Sparse signal recovery using sparse random projections

Wei Wang

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2009-169
http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-169.html

December 15, 2009

Sparse signal recovery using sparse random projections

by

Wei Wang

B.S. (Rice University) 2000
M.S. (University of California, Berkeley) 2002

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

Doctor of Philosophy

in

Engineering – Electrical Engineering and Computer Sciences

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Kannan Ramchandran, Chair
Professor Martin J. Wainwright
Professor Bin Yu

Fall 2009

The dissertation of Wei Wang is approved.

Chair Date

Date

Date

University of California, Berkeley

Sparse signal recovery using sparse random projections

by

Wei Wang

Abstract

Sparse signal recovery using sparse random projections

by

Wei Wang

Doctor of Philosophy in Engineering – Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Kannan Ramchandran, Chair

The problem of estimating a high-dimensional signal based on an incomplete set of noisy observations has broad applications. In remote sensing, network traffic measurement, and computational biology, the observation process makes it difficult or costly to obtain sample sizes larger than the ambient signal dimension. Signal recovery is in general intractable when the dimension of the signal is much larger than the number of observations. However, efficient recovery methods have been developed by imposing a sparsity constraint on the signal. There are different ways to impose sparsity, which has given rise to a diverse set of problems in sparse approximation, subset selection in regression, and graphical model selection.

This thesis makes several contributions. First, we examine the role of sparsity in the measurement matrix, representing the linear observation process through which we sample the signal. We develop a fast algorithm for approximation of compressible signals based on sparse random projections, where the signal is assumed to be well-approximated by a sparse vector in an orthonormal transform. We propose a novel distributed algorithm based on sparse random projections that enables refinable approximation in large-scale sensor networks. Furthermore, we analyze the information-theoretic limits of the sparse recovery problem, and study the effect of using dense versus sparse measurement matrices. Our analysis reveals that there is a fundamental limit on how sparse we can make the measurements before the number of observations required for recovery increases significantly. Finally, we develop a general framework for deriving information-theoretic lower bounds for sparse recovery. We use these methods to obtain sharp characterizations of the fundamental limits of sparse signal recovery and sparse graphical model selection.

Professor Kannan Ramchandran
Dissertation Committee Chair

To my family.

Contents

List of Figures

List of Tables

Acknowledgements

I am extremely grateful to have had the opportunity to interact with many people during the course of my Ph.D. whose ideas, enthusiasm and intellectual energy have truly inspired me. First of all, I would like to thank my thesis advisor Kannan Ramchandran whose support and guidance throughout my time at Berkeley has helped me find my way. Kannan's excitement and enthusiasm for research is infectious, and his ability to quickly see the key intuition and insights underlying a problem has been a great source of inspiration. I would also like to thank my collaborator Prof. Martin Wainwright, whom I was fortunate enough to have had the opportunity to interact with during the last years of my Ph.D. Martin's tremendous intellectual energy and impressive ability to see through abstract mathematical problems are both amazing and inspiring to observe. I have greatly benefited from his guidance and encouragement during much of the work in this thesis.

In addition, I would also like to thank my collaborator Prof. Minos Garofalakis, who introduced me to the literature on sketching algorithms and dimensionality reduction. Minos was a valuable resource of knowledge on the literature and I am grateful to him for sharing his enthusiasm and encouragement. I have also benefited from helpful interactions with Prof. Bin Yu, who served as a member of my Ph.D. thesis committee and was also a co-advisor for my M.S. I am grateful to Bin for her sage advice and insights.

Throughout my time at Berkeley, I have also learned much from interactions with members of the BASiCS group and Wireless Foundations. For that I am deeply grateful, and I would like to thank all my friends at Berkeley for making my time

there a memorable one. In particular, I would like to thank Vinod Prabhakaran, Prasad Santhanam, and Pradeep Ravikumar for many helpful research discussions.

Finally, I would like to thank my family for their love and support throughout the years. I cannot aptly express my debt of gratitude to them.

Chapter 1

Introduction

A fundamental problem in high-dimensional statistics is to estimate data based on noisy observations, when the dimensionality of the data is much larger than the number of samples. This problem arises in a wide variety of applications, including remote sensing, network traffic measurement, and computational biology. In many settings, the data is not directly accessible and the observation process makes it difficult or costly to obtain sample sizes greater than the ambient data dimension. However, many classical procedures assume that the problem dimension p is fixed while the number of samples n grows, and are known to break down when p/n does not go to zero. Moreover, recovery is frequently intractable when $p \gg n$ unless some additional structure is imposed on the data or underlying model. Accordingly, a line of recent research has focused on developing efficient recovery methods for high-dimensional estimation by imposing sparsity.

Sparsity can be exhibited in various forms in different problem settings. In subset selection in regression [42], the regression vector may include a large number of irrelevant variables, and the goal is to select the sparse subset of variables that linearly influence the observations. In sparse approximation [43, 23] and signal denoising [16], the data may be well-approximated by a sparse vector of coefficients in an orthonormal basis or overcomplete dictionary. Similarly, in compressed sensing [14, 25], the problem of interest is to recover a sparse vector that satisfies a set of observed linear constraints, where the sparsity on the signal may be imposed in the signal domain or the transform domain. Finally, in graphical model selection [41], sparsity may be imposed on an underlying graph that determines the conditional independence properties of a Markov random field, and the goal is to correctly estimate the graph structure.

Sparsity is an abstract concept, but a powerful one with diverse applications. The common underlying phenomenon in all these problems is that although the ambient dimensionality of the problem is high, the data actually lies in some low dimensional subspace. This sparse structure can be exploited to obtain computationally efficient recovery methods. The development of methods to solve such sparse recovery problems has varied from field to field. We now provide a broad overview of these developments in several areas.

1.1 Research areas related to sparse recovery

There is a long history in signal processing of studying sparse representations of signals, from classical Fourier analysis to wavelets and overcomplete dictionaries [22, 40]. Such representations are used in transform-based coding and compression schemes for images and audio [29]. Since signal decomposition in overcomplete dictionaries is not unique, sparse approximation algorithms [43, 23] were developed to find the sparsest representation with respect to a given dictionary. In particular, matching pursuit [39] uses greedy iterative algorithms, while basis pursuit [16] formulates the problem as an ℓ_1-minimization which can be solved using linear programming. A substantial body of recent work in compressive sensing [14, 25, 27, 13, 56] has analyzed the behavior of ℓ_1-relaxations for sparse approximation and established conditions on the signal sparsity and choice of measurement matrix under which they succeed.

In statistics, a great deal of work has similarly focused on ℓ_1 and other convex relaxations for high-dimensional inference. In particular, subset selection [42] refers to the problem of recovering the sparsity pattern of an unknown vector based on noisy observations. For applications in computational biology and group testing, it is the underlying support set that is of primary interest. Considerable research has analyzed the behavior of ℓ_1-constrained quadratic programming [13, 26, 56], also known as the Lasso [55, 41, 64, 58], for sparsity recovery. In graphical model selection, the problem of interest is to recover the underlying graph structure based on observed samples from a Markov random field. These problems have applications in image analysis, natural language processing, and computational biology. When the graph is known to be sparse, various ℓ_1-regularized methods [11, 63, 21, 46] have been shown to yield consistent estimators for high-dimensional graph selection.

Finally, in computer science and applied math, related problems in dimensionality reduction have been studied from an algorithmic perspective. The Johnson-

Lindenstrauss (JL) lemma [36] states that a point set in high-dimensional Euclidean space can be embedded in a low-dimensional space while preserving pairwise distances between the points. Most known constructions for JL embeddings use random projections [35, 20, 1], which provide both efficient algorithms for constructing such embeddings and simpler and sharper proofs of the lemma. A line of work in sketching [6, 34] has used such geometry-preserving embeddings to develop fast probabilistic algorithms for approximating frequency moments, histograms, and the top-k wavelet coefficients of streaming data. Such algorithms are used to estimate large flows in network traffic measurement and monitor statistics in massive databases. Furthermore, recent results [8] have shown that matrices which satisfy the Johnson-Lindenstrauss lemma also satisfy the restricted isometry property needed for ℓ_1-recovery, thereby establishing a mathematical connection between Johnson-Lindenstrauss and compressed sensing.

The deep connections between these seemingly disparate problem settings and research areas underscore the power and broad applicability of sparse recovery models. In this context, this thesis focuses on three main problems: estimation of compressible signals (i.e. signals which are approximately sparse in an orthonormal transform), exact support recovery of sparse vectors, and model selection for sparse graphical models. In particular, we address both computationally tractable recovery methods, as well as information-theoretic bounds on the performance of any recovery method.

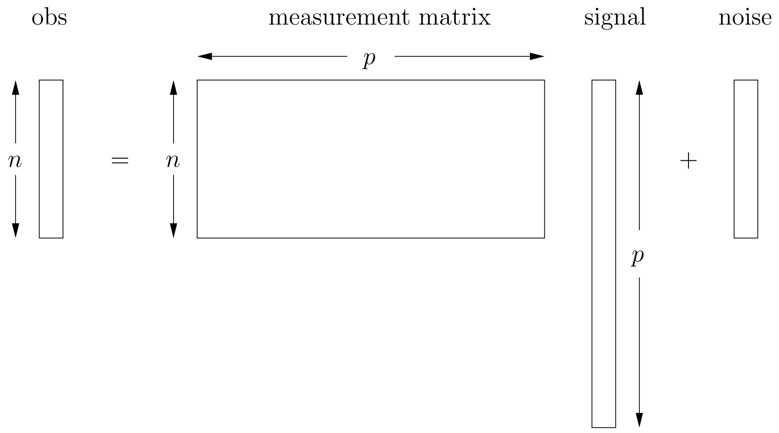


Figure 1.1. Sparse signal recovery refers to the problem of estimating an unknown signal based on n noisy observations, when the number of samples n is much less than the ambient signal dimension p.

1.2 Thesis overview

In this section, we provide an overview of the problems addressed in this thesis and discuss some of our contributions.

1.2.1 Sparse approximation in sensor networks

We first consider the problem of data approximation in large-scale distributed sensor networks. The traditional approach to data recovery in wireless sensor networks is to collect data from each sensor to one central server, and then process and compress the data centrally. The physical phenomena measured by sensor networks are typically smooth signals (e.g. temperature and humidity), and the sparse representation and compressibility of such classes of signals are well understood [40]. Consequently, the aggregate data vector containing all the sensor values can often be well-approximated by a sparse vector of coefficients in an orthonormal basis. In this setting, an interesting question is whether we can collect a set of samples from anywhere in the network, and recover an approximation of the data comparable to the best k-term approximation with respect to a given basis. Of course, this is possible when the sample size n is equal to the data dimension p; we are interested in regimes in which $n \ll p$.

As we discussed in the previous section, there is an extensive literature on recovery methods for classes of signals with some underlying sparse structure. However, when the data is distributed across many nodes in a large-scale network, new challenges arise. In particular, communication is a dominant cost in wireless networks when nodes are power-limited. Most known constructions in compressed sensing and related areas use dense measurement matrices, for example containing entries drawn from a standard Gaussian or Bernoulli$\{+1, -1\}$ distribution. Computing a single random projection of the sensor data using such dense measurements would require accessing the values at all the sensor nodes. The key idea we exploit in Chapter 2 is that sparse measurement matrices can greatly reduce the communication, storage and computational cost associated with sparse recovery.

Sparse recovery using sparse random projections

In Chapter 2, we show that a simple sketching decoder can recover compressible signals from sparse random projections in the presence of noise. We consider sparse measurement matrices with a γ-fraction of non-zero entries per row. Our results apply to general random ensembles of matrices that satisfy certain moment conditions, which includes γ-sparsified Bernoulli and γ-sparsified Gaussian matrices.

The sketching decoder is computationally cheaper than the ℓ_1-recovery methods used in compressed sensing, at the cost of requiring more measurements to recover at a given fidelity. In sensor network scenarios, this tradeoff may be desirable to allow resource-limited agents or sensors to recover coarse approximations of the network data. Our results establish conditions on the signal under which it is possible to have very sparse measurement matrices (e.g. with a constant number of nonzeros per row) without affecting the sampling efficiency of sketching. More generally, we characterize the tradeoff between measurement sparsity and sampling efficiency.

It is worth noting that compressed sensing results on ℓ_1 recovery of compressible signals (e.g. [13]) rely on certain mutual incoherence properties between the measurement matrix and the sparsifying basis. For example, measurement matrices with entries drawn from the standard Gaussian distribution are rotation invariant when multiplied by any orthonormal matrix. In this case, the problem of recovering compressible signals can be formulated equivalently as a problem of recovering sparse vectors. However, sparse measurement matrices are not generally spherically symmetric. To the best of our knowledge, our approach is the only method that uses sparse random projections to approximate compressible signals, when the measurements are made directly on the signal itself.

Distributed approximation for sensor networks

As we discussed in the previous section, we can recover an approximation of the data based on sparse random projections. A natural question, then, is whether the sensors can pre-process the data in a distributed manner to produce these random projections. The key ideas underlying Chapter 2 are that, (1) sparse random projections can be used to recover compressible data, and (2) sparsity can be exploited to reduce the amount of communication needed to pre-process the data in the network.

We propose a distributed algorithm based on sparse random projections, which has the useful properties of universality, refinability, and robustness. More specifically, the sensors need no knowledge of the data model, meaning the basis in which the data is approximately sparse and the value of the signal sparsity parameter k. Only the decoder needs to know this information, and thus the decoder can choose the number of sensors to query according to the desired quality of the approximation. The error of the approximation depends only on the number of measurements collected, and not on which sensors are queried.

1.2.2 Information-theoretic bounds for sparse recovery

Sparsity recovery refers to the problem of recovering the sparsity pattern or support set of an unknown vector based on noisy linear observations. This problem – also known as support recovery or variable selection – arises in subset selection in regression [42], graphical model selection [41], signal denoising [16] and compressed sensing [14, 25]. There is a large body of work focused on developing computationally efficient methods to solve this problem (e.g., [16, 14, 25, 55, 56, 58]); one prominent approach is the use of ℓ_1-relaxation methods which can be solved with convex optimization. The discovery of polynomial-time algorithms to solve this vastly underdetermined inverse problem has generated considerable excitement in the literature. However, the computational complexity can still be quite high: in the noiseless case,

ℓ_1-recovery methods using linear programming have complexity $O(p^3)$ in the signal dimension p. One way to reduce this complexity is to use sparse measurement matrices.

Most known constructions use dense measurement matrices, for example matrices with entries drawn from a standard Gaussian distribution or a Bernoulli$\{+1, -1\}$ distribution. These matrices consist entirely of nonzero entries with probability one, and as such are expensive to store and process. In contrast, sparse measurement matrices can directly reduce encoding complexity and storage costs, and can also lead to fast decoding algorithms by exploiting problem structure. Accordingly, a line of recent work has studied the use of measurement sparsity to reduce complexity [53, 60, 61, 44, 10]. On the other hand, measurement sparsity can potentially hurt performance by requiring more measurements to recover the signal. Intuitively, if both the signal and the measurement matrix are very sparse, then the random measurements will rarely align with the nonzero locations of the signal.

In this context, two key questions arise about the relationship between the sparsity of the measurement matrix and the number of measurements needed to recover the signal. First, how sparse can we make the measurement matrix without affecting the sampling efficiency? And second, when we increase the sparsity beyond this limit, what is the tradeoff between measurement sparsity and sampling efficiency?

Fundamental limits of sparsity

There is a substantial literature on computationally tractable methods for estimating high-dimensional sparse signals. Of complementary interest are the information-theoretic limits of the problem, which apply to the performance of any recovery method regardless of its computational complexity. Such analysis has two purposes: first, to demonstrate where known polynomial-time methods achieve the information-theoretic bounds, and second, to reveal situations in which current methods are sub-optimal.

Chapter 3 addresses the effect of the choice of measurement matrix on the information-theoretic limits of sparse signal recovery, in particular the effect of using dense versus sparse measurement matrices. Our analysis yields sharp characterizations of when the optimal decoder can recover for a general class of dense measurement matrices (including non-Gaussian ensembles). In addition, our results show the effect of measurement sparsity, and reveal that there is a critical threshold beyond which sparsity significantly increases the number of observations necessary for recovery. Surprisingly, this limit is fundamental, and not an artifact of a particular recovery method.

1.2.3 Graphical model selection bounds

Markov random fields or undirected graphical models are families of multivariate probability distributions whose factorization and conditional independence properties are characterized by the structure of an underlying graph. For example, in statistical image processing, the dependencies among the gray-scale values of the image pixels can be specified by a graphical model. For tasks such as image denoising and feature extraction, it is the structure of the underlying graph that is of interest. Graphical model selection refers to the problem of estimating the graph structure based on observed samples from an unknown Markov random field. For Gaussian Markov random fields, this problem is equivalent to estimating the sparsity pattern of the inverse covariance matrix. A line of recent work [63, 32, 21, 49, 46] has shown that ℓ_1-regularization methods provide consistent estimators for graphical model selection when the underlying graph is known to be sparse.

Chapter 4 focuses on the information-theoretic limits of this problem, which bound the performance of any algorithm regardless of its computational complexity. More specifically, our analysis yields a set of necessary conditions for consistent graphical model selection over Gaussian Markov random fields. Compared to previously known sufficient conditions using ℓ_1-penalized maximum likelihood [46], we obtain sharp characterizations in certain regimes of interest, while revealing a gap in other regimes. Furthermore, our results establish necessary conditions for estimation of the inverse covariance matrix with error measured in the elementwise ℓ_∞-norm, which implies similar conditions in other recovery norms as well.

At a high level, our general approach is to apply Fano's inequality [19] to restricted ensembles of graphical models, in which we view the observation process as a communication channel. The problem of establishing necessary conditions for recovery is then reduced to obtaining bounds on the mutual information between the observations and a random model index. From this perspective, our approach is related to a line of work on non-parametric estimation in statistics (e.g. [62]). In contrast to that literature, the spaces of possible codewords in our setting are not function spaces but instead classes of graphical models. We take a similar approach when deriving necessary conditions for subset selection in Chapter 3. Our analysis techniques may be more generally applicable to other recovery problems as well.

Chapter 2

Distributed approximation using sparse random projections

2.1 Introduction

Suppose a wireless sensor network measures data which is compressible in an orthonormal transform, so that p data values can be well-approximated using only $k \ll p$ transform coefficients. In this setting, an interesting question is whether we can pre-process the data in the network so that only k values need to be collected to recover the data with an acceptable approximation error. However, it is difficult to reliably compute a deterministic transform in a distributed manner over large-scale wireless networks. Furthermore, even if the data is sparse in the identity basis, one still must locate the largest non-zero coefficients in the network to recover the best k-term approximation.

There is a rich literature on the use of random projections to approximate functions of data. In particular, a large body of work in compressed sensing (e.g., [**?** 25]) and related areas has analyzed conditions under which sparse or compressible signals can be recovered from random linear projections of the data. Similarly, in the AMS sketching literature (e.g., [6, 34, 17]), random projections are used to approximate wavelet representations of streaming data. Moreover, random projections are also used in variations of the Johnson-Lindenstrauss (JL) lemma [36, 1, 38, 4] to perform geometry-preserving embeddings for dimensionality reduction. However, most of the known results in these fields rely on the use of *dense measurement matrices*, for example with entries drawn from a standard Gaussian distribution. Computing such matrices in a distributed setting would require $\Omega(p^2)$ communications, equivalent to flooding the network with data.

The focus of this chapter is on the use of *sparse measurement matrices* for approximation of compressible signals. First, we show that $O(k^2 \log p)$ sparse random projections are sufficient to recover a data approximation which is comparable to the optimal k-term approximation. Our analysis establishes conditions under which the average number of non-zeros in each random projection vector can be $O(1)$. More generally, we characterize the trade-off between the sparsity of the random projections and the number of random projections needed for recovery. Second, we present a distributed algorithm based on sparse random projections, which guarantees the recovery of a near-optimal approximation by querying any $O(k^2 \log p)$ sensors. Our algorithm effectively acts as an erasure code over real numbers, generating p sparse random projection coefficients out of which any subset of $O(k^2 \log p)$ is sufficient to decode. The communication cost can be reduced to a constant $O(1)$ number of packets per sensor, routed to randomly selected nodes in the network. There is a corresponding trade-off between the pre-processing communication cost and the number of sensors that need to be queried to recover an approximation at a given fidelity.

Our distributed algorithm has the interesting property that the decoder can choose how much or how little to query, depending on the desired approximation error. The sensors do not need any knowledge of the data model or the sparsifying transform, including the value of the sparsity parameter k. The decoder can choose k according to the desired quality of the approximation, and collect a sufficient number of random measurements from anywhere in the network. The error of the approximation depends only on the number of measurements collected, and not on which sensors are queried. Thus, our distributed algorithm enables robust refinable approximation.

The remainder of the chapter will be organized as follows. In Section 2.2, we define the problem setup and modeling assumptions, and discuss connections to previous work. In section 2.3, we state our main results on sparse approximation using sparse random projections. In Section 2.4, we then describe our distributed algorithm based on sparse random projections. Section 2.5 contains the analysis of the recovery method, while Section 2.6 contains some comparisons and numerical experiments.

2.2 Sparse approximation

We consider a wireless network of p sensors, each of which measures a real data value β_i. Suppose the aggregate data $\beta \in \mathbb{R}^p$ is compressible, so that it can be well-approximated using $k \ll p$ coefficients of some orthonormal transform. For simplicity, we assume that each sensor computes and stores one random projection. We want to be able to query any n sensors and recover an approximation of the p data values, with reconstruction error comparable to the best k-term approximation.

2.2.1 Compressible data

There is a long history in signal processing of studying sparse representations for classes of signals, including smooth signals with bounded derivatives and bounded variation signals [57, 40]. Sensor networks measuring a smooth temperature field, for example, may efficiently represent the data using only a few large transform coefficients, which record useful structure such as average temperature and sharp temperature changes. The remaining small transform coefficients may be discarded without much loss in the total signal energy.

We consider a real data vector $\beta \in \mathbb{R}^p$, and fix an orthonormal transform $\Psi \in \mathbb{R}^{p \times p}$ consisting of a set of orthonormal basis vectors $\left\{\psi^{(1)}, \ldots, \psi^{(p)}\right\}$. The transform Ψ can be, for example, a wavelet or a Fourier transform. The transform coefficients $\theta = \left(\beta^T \psi^{(1)}, \ldots, \beta^T \psi^{(p)}\right)^T$ of the data can be ordered in magnitude, so that $|\theta|_{(1)} \geq |\theta|_{(2)} \geq \cdots \geq |\theta|_{(p)}$. The best k-term approximation keeps the largest k transform coefficients and sets the remaining coefficients to zero. The corresponding approximation error is $\left\|\beta - \widehat{\beta}\right\|_2^2 = \left\|\theta - \widehat{\theta}\right\|_2^2 = \sum_{m=k+1}^{p} |\theta|_{(m)}^2$.

We now specify the model of compressible data as defined in the compressed sensing literature [13, 25]. We say that the data is compressible if the magnitude of its transform coefficients decay like a power law. That is, the mth largest transform coefficient satisfies

$$|\theta|_{(m)} \leq C_r \, m^{-r} \tag{2.1}$$

for each $1 \leq m \leq p$, where C_r is a constant, and $r \geq 1$. Note that r controls the compressibility (or rate of decay) of the transform coefficients. The approximation error obtained by taking the k largest transform coefficients and setting the remaining coefficients to zero, is then

$$\left\|\beta - \widehat{\beta}_k \right\|_2 = \left\|\theta - \widehat{\theta}_k \right\|_2 \leq C_r' \, k^{-r+1/2}$$

where C_r' is a constant that depends only on r.

2.2.2 Noisy observation setting

We study the problem of recovering an approximation of compressible signals based on noisy linear measurements. In particular, suppose we are given n noisy observations of the form

$$Y = X\beta + W \in \mathbb{R}^n, \tag{2.2}$$

where $W \in \mathbb{R}^n$ is the noise vector, and $X \in \mathbb{R}^{n \times p}$ is the measurement matrix. We consider the problem of estimating the signal β based on Y, where the quality of the approximation is measured with respect to an ℓ_2-norm error metric, $\|\beta - \widehat{\beta}\|_2$. We assume that the noise vector W has independent entries drawn from any distribution with mean zero and variance σ^2 (which includes for example the Gaussian distribution

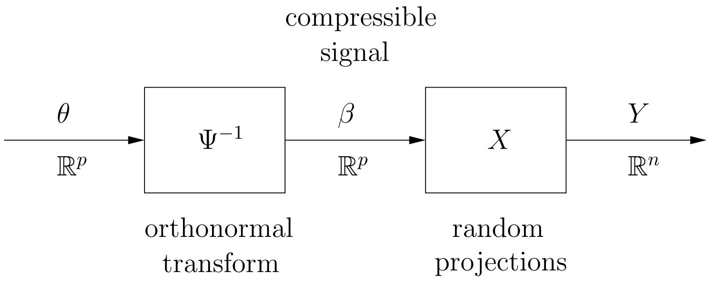


Figure 2.1. The compressible data model assumes that the largest k transform coefficients of θ in magnitude captures most of the signal energy.

$N(0, \sigma^2 I))$. Our results apply to general ensembles of measurement matrices which satisfy some moment conditions (defined in Section 2.3), which include γ-sparsified Bernoulli and γ-sparsified Gaussian matrices.

2.2.3 Random projections

Recent results in compressed sensing [15, 25] and related areas have shown that random projections of the data can be used to recover an approximation with error comparable to the best approximation using the k largest transform coefficients. More concretely, consider the random projection matrix $X \in \mathbb{R}^{k \times p}$ containing i.i.d. entries

$$X_{ij} \;=\; \begin{cases} +1 & \text{w.p. } \frac{1}{2} \\ -1 & \text{w.p. } \frac{1}{2} \end{cases} \tag{2.3}$$

Given k random projections $\frac{1}{\sqrt{p}} X\beta \in \mathbb{R}^k$, we can produce an approximation $\widehat{\beta}$ of the data β satisfying

$$\left\|\beta - \widehat{\beta}\,\right\|_2 \;\leq\; \alpha_r \left(\frac{k}{\log p}\right)^{-r+1/2},$$

with high probability, where the constant α_r depends on r. Compressed sensing decoding is performed by solving a linear program, which has $O(p^3)$ computational complexity. More recent work [53, 61, 10] in compressed sensing has examined the use of sparse measurement matrices to reduce decoding complexity. However, these results assume that the signal itself is sparse, and are not applicable to the compressible signal model considered here.

Random projections have also been used to recover approximate wavelet representations of streaming data in the AMS sketching literature (e.g. [6, 34, 17]). In sketching, the random projection matrix has entries X_{ij} defined as in (2.3), except only four-wise independence is required within each row. This relaxation allows the matrix to be generated pseudo-randomly and stored in small space. The decoding process estimates the largest k wavelet coefficients using random projections of the data and the wavelet bases. The sketching decoder requires $O(k^2 \log p)$ random projections to produce an approximation with error comparable to the best-k wavelet coefficients. However the decoding computational complexity is reduced to $O(np)$, where n is the number of random projections used. For some applications, it would be useful for sensors or other low-powered collectors to be able to decode a coarse approximation of the data cheaply and quickly. Meanwhile, collectors with greater resources can obtain more random projections and reconstruct a better approximation.

