

## REFINED LEAST SQUARES FOR SUPPORT RECOVERY

OFIR LINDENBAUM AND STEFAN STEINERBERGER

ABSTRACT. We study the problem of exact support recovery based on noisy observations and present Refined Least Squares (RLS). Given a set of noisy measurement

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta}^* + \boldsymbol{\omega},$$

and  $\mathbf{X} \in \mathbb{R}^{N \times D}$  which is a (known) Gaussian matrix and  $\boldsymbol{\omega} \in \mathbb{R}^N$  is an (unknown) Gaussian noise vector, our goal is to recover the support of the (unknown) sparse vector  $\boldsymbol{\theta}^* \in \{-1, 0, 1\}^D$ . To recover the support of the  $\boldsymbol{\theta}^*$  we use an average of multiple least squares solutions, each computed based on a subset of the full set of equations. The support is estimated by identifying the most significant coefficients of the average least squares solution. We demonstrate that in a wide variety of settings our method outperforms state-of-the-art support recovery algorithms.

## 1. INTRODUCTION

1.1. **The Problem.** The task of estimating the support of a sparse signal based on noisy measurements plays a key role in many applications, such as, image denoising [1, 2], communication [3], machine learning [4, 5, 6]. Here, we consider support recovery of an unknown ternary signal  $\boldsymbol{\theta}^* \in \{-1, 0, 1\}^D$ , given the noisy vector

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta}^* + \boldsymbol{\omega},$$

where  $\mathbf{X} \in \mathbb{R}^{N \times D}$  is a measurement matrix whose entries are independent random variables with mean 0 and variance 1, and the noise  $\boldsymbol{\omega} \in \mathbb{R}^N$  is Gaussian. Exploiting the known sparsity  $k$  of the unknown signal  $\boldsymbol{\theta}^*$ , our goal is to recover its support, i.e. all indices  $d$  such that  $\boldsymbol{\theta}_d^* \neq 0$ . We are interested in the case when the system is underdetermined and has more variables than equations  $D > N$ .

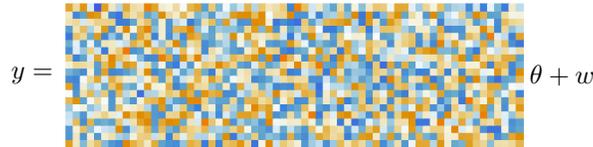


FIGURE 1. We try to recover the support of  $\theta$  from the observations  $X$  and  $y$ , where  $y = X\theta + \omega$ . The (known) matrix  $X$  is a Gaussian random matrix, so is the (unknown) noise  $\omega$ , we try to recover the support of  $\theta$  with few measurements.

---

S.S. is supported by the NSF (DMS-2123224) and the Alfred P. Sloan Foundation.

**1.2. Existing Results.** A popular approach for recovering the support of  $\boldsymbol{\theta}^*$  is using an  $\ell^1$  regularized quadratic programming problem, also known as the Least Absolute Shrinkage and Selection Operator (LASSO) [7]. The LASSO is cast as a convex problem, and can be solved using efficient optimization schemes [8, 9]. Several authors have introduced iterative algorithms improving the support recovery capabilities of the LASSO, these include: Iterative Re-weighted Least Squares (IRLS) [10], Iteratively Re-weighted  $\ell^1$  minimization (IRL1) [11], and Iterative Support Detection (ISD) [12]. Other solutions for support recovery include greedy algorithms such as Orthogonal Matching Pursuit (OMP) [13], and its extensions [14, 15, 16], or non convex schemes such as the Trimmed LASSO (TL) [17], smoothly clipped absolute deviation (SCAD) [18], or stochastic gates (STG) [19]. For a more complete survey of the existing methodologies, we refer the reader to [20, 21, 22, 23]. Following the approach of Randomly Aggregated Least Squares (RAWLS) proposed by the authors in [24], we introduce Refined Least Squares (RLS), a procedure for estimating the support of  $\boldsymbol{\theta}^*$  based on the noisy vector  $\mathbf{y}$  and the observations matrix  $\mathbf{X}$ . Our algorithm is greedy, relying on estimating the support (and sign) of one coefficient at a time. We demonstrate that Refined Least Squares (RLS) effectively recovers the support, specifically in the low information regime where  $N$  is much smaller than  $D$ , the regime where the noise is high and there is a large number of active coefficients ( $k$ ).

