Randomly Aggregated Least Squares for Support Recovery

Ofir Lindenbaum[1], Stefan Steinerberger[2,*]

Abstract

We study the problem of exact support recovery: given an (unknown) vector $\boldsymbol{\theta}^* \in \{-1, 0, 1\}^D$ with known sparsity $k = \|\boldsymbol{\theta}^*\|_0$, we are given access to the noisy measurement

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

where $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ is a (known) Gaussian matrix and the noise $\boldsymbol{\omega} \in \mathbb{R}^N$ is an (unknown) Gaussian vector. How small can N be for reliable recovery of the support of $\boldsymbol{\theta}^*$? We present RAWLS (**R**andomly **A**ggregated un**W**eighted **L**east Squares **S**upport Recovery): the main idea is to take random subsets of the N equations, perform least squares over this reduced bit of information, and average over many random subsets. We show that the proposed procedure can provably recover an approximation of $\boldsymbol{\theta}^*$ and demonstrate its use through numerical examples. We use numerical simulations to demonstrate that the proposed procedure is beneficial for the task of support recovery. Finally, we observe that RAWLS is at par with several strong baselines in the low information regime (i.e. N is small or k is large).

Keywords: Support Recovery, Compressed Sensing, Least Squares.

1. Introduction

The problem of support recovery, plays an important role in machine learning, signal processing, bioinformatics, and high dimensional statistics. In some applications, identifying the support leads to direct benefits such as reduction of memory and computational costs [1], identification of cancer risk genes [2]. In other tasks, such as image denoising [3], the coefficients $\boldsymbol{\theta}$ are of interest; based on the recovered support these could be estimated using least squares.

[1]Program in Applied Mathematics, Yale University, New Haven, CT 06511, USA
[2]Department of Mathematics, University of Washington, Seattle, WA 98195, USA

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

In the regime $N < D$, the support recovery problem (illustrated in Fig. 1) is under-determined: we have fewer equations N than variables D, and the observations are contaminated by additive noise $\boldsymbol{\omega}$. In this setting sparsity is a useful assumption and it would be natural to estimate $\boldsymbol{\theta}^*$ by minimizing

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\theta}\|_0 \leq k.$$

Since optimizing over this equation is intractable; several authors have replaced the ℓ^0 norm by the ℓ^1, which induces sparsity and leads to the well known Least Absolute Shrinkage and Selection Operator (LASSO) [4]. We note that the sparsity properties of ℓ^p norms, for $p > 1$ was studied in [5]. The LASSO, typically formulated using a regularized version of the problem, enjoys efficient optimization schemes [6, 7]. [8] showed that exact support recovery using the LASSO can occur with probability one if $N > 2k \log(D - k)$. Several iterative methods for support recovery have been proposed, including: Iterative Support Detection (ISD) [9], the iteratively reweighted least squares (IRLS) [10] and the iteratively reweighted ℓ^1 minimization (IRL1) [11]. The problem has also been addressed using greedy methods such as Orthogonal Matching Pursuit (OMP) [12], Random OMP [13] and other extensions [14, 15], or non convex schemes such as Trimmed LASSO (TL) [16] or smoothly clipped absolute deviation (SCAD) [17]. Recently, in [18], the authors proposed a constrained matching pursuit algorithm for support recovery. The importance of the problem has made it quite impossible to give an accurate, complete summary of the literature: we refer to the surveys [19, 20, 21, 22].

In this study, we propose RAWLS (**R**andomly **A**ggregated un**W**eighted **L**east Squares **S**upport Recovery), a simple scheme for support recovery from noisy measurements

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

where $\boldsymbol{\theta}^* \in \{-1, 0, 1\}^D$ is a sparse vector. The ternary model for $\boldsymbol{\theta}^*$ is motivated by several applications such as: compressing neural networks [23, 24] or representing biological signals [25]. RAWLS relies on subsampling the full set of equation and performing least squares on each subset. After averaging over the different solutions, we estimate the support using the most significant coefficients of the least squares solution. We prove a bound on the approximation of

$\boldsymbol{\theta}^*$ based on the proposed procedure. Finally, we demonstrate the applicability of RAWLS to the task of support recovery using different sparsity and noise levels. We observe, that our method outperforms several leading baselines in the low information regime.

The paper is structured as follows. In Section 2 we describe and motivate the proposed idea and our main results. Then, in Section 3 we provide a full description of RAWLS and demonstrate it efficacy using several examples. Finally, in Section 4 we prove our main result.

2. The Idea and the Main Result

2.1. The Idea.

Our idea is quite simple: to estimate $\boldsymbol{\theta}^*$, we will use least squares. This naive approach is a bad idea since

$$\widehat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta} \in \mathbb{R}^D} \|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y}\|_2^2$$

tends to require a fairly large number of queries N to recover $\boldsymbol{\theta}^*$ stably. The proposed scheme is based on the following observation: instead of running least squares on the full set of equations, we can use only a random subset of the equations. The underlying idea behind RAWLS (**R**andomly **A**ggregated Un**w**eighted **L**east **S**quares **S**upport Recovery)[3] is that none of the equations are distinguished: taking merely a subset of them amounts to a loss of information but provides a particularly unique point of view. However, since no particular subset of the equations is distinguished over any other subset, we average over a number of randomly selected subsets. Our analysis shows that this is indeed advantageous: while applying least squares using fewer equations leads to errors from the lack of information, these errors cancel (to some degree) when averaged. More precisely, let $A \subset \{1, \ldots, N\}$, we define $\boldsymbol{X}_A$ to be the restriction of $\boldsymbol{X}$ onto the rows whose index is in the set A and likewise for $\boldsymbol{y}_A$. We then find

$$\widehat{\boldsymbol{\theta}}_A = \arg \min_{\boldsymbol{\theta} \in \mathbb{R}^D} \|\boldsymbol{X}_A \boldsymbol{\theta} - \boldsymbol{y}_A\|_2^2. \tag{1}$$

We average this result over many subsets $(A_i)_{i=1}^m$ which we assume, for some fixed $n < \min(N, D)$, to be taken uniformly at random from all $n-$element subsets of $\{1, 2, \ldots, N\}$ and use this as our estimate for a rescaling of $\boldsymbol{\theta}^*$. We hope that

$$\frac{1}{m} \sum_{i=1}^m \widehat{\boldsymbol{\theta}}_{A_i} \sim \frac{n}{D} \boldsymbol{\theta}^*.$$

---

[3]'The natural distribution is neither just nor unjust; nor is it unjust that persons are born into society at some particular position. These are simply natural facts. What is just and unjust is the way that institutions deal with these facts.' (John Rawls, 'A Theory of Justice' [26]).
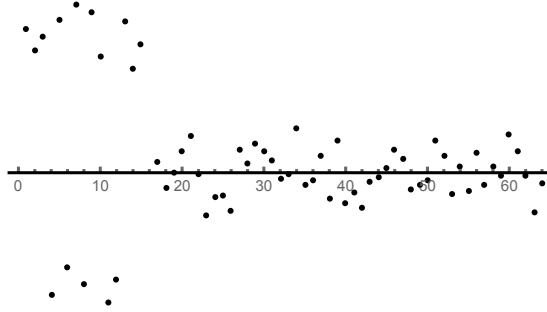
Figure 2: The reconstructed vector is much larger on the support of θ than off the support of θ and correctly identifies its sign.