Finally, random projections have also been used for geometry-preserving embeddings in dimensionality reduction. The Johnson-Lindenstrauss (JL) lemma [36, 1, 38, 4] states that any set of $p \geq d$ points can be mapped from $\mathbb{R}^d$ to $\mathbb{R}^k$ while preserving all pairwise distances within a factor of $(1 \pm \epsilon)$, where $k = O\left(\frac{\log p}{\epsilon^2}\right)$. A line of work [1, 38, 4] has explored the use of sparsity for efficient JL embedding.

2.2.4 Distributed data processing

Most of the known results in both compressed sensing and sketching use dense measurement matrices. The key idea of this chapter is that *sparse* random projections can reduce computational complexity, and in our distributed problem setting, minimize communication cost. Sparsity in the random measurement matrix may also be exploited to reduce decoding complexity.

Distributed compressed sensing schemes have been proposed in [9, 45, 7]. However, the problem formulations in these earlier works are very different from our set-up. In particular, the papers [9, 7] consider the scenario in which all sensors communicate directly to a central fusion center, without any in-network communication. The paper [9] defines a joint sparsity model on the data, and uses knowledge of this correlation structure to reduce communications from the sensors to the fusion center. The work in [7] uses uncoded coherent transmissions through an AWGN multiple access channel to simultaneously communicate and compute random projections from the sensors to the fusion center. Finally, the paper [45] poses the scenario where ultimately every sensor has a full approximation of the network data, by using gossip algorithms to compute each random projection.

2.3 Sparse random projections

For real data vectors $\beta \in \mathbb{R}^p$, our goal is to find the minimum number of observations n that is sufficient to recover an approximation of β with error comparable to the best k-term approximation. We consider γ-sparsified measurement matrices with

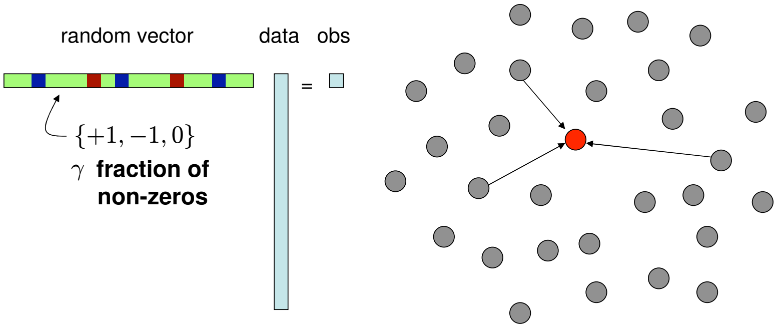


Figure 2.2. Sparsity of the random projection matrix leads to a more efficient distributed algorithm with fewer communications.

entries that are set to zero with probability $1 - \gamma$, so that on average there are γp non-zeros per row. Our results depend on the maximum value of the signal, defined as

$$\omega \quad := \quad \frac{\|\beta\|_\infty}{\|\beta\|_2} \tag{2.4}$$

so that it is invariant to rescaling of the data. This parameter bounds the ratio of the largest component of the data to the ℓ_2-norm, and guarantees that the total energy of the signal is not concentrated in a few elements. Intuitively, sparse random projections will not work well when the data is also very sparse. Interestingly, we can relate ω to the compressibility of the signal, as defined in (2.1).

2.3.1 Sufficient conditions for recovery

Our results apply to general ensembles of measurement matrices whose entries satisfy some moment conditions (defined in (2.6)), which allow for a variety of sparse matrices, including the γ-sparsified Gaussian ensemble or the γ-sparsified Bernoulli ensemble. In particular, consider the sparse random projection matrix $X \in \mathbb{R}^{n\times p}$ containing i.i.d. entries drawn accordingly to

$$X_{ij} \quad = \quad \frac{1}{\sqrt{\gamma}} \begin{cases} +1 & \text{w.p. } \frac{\gamma}{2} \\ 0 & \text{w.p. } 1-\gamma \\ -1 & \text{w.p. } \frac{\gamma}{2} \end{cases} . \tag{2.5}$$

Note that the parameter γ controls the sparsity of the random projections. If $\gamma = 1$, then the measurement matrix is dense. On the other hand, if $\gamma = \Theta(\frac{1}{p})$, then the average number of non-zeros in each row of the measurement matrix is $\Theta(1)$. Furthermore, our results hold more generally than the setting in which the entries of X are assumed to be i.i.d.. More specifically, we only need to assume that the entries within each row are four-wise independent, while the entries across different rows are fully independent. This limited independence assumption allows each random projection vector to be pseudo-randomly generated and stored in small space [6]. Note however, that we can directly exploit the sparsity of the measurement matrix to reduce storage costs, and hence we state our results using the i.i.d. assumption.

We first show a variant of the Johnson-Lindenstrauss embedding result for sparse measurement matrices, namely, that sparse random projections preserve inner products within an ϵ-interval. To do this, we show that pairwise inner products between a set of points are preserved in expectation under sparse random projections, and that they are concentrated about the mean using a standard Chebyshev-Chernoff argument. Lemma 3 states that an estimate of the inner product between two vectors, using only the random projections of those vectors, are correct in expectation and have bounded variance.

Lemma 1. *Consider a random matrix $X \in \mathbb{R}^{n\times p}$ with entries X_{ij} satisfying the following conditions:*

$$(a)\ X_{ij} \text{ are i.i.d.}, \qquad (b)\ \mathbb{E}[X_{ij}] = 0, \qquad (c)\ \mathbb{E}[X_{ij}^2] = 1, \qquad (d)\ \mathbb{E}[X_{ij}^4] = \tfrac{1}{\gamma}. \tag{2.6}$$

Suppose a random vector $W \in \mathbb{R}^n$ has independent entries satisfying $\mathbb{E}[W_i] = 0$ and $\mathbb{E}[W_i^2] = \sigma^2$. For any two vectors $\beta, \psi \in \mathbb{R}^p$, define the random projections of these vectors as $Y = X\beta + W$, $Z = X\psi \in \mathbb{R}^n$. Then the mean and variance of Z^TY/n are

$$\mathbb{E}\left[\frac{1}{n} Z^T Y\right] = \psi^T\beta \tag{2.7}$$

$$\mathrm{var}\left(\frac{1}{n} Z^T Y\right) = \frac{1}{n}\left\{(\psi^T\beta)^2 + \left(\|\beta\|_2^2 + \sigma^2\right)\|\psi\|_2^2 + \left(\frac{1}{\gamma} - 3\right)\sum_{j=1}^{p}\psi_j^2\beta_j^2\right\}. \tag{2.8}$$

Note that Lemma 3 and all subsequent results require only the sufficient conditions (2.6) on the random projection matrix. The sparse random projection matrix X defined in equation (2.5) satisfies the conditions (2.6), with the fourth moment $\mathbb{E}[X_{ij}^4]$ corresponding to the sparsity parameter of the matrix $1/\gamma$. It is interesting to note that these conditions also hold for other random projection matrices. For example, the non-sparse matrix containing Gaussian i.i.d. entries $X_{ij} \sim N(0,1)$ satisfies (2.6) with $\mathbb{E}[X_{ij}^4] = 3$. Similarly, $\mathbb{E}[X_{ij}^4] = 1$ for the non-sparse random projection matrix containing i.i.d. entries $X_{ij} = \pm 1$ as defined in equation (2.3).

Theorem 1 now states that sparse random projections of the data vector and any set of p vectors can produce estimates of their inner products to within a small error. Thus, sparse random projections can produce accurate estimates for the transform coefficients of the data, which are inner products between the data and the set of orthonormal bases.

Theorem 1 (Sparse-Noisy JL). *For any real data vector $\beta \in \mathbb{R}^p$, define the maximum value of β as*

$$\omega \;:=\; \frac{\|\beta\|_\infty}{\|\beta\|_2}. \tag{2.9}$$

In addition, let Ψ be a set of orthonormal basis vectors $\{\psi^{(1)}, \ldots, \psi^{(p)}\} \in \mathbb{R}^p$. Suppose a sparse random matrix $X \in \mathbb{R}^{n\times p}$ satisfies the conditions (2.6) *with sparsity parameter γ. For any ϵ and $\delta > 0$, if*

$$n \;\geq\; \frac{32(1+\delta)}{\epsilon^2}\left(2 + \frac{\sigma^2}{\|\beta\|_2^2} + \frac{\omega^2}{\gamma}\right)\log p,$$

then given the noisy projections $Y = X\beta + W$, one can produce estimates $\widehat{\theta}_m$ for $\theta_m = \beta^T\psi^{(m)}$ satisfying

$$\left|\widehat{\theta}_m - \theta_m\right| \;\leq\; \epsilon\,\|\beta\|_2$$

with probability greater than $1 - 1/p^\delta$ uniformly over all $m = 1, \ldots, p$.

Using these low-distortion embeddings, Theorem 2 now states our main result, that sparse random projections can produce a data approximation with error comparable to the best k-term approximation with high probabiliy.

Theorem 2. *Consider a real vector $\beta \in \mathbb{R}^p$ satisfying condition* (2.9)*, and a sparse random matrix $X \in \mathbb{R}^{n\times p}$ satisfying conditions* (2.6)*. Suppose that the best k-term approximation in an orthonormal transform Ψ has approximation error $\|\beta - \widehat{\beta}_k\|_2^2 \leq \eta\|\beta\|_2^2$. For any ϵ and $\delta > 0$, if the number of observations satisfies*

$$n \;\geq\; \frac{C(1+\delta)}{\epsilon^2}\left(2 + \frac{\sigma^2}{\|\beta\|_2^2} + \frac{\omega^2}{\gamma}\right)k^2\log p \tag{2.10}$$

for some constant C, then given the noisy observations $Y = X\beta + W$, one can produce an approximation $\widehat{\beta}$ satisfying

$$\|\beta - \widehat{\beta}\|_2^2 \;\leq\; (\epsilon + \eta)\|\beta\|_2^2$$

with probability at least $1 - 1/p^{\delta}$.

The signal-to-noise ratio for this noisy observation model is

$$SNR \;:=\; \frac{\mathbb{E}[\|X\beta\|_2^2]}{\mathbb{E}[\|W\|_2^2]} \;=\; \frac{\|\beta\|_2^2}{\sigma^2}. \tag{2.11}$$

If $\sigma^2 = 0$, then the problem reverts to the noiseless setting and the number of observations in (2.10) reduces to $n \geq \Omega\left(\left(2 + \frac{\omega^2}{\gamma}\right) k^2 \log p\right)$. If $\sigma^2 > 0$, then Theorem 2 implies that the number of observations that are sufficient for recovery behaves the same as in the noiseless case (in scaling terms), provided that the SNR does not go to zero.

The effect of measurement sparsity is the ω/γ term in (2.10). A straightforward calculation shows that the maximum value of the signal defined in (2.9) is bounded between

$$\frac{1}{p} \;\leq\; \omega^2 \;\leq\; 1. \tag{2.12}$$

In one extreme, $\omega = 1$ if and only if β has exactly one non-zero, and the remaining components are all equal to zero. In this case, if the measurement matrix is also very sparse (e.g. $\gamma \to 0$), then the non-zeros in the measurement matrix will rarely align with the non-zeros of the signal, making recovery very difficult. In the other extreme, $\omega = \frac{1}{p}$ if and only if β is constant (i.e. $\beta_i = c$ for all i). In this case, the

Compressibility	Measurement sparsity γ	Number of observations n
$r = 1$ $\omega = O\big(\frac{\log p}{\sqrt{p}}\big)$	$\gamma = 1$	$n = \Omega(k^2 \log p)$
	$\gamma = \frac{\log^2 p}{p}$	$n = \Omega(k^2 \log p)$
	$\gamma = \frac{\log p}{p}$	$n = \Omega(k^2 \log^2 p)$
	$\gamma = \frac{1}{p}$	$n = \Omega(k^2 \log^3 p)$
$r > 1$ $\omega = O\big(\frac{1}{\sqrt{p}}\big)$	$\gamma = \frac{1}{p}$	$n = \Omega(k^2 \log p)$

Table 2.1. The sufficient condition in Theorem 2 is shown under various scalings of the compressibility r of the signal and the measurement sparsity γ. The conditions shown in this table assume that the signal-to-noise ratio $SNR = \frac{\|\beta\|_2^2}{\sigma^2}$ does not go to zero.

measurement matrix can be made quite sparse (e.g. $\gamma = \frac{1}{p}$) without affecting the sampling efficiency. In general, ω is bounded between these two extremes. As the Lemma 2 shows, we can relate the maximum value ω in (2.9) to the compressibility of the data as defined in (2.1).

Lemma 2. *Suppose the vector β is compressible in the discrete Fourier transform as in* (2.1) *with compressibility parameter r.*

(a) If $r = 1$, then

$$\omega \;=\; O\left(\frac{\log p}{\sqrt{p}}\right). \tag{2.13a}$$

(b) If $r > 1$, then

$$\omega \;=\; O\left(\frac{1}{\sqrt{p}}\right). \tag{2.13b}$$

Sparsity in the measurement matrix produces an extra factor of $\frac{\omega^2}{\gamma}$ in the number of random projections that are sufficient for recovery. Consequently, there is an interesting trade-off between the number of random projections n, the average number of non-zeros γp in the random projections, and the peak-to-total energy ratio (or compressibility) of the data ω. For data compressible in the discrete Fourier transform (as in (2.1)) with $r = 1$, if the sparsity is $\gamma p = \log^2 p$, then $\frac{\omega^2}{\gamma} = O(1)$. In this case, there is no hit in the number of sparse random projections needed for approximation. If the sparsity is $\gamma p = \log p$, there is a hit of $\frac{\omega^2}{\gamma} = O(\log p)$ in the number of sparse random projections. If $\gamma p = 1$, then the hit in the number of projections is $\frac{\omega^2}{\gamma} = O(\log^2 p)$. For more compressible data with $r > 1$, if $\gamma p = 1$, then the hit in the number of sparse random projections is $\frac{\omega^2}{\gamma} = O(1)$.

We shall see in Section 2.4 that this trade-off, between the sparsity of the random projections and the number of projections, will have a corresponding trade-off in pre-processing communication cost and querying latency.

2.4 Distributed algorithms for sensor networks

We now describe an algorithm by which the p sensors of a wireless network each measure a data value β_i, and each computes and stores one sparse random projection of the aggregate data β. Consider an $p \times p$ sparse random matrix X with entries as defined in (2.5). For concreteness, let the probability of a non-zero entry be $\gamma = \frac{1}{p}$. Each sensor will compute and store the inner product $\sum_{j=1}^{p} X_{ij}\beta_j$ between the aggregate data β and one row of X. We think of this as generating a bipartite graph between the p data nodes and the p encoding nodes (see Figure 2.3).

2.4.1 Push-based algorithm

When the entries of X are independent and identically distributed, they can be generated at different sensor locations without any coordination between the sensors. To compute one random projection coefficient, each sensor j locally generates a random variable X_{ij}. If that random variable is zero the sensor does nothing, and if it's non-zero the sensor sends the product of X_{ij} with its own data β_j to one receiver sensor i. The receiver simply stores the sum of everything it receives, which is equal to the random projection coefficient $\sum_{j=1}^{p} X_{ij}\beta_j$. This process is repeated until every sensor has stored a random projection coefficient. Thus, computation of the sparse random projections can be achieved in a decentralized manner with the following *push-based* algorithm.

Distributed Algorithm I:

- Each data node j generates a set of independent random variables $\{X_{1j}, \ldots, X_{pj}\}$. For each i, if $X_{ij} \neq 0$, then data node j sends to encoding node i the value $X_{ij}\beta_j$. Repeat for all $1 \leq j \leq p$.
- Each encoding node i computes and stores the sum of the values it receives, which is equal to $\sum_{j=1}^{p} X_{ij}\beta_j$. Repeat for all $1 \leq i \leq p$.

Since the probability that $X_{ij} \neq 0$ is $\gamma = \frac{1}{p}$, each sensor independently and randomly sends its data to on average $O(1)$ sensors. Assuming that SNR does not go to zero, the decoder can query *any* $n = O\big(\big(1 + \frac{\omega^2}{\gamma}\big)\, k^2 \log p\big)$ sensors in the network and obtain a noisy version of $X_{n\times p}\beta$, where $X_{n\times p}$ is the matrix containing n

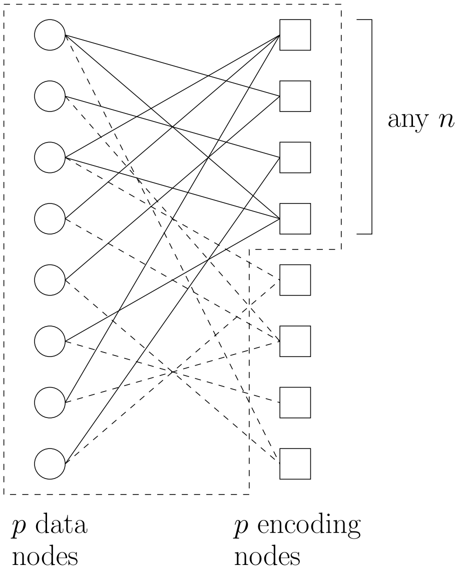


Figure 2.3. Every sensor stores a sparse random projection, so that a data approximation can be reconstructed by collecting coefficients from any k out of p sensors.

rows of $X \in \mathbb{R}^{p \times p}$. By Theorem 2, the decoder can then use the noisy observations $Y = X_{n \times p}\beta + W$, the measurement matrix $X_{n \times p}$, and orthonormal basis Ψ to recover a near-optimal approximation of the data β. The decoding algorithm proceeds as described in the proofs of Theorems 1 and 2.

2.4.2 Pull-based algorithm

We present an alternate, *pull-based*, distributed algorithm, which takes greater advantage of the limited independence of the sparse random projections. Each sensor i locally generates a set of four-wise independent random variables, corresponding to one row of the sparse random projection matrix. If a random variable X_{ij} is non-zero, sensor i sends a request for data to the associated data node j. Sensor j then sends its data β_j back to sensor i, who uses all the data thus collected to compute its random projection coefficient. Therefore, different sensors still act with complete independence.

Distributed Algorithm II:

- Each encoding node i generates a set of four-wise independent random variables $\{X_{i1}, \ldots, X_{ip}\}$. For each j, if $X_{ij} \neq 0$, then encoding node i sends a request for data to node j.
- If data node j receives a request for data from encoding node i, node j sends the value β_j to node i.
- Encoding node i computes and stores $\sum_{j=1}^{p} X_{ij}\beta_j$ using the values it receives. Repeat for all $1 \leq i \leq p$.

Since the average number of non-zeros per row of the sparse random projection matrix X is $\gamma p = 1$, the expected communication cost is still $O(1)$ packets per sensor, routed to random nodes. Algorithm II has twice the communication cost of Algorithm I, but the four-wise independence in Algorithm II allows the random projections to be generated pseudo-randomly. This further decreases the querying overhead cost for the collector seeking to reconstruct an approximation.

Both algorithms we described above perform a completely decentralized computation of p sparse random projections of the p distributed data values. In the end, collecting any subset of $O\big(\big(1+\frac{\omega^2}{\gamma}\big)\, k^2 \log p\big)$ sparse random projections will guarantee near-optimal signal recovery, as long as the SNR does not go to zero. Thus, our algorithms enables ubiquitous access to a compressed approximation of the data in a sensor network.

2.4.3 Trading off communication and query latency

In Section 2.3, we described the trade-off between the sparsity of the random projection matrix and the number of random projections needed for the desired approximation error. By Theorem 2, when the probability of a non-zero entry in the projection matrix is γ, the number of projections is $O\big(\big(1+\frac{\omega^2}{\gamma}\big)\, k^2 \log p\big)$ for non-vanishing SNR. In our distributed algorithms, the average number of packets transmitted per sensor is $O(\gamma p)$, while the number of sensors that need to be queried to recover an approximation is $O\big(\big(1 + \frac{\omega^2}{\gamma}\big)\, k^2 \log p\big)$. The average computation cost per sensor is also $O(\gamma p)$. Therefore, there is a trade-off between the amount of work performed by the sensors to pre-process the data in the network, and the number of sensors the decoder needs to query. Increasing the sparsity of the random projections decreases the pre-processing communication, but potentially increases the latency to recover a data approximation.

2.5 Analysis of sketching decoder

The intuition for our analysis is that sparse random projections preserve inner products within a small error, and hence we can use random projections of the data and the orthonormal bases to estimate the transform coefficients. Consequently, we can estimate all the transform coefficients to within a small error given only the sparse random projections of the data. However, we need to bound the sum squared error of our approximation over all the transform coefficients. If the data is compressible, and k of the transform coefficients are large and the others are close to zero, then we only need to accurately estimate k coefficients. The remaining small transform coefficients can be approximated as zero, incurring the same error as the best k-term approximation.

2.5.1 Moments of randomly projected vectors

We first derive the mean and variance of the randomly projected vectors given in Lemma 3. Let $X \in \mathbb{R}^{n \times p}$ be a random matrix satisfying the conditions in (2.6), and let $W \in \mathbb{R}^n$ have independent entries drawn from any distribution with mean zero and variance σ^2. We define the random variables

$$\begin{aligned} u_i &:= \left(\sum_{j=1}^{p} X_{ij}\psi_j\right)\left(\sum_{j=1}^{p} X_{ij}\beta_j\right) \\ v_i &:= \left(\sum_{j=1}^{p} X_{ij}\psi_j\right) W_i, \end{aligned}$$

so that we can express the inner products between the randomly projected vectors in the low-dimensional space as

$$\begin{aligned} Z^T Y &= \Psi^T X^T X \beta + \Psi^T X^T W \\ &= \sum_{i=1}^{n} u_i + \sum_{i=1}^{n} v_i. \end{aligned}$$

Note that by definition the u_i's are independent, and similarly the v_i's are independent.

We now compute the mean and variance of the first term. Using the moments of X, we compute the expectation of each u_i as

$$\begin{aligned}\mathbb{E}[u_i] &= \mathbb{E}\Big[\sum_{j=1}^{p} X_{ij}^2\psi_j\beta_j + \sum_{j\neq\ell} X_{ij}X_{i\ell}\psi_j\beta_\ell\Big] \\ &= \sum_{j=1}^{p}\mathbb{E}[X_{ij}^2]\,\psi_j\beta_j + \sum_{j\neq\ell}\mathbb{E}[X_{ij}]\,\mathbb{E}[X_{i\ell}]\,\psi_j\beta_\ell \\ &= \psi^T\beta.\end{aligned}$$

Hence we have that $\mathbb{E}\big[\sum_{i=1}^n u_i\big] = n\,\psi^T\beta$. Similarly, we compute the second moment of u_i as

$$\begin{aligned}\mathbb{E}[u_i^2] &= \mathbb{E}\Bigg[\Big(\sum_{j=1}^{p} X_{ij}^2\psi_j\beta_j\Big)^2 + \Big(\sum_{j\neq\ell} X_{ij}X_{i\ell}\psi_j\beta_\ell\Big)^2 + 2\Big(\sum_{j=1}^{p} X_{ij}^2\psi_j\beta_j\Big)\Big(\sum_{j\neq\ell} X_{ij}X_{i\ell}\psi_j\beta_\ell\Big)\Bigg] \\ &= \sum_{j=1}^{p}\mathbb{E}[X_{ij}^4]\,\psi_j^2\beta_j^2 + \sum_{j\neq\ell}\mathbb{E}[X_{ij}^2]\,\mathbb{E}[X_{i\ell}^2]\,\psi_j\beta_j\psi_\ell\beta_\ell + \sum_{j\neq\ell}\mathbb{E}[X_{ij}^2]\,\mathbb{E}[X_{i\ell}^2]\,\psi_j^2\beta_\ell^2 \\ &\quad + \sum_{j\neq\ell}\mathbb{E}[X_{ij}^2]\,\mathbb{E}[X_{i\ell}^2]\,\psi_j\beta_\ell\psi_\ell\beta_j \\ &= \frac{1}{\gamma}\sum_{j=1}^{p}\psi_j^2\beta_j^2 + 2\sum_{j\neq\ell}\psi_j\beta_j\psi_\ell\beta_\ell + \sum_{j\neq\ell}\psi_j^2\beta_\ell^2 \\ &= 2\Big(\sum_{j=1}^{p}\psi_j^2\beta_j^2 + \sum_{j\neq\ell}\psi_j\beta_j\psi_\ell\beta_\ell\Big) + \Big(\sum_{j=1}^{p}\psi_j^2\beta_j^2 + \sum_{j\neq\ell}\psi_j^2\beta_\ell^2\Big) + \Big(\frac{1}{\gamma} - 3\Big)\sum_{j=1}^{p}\psi_j^2\beta_j^2 \\ &= 2(\psi^T\beta)^2 + \|\psi\|_2^2\,\|\beta\|_2^2 + \Big(\frac{1}{\gamma} - 3\Big)\sum_{j=1}^{p}\psi_j^2\beta_j^2.\end{aligned}$$

This yields that the variance of each u_i is equal to $\mathrm{var}(u_i) = (\psi^T\beta)^2 + \|\psi\|_2^2\,\|\beta\|_2^2 + \big(\frac{1}{\gamma} - 3\big)\sum_{j=1}^{p}\psi_j^2\beta_j^2$. Using the fact that the u_i's are independent, we have

$$\begin{aligned}\mathrm{var}\Big(\sum_{i=1}^{n} u_i\Big) &= \sum_{i=1}^{n}\mathrm{var}(u_i) \\ &= n\left\{(\psi^T\beta)^2 + \|\psi\|_2^2\,\|\beta\|_2^2 + \Big(\frac{1}{\gamma} - 3\Big)\sum_{j=1}^{p}\psi_j^2\beta_j^2\right\}.\end{aligned}$$

Next, we compute the mean and variance of the v_i's. Using the moments of W and the fact that W is independent of X, we have

$$\mathbb{E}[v_i] \;=\; \sum_{j=1}^{p} \mathbb{E}[W_i]\,\mathbb{E}[X_{ij}]\,\psi_j \;=\; 0,$$

so that $\mathbb{E}\big[\sum_{i=1}^n v_i\big] = 0$. Similarly, we compute the second moment of v_i as

$$\begin{aligned}
\mathbb{E}[v_i^2] &= \mathbb{E}[W_i^2]\,\mathbb{E}\Bigg[\bigg(\sum_{j=1}^{p} X_{ij}\psi_j\bigg)^2\Bigg] \\
&= \sigma^2 \sum_{j=1}^{p} \mathbb{E}[X_{ij}^2]\,\psi_j^2 + \sigma^2 \sum_{j\neq\ell} \mathbb{E}[X_{ij}]\,\mathbb{E}[X_{i\ell}]\,\psi_j\psi_\ell \\
&= \sigma^2\,\|\psi\|_2^2.
\end{aligned}$$

Hence the variance of each v_i is $\mathrm{var}(v_i) = \sigma^2\,\|\psi\|_2^2$, and using the fact that the v_i's are independent, we have

$$\begin{aligned}
\mathrm{var}\bigg(\sum_{i=1}^{n} v_i\bigg) &= \sum_{i=1}^{n} \mathrm{var}(v_i) \\
&= n\,\sigma^2\,\|\psi\|_2^2.
\end{aligned}$$

To compute the covariance between u_i and v_i, note that

$$\mathbb{E}[u_i v_i] \;=\; \mathbb{E}[W_i]\,\mathbb{E}\Bigg[\bigg(\sum_{j=1}^{p} X_{ij}\psi_j\bigg)^2\bigg(\sum_{j=1}^{p} X_{ij}\beta_j\bigg)\Bigg] \;=\; 0,$$

and hence $\mathrm{cov}(u_i, v_i) = 0$. Note also that $\mathrm{cov}(u_i, v_j) = 0$ since u_i and v_j are independent for $i \neq j$. Thus we have,

$$\mathrm{cov}\bigg(\sum_{i=1}^{n} u_i,\, \sum_{i=1}^{n} v_i\bigg) \;=\; \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{cov}(u_i, v_j) \;=\; 0.$$

Putting together the pieces, we have

$$\mathbb{E}\bigg[\frac{1}{n}\,Z^T Y\bigg] \;=\; \frac{1}{n}\,\mathbb{E}\bigg[\sum_{i=1}^{n} u_i\bigg] + \frac{1}{n}\,\mathbb{E}\bigg[\sum_{i=1}^{n} v_i\bigg] \;=\; \psi^T\beta$$

and

$$\begin{aligned}
\operatorname{var}\left(\frac{1}{n} Z^T Y\right) &= \frac{1}{n^2}\operatorname{var}\left(\sum_{i=1}^n u_i\right) + \frac{1}{n^2}\operatorname{var}\left(\sum_{i=1}^n v_i\right) + \frac{2}{n^2}\operatorname{cov}\left(\sum_{i=1}^n u_i, \sum_{i=1}^n v_i\right) \\
&= \frac{1}{n}\left\{(\psi^T\beta)^2 + \left(\|\beta\|_2^2 + \sigma^2\right)\|\psi\|_2^2 + \left(\frac{1}{\gamma} - 3\right)\sum_{j=1}^p \psi_j^2\beta_j^2\right\}
\end{aligned}$$

as claimed.