## 2. REFINED LEAST SQUARES

**2.1. RAWLS.** The purpose of this section is to introduce the main idea behind Refined Least Squares (RLS). Our starting point is RAWLS (**R**andomly **A**ggregated **un**Weighted **L**east **S**quares), a method introduced by the authors in [24]. The main idea behind RAWLS is somewhat counter-intuitive: we are given  $N$  equations with  $D$  variables (with special emphasis on  $N < D$ )

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta}^* + \boldsymbol{\omega}.$$

It is easy to see that

$$\widehat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta}} \|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|$$

does not usually lead to very good results (in particular,  $\widehat{\boldsymbol{\theta}}$  is not typically sparse). The main idea behind RAWLS is to do a least squares regressions on a *subset* of the equations: for any arbitrary subset  $A \subset \{1, 2, \dots, N\}$ , we use  $\mathbf{X}_A$  to denote the reduction of  $\mathbf{X}$  to those rows indexed by  $A$  and likewise for the vector  $\mathbf{y}_A$ . We can then consider

$$\widehat{\boldsymbol{\theta}}_A = \arg \min_{\boldsymbol{\theta} \in \mathbb{R}^D} \|\mathbf{X}_A \boldsymbol{\theta} - \mathbf{y}_A\|.$$

This will typically lead to even worse results compared to using least squares on the full system since we work with less information. However, since there are *many* different subsets  $A \subset \{1, 2, \dots, N\}$ , we can obtain several different estimators  $\widehat{\boldsymbol{\theta}}_{A_1}, \widehat{\boldsymbol{\theta}}_{A_2}, \dots, \widehat{\boldsymbol{\theta}}_{A_m}$  and average them. RAWLS chooses  $m$  random subsets of size  $|A_i| = \lfloor 0.6 \min(D, N) \rfloor$  and averages over the different least squares solutions. The main observation that makes RAWLS work is that the coordinate (and the sign) of these averages with the largest (absolute) entry likely corresponds to a coordinate for which  $\boldsymbol{\theta}^*$  has a nonzero entry (with the same sign). This coordinate is removed in a greedy fashion and the procedure is reiterated on the remaining problem.

**2.2. Refined Least Squares.** Refined Least Squares (RLS) is a significant refinement over RAWLS and leads to a very effective algorithm (see §3 for results). As was already pointed out in [24], we believe that the idea of suitably averaging over least square solutions computed over *subsets* of equations to be a promising idea and not yet suitably explored. We present two additional ideas which we call

- (1) *averaged guessing* and
- (2) the *OMP flip*

They lead to Refined Least Squares (RLS) which achieves state-of-the-art results in several challenging regimes. We now explain these two new ideas.

**2.2.1. Averaged Guessing.** RAWLS produces a guess for a coordinate with nonzero support by averaging least square recoveries over random subsets of size  $|A_i| = \lfloor 0.6 \min(D, N) \rfloor$ . As demonstrated in Fig. 2, the constant 0.6 leads to the decent results in terms of support recovery but not in a very strict sense – any constant in the range  $[0.6, 0.9]$  seems to lead to rather comparable behavior (this is related to the distribution of singular values of random matrices, see below). The main new idea is to run several such predictions, for  $|A_i| \in [0.85 \min(D_k, N), 0.9 \min(D_k, N)]$ , where  $D_k$  is the number of variables at peeling step  $k = 1, \dots, D$ . Then, using a majority vote among the top selected components, we get an additional layer of stability leading to better results.

**2.2.2. The OMP flip.** The second new ingredient is based on the following idea. As RAWLS proceeds in greedily selecting coordinates, the problem tends to become more difficult as the leading variables (corresponding to the strongest coefficients) are removed because the signal to noise ratio shrinks. In particular, after  $k - 1$  steps, we have a system of  $N$  equations for  $D - k + 1$  unknowns which is mainly dominated by noise once  $k \sim D$ . The final step is as follows: for  $D$  Gaussian vectors  $\mathbf{g}_1, \dots, \mathbf{g}_D \in \mathbb{R}^N$  (the columns of the matrix  $\mathbf{X}$ ), we are given  $\mathbf{y} = (\pm 1) \cdot \mathbf{g}_i + \boldsymbol{\omega}$  and are asked to reconstruct  $i$  from knowing  $\mathbf{y}$  and the  $D$  vectors  $\mathbf{g}_1, \dots, \mathbf{g}_D \in \mathbb{R}^N$ . It is difficult to imagine an estimator better than

$$\hat{i} = \arg \max_{1 \leq i \leq D} |\langle \mathbf{y}, \mathbf{g}_i \rangle|.$$

This way of reconstructing coordinates is the main idea behind Orthogonal Matching Pursuit (OMP) [13]. Once there is a single coordinate left to reconstruct, (RLS) switches to this OMP heuristic. This also leads to additional stability.