An example (see Fig. 2) is as follows: let us define $\boldsymbol{\theta} \in \mathbb{R}^{64}$ by setting the first $k = 16$ entries to be ± 1 (randomly) and the rest to be 0. We take a random Gaussian matrix $\boldsymbol{X} \in \mathbb{R}^{64 \times 80}$, take subsets of size $n = 58$ equations and average the least-square recovery over $m = 100$ random choices of these 58 equations. We observe that the reconstructed vector is much larger on the actual support than it is off the support; moreover, it correctly identifies the sign of the entry of $\boldsymbol{\theta}$.
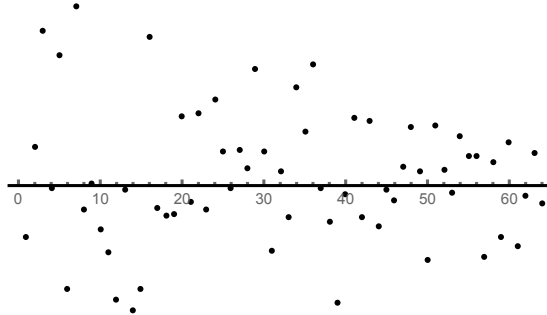

Figure 3: Reconstructing a noisy vector in $\mathbb{R}^D$, $D = 64$ (supported on the first 16 coordinates) using $N = 30$ equations (projected on $n = 18-$dimensional subsets) with RAWLS.

Once we go down to a smaller number of equations N, something remarkable happens. For simplicity of exposition, we consider the same problem as above (reconstruction of a vector in $D = 64$ dimensions) except now we only observe $N = 30$ equations and we average over random subsets of these equations of size $n = 18$. We emphasize that this quite the extreme setting; we are operating with very little information. This is reflected in the reconstructed vector (see Fig. 3): it is certainly not the case that the largest $k = 16$ entries (by absolute value)

4

correspond to the support of θ. *However*, what we observe in this setting is the largest entry is indeed located on the support of θ: for this particular choice of parameters $(D, N, n) = (64, 30, 18)$, this happens in $\sim 90\%$ of all cases. This motivated our RAWLS-based peeling algorithm discussed in §3, where we iteratively remove the coordinate corresponding to the largest reconstructed vector. We note that correctly identifying the first coordinate is the most difficult task; after that we have reduced the problem by decreasing the size of the support, one less dimension $D \to D - 1$, and the same number of equations N. This is an easier problem.

2.2. The Result

We can show that this yields provably good results. Before formally stating the result, we will quickly outline its meaning. Instead of trying to recover the vector $\boldsymbol{\theta}^*$, we will try to reconstruct its rescaled version $(n/D)\boldsymbol{\theta}^*$ via an

$$\text{average over random projections} \qquad \frac{1}{m} \sum_{i=1}^{m} \pi_{A_i} \boldsymbol{\theta}^*,$$

where π_{A_i} denotes the projection onto the subspace A_i and the A_i are, by an abuse of notation, subspaces of size n chosen uniformly at random (subspaces spanned by the rows of $\boldsymbol{X}$ indexed by A_i). However, we do not have access to $\boldsymbol{\theta}^* \in \mathbb{R}^D$, we only have access to $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}^* + \boldsymbol{\omega}$. Instead of taking a least squares projection of $\boldsymbol{y}$, we will use the least squares projections of $\boldsymbol{y}_{A_i}$ for random subsets of the equations in the hope that this approximately recovers $\boldsymbol{\theta}^*$

$$\frac{1}{m} \sum_{i=1}^{m} \widehat{\boldsymbol{\theta}}_{A_i} \sim \frac{n}{D} \boldsymbol{\theta}^*.$$

Theorem. *Let $\boldsymbol{\theta}^* \in \mathbb{R}^D$ be an arbitrary vector. Then, by projecting onto subsets of $n < 0.9 \cdot D$ equations of the N equations given by $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}^* + \boldsymbol{\omega}$, we have*

$$\mathbb{E}_{X,\omega} \left\| \frac{1}{m} \sum_{i=1}^{m} \pi_{A_i} \boldsymbol{\theta}^* - \frac{1}{m} \sum_{i=1}^{m} \widehat{\boldsymbol{\theta}}_{A_i} \right\|_{\ell^2} \lesssim \frac{n}{\sqrt{N}\sqrt{D-2}} + \frac{n}{D}.$$

Several remarks are in order.

1. The statement is *independent* of $\boldsymbol{\theta}^*$. In particular, there is no underlying assumption about the structure of $\boldsymbol{\theta}^*$ (and $\boldsymbol{\theta}^*$ need not be sparse). We also observe that the size of $\boldsymbol{\theta}^*$ does not appear on the right-hand side. This respects the problem setup where instead of $\boldsymbol{X}\boldsymbol{\theta}^*$ we are given the (additive) noisy version $\boldsymbol{X}\boldsymbol{\theta}^* + \boldsymbol{\omega}$, where $\boldsymbol{\omega}$ is a standard Gaussian vector $\omega_i \sim \mathcal{N}(0, 1)$.
2. In the setting $n \leq N \ll D$ our analysis is accurate down to constants, both quantities on the right-hand side are asymptotically correct (in the sense of having the correct constant, 1, in front) if the scales separate more and more (see Figs. 4, 5 and the Remark in §4.2).

5

3. The randomness in the choice of the A_i is not at all necessary. In fact, the proof suggests that one could simply pick completely deterministic subsets of the N equations as long as none of the individual dimensions are featured too prominently and all are represented roughly an equal number of times in the projections. This is also substantiated by numerical evidence. This poses the question of whether there are 'good' deterministic choices of subsets or whether there is a natural weight one could assign to the outcome resulting from each subset of equations (some 'measure of reliability').

4. The result suggests that using a smaller n leads to a smaller error. However, it also leads to a smaller projection. We can compensate for that by inserting the appropriate scaling in our result from which we obtain

$$\mathbb{E}_{X,\omega} \left\| \underbrace{\frac{1}{m}\frac{D}{n}\sum_{i=1}^{m}\pi_{A_i}\boldsymbol{\theta}^*}_{\approx\boldsymbol{\theta}^*} - \frac{1}{m}\frac{D}{n}\sum_{i=1}^{m}\widehat{\boldsymbol{\theta}}_{A_i} \right\|_{\ell^2} \lesssim \frac{\sqrt{D}}{\sqrt{N}} + 1.$$

This shows that there is some flexibility in the choice of n. In practice we have found that $n = 0.6\min(N, D)$ seems to be particularly suited (though not very different from, say, $n = 0.5\min(N, D)$). The precise role of n could be an interesting object for further study.

