Nearly Sharp Sufficient Conditions on Exact Sparsity Pattern Recovery

Kamiar Rahnama Rad

***Abstract*—Consider the n-dimensional vector $y = X\beta + \epsilon$ where $\beta \in \mathbb{R}^p$ has only k nonzero entries and $\epsilon \in \mathbb{R}^n$ is a Gaussian noise. This can be viewed as a linear system with sparsity constraints corrupted by noise, where the objective is to estimate the sparsity pattern of β given the observation vector y and the measurement matrix X. First, we derive a nonasymptotic upper bound on the probability that a specific wrong sparsity pattern is identified by the maximum-likelihood estimator. We find that this probability depends (inversely) exponentially on the difference of $\|X\beta\|_2$ and the ℓ_2-norm of $X\beta$ projected onto the range of columns of X indexed by the wrong sparsity pattern. Second, when X is randomly drawn from a Gaussian ensemble, we calculate a nonasymptotic upper bound on the probability of the maximum-likelihood decoder not declaring (partially) the true sparsity pattern. Consequently, we obtain sufficient conditions on the sample size n that guarantee almost surely the recovery of the true sparsity pattern. We find that the required growth rate of sample size n matches the growth rate of previously established necessary conditions.**

***Index Terms*—Hypothesis testing, random projections, sparsity pattern recovery, subset selection, underdetermined systems of equations.**

I. Introduction

FINDING solutions to underdetermined systems of equations arises in a wide array of problems in science and technology; examples include array signal processing [1], neural [2] and genomic data analysis [3], to name a few. In many of these applications, it is natural to seek for *sparse* solutions of such systems, i.e., solutions with few nonzero elements. A common setting is when we believe or we know *a priori* that only a *small subset* of the candidate sources, neurons, or genes influence the observations, but their location is unknown.

More concretely, the problem we consider is that of estimating the support of $\beta \in \mathbb{R}^p$ given the *a priori* knowledge that only k of its entries are nonzero based on the observational model

$$y = X\beta + \epsilon \tag{1}$$

where $X \in \mathbb{R}^{n\times p}$ is a collection of input measurement vectors, $y \in \mathbb{R}^n$ is the output measurement and $\epsilon \in \mathbb{R}^n$ is the additive measurement noise, assumed to be zero mean and with known covariance equal to $I_{n\times n}$[1]. Each row of X and the corresponding entry of y are viewed as an input and output measurement, respectively.

The output of the optimal (sparsity) decoder is defined as the support set of the sparse solution $\hat{\beta}$ with support size k that minimizes the residual sum of squares where

$$\hat{\beta} = \underset{|\mathrm{support}(\theta)|=k}{\arg\min} \|y - X\theta\|_2^2 \tag{2}$$

is the optimal estimate of β given the *a priori* information of sparseness. The support set of $\hat{\beta}$ is optimal in the sense of minimizing the probability of identifying a wrong sparsity pattern.

First, we are concerned with the likelihood of the sparsity pattern of $\hat{\beta}$ as a function of X and β. We obtain an upper bound on the probability that $\hat{\beta}$ has any specific sparsity pattern and find that this bound depends (inversely) exponentially on the difference of $\|X\beta\|_2$ and the ℓ_2-norm of $X\beta$ projected onto the range of columns of X indexed by the wrong sparsity pattern.

Second, when the entries of X are independent and identically distributed (i.i.d.) random variables we are concerned with establishing sufficient conditions that guarantee the reliability of sparsity pattern recovery. Ideally, we would like to characterize such conditions based on a minimal number of parameters including the sparsity level k, the signal dimension p, the number of measurements n and the signal-to-noise ratio(SNR) which is equal to

$$\mathrm{SNR} = \frac{\mathrm{E}\left[\|X\beta\|_2^2\right]}{\mathrm{E}\left[\|\epsilon\|_2^2\right]}. \tag{3}$$

Assume that the absolute value of the nonzero entries of β are lower bounded by $\beta_{\min}$[2]. Further, suppose that the variance of the entries of X is equal to one [1]1. Hence

$$\mathrm{SNR} \geq k\beta_{\min}^2$$

and therefore it is natural to ask, how does the ability to reliably estimate the sparsity pattern depend on $(n, p, k, \beta_{\min}^2)$.

We find that a nonasymptotic upper bound on the probability of the maximum-likelihood decoder not declaring the true sparsity pattern can be found when the entries of the measurement matrix are i.i.d. normal random variables. This allows us to obtain sufficient conditions on the number of measurements n as a function of $(p, k, \beta_{\min}^2)$ for reliable sparsity recovery. We show that our results strengthen earlier sufficient conditions

Manuscript received October 02, 2009; revised August 19, 2010; accepted January 28, 2011. Date of current version June 22, 2011. This work was presented at the 43rd Annual Conference on Information Sciences and Systems, March 2009.

The author is with the Department of Statistics, Columbia University, New York, NY 10027 USA (e-mail: kamiar@stat.columbia.edu).

Communicated by J. Romberg, Associate Editor for Signal Processing.

Digital Object Identifier 10.1109/TIT.2011.2145670

[1]This entails no loss of generality, by standard rescaling of β.

[2]To the best of our knowledge, Wainwright [4] was the first to formulate the information theoretic limitations of sparsity pattern recovery using $\beta_{\min}$ as one of the key parameters.

[4]–[7], and we show that the sufficient conditions on n match the growth rate of the necessary conditions in both the linear, i.e., $k = \Theta(p)$, and the sublinear, i.e., $k = o(p)$, regimes, as long as $\beta_{\min}^2$ is $\Omega\left(\frac{1}{k}\right)$ and $O(1)$.

A. Previous Work

A large body of recent work, including [4]–[10], analyzed reliable sparsity pattern recovery exploiting optimal and suboptimal decoders for large random Gaussian measurement matrices. The average error probability, necessary and sufficient conditions for sparsity pattern recovery for Gaussian measurement matrices were analyzed in [4] in terms of $\left(n, p, k, \beta_{\min}^2\right)$. As a generalization of the previous work, using the Fano inequality, necessary conditions for general random and sparse measurement matrices were presented in [8]. The sufficient conditions in [6] were obtained based on a simple maximum correlation algorithm and a closely related thresholding estimator discussed in [11]. In addition to the well-known formulation of the necessary and sufficient conditions based on $\left(n, p, k, \beta_{\min}^2\right)$, Fletcher *et al.* [6] included the maximum-to-average ratio[3] of β in their analysis. Necessary and sufficient conditions for fractional sparsity pattern recovery were analyzed in [5], [9].

We will discuss the relationship to this work below in more depth, after describing our analysis and results in more detail.

