Is "Capacity" Defined for Source Recovery & Learning Problems?

Ioannis Mitliagkas & Sriram Vishwanath
Lab. of Informatics, Networks and Communications (LINC),
Dept. of Electrical and Computer Engineering,
University of Texas, Austin.
{imitliagkas,sriramuniverse}@gmail.com

***Abstract*—This paper presents "strong" limits on the number of samples required for recovery in the domains of compressive sensing, matrix-completion and model-selection. In other words, the paper determines lower-bounds on systems parameters; where if these lower bounds are not met, then *any* recovery algorithm will be guaranteed to be erroneous with high probability. A common mathematical framework is presented which is subsequently used to develop such lower bounds for each of these three domains.**

These lower bounds, along with existing achievability arguments is used to argue that a notion of "capacity" analogous to Shannon capacity of communication systems exists for certain recovery/learning problems.

I. Introduction

There is a vast and growing body of literature on recovery algorithms for a variety of problem settings. In particular, there are three classes of recovery problems that are of interest to us in this paper - compressive sensing, matrix completion and model selection. Our goal is to characterize "strong" limits on the number of samples needed to enable *any form* of recovery for each of the three classes. By strong, we mean that, without as many samples, recovery is impossible with high probability. Such strong information-theoretic limits are important as they provide an understanding of settings where recovery is impossible, regardless of the computation complexity of the algorithm or the cleverness of its design.

The first of the three classes we study is compressive sensing. In this class, a sparse source is to be recovered given linearly-filtered (and noisy) observations [1]. This class has been extensively studied in literature, with multiple algorithms for recovery in existence [1], [2]. The second class is comprised of matrix completion problems [6]. Here, randomly sampled entries of (a possibly low-rank) matrix are available, and it is desired that the entire matrix be recovered [5], [4], [3]. The third and final class considered in this paper is the family of model selection problems. In this class, samples of random variables from a Markov random field are provided, and it is desired that the original graphical model be recovered from them.

As such, these three classes of recovery problems seem fairly distinct, and algorithms developed to enable efficient recovery for each class possess features unique to that particular class. However, existing tools used to analyze the fundamental limits of each class are strikingly similar. In particular, weak bounds have been developed for each class. In [8], [9], the authors determine weak bounds for compressive sensing using information-theoretic techniques. Similarly, in [10], the authors consider the matrix completion problem and again use information-theoretic techniques to obtain bounds. Finally, in [13], [11], the model selection problem is studied using similar tools. In particular, in all three classes of problems, a common tool used to determine bounds is Fano's inequality [14]. Indeed, Fano's inequality has been used to establish bounds for multiple applications such as channel capacity and compression [14] in addition to sparse signal recovery, matrix completion and model selection.

However, these existing results using Fano's inequality are limited in that it typically results in bounds in the *weak*-sense. Given n samples and the *average* probability of error P_e suffered in the recovery process, Fano's inequality provides a lower bound on n in terms of system parameters given that we desire that $P_e \to 0$ (i.e., that the average probability of error decay to zero). Fano's inequality can be written as [14]:

$$I(T;Y) \geq (1-P_e)|\mathcal{T}| - H_b(P_e) \tag{1}$$

where $I(;)$ denotes the mutual information function and H_b the binary entropy function. Further, Y denotes the n (possibly noisy) samples while the random variable T, which takes values from a set $\mathcal{T}$ denotes the source/model to be recovered. Note that Fano's as stated in (1) enables one to relate mutual information (which in turn determines n) with the cardinality of the input set for near-perfect sensing/completion/model selection (when $P_e \to 0$). It is not effective in showing strong limits. When $P_e \to 1$, the inequality (1) reduces to stating the fact that mutual information is non-negative. As our goal is to establish strong lower bounds on n, we adopt a strategy similar to strong converses for channel capacity in [16] instead of Fano's inequality.

Given that n represents the number of samples, the main results of this paper are:

- For the compressive sensing problem where we desire to recover an k-sparse p-dimensional vector, we require that $n = \Theta(k \log p/k)$ samples for any recovery to be possible.

- For the matrix completion problem where we wish to recover a p-dimensional matrix of rank r, $n = \Theta(rp)$ is essential.
- For a Gaussian model selection problem where we desire to reconstruct a graph from a set of possible graphs $\mathcal{T}$, we must have $n = \Theta(\log |\mathcal{T}|)$.

Note that algorithms already exist that operate at or above (within polylog factors of) each of these three limits. For compressive sensing, algorithms are known that reconstruct the vector with $n = \Theta(k \log p/k)$ samples [7]. A similar result is true for low-rank matrix completion [12], [6] and model selection [11]. Thus, whenever an achievable strategy can be found, these information theoretic bounds are *sharp thresholds* at or above which perfect recovery is possible with high probability and below which any recovery is impossible with high probability. Such a sharp threshold is highly analogous to the notion of *capacity* from information theory. Thus, our determination of strong limits is a stepping stone to developing an analog of channel capacity for source or model recovery problems.

