Supplementary material for *Estimating Unknown Sparsity in Compressed Sensing*

Abstract

In this supplement, we provide the proofs for the theoretical results in our submission *Estimating Unknown Sparsity in Compressed Sensing*.

Proposition 1.

Proof of Proposition 1. To prove the implication *(i)*, we calculate

$$\begin{aligned} \frac{1}{\sqrt{T}} \frac{\|x - x_T\|_1}{\|x\|_2} &= \frac{1}{\sqrt{T}} \frac{\|x\|_1 - \|x_T\|_1}{\|x\|_2} \\ &= \frac{\sqrt{s(x)}}{\sqrt{T}} - \frac{1}{\sqrt{T}} \frac{\|x_T\|_1}{\|x_T\|_2} \frac{\|x_T\|_2}{\|x\|_2} \\ &= \frac{\sqrt{s(x)}}{\sqrt{T}} - \frac{\sqrt{s(x_T)}}{\sqrt{T}} \frac{\|x_T\|_2}{\|x\|_2}. \end{aligned} \tag{1}$$

Since $s(x_T) \leq \|x_T\|_0 \leq T$, and $\frac{\|x_T\|_2}{\|x\|_2} \leq 1$, we obtain the lower bound

$$\frac{1}{\sqrt{T}} \frac{\|x - x_T\|_1}{\|x\|_2} \geq \frac{\sqrt{s(x)}}{\sqrt{T}} - 1.$$

Hence, if the left hand side is at most ε, we must have $T \geq \frac{s(x)}{(1+\varepsilon)^2}$, proving *(i)*. To prove the second implication, note that $T \geq c \log(p) \frac{\|x\|_1^2}{\|x\|_2^2}$ implies

$$\begin{aligned} \frac{1}{\sqrt{T}} \frac{\|x - x_T\|_1}{\|x\|_2} &\leq \frac{1}{\sqrt{c \log(p)}} \frac{\|x - x_T\|_1}{\|x\|_1} \\ &= \frac{1}{\sqrt{c \log(p)}} \left(1 - \frac{\|x_T\|_1}{\|x\|_1}\right). \end{aligned} \tag{2}$$

Next, consider the probability vectors $u, v \in \mathbb{R}^p$ defined by $u_i = 1/p$ and $v_i = |x|_{[i]}/\|x\|_1$ (that is, $v_1 \geq v_2 \geq \cdots \geq v_p$). It is a basic fact about the majorization ordering on $\mathbb{R}^p$ that u is majorized by any other probability vector (Marshall et al., 2010, p. 7). In particular, we have $\sum_{i=1}^T u_i \leq \sum_{i=1}^T v_i$ for any $T \in \{1, \ldots, p\}$, which is the same as

$$\frac{T}{p} \leq \frac{\|x_T\|_1}{\|x\|_1}.$$

Combining this with line (2) proves *(ii)*. $\square$

Theorem 1.

Proof of Theorem 1. Define the noiseless version of the measurement y_i to be

$$y_i^\circ := \langle a_i, x\rangle, \qquad i = 1, \dots, n_1 + n_2$$

and let the noiseless versions of the statistics $\widehat{T}_1$ and $\widehat{T}_2$ be given by

$$\tilde{T}_1 := \tfrac{1}{\gamma}\text{median}(|y_1^\circ|, \dots, |y_{n_1}^\circ|) \tag{3}$$

$$\tilde{T}_2^2 := \tfrac{1}{\gamma^2 n_2}\Big((y_{n_1+1}^\circ)^2 + \cdots + (y_{n_1+n_2}^\circ)^2\Big). \tag{4}$$

It is convenient to work in terms of these variables, since their limiting distributions may be computed exactly. Due to the fact that $\frac{y_1^\circ}{\gamma\|x\|_1}, \dots, \frac{y_{n_1}^\circ}{\gamma\|x\|_1}$ is an i.i.d. sample from the standard Cauchy distribution $C(0,1)$, the asymptotic normality of the sample median implies

$$\sqrt{n/2}\big(\tfrac{\tilde{T}_1}{\|x\|_1} - 1\big) \xrightarrow{\mathcal{L}} N(0, \tau_1^2), \tag{5}$$

where $\tau_1^2 = \pi^2/8$. Additional details may be found in (David, Theorem 9.2) and (Li et al., 2007, Lemma 3). Similarly, the variables $(\frac{y_{n_1+1}^\circ}{\gamma\|x\|_2})^2, \dots, (\frac{y_{n_1+n_2}^\circ}{\gamma\|x\|_2})^2$ are an i.i.d. sample from the chi-square distribution on one degree of freedom, and so it follows from the delta method that