*The Algorithm.* To estimate the support we use an ensemble of estimates, each based on a different subset size  $|A_i| = n_0$ . We fix the subset size

$$n_0 \in [0.85 \min(D_k, N), 0.9 \min(D_k, N)]$$

and start the peeling procedure to estimate the set of support indices  $\mathcal{I} \subset \{1, \dots, D\}$ . Using  $m$  (we use  $m = 100$ ) random subsets of the equations (of size  $n_0$  each), the RLS estimate is based on the following approximation

$$\bar{\boldsymbol{\theta}} = \frac{1}{m} \sum_{i=1}^m \hat{\boldsymbol{\theta}}_{A_i}.$$

We estimate a support coefficient as  $\hat{\theta}_\ell = \text{sign}(\max \bar{\boldsymbol{\theta}})$ , where  $\ell = \arg \max |\bar{\boldsymbol{\theta}}|$  is the index of the support variable. The process is repeated 5 times with different subset size  $n_0 \in [0.85 \min(D_k, N), 0.9 \min(D_k, N)]$ . Finally, the set of support indices  $\mathcal{I}$  is estimated as using a majority voting over the different sets identified in the 5 peeling

procedures: we pick the coordinate most frequently identified as most likely (in case of a tie, we choose randomly among the most frequently occurring ones). After that, we remove  $\mathbf{X}^\ell$  from  $\mathbf{X}$ , where  $\mathbf{X}^\ell$  is the  $\ell$ 's column of the matrix, and update  $\mathbf{y}$  by subtracting  $\mathbf{X}^\ell \widehat{\theta}_\ell$ . This process is repeated  $k - 1$  times for estimating the leading support indices, the last coefficient is estimated using OMP. A description of the peeling procedure is also presented in Algorithm 1.

---

**Algorithm 1** Refined Least Squares (RLS)

---

**Input:** Observations matrix  $\mathbf{X}$ , and target vector  $\mathbf{y}$ .

- 1: **for**  $d = 1 : k - 1$  **do**
  - 2:   **for**  $i = 1 : m$  **do**
  - 3:     Draw a set of equations  $A_i \subset \{1, \dots, N\}$  of size  $n_0$ .
  - 4:     Use least squares to find  $\bar{\theta}_{A_i} = \mathbf{X}_{A_i}^\dagger \mathbf{y}_{A_i}$
  - 5:   **end for**
  - 6:   Compute the approximation  $\bar{\theta} = \frac{1}{m} \sum_{i=1}^m \bar{\theta}_{A_i}$
  - 7:   Estimate  $\widehat{\theta}_\ell = \text{sign}(\max \bar{\theta})$ , where  $\ell = \arg \max |\bar{\theta}|$
  - 8:   Remove the  $\ell$  column from  $\mathbf{X}$  and update  $\mathbf{y}$  by subtracting  $\mathbf{X}^\ell \widehat{\theta}_\ell$ , where  $\mathbf{X}^\ell$  is the  $\ell$ 's column of the matrix.
  - 9: **end for**
  - 10: Estimate the  $k$ 'th coefficient using OMP.
- 

**2.3. The Theorem.** Our main theoretical contribution is a rigorous result explaining why, for this type of problem, it is actually advantageous to use a least squares approach over a *reduced* set of equations (something that is maybe counter-intuitive since we ‘throw away’ information). We will now state the theorem which is relevant for a large number of algorithms of this type. Our setting is as follows: we assume that  $N < D$  is given and that  $\mathbf{X} \in \mathbb{R}^{N \times D}$  is a random Gaussian matrix with each entry being an independent  $\mathcal{N}(0, 1)$  random variable. Let us assume that the vector  $\boldsymbol{\theta}^* \in \mathbb{R}^D$  is fixed and that  $\boldsymbol{\omega} \in \mathbb{R}^D$  is a Gaussian perturbation. We are given  $\mathbf{y} = \mathbf{X}\boldsymbol{\theta}^* + \boldsymbol{\omega}$ . We compute a least squares approximation  $\mathbf{X}^\dagger \mathbf{y}$ , where  $\mathbf{X}^\dagger = \mathbf{X}^T (\mathbf{X}\mathbf{X}^T)^{-1}$ , and are interested in how close this approximation is to  $\mathbf{X}^\dagger \boldsymbol{\theta}^*$ . This means that the main question is regarding the size of  $\|\mathbf{X}^\dagger \mathbf{y} - \mathbf{X}^\dagger \mathbf{X}\boldsymbol{\theta}^*\|$ : we hope that the least squares reconstruction of  $\mathbf{y}$  is very close to that of  $\mathbf{X}\boldsymbol{\theta}^*$ . The subsequent Theorem tells us that how close these two numbers are depends on the dimensions of  $\mathbf{X} \in \mathbb{R}^{N \times D}$ : their distance grows as  $N/D$  gets closer to 1.