2.5.2 Sparse noisy JL embeddings

We now prove a sparse noisy variant of the Johnson-Lindenstrauss lemma stated in Theorem 1. We first evaluate the claim for any pair of vectors β and $\psi^{(m)}$, and subsequently take the union bound over the set of all p orthonormal basis vectors in Ψ. Fix any vector $\psi \in \mathbb{R}^p$ with $\|\psi\|_2 = 1$. Let n_1 and n_2 be positive integers, which we will determine, and set $n = n_1 n_2$. Partition the $(n \times p)$ measurement matrix X into n_2 submatrices

$$X = \begin{bmatrix} X^{(1)} \\ \vdots \\ X^{(n_2)} \end{bmatrix},$$

where each $X^{(\ell)}$ is of size $(n_1 \times p)$. Accordingly, we partition the noisy observation vector Y into n_2 vectors $\left\{Y^{(1)}, \ldots, Y^{(n_2)}\right\}$, where each $Y^{(\ell)} \in \mathbb{R}^{n_1}$ is defined as

$$Y^{(\ell)} := X^{(\ell)}\beta + W^{(\ell)}$$

and $W^{(\ell)}$ denotes the corresponding subvector of the unknown noise W. In addition, we define the random projections of ψ as $Z^{(\ell)} := X^{(\ell)}\psi \in \mathbb{R}^{n_1}$, for $\ell = 1, \ldots, n_2$.

With this notation, we now form n_2 independent estimates $\{\alpha_1, \ldots, \alpha_{n_2}\}$ for the inner product $\psi^T \beta$, defined as

$$\alpha_\ell \;:=\; \frac{1}{n_1} \big(Z^{(\ell)}\big)^T Y^{(\ell)}.$$

Applying Lemma 3 to each α_ℓ yields that $\mathbb{E}[\alpha_\ell] = \psi^T \beta$ and

$$\mathrm{var}(\alpha_\ell) \;=\; \frac{1}{n_1} \left\{ (\psi^T \beta)^2 + \big(\|\beta\|_2^2 + \sigma^2\big) \|\psi\|_2^2 + \left(\frac{1}{\gamma} - 3\right) \sum_{j=1}^p \psi_j^2 \beta_j^2 \right\}.$$

For any $\epsilon > 0$, applying Chebyshev's inequality and using the fact that $\|\psi\|_2^2 = 1$, we have

$$\begin{aligned}
\mathbb{P}\big[|\alpha_\ell - \psi^T\beta| \geq \epsilon \|\beta\|_2\big] \;&\leq\; \frac{\mathrm{var}(\alpha_\ell)}{\epsilon^2 \|\beta\|_2^2} \\
&=\; \frac{1}{\epsilon^2 n_1} \left\{ \frac{(\psi^T\beta)^2}{\|\beta\|_2^2 \|\psi\|_2^2} + \frac{\|\beta\|_2^2 + \sigma^2}{\|\beta\|_2^2} + \left(\frac{1}{\gamma} - 3\right) \frac{\sum_{j=1}^p \psi_j^2 \beta_j^2}{\|\beta\|_2^2} \right\} \\
&\leq\; \frac{1}{\epsilon^2 n_1} \left\{ 2 + \frac{\sigma^2}{\|\beta\|_2^2} + \frac{\omega^2 \|\beta\|_2^2 \sum_{j=1}^p \psi_j^2}{\gamma \|\beta\|_2^2} \right\} \\
&=\; \frac{1}{\epsilon^2 n_1} \left(2 + \frac{\sigma^2}{\|\beta\|_2^2} + \frac{\omega^2}{\gamma} \right)
\end{aligned}$$

where the second inequality follows from the Cauchy-Schwarz inequality and the fact that β is component-wise upper bounded as $\beta_j^2 \leq \omega^2 \|\beta\|_2^2$. Let us define $q :$ $= \frac{1}{\epsilon^2 n_1} \left(2 + \frac{\sigma^2}{\|\beta\|_2^2} + \frac{\omega^2}{\gamma}\right)$. Consequently, we can guarantee that each estimate α_ℓ lies within an ϵ-interval around its mean with probability at least $1 - q$ by setting $n_1 = \frac{1}{q\epsilon^2} \left(2 + \frac{\sigma^2}{\|\beta\|_2^2} + \frac{\omega^2}{\gamma}\right)$.

From here, we define a new estimate $\bar{\alpha}$ as the median of the independent random variables $\{\alpha_1, \ldots, \alpha_{n_2}\}$. If the median $\bar{\alpha}$ lies outside the ϵ-interval, then at least half of the estimators α_ℓ must lie outside the interval. In particular, define the 0-1 indicator random variable $\xi_\ell := \mathbb{I}\big[|\alpha_\ell - \psi^T\beta| \geq \epsilon\|\beta\|_2\big]$, so that $\sum_{\ell=1}^{n_2} \xi_\ell$ is equal to the number of α_ℓ's that lie outside the ϵ-interval. Since $\xi_1, \ldots, \xi_{n_2}$ are independent and $\mu := \mathbb{E}\big[\frac{1}{n_2}\sum_{\ell=1}^{n_2} \xi_\ell\big] \leq q$, by Hoeffding's inequality we have

$$\mathbb{P}\left[\sum_{\ell=1}^{n_2} \xi_\ell \geq (\mu + t)n_2\right] \;\leq\; \exp(-2t^2 n_2)$$

for any $t \in (0, 1-\mu)$. Setting $q = \frac{1}{4}$ and $t = \frac{1}{4}$, we obtain the bound

$$\begin{aligned}\mathbb{P}\left[|\bar{\alpha} - \psi^T\beta| \geq \epsilon\|\beta\|_2\right] \;&\leq\; \mathbb{P}\left[\sum_{\ell=1}^{n_2} \xi_\ell \geq \frac{1}{2}n_2\right] \\ &\leq\; \exp\left(-\frac{n_2}{8}\right).\end{aligned}$$

Putting together the pieces, we have that for any pair of vectors β and $\psi^{(m)} \in \{\psi^{(1)}, \ldots, \psi^{(p)}\}$, we can use the above method to produce an unbiased estimate $\widehat{\theta}_m$ for $\beta^T\psi^{(m)}$ that lies outside an ϵ-interval around its mean with probability at most $\exp(-n_2/8)$. Taking the union bound over all p pairs, we have

$$\mathbb{P}\Big[\bigcup_{m=1,\ldots,p} \Big\{\big|\widehat{\theta}_m - \beta^T\psi^{(m)}\big| \geq \epsilon\|\beta\|_2\Big\}\Big] \;\leq\; p\,\exp\left(-\frac{n_2}{8}\right).$$

For any ϵ and $\delta > 0$, setting $n_1 = \frac{4}{\epsilon^2}\left(2 + \frac{\sigma^2}{\|\beta\|_2^2} + \frac{\omega^2}{\gamma}\right)$, $n_2 = 8(1+\delta)\log p$, and the number of observations $n = n_1 n_2$ gives the result as claimed.

2.5.3 Estimating top k coefficients

We now prove our main result in Theorem 2. Fix an orthonormal transform Ψ consisting of the p basis vectors $\{\psi^{(1)}, \ldots, \psi^{(p)}\}$ as rows, and let $\theta_m = \beta^T\psi^{(m)}$ for $m = 1, \ldots, p$. Note that by orthonormality, we have $\|\beta\|_2^2 = \|\theta\|_2^2$ and $\|\beta - \widehat{\beta}\|_2^2 = \|\theta - \Psi\widehat{\beta}\|_2^2$, and so the problem of estimating the data vector β is equivalent to the problem of estimating the coefficient vector θ. Recall that the best k-term approximation of θ, obtained by keeping the largest k transform coefficients and setting the rest to zero, is $\|\theta - \widehat{\theta}_{opt}\|_2^2 = \sum_{m=k+1}^{p} |\theta|_{(m)}^2$, and assume that $\|\theta - \widehat{\theta}_{opt}\|_2^2 \leq \eta \, \|\theta\|_2^2$ for some $\eta > 0$.

By Theorem 1, if the number of observations is bounded as $n \geq \left(\frac{32(1+\delta)}{\epsilon^2}\left(2 + \frac{\sigma^2}{\|\beta\|_2^2} + \frac{\omega^2}{\gamma}\right)\log p\right)$, then we can produce estimates $\{\widehat{\theta}_1, \ldots, \widehat{\theta}_p\}$ satisfying

$$|\widehat{\theta}_m - \theta_m| \quad \leq \quad \epsilon\|\theta\|_2$$

with high probability. Since by the triangle inequality we have that $\Big||\widehat{\theta}_m| - |\theta_m|\Big| \leq |\widehat{\theta}_m - \theta_m|$, the above condition implies

$$|\theta_m| - \epsilon\|\theta\|_2 \quad \leq \quad |\widehat{\theta}_m| \quad \leq \quad |\theta_m| + \epsilon\|\theta\|_2 \tag{2.14}$$

for all $m = 1, \ldots, p$.

We construct an approximation for β by estimating the largest k transform coefficients of β. More specifically, sort the estimates $\widehat{\theta}$ in decreasing order of magnitude, i.e. $|\widehat{\theta}|_{(1)} \geq |\widehat{\theta}|_{(2)} \geq \cdots \geq |\widehat{\theta}|_{(p)}$. Define a new vector $\widetilde{\theta}$ by keeping the k largest components of $\widehat{\theta}$ in magnitude, and setting the remaining components to zero. We then take the inverse transform of $\widetilde{\theta}$, and obtain the approximation vector $\widehat{\beta} = \Psi^T\widetilde{\theta}$.

There are two sources of error in our approximation: one is incorrectly estimating which indices are the top k transform coefficients, and the other is the error in approximating the transform coefficients that are kept. Let $\widetilde{S}$ be the index set of the k largest estimates $\widehat{\theta}_m$'s which we keep (and consequently $\widetilde{S}^C$ is the index set of the estimates we set to zero). Furthermore, let S be the true index set of the k

largest transform coefficients in θ. With this notation, the approximation error can be bounded as

$$\begin{aligned}\|\theta - \widetilde{\theta}\|_2^2 &= \sum_{m \in \widetilde{S}} |\theta_m - \widehat{\theta}_m|^2 + \sum_{m \in \widetilde{S}^C} |\theta_m|^2 \\ &\leq k\epsilon^2 \|\theta\|_2^2 + \sum_{m \in \widetilde{S}^C} |\theta_m|^2\end{aligned}$$

In the ideal case, if we correctly estimate the largest-k set $\widetilde{S} = S$, then the second term above would become $\sum_{m \in \widetilde{S}^C} |\theta_m|^2 = \sum_{m \in S^C} |\theta_m|^2$. If $\widetilde{S} \neq S$, then we must have chosen to keep the estimate of a transform coefficient which was not one of the k largest, and consequently set to zero the estimate of a coefficient which was one of the k largest. In other words, there exists some indices $m \in \widetilde{S}, m \notin S$ and $\ell \notin \widetilde{S}, \ell \in S$. This implies that $|\widehat{\theta}_m| > |\widehat{\theta}_\ell|$, but $|\theta_m| < |\theta_\ell|$. Since each estimate lies within an $(\epsilon\|\theta\|_2)$-interval around the corresponding transform coefficient (by (2.14)), this confusion can only happen if $|\theta_\ell| - |\theta_m| \leq 2\epsilon\|\theta\|_2$. Furthermore, note that $|\theta_\ell|^2 + |\theta_m|^2 \leq \|\theta\|_2^2$ implies that $|\theta_\ell| + |\theta_m| \leq \sqrt{3}\|\theta\|_2$. Hence we have that $|\theta_\ell|^2 - |\theta_m|^2 = (|\theta_\ell| - |\theta_m|)(|\theta_\ell| + |\theta_m|) \leq 2\sqrt{3}\epsilon\|\theta\|_2^2$. For each time this confusion happens, we get an additional error of $+|\theta_\ell|^2 - |\theta_m|^2$, and this confusion can happen at most k times. Therefore, we obtain the bound

$$\sum_{m \in \widetilde{S}^C} |\theta_m|^2 \leq \sum_{m \in S^C} |\theta_m|^2 + k\,(2\sqrt{3}\epsilon)\,\|\theta\|_2^2.$$

Putting everything together, the approximation error can then be bounded as

$$\begin{aligned}\|\theta - \widetilde{\theta}\|_2^2 &\leq k\,\epsilon^2\,\|\theta\|_2^2 + k\,(2\sqrt{3}\epsilon)\,\|\theta\|_2^2 + \sum_{m \in S^C} |\theta_m|^2 \\ &= k\,(\epsilon^2 + 2\sqrt{3}\epsilon)\,\|\theta\|_2^2 + \|\theta - \widehat{\theta}_{opt}\|_2^2 \\ &\leq k\,(\epsilon^2 + 2\sqrt{3}\epsilon)\,\|\theta\|_2^2 + \eta\|\theta\|_2^2\end{aligned}$$

Setting $\epsilon' = k(\epsilon^2 + 2\sqrt{3}\epsilon)$ and solving for the positive root, we have that $\epsilon = -\sqrt{3} + \sqrt{3 + \frac{\epsilon'}{k}} = O(\frac{\epsilon'}{k})$. Plugging this back into the number of observations, we have $n \geq \Omega\left(\frac{(1+\delta)}{(\epsilon')^2}\left(2 + \frac{\sigma^2}{\|\beta\|_2^2} + \frac{\omega^2}{\gamma}\right) k^2 \log p\right)$ as claimed.

2.5.4 Relating compressibility and ℓ_∞-norm

We show that the compressibility of a signal is related to its ℓ_∞-norm, as stated in Lemma 2. By the definition of the (orthonormal) inverse discrete Fourier transform, we have that each component of the vector β can be bounded as

$$\begin{aligned} |\beta_i| &\leq \frac{1}{\sqrt{p}} \sum_{m=1}^{p} |\theta_m| \left| \exp\left(\mathrm{j}\frac{2\pi(m-1)i}{p}\right)\right| \\ &= \frac{1}{\sqrt{p}} \sum_{m=1}^{p} |\theta_m| \;=\; \frac{1}{\sqrt{p}} \|\theta\|_1, \end{aligned}$$

for all $i = 1, \ldots, p$. This gives a bound on the ℓ_∞-norm of β of the form $\|\beta\|_\infty \leq \frac{1}{\sqrt{p}}\|\theta\|_1$.

For r-compressible signals, the DFT coefficients obey a power law decay as in (2.1), and consequently

$$\|\theta\|_1 \leq C \sum_{m=1}^{p} m^{-r}.$$

For $r = 1$, the summation is a Harmonic series, which diverges slowly as p grows and scales like $O(\log p)$. For $r > 1$, the summation becomes a p-series (or Riemann zeta function) which converges. In particular, we have $\sum_{m=1}^{p} m^{-r} \leq 1 + \int_1^p x^{-r}\, \mathrm{d}x = 1 + \left(\frac{1}{r-1}\right)\left(1 - \frac{1}{p^{r-1}}\right)$, which is upper bounded by a constant that depends only on r.

Therefore, if the data is compressible with $r = 1$, then $\|\theta\|_1 = O(\log p)$, and $\|\beta\|_\infty = O\left(\frac{\log p}{\sqrt{p}}\right)$. If $r > 1$, then $\|\theta\|_1 = O(1)$, and $\|\beta\|_\infty = O(\frac{1}{\sqrt{p}})$. Finally, we can verify that compressible signals have finite energy, since by orthonormality, we have that $\|\beta\|_2^2 = \|\theta\|_2^2 \leq C^2 \sum_{m=1}^{p} m^{-2r}$, and $\int_1^{p+1} x^{-2r}\,\mathrm{d}x \leq \sum_{m=1}^{p} m^{-2r} \leq 1 + \int_1^{p} x^{-2r}\,\mathrm{d}x$.

2.6 Comparisons and simulations

In this section, we give a numeric example comparing the approximation of piecewise polynomial data using wavelet transforms, sparse random projections, and the non-sparse schemes of AMS sketching and compressed sensing. We know analytically that compressed sensing requires only $O(k \log \frac{p}{k})$ random projections to obtain an approximation error comparable to the best k-term approximation, while sketching requires $O(k^2 \log p)$. However, the compressed sensing decoder has a computational complexity of $O(p^3)$ while the sketching decoding complexity is $O(np)$, where n is the number of random projections used. The low decoding complexity would make it possible for sensors and other low-powered collectors to query and decode a coarse approximation of the data cheaply and quickly. Collectors with greater resources can still query more sensors and recover a better approximation. Our sparse random projections recovery method is based on the low-complexity sketching decoder.

We have seen theoretically that there is a trade-off between the sparsity of the random projections and the number of random projections needed for a good approximation. The degree of sparsity corresponds to the number of packets per sensor that must be transmitted in the pre-processing stage. Sparse random projections can thus reduce the communication cost per sensor from $O(p)$ to $O(\log p)$ when compared to the non-sparse schemes.

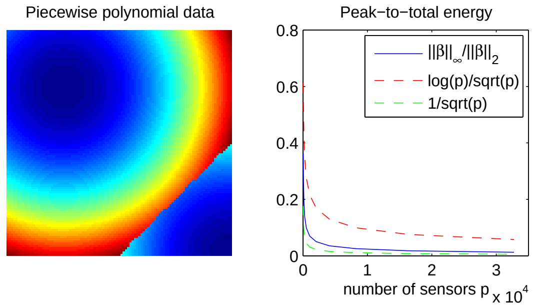


Figure 2.4. (a) Piecewise polynomial data. (b) Peak-to-total energy condition on data.

We now examine experimentally the effect of the sparsity of the random projections on data approximation. In our experimental setup, p sensors are placed randomly on a unit square, and measure piecewise polynomial data with two second-order polynomials separated by a line discontinuity, as shown in Figure 2.4 (a). In Figure 2.4 (b), we plot the peak-to-total energy (2.9) of the data and verify that it is bounded between $\frac{\log p}{\sqrt{p}}$ and $\frac{1}{\sqrt{p}}$.

Figure 2.5 compares the approximation error of sparse random projections to non-sparse AMS sketching and the optimal k-term approximation in the noiseless setting. The average approximation error using sparse random projections is as good as dense random projections, and very close to the optimal k-term approximation. However, the standard deviation of the approximation error increases with greater sparsity.

Figure 2.6 compares the approximation using sparse random projections for varying degrees of sparsity, along with the non-sparse schemes of sketching and compressed

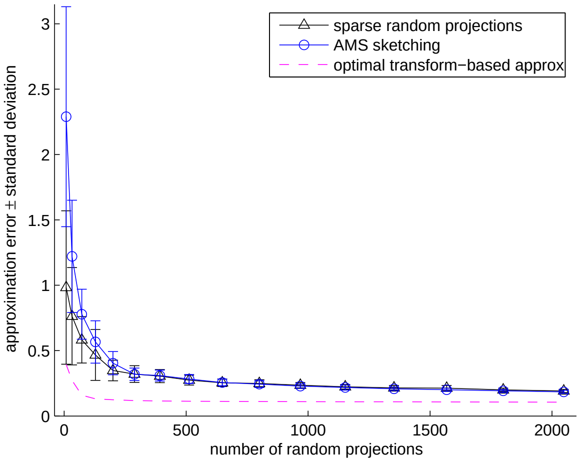


Figure 2.5. In the noiseless setting, a comparison of the approximation error of piecewise polynomial data using sparse random projections, non-sparse AMS sketching, and optimal Haar wavelet based approximation. The relative approximation error of the data $\frac{\|\beta-\widehat{\beta}\|_2^2}{\|\beta\|_2^2}$ is plotted versus the number of random projections $n = k^2 \log p$, for $p = 2048$ sensors. The error bars show the standard deviation of the approximation error.

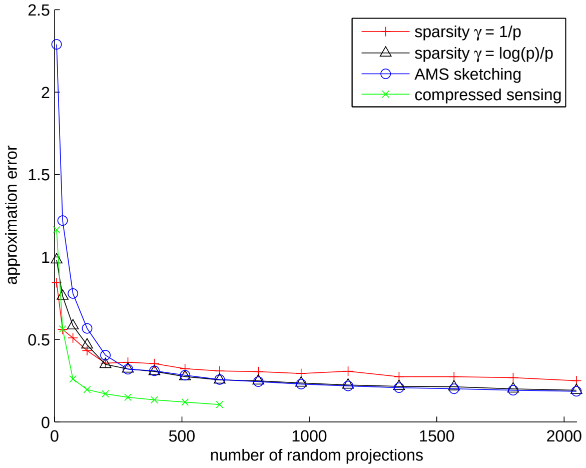


Figure 2.6. The effect of sparsity of the random projections on approximation error is illustrated in the noiseless setting. Varying degrees of sparsity in the random projections are compared against the dense projection methods of AMS sketching and compressed sensing. The relative approximation error of the data $\frac{\|\beta-\widehat{\beta}\|_2^2}{\|\beta\|_2^2}$ is plotted versus the number of random projections n, for $p = 2048$ sensors. The average number of non-zeros in the sparse random projections is γp.

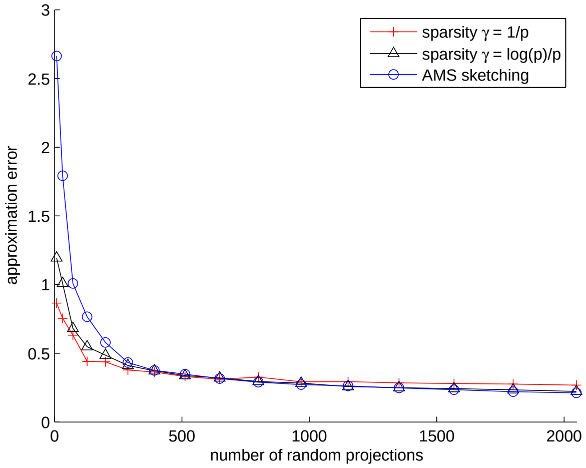


Figure 2.7. The effect of measurement sparsity of on approximation error is illustrated in the noisy setting using additive Gaussian noise. Varying degrees of measurement sparsity are compared against the dense projection methods of AMS sketching and compressed sensing. Again, the relative approximation error of the data $\frac{\|\beta-\widehat{\beta}\|_2^2}{\|\beta\|_2^2}$ is plotted versus the number of random projections n, for $p = 2048$ sensors.

sensing in the noiseless case. Sparse random projections with $O(\log p)$ non-zeros per row perform as well as (dense) sketching, while sparse random projections with $O(1)$ non-zeros per row perform slightly worse. As we would expect from the analysis, the compressed sensing decoder obtains better approximation error than the sketching decoder for the same number of random projections. But, the compressed sensing decoder has a higher computational complexity, which was appreciable in our simulations.

Figure 2.7 illustrates the effect of measurement sparsity on approximation error in the noisy observation setting, using additive Gaussian noise with mean zero and variance 50. The average approximation error with noise behaves very similarly to the

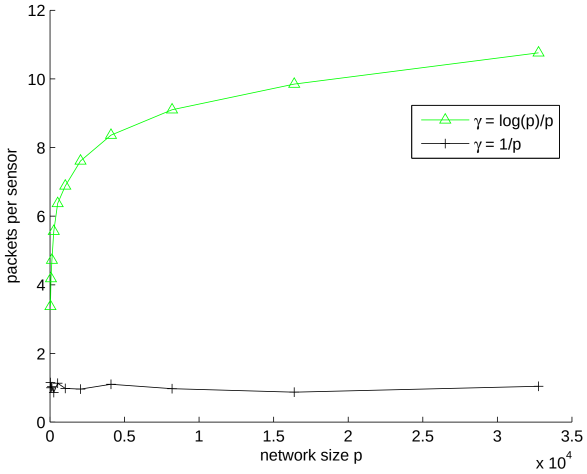


Figure 2.8. Communication cost for sparse random projections with varying degrees of sparsity. In comparison, compressed sensing and sketching both require $O(p)$ packets per sensor.

average approximation error without noise, thus confirming numerically the stability of our algorithm in the presence of noise.

Finally, Figure 2.8 shows the communication cost of computing sparse random projections in the network for varying degrees of sparsity. Both compressed sensing and sketching require $O(p)$ packets per sensor to compute the dense random projections in a network of size p. Thus sparse random projections greatly reduce the overall communication cost.

2.7 Discussion

In this chapter, we proposed distributed sparse random projections and showed that they can enable robust and refinable data approximation. In our framework, sensors store sparse random projections of the data, which allows any decoder to recover an approximation by querying a sufficient number of sensors from anywhere in the network. The communication cost to pre-process the data in the network is determined by the sparsity of the random projections. The quality of the approximation depends only on the number of random projections that are collected, and not which sensors are queried. Our results can be extended to scenarios in which random projections are collected only from sensors along the boundary of the network, or in which random projections are computed locally first in a multiresolution hierarchy (see Chapter 5 for a discussion of open problems).