$$\sqrt{n/2}\big(\tfrac{\tilde{T}_2}{\|x\|_2} - 1\big) \xrightarrow{\mathcal{L}} N(0, \tau_2^2), \tag{6}$$

where $\tau_2^2 = 1/2$. Note that in proving the last two limit statements, we intentionally scaled the variables y_i° in such a way that their distributions did not depend on any model parameters. It is for this reason that the limits hold even when the model parameters are allowed to depend on n. We conclude from the limits (5) and (6) that for any $\alpha \in (0, 1/2)$,

$$\mathbb{P}\Big(\tfrac{\tilde{T}_1}{\|x\|_1} \in [1 - \tfrac{\tau_1 z_{1-\alpha}}{\sqrt{n/2}}, 1 + \tfrac{\tau_1 z_{1-\alpha}}{\sqrt{n/2}}]\Big) = 1 - 2\alpha + o(1), \tag{7}$$

and

$$\mathbb{P}\Big(\tfrac{\tilde{T}_2}{\|x\|_2} \in [1 - \tfrac{\tau_2 z_{1-\alpha}}{\sqrt{n/2}}, 1 + \tfrac{\tau_2 z_{1-\alpha}}{\sqrt{n/2}}]\Big) = 1 - 2\alpha + o(1). \tag{8}$$

We now relate $\widehat{T}_1$ and $\widehat{T}_2$ in terms of intervals defined by $\tilde{T}_1$ and $\tilde{T}_2$. Since the noise variables are bounded by $|\epsilon_i| \le \sigma_0$, and $y_i = y_i^\circ + \epsilon_i$, it is easy to see that

$$\widehat{T}_1 \in [\tilde{T}_1 - \tfrac{\sigma_0}{\gamma}, \tilde{T}_1 + \tfrac{\sigma_0}{\gamma}].$$

Consequently, if we note that $\frac{\sigma_0}{\gamma\|x\|_1} \le \frac{\sigma_0}{\gamma\|x\|_2} = \rho$, then we may write

$$\tfrac{\widehat{T}_1}{\|x\|_1} \in \Big[\tfrac{\tilde{T}_1}{\|x\|_1} - \rho,\ \tfrac{\tilde{T}_1}{\|x\|_1} + \rho\Big]. \tag{9}$$

To derive a similar relationship involving $\widehat{T}_2$ and $\tilde{T}_2$, if we write $\widehat{T}_2$ in terms of $\|(y_{n_1}, \ldots, y_{n_1+n_2})\|_2$ and apply the triangle inequality, it follows that

$$\frac{\widehat{T}_2}{\|x\|_2} \in \Big[\frac{\tilde{T}_2}{\|x\|_2} - \rho, \ \frac{\tilde{T}_2}{\|x\|_2} + \rho\Big]. \tag{10}$$

The proof may now be completed by assembling the last several items. Recall the parameters δ_n and η_n, which are given by

$$\delta_n = \delta_n(\alpha, \rho) = \frac{\tau_1 z_{1-\alpha}}{\sqrt{n/2}} + \rho \tag{11}$$

$$\eta_n = \eta_n(\alpha, \rho) = \frac{\tau_2 z_{1-\alpha}}{\sqrt{n/2}} + \rho. \tag{12}$$

Combining the limits (7) and (8) with the intervals (9) and (10), we have the following asymptotic bounds for the statistics $\widehat{T}_1$ and $\widehat{T}_2$,

$$\mathbb{P}\Big(\frac{\widehat{T}_1}{\|x\|_1} \in [1 - \delta_n, 1 + \delta_n]\Big) \geq 1 - 2\alpha + o(1), \tag{13}$$

and

$$\mathbb{P}\Big(\frac{\widehat{T}_2}{\|x\|_2} \in [1 - \eta_n, 1 + \eta_n]\Big) \geq 1 - 2\alpha + o(1). \tag{14}$$

Due to the independence of $\widehat{T}_1$ and $\widehat{T}_2$, and the relation

$$\sqrt{\frac{\widehat{s}(x)}{s(x)}} = \frac{\widehat{T}_1/\|x\|_1}{\widehat{T}_2/\|x\|_2},$$

we conclude that

$$\mathbb{P}\Big(\sqrt{\frac{\widehat{s}(x)}{s(x)}} \in \big[\tfrac{1-\delta_n}{1+\eta_n}, \tfrac{1+\delta_n}{1-\eta_n}\big]\Big) \geq (1 - 2\alpha)^2 + o(1). \tag{15}$$

$\square$

Theorem 2.