B. Notation

The following conventions will remain in effect throughout this paper. Calligraphic letters are used to indicate sparsity patterns defined as a set of integers between 1 and p, with cardinality k. We say $\beta \in \mathbb{R}^p$ has sparsity pattern $\mathcal{T}$ if the entries with indices $i \in \mathcal{T}$ are nonzero. $\mathcal{T} - \mathcal{F}$ stands for the set of entries that are in $\mathcal{T}$ but not in $\mathcal{F}$ and $|\mathcal{T}|$ for the cardinality of $\mathcal{T}$. We denote by $X_{\mathcal{T}} \in \mathbb{R}^{n \times |\mathcal{T}|}$, the matrix obtained from X by extracting $|\mathcal{T}|$ columns with indices obeying $i \in \mathcal{T}$. Let $\mathcal{S}(\beta)$ stand for the sparsity pattern or support set of β. The matrix norm $\|.\|_{a,b}$ of a matrix A defined as

$$\|A\|_{a,b} := \max_{x \neq 0} \frac{\|Ax\|_a}{\|x\|_b}.$$

Note that if A is a positive semi-definite matrix then $\|A\|_{2,2}$ is equal to the top eigenvalue of A. Except for the matrix norm $\|\cdot\|_{2,2}$ all vector norms are ℓ_2, $\|\cdot\| = \|\cdot\|_2$. Finally, let the orthonormal operator projecting into the subspace spanned by the columns of $X_{\mathcal{F}}$ be defined as $\Pi_{\mathcal{F}} = X_{\mathcal{F}}(X_{\mathcal{F}}^T X_{\mathcal{F}})^{-1} X_{\mathcal{F}}^T$.

II. Results

For the observational model in (1), assume that the true sparsity model is $\mathcal{T}$; as a result

$$y = X_{\mathcal{T}} \beta_{\mathcal{T}} + \epsilon. \tag{4}$$

[3]The maximum-to-average ratio of β was defined as $k\beta_{\min}^2 / \|\beta\|_2^2$.

We first state a result on the probability of the event $\mathcal{S}(\hat{\beta}) = \mathcal{F}$, i.e., $\Pr[\mathcal{S}(\hat{\beta}) = \mathcal{F} | X, \beta, \mathcal{T}]$, for any $\mathcal{F} \neq \mathcal{T}$ and any measurement matrix X.

Theorem 1: For the observational model of (4) and estimate $\hat{\beta}$ in (2), the following bound holds:

$$\Pr\left[\mathcal{S}(\hat{\beta}) = \mathcal{F} | X, \beta, \mathcal{T}\right] \leq \exp\left\{-\frac{C}{2}\left\|(I - \Pi_{\mathcal{F}}) X_{\mathcal{T}-\mathcal{F}} \beta_{\mathcal{T}-\mathcal{F}}\right\|^2 + \frac{|\mathcal{T} - \mathcal{F}|}{2}\right\}$$

where $C = 3 - 2\sqrt{2}$.

The proof of Theorem 1, given in Section III, employs the Chernoff technique and the properties of the eigenvalues of the difference of projection matrices, to bound the probability of declaring a wrong sparsity pattern $\mathcal{F}$ instead of the true one $\mathcal{T}$ as function of the measurement matrix X and the true parameter β. The error rate decreases exponentially in the norm of the projection of $X_{\mathcal{T}-\mathcal{F}}\beta_{\mathcal{T}-\mathcal{F}}$ on the orthogonal subspace spanned by the columns of $X_{\mathcal{F}}$. This is in agreement with the intuition that the closer different subspaces corresponding to different sets of columns of X are, the harder it is to differentiate them, and hence the higher the error probability will be.

The theorem below gives a nonasymptotic bound on the probability of the event that the declared sparsity pattern $\mathcal{S}(\hat{\beta})$ differs from the true sparsity pattern $\mathcal{T}$ in no more than d indices, when the entries of the measurement matrix X are drawn i.i.d. from a standard normal distribution. It is clear that by letting $d = 1$ we obtain an upper bound on the error probability of exact sparsity pattern recovery.

Theorem 2: Suppose that for the observational model of (4) and the estimate $\hat{\beta}$ in (2) the entries of X are i.i.d. $\mathcal{N}(0,1)$ and $p > 2k$. If we have the equation shown at the bottom of the page, where

$$f_0(d, p, k, \beta_{\min}) := \frac{d \log\left(\frac{k(p-k)}{d^2}\right) + d}{\log(1 + Cd\beta_{\min}^2)}$$

$$f_1(k, \beta_{\min}) := 4k\left(1 + \frac{1}{Ck\beta_{\min}^2}\right)^2$$

$$f_2(p, k, \beta_{\min}) := \left(1 + \frac{1}{C\beta_{\min}^2}\right)[1 + 2\log(k(p-k))]$$

then

$$\Pr\left[|\mathcal{S}(\hat{\beta}) - \mathcal{T}| \geq d\right] < k \max\left\{\left[\frac{ek(p-k)}{d^2}\right]^{-B^\star d}, \left[\frac{e(p-k)}{k}\right]^{-B^\star k}\right\}$$

where $B^\star = \frac{B-5}{2}$ and $C = 3 - 2\sqrt{2}$.

The key elements in the proof include Theorem 1, application of union bounds (a fairly standard technique which has been

$$n - k > \max\left\{B\, f_0(d, p, k, \beta_{\min}), B\, f_0(k, p, k, \beta_{\min}), f_1(k, \beta_{\min}), f_2(p, k, \beta_{\min})\right\}$$

used before for this problem [4], [5], [7]), asymptotic behavior of binomial coefficients and properties of convex functions.

Note that in the linear regime, i.e., $k = \Theta(p)$, with $n = \Theta(p)$ and $k\beta_{\min}^2 = \Theta(1)$ the probability of misidentifying more than any fraction (less than one) goes to zero exponentially fast as $p \to \infty$. In words, if the SNR is fixed while the dimension of the signal increases unboundedly, it is still possible to recover reliably some fraction of the support. This is in agreement with previous results on partial sparsity pattern recovery [5], [9].

If we let $n(p)$, $k(p)$, and $\beta_{\min}(p)$ scale as a function of p, then the upper bound of $\Pr[\mathcal{S}(\hat{\beta}) \neq \mathcal{T}]$ scales like $k(p-k)^{-B^\star}$. For $B^\star > 2$ or equivalently $B > 9$, the probability of error as $p \to \infty$ is bounded above by p^{-D} for some $D > 1$. Therefore

$$\sum_{p=1}^{\infty} \Pr[\mathcal{S}(\hat{\beta}_{p\times 1}) \neq \mathcal{T}_p] \tag{5}$$

is finite and as a consequence of the Borel-Cantelli Lemma, for large enough p, the decoder declares the true sparsity pattern almost surely. In other words, the estimate $\hat{\beta}$ based on (2) achieves the same loss as an oracle which is supplied with perfect information about which coefficients of β are nonzero. The following corollary summarizes the aforementioned statements.