We conclude by showing Theorem 1 in a simple example. As mentioned above, the terms in the upper bound (without the implicit constant and constant 1 instead) correspond to the sharp asymptotic limiting case where $n \leq N \ll D$. We show the case where $D = 10000$, $N = 100$ and $1 \leq n \leq 100$. For each value of n, we sample over $m = 20$ random subsets of size n of the N equations. As for the vector θ, it does not actually play a role, we chose it to be a Gaussian vector in $\mathbb{R}^D$. We observe that the prediction is quite accurate (and the proof explains why this would be the case – various quantities start concentrating tightly around their expectation).

We also quickly illustrate that the restriction $n < 0.9 \cdot D$ is not just an artifact of the proof but, in fact, necessary (this also explains why RAWLS is better at recovering θ than an application of least squares to the full set of equations). We consider θ to be a unit vector (obtained from normalizing an instance of a Gaussian vector) in $D = 200$ dimensions. We take $N = 200$ equations and see what happens for $1 \leq n \leq 195$ (see Figure 5). What we observe is that the theoretical error bound (with the implicit constant assumed to be 1) nicely dominates the error until n starts getting very close to D (we plot it for $1 \leq n \leq 195 < 200 = D$). We see that the error starts exceeding the size of the vector by many orders of magnitude. The proof will explain this as a degeneracy of the smallest singular value of a rectangular Gaussian matrix which becomes approximately square.

The error observed around $n = 195 \sim N = D$ demonstrates why least squares
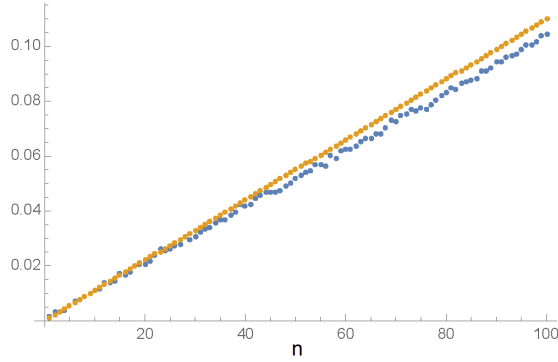
Figure 4: The error bounds in Theorem 1 (orange; ignoring the implicit constant) compared to the actual error (blue) for $D = 10000$, $N = 100$ and $1 \leq n \leq 100$.
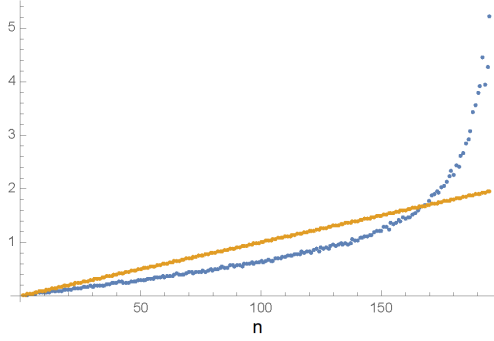

Figure 5: The error bounds in Theorem 1 (orange; ignoring the implicit constant) compared to the actual error (blue) for $D = 200 = N$ and $1 \leq n \leq 195$.

using the full set of equations does not work; we obtain similar results also for $N \ll D$, the averaging has a natural stabilizing effect. We refer to the Remark in §4.2. for a prediction for what one would expect the error to look like when, say, $n = 0.99D$.

2.3. Open Problems.

Theorem 1 raises a lot of open questions. Is there a particularly natural choice of subspaces on which to project? We are investigating the case of random projections but the proof does not seem to require this; are there natural 'adapted' subspaces that one can derive from a given matrix X?

We conclude with a particularly interesting question. We recall that the random projections reduce the size of the resulting vector. We can compensate for that

7

by inserting the appropriate scaling in our result from which we obtain

$$\mathbb{E}_{\mathrm{X},\omega} \left\| \underbrace{\frac{1}{m}\frac{D}{n}\sum_{i=1}^{m}\pi_{A_i}\theta}_{\approx\theta} - \frac{1}{m}\frac{D}{n}\sum_{i=1}^{m}x_{A_i}^* \right\|_{\ell^2} \lesssim \frac{\sqrt{D}}{\sqrt{N}}+1.$$

In the case where we assume $\theta \in \{-1,0,1\}^D$, we want to make sure that we are properly able to distinguish two different vectors of that type and this can then be seen to require $N \sim D$ (not entirely surprising, we are not making any assumptions on the sparsity of θ). However, a more refined approach is conceivable: ultimately, we are using the entries of our approximating vector to derive statements about the support. As such, the ℓ^2 is perhaps not the only interesting quantity and estimates on ℓ^∞ would be quite desirable. In particular, what we observe in practice (and what motivated the peeling algorithm) is that very large entries (either very large or very small) in the recovered approximation is a good indicator for θ having support in that coordinate. This simple observations forms the basis of the algorithm discussed in §3. It would be interesting to have results in that direction.

We also emphasize that the idea underlying RAWLS might have many other applications: it is ultimately an ℓ^2-based concept and as such many natural variations seem conceivable. One such applications, a nonlinear variant that is shown to work particularly well in the support recovery problem, is discussed in the next section. A second question, outside the scope of this paper, is whether other methods used for support recovery could conceivably be merged with our philosophy: running it on random subsets of the equations and hoping that the averaging effects compensates for the loss of information.

3. Support Recovery with RAWLS

3.1. The Idea.
If it is indeed the case that

$$\frac{1}{m}\sum_{i=1}^{m}\pi_{A_i}\boldsymbol{\theta}^* \approx \frac{n}{D}\boldsymbol{\theta}^* + \text{some error}$$

and *if* the error is nicely random (as one usually expects in these cases), then the largest (or smallest) entries of the vector should be contained in the support of $\boldsymbol{\theta}^*$. In a more elementary formulation, if we are given $\boldsymbol{v} \in \{-1,0,1\}^D$ (such that $\|\boldsymbol{v}\|_0$ is not too small compared to D) and add a random Gaussian vector $\boldsymbol{g}$ to it, then the largest (absolute) entry of $\boldsymbol{v}+\boldsymbol{g}$ will be attained (with high probability) on the support of $\boldsymbol{v}$. This is a simple consequence of the rapid decay of the Gaussian, and motivates the Algorithm proposed in the following subsection.

