapter 1

Introduction

Suppose that a vector $\mathbf{x}$ of length n is known to have a small number k of nonzero entries, but the values and locations of the nonzero entries are unknown and must be estimated from a set of m noisy linear projections (or samples) given by the vector

$$\mathbf{y} = A\mathbf{x} + \mathbf{w} \tag{1.1}$$

where A is a known $m \times n$ measurement matrix and $\mathbf{w}$ is additive white Gaussian noise. The problem of *sparsity pattern recovery* is to determine which entries in $\mathbf{x}$ are nonzero. This problem, which is known variously throughout the literature as support recovery or model selection, has applications in compressed sensing [8,11,20], sparse approximation [16], signal denoising [12], subset selection in regression [47], and structure estimation in graphical models [46].

A great deal of previous work [1,2,30,46,58,60,76–78,84], has focused on necessary and sufficient conditions for exact recovery of the sparsity pattern. By contrast, the primary focus of this thesis is on the tradeoff between the number of samples m and the number of detection errors. We consider the high-dimensional setting where the sparsity rate (i.e. the fraction of nonzero entries) and the per-sample signal-to-noise ratio (SNR) are finite constants, independent of the vector length n.

This thesis is outlined as follows:

- **Chapter 1:** In the remainder of this chapter, we overview the main contributions of this thesis, summarize previous work, and develop a framework for analyzing the problem of sparsity pattern recovery in terms of the *sampling rate-distortion* region.
- **Chapter 2:** We derive bounds on the sampling rate $\rho = m/n$ needed to attain a desired detection error rate D for several different recovery algorithms. These bounds are given explicitly in terms of the sparsity rate, the SNR, and various key properties of the unknown vector.

- **Chapter 3:** We derive information-theoretic necessary conditions on the sampling rate $\rho = m/n$ needed to attain a desired detection error rate D for the optimal recovery algorithm. These bounds are complementary to the bounds in Chapter 2.

- **Chapter 4:** We show how the bounds derived in Chapters 2 and 3 depend on the desired distortion D, the SNR, and various properties of the unknown vector. We then characterize problem regimes in which the behavior of the algorithms is near-optimal and other regimes in which the behavior is highly suboptimal. An illustration of the bounds is given in Fig. 1.1 below.

- **Chapter 5:** We extend the results and analysis developed in Chapters 2-4 to settings where one observes samples from multiple realizations of the nonzero values for the same sparsity pattern. We refer to this as "diversity" and show that the optimal amount of diversity significantly improves the behavior of the estimation problem for both optimal and computationally efficient algorithms.

- **Chapter 6:** We apply the insights developed in the previous chapters to analyze a vector wire-tap channel with multiplicative noise. This wire-tap channel has the surprising property that the secrecy capacity is nearly equal to the channel capacity, even if the eavesdropper observes as much as the intended receiver.

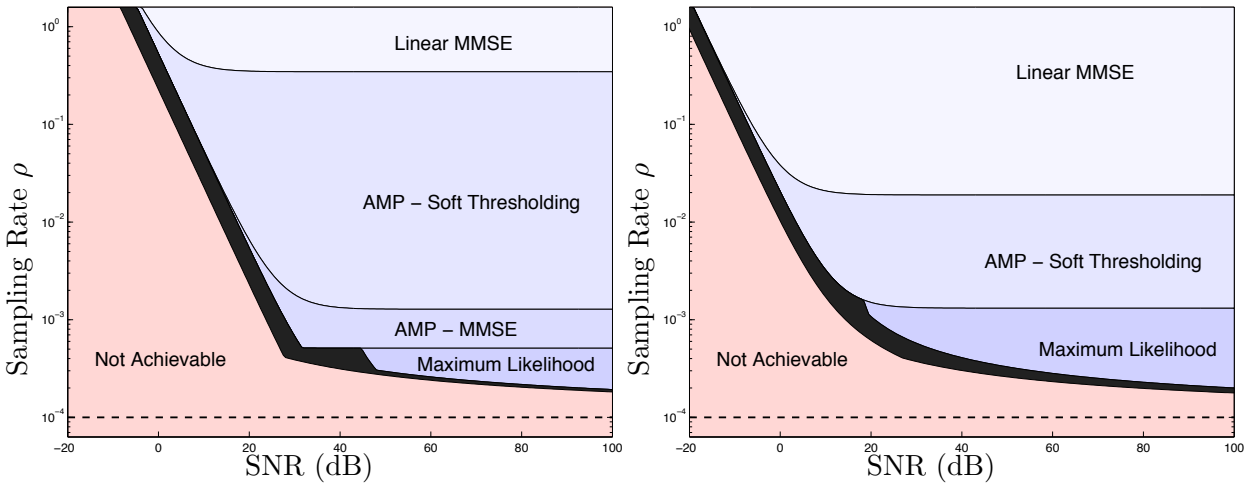

Figure 1.1: Bounds on the achievable sampling rate $\rho = m/n$ as a function of the SNR for various recovery algorithms when the desired sparsity pattern detection error rate is $D = 0.05$ (95% accuracy), the sparsity rate (i.e. fraction of nonzero entries in $\mathbf{x}$) is 10^{-4}, and the measurement matrices have i.i.d. Gaussian entries. In the left panel, the nonzero entries are i.i.d. zero-mean Gaussian. In the right panel, the nonzero entries are lower bounded in squared magnitude by 20% of their average power but are otherwise arbitrary.

1.1 Overview of Contributions

The main focus of this thesis is the high-dimensional setting where the measurement matrix A is generated randomly and independently of the vector $\mathbf{x}$ and the measurements are corrupted by additive white Gaussian noise. The main contributions of this thesis can be summarized as follows:

- **Fundamental Limits:** In Chapters 2 and 3 we derive upper and lower bounds on the sampling rate needed using optimal recovery algorithms. While previous work has focused on exact recovery [30,46,76–78,84] or the scaling behavior for approximate recovery [2], our work gives an explicit bound on the tradeoff between the sampling rate and the fraction of detection errors. These bounds provides a sharp characterization between what can and cannot be recovered.

- **Computationally Efficient Algorithms:** In addition to our analysis of the fundamental limits, we also derive matching upper and lower bounds on the sampling rate corresponding to several computationally efficient algorithms. These include the *matched filter* (MF) and the *linear minimum mean-squared error* (LMMSE) estimator, and a class of iterative recovery algorithms known collectively as *approximate message passing* (AMP) [7,21,22,48]. One special case of AMP corresponds to an approximation of the minimum MSE estimator and another special case corresponds to ℓ_1 penalized least squares regression (known also as Basis Pursuit or the LASSO). By comparison with our fundamental bounds, we show that these estimators are near-optimal in some parameter regimes, but highly suboptimal in others.

- **Statistical Physics Heuristics:** In Chapter 2, we derive a bound corresponding the minimum MSE estimator (MMSE) using the powerful but heuristic replica method from statistical physics [35, 36, 40, 50, 55, 67]. The close correspondence between our rigorous bounds and the behavior predicted using the replica method provides important evidence in support of the (currently unproven) assumptions underlying the validly of the replica analysis.

- **Phase Transitions:** In Chapter 4, we show that the low-distortion behavior depends primarily on the relative size of the smallest nonzero entries whereas the high SNR behavior depends primarily on the computational power of the recovery algorithm and the complexity of the underlying signal class, and we precisely characterize this dependence.

- **Universality:** It is shown that a fixed recovery algorithm can be universally near optimal over a large class of practically motivated signal models.

1.2 Previous Work

A great deal of previous work has focused on the approximation of sparse vectors with respect to mean squared error (MSE) [8–12, 17, 20, 23, 32, 38, 44, 45, 68, 69]. Two particularly relevant results from this literature are [9] and [23] which show that the vector $\mathbf{x}$ can be approximated with MSE inversely proportional to the SNR using $m = O(k \log(n/k))$ samples and a quadratic program known as Basis Pursuit [12]. With a few additional assumptions on the magnitude of the smallest nonzero entries in $\mathbf{x}$, these bounds on the MSE can be translated into bounds on the detection error rate. However, the resulting bounds correspond to adversarial noise and are thus loose in general.

Another line of previous work has focused directly on the problem of exact sparsity pattern recovery [30, 46, 76–78, 84]. It is now well understood that $m = \Theta(k \log n)$ samples are both necessary and sufficient for exact recovery when the SNR is finite and there exists a fixed lower bound on the magnitude of the smallest nonzero elements [30, 76, 78]. In contrast to the scaling required for bounded MSE, this scaling says that the ratio m/k must grow without bound as the vector length n becomes large. As a consequence, exact recovery is impossible in the setting considered in this thesis, when the sparsity rate, sampling rate, and SNR are finite constants, independent of the vector length n.

The fundamental limits of sparsity pattern recovery with a nonzero detection error rate have also been investigated. For the special case where the values of the nonzero entries are identical and known (throughout the system), Aeron et al. [1, Theorem V-2] showed that $m = C \cdot k \log(n/k)$ samples are necessary and sufficient for an ML decoder where the constant C is bounded explicitly in terms of the SNR and the desired detection error rate. In the general setting where the nonzero values are unknown, Akcakaya and Tarokh [2] showed that $m = C \cdot k \log(n/k)$ samples are necessary and sufficient for a joint typicality recovery algorithm where the constant C is finite, but otherwise unspecified. (It can also be shown that this same result is implied directly by the previous work of Candès et al. [9].) An important difference between these previous results and the results in this thesis is that we give an explicit and relatively tight characterization of the constant C for a broad class of signal models.

Our analysis of linear estimation is related to work by Verdú and Shamai [75] and Tse and Hanly [72] on linear multiuser detectors. Our analysis of AMP relies heavily on recent results by Donoho et al. [21, 22] and Bayati and Montanari [7] which characterize the limiting distribution of the AMP estimate under the assumption of i.i.d. Gaussian matrices. For an overview of related work and a generalization of the algorithm, see [57]. We note that similar results for message passing algorithms have also been shown under the assumption of sparse measurement matrices with locally tree-like properties [5, 35, 56].

The bounds in this thesis are compared to predictions made by the replica method from statistical physics. This is a powerful but nonrigorous heuristic that has been used previously in the context of multi-user detection [36, 40, 50, 67], and more recently in compressed sensing [35, 55].

In conjunction with the results outlined above, another line of research has focused on the fundamental limitations of sparse signal approximation that apply to any algorithm, regardless of computational complexity. For the special case of exact recovery in the noiseless setting, these limitations have been well understood: recovery of any k-sparse vector requires exactly $m = 2k$ samples for deterministic guarantees and only $m = k + 1$ samples for almost sure guarantees [27, 33, 74], regardless of the vector length n. In both cases, recovery corresponds to an NP-hard [51] exhaustive search through all possible sparsity patterns. In Section 3.4, we address the extent to which an even smaller number of samples are needed when there exists prior knowledge about the vector $\mathbf{x}$, or when only partial recovery is needed.

Although the noiseless setting provides insight into the limitations of sparse approximation that cannot be overcome simply by increasing the SNR, consideration of the noisy setting is crucial for cases where noise is intrinsic to the problem or where real-valued numbers are subject to rate constraints. From an information-theoretic perspective, a number of works have studied the rate-distortion behavior of sparse sources [28, 29, 31, 64, 79, 80]. Most closely related to this thesis, however, is work that has addressed sparsity pattern recovery directly. An initial necessary bound based on Fano's inequality was provided by Gastpar [33] who considered Gaussian signals and deterministic sampling vectors. Necessary and sufficient scalings of (n, k, m) were given by Wainwright [76] who considered deterministic vectors, characterized by the size of their smallest nonzero elements, and Gaussian sampling vectors. Wainwright's necessary bound was strengthened in our earlier work [58], for the special case where k scales proportionally with n, and for general scalings by Fletcher et al. [30] and Wang et al. [78].

A number of papers have also addressed extensions to approximate recovery: necessary and sufficient conditions were provided by Aeron et al. [1] for the special case of discrete vectors, and by Akcakaya and Tarokh [2] and our previous work [58] for general vectors.

1.3 Notation

When possible, we use the following conventions: a random variable X is denoted using uppercase and its realization x is denoted using lowercase; a random vector $\mathbf{V}$ is denoted using boldface uppercase and its realization $\mathbf{v}$ is denoted using boldface lowercase; and a random matrix $\mathbf{M}$ is denoted using boldface uppercase and its realization M is denoted using uppercase. We use $[n]$ to denote the set $\{1, 2, \cdots, n\}$. For a collection of vectors $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_L \in \mathbb{R}^n$, the empirical joint distribution of the entries in $\{\mathbf{v}_i\}_{i \in [L]}$ is the probability measure on $\mathbb{R}^L$ that puts point mass $1/n$ at each of the n points $(v_{1,i}, v_{2,i}, \cdots, v_{L,i})$. All logarithms are taken with respect to the natural base. Unspecified constants are denoted by C and are assumed to be positive and finite.

1.4 Problem Formulation

We now provide a precise problem formulation of approximate sparsity pattern recovery. This problem formulation is central to the results in Chapters 2-4. Different but related problems are considered in Chapters 5 and 6.

Let $\mathbf{x} \in \mathbb{R}^n$ be a fixed but unknown vector and consider the noisy linear observation model given by

$$\mathbf{Y} = \mathbf{A}\mathbf{x} + \frac{1}{\sqrt{\mathsf{snr}}}\mathbf{W} \tag{1.2}$$

where $\mathbf{A}$ is a random $m \times n$ matrix, $\mathsf{snr} \in (0, \infty)$ is a fixed scalar, and $\mathbf{W} \sim \mathcal{N}(0, I_{m\times m})$ is additive white Gaussian noise. Note that if $\mathbb{E}[\|\mathbf{A}\mathbf{x}\|^2] = m$, then snr corresponds to the per-sample signal-to-noise ratio of the problem.

The problem studied in this thesis is recovery of the sparsity pattern S^* of $\mathbf{x}$ which is given by

$$S^* = \{i \in [n] : x_i \neq 0\}. \tag{1.3}$$

We assume throughout that a recovery algorithm is given the vector $\mathbf{Y}$, the matrix $\mathbf{A}$, and a parameter κ corresponding to the fraction of nonzero entries in $\mathbf{x}$. The algorithm then returns an estimate $\hat{S}$ of size $\lceil \kappa\, n \rceil$. In some cases, additional prior information about the nonzero entries of $\mathbf{x}$ is also available. We use ALG to denote a generic recovery algorithm.

1.4.1 Distortion Measure

To assess the quality of an estimate $\hat{S}$ it is important to note that there are two types of errors. A *missed detection* occurs when an element in S^* is omitted from the estimate $\hat{S}$. The missed detection rate is given by

$$\mathrm{MDR}(S^*, \hat{S}) = \frac{1}{|S^*|} \sum_{i=1}^{n} \mathbf{1}(i \in S^*, i \notin \hat{S}). \tag{1.4}$$

Conversely, a *false alarm* occurs when an element not present in S^* is included in $\hat{S}$. The false alarm rate is given by

$$\mathrm{FAR}(S^*, \hat{S}) = \frac{1}{|\hat{S}|} \sum_{i=1}^{n} \mathbf{1}(i \notin S^*, i \in \hat{S}). \tag{1.5}$$

In general, various tradeoffs between the two errors types can be considered. In this thesis, however, we focus exclusively the distortion measure $d : S^* \times \hat{S} \mapsto [0, 1]$ given by

$$d(S^*, \hat{S}) = \max\Big(\mathrm{MDR}(S^*, \hat{S}),\, \mathrm{FAR}(S^*, \hat{S})\Big). \tag{1.6}$$

This distortion measure is a metric on subsets of $[n]$.

For any distortion $D \in [0,1]$ and recovery algorithm ALG we define the error probability

$$\varepsilon_n^{(\mathrm{ALG})}(D) = \Pr[d(S^*, \hat{S}) > D] \tag{1.7}$$

where the probability is taken with respect to the distribution on the matrix $\mathbf{A}$, the noise $\mathbf{W}$ and any additional randomness used by the recovery algorithm.

1.4.2 Signal and Measurement Models

We analyze a sequence of recovery problems $\{\mathbf{x}(n), \mathbf{A}(n), \mathbf{W}(n)\}_{n \geq 1}$ indexed by the vector length n.

Signal Assumptions: We consider a subset of the following assumptions on the sequence of vectors $\mathbf{x}(n) \in \mathbb{R}^n$.

S1 *Linear Sparsity:* The number of nonzero values $k(n)$ in each vector $\mathbf{x}(n)$ obeys

$$\lim_{n \to \infty} \frac{k(n)}{n} = \kappa \tag{1.8}$$

for some *sparsity rate* $\kappa \in (0, 1/2)$.

S2 *Convergence in Distribution:* The empirical distribution of the entries in $\mathbf{x}(n)$ converges weakly to the distribution p_X of a real-valued random variable X with $\mathbb{E}[X^2] = 1$ and $\Pr[X \neq 0] = \kappa$, i.e.

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(x_i(n) \leq x) = \Pr[X \leq x] \tag{1.9}$$

for all x such that $p_X(\{x\}) = 0$.

S3 *Average Power Constraint:* The empirical second moments of the entries in $\mathbf{x}(n)$ converge to one, i.e.

$$\lim_{n \to \infty} \frac{\|\mathbf{x}(n)\|^2}{n} = 1. \tag{1.10}$$

Assumption S1 says that all but a fraction κ of the entries are equal to zero, Assumption S2 characterizes the limiting distribution of the nonzero entries, and Assumption S3 prohibits the existence of a vanishing fraction of arbitrarily large nonzero values.

Measurement Assumptions: We consider a subset of the following assumptions on the sequence of measurement matrices $\mathbf{A}(n) \in \mathbb{R}^{m(n) \times n}$.

M1 *Non-Adaptive Measurements:* The distribution on $\mathbf{A}(n)$ is independent of the vector $\mathbf{x}(n)$ and the noise $\mathbf{W}(n)$.

M2 *Finite Sampling Rate:* The number of rows $m(n)$ obeys

$$\lim_{n\to\infty} \frac{m(n)}{n} = \rho \tag{1.11}$$

for some *sampling rate* $\rho \in (0, \infty)$.

M3 *Row Normalization:* The distribution on $\mathbf{A}(n)$ is normalized such that each of the $m(n)$ rows has unit magnitude on average, i.e.

$$\mathbb{E}\Big[\|\mathbf{A}(n)\|_F^2\Big] = m(n) \tag{1.12}$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

M4 *IID Entries:* The entries of $\mathbf{A}(n)$ are i.i.d. with mean zero and variance $1/n$.

M5 *Gaussian Entries:* The entries of $\mathbf{A}(n)$ are i.i.d. Gaussian $\mathcal{N}(0, 1/n)$.

Assumptions M1-M3 are used throughout the thesis. A sampling rate $\rho < 1$ corresponds to the *compressed* sensing setting where the number of equations m is less than the number of unknown signal values n. A sampling rate $\rho = 1$ corresponds to the number of linearly independent measurements that are needed to recover an arbitrary vector $\mathbf{x}$ in the absence of any measurement noise. Assumptions M4-M5 correspond to specific distributions on $\mathbf{A}(n)$ that are used for many of the results of Chapter 2.

1.4.3 The Sampling Rate-Distortion Function

Under Assumptions S1-S3 and M1-M3, the asymptotic recovery problem is characterized by the sampling rate ρ, limiting distribution p_X, and snr.

Definition 1.1. *A distortion D is* achievable *for a fixed tuple $(\rho, p_X, \mathsf{snr})$ and recovery algorithm ALG, if there exists a sequence of measurement matrices satisfying Assumptions M1-M3 such that*

$$\lim_{n\to\infty} \varepsilon_n^{(ALG)}(D) = 0 \tag{1.13}$$

for any sequence of vectors satisfying Assumptions S1-S3.

More generally, we may also consider problems characterized by a class of limiting distributions with the same sparsity rate κ. Let $\mathcal{P}(\kappa)$ denote the class of all probability measures obeying the conditions of Assumption S2, i.e.

$$\mathcal{P}(\kappa) = \Big\{p_X : p_X(\{0\}) = 1-\kappa,\ \textstyle\int x^2 p_X(dx) = 1\Big\}, \tag{1.14}$$

and let $\mathcal{P}_X$ be a subset of $\mathcal{P}(\kappa)$.

Definition 1.2. *A distortion D is* achievable *for a fixed tuple $(\rho, \mathcal{P}_X, \mathsf{snr})$ and recovery algorithm ALG, if there exists a sequence of measurement matrices satisfying Assumptions M1-M3 such that*

$$\lim_{n\to\infty} \varepsilon_n^{(ALG)}(D) = 0 \tag{1.15}$$

for any sequence of vectors satisfying Assumptions S1-S3 for some distribution $p_X \in \mathcal{P}_X$.

We emphasize that the recovery algorithm in Definition 1.2 is fixed and thus cannot be a function of the limiting distribution realized by an individual sequence of problems. It may however be optimized as a function of the class $\mathcal{P}_X$, thus attaining the minimax risk of the recovery problem.

Definition 1.3. *For a fixed tuple $(D, \mathcal{P}_X, \mathsf{snr})$ and recovery algorithm ALG the sampling rate-distortion function $\rho^{(ALG)}(D, \mathcal{P}_X, \mathsf{snr})$ is given by*

$$\rho^{(ALG)}(D, \mathcal{P}_X, \mathsf{snr}) = \inf\{\rho \geq 0 \,:\, D \text{ is achievable}\}. \tag{1.16}$$

The sampling rate-distortion function corresponding to the optimal recovery algorithm is denoted by $\rho^*(D, \mathcal{P}_X, \mathsf{snr})$.

To lighten the notation, we will denote the sampling rate-distortion function using $\rho^{(\mathrm{ALG})}$ where the dependence on the tuple $(D, \mathcal{P}_X, \mathsf{snr})$ is implicit.

1.4.4 Approximately Sparse Signal Models

The problem formulation given in the previous sections assumes that a large fraction of the entries in $\mathbf{x}$ are exactly equal to zero. More realistically though, it may be the case that many of these entries are only *approximately* equal to zero. This may occur, for instance, if a sparse vector is corrupted by a small amount of noise prior to being measured. In these cases, the vector $\mathbf{x}$ is not, strictly speaking, sparse, but recovery of the locations of the "significant" entries is still a meaningful task.

With these settings in mind, all of the bounds presented in Chapter 2 are first proved with respect to a *relaxed* sparsity pattern recovery task in which the goal is to recover the locations of the $\lceil \kappa\, n \rceil$ largest entries in $\mathbf{x}$. To prove achievability for this task, we assume that the weak converge of Assumption S2 holds (specifically the fact that all but a fraction κ of the entries in $\mathbf{x}$ are tending to zero as n becomes large) but do not require the strict sparsity constraint of Assumption S1.

The relaxed sparsity pattern recovery task is defined as follows. For any vector $\mathbf{x}$ and sparsity rate κ, let $\tilde{S}$ be a drawn uniformly at random from all subsets of $[n]$ of size $k = \lceil \kappa\, n \rceil$ obeying

$$\min_{i\in\tilde{S}} |x_i| \geq \max_{i\in[n]\setminus\tilde{S}} |x_i|. \tag{1.17}$$

The set $\tilde{S}$ corresponds to the k largest entries in $\mathbf{x}$ and is uniquely defined whenever the k'th largest entry of $\mathbf{x}$ is unique. For any distortion D and recovery algorithm ALG we define the relaxed sparsity pattern recovery error probability

$$\tilde{\varepsilon}_n^{(\mathrm{ALG})}(D) = \Pr[d(\tilde{S}, \hat{S}) > D] \tag{1.18}$$

where the probability is taken with respect to the distribution on $\tilde{S}$, the matrix $\mathbf{A}$, the noise $\mathbf{W}$, and any additional randomness in the recovery algorithm. The definition of achievability with respect to the error probability $\tilde{\varepsilon}_n(D)$ is exactly the same as for the error probability $\varepsilon_n(D)$ except that the strict sparsity of Assumption S1 is not required.

The following result shows that under the additional constraint of Assumption S1, achievability of relaxed sparsity pattern recovery implies achievability of the sparsity pattern in the strict sense.

Lemma 1.1. *Under Assumption S1,*

$$\lim_{n\to\infty} \left| d(\tilde{S}, \hat{S}) - d(S^*, \hat{S}) \right| = 0. \tag{1.19}$$

Proof. Two applications of the triangle inequality gives

$$\left| d(\tilde{S}, \hat{S}) - d(S^*, \hat{S}) \right| \leq d(\tilde{S}, S^*).$$

By the definition of $\tilde{S}$, it follows straightforwardly that $d(\tilde{S}, S^*) \to 0$ for any sequence of vectors obeying Assumption S1. □

Chapter 2

Upper Bounds for Algorithms

This chapter gives bounds on the sampling rate-distortion function $\rho^{(\mathrm{ALG})}$ for several different recovery algorithms. Each of the algorithms follows the same basic approach which is illustrated in Fig. 2.1 and consists of the following two stages:

- *Vector Estimation:* The first stage of recovery produces a random estimate $\hat{\mathbf{X}}$ of the unknown vector $\mathbf{x}$ based on the tuple $(\mathbf{Y}, \mathbf{A}, \kappa)$.
- *Componentwise Thresholding:* The second stage of recovery generates an estimate $\hat{S}$ of the unknown sparsity pattern S^* by thresholding the estimate $\hat{\mathbf{X}}$ generated in the first stage:

$$\hat{S} = \Big\{ i \in [n] : |\hat{X}_i| \geq T \Big\}.$$

The threshold T in the second stage provides a tradeoff between the two kinds of recovery errors: missed detections and false alarms. Throughout this chapter, we will assume that that T is chosen as a function of $(\hat{\mathbf{X}}, \kappa)$ such that the estimated sparsity pattern $\hat{S}$ has exactly $k = \lceil \kappa n \rceil$ elements. In practice, this is achieved by thresholding with the magnitude of the k'th largest entry in $\hat{\mathbf{X}}$, and using additional randomness to break ties whenever the k'th largest magnitude is not unique.

Conceptually, it is useful to think of the estimate $\hat{\mathbf{X}}$ generated in the first stage as a direct observation of the original signal that has been corrupted by additive noise, that is

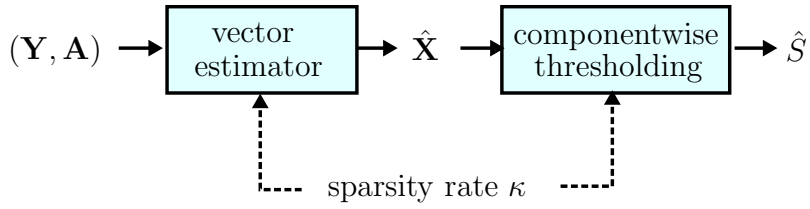

Figure 2.1: Illustration of the two-stage sparsity pattern recovery algorithm.

Table 2.1: Overview of the Sampling Rate-Distortion Bounds in Chapter 2

Recovery Algorithm			Bounds			
Vector Estimator	Parameters	Comp. Efficient	Result	Matrix Assump.	Unproven Assump.	Tight
ML	κ	no	Theorem 2.1	Gaussian	none	no
MF	κ	yes	Theorem 2.2	**i.i.d.**	none	yes
LMMSE	κ, snr	yes	Theorem 2.3	Gaussian	none	yes
AMP-MMSE	κ, snr, p_X	yes	Theorem 2.5	Gaussian	none	yes
AMP-ST	κ, snr, α	yes	Theorem 2.6	Gaussian	none	yes
MMSE	κ, snr, p_X	no	Theorem 2.7	**i.i.d.**	Replica Sym.	yes

we can write

$$\hat{\mathbf{X}} = \mathbf{x} + \tilde{\mathbf{W}}$$

where $\tilde{\mathbf{W}}$ is a vector of errors. Along the same lines, the componentwise thresholding in the second stage may be viewed as n independent hypothesis tests under the idealized assumption that the entries of $\tilde{\mathbf{W}}$ are i.i.d. and symmetric about the origin.

The main difference between the algorithms studied in this chapter is the vector estimator used in the first stage of recovery. In the following subsections, we give bounds on the sampling rate-distortion function corresponding to the maximum likelihood estimator, two different linear estimators (the matched filter and the MMSE), a class of estimators based on approximate message passing, and the MMSE estimator. Our results are summarized in Table 2.1. Analysis and illustrations are given in Chapter 4.

2.1 Maximum Likelihood

We begin with the method of *maximum likelihood* (ML). Conditioned on the realization of the matrix $\mathbf{A} = A$, the measurements $\mathbf{Y}$ have a multivariate Gaussian distribution with mean $A\mathbf{x}$ and covariance $\mathsf{snr}^{-1} I_{m\times m}$. Therefore, the ML estimate of sparsity $k = \lceil \kappa n \rceil$ is given by

$$\hat{\mathbf{x}}^{(\mathrm{ML})} = \arg \min_{\tilde{\mathbf{x}} \in \mathbb{R}^n \,:\, \|\tilde{\mathbf{x}}\|_0 = k} \|\mathbf{y} - A\tilde{\mathbf{x}}\| \tag{2.1}$$

where $\|\tilde{\mathbf{x}}\|_0$ denotes the number of nonzero entries in $\tilde{\mathbf{x}}$. If the minimizer of (2.1) is not unique, we will assume that the sparsity pattern estimate $\hat{S}$ in the second stage of the recovery algorithm is drawn uniformly at random from the set

$$\Big\{ S \,:\, S \text{ is the sparsity pattern of a minimizer of (2.1)} \Big\}.$$

This estimator has been studied previously for the task of exact sparsity pattern recovery by Wainwright [76] and Fletcher et al. [30].

Before we present our main result, two more definitions are needed. First, we define

$$\mathcal{H}(D;\kappa) = \kappa H_b(D) + (1-\kappa) H_b\Big(\frac{\kappa D}{1-\kappa}\Big) \tag{2.2}$$

where $H_b(p) = -p\log p - (1-p)\log(1-p)$ is the binary entropy function. In Lemma **??**, it is shown that the metric entropy rate for a sequence of sparsity patterns with sparsity rate κ under the distortion metric (1.6) is given by $H_b(\kappa) - \mathcal{H}(D;\kappa)$ for any $D \le 1-\kappa$.

Also, we define

$$P(D;p_X) = \int_0^\infty \Big(\Pr[X^2 > u] - (1-D)\kappa \Big)^+ du \tag{2.3}$$

where $(\cdot)^+ = \max(\cdot, 0)$. This function corresponds to the average power of the smallest fraction D of nonzero entries. It is a continuous and monotonically increasing function of D, with $P(0;p_X) = 0$ and $P(1;p_X) = 1$ for any $p_X \in \mathcal{P}(\kappa)$.

Our first result gives an upper bound on the sampling rate-distortion function corresponding to the ML estimator. The proof is given in Section 2.5.

Theorem 2.1. *Under Assumptions S1-S2 and M1-M5, a distortion D is achievable for the tuple $(\rho, p_X, \mathsf{snr})$ using the ML estimator if $\rho > \rho^{(ML\text{-}UB)}$ where*

$$\rho^{(ML\text{-}UB)} = \kappa + \max_{\tilde{D}\in[D,1]} \Lambda(\tilde{D}; p_X, \mathsf{snr}), \tag{2.4}$$

with $\Lambda(D;p_X,\mathsf{snr})$ given by

$$\Lambda(D;p_X,\mathsf{snr}) = \min\Big\{\Lambda_1(D;p_X,\mathsf{snr}), \Lambda_2(D;p_X,\mathsf{snr})\Big\} \tag{2.5}$$

where

$$\Lambda_1(D;p_X,\mathsf{snr}) = \frac{2\mathcal{H}(D;\kappa)}{\log(1+P(D;p_X)\,\mathsf{snr}) + (1+P(D;p_X)\,\mathsf{snr})^{-1} - 1}$$

$$\Lambda_2(D;p_X,\mathsf{snr}) = \min_{\theta,\mu\in(0,1)} \max\left(\frac{2\mathcal{H}(D;\kappa)}{\log(1+\frac{1}{4}(1-\theta)^2 P(D;p_X)\,\mathsf{snr})}, \frac{2\mathcal{H}(D;\kappa) - D\kappa\log(1-\mu^2)}{\log(1+\mu\theta P(D;p_X)\,\mathsf{snr})} \right).$$

Moreover, for any $\rho > \rho^{(ML\text{-}UB)}$ the error probability $\varepsilon_n^{(ML)}(D)$ decays at least exponentially rapidly with n, i.e. there exists a constant C such that

$$\varepsilon_n^{(ML)}(D) \le \exp(-C\,n). \tag{2.6}$$

Remark 2.1. *Theorem 2.1 does* not *require the convergence of the empirical second moments given in Assumption S3.*

Theorem 2.1 is a significant improvement over previous results in several respects. First, it applies generally to any distribution p_X. Second, the bound is given explicitly in terms of the problem parameters and is finite for any nonzero distortion D. Finally, as we will show in Sections 4.8.1, the behavior of the bound, in a scaling sense with respect to the SNR and distortion D, is optimal for a large class of distributions.

Corollary 2.1. *The statement of Theorem 2.1 holds if the function* $\Lambda(D; p_X, \mathsf{snr})$ *is replaced with any of the following upper bounds:*

$$\tilde{\Lambda}_1(D; p_X, \mathsf{snr}) = \frac{4\mathcal{H}(D;\kappa)}{\log\left(1 + \left[P(D; p_X)\,\mathsf{snr}/e\right]^2\right)} \tag{2.7}$$

$$\tilde{\Lambda}_2(D; p_X, \mathsf{snr}) = \frac{2\mathcal{H}(D;\kappa) + 2\log(5/3)\kappa D}{\log\left(1 + (4/25)P(D; p_X)\,\mathsf{snr}\right)} \tag{2.8}$$

$$\tilde{\Lambda}_3(D; p_X, \mathsf{snr}) = \min_{i\in\{1,2\}} \tilde{\Lambda}_i(D; p_X, \mathsf{snr}). \tag{2.9}$$

Proof. The bound $\tilde{\Lambda}_1(D; p_x, \mathsf{snr})$ follows from the first term in (2.5) and the simple fact that $\log(1+x) - x/(1+x) \geq (1/2)\log(1+x^2/e^2)$ for all $x \geq 0$. The bound $\tilde{\Lambda}_2(D; p_x, \mathsf{snr})$ follows from the second term in (2.5) evaluated with $\mu = 4/5$ and $\theta = 1/5$. □

2.2 Linear Estimation

Next, we consider two different linear estimators. The *matched filter* (MF) estimate is given by

$$\hat{\mathbf{x}}^{(\mathrm{MF})} = \left(\tfrac{n}{m}\right) A^T \mathbf{y} \tag{2.10}$$

and the *linear minimum mean-squared error* (LMMSE) estimate is given by

$$\hat{\mathbf{x}}^{(\mathrm{LMMSE})} = (A^T A + \mathsf{snr}\, I_{n\times n})^{-1} A^T \mathbf{y}. \tag{2.11}$$

These estimators are appealing in practice due to their low computational complexity. Their performance has been studied extensively in the context of multiuser detection with random spreading (see e.g. [72, 75]). More recently, the use of the matched filter for the task of sparsity pattern recovery was investigated by Fletcher et al. [30] and our previous work [61].

To characterize the behavior of the MF and LMMSE algorithms in the high-dimensional setting, it is useful to introduce a scalar equivalent model of the vector observation model given in (1.2).

Definition 2.1. *The* scalar equivalent model *of* (1.2) *is given by*

$$Z = X + \sigma W \tag{2.12}$$

where $X \sim p_X$ *and* $W \sim \mathcal{N}(0,1)$ *are independent and* $\sigma^2 \in (0, \infty)$ *is a fixed parameter called the noise power.*

In the context of the scalar model, the problem of support recovery is to determine whether or not X is equal to zero. Let $S = \mathbf{1}(X \neq 0)$ be the indicator of this event and let $\hat{S}$ be an estimate of the form $\hat{S} = \mathbf{1}(|Z| > t)$. Then, the detection error probability corresponding to the distortion measure defined in Section 1.4.1 is given by

$$p_D(t) = \max\Big(\Pr[\hat{S} = 0|S = 1], \Pr[S = 0|\hat{S} = 1]\Big). \tag{2.13}$$

We define

$$D_{\text{awgn}}(\sigma^2; p_X) = \min_t p_D(t) \tag{2.14}$$

to be a mapping between the noise power σ^2 and the minimal detection error probability $p_D(t)$ achieved by $\hat{S}$. We also define

$$\sigma^2_{\text{awgn}}(D; p_X) = \sup\{\sigma^2 \geq 0 \,:\, D_{\text{awgn}}(\sigma^2; p_X) \leq D\} \tag{2.15}$$

to be the inverse mapping. Here, we use the subscript "awgn" to emphasize the fact that this error probability corresponds to additive noise W that is Gaussian and independent of X.

The following results give an explicit expression for the sampling rate-distortion function of the MF and LMMSE recovery algorithms. Their proofs are given in Appendices 2.6.2 and 2.6.3 respectively.

Theorem 2.2. *Under Assumptions S1-S3 and M1-M4, the sampling rate-distortion function corresponding to the MF estimator is given by*

$$\rho^{(MF)} = \frac{1}{\sigma^2\,\mathsf{snr}} + \frac{1}{\sigma^2} \tag{2.16}$$

where $\sigma^2 = \sigma^2_{awgn}(D; p_X)$.

Remark 2.2. *Theorem 2.2 does* not *require the measurement matrix* $\mathbf{A}(n)$ *to be Gaussian.*

Theorem 2.3. *Under Assumptions S1-S3 and M1-M5, the sampling rate-distortion function corresponding to the LMMSE estimator is given by*

$$\rho^{(LMMSE)} = \frac{1}{\sigma^2\,\mathsf{snr}} + \frac{1}{1+\sigma^2} \tag{2.17}$$

where $\sigma^2 = \sigma^2_{awgn}(D; p_X)$.

Recall that our definition of achievability says that the probability that the distortion $d(S^*, \hat{S})$ exceeds a threshold D must tend to zero as n becomes large. For the MF and LMMSE estimators, convergence of the expected distortion $\mathbb{E}[d(S^*, \hat{S})]$ can be established straightforwardly using results in [75] and [72]. Therefore, the key contribution of Theorems 2.2 and 2.3 is to show that this convergence holds also in probability. For the MF estimator, this is achieved using a general decoupling result which applies generally for any i.i.d. distribution on the measurement matrix. For the LMMSE estimator, we use the fact that the LMMSE can be computed using the AMP algorithm discussed in the next section.

2.3 Approximate Message Passing

We now consider estimation using *approximate message passing* (AMP) [21]. The AMP algorithm is characterized in terms of a scalar de-noising function $\eta(z, \sigma^2)$ which is assumed to be Lipschitz continuous with respect to its first argument and continuous with respect to its second argument. Starting with initial conditions $\mathbf{x}^0 = \mathbf{0}_{n\times 1}$, $\mathbf{u}^0 = (\frac{n}{m})\mathbf{y}$ and $\hat{\sigma}_0^2 = (\mathsf{snr}^{-1} + 1)/\rho$, the algorithm proceeds for iterations $t = 1, 2, \cdots$ according to

$$\mathbf{x}^t = \eta\Big(A^T\mathbf{u}^{t-1} + \mathbf{x}^{t-1}, \hat{\sigma}_{t-1}^2\Big) \tag{2.18}$$

$$\mathbf{u}^t = \Big(\tfrac{n}{m}\Big)\Big[\mathbf{y} - A\mathbf{x}^t + \mathbf{u}^{t-1}\frac{1}{n}\sum_{i=1}^{n}\eta'\Big(\Big(A^T\mathbf{u}^{t-1} + \mathbf{x}^{t-1}\Big)_i, \hat{\sigma}_{t-1}^2\Big)\Big] \tag{2.19}$$

$$\hat{\sigma}_t^2 = \frac{1}{n}\|\mathbf{u}^t\|^2, \tag{2.20}$$

where $\eta'(z, \sigma^2)$ denotes the partial derivative of $\eta(z, \sigma^2)$ with respect to z, and, for any vector $\mathbf{z}$, $\eta(\mathbf{z}, \sigma^2)$ denotes the vector obtained by applying the function $\eta(z, \sigma^2)$ componentwise.

The AMP algorithm is said to succeed if the tuple $(\mathbf{x}^t, \mathbf{u}^t, \hat{\sigma}_t^2)$ converges to a fixed point $(\mathbf{x}^\infty, \mathbf{u}^\infty, \hat{\sigma}_\infty^2)$. Various stability assumptions guaranteeing convergence of the algorithm are discussed in [21, 22]. In some cases, the rate of convergence is exponential in the number of iterations.

Remark 2.3. *Our update equations for the AMP algorithm differ slightly from those given in [7, 21, 22]. This difference is due to the fact that this thesis considers row normalization of the measurement matrix (Assumption M3) whereas the previous work considers column normalization.*

Conceptually, it is useful to think of the vector $\mathbf{x}^t$, generated in the t'th iteration of the AMP algorithm, as a noisy version of the original vector $\mathbf{x}$ that has been passed through the scalar de-noising function $\eta(\cdot, \hat{\sigma}_{t-1}^2)$. More specifically, we can write

$$\mathbf{x}^t = \eta(\mathbf{x} + \tilde{\mathbf{w}}^{t-1}; \hat{\sigma}_{t-1}^2) \tag{2.21}$$

where

$$\tilde{\mathbf{w}}^{t-1} = A^T\mathbf{u}^{t-1} + \mathbf{x}^{t-1} - \mathbf{x} \tag{2.22}$$

is a vector of errors.

