Recovering the Sparsest Element in a Subspace

Laurent Demanet and Paul Hand

Massachusetts Institute of Technology, Department of Mathematics,
77 Massachusetts Avenue, Cambridge, MA 02139

October 2013

Abstract

We address the problem of recovering a sparse n-vector from an arbitrary basis of a subspace spanned by the vector and k random vectors. We prove that the sparse vector will be the output to one of n linear programs with high probability if its support size s satisfies $s \lesssim n/\sqrt{k \log n}$. The scaling law still holds when the desired vector is approximately sparse. To get a single estimate for the sparse vector from the n linear programs, we must select which output is the sparsest. This selection process can be based on any proxy for sparsity and has the potential to improve or worsen the scaling law. If sparsity is interpreted in an ℓ_1/ℓ_∞ sense, then the scaling law can not be better than $s \lesssim n/\sqrt{k}$. Computer simulations show that selecting the sparsest output in the ℓ_1/ℓ_2 or thresholded-ℓ_0 senses can lead to a larger parameter range for successful recovery than that given by the ℓ_1/ℓ_∞ sense.

Acknowledgements. PH is funded by an NSF Mathematical Sciences Postdoctoral Research Fellowship. LD acknowledges funding from the NSF, the Alfred P. Sloan Foundation, TOTAL S.A., AFOSR, and ONR. The authors would also like to thank Jonathan Kelner and Vladislav Voroninski for helpful discussions.

1 Introduction

We are interested in finding the sparsest nonzero element in a given subspace of $\mathbb{R}^n$. Such a task is interesting because it can be used to construct a basis of sparse vectors, which is an important part of many problems in dictionary learning [2, 14], blind source separation [18], and optimization theory [6]. Finding a sparse nonzero vector in a nullspace also has applications in spectral estimation, such as with Prony's method [3].

Minimizing sparsity in a strict-ℓ_0 sense over the nonzero elements of a subspace is NP-hard [6]. Instead, it is natural to attempt to minimize a proxy for sparsity based on the ℓ_1 norm. To prevent the zero solution from being feasible, one could introduce a constraint that the Euclidean length is unity [18]. Unfortunately, the resulting problem is nonconvex. Alternatively, one could simply set one of the coefficients to be unity. This is the approach presented in Spielman et al. [14], which provides an algorithm for finding a basis of sparse vectors by solving n linear programs. Under an appropriate scaling, and in the case where there is a basis in which every vector is exactly sparse, they show that all the basis vectors will be recovered by at least one of the linear programs. They do not study the case of approximate sparsity.

In the present paper, we will study the same approach as [14] for a different subspace model. We will establish a scaling law for recovery within a subspace spanned by a single sparse vector along with several random vectors. We will study when the desired vector is recovered by one of n linear programs. Additionally, we will show that the method is robust when the special vector is only approximately sparse. We will also comment on the process of selecting a single output from the n linear programs.

Note that the task of finding the sparsest element in a subspace is different from the standard problem of compressed sensing. First, the sparsest element is necessarily zero, which we do not allow. To get around this pathology, we consider many linear programs, each of which could be viewed as a compressed sensing problem. In these problems, the sensing matrix would have many rows that are orthogonal to the planted sparse vector, hence complicating an analysis based on restricted isometries. Because of this difficulty, we pursue an alternative proof approach.

A second major difference with compressed sensing is that there are multiple senses in which sparsity may be understood. Each sense potentially gives rise to a different scaling law, which can have a form different from that of compressed sensing. For example, if sparsity is minimized in the sense of the ratio of the ℓ_1 to ℓ_∞ norms, we show that the scaling law can not be better than $s \lesssim n/\sqrt{k}$. In contrast, the scaling law for compressed sensing says that the number of measurements required for recovery is proportional to the support size of the desired vector [5]. The dimensionality of the ambient space has only a weak effect through a logarithmic factor. In the subspace problem of this paper, the analogous result would be for the maximum recoverable sparsity to scale like $n - k$.