3.2. Peeling with RAWLS

1. Compute the approximation

$$\widehat{\boldsymbol{\theta}} = \frac{1}{m} \sum_{i=1}^{m} \widehat{\boldsymbol{\theta}}_{A_i},$$

where $\hat{\boldsymbol{\theta}}$ is estimated based on Eq. 1.

2. Find the element of largest absolute value of $\widehat{\boldsymbol{\theta}}$. If this element $\widehat{\theta}_{\ell}$ is positive, then we assume that $\theta_{\ell}^* = 1$; if negative, we assume $\theta_{\ell}^* = -1$.

3. Remove the corresponding column from the matrix $\boldsymbol{X}$ and update the right-hand side $\boldsymbol{y}$ by subtracting the $\boldsymbol{X}$ projected onto the estimated co-ordinate $\widehat{\theta}_{\ell}$.

4. Return to (1) until k non-zero entries of $\widehat{\boldsymbol{\theta}}$ are estimated.

This algorithm is thus a fairly simple greedy algorithm that identifies likely candidates for the support of $\boldsymbol{\theta}^*$ by looking for particularly large entries in the RAWLS-reconstruction of $\boldsymbol{\theta}^*$. We emphasize that for this type of iterative algorithm, each step is more difficult than the next one: having found a correct entry, the problem is reduced to a simpler problem $D \to D-1$ while maintaining the same amount of information $N \to N$. We point out that the method, just as other methods, should also be suitable for *partial* recovery: finding a set of k entries that has a large overlap with the ground truth, we do not pursue this here. We do not have any theoretical guarantees for the success rate of the peeling algorithm at this point and consider this to be an interesting problem. Perhaps the most interesting question at this stage is whether there are other implementations of these underlying ideas that can yield even better results.

3.3. Numerical Performance

In this section we support the effectiveness of RAWLS using numerical simulations. We focus on the task of exact support recovery using a random Gaussian design matrix $\boldsymbol{X}$ with values drawn independently from $N(0,1)$ and random additive Gaussian noise $\boldsymbol{\omega}$ with values drawn independently from $N(0, \sigma^2)$. As baselines, we compare the method to LASSO [4], IRL1 [11], TL [16], OMP [12], RandOMP [13] and STG [27]. To evaluate the probability of exact support recovery we run each method 100 times and count the portion of successful estimations. A successful estimation of the support is counted if $S(\boldsymbol{\theta}) = S(\widehat{\boldsymbol{\theta}})$, where $S(\boldsymbol{\theta}) := \{i \in 1, ..., D | \boldsymbol{\theta}_i \neq 0\}$. To improve the stability of LASSO, after each run we select the top k coefficients of $\boldsymbol{\theta}$ as the estimated support.

First, in Figure. 6 we present the probability of successful support recovery using $D = 64$ variables, a fixed sparsity of $k = 10$ and different number of measurements N. This example demonstrates that RAWLS can successfully recover the unknown support with fewer measurement compared with OMP,RandOMP and the LASSO.

Next, we demonstrate how the sparsity level k affects the success of RAWLS in recovering the support of $\boldsymbol{\theta}^*$. We use 100 simulations with design matrix $\boldsymbol{X}$ and
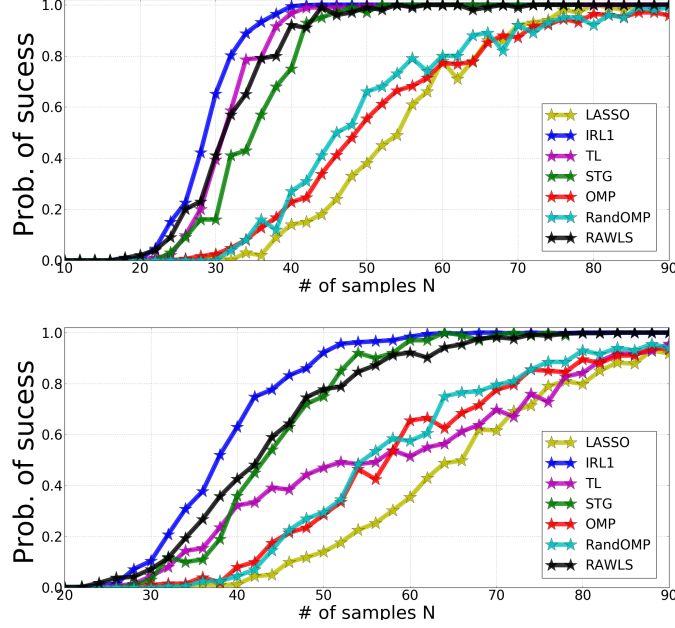
Figure 6: Numerical evaluation for the probability of exact support recovery vs. number of measurements N. We compare Peeling with RAWLS to several baselines for: $\sigma = 0.5$ (top panel) and $\sigma = 1$ (bottom panel).

Gaussian noise $\boldsymbol{\omega}$ defined as in the previous example and evaluate the performance of RAWLS for sparsity levels k in $\{2, 4, ..., 34\}$. In Figure 7 demonstrate that RAWLS success rate is comparable to IRL1, STG and TL for large values of k.

We further evaluate the performance of RAWLS for other type of measurement matrices. Specifically, we generate a Bernoulli design matrix $\boldsymbol{X}$ taking values $\{-1, 1\}$ with equal probability. We use additive Gaussian noise with zero mean and standard deviation of $\sigma = 0.5$. Here, we also compare RAWLS to stochastic resonance OMP (SR_OMP) [28]. In Fig. 8 we present the probability of support recovery using RAWLS and several other baselines. In this example, the sparsity level is $k = 10$ with dimension $D = 64$ and the results are based on 100 simulations.

Finally, we evaluate the performance of RAWLS in the regime of low information. Specifically, we use a vector with a sparsity level $k = 30$, with $D = 64$ variables, and focus on the regime of $30 \leq N \leq 90$. Here, we restrict our comparison to the leading baselines, namely to IRL1 and TL. As observed in Fig. 9, RAWLS outperforms IRL1 and TL, in the regime of low information. Precisely, for $N \leq 55$, RALWS recovers the support with a higher probability compared with the competing methods.
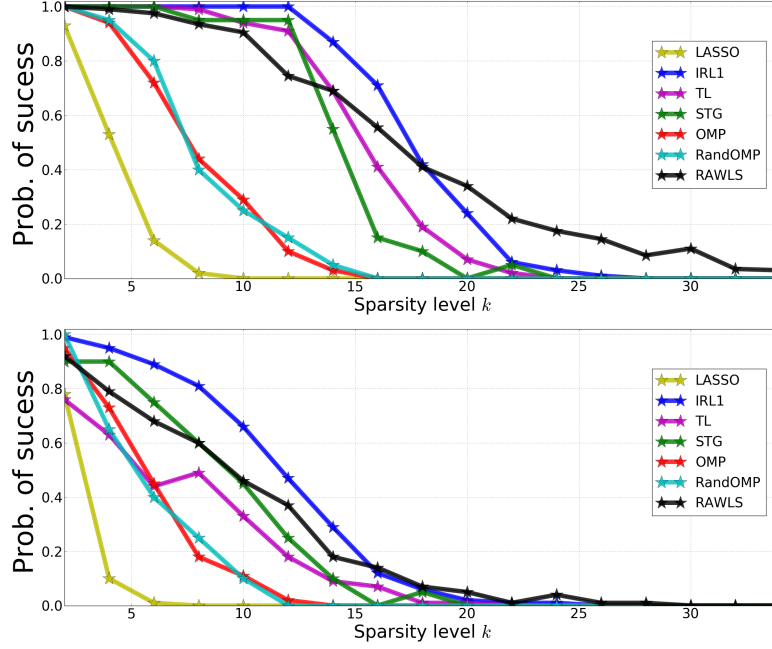
10