In [21, 22], it is shown, both heuristically and empirically, that, under Assumptions S1-S3 and M1-M5, the error vector $\mathbf{w}^{t-1}$ defined in (2.22) behaves similarly to additive white Gaussian noise with mean zero and variance $\hat{\sigma}_{t-1}^2$. A precise statement of this behavior, corresponding to the empirical marginal distribution of the tuple $(\mathbf{x}, \mathbf{x}^t, \tilde{\mathbf{w}}^t)$, is proved in ensuing work by Bayati and Montanari [7]. See Section 2.6 for more details.

At this point, we are faced with the following question: based on the output $(\mathbf{x}^\infty, \mathbf{u}^\infty, \hat{\sigma}^2_\infty)$ of the AMP algorithm, what should we choose as an estimate $\hat{\mathbf{x}}$ of the unknown vector $\mathbf{x}$? In previous work, where the primary objective is to minimize the MSE, the output $\hat{\mathbf{x}}^\infty$ is used as an estimate of $\mathbf{x}$ (see e.g. [6]). The main reason for using this estimate is that the function $\eta(\cdot, \sigma^2)$ provides a scalar de-noising step that reduces the effect of the additive error $\tilde{\mathbf{w}}$.

In this thesis, however, our primary objective to is generate an estimate of $\mathbf{x}$ that leads to an accurate estimate of the sparsity pattern in the second stage of estimation. As such, the final scalar de-noising step is unnecessary, and potentially counterproductive. To see why, note that the componentwise thresholding in the second stage of recovery depends entirely on the relative magnitudes of the entries in $\hat{\mathbf{x}}$. If the denoiser does not preserve the ranking of these magnitudes (e.g. if many nonzero values are mapped to zero), then relevant information about the sparsity pattern is lost.

Accordingly, we use the vector estimate given by

$$\hat{\mathbf{x}}^{(\mathrm{AMP})} = A^T \mathbf{u}^\infty + \mathbf{x}^\infty. \tag{2.23}$$

Since the AMP output $(\mathbf{x}^\infty, \mathbf{u}^\infty, \hat{\sigma}^2_\infty)$ satisfies the fixed point equation

$$\mathbf{x}^\infty = \eta(A^T \mathbf{u}^\infty + \mathbf{x}^\infty, \hat{\sigma}^2_\infty),$$

we see that our estimate corresponds directly to the signal-plus-noise estimate $\mathbf{x} + \tilde{\mathbf{w}}^\infty$ prior to the scalar de-noising.

To characterize the behavior of AMP in the high-dimensional setting, we return to the scalar equivalent model given in Definition 2.1. We define the scalar mean-squared error function

$$\mathsf{mse}(\sigma^2; p_X, \eta) = \mathbb{E}\Big[\Big|X - \eta(X + \sigma W, \sigma^2)\Big|^2\Big] \tag{2.24}$$

where $X \sim p_X$ and $W \sim \mathcal{N}(0,1)$ are independent, and let $\{\sigma_t^2\}_{t \geq 1}$ be a sequence of noise powers defined by the recursion

$$\sigma_t^2 = \frac{\mathsf{snr}^{-1} + \mathsf{mse}(\sigma_{t-1}^2; p_X, \eta)}{\rho} \tag{2.25}$$

where $\sigma_0^2 = (\mathsf{snr}^{-1} + 1)/\rho$. This recursion is referred to as *state evolution* [21].

The following result shows that the distortion corresponding to the AMP estimate is characterized by the state evolution recursion. In Section 2.6.4, it is shown how this result follows straightforwardly from recent work of Bayati and Montanari [7].

Theorem 2.4. *Suppose that the noise powers defined by the state evolution recursion* (2.25) *converge to a finite limit*

$$\sigma_\infty^2 = \lim_{t \to \infty} \sigma_t^2. \tag{2.26}$$

Then, under Assumptions S1-S3 and M1-M5, the distortion $d(S^*, \hat{S})$ corresponding to the AMP estimator converges in probability as $n \to \infty$ to the limit $D_{awgn}(\sigma_\infty^2; p_X)$.

Remark 2.4. *We note that the limiting noise power σ_∞^2 is a function of the tuple $(\rho, p_X, \mathsf{snr})$ and the function $\eta(z, \sigma^2)$. In some cases, it is possible that σ_∞^2 is an increasing function of ρ, and thus increasing the sampling rate increases the distortion.*

In the following subsections, two special cases of the AMP estimator are considered.

2.3.1 Optimized AMP

If the limiting distribution p_X is known, then the limiting noise power σ_∞^2 is minimized when $\eta(z, \sigma^2)$ is given by the conditional expectation

$$\eta^{(\text{MMSE})}(z, \sigma^2; p_X) = \mathbb{E}[X | X + \sigma W = z] \tag{2.27}$$

corresponding to the distribution p_X. We will refer to this version of the AMP algorithm as AMP-MMSE, and we define the corresponding mean-squared error function

$$\mathsf{mmse}(\sigma^2; p_X) = \mathbb{E}\Big[\Big|X - \mathbb{E}[X | X + \sigma W]\Big|^2\Big]. \tag{2.28}$$

Theorem 2.5. *Under Assumptions S1-S3 and M1-M5, the sampling rate-distortion function corresponding to the AMP-MMSE estimator is given by*

$$\rho^{(\textit{AMP-MMSE})} = \sup_{\tau \geq \sigma^2} \left\{ \frac{\mathsf{snr}^{-1} + \mathsf{mmse}(\tau; p_X)}{\tau} \right\} \tag{2.29}$$

where $\sigma^2 = \sigma_{awgn}^2(D; p_X)$.

Proof. By the definition of the MMSE, we have $\mathsf{mmse}(\sigma^2; p_X) < \mathbb{E}[X^2] = 1$ for all $\sigma^2 < \infty$. Therefore, any solution σ^2 to the fixed point equation

$$\sigma^2 = \frac{\mathsf{snr}^{-1} + \mathsf{mmse}(\sigma^2; p_X)}{\rho} \tag{2.30}$$

is strictly less than the initial noise power σ_0^2. Since $\mathsf{mmse}(\sigma^2; p_X)$ is a strictly decreasing function of σ^2, it thus follows that the limit σ_∞^2 always exists and is given by the largest solution to (2.30), i.e.

$$\sigma_\infty^2 = \sup \left\{ \tau \geq 0 \, : \rho = \frac{\mathsf{snr}^{-1} + \mathsf{mmse}(\tau; p_X)}{\tau} \right\}. \tag{2.31}$$

Since the right hand side of (2.31) is a strictly decreasing function of ρ, Theorem 2.5 follows directly from Theorem 2.4 and the definition of the sampling rate-distortion function. □

It is important to note that the AMP-MMSE estimate is a function of the distribution p_X. If this distribution is unknown and the estimate is made using a postulated distribution that differs from the true one, then the performance of the algorithm could be highly suboptimal.

2.3.2 Soft Thresholding

Another special case of the AMP algorithm is when $\eta(z, \sigma^2)$ is given by the soft thresholding function

$$\eta^{(\mathrm{ST})}(z, \sigma^2; \alpha) = \begin{cases} z + \alpha\sigma, & \text{if } z < -\alpha\sigma \\ 0, & \text{if } |z| \leq \alpha\sigma \\ z - \alpha\sigma, & \text{if } z \geq \alpha\sigma \end{cases} \tag{2.32}$$

for some threshold $\alpha \geq 0$. We will refer to this algorithm as AMP-ST.

Remark 2.5. *It is argued in [22] and shown rigorously in [6] that, for a fixed set (p_X, snr), the behavior of AMP-ST is equivalent to that of LASSO [68] under an appropriate calibration between the threshold α and the regularization parameter of LASSO.*

To characterize the behavior of AMP-ST, we follow the steps outlined by Donoho et al. [22] and define the noise sensitivity

$$\mathcal{M}(\sigma^2, \alpha; p_X) = \frac{\mathsf{mse}(\sigma^2; p_X, \eta^{(\mathrm{ST})})}{\sigma^2}. \tag{2.33}$$

Theorem 2.6. *Under Assumptions S1-S3 and M1-M5, the sampling rate-distortion function corresponding to the AMP-ST estimator is given by*

$$\rho^{(\textit{AMP-ST})} = \frac{1}{\sigma^2 \mathsf{snr}} + \mathcal{M}(\sigma^2, \alpha; p_X) \tag{2.34}$$

where $\sigma^2 = \sigma^2_{\textit{awgn}}(D; p_X)$.

Proof. This result is an immediate consequence of Theorem 2.4 and [22, Lemma 4.1] which shows that σ^2_∞ exists and is given by the unique solution to the fixed point equation

$$\rho = \frac{1}{\sigma^2_\infty \, \mathsf{snr}} + \mathcal{M}(\sigma^2_\infty, \alpha; p_X). \tag{2.35}$$

□

We note that Theorem 2.6 can be used to find the optimal value for the soft-thresholding parameter α. If, for example, the goal is to minimize the sampling rate ρ as a function of the tuple (D, p_X, snr), then the optimal value of α is given by the minimizer of $\mathcal{M}(\sigma^2, \alpha; p_X)$. Conversely, if the goal is to minimize the distortion D as a function of the tuple $(\rho, p_X, \mathsf{snr})$, then the optimal value of α is one that minimizes the value of σ^2_∞ in the fixed point equation (2.35).

We emphasize that the soft-thresholding function is, in general, suboptimal for a given distribution p_X (recall that the optimal version of AMP is given by AMP-MMSE). The

main reason that we study soft-thresholding is to deal with settings where the distribution p_X is unknown. In Section 4.9, it is shown how the function $\mathcal{M}(\sigma^2, \alpha; p_X)$ can be upper bounded uniformly over the class of distributions $\mathcal{P}_\kappa$, and how combining this upper bound with Theorem 2.6 gives bounds on the sampling rate-distortion function that hold uniformly over any class of distributions $\mathcal{P}_X \subset \mathcal{P}(\kappa)$.

2.4 MMSE via the Replica Method

Lastly, we consider the performance of the *minimum mean-squared error* (MMSE) estimator. For a known distribution p_X, this estimator is given by the conditional expectation

$$\mathbf{x}^{(\mathrm{MMSE})} = \mathbb{E}[\mathbf{X} | A\mathbf{X} + \mathsf{snr}^{-1/2}\mathbf{W} = \mathbf{y}], \tag{2.36}$$

where the entries of $\mathbf{X}$ are i.i.d. p_X.

To analyze the behavior of the MMSE estimator, we develop a result based on the powerful but heuristic *replica method* from statistical physics. This method was developed originally in the context of spin glasses [25] and has been applied to the vector estimation problem studied in this thesis by a series of recent papers [35, 36, 40, 50, 55, 67].

In the replica analysis, the unknown vector is modeled as a random vector $\mathbf{X}$ whose entries are i.i.d. p_X. Accordingly, each realization of the measurement matrix $\mathbf{A} = A$, induces a joint probability measure on the random input-output pair $(\mathbf{X}, \mathbf{Y})$, or equivalently on the random input-estimate pair $(\mathbf{X}, \hat{\mathbf{X}})$. At this point, the key argument exploited by the replica method is that, due to a certain type of "replica symmetry" in the problem, the joint probability measure on $(\mathbf{X}, \hat{\mathbf{X}})$ behaves similarly for all typical realizations of the measurement matrix $\mathbf{A}$ in the high-dimensional setting. Based on this assumption, it can then be argued that the marginal joint distribution on the entries in $(\mathbf{X}, \hat{\mathbf{X}})$ converges to a nonrandom limit, characterized by the tuple $(\rho, p_X, \mathsf{snr})$.

A detailed explanation of the replica analysis is beyond the scope of this thesis. The assumptions needed for our results are summarized below.

Replica Analysis Assumptions: The key assumptions underlying the replica analysis are stated explicitly by Guo and Verdú in [36]. A concise summary can also be found in [55, Appendix A]. Two assumptions that are used—and generally accepted throughout the literature—are the validity the "replica trick" and the self averaging property of a certain function defined on the random matrix $\mathbf{A}$. A further assumption that is also required is that of *replica symmetry*. This last assumption is problematic, however, since it is known that there are cases where it does not hold, and there is currently no test to determine whether or not it holds in the setting of this thesis.

The following result characterizes the sampling rate-distortion function corresponding to the MMSE estimator under the condition that the replica assumptions are valid. The proof is given in chapter 2.6.5.

Theorem 2.7. *Assume that the replica analysis assumptions hold. Under Assumptions S1-S3 and M1-M4, the distortion $d(S^*, \hat{S})$ corresponding to the MMSE estimator converges in probability as $n \to \infty$ to the limit $D_{awgn}(\tau^*; p_X)$ where*

$$\tau^* = \arg\min_{\tau > 0} \left\{ \rho \log \tau + \frac{1}{\tau \, \mathsf{snr}} + 2I(X; X + \sqrt{\tau} W) \right\} \tag{2.37}$$

with $X \sim p_X$ and $W \sim \mathcal{N}(0,1)$ independent.

We emphasize that a key difference between Theorem 2.7 and the previous bounds in this chapter is that the replica analysis assumptions on which it is based are currently unproven. In the context of the recovery problem outlined in this paper, this means that Theorem 2.7 provides only a *heuristic prediction* for the true behavior of the MMSE estimator. The validity of this prediction for the setting of the paper depends entirely on the validity of the replica assumptions.

In the next section, we will see that there are many parameter regimes in which the replica prediction for the MMSE estimator is tightly sandwiched between the rigorous upper bounds given earlier in this paper and the information-theoretic lower bounds in Chapter 3. Thus, beyond the context of sparsity pattern recovery, a significant contribution of this paper is that we provide strong evidence in support of the replica analysis assumptions.

Remark 2.6. *One interesting implication of Theorem 2.7 is that the AMP-MMSE estimate is equivalent to the MMSE estimate whenever the noise power τ^* defined in* (2.37) *is equal to the limit σ^2_∞ defined in* (2.31)*. This suggests that the MMSE estimate can be computed efficiently in some problem regimes.*

Finally, it is important to note that the MMSE estimator is a function of the limiting distribution p_X. If this distribution is unknown and the estimate is made using a postulated distribution that differs from the true one, then the performance could be highly suboptimal. Using further results developed in [36] it is possible to characterize the sampling rate in terms of an arbitrary postulated prior and true limiting distribution. Such analysis, however, is beyond the scope of this paper.

2.5 Proof of ML Upper Bound

Following the discussion in Section 1.4.4, we first prove achievability with respect to relaxed sparsity pattern recovery.

Theorem 2.8. *Under Assumptions S2 and M1-M5, the statement of Theorem 2.1 holds with respect to the relaxed sparsity pattern recovery error probability $\tilde{\varepsilon}_n(D)$ defined in* (1.18)*.*

Combining Theorem 2.8 with Lemma 1.1 and the fact that $\rho^{(\text{ML-UB})}$ is a continuous and monotonically decreasing function of D completes the proof of Theorem 2.1.

The remainder of this section is dedicated to the proof of Theorem 2.8. We begin with a general bound for the non-asymptotic setting in Section 2.5.1 and then extend this bound to the asymptotic setting in Section 2.5.2.

Throughout the proof we use $\mathcal{S}_k^n$ to denote the set of all subsets of $[n]$ of size k, and for any set $\mathbf{s} \subset [n]$, we use $\mathbf{s}^c$ to denote its complement $[n] \backslash \mathbf{s}$.

2.5.1 A Non-Asymptotic Bound

Consider the measurement model given in (1.2) where $\mathbf{x} \in \mathbb{R}^n$ is an arbitrary vector whose true sparsity is unknown. For a given parameter κ, let $k = \lceil \kappa\, n \rceil$, let $\tilde{S}$ be drawn uniformly at random from all subsets of $[n]$ of size $k = \lceil \kappa\, n \rceil$ obeying (1.17), and let $\hat{S}$ be the output of the ML recovery algorithm.

Also, for each integer $b \in \{0, 1, \cdots, k\}$, define

$$M(b) = \min_{\mathbf{s} \in \mathcal{S}_k^n : |\mathbf{s} \backslash \tilde{\mathbf{s}}| = b} \frac{1}{n} \|\mathbf{x}_{\mathbf{s}^c}\|^2 \cdot \mathsf{snr}. \tag{2.38}$$

By the definition of $\tilde{S}$, it is straightforward to see that $M(b)$ is defined uniquely by $\mathbf{x}$ and b (i.e. it does not depend on the realization of $\tilde{\mathbf{s}}$).

The following result gives an upper bound on $\tilde{\varepsilon}_n(D)$ that depends only on the distortion D, the dimensions n, m, k, and the function $M(b)$.

Lemma 2.1. *If the entries of the measurement matrix* $\mathbf{A}$ *are i.i.d.* $\mathcal{N}(0, 1/n)$*, then the following bounds hold for any distortion* $D \in [0, 1]$ *and integer* $k < m$*,*

$$\tilde{\varepsilon}_n^{(ML)}(D) \leq \sum_{b=\lfloor Dk+1 \rfloor}^{k} \min(\Psi_1(b), \Psi_2(b)) \tag{2.39}$$

where

$$\Psi_1(b) = \min_{\lambda \in [0,1]} \left[\left(2 \log \left[\sqrt{\frac{1 + \lambda M(b) + (1-\lambda) M(0)}{1 + M(0)}} - 1 \right] \right)^{-\frac{m-k}{4}} + \binom{k}{b} \binom{n-k}{b} \left(\frac{1 + M(b)}{1 + \lambda M(b) + (1-\lambda) M(0)} \right)^{-\frac{m-k}{2}} \right] \tag{2.40}$$

$$\Psi_2(b) = \min_{\substack{\theta, \mu \in (0,1) \\ \epsilon > 0}} \binom{k}{b} \binom{n-k}{b} \left[\left(1 + \frac{1}{4}(1-\theta)^2 M(b) \right)^{-\frac{m-k}{2}} + \left(\frac{\exp(\epsilon)}{2M(0)} \right)^{-\frac{m-k}{2}} + \left(\frac{1 + \mu\theta M(b)}{\exp(\epsilon M(0))} \right)^{-\frac{m-k}{2}} (1 - \mu^2)^{-\frac{b}{2}} \right]. \tag{2.41}$$

Proof. For each $S \in \mathcal{S}_k^n$ let $\Pi(\mathbf{A_s})$ denote the $m \times m$ orthonormal projection onto the null space of the $m \times k$ matrix $\mathbf{A_s}$. If $\mathbf{A}_S$ is full rank (an event that occurs with probability one over the assumed distribution on $\mathbf{A}$) then this projection is given by

$$\Pi(\mathbf{A_s}) = I_{m\times m} - \mathbf{A_s}(\mathbf{A_s}^T\mathbf{A_s})^{-1}\mathbf{A_s}^T. \tag{2.42}$$

Since

$$\min_{\mathbf{u_s}\in\mathbb{R}^k} \|\mathbf{A_s u_s} - \mathbf{Y}\| = \|\Pi(\mathbf{A_s})\mathbf{Y}\|, \tag{2.43}$$

it can be easily verified that the ML estimate of size k is an element of the set

$$\arg\min_{\mathbf{s}\in\mathcal{S}_k^n} \|\Pi(\mathbf{A_s})\mathbf{Y}\|. \tag{2.44}$$

Now, for each integer $b \in \{0, 1, 2, \cdots, k\}$, define the event

$$\mathcal{E}(b) = \left\{ \min_{\mathbf{s}\in B_b} \|\Pi(\mathbf{A_s})\mathbf{Y}\| > \|\Pi(\mathbf{A_{\tilde{s}}})\mathbf{Y}\| \right\} \tag{2.45}$$

where $B_b = \{\mathbf{s} \in \mathcal{S}_k^n : |\mathbf{s}\backslash\tilde{\mathbf{s}}| = b\}$. In words, the event $\mathcal{E}(b)$ guarantees that the set of minimizers in (2.44) will not contain any set S for which $d(S, \tilde{S}) = b/k$. Thus, a sufficient condition for recovery is given by the event $\bigcap_{\lfloor Dk+1\rfloor}^k \mathcal{E}(b)$, and by the union bound we have

$$\tilde{\varepsilon}_n^{(\mathrm{ML})}(D) \leq \sum_{b=\lfloor Dk+1\rfloor}^{k} \Pr[\mathcal{E}^c(b)] \tag{2.46}$$

where $\mathcal{E}^c(b)$ denotes the complement $\mathcal{E}(b)$.

The bounds $\Pr[\mathcal{E}^c(b)] \leq \Psi_1(b)$ and $\Pr[\mathcal{E}^c(b)] \leq \Psi_2(b)$ are proved in Sections 2.5.3 and 2.5.4 respectively. □

Remark 2.7. *Lemma 2.1 is general in the sense that it makes no assumptions about the sparsity of* $\mathbf{x}$ *or the size of* $\tilde{S}$*. Therefore, it can be used to address a variety of recovery tasks such as recovering a subset or superset of the true support.*

Remark 2.8. *If* $M(0) < M(\lfloor Dk+1\rfloor)$*, then the upper bound decreases exponentially rapidly with* m*, i.e. there exists a constant* C *such that* $\tilde{\varepsilon}_n^{(ML)}(D) \leq \exp(-C\,m)$*.*

2.5.2 The Asymptotic Setting

We now prove Theorem 2.8 by applying Lemma 2.1 to a sequence of problems obeying Assumptions S2 and M1-M5. For each problem of size n let $k_n = \lceil \kappa n \rceil$. Beginning with

(2.39), we have

$$\tilde{\varepsilon}_n(D) \leq \sum_{b=\lceil Dk_n \rceil}^{k_n} \min(\Psi_1(b), \Psi_2(b)) \tag{2.47}$$

$$\leq n \max_{\lceil Dk_n \rceil \leq b \leq k_n} \min(\Psi_1(b), \Psi_2(b)) \tag{2.48}$$

$$= n \exp\Big(-n \min_{\beta \in [D,1]} \psi_n(\beta) \Big) \tag{2.49}$$

where $\psi_n(\beta) = -\frac{1}{n} \min_{i \in \{1,2\}} \log \Psi_i(\lceil \beta k_n \rceil)$. To study the asymptotic behavior of this bound we need the following lemma. The proof is given in Section 2.5.5.

Lemma 2.2. *Under Assumption S2, the sequence of functions $\{\psi_n(\beta)\}_{n \geq 1}$ is uniformly asymptotically lower bounded in the following sense*

$$\limsup_{n \to \infty} \max_{\beta \in [0,1]} \Big(\psi(\beta) - \psi_n(\beta) \Big) \leq 0 \tag{2.50}$$

where $\psi(\beta) = \max_{i \in \{1,2\}} \psi_i(\beta)$ and

$$\psi_1(\beta) = \max_{\lambda \in [0,1]} \min \Bigg\{ \frac{\rho - \kappa}{4} \Big(\sqrt{1 + \lambda P(\beta) \mathsf{snr}} - 1 \Big)^2, \\ \frac{\rho - \kappa}{2} \Big[\log \Big(\frac{1 + P(\beta)\mathsf{snr}}{1 + \lambda P(\beta)\mathsf{snr}} \Big) - \frac{(1-\lambda) P(\beta)\mathsf{snr}}{1 + P(\beta)\mathsf{snr}} \Big] - \mathcal{H}(\beta; \kappa) \Bigg\} \tag{2.51}$$

$$\psi_2(\beta) = \max_{\theta, \mu \in [0,1]} \min \Bigg\{ \frac{\rho - \kappa}{2} \log \Big(1 + \tfrac{1}{4} P(\beta)\mathsf{snr} \Big), \\ \frac{\rho - \kappa}{2} \log \Big(1 + \theta \mu P(\beta)\mathsf{snr} \Big) + \frac{\beta \kappa}{2} \log \Big(1 - \mu^2 \Big) \Bigg\} - \mathcal{H}(\beta; \kappa). \tag{2.52}$$

Remark 2.9. *Under the additional constraint of Assumption S3, the bound* (2.50) *holds with respect to the absolute difference $|\psi(\beta) - \psi_n(\beta)|$. For the proof of Theorem 2.8, however, only the lower bound is needed.*

Returning to (2.49), we can now write

$$\liminf_{n \to \infty} -\frac{1}{n} \log \tilde{\varepsilon}_n(D) \geq \liminf_{n \to \infty} \min_{\beta \in [D,1]} \psi_n(\beta) \\ \geq \min_{\beta \in [D,1]} \psi(\beta) \tag{2.53}$$

where the swapping of the limit and the minimum in (2.53) is justified by Lemma 2.2.

With a bit of algebra, it can be verified that

$$\kappa + \Lambda(\beta; p_X, \mathsf{snr}) = \inf \Big\{ \rho : \psi(\beta) > 0 \Big\}, \tag{2.54}$$

and thus

$$\rho^{(\text{ML-UB})} = \inf\Big\{\rho : \min_{\beta\in[D,1]} \psi(\beta) > 0\Big\}. \tag{2.55}$$

Since $\psi(\beta)$ is a continuous and monotonically increasing function of ρ, it follows that the right hand side of (2.53) is strictly positive for any $\rho > \rho^{(\text{ML})}$. This concludes the proof of Theorem 2.8.

2.5.3 Proof of the bound $\Psi_1(b)$ in Lemma 2.1

We begin with a bounding technique used previously by Wainwright [76] for the study of exact sparsity pattern recovery. For notational simplicity, we define the random variable

$$Z_{\mathbf{s}} = \mathsf{snr}\,\|\Pi(\mathbf{A}_{\mathbf{s}})\mathbf{Y}\|^2 \tag{2.56}$$

which corresponds to the distance between the samples $\mathbf{Y}$ and subspace spanned by $\mathbf{A}_{\mathbf{s}}$.

For any $t \in \mathbb{R}$, we can write

$$\begin{aligned}\Pr[\mathcal{E}^c(b)] &= \Pr[\mathcal{E}^c(b) \cap \{Z_{\tilde{\mathbf{s}}} > t\}] + \Pr[\mathcal{E}^c(b) \cap \{Z_{\tilde{\mathbf{s}}} < t\}] \\ &\leq \Pr[Z_{\tilde{\mathbf{s}}} > t] + \Pr[\mathcal{E}^c(b) \cap \{Z_{\tilde{\mathbf{s}}} < t\}]. \end{aligned}\tag{2.57}$$

Furthermore,

$$\begin{aligned}\Pr[\mathcal{E}^c(b) \cap \{Z_{\tilde{\mathbf{s}}} \leq t\}] &= \Pr\Big[\{\min_{\mathbf{s}\in B_b} Z_{\mathbf{z}} \leq Z_{\tilde{\mathbf{s}}}\} \cap \{Z_{\tilde{\mathbf{s}}} \leq t\}\Big] \\ &\leq \Pr\Big[\min_{\mathbf{s}\in B_b} Z_{\mathbf{s}} \leq t\Big] \\ &\leq \sum_{\mathbf{s}\in B_b} \Pr[Z_{\mathbf{s}} \leq t], \end{aligned}\tag{2.58}$$

where (2.58) follows from the union bound. Plugging (2.58) back into (2.57) gives

$$\Pr[\mathcal{E}^c(b)] \leq \Pr[Z_{\tilde{\mathbf{s}}} > t] + \sum_{\mathbf{s}\in B_b} \Pr[Z_{\mathbf{s}} \leq t]. \tag{2.59}$$

Note that $\Pr[Z_{\mathbf{s}} \leq t]$ depends only on the marginal distributions of the random variable $Z_{\mathbf{s}}$. In Wainwright's analysis [76], this probability is upper bounded in terms of a noncentral chi-squared random variable whose noncentrality parameter is unknown but bounded. In this proof however, we begin with the exact distribution on $Z_{\mathbf{s}}$.

Lemma 2.3. *For each $\mathbf{s} \in \mathcal{S}_k^n$, the random variable*

$$\frac{Z_{\mathbf{s}}}{1 + \frac{1}{n}\|\mathbf{x}_{\mathbf{s}^c}\|^2 \mathsf{snr}}$$

has a chi-squared distribution with $m - k$ degrees of freedom.

Proof. Since $\mathbf{A_s x_s}$ lies in the range space of $\mathbf{A_s}$, we can write

$$\begin{aligned} Z_{\mathbf{s}} &= \|\Pi(\mathbf{A_s})(\sqrt{\mathsf{snr}}\mathbf{A}\mathbf{x} + \mathbf{W})\|^2 \\ &= \|\Pi(\mathbf{A_s})(\sqrt{\mathsf{snr}}\mathbf{A}_{\mathbf{s}^c}\mathbf{x}_{\mathbf{s}^c} + \mathbf{W})\|^2. \end{aligned}$$

The vector $\sqrt{\mathsf{snr}}\mathbf{A_s x_s} + \mathbf{W}$ is independent of $\Pi(\mathbf{A_s})$ and has i.i.d. Gaussian entries with mean zero and variance $1 + \frac{1}{n}\|\mathbf{x}_{\mathbf{s}^c}\|^2\,\mathsf{snr}$. Also, with probability one over the distribution $\mathbf{A}$, the matrix $\Pi(\mathbf{A_s})$ has exactly $n-k$ singular values equal to 1 and k singular values equal to 0. Therefore, the stated result follows immediately from the rotational invariance of the Gaussian distribution. □

To proceed, let V denote a chi-squared random variable with $m-k$ degrees of freedom and let $t = (1+\bar{M})(m-k)$ where $\bar{M} = \lambda M(b) + (1-\lambda)M(0)$ for some $\lambda \in (0,1)$. Then, by Lemma 2.3,

$$\Pr[Z_{\tilde{\mathbf{s}}} > t] = \Pr\left[\Big(\frac{1}{m-k}\Big)V > \frac{1+\bar{M}}{1+M(0)}\right] \tag{2.60}$$

and

$$\begin{aligned} \Pr[Z_{\mathbf{s}} \le t] &= \Pr\left[\Big(\frac{1}{m-k}\Big)V \le \frac{1+\bar{M}}{1+\frac{1}{n}\|\mathbf{x_s}\|^2\mathsf{snr}}\right] &(2.61)\\ &\le \Pr\left[\Big(\frac{1}{m-k}\Big)V \le \frac{1+\bar{M}}{1+M(b)}\right] &(2.62) \end{aligned}$$

where (2.62) follows from the definition of $M(b)$.

Both (2.60) and (2.62) can be upper bounded using the chi-squared large deviations bounds given in Lemma 2.4 below. Combining these bounds with (2.59) and the simple fact that

$$|B_b| = \binom{k}{b}\binom{n-k}{b}, \tag{2.63}$$

shows that $\Pr[\mathcal{E}^c(b)] \le \Psi_1(b)$, which completes the proof.

Lemma 2.4. *Let V be a chi-squared random variable with d degrees of freedom. For any $x > 1$,*

$$\begin{aligned} \Pr[V \ge d\,x\big] &\le \exp\Big(-d\,\tfrac{1}{4}(\sqrt{2x-1}-1)^2\Big), &(2.64)\\ \Pr[V \le d/x] &\le \exp\Big(-d\tfrac{1}{2}[\log x + 1/x - 1]\Big). &(2.65) \end{aligned}$$

Proof. The upper bound (2.64) follows directly from Laurent and Massart [41, pp. 1325].

For the lower bound (2.65), observe that for any $\mu > 0$,

$$\begin{aligned}\Pr[V \leq (\tfrac{1}{x})d] &= \Pr\big[\exp(-\mu V) \geq \exp(-\mu(\tfrac{1}{x})d)\big] \\ &\leq \mathbb{E}[\exp(-\mu X)\exp(\mu(\tfrac{1}{x})d)] \qquad (2.66) \\ &= \exp(-d[\tfrac{1}{2}\log(1+2\mu) - \mu(\tfrac{1}{x})]) \qquad (2.67)\end{aligned}$$

where (2.66) follows from Markov's inequality and (2.67) follows from the moment generating function of a chi-squared distribution. Letting $\mu = (x-1)/2$ completes the proof. □

2.5.4 Proof of the bound $\Psi_2(b)$ in Lemma 2.1

This proof uses a new decomposition of the error event to obtain a different upper bound on $\Pr[\mathcal{E}^c(b)]$. In some problem regimes, this bound improves significantly over the bound derived in the previous section. As before, we use the definition of $Z_{\mathbf{s}}$ given in (2.56).

To motivate our proof strategy, observe that one weakness of the bound (2.59) is that the threshold t is a fixed constant whereas the event $\mathcal{E}(b)$ depends on the *relative* magnitudes of the variables $Z_{\mathbf{S}}$.

In this proof, we begin with the union bound as follows

$$\Pr[\mathcal{E}^c(b)] \leq \sum_{\mathbf{s}\in B_a} \Pr[Z_{\mathbf{s}} \leq Z_{\tilde{\mathbf{s}}}]. \qquad (2.68)$$

Unlike (2.59), each probability on the right hand side of (2.68) depends on the relative magnitudes of Z_S and $Z_{\tilde{S}}$. In the remainder of this proof, our goal is to derive an upper bound on $\Pr[Z_{\mathbf{s}} \leq Z_{\tilde{\mathbf{s}}}]$ that exploits the dependence between $Z_{\tilde{\mathbf{s}}}$ and $Z_{\mathbf{s}}$. A key step is the following lemma.

Lemma 2.5. *For any $\mathbf{s} \in \mathcal{S}_k^n$, define the random variables*

$$\begin{aligned}T_{\mathbf{s}} &= \sqrt{\mathsf{snr}}\,\|\Pi(\mathbf{A}_{\mathbf{s}})\mathbf{A}_{\mathbf{s}^c}\mathbf{x}_{\mathbf{s}^c}\| \\ U_{\mathbf{s}} &= \frac{\langle \Pi(\mathbf{A}_{\mathbf{s}})\mathbf{A}_{\mathbf{s}^c}\mathbf{x}_{\mathbf{s}^c}, \mathbf{W}\rangle}{\|\Pi(\mathbf{A}_{\mathbf{s}})\mathbf{A}_{\mathbf{s}^c}\mathbf{x}_{\mathbf{s}^c}\|} \\ V_{\mathbf{s}} &= \|\Pi(\mathbf{A}_{\mathbf{s}})\mathbf{W}\|.\end{aligned}$$

The following statements hold:

(a) $Z_{\mathbf{s}} = T_{\mathbf{s}}^2 + 2T_{\mathbf{s}}U_{\mathbf{s}} + V_{\mathbf{s}}^2$

(b) $T_{\mathbf{s}}^2/(\frac{1}{n}\|\mathbf{x}_{\mathbf{s}^c}\|^2\,\mathsf{snr})$ *has a chi-squared distribution with $m-k$ degrees of freedom.*

(c) $U_{\mathbf{s}}$ *is independent of* $T_{\mathbf{s}}$ *and has a Gaussian distribution with mean zero and variance one.*

(d) $V_{\mathbf{s}}$ is independent of $T_{\mathbf{s}'}$ for any $\mathbf{s}, \mathbf{s}' \in \mathcal{S}_k^n$.

Proof. To prove part (a) observe that

$$\begin{aligned} Z_{\mathbf{s}} &= \|\Pi(\mathbf{A}_{\mathbf{s}})(\sqrt{\mathsf{snr}}\mathbf{A}_{\mathbf{s}^c}\mathbf{x}_{\mathbf{s}^c} + \mathbf{W})\|^2 \\ &= \mathsf{snr}\,\|\Pi(\mathbf{A}_{\mathbf{s}})\mathbf{A}_{\mathbf{s}^c}\mathbf{x}_{\mathbf{s}^c}\|^2 + \|\Pi(\mathbf{A}_{\mathbf{s}})\mathbf{W}\|^2 \\ &\quad + 2\sqrt{\mathsf{snr}}\,\langle \Pi(\mathbf{A}_{\mathbf{s}})\mathbf{A}_{\mathbf{s}^c}\mathbf{x}_{\mathbf{s}^c}, \Pi(\mathbf{A}_{\mathbf{s}})\mathbf{W}\rangle \end{aligned} \tag{2.69}$$

Part (b) follows from the proof of Lemma 2.3. Part (c) follows from the fact that the vector $\Pi(\mathbf{A}_{\mathbf{s}})\mathbf{A}_{\mathbf{s}^c}\mathbf{x}_{\mathbf{s}^c}$ is independent of $\mathbf{W}$ and is nonzero with probability one. Part (d) follows from the fact that $\Pi(\mathbf{A}_{\mathbf{s}})$, $\mathbf{A}_{\mathbf{s}^c}\mathbf{x}_{\mathbf{s}^c}$, and $\mathbf{W}$ are mutually independent and $\Pi(\mathbf{A}_{\mathbf{s}})$ has rank $m-k$ with probability one. □

To proceed, fix any $\theta \in (0,1)$ and $\epsilon > 0$ and define the event $\mathcal{A} = \cap_{i=1}^3 \mathcal{A}_i$ where

$$\mathcal{A}_1 = \left\{T_{\mathbf{s}}^2 + 2T_{\tilde{\mathbf{s}}}U_{\mathbf{s}} \geq \theta T_{\mathbf{s}}^2\right\} \tag{2.70}$$

$$\mathcal{A}_2 = \left\{T_{\tilde{\mathbf{s}}}^2 + 2T_{\tilde{\mathbf{s}}}U_{\tilde{\mathbf{s}}} \leq \epsilon(m-k)\right\} \tag{2.71}$$

$$\mathcal{A}_3 = \left\{\theta T_{\mathbf{s}}^2 + V_{\mathbf{s}}^2 - V_{\tilde{\mathbf{s}}}^2 > \epsilon(m-k)\right\}. \tag{2.72}$$

Using part (a) of Lemma 2.5 it can be verified that $\{Z_{\mathbf{s}} \leq Z_{\tilde{\mathbf{s}}}\} \cap \mathcal{A} = \{\emptyset\}$, and thus

$$\begin{aligned} \Pr[Z_{\mathbf{s}} \leq Z_{\tilde{\mathbf{s}}}] &= \Pr[\{Z_{\mathbf{s}} \leq Z_{\tilde{\mathbf{s}}}\} \cap \mathcal{A}^c] \\ &\leq \sum_{i=1}^3 \Pr[\mathcal{A}_i^c] \end{aligned} \tag{2.73}$$

where (2.73) follows from the union bound. In the following three subsections, we prove upper bounds on the probabilities $\Pr[\mathcal{A}_j^c]$, $j \in \{1,2,3\}$. Plugging these bounds back into (2.73) and using the fact that the cardinality of B_b is given by (2.63) completes the proof.

Upper Bound on $\Pr[\mathcal{A}_1^c]$

The first error event is relatively straightforward to bound. Observe that

$$\begin{aligned} \Pr[\mathcal{A}_1^c] &= \Pr[\tfrac{(1-\theta)^2}{4}T_{\mathbf{s}}^2 + \tfrac{(1-\theta)}{2}T_{\mathbf{s}}U_{\mathbf{s}} < 0] \\ &= \Pr[\exp(-\tfrac{(1-\theta)^2}{4}T_{\mathbf{s}}^2 - \tfrac{(1-\theta)}{2}T_{\mathbf{s}}U_{\mathbf{s}}) \geq 1] \\ &\leq \mathbb{E}[\exp(-\tfrac{(1-\theta)^2}{4}T_{\mathbf{s}}^2 - \tfrac{(1-\theta)}{2}T_{\mathbf{s}}U_{\mathbf{s}})] && (2.74) \\ &= \mathbb{E}[\exp(-\tfrac{(1-\theta)^2}{8}T_{\mathbf{s}}^2)] && (2.75) \\ &= \left(1 - \tfrac{(1-\theta)^2}{4}\tfrac{1}{n}\|\mathbf{x}_{\mathbf{s}^c}\|^2\,\mathsf{snr}\right)^{-(m-k)/2} && (2.76) \\ &\leq \left(1 - \tfrac{(1-\theta)^2}{4}M(b)\right)^{-(m-k)/2} && (2.77) \end{aligned}$$

where: (2.74) follows from Markov's inequality; (2.75) follows from part (c) of Lemma 2.5 and the moment generating function of the Gaussian distribution; (2.76) follows from part (b) of Lemma 2.5 and the moment generating function of the chi-squared distribution; and (2.77) follows from the definition of $M(b)$.

Upper Bound on $\Pr[\mathcal{A}_2^c]$

The second error event is similar to the first one, except that the inequality is in the other direction and there is a constant term to deal with. If we let $t = \epsilon M(0)(m-k)$ and $\lambda = 1/(2M(0))$, then

$$\begin{aligned}
\Pr[\mathcal{A}_2^c] &= \Pr[\lambda(-t + T_{\tilde{\mathbf{s}}}^2 + 2T_{\tilde{\mathbf{s}}}U_{\tilde{\mathbf{s}}}) > 0] \\
&= \Pr[\exp(-\lambda t + \lambda T_{\tilde{\mathbf{s}}}^2 + 2\lambda T_{\tilde{\mathbf{s}}}U_{\tilde{\mathbf{s}}}) > 1] \\
&\leq \mathbb{E}[\exp(-\lambda t + \lambda T_{\tilde{\mathbf{s}}}^2 + 2\lambda T_{\tilde{\mathbf{s}}}U_{\tilde{\mathbf{s}}})] && (2.78) \\
&= \mathbb{E}[\exp(-\lambda t + (\lambda - 2\lambda^2)T_{\tilde{\mathbf{s}}}^2)] && (2.79) \\
&= \exp(-\lambda t)\Big(1 - 2(\lambda - 2\lambda^2)M(0)\Big)^{-\frac{m-k}{2}} && (2.80) \\
&= \Big(\frac{\exp(\epsilon)}{2M(0)}\Big)^{-\frac{m-k}{2}} && (2.81)
\end{aligned}$$

where: (2.78) follows from Markov's inequality; (2.79) follows from part (c) of Lemma 2.5 and the moment generating function of the Gaussian distribution; (2.80) follows from part (b) of Lemma 2.5 and the moment generating function of the chi-squared distribution; and (2.77) follows from plugging in the definitions of t and λ.