To reiterate, our approach to finding strong bounds is based on drawing parallels between each recovery problem and channel capacity analysis in Shannon theory. In many recovery problem settings, including compressed sensing and matrix completion, we have a real-valued source S belonging to a set $\mathcal{S}$ of uncountably infinite possibilities. However, channel capacity analysis typically assumes that the "message" being communicated belongs to a discrete set. We resolve this by identifying an underlying discrete set $\mathcal{T}$ that is being recovered. The choice for $\mathcal{T}$ could sometimes be intuitive. For instance, in compressed sensing, it could be the sparsity pattern of the source. More generally, it could be a quantized version of the source alphabet to within an (arbitrarily small) distortion $\delta > 0$.

The remainder of this paper is structured as follows: the next subsection summarizes the notation used in this paper. Section II presents a limited background on existing strong bounds in information theory. Next, Section III discusses strong bounds for the noisy compressive sensing problem. Similarly, Sections IV and V develop strong bounds for the matrix completion and model selection problems respectively. Finally, the paper concludes with Section VI.

A. Notation

As partly covered in the introduction, $\mathcal{X}$ denotes a set and $|\mathcal{X}|$ its cardinality. For any pair of random variables X, Y, x, y denote instantiations and $H(X)$ ($h(X)$) denotes the entropy in case of discrete (and differential entropy when continuous) in the information-theoretic sense [14]. Similarly $I(X;Y)$ denotes mutual information.

For matrices M, M^t, M^{-1} and $|M|$ denote its transpose, inverse and determinant respectively. All vectors are assumed to be column vectors unless specified otherwise.

II. Strong Bounds: Background

Strong limits (in particular, strong-converses) have found applications in the domains of channel capacity and compression. Wolfowitz [15] developed a strong converse for the capacity of discrete memoryless channels (DMCs) where channel capacity is defined as:

$$C = \max_{p(x)} I(X;Y).$$

In [15], the author shows that if the rate of transmission R exceeds C, then the average block error probability $P_e \overset{n\to\infty}{\longrightarrow} 1$.

Since we know that, for all $R < C$, $P_e \overset{n\to\infty}{\longrightarrow} 0$, this indicates that channel capacity C is a sharp threshold below which arbitrarily reliable communication is possible and above which communication is almost guaranteed to be erroneous.

In this paper, we do not find achievable strategies (algorithms for recovery) as many sophisticated ones already exist, and instead focus our energy on strong lower bounds on n for recovery. Note that our lower bounds on n hold in the scaling sense and in that way different than channel capacity, which is a single value.

Next, we present the problem setting and analysis for each of our three classes of problems.

III. Class 1: Compressive Sensing

A. Problem Setting

The setting adopted here is similar to that in [8]. Consider a vector S to be estimated through a noisy linear-filter:

$$Y = XS + W \tag{2}$$

where S is a p-dimensional k-sparse vector, X is the $n \times p$ measurement matrix and W i.i.d. Gaussian noise with unit variance. We impose a measure on the source set $\mathcal{S}$ such that the covariance of S is bounded by a semi-definite matrix Q and assume all source values are bounded. As before, our aim is to find a lower bound on n below which recovery of the source S is not possible with high probability.

B. Analysis

Our analysis of lower bounds will focus on the recovering the support of the sparse vector S instead of the vector itself. As the support is a (lossy) function of the vector S, the vector cannot be recovered if its support cannot be identified with high probability. Therefore, consider the support set T defined as:

$$T \triangleq \{i : S_i \neq 0\}$$

Let $\mathcal{T}$ denote a set comprised of all possible support sets. Note that $|\mathcal{T}|$ equals $\binom{p}{k}$. The support recovery problem closely resembles a communication channel where the "transmission codebook" consists of $\mathcal{T}$, and the "receiver" observes Y. Thus, from our understanding of channel capacity from information theory, the fundamental limit for recovering T (and thus, S) must (intuitively) be of the form $I(S;Y)$. As such $I(S;Y)$ is not necessarily a computationally tractable

expression. To obtain an explicit characterization, we upper bound $I(S;Y)$ as follows:

$$\begin{aligned} I(S;Y) &\overset{(a)}{=} h(Y) - h(Y|S) \\ &\overset{(b)}{\leq} \frac{1}{2}\log(2\pi e)^n |XQX^t + I| - h(XS + W|S) \\ &\overset{(c)}{=} \frac{1}{2}\log(2\pi e)^n |XQX^t + I| - h(W) \\ &\overset{(d)}{=} \frac{1}{2}\log|XQX^t + I| \end{aligned}$$

Here, (a) follows from the definition of mutual information. (b) follows from the fact that Gaussian distributions maximize differential entropy given a covariance constraint [14]. The noise W is independent of S, and therefore (c) is obtained. (d) follows as the noise is given to be Gaussian.

Next, we show that the upper bound on $I(S;Y)$ in (d) above is indeed an upper limit on "rate" for the compressive sensing system:

$$C_{cs} \triangleq \frac{1}{2}\log|XQX^t + I| \tag{3}$$

In other words, if the "rate" $R = \log\binom{p}{k}$ exceeds C_{cs}, then any form of recovery is impossible with high probability.