Corollary 3: For the observational model of (4) and the estimate $\hat{\beta}$ in (2), let n, k and $\beta_{\min}^2$ scale as a function of p. Then there exists a constant $C^\star$ such that if $\beta_{\min}^2$ is $\Omega\left(\frac{1}{k}\right)$ and $O(1)$, and

$$n > C^\star \max\left\{\frac{\log(p-k)}{\log\left(1+\beta_{\min}^2\right)}, \frac{k\log\left(\frac{p}{k}\right)}{\log(1+k\beta_{\min}^2)}, k\right\}$$

then a.s. for large enough p, $\hat{\beta}$ achieves the same performance loss as an oracle which is supplied with perfect information about which coefficients of β are nonzero and $\mathcal{S}(\hat{\beta}) = \mathcal{T}$.

Remarks:

- $\beta_{\min}^2 = O(1)$ is required to ensure that for a sufficiently large $C^\star$, we have $C^\star f_0(1, p, k, \beta_{\min}) > f_2(p, k, \beta_{\min})$ where f_0 and f_2 are defined in Theorem 1.
- $\beta_{\min}^2 = \Omega\left(\frac{1}{k}\right)$ is required to ensure that for a sufficiently large $C^\star$, we have $C^\star k > f_1(k, \beta_{\min})$ where f_1 is defined in Theorem 1.

The sufficient conditions in Corollary 3 can be compared against similar conditions for exact sparsity pattern recovery in [4]–[7]; for example, in the sublinear regime $k = o(p)$, when $\beta_{\min}^2 = \Theta(1)$, [4], [7] proved that $n = \Theta(k\log\left(\frac{p}{k}\right))$ is sufficient, and [5], [6] proved that $n = \Theta(k\log(p-k))$ is sufficient. In that vein, according to Corollary 3

$$n = \max\left\{\Theta\left(\frac{k\log\left(\frac{p}{k}\right)}{\log k}\right), \Theta(k)\right\}$$

suffices to ensure exact sparsity pattern recovery; therefore, it strengthens these earlier results.

What remains is to see whether the sufficient conditions in Corollary 3 match the necessary conditions proved in [8]:

Theorem 4 [8]: Suppose that the entries of the measurement matrix $X \in \mathbb{R}^{n\times p}$ are drawn i.i.d. from any distribution with zero-mean and variance one. Then a necessary condition for asymptotically reliable recovery is that

$$n > \max\left\{f_1\left(k, p, \beta_{\min}^2\right), f_2\left(k, p, \beta_{\min}^2\right), k-1\right\}$$

where

$$f_1\left(k, p, \beta_{\min}^2\right) = \frac{\log\binom{p}{k} - 1}{\frac{1}{2}\log\left(1 + k\beta_{\min}^2\left(1 - \frac{k}{p}\right)\right)}$$

$$f_2\left(k, p, \beta_{\min}^2\right) = \frac{\log(p-k+1) - 1}{\frac{1}{2}\log\left(1 + \beta_{\min}^2\left(1 - \frac{1}{p-k+1}\right)\right)}.$$

The necessary condition in Theorem 4 asymptotically resembles the sufficient condition in Corollary 3; recall that $\log\binom{p}{k} < k\log\left(\frac{ep}{k}\right)$. The sufficient conditions of Corollary 3 can be compared against the necessary conditions in [8] for exact sparsity pattern recovery, as shown in Table I. The first paper to establish the sufficient conditions in row 1 and row 4 of Table I is [10]. The sufficient conditions presented in the first four rows of Table I are a consequence of past work [4], also recovered by Corollary 3. The new stronger result in this paper provides the sufficient conditions in row 5 and 6, which did not appear in previous studies [4]–[7], and match the previous necessary conditions presented in [8]. (It is worth reminding that these results are restricted to $\beta_{\min}^2 = O(1)$ and $\beta_{\min}^2 = \Omega\left(\frac{1}{k}\right)$.).

III. Proof of Theorem 1

We first state three basic lemmas.

Lemma 5: If any $2k$ columns of the $n \times p$ matrix X are linearly independent then for any sparsity pattern $\mathcal{T}$ and $\mathcal{F}$ such that $|\mathcal{T}| = |\mathcal{F}| = k$ the difference of projection matrices $\Pi_{\mathcal{F}} - \Pi_{\mathcal{T}}$ has $d = |\mathcal{T} - \mathcal{F}|$ pairs of nonzero positive and negative eigenvalues, bounded above by one and bounded below by negative one, respectively, and equal in magnitude.

Lemma 6: For $y \sim \mathcal{N}(\mu, I)$ and $\|2t\Psi\|_{2,2} < 1$, we have

$$\mathrm{E}[e^{ty^T\Psi y}] = \frac{e^{t\mu^T\Psi\mu + 2t^2\mu^T\Psi(I-2t\Psi)^{-1}\Psi\mu}}{\det(I-2t\Psi)^{\frac{1}{2}}}.$$

Lemma 7: For $\Psi = \Pi_{\mathcal{F}} - \Pi_{\mathcal{T}}$ and $d = |\mathcal{T} - \mathcal{F}|$, we have

$$\log\det(I-2t\Psi) \geq d\log(1-4t^2)$$
$$\|(I-2t\Psi)^{-1/2}\|_{2,2}^2 \leq (1-2t)^{-1}.$$

We defer the proofs of the lemmas 5 and 7 to after the proof of Theorem 1. Lemma 6 follows standard Gaussian integrals [12].

A. Proof of Theorem 1

For a given sparsity pattern $\mathcal{F}$, the minimum residual sum of squares is achieved by

$$\min_{\theta_{\mathcal{F}}\in\mathbb{R}^k} \|y - X_{\mathcal{F}}\theta_{\mathcal{F}}\|^2 = \|y - \Pi_{\mathcal{F}}y\|^2$$

where $\Pi_{\mathcal{F}}$ denotes the orthogonal projection operator into the column space of $X_{\mathcal{F}}$; that is, among all sparsity patterns with size k, the optimum decoder declares

$$\hat{\mathcal{T}}(y, X) = \arg\min_{|\mathcal{F}|=k} \|y - \Pi_{\mathcal{F}}y\|^2$$

TABLE I
NECESSARY AND SUFFICIENT CONDITIONS ON THE NUMBER OF MEASUREMENTS n REQUIRED FOR RELIABLE SUPPORT RECOVERY IN THE LINEAR AND THE SUBLINEAR REGIME. THE SUFFICIENT CONDITIONS PRESENTED IN THE FIRST FOUR ROWS ARE A CONSEQUENCE OF PAST WORK [4], ALSO RECOVERED BY COROLLARY 3. THE NEW STRONGER RESULT IN THIS PAPER PROVIDES THE SUFFICIENT CONDITIONS IN ROW 5 AND 6, WHICH DID NOT APPEAR IN PREVIOUS STUDIES [4]–[7], AND MATCH THE NECESSARY CONDITIONS PRESENTED IN [8]

