Sharp Upper Bound on Error Probability of Exact Sparsity Recovery

Kamiar Rahnama Rad
Department of Statistics and Center for Theoretical Neuroscience
Columbia University
kamiar@stat.columbia.edu

***Abstract*—Imagine the vector $y = X\beta + \epsilon$ where $\beta \in \mathbb{R}^m$ has only k non zero entries and $\epsilon \in \mathbb{R}^n$ is a Gaussian noise. This can be viewed as a linear system with sparsity constraints corrupted with noise. We find a non-asymptotic upper bound on the error probability of exact recovery of the sparsity pattern given any generic measurement matrix X. By drawing X from a Gaussian ensemble, as an example, to ensure exact recovery, we obtain asymptotically sharp sufficient conditions on n which meet the necessary conditions introduced in (Wang et al., 2008).**

I. Introduction

We study the performance of the maximum likelihood estimate of $\beta \in \mathbb{R}^m$, with k nonzero entries, based on the observation of the following linear model:

$$y = X\beta + \epsilon,$$

where $X \in \mathbb{R}^{n\times m}$ is a set of perturbation 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 matrix equal to $\sigma^2 I_{n\times n}$. Each row of X and the corresponding entry of y is viewed as an input and output measurement, respectively. And, because of that, n is the number of measurements. Set a lower bound β_{min} on the absolute value of the non-zero entries of β and assume $\sigma = 1$; this entails no loss of generality, by the rescaling of β_{min}. Finally, $\|.\|$ stands for the ℓ_2-norm for the remainder of this paper.

The following definition will be useful for the remainder of this discussion:

Definition 1. *Consider a set of integers $\mathcal{T} = \{t_1, \ldots, t_{|\mathcal{T}|}\}$, such that:*

$$1 \leq t_i \leq m, \qquad i = 1, \ldots, |\mathcal{T}|.$$

We say $\beta \in \mathbb{R}^m$ has sparsity pattern *$\mathcal{T}$ if only entries with indices $i \in \mathcal{T}$ are nonzero. We shall generally denote by $X_\mathcal{T} \in \mathbb{R}^{n\times|\mathcal{T}|}$, the matrix obtained from X by extracting $|\mathcal{T}|$ columns with indices i obeying $i \in \mathcal{T}$. Further, let $\mathcal{S}_\mathcal{T}$ and $\Pi_{\mathcal{S}_\mathcal{T}}$ denote the corresponding column space of $X_\mathcal{T}$ and projection matrix into $\mathcal{S}_\mathcal{T}$, respectively.*

Assuming $\mathcal{T}$ is the true sparsity pattern we can use $X\beta$ and $X_\mathcal{T}\beta_\mathcal{T}$, interchangeably. Furthermore, we consider measurements X such that any k column of X are linearly independent. This ensures that for any $\mathcal{F} \neq \mathcal{F}'$ the subspaces $\mathcal{S}_\mathcal{F}$ and $\mathcal{S}_\mathcal{F}$ are different and have dimension equal to $|\mathcal{F}| = k$ which is an identifiability criterion. Therefore, $\dim(\mathcal{S}_\mathcal{F}) = |\mathcal{F}|$ for any $\mathcal{F}$ such that $|\mathcal{F}| \leq k$. The optimum maximum likelihood decoder is defined as:

$$\hat{\mathcal{T}} = \arg\min_{\mathcal{F}} \|(I - \Pi_{\mathcal{S}_\mathcal{F}})y\|,$$

for all $|\mathcal{F}| = k$.

We present an upper bound on the probability of declaring a wrong sparsity pattern based on the optimum maximum likelihood decoder for any measurement matrix X. We say that sparse pattern recovery is reliable when the probability that the optimum decoder declares the true sparsity pattern goes to one when $n \to \infty$. Asymptotic analysis has been done for random measurement matrices (Wainwright, 2007; Akcakaya and Tarokh, 2008; Fletcher et al., 2008; Wang et al., 2008). Our methods is different than those used in (Wainwright, 2007; Akcakaya and Tarokh, 2008; Fletcher et al., 2008; Wang et al., 2008) and applies to any generic measurement matrix X. Wainwright analyzed the average error probability for Gaussian measurements (Wainwright, 2007) and different error metrics were analyzed in (Akcakaya and Tarokh, 2008). The necessary conditions for exact reliable sparse recovery for sparse versus dense random measurements is examined in (Wang et al., 2008). Furthermore, they generalized the necessary conditions established for Gaussian measurements in (Wainwright, 2007) to random measurement matrices with i.i.d entries and bounded variance.