Theorem 1: Given the noisy compressive sensing problem as formulated by (2), the probability of error $P_e^{(n)}$ for any recovery algorithm can be lower bounded as:

$$P_e^{(n)} \geq 1 - \frac{nA_{cs}}{(R - C_{cs})^2} - 2^{(R - C_{cs})}$$

for a real-valued constant $A_{cs} > 0$, where C_{cs} is as defined by (3) and

$$R = \log\binom{p}{k}$$

In particular, for any $\alpha, \gamma > 0$, if

$$R > C_{cs} + n^{0.5+\alpha}\gamma \quad \Rightarrow \quad P_e \overset{n\to\infty}{\longrightarrow} 1 \tag{4}$$

A proof of this theorem is provided in the Appendix. In order to transform this result into a bound in terms of system parameters, we return to the definition of C_{cs} in (3). With some basic matrix manipulations, it is easy to show that C_{cs} grows as $\Theta(n)$. Therefore, rewriting (4), we get a lower bound on n required for recovery given by:

$$n = \Theta(k\log(p/k))$$

Note that this bound resembles bounds based on Fano's inequality as studied in [8, Theorem 2] and related literature. As shown in [8], algorithms are known to exist that operate at or above $n = \Theta(k\log(p/k))$. In all cases where algorithms exist, this scaling behavior in n represents a critical threshold, below which any recovery (even imperfect) is not possible with high probability, and above which reliable recovery can be achieved.

IV. Class 2: Matrix Completion

A. Problem Setting

The matrix completion problem setting follows the model studied in [12], [6]. $\mathcal{S}$ denotes the set of all rank r matrices S of the form:

$$S = UV$$

where U and V are $m \times r$ and $r \times m$ full rank matrices. The sampling strategy Z is comprised of n positions chosen uniformly at random within an $m \times m$ grid. The outcome of the sampling process is

$$Y = S_Z + W \tag{5}$$

Here, S_Z represents a vector formed by the values obtained with sampling strategy Z is applied, and W is i.i.d. Gaussian noise with identity covariance. As in Section III, we impose a measure on the set $\mathcal{S}$ such that the covariance of S is upper bounded by Q, and assume bounded source values.

The primary question in this setting is to determine a strong bound on the number of samples n to enable recovery to a finite set $\mathcal{T} \subseteq \mathcal{S}$. The case where $\mathcal{T}$ is a strict subset of $\mathcal{S}$ represents (lossy) compression where we wish to construct a matrix "closest" (given a suitable distortion measure) to the original matrix S.

B. Analysis

In obtaining a strong bound, it is worthwhile to first intuitively understand its connection with the domain of channel capacity. Here, the matrix $T \in \mathcal{T}$ to be recovered represents the "transmission codebook", and Y is the received vector from which we must recover T. In this analogy, the sampling strategy Z is known both to the transmitter and the receiver. Thus, we might expect the fundamental limit on matrix completion to closely resemble $I(S;Y|Z)$. Again, to obtain a computationally tractable bound, we take the following steps:

$$\begin{aligned} I(S;Y|Z) &\overset{(a)}{=} h(Y|Z) - h(Y|S,Z) && (6)\\ &\overset{(b)}{\leq} \frac{1}{2}\log(2\pi e)^n |Q_Z + I| - h(Y|S,Z) && (7)\\ &\overset{(c)}{\leq} \frac{1}{2}\log(2\pi e)^n |Q_Z + I| - h(W) && (8)\\ &\overset{(d)}{\leq} \frac{1}{2}\log|Q_Z + I| && (9) \end{aligned}$$

Here, Q_Z denotes the covariance matrix for sampled values of S and the reasons behind Equations $(a) - (d)$ are identical to those in Section III. We define:

$$C_{mr} \triangleq \frac{1}{2}\log|Q_Z + I| \tag{10}$$

as the upper limit ("capacity") of the matrix completion system. Thus, all rates $R = \log|\mathcal{T}|$ must be less that C_{mr} for recovery, and for any $R > C$, recovery is impossible with high probability. This notion is formalized by the following theorem:

Theorem 2: Given the noisy matrix completion problem as formulated by (5), the probability of error $P_e^{(n)}$ incurred by any matrix completion algorithm can be lower bounded as:

$$P_e^{(n)} \geq 1 - \frac{nA_{mr}}{(R - C_{mr})^2} - 2^{(R - C_{mr})}$$

for a real-valued constant $A_{mr} > 0$, where C_{mr} is as defined by (10) and $R = log|\mathcal{T}|$. In particular, for any $\alpha, \gamma > 0$, if

$$R > C_{mr} + n^{0.5+\alpha}\gamma \quad \Rightarrow \quad P_e \overset{n\to\infty}{\longrightarrow} 1 \tag{11}$$

Note that C_{mr} in fact grows as $\Theta(n)$ and $R = \Theta(rm)$ for a distortion metric that does not grow with m [10]. Thus, the strong lower bound on n is:

$$n = \Theta(rm) \tag{12}$$

V. Class 3: Gaussian Model Selection

A. Problem Setting

The setting we consider is similar to the one studied in [11]. Let $\mathcal{T}$ represent a class of graphs from which one represents the model corresponding to a Gaussian Markov random field (MRF). For instance, this class could be $\mathcal{G}_{p,d}$, the set of all graphs over p vertices with at most a degree d. In this section, we do not focus on the nature of the set $\mathcal{T}$, but instead investigate order-wise bounds on n in terms of $|\mathcal{T}|$ below which recovery is not possible with high probability.