Scaling	Sufficient condition Corollary 3	Necessary condition Theorem 4 [8]
$k=\Theta(p)$ $\beta_{\min}^2=\Theta(\frac{1}{k})$	$n=\Theta(p\log p)$	$n=\Theta(p\log p)$
$k=\Theta(p)$ $\beta_{\min}^2=\Theta(\frac{\log k}{k})$	$n=\Theta(p)$	$n=\Theta(p)$
$k=\Theta(p)$ $\beta_{\min}^2=\Theta(1)$	$n=\Theta(p)$	$n=\Theta(p)$
$k=o(p)$ $\beta_{\min}^2=\Theta(\frac{1}{k})$	$n=\Theta(k\log(p-k))$	$n=\Theta(k\log(p-k))$
$k=o(p)$ $\beta_{\min}^2=\Theta(\frac{\log k}{k})$	$n=\max\left\{\Theta(\frac{k\log(p-k)}{\log k}),\Theta\left(\frac{k\log(\frac{p}{k})}{\log\log k}\right)\right\}$	$n=\max\left\{\Theta(\frac{k\log(p-k)}{\log k}),\Theta\left(\frac{k\log(\frac{p}{k})}{\log\log k}\right)\right\}$
$k=o(p)$ $\beta_{\min}^2=\Theta(1)$	$n=\max\left\{\Theta\left(\frac{k\log(\frac{p}{k})}{\log k}\right),\Theta(k)\right\}$	$n=\max\left\{\Theta\left(\frac{k\log(\frac{p}{k})}{\log k}\right),\Theta(k)\right\}$

as the optimum estimate of the true sparsity pattern in terms of minimum error probability. Recall the definition of $\hat{\beta}$ in (2) and note that $\mathcal{S}(\hat{\beta})=\hat{\mathcal{T}}(y,X)$. If the decoder incorrectly declares $\mathcal{F}$ instead of the true sparsity pattern (namely $\mathcal{T}$), then

$$\|y-\Pi_{\mathcal{F}}y\|^2<\|y-\Pi_{\mathcal{T}}y\|^2$$

or equivalently

$$Z_{\mathcal{F}}:=y^T(\Pi_{\mathcal{F}}-\Pi_{\mathcal{T}})y>0.$$

The probability that the optimal decoder declares wrongly the sparsity pattern $\mathcal{F}$ instead of the true sparsity pattern $\mathcal{T}$ is less than the probability that $Z_{\mathcal{F}}>0$. With the aid of the Chernoff technique an upper bound on the probability that $Z_{\mathcal{F}}>0$ is obtained

$$\Pr[Z_{\mathcal{F}}>0|X,\mathcal{T},\beta]\leq\inf_{|t|<1/2}\mathrm{E}[e^{Z_{\mathcal{F}}t}|X,\mathcal{T},\beta].$$

Note that $Z_{\mathcal{F}}$ is a random variable that has a quadratic form in Gaussian random vectors. This allows us to use standard Gaussian integrals to calculate $\mathrm{E}[e^{Z_{\mathcal{F}}t}|X,\mathcal{T},\beta]$. In order to bound the expectation, $|t|$ is required to be bounded which is a necessary condition in Lemma 6. From Lemma 6, we learned that

$$\log\mathrm{E}[e^{Z_{\mathcal{F}}t}]$$
$$=2t^2\mu^T\Psi(I-2t\Psi)^{-1}\Psi\mu+t\mu^T\Psi\mu-\frac{1}{2}\log\det(I-2t\Psi)\quad(6)$$

where we made the following abbreviations:

$$\mu=X_{\mathcal{T}}\beta_{\mathcal{T}}$$
$$\Psi=\Pi_{\mathcal{F}}-\Pi_{\mathcal{T}}.$$

For Lemma 6, we need $\|2t\Psi\|_{2,2}<1$ and we prove in Lemma 5 that the eigenvalues of Ψ are bounded in absolute value by one; consequently, (6) holds for $|2t|<1$.

With the aid of the definition of the ℓ_2 norm of matrices and applying it to $\|(I-2t\Psi)^{-1/2}\Psi\mu\|^2$ the first term in the r.h.s. of (6) can be bounded as follows:

$$2t^2\mu^T\Psi(I-2t\Psi)^{-1}\Psi\mu\leq 2t^2\|(I-2t\Psi)^{-1/2}\|_{2,2}^2\mu^T\Psi^2\mu.\quad(7)$$

Since μ lies in the subspace spanned by the columns of $X_{\mathcal{T}}$ we have

$$\begin{aligned}\Pi_{\mathcal{T}}\mu&=\mu\text{ and}\\(\Pi_{\mathcal{T}}-\Pi_{\mathcal{F}})\mu&=(\Pi_{\mathcal{T}}-\Pi_{\mathcal{F}})X_{\mathcal{T}-\mathcal{F}}\beta_{\mathcal{T}-\mathcal{F}}\\&=(I-\Pi_{\mathcal{F}})X_{\mathcal{T}-\mathcal{F}}\beta_{\mathcal{T}-\mathcal{F}}\end{aligned}$$

which yields the following:

$$\begin{aligned}\mu^T\Psi\mu&=-\|(I-\Pi_{\mathcal{F}})X_{\mathcal{T}}\beta_{\mathcal{T}}\|^2\\&=-\|(I-\Pi_{\mathcal{F}})X_{\mathcal{T}-\mathcal{F}}\beta_{\mathcal{T}-\mathcal{F}}\|^2\end{aligned}$$

and similarly

$$\mu^T\Psi^2\mu=\|(I-\Pi_{\mathcal{F}})X_{\mathcal{T}-\mathcal{F}}\beta_{\mathcal{T}-\mathcal{F}}\|^2.$$

The aforementioned equations and the inequality (7) yields the upper bound shown in (8), as found at the bottom of the page. Lemma 7 introduces an upper bound for $\|(I-2t\Psi)^{-1/2}\|_{2,2}^2$ and a lower bound for $\log\det(I-2t\Psi)$ that can be used to further simplify the upper bound of $\log\mathrm{E}[e^{Z_{\mathcal{F}}t}]$. The main ingredient in the proof of Lemma 7 is the eigenvalue properties of $\Pi_{\mathcal{F}}-$

$$\begin{aligned}\log\mathrm{E}[e^{Z_{\mathcal{F}}t}]&\leq 2t^2\|(I-2t\Psi)^{-1/2}\|_{2,2}^2\mu^T\Psi^2\mu+t\mu^T\Psi\mu-\frac{1}{2}\log\det(I-2t\Psi)\\&=\left\{2t^2\|(I-2t\Psi)^{-1/2}\|_{2,2}^2-t\right\}\|(I-\Pi_{\mathcal{F}})X_{\mathcal{T}-\mathcal{F}}\beta_{\mathcal{T}-\mathcal{F}}\|^2-\frac{1}{2}\log\det(I-2t\Psi)\end{aligned}\quad(8)$$

$\Pi_{\mathcal{T}}$ that were established in Lemma 5. Substituting the bounds obtained in Lemma 7 in (8), we have

$$\log \mathrm{E}[e^{Z_{\mathcal{F}} t}] \leq \left[\frac{2t^2}{1-2t} - t\right] \|(I-\Pi_{\mathcal{F}})X_{\mathcal{T}-\mathcal{F}}\beta_{\mathcal{T}-\mathcal{F}}\|^2 - \frac{d}{2}\log(1-4t^2). \quad (9)$$

Finally, to prove Theorem 1, we take the infimum of $\frac{2t^2}{1-2t} - t$ over $|t| < 1/2$ which is equal to $\sqrt{2} - 3/2$ at $t^\star = 1/2(1 - \sqrt{2}/2)$ and obtain the desired bound as shown in the equation at the bottom of the page. □

Now we prove the remaining lemmas.