Upper Bound on $\Pr[\mathcal{A}_3^c]$

The third error event requires the most work. Part of the difficulty is that the random variables $V_{\mathbf{s}}^2$ and $V_{\tilde{\mathbf{s}}}^2$ are not independent. The following result uses the fact that they are positively correlated to obtain a nontrivial upper bound on the moment generating function of their difference; the proof is given in Section 2.5.6.

Lemma 2.6. *For any* $\mu \in (0, 1)$,

$$\mathbb{E}[\exp(\tfrac{\mu}{2}[V_{\tilde{\mathbf{s}}}^2 - V_{\mathbf{s}}^2])] \leq (1-\mu^2)^{-b/2}. \tag{2.82}$$

We remark that the exponent in (2.82) is proportional to the overlap b. If $V_{\mathbf{s}}^2$ and $V_{\tilde{\mathbf{s}}}^2$ were independent, then the exponent would be proportional to k. This difference in the exponents is the key reason why this bounding technique works well in settings where the previous technique failed.

With Lemma 2.6 in hand, we are now ready to upper bound the probability $\Pr[\mathcal{A}_3^c]$. Let $t = \epsilon P_n(0)(m-k)$ and fix any $\mu \in (0,1)$. Then,

$$\begin{aligned}
\Pr[\mathcal{A}_3^c] &= \Pr[\tfrac{\mu}{2}(t - \theta T_{\mathbf{s}}^2 - V_{\mathbf{s}}^2 + V_{\tilde{\mathbf{s}}}^2) \geq 0] \\
&= \Pr[\exp(\tfrac{\mu}{2}(t - \theta T_{\mathbf{s}}^2 - V_{\mathbf{s}}^2 + V_{\tilde{\mathbf{s}}}^2)) \geq 1] \\
&\leq \mathbb{E}[\exp(\tfrac{\mu}{2}(t - \theta T_{\mathbf{s}}^2 - V_{\mathbf{s}}^2 + V_{\tilde{\mathbf{s}}}^2))] && (2.83) \\
&= \mathbb{E}[\exp(\tfrac{\mu}{2}[t - \theta T_{\mathbf{s}}^2])]\mathbb{E}[\exp(\tfrac{\mu}{2}[V_{\tilde{\mathbf{s}}}^2 - V_{\mathbf{s}}^2])] && (2.84) \\
&= e^{\frac{\mu}{2}t}\Big(1 + \mu\theta\|\mathbf{x}_{\mathbf{s}^c}\|^2\,\mathsf{snr}\Big)^{-\frac{m-k}{2}}(1-\mu^2)^{-\frac{b}{2}} && (2.85) \\
&\leq e^{\frac{\mu}{2}t}\Big(1 + \mu\theta M(b)\Big)^{-\frac{m-k}{2}}(1-\mu^2)^{-\frac{b}{2}} && (2.86) \\
&= \Big(\frac{1+\mu\theta M(b)}{\exp(\epsilon P_n(0))}\Big)^{-\frac{m-k}{2}}(1-\mu^2)^{-\frac{b}{2}} && (2.87)
\end{aligned}$$

where: (2.83) follows from Markov's inequality; (2.84) follows from part (d) of Lemma 2.5; (2.85) follows from part (b) of Lemma 2.5, the moment generating function of the chi-squared distribution, and Lemma 2.6; (2.86) follows from the definition of $M(b)$; and (2.87) follows from the definition of t.

2.5.5 Proof of Lemma 2.2

To simplify notation we will write k instead of k_n where the dependence on n is implicit.

Since $\Psi_1(b)$ and $\Psi_2(b)$ are non-increasing functions of $M(b)$, it is sufficient to show that the following limits hold:

$$\lim_{n\to\infty} \sup_{\beta\in[0,1]} \left| \mathcal{H}(\beta,\kappa) - \frac{1}{n}\log\binom{k}{\lceil\beta k\rceil}\binom{n-k}{\lceil\beta k\rceil} \right| = 0 \tag{2.88}$$

$$\lim_{n\to\infty} M(0) = 0 \tag{2.89}$$

$$\limsup_{n\to\infty} \max_{\beta\in[0,1]} \Big(P(\beta)\,\mathsf{snr} - M(\lceil\beta k\rceil) \Big) < 0. \tag{2.90}$$

Then, it follows immediately that

$$x \limsup_{n\to\infty} \max_{\beta\in[0,1]} \left(\psi_i(\beta) + \frac{1}{n}\log(\Psi_i(\lceil\beta k_n\rceil) \right) < 0 \tag{2.91}$$

for $i \in \{1,2\}$, which proves the desired result.

To begin, note that (2.88) follows directly from a strong form of Stirling's approximation [14, Lemma 17.5.1].

Next, we consider the term $M(0)$. For each problem of size n, let $\{n_i\}_{i\in[n]}$ be a permutation of $[n]$ such that $x_{n_1}^2 \leq x_{n_2}^2 \leq \cdots \leq x_{n_n}^2$. Starting with the definition of $\tilde{\mathbf{s}}$, we can

write

$$\mathsf{snr}^{-1} M(0) = \min_{\mathbf{s}\in\mathcal{S}_k^n} \frac{1}{n}\|\mathbf{x}_{\mathbf{s}^c}\|^2 \tag{2.92}$$

$$= \frac{1}{n}\sum_{i=1}^{n-k} x_{n_i}^2 \tag{2.93}$$

$$= \int_0^\infty \Big(\frac{n-k}{n} - \frac{1}{n}\sum_{i=1}^{n-k} \mathbf{1}(x_{n_i}^2 \le u)\Big) du \tag{2.94}$$

$$= \int_0^\infty \max\Big(1 - G_n(u) - \tfrac{k}{n}, 0\Big) du, \tag{2.95}$$

where $G_n(u)$ denotes the empirical distribution function of $\{x_i^2\}_{i\in[n]}$, Thus, for any $\epsilon > 0$,

$$\begin{aligned}\mathsf{snr}^{-1} M(0) &= \int_0^\epsilon \max\Big(1 - G_n(u) - \tfrac{k}{n}, 0\Big) du \\ &\quad + \int_\epsilon^\infty \max\Big(1 - G_n(u) - \tfrac{k}{n}, 0\Big) du \\ &\le \epsilon + \max\Big(1 - G_n(\epsilon) - \tfrac{k}{n}, 0\Big).\end{aligned} \tag{2.96}$$

By the weak convergence of Assumption S2, it follows that the second term on the right hand side of (2.96) converges to zero as $n \to \infty$. Since epsilon is arbitrary, we conclude that $\lim_{n\to\infty} M(0) = 0$.

We now consider the final term $M(b)$. Since

$$\begin{aligned} n\,\mathsf{snr}^{-1} M(b) &= \|\mathbf{x}\|^2 - \max_{\mathbf{s}\in B_b} \|\mathbf{x}_{\mathbf{s}}\|^2 \\ &\ge \|\mathbf{x}\|^2 - \max_{\mathbf{s}\in B_b} \Big(\|\mathbf{x}_{\mathbf{s}}\|^2 + \|\mathbf{x}_{\mathbf{s}^c\cap\tilde{\mathbf{s}}^c}\|^2\Big) \\ &= \|\mathbf{x}\|^2 - \max_{\mathbf{s}\in B_b} \|\mathbf{x}_{\mathbf{s}\cap\tilde{\mathbf{s}}}\|^2 - \|\mathbf{x}_{\tilde{\mathbf{s}}^c}\|^2 \\ &= \min_{\mathbf{s}\in\mathcal{S}_{k-b}^n} \|\mathbf{x}_{\mathbf{s}^c}\|^2 - n\,\mathsf{snr}^{-1} M(0) \end{aligned}$$

it is sufficient to show that

$$\limsup_{n\to\infty} \max_{\beta\in[0,1]} \Big(P(\beta) - P_n(\beta)\Big) < 0 \tag{2.97}$$

where $P_n(\beta) = \frac{1}{n}\min_{\mathbf{s}\in\mathcal{S}_{k-b}^n} \|\mathbf{x}_{\mathbf{s}^c}\|^2$.

Following the same steps we used for $M(0)$, we have

$$P_n(\beta) = \int_0^\infty \max\Big(1 - G_n(u) - \tfrac{k-\lceil\beta k\rceil}{n}, 0\Big) du.$$

Also, by definition

$$P(\beta) = \int_0^\infty \max\Big(1 - G(u) - (1-\beta)\kappa, 0\Big) du \tag{2.98}$$

where $G(u) = \Pr[X^2 \leq u]$. Thus, for any $\epsilon > 0$ we have

$$\begin{aligned} P(\beta) - P_n(\beta) &= \int_0^\epsilon \max\Big(1 - G(u) - (1-\beta)\kappa, 0\Big) du + \int_0^\infty \varphi_n(u) du \\ &\leq \epsilon + \int_0^\infty \max(\varphi_n(u), 0) du \end{aligned} \tag{2.99}$$

where

$$\varphi_n(u) = \Big[\max\Big(1 - G(u+\epsilon) - (1-\beta)\kappa, 0\Big) - \max\Big(1 - G_n(u) - \tfrac{k - \lceil \beta k \rceil}{n}, 0\Big)\Big].$$

To deal with the second term in (2.99), observe that

$$\varphi_n(u) \leq \Big|(1-\beta)\kappa - \tfrac{k - \lceil \beta k \rceil}{n}\Big| + G_n(u) - G(u+\epsilon). \tag{2.100}$$

Thus, by the weak convergence of Assumption S2,

$$\lim_{n\to\infty} \max_{\beta \in [0,1]} \max(\varphi_n(u), 0) = 0 \tag{2.101}$$

for every $u \in \mathbb{R}$. Since $\varphi_n(u)$ is upper bounded by the integrable function $1 - G(u+\epsilon)$, it follows from the bounded convergence theorem that the second term in (2.99) is equal to zero. Since ϵ is arbitrary, this proves (2.97) and hence (2.90).

2.5.6 Proof of Lemma 2.6

The key to this proof is to consider the eigenvalues of the matrix $M = \Pi(\mathbf{A}_{\tilde{\mathbf{s}}}) - \Pi(\mathbf{A}_{\mathbf{s}})$. Since M is symmetric, it can be expressed as $M = Q \Lambda Q^T$ where Q is an $m \times m$ orthonormal matrix and Λ is a real valued diagonal matrix whose diagonal entries obey $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_m$. Letting $\tilde{\mathbf{W}} = Q^T \mathbf{W}$, we have

$$V_{\tilde{\mathbf{s}}}^2 - V_{\mathbf{s}}^2 = \mathbf{W}^T M \mathbf{W}^T = \sum_{i=1}^m \lambda_i \tilde{W}_i^2 \tag{2.102}$$

where $\tilde{W}_1, \tilde{W}_2, \cdots, \tilde{W}_m$ are i.i.d. Gaussian $\mathcal{N}(0,1)$, and thus

$$\mathbb{E}[\exp(\tfrac{\mu}{2}[V_{\tilde{\mathbf{s}}}^2 - V_{\mathbf{s}}^2])] = \prod_{i=1}^m \mathbb{E}[\exp(\tfrac{\mu}{2}\lambda_i \tilde{W}_i^2)] \tag{2.103}$$

$$= \prod_{i=1}^m (1 - \mu\lambda_i)^{-1/2}. \tag{2.104}$$

To characterize the eigenvalues, we now consider two cases. If $m \geq 2k$, then

$$\begin{aligned} \operatorname{rank}(M) &= \operatorname{rank}\Big([I - \Pi(\mathbf{A}_{\mathbf{s}})] - [I - \Pi(\mathbf{A}_{\tilde{\mathbf{s}}})]\Big) \\ &\leq \operatorname{rank}(I - \Pi(\mathbf{A}_{\mathbf{s}})) + \operatorname{rank}(I - \Pi(\mathbf{A}_{\tilde{\mathbf{s}}})) \\ &\leq 2k, \end{aligned}$$

and so at least $m - 2k$ eigenvalues are equal to zero. It can be shown (see [53, p. 8]), that the remaining $2k$ singular values are given by $\lambda_i = \sin\theta_i$ and $\lambda_{m-i+1} = -\sin\theta_i$ for $i = 1, 2, \cdots, k$ where $\pi/2 \geq \theta_1 \geq \theta_2 \geq \cdots \geq \theta_k \geq 0$ are known as the *principal angles* between the k-dimensional subspaces $\mathcal{R}(\mathbf{A_s})$ and $\mathcal{R}(\mathbf{A_{\tilde{s}}})$ spanned by the columns of $\mathbf{A_s}$ and $\mathbf{A_{\tilde{s}}}$ respectively. Since the number of principal angles that are equal to zero is given by the dimension of the intersection of the two subspaces, it follows that

$$\begin{aligned} |\{i : \theta_i = 0\}| &= \dim\Big(\mathcal{R}(\mathbf{A_s}) \cap \mathcal{R}(\mathbf{A_{\tilde{s}}})\Big) \\ &\geq \dim\Big(\mathcal{R}(\mathbf{A_{s\cap\tilde{s}}})\Big) \\ &= k - b \end{aligned}$$

where the last equality holds with probability one over the distribution on $\mathbf{A}$.

Returning to (2.104), we can now write

$$\mathbb{E}[\exp(\tfrac{\mu}{2}[V_{\tilde{\mathbf{s}}}^2 - V_{\mathbf{s}}^2])] = \prod_{i=1}^{b}(1 - \mu^2\sin^2\theta_i)^{-1/2} \tag{2.105}$$

$$\leq (1 - \mu^2)^{-b/2} \tag{2.106}$$

where (2.106) follows from the fact that $0 \leq \sin^2\theta_i \leq 1$.

For the case $m < 2k$ we use similar arguments. Since

$$\begin{aligned} \operatorname{rank}(M) &\leq \operatorname{rank}(\Pi(\mathbf{A_s})) + \operatorname{rank}(\Pi(\mathbf{A_{\tilde{s}}})) \\ &\leq 2(m - k), \end{aligned}$$

at least $2k - m$ eigenvalues of M are equal to zero. The remaining $2(m - k)$ eigenvalues are given by $\lambda_i = \sin\theta_i$ and $\lambda_{m-i+1} = -\sin\theta_i$ for $i = 1, 2, \cdots, m - k$ where the θ_i are the principal angles between the $m-k$ dimensional subspaces $\mathcal{N}(\mathbf{A_s})$ and $\mathcal{N}(\mathbf{A_{\tilde{s}}})$ corresponding to the orthogonal complements of $\mathcal{R}(\mathbf{A_s})$ and $\mathcal{R}(\mathbf{A_{\tilde{s}}})$ respectively. Thus, we have

$$\begin{aligned} |\{i : \theta_i = 0\}| &= \dim\Big(\mathcal{N}(\mathbf{A_s}) \cap \mathcal{N}(\mathbf{A_{\tilde{s}}})\Big) \\ &= m - \dim\Big(\mathcal{R}(\mathbf{A_s}) + \mathcal{R}(\mathbf{A_{\tilde{s}}})\Big) \\ &\geq \max\Big(0, m - 2k + \dim\Big(\mathcal{R}(\mathbf{A_{s\cap\tilde{s}}})\Big)\Big) \\ &= \max\Big(0, m - k - b\Big) \end{aligned}$$

where the last equality holds with probability one over the distribution on $\mathbf{A}$. Therefore, there are at most b nonzero principle angles. Following the same steps used in the previous case, leads again to the upper bound (2.106). This concludes the proof of Lemma 2.6.

2.6 Proofs of Remaining Upper Bounds

We now give the proofs of Theorems 2.2, 2.3, 2.4, and 2.7. Each of these proofs follows a similar outline. First, we establish convergence of the empirical joint distribution on the entries in $\mathbf{x}$ and the vector estimate $\hat{\mathbf{X}}$ generated in the first stage recovery (see Fig 2.1). Then, we show that this convergence characterizes the limiting distortion with respect to the relaxed sparsity pattern recovery task described in Section 1.4.4.

In these proofs, we use the superscripts $\stackrel{prob}{\rightarrow}$ and $\stackrel{dist}{\rightarrow}$ to denote convergence in probability and distribution, respectively.

2.6.1 From Convergence in Distribution to Relaxed Recovery

For each problem of size n, let $\hat{\mathbf{X}}$ denote the estimate of the unknown vector $\mathbf{x}$ generated in the first stage of sparsity pattern recovery and let $F_n(x, \hat{x})$ denote the cumulative distribution function (CDF) of the empirical joint distribution on the entries in $(\mathbf{x}, \hat{\mathbf{X}})$, i.e.

$$F_n(x, \hat{x}) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\Big(x_i \leq x, \hat{X}_i \leq \hat{x}\Big). \tag{2.107}$$

Note that $F_n(x, \hat{x})$ is a random function due to the randomness in $\hat{\mathbf{X}}$. Also, let $F(x, z)$ denote the CDF of the random pair (X, Z) given in Definition 2.1, i.e.

$$F(x, z) = \Pr[X \leq x, Z \leq z]. \tag{2.108}$$

According to standard terminology, $F_n(x, \hat{x})$ converges weakly in probability to the limit $F(x, z)$ if

$$\lim_{n \to \infty} \Pr\Big[\Big|F_n(x, z) - F(x, z)\Big| > \epsilon\Big] = 0 \tag{2.109}$$

for any fixed $\epsilon > 0$ and $(x, z) \in \mathbb{R}^2$ such that (x, z) are continuity points of $F(x, z)$. Since Z is a continuous random variable, the last constraint simplifies to all $(x, z) \in \mathbb{R}^2$ such that $p_X(\{x\}) = 0$.

Lemma 2.7. *If $F_n(x, \hat{x})$ convergence weakly in probability to a limit $F(x, z)$ characterized by a distribution p_X and noise power $\sigma^2 > 0$, then the distortion between the sparsity pattern estimate $\hat{S}$ generated in the second stage of recovery and the set $\tilde{S}$ described in Section 1.4.4 obeys*

$$\lim_{n \to \infty} d(\tilde{S}, \hat{S}) \stackrel{prob}{=} D_{awgn}(\sigma^2; p_X) \tag{2.110}$$

where $D_{awgn}(\sigma^2; p_X)$ is given by (2.14).

Proof. For each problem of size n, define

$$\tilde{U} = \{i \in [n] : |x_i| > \delta\} \quad \text{and} \quad \hat{U} = \{i \in [n] : |\hat{X}_i| > t\},$$

where $\delta > 0$ satisfies $\Pr[|X| = \delta] = 0$ and t is the unique solution to $\Pr[|Z| \geq t] = \kappa$. Note that t corresponds to the minimizer of the right hand side of (2.14).

By the triangle inequality, we have

$$\left| d(\tilde{\mathbf{s}}, \hat{\mathbf{s}}) - d(\tilde{U}, \hat{U}) \right| \leq d(\tilde{\mathbf{s}}, \tilde{U}) + d(\hat{U}, \hat{\mathbf{s}}). \tag{2.111}$$

Furthermore, by the weak convergence of $F_n(x, \hat{x})$ to $F(x, z)$ and the definitions of $\tilde{S}$ and $\hat{S}$, it can be shown that,

$$\lim_{n\to\infty} d(\tilde{\mathbf{s}}, \tilde{U}) = \Pr[|X| \leq \delta | X \neq 0] \tag{2.112}$$

$$\lim_{n\to\infty} d(\tilde{U}, \hat{U}) \overset{prob}{=} \Pr[|X| \leq \delta \,|\, |Z| > t] \tag{2.113}$$

$$\lim_{n\to\infty} d(\hat{U}, \hat{\mathbf{s}}) \overset{prob}{=} 0, \tag{2.114}$$

where (2.113) and (2.114) follow from the definition of t.

By the assumptions on p_X and the definition of $D_{\text{awgn}}(\sigma^2; p_X)$, there exists, for any $\epsilon > 0$, a $\delta > 0$ such that $\Pr[|X| = \delta] = 0$ and

$$\Pr[|X| \leq \delta | X \neq 0] \leq \epsilon \tag{2.115}$$

$$\left| \Pr[|X| \leq \delta \,|\, |Z| > t] - D_{\text{awgn}}(\sigma^2; p_X) \right| \leq \epsilon. \tag{2.116}$$

Hence, we have shown that

$$\lim_{n\to\infty} \Pr\left[\left| d(\tilde{S}, \hat{S}) - D_{\text{awgn}}(\sigma^2; p_X) \right| > \epsilon' \right] = 0 \tag{2.117}$$

for any $\epsilon' > 0$ which completes the proof. □

2.6.2 Proof of Theorem 2.2

In this section, we prove convergence of the empirical CDF $F_n(x, \hat{x})$ corresponding to the MF estimate. Theorem 2.2 then follows immediately from Lemmas 1.1 and 2.7.

The crucial step in this proof is the following result which characterizes the limiting joint distribution of a randomly chosen subset of the entries in $(\mathbf{x}, \hat{\mathbf{X}}^{(\text{MF})})$. Due to the simplicity of the MF estimate, we are able to prove this convergence generally for any i.i.d. distribution on the entries of the measurement matrix $\mathbf{A}$.

Lemma 2.8. *Let L be a fixed integer. For each problem of size $n \geq L$, let $\mathcal{L}$ be distributed uniformly over all subsets of $[n]$ of size L. Then, under Assumptions S2-S3 and M1-M4,*

the joint distribution on $\{(x_\ell, \hat{X}_\ell^{(MF)})\}_{\ell\in\mathcal{L}}$ *converges weakly to the joint distribution on* L *independent copies of the random pair* (X, Z) *given in Definition 2.1 where* σ^2 *is given by*

$$\sigma^2 = \frac{\mathsf{snr}^{-1} + 1}{\rho}. \tag{2.118}$$

Proof. To gain intuition, observe that the entries in the MF estimate indexed by $\mathcal{L}$ can be decomposed as follows:

$$\hat{\mathbf{X}}_{\mathcal{L}}^{(\mathrm{MF})} = \left(\tfrac{n}{m}\right)\mathbf{A}_{\mathcal{L}}^T\mathbf{A}_{\mathcal{L}}\mathbf{x}_{\mathcal{L}} + \left(\tfrac{n}{m}\right)\mathbf{A}_{\mathcal{L}}^T\Big(\mathbf{A}_{\mathcal{L}^c}\mathbf{x}_{\mathcal{L}^c} + \tfrac{1}{\sqrt{\mathsf{snr}}}\mathbf{W}\Big). \tag{2.119}$$

By the law of large numbers, it is straightforward to show that the first term on the right hand side of (2.119) converges in distribution to random vector $\mathbf{X}$ whose entries are i.i.d. copies of X. Also, by the central limit theorem, it is straightforward to show that the second term converges in distribution to a vector whose entries are i.i.d. $\mathcal{N}(0, \sigma^2)$. However, since the terms in (2.119) are *not* mutually independent, these arguments are, by themselves, insufficient to prove Lemma 2.8.

To proceed, we will introduce an additional term that allows us to decompose $\hat{\mathbf{X}}_{\mathcal{L}}^{(\mathrm{MF})}$ into independent components. Specifically, for each problem of size n, let $\tilde{\mathbf{A}}$ be an $m \times L$ random matrix whose columns are independent copies of the columns of $\mathbf{A}$ and define the random vectors

$$\mathbf{U} = \Big[\left(\tfrac{n}{m}\right)\mathbf{A}_{\mathcal{L}}^T\Big(\mathbf{A}_{\mathcal{L}} - \tilde{\mathbf{A}}\Big) - I_{L\times L}\Big]\mathbf{x}_{\mathcal{L}} \tag{2.120}$$

$$\mathbf{V} = \left(\tfrac{n}{m}\right)\mathbf{A}_{\mathcal{L}}^T\Big(\tilde{\mathbf{A}}\mathbf{x}_{\mathcal{L}} + \mathbf{A}_{\mathcal{L}^c}\mathbf{x}_{\mathcal{L}^c} + \mathsf{snr}^{-1/2}\mathbf{W}\Big). \tag{2.121}$$

Then, we can write

$$\hat{\mathbf{X}}_{\mathcal{L}}^{(\mathrm{MF})} = \mathbf{x}_{\mathcal{L}} + \mathbf{U} + \mathbf{V} \tag{2.122}$$

where the vectors $\mathbf{x}_{\mathcal{L}}$ and $\mathbf{V}$ are independent.

From here, the proof is straightforward. If the following limits hold,

$$\lim_{n\to\infty} \mathbf{x}_{\mathcal{L}} \overset{dist}{=} \mathbf{X} \tag{2.123}$$

$$\lim_{n\to\infty} \mathbf{U} \overset{prob}{=} \mathbf{0}_{L\times 1} \tag{2.124}$$

$$\lim_{n\to\infty} \mathbf{V} \overset{dist}{=} \mathcal{N}(0, \sigma^2 I_{L\times L}), \tag{2.125}$$

then the desired convergence follows immediately from Slutsky's theorem.

The limit (2.123) follows from Assumption S2, and the fact that L is finite. To prove (2.124), observe that by Assumptions M1-M4 and the weak law of large numbers, $\mathbf{A}_{\mathcal{L}}^T\mathbf{A}_{\mathcal{L}} \to (m/n)I_{L\times L}$ and $\mathbf{A}_{\mathcal{L}}^T\tilde{\mathbf{A}}_{\mathcal{L}} \to \mathbf{0}_{L\times L}$ in probability as $n \to \infty$. Combining these facts with (2.123) shows that $\mathbf{U}$ converges to $\mathbf{0}_{L\times 1}$ in distribution, and thus also in probability.

Finally, to prove (2.125), observe that $\mathbf{V} = \sum_{i=1}^{m} \mathbf{V}_i$ where

$$\mathbf{V}_i = \left(\tfrac{n}{m}\right)(\mathbf{A}_{\mathcal{L}}^T)_i\Big(\tilde{\mathbf{A}}\mathbf{x}_{\mathcal{L}} + \mathbf{A}_{\mathcal{L}^c}\mathbf{x}_{\mathcal{L}^c} + \mathsf{snr}^{-1/2}\mathbf{W}\Big)_i \tag{2.126}$$

and $(\mathbf{A}_{\mathcal{L}}^T)_i$ denotes the i'th column of the $L \times m$ matrix $\mathbf{A}_{\mathcal{L}}^T$. Since the entries in $\mathbf{A}$, $\tilde{\mathbf{A}}$, and $\mathbf{W}$ are mutually independent, it can be verified that the vectors $\{\mathbf{V}_i\}_{i\in[m]}$ are i.i.d. with mean zero and covariance

$$\mathbb{E}[\mathbf{V}_i\mathbf{V}_i^T] = \left(\frac{n}{m^2}\right)\left(\frac{1}{n}\|\mathbf{x}\|^2 + \mathsf{snr}^{-1}\right)I_{L\times L}. \tag{2.127}$$

Therefore, the limit (2.125) follows from the multivariate central limit theorem and Assumption S3. □

With Lemma 2.8 in hand, we can now prove convergence of the empirical CDF $F_n(x,\hat{x})$ directly from Chebyshev's inequality.

Lemma 2.9. *Under Assumptions S2-S3 and M1-M4, the empirical CDF $F_n(x,\hat{x})$ corresponding to the MF estimate converges weakly in probability to a limit $F(x,z)$ with noise power σ^2 given by* (2.118).

Proof. Beginning with Chebyshev's inequality, we have

$$\begin{aligned}\Pr\Big[\Big|F_n(x,\hat{x}) - F(x,\hat{x})\Big| > \epsilon\Big] &\le \epsilon^{-2}\,\mathbb{E}\Big[\Big|F_n(x,\hat{x}) - F(x,\hat{x})\Big|^2\Big] \\ &= \epsilon^{-2}\left|\mathbb{E}\Big[F_n^2(x,\hat{x})] - F^2(x,\hat{x})\Big|\right. - \epsilon^{-2}2\Big|\mathbb{E}[F_n(x,\hat{x})] - F(x,\hat{x})\Big|\end{aligned} \tag{2.128}$$

for any $\epsilon > 0$. By the linearity of expectation, we can write

$$\mathbb{E}[F_n(x,\hat{x})] = \Pr[x_{\ell_1} \le x, \hat{X}_{\ell_1} \le \hat{x}] \tag{2.129}$$

$$\begin{aligned}\mathbb{E}[F_n^2(x,\hat{x})] = \tfrac{n-1}{n}\Pr[x_{\ell_1} \le x, \hat{X}_{\ell_1} \le \hat{x}, x_{\ell_2} \le x, \hat{X}_{\ell_2} \le \hat{x}] \\ + \tfrac{1}{n}\Pr[x_{\ell_1} \le x, \hat{X}_{\ell_1} \le \hat{x}]\end{aligned} \tag{2.130}$$

where ℓ_1 and ℓ_2 are drawn uniformly at random without replacement from $[n]$. Hence, by Lemma 2.8, it follows that

$$\begin{aligned}\lim_{n\to\infty}\mathbb{E}[F_n(x,\hat{x})] &= F(x,\hat{x}) \\ \lim_{n\to\infty}\mathbb{E}[F_n^2(x,\hat{x})] &= F^2(x,\hat{x}).\end{aligned}$$

Therefore, both terms on the right hand side of (2.128) converge to zero as $n \to \infty$, thus completing the proof. □

2.6.3 Proof of Theorem 2.3

For this proof, we use the well known fact (see e.g. [34]) that matrix inversion can be performed using iterative methods. Specifically, for a fixed tuple $(\mathbf{y}, A, \mathsf{snr})$, let γ be the unique positive solution to quadratic equation

$$\mathsf{snr} = \gamma\Big(\frac{m}{n} - \frac{1}{1+\gamma}\Big), \tag{2.131}$$

and consider the AMP algorithm with $\eta(z, \sigma^2) = z/(1+\gamma)$. If the sequences $\{\mathbf{x}^t\}_{t\geq 1}$ and $\{\mathbf{u}^t\}_{t\geq 1}$ converge to a fixed point $(\mathbf{x}^\infty, \mathbf{u}^\infty)$, then it can be verified by checking the update equations (2.18) and (2.19) that $\mathbf{x}^\infty = \mathbf{x}^{(\mathrm{LMMSE})}$, $A^T\mathbf{u}^\infty = \gamma\mathbf{x}^{(\mathrm{LMMSE})}$, and thus

$$\mathbf{x}^{(\mathrm{AMP})} = (1+\gamma)\mathbf{x}^{(\mathrm{LMMSE})}. \tag{2.132}$$

Therefore, the LMMSE estimate can be computed using the appropriate linear version of AMP, provided that the AMP algorithm converges.

We now use the analysis of Bayati and Montanari to characterize the limiting behavior of the AMP estimate. For each problem of size n let $\hat{\mathbf{X}}^{(\mathrm{AMP})}$ denote the output of the AMP algorithm corresponding to the function $\eta(z, \sigma^2) = z/(1+\gamma_n)$ where γ_n is the unique positive solution to (2.131). Then, under Assumptions S2-S3 and M1-M5, it follows from part (b) of [7, Lemma 1] that the empirical CDF corresponding to $\hat{\mathbf{X}}^{(\mathrm{AMP})}$ converges weakly almost surely to a limit $F(x, z)$ with a noise power σ_∞^2 that is the unique solution to the quadratic equation

$$\rho = \frac{1}{\sigma_\infty^2\,\mathsf{snr}} + \frac{1}{1+\sigma_\infty^2}. \tag{2.133}$$

Since the LMMSE estimate is proportional to the AMP estimate, this result, along with Lemmas 1.1 and 2.7, completes the proof of Theorem 2.3.

2.6.4 Proof of Theorem 2.4

To begin, consider a modified version of the AMP algorithm in which the sequence of noise power estimates $\{\hat{\sigma}_t^2\}_{t\geq 1}$ is replaced with the sequence of noise powers $\{\sigma_t^2\}_{t\geq 1}$ defined by the state evolution recursion (2.25). (Note that this modified algorithm depends explicitly on the distribution p_X.) For each problem of size n, let

$$\hat{\mathbf{X}}^t = \mathbf{A}^T\mathbf{U}^t + \mathbf{X}^t \tag{2.134}$$

denote the modified AMP estimate at iteration t. Then, under Assumptions S2-S3 and M1-M5, it follows from part (b) of [7, Lemma 1] that the empirical CDF corresponding to $\hat{\mathbf{X}}^t$ converges weakly almost surely to a limit $F(x, z)$ with noise power σ_t^2.

Moreover, by part (c) of [7, Lemma 1] it can be shown that, under the same assumptions, the residuals $\mathbf{U}^t$ corresponding to the modified AMP algorithm obey

$$\lim_{n\to\infty} \tfrac{1}{n}\|\mathbf{U}^t(n)\|^2 = \sigma_t^2 \tag{2.135}$$

almost surely. Thus, by the continuity of $\eta(z,\sigma^2)$ with respect to σ^2, it follows that the AMP algorithm using the empirical estimates $\hat{\sigma}_t^2$ has the same limiting behavior as the modified AMP algorithm.

By the above arguments, the empirical CDF $F_n(x,\hat{x})$ corresponding to the AMP estimate (2.23) converges weakly almost surely, and hence also in probability, to a limit $F(x,z)$ with noise power σ_∞^2 given in (2.26). Combining this result with Lemmas 1.1 and 2.7 completes the proof of Theorem 2.4.

2.6.5 Proof of Theorem 2.7

This proof follows along the same lines as the proof of Theorem 2.2. The key step is the following result which is analogous to Lemma 2.8 except that it also requires the replica analysis assumptions. This result is stated as Claim 3 in [35], and its proof follows directly from the analysis in [36, Section IV-B].

Lemma 2.10. *Assume that the replica analysis assumptions hold. Let L be a fixed integer. For each problem of size $n \geq L$, let $\mathcal{L}$ be distributed uniformly over all subsets of $[n]$ of size L. Then, under Assumptions S2-S3 and M1-M4, the joint distribution on $\{(x_\ell, \hat{X}_\ell^{(MMSE)})\}_{\ell\in\mathcal{L}}$ converges weakly to the joint distribution on L independent copies of the random pair (X,Z) given in Definition 2.1 where σ^2 is given by the noise power τ^* defined in* (2.37).

From here, the rest of the proof follows immediately from Chebyshev's inequality (see Lemma 2.9).

Chapter 3

Information-Theoretic Lower Bounds

This chapter gives lower bounds on the fundamental sampling rate distortion-function. These bounds consist of necessary conditions which apply generally to any possible recovery algorithm. We begin by describing a stochastic signal model in Section 3.1. In Section 3.2 we derive general lower bounds which hold for any sequence of matrices obeying Assumptions M1-M3. In Section 3.3 we strengthen these results for matrices whose entries are also i.i.d. (Assumption M4). Additional results for the noiseless setting are considered in Section 3.4, and proofs are given in Section 3.5.

3.1 Stochastic Signal Model

Throughout this chapter, the unknown vector is modeled as a random vector $\mathbf{X}$. Accordingly, the linear observation model described in Section 1.4 is expressed as

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \frac{1}{\sqrt{\mathsf{snr}}}\mathbf{W}. \tag{3.1}$$

To characterize a sequence of recovery problems $\{\mathbf{X}(n), \mathbf{A}(n), \mathbf{W}(n)\}_{n\geq 1}$ indexed by the vector length n, we use the following stochastic signal assumptions.

Stochastic Signal Assumptions: We consider the following assumptions on a sequence of random vectors $\mathbf{X}(n) \in \mathbb{R}^n$.

SS1 *Linear Sparsity:* The sparsity pattern S^* is distributed uniformly over all subsets of $\{1, 2, \cdots, n\}$ of size $k(n)$ where

$$\lim_{n\to\infty} \frac{k(n)}{n} = \kappa \tag{3.2}$$

for some *sparsity rate* $\kappa \in (0, 1/2)$.

SS2 *I.I.D. Entries:* The nonzero entries $\{X_i : i \in \mathcal{S}^*\}$ are i.i.d. p_U where p_U is a probability measure with zero mass at 0, i.e. $\Pr[U = 0] = 0$. We use p_X to denote the probability measure given by

$$p_X = (1 - \kappa)\delta_0 + \kappa\, p_U$$

where δ_0 denotes a point-mass at $x = 0$.

SS3 *Normalization:* The distribution p_X has second moment equal to one.

The stochastic signal assumptions are closely related to the deterministic signal assumptions given in Section 1.4.2. One difference is that under Assumptions SS1-SS2 we may consider distributions p_X without a second moment constraint. This extra degree of freedom gives us greater flexibility in stating our lower bounds. In all cases, Assumption SS3 can be enforced by rescaling the parameter snr appropriately.

The definition of achievability under the stochastic signal assumptions SS1-SS3 is the same as the definition of achievability under the deterministic signal assumptions S1-S3 except that the error probability $\varepsilon_n^{(\mathrm{ALG})}$ given in (1.7) is taken with respect to the probability measure on $\mathbf{X}$. Also we assume that the number of nonzero entries $k = |S^*|$ is known throughout the system. Under these assumptions, the optimal recovery algorithm can be stated explicitly as

$$\hat{S}^{(\mathrm{OPT})} = \arg \min_{S:|S|=k} \Pr[d(S^*, S) > D | A\mathbf{X} + \mathsf{snr}^{-1/2}\mathbf{W} = \mathbf{y}]. \tag{3.3}$$

The following result shows that a necessary condition for the stochastic setting implies a necessary condition for the deterministic setting.

Lemma 3.1. *If a distortion D is not achievable for the tuple $(\rho, p_X, \mathsf{snr})$ under Assumptions SS1-SS3, then it is not achievable under Assumptions S1-S3.*

Proof. This results follows immediately from the fact that a random sequence of vectors $\{\mathbf{X}(n)\}_{n\geq 1}$ distributed according to Assumptions SS1-SS3 obeys Assumptions of S1-S3 with probability one. □

3.2 Bounds for Arbitrary Measurement Matrices

This section derives lower bounds on the fundamental sampling rate distortion function that apply generally to any sequence of measurement matrices obeying Assumptions M1-M3.

Before we present our bounds, two more definitions are needed. First, we use the notation

$$V_X = \mathrm{Var}(X) \tag{3.4}$$

to denote the variance of the distribution p_X. Note that $(1-\kappa) \leq V_X \leq 1$ for any distribution p_X obeying the constraints of Assumptions SS1-SS3.

Also, we define

$$R(D;\kappa) = \begin{cases} H(\kappa) - \kappa H(D) - (1-\kappa) H\Big(\frac{\kappa D}{1-\kappa}\Big), & \text{if } D < 1-\kappa \\ 0, & \text{if } D \geq 1-\kappa \end{cases} \tag{3.5}$$

where $H(p) = -p \log p - (1-p)\log(1-p)$ is binary entropy. It is straightforward to show that $R(D;\kappa)$ corresponds to the information rate (given in nats per dimension) required to encode a sparsity pattern to within distortion D.

Our first lower bound is general in the sense that it depends only on the variance of the distribution p_X. This result serves as a building block for our stronger bounds.

Theorem 3.1. *Under Assumptions SS1-SS2 and M1-M3, a distortion D is not achievable for the tuple $(\rho, p_X, \mathsf{snr})$ if*

$$\min(1,\rho) \log\Big(1 + \max(1,\rho) V_X \, \mathsf{snr}\Big) < 2R(D;\kappa). \tag{3.6}$$

Proof. See Section 3.5.1. □

The following Corollary is equivalent to Theorem 3.1 in the under-sampled setting $\rho < 1$. This result has been derived previously in the special case of exact recovery [33, 64, 78], as well as for approximate recovery in the special case of binary signals [1].

Corollary 3.1. *Under Assumptions SS1-SS2 and M1-M3, a distortion D is not achievable for the tuple $(\rho, p_X, \mathsf{snr})$ if*

$$\rho < \frac{2R(D;\kappa)}{\log(1 + V_X \, \mathsf{snr})}. \tag{3.7}$$

Proof. This result follows from the fact that $\log(1+\rho\gamma) \leq \rho \log(1+\gamma)$ for all $\rho \geq 1$. □

Theorem 3.1 is remarkable in that it holds for any possible recovery algorithm. Moreover, it shows that a nonzero sampling rate ρ is necessary in the presence of noise.

One critical weakness of Theorem 3.1, however, is that it does not reflect the true difficulty of sparsity recovery when the desired distortion D is small. For example, if $D = 0$, then the lower bound on sampling rate is finite even though it has been shown that an infinite sampling rate is needed in the presence of noise [58]. Among other things, this discrepancy leaves open the possibility that the total number of recovery errors could grow sublinearly with the length n such that the fraction of errors is asymptotically zero.

Our next result allows us to lower bound the distortion corresponding to a distribution p_X in terms of a different but related distribution p_Z. This result is extremely powerful since it allows us to isolate the key aspects of the recovery problem that make recovery difficult.