**Theorem.** *If the size of the matrix  $\mathbf{X}$  tends to infinity, with  $N/D < 1$  fixed, and  $\omega_i \sim \mathcal{N}(0, 1)$ , we have*

$$\mathbb{E}_{\mathbf{X}, \boldsymbol{\omega}} \|\mathbf{X}^\dagger \mathbf{y} - \mathbf{X}^\dagger \mathbf{X}\boldsymbol{\theta}^*\| = (1 + o(1)) \sqrt{\frac{N}{D - N}}.$$

We note that  $\mathbb{E}\|\mathbf{X}^\dagger \mathbf{X}\boldsymbol{\theta}^*\|$  may be thought of as the projection onto a random hyperplane and we may thus assume that

$$\mathbb{E}\|\mathbf{X}^\dagger \mathbf{X}\boldsymbol{\theta}^*\| \sim \sqrt{\frac{N}{D}} \|\mathbf{X}\boldsymbol{\theta}^*\|.$$

This shows that both  $\|\mathbf{X}^\dagger \mathbf{X}\boldsymbol{\theta}^*\|$  and  $\|\mathbf{X}^\dagger \mathbf{y} - \mathbf{X}^\dagger \mathbf{X}\boldsymbol{\theta}^*\|$  have a priori same dependence on the ratio of the dimensions of  $\mathbf{X}$ .  $N/D$  should not be too small because

this amounts to throwing away a lot of information. However, it also should not be too close to 1 either since in that case the size of the error starts growing dramatically: this is because the matrix  $\mathbf{X}$  suddenly starts having singular values closer to 0 which badly influence the inversion problem. The main ingredient is the Marchenko-Pastur Theorem which holds for general random rectangular matrices whose random variables are independent and have mean value 0 and variance 1: the Theorem also applies to measurement matrices that are not Gaussian (as we demonstrate in Section 3). A numerical evaluation supporting the Theorem appears in Fig. 3.

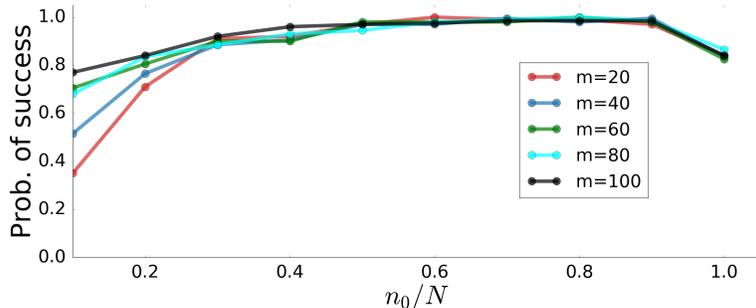


FIGURE 2. Probability of successfully recovering the support of the signal using RLS with a subset of size  $n_0$  (x-axis presents the normalized subset size  $n_0/N$ ). In this experiment, we use  $D = 64$  variables  $N = 50$  equations and a sparsity of  $k = 20$ . We use several values for  $m$ , the number of subsets used for estimating the support via Algorithm 1.

### 3. EXPERIMENTAL RESULTS

In the following subsection we evaluate the support recovery capabilities of RLS in different settings. We compare the performance of RLS to several strong baselines, such as: the LASSO [7], IRL1 [11], TL [17], OMP [13], ISD [12] and RAWLS [24]. We apply RLS with  $m = 100$  which was shown stable across different settings (as demonstrated in Fig. 2 smaller values of  $m$  may also work in practice). We evaluate performance in terms of the empirical probability for perfect support recovery. This probability is estimated as the portion of simulations which obtained perfect support recovery out of 200 runs. Perfect support recovery is counted only if  $S(\hat{\boldsymbol{\theta}}) = S(\boldsymbol{\theta})$ , where  $S(\boldsymbol{\theta}) := \{i \in 1, \dots, D | \boldsymbol{\theta}_i \neq 0\}$ .

For the first example (see Fig. 4), we use a design matrix  $\mathbf{X}$  with values drawn independently from  $N(0, 1)$ , the number of variables is  $D = 64$  with sparsity  $k = 30$ . This example demonstrates the advantage of RLS over several strong baselines for different noise levels ( $\sigma = 0.5$  and  $\sigma = 1$ ). Then (see Fig. 5), we repeat the experiment but use a correlated design matrix  $\mathbf{X}$ . To generate  $\mathbf{X}$ , we first construct a Toeplitz covariance matrix  $\boldsymbol{\Sigma}$  with values  $\Sigma_{i,j} = 0.3^{|i-j|}$ , where  $i, j = 1, \dots, D$ , then we generate  $\mathbf{X}$  by drawing its values from  $N(0, \boldsymbol{\Sigma})$ . In the next experiment (see Fig. 6), we investigate what happens when the design matrix  $\mathbf{X}$  takes its entries from a fair Bernoulli distribution. Each entry  $\mathbf{X}_{ij}$  is taken independently