The proof of Theorem 2 is almost the same as the proof of Theorem 1 and we omit the details. One point of difference is that in Theorem 1, the bounding probability is $(1 - 2\alpha)^2$, whereas in Theorem 2 it is $(1 - 2\alpha)$. The reason is that in the case of Theorem 2, the condition $\breve{T}_1/\|x\|_1 \in [1 - \varrho, 1 + \varrho]$ holds with probability 1, whereas the analogous statement $\widehat{T}_1/\|x\|_1 \in [1 - \rho, 1 + \rho]$ holds with probability $1 - 2\alpha$ in the case of Theorem 1.

Theorem 3.

The following lemma illustrates the essential reason why estimating $s(x)$ is difficult in the deterministic case. The idea is that for any measurement matrix A, it is possible to find two signals that are indistinguishable with respect to A, and yet have very different sparsity levels in terms of $s(\cdot)$. We prove Theorem 3 after giving the proof of the lemma.

Lemma 1. *Let $A \in \mathbb{R}^{n\times p}$ be an arbitrary matrix, and let $x \in \mathbb{R}^p$ be an arbitrary signal. Then, there exists a non-zero vector $\tilde{x} \in \mathbb{R}^p$ satisfying $Ax = A\tilde{x}$, and*

$$s(\tilde{x}) \geq \frac{p-n}{(1+2\sqrt{2\log(2p)})^2}. \tag{16}$$

Proof of Lemma 1. By Hölder's inequality, $\|\tilde{x}\|_1^2/\|\tilde{x}\|_2^2 \geq \|\tilde{x}\|_2^2/\|\tilde{x}\|_\infty^2$, and so it suffices to lower-bound the second ratio. The overall approach to finding a dense vector $\tilde{x}$ is to use the probabilistic method. Let $B \in \mathbb{R}^{p\times(p-r)}$ be a matrix whose columns are an orthonormal basis for the null space of A, where $r = \text{rank}(A)$. Also define the scaled matrix $\tilde{B} := \|x\|_\infty B$. Letting $z \in \mathbb{R}^{p-r}$ be a standard Gaussian vector, we will consider $\tilde{x} := x + \tilde{B}z$, which satisfies $Ax = A\tilde{x}$ for all realizations of z. We begin the argument by defining the function

$$f(z) := \|x + \tilde{B}z\|_2 - c(n,p)\cdot\|x + \tilde{B}z\|_\infty, \tag{17}$$

where

$$c(n,p) := \frac{\sqrt{p-n}}{1+2\sqrt{2\log(2p)}}.$$

The proof amounts to showing that the event $\{f(z) > 0\}$ holds with positive probability. To see this, notice that the event $\{f(z) > 0\}$ is equivalent to

$$\frac{\|\tilde{x}\|_2}{\|\tilde{x}\|_\infty} = \frac{\|x+\tilde{B}z\|_2}{\|x+\tilde{B}z\|_\infty} > \frac{\sqrt{p-n}}{1+2\sqrt{2\log(2p)}}.$$

We will prove that $\mathbb{P}(f(z) > 0)$ is positive by showing that $\mathbb{E}[f(z)] > 0$, and this will be accomplished by lower-bounding the expected value of $\|x + \tilde{B}z\|_2$, and upper-bounding the expected value of $\|x + \tilde{B}z\|_\infty$.

First, to lower-bound $\|x + \tilde{B}z\|_2$, we begin by considering the variance of $\|x + \tilde{B}z\|_2$, and use the fact that $\|\tilde{B}z\|_2^2 = z^\top \tilde{B}^\top \tilde{B}z = \|x\|_\infty^2 \, \|z\|_2^2$, obtaining

$$\begin{aligned} \mathbb{E}\|x+\tilde{B}z\|_2 &= \sqrt{\mathbb{E}\|x+\tilde{B}z\|_2^2 - \text{var}\,\|x+\tilde{B}z\|_2} \\ &= \sqrt{\|x\|_2^2 + \|x\|_\infty^2\,(p-r) - \text{var}\,\|x+\tilde{B}z\|_2}. \end{aligned} \tag{18}$$

To upper-bound the variance, we use the Poincaré inequality for the standard Gaussian measure on $\mathbb{R}^{p-r}$ Beckner (1989). Since the function $g(z) := \|x + \tilde{B}z\|_2$ has a Lipschitz constant equal to $\|\tilde{B}\|_{\text{op}} = \|x\|_\infty$ with respect to the Euclidean norm, it follows that $\|\nabla g(z)\|_2 \leq \|x\|_\infty$. Consequently, the Poincaré inequality implies

$$\text{var}\,\|x+\tilde{B}z\|_2 \leq \|x\|_\infty^2.$$

Using this in conjunction with the inequality (18), and the fact that $r = \text{rank}(A)$ is at most n, we obtain the lower bound

$$\mathbb{E}\|x+Bz\|_2 \geq \sqrt{\|x\|_2^2 + \|x\|_\infty^2(p-n) - \|x\|_\infty^2}. \tag{19}$$