Theorem 3.2. *Let p_X and p_Z be probability measures with the following properties:*

$$0 < \kappa_Z \leq \kappa_X \tag{3.8}$$

$$\frac{p_Z(A)}{1-\kappa_Z} \leq \frac{p_Z(A)}{1-\kappa_X} \quad \text{for all } A \subseteq \mathbb{R}\backslash\{0\}. \tag{3.9}$$

where $\kappa_X = 1 - p_X(\{0\})$ and $\kappa_Z = 1 - p_X(\{0\})$. For a given tuple (D, ρ, snr) define

$$\tilde{D} = \Big(\frac{1-\kappa_Z}{1-\kappa_X}\Big)\Big(\frac{\kappa_X}{\kappa_Z}\Big)D \tag{3.10}$$

$$\tilde{\rho} = \Big(\frac{1-\kappa_Z}{1-\kappa_X}\Big)\rho \tag{3.11}$$

$$\tilde{\mathsf{snr}} = \Big(\frac{1-\kappa_Z}{1-\kappa_X}\Big)\mathsf{snr}. \tag{3.12}$$

Under Assumptions SS1-SS2 and M1-M3, the following statement holds: If the distortion $\tilde{D}$ is not achievable for the tuple $(\tilde{\rho}, p_Z, \tilde{\mathsf{snr}})$, then the distortion D is not achievable for the tuple $(\rho, p_X, \mathsf{snr})$.

Proof. The proof is based on a genie argument and is given in Section 3.5.2. □

The reason that Theorem 3.2 is useful is that it allows us to isolate the nonzero entries in $\mathbf{X}$ whose locations are difficult to identify. Starting with an initial distribution p_X, one way to create an appropriate distribution p_Z is via truncation and re-normalization. For example, for any set $\{0\} \subset T \subseteq \mathbb{R}$, the distribution

$$p_Z(A) = \frac{p_X(A \cap T)}{p_X(T)}$$

satisfies the constraints of Theorem 3.2 with $\kappa_Z = 1 - (1-\kappa_X)/p_X(T)$.

Our next result combines Theorems 3.1 and 3.2 to give a bound that overcomes the weakness of Theorem 3.1 and accurately characterizes the difficulty of recovery when the distortion D is small.

Theorem 3.3. *Under Assumptions SS1-SS2 and M1-M3, a distortion D is not achievable for the tuple $(\rho, p_X, \mathsf{snr})$ if there exists a tuple $(\tilde{\rho}, p_Z, \tilde{\mathsf{snr}})$ satisfying the assumptions of Theorem 3.2 such that*

$$\min(1, \tilde{\rho}) \log(1 + \max(1, \tilde{\rho}) V_Z \, \tilde{\mathsf{snr}}) < 2R(\tilde{D}; \kappa_Z). \tag{3.13}$$

Proof. This result follows directly from Theorems 3.1 and 3.2. □

By the constraints of Theorem 3.2, it can be verified that (3.13) gives a nonzero lower bound only if the distortion $\tilde{D}$ defined in (3.10) obeys $D \leq \tilde{D} \leq 1$. By characterizing an explicit mapping between p_X and a distribution p_Z parameterized terms of the fraction $D' = D/\tilde{D}$, we obtain the following lower bound which is similar in style to the ML upper bound given Theorem 2.1.

Corollary 3.2. *Under Assumptions SS1-SS3 and M1-M3, a distortion D is not achievable for the tuple $(\rho, p_X, \mathsf{snr})$ if*

$$\rho < \max_{D' \in [D,1]} \frac{2(1-\kappa+\kappa D')R\left(\frac{D}{D'}; \frac{\kappa D'}{1-\kappa+\kappa D'}\right)}{\log(1+P(D';p_X)\,\mathsf{snr})} \tag{3.14}$$

where $P(D;p_X)$ is given in (2.3).

Proof. Fix any $D < D' < 1$. Starting with p_X let p_Z be the distribution that minimizes $\mathbb{E}[Z^2]$ subject to the constraints (3.8) and (3.9) with $\kappa_Z = \kappa_X D'/(1-\kappa_X D')$. As a simple exercise, it can then be verified that $V_Z\,\tilde{\mathsf{snr}} = P(D';p_X)\,\mathsf{snr}$. Parameterizing the bound in terms of D', using the approximation of Corollary 3.1, and maximizing over D' leads to (3.14). □

The following result gives a further simplification of Theorem 3.3.

Corollary 3.3. *Fix any $\alpha > 1$. Under Assumptions S1-S3 and M1-M3, a distortion $D \leq 1/\alpha$ is not achievable for the tuple $(\rho, p_X, \mathsf{snr})$ if*

$$\rho < \frac{2R\left(\frac{1}{\alpha}; \frac{\kappa\alpha D}{1-\kappa+\kappa\alpha D}\right)}{\log(1+P(\alpha D;p_X)\,\mathsf{snr})}. \tag{3.15}$$

Proof. This bound follows from evaluating the right hand side of (3.14) with $D' = \alpha D$. □

As the distortion D becomes small, it can be shown that the right hand side of (3.15) tends to infinity. Therefore, one important contribution of Theorem 3.3 is that it is not possible to have a vanishing fraction of errors if both the sampling rate and SNR are finite. In Chapter 4, we will show, by comparison with the upper bounds of Chapter 2, that Theorem 3.3 is relatively tight in the low distortion setting.

3.3 Bounds for IID Measurement Matrices

We now derive stronger lower bounds for measurement matrices whose entries are i.i.d. (Assumption M4). Unlike the bounds given in the previous section, these bounds capture the fact that the nonzero entries of $\mathbf{X}$ are unknown.

We define the nonzero entropy power of a random variable $X \sim p_X$ to be

$$N_X = \begin{cases} \dfrac{\kappa \exp\left(2h(X|X \neq 0)\right)}{2\pi e}, & \text{if } h(X|X \neq 0) \text{ exists} \\ 0, & \text{otherwise} \end{cases} \tag{3.16}$$

where $h(X|X \neq 0)$ denotes the differential entropy of the nonzero part of p_X. The nonzero entropy power allows us to assess the relative uncertainty about the nonzero entries.

The following result gives an improved lower bound in terms of the variance V_X and the nonzero entropy power N_X. The proof of this result relies heavily on the entropy power inequality and the spectral convergence of i.i.d. random matrices.

Theorem 3.4. *Under Assumptions SS1-SS3 and M1-M4, a distortion D is not achievable for the tuple $(\rho, p_X, \mathsf{snr})$ if*

$$\mathcal{V}(\rho, V_X\,\mathsf{snr}) - \kappa\mathcal{V}_{LB}(\rho/\kappa, N_X\,\mathsf{snr}) < 2R(D;\kappa) \tag{3.17}$$

where

$$\mathcal{V}(r,\gamma) = r\log\Big(1+\gamma-\mathcal{F}(r,\gamma)\Big) + \log\Big(1+r\,\gamma-\mathcal{F}(r,\gamma)\Big) - \frac{\mathcal{F}(r,\gamma)}{\gamma} \tag{3.18}$$

with

$$\mathcal{F}(r,\gamma) = \frac{1}{4}\left(\sqrt{\gamma\,(\sqrt{r}+1)^2+1} - \sqrt{\gamma\,(\sqrt{r}-1)^2+1}\right)^2 \tag{3.19}$$

and

$$\mathcal{V}_{LB}(r,\gamma) = \begin{cases} r\log\left(1+\gamma\Big(\dfrac{1}{1-r}\Big)^{1/r-1}\dfrac{1}{e}\right), & \text{if } r<1 \\ \log\left(1+\gamma\dfrac{1}{e}\right), & \text{if } r=1\,. \\ \log\left(1+\gamma\,r\Big(\dfrac{r}{r-1}\Big)^{r-1}\dfrac{1}{e}\right), & \text{if } r>1 \end{cases} \tag{3.20}$$

Proof. See section 3.5.3. □

Remark 3.1. *In the special case where the nonzero part of the distribution p_X is Gaussian, the function $\mathcal{V}_{LB}(r,\gamma)$ in the second term on the left hand side of (3.17) can be replaced with the function $\mathcal{V}(r,\gamma)$, thus providing a slightly stronger condition.*

The next result gives a simplified, and necessarily weaker, version of Theorem 3.4.

Corollary 3.4. *Under Assumptions SS1-SS3 and M1-M4, a distortion D is not achievable for the tuple $(\rho, p_X, \mathsf{snr})$ if*

$$\rho < \frac{\min(\rho,\kappa)\log(1+(N_X/e)\,\mathsf{snr}) + 2R(D,\kappa)}{\log(1+V_X\,\mathsf{snr})}. \tag{3.21}$$

Proof. This result follows immediately from (3.17) and the bounds $\mathcal{V}(r,\gamma) \le r\log(1+\gamma)$ and $\mathcal{V}_{LB}(r,\gamma) \ge \min(r,1)\log(1+\gamma/e)$. □

The key difference between Theorem 3.4 and the previous bounds occurs in the high SNR setting. For example, if D is small relative to N_X, then the lower bound on the fundamental sampling rate distortion function ρ^* given by (3.21) behaves like

$$\rho^* \geq \kappa + \frac{C}{\log(1 + \mathsf{snr})}$$

in the high SNR setting. The next result gives a lower bound on the sampling rate that is bounded away from zero for all SNR.

Corollary 3.5. *Under Assumptions SS1-SS3 and M1-M4, a distortion D is not achievable for the tuple $(\rho, p_X, \mathsf{snr})$ if*

$$\rho < \min\left(\kappa, \frac{2R(D;\kappa)}{(\frac{1-\kappa}{\kappa})\log(\frac{1}{1-\kappa}) + \log(V_X/N_X)}\right). \tag{3.22}$$

Proof. Since the fundamental sampling rate-distortion function is a non-increasing function of the SNR, the infinite SNR limit of (3.17) gives a necessary condition for any finite SNR. □

One weakness of Theorem 3.4, however, is that it does not improve significantly upon Theorem 3.1 when the nonzero entropy power N_X is equal to zero. Another weakness is that it does not accurately reflect the difficulty of recovery when D is small. To fix these weaknesses, we combine Theorem 3.4 with Theorem 3.2 to obtain the following bound.

Theorem 3.5. *Under Assumptions SS1-SS3 and M1-M4, a distortion D is not achievable for the tuple $(\rho, p_X, \mathsf{snr})$ if there exists a tuple $(\tilde{D}, \tilde{\rho}, p_Z, \tilde{\mathsf{snr}})$ satisfying the assumptions of Theorem 3.2 such that*

$$\mathcal{V}(\rho, V_Z\,\tilde{\mathsf{snr}}) - \kappa_Z \mathcal{V}_{LB}(\rho/\kappa_Z, N_Z\,\tilde{\mathsf{snr}}) < 2R(\tilde{D};\kappa_Z). \tag{3.23}$$

Proof. This result follows immediately from Theorems 3.2 and 3.4. □

One significant advantage of Theorem 3.5 over Theorem 3.4 is that we can now give a nontrivial high SNR bound for any distribution p_X whose nonzero part is a discrete-continuous mixture. As the following result shows, the high SNR behavior is dominated by the weight of the continuous part of the distribution.

Corollary 3.6. *Suppose that p_X can be expressed as*

$$p_X = (1-\kappa)\,\delta_0 + \omega_c\, p_{X_c} + (\kappa - \omega_c)\, p_{X_d} \tag{3.24}$$

where X_c is continuous with finite differential entropy and X_d is discrete. Under Assumptions SS1-SS2 and M1-M4, a distortion D is not achievable for the tuple $(\rho, p_X, \mathsf{snr})$ in the noiseless

setting if $\rho < \omega_c$ *and* (3.23) *holds for the tuple* $(\tilde{D}, \tilde{\rho}, p_Z, \tilde{\mathsf{snr}})$ *given by*

$$\tilde{D} = \frac{\kappa}{\omega_c} D \tag{3.25}$$

$$\tilde{\rho} = \left(\frac{1}{1-\kappa+\omega_c}\right)\rho \tag{3.26}$$

$$p_Z = \left(\frac{1-\kappa}{1-\kappa+\omega_c}\right)\delta_0 + \left(\frac{\omega_c}{1-\kappa+\omega_c}\right)p_{X_c} \tag{3.27}$$

$$\tilde{\mathsf{snr}} = \left(\frac{1}{1-\kappa+\omega_c}\right)\mathsf{snr}. \tag{3.28}$$

3.4 The Noiseless Setting

The previous sections provided lower bounds for the noisy setting. In this section, we address lower bounds for the setting where there is no measurement noise. Accordingly, we consider the linear observation model given by

$$\mathbf{Y} = \mathbf{A}\mathbf{X}. \tag{3.29}$$

The following result highlights the fact that recovery in the noiseless setting is very different in nature than recovery in the presence of noise.

Proposition 3.1. *Let* $\mathbf{x}$ *be an* $n \times 1$ *vector whose entries are supported on a countable set* $\mathcal{X} \subset \mathbb{R}$ *and let* $\mathbf{A}$ *be a* $1 \times n$ *random vector whose entries are i.i.d. from a distribution that is absolutely continuous with respect to Lebesgue measure. With probability one,* $\mathbf{x}$ *can be recovered uniquely from the tuple* $(\mathcal{X}, \mathbf{A}, \mathbf{A}\mathbf{x})$.

Proof. With probability one, the projection $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$ maps each possible realization of $\mathbf{x}$ to a unique real number. □

Corollary 3.7. *Suppose that* p_X *is a discrete distribution. Under Assumptions SS1-SS2 and M1-M3, the fundamental sampling rate distortion function for the noiseless setting is given by*

$$\rho^* = 0 \tag{3.30}$$

for all distortions $0 \leq D \leq 1$.

The proof of the following result follows directly from Theorem 3.1 and the proof of Thoerem 6.3 in Chapter 6. Alternatively, it can also be shown that this result corresponds to the infinite SNR limit of Theorem 3.4.

Theorem 3.6. *Consider the noiseless measurement model given in* (3.29)*. Under Assumptions SS1-SS2 and M1-M4, a distortion D is not achievable for the pair (ρ, p_X) if $\rho < \kappa$ and*

$$\rho \log\left(\frac{V_X}{N_X}\right) + (1-\rho)\log\left(\frac{1}{1-\rho}\right) - (\kappa-\rho)\log\left(\frac{\kappa}{\kappa-\rho}\right) < 2R(D;\kappa). \tag{3.31}$$

Theorem 3.7. *Suppose that p_X can be expressed as*

$$p_X = (1-\kappa)\,\delta_0 + \omega_c\, p_{X_c} + (\kappa-\omega_c)\, p_{X_d} \tag{3.32}$$

where X_c is continuous with finite differential entropy and X_d is discrete. Under Assumptions SS1-SS2 and M1-M4, a distortion D is not achievable for the pair (ρ, p_X) in the noiseless setting if $\rho < \omega_c$ and (3.31) *holds for the tuple $(\tilde{D}, \tilde{\rho}, p_Z)$ given by*

$$\tilde{D} = \frac{\kappa}{\omega_c} D \tag{3.33}$$

$$\tilde{\rho} = \left(\frac{1}{1-\kappa+\omega_c}\right)\rho \tag{3.34}$$

$$p_Z = \left(\frac{1-\kappa}{1-\kappa+\omega_c}\right)\delta_0 + \left(\frac{\omega_c}{1-\kappa+\omega_c}\right)p_{X_c} \tag{3.35}$$

Proof. This result follows directly from Theorems 3.6 and 3.2. □

Corollary 3.8. *Under the assumptions of Corollary 3.7 the fundamental sampling rate distortion function for the noiseless setting is given by*

$$\rho^* = \omega_c \tag{3.36}$$

for all distortions D such that

$$\left(\frac{\omega_c}{1-\kappa+\omega_c}\right)\log\left(\frac{V_Z}{N_Z}\right) + \left(\frac{1-\kappa}{1-\kappa+\omega_c}\right)\log\left(\frac{1-\kappa+\omega_c}{1-\kappa}\right) < 2R\Big(\frac{\kappa}{\omega_c}D; \frac{\omega_c}{1-\kappa+\omega_c}\Big) \tag{3.37}$$

where p_Z is given by (3.35)

3.5 Proofs of Lower Bounds

This section gives the proofs of our information-theoretic lower bounds.

3.5.1 Proof of Theorem 3.1

The cornerstone of this proof is Fano's inequality which gives a lower bound on the error probability for any possible recovery algorithm in terms of the mutual information between S^* and the pair $(\mathbf{Y}, \mathbf{A})$. We assume that the tuple (D, p_X, snr) is known throughout the system.

Lemma 3.2 (Fano's Inequality)**.** *Let S^* be distributed uniformly over all subsets of $[n]$ of size $k < n/2$. If $S^* \to (\mathbf{Y}, \mathbf{A}) \to \hat{S}$ forms a Markov chain then*

$$\Pr[d(S^*, \hat{S}) > D] \geq 1 - \frac{I(S^*; \mathbf{Y}, \mathbf{A}) + \log(2)}{\log\binom{n}{k} - \log\left(\sum_{\ell=0}^{\lceil Dk \rceil}\binom{k}{\ell}\binom{n-k}{\ell}\right)} \tag{3.38}$$

for all $0 \leq D \leq 1$.

Proof. We follow the proof of Fano's inequality given in [14] with some modifications to handle our error criterion. To begin, we define the random variable

$$E = \begin{cases} 1, & \text{if } d(S^*, \hat{S}) > D \\ 0, & \text{if } d(S^*, \hat{S}) \leq D \end{cases},$$

and note that $\Pr[E = 1] = \Pr[d(S^*, \hat{S}) > D]$.

Using the chain rule for entropy, $H(E, S^*|\mathbf{Y}, \mathbf{A}, \hat{S})$ can be written two ways as

$$H(E, S^*|\mathbf{Y}, \mathbf{A}, \hat{S}) = H(S^*|\mathbf{Y}, \mathbf{A}, \hat{S}) + H(E|S^*, \mathbf{Y}, \mathbf{A}, \hat{S}) \tag{3.39}$$

$$= H(E|\mathbf{Y}, \mathbf{A}, \hat{S}) + H(S^*|E, \mathbf{Y}, \mathbf{A}, \hat{S}). \tag{3.40}$$

By the Markov property, $H(S^*|\mathbf{Y}, \mathbf{A}, \hat{S}) = H(S^*|\mathbf{Y}, \mathbf{A})$. Since entropy is nonnegative, $H(E|S^*, \mathbf{Y}, \mathbf{A}, \hat{S}) \geq 0$. Also, since conditioning cannot increase entropy, $H(E|\mathbf{Y}, \mathbf{A}, \hat{S}) \leq H(E) \leq \log(2)$ and $H(S^*|E, \mathbf{Y}, \mathbf{A}, \hat{S}) \leq H(S^*|E, \hat{S})$. Putting everything together we obtain

$$H(S^*|\mathbf{Y}, \mathbf{A}) - \log 2 \leq H(S^*|E, \hat{S}) \tag{3.41}$$

$$= \Pr[E = 1]H(S^*|E = 1, \hat{S}) + \Pr[E = 0]H(S^*|E = 0, \hat{S}) \tag{3.42}$$

Since the uniform distribution maximizes the entropy of S^*,

$$H(S^*|E = 1, \hat{S}) \leq \log\binom{n}{k}. \tag{3.43}$$

Also, since the distortion measure $d(\cdot, \cdot)$ corresponds to the maximum of the two detection error rates, we may assume without any loss of generality that $\hat{S}$ has cardinality k. Therefore, a simple counting argument gives

$$H(S^*|E = 0, \hat{S}) \leq \log\left(\sum_{\ell=0}^{\lfloor Dk \rfloor}\binom{k}{\ell}\binom{n-k}{\ell}\right). \tag{3.44}$$

Plugging (3.43) and (3.44) back into (3.42), multiplying the expression by negative one, and adding $H(S^*) = \log\binom{n}{k}$ to each side gives:

$$I(S^*|\mathbf{Y}, \mathbf{A}) + \log 2 \geq \Big(1 - \Pr[E = 1]\Big)\left[\log\binom{n}{k} - \log\left(\sum_{\ell=0}^{\lfloor Dk \rfloor}\binom{k}{\ell}\binom{n-k}{\ell}\right)\right]. \tag{3.45}$$

Solving for $\Pr[E = 1]$ completes the proof. □

The next step in the proof is to verify that the right hand side of (3.38) is bounded away from zero for all sequences of problems obeying the assumptions of Theorem 3.1. For each problem of size n, let $k = \lceil \kappa\, n \rceil$ where the dependence on n is implicit. Using Stirling's approximation [14, Lemma 17.5.1], it is straight forward to verify that

$$\lim_{n\to\infty} \frac{1}{n} \left[\log \binom{n}{k} \log \left(\sum_{\ell=0}^{\lceil Dk \rceil} \binom{k}{\ell} \binom{n-k}{\ell} \right) \right] = R(D;\kappa) \tag{3.46}$$

where $R(D;\kappa)$ is given in (3.5).

Combining (3.38) and (3.46) it follows that a distortion D is not achievable if

$$\limsup_{n\to\infty} \frac{1}{n} I(S^*; \mathbf{Y}, \mathbf{A}) < R(D;\kappa). \tag{3.47}$$

The remainder of the proof is dedicated to upper bounding the left hand side of (3.47). Starting with the chain rule for mutual information, we have

$$\begin{aligned} I(S^*; \mathbf{Y}, \mathbf{A}) &= I(S^*; \mathbf{Y}|\mathbf{A}) + I(S^*; \mathbf{A}) && (3.48)\\ &= I(S^*; \mathbf{Y}|\mathbf{A}) && (3.49)\\ &\le I(\mathbf{X}; \mathbf{Y}|\mathbf{A}) && (3.50) \end{aligned}$$

where (3.49) follows from the independence of Assumption M3 and (3.50) follows from the data processing inequality and the fact that $S^* \to \mathbf{X} \to \mathbf{Y}$ forms a Markov chain.

Next, we can write

$$\begin{aligned} I(\mathbf{X}; \mathbf{Y}|\mathbf{A} = A) &= I(\mathbf{X}; A\mathbf{X} + \mathsf{snr}^{-1/2}\mathbf{W}) && (3.51)\\ &= I\Big(\mathbf{X} - \mathbb{E}[\mathbf{X}]; A(\mathbf{X} - \mathbb{E}[\mathbf{X}]) + \mathsf{snr}^{-1/2}\mathbf{W}\Big) && (3.52)\\ &\le \max_{Z} I(A\mathbf{Z}; A\mathbf{Z} + \mathsf{snr}^{-1/2}\mathbf{W}) && (3.53) \end{aligned}$$

where the maximum is over all n-dimensional random vectors $\mathbf{Z}$ obeying the power constraint

$$\mathbb{E}[\mathbf{Z}\mathbf{Z}^T] = \mathbb{E}\Big[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T\Big] = V_X I_{n\times n}. \tag{3.54}$$

It is well known (see e.g. [14]) that the maximum of (3.53) is attained when the entries of $\mathbf{Z}$ are i.i.d. $\mathcal{N}(0, V_X)$, and thus we obtain

$$I(\mathbf{X}; \mathbf{Y}|\mathbf{A} = A) \le \frac{1}{2} \log\det(I_{m\times m} + \mathsf{snr}\, V_X A A^T). \tag{3.55}$$

By the concavity of the log determinant, Hadamard's inequality, and Jensen's inequality we can bound the expectation of (3.55) with respect to a random matrix $\mathbf{A}$ obeying the normalization of Assumption M3 as follows:

$$\begin{aligned} \mathbb{E}\Big[\frac{1}{2} \log\det(I_{m\times m} + \mathsf{snr}\, V_X \mathbf{A}\mathbf{A}^T)\Big] &\le \frac{1}{2} \log\det\Big(I_{m\times m} + \mathsf{snr}\, V_X \mathbb{E}\Big[\mathbf{A}\mathbf{A}^T\Big]\Big) && (3.56)\\ &= \frac{m}{2} \log(1 + \mathsf{snr}\, V_X). && (3.57) \end{aligned}$$

Alternatively, starting with Sylvester's determinant theorem, we have

$$\mathbb{E}\Big[\frac{1}{2}\log\det(I_{m\times m}+\mathsf{snr}\,V_X\mathbf{A}\mathbf{A}^T)\Big] = \mathbb{E}\Big[\frac{1}{2}\log\det(I_{n\times n}+\mathsf{snr}\,V_X\mathbf{A}^T\mathbf{A})\Big] \tag{3.58}$$

$$\leq \frac{1}{2}\log\det\Big(I_{n\times n}+\mathsf{snr}\,V_X\mathbb{E}\Big[\mathbf{A}^T\mathbf{A}\Big]\Big) \tag{3.59}$$

$$= \frac{n}{2}\log\Big(1+\frac{m}{n}\mathsf{snr}\,V_X\Big). \tag{3.60}$$

Combining (3.55), (3.57), and (3.60) gives

$$I(\mathbf{X};\mathbf{Y}|\mathbf{A}) \leq \left(\frac{m}{2}\log(1+\mathsf{snr}\,V_X), \frac{n}{2}\log\Big(1+\frac{m}{n}\mathsf{snr}\,V_X\Big)\right), \tag{3.61}$$

and hence

$$\limsup_{n\to\infty}\frac{1}{n}I(S^*;\mathbf{Y},\mathbf{A}) < \frac{\min(1,\rho)}{2}\log\Big(1+\max(1,\rho)V_X\,\mathsf{snr}\Big), \tag{3.62}$$

for any sequence of matrices obeying Assumptions M1-M3. Combining (3.47) and (3.62) completes the proof of Theorem 3.1.

3.5.2 Proof of Theorem 3.2

This proof is based on a genie argument. Suppose that a genie provides the recovery algorithm with the pair $(G, \mathbf{X}_G)$ where G is a subset of the sparsity pattern S^* and $\mathbf{X}_G$ is a $|G|$-dimensional vector corresponding to the entries of $\mathbf{X}$ indexed by G. Given this extra information, the recovery algorithm must then determine which of the remaining unknown entries $\{X_i : i \notin G\}$ are nonzero. Clearly, any lower bound on the achievable distortion D in the genie-aided setting is also a lower bound on the achievable distortion in the original setting.

In the following sections, we first describe how the genie selects the index set G. We then show that the resulting recovery problem is equivalent to the original recovery problem with altered parameters.

Genie Selection Strategy

The set G is constructed as follows: each index $i = 1, 2, \cdots, n$ is reported, independently of the other indices, with probability $q(X_i)$ where the function $0 \leq q(x) \leq 1$ is chosen such that

$$\Pr[X_i \leq t|\ i \text{ is not reported}] = \Pr[Z \leq t]$$

where $Z \sim p_Z$. In words, the genie "prunes" the entries of $\mathbf{X}$ in a way such that the unreported entries are marginally distributed according to the distribution p_Z. By the constraints (3.8) and (3.9) it can be verified that the function $q(x)$ exists and that $q(0) = 0$.

We now make several observations. First, since $q(0) = 0$, only nonzero entries are reported and so $G \subseteq S^*$. Second, since the indices are selected independently, the remaining nonzero entries $\{X_i : i \in S^* \backslash G\}$ are i.i.d. according to the nonzero part of p_Z. Finally, conditioned on the cardinality $|G|$, the set $S^* \backslash G$ is distributed uniformly over all subsets of $[n] \backslash G$ of size $|S^*| - |G|$.

As a consequence of the above observations, the sequence of vectors corresponding to $\mathbf{X}_{[n]\backslash G}$ satisfies Assumptions SS1-SS2 with distribution p_Z. Moreover, if we let $\tilde{\mathbf{Y}}$ denote the measurements corresponding to the vector $\mathbf{X}_{[n]\backslash G}$ and measurement matrix $\mathbf{A}_{[n]\backslash G}$, i.e.

$$\tilde{\mathbf{Y}} = \mathbf{A}_{[n]\backslash G}\mathbf{X}_{[n]\backslash G} + \frac{1}{\sqrt{\mathsf{snr}}}\mathbf{W}, \tag{3.63}$$

then it is straightforward to show that an appropriately normalized version of the measurement model given by (3.63) obeys Assumptions MM1-MM3 with sampling rate $\tilde{\rho}$ and signal-to-noise ratio $\tilde{\mathsf{snr}}$.

Lower Bound on Genie-Aided Recovery

We now derive a necessary condition for recovery in the genie-aided setting. We begin with the following key fact: if the set G is chosen according to the selection strategy outlined above, the tuple $(\tilde{\mathbf{Y}}, \mathbf{A}, G)$ is a sufficient statistic for estimation of S^*. To see why, observe that

$$\begin{aligned} I(S^*; \mathbf{Y}, \mathbf{A}, G, \mathbf{X}_G) &= I(S^*; \mathbf{Y} - \mathbf{A}_G\mathbf{X}_G, \mathbf{A}, G, \mathbf{X}_G) && (3.64)\\ &= I(S^*; \tilde{\mathbf{Y}}, \mathbf{A}, G, \mathbf{X}_G) && (3.65)\\ &= I(S^*; \tilde{\mathbf{Y}}, \mathbf{A}, G) + I(S^*; \mathbf{X}_G | \tilde{\mathbf{Y}}, \mathbf{A}, G) && (3.66)\\ &= I(S^*; \tilde{\mathbf{Y}}, \mathbf{A}, G) && (3.67) \end{aligned}$$

where: (3.65) follow from the definition of $\tilde{Y}$; (3.66) follows from the chain rule for mutual information; and (3.66) follows from the fact that S^* and $\mathbf{X}_G$ are conditionally independent given the pair $(\tilde{\mathbf{Y}}, \mathbf{A}, G)$.

Let $\hat{S}$ denote the optimal estimate of the sparsity pattern in the genie-aided setting (i.e. the sparsity pattern estimate that minimizes the error probability). By the arguments above, we know that

$$S^* \to (\tilde{\mathbf{Y}}, \mathbf{A}, G) \to \hat{S} \tag{3.68}$$

forms a Markov chain. Also, by the optimality of $\hat{S}$ and the fact that distortion measure $d(\cdot, \cdot)$ corresponds to the maximum of the two detection error rates, it can also be shown that $\hat{S}$ contains the set G and has the same cardinality as S^*. Therefore, the sparsity pattern distortion can be expressed as

$$d(S^*, \hat{S}) = \Big(\frac{|S^*| - |G|}{|S^*|}\Big) d(S^* \backslash G, \hat{S} \backslash G). \tag{3.69}$$

Note that

$$\lim_{n\to\infty}\left(\frac{|S^*|-|G|}{|S^*|}\right)=\left(\frac{1-\kappa_X}{\kappa_X}\right)\left(\frac{\kappa_Z}{1-\kappa_Z}\right) \tag{3.70}$$

almost surely under Assumptions SS1-SS2.

We now arrive at the crux of the argument. Suppose that the distortion $\tilde{D}$ is not achievable for the tuple $(\tilde{\rho}, p_Z, \tilde{\mathsf{snr}})$. By (3.68) and the fact that the observation model given in (3.63) corresponds to the tuple $(\tilde{\rho}, p_Z, \tilde{\mathsf{snr}})$, it follows that the error probability

$$\Pr[d(S^*\backslash G, \hat{S}\backslash G)\geq \tilde{D}]$$

corresponding to the genie-aided setting is bounded away from zero for all n. By (3.69) and (3.70), it then follows that the distortion D is not achievable for the tuple $(\rho, p_X, \mathsf{snr})$. This concludes the proof of Theorem 3.2.

3.5.3 Proof of Theorem 3.4

One weakness of the proof of Theorem 3.1 is that the data processing inequality used to upper bound the mutual information $I(S^*;\mathbf{Y}|\mathbf{A})$ in (3.50) is not tight. In this proof, we derive a stronger upper bound that takes into account the fact that values of the nonzero elements are unknown. We assume throughout the proof that the nonzero entropy power N_X is strictly positive.

Using the chain rule for mutual information, $I(\mathbf{A}, S^*;\mathbf{Y}|\mathbf{A})$ can be written two ways as

$$\begin{aligned} I(S^*,\mathbf{X};\mathbf{Y}|\mathbf{A}) &= I(S^*;\mathbf{Y}|\mathbf{A})+I(\mathbf{X};\mathbf{Y}|S^*,\mathbf{A}) \\ &= I(\mathbf{X};\mathbf{Y},\mathbf{A})+I(S^*;\mathbf{Y}|\mathbf{X},\mathbf{A}). \end{aligned}$$

Since $\mathbf{S}\to\mathbf{X}\to(\mathbf{Y},\mathbf{A})$ forms a Markov chain, $I(S^*;\mathbf{Y}|\mathbf{X},\mathbf{A})$ is equal to zero and

$$I(S^*;\mathbf{Y}|\mathbf{A}) = I(\mathbf{X};\mathbf{Y}|\mathbf{A}) - I(\mathbf{X};\mathbf{Y}|S^*,\mathbf{A}). \tag{3.71}$$

Conceptually, term $I(\mathbf{X};\mathbf{Y}|S^*,\mathbf{A})$ quantifies the amount of $I(\mathbf{X};\mathbf{Y}|\mathbf{A})$ that is "used up" describing the values of the nonzero elements, and hence cannot contribute to estimation of the sparsity pattern.

Following the proof of Theorem 3.1, the first term on the right hand side of (3.71) can be upper bounded as

$$I(\mathbf{X};\mathbf{Y}|\mathbf{A})\leq\frac{1}{2}\mathbb{E}\Big[\log\det\left(I_{m\times m}+\mathsf{snr}\,V_X\mathbf{A}\mathbf{A}^T\right)\Big] \tag{3.72}$$

where the expectation is taken with respect to the random matrix $\mathbf{A}$.

To deal with the second term on the right hand side of (3.71) we first consider the case $m \leq k$. If we let

$$N(\mathbf{Z}) = \frac{1}{2\pi e} \exp\left(\frac{2}{m} h(\mathbf{Z})\right) \tag{3.73}$$

denote the entropy power of an m-dimensional random vector $\mathbf{Z}$, then it follows straightforwardly that

$$\begin{aligned} I(\mathbf{X}; \mathbf{Y} | S^* = S, \mathbf{A} = A) &= I(\mathbf{X}_S; \sqrt{\mathsf{snr}} A_S \mathbf{X}_S + \mathbf{W}) && (3.74)\\ &= h(\sqrt{\mathsf{snr}} A_S \mathbf{X}_S + \mathbf{W}) - h(\sqrt{\mathsf{snr}} A_S \mathbf{X}_S + \mathbf{W} | \mathbf{X}_S) && (3.75)\\ &= \frac{m}{2} \log\left(2\pi e \, N(\sqrt{\mathsf{snr}} A_S \mathbf{X}_S + \mathbf{W})\right) - \frac{m}{2} \log(2\pi e) && (3.76)\\ &= \frac{m}{2} \log\left(N(\sqrt{\mathsf{snr}} A_S \mathbf{X}_S + \mathbf{W})\right). && (3.77) \end{aligned}$$

Using two applications of the entropy power inequality (see e.g. [14]) we can write

$$\begin{aligned} N(\sqrt{\mathsf{snr}} A_S \mathbf{X}_S + \mathbf{W}) &\leq N(\sqrt{\mathsf{snr}} A_S \mathbf{X}_S) + N(\mathbf{W}) && (3.78)\\ &\leq \mathsf{snr} \left(\frac{N_X}{\kappa}\right) \det(A_S A_S^T)^{1/m} + 1, && (3.79) \end{aligned}$$

where $N_X = \kappa N(X_i | i \in S^*)$ denotes the nonzero entropy power of p_X. Note that the assumption $m \leq k$ is critical here since the determinant $A_S A_S^T$ is equal to zero for all $m < k$.

Plugging (3.79) back into (3.77) leads to

$$I(\mathbf{X}; \mathbf{Y} | S^*, \mathbf{A}) \geq \frac{m}{2} \mathbb{E}\left[\log\left(1 + \mathsf{snr} \, N_X \, \kappa^{-1} \det(\mathbf{A}_{S^*} \mathbf{A}_{S^*}^T)^{1/m}\right)\right] \tag{3.80}$$

where the expectation is with respect to the random matrix $\mathbf{A}_{\mathbf{S}^*}$.

Next we consider the case $m > k$. If the matrix A_S is full rank and we let $A_S^\dagger$ denote its Moore-Penrose pseudoinverse, we can write

$$\begin{aligned} I(\mathbf{X}; \mathbf{Y} | S^* = S, \mathbf{A} = A) &= I(\mathbf{X}_S; \sqrt{\mathsf{snr}} \, \mathbf{X}_S + A_S^\dagger \mathbf{W}) && (3.81)\\ &= h(\sqrt{\mathsf{snr}} \, \mathbf{X}_S + A_S^\dagger \mathbf{W}) - h(\sqrt{\mathsf{snr}} \mathbf{X}_S + A_S^\dagger \mathbf{W} | \mathbf{X}_S) && (3.82)\\ &= \frac{k}{2} \log\left(N(\sqrt{\mathsf{snr}} \, \mathbf{X}_S + A_S^\dagger \mathbf{W})\right) + \frac{1}{2} \log\det(A_S^T A_S) && (3.83)\\ &= \frac{k}{2} \log\left(1 + \mathsf{snr} \, N_X \, \kappa^{-1} \det(A_S^T A_S^T)^{1/k}\right) && (3.84) \end{aligned}$$

where (3.84) follows again from the entropy power inequality. Thus, we obtain

$$I(\mathbf{X}; \mathbf{Y} | S^*, \mathbf{A}) \geq \frac{k}{2} \mathbb{E}\left[\log\left(1 + \mathsf{snr} \, N_X \, \kappa^{-1} \det(\mathbf{A}_{S^*}^T \mathbf{A}_{S^*})^{1/k}\right)\right], \tag{3.85}$$

where the expectation is with respect to the random matrix $\mathbf{A}_{\mathbf{S}^*}$.

To characterize the asymptotic behavior of the bounds in (3.72), (3.80), and (3.85), we use the following results from random matrix theory.

Lemma 3.3. *[75] Let* $\mathbf{A}$ *denote an* $m \times n$ *random matrix whose entries are i.i.d. with mean zero and variance* $1/n$*. If* $m/n \to r$ *as* $n \to \infty$*, then*

$$\lim_{n\to\infty} \frac{1}{n} \log\det\left(I_{m\times m} + \gamma \mathbf{A}\mathbf{A}^T\right) = \mathcal{V}(r,\gamma) \tag{3.86}$$

almost surely where $\mathcal{V}(r,\gamma)$ *is given by* (3.18)*.*

Lemma 3.4. *[63] Let* $\mathbf{A}$ *denote an* $m \times n$ *random matrix whose entries are i.i.d. with mean zero and variance* $1/n$*. If* $m/n \to r$ *as* $n \to \infty$*, then*

$$\lim_{n\to\infty} \left(\det(\mathbf{A}\mathbf{A}^T)\right)^{1/m} = \left(\frac{1}{1-r}\right)^{1/r-1} \frac{1}{e}, \qquad \text{if } r < 1 \tag{3.87}$$

$$\lim_{n\to\infty} \left(\det(\mathbf{A}^T\mathbf{A})\right)^{1/n} = \left(\frac{r}{r-1}\right)^{r-1} \frac{1}{e}, \qquad \text{if } r > 1 \tag{3.88}$$

almost surely.

Combining Lemma 3.3 with the upper bound (3.72) leads immediately to

$$\limsup_{n\to\infty} \frac{1}{n} I(\mathbf{X};\mathbf{Y}|\mathbf{A}) \leq \frac{1}{2}\mathcal{V}(\rho, V_X\,\mathsf{snr}). \tag{3.89}$$

Similarly, combining Lemma 3.4 with the lower bounds (3.80) and (3.85) leads to

$$\liminf_{n\to\infty} \frac{1}{n} I(\mathbf{X};\mathbf{Y}|\mathbf{A}) \geq \frac{1}{2}\kappa\mathcal{V}_{LB}(\rho/\kappa, N_X\,\mathsf{snr}) \tag{3.90}$$

where $\mathcal{V}_{LB}(r,\gamma)$ is given by (3.20). Plugging these limits back into (3.71) and (3.47) completes the proof of Theorem 3.4

Chapter 4

Analysis of the Sampling Rate-Distortion Function

In this chapter, we show how the bounds on the sampling rate-distortion functions given Chapters 2 and 3 depend on the desired distortion D, the SNR, and various properties of the distribution p_X. We provide illustrations of the bounds and we characterize problem regimes in which the behavior of the algorithms is near-optimal and other regimes in which the behavior is highly suboptimal.

4.1 Preliminaries

4.1.1 Signal Classes

Following the problem formulation outlined in Section 1.4.2, a class of signals can be characterized by a class of limiting distributions $\mathcal{P}_X \subset \mathcal{P}(\kappa)$ where $\mathcal{P}(\kappa)$ is the class of all distributions with second moment equal to one and probability mass $1-\kappa$ at zero. To facilitate our analysis in the following sections, we introduce the following three classes:

- *Bounded:* We use $\mathcal{P}_{\text{Bounded}}(\kappa, B)$ to denote the class of all distributions $p_X \in \mathcal{P}(\kappa)$ such that
$$\Pr[|X| < B | X \neq 0] = 0$$
for some *lower bound* $B > 0$. Due to the second moment constraint, the lower bound B cannot exceed $1/\sqrt{\kappa}$.

- *Polynomial Decay:* We use $\mathcal{P}_{\text{Poly.}}(\kappa, L, \tau)$ to denote the class of all distributions $p_X \in \mathcal{P}(\kappa)$ such that
$$\lim_{x \to 0} \frac{\Pr[|X| \leq x | X \neq 0]}{x^L} = \tau$$
for some *polynomial decay rate* $L > 0$ and limiting constant $\tau \in (0, \infty)$.