B. Analysis

The analogy to communication in this setting is fairly straightforward. The graphical model T forms the transmission codebook, and Y represents the received vector from which T is to be determined. As in the previous sections, we associate a probability measure with the set $\mathcal{T}$. Again, it is intuitive that the fundamental bound on rate (defined as the number of distinguishable graphs) be related to $I(T;Y)$. Computationally tractable expressions for $I(T;Y)$ can be obtained, along the same lines as [11], as:

$$I(T;Y) = \frac{1}{2}\log|\mathbb{E}_T Q| - \mathbb{E}_T \log|Q|$$

where Q is the covariance of the Gaussian MRF for a particular realization of T. Further simplifications of this and other ways of expressing $I(T;Y)$ for model selection can be found in [11].

Defining

$$C_{gm} \triangleq \frac{1}{2}\log|\mathbb{E}_T Q| - \mathbb{E}_T \log|Q|, \tag{13}$$

we have the following theorem:

Theorem 3: Given the model selection problem as defined by (13), the probability of error for any selection algorithm can be lower bounded as:

$$P_e^{(n)} \geq 1 - \frac{nD}{(R - C_{gm})^2} - 2^{(R - C_{gm})}$$

where D is a positive constant and $R = \log|\mathcal{T}|$.

The proof of this theorem is in Appendix-*B*. In general, evaluating both C_{gm} and $|\mathcal{T}|$ in terms of system parameters can be non-trivial. [11], multiple bounds are presented for a Gaussian Markov random field. An analogous approach (but with different bounds) for binary MRFs can be found in [13].

VI. Conclusion

The main goal of this paper is to establish a general mechanism for finding strong limits for source/model recovery problems. The general structure for doing so is very similar to channel capacity analysis:

- Identify a mutual information expression that represents the information shared between source and samples
- Upper bound this mutual information term to obtain a computable expression in terms of system parameters
- Use a standardized suite of bounding techniques to show lower bounds on probability of error.

In all cases where achievable algorithms exist that operate in the same order, we have the order-wise capacity of the recovery problem.

VII. Appendix

A. Proof of Theorems 1 & 2

Here, we prove both Theorems 1 and 2 using a common framework based of Gallager's original strong converse proof [16]. The general mathematical model of the system for both cases (compressive sensing and matrix completion) can be written as

$$Y^n = XS_Z + W^n$$

where $W \sim \mathcal{N}(0, I)$. In the case of compressive sensing $S_Z = S$ and in the case of matrix completion $X = I$. In either case, a unitary transformation U exists such that

$$\tilde{Y^n} = UY^n = UXS_Z + UW = \tilde{S}_Z + \tilde{W^n}$$

where the covariance constraint on $\tilde{S}_Z$ is diagonal while the covariance on noise remains an identity. Henceforth, with a slight abuse of notation, we will assume that $Y^n = S_Z + W^n$ where

$$\mathbb{E}[S_Z S_Z^t] \leq diag(\lambda_1, \ldots, \lambda_m)$$

for some real valued $\lambda_i > 0 \ \forall \ i$.

Define for given source $s \in \mathcal{S}$ and sequence z^n

$$I(s; Y^n|z^n) \triangleq E_{Y^n|s,z^n} \log \frac{p(Y^n|z^n, s)}{p(Y^n|z^n)}$$

Note that, in compressive sensing, z^n is any constant sequence (unrelated to the problem), while in matrix completion, it represents a realization of the sampling sequence Z.

Next, we have:

$$\begin{aligned}
I(s;Y^n|z^n) &= \mathbb{E}_{Y^n|s,z^n} \log \frac{p(Y^n|z^n,s)}{p(Y^n|z^n)} \\
&= \mathbb{E}_{Y^n|s,z^n} \log \frac{p(Y^n|z^n,s)}{q(Y^n|z^n)} \frac{q(Y^n|z^n)}{p(Y^n|z^n)} \\
&= \mathbb{E}_{Y^n|s,z^n} \log \frac{p(Y^n|z^n,s)}{q(Y^n|z^n)} \\
&\quad - \mathbb{E}_{Y^n|s,z^n} \log \frac{p(Y^n|z^n)}{q(Y^n|z^n)} \\
&= \sum_{i=1}^{n} \mathbb{E}_{Y_i|s,z_i} \log \frac{p(Y_i|z_i,s)}{q(Y_i|z_i)} \\
&\quad - \mathbb{E}_{Y^n|s,z^n} \log \frac{p(Y^n|z^n)}{q(Y^n|z^n)}
\end{aligned}$$

where the introduced q is any measure such that $q(Y^n|z^n) = \prod_i q(Y_i|z_i)$. Define $C \triangleq \sum_{i=1}^{n} C_i$ where