In addition, we showed that a fast sketching decoder can recover compressible signals based on sparse random projections, and that the approximation is stable in the presence of noise. Our results apply to general measurement matrices for which we have control over the first four moments, and include as special cases the γ-sparsified Gaussian ensemble and the γ-sparsified Bernoulli ensemble. Our analysis reveals that the effect of measurement sparsity on the sampling efficiency of the sketching decoder is characterized by the quantity ω/γ. Intuitively, if the signal and the measurement matrix are both be very sparse, then the measurements will rarely hit the non-zero locations of the signal. But this should impede the ability of any decoder to recover the signal. In Chapter 3, we examine the effect of measurement sparsity on the information-theoretic limits of the sparse recovery problem.

Chapter 3

Information-theoretic limits on sparse signal recovery

3.1 Introduction

Sparsity recovery refers to the problem of estimating the support of a p-dimensional but k-sparse vector $\beta \in \mathbb{R}^p$, based on a set of n noisy linear observations. The sparsity recovery problem is of broad interest, arising in subset selection in regression [42], model selection in sparse graphs [41], group testing, signal denoising [16], sparse approximation [43], and compressive sensing [25, 14]. A large body of work (e.g., [16, 25, 28, 14, 13, 41, 55, 56, 58]) has analyzed the performance of computationally tractable methods, in particular based on ℓ_1 or other convex relaxations, for estimating high-dimensional sparse signals. Such results have established conditions, on signal sparsity and the choice of measurement matrices, under which a given recovery method succeeds with high probability.

Of complementary interest are the information-theoretic limits of the sparsity recovery problem, which apply to the performance of any procedure regardless of its computational complexity. Such analysis has two purposes: first, to demonstrate where known polynomial-time methods achieve the information-theoretic bounds, and second, to reveal situations in which current methods are sub-optimal. An interesting question which arises in this context is the effect of the choice of measurement matrix on the information-theoretic limits. As we will see, the standard Gaussian measurement ensemble achieves an optimal scaling of the number of observations required for recovery. However, this choice produces highly dense matrices, which may lead to prohibitively high computational complexity and storage requirements[1]. In contrast, sparse measurement matrices directly reduce encoding and storage costs, and can also lead to fast decoding algorithms by exploiting problem structure (see Section 3.2.3 for a brief overview of the growing literature in this area). In addition, measurement sparsity can be used to lower communication cost and latency in distributed sensor network and streaming applications. On the other hand, measurement sparsity can potentially reduce statistical efficiency by requiring more observations to recover the signal. Intuitively, the non-zeros in the signal may rarely align with the non-zeros in a sparse measurement matrix[2]. Therefore, an important question is to characterize the trade-off between measurement sparsity and statistical efficiency.

[1] For example, ℓ_1-recovery methods based on linear programming have complexity $O(p^3)$ in the signal dimension p.

[2] Note however that misalignments between the measurements and the signal still reveal *some* information about the locations of the non-zeros in the signal.

This chapter provides two classes of information-theoretic bounds. First, we derive sharper necessary conditions for exact support recovery, applicable to a general class of dense measurement matrices (including non-Gaussian ensembles). In conjunction with the sufficient conditions from previous work [59], this analysis provides a sharp characterization of necessary and sufficient conditions for various sparsity regimes. Second, we address the effect of measurement sparsity, meaning the fraction $\gamma \in (0, 1]$ of non-zeros per row in the matrices used to collect measurements. We derive lower bounds on the number of observations required for exact sparsity recovery, as a function of the signal dimension p, signal sparsity k, and measurement sparsity γ. This analysis highlights a trade-off between the statistical efficiency of a measurement ensemble and the computational complexity associated with storing and manipulating it.

The remainder of the chapter is organized as follows. We first define the sparsity recovery problem in Section 3.2, and then discuss our contributions and some connections to related work in Section 3.2.3. Section 3.3 provides precise statements of our main results, as well as a discussion of their consequences. Section 3.4 describes our general approach based on Fano's method, while Sections 3.5 and 3.6 provide proofs of the necessary conditions for various classes of measurement matrices. Finally, we conclude and discuss open problems in Section 3.7.

3.2 Exact support recovery

Let $\beta \in \mathbb{R}^p$ be a fixed but unknown vector, with the support set of β defined as

$$S(\beta) \;:=\; \{i \in \{1, \ldots, p\} \mid \beta_i \neq 0\}. \tag{3.1}$$

We refer to $k := |S(\beta)|$ as the *signal sparsity*, and p as the *signal dimension*. Suppose we are given a vector of n noisy observations $Y \in \mathbb{R}^n$, of the form

$$Y \;=\; X\beta + W, \tag{3.2}$$

where $X \in \mathbb{R}^{n \times p}$ is the known measurement matrix, and $W \sim N(0, \sigma^2 I_{n \times n})$ is additive Gaussian noise. Our goal is to perform exact recovery of the underlying sparsity pattern $S(\beta)$, which we refer to as the *sparsity recovery problem*. The focus of this chapter is to find conditions on the model parameters (n, p, k) that are necessary for any method to successfully recover the support set $S(\beta)$. Our results apply to various classes of dense and γ-sparsified measurement matrices, which will be defined in Section 3.3.

3.2.1 Classes of signals

The difficulty of sparsity recovery from noisy measurements naturally depends on the minimum value of β on its support, defined by the function

$$\lambda^*(\beta) \;:=\; \min_{i \in S(\beta)} |\beta_i|. \tag{3.3}$$

We study the class of signals parameterized by a lower bound λ on the minimum value

$$\mathcal{C}_{p,k}(\lambda) \;:=\; \{\beta \in \mathbb{R}^p \mid |S(\beta)| = k,\ \lambda^*(\beta) \geq \lambda\}. \tag{3.4}$$

The associated class of sparsity patterns $\mathcal{C}_{p,k}$ is the collection of all $N = \binom{p}{k}$ possible subsets of size k. We assume without loss of generality that the noise variance $\sigma^2 = 1$, since any scaling of σ can be accounted for in the scaling of β.

3.2.2 Decoders and error criterion

Suppose that nature chooses some vector β from the signal class $\mathcal{C}_{p,k}(\lambda)$. The statistician observes n samples $Y = X\beta + W \in \mathbb{R}^n$ and tries to infer the underlying sparsity pattern $S(\beta)$. Our analysis applies to arbitrary decoders. A decoder is a mapping $g : \mathbb{R}^n \to \mathcal{C}_{p,k}$ from the observations Y to an estimated subset $\widehat{S} = g(Y)$. We measure the error between the estimate $\widehat{S}$ and the true support $S(\beta)$ using the $\{0,1\}$-valued loss function $\mathbb{I}[g(Y) \neq S(\beta)]$, which corresponds to a standard model selection error criterion. The probability of incorrect subset selection is then the associated 0-1 risk $\mathbb{P}[g(Y) \neq S \mid S(\beta) = S]$, where the probability is taken over the measurement noise W and the choice of random measurement matrix X. We define the maximal probability of error over the class $\mathcal{C}_{p,k}(\lambda)$ as

$$\omega(g) \;:=\; \max_{\beta \in \mathcal{C}_{p,k}(\lambda)} \mathbb{P}[g(Y) \neq S \mid S(\beta) = S]. \tag{3.5}$$

We say that sparsity recovery is asymptotically reliable over the signal class $\mathcal{C}_{p,k}(\lambda)$ if $\omega(g) \to 0$ as $n \to \infty$.

With this set-up, our goal is to find necessary conditions on the parameters $(n, p, k, \lambda, \gamma)$ that any decoder, regardless of its computational complexity, must satisfy for asymptotically reliable recovery to be possible. We are interested in lower bounds on the number of measurements n, in general settings where both the signal sparsity k and the measurement sparsity γ are allowed to scale with the signal dimension p.

3.2.3 Related work

One body of past work [31, 52, 2] has focused on the information-theoretic limits of sparse estimation under ℓ_2 and other distortion metrics, using power-based SNR measures of the form

$$\text{SNR} \;:=\; \frac{\mathbb{E}[\|X\beta\|_2^2]}{\mathbb{E}[\|W\|_2^2]} \;=\; \|\beta\|_2^2. \tag{3.6}$$

(Note that the second equality assumes that the noise variance $\sigma^2 = 1$, and that the measurement matrix is standardized, with each element X_{ij} having zero mean and variance one.) It is important to note that the power-based SNR (3.6), though appropriate for ℓ_2-distortion, is not suitable for the support recovery problem. Although the minimum value is related to this power-based measure by the inequality $k\lambda^2 \leq \text{SNR}$, for the ensemble of signals $\mathcal{C}_{p,k}(\lambda)$ defined in equation (3.4), the ℓ_2-based SNR (3.6) can be made arbitrarily large while still having one coefficient β_i equal to the minimum value (assuming that $k > 1$). Consequently, as our results show, it is possible to generate problem instances for which support recovery is arbitrarily difficult—in particular, by sending $\lambda \to 0$ at an arbitrarily rapid rate—even as the power-based SNR (3.6) becomes arbitrarily large.

The paper [59] was the first to consider the information-theoretic limits of exact subset recovery using standard Gaussian measurement ensembles, explicitly identifying the minimum value λ as the key parameter. This analysis yielded necessary and sufficient conditions on general quadruples (n, p, k, λ) for asymptotically reliable recovery. Subsequent work on the problem has yielded sharper conditions for standard Gaussian ensembles [47, 5, 30, 3], and extended this type of analysis to the criterion of partial support recovery [5, 47]. We consider only exact support recovery, but provide results for general dense measurement ensembles, including non-Gaussian matrices. In conjunction with known sufficient conditions [59], one consequence of our first main result (Theorem 3, below) is a set of sharp necessary and sufficient conditions for the optimal decoder to recover the support of a signal with linear sparsity ($k = \Theta(p)$), using only a linear fraction of observations ($n = \Theta(p)$). As we discuss at more length in Section 3.3.1, for the special case of the standard Gaussian ensemble, Theorem 3 also recovers some results independently obtained in past work by Reeves [47], and concurrent work by Fletcher et al. [30] and Aeron et al. [3].

In addition, we study the effect of measurement sparsity, which we assess in terms of the fraction $\gamma \in (0, 1]$ of non-zeros per row of the the measurement matrix X. In the noiseless setting, a growing body of work has examined computationally efficient recovery methods based on sparse measurement matrices, including work inspired by expander graphs and coding theory [53, 61, 10], as well as dimension-reducing embeddings and sketching [18, 33, 60]. In addition, some results have been shown to be stable in the ℓ_2 or ℓ_1 norm in the presence of noise [18, 10]; note however that ℓ_2/ℓ_1 stability does not guarantee exact recovery of the support set. In the noisy setting, the paper [2] provides results for sparse measurements and distortion-type error metrics, using a power-based SNR that is not appropriate for the subset recovery problem. For the noisy observation model (3.2), some concurrent work [44] provides sufficient conditions for support recovery using the Lasso (i.e. ℓ_1-constrained quadratic programming) for appropriately sparsified ensembles. These results can be viewed as complementary to the information-theoretic analysis presented here, in which we characterize the inherent trade-off between measurement sparsity and statistical efficiency. More specifically, our second main result (Theorem 4, below) provides necessary conditions for exact support recovery using γ-sparsified Gaussian measurement matrices (see equation (3.7)), for general scalings of the parameters $(n, p, k, \lambda, \gamma)$. This analysis reveals three regimes of interest, corresponding to whether measurement sparsity has no effect, a small effect, or a significant effect on the number of measurements necessary for recovery. Thus, there exist regimes in which measurement sparsity fundamentally alters the ability of any method to decode.

3.3 Necessary conditions for sparse recovery

In this section, we state our main results, and discuss some of their consequences. Our analysis applies to random ensembles of measurement matrices $X \in \mathbb{R}^{n \times p}$, where each entry X_{ij} is drawn i.i.d. from some underlying distribution. The most commonly studied random ensemble is the standard Gaussian case, in which each $X_{ij} \sim N(0, 1)$. Note that this choice generates a highly dense measurement matrix X, with np non-zero entries. Our first result (Theorem 3) applies to more general ensembles that satisfy the moment conditions $\mathbb{E}[X_{ij}] = 0$ and $\text{var}(X_{ij}) = 1$, which allows for a variety of non-Gaussian distributions (e.g., uniform, Bernoulli etc.). In addition, we also derive results (Theorem 4) for γ-sparsified matrices X, in which each entry X_{ij}

is i.i.d. drawn according to

$$X_{ij} = \begin{cases} N(0, \frac{1}{\gamma}) & \text{w.p. } \gamma \\ 0 & \text{w.p. } 1-\gamma \end{cases}. \tag{3.7}$$

Note that when $\gamma = 1$, the distribution in (3.7) is exactly the standard Gaussian ensemble. We refer to the sparsification parameter $\gamma \in (0,1]$ as the *measurement sparsity.* Our analysis allows this parameter to vary as a function of (n, p, k).

3.3.1 Bounds on dense ensembles

We begin by stating a set of necessary conditions on (n, p, k, λ) for asymptotically reliable recovery with any method, which apply to general ensembles of zero-mean and unit-variance measurement matrices. In addition to the standard Gaussian ensemble ($X_{ij} \sim N(0,1)$), this result also covers matrices from other common ensembles (e.g., Bernoulli $X_{ij} \in \{-1, +1\}$). Furthermore, our analysis can be extended to matrices with independent rows drawn from any distribution with zero mean and covariance matrix Σ.

Theorem 3 (General ensembles)**.** *Let the measurement matrix $X \in \mathbb{R}^{n \times p}$ be drawn with i.i.d. elements from any distribution with zero mean and unit variance. Then a necessary condition for asymptotically reliable recovery over the signal class $\mathcal{C}_{p,k}(\lambda)$ is*

$$n \quad > \quad \max\big\{f_1(p,k,\lambda), \quad \ldots, \quad f_k(p,k,\lambda), \quad k\big\}, \tag{3.8}$$

where

$$f_m(p,k,\lambda) \quad := \quad \frac{\log\binom{p-k+m}{m} - 1}{\frac{1}{2}\log\Big(1 + m\lambda^2\big(1 - \frac{m}{p-k+m}\big)\Big)} \tag{3.9}$$

for $m = 1, \ldots, k$.

The proof of Theorem 3, given in Section 3.5, uses Fano's inequality [19] to bound the probability of error in a restricted ensemble, which can then be viewed as a type of channel coding problem. Moreover, the proof constructs a family of restricted ensembles that sweeps the range of possible overlaps between subsets, and tries to capture the difficulty of distinguishing between subsets at various distances.

We now consider some consequences of the necessary conditions in Theorem 3 under two scalings of the signal sparsity: the regime of linear signal sparsity, in which $k/p = \alpha$ for some $\alpha \in (0, 1)$, and the regime of sublinear signal sparsity, meaning $k/p \to 0$. In particular, the necessary conditions in Theorem 3 can be compared against the sufficient conditions in Wainwright [59] for exact support recovery using the standard Gaussian ensemble, as shown in Table 3.1. This comparison reveals that Theorem 3 generalizes and strengthens earlier results on necessary conditions for subset recovery [59]. We obtain tight scalings of the necessary and sufficient conditions in the regime of linear signal sparsity (meaning $k/p = \alpha$), under various scalings of the minimum value λ (shown in the first three rows of Table 3.1). We also obtain tight scaling conditions in the regime of sublinear signal sparsity (in which $k/p \to 0$), when $k\lambda^2 = \Theta(1)$ (as shown in row 4 of Table 3.1). There remains a slight gap, however, in the sublinear sparsity regime when $k\lambda^2 \to \infty$ (see bottom two rows in Table 3.1).

In the regime of linear sparsity, Wainwright [59] showed, by direct analysis of the optimal decoder, that the scaling $\lambda^2 = \Omega(\log(k)/k)$ is sufficient for exact support

	Necessary conditions (Theorem 3)	Sufficient conditions (Wainwright [59])
$k = \Theta(p)$ $\lambda^2 = \Theta(\frac{1}{k})$	$\Theta(p \log p)$	$\Theta(p \log p)$
$k = \Theta(p)$ $\lambda^2 = \Theta(\frac{\log k}{k})$	$\Theta(p)$	$\Theta(p)$
$k = \Theta(p)$ $\lambda^2 = \Theta(1)$	$\Theta(p)$	$\Theta(p)$
$k = o(p)$ $\lambda^2 = \Theta(\frac{1}{k})$	$\Theta(k \log(p-k))$	$\Theta(k \log(p-k))$
$k = o(p)$ $\lambda^2 = \Theta(\frac{\log k}{k})$	$\max\left\{\Theta\left(\frac{k \log \frac{p}{k}}{\log \log k}\right), \Theta\left(\frac{k \log(p-k)}{\log k}\right)\right\}$	$\Theta\left(k \log \frac{p}{k}\right)$
$k = o(p)$ $\lambda^2 = \Theta(1)$	$\max\left\{\Theta\left(\frac{k \log \frac{p}{k}}{\log k}\right), \Theta(k)\right\}$	$\Theta\left(k \log \frac{p}{k}\right)$

Table 3.1. Tight scalings of the necessary and sufficient conditions on the number of observations n required for exact support recovery are obtained in several regimes of interest.

recovery using a linear fraction $n = \Theta(p)$ of observations. Combined with the necessary condition in Theorem 3, we obtain the following corollary that provides a sharp characterization of the linear-linear regime:

Corollary 1. *Consider the regime of linear sparsity, meaning that $k/p = \alpha \in (0, 1)$, and suppose that a linear fraction $n = \Theta(p)$ of observations are made. Then the optimal decoder can recover the support exactly if and only if $\lambda^2 = \Omega(\log k/k)$.*

Theorem 3 has some consequences related to results proved in recent and concurrent work. Reeves and Gastpar [47] have shown that in the regime of linear sparsity $k/p = \alpha > 0$, and for standard Gaussian measurements, if any decoder is given only a linear fraction sample size (meaning that $n = \Theta(p)$), then one must have $k\lambda^2 \to +\infty$ in order to recover the support exactly. This result is one corollary of Theorem 3, since if $\lambda^2 = \Theta(1/k)$, then we have

$$n \;>\; \frac{\log(p-k+1)-1}{\frac{1}{2}\log(1+\Theta(1/k))} \;=\; \Omega(k\log(p-k)) \;\gg \Theta(p),$$

so that the scaling $n = \Theta(p)$ is precluded. In concurrent work, Fletcher et al. [30] used direct methods to show that for the special case of the standard Gaussian ensemble, the number of observations must satisfy $n > \Omega\big(\frac{\log(p-k)}{\lambda^2}\big)$. The qualitative form of this bound follows from our lower bound $f_1(p,k,\lambda)$, which holds for standard Gaussian ensembles as well as more general (non-Gaussian) ensembles. However, we note that the direct methods used by Fletcher et al. [30] yield better control of the constant pre-factors for the standard Gaussian ensemble. Similarly, concurrent work by Aeron et al. [3] showed that in the regime of linear sparsity (i.e., $k = \Theta(p)$) and for standard Gaussian measurements, the number of observations must satisfy $n > \Omega\left(\frac{\log p}{\lambda^2}\right)$. This result also follows as a consequence of our lower bound $f_1(p,k,\lambda)$.

The results in Theorem 3 can also be compared to an intuitive bound based on classical channel capacity results, as pointed out previously by various researchers (e.g., [52, 5]). Consider a restricted problem, in which the values associated with each possible sparsity pattern on β are fixed and known at the decoder. Then support recovery can be viewed as a type of channel coding problem, in which the $N = \binom{p}{k}$ possible support sets of β correspond to messages to be sent over a Gaussian channel. Suppose each support set S is encoded as the codeword $X\beta$, where X has i.i.d. Gaussian entries. The effective code rate is then $R = \frac{\log\binom{p}{k}}{n}$, and by standard Gaussian channel capacity results, we have the lower bound,

$$n \;>\; \frac{\log\binom{p}{k}}{\frac{1}{2}\log\left(1+\|\beta\|_2^2\right)}. \tag{3.10}$$

This bound is tight for $k = 1$ and Gaussian measurements, but loose in general. As Theorem 3 clarifies, there are additional elements in the support recovery problem that distinguish it from a standard Gaussian coding problem: first, the signal power $\|\beta\|_2^2$ does not capture the inherent problem difficulty for $k > 1$, and second, there is overlap between support sets for $k > 1$. Note that $\|\beta\|_2^2 \geq k\lambda^2$ (with equality in the case when $|\beta_j| = \lambda$ for all indices $j \in S$), so that Theorem 3 is strictly tighter than the intuitive bound (3.10). Moreover, by fixing the value of β at $(k-1)$ indices to λ and allowing the last component of β to tend to infinity, we can drive the power $\|\beta\|_2^2$ to infinity, while still having the minimum λ enter the lower bound.

3.3.2 Effect of measurement sparsity

We now turn to the effect of measurement sparsity on subset recovery, considering in particular the γ-sparsified ensemble (3.7). Since each X_{ij} has zero mean and unit variance for all choices of γ by construction, Theorem 3 applies to the γ-sparsified Gaussian ensemble (3.7); however, it yields necessary conditions that are independent of γ. Intuitively, it is clear that the procedure of γ-sparsification should cause deterioration in support recovery. Indeed, the following result provides more refined bounds that capture the effects of γ-sparsification. We first state a set of necessary conditions on $(n, p, k, \lambda, \gamma)$ in general form, and subsequently bound these conditions in different regimes of sparsity. Let $\phi(\mu, \sigma^2)$ denote the Gaussian density with mean μ and variance σ^2, and define the family of mixture distributions $\{\overline{\psi}_m\}_{m=1,\ldots,k}$ with

$$\overline{\psi}_m \;:=\; \sum_{\ell=0}^{m} \binom{m}{\ell} \gamma^{\ell}(1-\gamma)^{m-\ell}\, \phi\left(0, 1 + \frac{\ell\lambda^2}{\gamma}\right). \tag{3.11}$$

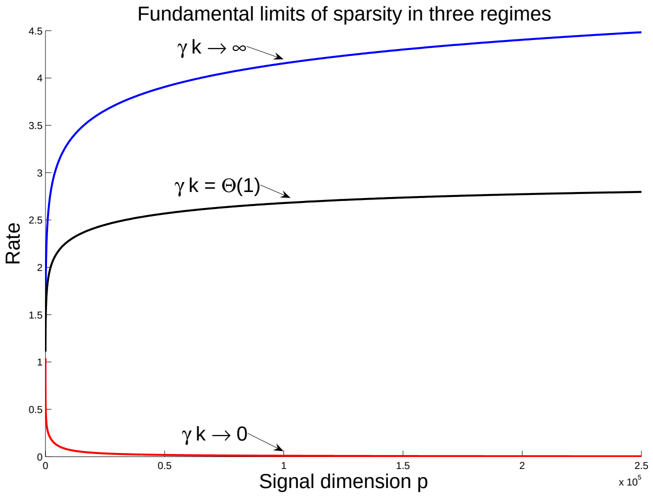


Figure 3.1. The rate $R = \frac{\log \binom{p}{k}}{n}$, defined as the logarithm of the number of possible subsets the decoder can reliably estimate based on n observations, is plotted using equation (3.12) in three regimes, depending on how the quantity γk scales. In particular, γk corresponds to the average number of non-zeros in β that align with the non-zeros in each row of the measurement matrix.

Furthermore, let $h(\cdot)$ denote the differential entropy functional. With this notation, we have the following result.

Theorem 4 (Sparse ensembles)**.** *Let the measurement matrix $X \in \mathbb{R}^{n \times p}$ be drawn with i.i.d. elements from the γ-sparsified Gaussian ensemble* (3.7)*. Then a necessary condition for asymptotically reliable recovery over the signal class $\mathcal{C}_{p,k}(\lambda)$ is*

$$n \quad > \quad \max\left\{g_1(p,k,\lambda,\gamma), \quad \ldots, \quad g_k(p,k,\lambda,\gamma), \quad k\right\}, \tag{3.12}$$

where

$$g_m(p,k,\lambda,\gamma) \quad := \quad \frac{\log\binom{p-k+m}{m} - 1}{h(\overline{\psi}_m) - \frac{1}{2}\log(2\pi e)} \tag{3.13}$$

for $m = 1, \ldots, k$.

The proof of Theorem 4, given in Section 3.6, again uses Fano's inequality, but explicitly analyzes the effect of measurement sparsification on the entropy of the observations. The necessary condition in Theorem 4 is plotted in Figure 3.1, showing distinct regimes of behavior depending on how the quantity γk scales, where $\gamma \in (0, 1]$ is the measurement sparsification parameter and k is the signal sparsity index. In order to characterize the regimes in which measurement sparsity begins to degrade the recovery performance of any decoder, Corollary 2 below further bounds the necessary conditions in Theorem 4 in three cases. For any scalar γ, let $H_{binary}(\gamma)$ denote the entropy of a Bernoulli(γ) variate.

Corollary 2 (Three regimes)**.** *The necessary conditions in Theorem 4 can be simplified as follows.*

(a) If $\gamma m \to \infty$, then

$$g_m(p, k, \lambda, \gamma) \;\; \geq \;\; \frac{\log \binom{p-k+m}{m} - 1}{\frac{1}{2}\log\left(1 + m\lambda^2\right)}. \tag{3.14a}$$

(b) If $\gamma m = \tau$ for some constant τ, then

$$g_m(p, k, \lambda, \gamma) \;\; \geq \;\; \frac{\log \binom{p-k+m}{m} - 1}{\frac{1}{2}\tau\log\left(1 + \frac{m\lambda^2}{\tau}\right) + C}, \tag{3.14b}$$

where $C = \frac{1}{2}\log(2\pi \mathrm{e}(\tau + \frac{1}{12}))$ is a constant.

(c) If $\gamma m \to 0$, then

$$g_m(p, k, \lambda, \gamma) \;\; \geq \;\; \frac{\log \binom{p-k+m}{m} - 1}{\frac{1}{2}\gamma m \log\left(1 + \frac{\lambda^2}{\gamma}\right) + m H_{binary}(\gamma)}. \tag{3.14c}$$

Necessary conditions (Theorem 4)	$k = o(p)$	$k = \Theta(p)$
$\lambda^2 = \Theta(\frac{1}{k})$ $\gamma = o(\frac{1}{k \log k})$	$\Theta\left(\frac{k \log(p-k)}{\gamma k \log \frac{1}{\gamma}}\right)$	$\Theta\left(\frac{p \log p}{\gamma p \log \frac{1}{\gamma}}\right)$
$\lambda^2 = \Theta(\frac{1}{k})$ $\gamma = \Omega(\frac{1}{k \log k})$	$\Theta(k \log(p-k))$	$\Theta(p \log p)$
$\lambda^2 = \Theta(\frac{\log k}{k})$ $\gamma = o(\frac{1}{k \log k})$	$\Theta\left(\frac{k \log(p-k)}{\gamma k \log \frac{1}{\gamma}}\right)$	$\Theta\left(\frac{p \log p}{\gamma p \log \frac{1}{\gamma}}\right)$
$\lambda^2 = \Theta(\frac{\log k}{k})$ $\gamma = \Theta(\frac{1}{k \log k})$	$\Theta(k \log(p-k))$	$\Theta(p \log p)$
$\lambda^2 = \Theta(\frac{\log k}{k})$ $\gamma = \Omega(\frac{1}{k})$	$\max \left\{ \Theta\left(\frac{k \log \frac{p}{k}}{\log \log k}\right), \Theta\left(\frac{k \log(p-k)}{\log k}\right) \right\}$	$\Theta(p)$

Table 3.2. Necessary conditions on the number of observations n required for exact support recovery is shown in different regimes of the parameters (p, k, λ, γ).