978-1-4244-2734-5/09/$25.00 ©2009 IEEE

II. General Measurement

The probability that we declare the sparsity pattern $\mathcal{F}$ when the true sparsity pattern is $\mathcal{T}$ is denoted $P_{\mathcal{T}}[\mathcal{F}|X]$. We introduced it as a conditional probability given the measurement matrix X. Later we consider random matrix measurement as a special case of our result.

Theorem 2. *The probability* $\Pr_{\mathcal{T}}[\mathcal{F}|X]$ *that the optimum decoder declares* $\mathcal{F}$ *when* $\mathcal{T}$ *is the true sparsity pattern, given any generic measurment matrix* X*, is upper bounded by* $\rho^d e^{-\alpha\|(I-\Pi_{\mathcal{S}_{\mathcal{F}}})X_{\mathcal{T}-\mathcal{F}}\beta_{\mathcal{T}-\mathcal{F}}\|^2}$*, where* $\alpha = \frac{3-2\sqrt{2}}{2}$*,* $\rho = \frac{2}{2\sqrt{2}-1}$ *and* $|\mathcal{F}-\mathcal{T}| = d$*.*

It is interesting to see that error rate depends exponentially on the projection of $X_{\mathcal{T}-\mathcal{F}}\beta_{\mathcal{T}-\mathcal{F}}$ to $\mathcal{S}_{\mathcal{F}}{}^c$. Furthermore, note $\|(I - \Pi_{\mathcal{S}_{\mathcal{F}}})X_{\mathcal{T}-\mathcal{F}}\beta_{\mathcal{T}-\mathcal{F}}\|^2 \geq \sin^2\theta_{\mathcal{F},\mathcal{T}-\mathcal{F}}\|X_{\mathcal{T}-\mathcal{F}}\beta_{\mathcal{T}-\mathcal{F}}\|^2$; the smallest canonical or principle angle between $\mathcal{S}_{\mathcal{F}}$ and $\mathcal{S}_{\mathcal{T}-\mathcal{F}}$ defined as:

$$\cos\theta_{\mathcal{F},\mathcal{T}-\mathcal{F}} = \max_{v\in\mathcal{S}_{\mathcal{F}},\|v\|^2=1}\ \max_{u\in\mathcal{S}_{\mathcal{T}-\mathcal{F}},\|u\|^2=1} v^T u.$$

The probability that there exist a sparsity pattern $\mathcal{F} \neq \mathcal{T}$ which is declared by the optimal decoder when the true sparsity pattern is $\mathcal{T}$ is upper bounded by the union bound $\sum_{\mathcal{F}\neq\mathcal{T}} P_{\mathcal{T}}[\mathcal{F}|X]$. For any $1 \geq d \geq k$ there exist $\binom{k}{d}\binom{m-k}{d}$ sparsity patterns $\mathcal{F}$ such that $|\mathcal{F}-\mathcal{T}| = d$ and $|\mathcal{F}| = |\mathcal{T}| = k$.

A. Proof of Theorem 2

Lemma 3. *Define* $Z = y^T(\Pi_{\mathcal{S}_{\mathcal{F}}} - \Pi_{\mathcal{S}_{\mathcal{T}}})y$*, where* $y \sim \mathcal{N}(X_{\mathcal{T}}\beta_{\mathcal{T}}, I_{n\times n})$ *and let* $|\mathcal{F}-\mathcal{T}| = d$*, then*

$$\begin{aligned}\log \mathrm{E}\exp Zt \quad &\leq \quad -d\log(\sqrt{2}-\frac{1}{2}) \\ -\frac{\|\Pi_{\mathcal{S}_{\mathcal{F}}{}^c}X_{\mathcal{T}-\mathcal{F}}\beta_{\mathcal{T}-\mathcal{F}}\|^2}{2}&(3 \quad - \quad 2\sqrt{2}).\end{aligned}$$