Proof of Lemma 5: Before we prove the result, let us introduce some notations.

- For any $\mathcal{F} \in \{1, 2, \ldots, p\}$, $V_{\mathcal{F}}$ is defined as the linear subspace spanned by the columns of $X_{\mathcal{F}}$,
- $V_{\mathcal{F}}^{\perp}$ stands for the subspace orthogonal to $V_{\mathcal{F}}$,
- $\tilde{V}_{\mathcal{F}}$ and $\tilde{V}_{\mathcal{T}}$ stand for $V_{\mathcal{F}} \cap (V_{\mathcal{F}} \cap V_{\mathcal{T}})^{\perp}$ and $V_{\mathcal{T}} \cap (V_{\mathcal{T}} \cap V_{\mathcal{F}})^{\perp}$, respectively,
- and finally for any subspace V, Π_V designates the orthogonal projection onto V. (With a slight abuse of notation, for any sparsity pattern $\mathcal{F}$, we use $\Pi_{\mathcal{F}}$ instead of $\Pi_{V_{\mathcal{F}}}$).

It is worthwhile noting that $\tilde{V}_{\mathcal{F}} \cap \tilde{V}_{\mathcal{T}}$ is empty. From [13, Lemma 4.1], for any $V_{\mathcal{T}}$ and $V_{\mathcal{F}}$ in $\mathbb{R}^n$, it holds that

$$V_{\mathcal{T}} = \tilde{V}_{\mathcal{T}} \oplus (V_{\mathcal{T}} \cap V_{\mathcal{F}})$$
$$V_{\mathcal{F}} \cup V_{\mathcal{T}} = \tilde{V}_{\mathcal{T}} \oplus \tilde{V}_{\mathcal{F}} \oplus (V_{\mathcal{T}} \cap V_{\mathcal{F}}) \quad (10)$$
$$V_{\mathcal{T}} \cap V_{\mathcal{F}} \perp \tilde{V}_{\mathcal{F}} \oplus \tilde{V}_{\mathcal{T}} \quad (11)$$

which yields

$$\Pi_{\mathcal{F}} = \Pi_{\tilde{V}_{\mathcal{F}}} + \Pi_{V_{\mathcal{F}} \cap V_{\mathcal{T}}}$$
$$\Pi_{\mathcal{T}} = \Pi_{\tilde{V}_{\mathcal{T}}} + \Pi_{V_{\mathcal{F}} \cap V_{\mathcal{T}}}.$$

Consequently

$$\Pi_{\mathcal{F}} - \Pi_{\mathcal{T}} = \Pi_{\tilde{V}_{\mathcal{F}}} - \Pi_{\tilde{V}_{\mathcal{T}}}. \quad (12)$$

Since any set of columns of X with size less or equal to $2k$ are independent, for any $\mathcal{T}$ and $\mathcal{F}$ such that $|\mathcal{T}| = |\mathcal{F}| = k$ and $|\mathcal{F} - \mathcal{T}| = d$, we have

$$V_{\mathcal{F}} \cap V_{\mathcal{T}} = V_{\mathcal{F} \cap \mathcal{T}},$$
$$V_{\mathcal{F}} \cup V_{\mathcal{T}} = V_{\mathcal{F} \cup \mathcal{T}}$$

and

$$\dim(V_{\mathcal{F} \cap \mathcal{T}}) = |\mathcal{F} \cap \mathcal{T}| = k - d \quad (13)$$
$$\dim(V_{\mathcal{F} \cup \mathcal{T}}) = |\mathcal{F} \cup \mathcal{T}| = k + d \quad (14)$$

therefore

$$\begin{aligned}\dim(\tilde{V}_{\mathcal{F}}) &= \dim(V_{\mathcal{F}}) - \dim(V_{\mathcal{F}} \cap V_{\mathcal{T}}) \\ &= \dim(V_{\mathcal{F}}) - \dim(V_{\mathcal{F} \cap \mathcal{T}}) = k - (k - d) \\ &= d = \dim(\tilde{V}_{\mathcal{T}}).\end{aligned}$$

The dimension of $(\tilde{V}_{\mathcal{F}} \cup \tilde{V}_{\mathcal{T}})^{\perp}$ which is the null space of $\Pi_{\mathcal{F}} - \Pi_{\mathcal{T}}$ is equal to

$$\dim\left((\tilde{V}_{\mathcal{F}} \cup \tilde{V}_{\mathcal{T}})^{\perp}\right) = n - \dim(\tilde{V}_{\mathcal{F}}) - \dim(\tilde{V}_{\mathcal{T}}) = n - 2d.$$

We just proved that $\Pi_{\mathcal{F}} - \Pi_{\mathcal{T}}$ has $n - 2d$ eigenvalues with eigenvalue zero. The range of $\Pi_{\mathcal{F}} - \Pi_{\mathcal{T}}$ is the $2d$ dimensional space $\tilde{V}_{\mathcal{F}} \cup \tilde{V}_{\mathcal{T}}$. Therefore, $\Pi_{\mathcal{F}} - \Pi_{\mathcal{T}}$ has $2d$ nonzero eigenvalues with absolute value less or equal to one (The eigenvalues of $\Pi_{\mathcal{F}} - \Pi_{\mathcal{T}}$ are equal to one only if $\tilde{V}_{\mathcal{F}} \perp \tilde{V}_{\mathcal{T}}$.)

If $v_{(\lambda)}$ is an eigenvector of $\Pi_{\mathcal{F}} - \Pi_{\mathcal{T}}$ with eigenvalue λ, then we have

$$(\Pi_{\mathcal{F}} - \Pi_{\mathcal{T}})v_{(\lambda)} = \lambda v_{(\lambda)}.$$