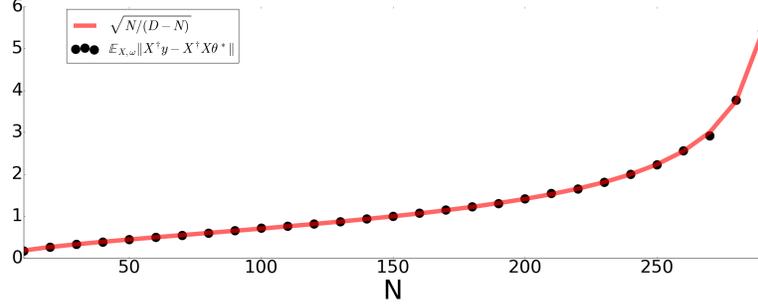


FIGURE 3. Numerical evaluation of the relation predicted by Theorem 1 for  $D = 300$  and  $10 \leq N \leq 290$ . As the number of equations  $N$  tends to  $D$  (the number of variables), the expected  $\ell^2$  norm of  $\mathbf{X}^\dagger \mathbf{y} - \mathbf{X}^\dagger \mathbf{X} \boldsymbol{\theta}^*$  (black dots) grows like  $N^{1/2}(D-N)^{-1/2}$  (red line).

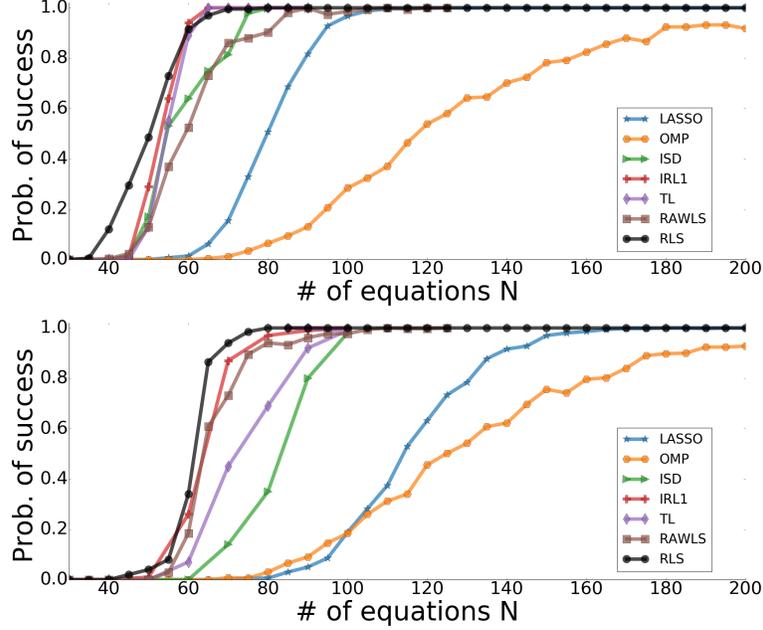


FIGURE 4. Evaluating the probability of exact support recovery vs. the number of measurements  $N$ . We use  $D = 64$  variables, with a sparsity  $k = 30$ , and a design matrix  $\mathbf{X}$  with values drawn from  $N(0, 1)$ . We compare the proposed procedure (RAWLS) to several baselines for:  $\sigma = 0.5$  (top panel) and  $\sigma = 1$  (bottom panel).

and uniformly at random from  $\{-1, 1\}$ . In the last experiment (see Fig. 7), we investigate the behavior of these methods with respect to sparsity. We fix  $D = 64$  variables,  $N = 40$  equations and assume the error on the right-hand side to be of scale  $\sigma = 1$ . We observe that RLS can recover the true support with some nonzero (albeit small) likelihood even when the solution is far from sparse ( $k = 30$ ).

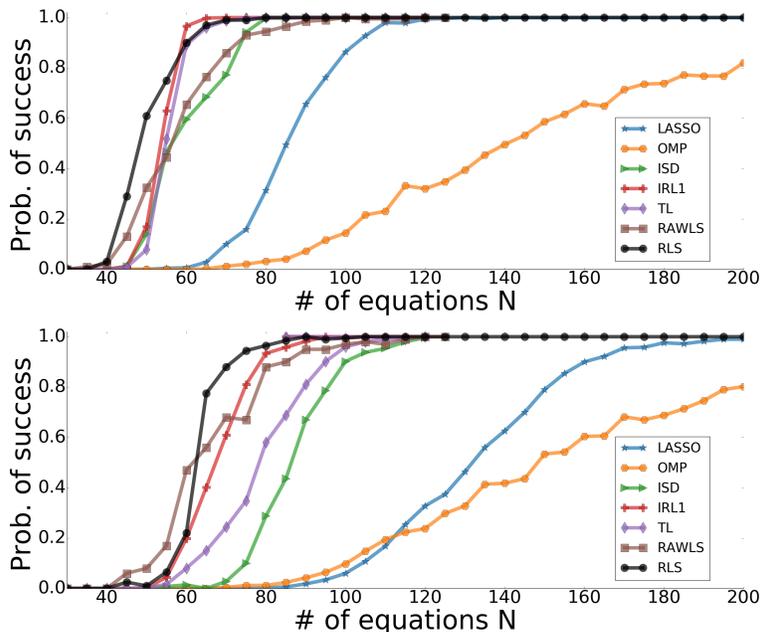


FIGURE 5. Evaluating the probability of exact support recovery vs. the number of measurements  $N$ . We use  $D = 64$  variables, with a sparsity  $k = 30$ , and a design matrix  $\mathbf{X}$  with values drawn from  $\mathcal{N}(0, \Sigma)$ , where  $\Sigma_{i,j} = 0.3^{|i-j|}$ . We compare the proposed procedure (RAWLS) to several baselines for:  $\sigma = 0.5$  (top panel) and  $\sigma = 1$  (bottom panel).