Corollary 2 reveals three regimes of behavior, defined by the scaling of the measurement sparsity γ and the signal sparsity k. Intuitively, γk is the average number of non-zeros in β that align with the non-zeros in each row of the measurement matrix. If $\gamma k \to \infty$ as $p \to \infty$, then the recovery threshold (3.14a) is of the same order as the threshold for dense measurement ensembles. In this regime, sparsifying the measurement ensemble has no asymptotic effect on performance. In sharp contrast, if $\gamma k \to 0$ sufficiently fast as $p \to \infty$, then the denominator in (3.14c) goes to zero, and the recovery threshold changes fundamentally compared to the dense case. Hence, the number of measurements that any decoder needs in order to reliably recover increases dramatically in this regime. Finally, if $\gamma k = \Theta(1)$, then the recovery threshold (3.14b) transitions between the two extremes. Using the bounds in Corollary 2, the necessary conditions in Theorem 4 are shown in Table 3.2 under different scalings of the parameters $(n, p, k, \lambda, \gamma)$. In particular, if $\gamma = o(\frac{1}{k \log k})$ and the minimum value λ^2 does not increase with k, then the denominator $\gamma k \log \frac{1}{\gamma}$ goes to zero.

3.4 Fano's method

In this section, we describe a general framework for deriving necessary conditions, which sets the stage for the proofs of Theorems 3 and 4 in later sections. Establishing necessary conditions for exact sparsity recovery amounts to finding conditions on (n, p, k, λ) (and possibly γ) under which the probability of error of any recovery method stays bounded away from zero as $n \to \infty$. At a high-level, our general approach is quite simple: we consider restricted problems in which the decoder has been given some additional side information, and then apply Fano's inequality [19] to lower bound the probability of error. In order to establish the collection of necessary conditions (e.g., $\big\{f_1(p,k,\lambda), \ldots, f_k(p,k,\lambda)\big\}$), we construct a family of restricted ensembles which sweeps the range of possible overlaps between support sets. At the extremes of this family are two classes of ensembles: one which captures the bulk effect of having many competing subsets at large distances, and the other which captures the effect of a smaller number of subsets at very close distances (this is illustrated in Figure (3.2a)). Accordingly, we consider the family of ensembles $\big\{\widetilde{\mathcal{C}}_{p-k+m,m}(\lambda)\big\}_{m=1,\ldots,k}$, where the m^{th} restricted ensemble is defined as follows.

We use the notation $X_j \in \mathbb{R}^n$ to denote column j of the matrix X, and $X_U \in \mathbb{R}^{n \times |U|}$ to denote the submatrix containing columns indexed by set U. Similarly, let $\beta_U \in \mathbb{R}^{|U|}$ denote the subvector of β corresponding to the index set U. In addition, let $H(\cdot)$ and $h(\cdot)$ denote the entropy and differential entropy functionals, respectively.

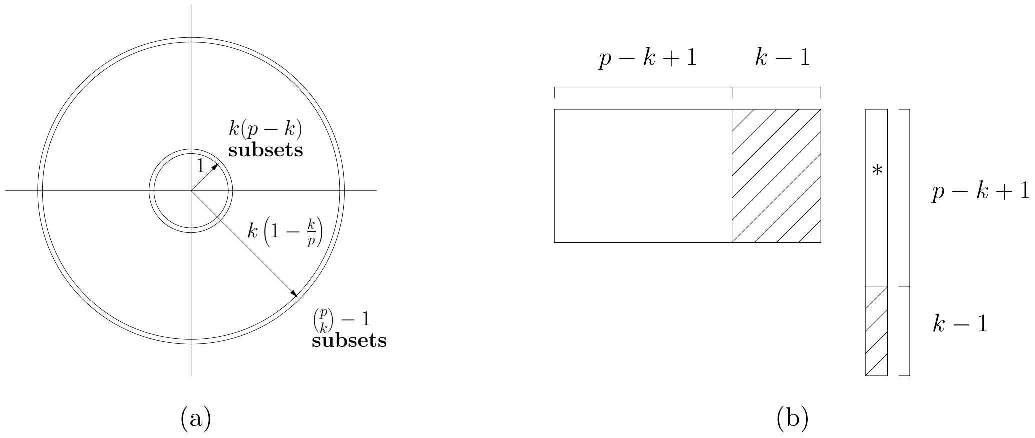


Figure 3.2. Illustration of restricted ensembles. (a) In restricted ensemble $\widetilde{\mathcal{C}}_{p,k}(\lambda)$, the decoder must distinguish between $\binom{p}{k}$ support sets with an average overlap of size $\frac{k^2}{p}$, whereas in restricted ensemble $\widetilde{\mathcal{C}}_{p-k+1,1}(\lambda)$, it must decode amongst a subset of the $k(p-k)+1$ supports with overlap $k-1$. (b) In restricted ensemble $\widetilde{\mathcal{C}}_{p-k+1,1}(\lambda)$, the decoder is given the locations of the $k-1$ largest non-zeros, and it must estimate the location of the smallest non-zero from the $p-k+1$ remaining possible indices.

3.4.1 Constructing restricted ensembles

Suppose that the decoder is given the locations of all but the m smallest non-zero values of the vector β, as well as the values of β on its support. More precisely, let S represent the true underlying support of β and let T denote the set of revealed indices, which has size $|T| = k - m$. Let $U = S \setminus T$ denote the set of unknown locations, and assume that $\beta_j = \lambda$ for all $j \in U$. Given knowledge of (T, β_T, λ), the decoder may simply subtract $X_T\beta_T = \sum_{j\in T} X_j\beta_j$ from Y, so that it is left with the modified n-vector of observations

$$\widetilde{Y} \;:=\; \sum_{j\in U} X_j\lambda + W. \tag{3.15}$$

By re-ordering indices as need be, we may assume without loss of generality that $T = \{p-k+m+1, \ldots, p\}$, so that $U \subset \{1, \ldots, p-k+m\}$. The remaining subproblem is to determine, given the observations $\widetilde{Y}$, the locations of the m non-zeros in U. Note that when we assume the support of β is uniformly chosen over all $\binom{p}{k}$ possible subsets of size k, then given T, the remaining subset U is uniformly distributed over the $\binom{p-k+m}{m}$ possible subsets of size m.

We will now argue that analyzing the probability of error of this restricted problem gives us a lower bound on the probability of error in the original problem. Consider the restricted signal class $\widetilde{\mathcal{C}}_{p-k+m,m}(\lambda)$ defined as

$$\widetilde{\mathcal{C}}_{p-k+m,m}(\lambda) \quad := \quad \left\{ \widetilde{\beta} \in \mathbb{R}^{p-k+m} \;\middle|\; |U(\widetilde{\beta})| = m, \; \widetilde{\beta}_j = \lambda \; \forall j \in U(\widetilde{\beta}) \right\} \tag{3.16}$$

where we denote the support set of vector $\widetilde{\beta}$ as $U(\widetilde{\beta}) := \{ j \mid \widetilde{\beta}_j \neq 0 \}$. For any $\widetilde{\beta} \in \widetilde{\mathcal{C}}_{p-k+m,m}(\lambda)$, we can concatenate $\widetilde{\beta}$ with a vector v of $k-m$ non-zeros (with $\min_j |v_j| \geq \lambda$) at the end to obtain a p-dimensional vector. If a decoder can recover the support of any p-dimensional k-sparse vector $\beta \in \mathcal{C}_{p,k}(\lambda)$, then it can recover the support of the augmented $\widetilde{\beta}$, and hence the support of $\widetilde{\beta}$. Furthermore, providing the decoder with the non-zero values of β cannot increase the probability of error. Thus, we can apply Fano's inequality to lower bound the probability of error in the restricted problem, and so obtain a lower bound on the probability of error for the general problem.

3.4.2 Applying Fano to restricted ensembles

Consider the class of signals $\widetilde{\mathcal{C}}_{p-k+m,m}(\lambda)$ defined in (3.16), which consists of $M = \binom{p-k+m}{m}$ models $\{\widetilde{\beta}^{(1)}, \ldots, \widetilde{\beta}^{(M)}\}$ corresponding to the M possible subsets $U \subset \{1, \ldots, p-k+m\}$ of size k. Suppose that a model index θ is chosen uniformly at random from $\{1, \ldots, M\}$, and we sample n observations $\widetilde{Y} \in \mathbb{R}^n$ via the measurement matrix $\widetilde{X} \in \mathbb{R}^{n \times (p-k+m)}$. For any decoding function $f : \mathbb{R}^n \to \{1, \ldots, M\}$, the average probability of error is defined as

$$p_{err}(f) \;=\; \frac{1}{M} \sum_{i=1}^{M} \mathbb{P}\big[f(\widetilde{Y}) \neq i \;\mid\; \theta = i\big],$$

while the maximal probability of error over the class $\widetilde{\mathcal{C}}_{p-k+m,m}(\lambda)$ is defined as

$$\omega(f) \;=\; \max_{i=1,\ldots,M} \mathbb{P}\big[f(\widetilde{Y}) \neq i \;\mid\; \theta = i\big].$$

We first apply Fano's lemma [19] to bound the error probability over $\widetilde{\mathcal{C}}_{p-k+m,m}(\lambda)$ for a particular instance of the random measurement matrix $\widetilde{X}$, and subsequently average over the ensemble of matrices. Thus by Fano's inequality, the average probability of error, and hence also the maximal probability of error, is lower bounded as

$$p_{err}(f) \;\geq\; \frac{H\big(\theta \big| \widetilde{Y}, \widetilde{X}\big) - 1}{\log M} \;=\; 1 - \frac{I\big(\theta; \widetilde{Y} \big| \widetilde{X}\big) + 1}{\log M}. \tag{3.17}$$

Consequently, the problem of establishing necessary conditions for asymptotically reliable recovery is reduced to obtaining upper bounds on the conditional mutual information $I\big(\theta; \widetilde{Y} \big| \widetilde{X}\big)$.

3.5 Analysis of general measurement ensembles

In this section, we derive the necessary conditions stated in Theorem 3 for the general class of measurement matrices, by applying Fano's inequality to bound the probability of decoding error in each of the k restricted ensembles in the family $\big\{\widetilde{\mathcal{C}}_{p-k+m,m}(\lambda)\big\}_{m=1,\ldots,k}$.

We begin by performing our analysis of the error probability over $\widetilde{\mathcal{C}}_{p-k+m,m}(\lambda)$ for any $m \in \{1, \ldots, k\}$. Let $\widetilde{X} \in \mathbb{R}^{n\times(p-k+m)}$ be a matrix with independent, zero-mean and unit-variance entries. Conditioned on the event that U is the true underlying support of $\widetilde{\beta}$, the vector of n observations can be written as

$$\widetilde{Y} \;:=\; \widetilde{X}_U\widetilde{\beta}_U + W \;=\; \lambda \sum_{j\in U} \widetilde{X}_j + W.$$

Accordingly, the conditional mutual information in equation (4.15) can be expanded as

$$I\big(\theta; \widetilde{Y}\big|\widetilde{X}\big) \;=\; h\big(\widetilde{Y}\big|\widetilde{X}\big) - h\big(\widetilde{Y}\big|\theta, \widetilde{X}\big) \;=\; h\big(\widetilde{Y}\big|\widetilde{X}\big) - h(W).$$

We bound the first term using the fact that the differential entropy of the observation vector $\widetilde{Y}$ for a particular instance of matrix $\widetilde{X}$ is maximized by the Gaussian distribution with a matched variance. More specifically, for a fixed $\widetilde{X}$, the distribution of $\widetilde{Y}$ is a Gaussian mixture with density $\psi\big(y \,\big|\, \widetilde{X}\big) = \frac{1}{\binom{p-k+m}{m}} \sum_U \phi\big(\widetilde{X}_U\widetilde{\beta}_U, I\big)$, where we are using ϕ to denote the density of a Gaussian random vector with mean $\widetilde{X}_U\widetilde{\beta}_U$ and covariance I. Let $\Lambda\big(\widetilde{X}\big)$ denote the covariance matrix of $\widetilde{Y}$ conditioned on $\widetilde{X}$. (Hence entry $\Lambda_{ii}\big(\widetilde{X}\big)$ on the diagonal represents the variance of $\widetilde{Y}_i$ given $\widetilde{X}$.) With this notation, the entropy associated with the marginal density $\psi\big(y_i \,\big|\, \widetilde{X}\big)$ is upper bounded by $\frac{1}{2}\log\big(2\pi\mathrm{e}\,\Lambda_{ii}\big(\widetilde{X}\big)\big)$. When $\widetilde{X}$ is randomly chosen, the conditional entropy of $\widetilde{Y}$ given $\widetilde{X}$ (averaged over the choice of $\widetilde{X}$) can be bounded as

$$\begin{aligned} h\big(\widetilde{Y}\big|\widetilde{X}\big) \;&\leq\; \sum_{i=1}^{n} h\big(\widetilde{Y}_i\big|\widetilde{X}\big) \\ &\leq\; \sum_{i=1}^{n} \mathbb{E}_{\widetilde{X}}\left[\frac{1}{2}\log\big(2\pi\mathrm{e}\,\Lambda_{ii}\big(\widetilde{X}\big)\big)\right]. \end{aligned}$$

The conditional entropy can be further bounded by exploiting the concavity of the logarithm and applying Jensen's inequality, as

$$h\big(\widetilde{Y}\big|\widetilde{X}\big) \;\leq\; \sum_{i=1}^{n} \frac{1}{2} \log\Big(2\pi \mathrm{e}\, \mathbb{E}_{\widetilde{X}}\big[\Lambda_{ii}\big(\widetilde{X}\big)\big]\Big).$$

Next, the entropy of the Gaussian noise vector $W \sim N(0, I_{n\times n})$ can be computed as $h(W) = \frac{n}{2}\log(2\pi \mathrm{e})$. Combining these two terms, we then obtain the following bound on the conditional mutual information,

$$I\big(\theta; \widetilde{Y}\big|\widetilde{X}\big) \;\leq\; \sum_{i=1}^{n} \frac{1}{2} \log\Big(\mathbb{E}_{\widetilde{X}}\big[\Lambda_{ii}\big(\widetilde{X}\big)\big]\Big).$$

It remains to compute the expectation $\mathbb{E}_{\widetilde{X}}\big[\Lambda_{ii}\big(\widetilde{X}\big)\big]$, over the ensemble of matrices $\widetilde{X}$ drawn with i.i.d. entries from any distribution with zero mean and unit variance. The proof of the following lemma involves some relatively straightforward but lengthy calculation, and is given in Section 3.5.1.

Lemma 3. *Given i.i.d. $\widetilde{X}_{ij}$ with zero mean and unit variance, the averaged covariance matrix of $\widetilde{Y}$ given $\widetilde{X}$ is*

$$\mathbb{E}_{\widetilde{X}}\big[\Lambda\big(\widetilde{X}\big)\big] \;=\; \left(1 + m\lambda^2\left(1 - \frac{m}{p-k+m}\right)\right) I_{n\times n}. \tag{3.18}$$

Finally, combining Lemma 3 with equation (4.15), we obtain that the average probability of error is bounded away from zero if

$$n \;<\; \frac{\log\binom{p-k+m}{m} - 1}{\frac{1}{2}\log\left(1 + m\lambda^2\left(1 - \frac{m}{p-k+m}\right)\right)},$$

as claimed.

3.5.1 Averaging over the ensemble

We now derive Lemma 3. We begin by defining some additional notation. Recall that for a given instance of the matrix $\widetilde{X}$, the observation vector $\widetilde{Y}$ has a Gaussian mixture distribution with density $\psi\big(y \,\big|\, \widetilde{X}\big) = \frac{1}{\binom{p-k+m}{m}} \sum_U \phi\big(\widetilde{X}_U \widetilde{\beta}_U, I\big)$, where ϕ denotes the Gaussian density with mean $\widetilde{X}_U\widetilde{\beta}_U$ and covariance I. Let $\mu\big(\widetilde{X}\big) = \mathbb{E}\big[\widetilde{Y}\big|\widetilde{X}\big] \in \mathbb{R}^n$ and $\Lambda\big(\widetilde{X}\big) = \mathbb{E}\big[\widetilde{Y}\widetilde{Y}^T\big|\widetilde{X}\big] - \mu\big(\widetilde{X}\big)\mu\big(\widetilde{X}\big)^T \in \mathbb{R}^{n\times n}$ be the mean vector and covariance matrix of $\widetilde{Y}$ given $\widetilde{X}$, respectively. Accordingly we have

$$\mu\big(\widetilde{X}\big) \;=\; \frac{1}{\binom{p-k+m}{m}} \sum_U \widetilde{X}_U \widetilde{\beta}_U$$

and

$$\mathbb{E}\big[\widetilde{Y}\widetilde{Y}^T\big|\widetilde{X}\big] \;=\; \frac{1}{\binom{p-k+m}{m}} \sum_U \big(\widetilde{X}_U\widetilde{\beta}_U\big)\big(\widetilde{X}_U\widetilde{\beta}_U\big)^T \;+\; I.$$

With this notation, we can now compute the expectation of the covariance matrix $\mathbb{E}_{\widetilde{X}}\big[\Lambda\big(\widetilde{X}\big)\big]$, averaged over any distribution on $\widetilde{X}$ with independent, zero-mean and unit-variance entries. To compute the first term, we have

$$\begin{aligned}
\mathbb{E}_{\widetilde{X}}\Big[\mathbb{E}\big[\widetilde{Y}\widetilde{Y}^T\big|\widetilde{X}\big]\Big] \;&=\; \frac{\lambda^2}{\binom{p-k+m}{m}} \sum_U \mathbb{E}_{\widetilde{X}}\Bigg[\sum_{j\in U} \widetilde{X}_j\widetilde{X}_j^T + \sum_{i\neq j\in U} \widetilde{X}_i\widetilde{X}_j^T\Bigg] \;+\; I \\
&=\; \frac{\lambda^2}{\binom{p-k+m}{m}} \sum_U \sum_{j\in U} I \;+\; I \\
&=\; \big(1 + m\lambda^2\big)\; I
\end{aligned}$$

where the second equality uses the fact that $\mathbb{E}_{\widetilde{X}}\big[\widetilde{X}_j\widetilde{X}_j^T\big] = I$, and $\mathbb{E}_{\widetilde{X}}\big[\widetilde{X}_i\widetilde{X}_j^T\big] = 0$ for

$i \neq j$. Next, we compute the second term as,

$$\begin{aligned}
\mathbb{E}_{\widetilde{X}}\left[\mu(\widetilde{X})\mu(\widetilde{X})^T\right] &= \left(\frac{\lambda}{\binom{p-k+m}{m}}\right)^2 \mathbb{E}_{\widetilde{X}}\left[\sum_{U,V}\sum_{j\in U\cap V}\widetilde{X}_j\widetilde{X}_j^T + \sum_{U,V}\sum_{\substack{i\in U, j\in V\\ i\neq j}}\widetilde{X}_i\widetilde{X}_j^T\right] \\
&= \left(\frac{\lambda}{\binom{p-k+m}{m}}\right)^2 \sum_{U,V}\sum_{j\in U\cap V} I \\
&= \left(\left(\frac{\lambda}{\binom{p-k+m}{m}}\right)^2 \sum_{U,V}|U\cap V|\right) I.
\end{aligned}$$

From here, note that there are $\binom{p-k+m}{m}$ possible subsets U of size m. For each U, a counting argument reveals that there are $\binom{m}{\delta}\binom{p-k}{m-\delta}$ subsets V of size m which have $|U\cap V| = \delta$ overlaps with U. Thus the scalar multiplicative factor above can be written as

$$\left(\frac{\lambda}{\binom{p-k+m}{m}}\right)^2 \sum_{U,V}|U\cap V| = \frac{\lambda^2}{\binom{p-k+m}{m}}\sum_{\delta=1}^{m}\binom{m}{\delta}\binom{p-k}{m-\delta}\delta.$$

Finally, using a substitution of variables (by setting $\delta' = \delta - 1$) and applying Vandermonde's identity [48], we have

$$\begin{aligned}
\left(\frac{\lambda}{\binom{p-k+m}{m}}\right)^2 \sum_{U,V}|U\cap V| &= \frac{\lambda^2}{\binom{p-k+m}{m}}\, m\sum_{\delta'=0}^{m-1}\binom{m-1}{\delta'}\binom{p-k}{m-\delta'-1} \\
&= \frac{\lambda^2}{\binom{p-k+m}{m}}\, m\binom{p-k+m-1}{m-1} \\
&= \frac{m^2\lambda^2}{p-k+m}.
\end{aligned}$$

Combining these terms, we conclude that

$$\mathbb{E}_{\widetilde{X}}[\Lambda(\widetilde{X})] = \left(1 + m\lambda^2\left(1 - \frac{m}{p-k+m}\right)\right) I.$$

3.6 Analysis of sparse measurement ensembles

This section contains proofs of the necessary conditions in Theorem 4 for the γ-sparsified Gaussian measurement ensemble (3.7). We proceed as before, applying Fano's inequality to each restricted class in the family $\{\widetilde{\mathcal{C}}_{p-k+m,m}(\lambda)\}_{m=1,\ldots,k}$, in order to derive the corresponding k conditions in Theorem 4.

In analyzing the probability of error over $\widetilde{\mathcal{C}}_{p-k+m,m}(\lambda)$, the initial steps proceed as in the proof of Theorem 3, by expanding the conditional mutual information in equation (4.15) as

$$\begin{aligned} I\big(\theta;\widetilde{Y}\big|\widetilde{X}\big) &= h\big(\widetilde{Y}\big|\widetilde{X}\big) - h(W) \\ &\leq \sum_{i=1}^{n} h\big(\widetilde{Y}_i\big|\widetilde{X}\big) - \frac{n}{2}\log(2\pi\mathrm{e}), \end{aligned}$$

using the Gaussian entropy for $W \sim N(0, I_{n\times n})$.

From this point, the key subproblem is to compute the conditional entropy of $\widetilde{Y}_i = \lambda\sum_{j\in U(\widetilde{\beta})}\widetilde{X}_{ij} + W_i$, when the support of $\widetilde{\beta}$ is uniformly chosen over all $\binom{p-k+m}{m}$ possible subsets of size m. To characterize the limiting behavior of the random variable $\widetilde{Y}_i$, note that for a fixed matrix $\widetilde{X}$, each $\widetilde{Y}_i$ is distributed according to the density defined as

$$\psi_m\big(y_i \,\big|\, \widetilde{X}\big) \;=\; \frac{1}{\binom{p-k+m}{m}}\sum_{U}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\Big(y_i - \lambda\sum_{j\in U}\widetilde{X}_{ij}\Big)^2\right).$$

This density is a mixture of Gaussians with unit variances and means that depend on the values of $\{\widetilde{X}_{i1},\ldots,\widetilde{X}_{i(p-k+m)}\}$, summed over subsets $U \subset \{1,\ldots,p-k+m\}$ with $|U| = m$. At a high-level, our immediate goal is to characterize the entropy $h(\psi_m)$.

Note that as $\widetilde{X}$ varies over the sparse ensemble (3.7), the sequence $\big\{\psi_m\big(y_i \,\big|\, \widetilde{X}\big)\big\}_p$, indexed by the signal dimension p, is actually a sequence of random densities. As an intermediate step, the following lemma characterizes the average pointwise behavior of this random sequence of densities, and is proven in Section 3.6.1.

Lemma 4. *Let $\widetilde{X}$ be drawn with i.i.d. entries from the γ-sparsified Gaussian ensemble* (3.7)*. For any fixed y_i and m, $\mathbb{E}_{\widetilde{X}}\big[\psi_m\big(y_i \,\big|\, \widetilde{X}\big)\big] = \overline{\psi}_m(y_i)$, where*

$$\overline{\psi}_m(y_i) \;\;=\;\; \mathbb{E}_L\left[\frac{1}{\sqrt{2\pi\big(1+\frac{L\lambda^2}{\gamma}\big)}}\exp\left(-\frac{y_i^2}{2\big(1+\frac{L\lambda^2}{\gamma}\big)}\right)\right] \tag{3.19}$$

is a mixture of Gaussians with binomial weights $L \sim \mathrm{Binomial}(m,\gamma)$.

For certain scalings, we can use concentration results for U-statistics [54] to prove that ψ_m converges uniformly to $\overline{\psi}_m$, and from there that $h(\psi_m) \xrightarrow{p} h(\overline{\psi}_m)$. In general, however, we always have an upper bound, which is sufficient for our purposes. Indeed, since differential entropy $h(\psi_m)$ is a concave function of ψ_m, by Jensen's inequality and Lemma 4, we have

$$\mathbb{E}_{\widetilde{X}}[h(\psi_m)] \;\;\leq\;\; h\left(\mathbb{E}_{\widetilde{X}}[\psi_m]\right) \;\;=\;\; h(\overline{\psi}_m).$$

With these ingredients, we conclude that the conditional mutual information in equation (4.15) is upper bounded by

$$\begin{aligned} I\big(\theta;\widetilde{Y}\big|\widetilde{X}\big) \;\;&\leq\;\; \sum_{i=1}^{n} h\big(\widetilde{Y}_i\big|\widetilde{X}\big) - \frac{n}{2}\log(2\pi\mathrm{e}) \\ &=\;\; \sum_{i=1}^{n} \mathbb{E}_{\widetilde{X}}[h(\psi_m)] - \frac{n}{2}\log(2\pi\mathrm{e}) \\ &\leq\;\; nh(\overline{\psi}_m) - \frac{n}{2}\log(2\pi\mathrm{e}), \end{aligned}$$

where the last inequality uses the fact that the entropies $h(\overline{\psi}_m)$ associated with the densities $\overline{\psi}_m(y_i)$ are the same for all i. Therefore, the probability of decoding error, averaged over the sparsified Gaussian measurement ensemble, is bounded away from zero if

$$n \;<\; \frac{\log\binom{p-k+m}{m} - 1}{H(\overline{\psi}_m) - \frac{1}{2}\log(2\pi\mathrm{e})},$$

as claimed.