Next, we prove that the vector

$$v_{(-\lambda)} = v_{(\lambda)} - (\Pi_{\mathcal{F}} + \Pi_{\mathcal{T}})v_{(\lambda)}$$

is an eigenvector of $\Pi_{\mathcal{F}} - \Pi_{\mathcal{T}}$ with eigenvalue $-\lambda$. The proof presented in the following exploits the definition of the eigenvector $v_{(\lambda)}$:

$$\begin{aligned}(\Pi_{\mathcal{F}} - \Pi_{\mathcal{T}})v_{(-\lambda)} &= (\Pi_{\mathcal{F}} - \Pi_{\mathcal{T}})\left(v_{(\lambda)} - (\Pi_{\mathcal{F}} + \Pi_{\mathcal{T}})v_{(\lambda)}\right) \\ &= (\Pi_{\mathcal{F}} - \Pi_{\mathcal{T}})v_{(\lambda)} - (\Pi_{\mathcal{F}} - \Pi_{\mathcal{T}})(\Pi_{\mathcal{F}} + \Pi_{\mathcal{T}})v_{(\lambda)} \\ &= \lambda v_{(\lambda)} - (\Pi_{\mathcal{F}} + \Pi_{\mathcal{F}}\Pi_{\mathcal{T}} - \Pi_{\mathcal{T}}\Pi_{\mathcal{F}} - \Pi_{\mathcal{T}})\, v_{(\lambda)} \\ &= -\Pi_{\mathcal{F}}\Pi_{\mathcal{T}}v_{(\lambda)} + \Pi_{\mathcal{T}}\Pi_{\mathcal{F}}v_{(\lambda)} \\ &= -\Pi_{\mathcal{F}}(\Pi_{\mathcal{F}} - \lambda)v_{(\lambda)} + \Pi_{\mathcal{T}}(\Pi_{\mathcal{T}} + \lambda)v_{(\lambda)} \\ &= -\Pi_{\mathcal{F}}(1 - \lambda)v_{(\lambda)} + \Pi_{\mathcal{T}}(1 + \lambda)v_{(\lambda)} \\ &= -(\Pi_{\mathcal{F}} - \Pi_{\mathcal{T}})v_{(\lambda)} + \lambda(\Pi_{\mathcal{F}} + \Pi_{\mathcal{T}})v_{(\lambda)} \\ &= -\lambda v_{(\lambda)} + \lambda(\Pi_{\mathcal{F}} + \Pi_{\mathcal{T}})v_{(\lambda)} \\ &= -\lambda v_{(-\lambda)}.\end{aligned}$$

This means that for every eigenvector $v_{(\lambda)}$ with eigenvalue λ there exist another eigenvector $v_{(-\lambda)}$ with eigenvalue $-\lambda$.

Proof of Lemma 7: From Lemma 5, we know that $\Pi_{\mathcal{F}} - \Pi_{\mathcal{T}}$ has d pairs of nonzero positive and negative eigenvalues, whose magnitudes are equal. Let the positive eigenvalues be denoted by $\lambda_1, \ldots, \lambda_d$, then

$$\begin{aligned}\log\det(I - 2t\Psi) &= \sum_{i=1}^{d}\{\log(1 - 2t\lambda_i) + \log(1 + 2t\lambda_i)\} \\ &= \sum_{i=1}^{d}\log(1 - 4t^2\lambda_i^2).\end{aligned}$$

Since, the eigenvalues are bounded by one, again by Lemma 5, $\log(1 - 4t^2\lambda_i^2)$ is lower bounded by $\log(1 - 4t^2)$; consequently

$$\log\det(I - 2t\Psi) \leq d\log(1 - 4t^2).$$

$$\begin{aligned}\inf_{|t|<1/2} \log \mathrm{E}[e^{Z_{\mathcal{F}} t}] &\leq -\frac{3 - 2\sqrt{2}}{2}\|(I - \Pi_{\mathcal{F}})X_{\mathcal{T}-\mathcal{F}}\beta_{\mathcal{T}-\mathcal{F}}\|^2 - \frac{d}{2}\log(\sqrt{2} - 1/2) \\ &\leq -\frac{3 - 2\sqrt{2}}{2}\|(I - \Pi_{\mathcal{F}})X_{\mathcal{T}-\mathcal{F}}\beta_{\mathcal{T}-\mathcal{F}}\|^2 + \frac{d}{2}\end{aligned}$$

To prove $\|(I-2t\Psi)^{-1/2}\|_{2,2}^2 \le (1-2t)^{-1}$, note that $(I-2t\Psi)^{-1/2}$ has:

- d eigenvalues equal to $(1-2t\lambda_1)^{-1/2},\dots,(1-2t\lambda_d)^{-1/2}$;
- d eigenvalues equal to $(1+2t\lambda_1)^{-1/2},\dots,(1+2t\lambda_d)^{-1/2}$;
- and $n-2d$ eigenvalues equal to one.

It is not hard to see that because $2t<1$ and $\lambda_i<1$ the top eigenvalue of $(I-2t\Psi)^{-1/2}$ is bounded above by $(I-2t)^{-1/2}$ and hence,

$$\|(I-2t\Psi)^{-1/2}\|_{2,2}^2 \le (1-2t)^{-1}.$$

IV. Proof of Theorem 2

We state two simple lemmas used to prove Theorem 2.

Lemma 8: For Gaussian measurement matrices, with $X_{ij}\sim\mathcal{N}(0,1)$ the average error probability that the optimum decoder declares $\mathcal{F}$ is bounded by

$$\Pr[\hat{\mathcal{T}}(y,X)=\mathcal{F}|\beta,\mathcal{T}] \le \exp\left\{-\frac{n-k}{2}\log\left(1+C\|\beta_{\mathcal{T}-\mathcal{F}}\|^2\right)+\frac{|\mathcal{T}-\mathcal{F}|}{2}\right\}$$

where $C=3-2\sqrt{2}$.

Lemma 9: For the function

$$g(r):=r\left[\frac{5}{2}+\log\left(\frac{k(p-k)}{r^2}\right)\right]-\frac{n-k}{2}\log(1+r\gamma)$$

defined on positive integers if

$$n-k>\max\left\{4k\left(1+\frac{1}{k\gamma}\right)^2,\left(1+\frac{1}{\gamma}\right)[1+2\log(k(p-k))]\right\} \tag{15}$$

then

$$\max_{r=d,\dots,k} g(r)\le\max\{g(d),g(k)\}.$$

Before we prove the two lemmas, let us see how they imply Theorem 2.

A. Proof of Theorem 2

In order to find conditions under which $\Pr[|\mathcal{S}(\hat\beta)-\mathcal{T}|\ge d]$ asymptotically goes to zero, we exploit the union bound in conjunction with counting arguments and the previously stated two lemmas.