The second main portion of the proof is to upper-bound $\mathbb{E}\|x + \tilde{B}z\|_\infty$. Since $\|x + \tilde{B}z\|_\infty \leq \|x\|_\infty + \|\tilde{B}z\|_\infty$, it is enough to upper-bound $\mathbb{E}\|\tilde{B}z\|_\infty$, and we will do this using a version of Slepian's inequality. If $\tilde{b}_i$ denotes the i^{th} row of $\tilde{B}$, define $g_i = \langle \tilde{b}_i, z\rangle$, and let $w_1, \ldots, w_p$ be i.i.d. $N(0,1)$ variables. The idea is to compare the Gaussian process g_i with the Gaussian process $\|x\|_\infty w_i$. By Proposition A.2.6 in van der Vaart and Wellner van der Vaart & Wellner (1996), the inequality

$$\mathbb{E}\|\tilde{B}z\|_\infty = \mathbb{E}\left[\max_{i=1,\ldots,p} |g_i|\right] \leq 2\|x\|_\infty\, \mathbb{E}\left[\max_{i=1,\ldots,p} |w_i|\right],$$

holds as long as the condition $\mathbb{E}(g_i - g_j)^2 \leq \|x\|_\infty^2\, \mathbb{E}(w_i - w_j)^2$ is satisfied for all $i, j \in \{1, \ldots, p\}$, and this is simple to verify. To finish the proof, we make use of a standard bound for the expectation of Gaussian maxima

$$\mathbb{E}\left[\max_{i=1,\ldots,p} |w_i|\right] < \sqrt{2\log(2p)},$$

which follows from a modification of the proof of Massart's finite class lemma (Massart, 2000, Lemma 5.2)[1]. Combining the last two steps, we obtain

$$\mathbb{E}\|x + Bz\|_\infty < \|x\|_\infty + 2\|x\|_\infty\sqrt{2\log(2p)}. \tag{20}$$

Finally, applying the bounds (19) and (20) to the definition of the function f in (17), we have

$$\begin{aligned}
\frac{\mathbb{E}\|x + Bz\|_2}{\mathbb{E}\|x + Bz\|_\infty} &> \frac{\sqrt{\|x\|_2^2 + \|x\|_\infty^2(p-n) - \|x\|_\infty^2}}{\|x\|_\infty + 2\|x\|_\infty\sqrt{2\log(2p)}} \\
&= \frac{\sqrt{\frac{\|x\|_2^2}{\|x|_\infty^2} + (p-n) - 1}}{1 + 2\sqrt{2\log(2p)}} \qquad (21) \\
&\geq \frac{\sqrt{p-n}}{1 + 2\sqrt{2\log(2p)}},
\end{aligned}$$

which proves $\mathbb{E}[f(z)] > 0$, as needed. □

We now apply Lemma 3 to prove Theorem 3.

Proof of Theorem 3. We begin by making several reductions. First, it is enough to show that

$$\inf_{A\in\mathbb{R}^{n\times p}} \inf_{\delta:\mathbb{R}^n\to\mathbb{R}} \sup_{x\in\mathbb{R}^p\setminus\{0\}} \Big|\delta(Ax) - s(x)\Big| \geq \frac{p-n-1}{2(1+2\sqrt{2\log(2p)})^2}. \tag{22}$$

To see this, note that the general inequality $s(x) \leq p$ implies

$$\left|\tfrac{\delta(Ax)}{s(x)} - 1\right| \geq \tfrac{1}{p}\big|\delta(Ax) - s(x)\big|,$$

[1]The "extra" factor of 2 inside the logarithm arises from taking the absolute value of the w_i.