3.6.1 Limiting behavior

We now provide the proof of Lemma 4. Consider the following sequences of densities,

$$\psi_m\big(y_i \,\big|\, \widetilde{X}\big) \;=\; \frac{1}{\binom{p-k+m}{m}} \sum_U \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\Big(y_i - \lambda \sum_{j\in U} \widetilde{X}_{ij}\Big)^2\right)$$

and

$$\overline{\psi}_m(y_i) \;=\; \mathbb{E}_L\left[\frac{1}{\sqrt{2\pi\big(1+\frac{L\lambda^2}{\gamma}\big)}} \exp\left(-\frac{y_i^2}{2\big(1+\frac{L\lambda^2}{\gamma}\big)}\right)\right],$$

where $L \sim \text{Binomial}(m,\gamma)$. Our goal is to show that for any fixed y_i, the pointwise average of the stochastic sequence of densities ψ_m over the ensemble of matrices $\widetilde{X}$ satisfies $\mathbb{E}_{\widetilde{X}}\big[\psi_m\big(y_i \,\big|\, \widetilde{X}\big)\big] = \overline{\psi}_m(y_i)$.

By symmetry of the random measurement matrix $\widetilde{X}$, it is sufficient to compute this expectation for the subset $U = \{1,\ldots,m\}$. When each $\widetilde{X}_{ij}$ is i.i.d. drawn according to the γ-sparsified ensemble (3.7), the random variable $Z := \Big(y_i - \lambda\sum_{j=1}^m \widetilde{X}_{ij}\Big)$ has a Gaussian mixture distribution which can be described as follows. Denoting the mixture label by L, then $Z \sim N\big(y_i, \frac{\ell\lambda^2}{\gamma}\big)$ if $L = \ell$, for $\ell = 0, \ldots, m$. Moreover, define the modified random variable $\tilde{Z} := \frac{\gamma}{L\lambda^2}\Big(y_i - \lambda\sum_{j=1}^m \widetilde{X}_{ij}\Big)^2$. Then, conditioned on the mixture label $L = \ell$, the random variable $\tilde{Z}$ has a noncentral chi-square distribution with 1 degree of freedom and parameter $\frac{\gamma y_i^2}{\ell\lambda^2}$. Letting $M_\ell(t) = \mathbb{E}\big[\exp(t\tilde{Z})\,|\,L = \ell\big]$ denote the ℓ^{th} moment-generating function of $\tilde{Z}$, we have

$$\begin{aligned}
&\mathbb{E}_{\widetilde{X}}\left[\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\Big(y_i - \lambda\sum_{j=1}^m \widetilde{X}_{ij}\Big)^2\right)\right] \\
&= \sum_{\ell=0}^m \frac{1}{\sqrt{2\pi}}\,\mathbb{E}_{\widetilde{X}}\left[\exp\left(-\frac{1}{2}\Big(y_i - \lambda\sum_{j=1}^m \widetilde{X}_{ij}\Big)^2\right)\;\Bigg|\; L = \ell\right]\,\mathbb{P}(L = \ell) \\
&= \sum_{\ell=0}^m \,\frac{1}{\sqrt{2\pi}}M_\ell\left(-\frac{\ell\lambda^2}{2\gamma}\right)\,\mathbb{P}(L = \ell)
\end{aligned}$$

Evaluating the moment generating function [12] of a noncentral chi-square random variable then gives the desired quantity,

$$\mathbb{E}_{\widetilde{X}}\big[\psi_m\big(y_i\,\big|\,\widetilde{X}\big)\big] \;=\; \mathbb{E}_L\left[\frac{1}{\sqrt{2\pi\big(1 + \frac{L\lambda^2}{\gamma}\big)}}\exp\left(-\frac{y_i^2}{2\big(1 + \frac{L\lambda^2}{\gamma}\big)}\right)\right]$$

as claimed.

3.6.2 Bounding three regimes

In this section, we derive the bounds in Corollary 2 on the necessary conditions $g_m(p, k, \lambda, \gamma)$ stated in Theorem 4. We begin by applying a simple yet general bound on the entropy of the Gaussian mixture distribution with density $\overline{\psi}_m$ defined in (3.11). The variance associated with the density $\overline{\psi}_m$ is equal to $\sigma_m^2 = 1 + m\lambda^2$, and so $h(\overline{\psi}_m)$

is bounded by the entropy of a Gaussian distribution with variance σ_m^2, as

$$h(\overline{\psi}_m) \;\; \leq \;\; \frac{1}{2}\log(2\pi\mathrm{e}(1+m\lambda^2)).$$

This yields the first set of bounds in (3.14a).

Next, to derive more refined bounds which capture the effects of measurement sparsity, we will make use of the following lemma to bound the entropy associated with the mixture density $\overline{\psi}_m$.

Lemma 5. *For the Gaussian mixture distribution with density $\overline{\psi}_m$ defined in* (3.11)*,*

$$h(\overline{\psi}_m) \;\; \leq \;\; \mathbb{E}_L\left[\frac{1}{2}\log\left(1+\frac{L\lambda^2}{\gamma}\right)\right] + H(L) + \frac{1}{2}\log(2\pi\mathrm{e}),$$

where $L \sim$ Binomial(m,γ).

Proof. Let Z be a random variable distributed according to the density (3.19) with mixture label $L \sim \text{Binomial}(m,\gamma)$. To compute the entropy of Z, we expand the mutual information $I(Z;L)$ and obtain

$$h(Z) \;\; = \;\; h(Z|L) + H(L) - H(L|Z).$$

The conditional distribution of Z given that $L=\ell$ is Gaussian, and so the conditional entropy of Z given L can be written as

$$h(Z|L) \;\; = \;\; \mathbb{E}_L\left[\frac{1}{2}\log\left(2\pi\mathrm{e}\left(1+\frac{L\lambda^2}{\gamma}\right)\right)\right].$$

Using the fact that $0 \leq H(L|Z) \leq H(L)$, we obtain the upper and lower bounds on $h(Z)$,

$$h(Z|L) \;\; \leq \;\; h(Z) \;\; \leq \;\; h(Z|L) + H(L),$$

as claimed.

□

We can further bound the expression in Lemma 5 in three cases, delineated by the quantity γm. The proof of the following claim in given in Section 3.6.3.

Lemma 6. *Let* $E := \mathbb{E}_L\left[\frac{1}{2}\log\left(1+\frac{L\lambda^2}{\gamma}\right)\right]$, *where* $L \sim \text{Binomial}(m,\gamma)$.

(a) If $\gamma m > 3$, *then*

$$\frac{1}{4}\log\left(1+\frac{m\lambda^2}{3}\right) \;\leq\; E \;\leq\; \frac{1}{2}\log\left(1+m\lambda^2\right). \tag{3.20}$$

(b) If $\gamma m = \tau$ *for some constant* τ, *then*

$$\frac{1}{2}(1-\mathrm{e}^{-\tau})\log\left(1+\frac{m\lambda^2}{\tau}\right) \;\leq\; E \;\leq\; \frac{1}{2}\tau\log\left(1+\frac{m\lambda^2}{\tau}\right). \tag{3.21}$$

(c) If $\gamma m \leq 1$, *then*

$$\frac{1}{4}\gamma m\log\left(1+\frac{\lambda^2}{\gamma}\right) \;\leq\; E \;\leq\; \frac{1}{2}\gamma m\log\left(1+\frac{\lambda^2}{\gamma}\right). \tag{3.22}$$

Finally, combining Lemmas 5 and 6 with some simple bounds on the entropy of the binomial variate L (summarized in Lemmas 7 and 8 in Section 3.6.4), we obtain the bounds on $g_m(p,k,\lambda,\gamma)$ in equations (3.14b) and (3.14c).

3.6.3 Binomial concentration

We now derive the bounds in Lemma 6 on the expectation $E := \mathbb{E}_L\big[\frac{1}{2}\log\big(1+\frac{L\lambda^2}{\gamma}\big)\big]$, where $L \sim \text{Binomial}(m,\gamma)$. We first derive a general upper bound on E and then show that this bound is reasonably tight in the case when $\gamma m \leq 1$. We can rewrite the binomial probability as

$$p(\ell) \ := \ \binom{m}{\ell}\gamma^{\ell}(1-\gamma)^{m-\ell} \ = \ \frac{\gamma m}{\ell}\binom{m-1}{\ell-1}\gamma^{\ell-1}(1-\gamma)^{m-\ell}$$

and hence

$$E \ = \ \frac{1}{2}\gamma m \sum_{\ell=1}^{m} \frac{\log\left(1+\frac{\ell\lambda^2}{\gamma}\right)}{\ell}\binom{m-1}{\ell-1}\gamma^{\ell-1}(1-\gamma)^{m-\ell}.$$

Taking the first two terms of the binomial expansion of $\left(1+\frac{\lambda^2}{\gamma}\right)^{\ell}$ and noting that all the terms are non-negative, we obtain the inequality

$$\left(1+\frac{\lambda^2}{\gamma}\right)^{\ell} \ \geq \ 1+\frac{\ell\lambda^2}{\gamma}$$

and consequently $\log\left(1+\frac{\lambda^2}{\gamma}\right) \geq \frac{1}{\ell}\log\left(1+\frac{\ell\lambda^2}{\gamma}\right)$. Using a change of variables (by setting $\ell' = \ell - 1$) and applying the binomial theorem, we thus obtain the upper bound

$$\begin{aligned} E \ &\leq \ \frac{1}{2}\gamma m \sum_{\ell=1}^{m} \log\left(1+\frac{\lambda^2}{\gamma}\right)\binom{m-1}{\ell-1}\gamma^{\ell-1}(1-\gamma)^{m-\ell} \\ &= \ \frac{1}{2}\gamma m \log\left(1+\frac{\lambda^2}{\gamma}\right)\sum_{\ell'=0}^{m-1}\binom{m-1}{\ell'}\gamma^{\ell'}(1-\gamma)^{m-\ell'-1} \\ &= \ \frac{1}{2}\gamma m \log\left(1+\frac{\lambda^2}{\gamma}\right). \end{aligned}$$

In the case when $\gamma m \leq 1$, we can derive a similar lower bound by first bounding E as

$$\begin{aligned} E \ &\geq \ \frac{1}{2}\log\left(1+\frac{\lambda^2}{\gamma}\right)\sum_{\ell=1}^{m} p(\ell) \\ &= \ \frac{1}{2}\log\left(1+\frac{\lambda^2}{\gamma}\right)(1-(1-\gamma)^m). \end{aligned}$$

Now using the fact that $1+x \leq \mathrm{e}^x$ for all $x \in \mathbb{R}$, and $\mathrm{e}^{-x} \leq 1 - \frac{x}{2}$ for $x \in [0,1]$, we have

$$\begin{aligned} E &\geq \frac{1}{2}\log\left(1+\frac{\lambda^2}{\gamma}\right)(1-\mathrm{e}^{-\gamma m}) \\ &\overset{(a)}{\geq} \frac{1}{2}\log\left(1+\frac{\lambda^2}{\gamma}\right)\left(\frac{\gamma m}{2}\right). \end{aligned}$$

This yields the upper and lower bounds in (3.22).

Next, we examine the case when $\gamma m = \tau$ for some constant τ. The derivation of the upper bound for the $\gamma m \leq 1$ case holds when $\gamma m = \tau$ as well. The proof of the lower bound follows the same steps as in the $\gamma m \leq 1$ case, except that we stop before applying the last inequality (a). This gives the bounds in (3.21).

Finally, we derive bounds in the case when $\gamma m > 3$. Since the mean of a $L \sim$ Binomial(m,γ) random variable is γm, by Jensen's inequality the following upper bound always holds,

$$\mathbb{E}_L\left[\frac{1}{2}\log\left(1+\frac{L\lambda^2}{\gamma}\right)\right] \leq \frac{1}{2}\log(1+m\lambda^2).$$

To derive a matching lower bound, we use the fact that the median of a Binomial(m,γ) distribution is one of $\{\lfloor\gamma m\rfloor - 1, \lfloor\gamma m\rfloor, \lfloor\gamma m\rfloor + 1\}$. This allows us to bound

$$\begin{aligned} E &\geq \frac{1}{2}\sum_{\ell=\lfloor\gamma m\rfloor-1}^{m}\log\left(1+\frac{\ell\lambda^2}{\gamma}\right)p(\ell) \\ &\geq \frac{1}{2}\log\left(1+\frac{(\lfloor\gamma m\rfloor-1)\lambda^2}{\gamma}\right)\sum_{\ell=\lfloor\gamma m\rfloor-1}^{m} p(\ell) \\ &\geq \frac{1}{4}\log\left(1+\frac{m\lambda^2}{3}\right) \end{aligned}$$

where in the last step we used the fact that $\frac{(\lfloor\gamma m\rfloor-1)\lambda^2}{\gamma} \geq \frac{(\gamma m-2)\lambda^2}{\gamma} \geq \frac{m\lambda^2}{3}$ for $\gamma m > 3$, and $\sum_{\ell=\mathrm{median}}^{m} p(\ell) \geq \frac{1}{2}$. Thus we obtain the bounds in (3.20).

3.6.4 Bounds on binomial entropy

Lemma 7. *Let* $L \sim \text{Binomial}(m, \gamma)$*, then*

$$H(L) \;\; \leq \;\; \frac{1}{2} \log \left(2\pi \mathrm{e} \left(m\gamma(1-\gamma) + \frac{1}{12} \right) \right).$$

Proof. We immediately obtain this bound by applying the differential entropy bound on discrete entropy [19]. As detailed in [19], the proof follows by relating the entropy of the discrete random variable L to the differential entropy of a particular continuous random variable, and then upper bounding the latter by the entropy of a Gaussian random variable. □

Lemma 8. *The entropy of a binomial random variable* $L \sim \text{Binomial}(m, \gamma)$ *is bounded by*

$$H(L) \;\; \leq \;\; mH_{binary}(\gamma).$$

Proof. We can express the binomial variate as $L = \sum_{i=1}^{m} Z_i$, where $Z_i \sim \text{Bernoulli}(\gamma)$ i.i.d. Since $H(g(Z_1, \ldots, Z_m)) \leq H(Z_1, \ldots, Z_m)$, we have

$$H(L) \;\; \leq \;\; H(Z_1, \ldots, Z_m) \;\; = \;\; mH_{binary}(\gamma).$$

□

Lemma 9. *If* $\gamma = o\left(\frac{1}{m \log m}\right)$*, then* $mH_{binary}(\gamma) \to 0$ *as* $m \to \infty$*.*

Proof. To find the limit of $mH_{binary}(\gamma) = m\gamma \log \frac{1}{\gamma} + m(1-\gamma) \log \frac{1}{1-\gamma}$, let $\gamma = \frac{1}{mf(m)}$ for some function f, and assume that $f(m) = \omega(\log m)$. We can expand the first term as

$$m\gamma \log \frac{1}{\gamma} \;=\; \frac{1}{f(m)} \log(mf(m)) \;=\; \frac{\log m}{f(m)} + \frac{\log f(m)}{f(m)},$$

and so $\lim_{m\to\infty} m\gamma \log \frac{1}{\gamma} = 0$. The second term can also be expanded as

$$\begin{aligned} -m(1-\gamma)\log(1-\gamma) &= -m \log\left(1 - \frac{1}{mf(m)}\right) + \frac{1}{f(m)} \log\left(1 - \frac{1}{mf(m)}\right) \\ &= -\log\left(1 - \frac{1}{mf(m)}\right)^m + \frac{1}{f(m)} \log\left(1 - \frac{1}{mf(m)}\right). \end{aligned}$$

Since $f(m) \to \infty$ as $m \to \infty$, we have the limits

$$\lim_{m\to\infty} \left(1 - \frac{1}{mf(m)}\right)^m \;=\; 1 \qquad \text{and} \qquad \lim_{m\to\infty} \left(1 - \frac{1}{mf(m)}\right) \;=\; 1,$$

which in turn imply that

$$\lim_{m\to\infty} \log\left(1 - \frac{1}{mf(m)}\right)^m \;=\; 0 \qquad \text{and} \qquad \lim_{m\to\infty} \frac{1}{f(m)} \log\left(1 - \frac{1}{mf(m)}\right) \;=\; 0.$$

□

3.7 Discussion

In this chapter, we have studied the information-theoretic limits of exact support recovery for general scalings of the parameters $(n, p, k, \lambda, \gamma)$. Our first result (Theorem 3) applies generally to measurement matrices with zero-mean and unit-variance entries. It strengthens previously known bounds, and combined with known sufficient conditions [59], yields a sharp characterization of recovering signals with linear sparsity with a linear fraction of observations (Corollary 1). Our second result (Theorem 4) applies to γ-sparsified Gaussian measurement ensembles, and reveals three different regimes of measurement sparsity, depending on how significantly they impair statistical efficiency. For linear signal sparsity, Theorem 4 is not a sharp result (by a constant factor in comparison to Theorem 3 in the dense case); however, its tightness for sublinear signal sparsity is an interesting open problem. Finally, Theorem 3 implies that no measurement ensemble with zero-mean and unit-variance entries can further reduce the number of observations necessary for recovery, while [59] shows that the standard Gaussian ensemble can achieve the same scaling. This raises an interesting open question on the design of other, more computationally friendly, measurement matrices which achieve the same information-theoretic bounds.

Chapter 4

Model selection bounds for Gaussian Markov random fields

4.1 Introduction

Markov random fields or undirected graphical models are families of probability distributions whose factorization and conditional independence properties are characterized by the structure of an underlying graph. Graphical model selection refers to the problem of estimating the graph structure based on observed samples from a Markov random field. This problem arises in a wide variety of settings, including statistical image analysis, natural language processing, and computational biology. In many applications, this problem is of interest in the high-dimensional setting, in which both the graph size p and the number of samples n are large. Classical methods are known to break down when p/n does not go to zero. Moreover, without additional structure, the problem is often intractable when $p \gg n$. A line of recent work has focused on developing computationally efficient methods to solve this problem by imposing sparsity on the underlying graph structure. In particular, methods based on ℓ_1-regularization [63, 32, 21, 49, 46] have been shown to yield consistent estimators for high-dimensional graph selection.

Complementary in nature are the information-theoretic limits associated with any procedure for graphical model selection. Such analysis can serve two purposes. First, it can demonstrate when known polynomial-time algorithms achieve the information-theoretic bounds. Second, it can reveal regimes in which there exists a gap between the performance of current methods and the fundamental limits. With this motivation, previous work [50] has analyzed the fundamental limits of graphical model selection for binary Markov random fields.

The focus of this chapter is on the information-theoretic limits of Gaussian graphical model selection, in which the observed random vector has a multivariate Gaussian distribution. For Gaussian Markov random fields, the model selection problem is equivalent to estimating the off-diagonal sparsity pattern of the inverse covariance matrix. This chapter contains two types of results. Our first result is to derive conditions on the sample size n, graph size p, and maximum node degree d, that are necessary for any method to correctly recover the underlying graph with probability of error going to zero. Our second result addresses the problem of estimating the inverse covariance matrix Θ, and establishes necessary conditions for any method to produce an estimate $\widehat{\Theta}$ satisfying $\|\widehat{\Theta} - \Theta\| < \delta$. Our results can be compared against known sufficient conditions for graph selection and covariance estimation using ℓ_1-penalized maximum likelihood [46].

4.2 Graphical models

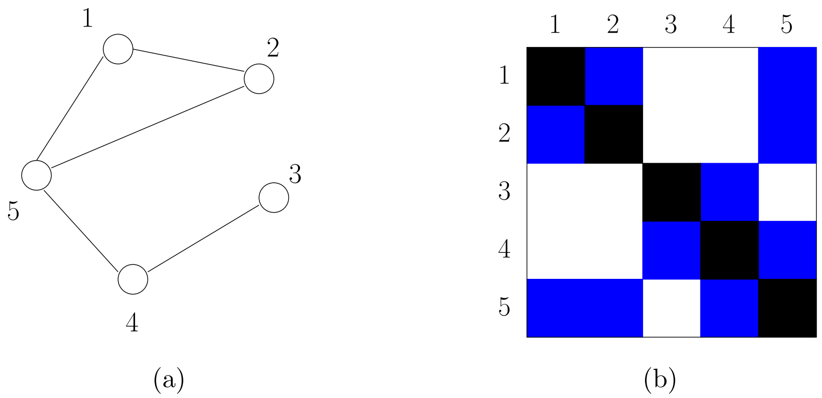
We begin with some background on Gaussian Markov random fields. We then formulate the graphical model selection problem, which for Gaussian models is directly related to estimation of the inverse covariance matrix. Our goal is to derive information-theoretic lower bounds on the number of samples required for recovery, which apply to any procedure regardless of its computational complexity.

4.2.1 Gaussian Markov random fields

Let $X = (X_1, \ldots, X_p)$ be a multivariate Gaussian random vector with zero mean and covariance matrix Σ. Accordingly, its density is determined completely by the inverse covariance matrix $\Theta = \Sigma^{-1}$, and has the form

$$f(x_1, \ldots, x_p) \;=\; \frac{1}{\sqrt{(2\pi)^p \det(\Theta^{-1})}} \exp\{-\frac{1}{2} x^T \Theta x\}. \tag{4.1}$$

For a given undirected graph $G = (V, E)$ with vertex set V and edge set $E \subset V \times V$, we associate a random variable X_i with each vertex $i \in V$. The Gaussian Markov random field associated with the graph G is the family of Gaussian distributions that respect the Markov properties of G. In particular, the off-diagonal sparsity pattern of the inverse covariance matrix Θ is specified by the edge structure of the graph, such that $\Theta_{ij} = 0$ if $(i, j) \notin E$ (see Figure).



(a) (b)

Figure 4.1. Illustration of Gaussian Markov random fields. (a) Given an undirected graph, associate a random variable X_i with each vertex i in the graph. A GMRF is the family of probability distributions over the vector X that respect the structure of the graph. (b) Sparsity pattern of the inverse covariance matrix Θ associated with the GMRF in (a).

Given i.i.d. samples from an unknown Markov random field, the problem of estimating the inverse covariance matrix Θ corresponds to recovering the graphical model instance, while the problem of estimating the underlying graph G corresponds to graphical model selection. We define the maximum degree of the graph as

$$d \ := \ \max_{i \in V} \Big| \{ j \in V \ | \ (i,j) \in E \} \Big|, \tag{4.2}$$

which is equal to the maximum number of non-zeros per row of the inverse covariance matrix Θ. Note that we are not including self-loops at each vertex in the degree count, corresponding to the diagonal entries Θ_{ii}. We often write $\Theta(G)$ to emphasize the graph-based structure of Θ.

4.2.2 Classes of graphical models

Let $\mathcal{G}_{p,d}$ be a family of undirected graphs on p vertices with edge sets that have degree at most d. For a given graph $G \in \mathcal{G}_{p,d}$, let $\Sigma(G)$ be the covariance matrix of a Gaussian Markov random field (GMRF) defined by the graph G. By definition, the inverse covariance matrix $\Theta(G)$ must have non-zeros only in positions corresponding to edges in E. In addition to graph structure, the difficulty of graphical model selection also depends on properties of the inverse covariance matrix entries. We define the minimum value of each matrix $\Theta(G)$ by the function

$$\lambda^*(\Theta(G)) \quad := \quad \min_{(s,t)\in E} \frac{|\Theta_{st}|}{\sqrt{\Theta_{ss}\Theta_{tt}}}, \tag{4.3}$$

so that it is invariant to rescaling of the data. We study the class of Gaussian Markov random fields parameterized by a lower bound on the minimum value, defined as

$$\mathcal{G}_{p,d}(\lambda) \quad := \quad \big\{\phi_{\Theta(G)} \;\big|\; G \in \mathcal{G}_{p,d},\ \Theta_{st} = 0 \text{ if } (s,t) \notin E,\ \lambda^*(\Theta(G)) \geq \lambda\big\}, \tag{4.4}$$

which consists of probability distributions of the form $\phi_{\Theta(G)} = \phi(0, \Sigma(G))$.

4.2.3 Decoders and error metrics

Suppose we are given n i.i.d. samples $X_1^n = \big(X^{(1)}, \dots, X^{(n)}\big) \in \mathbb{R}^{n\times p}$ from an unknown distribution $\phi_{\Theta(G)}$ in the class $\mathcal{G}_{p,d}(\lambda)$. Graphical model selection refers to the problem of estimating the underlying graph G based on the observations X_1^n. A decoder $\psi : \mathbb{R}^{n\times p} \to \mathcal{G}_{p,d}$ maps from the observations X_1^n to an estimated graph $\widehat{G} = \psi(X_1^n)$. We define the error metric between the estimate $\widehat{G}$ and the true underlying graph G using the 0-1 loss function $\mathbb{I}[\psi(X_1^n) \neq G]$. For any decoder ψ, we define the maximal probability of error over the class $\mathcal{G}_{p,d}(\lambda)$ as

$$p_{err}(\psi) \quad := \quad \max_{\phi_{\Theta(G)} \in \mathcal{G}_{p,d}(\lambda)} \mathbb{P}_{\Theta(G)}\big[\psi(X_1^n) \neq G\big], \tag{4.5}$$

where the error probability $\mathbb{P}_{\Theta(G)}\big[\psi(X_1^n) \neq G\big] = \mathbb{E}_{\Theta(G)}\big[\mathbb{I}[\psi(X_1^n) \neq G]\big]$ is taken with respect to the product distribution $\mathbb{P}_{\Theta(G)} = \phi(0, \Sigma(G))^n$ over n i.i.d. samples.

While graphical model selection corresponds to recovering the support set of $\Theta(G)$, the goal of inverse covariance estimation is to recover the entries of the inverse covariance matrix. More precisely, a decoder $\bar{\psi} : \mathbb{R}^{n \times p} \to \mathcal{G}_{p,d}(\lambda)$ maps from the samples X_1^n to an estimate $\widehat{\Theta} = \bar{\psi}(X_1^n)$. We measure the error between the estimate $\widehat{\Theta}$ and the true inverse covariance matrix Θ using the elementwise ℓ_∞-norm $\|\widehat{\Theta} - \Theta\|_\infty$, and define the probability of error $\mathbb{P}_{\Theta(G)}\big[\|\widehat{\Theta} - \Theta\|_\infty \geq \delta/2\big]$ with respect to the product distribution $\mathbb{P}_{\Theta(G)} = \phi(0, \Sigma(G))^n$. The maximal probability of error over the model class $\mathcal{G}_{p,d}(\lambda)$ is then defined as

$$p_{err}(\bar{\psi}) \quad := \quad \max_{\phi_{\Theta(G)} \in \mathcal{G}_{p,d}(\lambda)} \mathbb{P}_{\Theta(G)}\big[\|\widehat{\Theta} - \Theta\|_\infty \geq \delta/2\big]. \tag{4.6}$$

Although the error metrics for graphical model selection and inverse covariance estimation are closely related, neither recovery guarantee is strictly stronger than the other. In particular, it is possible to recover an estimate $\widehat{G} = G$ when $\|\widehat{\Theta} - \Theta\|_\infty \geq \delta/2$; conversely, it is also possible to recover an estimate satisfying $\|\widehat{\Theta} - \Theta\|_\infty < \delta/2$ when $\widehat{G} \neq G$.