$$C_i \triangleq \mathbb{E} \log \frac{p(Y_i|z_i,s)}{q(Y_i|z_i)},$$

then

$$\begin{aligned}
I(S;Y^n|Z^n) &= \mathbb{E}_{s,z^n} I(s;Y^n|z^n) \\
&= \mathbb{E}_{s,z^n} \sum_{i=1}^{n} \mathbb{E}_{Y_i|s,z_i} \log \frac{p(Y_i|z_i,s)}{q(Y_i|z_i)} \\
&\quad - \mathbb{E}_{s,z^n} \mathbb{E}_{Y^n|s,z^n} \log \frac{p(Y^n|z^n)}{q(Y^n|z^n)} \\
&= \mathbb{E}_{s,z^n} \sum_{i=1}^{n} \mathbb{E}_{Y_i|s,z_i} \log \frac{p(Y_i|z_i,s)}{q(Y_i|z_i)} \\
&\quad - \mathbb{E}_{z^n} D(p(Y^n|z^n)||q(Y^n||z^n)) \\
&\leq \sum_{i=1}^{n} \mathbb{E} \log \frac{p(Y_i|z_i,s)}{q(Y_i|z_i)} \\
&= C
\end{aligned}$$

as the K-L divergence between any two distributions is always non-negative. Since the source set $\mathcal{S}$ is, in the general case, infinite and often uncountable, we cannot demand exact recovery. We define a discrete subset of $\mathcal{S}$, called $\mathcal{T}$, over which we make the recovery decision, and sets D_t that constitute a partition of $\mathcal{S}$ (that is, $\bigcup_{t\in\mathcal{T}} D_t = \mathcal{S}$ and $D_x \bigcap D_y = \emptyset$ for $x, y \in \mathcal{T}$, $x \neq y$). Now define

$$B(s,z^n) = \left[y^n : \sum_{i=1}^{n} \log \frac{p(y_i|z_i,s)}{q(y_i|z_i)} > \sum_{i=1}^{n} (C_i + \epsilon) \right],$$

and consider an arbitrary recovery function g. The average probability of correct decoding is given by

$$P_c = \frac{1}{|\mathcal{Z}^n|} \sum_{z^n \in \mathcal{Z}^n} \sum_{t\in\mathcal{T}} \int_{s\in D_t} \int_{\substack{y^n \\ g(y^n,z^n)=t}} p(y^n|s,z^n) dy^n p(s) ds$$

Due to symmetry, the probability of correct decoding is the same for all typical sequences z^n [14], and so we express it in terms of a particular sequence z^n as:

$$\begin{aligned}
P_c &= \sum_{t\in\mathcal{T}} \int_{s\in D_t} \int_{y^n:g(y^n,z^n)=t} p(y^n|s,z^n) dy^n \; p(s) ds \\
&= \sum_{t\in\mathcal{T}} \int_{s\in D_t} \int_{\substack{y^n:g(y^n,z^n)=t \\ y^n \notin B(s,z^n)}} p(y^n|s,z^n) dy^n \; p(s) ds \\
&\quad + \sum_{t\in\mathcal{T}} \int_{s\in D_t} \int_{\substack{y^n:g(y^n,z^n)=t \\ y^n \in B(s,z^n)}} p(y^n|s,z^n) dy^n \; p(s) ds
\end{aligned}$$

We bound the first term as follows:

$$\begin{aligned}
&\sum_{t\in\mathcal{T}} \int_{s\in D_t} \int_{\substack{y^n:g(y^n,z^n)=t \\ y^n \notin B(s,z^n)}} p(y^n|s,z^n) dy^n \; p(s) ds \\
\leq &\sum_{t\in\mathcal{T}} \int_{s\in D_t} \int_{\substack{y^n:g(y^n,z^n)=t \\ y^n \notin B(s,z^n)}} q(y^n|z^n) 2^{\sum_{i=1}^{n}(C_i+\epsilon)} dy^n \; p(s) ds \\
\leq &\sum_{t\in\mathcal{T}} \int_{s\in D_t} \int_{y^n:g(y^n,z^n)=t} q(y^n|z^n) 2^{\sum_{i=1}^{n}(C_i+\epsilon)} dy^n \; p(s) ds \\
\leq &2^{\sum_{i=1}^{n}(C_i+\epsilon)} \sum_{t\in\mathcal{T}} \int_{s\in D_t} \int_{y^n:g(y^n,z^n)=t} q(y^n|z^n) dy^n \; p(s) ds \\
\leq &2^{\sum_{i=1}^{n}(C_i+\epsilon)} \sum_{t\in\mathcal{T}} \left(\int_{s\in D_t} p(s) ds \right) \left(\int_{y^n:g(y^n)=t} q(y^n) dy^n \right) \\
= &2^{\sum_{i=1}^{n}(C_i+\epsilon)} \sum_{t\in\mathcal{T}} \Pr[s \in D_t] \Pr[g(y^n) = t] \\
\leq &2^{\sum_{i=1}^{n}(C_i+\epsilon)} \max_{t\in\mathcal{T}} \Pr[s \in D_t]
\end{aligned}$$