Figure 7: Numerical evaluation for the probability of exact support recovery vs. sparsity level k. Here the number of variables and measurements are fixed, specifically $D = 64$, and $N = 40$. We compare Peeling with RAWLS to several baselines for: $\sigma = 0.5$ (top panel) and $\sigma = 1$ (bottom panel).

Figure 8: Numerical evaluation for the probability of exact support recovery vs. number of measurements N. Here the number of variables and sparsity are fixed, specifically $D = 64$, and $k = 10$. We compare Peeling with RAWLS to several baselines for a binary design matrix with values drawn from a fair Bernoulli distribution. The additive noise is Gaussian with zero mean and standard deviation $\sigma = 0.5$
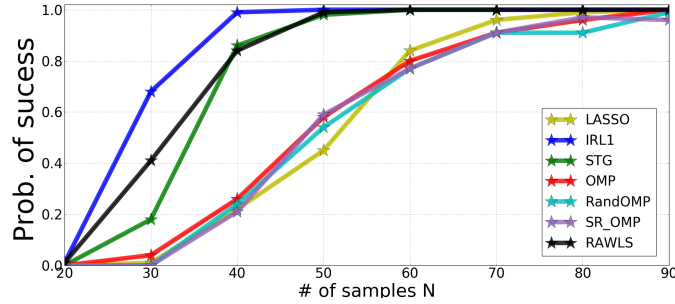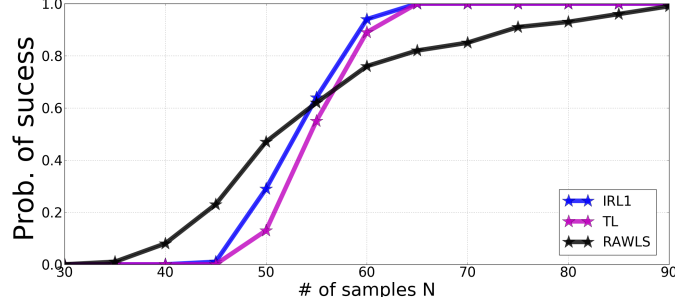
11

Figure 9: Numerical evaluation for the probability of exact support recovery vs. number of measurements N. Here the number of variables and sparsity are fixed, specifically $D = 64$, and $k = 30$. The additive noise is Gaussian with zero mean and standard deviation $\sigma = 0.5$. RAWLS outperforms state of the art method in the low information regime.

4. Proof of the Theorem

4.1. Setup.

Let $\boldsymbol{\theta}^* \in \{-1, 0, 1\}^D$ be a sparse vector with support $\|\boldsymbol{\theta}^*\|_0 = k$ and let $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ be a design matrix all of whose entries are i.i.d. random variables drawn from $\mathcal{N}(0, 1)$. We will also use the notation $g = (\boldsymbol{g}_i)_{i=1}^N$ to denote the Gaussian vectors in $\mathbb{R}^D$ dimensions that are forming the rows. We are given

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

where each entry of $\boldsymbol{\omega}$ is i.i.d. normally distributed $\boldsymbol{\omega}_i \sim \mathcal{N}(0, 1)$. We try to understand how our algorithm performs on this data. Let $A \subset \{1, \dots, N\}$ be a random subset of size $|A| = n$. We are trying to understand the least squares solution in Eq. 1 where $\boldsymbol{X}_A$ denotes the restrictions onto the rows of $\boldsymbol{X}$ indexed by A and likewise for $\boldsymbol{y}_A$. If $n \leq D$, then the system has more variables than equations and always has a solution: we are interested in the solution with the smallest ℓ^2-norm and will denote it by $\widehat{\boldsymbol{\theta}}_A$.

4.2. A Single Projection.

The purpose of this statement is to provide the analysis of a single projection onto a random subspace spanned by a random subset of the rows. The main insight is that this projection can be approximately deconstructed into the projection of the ground truth, a highly structured Gaussian error on top of that and a relatively small error term.

Lemma. *Let $\boldsymbol{\theta}^* \in \mathbb{R}^D$ be fixed, let $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ be a random Gaussian matrix and let $A \subset \{1, 2, \dots, N\}$ be a randomly chosen subset of size $|A| = n < 0.9 \cdot D$. Then the orthogonal (noisy) projection of $\boldsymbol{\theta}^*$ onto the subspace spanned by the*

12

rows indexed by A (given by $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}^* + \boldsymbol{\omega}$) satisfies

$$\widehat{\boldsymbol{\theta}}_A = \pi_A \boldsymbol{\theta}^* + \left(\sum_{a \in A} \frac{\boldsymbol{g}_a}{\|\boldsymbol{g}_a\|^2} \boldsymbol{\omega}_a \right) + \boldsymbol{e},$$

where g_a is the a-th row of the matrix $\boldsymbol{X}$, and $\boldsymbol{e}$ satisfies, with high likelihood, $\mathbb{E}_{\mathrm{X},\omega} \|\boldsymbol{e}\| \lesssim \frac{n}{D}$.

The purpose of this Lemma is to show that the (noisy) projection of $\boldsymbol{\theta}^*$ onto a random subspace (this is one interpretation of $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}^* + \boldsymbol{\omega}$) leads to substantial distortions; however, these distortions are *not* arbitrary and follow a fairly regular pattern up to a small error. The second term is not necessarily that small; however, its form will allow us to show that averaging it over multiple subspaces will further decrease the size. We emphasize that in the case $n \ll D$ our estimate is sharp and we expect $\|\boldsymbol{e}\| \sim n/D$ with tight concentration and a small error (this could be made precise when D/n becomes large).