With this set-up, our goal is to derive necessary conditions on the sample size $n(p, d, \lambda)$ for any decoder to reliably recover the underlying graph (or estimate the inverse covariance matrix). We say that recovery is asymptotically reliable over the graphical model class $\mathcal{G}_{p,d}(\lambda)$ if $p_{err} \to 0$ as $n \to 0$. Our analysis is high-dimensional in nature, in which the number of samples n, graph size p, and maximum degree of the graph d are all allowed to tend to infinity in a general manner.

4.3 Main results and consequences

In this section, we state our main results on the information-theoretic limits of Gaussian graphical model selection and inverse covariance estimation, and then discuss some of their consequences.

4.3.1 Graphical model selection

We begin with a set of necessary conditions for graphical model selection, applicable to any recovery method regardless of its computational complexity.

Theorem 5. *Consider the family $\mathcal{G}_{p,d}(\lambda)$ of Gaussian Markov random fields with $\lambda \in [0,1)$. A necessary condition for asymptotically reliable graphical model selection over the class $\mathcal{G}_{p,d}(\lambda)$ is*

$$n \quad > \quad \max\left\{ \frac{\log\binom{p-d+2}{2} - 1}{2\lambda^2}, \quad \frac{\log\binom{p}{d} - 1}{\frac{1}{2}\big(\log(1+d\lambda) - \frac{d\lambda}{1+d\lambda}\big)} \right\}. \tag{4.7}$$

The proof of Theorem 5, given in Section 4.4, constructs restricted ensembles of graphical models and then, viewing the observation process as a communication channel, uses Fano's inequality to bound the probability of error.

The first bound in Theorem 5 captures how the sample size must grow with graph size p and minimum value λ. In particular, if the minimum value scales as $\lambda = \Theta(\frac{1}{d})$, then Theorem 5 implies that the sample size must scale as $n = \Omega(d^2 \log(p-d))$. For any constant $\lambda \in [0,1)$, the second bound in Theorem 5 scales as $n = \Omega\big(\frac{d \log(p/d)}{\log(1+d\lambda)}\big)$. Moreover, it implies that $n = \Omega(d^{1-\epsilon} \log(\frac{p}{d}))$ for any $\epsilon > 0$.

The necessary conditions in Theorem 5 can be compared with previous work on polynomial-time methods for consistent graph selection. In particular, the inverse covariance matrix Θ can be estimated by solving the ℓ_1-regularized log-determinant program

$$\widehat{\Theta} \;:=\; \arg\min_{\Theta \succ 0} \big\{ \langle\!\langle \Theta, \widehat{\Sigma}^n \rangle\!\rangle - \log\det(\Theta) + \lambda_n \|\Theta\|_{1,\text{off}} \big\} \tag{4.8}$$

where $\widehat{\Sigma}^n$ denotes the sample covariance matrix, $\langle\!\langle A, B \rangle\!\rangle := \sum_{i,j} A_{ij} B_{ij}$ denotes the trace inner product, $\lambda_n > 0$ is a regularization parameter, and $\|\Theta\|_{1,\text{off}} := \sum_{i \neq j} |\Theta_{ij}|$ is the off-diagonal ℓ_1 regularizer. The underlying graph structure can then be estimated by the edge set $E\big(\widehat{\Theta}\big) = \{(i,j) \mid i \neq j, \widehat{\Theta}_{ij} \neq 0\}$. For Gaussian Markov random fields, the problem (4.8) is equivalent to ℓ_1-regularized maximum likelihood. Ravikumar et al. [46] showed that a sufficient condition for the problem (4.8) to consistently estimate the underlying graph is

$$n \;=\; \Omega((d^2 + \lambda^{-2}) \log p). \tag{4.9}$$

In the regime in which $\lambda = \Theta(\frac{1}{d})$, this scaling matches the information-theoretic bounds in Theorem 5.

4.3.2 Inverse covariance estimation

We now state some necessary conditions for the closely related problem of inverse covariance estimation. Let $\|A\|_\infty := \max_{ij} |A_{ij}|$ to denote the element-wise ℓ_∞-norm.

Theorem 6. *Consider the class of Gaussian Markov random fields $\mathcal{G}_{p,d}(\lambda)$. If any decoder can recover an estimate of the inverse covariance matrix satisfying $\|\widehat{\Theta} - \Theta\|_\infty < \delta/2$ in the element-wise ℓ_∞-norm with probability of error going to zero, then the number of samples must be greater than*

$$n \;>\; \frac{\log\left(\frac{pd}{4}\right) - 1}{2\delta^2}. \tag{4.10}$$

Theorem 6 captures how the sample size must grow with the minimum separation between models δ. A consequence of Theorem 6 is that if the recovery error decays at rate $\delta = 1/d$, then the sample size must scale as $n > d^2\big(\log\left(\frac{pd}{4}\right) - 1\big)/2$. Furthermore, Theorem 6 implies that the same necessary condition holds for inverse covariance estimation with other error metrics as well. In particular, let $\|\!|A|\!\|_F := (\sum_{ij} A_{ij}^2)^{1/2}$ denote the Frobenius norm.

Corollary 3. *A necessary condition for asymptotically reliable inverse covariance estimation, with recovery error at most $\delta/2$ measured in the Frobenius norm, is $n > \frac{\log\left(\frac{pd}{4}\right) - 1}{2\delta^2}$.*

Proof. For any two matrices $\widehat{\Theta}$ and Θ, the elementwise ℓ_∞ norm is upper bounded by the Frobenius norm as $\|\widehat{\Theta} - \Theta\|_\infty \leq \|\!|\widehat{\Theta} - \Theta|\!\|_F$. Consequently, the probability of error can be bounded as

$$\mathbb{P}_{\Theta(G)}\big[\|\widehat{\Theta} - \Theta\|_\infty \geq \delta/2\big] \;\leq\; \mathbb{P}_{\Theta(G)}\big[\|\!|\widehat{\Theta} - \Theta|\!\|_F \geq \delta/2\big]. \tag{4.11}$$

The necessary condition then follows from Theorem 6. □

The necessary condition in Theorem 6 can be compared to known sufficient conditions for ℓ_1-regularized maximum likelihood (4.8) to consistently estimate the inverse covariance matrix. Ravikumar et al. [46] showed that if the sample size is bounded as

$$n \;=\; \Omega(d^2 \log p), \tag{4.12}$$

then with probability going to one the program (4.8) can recover an estimate $\widehat{\Theta}$ satisfying

$$\|\widehat{\Theta} - \Theta\|_\infty \;=\; O\left(\sqrt{\frac{\log p}{n}}\right). \tag{4.13}$$

Consequently, the performance of the polynomial-time algorithm (4.8) matches the scaling of the information-theoretic bound in Theorem 6.

4.4 Applying Fano's method

4.4.1 Fano's method

Our general approach is to construct restricted ensembles of graphical models, and then use Fano's method to lower bound the probability of error in each restricted ensemble. Consider a restricted ensemble $\widetilde{\mathcal{G}}$ consisting of $M = \left|\widetilde{\mathcal{G}}\right|$ models, and let model index θ be chosen uniformly at random from $\{1, \dots, M\}$. Given the observations $\widetilde{X}_1^n \in \mathbb{R}^{n\times\nu}$, the decoder $\widetilde{\psi}$ estimates the underlying graph structure with probability of decoding error defined as

$$p_{err}(\widetilde{\psi}) \;=\; \max_{j=1,\dots,M} \mathbb{P}_{\widetilde{\Theta}(\widetilde{G}_j)}\left[\widetilde{\psi}\left(\widetilde{X}_1^n\right) \neq \widetilde{G}_j\right]. \tag{4.14}$$

By Fano's inequality, the maximal probability of error over $\widetilde{\mathcal{G}}$ can be lower bounded as

$$p_{err}(\widetilde{\psi}) \quad \geq \quad 1 - \frac{I\big(\theta; \widetilde{X}_1^n\big) + 1}{\log M}. \tag{4.15}$$

In order to make use of the Fano bound, the key is to design ensembles of matrices for which $\log M$ is large, while the mutual information $I\big(\theta; \widetilde{X}_1^n\big)$ is relatively small. Since it is typically difficult to evaluate the mutual information exactly, we discuss some upper bounds on it.

1. **Entropy-based bound**: Define the averaged covariance matrix

$$\bar{\Sigma} \quad := \quad \frac{1}{M} \sum_{j=1}^{M} \widetilde{\Sigma}\big(\widetilde{G}_j\big). \tag{4.16}$$

The mutual information is upper bounded by $I\big(\theta; \widetilde{X}_1^n\big) \leq \frac{n}{2} F\big(\widetilde{\mathcal{G}}\big)$, where

$$F\big(\widetilde{\mathcal{G}}\big) \quad := \quad \log\det \bar{\Sigma} - \frac{1}{M} \sum_{j=1}^{M} \log\det \widetilde{\Sigma}\big(\widetilde{G}_j\big). \tag{4.17}$$

Proof. See Section 4.4.2. □

2. **KL-based bound**: Let $\mathbb{P}_j = f\big(\widetilde{X}_1^n \big| \theta = j\big) = \phi\big(0, \widetilde{\Sigma}\big(\widetilde{G}_j\big)\big)^n$ for $j = 1, \ldots, M$. An alternative bound on the mutual information is given by

$$I\big(\theta; \widetilde{X}_1^n\big) \quad = \quad \mathbb{E}_\theta\left[D\left(\mathbb{P}_\theta \middle\| \frac{1}{M} \sum_{j=1}^{M} \mathbb{P}_j\right)\right] \tag{4.18}$$

$$\leq \quad \mathbb{E}_\theta[D(\mathbb{P}_\theta \| \mathbb{Q})] \tag{4.19}$$

for any distribution $\mathbb{Q}$ over $\widetilde{X}_1^n$. This upper bound follows from the fact that the mixture distribution $\frac{1}{M}\sum_{j=1}^{M} \mathbb{P}_j$ minimizes the averaged Kullback-Leibler distance over the family. Setting $\mathbb{Q} = \phi(0, I_{\nu\times\nu})^n$, the KL distance can be expressed as

$$D(\mathbb{P}_j \| \mathbb{Q}) \;=\; \frac{n}{2}\Big\{ \log\det\widetilde{\Theta}\big(\widetilde{G}_j\big) + \operatorname{trace}\big(\widetilde{\Sigma}\big(\widetilde{G}_j\big)\big) - \nu \Big\}. \tag{4.20}$$

Note that we are assuming $\log_e$ throughout; using $\log_2$ instead would change our results by a multiplicative constant of $\frac{1}{\log_e 2}$.

4.4.2 Bounds on mutual information

In this section we derive bounds on the mutual information that arises in Fano's inequality. Consider a restricted ensemble consisting of M models, and let model index θ be chosen uniformly at random from $\{1, \ldots, M\}$. Conditioned on the event $\theta = j$, we obtain n observations $\widetilde{X}_1^n \in \mathbb{R}^{n\times\nu}$, where each observation vector is sampled i.i.d. according to $\widetilde{X}^{(i)} \sim \phi\big(0, \widetilde{\Sigma}\big(\widetilde{G}_j\big)\big)$.

We begin by expanding the mutual information as

$$I\big(\theta; \widetilde{X}_1^n\big) \;=\; h\big(\widetilde{X}_1^n\big) - h\big(\widetilde{X}_1^n \big| \theta\big).$$

We bound the first term using the fact that differential entropy is maximized by the Gaussian distribution with a matched covariance matrix. Since each observation vector has covariance $\operatorname{cov}(\widetilde{X}^{(i)}) = \frac{1}{M}\sum_{j=1}^{M} \widetilde{\Sigma}\big(\widetilde{G}_j\big)$, we have

$$h\big(\widetilde{X}^{(i)}\big) \;\leq\; \frac{1}{2}\log\left((2\pi \mathrm{e})^{\nu} \det\left(\frac{1}{M}\sum_{j=1}^{M} \widetilde{\Sigma}\big(\widetilde{G}_j\big) \right) \right)$$

and consequently,

$$\begin{aligned} h\big(\widetilde{X}_1^n\big) &\le \sum_{i=1}^{n} h\big(\widetilde{X}^{(i)}\big) \\ &\le \frac{n}{2}\log(2\pi \mathrm{e})^{\nu} + \frac{n}{2}\log\det\left(\frac{1}{M}\sum_{j=1}^{M}\widetilde{\Sigma}\big(\widetilde{G}_j\big)\right). \end{aligned}$$

To compute the second term, we expand the conditional differential entropy of each observation vector as

$$h\big(\widetilde{X}^{(i)}|\theta\big) = \frac{1}{M}\sum_{j=1}^{M}\frac{1}{2}\log\big((2\pi\mathrm{e})^{\nu}\det\big(\widetilde{\Sigma}\big(\widetilde{G}_j\big)\big)\big).$$

Since the $\widetilde{X}^{(i)}$'s are conditionally independent given θ, we have

$$\begin{aligned} h\big(\widetilde{X}_1^n|\theta\big) &= \sum_{i=1}^{n} h\big(\widetilde{X}^{(i)}|\theta\big) \\ &= \frac{n}{2}\log(2\pi\mathrm{e})^{\nu} + \frac{n}{2M}\sum_{j=1}^{M}\log\det\big(\widetilde{\Sigma}\big(\widetilde{G}_j\big)\big). \end{aligned}$$

Combining these terms, we obtain the bound on the mutual information

$$I\big(\theta;\widetilde{X}_1^n\big) \le \frac{n}{2}\left\{\log\det\left(\frac{1}{M}\sum_{j=1}^{M}\widetilde{\Sigma}\big(\widetilde{G}_j\big)\right) - \frac{1}{M}\sum_{j=1}^{M}\log\det\big(\widetilde{\Sigma}\big(\widetilde{G}_j\big)\big)\right\},$$

as claimed.

4.4.3 Comparing bounds on mutual information

We now show that the entropy-based bound on mutual information in (4.17) is always tighter than the KL-based bound in (4.20). However, the two bounds are surprisingly close, and the KL-based bound is sometimes easier to compute. Recall that the entropy-based bound states that

$$I\big(\theta; \widetilde{X}_1^n\big) \;\; \leq \;\; \frac{n}{2}\left\{\log\det\bar{\Sigma} - \frac{1}{M}\sum_{j=1}^{M}\log\det\widetilde{\Sigma}\big(\widetilde{G}_j\big)\right\}, \tag{4.21}$$

while the KL-based bound gives

$$I\big(\theta; \widetilde{X}_1^n\big) \;\; \leq \;\; \frac{n}{2}\left\{\frac{1}{M}\sum_{j=1}^{M}\log\det\Theta\big(\widetilde{G}_j\big) + \frac{1}{M}\sum_{j=1}^{M}\operatorname{trace}\big(\widetilde{\Sigma}\big(\widetilde{G}_j\big)\big) - \nu\right\}. \tag{4.22}$$

First, note that two of the terms are the same, namely

$$-\frac{1}{M}\sum_{j=1}^{M}\log\det\widetilde{\Sigma}\big(\widetilde{G}_j\big) \;\; = \;\; \frac{1}{M}\sum_{j=1}^{M}\log\det\Theta\big(\widetilde{G}_j\big). \tag{4.23}$$

Next, since trace is a linear function, we have

$$\frac{1}{M}\sum_{j=1}^{M}\operatorname{trace}\big(\widetilde{\Sigma}\big(\widetilde{G}_j\big)\big) \;\; = \;\; \operatorname{trace}\left(\frac{1}{M}\sum_{j=1}^{M}\widetilde{\Sigma}\big(\widetilde{G}_j\big)\right) \;\; = \;\; \operatorname{trace}\big(\bar{\Sigma}\big). \tag{4.24}$$

Letting $\{\lambda_1, \ldots, \lambda_\nu\}$ denote the eigenvalues of $\bar{\Sigma}$, we can then compare the remaining terms in (4.22)

$$\operatorname{trace}\big(\bar{\Sigma}\big) - \nu \;\; = \;\; \sum_{i=1}^{\nu}(\lambda_i - 1) \tag{4.25}$$

to the remaining term in (4.21)

$$\log\det\bar{\Sigma} \;\; = \;\; \sum_{i=1}^{\nu}\log\lambda_i \;\; = \;\; \sum_{i=1}^{\nu}\log(1 + (\lambda_i - 1)). \tag{4.26}$$

Using the fact that $\log(x) \leq x$ for all x, we have that $\log\det\bar{\Sigma} \leq \operatorname{trace}\big(\bar{\Sigma}\big) - \nu$. Hence the entropy-based bound (4.17) is always tighter than the KL-based bound (4.20).

4.5 Analysis for graphical model selection

We now use these methods to derive the necessary conditions (stated in Theorem 5) on the sample size n as a function of the number of vertices p, degree d and minimum value λ. We obtain two necessary conditions, which can be seen as end points of an entire family of bounds, by analyzing ensembles of graphs in which a subset S of d nodes form a d-clique (i.e. fully connected subset), and the remaining nodes are all isolated.

4.5.1 Restricted ensemble A

We begin by deriving the following bound, which captures how the sample size must grow with the minimum value λ. Consider a family of graphs on p vertices, in which each edge set $E(S) = \{(s,t) \,|\, s,t \in S\}$ defines a clique over a subset S of size d. For a given graph $G = (V, E(S))$ and a parameter $a \geq 0$, we define the inverse covariance matrix

$$\Theta(G) \;:=\; I + a\mathbf{1}_S\mathbf{1}_S^T, \tag{4.27}$$

where $\mathbf{1}_S$ is the indicator vector of set S. The covariance matrix can then be computed as

$$\Sigma(G) \;=\; (\Theta(G))^{-1} \;=\; I - \frac{a}{1+da}\mathbf{1}_S\mathbf{1}_S^T. \tag{4.28}$$

The resulting class of graphical models is a subset of $\mathcal{G}_{p,d}(\lambda)$ if $\lambda^*(\Theta(G)) = \frac{a}{1+a} \geq \lambda$.

Suppose the decoder is given the indices of $(d-2)$ vertices in S, and the parameter value a. Estimating the underlying graph structure G now amounts to finding the remaining pair of nodes in S, out of $\binom{p-d+2}{2}$ possibilities. More precisely, let $T \subset S$ denote the set of revealed vertices and let $\{s,t\}$ denote the unknown indices of the remaining two nodes in S. Given (T, a), the decoder can extract the submatrix of observations $\widetilde{X}_1^n := (X_1^n)_{T^C} \in \mathbb{R}^{n\times(p-d+2)}$. When the original observations are sampled i.i.d. from the distribution $X^{(i)} \sim N(0, \Sigma)$, the modified observations are distributed according to $\widetilde{X}^{(i)} \sim N(0, \Sigma_{T^C T^C})$. Since the modified covariance matrix is of the form

$$\widetilde{\Sigma}\big(\widetilde{G}\big) \;:=\; \Sigma_{T^C T^C} \;=\; I - \frac{a}{1+da}\mathbf{1}_{st}\mathbf{1}_{st}^T, \tag{4.29}$$

the inverse covariance matrix becomes

$$\widetilde{\Theta}\big(\widetilde{G}\big) \;=\; \big(\widetilde{\Sigma}\big(\widetilde{G}\big)\big)^{-1} \;=\; I + \frac{a}{1+(d-2)a}\mathbf{1}_{st}\mathbf{1}_{st}^T. \tag{4.30}$$

Note that the underlying graph associated with $\widetilde{\Theta}\big(\widetilde{G}\big)$ is $\widetilde{G} := G \setminus T$ (i.e. the graph obtained by removing the vertices in set T and all edges connected to T from graph G). The remaining sub-problem is to determine, given the observations $\widetilde{X}_1^n$, the single edge graph on $(p-d+2)$ vertices.

Let $\widetilde{\mathcal{G}}$ denote the set of graphs on $(p-d+2)$ vertices with a single edge, and let $\widetilde{\mathcal{G}}(\lambda)$ denote the associated class of Gaussian Markov random fields with inverse covariance matrices defined as in (4.30). For this restricted ensemble, each matrix $\widetilde{\Sigma}\big(\widetilde{G}\big)$ has $(p-d+1)$ eigenvalues equal to one, and one eigenvalue equal to $1 - \frac{2a}{1+da}$. Consequently, for any $\widetilde{G} \in \widetilde{\mathcal{G}}$, we have

$$\log\det\big(\widetilde{\Sigma}\big(\widetilde{G}\big)\big) \;=\; \log\left(1 - \frac{2a}{1+da}\right). \tag{4.31}$$

We now calculate the averaged covariance matrix $\bar{\Sigma} = \frac{1}{\binom{p-d+2}{2}} \sum_{\widetilde{G}\in\widetilde{\mathcal{G}}} \widetilde{\Sigma}\big(\widetilde{G}\big)$. When averaged over all $\binom{p-d+2}{2}$ elements of $\widetilde{\mathcal{G}}$, each diagonal entry is equal to $1 - \frac{a}{1+da}$ with probability $\frac{p-d+1}{\binom{p-d+2}{2}} = \frac{2}{p-d+2}$, and 1 with probability $1 - \frac{2}{p-d+2}$. Each off-diagonal entry is equal to $-\frac{a}{1+da}$ with probability $\frac{1}{\binom{p-d+2}{2}}$, and 0 with probability $1-\frac{1}{\binom{p-d+2}{2}}$. Consequently, we have

$$\bar{\Sigma} = \left(1-\frac{2a}{(1+da)(p-d+2)}+\frac{a}{(1+da)\binom{p-d+2}{2}}\right)I-\frac{a}{(1+da)\binom{p-d+2}{2}}\mathbf{1}\mathbf{1}^T \tag{4.32}$$

Let us define $\gamma(p,d,a) := 1-\frac{2a}{(1+da)(p-d+2)}+\frac{a}{(1+da)\binom{p-d+2}{2}}$. The matrix $\bar{\Sigma}$ has $(p-d+1)$ eigenvalues with value $\gamma(p,d,a)$, and one with value $\gamma(p,d,a)-\frac{2a}{(1+da)(p-d+1)}$.

Putting together the pieces, we have that the entropy-based bound on mutual information (4.17) can be expressed as

$$F\big(\widetilde{\mathcal{G}}(\lambda)\big) = (p-d+1)\log\gamma(p,d,a)+\log\left(\gamma(p,d,a)-\frac{2a}{(1+da)(p-d+1)}\right) \tag{4.33}$$

$$-\log\left(1-\frac{2a}{1+da}\right). \tag{4.34}$$

We will bound each term of $F\big(\widetilde{\mathcal{G}}(\lambda)\big)$ using the fact that $\log(1+x)\leq x$ for all x. Accordingly, the first term can be bounded as

$$(p-d+1)\log\gamma(p,d,a) = (p-d+1)\log\left(1-\frac{2a(p-d)}{(1+da)(p-d+2)(p-d+1)}\right) \tag{4.35}$$

$$\leq -\frac{2a(p-d)}{(1+da)(p-d+2)}. \tag{4.36}$$

Similarly, the second term can be bounded as

$$\log\left(\gamma(p,d,a)-\frac{2a}{(1+da)(p-d+1)}\right) = \log\left(1-\frac{4a}{(1+da)(p-d+2)}\right) \tag{4.37}$$

$$\leq -\frac{4a}{(1+da)(p-d+2)}. \tag{4.38}$$

Finally, the last term can be bounded as

$$-\log\left(1-\frac{2a}{1+da}\right) = \log\left(1+\frac{2a}{1+(d-2)a}\right) \tag{4.39}$$

$$\leq \frac{2a}{1+(d-2)a}. \tag{4.40}$$

Combining these terms, we have the following bound

$$F\big(\widetilde{\mathcal{G}}(\lambda)\big) \;\leq\; \frac{4a^2}{(1+da)(1+(d-2)a)} \tag{4.41}$$

$$\leq\; \frac{4a^2}{(1+a)^2} \tag{4.42}$$

for $d \geq 3$. Recalling that $\lambda^*(\Theta(G)) = \frac{a}{1+a}$ and setting $a = \frac{\lambda}{1-\lambda}$, we obtain the bound $F\big(\widetilde{\mathcal{G}}(\lambda)\big) \leq 4\lambda^2$. Consequently, applying the Fano bound (4.15), we obtain that the probability of decoding error is bounded away from zero if

$$n \;<\; \frac{\log \binom{p-d+2}{2} - 1}{2\lambda^2} \tag{4.43}$$

as claimed.

4.5.2 Restricted ensemble B

We now derive a second lower bound, again using the ensemble of d-clique graphs and the entropy-based bound on mutual information (4.17). Consider the ensemble of graphs consisting of edge sets $E(S) = \{(s,t) \,|\, s,t \in S\}$ with $|S| = d$. For a given edge set $E(S)$ and paramter $a \geq 0$, define the inverse covaraince matrix

$$\Theta(G) \;:=\; I + a\mathbf{1}_S\mathbf{1}_S^T \tag{4.44}$$

and associated covariance matrix

$$\Sigma(G) \;=\; (\Theta(G))^{-1} \;=\; I - \frac{a}{1+da}\mathbf{1}_S\mathbf{1}_S^T. \tag{4.45}$$

Each matrix $\Theta(G)$ has $(p-1)$ eigenvalues equal to 1 and one eigenvalue equal to $1+da$, and thus we have

$$\log\det(\Sigma(G)) \;=\; -\log\det(\Theta(G)) \;=\; -\log(1+da). \tag{4.46}$$

We compute the averaged covariance matrix $\bar{\Sigma}$ over this ensemble as follows. Each diagonal entry is equal to $1 - \frac{a}{1+da}$ with probability $\binom{p-1}{d-1}/\binom{p}{d} = \frac{d}{p}$, and 1 with probability $1 - \frac{d}{p}$. Each off-diagonal entry is equal to $-\frac{a}{1+da}$ with probability $\binom{p-2}{d-2}/\binom{p}{d} = \frac{d(d-1)}{p(p-1)}$, and 0 with probability $1 - \frac{d(d-1)}{p(p-1)}$. The averaged covariance matrix can thus be written as

$$\bar{\Sigma} = \left(1 - \frac{da}{(1+da)p} + \frac{da(d-1)}{(1+da)p(p-1)}\right) I - \frac{da(d-1)}{(1+da)p(p-1)} \mathbf{1}\mathbf{1}^T. \tag{4.47}$$

The matrix $\bar{\Sigma}$ has $(p-1)$ eigenvalues equal to $\gamma(p,d,a) := 1 - \frac{da}{(1+da)p} + \frac{da(d-1)}{(1+da)p(p-1)}$, and one eigenvalue equal to $\gamma(p,d,a) - \frac{da(d-1)}{(1+da)(p-1)}$.