Proof. First note that for $y \sim \mathcal{N}(\mu, I_{n\times n})$, we have:

$$\begin{aligned}\mathrm{E}e^{ty^T\Psi y} &= \int e^{t(\mu+\epsilon)^T\Psi(\mu+\epsilon)}\frac{e^{-\frac{\|\epsilon\|^2}{2}}}{(2\pi)^{n/2}}d\epsilon \\ &= \frac{e^{t\mu^T\Psi\mu+2t^2\mu^T\Psi(I-2t\Psi)^{-1}\Psi\mu}}{\det(I-2t\Psi)^{\frac{1}{2}}} \\ &\times \int \frac{e^{-\frac{\|(I-2t\Psi)^{1/2}(\epsilon-\epsilon_0)\|^2}{2}}}{(2\pi)^{n/2}\det(I-2t\Psi)^{-\frac{1}{2}}}d\epsilon,\end{aligned}$$

where $\epsilon_0 = 2t(I-2t\Psi)^{-1}\Psi\mu$. Therefore, for $\mu = X\beta$ and $\Psi = \Pi_{\mathcal{S}_{\mathcal{F}}} - \Pi_{\mathcal{S}_{\mathcal{T}}}$:

$$\begin{aligned}\log \mathrm{E}\exp(Zt) \quad &= \quad 2t^2\mu^T\Psi(I-2t\Psi)^{-1}\Psi\mu \\ +t\mu^T\Psi\mu \quad &- \quad \frac{1}{2}\log\det(I-2t\Psi),\end{aligned}$$

for $|t| < 1/2$. If $|\mathcal{T}-\mathcal{F}| = d$ then $\Pi_{\mathcal{S}_{\mathcal{F}}} - \Pi_{\mathcal{S}_{\mathcal{T}}}$ has d positive eigenvalues and d negative eigenvalues. For every positive eigenvalue there exist a corresponding negative eigenvalue with the same magnitude. Furthermore, the eigenvalues of the difference of two projection matrices are upper bounded by one. And, because of that, $(I - 2t\Psi)$ is positive definite for $t < 1/2$, $\|(I-2t\Psi)^{-\frac{1}{2}}\|_2 \leq (1-2t)^{-1}$ and $I - 2t\Psi$ has $2d$ eigenvalues lower bounded by $1-2t$ whereas the rest $n-2d$ eigenvalues are equal to one. Finally, note that because $\mu = X\beta \in \mathcal{S}_{\mathcal{T}}$, we have:

$$\begin{aligned}\mu^T\Psi\mu &= -\|\Pi_{\mathcal{S}_{\mathcal{F}}{}^c}X\beta\|^2 \\ \mu^T\Psi^2\mu &= \|\Pi_{\mathcal{S}_{\mathcal{F}}{}^c}X\beta\|^2,\end{aligned}$$

where $\Pi_{\mathcal{S}_{\mathcal{F}}{}^c} = I - \Pi_{\mathcal{S}_{\mathcal{F}}}$. As a consequence of the points made in the previous paragraph and the form of the cumulant distribution, we obtain the following bound on the cumulant distribution function :

$$\begin{aligned}\log \mathrm{E}\exp(Zt) \quad &\leq \quad -d\log(1-u^2) \\ &- \quad \frac{\|\Pi_{\mathcal{S}_{\mathcal{F}}{}^c}X\beta\|^2}{2}(u - \frac{u^2}{1-u}) \\ ,\end{aligned}$$

for $u = 2t$. Since this inequality is true for any $|u| < 1$ we take the supremum of $u - \frac{u^2}{1-u}$ which is equal to $3 - 2\sqrt{2}$ at $u = 1 - \frac{\sqrt{2}}{2}$ and obtain the following bound:

$$\begin{aligned}\log \mathrm{E}\exp(Zt) \quad &\leq \quad -\frac{\|\Pi_{\mathcal{S}_{\mathcal{F}}{}^c}X\beta\|^2}{2}(3-2\sqrt{2}) \\ &- \quad d\log(\sqrt{2}-\frac{1}{2}).\end{aligned}$$