1.1 Sparse Recovery by n linear programs

Consider the task of finding a sparse or approximately sparse nonzero vector $v \in \mathbb{R}^n$ given a subspace W spanned by it and k random vectors. Let $\tilde{v}_1, \ldots, \tilde{v}_k$ be i.i.d. $\mathcal{N}(0, I_n)$ random vectors, and suppose W is presented to us in the form of an arbitrary basis. Concisely, our problem is as follows:

$$\begin{aligned} \text{Given} &: \{w_1, \ldots, w_{k+1}\} \text{ a basis for } W = \text{span}\{v, \tilde{v}_1, \ldots, \tilde{v}_k\} \\ \text{Find} &: v \end{aligned}$$

Recovery of v is at best possible up to a multiplicative constant that may be negative. As in [14], we attempt to recover v by solving the collection of n linear programs

$$\min \|z\|_1 \text{ such that } z \in W, z(i) = 1, \tag{1}$$

for each $1 \leq i \leq n$. We note that the constraint that $z \in W$ is equivalent to the assertion that z is orthogonal to each element of the orthogonal complement $W^\perp$. Standard linear algebra decompositions allow us to compute an orthonormal basis of $W^\perp$ from the basis of W. Hence, the subspace constraint in (1) can be written as $n - k - 1$ homogeneous linear equalities that are easily computable from any basis of W.

In order to get a single estimate for v from these n outputs, we need a selector that returns the one that is the 'sparsest.' Here, the sparsity of a vector z may be interpreted in one many precise senses, such as the strict-ℓ_0 sense with $s = \|z\|_0$; a thresholded-ℓ_0 sense with $s = \#\{i \mid |z(i)| > \epsilon\}$; the ℓ_1/ℓ_∞ sense with $s = \|z\|_1/\|z\|_\infty$; or the ℓ_1/ℓ_2 sense with $s = \|z\|_1/\|z\|_2$. In principle, the interpretation of sparsity may greatly affect the performance of the overall recovery process.

To correctly recover v, we need at least one of the linear programs (1) to return v. Further, we need to select v as being sparser than the outputs of the other linear programs. In the rest of this section, we present recovery theorems that provide a scaling law for recovery of v in at least one of the linear programs. We do not rigorously analyze the process of selecting the sparsest of the n outputs. Instead, we will present computer simulations in Section 3 that explore the effect of different sparsity selectors.

1.2 Exact Recovery by a Single Program

If v is sparse enough in a strict-ℓ_0 sense, we show that at least one of the n linear programs (1) will return v exactly. Note that $v/v(i)$ is feasible for (1). If i is one of the coefficients on which v is small, we expect recovery to fail because this feasible vector will be large in ℓ_1. Similarly, if i is one of the largest coefficients of v, recovery will be most likely. Thus, we will study the case where $i^* \in \text{argmax}_i |v(i)|$. The following theorem provides a scaling law under which v is the exactly recovered by at least this instance of the program (1).

Theorem 1. *Let $k \leq n/32$. There exists a universal constant c such that for sufficiently large n,*

$$\|v\|_0 \leq c\frac{n/\sqrt{\log n}}{\sqrt{k}} \Rightarrow \frac{v}{v(i^*)} \text{ is the unique solution to (1) for } i = i^*, \tag{2}$$

with probability at least $1 - 2e^{-n/64} - \gamma_1 e^{-\gamma_2 n/2} - ke^{-\lfloor c\sqrt{n/\log n}\rfloor} - \frac{k}{n^2}$. Here, γ_1 and γ_2 are universal constants.

From the scaling law, we observe the following scaling limits on the permissible sparsity in terms of the dimensionality of the search space:

$$k \sim 1 \Longrightarrow \|v\|_0 \lesssim n/\sqrt{\log n} \tag{3}$$

$$k \sim n \Longrightarrow \|v\|_0 \lesssim \sqrt{n}/\sqrt{\log n} \tag{4}$$

That is, a search space of constant size permits the discovery of a vector whose support size is almost a constant fraction of n. Similarly, a search space of fixed and sufficiently small fraction of the ambient dimension allows recovery of a vector whose support size is almost on the the order of the square root of that dimension.