#### 4. PROOF OF THE THEOREM

*Proof.* We will use  $\mathbf{X}^\dagger \mathbf{y}$  to denote the  $\ell^2$ -smallest vector satisfying

$$\mathbf{X}^\dagger \mathbf{y} = \arg \min_{\boldsymbol{\theta} \in \mathbb{R}^D} \|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|.$$

This solution is contained in the vector space  $V$  spanned by the rows of the matrix (if  $\boldsymbol{\theta}$  had a component that was orthogonal to these rows, then it would not have any effect in the matrix multiplication  $\mathbf{X}\boldsymbol{\theta}$  and removing that component would result in a smaller  $\ell^2$ -norm). Since the number of variables,  $D$ , is larger than the number of equations,  $N$ , and  $\mathbf{X}$  is Gaussian we know that the minimum is 0 with likelihood 1 (the system is underdetermined and thus has a solution with likelihood 1). By linearity, we have  $\mathbf{X}^\dagger(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}^*) = \mathbf{X}^\dagger \boldsymbol{\omega}$  and thus  $\|\mathbf{X}^\dagger \mathbf{y} - \mathbf{X}^\dagger \mathbf{X}\boldsymbol{\theta}^*\| = \|\mathbf{X}^\dagger \boldsymbol{\omega}^*\|$ . At this point, we fix the random matrix  $\mathbf{X}$  and take an expectation over  $\boldsymbol{\omega}$ . We observe that each entry of  $\boldsymbol{\omega}$  is an independent  $\mathcal{N}(0, 1)$  Gaussian. Let us now assume that the singular value decomposition of  $\mathbf{X}$  is given by  $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^*$  and thus  $\mathbf{X}^\dagger = \mathbf{V}\boldsymbol{\Sigma}^\dagger\mathbf{U}^*$ . The matrices  $\mathbf{U}$  and  $\mathbf{V}$  are unitary and the distribution of Gaussian vectors is invariant under an application of orthogonal matrices, thus  $\mathbb{E}_\omega \|\mathbf{X}^\dagger \boldsymbol{\omega}\| = \mathbb{E}_\omega \|\boldsymbol{\Sigma}^\dagger \boldsymbol{\omega}\|$ . Writing  $\boldsymbol{\omega} = (\omega_1, \dots, \omega_N)$  and writing  $\sigma_1, \dots, \sigma_N$  for the

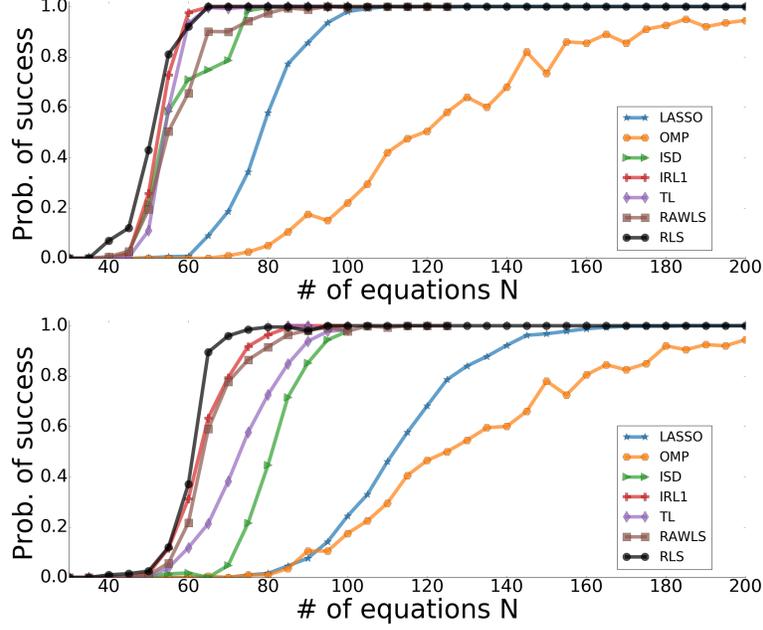


FIGURE 6. Evaluating the probability of exact support recovery vs. the number of measurements  $N$ . We use  $D = 64$  variables, with a sparsity  $k = 30$ , and a design matrix  $\mathbf{X}$  with values drawn from a fair Bernoulli distribution. We compare the proposed procedure (RAWLS) to several baselines for:  $\sigma = 0.5$  (top panel) and  $\sigma = 1$  (bottom panel).