The proof makes use of the following basic fact in linear algebra that we recall for the convenience of the reader: let $(\boldsymbol{g}_a)_{a=1}^n$ be n vectors in $\mathbb{R}^D$ with $D > n$ and let $\boldsymbol{v} \in \mathrm{span}\{\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n\}$. Then

$$\sigma_{\min}(\boldsymbol{G})^2 \|\boldsymbol{v}\|^2 \leq \sum_{a=1}^n |\langle \boldsymbol{g}_a, \boldsymbol{v} \rangle|^2 \leq \sigma_{\max}(\boldsymbol{G})^2 \|\boldsymbol{v}\|^2,$$

where σ denotes the singular values of the matrix $\boldsymbol{G}$ obtained by collecting $\{\boldsymbol{g}_a\}_{a=1}^n$ as column vectors (or, alternatively, the largest and smallest eigenvectors of $\boldsymbol{G}^T \boldsymbol{G}$). This follows easily from observing that

$$\sum_{a=1}^n |\langle \boldsymbol{g}_a, \boldsymbol{v} \rangle|^2 = \|\boldsymbol{G}^T \boldsymbol{v}\|^2.$$

This is well-known in frame theory: the frame constants for finite-dimensional problems are given by the singular values of the associated matrix.

Proof of the Lemma. We will use $\widehat{\boldsymbol{\theta}}_A$ to denote the ℓ^2-smallest vector satisfying Eq. 1. This solutions is contained in the vector space $V = \mathrm{span}\{\boldsymbol{g}_a : a \in A\}$ (if $\boldsymbol{\theta}$ had a component that was orthogonal to these rows, then it would not have any effect in the matrix multiplication $\boldsymbol{X}_A \boldsymbol{\theta}$ and removing that component would result in a smaller ℓ^2-norm). Since the number of variables, D, is larger than the number of equations, n, and $\boldsymbol{X}$ is Gaussian we know that the minimum is 0 with likelihood 1. Thus $\boldsymbol{X}_A \widehat{\boldsymbol{\theta}}_A = \boldsymbol{y}_A = \boldsymbol{X}_A \boldsymbol{\theta}^* + \boldsymbol{\omega}_A$. We will analyze this equation for a single row. For any $a \in A$,

$$\left\langle \boldsymbol{g}_a, \widehat{\boldsymbol{\theta}}_A \right\rangle = \langle \boldsymbol{g}_a, \boldsymbol{\theta}^* \rangle + \boldsymbol{\omega}_a = \left\langle \boldsymbol{g}_a, \boldsymbol{\theta}^* + \frac{\boldsymbol{\omega}_a}{\|\boldsymbol{g}_a\|^2} \boldsymbol{g}_a \right\rangle.$$

We will use this equation for all $a \in A$. By definition of the orthogonal projec-

tion, we have, for all $a \in A$, $\langle \boldsymbol{g}_a, \boldsymbol{\theta}^* \rangle = \langle \boldsymbol{g}_a, \pi_A \boldsymbol{\theta}^* \rangle$, and thus the identity

$$\left\langle \boldsymbol{g}_a, \widehat{\boldsymbol{\theta}}_A \right\rangle = \left\langle \boldsymbol{g}_a, \pi_A \boldsymbol{\theta}^* + \frac{\boldsymbol{\omega}_a}{\|\boldsymbol{g}_a\|^2} \boldsymbol{g}_a \right\rangle. \tag{2}$$

This is an interesting way of interpreting the introduction of additive noise: the error that we are given makes it seem as if the inner product was not with $\pi_A \boldsymbol{\theta}^*$ but instead with $\pi_A \boldsymbol{\theta}^*$ and a small additional multiple of $\boldsymbol{g}_a$. In practice, if $n \ll D$, then the Gaussian vectors are "almost" orthogonal and "almost" form an orthogonal basis of the space that they span. This motivates the *ansatz*

$$\widehat{\boldsymbol{\theta}}_A = \pi_A \boldsymbol{\theta}^* + \left(\sum_{a \in A} \frac{\boldsymbol{g}_a}{\|\boldsymbol{g}_a\|^2} \boldsymbol{\omega}_a \right) + \boldsymbol{e},$$

where π_A is the orthogonal projection onto the vector space $V = \mathrm{span}\,\{\boldsymbol{g}_a : a \in A\}$ and $\boldsymbol{e} \in \mathbb{R}^D$ is an error term whose size we try to investigate. We plug in our ansatz in to Eq. 2 and obtain, for all $a \in A$,

$$\langle \boldsymbol{g}_a, \boldsymbol{e} \rangle = - \left\langle \boldsymbol{g}_a, \sum_{a \neq i \in A} \frac{\boldsymbol{g}_i}{\|\boldsymbol{g}_i\|^2} \boldsymbol{\omega}_i \right\rangle.$$

We emphasize that, since the $\boldsymbol{g}_a$ span V with probability 1, these n equations uniquely identify $\boldsymbol{e} \in V$ with probability 1. We first try to understand the quantity on the right-hand side. We have

$$\left\langle \boldsymbol{g}_a, \sum_{a \neq i \in A} \frac{\boldsymbol{g}_i}{\|\boldsymbol{g}_i\|^2} \boldsymbol{\omega}_i \right\rangle = \sum_{a \neq i \in A} \frac{\langle \boldsymbol{g}_i, \boldsymbol{g}_a \rangle}{\|\boldsymbol{g}_i\|^2} \boldsymbol{\omega}_i.$$

The inner product of two random Gaussians is a random variable at scale $\langle \boldsymbol{g}_i, \boldsymbol{g}_a \rangle \sim \sqrt{D}$, the size of an individual Gaussian vector is at scale $\mathbb{E}\|\boldsymbol{g}_i\|^2 \sim D + \mathcal{O}(\sqrt{D})$ with high likelihood. The $\boldsymbol{\omega}_i \sim \mathcal{N}(0, 1)$ have an additional randomization effect. The sum runs over $n - 1$ elements. Altogether, we expect the quantity to be a random variable at scale

$$\left| \sum_{a \neq i \in A} \frac{\langle \boldsymbol{g}_i, \boldsymbol{g}_a \rangle}{\|\boldsymbol{g}_i\|^2} \boldsymbol{\omega}_i \right| \sim \frac{\sqrt{n}}{\sqrt{D}}.$$

An explicit computation shows that

$$\mathbb{E} \left| \sum_{a \neq i \in A} \frac{\langle \boldsymbol{g}_i, \boldsymbol{g}_a \rangle}{\|\boldsymbol{g}_i\|^2} \boldsymbol{\omega}_i \right|^2 = \mathbb{E} \sum_{a \neq i \in A} \frac{\langle \boldsymbol{g}_i, \boldsymbol{g}_a \rangle^2}{\|\boldsymbol{g}_i\|^4} \boldsymbol{\omega}_i^2$$

$$+ \mathbb{E} \sum_{a \neq i_1 \neq i_2 \in A} \frac{\langle \boldsymbol{g}_{i_1}, \boldsymbol{g}_a \rangle}{\|\boldsymbol{g}_{i_1}\|^2} \frac{\langle \boldsymbol{g}_{i_2}, \boldsymbol{g}_a \rangle}{\|\boldsymbol{g}_{i_2}\|^2} \boldsymbol{\omega}_{i_1} \boldsymbol{\omega}_{i_2}.$$