- *Bernoulli-Gaussian:* We say that a distribution p_X is Bernoulli-Gaussian with sparsity κ if the nonzero part of p_X is zero-mean Gaussian, i.e. if

$$X \sim \begin{cases} 0, & \text{with probability } 1-\kappa \\ \mathcal{N}(0, \frac{1}{\kappa}), & \text{with probability } \kappa \end{cases}.$$

The bounded class corresponds to the setting where the nonzero entries in $\mathbf{x}$ have a fixed lower bound B on their magnitudes, independent of the vector length n. By contrast, the polynomial decay class corresponds to the setting where the magnitude of the $\lceil \beta\, k \rceil$'th smallest nonzero entry is proportional to $\beta^{1/L}$ for small β. Note that in the case of polynomial decay, a vanishing fraction of the nonzero entries are tending to zero as the vector length n becomes large.

The Bernoulli-Gaussian distribution is an example of a distribution with polynomial decay rate $L = 1$ and limiting constant $\tau = \sqrt{2/(\pi\kappa)}$.

4.1.2 Illustrations

In the following sections we provide illustrations of the bounds derived in Chapters 2 and 3 corresponding to either the Bernoulli-Gaussian distribution or the class of bounded distributions $\mathcal{P}_{\text{Bounded}}(\kappa, B)$ with lower bound $B = \sqrt{0.2/\kappa}$. Note that this choice of B means that the nonzero entries in $\mathbf{x}$ are lower bounded in squared magnitude by 20% of their average power.

The bounds corresponding to the Bernoulli-Gaussian distribution are optimized as a function of the relevant parameters. For the AMP-MMSE and MMSE bounds, this means that the true distribution p_X is used to define the conditional expectations. For the AMP-ST bound, this means that the threshold α is chosen to either minimize the distortion as a function of the sampling rate or to minimize the sampling rate as a function of the distortion.

In order to derive uniform bounds for the class of bounded distributions $\mathcal{P}_{\text{Bounded}}(\kappa, B)$, it is necessary to consider the worst-case distribution in the class. For the ML and linear estimators, these bounds are obtained straightforwardly by lower bounding the functions $P(D; p_X)$ and $\sigma^2_{\text{awgn}}(D; p_X)$ (see Proposition 4.5 below). For the AMP-ST we obtain a uniform bound by replacing the noise sensitivity $\mathcal{M}(\sigma^2, \alpha; p_X)$ in Theorem 2.6 with the upper bound $\mathcal{M}^*(\sigma^2, \alpha, \kappa)$ given in Section 4.9, and then optimizing the resulting expression as a function of the threshold α. Uniform bounds corresponding to the AMP-MMSE and MMSE cannot be derived using the results in this thesis, since these estimators depend on the true underlying distribution p_X.

In all illustrations, the lower bound is given by Theorem 3.5 from Chapter 3. We note that this bound the performance of the optimal recovery algorithm under Assumptions M1-M4.

All illustrations correspond to a sampling rate of $\kappa = 10^4$. The qualitative behavior of the bounds does not change significantly for sparsity rates within several orders of magnitude of this value.

4.2 Sampling Rate versus SNR

We begin our analysis of the bounds by studying the tradeoff between sampling rate and SNR. For a given recovery algorithm ALG, we use $\rho_\infty^{(\mathrm{ALG})}$ to denote the infinite SNR limit of the sampling rate-distortion function:

$$\rho_\infty^{(\mathrm{ALG})} = \lim_{\mathsf{snr}\to\infty} \rho^{(\mathrm{ALG})}. \tag{4.1}$$

This limit is a function of the pair (D, p_X) and may be interpreted as the sampling rate required in the absence of noise.

For the ML estimator, the infinite SNR limit of the upper bound in Theorem 2.1 is given by the sparsity rate κ, regardless of the distribution p_X and distortion D. Since it can be shown that the ML estimate is equivalent to random guessing whenever $\rho < \kappa$, we thus conclude that the infinite SNR limit of $\rho^{(\mathrm{ML})}$ is given explicitly by the piecewise constant function

$$\rho_\infty^{(\mathrm{ML})} = \begin{cases} \kappa, & \text{if } D \leq 1-\kappa \\ 0, & \text{if } D > 1-\kappa \end{cases}. \tag{4.2}$$

An upper bound on the rate at which $\rho^{(\mathrm{ML})}$ approaches its infinite SNR limit is given by the following result. The proof follows directly from the analysis of Theorem 2.1 given in Section 4.8.1.

Proposition 4.1. *For any nonzero distortion D and distribution p_X, there exists a constant C such that*

$$\rho^{(ML)} \leq \kappa + \frac{C}{\log(1+\mathsf{snr})}. \tag{4.3}$$

The following result is a consequence of Corollary 3.6 and shows that, under some additional assumptions on the pair (D, p_X), Proposition 4.1 is tight, in a scaling sense, with respect to the SNR.

Proposition 4.2. *Suppose that p_X can be expressed as*

$$p_X = (1-\kappa)\,\delta_0 + \omega_c\, p_{X_c} + (\kappa - \omega_c)\, p_{X_d} \tag{4.4}$$

where X_c is continuous with finite differential entropy $h(X_c)$ and X_d is discrete. Let $D < 1-\kappa$ be any distortion that satisfies

$$2H_b(\kappa_c) - 2\mathcal{H}\Big(\frac{\kappa}{\omega_c}D; \kappa_c\Big) > \kappa_c \log\left(\frac{\mathbb{E}[X_c^2] - \kappa_c(\mathbb{E}[X_c])^2}{N(X_c)}\right) + (1-\kappa_c)\log\left(\frac{1}{1-\kappa_c}\right) \tag{4.5}$$

where $\kappa_c = \omega_c/(1-\kappa+\omega_c)$ *and* $N(X_c) = (2\pi e)^{-1}\exp(2h(X_c))$. *Then, under Assumptions S1-S2 and M1-M4, there exists a constant* C *such that*

$$\rho > \omega_c + \frac{C}{\log(1+\mathsf{snr})} \tag{4.6}$$

is a necessary condition for any recovery algorithm.

Note that the constant ω_c in Proposition 4.2 is equal to the sparsity rate κ whenever the nonzero part of p_X is absolutely continuous with respect to Lebesgue measure. When this occurs, Propositions 4.1 and 4.2 characterize the fundamental behavior of the recovery problem for any distortion D satisfying (4.5).

For the linear and AMP estimators, it is straightforward to show that the infinite SNR limits can be expressed as

$$\rho_\infty^{(\mathrm{MF})} = \frac{1}{\sigma^2} \tag{4.7}$$

$$\rho_\infty^{(\mathrm{LMMSE})} = \frac{1}{1+\sigma^2} \tag{4.8}$$

$$\rho_\infty^{(\mathrm{AMP\text{-}ST})} = \mathcal{M}(\sigma^2, \alpha, p_x) \tag{4.9}$$

$$\rho_\infty^{(\mathrm{AMP\text{-}MMSE})} = \sup_{\tau > \sigma^2} \frac{\mathsf{mmse}(\tau; p_X)}{\tau} \tag{4.10}$$

where $\sigma^2 = \sigma^2_{\mathrm{awgn}}(D; p_X)$. By comparison with the ML limit, we see that each of these algorithms is strictly suboptimal at high SNR whenever its limit exceeds the sparsity rate κ.

For the MMSE estimator, the infinite SNR limit of the sampling rate predicted by the replica method in Theorem 2.7 is characterized by the infinite SNR limit of the noise power τ^* given in (2.37). It is easy to check that this limit is always less than or equal to κ, and thus the predicted MMSE infinite SNR limit is upper bounded by the ML infinite SNR limit.

The rate at which the achievable sampling rates converge to their infinite SNR limits is illustrated in Fig 4.1 for the Bernoulli-Gaussian distribution. The relative tightness of the ML upper bound and the information-theoretic lower bound from Chapter 3 provides rigorous verification of the MMSE behavior derived heuristically using the replica method. Moreover, as the SNR becomes large, the bounds corresponding to the AMP and linear estimate are significantly greater than the ML bounds, thus indicating that these methods are highly suboptimal at high SNR.

In Fig. 4.2, the infinite SNR limits corresponding to the Bernoulli-Gaussian distribution are shown as a function of the distortion. For this distribution, the MMSE limit is equal to the minimum of the ML and AMP-MMSE limits. When the distortion is relatively small (i.e. less than ≈ 0.9), the limits for ML, MMSE, and the information-theoretic lower bound are equal to the sparsity rate κ. When the distortion is relatively large, all of the bounds except for the ML bound converge to zero. If the goal is to minimize the distortion D as a function of the sampling rate ρ, then this behavior shows that ML is strictly suboptimal whenever the sampling rate ρ is strictly less than the sparsity rate κ.

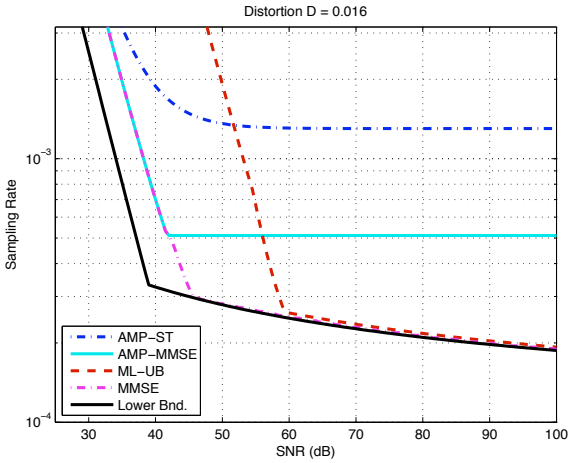

Figure 4.1: Bounds on the achievable sampling rate ρ as a function of the SNR when the nonzero entries are i.i.d. zero-mean Gaussian and the sparsity rate is $\kappa = 10^{-4}$.

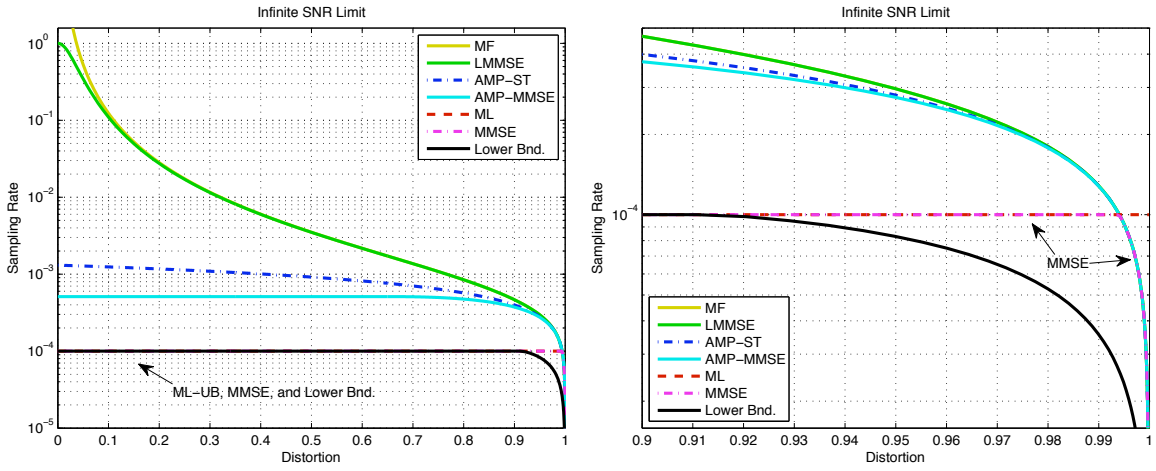

Figure 4.2: Bounds on the infinite SNR limit of the achievable sampling rate ρ as a function of the distortion D when the nonzero entries are i.i.d. zero-mean Gaussian and the sparsity rate is $\kappa = 10^{-4}$. The right panel highlights the large distortion behavior.

4.3 Stability Thresholds

For a given recovery algorithm ALG, we define the *stability threshold* as follows:

$$\varrho^{(\mathrm{ALG})} = \lim_{D \to 0} \lim_{\mathsf{snr} \to \infty} \rho^{(\mathrm{ALG})}. \tag{4.11}$$

This threshold is a function of the distribution p_X and may be interpreted as the sampling rate required for exact recovery in the absence of noise. For future reference, its significance is summarized in the following result.

Proposition 4.3. *Consider a fixed recovery algorithm ALG and distribution p_X with stability threshold $\varrho^{(ALG)}$.*

(a) If $\rho > \varrho^{(ALG)}$, then recovery is stable *in the sense that the distortion D can be made arbitrarily small by increasing the SNR.*

(b) If $\rho < \varrho^{(ALG)}$, then there exists a fixed lower bound on the achievable distortion D, regardless of the SNR.

Proof. This result follows immediately from the definition of the sampling rate-distortion function and the definition of the stability threshold in (4.11). □

Starting with the infinite SNR limits given in Section 4.2, it is straightforward to show that the stability thresholds of the recovery algorithms studied in Chapter 2 are given by

$$\varrho^{(\mathrm{ML})} = \kappa \tag{4.12}$$

$$\varrho^{(\mathrm{MF})} = \infty \tag{4.13}$$

$$\varrho^{(\mathrm{LMMSE})} = 1 \tag{4.14}$$

$$\varrho^{(\text{AMP-ST})} = \mathcal{M}_0(\alpha, \kappa) \tag{4.15}$$

$$\varrho^{(\text{AMP-MMSE})} = \sup_{\tau > 0} \frac{\mathsf{mmse}(\tau; p_X)}{\tau} \tag{4.16}$$

where $M_0(\alpha, \kappa)$ is given by Eq. (4.72) in Section 4.9.

The ML stability threshold corresponds to the well known fact that $m = k + 1$ random linear projections are, with probability one, sufficient to recover an arbitrary k-sparse vector. The LMMSE stability threshold corresponds to the fact that $m = n$ linearly independent projections are sufficient to recover an arbitrary vector of length n. The AMP-ST stability threshold, which depends only on the sparsity rate of the distribution p_X, has been studied previously in [22] where it is shown that $\min_\alpha \mathcal{M}_0(\alpha, \kappa)$ corresponds to the ℓ_1/ℓ_0 equivalence threshold of Donoho and Tanner [19]. The AMP-MMSE threshold has, to the best of our knowledge, not been studied previously.

Starting with Proposition 4.2, it can also be shown that the stability threshold of the optimal recovery algorithm is lower bounded by

$$\varrho^{(\text{Lower Bnd.})} = \omega_c \tag{4.17}$$

for any distribution p_X for which the strict inequality in (4.5) holds with $D = 0$. In many cases, this lower bound is equal to the sparsity rate κ.

Finally, using the analysis of the MMSE bound provided in Section 4.8.2, it can be shown that the stability threshold of the MMSE estimator, as predicted by the replica method, is given by

$$\varrho^{(\text{MMSE})} = \lim_{\tau \to 0} \frac{\mathsf{mmse}(\tau; p_X)}{\tau} \tag{4.18}$$

when the limit exists. The right hand side of (4.18) is referred to as the *MMSE dimension* of the distribution p_X by the authors in [81], and it is equal to the weight on the continuous part of p_X whenever p_X is a purely continuous-discrete mixture.

In Fig. 4.2, the stability thresholds corresponding to the Bernoulli-Gaussian distribution correspond to the zero distortion limit (i.e. the intersection with the y-axis).

4.4 Distortion versus Sampling Rate

We now turn our attention to the tradeoff between the achievable distortion and the sampling rate. We begin with a precise characterization of the low-distortion behavior.

Proposition 4.4. *The low-distortion behavior corresponding to a fixed pair* (snr, p_X) *is given by*

$$\lim_{D \to 0} \left(\frac{P(D; p_X)}{\mathcal{H}(D; \kappa)} \right) \rho^{(ML\text{-}UB)} = \left(\frac{2}{3 - \sqrt{8}} \right) \frac{1}{\mathsf{snr}} \tag{4.19}$$

$$\lim_{D \to 0} \sigma^2_{awgn}(D, p_X)\, \rho^{(MF)} = \frac{1}{\mathsf{snr}} + 1 \tag{4.20}$$

$$\lim_{D \to 0} \sigma^2_{awgn}(D, p_X)\, \rho^{(ALG)} = \frac{1}{\mathsf{snr}} \tag{4.21}$$

where (4.21) *holds for the LMMSE, AMP-MMSE, AMP-ST, and MMSE recovery algorithms.*

Proof. The limits corresponding to the ML and MMSE estimators are proved in Appendices 4.8.1 and 4.8.2 respectively. The limits corresponding to the linear estimators follow immediately from the fact that $\sigma^2_{\text{awgn}}(D, p_X) \to 0$ as $D \to 0$. For the AMP-ST estimator, we use the additional fact that the noise sensitivity $\mathcal{M}(\sigma^2, \alpha, p_X)$ is bounded (see Section 4.9)

and hence $\sigma^2 \mathcal{M}(\sigma^2, \alpha, p_X) \to 0$ as $\sigma^2 \to 0$. For the AMP-MMSE estimator, we use the bound

$$\left| \sigma^2(D; p_X)\rho^{(\text{AMP-MMSE})} - \frac{1}{\mathsf{snr}} \right| \leq \sigma^2(D; p_X) \sup_{\tau > 0} \left\{ \frac{\mathsf{mmse}(\tau; p_X)}{\tau} \right\} \tag{4.22}$$

and note that the right hand side of (4.22) becomes arbitrarily small as $D \to 0$. □

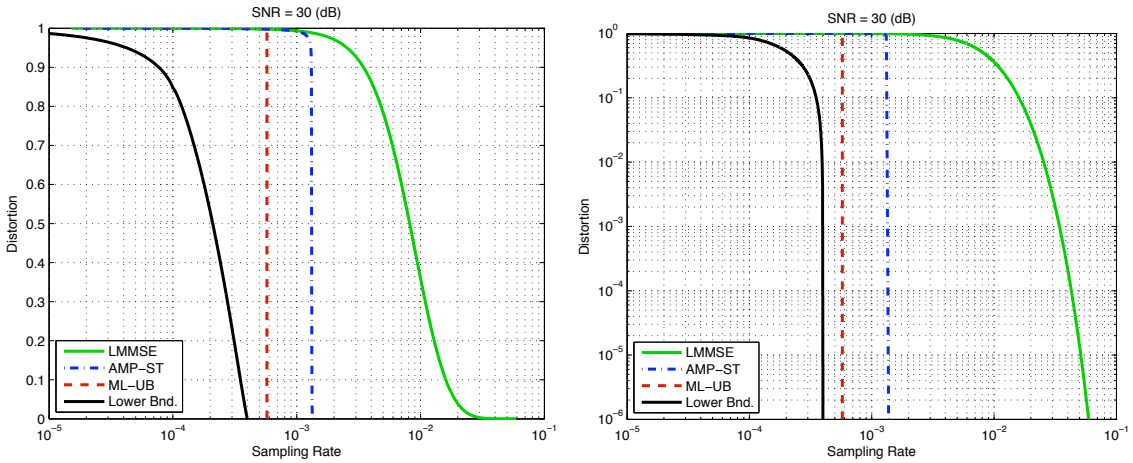

Figure 4.3: Bounds on the achievable distortion D as a function of the sampling rate ρ when the nonzero entries are lower bounded in squared magnitude by 20% of their average power, but are otherwise arbitrary and the sparsity rate is $\kappa = 10^{-4}$. The MF bound is comparable to the LMMSE bound and is not shown.

In words, Proposition 4.4 says that as the desired distortion D becomes small, the ML upper bound is inversely proportional to the ratio $P(D; p_X)/\mathcal{H}(D; \kappa)$ whereas the low distortion behavior of the remaining bounds is inversely proportional to the function $\sigma^2_{\text{awgn}}(D; p_X)$. The behavior of these terms is characterized for the bounded and polynomial decay signal classes in the following results.

Proposition 4.5 (Bounded)**.** *If* $p_X \in \mathcal{P}_{Bounded}(\kappa, B)$*, then*

$$\frac{P(D; p_X)}{\mathcal{H}(D; \kappa)} \geq \frac{B^2}{2[\log\left(1/D\right) + 1 + \log\left(\frac{1-\kappa}{\kappa}\right)]} \tag{4.23}$$

and

$$\sigma^2_{awgn}(D; p_X) \geq \frac{B^2}{8[\log(1/D) + \log(\frac{1-\kappa}{\kappa})]}. \tag{4.24}$$

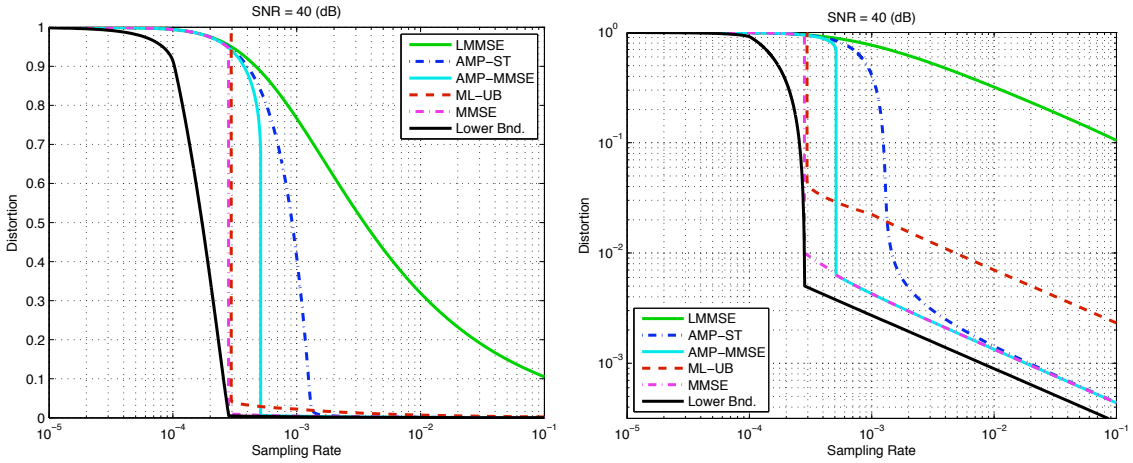

Figure 4.4: Bounds on the achievable distortion D as a function of the sampling rate ρ when the nonzero entries are i.i.d. zero-mean Gaussian and the sparsity rate is $\kappa = 10^{-4}$. The MF bound is comparable to the LMMSE bound and is not shown.

Proposition 4.6 (Polynomial Decay)**.** *If* $p_X \in \mathcal{P}_{Poly.}(\kappa, L, \tau)$*, then*

$$\lim_{D\to 0}\left(\frac{\log(1/D)}{D^{2/L}}\right)\frac{P(D;p_X)}{\mathcal{H}(D;\kappa)} = \frac{\tau^{-2/L}}{2(1+2/L)} \tag{4.25}$$

and

$$\lim_{D\to 0}\left(\frac{\log(1/D)}{D^{2/L}}\right)\sigma^2_{awgn}(D;p_X) = \frac{\tau^{-2/L}}{2}. \tag{4.26}$$

The proofs of Propositions 4.5 and 4.6 are given in Appendices 4.8.3 and 4.8.4 respectively.

One way to interpret these results is to think of the achievable distortion as a function of the sampling rate ρ. For a given tuple $(\rho, p_X, \mathsf{snr})$ and recovery algorithm ALG, we use $D^{(\mathrm{ALG})}$ to denote the smallest achievable distortion, i.e.

$$D^{(\mathrm{ALG})} = \inf\{D \geq 0 \,:\, D \text{ is achievable}\}. \tag{4.27}$$

An upper bound on the rate at which $D^{(\mathrm{ALG})}$ decreases as the sampling rate becomes large is given in the following result, which is an immediate consequence of Propositions 4.4, 4.5, and 4.6.

Proposition 4.7. *Consider a fixed pair* (snr, p_X)*, and let ALG denote one of the ML, MF, LMMSE, AMP-MMSE, AMP-ST, MMSE recovery algorithms.*

(a) If $p_X \in \mathcal{P}_{Bounded}(\kappa, B)$ *then there exists a constant* C *such that*

$$D^{(ALG)} \leq \exp(-C\,\rho) \tag{4.28}$$

for all sampling rates $\rho > 0$*.*

(b) If $p_X \in \mathcal{P}_{Poly.}(\kappa, L, \tau)$ then there exists a constant C such that

$$\left(\frac{1}{D^{(ALG)}}\right)^{2/L} \log\left(\frac{1}{D^{(ALG)}}\right) \leq C\,\rho \tag{4.29}$$

for all sampling rates $\rho > 0$.

Proposition 4.7 shows that the low-distortion behavior depends critically on the behavior of the distribution p_X around the point $x = 0$. If the nonzero part of the distribution is bounded away from zero, then the distortion decays exponentially rapidly with the sampling rate. Conversely, if the nonzero part of p_X has a polynomial decay rate $L > 0$, then the distortion decays polynomially rapidly with the sampling rate, with an exponent that converges to $L/2$.

Using Corollary 3.3, it can be shown that the scaling behavior in Proposition 4.7 is optimal in the sense that, up to constants, no recovery algorithm can do any better. Consequently, each of the algorithms presented in this Chapter 2 is optimal in a scaling sense as the SNR becomes large whenever the sampling rate is strictly greater than the stability threshold.

The behavior of the achievable distortion D as a function of the sampling rate ρ is illustrated in Fig. 4.3 for the class of bounded distributions $\mathcal{P}_{\text{Bounded}}(\kappa, B)$ with $B = \sqrt{0.2/\kappa}$. In accordance with part (a) of Proposition 4.7, the LMMSE bound decays exponentially rapidly as a function the sampling rate. The same scaling behavior also occurs for the ML and AMP-ST bounds as well as the lower bound from Chapter 3. However, due to the relatively large SNR, this behavior occurs only for distortions much less than 10^{-6} and is therefore not visible in the range of distortions plotted in Fig. 4.3.

For comparison, the same behavior is illustrated in Fig. 4.4 for a Bernoulli-Gaussian distribution which has decay rate $L = 1$. In accordance with part (b) of Proposition 4.7, the distortion decays polynomially with rate $1/2$. Interestingly, the AMP-MMSE and AMP-ST bounds converge to the MMSE bound, and are within a constant factor ≈ 1.18 of the lower bound. This behavior shows that these algorithms are near-optimal when the sampling rate is relatively large. We suspect that the gap between these algorithms and the ML upper bound is due primarily to looseness in our bounding technique.

4.5 Distortion versus SNR

The previous section showed that computationally efficient algorithms can be near-optimal when the sampling rate is large. In the context of compressed sensing, a more interesting question is whether or not these same algorithms can be near-optimal when the sampling rate is fixed, and much less than one. In this section, we show that the answer to this question is 'yes', provided that the sampling rate is strictly greater than the stability threshold of the algorithm.

For a given tuple (D, ρ, p_X) and recovery algorithm ALG, we let $\mathsf{snr}^{(\mathrm{ALG})}$ denote the infimum over all $\mathsf{snr} \geq 0$ such that D is achievable, i.e.

$$\mathsf{snr}^{(\mathrm{ALG})} = \inf\{\mathsf{snr} \geq 0 \,:\, D \text{ is achievable}\}.$$

The following result characterizes the low-distortion behavior with respect to the SNR.

Proposition 4.8. *The low-distortion behavior corresponding to a fixed pair (ρ, p_X) is given by*

$$\lim_{D \to 0} \left(\frac{P(D; p_X)}{\mathcal{H}(D; \kappa)} \right) \mathsf{snr}^{(\textit{ML-UB})} = \left(\frac{2}{3 - \sqrt{8}} \right) \frac{1}{\rho - \kappa} \tag{4.30}$$

if $\rho > \kappa$, and

$$\lim_{D \to 0} \sigma^2_{\textit{awgn}}(D, p_X)\, \mathsf{snr}^{(\textit{ALG})} = \frac{1}{\rho - \varrho^{(\textit{ALG})}} \tag{4.31}$$

if $\rho > \varrho^{(\textit{ALG})}$ where (4.31) holds for the LMMSE, AMP-MMSE, AMP-ST, and MMSE recovery algorithms.

Proof. The limits corresponding to the ML and MMSE recovery algorithms are proved in Appendices 4.8.1 and 4.8.2 respectively. The limits corresponding to the LMMSE and AMP recovery algorithms follow straightforwardly along the same lines as the proof of Proposition 4.4. □

Proposition 4.8 is analogous to Proposition 4.4 except that it is valid only if the sampling rate ρ exceeds the stability threshold. The reason that Proposition 4.8 does not provide a bound for the MF estimator is that the stability threshold of the MF estimator is infinite, and thus the corresponding limit in (4.31) is not defined.

Combining Proposition 4.8 with Propositions 4.5 and 4.6 leads to the following result, which bounds the rate at which $D^{(\mathrm{ALG})}$ decreases as the SNR becomes large.

Proposition 4.9. *Consider a fixed pair (ρ, p_X), and let ALG denote one of the ML, LMMSE, AMP-MMSE, AMP-ST, MMSE recovery algorithms.*

(a) If $p_X \in \mathcal{P}_{\textit{Bounded}}(\kappa, B)$ and $\rho > \varrho^{(\textit{ALG})}$, then there exists a constant C such that

$$D^{(\textit{ALG})} \leq \exp(-C\, \mathsf{snr}) \tag{4.32}$$

for all $\mathsf{snr} > 0$.

(b) If $p_X \in \mathcal{P}_{\textit{Poly.}}(\kappa, L, \tau)$ and $\rho > \varrho^{(\textit{ALG})}$, then there exists a constant C such that

$$\left(\frac{1}{D^{(\textit{ALG})}} \right)^{2/L} \log \left(\frac{1}{D^{(\textit{ALG})}} \right) \leq C\, \mathsf{snr} \tag{4.33}$$

for all $\mathsf{snr} > 0$.

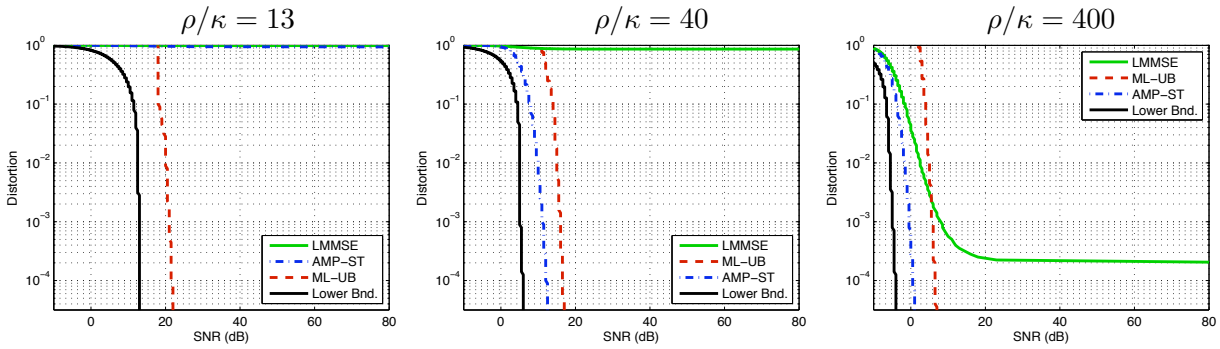

Figure 4.5: Bounds on the achievable distortion D as a function of the SNR for three different sampling rates ρ when the nonzero entries are lower bounded in squared magnitude by 20% of their average power, but are otherwise arbitrary and the sparsity rate is $\kappa = 10^{-4}$. The MF bound is comparable to the LMMSE bound and is not shown.

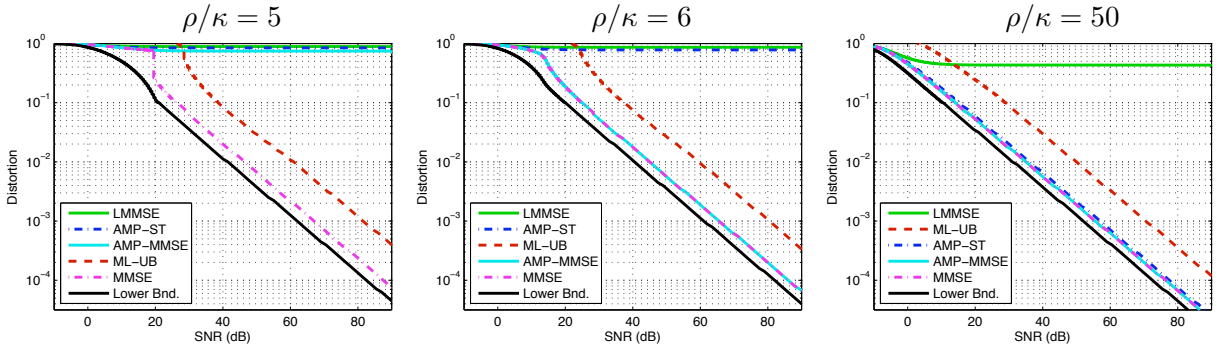

Figure 4.6: Bounds on the achievable distortion D as a function of the SNR for three different sampling rates ρ when the nonzero entries are i.i.d. zero-mean Gaussian and the sparsity rate is $\kappa = 10^{-4}$. The MF bound is comparable to the LMMSE bound and is not shown.

Using Corollary 3.3 in Chapter 3, it can be shown that the scaling behavior in Proposition 4.9 is optimal in the sense that, up to constants, no recovery algorithm can do any better. Consequently, each of the algorithms presented in this Chapter 2 (except for the MF estimator) is optimal in a scaling sense as the SNR becomes large whenever the sampling rate is strictly greater than the stability threshold.

The behavior of the achievable distortion $D^{(\mathrm{ALG})}$ as a function of the SNR is illustrated in Fig. 4.5 for three different sampling rates ρ and the class of bounded distributions $\mathcal{P}_{\mathrm{Bounded}}(\kappa, B)$ with $B = \sqrt{0.2/\kappa}$. In the left panel, the sampling rate is greater than $\varrho^{(\mathrm{ML})}$ but less than $\varrho^{(\mathrm{AMP\text{-}ST})}$ and $\varrho^{(\mathrm{LMMSE})}$. In accordance with part (a) of Proposition 4.9, the ML distortion decays exponentially rapidly whereas the AMP-ST and LMMSE distortions are bounded away from zero. In the second panel, the sampling rate is greater than $\varrho^{(\mathrm{ML})}$ and $\varrho^{(\mathrm{AMP\text{-}ST})}$ but less than $\varrho^{(\mathrm{LMMSE})}$, and hence the AMP-ST distortion also decays exponentially rapidly. In the third panel, ρ is relatively large but still less than $\varrho^{(\mathrm{LMMSE})}$. Thus, even though the LMMSE distortion is less than it was before, it is still bounded away from zero.

For comparison, the same behavior is illustrated in Fig. 4.6 for a Bernoulli-Gaussian distribution which has decay rate $L = 1$. In accordance with part (b) of Proposition 4.9, the distortion of each algorithm decays polynomially with rate $1/2$ whenever the sampling rate is greater than the stability threshold of the algorithm. It is interesting to note that the relatively small difference in sampling rates between the left and middle panels marks the boundary between the setting where all of the computationally feasible algorithms studied in this Chapter 2 are highly suboptimal and the setting where the distortion of the computationally feasible AMP-MMSE algorithm, is within a constant factor ≈ 1.75 of the lower bound.

4.6 Rate-Sharing Matrices

All of the bounds presented in Chapter 2 assume that the measurement matrix $\mathbf{A}$ has i.i.d. entires (Assumption M4). A natural question then, is whether relaxing this assumption can lead to better performance. Interestingly, the answer to this question can be 'yes'. In this section, we show that certain *rate-sharing* matrices can achieve points in the sampling rate-distortion region that are impossible using i.i.d. matrices.

The concept of rate-sharing is analogous to the idea of time-sharing in communications and can be summarized as follows. By using an appropriately constructed block-diagonal measurement matrix it is possible to separate the recovery problem into two subproblems, each of which is statistically identical to the original problem. By assigning different sampling rates to each of the subproblems and then combining the resulting sparsity pattern estimates, it is possible to achieve an effective sampling rate-distortion pair (ρ, D) that is a *linear combination* of the sampling rate-distortion pairs for each of the subproblems.

Construction of a rate-sharing matrix: For a fixed pair (snr, p_X) and recovery algorithm ALG, let (ρ_1, D_1) and (ρ_2, D_2) be two achievable sampling rate-distortion pairs. Let $\{\mathbf{A}_1(n)\}_{n\geq 1}$ and $\{\mathbf{A}_2(n)\}_{n\geq 1}$ be sequences of measurement matrices obeying Assumptions M1-M3 that achieve these rates. Then, for any $\lambda \in [0,1]$, a sequence of *rate-sharing* matrices is given by

$$\mathbf{A}(n) = \begin{bmatrix} \mathbf{A}_1(\lceil \lambda\, n \rceil) & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2(n - \lceil \lambda\, n \rceil) \end{bmatrix} \mathbf{P}(n) \tag{4.34}$$

where $\mathbf{0}$ denotes a matrix of zeros and $\mathbf{P}(n)$ is a random matrix distributed uniformly over the set of $n \times n$ permutation matrices.

Recovery using a rate-sharing matrix: For a problem of size n, the measurements $\mathbf{Y}$ made using the rate-sharing matrix $\mathbf{A}$ can be expressed as

$$\begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{X}}_1 \\ \tilde{\mathbf{X}}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \end{bmatrix}$$

where $\tilde{\mathbf{X}} = [\tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2]^T = \mathbf{P}\mathbf{x}$ corresponds to a random permutation of the entries in $\mathbf{x}$. To recover the sparsity pattern of $\mathbf{x}$ from these measurements, the recovery algorithm performs the following two steps:

(1) Individually estimate the sparsity patterns of $\tilde{\mathbf{X}}_1$ and $\tilde{\mathbf{X}}_2$ assuming a sparsity rate of κ for each vector.

(2) Use these estimates to produce an estimate $\hat{S}$ of the sparsity pattern of $\mathbf{x}$.

Proposition 4.10 (Rate-Sharing)**.** *For a fixed pair* (snr, p_X) *and algorithm ALG, let* (ρ_1, D_1) *and* (ρ_2, D_2) *be two achievable sampling rate-distortion pairs. Then, for any parameter* $\lambda \in [0,1]$*, the sampling rate-distortion pair* (ρ, D) *given by*

$$\rho = \lambda\rho_1 + (1-\lambda)\rho_2 \tag{4.35}$$

$$D = \lambda D_1 + (1-\lambda)D_2 \tag{4.36}$$

is achievable using the rate-sharing strategy outlined above.

Proof. Based on the assumptions on $\mathbf{A}_1(n)$ and $\mathbf{A}_2(n)$ and the fact that

$$\|\mathbf{A}(n)\|_F^2 = \|\mathbf{A}_1(\lceil \lambda\, n \rceil)\|_F^2 + \|\mathbf{A}_2(n - \lceil \lambda\, n \rceil)\|_F^2,$$

it is straightforward to verify that the sequence of rate-sharing matrices $\{\mathbf{A}(n)\}_{n\geq 1}$ defined by (4.34) satisfies Assumptions M1-M3 with sampling rate $\rho = \lambda\rho_1 + (1-\lambda)\rho_2$.

The next step is to verify that the distortion D is achievable. Since each permutation $\mathbf{P}(n)$ is independent of the vector $\mathbf{x}(n)$, the random sequences $\{\tilde{\mathbf{X}}_1(n)\}_{n\geq 1}$ and $\{\tilde{\mathbf{X}}_2(n)\}_{n\geq 1}$

obey Assumptions S1-S3 with probability one. Since the pairs (ρ_1, D_1) and (ρ_2, D_2) are achievable, it thus follows that the distortions D_1 and D_2 are achievable for the individual sparsity pattern estimates made in step (1).

Now, for a given problem of size n, let S_1^*, $\hat{S}_1$, S_2^*, $\hat{S}_2$ denote the true and estimated sparsity patterns corresponding to the vectors $\tilde{\mathbf{X}}_1$ and $\tilde{\mathbf{X}}_2$, and let S^* and $\hat{S}$ denote the true and estimated sparsity pattern of $\mathbf{x}$. As a simple exercise, it can be verified that

$$d(S^*, \hat{S}) \leq \lambda_n d(S_1^*, \hat{S}_1) + (1 - \lambda_n) d(S_2^*, \hat{S}_2)$$

where

$$\lambda_n = \frac{\max(|S_1^*|, |\hat{S}_1|)}{\max(|S_1^*|, |\hat{S}_1|) + \max(|S_2^*|, |\hat{S}_2|)}.$$

Using the arguments outlined above, it can then be verified that $\lambda_n \to \lambda$ almost surely as $n \to \infty$, and thus we conclude that the distortion $D = \lambda D_1 + (1 - \lambda) D_2$ is achievable. □

As an immediate consequence of Proposition 4.10, we have the following result.

Corollary 4.1. *For a fixed pair* (snr, p_X) *and algorithm ALG the sampling rate-distortion function is a convex function of the distortion* D*.*

By comparing the convexified versions of the achievable bounds in Chapter 2 with the lower bounds developed in Chapter 3 for matrices obeying Assumptions M1-M4, it can be verified that there are cases where rate-sharing (even with a potentially suboptimal recovery algorithm) is strictly better than using an i.i.d. matrix and the optimal recovery algorithm. This difference is most dramatic in the high SNR setting when the sampling rate is relatively small compared to the sparsity rate.

4.7 Discussion of Bounds

In this section, we review the main contributions Chapters 1-3 and discuss various implications of our analysis.

4.7.1 Fundamental Behavior of Sparsity Pattern Recovery

The achievable bounds derived in Chapter 2, in conjunction with the information-theoretic lower bounds in Chapter 3 characterize the fundamental limit of what cannot be recovered in presence of noise. A major technical contribution of this Chapter 2 is the upper bound on the sampling rate-distortion function for the maximum likelihood estimator (Theorem 2.1). To our knowledge, this is the only achievable bound in the literature that converges to the noiseless limit as the SNR becomes large and correctly characterizes the high SNR behavior.