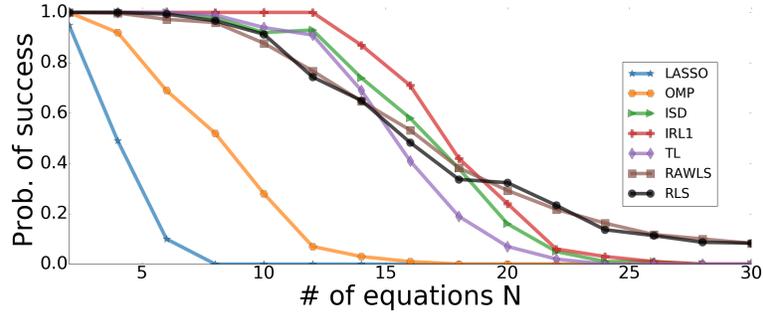


FIGURE 7. Evaluating the probability of exact support recovery vs. sparsity level  $k$ . We use  $D = 64$  variables, with  $N = 40$  observations,  $\sigma = 1$  and different sparsity levels.

singular values of  $\mathbf{X}$ , we can use the explicit form of  $\Sigma^\dagger$  to compute

$$\mathbb{E}_\omega \|\Sigma^\dagger \boldsymbol{\omega}\| = \mathbb{E}_\omega \left( \sum_{i=1}^N \frac{\omega_i^2}{\sigma_i^2} \right)^{1/2}.$$

Jensen's inequality leads to

$$\begin{aligned} \mathbb{E}_\omega \left( \sum_{i=1}^N \frac{\omega_i^2}{\sigma_i^2} \right)^{1/2} &\leq \left( \mathbb{E}_\omega \sum_{i=1}^N \frac{\omega_i^2}{\sigma_i^2} \right)^{1/2} \\ &= \left( \frac{1}{N} \sum_{i=1}^N \frac{1}{\sigma_i^2} \right)^{1/2} \sqrt{N}. \end{aligned}$$

We note that this inequality is strict, however, standard deviation inequalities imply that the inequality becomes asymptotically sharp because the sum starts to tightly concentrate around its mean value as the dimension of the matrix increases. The distribution of singular values of rectangular matrices of dimension  $N \times D$  is given by the Marchenko-Pastur distribution with parameter  $0 \leq N/D \leq 1$ . Using  $f_{N/D}$  to denote the density of the Marchenko-Pastur distribution, we see that, in the limit as  $N, D \rightarrow \infty$  with  $N/D$  fixed,

$$\mathbb{E}_X \frac{1}{N} \sum_{i=1}^N \frac{1}{\sigma_i^2} = (1 + o(1)) \frac{1}{D} \int_0^\infty \frac{f_{N/D}(x)}{x} dx.$$

It remains to analyze the integral. The Marchenko-Pastur distribution is given by

$$f_\lambda(x) = \frac{1}{2\pi} \frac{\sqrt{(\lambda_+ - x)(x - \lambda_-)}}{\lambda x}$$

where  $\lambda_\pm = (1 \pm \sqrt{\lambda})^2$  and  $0 < \lambda < 1$  is the ratio of the random matrix. The antiderivative of this function when weighted with  $1/x$  happens to have a closed form

$$\begin{aligned} &\int \frac{\sqrt{(\lambda_+ - x)(x - \lambda_-)}}{x^2} dx = \\ &= \frac{1}{x} \sqrt{-\lambda^2 - (x-1)^2 + 2\lambda(1+x)} \\ &+ \arctan \left( \frac{1 + \lambda - x}{\sqrt{-\lambda^2 - (x-1)^2 + 2\lambda(1+x)}} \right) \\ &- \frac{1 + \lambda}{1 - \lambda} \arctan \left( \frac{x + \lambda(2+x) - \lambda^2 - 1}{(\lambda - 1)\sqrt{\lambda^2 + 2\lambda(1+x) - (1+x)^2}} \right). \end{aligned}$$

Taking appropriate limits  $x \rightarrow \pm\lambda_\pm$ , we obtain

$$\int_{\lambda_-}^{\lambda_+} \frac{\sqrt{(\lambda_+ - x)(x - \lambda_-)}}{x^2} dx = \frac{2\lambda\pi}{1 - \lambda}$$

from which we deduce

$$\int_0^\infty \frac{f_{N/D}(x)}{x} dx = \frac{1}{1 - ND^{-1}} = \frac{D}{D - N}.$$