and we can optimize over both sides with p being a constant. Next, for any fixed matrix $A \in \mathbb{R}^{n\times p}$, it is enough to show that

$$\inf_{\delta:\mathbb{R}^n\to\mathbb{R}} \sup_{x\in\mathbb{R}^p\setminus\{0\}} \Big|\delta(Ax) - s(x)\Big| \geq \frac{p-n-1}{2(1+2\sqrt{2\log(2p)})^2}, \tag{23}$$

as we may take the infimum over all matrices A without affecting the right hand side. To make a third reduction, it is enough to prove the same bound when $\mathbb{R}^p\setminus\{0\}$ is replaced with any smaller set, as this can only make the supremum smaller. In particular, we will replace $\mathbb{R}^p \setminus \{0\}$ with a two-point subset $\{x^\circ, \tilde{x}\} \subset \mathbb{R}^p \setminus \{0\}$, where by Lemma 1, we may choose $\tilde{x}$ and x° to satisfy $Ax^\circ = A\tilde{x}$, as well as

$$s(x^\circ) = 1, \quad \text{and} \quad s(\tilde{x}) \geq \frac{p-n}{2(1+2\sqrt{2\log(2p)})^2}.$$

We now aim to prove that

$$\inf_{\delta:\mathbb{R}^n\to\mathbb{R}} \sup_{x\in\{x^\circ,\tilde{x}\}} \Big|\delta(Ax) - s(x)\Big| \geq \frac{p-n-1}{2(1+2\sqrt{2\log(2p)})^2}, \tag{24}$$

and we will accomplish this using the classical technique of constructing a Bayes procedure with constant risk. For any decision rule $\delta : \mathbb{R}^n \to \mathbb{R}$ and any point $x \in \{x^\circ, \tilde{x}\}$, define the (deterministic) risk function

$$R(x,\delta) := \Big|\delta(Ax) - s(x)\Big|.$$

Also, for any prior π on $\{x^\circ, \tilde{x}\}$, define

$$r(\pi,\delta) := \int R(x,\delta)d\pi(x).$$

By Propositions 3.3.1 and 3.3.2 of Bickel & Doksum (2001), the inequality (24) holds if there exists a prior distribution π^* on $\{x^\circ, \tilde{x}\}$ and a decision rule $\delta^* : \mathbb{R}^n \to \mathbb{R}$ with the following three properties:

1. The rule δ^* is Bayes for π^*, i.e. $r(\pi^*, \delta^*) = \inf_\delta r(\pi^*, \delta)$.

2. The rule δ^* has constant risk over $\{x^\circ, \tilde{x}\}$, i.e. $R(x^\circ, \delta^*) = R(\tilde{x}, \delta^*)$.

3. The constant value of the risk of δ^* is at least $\frac{p-n-1}{2(1+2\sqrt{2\log(2p)})^2}$.

To exhibit π^* and δ^* with these properties, we define π^* to be the two-point prior that puts equal mass at x° and $\tilde{x}$, and we define δ^* to be the trivial decision rule that always returns the average of the two possibilities, namely $\delta^*(Ax) = \frac{1}{2}(s(\tilde{x}) + s(x^\circ))$. It is simple to check the second and third properties, namely that δ^* has constant risk equal to $\frac{1}{2}|s(\tilde{x}) - s(x^\circ)|$, and that this risk is at least $\frac{p-n-1}{2(1+2\sqrt{2\log(2p)})^2}$. It remains to check that δ^* is Bayes for π^*. This follows easily from the triangle inequality, and the fact that $\delta(A\tilde{x}) = \delta(Ax^\circ)$ holds for all δ. Namely,

$$\begin{aligned} r(\pi^*, \delta) &= \tfrac{1}{2}\Big|\delta(A\tilde{x}) - s(\tilde{x})\Big| + \tfrac{1}{2}\Big|\delta(Ax^\circ) - s(x^\circ)\Big|, \\ &\geq \tfrac{1}{2}\Big|s(\tilde{x}) - s(x^\circ)\Big| \\ &= r(\pi^*, \delta^*). \end{aligned} \tag{25}$$

□

References

Beckner, W. A generalized Poincaré inequality for Gaussian measures. *Proceedings of the American Mathematical Society*, 105(2):397–400, 1989.

Bickel, P.J. and Doksum, K.A. *Mathematical Statistics, volume I.* Prentice Hall, 2001.

David, HA. Order statistics. 1981. *J. Wiley.*

Li, P., Hastie, T., and Church, K. Nonlinear estimators and tail bounds for dimension reduction in l 1 using cauchy random projections. *Journal of Machine Learning Research*, pp. 2497–2532, 2007.

Marshall, A.W., Olkin, I., and Arnold, B.C. *Inequalities: theory of majorization and its applications.* Springer, 2010.

Massart, P. Some applications of concentration inequalities to statistics. In *Annales-Faculte des Sciences Toulouse Mathematiques*, volume 9, pp. 245–303. Université Paul Sabatier, 2000.

van der Vaart, A.W. and Wellner, J.A. *Weak convergence and empirical processes.* Springer Verlag, 1996.