First, note that the event $|\mathcal{S}(\hat\beta)-\mathcal{T}|\ge d$ can be written as the union of the events $\mathcal{S}(\hat\beta)=\mathcal{F}$ for all sparsity patterns $\mathcal{F}$ such that $|\mathcal{F}-\mathcal{T}|\ge d$. The union bound allows us to bound the probability of the event $\cup_{|\mathcal{F}-\mathcal{T}|\ge d}\{\mathcal{S}(\hat\beta)=\mathcal{F}\}$ by the sum of probabilities of events like $\mathcal{S}(\hat\beta)=\mathcal{F}$. In mathematical terms

$$\Pr[|\mathcal{S}(\hat\beta)-\mathcal{T}|\ge d]=\Pr\left[\cup_{|\mathcal{F}-\mathcal{T}|\ge d}\{\mathcal{S}(\hat\beta)=\mathcal{F}\}\right] \le\sum_{r=d}^{k}\sum_{|\mathcal{F}-\mathcal{T}|=r}\Pr\left[\mathcal{S}(\hat\beta)=\mathcal{F}\right].$$

Lemma 8 which is based on generating functions of chi-square distributions introduces an upper bound for the event $\mathcal{S}(\hat\beta)=\mathcal{F}$; namely

$$\Pr[\hat{\mathcal{T}}(y,X)=\mathcal{F}|\beta,\mathcal{T}]\le e^{-\frac{n-k}{2}\log(1+C\|\beta_{\mathcal{T}-\mathcal{F}}\|^2)+\frac{|\mathcal{T}-\mathcal{F}|}{2}}$$

with $C=3-2\sqrt{2}$. If we replace $C\|\beta_{\mathcal{T}-\mathcal{F}}\|^2$ with the lower bound $C|\mathcal{T}-\mathcal{F}|\beta_{\min}^2$ which follows the definition of $\beta_{\min}$ we obtain an upper bound for the event $\mathcal{S}(\hat\beta)=\mathcal{F}$ that does not depend on $\mathcal{F}$ as long as $|\mathcal{F}-\mathcal{T}|$ is fixed. The number of sparsity patterns $\mathcal{F}$ that are different from $\mathcal{T}$ in exactly r elements is $\binom{k}{r}\binom{p-k}{r}$. Therefore, we can bound $\sum_{|\mathcal{F}-\mathcal{T}|=r}\Pr\left[\mathcal{S}(\hat\beta)=\mathcal{F}\right]$ by $\binom{k}{r}\binom{p-k}{r}e^{-\frac{n-k}{2}\log(1+Cr\beta_{\min}^2)+\frac{r}{2}}$. To summarize, exploiting inequality $\log\binom{a}{b}<b\log\left(\frac{ae}{b}\right)$, we have

$$\Pr[|\mathcal{S}(\hat\beta)-\mathcal{T}|\ge d]\le\sum_{r=d}^{k}e^{r\left[\frac{5}{2}+\log\left(\frac{k(p-k)}{r^2}\right)\right]-\frac{n-k}{2}\log\left(1+Cr\beta_{\min}^2\right)}. \tag{16}$$

Let $g(r)$ stand for the exponent in the previous equation

$$g(r):=r\left[\frac{5}{2}+\log\left(\frac{k(p-k)}{r^2}\right)\right]-\frac{n-k}{2}\log(1+r\gamma)$$

where we defined

$$\gamma:=C\beta_{\min}^2.$$

From Lemma 9, we know that if

$$n-k>\max\left\{4k\left(1+\frac{1}{k\gamma}\right)^2,\left(1+\frac{1}{\gamma}\right)[1+2\log(k(p-k))]\right\} \tag{17}$$

then $\max_{r=d,\dots,k}g(r)\le\max\{g(d),g(k)\}$ and therefore

$$\Pr[|\mathcal{S}(\hat\beta)-\mathcal{T}|\ge d]\le\sum_{r=d}^{k}e^{g(r)} \le ke^{\max\{g(d),g(k)\}}. \tag{18}$$

For $\Pr[|\mathcal{S}(\hat\beta)-\mathcal{T}|\ge d]\to 0$, it suffices that $g(d)$ and $g(k)$ go to $-\infty$ fast enough. In the statement of Theorem 2, we have the following condition:

$$n-k>B\frac{d\log\left(\frac{k(p-k)}{d^2}\right)+d}{\log(1+d\gamma)}$$

that results in the following upper bound:

$$g(d)=d\left[\frac{5}{2}+\log\left(\frac{k(p-k)}{d^2}\right)\right]-\frac{n-k}{2}\log(1+d\gamma) \tag{19}$$

$$\le d\left[\frac{5}{2}+\log\left(\frac{k(p-k)}{d^2}\right)\right]-\frac{B}{2}\left[d\log\left(\frac{k(p-k)}{d^2}\right)+d\right]$$

$$\le-\frac{B-5}{2}\left[d\log\left(\frac{k(p-k)}{d^2}\right)+d\right]. \tag{20}$$

Hence, if

$$n-k > B\max\left\{\frac{d\log\left(\frac{k(p-k)}{d^2}\right)+d}{\log(1+d\gamma)}, \frac{k\log\left(\frac{k(p-k)}{k^2}\right)+k}{\log(1+k\gamma)}\right\} \tag{21}$$

then we get the equation shown at the bottom of the page. Therefore, inequalities (17) and (21), which are the main conditions in Theorem 1, imply that

$$\Pr[|\mathcal{S}(\hat{\beta})-\mathcal{T}|\geq d] < k\max\left\{\left[\frac{ek(p-k)}{d^2}\right]^{-dB^\star}, \left[\frac{e(p-k)}{k}\right]^{-kB^\star}\right\}$$

where $B^\star=\frac{B-5}{2}$. □

Now we prove the remaining lemmas.

Proof of Lemma 8: The columns of $X_{\mathcal{F}}$ and $X_{\mathcal{T}-\mathcal{F}}$ are, by definition, disjoint and therefore independent Gaussian random matrices with column spaces spanning random independent $|\mathcal{F}|$-and $|\mathcal{T}-\mathcal{F}|$-dimensional subspaces, respectively. The Gaussian random vector $X_{\mathcal{T}-\mathcal{F}}\beta_{\mathcal{T}-\mathcal{F}}$ has i.i.d. Gaussian entries with variance $\|\beta_{\mathcal{T}-\mathcal{F}}\|^2$. Therefore, we conclude that, since the random Gaussian vector $X_{\mathcal{T}-\mathcal{F}}\beta_{\mathcal{T}-\mathcal{F}}$ is projected onto the subspace orthogonal to the random column space of $X_{\mathcal{F}}$, the quantity $\|(I-\Pi_{\mathcal{F}})X_{\mathcal{T}-\mathcal{F}}\beta_{\mathcal{T}-\mathcal{F}}\|^2/\|\beta_{\mathcal{T}-\mathcal{F}}\|^2$ is a chi-square random variable with $n-k$ degrees of freedom. Thus

$$\begin{aligned}\Pr[\hat{\mathcal{T}}(y,X)=\mathcal{F}|\beta,\mathcal{T}] &= \mathrm{E}_X\left\{\Pr[\hat{\mathcal{T}}(y,X)=\mathcal{F}|X,\beta,\mathcal{T}]\right\}\\ &\overset{1}{\leq} \mathrm{E}_X\left\{e^{-\frac{C}{2}\|(I-\Pi_{\mathcal{F}})X_{\mathcal{T}-\mathcal{F}}\beta_{\mathcal{T}-\mathcal{F}}\|^2+\frac{d}{2}}\right\}\\ &= \mathrm{E}_{W\sim\chi^2_{n-k}} e^{-\frac{C}{2}W\|\beta_{\mathcal{T}-\mathcal{F}}\|^2+\frac{d}{2}}\\ &\overset{2}{=} e^{-\frac{n-k}{2}\log(1+C\|\beta_{\mathcal{T}-\mathcal{F}}\|^2)+\frac{d}{2}}.\end{aligned}$$

The first inequality follows from Theorem 1 and the second equality comes from the well-known formula (see for example [12]) for the moment-generating function of a chi-square random variable; that is, $\mathrm{E}_{W\sim\chi^2_{n-k}}e^{tW}=(1-2t)^{-\frac{n-k}{2}}$ for $2t<1$.