For the second term:

$$\begin{aligned}
&\sum_{t\in\mathcal{T}} \int_{s\in D_t} \int_{\substack{y^n:g(y^n,z^n)=t \\ y^n \in B(s,z^n)}} p(y^n|s,z^n) dy^n p(s) ds \\
\leq &\sum_{t\in\mathcal{T}} \int_{s\in D_t} \int_{y^n\in B(s,z^n)} p(y^n|s,z^n) dy^n p(s) ds \\
\leq &\sum_{t\in\mathcal{T}} \int_{s\in D_t} \Pr(B(s,z^n)) p(s) ds \\
\leq &\sum_{t\in\mathcal{T}} \int_{s\in D_t} \Pr\left(\sum_{i=1}^{n} \log \frac{p(Y_i|z_i,s)}{q(Y_i|z_i)} > \sum_{i=1}^{n} (C_i + \epsilon) \right) p(s) ds \\
\leq &\sum_{t\in\mathcal{T}} \int_{s\in D_t} \Pr\left(\sum_{i=1}^{n} \log \frac{p(Y_i|z_i,s)}{q(Y_i|z_i)} - \sum_{i=1}^{n} C_i > n\epsilon \right) p(s) ds
\end{aligned}$$

Notice that by definition

$$\mathrm{E} \sum_{i=1}^{n} \log \frac{p(Y_i|z_i,s)}{q(Y_i|z_i)} = \sum_{i=1}^{n} C_i \ ,$$

and by independence

$$\mathrm{var}\left(\sum_{i=1}^{n} \log \frac{p(y_i|z_i,s)}{q(y_i|z_i)} \right) = \sum_{i=1}^{n} \mathrm{var}\left(\log \frac{p(y_i|z_i,s)}{q(y_i|z_i)} \right).$$

Taking $q(y_i|z_i)$ to be zero mean Gaussians with variance $\tilde{\lambda}_i + 1$, with $\tilde{\lambda}_i \leq \lambda_i$, the variance of individual terms can be calculated as

$$\text{var}\left(\log \frac{p(y_i|z_i,s)}{q(y_i|z_i)}\right) = \frac{\tilde{\lambda_i}^2 + 2s^2(z_i)}{2(\tilde{\lambda}_i+1)^2}.$$

where $s(z_i)$ denotes the element of s indexed by z_i. Since all variances and source values are finite, we can define

$$A = \max_i \text{var}\left(\log \frac{p(y_i|z_i,s)}{q(y_i|z_i)}\right).$$

A straightforward application of the Chebyshev inequality gives us, for any s,

$$\Pr\left(\sum_{i=1}^n \log \frac{p(Y_i|z_i,s)}{q(Y_i|z_i)} - \sum_{i=1}^n C_i > n\epsilon\right) \leq \frac{nA}{(n\epsilon)^2} = \frac{A}{n\epsilon^2}.$$

Our second term is a convex combination of probabilities like the one we bounded, so

$$\sum_{t\in\mathcal{T}} \int_{s\in D_t} \int_{\substack{y^n: g(y^n,z^n)=t \\ y^n \in B(s,z^n)}} p(y^n|s,z^n)dy^n \; p(s)ds \leq \frac{A}{n\epsilon^2}.$$

From these two bounds on the terms of P_c we get

$$P_c \leq \frac{A}{n\epsilon^2} + 2^{\sum_{i=1}^n (C_i+\epsilon)} \max_{t\in\mathcal{T}} \Pr[s \in D_t].$$

For partitions that contain sets of uniform mass this becomes

$$P_c \leq \frac{A}{n\epsilon^2} + 2^{\sum_{i=1}^n (C_i+\epsilon) - \log|\mathcal{T}|}.$$

B. Proof of Theorem 3

Now, we prove Theorem 3 using the same proof technique as before. We opted to give this proof separately to make it cleaner. A total of n samples from the GMRF are available

$$Y^n = S^n$$

where $S_i \sim \mathcal{N}(0,\Sigma)$. Let $\Theta = \Sigma^{-1}$ be the inverse covariance matrix. In the problem of inverse covariance estimation we want to recover Θ from Y^n, and in the problem of model selection we are only interested in the underlying graph (i.e. the sparsity pattern of Θ). We treat the latter as a subcase of the former, since we can partition the space $\mathcal{S}$ of all inverse covariance matrices into sets of matrices with the same sparsity structure. For now, we consider a general partition of $\mathcal{S}$ consisting of a finite number of sets D_t, with representative elements $t \in \mathcal{T}$.