Using the entropy-based bound on mutual information in equation (4.17), we obtain

$$F(\widetilde{\mathcal{G}}) = (p-1)\log\gamma(p,d,a) + \log\left(\gamma(p,d,a) - \frac{da(d-1)}{(1+da)(p-1)}\right) + \log(1+da) \tag{4.48}$$

We now bound each term of $F(\widetilde{\mathcal{G}})$ in turn, using the fact that $\log(1+x) \leq x$ for all x. The first term can be bounded as

$$\begin{aligned} (p-1)\log\gamma(p,d,a) &= (p-1)\log\left(1 - \frac{da(p-d)}{(1+da)p(p-1)}\right) && (4.49)\\ &\leq -\frac{da(p-d)}{(1+da)p}. && (4.50) \end{aligned}$$

Similarly, the second term can be bounded as

$$\begin{aligned} \log\left(\gamma(p,d,a) - \frac{da(d-1)}{(1+da)(p-1)}\right) &= \log\left(1 - \frac{d^2a}{(1+da)p}\right) && (4.51)\\ &\leq -\frac{d^2a}{(1+da)p}. && (4.52) \end{aligned}$$

Combining these terms, we obtain the following bound on mutual information,

$$F(\widetilde{\mathcal{G}}) \leq \log(1+da) - \frac{da}{1+da}. \tag{4.53}$$

Note that the above bound is of the same form as the KL-based bound on mutual information (4.20), since $\text{trace}(\Sigma(G)) = p - \frac{da}{1+da}$. Recall that the minimum value for this ensemble must be bounded as $\lambda^*(\Theta(G)) = \frac{a}{1+a} \geq \lambda$. Assuming that $\lambda \in [0,1)$, we can set $a = \lambda$. Applying Fano's inequality (4.15) then gives that the probability of error stays bounded away from zero if

$$n \;<\; \frac{\log \binom{p}{d} - 1}{\frac{1}{2}\big(\log(1+d\lambda) - \frac{d\lambda}{1+d\lambda}\big)} \tag{4.54}$$

as claimed.

4.6 Analysis for inverse covariance estimation

We derive a new set of necessary conditions using an ensemble of graphical models which share the same underlying graph, but vary by perturbing a single edge weight. These bounds capture the difficulty of distinguishing between models with inverse covariance matrices that are δ-close, e.g in the element-wise ℓ_∞-norm. Note that for any two models $\Theta^{(i)}$ and $\Theta^{(j)}$ in our ensemble, since $\|\Theta^{(i)} - \Theta^{(j)}\|_\infty = \delta$ by construction, there does not exist a matrix $\widehat{\Theta}$ satisfying both $\|\widehat{\Theta} - \Theta^{(i)}\|_\infty < \delta/2$ and $\|\widehat{\Theta} - \Theta^{(j)}\|_\infty < \delta/2$. Consequently, we can apply Fano's inequality (4.15) to bound the probability of error in the restricted ensemble, and the problem is reduced to bounding the mutual information between the model index and the observations.

4.6.1 Alternate KL bound

We begin by stating a variant of the KL-based bound on mutual information in (4.20), using KL distances between all pairs of models in the class, instead of KL distances between each model and the standard Gaussian distribution.

Pairwise KL-based bound: By convexity of the Kullback-Leibler divergence, we have the following bound on mutual information

$$\begin{aligned} I\big(\theta; \widetilde{X}_1^n\big) &= \mathbb{E}_\theta\Big[D\Big(\mathbb{P}_\theta \Big\| \frac{1}{M}\sum_{j=1}^{M} \mathbb{P}_j\Big)\Big] && (4.55) \\ &\leq \frac{1}{M^2}\sum_{i=1}^{M}\sum_{j=1}^{M} D(\mathbb{P}_i \| \mathbb{P}_j) && (4.56) \end{aligned}$$

We define the symmetrized Kullback-Leibler divergence,

$$S(\mathbb{P}_i \| \mathbb{P}_j) \;:=\; D(\mathbb{P}_i \| \mathbb{P}_j) + D(\mathbb{P}_j \| \mathbb{P}_i), \tag{4.57}$$

and rewrite the bound on mutual information as

$$I\big(\theta; \widetilde{X}_1^n\big) \;\leq\; \frac{1}{M^2}\sum_{i=1}^{M}\sum_{j=i+1}^{M} S(\mathbb{P}_i \| \mathbb{P}_j). \tag{4.58}$$

For Gaussian Markov random fields, a straightforward calculation shows that the symmetrized KL distance is equal to

$$S(\mathbb{P}_i \| \mathbb{P}_j) \;=\; \frac{n}{2}\sum_{\ell=1}^{p}\sum_{m=1}^{p}\Big(\Theta_{\ell m}^{(i)} - \Theta_{\ell m}^{(j)}\Big)\Big(\Sigma_{\ell m}^{(j)} - \Sigma_{\ell m}^{(i)}\Big) \tag{4.59}$$

4.6.2 Restricted ensemble C

We now use these methods to derive necessary conditions for inverse covariance estimation (stated in Theorem 6), which capture how the sample size must grow with the minimum separation between models δ. Consider a graph on p vertices consisting of $\lfloor \frac{p}{d+1} \rfloor$ cliques, where each clique is of size $(d+1)$. Let $N = \lfloor \frac{p}{d+1} \rfloor$, and let $\{S_1, \ldots, S_N\}$ denote the N cliques with $|S_i| = d+1$. We define the inverse covariance matrix associated with this graph as

$$\bar{\Theta} \;:=\; I + a \sum_{i=1}^{N} \mathbf{1}_{S_i} \mathbf{1}_{S_i}^T \tag{4.60}$$

for some parameter $a \geq 0$. From this base model, we generate an ensemble of Gaussian Markov random fields in which each model perturbs the weight associated with one edge. Thus the model obtained by perturbing the weight on edge (s,t) is defined by the inverse covariance matrix

$$\Theta^{(i)} \;:=\; \bar{\Theta} + \delta(\mathbf{1}_{st}\mathbf{1}_{st}^T - I_{st}) \tag{4.61}$$

for some parameter $\delta \in (0, \frac{1}{2}]$. Note that we are using $(\mathbf{1}_{st}\mathbf{1}_{st}^T - I_{st})$ to denote the matrix with ones in locations (s,t) and (t,s), and zeros elsewhere. The resulting ensemble of graphical models has cardinality $M = \lfloor \frac{p}{d+1} \rfloor \binom{d+1}{2} \geq \frac{pd}{4}$.

For any two models in this ensemble, the perturbation matrix $E = \Theta^{(i)} - \Theta^{(j)}$ has exactly four non-zero entries, namely $\{\delta, \delta, -\delta, -\delta\}$. Suppose matrix $\Theta^{(i)}$ has perturbed weights corresponding to edge (s,t), and matrix $\Theta^{(j)}$ has perturbed weights corresponding to edge (u,v). Accordingly, the symmetrized KL distance (4.59) between these two models reduces to

$$S(\mathbb{P}_i \| \mathbb{P}_j) \;=\; \frac{n\delta}{2} \left\{ \left(\Sigma_{st}^{(j)} - \Sigma_{st}^{(i)}\right) + \left(\Sigma_{ts}^{(j)} - \Sigma_{ts}^{(i)}\right) - \left(\Sigma_{uv}^{(j)} - \Sigma_{uv}^{(i)}\right) - \left(\Sigma_{vu}^{(j)} - \Sigma_{vu}^{(i)}\right) \right\} \tag{4.62}$$

In order to compute the covariance matrix $\Sigma^{(i)}$, we will use the fact that $\Theta^{(i)}$ is a block diagonal matrix. In particular, define the submatrices

$$A \;:=\; I + a\mathbf{1}\mathbf{1}^T \;\in\; \mathbb{R}^{(d+1)\times(d+1)} \tag{4.63}$$

$$B \;:=\; A + \delta(\mathbf{1}_{\ell m}\mathbf{1}_{\ell m}^T - I_{\ell m}) \;\in\; \mathbb{R}^{(d+1)\times(d+1)}. \tag{4.64}$$

With this notation, the inverse covariance matrix $\Theta^{(i)}$ has $(N-1)$ blocks along the diagonal equal to A, and one block equal to B. In canonical form, the inverse covariance matrix can be written as

$$\Theta^{(i)} \;=\; \begin{bmatrix} A & & & 0 \\ & \ddots & & \\ & & A & \\ 0 & & & B \end{bmatrix}, \tag{4.65}$$

and the corresponding covariance matrix can be expressed as

$$\Sigma^{(i)} \;=\; \left(\Theta^{(i)}\right)^{-1} \;=\; \begin{bmatrix} A^{-1} & & & 0 \\ & \ddots & & \\ & & A^{-1} & \\ 0 & & & B^{-1} \end{bmatrix}. \tag{4.66}$$

A straightforward calculation yields that the inverse of A is equal to

$$A^{-1} \;=\; I - \frac{a}{1+(d+1)a}\mathbf{1}\mathbf{1}^T. \tag{4.67}$$

We can compute the inverse of B in canonical form by setting $\{\ell, m\} = \{d, d+1\}$, since for any permutation matrix P, if $\widetilde{B} = PBP^T$ then $\widetilde{B}^{-1} = PB^{-1}P^T$. Let us define the scalar parameters

$$\alpha \;:=\; \frac{1}{1-\delta}\left(\frac{a}{1+(d-1)a} + \delta\right) \tag{4.68}$$

$$\gamma \;:=\; \frac{a}{2a + (1+\delta)(1+(d-1)a)}. \tag{4.69}$$

Using the matrix inversion formula for block matrices, a little calculation shows that

$$B^{-1} = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}, \tag{4.70}$$

where

$$\begin{aligned} B_{11} &= I - \gamma(1+\delta)\mathbf{1}\mathbf{1}^T \in \mathbb{R}^{(d-1)\times(d-1)} & (4.71)\\ B_{12} &= B_{21}^T = -\gamma\mathbf{1}\mathbf{1}^T \in \mathbb{R}^{(d-1)\times 2} & (4.72)\\ B_{22} &= \frac{1}{1-\delta}\left(I - \frac{\alpha}{1+2\alpha}\mathbf{1}\mathbf{1}^T\right) \in \mathbb{R}^{2\times 2}. & (4.73) \end{aligned}$$

Putting together the pieces, we now compute the symmetrized Kullback-Leibler divergence (4.59) in three possible cases:

1. In the first case, the perturbed edge weights in $\Theta^{(i)}$ and $\Theta^{(j)}$ are in different cliques, and consequently we have

$$S(\mathbb{P}_i \| \mathbb{P}_j) = \frac{4n\delta}{2}\left\{-\frac{a}{1+(d+1)a} + \frac{\alpha}{(1-\delta)(1+2\alpha)}\right\}. \tag{4.74}$$

2. In the second case, the perturbed edges are in the same clique, but do not share a vertex, and we have

$$S(\mathbb{P}_i \| \mathbb{P}_j) = \frac{4n\delta}{2}\left\{-\gamma(1+\delta) + \frac{\alpha}{(1-\delta)(1+2\alpha)}\right\}. \tag{4.75}$$

3. In the third case, the perturbed edges are in the same clique, and they overlap in exactly one vertex, so that

$$S(\mathbb{P}_i \| \mathbb{P}_j) = \frac{4n\delta}{2}\left\{-\gamma + \frac{\alpha}{(1-\delta)(1+2\alpha)}\right\}. \tag{4.76}$$

Note that $(1+(d+1)a) \leq (1+(d+1)a+\delta(1+(d-1)a)) = (2a+(1+\delta)(1+(d-1)a))$, and consequently the bound in (4.74) is less than or equal to the bound in (4.76). Similarly, since $-\gamma \leq 0$ and $(1+\delta) \geq 1$, the bound in (4.75) is less than or equal to the bound in (4.76). Finally, the bound in (4.76) can be simplified to

$$\begin{aligned} S(\mathbb{P}_i \| \mathbb{P}_j) &= 2n\delta \left\{ \frac{\delta}{(1-\delta)} \left(\frac{1+da}{1+(d+1)a+\delta(1+(d-1)a)} \right) \right\} && (4.77) \\ &\leq 2n\delta \left\{ \frac{\delta}{1-\delta} \right\} && (4.78) \\ &\leq 4n\delta^2 && (4.79) \end{aligned}$$

if $\delta \in (0, \frac{1}{2}]$. Consequently, the mutual information in (4.58) can be bounded as

$$I\big(\theta; \widetilde{X}_1^n\big) \leq 2n\delta^2, \tag{4.80}$$

and applying the Fano bound (4.15) over this ensemble then gives the result as claimed.

4.7 Discussion

In this chapter, we have studied the information-theoretic limitations of Gaussian graphical model selection in the high-dimensional setting. Our analysis yielded a set of necessary conditions for consistent graph selection with any method, which matches the scaling of known sufficient conditions for ℓ_1-regularized maximum likelihood in certain regimes. Furthermore, we derived a set of necessary conditions for inverse covariance estimation, which similarly matches the performance of polynomial-time recovery methods. At a high-level, our analsis is based on a general framework for deriving information-theoretic bounds in which we view the observation process as a communication channel. This framework also underlies the analysis of the information-theoretic limits of sparse signal recovery in Chapter 3.

Chapter 5

Conclusions and future work

This thesis examined several problems in the area of high-dimensional sparse recovery, namely sparse approximation, subset selection, and graphical model selection. The common phenomenon in all these problems is that signal recovery is often intractable in the high-dimensional setting, but efficient recovery methods can be obtained by imposing sparsity on the underlying model. In this context, this thesis focused on two major themes:

- highlighting the power of sparse random projections for sparse signal recovery, in particular, for reducing computational complexity and storage costs as well as minimizing communication in distributed network applications

- characterizing the information-theoretic limits of sparse recovery problems, and providing a general framework for studying the fundamental limitations in such classes of problems.

More specifically, in Chapter 2, we analyzed a fast sketching algorithm for approximation of compressible signals based on sparse random projections in the presence of noise. We proposed a novel distributed algorithm using sparse random projections that enables robust refinable approximation in sensor networks. In Chapter 3, we showed the effect of using dense versus sparse measurement matrices on the information-theoretic limits of the sparsity recovery problem. Our analysis revealed that there is a fundamental trade-off between the sparsity of the measurement matrix and the number of samples needed for recovery. In Chapter 4, we studied the fundamental limits of graphical model selection and inverse covariance estimation, obtaining sharp characterizations in several regimes of interest.

5.1 Open problems and future research

We now discuss some open problems which arise from the work in this thesis. One interesting direction is the design of distributed multiresolution representations for large-scale networks. In Chapter 2, we proposed the use of distributed sparse random projections which allows approximations of the network data to be recovered by collecting a sufficient number of random projections from anywhere in the network. Computing these sparse random projections can be accomplished by communicating with as few as a constant number of sensors per random measurement. However, the data values must be routed between sensors over some underlying communication graph. Information could be disseminated more efficiently if the data is mixed locally first. Accordingly, ideas from gossip algorithms [37, 51, 24] and network coding could potentially be used in a hierarchical manner to efficiently combine data in the network. Such hierarchical structure could further be exploited to enable multiresolution reconstructions, where by querying sensors in a local region the decoder could obtain a coarse global approximation with fine local detail.

Our work also has implications for data streaming applications. Under the streaming model [6], large quantities of data arrive at high speeds, and it is infeasible to record in real time information like the most frequent k items. This problem arises in settings like internet traffic measurement, data center networks, and databases. Random linear projections of the data can be used to estimate the top k items, and can be computed over massive data streams in one pass and stored in small space. The use of sparse random projections could reduce storage and update times, and could also enable the design of distributed algorithms to infer network-wide traffic patterns from measurements at multiple routers.

Another interesting open problem is the development of recovery methods for compressible signals which have linear decoding complexity and optimal sampling efficiency. The use of sparse measurement matrices is one promising approach to developing faster recovery methods, for example by using message-passing algorithms on sparse graphs (e.g. [53, 61]). However, current methods – including the sketching recovery method analyzed in Chapter 2 – either do not achieve one of those performance goals, or cannot be applied to compressible signals.

Complementary to the development of efficient recovery methods are the information-theoretic bounds on sparse recovery. In Chapter 3, we presented a set of necessary conditions for sparse support recovery using sparse measurement matrices.

An interesting question is the tightness of these necessary conditions for sparse ensembles, i.e. whether there exist matching achievable bounds. Moreover, the comparison of the fundamental limits using dense versus sparse measurement matrices in Chapter 3 highlights the issue of designing more computationally friendly measurement ensembles which achieve the information-theoretic bounds. Furthermore, necessary and sufficient conditions for recovery in other error metrics (e.g. approximation in various ℓ_q-norms, partial support recovery, and prediction) remain an open question for general measurement ensembles.

Finally, we presented necessary conditions for Gaussian graphical model selection and inverse covariance estimation in Chapter **??** which matched known sufficient conditions in certain regimes of interest. A question for future research is the tightness of the bounds in regimes where there currently exists a gap, either by analysis of direct methods for graph selection (e.g. using exhaustive search) or by obtaining tighter necessary conditions. Another interesting direction is generalizations of our analysis to observation models with additive Gaussian noise, which could have applications in cryptography and image processing.

Bibliography

[1] D. Achlioptas, “Database-friendly random projections: Johnson-lindenstrauss with binary coins,” *Journal of Computer and System Sciences*, vol. 66, no. 4, pp. 671–687, 2003.

[2] S. Aeron, M. Zhao, and V. Saligrama, “Information-theoretic bounds to sensing capacity of sensor networks under fixed snr,” in *Information Theory Workshop*, September 2007.

[3] ——, “Fundamental limits on sensing capacity for sensor networks and compressed sensing, Tech. Rep. arXiv:0804.3439v1 [cs.IT], April 2008.

[4] N. Ailon and B. Chazelle, “Approximate nearest neighbors and the fast johnson-lindenstrauss transform,” in *ACM Symposium on Theory of Computing (STOC)*, 2006.

[5] M. Akcakaya and V. Tarokh, “Shannon theoretic limits on noisy compressive sampling, Tech. Rep. arXiv:0711.0366v1 [cs.IT], November 2007.

[6] N. Alon, Y. Matias, and M. Szegedy, “The space complexity of approximating the frequency moments,” in *ACM Symposium on Theory of Computing (STOC)*, 1996.

[7] W. Bajwa, J. Haupt, A. Sayeed, and R. Nowak, “Compressive wireless sensing,” in *International Conference on Information Processing in Sensor Networks (IPSN)*, 2006.

[8] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, “A simple proof of the restricted isometry property for random matrices,” *Constructive Approximation*, vol. 28, no. 3, pp. 253–263, December 2008.

[9] D. Baron, M. F. Duarte, S. Sarvotham, M. B. Wakin, and R. G. Baraniuk, “An information-theoretic approach to distributed compressed sensing,” in *Allerton*

Conference Communication, Control, and Computing, Monticello, IL, September 2005.

[10] R. Berinde, A. C. Gilbert, P. Indyk, H. Karloff, and M. J. Strauss, "Combining geometry and combinatorics: A unified approach to sparse signal recovery," in *Proc. Allerton Conference on Communication, Control and Computing*, Monticello, IL, September 2008.

[11] P. J. Bickel and E. Levina, "Regularized estimation of large covariance matrices," *Annals of Statistics*, vol. 36, no. 1, pp. 199–227, 2008.

[12] L. Birgé, "An alternative point of view on Lepski's method," in *State of the Art in Probability and Statistics*, ser. IMS Lecture Notes. Institute of Mathematical Statistics, 2001, no. 37, pp. 113–133.

[13] E. Candes, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, August 2006.

[14] E. Candes and T. Tao, "Decoding by linear programming," *IEEE Trans. Info Theory*, vol. 51, no. 12, pp. 4203–4215, December 2005.

[15] ——, "Near optimal signal recovery from random projections: Universal encoding strategies," *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5406–5425, December 2006.

[16] S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Computing*, vol. 20, no. 1, pp. 33–61, 1998.

[17] G. Cormode, M. Garofalakis, and D. Sacharidis, "Fast approximate wavelet tracking on streams," in *International Conference on Extending Database Technology (EDBT)*, 2006.

[18] G. Cormode and S. Muthukrishnan, "Towards an algorithmic theory of compressed sensing," Rutgers University, Tech. Rep., July 2005.

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

[20] S. Dasgupta and A. Gupta, "An elementary proof of the Johnson-Lindenstrauss lemma," UC Berkeley, Tech. Rep. 99-006, March 1999.

[21] A. d'Aspremont, O. Banerjee, and L. E. Ghaoui, "First order methods for sparse covariance selection," *SIAM Journal on Matrix Analysis and its Applications*, vol. 30, no. 1, pp. 56–66, 2008.

[22] I. Daubechies, "Ten lectures on wavelets," in *Regional Conference: Society for Industrial and Applied Mathematics*, Philadelphia, PA, 1992.

[23] R. A. DeVore, "Nonlinear approximation," *Acta Numerica*, vol. 7, pp. 51–150, 1998.

[24] A. Dimakis, A. Sarwate, and M. Wainwright, "Geographic gossip: Efficient averaging for sensor networks," *IEEE Transactions on Signal Processing*, vol. 56, no. 3, pp. 1205–1216, March 2008.

[25] D. Donoho, "Compressed sensing," *IEEE Trans. Info. Theory*, vol. 52, no. 4, pp. 1289–1306, April 2006.

[26] ——, "For most large underdetermined systems of linear equations, the minimal ℓ_1-norm near-solution approximates the sparsest near-solution," *Communications on Pure and Applied Mathematics*, vol. 59, no. 7, pp. 907–934, July 2006.

[27] ——, "For most large underdetermined systems of linear equations, the minimal ℓ_1-norm solution is also the sparsest solution," *Communications on Pure and Applied Mathematics*, vol. 59, no. 6, pp. 797–829, June 2006.

[28] D. Donoho, M. Elad, and V. M. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Info Theory*, vol. 52, no. 1, pp. 6–18, January 2006.

[29] D. L. Donoho, M. Vetterli, R. A. DeVore, and I. Daubechies, "Data compression and harmonic analysis," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2435–2476, October 1998.

[30] A. K. Fletcher, S. Rangan, and V. K. Goyal, "Necessary and sufficient conditions on sparsity pattern recovery, Tech. Rep. arXiv:0804.1839v1 [cs.IT], April 2008.

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

[32] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2007.

[33] A. Gilbert, M. Strauss, J. Tropp, and R. Vershynin, "Algorithmic linear dimension reduction in the ℓ_1-norm for sparse vectors," in *Proc. Allerton Conference on Communication, Control and Computing*, Allerton, IL, September 2006.

[34] A. Gilbert, Y. Kotidis, S. Muthukrishnan, and M. Strauss, "One-pass wavelet decompositions of data streams," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 3, pp. 541–554, May 2003.

[35] P. Indyk and R. Motwani, "Approximate nearest neighbors: towards removing the curse of dimensionality," in *ACM Symposium on Theory of Computing*, Dallas, TX, 1998.

[36] W. B. Johnson and J. Lindenstrauss, "Extensions of lipschitz mappings into a hilbert space," in *Conference in Modern Analysis and Probability*, 1984.

[37] D. Kempe, J. Kleinberg, and A. Demers, "Spatial gossip and resource location protocols," in *ACM Symposium on Theory of Computing*, 2001.

[38] P. Li, T. Hastie, and K. Church, "Very sparse random projections," in *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, 2006.

[39] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Processing*, vol. 41, no. 12, pp. 3397–3415, December 1993.

[40] S. G. Mallat, *A wavelet tour of signal processing*. New York: Academic Press, 1998.

[41] N. Meinshausen and P. Buhlmann, "High-dimensional graphs and variable selection with the lasso," *Annals of Statistics*, vol. 34, pp. 1436–1462, 2006.

[42] A. J. Miller, *Subset selection in regression*. New York, NY: Chapman-Hall, 1990.

[43] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Computing*, vol. 24, no. 2, pp. 227–234, 1995.

[44] D. Omidiran and M. J. Wainwright, "High-dimensional subset recovery in noise: Sparsified measurements without loss of statistical efficiency," Department of Statistics, UC Berkeley, Tech. Rep., April 2008, short version presented at Int. Symp. Info. Theory, July 2008.

[45] M. Rabbat, J. Haupt, A. Singh, and R. Nowak, "Decentralized compression and predistribution via randomized gossiping," in *International Conference on Information Processing in Sensor Networks (IPSN)*, 2006.

[46] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu, "High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence," Department of Statistics, UC Berkeley, Tech. Rep. 767, November 2008.

[47] G. Reeves and M. Gastpar, "Sampling bounds for sparse support recovery in the presence of noise," in *International Symposium on Information Theory*, Toronto, Canada, July 2008.

[48] J. Riordan, *Combinatorial Identities*, ser. Wiley Series in Probability and Mathematical Statistics. New York: Wiley, 1968.

[49] A. J. Rothman, P. J. Bickel, E. Levina, and J. Zhu, "Sparse permutation invariant covariance estimation," *Electronic Journal of Statistics*, vol. 2, pp. 494–515, 2008.

[50] N. Santhanam and M. J. Wainwright, "Information-theoretic limits of selecting binary graphical models in high dimensions," in *International Symposium on Information Theory (ISIT)*, Toronto, Canada, July 2008.

[51] R. Sarkar, X. Zhu, and J. Gao, "Hierarchical spatial gossip for multi-resolution representations in sensor networks," in *International Conference on Information Processing in Sensor Networks*, 2007.

[52] S. Sarvotham, D. Baron, and R. G. Baraniuk, "Measurements versus bits: Compressed sensing meets information theory," in *Proc. Allerton Conference on Control, Communication and Computing*, September 2006.

[53] ——, "Sudocodes: Fast measurement and reconstruction of sparse signals," in *Int. Symposium on Information Theory*, Seattle, WA, July 2006.

[54] R. J. Serfling, *Approximation Theorems of Mathematical Statistics*, ser. Wiley Series in Probability and Statistics. Wiley, 1980.

[55] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[56] J. Tropp, "Just relax: Convex programming methods for identifying sparse signals in noise," *IEEE Trans. Info Theory*, vol. 52, no. 3, pp. 1030–1051, March 2006.

[57] M. Vetterli and J. Kovačević, *Wavelets and Subband Coding.* Englewood Cliffs: Prentice Hall, 1995.

[58] M. J. Wainwright, “Sharp thresholds for high-dimensional and noisy recovery of sparsity using ℓ_1-constrained quadratic programs,” Department of Statistics, UC Berkeley, Tech. Rep. 709, 2006.

[59] ——, “Information-theoretic bounds for sparsity recovery in the high-dimensional and noisy setting,” Department of Statistics, UC Berkeley, Tech. Rep. 725, January 2007, presented at International Symposium on Information Theory, June 2007.

[60] W. Wang, M. Garofalakis, and K. Ramchandran, “Distributed sparse random projections for refinable approximation,” in *International Conference on Information Processing in Sensor Networks*, Nashville, TN, April 2007.

[61] W. Xu and B. Hassibi, “Efficient compressed sensing with deterministic guarantees using expander graphs,” in *Information Theory Workshop (ITW)*, September 2007.

[62] Y. Yang and A. Barron, “Information-theoretic determination of minimax rates of convergence,” *Annals of Statistics*, vol. 27, no. 5, pp. 1564–1599, 1999.

[63] M. Yuan and Y. Lin, “Model selection and estimation in the Gaussian graphical model,” *Biometrika*, vol. 94, no. 1, pp. 19–35, 2007.

[64] P. Zhao and B. Yu, “On model selection consistency of lasso,” *Journal of Machine Learning Research*, vol. 7, pp. 2541–2563, November 2006.