The rest follows from:

$$(\Pi_{\mathcal{S}_{\mathcal{T}}} - \Pi_{\mathcal{S}_{\mathcal{F}}})X_{\mathcal{T}}\beta_{\mathcal{T}} \quad = \quad (I - \Pi_{\mathcal{S}_{\mathcal{F}}})X_{\mathcal{T}-\mathcal{F}}\beta_{\mathcal{T}-\mathcal{F}}.$$

The optimum decoder declares $\mathcal{F}$ when the true sparsity pattern is $\mathcal{T}$ only if $\|(I-\Pi_{\mathcal{S}_{\mathcal{T}}})y\| > \|(I-\Pi_{\mathcal{S}_{\mathcal{F}}})y\|$, therefore, by defining $Z = y^T(\Pi_{\mathcal{S}_{\mathcal{F}}} - \Pi_{\mathcal{S}_{\mathcal{T}}})y$ and exploiting the Chernoff inequality, we obtain the following bound:

$$\begin{aligned}P_{\mathcal{T}}[\mathcal{F}|X] &= P_{\mathcal{T}}[y^T(\Pi_{\mathcal{S}_{\mathcal{F}}} - \Pi_{\mathcal{S}_{\mathcal{T}}})y > 0|X] \\ &= P_{\mathcal{T}}[e^{Zt} > 1|X] \\ &\leq \inf_{|t|\leq\frac{1}{2}} \mathrm{E}\exp(Zt) \\ &\leq \rho^d e^{-\alpha\|\Pi_{\mathcal{S}_{\mathcal{F}}{}^c}X\beta\|^2},\end{aligned}$$

for $\alpha = \frac{3-2\sqrt{2}}{2}$ and $\rho = \frac{2}{2\sqrt{2}-1}$.

978-1-4244-2734-5/09/$25.00 ©2009 IEEE

III. Gaussian Measurement

In this section we present the average error probability for random measurement matrices and the asymptotic upper bound on n to ensure exact sparsity recovery for the model defined in section I.

Lemma 4. *For Gaussian measurement matrices, with $X_{ij} \sim \mathcal{N}(0,1)$ the average probability that the optimum decoder declares $\mathcal{F}$ is upper bounded by $e^{-\frac{n-k}{2}\log(1+2\alpha d\beta_{min}^2)+d}$, where $\mathcal{T}$ the true sparsity pattern, $|\mathcal{F}-\mathcal{T}| = d$ and $\alpha = \frac{3-2\sqrt{2}}{2}$.*

Proof is given in section V. It is worth mentioning other known upper bounds for $P_{\mathcal{T}}[\mathcal{F}]$. Wainwright found the following bound (Wainwright, 2007):

$$P_{\mathcal{T}}[\mathcal{F}] \quad \leq \quad 3e^{-\frac{(n-k)d\beta_{min}^2}{12(d\beta_{min}^2+8)}}.$$

Theorem 5. *If*

$$n \quad > \quad C^\star \max\Big\{k + \frac{\log k(m-k)}{\log(1+\beta_{min}^2)}, \\ k + \frac{k\log\frac{m-k}{k}}{\log(1+\beta_{min}^2 k)}\Big\},$$

for some fixed constant $C^\star > 0$ then sparse pattern recovery is reliable.

Proof is given in section VI. We consider Theorem 5 at six different scaling regimes for k and β_{min}. For any of the six scaling regimes we obtain the sufficient scaling of the number of measurements n to guarantee reliable exact recovery. The following corollary considers the sufficient conditions for exact recovery in several regimes of interest.