Our bounds show that the tradeoffs between the sampling rate ρ, the distortion D, and the SNR can be characterized in terms of several key properties of the limiting distribution

p_X. Roughly speaking, the high-SNR behavior is characterized by the differential entropy of the nonzero part of p_X whereas the low-distortion behavior is characterized by the behavior of the distribution around the point $x = 0$. These dependencies can be summarized as follows:

- *High-SNR Behavior:* If the nonzero part of p_X has a relatively large differential entropy, then the tradeoff between sampling rate and SNR is given by

$$\rho \approx \kappa + \frac{C}{\log(\mathsf{snr})}.$$

- *Low-SNR Behavior:* If the nonzero part of p_X has a polynomial decay L, then the tradeoff between sampling rate and distortion is given by

$$\rho \approx C \cdot (\tfrac{1}{D})^{1/L} \log(\tfrac{1}{D}),$$

  and the tradeoff between SNR and distortion is given by

$$\mathsf{snr} \approx C \cdot (\tfrac{1}{D})^{1/L} \log(\tfrac{1}{D}) \quad \text{if} \quad \rho > \kappa,$$

  where the condition $\rho > \kappa$ is necessary if the nonzero part of p_X has a relatively large differential entropy. Note that $L = 0$ if the nonzero part of p_X is bounded away from zero.

The high-SNR behavior of the bounds is illustrated in Figures 1.1, 4.1, and 4.2. The low-distortion behavior is illustrated in Figures 4.3, 4.4, 4.5, and 4.6.

4.7.2 Near-Optimality of Efficient Algorithms

From a practical standpoint, a key question is whether or not a particular computationally efficient algorithm is near-optimal. A positive answer to this question means that more complicated algorithms are unnecessary. A negative answer, however, suggests that it is worth investing resources in the design and implementation of better algorithms.

In the absence of measurement noise, the tradeoffs for existing algorithms have been relatively well understood. For example, the number of measurements m needed for exact recovery of a k-sparse vector of length n can be summarized as follows: linear recovery (i.e. solving a system of full rank linear equations) requires $m \geq n$; linear programming requires $m \geq C \cdot k \log(n/k)$ for some constant C; and an NP-hard exhaustive search requires $m \geq k + 1$.

One of the contributions of this thesis, has been to extend the understanding of these tradeoffs to practically motivated settings where, due to measurement noise, only approximate recovery is possible. Interestingly, our results show that there are problem regimes

where existing computationally efficient algorithms—such as linear estimation or approximate message passing—are near-optimal and other regimes where they are highly suboptimal.

For example, the dependence of the sampling rate on the SNR illustrated in Fig. 1.1 shows that computationally simple algorithms are near-optimal at low SNR, but suggests that increasing sophistication is required as the SNR increases.

Moreover, the bounds illustrated in Fig. 4.6 show that a small change in the sampling rate can make the crucial difference between whether or not approximate message passing achieves the optimal tradeoff between SNR and distortion.

4.7.3 Comparison with Replica Predictions

In this thesis, we provide a comparison of rigorous bounds with the nonrigourous analysis of the replica method (Theorem 2.7). Since the predictions of the replica method are sharp, they provide valuable insights about where our bounds are tight and where they can be improved. For example, in Fig. 4.1 there exists a gap between the upper and lower bounds for SNR in the range of 45 to 60 dB. In this region, the replica prediction suggests that the information-theoretic lower bound from Chapter 3 is essentially correct and that the ML upper bound is loose.

An additional contribution of this comparison, is that the relative tightness of our rigorous bounds provides evidence in support of the unproven replica assumptions. For example, in Fig. 4.1, the upper and lower bounds are extremely close and sandwich the replica prediction for all SNR greater than 60 dB. Despite a vast amount of work on this topic, such evidence has been notoriously difficult to come by.

4.7.4 Universality of Bounds

To characterize the limiting behavior of a sequence of vectors we assume convergence of the empirical distributions (Assumption S2). If the limiting distribution is known, it is possible to use optimized recovery algorithms based on the distribution (e.g. the AMP-MMSE and MMSE recovery algorithms). In many cases, however, the limiting distribution is unknown. To address these settings, we develop bounds for fixed estimators which hold uniformly over a class of limiting distributions such as the class of all distributions bounded away from zero or the class of all distributions with polynomial decay (see Section 4.1.1) .

Our results show that, in many cases, prior information about the limiting distribution does not help significantly. For example, in the right panel of Fig. 1.1, the upper and lower bounds on the sampling rate-distortion function are relatively tight, uniformly over the class of distributions bounded away from zero. Another example is given by Propositions 4.4 and 4.8 which show that the low distortion behavior depends entirely on certain properties of the underlying distributions (specifically, the behavior of the distribution around the point $x = 0$).

We remark that an important counterexample occurs if the limiting distribution is supported on a finite subset of the real line (see Corollary 3.7). Then, the high-SNR sampling rate-distortion behavior can depend crucially on prior information about the distribution.

4.7.5 Role of Model Assumptions

This thesis focusses on the setting where a constant fraction of the entries are nonzero (Assumption S1). In Section 1.4.4 it is shown that the results in this thesis still hold when all but a fraction κ of the entries in $\mathbf{x}$ are tending to zero as n becomes large. In principle, many of the tools developed in the thesis could also be used to address settings where the number of nonzero entries grows sub-linearly with the vector length, and hence there is a vanishing fraction of nonzero entries.

Our use of row normalization (Assumption M3) differs from many related works which use column normalization. The reason for our scaling is that, from a sampling perspective, one way to decrease the effect of noise is to take additional samples (all at a fixed per-measurement SNR). If the column norms of the measurement matrix are constrained, then this is not possible since the per-measurement SNR will necessarily decrease as the number of measurements increases. Since it is assumed throughout that the sampling rate ρ is a fixed constant, all results in this thesis can be compared to existing works under an appropriate rescaling of the SNR.

The proofs of our upper bounds rely heavily on the assumption that the measurement matrices have i.i.d. entries (Assumption M4). The proofs of Theorem 2.1 and 2.4 further assume that these entries are Gaussian (Assumption M5). The extent to which these assumptions can be relaxed is an important direction for future research.

In Section 4.6 it is shown that rate-sharing matrices (which are not i.i.d.) can convexify the sampling rate-distortion region, thus leading to better performance. This result shows that i.i.d. matrices are strictly suboptimal in some settings.

4.8 Scaling Behavior

This section provides additional analysis of the sampling rate-distortion bounds presented in Chapter 2.

4.8.1 Behavior of the ML Upper Bound

This section studies the scaling behavior of the upper bound $\rho^{(\text{ML-UB})}$ given in Theorem 2.1. For notational simplicity, we will use the notation $\Lambda(D)$, $P(D)$ and $\mathcal{H}(D)$ where the dependence on snr and p_X is implicit. Recall that the upper bound is given by

$$\rho^{(\text{ML-UB})} = \kappa + \max_{\tilde{D} \in [D,1]} \Lambda(D).$$

We first consider the behavior as $D \to 0$. Note that the function $\Lambda(D)$ is finite for all $D > 0$ but grows without bound as $D \to 0$, and hence

$$\lim_{D\to 0} \frac{P(D)}{\mathcal{H}(D)} \max_{\tilde{D}\in[D,1]} \Lambda(\tilde{D}) = \lim_{D\to 0} \frac{P(D)}{\mathcal{H}(D)} \Lambda(D). \tag{4.37}$$

Starting with the definition of $\Lambda(D)$ given in (2.5), it is straightforward to show that

$$\lim_{D\to 0} \frac{P(D)}{\mathcal{H}(D)} \Lambda(D) = \frac{2}{\mathsf{snr}} \lim_{D\to 0} \lambda(D) \tag{4.38}$$

where

$$\lambda(D) = \min_{\theta,\mu\in(0,1)} \max\left\{ \frac{4}{(1-\theta)^2}, \frac{1}{\mu\theta} - \frac{D\kappa \log(1-\mu^2)}{2\mu\theta\mathcal{H}(D)} \right\}. \tag{4.39}$$

Using the fact that $D/\mathcal{H}(D) \to 0$ as $D \to 0$ gives

$$\lim_{D\to 0} \lambda(D) = \min_{\theta\in(0,1)} \max\left\{ \frac{4}{(1-\theta)^2}, \frac{1}{\theta} \right\} = \frac{1}{3-\sqrt{8}},$$

and putting everything together gives

$$\lim_{D\to 0} \frac{P(D)}{\mathcal{H}(D)} \left[\rho^{(\text{ML-UB})} - \kappa \right] = \left(\frac{2}{3-\sqrt{8}} \right) \frac{1}{\mathsf{snr}}.$$

We next consider the behavior as a function of the SNR. For any $D > 0$ it is easy to verify that $\Lambda(D) \to 0$ as $\mathsf{snr} \to \infty$ and hence the infinite SNR limit is given by

$$\lim_{\mathsf{snr}\to\infty} \rho^{(\text{ML-UB})} = \kappa. \tag{4.40}$$

To characterize the rate at which the upper bound approaches this limit, let $D > 0$ be fixed and observe that

$$\begin{aligned}
&\lim_{\mathsf{snr}\to\infty} \log(\mathsf{snr}) \left[\rho^{(\text{ML-UB})} - \kappa \right] \\
&= \lim_{\mathsf{snr}\to\infty} \log(\mathsf{snr}) \max_{\tilde{D}\in[D,1]} \Lambda(\tilde{D}) \\
&= \max_{\tilde{D}\in[D,1]} 2\mathcal{H}(\tilde{D}) \qquad (4.41) \\
&= 2H_b(\kappa) \qquad (4.42)
\end{aligned}$$

where (4.41) follows from the fact that $P(D; p_X)$ is strictly positive for any $D > 0$.

Alternatively, with a bit of work it can be shown that the low SNR behavior is given by

$$\begin{aligned}
&\lim_{\mathsf{snr}\to 0} \mathsf{snr}\left[\rho^{(\text{ML-UB})} - \kappa\right] \\
&= \lim_{\mathsf{snr}\to 0} \mathsf{snr} \max_{\tilde{D}\in[D,1]} \Lambda(\tilde{D}) \\
&= \max_{\tilde{D}\in[D,1]} \frac{\mathcal{H}(\tilde{D})}{P(\tilde{D})} 2\lambda(\tilde{D}),
\end{aligned} \tag{4.43}$$

where $\lambda(D)$ is given by (4.39). Note that this limit is strictly positive for any $D > 0$.

Combining (4.42) and (4.43) shows that there exists, for each fixed pair (D, p_X), a constant C such that

$$\rho^{(\text{ML-UB})} \leq \kappa + \frac{C}{\log(1+\mathsf{snr})} \tag{4.44}$$

for all snr.

Lastly, we consider the tradeoff between the distortion D and the SNR. For a given tuple $(\rho, \mathsf{snr}, p_X)$, let $D^{(\text{ML-UB})}$ denote the infimum over all distortions $D \geq 0$ such that $\rho^{(\text{ML-UB})} \leq \rho$. If $\rho > \kappa$, then the analysis given above shows that $D^{(\text{ML-UB})} \to 0$ as $\mathsf{snr} \to \infty$. Since $\Lambda(D)$ is finite for all $D > 0$ but grows without bound as $D \to 0$, this means that the following limit must be satisfied:

$$\lim_{\mathsf{snr}\to\infty} \Lambda(D^{(\text{ML-UB})}) = \rho - \kappa. \tag{4.45}$$

Starting with the definition of $\Lambda(D)$ given in (2.5), it is straightforward to show that (4.45) is satisfied if and only if

$$\lim_{\mathsf{snr}\to\infty} \mathsf{snr}\, \frac{P(D^{(\text{ML-UB})})}{\mathcal{H}(D^{(\text{ML-UB})})} = \left(\frac{2}{3-\sqrt{8}}\right) \frac{1}{\rho - \kappa}. \tag{4.46}$$

4.8.2 Behavior of the MMSE Noise Power

This section studies the behavior of the effective noise power τ^* defined in Theorem 2.7. Since there is a one-to-one correspondence between τ^* and the resulting distortion D, the results in this section immediately extend to the behavior of the distortion.

Starting with the definition in (2.37), this noise power can be expressed as

$$\tau^* = \arg\min_{\tau>0} \Gamma(\tau)$$

where

$$\Gamma(\tau) = \rho \log(\tau) + \frac{1}{\tau\, \mathsf{snr}} + 2I(X; X + \sqrt{\tau}W).$$

For any fixed tuple $(\rho, \mathsf{snr}, p_X)$, the function $\Gamma(\tau)$ grows without bound as either $\tau \to 0$ or $\tau \to \infty$. Therefore, the minimizer τ^* must be a solution to $\Gamma'(\tau^*, \mathsf{snr}) = 0$ where $\Gamma'(\cdot, \cdot)$ denotes the derivative of $\Gamma(\cdot, \cdot)$ with respect to the first argument. Using the following result of Guo et al. [37]:

$$\frac{d}{d\gamma} 2I(X; X + \sqrt{1/\gamma} W) = \mathsf{mmse}(1/\gamma; p_X), \tag{4.47}$$

it is straightforward to show that the condition $\Gamma'(\tau^*, \mathsf{snr}) = 0$ is equivalent to

$$\rho\, \tau^* = \frac{1}{\mathsf{snr}} + \mathsf{mmse}(\tau^*; p_X). \tag{4.48}$$

Note that (4.48) may have additional fixed point solutions (other than τ^*) corresponding to local minima or maxima of the function $\Gamma(\tau)$.

We first consider the behavior as $\rho \to \infty$. By the optimality of the MMSE estimate (with respect to mean squared error) the noise power τ^* is a non-increasing function ρ, and thus $\mathsf{mmse}(\tau^*, p_X)$ is a non-increasing function of ρ. Combining this fact with (4.48) shows that $\tau^* \to 0$ as $\rho \to \infty$. Since $\mathsf{mmse}(\tau, p_X) \to 0$ as $\tau \to 0$, we obtain the limit

$$\lim_{\rho \to \infty} \rho\, \tau^* = \frac{1}{\mathsf{snr}}. \tag{4.49}$$

We next consider the behavior as $\mathsf{snr} \to \infty$. If τ is a fixed constant, independent of snr, then $\Gamma(\tau)$ converges to a finite constant. However, if $\tau = \tau(\mathsf{snr})$ scales with snr in such a way that $\tau(\mathsf{snr}) \to 0$ then

$$\begin{aligned}
&\lim_{\mathsf{snr} \to \infty} \frac{1}{\log \tau(\mathsf{snr})} \left[\Gamma(\tau(\mathsf{snr})) - \frac{1}{\tau(\mathsf{snr})\, \mathsf{snr}} \right] \\
&= \lim_{\mathsf{snr} \to \infty} \frac{1}{\log \tau(\mathsf{snr})} \left[\rho \log \tau(\mathsf{snr}) + I\left(X; X + \sqrt{\tfrac{1}{\mathsf{snr}}} W \right) \right] \\
&= \rho - \lim_{\epsilon \to 0} \frac{2I(X; X + \sqrt{\epsilon}\, W)}{\log(1/\epsilon)} \\
&= \rho - \lim_{\epsilon \to 0} \frac{\mathsf{mmse}(\epsilon; p_X)}{\epsilon}
\end{aligned} \tag{4.50}$$

where (4.50) follows from L'Hopital's rule and (4.47).

We consider two cases. If the right hand side of (4.50) is strictly positive, then there exists a scaling $\tau(\mathsf{snr})$ such that $\Gamma(\tau(\mathsf{snr}))$ decreases without bound. Since $\Gamma(\tau)$ is finite for fixed τ and grows without bound as $\tau \to \infty$, this means that $\tau^* \to 0$ as $\mathsf{snr} \to \infty$. Conversely, if the right hand side of (4.50) is strictly negative, then $\Gamma(\tau(\mathsf{snr}))$ increases without bound for any scaling where $\tau(\mathsf{snr}) \to 0$. This means that τ^* is bounded away from zero for all snr.

Combining these cases, we can conclude that the stability threshold $\varrho^{(\mathrm{MMSE})}$ of the MMSE estimator is given by

$$\varrho^{(\mathrm{MMSE})} = \lim_{\epsilon \to 0} \frac{\mathsf{mmse}(\epsilon; p_X)}{\epsilon}. \tag{4.51}$$

To characterize the rate at which τ^* decreases as $\mathsf{snr} \to \infty$, we rearrange (4.48) to obtain

$$\mathsf{snr}\,\tau^* = \left[\rho - \frac{\mathsf{mmse}(\tau^*; p_X)}{\tau^*}\right]^{-1}. \tag{4.52}$$

If $\rho > \varrho^{(\mathrm{MMSE})}$, then $\tau^* \to 0$ as $\mathsf{snr} \to \infty$. Hence, by (4.52) and the definition of $\varrho^{(\mathrm{MMSE})}$, we obtain the limit

$$\lim_{\mathsf{snr} \to \infty} \mathsf{snr}\,\tau^* = \frac{1}{\rho - \varrho^{(\mathrm{MMSE})}}. \tag{4.53}$$

4.8.3 Proof of Proposition 4.5

Using the bound $H_b(p) \le p\log(1/p) + p$ we obtain

$$\mathcal{H}(D;\kappa) \le 2\kappa D[\log(1/D) + 1 + \log(\tfrac{1-\kappa}{\kappa})]. \tag{4.54}$$

Using the definition of $P(D; p_X)$ and the fact that X is lower bounded, we obtain

$$\begin{aligned} P(D;p_X) &= \int_0^\infty \Big(\Pr[X^2 \ge u] - (1-D)\kappa)\Big)_+ du \\ &\ge \int_0^\infty \Big(\kappa \mathbf{1}(u < B^2) - (1-D)\kappa)\Big)_+ du \\ &= \kappa D B^2. \end{aligned} \tag{4.55}$$

Combining (4.54) and (4.55) completes the proof of (4.23).

The bound (4.24) follows immediately from the upper bound

$$\begin{aligned} D_{\mathrm{awgn}}(\sigma^2; p_X) &= \min_{t \ge 0} \max\Big(\Pr[|X + \sigma W| \le t], \tfrac{1-\kappa}{\kappa}\Pr[|\sigma W| > t]\Big) \\ &\le \min_{t \ge 0} \max\Big(\Pr[|B + \sigma W| \le t], \tfrac{1-\kappa}{\kappa}\Pr[|\sigma W| > t]\Big) \\ &\le \max\Big(\Pr[|B + \sigma W| \le \tfrac{B}{2}], \tfrac{1-\kappa}{\kappa}\Pr[|\sigma W| > \tfrac{B}{2}]\Big) \\ &\le \Big(\tfrac{1-\kappa}{\kappa}\Big)\Pr[|\sigma W| > \tfrac{B}{2}] \qquad (4.56) \\ &\le \Big(\tfrac{1-\kappa}{\kappa}\Big)\exp\Big(-\tfrac{B^2}{8\sigma^2}\Big) \qquad (4.57) \end{aligned}$$

where (4.56) follows from the triangle inequality and (4.57) follows from the well known upper bound (see e.g. [71]) $\Pr[|W| > t] \le \exp(-t^2/2)$.

4.8.4 Proof of Proposition 4.6

For this proof, it is convenient to define the quantile function

$$\xi(D) = \inf\{t \geq 0 : \Pr[|X|^2 \leq t | X \neq 0] \geq D\},$$

and note that

$$\lim_{D\to 0} \frac{\xi(D)}{D^{2/L}} = \tau^{-2/L}. \tag{4.58}$$

We first consider (4.25). Using the bounds $p\log(1/p) \leq H_b(p) \leq p\log(1/p)+p$, we obtain

$$\lim_{D\to 0} \frac{\mathcal{H}(D;\kappa)}{D\log(1/D)} = 2\kappa. \tag{4.59}$$

Next, starting from the definition of $P(D;p_X)$ and using a change of variables leads to the expression

$$P(D;p_X) = \kappa D \int_0^1 \xi(\beta D) d\beta.$$

Thus, we can write

$$\begin{aligned}\lim_{D\to 0} \frac{P(D;p_X)}{D^{1+2/L}} &= \kappa \int_0^1 \lim_{D\to 0} \frac{\xi(\beta D)}{D^{2/L}} d\beta \\ &= \kappa\,\tau^{-2/L} \int_0^1 \beta^{2/L} d\beta \\ &= \frac{\kappa\,\tau^{-2/L}}{1+2/L}\end{aligned} \tag{4.60}$$

where swapping the limit and the integral is justified by the fact that $\xi(D)$ is continuous and monotonically increasing. Combining (4.59) and (4.60) completes the proof of (4.25).

We next consider (4.26). Let w_D be the unique solution to $\Pr[|W| > w_D] = \kappa D/(1-\kappa)$. Using standard bounds on the cumulative distribution function of the Gaussian distribution (see e.g. [71]) it can be verified that

$$\lim_{D\to 0} \frac{w_D^2}{\log(1/D)} = 2. \tag{4.61}$$

Therefore, by (4.58) and (4.61), the limit (4.26) follows immediately if we can show that

$$\lim_{D\to 0} \left(\frac{\xi(D)}{w_D^2}\right)^{-1} \sigma^2_{\text{awgn}}(D;p_X) = 1. \tag{4.62}$$

To proceed, define the probabilities

$$p_1(\theta) = \Pr\left[\left|\frac{X}{\sqrt{\xi(D)}} + \frac{W}{w_D}\right| \le \theta | X \neq 0\right]$$
$$p_2(\theta) = \left(\tfrac{1-\kappa}{\kappa}\right) \Pr[|\tfrac{W}{w_D}| > \theta],$$

and note that

$$D_{\text{awgn}}\left(\tfrac{\xi(D)}{w_D^2}; p_X\right) = \inf_{\theta \in \mathbb{R}} \max\left(p_1(\theta), p_2(\theta)\right). \tag{4.63}$$

By a change of variables, we can write

$$p_1(1) = D \int_0^\infty \mathbf{1}(\beta \le \tfrac{1}{D}) \Pr\left[\left|\sqrt{\tfrac{\xi(\beta D)}{\xi(D)}} + \tfrac{W}{w_D}\right| \le 1\right] d\beta.$$

Using (4.58) and the fact that $\xi(D)$ is a strictly decreasing function when D is small, it can be shown that the integrand of the above expression converges pointwise to $\mathbf{1}(\beta \le 1)$ and hence

$$\lim_{D \to 0} D^{-1} p_1(1) = 1. \tag{4.64}$$

Since $p_1(\theta)$ is a strictly increasing function of θ and $p_2(\theta)$ is a strictly decreasing function of θ with $p_2(1) = D$, it thus follows that

$$\lim_{D \to 0} D^{-1} D_{\text{awgn}}\left(\tfrac{\xi(D)}{w_D^2}; p_X\right) = 1.$$

Since $D_{\text{awgn}}(\sigma^2; p_X)$ is a strictly increasing function of σ^2, this proves the limit (4.62), and thus completes the proof of (4.26).

4.9 Properties of Soft Thresholding

This Section reviews several useful properties of the soft-thresholding noise sensitivity $\mathcal{M}(\sigma^2, \alpha, p_X)$ introduced in Section 2.3.

To begin, observe that the noise sensitivity defined in (2.33) can be expressed as

$$\mathcal{M}(\sigma^2, \alpha, p_X) = \frac{\mathbb{E}\left[|\eta^{(\text{ST})}(X + \sigma W, \sigma^2; \alpha) - X|^2\right]}{\sigma^2} \tag{4.65}$$
$$= \mathbb{E}[\mu(X/\sigma, \alpha)] \tag{4.66}$$

where $\mu(z, \alpha)$ is given by

$$\mu(z, \alpha) = \mathbb{E}\left[|\eta^{(\text{ST})}(z + W, 1; \alpha) - z|^2\right]. \tag{4.67}$$

With a bit of calculus, it can then be verified that

$$\begin{aligned}\mu(z,\alpha) = z^2\Big[1 - \Phi(-\alpha+z) - \Phi(-\alpha-z)\Big] \\ + (1+\alpha^2)\Big[\Phi(-\alpha+z) + \Phi(-\alpha-z)\Big] \\ - (\alpha+z)\phi(\alpha-z) - (\alpha-z)\phi(\alpha+z),\end{aligned} \tag{4.68}$$

where $\phi(x) = (2\pi)^{-1/2}e^{-x^2/2}$ and $\Phi(x) = \int_{-\infty}^{x}\phi(t)dt$.

If we let $\tilde{X}$ be distributed according to the nonzero part of p_X, then we obtain the general expression

$$\mathcal{M}(\sigma^2,\alpha,p_X) = (1-\kappa)\mu(0,\alpha) + \kappa\,\mathbb{E}[\mu(\tfrac{1}{\sigma}\tilde{X},\alpha)]. \tag{4.69}$$

4.9.1 Infinite SNR Limit

The infinite SNR limit of the AMP-ST bound corresponds to the limit of $\mathcal{M}(\sigma^2,\alpha,p_X)$ as the noise power σ^2 tends to zero. A simple exercise shows that

$$\lim_{\sigma^2\to 0}\mathbb{E}[\mu(\tfrac{1}{\sigma}\tilde{X},\alpha)] = (1+\alpha^2) \tag{4.70}$$

for any random variable $\tilde{X}$ with $\Pr[\tilde{X}=0]=0$. Therefore, for any distribution $p_X \in \mathcal{P}(\kappa)$, we obtain the general limit

$$\lim_{\sigma^2\to 0}\mathcal{M}(\sigma^2,\alpha,p_X) = \mathcal{M}_0(\alpha,\kappa) \tag{4.71}$$

where

$$\mathcal{M}_0(\alpha,\kappa) = \kappa(1+\alpha^2) + (1-\kappa)2\Big[(1+\alpha^2)\Phi(-\alpha) - \alpha\phi(\alpha)\Big]. \tag{4.72}$$

Minimizing $\mathcal{M}_0(\alpha,\kappa)$ as a function of α recovers the ℓ_1/ℓ_0 equivalence threshold of Donoho and Tanner [19].

4.9.2 Universal Bounds

In [18], it is shown that, over the class of distributions $\mathcal{P}(\kappa)$, the noise sensitivity is maximized at a "three-point" distribution that places all of its nonzero mass at $\pm 1/\sqrt{\kappa}$. Combining this result with (4.69) leads to the uniform upper bound

$$\sup_{p_X\in\mathcal{P}(\kappa)}\mathcal{M}(\sigma^2,\alpha,p_X) = \mathcal{M}^*(\sigma^2,\alpha,\kappa) \tag{4.73}$$

where

$$\mathcal{M}^*(\sigma^2,\alpha,\kappa) = (1-\kappa)\mu(0,\alpha) + \kappa\mu(\tfrac{1}{\sigma\sqrt{\kappa}},\alpha). \tag{4.74}$$

Using (4.73) it is now possible to extend the bound given in Theorem 2.6 to a given class of distributions $\mathcal{P}_X \subset \mathcal{P}(\kappa)$. Specifically, we can conclude that a distortion D is achievable for a tuple $(\rho, \mathcal{P}_X, \mathsf{snr})$ if

$$\rho > \frac{1}{\sigma^2 \, \mathsf{snr}} + \mathcal{M}^*(\sigma^2, \alpha, \kappa) \tag{4.75}$$

where

$$\sigma^2 = \min_{p_X \in \mathcal{P}_X} \sigma^2_{\text{awgn}}(D; p_X). \tag{4.76}$$

We note that the bounds (4.75) and (4.76) can be used to find a value of the soft-thresholding parameter α that works well uniformly over the class $\mathcal{P}_X$. However, since these universal bounds are not tight, we cannot conclude that the resulting value of α is minimax optimal.

Chapter 5

The Role of Diversity

In the previous chapters we have seen that a major challenge in sparsity pattern recovery is that small nonzero values are difficult to detect in the presence of noise. In this chapter, we show how this problem can be alleviated if one can observe samples from multiple realizations of the nonzero values for the same sparsity pattern.

5.1 Joint Sparsity Pattern Recovery

It well known that the presence of additional structure, beyond sparsity, can significantly alter the problem of sparsity pattern recovery. Various examples include distributed or model-based compressed sensing [3,4,73], estimation from multiple measurement vectors [13], simultaneous sparse approximation [70], model selection [83], union support recovery [52], multi-task learning [43], and estimation of block-sparse signals [26, 66].

In this chapter, we consider a joint sparsity pattern estimation framework motivated in part by the following engineering problem. Suppose that one wishes to estimate the sparsity patten of an unknown vector and is allowed to take either M noisy linear measurements of the vector itself, or spread the same number measurements amongst multiple vectors with same sparsity pattern as the original vector, but different nonzero values. This type of problem arises, for example, in magnetic resonance imaging where the vectors correspond to images of the same body part (common sparsity pattern) viewed with different contrasting agents (different nonzero values).

On one hand, splitting measurements across different vectors increases the number of unknown values, potentially making estimation more difficult. On the other hand, using all measurements on a single vector has the risk that nonzero values with small magnitudes will not be detected. To understand this tradeoff, this chapter bounds the accuracy of various estimators for the estimation problem illustrated in Figure 5.1. We refer to the number of vectors J as the "diversity".

The results in this chapter show that the right amount of diversity is beneficial, but too

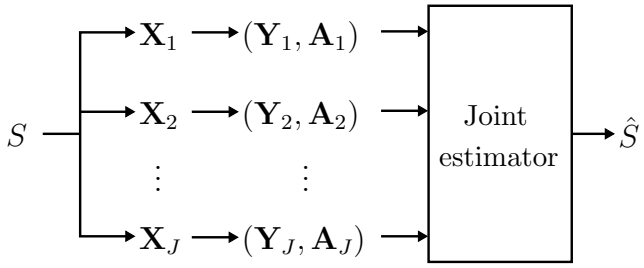

Figure 5.1: Illustration of joint sparsity pattern estimation. The vectors $\mathbf{X}_j$ share a common sparsity pattern S but have independent nonzero values. The sparsity pattern S is estimated jointly using measurements vectors $\mathbf{Y}_i$ corresponding to different measurement matrices $\mathbf{A}_j$.

much or too little can be detrimental (when the total number of measurements is fixed). Moreover, we show that diversity can significantly reduce the gap in performance between computationally efficient estimators, such the matched filter or LASSO, and estimators without any computational constraints.

5.1.1 Problem Formulation

Let $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_J \in \mathbb{R}^n$ be a set of jointly random sparse vectors whose nonzero values are indexed by a common sparsity pattern S

$$S = \{i : X_j(i) \neq 0\}, \quad \text{for } j = 1, 2, \cdots, J. \tag{5.1}$$

We assume that S is distributed uniformly over all subsets of $\{1, 2, \cdots, n\}$ of size k where k is known. For simplicity, we focus exclusively on the setting where the nonzero entries are i.i.d. Gaussian with zero mean.

We consider estimation of S from measurement vectors $\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_J \in \mathbb{R}^m$ of the form

$$\mathbf{Y}_j = \mathbf{A}_j \mathbf{X}_j + \frac{1}{\sqrt{\mathsf{snr}}} \mathbf{W}_j \quad \text{for } j = 1, 2, \cdots, J \tag{5.2}$$

where each $\mathbf{A}_j \in \mathbb{R}^{m \times n}$ is a known matrix whose elements are i.i.d. $\mathcal{N}(0, 1)$ and $\mathbf{W}_j \sim \mathcal{N}(0, I_{m \times m})$ is unknown noise. The estimation problem is depicted in Figure 5.1. The accuracy of an estimate $\hat{S}$ is assessed using the distortion metric $d(S^*, \hat{S})$ given in (1.6).

Our analysis considers the high dimensional setting where the diversity J is fixed but the vector length n, sparsity k, and number of measurements per vector m tend to infinity. We focus exclusively on the setting of linear sparsity where $k/n \to \kappa$ for some fixed *sparsity rate* $\kappa \in (0, 1/2)$ and $m/n \to r$ for some fixed per-vector *sampling rate* $r > 0$. The total number of measurements is given by $M = mJ$, and we use $\rho = Jr$ to denote the total sampling rate. We say that a distortion D is *achievable* for an estimator $\hat{S}$ if $\Pr[d(S, \hat{S}) > D] \to 0$ as $n \to \infty$. The case $D = 0$ corresponds to exact recovery and the case $D > 0$ corresponds to a constant fraction of errors.

We remark that we note that the joint estimation problem in this chapter is closely related to the multiple measurement vector problem [13], except that each vector is measured using a different matrix. Alternatively, our problem is a special case of block-sparsity [26,66] with a block-sparse measurement matrix. Versions of our bounds for block-sparsity with dense measurement matrices can also be derived.

5.1.2 Notations

For a matrix A and set of integers S we use $A(S)$ to denote the matrix formed by concatenating the columns of A indexed by S. We use $H_b(p) = -p\log p - (1-p)\log(1-p)$ to denote binary entropy and all logarithms are natural.

5.2 Recovery Bounds

This section gives necessary and sufficient conditions for the joint sparsity pattern estimation problem depicted in Figure 5.1.

One important property of the estimation problem is the relative size of the smallest nonzero values, averaged across realizations. For a given fraction $\beta \in [0,1]$, we define random variable

$$P_J^{(n)}(D) = \min_{\Delta \subset S\,:\,|\Delta|=Dk} \frac{1}{J}\sum_{j=1}^{J} \frac{k}{n}\|\mathbf{X}_j(\Delta)\|^2. \tag{5.3}$$

By the Glivenko-Cantelli theorem, $P_J^{(n)}(D)$ converges almost surely to a nonrandom limit $P_J(D)$. We will refer to this limit as the *diversity power*. If the nonzero values are Gaussian, as is assumed in this chapter, it can be shown that

$$P_J(D) = \textstyle\int_0^\alpha \xi_J(p)dp \tag{5.4}$$

where

$$\xi_J(p) = \left\{t : \Pr[\tfrac{1}{J}\chi_J^2 \le t] = p\right\} \tag{5.5}$$

denotes the quantile function of a normalized chi-square random variable with J degrees of freedom.

Another important property is the metric entropy rate (in nats per vector length) of S with respect to our distortion function $d(S,\hat{S})$. In Chapter 3, it is shown that this rate is given by

$$R(D;\kappa) = H(\kappa) - \kappa H_b(D) - (1-\kappa)H_b(\tfrac{\kappa D}{1-\kappa}) \tag{5.6}$$

for all $D < 1-\kappa$ and is equal to zero otherwise.

5.2.1 Joint ML Upper Bound

We first consider the ML recovery algorithm which is given by

$$\hat{S}^{\mathrm{NS}} = \arg \min_{S \,:\, |S|=k} \sum_{j=1}^{J} \mathrm{dist}(\mathbf{Y}_j, \mathbf{A}_j(S))^2 \tag{5.7}$$

where $\mathrm{dist}(\mathbf{Y}_j, \mathbf{A}_j(S))$ denotes the euclidean distance between $\mathbf{Y}_j$ and the linear subspace spanned by the columns of $\mathbf{A}_j(S)$. (For the case $J = 1$, this estimator corresponds to the ML estimator studied in Chapter 2.)

Theorem 5.1. *A distortion D is achievable for the tuple $(\kappa, \mathsf{snr}, \rho, J)$ using the ML recovery algorithm if*

$$\rho > \kappa J + \max_{\tilde{D} \in [D,1]} \min \Big(E_1(\tilde{D}), E_2(\tilde{D}) \Big) \tag{5.8}$$

where

$$E_1(D) = \frac{2H_b(\kappa) - 2R(D;\kappa) + 2D\kappa J \log(5/3)}{\frac{1}{J}\log\Big(1 + \frac{4}{25} J P_J(D)\,\mathsf{snr}\Big)} \tag{5.9}$$

$$E_2(D) = \frac{2H_b(\kappa) - 2R(D;\kappa)}{\log\Big(1 + P_1(D)\,\mathsf{snr}\Big) + 1/\Big(P_1(D)\,\mathsf{snr}\Big) - 1}. \tag{5.10}$$

For the case $J = 1$, the functions $E_1(D)$ and $E_2(D)$ correspond to the functions $\Lambda_1(D)$ and $\Lambda_2(D)$ given in Theorem 2.1. To extend these bounds to the setting $J > 1$ requires a large deviations bound on random variable $P_J^{(n)}(D)$. The full proof of this Theorem 5.1 is given in [59].

Theorem 5.1 is a combination of two bounds. The part due to $E_1(D)$ determines the scaling behavior at low distortions and low SNR and the part due to $E_2(D)$ determines the scaling behavior at high SNR.

5.2.2 Information-Theoretic Lower Bound

We next consider an information-theoretic lower bound on the distortion for any estimator. This bound depends on the entropy of the smallest nonzero values. For a given fraction $D \in [0, 1]$, we define the *conditional entropy power*

$$\mathcal{N}(D) = \frac{1}{2\pi e} \exp\Big\{ -2h\Big(U | U^2 \le \xi_1(D)\Big)\Big\} \tag{5.11}$$

where $h(\cdot)$ is differential entropy and $U \sim \mathcal{N}(0, 1)$.

The following result gives a necessary condition for any possible recovery algorithm. The proof is outlined in Section 5.4.1.

Theorem 5.2. *A distortion D is not achievable for the tuple $(\kappa, \mathsf{snr}, \rho, J)$ if there exists a distortion $\tilde{D} \in [D, 1]$ for which that at least one of the following inequalities is satisfied:*

$$2R\Big(\frac{D}{\tilde{D}}; \frac{\tilde{D}\kappa}{1-\kappa+\tilde{D}\kappa}\Big) > J\,\mathcal{V}_{UB}\Big(\frac{\rho}{1-\kappa+\tilde{D}\kappa}, P_J^2(D)\mathsf{snr}\Big) \tag{5.12}$$

$$\begin{aligned} 2R\Big(\frac{D}{\tilde{D}}; \frac{\tilde{D}\kappa}{1-\kappa+\tilde{D}\kappa}\Big) &> \mathcal{V}_{UB}\Big(\frac{\rho}{1-\kappa+\tilde{D}\kappa}, \tilde{D}^{1-1/J} P_1(\tilde{D}^{1/J})\,\mathsf{snr}\Big) \\ &\quad - \Big(\frac{\tilde{D}\kappa}{1-\kappa+\tilde{D}\kappa}\Big)\mathcal{V}_{LB}\Big(\frac{\rho}{\tilde{D}\kappa}, \tilde{D}\,\mathcal{N}(\tilde{D}^{1/J})\,\mathsf{snr}\Big) \end{aligned} \tag{5.13}$$

where $\mathcal{V}_{LB}(r,\gamma)$ is given by (3.20) *and*

$$\mathcal{V}_{UB}(r,\gamma) = \begin{cases} \dfrac{r}{2}\log(1+\gamma), & \text{if } r \le 1 \\ \dfrac{1}{2}\log(1+r\,\gamma), & \text{if } r > 1 \end{cases}. \tag{5.14}$$

As was the case for the ML upper bound, the bound in Theorem 5.2 is inversely proportional to the effective power $P_J(\beta)\,\mathsf{snr}$ when the effective power is small.

5.2.3 Two-Stage Recovery Bounds

This section gives bounds for the two-stage estimation architecture depicted in Figure 5.2. In the first stage, each vector $\mathbf{X}_j$ is estimated from its measurements $\mathbf{Y}_j$. In the second stage, the sparsity pattern S is estimated by jointly thresholding estimates $\hat{\mathbf{X}}_1, \hat{\mathbf{X}}_2, \cdots, \hat{\mathbf{X}}_J$. One advantage of this architecture is that the estimation in the first stage can be done in parallel. We will see that this architecture can be near optimal in some settings but is highly suboptimal in others.

Single-Vector Estimation

For the first stage of estimation, we may use the results derived in Chapter 2 for the MF, LMMSE, AMP, and MMSE estimators. Throughout this chapter, we we use the fact that the AMP-ST is equivalent to LASSO under a proper calibration of the regularization parameters.

Thresholding

For the second stage of recovery we consider the *joint thresholding* of the form

$$\hat{S}^{\text{TH}} = \Big\{i : \textstyle\sum_{j=1}^{J} \hat{X}_j^2(i) \ge t\Big\} \tag{5.15}$$

where the threshold $t \ge 0$ is chosen to minimize the expected distortion. Since each index $i \in \{1, 2, \cdots, n\}$ is evaluated independently, and since the estimated vectors $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_j$

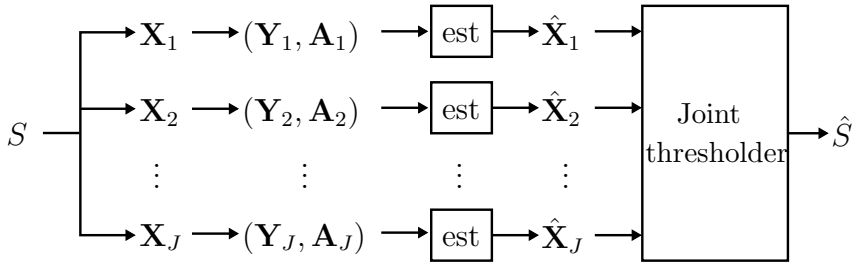

Figure 5.2: Illustration of single-vector estimation followed by joint thresholding.

are conditionally independent given the sparsity pattern S, the distribution on the distortion $d(S, \hat{S}^{\mathrm{TH}})$ can be characterized by joint distribution on $(X_1(1), \hat{X}_1(1))$.

The following result describes the relationship between the distortion α, the diversity J, and the effective noise power σ^2. The proof is given in Section 5.4.2.