□

## REFERENCES

- [1] M. Elad and M. Aharon, “Image denoising via learned dictionaries and sparse representation,” in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06)*, vol. 1. IEEE, 2006, pp. 895–900.
- [2] Y.-Q. Zhao and J. Yang, “Hyperspectral image denoising via sparse representation and low-rank constraint,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 1, pp. 296–308, 2014.
- [3] Z. Qin, J. Fan, Y. Liu, Y. Gao, and G. Y. Li, “Sparse representation for wireless communications: A compressive sensing approach,” *IEEE Signal Processing Magazine*, vol. 35, no. 3, pp. 40–58, 2018.
- [4] K. Guo, P. Ishwar, and J. Konrad, “Action recognition using sparse representation on covariance manifolds of optical flow,” in *2010 7th IEEE international conference on advanced video and signal based surveillance*. IEEE, 2010, pp. 188–195.
- [5] J. Yang, J. Wang, and T. Huang, “Learning the sparse representation for classification,” in *2011 IEEE International Conference on Multimedia and Expo*. IEEE, 2011, pp. 1–6.
- [6] O. Lindenbaum, U. Shaham, J. Svirsky, E. Peterfreund, and Y. Kluger, “Let the data choose its features: Differentiable unsupervised feature selection,” *arXiv preprint arXiv:2007.04728*, 2020.
- [7] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.
- [8] Y. Nesterov, “Gradient methods for minimizing composite functions,” *Mathematical Programming*, vol. 140, no. 1, pp. 125–161, 2013.
- [9] J. Qian, W. Du, Y. Tanigawa, M. Aguirre, R. Tibshirani, M. A. Rivas, and T. Hastie, “A fast and flexible algorithm for solving the lasso in large-scale and ultrahigh-dimensional problems,” *BioRxiv*, p. 630079, 2019.
- [10] I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk, “Iteratively reweighted least squares minimization for sparse recovery,” *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, vol. 63, no. 1, pp. 1–38, 2010.
- [11] E. J. Candes, M. B. Wakin, and S. P. Boyd, “Enhancing sparsity by reweighted  $\ell_1$  minimization,” *Journal of Fourier analysis and applications*, vol. 14, no. 5-6, pp. 877–905, 2008.
- [12] Y. Wang and W. Yin, “Sparse signal reconstruction via iterative support detection,” *SIAM Journal on Imaging Sciences*, vol. 3, no. 3, pp. 462–491, 2010.
- [13] J. A. Tropp and A. C. Gilbert, “Signal recovery from random measurements via orthogonal matching pursuit,” *IEEE Transactions on information theory*, vol. 53, no. 12, pp. 4655–4666, 2007.
- [14] M. Elad and I. Yavneh, “A plurality of sparse representations is better than the sparsest one alone,” *IEEE Transactions on Information Theory*, vol. 55, no. 10, pp. 4701–4714, 2009.
- [15] D. L. Donoho, Y. Tsaig, I. Drori, and J.-L. Starck, “Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit,” *IEEE transactions on Information Theory*, vol. 58, no. 2, pp. 1094–1121, 2012.
- [16] D. Needell and J. A. Tropp, “Cosamp: Iterative signal recovery from incomplete and inaccurate samples,” *Applied and computational harmonic analysis*, vol. 26, no. 3, pp. 301–321, 2009.
- [17] D. Bertsimas, M. S. Copenhaver, and R. Mazumder, “The trimmed lasso: Sparsity and robustness,” *arXiv preprint arXiv:1708.04527*, 2017.
- [18] J. Fan and R. Li, “Variable selection via nonconcave penalized likelihood and its oracle properties,” *Journal of the American statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.
- [19] Y. Yamada, O. Lindenbaum, S. Negahban, and Y. Kluger, “Feature selection using stochastic gates,” *arXiv preprint arXiv:1810.04247*, 2018.
- [20] Y. Arjoune, N. Kaabouch, H. El Ghazi, and A. Tamtaoui, “Compressive sensing: Performance comparison of sparse recovery algorithms,” in *2017 IEEE 7th annual computing and communication workshop and conference (CCWC)*. IEEE, 2017, pp. 1–7.
- [21] A. M. Bruckstein, D. L. Donoho, and M. Elad, “From sparse solutions of systems of equations to sparse modeling of signals and images,” *SIAM review*, vol. 51, no. 1, pp. 34–81, 2009.

- [22] E. C. Marques, N. Maciel, L. Naviner, H. Cai, and J. Yang, “A review of sparse recovery algorithms,” *IEEE Access*, vol. 7, pp. 1300–1322, 2018.
- [23] S. Mousavi, M. M. R. Taghiabadi, and R. Ayanzadeh, “A survey on compressive sensing: Classical results and recent advancements,” *arXiv preprint arXiv:1908.01014*, 2019.
- [24] O. Lindenbaum and S. Steinerberger, “Randomly aggregated least squares for support recovery,” *Signal Processing*, vol. 180, p. 107858, 2021.

PROGRAM IN APPLIED MATHEMATICS, YALE UNIVERSITY, NEW HAVEN, CT 06511, USA  
*Email address:* `ofir.lindenbaum@yale.edu`

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF WASHINGTON, SEATTLE, WA 98195, USA  
*Email address:* `steinerb@uw.edu`