Corollary 6. *In the following we present sharp sufficient conditions on n to ensure exact sparsity recovery under different scaling regimes of k and β_{min}.*

1) *For $k = \Theta(m)$ and $\beta_{min}^2 = \Theta(\frac{1}{k})$, $n = \Theta(m\log m)$*
2) *For $k = \Theta(m)$ and $\beta_{min}^2 = \Theta(\frac{\log k}{k})$, $n = \Theta(m)$*
3) *For $k = \Theta(m)$ and $\beta_{min}^2 = \Theta(1)$, $n = \Theta(m)$*
4) *For $k = o(m)$ and $\beta_{min}^2 = \Theta(\frac{1}{k})$, $n = \Theta(k\log(m-k))$*
5) *For $k = o(m)$ and $\beta_{min}^2 = \Theta(\frac{\log k}{k})$, $n = \Theta(\frac{k\log\frac{m}{k}}{\log\log k})$*
6) *For $k = o(m)$ and $\beta_{min}^2 = \Theta(1)$, $n = \Theta(m)$, $n = \Theta(\frac{k\log\frac{m}{k}}{\log k})$*

Note that for all different scalings of n the inequality (2) is satisfied. All except the last two have been shown in (Wang et al., 2008) to be sharp in the sense that they are equal to the known necessary conditions. Since the last two upper bounds are similar to the lower bounds found in (Wang et al., 2008), all the above upper bounds are sharp.

IV. Discussion

We analyzed the probability that the optimal decoder declares a wrong sparsity pattern. We assumed a linear model with Gaussian noise. We obtained a sharp upper bound on the error probability for any generic measurement matrix. This allows us to calculate the expected value of the error probability when we are dealing random measurements. In the special when the entries of the measurement are i.i.d. Gaussian random variables we found an upper bound on the expected error probability. We found sufficient conditions on the number of measurements that are sharp because they match the known necessary conditions. An interesting open problem is how to extend the Gaussian measurement results to other random matrices.

V. Proof of Lemma 4

We have:

$$\begin{aligned} \|\Pi_{\mathcal{S}_{\mathcal{F}}^c}\mu\|^2 &= \|(I-\Pi_{\mathcal{S}_{\mathcal{F}}})X_{\mathcal{T}-\mathcal{F}}\beta_{\mathcal{T}-\mathcal{F}}\|^2 \\ &\geq \beta_{min}^2\|(I-\Pi_{\mathcal{S}_{\mathcal{F}}})(\sum_{i\in\mathcal{T}-\mathcal{F}} X_i)\|^2. \end{aligned}$$

Note that the random subspace spanned by the columns of $X_{\mathcal{F}}$ is independent of the random vector $\sum_{i\in\mathcal{T}-\mathcal{F}} X_i$. Furthermore, recall that $\dim(\mathcal{S}_{\mathcal{F}}) = k$, therefore we conclude that $W \equiv d^{-1}\|(I-\Pi_{\mathcal{S}_{\mathcal{F}}})(\sum_{i\in\mathcal{T}-\mathcal{F}} X_i)\|^2$ is a chi-square random variable with $n-k$ degrees of freedom. Therefore,

$$\begin{aligned} P_{\mathcal{T}}[\mathcal{F}] &= E_X P_{\mathcal{T}}[\mathcal{F}|X] \\ &\leq \rho^d E_X e^{-\alpha\|\Pi_{\mathcal{S}_{\mathcal{F}}^c}\mu\|^2} \\ &\leq \rho^d E_{W\sim\chi_{n-k}^2} e^{-\alpha d\beta_{min}^2 W} \\ &= e^{-\frac{n-k}{2}\log(1+2\alpha d\beta_{min}^2)+d} \end{aligned}$$

where α and ρ were defined in theorem (2) and the last equation is a result of $E_{W\sim\chi_{n-k}^2} e^{tW} = (1-2t)^{-\frac{n-k}{2}}$ for $2t < 1$.

VI. Proof of Theorem 5

The total error probability is upper bounded by the sum of $P_{\mathcal{T}}[\mathcal{F}]$ for every $\mathcal{F} \neq \mathcal{T}$. There exist

$\binom{k}{d}\binom{m-k}{d}$ sparsity patterns $\mathcal{F}$ such that $|\mathcal{F}-\mathcal{T}| = d$. Note that

$$\begin{aligned} \binom{k}{d} &\le \exp\{d\log\frac{ke}{d}\}, \\ \binom{m-k}{d} &\le \exp\{d\log\frac{(m-k)e}{d}\}. \end{aligned}$$