Except for the logarithmic factor, the scaling law between n, k, and $\|v\|_0$ in Theorem 1 can not be improved. To see this, recall that $|v(i^*)| = \|v\|_\infty$ and $v/v(i^*)$ is feasible for (1) with $i = i^*$. A necessary condition for successful recovery is that the sparse vector is smaller in ℓ_1 than the minimum value attainable by the span of the random vectors:

$$\frac{\|v\|_1}{\|v\|_\infty} \leq \min \|z\|_1 \text{ such that } z \in \text{span}\{\tilde{v}_1, \ldots, \tilde{v}_k\}, z(i^*) = 1, \tag{5}$$

We will show in Section 2.1 that the right hand side of (5) scales like $n/\sqrt{k}$ when k is at most some constant fraction of n. As $\|v\|_1/\|v\|_\infty \leq \|v\|_0$ for all v, and the equality is attained for some v, we conclude that high probability recovery of arbitrary v is possible only if $\|v\|_0 \lesssim n/\sqrt{k}$. We remark that the scaling may be improvable in the context where we only seek some i for which $v/v(i)$ is the unique solution to (1).

1.3 Stable Recovery by a Single Program

The linear programs (1) can also recover an approximately sparse v. That is, if v is close to a vector that is sparse enough in a strict ℓ_0 sense, we expect to recover something close to v by solving (1) with $i = i^*$. Let v_s be the best s-sparse approximation of v. The following theorem provides a scaling law under which v is approximately the output of at least this instance of the program (1).

Theorem 2. *Let $k \leq n/32$. There exists universal constants c, C such that for sufficiently large n, for $s = \lfloor c\frac{n/\sqrt{\log n}}{\sqrt{k}} \rfloor$, and for $i = i^*$, any minimizer $z^\#$ of (1) satisfies*

$$\left\| z^\# - \frac{v}{v(i^*)} \right\|_2 \leq C \frac{\sqrt{k \log n}}{\sqrt{n}} \frac{\|v - v_s\|_1}{\|v\|_\infty} \tag{6}$$

with probability at least $1 - 2e^{-n/64} - 2e^{-n/32} - \gamma_1 e^{-\gamma_2 n/2} - ke^{-\lfloor c\sqrt{n/\log n} \rfloor} - k/n^2$.

The dependence on k and n in (6) is favorable provided that $k \lesssim n/\log n$. In the case that $k \sim n$, the error bound has a mildly unfavorable constant, growing like $\sqrt{\log n}$. The $\sqrt{n/k}$ behavior of the error constant plays the roll of the $1/\sqrt{s}$ term that arises in the noisy compressed sensing problem [5]. The estimate (6) is slightly worse, as $\sqrt{k/n} \sim k^{1/4}/\sqrt{s}$, ignoring logarithmic factors. We believe that the k and n dependence of the error bound could be improved.

1.4 Discussion

As mentioned before, successful recovery of v by the process described in Section 1.1 requires both that v is the output to (1) for some i, and that v is selected as the 'sparsest' output among all n linear programs. Computer simulations in Section 3 show that the precise sense in which sparsity is interpreted can substantially affect recovery performance, especially in the low k regime. Among the three senses of sparsity given by ℓ_1/ℓ_∞, ℓ_1/ℓ_2, and thresholded-ℓ_0, the worst empirical performance is exhibited by ℓ_1/ℓ_∞. The best performance is by thresholded-ℓ_0, though this method has the drawback of requiring a threshold parameter.

Because ℓ_1/ℓ_∞ is a proxy for sparsity, it is natural to try to find the sparsest nonzero element in the subspace W by considering

$$\min \frac{\|z\|_1}{\|z\|_\infty} \text{ such that } z \in W, z \neq 0. \tag{7}$$

We note that the scaling for successful recovery with (7) can be no better than than $s \lesssim n/\sqrt{k}$. To see this, observe that the right hand side of (5) provides an upper bound on the minimal value of $\|z\|_1/\|z\|_\infty$ for $z \in \text{span}\{\tilde{v}_1, \ldots, \tilde{v}_k\}$. Hence, outside the scaling $s \lesssim n/\sqrt{k}$, there would be random vectors that are sparser than v in the ℓ_1/ℓ_∞ sense .

We further note that solving the nonconvex program (7) is equivalent to solving all n programs (1) and selecting the one that is sparsest in the ℓ_1/ℓ_∞ sense. We do not recommend simply solving (7) as an approach for finding the sparsest vector in a subspace because computer simulations reveal that recovery can be improved merely by changing the sparsity selector to ℓ_1/ℓ_2 or thresholded-ℓ_0.