The second expectation is clearly 0 since $\boldsymbol{\omega}_i \sim \mathcal{N}(0,1)$ and these are independent of each other. It remains to evaluate the first expectation. Since $\boldsymbol{\omega}_i$ are independent of $\boldsymbol{g}_i$, we get

$$\mathbb{E}_{X,\omega} \sum_{a \neq i \in A} \frac{\langle \boldsymbol{g}_i, \boldsymbol{g}_a \rangle^2}{\|\boldsymbol{g}_i\|^4} \boldsymbol{\omega}_i^2 = \mathbb{E}_X \sum_{a \neq i \in A} \frac{\langle \boldsymbol{g}_i, \boldsymbol{g}_a \rangle^2}{\|\boldsymbol{g}_i\|^4}.$$

This sum can be decoupled into two parts

$$\mathbb{E}_X \sum_{a \neq i \in A} \frac{\langle \boldsymbol{g}_i, \boldsymbol{g}_a \rangle^2}{\|\boldsymbol{g}_i\|^4} = \mathbb{E}_X \sum_{a \neq i \in A} \left\langle \frac{\boldsymbol{g}_i}{\|\boldsymbol{g}_i\|}, \boldsymbol{g}_a \right\rangle^2 \frac{1}{\|\boldsymbol{g}_i\|^2}.$$

We observe that $\boldsymbol{g}_i/\|\boldsymbol{g}_i\|$ is a random vector on the unit sphere (this follows from the rotational symmetry of Gaussian vectors); as such, it is completely independent of its length $\|\boldsymbol{g}_i\|$ allowing us to treat both quantities as independent random variables. However, the first term is simply an inner product of a Gaussian vector against a unit length vector, thus

$$\left\langle \frac{\boldsymbol{g}_i}{\|\boldsymbol{g}_i\|}, \boldsymbol{g}_a \right\rangle \text{ is a Gaussian variable and } \mathbb{E} \left\langle \frac{\boldsymbol{g}_i}{\|\boldsymbol{g}_i\|}, \boldsymbol{g}_a \right\rangle^2 = 1.$$

The remaining quantity is the mean of an inverse $\chi-$distribution which is $1/(D-2)$ for $D \geq 3$ and thus

$$\mathbb{E}_X \sum_{a \neq i \in A} \frac{\langle \boldsymbol{g}_i, \boldsymbol{g}_a \rangle^2}{\|\boldsymbol{g}_i\|^4} = \mathbb{E}_X \sum_{a \neq i \in A} \frac{1}{\|\boldsymbol{g}_i\|^2} = \frac{n-1}{D-2} \lesssim \frac{n}{D}.$$

Summing up, we obtain $\mathbb{E} \sum_{a \in A} |\langle \boldsymbol{g}_a, e \rangle|^2 \lesssim \frac{n^2}{D}$. However, since $e \in \text{span}\{\boldsymbol{g}_a : a \in A\}$, we have the basic inequality

$$\sigma_{\min}^2 \|e\|^2 \leq \sum_{a \in A} |\langle \boldsymbol{g}_a, e \rangle|^2 \leq \sigma_{\max}^2 \|e\|^2.$$

The smallest singular value of a random rectangular Gaussian matrix was determined by Silverstein [29] who showed that we can expect, in the limit, that $\sigma_{\min} \sim \sqrt{D} - \sqrt{n}$. Combining all these results shows that we expect, in the regime where d has a bounded gap from D, say $n \leq 0.9 \cdot D$, that

$$\|e\| \lesssim \frac{n}{D}.$$

$\square$

Remark. We observe that the first part of the argument is fairly tight, in

particular, we expect

$$\mathbb{E} \sum_{a \in A} |\langle \boldsymbol{g}_a, \boldsymbol{e} \rangle|^2 \sim \frac{n^2}{D}$$

with tight concentration. The second part of the argument is not precise down to constants but it becomes tight if we have $n \ll D$. We observe that if $n \ll D$, then we actually have $\sigma_{\min} \sim \sigma_{\max}$ since the singular are expected to be in the interval $[\sqrt{D} - \sqrt{n}, \sqrt{D} + \sqrt{n}]$. Since all the estimates we carried out are actually quite tightly concentrated, we thus expect, with a fair degree of accuracy,

$$\|\boldsymbol{e}\| \sim \frac{n}{D}.$$

More precise, estimates are conceivable: if $\boldsymbol{e}$ is uniformly distributed across all singular vectors, then we could hope that

$$\frac{1}{\|\boldsymbol{e}\|^2} \sum_{a \in A} |\langle \boldsymbol{g}_a, \boldsymbol{e} \rangle|^2 \sim Z^2,$$

where Z is the Marchenko-Pastur distribution modeling the singular values of the random matrix X. When $n \ll D$, then $Z \sim \sqrt{D} \pm \sqrt{n} \sim \sqrt{D}$ and we recover the usual estimate. As soon as n starts approaching D, the distribution of Z gets closer and closer to 0 and the inverse distribution $1/Z^2$ spreads over many scales. However, in principle, if $\boldsymbol{e}$ is uniformly distributed over the singular vectors, then one could use this heuristic to predict the sharp constant to be expected when, for example $n = 0.99 \cdot D$. Basic numerics seems to indicate that this is a reasonable assumption.

4.3. Multiple Projections.
We now discuss the effect of averaging quantities like

$$\sum_{a \in A} \frac{\boldsymbol{g}_a}{\|\boldsymbol{g}_a\|^2} \boldsymbol{\omega}_a$$

over multiple randomly chosen sets A.