Theorem 5.3. *Suppose that for $j = 1, 2, \cdots, J$, the empirical joint distributions on the elements of $(\mathbf{X}_j, \hat{\mathbf{X}}_j)$ converge weakly to the distribution on the pair (X, Z) in the scalar equivalent model given in Definition 2.1 with noise power σ^2. Then, a distortion D is a achievable if $\sigma^2 < \sigma_J^2(D)$ and not achievable if $\sigma^2 > \sigma_J^2(D)$ where*

$$\sigma_J^2(D) = \frac{\xi_J(D)}{\xi_J(1 - \frac{D\kappa}{1-\kappa}) - \xi_J(D)} \tag{5.16}$$

with $\xi_J(D)$ given by (5.5).

Theorem 5.3 shows that the relationship between D and J is encapsulated by the term $\sigma_J^2(D)$. With a bit of work it can be shown that the numerator and denominator in (5.16) scale like $D^{-1}P_J(D)$ and $D^{-1}R(D;\kappa)$ respectively when D is small. Thus, plugging $\sigma_J^2(D)$ into the equivalent noise expression of the matched filter given in (2.118) shows that bounds attained using Theorem 5.3 have similar low distortion behavior to the bounds in Section 5.2.

One advantageous property of Theorem 5.3 is that the bounds are exact. As a consequence, these bounds are sometimes lower than the upper bound in Theorem 5.1, which is loose in general. One shortcoming however, is that the two-stage architecture does not take full advantage of the joint structure during the first stage of estimation. As a consequence, the performance of these estimators can be highly suboptimal, especially at high SNR.

5.3 Sampling Rate-Diversity Tradeoff

In this section, we analyze various behaviors of the bounds in Theorems 5.1, 5.2, and 5.3, with an emphasis on the tradeoff provided by the diversity J. The following results characterize the high SNR and low distortion behavior of optimal estimation. Their proofs are given in [59].

Proposition 5.1 (High SNR)**.** *Let (κ, J, D), be fixed and let $\rho(\mathsf{snr})$ denote the infimum over sampling rates ρ such that α is achievable for the optimal estimator. Fix any $\epsilon > 0$.*

(a) If $D > 0$, then

$$\rho(\mathsf{snr}) \leq J\kappa + \frac{2H_b(\kappa)(1+\epsilon)}{\log(\mathsf{snr})} \tag{5.17}$$

for all snr large enough.

(b) If $2R(D;\kappa) > J\kappa$, then

$$\rho(\mathsf{snr}) \geq J\kappa + \frac{2R(\kappa,\alpha)(1-\epsilon)}{\log(\mathsf{snr})} \tag{5.18}$$

for all snr large enough.

Proposition 5.2 (Low Distortion)**.** *Let $(\kappa, J, \mathsf{snr})$ be fixed and let $\rho(D)$ denote the infimum over sampling rates ρ such that D is achievable for the optimal estimator. There exist constants $0 < C^- \leq C^+ < \infty$ such that*

$$C^-\left(\tfrac{1}{D}\right)^{2/J}\log(\tfrac{1}{D}) \leq \rho(D) \leq C^+\left(\tfrac{1}{D}\right)^{2/J}\log(\tfrac{1}{D}) \tag{5.19}$$

for all D small enough.

Propositions 5.1 and 5.2 illustrate a tradeoff. At high SNR, the difficulty of estimation is dominated by the uncertainty about the nonzero values. Accordingly, the number of measurements is minimized by letting $J = 1$. As the desired distortion becomes small however, the opposite behavior occurs. Since estimation is limited by the size of the smallest nonzero values, it is optimal to choose J large to increase the diversity power. This behavior can be seen, for example, in Figures 5.3-5.6.

A natural question then, is how does one best choose the diversity J? The following result shows that the right amount of diversity can significantly improve performance. The proof is given in [59].

Proposition 5.3. *Let (κ, snr) be fixed and let $\rho(D, J)$ denote the infimum over sampling rates ρ such that D is achievable with diversity J. Then,*

$$\rho(D, J) \leq \kappa J + O\left(\tfrac{D}{P(D,J)}\right). \tag{5.20}$$

Moreover, if $J = J^*(D) = \Theta(\log(1/D)$ then

$$\rho(D, J^*(D)) = \Theta(\log(1/D)). \tag{5.21}$$

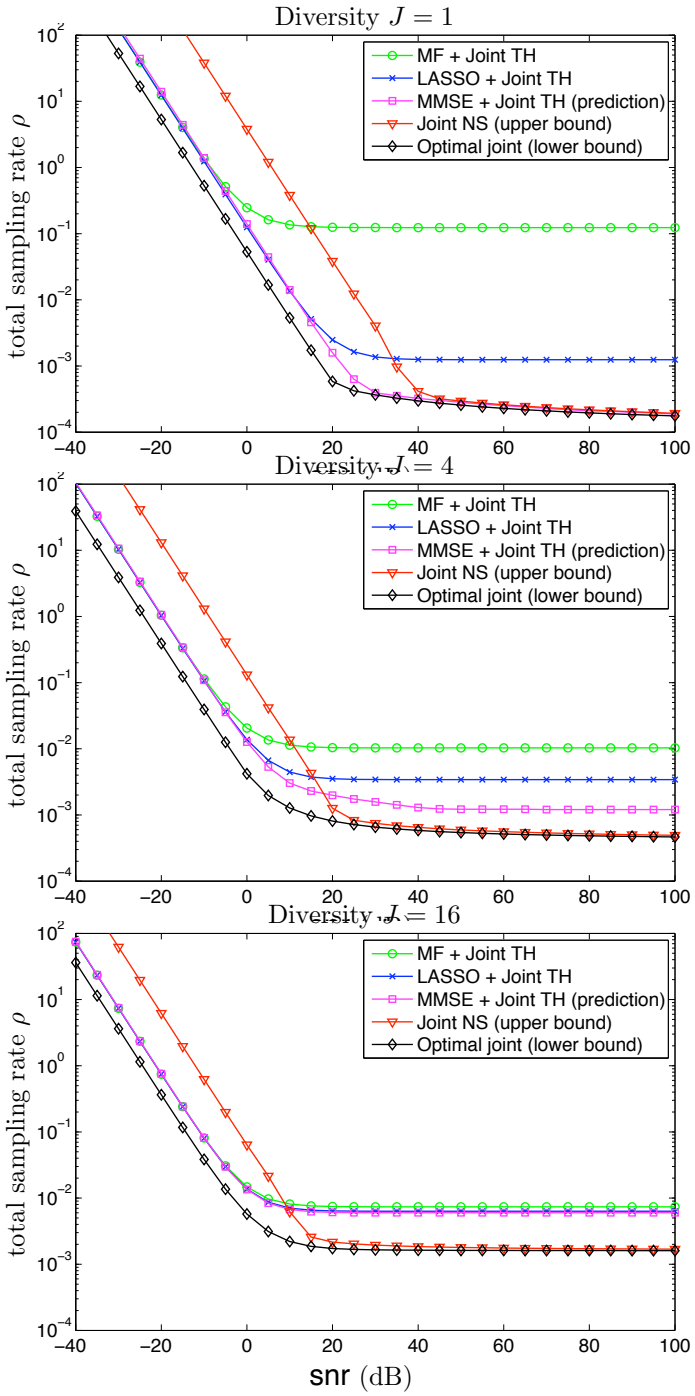

Figure 5.3: Bounds on the total sampling rate $\rho = Jr$ as a function of snr for various J when $D = 0.1$ and $\kappa = 10^{-4}$.

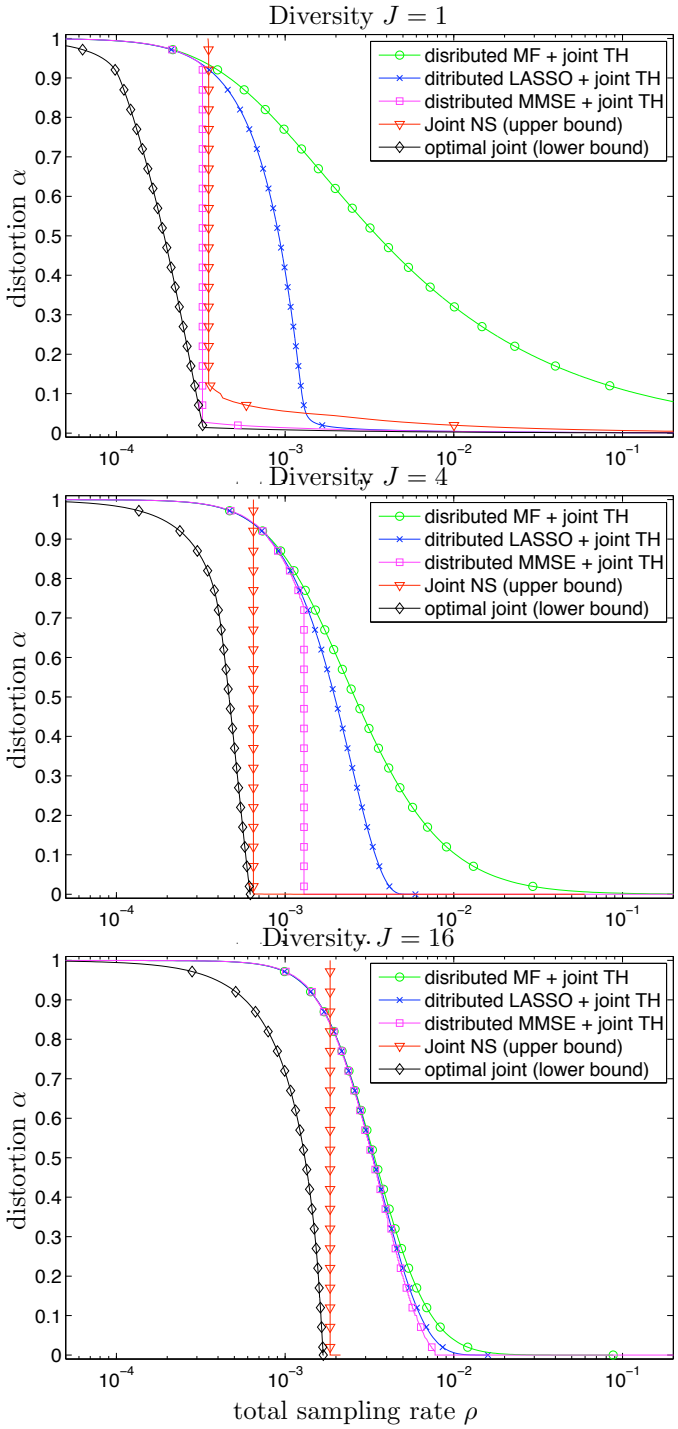

Figure 5.4: Bounds on the distortion D as a function of the total sampling rate $\rho = Jr$ for various J when $\mathsf{snr} = 40$ dB and $\kappa = 10^{-4}$.

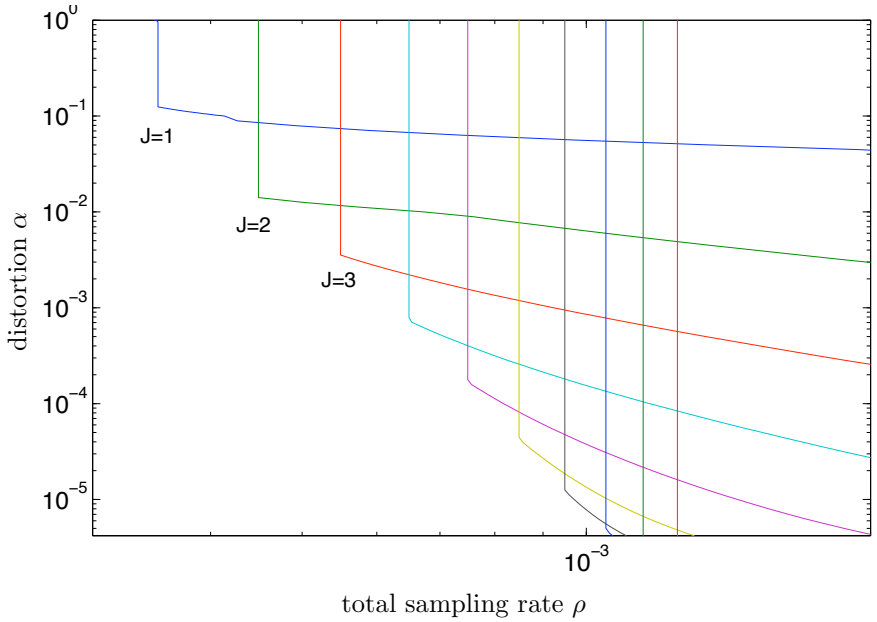

Figure 5.5: The upper bound (Theorem 5.1) on the total sampling rate $\rho = Jr$ of the nearest subspace estimator as a function of the distortion α for various J when snr $= 40$ dB and $\kappa = 10^{-4}$.

An important implication of Proposition 5.3 is that the optimal choice of J allows the distortion to decay *exponentially* rapidly with the sampling rate ρ. Note that the rate of decay is only polynomial if J is fixed. Interestingly, it can also be shown that the same exponential boost can be obtained using non-optimal estimators, albeit with smaller constants in the exponent.

The effect of the diversity J is illustrated in Fig. 5.5 for the nearest subspace estimator and in Fig. 5.6 for Lasso + thresholding. In both cases, the bounds show the same qualitative behavior–each value of the diversity J traces out a different curve in the sampling rate distortion region. It is important to note however, that due to the sub-optimality of the two stage architecture and the LASSO estimator, these similar behaviors occur only at different SNRs and with an order of magnitude difference in the sampling rate.

5.4 Proofs

5.4.1 Proof Outline of Theorem 5.2

The section outlines the proof of Theorem 5.2; the full proof is given in [59].

The proof of Theorem 5.2 uses many of the ideas developed in Chapter 3 for the single-vector setting. To apply these bounds, we cast the multi-vector problem as a version of the single-vector problem where the vector length is Jn and the number of nonzero elements is

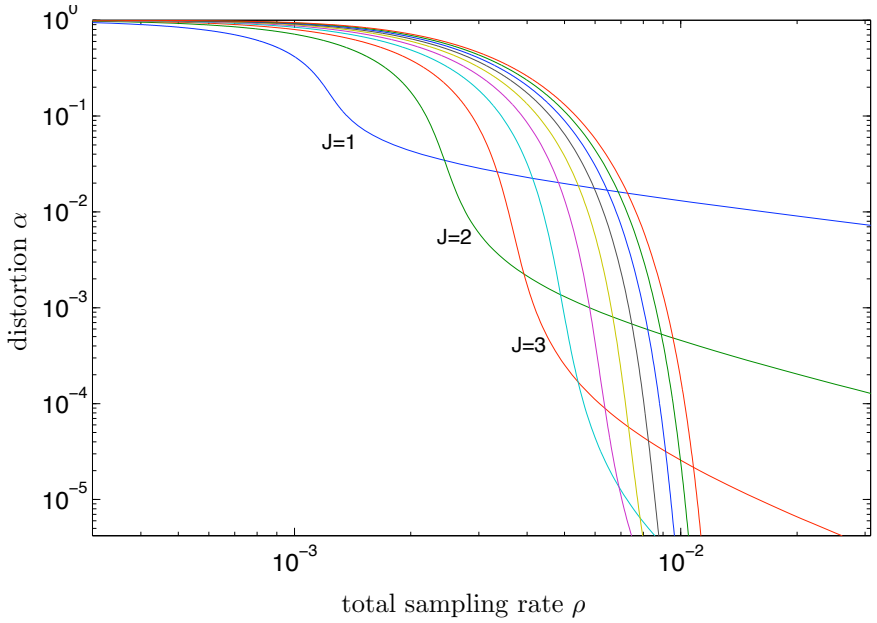

Figure 5.6: The upper bound (Theorem 5.3) on the total sampling rate $\rho = Jr$ of LASSO + Joint Thresholding as a function of the distortion α for various J when $\mathsf{snr} = 30$ dB and $\kappa = 10^{-4}$.

Jk:

$$\begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_J \end{bmatrix} = \begin{bmatrix} \mathbf{A}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{A}_2 & \ddots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{A}_J \end{bmatrix} \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_J \end{bmatrix} + \begin{bmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \\ \vdots \\ \mathbf{W}_J \end{bmatrix}.$$

Taking the joint sparsity constraint into account shows that the metric entropy rate of S with respect to the new problem is given by $\frac{1}{J}R(\kappa, \alpha)$.

The remaining challenge in this proof is that the low distortion behavior relies on the concept of a genie, who provides the estimator indices and values of the largest nonzero elements. This genie trick, is used to isolate the effect of the smallest nonzero values. The new difficultly in the multi-vector setting is the following: if the genie chooses which indices i he reveals based on the average magnitude given by $\sum_{j=1}^{J} X_j^2(i)$, then the values $X_1(i), X_2(i), \cdots, X_J(i)$ conditioned on the genies decision are no longer independent. As a consequence, it is not possible to compute their joint entropy and the sum of their individual entropy's.

To resolve this issue, we develop two different bounds. For the first bound, the genie selects indices according the average magnitude and we ignore the entropy of the remaining (dependent) unknown values. This bound leads to the necessary condition (5.12). For the second bound, the genie select indices according to the largest magnitude, i.e. $\max_{1 \le j \le J} X^2(i)$.

This selection strategies preserves the conditional independence and leads to the necessary condition (5.13).

5.4.2 Proof of Theorem 5.3

For each index i, the random variables $\hat{X}_1(i), \hat{X}_2(i), \cdots, \hat{X}_J(i)$ are asymptotically i.i.d. $\mathcal{N}(0, \sigma^2)$ conditioned on $i \notin S$ and i.i.d. $\mathcal{N}(0, 1+\sigma^2)$ conditioned on $i \in S$. Thus, the total magnitude $Z = \sum_{j=1}^{J} X_j^2$ is a sufficient statistic for estimation of $\mathbf{1}(i \in S)$, and it is straightforward to show that the optimal estimator has the form $\hat{H} = \mathbf{1}(Z > t^*)$ where

$$t^* = \arg\min_t \max(\Pr[i \in S, Z < t], \Pr[i \notin S, Z \geq t]). \tag{5.22}$$

With a bit of work, it can be verified that this occurs when

$$(1-\kappa)\Pr[\sigma^2 \chi_J^2 > t^*] = \kappa \Pr[(1+\sigma^2)\chi_J^2 \leq t^*]. \tag{5.23}$$

Using the fact that the limiting distortion is given by

$$D = \tfrac{1}{\kappa} \max\left(\Pr[i \in S, Z < t^*], \Pr[i \notin S, Z \geq t^*]\right) \tag{5.24}$$

and solving for t^* completes the proof.

Chapter 6

A Compressed Sensing Wire-Tap Channel

This chapter studies a multiplicative Gaussian wire-tap channel inspired by compressed sensing. Lower and upper bounds on the secrecy capacity are derived, and shown to be relatively tight in the large system limit for a large class of compressed sensing matrices. Surprisingly, it is shown that the secrecy capacity of this channel is nearly equal to the capacity without any secrecy constraint provided that the channel of the eavesdropper is strictly worse than the channel of the intended receiver. In other words, the eavesdropper can see almost everything and yet learn almost nothing. This behavior, which contrasts sharply with that of many commonly studied wiretap channels, is made possible by the fact that a small number of linear projections can make a crucial difference in the ability to estimate sparse vectors.

6.1 Secrecy and Compressed Sensing

Following Shannon's theory in 1949 of information-theoretic secrecy [65], Wyner introduced the wiretap channel in 1975 [82]. In the wiretap setting, a sender Alice wishes to communicate a message to a receiver Bob over a main channel but her transmissions are intercepted by an eavesdropper Eve through a secondary wiretap channel. This chapter analyzes a multiplicative Gaussian wiretap channel inspired by compressed sensing. The input to the channel is a p-length binary vector. The channel output is a linear transform of the input after it has first been corrupted by multiplicative white Gaussian noise. We analyze the setting where Bob and Eve observe different linear transforms characterized by two different channel matrices.

Secrecy via compressed sensing schemes has received little attention from an information-theoretic viewpoint. In prior work, authors consider using a sensing matrix as a key (unknown to the eavesdropper) for both encryption and compression [54]. Privacy via compressed

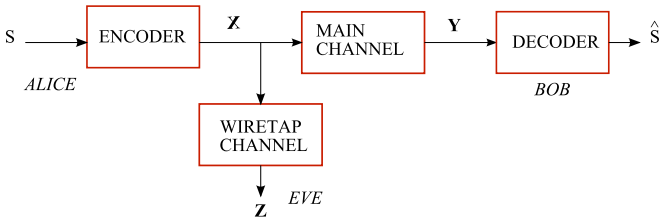

Figure 6.1: *(Multiplicative Gaussian Wiretap Channel)* For each block length n, Alice transmits a sequence of n binary valued support vectors $\mathbf{X} \in \{0,1\}^p$ over a main channel characterized by a matrix transform so that Bob receives $\mathbf{Y} = A_b W \mathbf{X}$. The eavesdropper Eve receives $\mathbf{Z} = A_e W \mathbf{X}$.

sensing and linear programming decoding was explored in [24]. By contrast, this chapter assumes that the sensing matrices are known (non-secret); as a special case, Eve's sensing matrix might correspond to a subset of the rows of Bob's channel matrix. Our analysis shows that certain channel matrices, inspired by compressed sensing, allow for secrecy rates that are nearly equal to the main channel capacity even if Eve's capacity is large.

6.1.1 Channel Model

Outlined in Fig. 6.1, the multiplicative Gaussian wiretap channel with binary vector input is characterized by

$$\mathbf{Y} = A_b W \mathbf{X}, \tag{6.1}$$
$$\mathbf{Z} = A_e W \mathbf{X}, \tag{6.2}$$

where $\mathbf{X} \in \{0,1\}^p$ is the transmitted signal, and $\mathbf{Y} \in \mathbb{R}^{m_b}$, $\mathbf{Z} \in \mathbb{R}^{m_e}$ are the received real-valued signals at the legitimate user and eavesdropper, respectively. A related channel model in [42] also involves a wire-tap setting with binary input and real-valued output. The dimensions of the channel satisfy $0 \leq m_e < m_b < p/2$. The linear mixing parameters, matrices $A_b \in \mathbb{R}^{m_b \times p}$ and $A_e \in \mathbb{R}^{m_e \times p}$, are fixed and known to all parties. The randomness of the channel is derived from $W \in \mathbb{R}^{p \times p}$, a diagonal matrix whose values are i.i.d. Gaussian random variables with mean zero and variance one. The channel is assumed to be memoryless between channel uses.

6.1.2 Secrecy Capacity

Alice selects a message $S_n \in [1 : 2^{npR}]$, where R represents a normalized rate, and wishes to communicate reliably with Bob while keeping the message secret from Eve. A $(2^{npR}, n)$ secrecy code for the multiplicative wiretap channel consists of the following: (1) A message set $[1 : 2^{npR}]$; (2) A randomized encoder that generates a codeword $\mathbf{X}^n(S_n)$, $S_n \in [1 : 2^{npR}]$,

according to $P_{\mathbf{X}^n|S_n}$; (3) A decoder that assigns a message $\hat{S}_n(\mathbf{Y}^n)$ to each received sequence $\mathbf{Y}^n \in \mathcal{Y}^n$. The message S_n is a random variable with entropy satisfying

$$\lim_{n\to\infty} \frac{H(S_n)}{np} = R. \tag{6.3}$$

A secrecy code is reliable if

$$\lim_{n\to\infty} \Pr[\hat{S}_n(\mathbf{Y}^n) \neq S_n] = 0. \tag{6.4}$$

A secrecy code is secret if the information leakage rate tends to zero as block length $n \to \infty$,

$$\lim_{n\to\infty} \frac{I(\mathbf{Z}^n; S_n)}{n} = 0. \tag{6.5}$$

Note that this leakage rate is not normalized by p. A normalized rate R is achievable if there exists a sequence of $(2^{npR}, n)$ secrecy codes satisfying both Eqn. (6.4) and Eqn. (6.5). The secrecy capacity C_s is the supremum over all achievable rates.

6.1.3 Outline of Results

To analyze the secrecy capacity C_s, we first develop bounds as a function of the channel matrices A_b and A_e. We then analyze these bounds for certain random matrices in the large system limit where $m_b/p \to \rho_b$ and $m_e/p \to \rho_e$ as $p \to \infty$ for fixed constants $0 \leq \rho_e \leq \rho_b \leq 1/2$. Lower bounds on the secrecy capacity, corresponding to Wyner's coding strategy for discrete memoryless channels are developed in Section 6.2.1. Corresponding upper bounds are derived in Section 6.2.2. Section 6.2.3 provides an improved upper bound under a certain encoding constraint on Alice. Proofs are given in Section 6.3.

6.1.4 Notations

For a matrix $A \in \mathbb{R}^{m\times p}$ and vector $\mathbf{x} \in \{0,1\}^p$, we use $A(\mathbf{x})$ to denote the matrix formed by concatenating the columns indexed by $\mathbf{x}$, and we use $A(i)$ to denote the ith column of A. Also, we use $\mathcal{X}_k^p$ to denote the set of all binary vectors $\mathbf{x} \in \{0,1\}^p$ with exactly k ones. We use $H_2(x) = -x\log x - (1-x)\log(1-x)$ to denote binary the binary entropy function. We use log to denote the logarithm with base two and ln to denote the logarithm with the natural base.

6.2 Bounds on the Secrecy Capacity

Csiszar and Korner showed in [15] that the secrecy capacity of a discrete memoryless wiretap channel is given by

$$C_s = \max_{(U,\mathbf{X})} \left[\tfrac{1}{p} I(U;\mathbf{Y}) - \tfrac{1}{p} I(U;\mathbf{Z})\right] \tag{6.6}$$

where the auxiliary random variable U satisfies the Markov chain relationship: $U \to \mathbf{X} \to (\mathbf{Y}, \mathbf{Z})$. It can be verified that this is also the secrecy capacity when the channels have discrete inputs and continuous outputs (see e.g. [42]).

In some special cases, the secrecy capacity can be computed easily from (6.6). For example, if A_b and A_e correspond to the first m_b and m_e rows of the $p \times p$ identity matrix respectively, then it is straightforward to show that

$$C_s = \begin{cases} \frac{m_b}{p} - \frac{m_e}{p}, & \text{if } m_e < m_b \\ 0 & \text{if } m_e \geq m_b \end{cases}. \tag{6.7}$$

In this case, the secrecy capacity happens to be the difference of the individual channel capacities; thus as m_e approaches m_b the secrecy capacity tends to zero. In the following sections we will develop bounds for a class of matrices inspired by compressed sensing. Interestingly, we will see that the secrecy behavior of these matrices differs greatly from the behavior shown in (6.7).

6.2.1 Lower Bounds

We say that a matrix $A \in \mathbb{R}^{m \times p}$ is *fully linearly independent* (FLI) if the span of each submatrix $\{A(\mathbf{x}) \in \mathbb{R}^{m \times m-1} \ : \ \mathbf{x} \in \mathcal{X}^p_{m-1}\}$ defines a unique linear subspace of $\mathbb{R}^m$. Examples of FLI matrices include the first m rows of the $p \times p$ discrete cosine transform matrix or, with probability one, any matrix whose entries are drawn i.i.d. from a continuous distribution. A counter example is given by the first m rows of the $p \times p$ identity matrix.

Our first result, which is proved in Section 6.3.1, gives a general lower bound on the secrecy capacity for any FLI matrices.

Theorem 6.1. *Suppose that A_b and A_e are fully linearly independent. If $m_e < m_b$, then the secrecy capacity is lower bounded by*

$$\begin{aligned} C_s \geq & \tfrac{1}{p} \log \tbinom{p}{m_b - 1} - \tfrac{1}{2p} \log \det(\tfrac{1}{p} A_e A_e^T) \\ & + \sum_{\mathbf{x} \in \mathcal{X}^p_{m_b-1}} \tfrac{1}{\binom{p}{m_b-1} 2p} \log \det \left(\tfrac{1}{m_b - 1} A_e(\mathbf{x}) A_e(\mathbf{x})^T \right). \end{aligned} \tag{6.8}$$

The lower bound in Theorem 6.1 is derived by evaluating the right hand side of (6.6) when $\mathbf{X}$ is distributed uniformly over the set $\mathcal{X}^p_{m_b-1}$. We note that the condition $m_e < m_b$ is necessary to obtain a nontrivial lower bound since the secrecy capacity may be equal to zero otherwise.

Unfortunately, the bound in Theorem 6.1 is difficult to compute if m_b and p are large. One way to address this issue is to analyze the behavior for a random matrix (random matrices are denoted via boldface, uppercase letters). The following result is proved in Section 6.3.2.

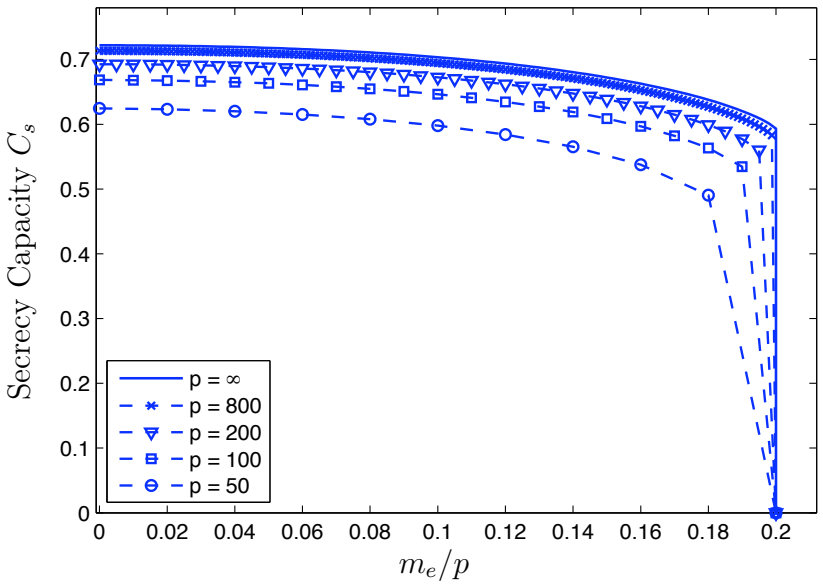

Figure 6.2: Illustration of the lower bound in Theorem 6.2 on the expected secrecy capacity $\mathbb{E}_{\mathbf{A}_e}[C_s]$ as a function of m_e for various values of p when A_b is fully linearly independent, $\mathbf{A}_e$ is a random matrix whose elements are i.i.d. $\mathcal{N}(0,1)$, and $m_b/p = 0.2$.

Theorem 6.2. *Suppose that A_b is fully linearly independent and $\mathbf{A}_e$ is a random matrix whose elements are i.i.d. $\mathcal{N}(0,1)$. If $m_e < m_b$ then the expectation of the secrecy capacity is lower bounded by*

$$\begin{aligned}\mathbb{E}_{\mathbf{A}_e}[C_s] \geq\ & \tfrac{1}{p}\log\binom{p}{m_b-1} - \tfrac{m_e}{2p}\log\left(\tfrac{m_b-1}{p}\right) \\ & - \tfrac{\log e}{2p}\sum_{i=1}^{m_e}\left[\psi\left(\tfrac{p-i+1}{2}\right) - \psi\left(\tfrac{m_b-i}{2}\right)\right] \end{aligned} \tag{6.9}$$

where $\psi(x) = \Gamma'(x)/\Gamma(x)$ is Euler's digamma function.

One benefit of Theorem 6.2 is that the bound is independent of the realization of the matrix $\mathbf{A}_e$ and can be analyzed directly. An illustration of the bound is shown in Figure 6.2 as a function of m_e/p for various values of p with m_b/p held fixed. Remarkably, as p becomes large, the lower bound in Theorem 6.2 remains bounded away from zero for all values of m_e strictly less than m_b. This behavior is in stark contrast to the secrecy capacity shown in (6.7).

One shortcoming of Theorem 6.2, is that the bound holds only in expectation, and it is possible that it is violated for a constant fraction of matrices $\mathbf{A}_e$. The next result, which is proved in Section 6.3.3, shows that, in the asymptotic setting, the limit of the bound (6.9) holds for almost every realization of $\mathbf{A}_e$. We use the notation $\{A^{(p)} \in \mathbb{R}^{m^{(p)}\times p}\}$ to denote a sequence of matrices indexed by the number of columns p.

Theorem 6.3. *Suppose that* $\{A_b^{(p)} \in \mathbb{R}^{m_b^{(p)} \times p}\}$ *is a sequence of linearly independent matrices and* $\{\mathbf{A}_e^{(p)} \in \mathbb{R}^{m_e^{(p)} \times p}\}$ *is a sequence of random matrices whose elements are i.i.d.* $\mathcal{N}(0,1)$*. If* $m_e^{(p)} > m_b^{(p)}$ *and* $m_b^{(p)}/p \to \rho_b$ *and* $m_e^{(p)}/p \to \rho_e$ *as* $p \to \infty$ *where* $0 \le \rho_e \le \rho_b \le 1/2$*, then the asymptotic secrecy capacity is lower bounded by*

$$\liminf_{p\to\infty} C_s \ge H_2(\rho_b) - \tfrac{1}{2}\Big[(1-\rho_e)\log\left(\tfrac{1}{1-\rho_e}\right) - (\rho_b-\rho_e)\log\left(\tfrac{\rho_b}{\rho_b-\rho_e}\right)\Big] \tag{6.10}$$

almost surely.

Theorem 6.3 provides a concise characterization of the lower bound in the asymptotic setting. The bound is illustrated in Figure 6.2 in the case $p = \infty$. Since the secrecy capacity can be equal to zero if $m_e^{(p)} = m_b^{(p)}$, Theorem 6.3 shows that there is a *discontinuity* in the asymptotic secrecy capacity as a function of ρ_e.

6.2.2 Upper Bounds via Channel Capacity

This section considers the capacity of Bob's channel which is denoted C_b. We note that this capacity gives us an upper bound on the secrecy capacity.

Upper bounding the capacity is more technically challenging than lower bounding the secrecy capacity, since the optimal distribution on $\mathbf{X}$ may depend nontrivially on channel matrix A_b. The following result, which is proved in Section 6.3.4, serves as a starting point.

Theorem 6.4. *If* A_b *is fully linearly independent, then the channel capacity of Bob's channel is upper bounded by*

$$C_b \le \tfrac{1}{p}\max\left(\log\binom{p}{m_b-1}, \max_{m_b \le k \le p} \tilde{c}(k)\right) + \tfrac{\log p}{p} \tag{6.11}$$

where

$$\begin{aligned}\tilde{c}(k) = &\max_{1\le i\le p} \tfrac{m_b}{2}\log\left(\tfrac{1}{m_b}\|A_b(i)\|^2\right) \\ &- \max_{\mathbf{x}\in\mathcal{X}_k^p} \tfrac{1}{2}\log\det(\tfrac{1}{k}A_b(\mathbf{x})A_b(\mathbf{x})^T).\end{aligned} \tag{6.12}$$

Although it is tempting to consider the expectation of (6.11) with respect to a random matrix (as we did for Theorem 6.2), this is difficult since the maximization in (6.12) occurs inside the expectation.

Our next result, which is proved in Section 6.3.5, leverages the strong concentration properties of the Gaussian distribution to characterize the asymptotic capacity for Gaussian matrices.

Theorem 6.5. *Suppose that* $\{\mathbf{A}_b^{(p)} \in \mathbb{R}^{m_b^{(p)} \times p}\}$ *is a sequence of random matrices whose elements are i.i.d.* $\mathcal{N}(0,1)$*. If* $m_b^{(p)}/p \to \rho_b$ *where* $0 < \rho_b \leq 1/2$*, then the asymptotic channel capacity of Bob's channel is given by*

$$\lim_{p\to\infty} C_b = H_2(\rho_b) \tag{6.13}$$

almost surely.

Theorem 6.5 shows that the strategy used in our lower bounds, namely choosing $\mathbf{X}$ uniformly over $\mathcal{X}_{m_b-1}^p$ achieves the capacity of Bob's channel in the asymptotic setting. What is remarkable is that for this same input distribution, Eve learns very little about what is being sent, even if her channel matrix is equal to the first m_e rows of Bob's channel matrix.

6.2.3 Improved Upper Bound for a Restricted Setting

We say that the distribution is *symmetric* if $\Pr[\mathbf{X} = \mathbf{x}] = \Pr[\mathbf{X} = \tilde{\mathbf{x}}]$ for all $\mathbf{x}, \tilde{\mathbf{x}}$ such that $\sum_{i=1}^p x_i = \sum_{i=1}^p \tilde{x}_i$. The following result is proved in Section 6.3.6.

Theorem 6.6. *Suppose that* $\{\mathbf{A}_b^{(p)} \in \mathbb{R}^{m_b^{(p)} \times p}\}$ *and* $\{\mathbf{A}_e^{(p)} \in \mathbb{R}^{m_e^{(p)} \times p}\}$ *are sequences of random matrices whose elements are i.i.d.* $\mathcal{N}(0,1)$*. If* $m_b^{(p)} > m_e^{(p)}$ *and* $m_b^{(p)}/p \to \rho_b$*, and* $m_e^{(p)}/p \to \rho_e$ *where* $0 \leq \rho_e \leq \rho_b \leq 1/2$*, and if Alice is restricted to use coding strategies that induce a symmetric distribution on* $\mathbf{X}$*, then the asymptotic secrecy capacity is upper bounded by*

$$\limsup_{p\to\infty} C_s \leq H(X|WX + \sqrt{\rho_b/\rho_e}V) \tag{6.14}$$

almost surely where $X \sim$ *Bernoulli*(ρ_b)*,* $W \sim \mathcal{N}(0,1)$ *and* $V \sim \mathcal{N}(0,1)$ *are independent random variables.*

The bound in Theorem 6.6 is strictly less than the channel capacity $H_2(\rho_b)$ for all $\rho_e > 0$, and can be computed easily using numerical integration. We suspect that this result also holds without the symmetry restriction on $\mathbf{X}$.

6.2.4 Illustration of Bounds

The bounds on the asymptotic secrecy capacity given in Theorems 6.3, 6.5, and 6.6 are illustrated in Fig. 6.3 as a function of the size parameter ρ_e of the eavesdropper channel. The bounds correspond to the setting where the elements of the matrices are i.i.d. Gaussian. Note that the lower bound on the secrecy capacity is nearly equal to that of the main channel for all $\rho_e < \rho_b$.

For comparison, the secrecy capacity for the special case where A_b and A_e correspond to the first rows of the $p \times p$ identity matrix are shown in Fig. 6.4. In this case, the secrecy capacity is equal to the difference between the main channel capacity and the eavesdropper capacity.

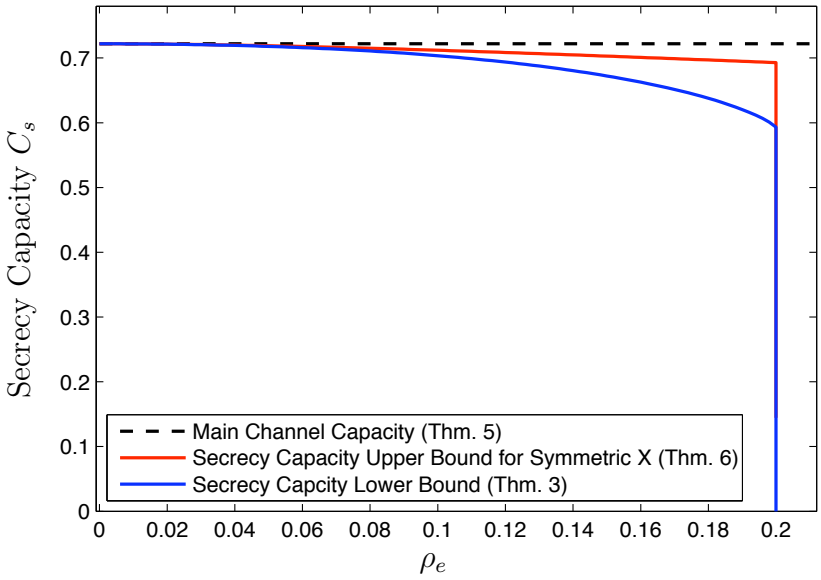

Figure 6.3: Bounds on the asymptotic (normalized) secrecy capacity C_s of the multiplicative Gaussian wiretap channel as a function of ρ_e when $\rho_b = 0.2$ and $\mathbf{A}_b$ and $\mathbf{A}_e$ are random matrices whose elements are i.i.d. $\mathcal{N}(0, 1)$.

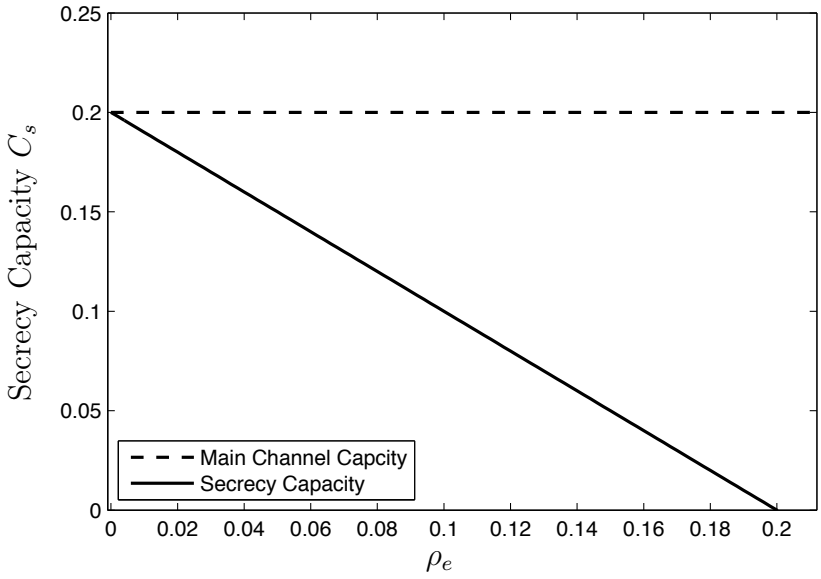

Figure 6.4: The (normalized) secrecy capacity C_s of the multiplicative Gaussian wiretap channel as a function of ρ_e when $\rho_b = 0.2$ and A_b and A_e correspond to the first rows of the identity matrix.