Hence, the total error probability is upper bounded by

$$p_e = \sum_{d=1}^{k} e^{d[3+\log\frac{k}{d}+\log\frac{m-k}{d}]-\frac{n-k}{2}\log(1+2\alpha\beta_{min}^2 d)}. \tag{1}$$

In Appendix, we show that $f(d) = d[3+\log\frac{k}{d}+\log\frac{m-k}{d}] - \frac{n-k}{2}\log(1+2\alpha\beta_{min}^2 d)$ is convex in d if

$$(n-k)\beta_{min}^2\alpha > 4 + \frac{1}{k\beta_{min}^2\alpha} + 4k\beta_{min}^2\alpha. \tag{2}$$

Since $f(d)$ is convex, its maximum is achieved at the boundaries. Therefore, we have $f(d) \le \max(f(k), f(1))$. Therefore, $p_e \le e^{\log k + \max(f(1),f(k))}$ and for

$$\begin{aligned} n &> \max\Big\{k + \frac{6+2\log k(m-k)}{\log(1+2\alpha\beta_{min}^2)}, \\ &\qquad k + \frac{2k[3+\log\frac{m-k}{k}]}{\log(1+2\alpha\beta_{min}^2 k)}\Big\}, \end{aligned}$$

as $n \to \infty$ the error probability goes to zero because of $\log k + \max(f(1), f(k)) \to -\infty$. Note that the inequality (2) is satisfied. Therefore, asymptotically speaking, the total error probability goes to zero if

$$\begin{aligned} n &> C^\star \max\Big\{k + \frac{\log k(m-k)}{\log(1+\beta_{min}^2)}, \\ &\qquad k + \frac{k\log\frac{m-k}{k}}{\log(1+\beta_{min}^2 k)}\Big\} \end{aligned}$$

for a constant $C^\star$ independent of (m, n, k, β_{min}).

VII. Appendix

We have

$$\begin{aligned} f(d) &= d[3+\log\frac{k}{d}+\log\frac{m-k}{d}] \\ &- \frac{n-k}{2}\log(1+2\alpha\beta_{min}^2 d), \\ \frac{\partial f(d)}{\partial d} &= 1 + 2\log\sqrt{k(m-k)} \\ &- 2\log d - \frac{\alpha\beta_{min}^2(n-k)}{1+2\alpha\beta_{min}^2 d}, \\ \frac{\partial^2 f(d)}{\partial d^2} &= -\frac{2}{d} + \frac{2\alpha^2\beta_{min}^4(n-k)}{(1+2\alpha\beta_{min}^2 d)^2}. \end{aligned}$$

If for every $1 \ge d \ge k$ we have

$$(n-k)\beta_{min}^2\alpha > 4 + \frac{1}{d\beta_{min}^2\alpha} + 4d\beta_{min}^2\alpha,$$

then $\frac{\partial^2 f(d)}{\partial d^2} > 0$ and hence $f(d)$ is convex in d. The r.h.s is maximized when $d = k$ which yields that if $(n-k)\beta_{min}^2\alpha > 4 + \frac{1}{k\beta_{min}^2\alpha} + 4k\beta_{min}^2\alpha$, $f(d)$ is convex.

Acknowledgment

The author would like to thank Liam Paninski for helpful conversions and and Vwani Roychowdhury for introducing the problem to the author. The author is grateful to Yuriy Mishchenko for carefully reading the manuscript and noticing mistakes in the earlier version.

References

Akcakaya, M. and Tarokh, V. (2008). Noisy compressive sampling limits in linear and sublinear regimes. *42nd Annual Conference on Information Sciences and Systems*, pages 1–4.

Fletcher, A., Rangan, S., and Goyal, V. (2008). Necessary and sufficient conditions on sparsity pattern recovery. *Arxiv preprint arXiv0804.1839*.

Wainwright, M. J. (2007). Information-theoretic bounds on sparsity recovery in the high-dimensional and noisy setting. *IEEE International Symposium on Information Theory*, pages 961–965.

Wang, W., Wainwright, M., and Ramchandran, K. (2008). Information-theoretic limits on sparse signal recovery: Dense versus sparse measurement matrices. *Arxiv preprint arXiv:0806.0604*.