Lemma. *Let $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ be a matrix with i.i.d. standard $\mathcal{N}(0,1)$ entries and let $\boldsymbol{\omega} \in \mathbb{R}^N$ be a random vector all of whose entries are i.i.d. $\mathcal{N}(0,1)$. Let $A \subset \{1, \dots, N\}$ denote a random set of size n (chosen uniformly at random among all $n-$element subsets of A). Then*

$$\mathbb{E}_{X,\omega} \lim_{\ell \to \infty} \left\| \frac{1}{\ell} \sum_{i=1}^{\ell} \sum_{a \in A_i} \frac{\boldsymbol{g}_a}{\|\boldsymbol{g}_a\|^2} \boldsymbol{\omega}_a \right\| \leq \frac{n}{N\sqrt{D-2}}.$$

Proof. We observe that the vectors $\boldsymbol{g}_a$ are, albeit Gaussian random vectors, fixed once given and so are the $\boldsymbol{\omega}_a$. Thus, the law of large numbers implies that averaging over many randomly chosen subsets $A \subset \{1, 2, \dots, N\}$ of size A

16

results, ultimately, in each coordinate being picked the same number of times and thus

$$\lim_{\ell \to \infty} \frac{1}{\ell} \sum_{i=1}^{\ell} \sum_{a \in A_i} \frac{\boldsymbol{g}_a}{\|\boldsymbol{g}_a\|^2} \boldsymbol{\omega}_a = \frac{n}{N} \sum_{a=1}^{N} \frac{\boldsymbol{g}_a}{\|\boldsymbol{g}_a\|^2} \boldsymbol{\omega}_a.$$

We have

$$\frac{n}{N} \sum_{a=1}^{N} \frac{\boldsymbol{g}_a}{\|\boldsymbol{g}_a\|^2} \boldsymbol{\omega}_a = \frac{n}{N} \sum_{a=1}^{N} \frac{\boldsymbol{g}_a}{\|\boldsymbol{g}_a\|} \frac{\boldsymbol{\omega}_a}{\|\boldsymbol{g}_a\|}.$$

We interpret this as follows: the vector $\boldsymbol{g}_a/\|\boldsymbol{g}_a\|$ is uniformly distributed over the unit sphere in $\mathbb{R}^D$ (a consequence of the radial symmetry of the Gaussian distribution), the vector $\boldsymbol{\omega}^* = (\boldsymbol{\omega}_a/\|\boldsymbol{g}_a\|)_{a=1}^{N}$ is interpreted as a random vector. Again, as a consequence of the radial symmetry, the vector $\boldsymbol{g}_a/\|\boldsymbol{g}_a\|$ and the size $\|\boldsymbol{g}_a\|$ can be interpreted as independent random variables. We compute $\mathbb{E}_{X,\omega} \left\| \sum_{a=1}^{N} \frac{\boldsymbol{g}_a}{\|\boldsymbol{g}_a\|} \frac{\boldsymbol{\omega}_a}{\|\boldsymbol{g}_a\|} \right\|^2$ as

$$\sum_{a_1,a_2=1}^{N} \mathbb{E}_{X,\omega} \left\langle \frac{g_{a_1}}{\|g_{a_1}\|} \frac{\boldsymbol{\omega}_{a_1}}{\|\boldsymbol{g}_{a_1}\|}, \frac{\boldsymbol{g}_{a_2}}{\|\boldsymbol{g}_{a_2}\|} \frac{\boldsymbol{\omega}_{a_2}}{\|\boldsymbol{g}_{a_2}\|} \right\rangle$$

$$= \sum_{a=1}^{N} \mathbb{E}_{X,\omega} \frac{\boldsymbol{\omega}_a^2}{\|\boldsymbol{g}_a\|^2} = \sum_{a=1}^{N} \mathbb{E}_X \frac{1}{\|\boldsymbol{g}_a\|^2} = \mathbb{E} \frac{N}{\|\boldsymbol{g}\|^2}.$$

This quantity is the mean of an inverse $\chi-$distribution which is $1/(D-2)$ for $D \geq 3$. Thus, using the Cauchy-Schwarz inequality, we get

$$\mathbb{E}_{X,\omega} \frac{n}{N} \left\| \sum_{a=1}^{N} \frac{\boldsymbol{g}_a}{\|\boldsymbol{g}_a\|} \frac{\boldsymbol{\omega}_a}{\|\boldsymbol{g}_a\|} \right\| \leq \frac{n}{N} \frac{\sqrt{N}}{\sqrt{D-2}} = \frac{n}{\sqrt{N}\sqrt{D-2}}.$$

$\square$

5. Acknowledgement

References

[1] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.

[2] O. Kohannim, D. P. Hibar, J. L. Stein, N. Jahanshad, X. Hua, P. Rajagopalan, A. Toga, C. R. Jack Jr, M. W. Weiner, G. I. De Zubicaray *et al.*, "Discovery and replication of gene influences on brain structure using lasso regression," *Frontiers in neuroscience*, vol. 6, p. 115, 2012.

[3] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image processing*, vol. 15, no. 12, pp. 3736–3745, 2006.

[4] 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.

[5] J. Shen and S. Mousavi, "Least sparsity of p-norm based optimization problems with p¿1," *SIAM Journal on Optimization*, vol. 28, no. 3, pp. 2721–2751, 2018.

[6] Y. Nesterov, "Gradient methods for minimizing composite functions," *Mathematical Programming*, vol. 140, no. 1, pp. 125–161, 2013.

[7] 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.

[8] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso)," *IEEE transactions on information theory*, vol. 55, no. 5, pp. 2183–2202, 2009.

[9] 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.

[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 ℓ_1 minimization," *Journal of Fourier analysis and applications*, vol. 14, no. 5-6, pp. 877–905, 2008.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] D. Bertsimas, M. S. Copenhaver, and R. Mazumder, "The trimmed lasso: Sparsity and robustness," *arXiv preprint arXiv:1708.04527*, 2017.

[17] 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.

[18] J. Shen and S. Mousavi, "Exact support and vector recovery of constrained sparse vectors via constrained matching pursuit," *arXiv preprint arXiv:1903.07236*, 2019.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] M. Yan, M. Li, P. Roccaro, and G. V. Korshin, "Ternary model of the speciation of i-/br-/cl-trihalomethanes formed in chloraminated surface waters," *Environmental Science & Technology*, vol. 50, no. 8, pp. 4468–4475, 2016.

[24] G. Di Guglielmo, J. Duarte, P. Harris, D. Hoang, S. Jindariani, E. Kreinar, M. Liu, V. Loncar, J. Ngadiuba, K. Pedro *et al.*, "Compressing deep neural networks on fpgas to binary and ternary precision with hls4ml," *arXiv preprint arXiv:2003.06308*, 2020.

[25] H. Alemdar, V. Leroy, A. Prost-Boucle, and F. Pétrot, "Ternary neural networks for resource-efficient ai applications," in *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, pp. 2547–2554.

[26] J. Rawls, "A theory of justice," *An Imprint of Harvard University Press; 2 edition*, 1999.

[27] Y. Yamada, O. Lindenbaum, S. Negahban, and Y. Kluger, "Feature selection using stochastic gates," *arXiv preprint arXiv:1810.04247*, 2018.

[28] D. Simon, J. Sulam, Y. Romano, Y. M. Lu, and M. Elad, "Mmse approximation for sparse coding algorithms using stochastic resonance," *IEEE Transactions on Signal Processing*, vol. 67, no. 17, pp. 4597–4610, 2019.

[29] J. W. Silverstein *et al.*, "The smallest eigenvalue of a large dimensional wishart matrix," *The Annals of Probability*, vol. 13, no. 4, pp. 1364–1368, 1985.