6.3 Proofs

6.3.1 Proof of Theorem 6.1

Let $U = \mathbf{X}$ where $\mathbf{X}$ is distributed uniformly over $\mathcal{X}^p_{m_b-1}$. Since A_b is fully linearly independent, the probability that $\mathbf{Y}$ is in the range space of $A_b(\tilde{\mathbf{x}})$ for any $\tilde{\mathbf{x}} \in \mathcal{X}^p_{m_b-1}$ not equal to the true vector $\mathbf{X}$ is equal to zero. Thus, $H(\mathbf{X}|\mathbf{Y}) = 0$ and

$$I(U;\mathbf{Y}) = I(\mathbf{X};\mathbf{Y}) = H(\mathbf{X}) = \log \tbinom{p}{m_b-1}. \tag{6.15}$$

Next, since A_e is fully linearly independent and the number of nonzero values in $\mathbf{X}$ is strictly greater than the rank of A_e, it can be verified that both $\mathbf{Z}$ and $\mathbf{Z}|\mathbf{X}$ have probability densities. (Note that the condition $m_b > m_e$ is critical here, since $\mathbf{Z}$ does not have a density otherwise.) Thus we can write

$$I(U;\mathbf{Z}) = I(\mathbf{X};\mathbf{Z}) = h(\mathbf{Z}) - h(\mathbf{Z}|\mathbf{X}) \tag{6.16}$$

where $h(\cdot)$ denotes differential entropy (see e.g. [14]). The entropy $h(\mathbf{Z})$ can be upper bounded as

$$\begin{aligned} h(\mathbf{Z}) &\leq \max_{\tilde{\mathbf{Z}}\,:\,\mathbb{E}[\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}]=\mathbb{E}[\mathbf{Z}\mathbf{Z}^T]} h(\tilde{\mathbf{Z}}) \\ &\leq \tfrac{1}{2}\log\Big((2\pi e)^{m_e}\det(\mathbb{E}[\mathbf{Z}\mathbf{Z}^T])\Big) \qquad (6.17) \\ &= \tfrac{1}{2}\log\Big((2\pi e\,(\tfrac{m_b-1}{p}))^{m_e}\det(A_eA_e^T)\Big) \qquad (6.18) \end{aligned}$$

where (6.17) follows from the fact that the Gaussian distribution maximizes the differential entropy and (6.18) follows from the fact that

$$\mathbb{E}[\mathbf{Z}\mathbf{Z}^T] = \tfrac{m_b-1}{p}A_eA_e^T. \tag{6.19}$$

The conditional entropy $h(\mathbf{Z}|\mathbf{X})$ is given by

$$h(\mathbf{Z}|\mathbf{X}) = \mathbb{E}\left[\tfrac{1}{2}\log\Big((2\pi e)^{m_e}\det(A_e(\mathbf{X})A_e(\mathbf{X})^T)\Big)\right] \tag{6.20}$$

where we used the fact that, conditioned on any realization $\mathbf{X} = \mathbf{x}$, $\mathbf{Z}$ is a (non-degenerate) Gaussian random vector with covariance matrix $A_e(\mathbf{x})A_e(\mathbf{x})^T$. Combining (6.15), (6.16), (6.18), and (6.20) with the expression of the secrecy capacity given in (6.6) completes the proof of Theorem 6.1.

6.3.2 Proof of Theorem 6.2

It is straightforward to show that $\mathbf{A}_e$ is fully linearly independent with probability one. Since the secrecy rate is bounded, it thus follows from Theorem 6.1 and the linearity of expectation

that

$$\begin{aligned}\mathbb{E}[C_s] \geq & \tfrac{1}{p}\log\tbinom{p}{m_b-1} - \tfrac{m_e}{2p}\log\left(\tfrac{m_b-1}{p}\right) \\ & - \frac{1}{2p}\mathbb{E}\Big[\log\det(\mathbf{A}_e\mathbf{A}_e^T)\Big] \\ & + \sum_{\mathbf{x}\in\mathcal{X}^p_{m_b-1}} \tfrac{1}{\binom{p}{m_b-1}2p}\mathbb{E}\Big[\log\det\Big(\mathbf{A}_e(\mathbf{x})\mathbf{A}_e(\mathbf{x})^T\Big)\Big]. \end{aligned} \tag{6.21}$$

Using well known properties of random Gaussian matrices (see e.g. [49, pp. 99-103]) shows that

$$\begin{aligned}\mathbb{E}\Big[\log\det(\mathbf{A}_e\mathbf{A}_e^T)] &= m_e + \log e\sum_{i=1}^{m_e}\psi(\tfrac{p-i+1}{2}) \\ \mathbb{E}\Big[\log\det(\mathbf{A}_e(\mathbf{x})\mathbf{A}_e(\mathbf{x})^T)] &= m_e + \log e\sum_{i=1}^{m_e}\psi(\tfrac{m_b-i}{2})\end{aligned}$$

where the second equality holds for every $\mathbf{x}\in\mathcal{X}^p_{m_b-1}$. Plugging these expressions into (6.21) completes the proof.

6.3.3 Proof of Theorem 6.3

Since $\mathbf{A}_e$ is fully linearly independent with probability one, it is sufficient to consider the asymptotic behavior of the bound in Theorem 6.1. If $\mathbf{X}$ is a random vector distributed uniformly over $\mathcal{X}^p_{m_b-1}$ then $\mathbf{A}_e\mathbf{A}_e^T$ and $\mathbf{A}_e(\mathbf{X})\mathbf{A}_e(\mathbf{X})^T$ are $m_e\times m_e$ Wishart matrices with p and m_b-1 degrees of freedom respectively. Using Lemma 6.6 gives

$$\begin{aligned}\lim_{p\to\infty}\tfrac{1}{p}\log\det(\tfrac{1}{p}\mathbf{A}_e\mathbf{A}_e^T) &= \mu(\rho_e) \\ \lim_{p\to\infty}\tfrac{1}{p}\log\det(\tfrac{1}{m_b-1}\mathbf{A}_e(\mathbf{X})\mathbf{A}_e^T(\mathbf{X})) &= \rho_b\,\mu(\rho_e/\rho_b)\end{aligned}$$

almost surely where $\mu(r) = (1-r)\ln\left(\frac{1}{1-r}\right) - r\log e$. Thus,

$$\begin{aligned}
&\lim_{p\to\infty}\Bigg[\tfrac{m_e}{p}\log\left(\tfrac{m_b-1}{p}\right)\tfrac{1}{p}\log\det(\mathbf{A}_e\mathbf{A}_e^T) \\
&\quad - \sum_{\mathbf{x}\in\mathcal{X}_{m_b-1}^p} \tfrac{1}{\binom{p}{m_b-1}p}\log\det(\mathbf{A}_e(\mathbf{x})\mathbf{A}_e(\mathbf{x})^T)\Bigg] \\
&\quad = \lim_{p\to\infty}\Bigg[\tfrac{m_e}{p}\log\left(\tfrac{m_b-1}{p}\right)\tfrac{1}{p}\log\det(\mathbf{A}_e\mathbf{A}_e^T) \\
&\qquad - \tfrac{1}{p}\log\det(\mathbf{A}_e(\mathbf{X})\mathbf{A}_e(\mathbf{X})^T)\Bigg] \qquad (6.22)\\
&\quad = \mu(\rho_b) - \rho_b\,\mu(\rho_e/\rho_b) \\
&\quad = (1-\rho_e)\log\left(\tfrac{1}{1-\rho_e}\right) - (\rho_b-\rho_e)\log\left(\tfrac{\rho_b}{\rho_b-\rho_e}\right) \qquad (6.23)
\end{aligned}$$

almost surely where the substitution in (6.22) is justified by the fact that the expectation of $\frac{1}{m_b}\log\det(\frac{1}{m_b}\mathbf{A}(\mathbf{X})\mathbf{A}_e(\mathbf{X})^T)$ with respect to both $\mathbf{A}$ and $\mathbf{X}$ is bounded uniformly for all p (see the proof of Theorem 6.2).

Combining (6.23) with the well known fact that

$$\lim_{p\to\infty}\tfrac{1}{p}\log\binom{p}{m_b-1} = H_2(\rho_b) \qquad (6.24)$$

completes the proof of Theorem 6.3.

6.3.4 Proof of Theorem 6.4

Let $K = \sum_{i=1}^p X_i$ denote the number of ones in $\mathbf{X}$. Then,

$$\begin{aligned}
I(\mathbf{X};\mathbf{Y}) &= I(\mathbf{X};\mathbf{Y}|K) + I(\mathbf{Y};K) &(6.25)\\
&\le I(\mathbf{X};\mathbf{Y}|K) + \log p &(6.26)\\
&\le \max_{0\le k\le p} I(\mathbf{X};\mathbf{Y}|K=k) + \log p &(6.27)
\end{aligned}$$

where (6.25) follows from the chain rule for mutual information, (6.26) follows from the fact that $I(\mathbf{Y};K) \le H(k) \le \log p$, and (6.27) follows from expanding the term $I(\mathbf{X};\mathbf{Y}|K)$. If we define

$$c(k) = \max_{\mathbf{X}\in\mathcal{X}_k^p} I(\mathbf{X};\mathbf{Y})$$

then we have

$$\max_{\mathbf{X}} I(\mathbf{X};\mathbf{Y}) \le \max_{0\le k\le p} c(k) + \log p. \qquad (6.28)$$

To complete the proof, we split the maximization over k into two cases. For $0 \leq k < m_b$ we use the simple bound

$$\max_{0\leq k<m_b} c(k) \leq \max_{\mathbf{X}\in\mathcal{X}_k\,:\,0\leq k<m_b} H(\mathbf{X}) = \log\tbinom{p}{m_b-1}.$$

For $m_b \leq k \leq p$ we use the following lemma.

Lemma 6.1. *If $m_b \leq k \leq p$, then $c(k) \leq \tilde{c}(k)$ where $\tilde{c}(k)$ is given in* (6.12)*.*

Proof. Let $\mathbf{X}$ have any distribution on $\mathcal{X}_k^p$ where $m_b \leq k \leq p$. Since A_e is fully linearly independent, both $\mathbf{Y}$ and $\mathbf{Y}|\mathbf{X}$ have probability densities and we can write

$$I(\mathbf{X};\mathbf{Y}) = h(\mathbf{Y}) - h(\mathbf{Y}|\mathbf{X}) \tag{6.29}$$

where $h(\cdot)$ denotes differential entropy (see e.g. [14]). The entropy $h(\mathbf{Y})$ can be upper bounded as

$$\begin{aligned} h(\mathbf{Y}) &\leq \max_{\tilde{\mathbf{Y}}\,:\,\mathbb{E}[\|\tilde{\mathbf{Y}}\|^2]=\mathbb{E}[\|\mathbf{Y}\|^2]} h(\tilde{\mathbf{Y}}) \\ &= \tfrac{m_b}{2}\log\Big(2\pi e\,\tfrac{1}{m_b}\mathbb{E}[\|\mathbf{Y}\|^2]\Big) &(6.30)\\ &\leq \max_{1\leq i\leq p} \tfrac{m_b}{2}\log\Big(2\pi e\tfrac{k}{m_b}\|A_b(i)\|^2\Big) &(6.31)\end{aligned}$$

where (6.30) follows from the fact that an isotropic Gaussian vector maximizes differential entropy, and (6.31) follows from the fact that

$$\begin{aligned} \mathbb{E}[\|\mathbf{Y}\|^2] &= \mathbb{E}\Big[\mathbb{E}[\|\mathbf{Y}\|^2|\mathbf{X}]\Big] \\ &\leq \max_{\mathbf{x}} \mathbb{E}[\|\mathbf{Y}\|^2|\mathbf{X}=\mathbf{x}] \\ &= \max_{\mathbf{x}} \operatorname{tr}\Big(A_b(\mathbf{x})A_b(\mathbf{x})^T\Big) \\ &\leq \max_{1\leq i\leq p} k\|A_b(i)\|^2. \end{aligned}$$

The conditional entropy $h(\mathbf{Y}|\mathbf{X})$ is lower bounded by

$$\begin{aligned} h(\mathbf{Y}|\mathbf{X}) &= \mathbb{E}\left[\tfrac{1}{2}\log\Big((2\pi e)^{m_b}\det(A_b(\mathbf{X})A_b(\mathbf{X})^T)\Big)\right] \\ &\geq \min_{\mathbf{x}\in\mathcal{X}_k^p} \tfrac{1}{2}\log\Big((2\pi e)^{m_b}\det(A_b(\mathbf{x})A_b(\mathbf{x})^T)\Big) &(6.32)\end{aligned}$$

where we used the fact that, conditioned on any realization $\mathbf{X}=\mathbf{x}$, $\mathbf{Y}$ is a (non-degenerate) Gaussian random vector with covariance matrix $A_b(\mathbf{x})A_b(\mathbf{x})^T$. Combining (6.29), (6.31) and (6.32) completes the proof of Theorem 6.4. ∎

6.3.5 Proof of Theorem 6.5

Since $\mathbf{A}_e$ is fully linearly independent with probability one, it is sufficient to consider the asymptotic behavior of the bound in Theorem 6.4. The limit of the first term in the maximization is given by (6.24). To evaluate the second term, we use the following technical lemmas whose proofs are given in the Appendices 6.3.7 and 6.3.8.

Lemma 6.2.

$$\limsup_{p\to\infty} \max_{1\le i\le p} \frac{1}{m_b}\|\mathbf{A}_b(i)\|^2 \le 1 \tag{6.33}$$

almost surely.

Lemma 6.3. *If $k \ge m_b$ and $k/p \to \kappa$ where $\rho_b \le \kappa \le 1$, then*

$$\liminf_{p\to\infty} \min_{\mathbf{x}\in\mathcal{X}_k^p} \tfrac{1}{p}\log\det(\tfrac{1}{k}\mathbf{A}_b(\mathbf{x})\mathbf{A}_b(\mathbf{x})^T) \ge \kappa\,\mu(\rho_b/\kappa) \tag{6.34}$$

almost surely where $\mu(r) = (1-r)\log\left(\frac{1}{1-r}\right) - r\log e$.

The convergence show in Lemmas 6.2 and 6.3 leads immediately to the following asymptotic upper bound on the term $\tilde{c}(k)$ defined in (6.12):

$$\begin{aligned}
&\limsup_{p\to\infty} \max_{m_b<k\le p} \tfrac{1}{p}\tilde{c}(k)\\
&\le \max_{\rho_b\le\kappa\le 1} \tfrac{1}{2}\Big[\rho_b\log e - (\kappa-\rho_b)\log\left(\tfrac{\kappa}{\kappa-\rho_b}\right)\Big]\\
&= \tfrac{1}{2}\rho_b\log e
\end{aligned}$$

almost surely. Since $H_2(\rho_b) > \frac{1}{2}\rho_b\log e$ for all $\rho_b \in (0, 1/2)$, we conclude that the asymptotic capacity is upper bounded by $H_2(\rho_b)$. The achievable strategy outlined in the proof of Theorem 6.1 shows that $H_2(\rho_b)$ is also achievable which concludes the proof of Theorem 6.5.

6.3.6 Proof of Theorem 6.6

Let $K = \sum_{i=1}^p X_i$. Then, for any pair $(U, \mathbf{X})$ such that $U \to \mathbf{X} \to (\mathbf{Y}, \mathbf{Z})$, we have

$$\begin{aligned}
&I(U;\mathbf{Y}) - I(U;\mathbf{Z})\\
&= I(\mathbf{X};\mathbf{Y}) - I(\mathbf{X};\mathbf{Z}) + I(\mathbf{X};\mathbf{Z}|U) - I(\mathbf{X};\mathbf{Y}|U) && (6.35)\\
&\le I(\mathbf{X};\mathbf{Y}|K) - I(\mathbf{X};\mathbf{Z}|K)\\
&\quad + I(\mathbf{X};\mathbf{Z}|U,K) - I(\mathbf{X};\mathbf{Y}|U,K) + 2\log p && (6.36)\\
&\le \max_{0\le k\le p} \Delta(k) + 2\log p && (6.37)
\end{aligned}$$

where

$$\begin{aligned}\Delta(k) = I(\mathbf{X};\mathbf{Y}|K=k) - I(\mathbf{X};\mathbf{Z}|K=k) \\ + I(\mathbf{X};\mathbf{Z}|U,K=k) - I(\mathbf{X};\mathbf{Y}|U,K=k).\end{aligned}$$

We now consider two cases. For the case $m_b \leq k \leq p$, we use the upper bound

$$\begin{aligned}\Delta(k) &\leq I(\mathbf{X};\mathbf{Y}|K=k) + I(\mathbf{X};\mathbf{Z}|U,K=k) \\ &\leq I(\mathbf{X};\mathbf{Y}|K=k) + I(\mathbf{X};\mathbf{Z}|K=k)\end{aligned}$$

which follows from the non-negativity of mutual information and the data processing inequality. Following the steps outlined in the proofs of Theorems 6.4 and 6.5 shows that

$$\limsup_{p\to\infty} \max_{\mathbf{X}\in\mathcal{X}_k^p \,:\, m_b\leq k\leq p} \tfrac{1}{p}\Delta(k) \leq \tfrac{1}{2}(\rho_b+\rho_e)\log e \leq \rho_b \log e \tag{6.38}$$

almost surely. (Note that this step does not require the symmetry assumption.)

Alternatively, for the case $0 \leq k < m_b$, we use the bound

$$\Delta(k) \leq H(\mathbf{X}|\mathbf{Z},K=k) + H(\mathbf{X}|\mathbf{Y},K=k)$$

which follows from the non-negativity of entropy and the fact that conditioning cannot increase entropy. Since $\mathbf{A}_b$ is fully linearly independent almost surely, it follows from the proof of Theorem 6.1 that $H(\mathbf{X}|\mathbf{Y},K=k)$ is equal to zero almost surely. To characterize the asymptotic behavior of the remaining term, $H(\mathbf{X}|\mathbf{Z},K=k)$, we use the following lemma which is proved in Section 6.3.9.

Lemma 6.4. *Suppose that* $\mathbf{X}$ *is symmetric. If* $0 \leq k < m_b$ *and* $k/p \to \kappa$ *where* $0 \leq \kappa \leq \rho_b$, *then*

$$\limsup_{p\to\infty} \tfrac{1}{p} H(\mathbf{X}|\mathbf{Z},K=k) \leq g(\kappa,\rho_e) \tag{6.39}$$

almost surely where

$$g(\kappa,\rho_e) = H(X|WX+\sqrt{\kappa/\rho_e}V) \tag{6.40}$$

and $X \sim$ *Bernoulli*(κ), $W \sim \mathcal{N}(0,1)$ *and* $V \sim \mathcal{N}(0,1)$ *are independent random variables.*

Noting that $g(\kappa,\rho_e)$ is nondecreasing in κ, we obtain the asymptotic upper bound

$$\limsup_{p\to\infty} C_s \leq \max\Big(g(\rho_b,\rho_e),\rho_b\Big). \tag{6.41}$$

It can be verified numerically that this maximum occurs at $g(\rho_b,\rho_e)$ for all $\rho_b \in (0,1/2)$ which completes the proof of Theorem 6.6.

6.3.7 Proof of Lemma 6.2

Note that the column magnitudes $\|\mathbf{A}_b(i)\|^2, i = 1, 2, \cdots, p$ are i.i.d. chi-square random variables with m_b degrees of freedom. Thus for any $\epsilon \in (0, 1/2)$, the chi-square concentration inequality in Lemma 6.5 gives

$$\Pr[\max_{1\le i\le p} \tfrac{1}{m_b}\|\mathbf{A}_b(i)\|^2 \ge 1+\epsilon] \le p\exp(-\tfrac{3}{16}\lceil\rho_b p\rceil\epsilon^2) \tag{6.42}$$

which decays exponentially rapidly with p as $p \to \infty$.

6.3.8 Proof of Lemma 6.3

For each $\mathbf{x} \in \mathcal{X}_k^p$, let

$$N(\mathbf{A}_b, \mathbf{x}) = \tfrac{1}{k}\log\det\left(\tfrac{1}{k}\mathbf{A}(\mathbf{x})\mathbf{A}(\mathbf{x})^T\right). \tag{6.43}$$

By the union bound, and the symmetry of $\mathbf{A}_b$ we have

$$\Pr\left[\min_{\mathbf{x}\in\mathcal{X}_k^p} N(\mathbf{A}_b, \mathbf{x}) \le t\right] \le \binom{p}{k}\Pr[N(\mathbf{A}, \mathbf{x}) \le t], \tag{6.44}$$

for any arbitrary $\mathbf{x} \in \mathcal{X}_k^p$. Using the bound $\binom{p}{k} \le (pe/k)^k$ and Lemma 6.6, shows that for any $\epsilon > 0$,

$$\limsup_{p\to\infty} \tfrac{1}{p\ln p}\ln\Pr[N(\mathbf{A}, \mathbf{x}) \le \mu(\rho_b/\kappa) - \epsilon] \le -\epsilon$$

which suffices to prove almost sure convergence.

6.3.9 Proof of Lemma 6.4

Let $\hat{\mathbf{X}} = \frac{1}{m_e}\mathbf{A}_e^T\mathbf{Z}$. Then, we have

$$\begin{aligned} H(\mathbf{X}|\mathbf{Z}, K=k) &\le H(\mathbf{X}|\hat{\mathbf{X}}, K=k) && (6.45)\\ &\le \sum_{i=1}^{p} H(X_i|\hat{\mathbf{X}}, K=k) && (6.46)\\ &\le \sum_{i=1}^{p} H(X_i|\hat{X}_i, K=k) && (6.47)\end{aligned}$$

where (6.45) follows from the data processing inequality, (6.46) follows from the chain rule and (6.47) follows from the fact that conditioning cannot increase entropy.

Next, we observe that $\hat{X}_i$ can be written as

$$\hat{X}_i = \frac{1}{m_e}\sum_{i=1}^{p}\langle \mathbf{A}_e(i), \mathbf{A}_e(j)\rangle W_j X_j \tag{6.48}$$

$$= \frac{\|\mathbf{A}_e(i)\|^2}{m_e} W_i X_i + \sigma_i(\mathbf{A}_e, \mathbf{X})V \tag{6.49}$$

where $V \sim \mathcal{N}(0,1)$ is independent of W_i, X_i and $\sigma_i^2(\mathbf{A}_e, \mathbf{X})$ where

$$\sigma_i^2(\mathbf{A}_e, \mathbf{X}) = \frac{1}{m_e}\sum_{j\neq i}\langle \mathbf{A}_e(i), \mathbf{A}_e(j)\rangle^2 X_j. \tag{6.50}$$

Using standard chi-square inequalities, it is straightforward to show that

$$\lim_{p\to\infty}\max_{1\leq i\leq p} |\tfrac{\|\mathbf{A}_e(i)\|^2}{m_e} - 1| = 0 \tag{6.51}$$

almost surely. With a bit more work, and the use of the fact that, by the symmetry constraint, $\mathbf{X}$ is distributed uniformly over $\mathcal{X}_k^p$ it can also be shown that

$$\lim_{p\to\infty}\max_{1\leq i\leq p} |\sigma_j^2(\mathbf{A}_e, \mathbf{X}) - \tfrac{\kappa}{\rho_e}| = 0 \tag{6.52}$$

almost surely. Thus, we conclude that the empirical distribution of the pairs $(X_i, \hat{X}_i)$ converges weakly almost surely to the distribution on $(X, WX + \sqrt{\kappa/\rho_e}V)$, which concludes the proof.

6.3.10 Technical Lemmas

Lemma 6.5 ([39]). *If X is a chi-square random variable with n degrees of freedom then for all $\epsilon \in (0, 1/2)$,*

$$\Pr[X \geq d(1+\epsilon)] \leq \exp\Big(-\tfrac{3}{16}d\epsilon^2\Big). \tag{6.53}$$

Lemma 6.6. *Let $\mathbf{W}$ be an $m \times m$ Wishart random matrix with $n \geq m$ degrees of freedom. If $m/n \to \rho \in (0,1]$ as $n \to \infty$, then for any $\epsilon > 0$,*

$$\limsup_{n\to\infty} \tfrac{1}{n\ln n}\ln \Pr\Big[\Big|\tfrac{1}{n}\log\det(\tfrac{1}{n}\mathbf{W}) - \mu(\rho)\Big| > \epsilon\Big] \leq -\tfrac{\epsilon}{\log e}$$

where

$$\mu(\rho) = \begin{cases} (1-\rho)\log\left(\frac{1}{1-\rho}\right) - \rho\log e, & \text{if } 0<\rho<1 \\ -\log e, & \text{if } \rho = 1 \end{cases}. \tag{6.54}$$

Proof. We begin with a one-sided bound. For any $r > 0$ we have

$$\begin{aligned}
&\Pr\left[\tfrac{1}{n}\ln\det\left(\tfrac{1}{n}\mathbf{W}\right) \le t\right] \\
&= \Pr\left[\ln\det\left(\mathbf{W}\right) \le nt + m\log n\right] \\
&= \Pr\left[\Big(\det(\mathbf{W})\Big)^{-r} \ge \exp(-rnt + m\log n)\right] \\
&\le \exp(rnt + rm\log n)\,\mathbb{E}\left[\Big(\det(\mathbf{W})\Big)^{-r}\right] \qquad (6.55)
\end{aligned}$$

where (6.55) follows from Markov's inequality. If $r < (n-m)/2$, then it can be shown (see e.g. [49, pp. 99-103]) that

$$\mathbb{E}\left[\Big(\det(\mathbf{W})\Big)^{-r}\right] = \exp(-M(r))$$

where

$$M(r) = rm\ln(2) + \sum_{i=0}^{m-1}\left[\ln\Gamma(\tfrac{n-i}{2}) - \ln\Gamma(\tfrac{n-i}{2} - r)\right].$$

If r is an integer then we use the relation

$$\ln\Gamma(z) - \ln\Gamma(z-r) = \sum_{i=1}^{r}\ln(z-i)$$

to obtain

$$\begin{aligned}
M(r) &= rm\ln(2) + \sum_{i=0}^{m-1}\sum_{j=1}^{r}\ln\left(\tfrac{n-i}{2} - j\right) \\
&= rm\ln n + \sum_{i=0}^{m-1}\sum_{j=1}^{r}\ln\left(\tfrac{n-i-2j}{n}\right).
\end{aligned}$$

Plugging this back into (6.55) gives

$$\ln\Pr\left[\tfrac{1}{n}\ln\det(\tfrac{1}{n}\mathbf{W}) \le t\right] < rnt - \sum_{i=0}^{m-1}\sum_{j=1}^{r}\ln\left(\tfrac{n-i-2j}{n}\right).$$

We now consider what happens as $n \to \infty$. If $r = \ln n$ then it is straightforward to show that

$$\lim_{n\to\infty}\frac{1}{rn}\sum_{i=0}^{m-1}\sum_{j=1}^{r}\ln\left(\tfrac{n-i-2j}{n}\right) = \frac{\mu(\rho)}{\log e},$$

and thus

$$\limsup_{n\to\infty} \tfrac{1}{n \ln n} \ln \Pr\left[\tfrac{1}{n} \ln\det\left(\tfrac{1}{n}\mathbf{W}\right) \leq t\right] \leq t - \tfrac{\mu(\rho)}{\log e}.$$

To prove the other side of the bound, we use the same steps as before to obtain

$$\ln \Pr\left[\tfrac{1}{n} \ln\det(\tfrac{1}{n}\mathbf{W}) \geq t\right] < \sum_{i=0}^{m-1} \sum_{j=1}^{r} \log\left(\tfrac{n-i+2j}{n}\right) - rnt.$$

Letting $r = \ln n$ leads to

$$\limsup_{n\to\infty} \tfrac{1}{n \ln n} \ln \Pr\left[\tfrac{1}{n} \ln\det\left(\tfrac{1}{n}\mathbf{W}\right) \geq t\right] \leq \tfrac{\mu(\rho)}{\log e}, -t.$$

Changing the base of the logarithms concludes the proof of Lemma 6.6. □

Bibliography

[1] S. Aeron, V. Saligrama, and M. Zhao, "Information theoretic bounds for compressed sensing," *IEEE Trans. Inf. Theory*, vol. 56, no. 10, pp. 5111–5130, Oct. 2010.

[2] M. Akcakaya and V. Tarokh, "Shannon theoretic limits on noisy compressive sampling," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 492–504, Jan. 2010.

[3] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, "Model-based Compressive Sensing," *IEEE Trans. Inf. Theory*, vol. 56, pp. 1982–2001, Apr. 2010.

[4] D. Baron, M. F. Duarte, M. B. Wakin, S. Sarvotham, and R. G. Baraniuk, "Distributed compressed sensing," Rice University, Department of Electrical and Computer Engineering, Tech. Rep. TREE-0612, Nov. 2006.

[5] D. Baron, S. Sarvotham, and R. G. Baraniuk, "Bayesian compressive sensing via belief propagation," *IEEE Trans. Signal Process.*, vol. 58, no. 1, pp. 269–280, Jan. 2010.

[6] M. Bayati and A. Montanari, "The lasso risk for gaussian matrices," Aug. 2010, arXiv:1008.2581v1 [math.ST].

[7] ——, "The dynamics of message passing on dense graphs, with applications to compressed sensing," *IEEE Trans. Inf. Theory*, vol. 57, no. 2, pp. 764–785, Feb. 2011.

[8] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[9] ——, "Stable signal recovery from incomplete and inaccurate measurements," *Comm. on Pure and Applied Math.*, vol. 59, pp. 1207–1223, Feb. 2006.

[10] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.

[11] ——, "Near optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.

[12] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. of Sci. Comp.*, vol. 20, no. 1, pp. 33–61, 1999.

[13] S. F. Cotter, B. D. Rao, K. Engan, and K. Kreutz-delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Trans. Signal Process.*, vol. 53, no. 7, pp. 2477–2488, Jul. 2005.

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

[15] I. Csiszar and J. Korner, "Broadcast channels with confidential messages," *IEEE Transactions on Information Theory*, vol. 24, no. 3, pp. 339–348, 1978.

[16] R. A. DeVore and G. G. Lorentz, *Constructive Approximation*. New York, NY: Springer Verlag, 1993.

[17] D. L. Donoho, "For most large underdetermined systems of linear equations, the minimal ℓ_1-norm solution is also the sparsest solution," *Comm. on Pure and Applied Math.*, vol. 59, no. 6, pp. 797–829, Jun. 2006.

[18] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, 1994.

[19] D. L. Donoho and J. Tanner, "Counting faces of randomly-projected polytopes when the projection radically lowers dimension," *J. Amer. Match. Soc.*, vol. 22, no. 1, pp. 1–53, Jan. 2009.

[20] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[21] D. L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing," in *Proc. Nat. Acad. Sci.*, vol. 106, 2009, pp. 18 914–18 919.

[22] ——, "The noise-sensitivity phase transition in compressed sensing," Apr. 2010. [Online]. Available: http://arxiv.org/abs/1004.1218

[23] D. Donoho, M. Elad, and V. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Inf. Theory.*, vol. 52, no. 1, pp. 6–18, Jan. 2006.

[24] C. Dwork, F. McSherry, and K. Talwar, "The Price of Privacy and the Limits of LP Decoding," *Annual ACM Symposium on Theory of Computing*, vol. 39, pp. 85–94, 2007.

[25] S. F. Edwards and P. W. Anderson, "Theory of spin glasses," *J. Phys. F: Metal Physics*, vol. 5, pp. 965–974, 1975.

[26] Y. C. Eldar, P. Kuppinger, and H. Bolcskei, "Block-sparse signals: Uncertainty relations and efficient recovery," *IEEE Trans. Signal Process.*, vol. 58, no. 6, pp. 3042–3054, Jun. 2010.

[27] 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.

[28] A. K. Fletcher, S. Rangan, and V. K. Goyal, "Rate-distortion bounds for sparse approximation," in *Proc. IEEE Statist. Sig. Process. Workshop*, Madison, WI, Aug. 2007, pp. 254–258.

[29] ——, "Compressive sampling and lossy compression," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 48–56, Mar. 2008.

[30] ——, "Necessary and sufficient conditions for sparsity pattern recovery," *IEEE Trans. Inf. Theory*, vol. 55, no. 12, pp. 5758–5772, Dec. 2009.

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

[32] J. J. Fuchs, "Recovery of exact sparse representations in the presence of noise," *IEEE Trans. Inf. Theory*, vol. 51, no. 10, pp. 3601–3608, Oct. 2005.

[33] 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.

[34] A. Grant and C. Schlegel, "Convergence of linear interference cancellation multiuser receivers," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 1823–1834, Oct. 2001.

[35] D. Guo, D. Baron, and S. Shamai, "A single-letter characterization of optimal noisy compressed sensing," in *Proc. Allerton Conf. on Comm., Control, and Computing*, Monticello, IL, Sep. 2009.

[36] D. Guo and S. Verdú, "Randomly spread CDMA: Asymptotics via statistical physics," *IEEE Trans. Inf. Theory*, vol. 51, no. 6, pp. 1983–2010, Jun. 2005.

[37] D. Guo, S. Shamai, and S. Verdú, "Mutual information and minimum mean-square error in Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 51, pp. 1261–1282, Apr. 2005.

[38] J. Haupt and R. Nowak, "Signal reconstruction from noisy random projections," *IEEE Trans. Inf. Theory*, vol. 59, no. 8, pp. 1207–1223, Aug. 2006.

[39] I. M. Johnstone, "Chi-square oracle inequalities," *State of the Art in Probability and Statistics*, vol. 37, pp. 399–418, 2001.

[40] Y. Kabashima, T. Wadayama, and T. Tanaka, "A typical reconstruction limit of compressed sensing based on lp norm minimization," *J. Stat. Mech.*, 2009.

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

[42] R. Liu, Y. Liang, H. V. Poor, and P. Spasojevic, "Secure Nested Codes for Type-II Wire-tap Channels," *Proc. of IEEE Information Theory Workshop (ITW)*, 2007.

[43] K. Lounici, M. Pontil, A. B. Tsybakov, and S. van de Geer, "Taking advantage of sparsity in multi-task learning," arXiv:0903.1468v1 [stat.ML], Mar. 2009.

[44] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.

[45] P. Massart, *Concentration Inequalities and Model Selection*. Springer, 2007.

[46] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the lasso," *Annals of Stat.*, vol. 34, pp. 1436–1462, 2006.

[47] A. J. Miller, *Subset selection in regression*. New York, NY: Chapman-Hall, 1990.

[48] A. Montanari. (2011, Mar.) Graphical model concepts in compressed sensing. [Online]. Available: http://arxiv.org/abs/1011.4328

[49] R. J. Muirhead, *Aspects of Multivariate Statistical Theory*. New York, NY: John Wiley and Sons, 1982.

[50] R. R. Muller, "Channel capacity and minimum probability of error in large dual antenna array systems with binary modulation," *IEEE Trans. Inf. Theory*, vol. 51, no. 11, pp. 2821–2822, Nov. 2003.

[51] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Computing*, vol. 24, no. 2, pp. 227–234, Apr. 1995.

[52] G. Obozinski, M. J. Wainwright, and M. I. Jordan, "Union support recovery in high-dimensional multivariate regression," Department of Statistics, UC Berkeley, Tech. Rep., Aug. 2008.

[53] C. C. Paige and M. Wei, "History and generality of the CS decomposition," *Linear Algebra Appl.*, vol. 208-209, pp. 303–326, 1994.

[54] Y. Rachlin and D. Baron, "The secrecy of compressed sensing measurements," in *Communication, Control, and Computing, 2008 46th Annual Allerton Conference*. IEEE, 2009, pp. 813–817.

[55] S. Rangan, A. K. Fletcher, and V. K. Goyal, "Asymptotic analysis of map estimation via the replica method and applications to compressed sensing," in *Proc. Neural Information Processing Systems Conf.*, vol. 22, Vancouver, CA, Dec. 2009, pp. 1545–1553.

[56] S. Rangan, "Estimation with random linear mixing, belief propagation and compressed sensing," Jan. 2010. [Online]. Available: http://arxiv.org/abs/1001.2228

[57] ——, "Generalized approximate message passing for estimation with random linear mixing," Oct. 2010. [Online]. Available: http://arxiv.org/abs/1010.5141

[58] G. Reeves, "Sparse signal sampling using noisy linear projections," Department of EECS, UC Berkeley, Tech. Rep. UCB/EECS-2008-3, Jan. 2008.

[59] G. Reeves and M. Gastpar, "Sapling rate-distortion tradeoffs for structured sparsity," To appear.

[60] ——, "Sampling bounds for sparse support recovery in the presence of noise," in *Proc. IEEE Int. Symp. on Inf. Theory*, Toronto, Canada, Jul. 2008.

[61] ——, "Efficient sparsity pattern recovery," in *Proc. 30th Symposium on Information Theory*, Eindhoven, May. 2009.

[62] ——, "Approximate sparsity pattern recovery: Information-theoretic lower bounds," Feb. 2010, arXiv:1002.4458v1 [cs.IT].

[63] J. Salo, D. Seethaler, and A. Skupch, "On the asymptotic geometric mean of mimo channel eigenvalues," in *Proc. IEEE Int. Symp. on Inform. Theory*, Seattle, WA, Jul. 2006.

[64] S. Sarvotham, D. Baron, and R. G. Baraniuk, "Measurements vs. bits: Compressed sensing meets information theory," in *Proc. Allerton Conf. on Comm., Control, and Computing*, Monticello, IL, Sep. 2006.

[65] C. E. Shannon, "Communication Theory of Secrecy Systems," *Bell System Technical Journal*, vol. 28, pp. 656–715, 1949.

[66] M. Stojnic, F. Parvaresh, and B. Hassibi, "On the reconstruction of block-sparse signals with an optimal number of measurements," *IEEE Trans. Signal Process.*, vol. 57, no. 8, pp. 3075–3085, Aug. 2009.

[67] T. Tanaka, "A statistical-mechanics approach to large-system analysis of cdma multiuser detectors," *IEEE Trans. Inf. Theory*, vol. 48, no. 11, pp. 2888–2910, Nov. 2002.

[68] R. Tibshirani, "Regression shrinkage and selection via the lasso,," *J. Royal Stat. Soc., Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.

[69] J. Tropp, "Just relax: Convex programming methods for identifying sparse signals in noise," *IEEE Trans. Inf. Theory*, vol. 52, no. 3, pp. 1030–1051, Mar. 2006.

[70] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Algorithms for simultaneous sparse approximation part I: Greedy pursuit," *Signal Processing*, vol. 86, pp. 572–588, Apr. 2006.

[71] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.

[72] D. N. C. Tse and S. V. Hanly, "Linear multiuser receivers: Effective interference, effective bandwidth and user capacity," *IEEE Trans. Inf. Theory*, vol. 45, no. 2, pp. 641–657, Mar. 1999.

[73] N. Vaswani and W. Lu, "Modified-cs: Modifying compressive sensing for problems with partially known support," *IEEE Trans. Signal Process.*, vol. 58, no. 9, pp. 4595–4607, Sep. 2010.

[74] 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.

[75] S. Verdú and S. Shamai, "Spectral efficiency of CDMA with random spreading," *IEEE Trans. Inf. Theory*, vol. 45, no. 2, pp. 622–640, Mar. 1999.

[76] M. J. Wainwright, "Information-theoretic limitations on sparsity recovery in the high-dimensional and noisy setting." *IEEE Trans. Inf. Theory*, vol. 55, pp. 5728–5741, Dec. 2009.

[77] ——, "Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming (lasso)," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2183–2202, May 2009.

[78] W. Wang, M. J. Wainwright, and K. Ramchandran, "Information-theoretic limits on sparse signal recovery: Dense versus sparse measurement matrices," *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2967–2979, Jun. 2010.

[79] C. Weidmann, "Oligoquantization in low-rate lossy source coding," Ph.D. dissertation, EPFL, Lausanne, Switzerland, Jul. 2000.

[80] C. Weidmann and M. Vetterli, "Rate distortion behavior of sparse sources," Dec. 2008, submitted to IEEE Trans. Inf. Thoery.

[81] Y. Wu and S. Verdú, "Renyi information dimension: Fundamental limits of almost lossless analog compression," *IEEE Trans. Inf. Theory*, vol. 56, no. 8, pp. 3721–3748, Aug. 2010.

[82] A. Wyner, "The wire-tap channel," *Bell System Technical Journal*, vol. 54, no. 8, pp. 1355–1387, 1975.

[83] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. of the Royal Stat. Soc. Series B*, vol. 68, no. 1, pp. 49–67, 2006.

[84] P. Zhao and B. Yu, "On model selection consistency of lasso," *J. of Machine Learning Research*, vol. 51, no. 10, pp. 2541–2563, Nov. 2006.