Proof of Lemma 9: Let us first explain the idea behind this Lemma. We aim to prove that under certain conditions, for some $r_0\in[1,k]$, $g(r)$ is a decreasing function for $r\in[1,r_0]$ and an increasing function for $r\in[r_0,k]$. This yields the desired upper bound

$$\max_{r\in[d,k]} g(r)\leq\max\{g(d),g(k)\}. \tag{22}$$

We begin by taking derivatives of $g(r)$ to prove the aforementioned claim

$$\begin{aligned}g'(r)&=\frac{1}{2}+\log\left(\frac{k(p-k)}{r^2}\right)-\frac{\gamma(n-k)}{2(1+r\gamma)}\\ g''(r)&=\frac{-4(1+r\gamma)^2+r\gamma^2(n-k)}{2r(1+r\gamma)^2}.\end{aligned}$$

Note that in the following steps, we use inequality (15), i.e.,

$$n-k>\max\left\{4k\left(1+\frac{1}{k\gamma}\right)^2, \left(1+\frac{1}{\gamma}\right)[1+2\log(k(p-k))]\right\}$$

to prove inequality (22).

1. $g''(r)=0$ has two solutions $r_1^\star$ and $r_2^\star$ such that $r_1^\star<r_2^\star$. Due to the positivity of the denominator and the quadratic and concave nature of the numerator of $g''(r)$, we have:
   (a) $g''(r)<0$ for $r<r_1^\star$;
   (b) $g''(r)>0$ for $r_1^\star<r<r_2^\star$;
   (c) $g''(r)<0$ for $r_2^\star<r$.
2. From inequality (15), we have $n-k>4k\left(1+\frac{1}{k\gamma}\right)^2$ which ensures that $g''(k)>0$. Therefore, we have $r_1^\star<k<r_2^\star$. This implies the convexity of $g(r)$ for $r\in[r_1^\star,k]$ and the negativity of $g''(r)$ for $r<r_1^\star$. We have two situations depending on whether $1<r_1^\star$ or not:
   1. $1<r_1^\star$: From inequality (15) we have $n-k>\frac{1+\gamma}{\gamma}[1+2\log(k(p-k))]$ which implies that $g'(1)<0$. This, in conjunction with $g''(r)<0$ for $r<r_1^\star$, implies that $g(r)$ is decreasing for $r\in[1,r_1^\star]$.
   2. $r_1^\star\leq 1$: $g(r)$ is convex for $r\in[1,k]$.
   3. Either case, i.e., $g(r)$ is convex for $r\in[1,k]$ or decreasing for all $r\in[1,r_1^\star]$ and convex for $r\in[r_1^\star,k]$, proves the desired inequality (22).

V. Conclusion

In this paper, we examined the probability that the optimal decoder declares an incorrect sparsity pattern. We obtained an upper bound for any generic measurement matrix, and this allowed us to calculate the error probability in the case of random measurement matrices. In the special case when the entries of the measurement matrix are i.i.d. normal random variables, we computed an upper bound on the expected error probability. Sufficient conditions on exact sparsity pattern recovery were obtained, and they were shown to improve the previous results [4]–[7]. Moreover, these results asymptotically match (in terms of growth rate) the corresponding necessary condition presented in [8]. An interesting open problem is to extend the sufficient conditions derived in this work to non-Gaussian and sparse measurement matrices.

$$\max\{g(d),g(k)\}\geq\frac{B-5}{2}\max\left\{\left[d\log\left(\frac{k(p-k)}{d^2}\right)+d\right],\left[k\log\left(\frac{k(p-k)}{k^2}\right)+k\right]\right\}.$$

ACKNOWLEDGMENT

The author would like to express his gratitude to V. Roychowdhury for introducing him to this subject. The author is grateful to I. Kontoyiannis, L. Paninski, X. Pitkov, and Y. Mishchenko for careful reading of the manuscript and fruitful discussions, and to the referees for their critical comments that improved the presentation of the manuscript.

REFERENCES

[1] M. Zibulevsky and B. Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionary," *Neur. Comput.*, vol. 13, pp. 863–882, 2001.

[2] W. Vinje and J. Gallant, "Sparse coding and decorrelation in primary visual cortex during natural vision," *Science*, vol. 287, no. 5456, pp. 1273–1276, 2000.

[3] D. di Bernardo, M. J. Thompson, T. Gardner, S. E. Chobot, E. L. Eastwood, A. P. Wojtovich, S. J. Elliott, S. Schaus, and J. J. Collins, "Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks," *Nat. Biotech*, vol. 23, pp. 377–383, Mar. 2005.

[4] M. Wainwright, "Information-theoretic limitations on sparsity recovery in the high-dimensional and noisy setting," *IEEE Trans. Inf. Theory*, vol. 55, no. 12, pp. 5728–5741, Dec. 2009.

[5] M. Akcakaya and V. Tarokh, "Shannon-theoretic limits on noisy compressive sampling," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 492–504, Jan. 2010.

[6] A. Fletcher, S. Rangan, and V. Goyal, "Necessary and sufficient conditions on sparsity pattern recovery," *IEEE Trans. Inf. Theory*, vol. 55, no. 12, pp. 5758–5772, Dec. 2009.

[7] A. Karbasi, A. Hormati, S. Mohajer, and M. Vetterli, "Support recovery in compressed sensing: An estimation theoretic approach," in *Proc. 2009 IEEE Int. Symp. Information Theory*, 2009.

[8] W. Wang, M. Wainwright, and K. Ramchandran, "Information-theoretic limits on sparse signal recovery: Dense versus sparse measurement matrices," *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2967–2979, Jun. 2010.

[9] G. Reeves and M. Gastpar, "Sampling bounds for sparse support recovery in the presence of noise," in *Proc. IEEE Int. Symp. Information Theory*, 2008, pp. 2187–2191.

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

[11] H. Rauhut, K. Schnass, and P. Vandergheynst, "Compressed sensing and redundant dictionaries," *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 2210–2219, May 2008.

[12] T. A. Severini, *Elements of Distribution Theory*. Cambridge, U.K.: Cambridge University Press, 2005.

[13] P. Bjorstad and J. Mandel, "On the spectra of sums of orthogonal projections with applications to parallel computing," *BIT Numer. Math.*, vol. 31, pp. 76–88, 1991.

Kamiar Rahnama Rad was born in Darmstadt, Germany. He received the B.Sc. degree in electrical engineering from Sharif University of Technology, Tehran, in 2004 and the M.Sc. degree in electrical engineering from University of California, Los Angeles, in 2006. He is currently pursuing the Ph.D. degree in statistics at Columbia University, New York.

His research interests include information theory, computational neuroscience, and social learning theory.