Define for given inverse covariance matrix $\theta \in \mathcal{S}$

$$I(\theta; Y^n) \triangleq E_{Y^n|\theta} \log \frac{p(Y^n|\theta)}{p(Y^n)}.$$

Next, we have:

$$\begin{aligned}
I(\theta;Y^n) &= \mathbb{E}_{Y^n|\theta} \log \frac{p(Y^n|\theta)}{p(Y^n)} \\
&= \mathbb{E}_{Y^n|\theta} \log \frac{p(Y^n|\theta)}{q(Y^n)} \frac{q(Y^n)}{p(Y^n)} \\
&= \mathbb{E}_{Y^n|\theta} \log \frac{p(Y^n|\theta)}{q(Y^n)} - \mathbb{E}_{Y^n|\theta} \log \frac{p(Y^n)}{q(Y^n)} \\
&= \sum_{i=1}^n \mathbb{E}_{Y_i|\theta} \log \frac{p(Y_i|\theta)}{q(Y_i)} - \mathbb{E}_{Y^n|\theta} \log \frac{p(Y^n)}{q(Y^n)}
\end{aligned}$$

where the introduced q is any measure such that $q(Y^n) = \prod_i q(Y_i)$. Define $C \triangleq \sum_{i=1}^n C_i$ where

$$C_i \triangleq \mathbb{E} \log \frac{p(Y_i|\theta)}{q(Y_i)},$$

then

$$\begin{aligned}
I(\Theta;Y^n) &= \mathbb{E}_\theta I(\theta;Y^n) \\
&= \mathbb{E}_\theta \left(\sum_{i=1}^n \mathbb{E}_{Y_i|\theta} \log \frac{p(Y_i|\theta)}{q(Y_i)} - \mathbb{E}_{Y^n|\theta} \log \frac{p(Y^n)}{q(Y^n)} \right) \\
&= \sum_{i=1}^n \mathbb{E} \log \frac{p(Y_i|\theta)}{q(Y_i)} - \mathbb{E}_{Y^n} \log \frac{p(Y^n)}{q(Y^n)} \\
&= \sum_{i=1}^n \mathbb{E} \log \frac{p(Y_i|\theta)}{q(Y_i)} - D(p(Y^n)||q(Y^n)) \\
&\leq \sum_{i=1}^n \mathbb{E} \log \frac{p(Y_i|\theta)}{q(Y_i)} \\
&= C
\end{aligned}$$

as the K-L divergence between any two distributions is always non-negative. Now define

$$B(\theta) = \left[y^n : \sum_{i=1}^n \log \frac{p(y_i|\theta)}{q(y_i)} > \sum_{i=1}^n (C_i + \epsilon) \right],$$

and consider an arbitrary recovery function g. The probability of correct recovery is given by

$$\begin{aligned}
P_c &= \sum_{t\in\mathcal{T}} \int_{\theta\in D_t} \int_{y^n: g(y^n)=t} p(y^n|\theta)dy^n \; p(\theta)d\theta \\
&= \sum_{t\in\mathcal{T}} \int_{\theta\in D_t} \int_{\substack{y^n: g(y^n)=t \\ y^n \notin B(\theta)}} p(y^n|\theta)dy^n \; p(\theta)d\theta \\
&\quad + \sum_{t\in\mathcal{T}} \int_{\theta\in D_t} \int_{\substack{y^n: g(y^n)=t \\ y^n \in B(\theta)}} p(y^n|\theta)dy^n \; p(\theta)d\theta
\end{aligned}$$

We bound the first term as follows:

$$\begin{aligned}
&\sum_{t\in\mathcal{T}}\int_{\theta\in D_t}\int_{\substack{y^n:g(y^n)=t\\ y^n\notin B(\theta)}} p(y^n|\theta)dy^n\ p(\theta)d\theta\\
\leq&\sum_{t\in\mathcal{T}}\int_{\theta\in D_t}\int_{\substack{y^n:g(y^n)=t\\ y^n\notin B(\theta)}} q(y^n)2^{\sum_{i=1}^n(C_i+\epsilon)}dy^n\ p(\theta)d\theta\\
\leq&\sum_{t\in\mathcal{T}}\int_{\theta\in D_t}\int_{y^n:g(y^n)=t} q(y^n)2^{\sum_{i=1}^n(C_i+\epsilon)}dy^n\ p(\theta)d\theta\\
\leq&2^{\sum_{i=1}^n(C_i+\epsilon)}\sum_{t\in\mathcal{T}}\int_{\theta\in D_t}\int_{y^n:g(y^n)=t} q(y^n)dy^n\ p(\theta)d\theta\\
\leq&2^{\sum_{i=1}^n(C_i+\epsilon)}\sum_{t\in\mathcal{T}}\left(\int_{\theta\in D_t} p(\theta)d\theta\right)\left(\int_{y^n:g(y^n)=t} q(y^n)dy^n\right)\\
=&2^{\sum_{i=1}^n(C_i+\epsilon)}\sum_{t\in\mathcal{T}}\Pr[\theta\in D_t]\Pr[g(y^n)=t]\\
\leq&2^{\sum_{i=1}^n(C_i+\epsilon)}\max_{t\in\mathcal{T}}\Pr[\theta\in D_t]
\end{aligned}$$

For the second term:

$$\begin{aligned}
&\sum_{t\in\mathcal{T}}\int_{\theta\in D_t}\int_{\substack{y^n:g(y^n)=t\\ y^n\in B(\theta)}} p(y^n|\theta)dy^n p(\theta)d\theta\\
\leq&\sum_{t\in\mathcal{T}}\int_{\theta\in D_t}\int_{y^n\in B(\theta)} p(y^n|\theta)dy^n p(\theta)d\theta\\
\leq&\sum_{t\in\mathcal{T}}\int_{\theta\in D_t}\Pr(B(\theta))p(\theta)d\theta\\
\leq&\sum_{t\in\mathcal{T}}\int_{\theta\in D_t}\Pr\left(\sum_{i=1}^n\log\frac{p(Y_i|\theta)}{q(Y_i)}>\sum_{i=1}^n(C_i+\epsilon)\right)p(\theta)d\theta\\
\leq&\sum_{t\in\mathcal{T}}\int_{\theta\in D_t}\Pr\left(\sum_{i=1}^n\log\frac{p(Y_i|\theta)}{q(Y_i)}-\sum_{i=1}^nC_i>n\epsilon\right)p(\theta)d\theta
\end{aligned}$$

Notice that by definition

$$\mathrm{E}\sum_{i=1}^n\log\frac{p(Y_i|\theta)}{q(Y_i)}=\sum_{i=1}^nC_i\ ,$$

and by independence

$$\mathrm{var}\left(\sum_{i=1}^n\log\frac{p(y_i|\theta)}{q(y_i)}\right)=\sum_{i=1}^n\mathrm{var}\left(\log\frac{p(y_i|\theta)}{q(y_i)}\right).$$

The variance of individual terms is bounded and independent of n provided all elements of covariance matrix θ^{-1} are finite, and for $q(y_i)$ we also select a distribution with finite variance. Then, we can define

$$A=\max_i\mathrm{var}\left(\log\frac{p(y_i|\theta)}{q(y_i)}\right).$$

Invoking the Chebyshev inequality we get, for any θ,

$$\Pr\left(\sum_{i=1}^n\log\frac{p(Y_i|\theta)}{q(Y_i)}-\sum_{i=1}^nC_i>n\epsilon\right)\leq\frac{nA}{(n\epsilon)^2}=\frac{A}{n\epsilon^2}.$$

Our second term is a convex combination of probabilities like the one we bounded, so

$$\sum_{t\in\mathcal{T}}\int_{\theta\in D_t}\int_{\substack{y^n:g(y^n)=t\\ y^n\in B(\theta)}} p(y^n|\theta)dy^n\ p(\theta)d\theta\leq\frac{A}{n\epsilon^2}.$$

From these two bounds on the terms of P_c we get

$$P_c\leq\frac{A}{n\epsilon^2}+2^{\sum_{i=1}^n(C_i+\epsilon)}\max_{t\in\mathcal{T}}\Pr[\theta\in D_t].$$

For partitions that contain sets of uniform mass this becomes

$$P_c\leq\frac{A}{n\epsilon^2}+2^{\sum_{i=1}^n(C_i+\epsilon)-\log|\mathcal{T}|}.$$

REFERENCES

[1] D. Donoho, "Compressed sensing," IEEE Trans. Inf.. Theory, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[2] E. Candés and T. Tao, "Decoding by linear programming", IEEE Trans. Inf. Theory, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.

[3] A. Singer and M. Cucuringu, "Uniqueness of Low-Rank Matrix Comple- tion by Rigidity Theory", submitted 2009. arXiv:0902.3846.

[4] J-F. Cai, E. J. Candés, and Z. Shen, "A singular value thresholding algorithm for matrix completion", Technical report, 2008. arXiv:0810.3286.

[5] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum rank solutions to linear matrix equations via nuclear norm minimization", preprint (2007), submitted to SIAM Review. arXiv:0706.4138.

[6] E. J. Candés and T. Tao, "The power of convex relaxation: Near- optimal matrix completion", IEEE Trans. Inform. Theory, to appear. arXiv:0903.1476.

[7] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1–constrained quadratic programming (Lasso)," IEEE Trans. Inf. Theory, vol. 55, no. 5, pp. 2183–2202, May 2009.

[8] M. J. Wainwright, "Information-theoretic limitations on sparsity recovery in the high-dimensional and noisy setting", IEEE Transactions on Information Theory, December 2009.

[9] W. Wang, M. J. Wainwright and K. Ramchandran, "Information-theoretic limits on sparse signal recovery: Dense versus sparse measurement matrices", IEEE Transactions on Information Theory, June 2010.

[10] S. Vishwanath,"Information-theoretic bounds on low-rank matrix completion", Proceedings of the International Symposium on Information Theory (ISIT), Austin, TX 2010.

[11] W. Wang, M. Wainwright and K. Ramchandran, "Information-theoretic bounds on model selection for Gaussian Markov random fields", Proceedings of the International Symposium on Information Theory (ISIT), Austin, TX 2010.

[12] R. Kesavan, A. Montanari and S. O h, "Matrix Completion from a few entries", arXiv:0901.3150, 2009.

[13] N. Santhanam and M. J. Wainwright, "Information-theoretic limits of selecting binary graphical models in high dimensions", Technical report.

[14] T. Cover and J. Thomas, Elements of Information Theory, Wiley 1991.

[15] J. Wolfowitz. Strong converse of the coding theorem for the general discrete finite-memory channel. *Inform. and Control*, 3:89–93, 1960.

[16] Robert G. Gallager. *Information Theory and Reliable Communication*. John Wiley & Sons, Inc., New York, NY, USA, 1968.