It would be desirable if we could recover v by directly solving

$$\min \frac{\|z\|_1}{\|z\|_2} \text{ such that } z \in W, z \neq 0. \tag{8}$$

The lack of convexity makes this optimization problem difficult to solve directly. A lifting procedure similar to [4] can relax (8) to an $n \times n$ semidefinite program. While the procedure squares the dimensionality, it may give rise to provable recovery guarantees under a scaling law. We are unaware of any such results in the literature. In a sense, the present paper can be viewed as a simplification of this $n \times n$ matrix recovery problem into n linear programs on vectors, provided that we are willing to change the precise proxy for sparsity that we are optimizing.

2 Proofs

To prove the theorems, we note that (1), (2), (6), and the value of i^* are all invariant to any rescaling of v. Without loss of generality, it suffices to take $\|v\|_\infty = 1$ and $v(i^*) = 1$.

We begin with some notation. Let $v(i)$ be the ith component of the vector v. Let $V = [v, \tilde{V}]$, where $\tilde{V} = [\tilde{v}_1, \tilde{v}_2, \ldots, \tilde{v}_k]$. Let $V_{i^*,:}$ be the i^*-th row of V. Write $V_{i^*,:} = [1, \tilde{a}^t]$, where $\tilde{a} \in \mathbb{R}^k$. For a set S, write $\tilde{V}_S$ and $\tilde{V}_{S^c}$ as the restrictions of $\tilde{V}$ to the rows given by S and S^c, respectively.

Our aim is to prove that v is or is near the solution to (1) when $i = i^*$. We begin by noting that $W = \text{range}(V)$. Hence, changing variables by $z = Vx$, (1) is equivalent to

$$\min \|Vx\|_1 \text{ such that } Vx(i^*) = 1 \tag{9}$$

when $i = i^*$. We will show that $x = e_1$ is the solution to (9) in the exact case and is near the solution in the noisy case. Write $x = [x(1), \tilde{x}]$ in order to separately study the behavior of x on and away from the first coefficient. Our overall proof approach is to show that if n is larger than the given scaling, a nonzero $\tilde{x}$ gives rise to a large contribution to the ℓ_1 norm of Vx from coefficients off the support of v.

2.1 Scaling Without a Sparse Vector

In this section, we derive the scaling law for the minimal ℓ_1 norm attainable in a random subspace when one of the components is set to unity. This law is the key part of the justification that the scaling in the theorems can not be improved except for the logarithmic factor. As per Section 1.4, it also provides the proof of the best possible scaling when minimizing sparsity in an ℓ_1/ℓ_∞ sense. We also present the derivation for pedagogical purposes, as it contains the key ideas and probabilistic tools we will use when proving the theorems. The rest of this section will prove the following lemma.

Lemma 3. *Let $\tilde{V}$ be an $n \times k$ matrix with i.i.d. $\mathcal{N}(0,1)$ entries, where $k \leq n/16$. With high probability,*

$$\frac{n}{\sqrt{k}} \approx \min \|\tilde{V}\tilde{x}\|_1 \text{ such that } \tilde{V}\tilde{x}(i^*) = 1. \tag{10}$$

The failure probability is exponentially small in n and k.

Proof. Because $\text{range}(\tilde{V})$ is a k-dimensional random subspace, we can appeal to the uniform equivalence of the ℓ_1 and ℓ_2 norms, as given by the following lemma.

Lemma 4. *Fix $\eta < 1$. For every y in a randomly chosen (with respect to the natural Grassmannian measure) ηn-dimensional subspace of $\mathbb{R}^n$,*

$$c_\eta \sqrt{n} \|y\|_2 \leq \|y\|_1 \leq \sqrt{n}\|y\|_2$$

with probability $1 - \gamma_1 e^{-\gamma_2 n}$ *for universal constants* γ_1, γ_2.

This result is well known [7, 10]. Related results with different types of random subspaces can be found at [13, 1, 11, 9]. Thus, with high probability,

$$\|\tilde{V}\tilde{x}\|_1 \approx \sqrt{n}\|\tilde{V}\tilde{x}\|_2 \text{ for all } \tilde{x}. \tag{11}$$

We now appeal to nonasymptotic estimates of the singular values of $\tilde{V}$. Corollary 5.35 in [17] gives that for a matrix $A \in \mathbb{R}^{n\times k}$ with $k \leq n/16$ and i.i.d. $\mathcal{N}(0,1)$ entries,

$$\mathbb{P}\Big(\frac{\sqrt{n}}{2} \leq \sigma_{\min}(A) \leq \sigma_{\max}(A) \leq \frac{3\sqrt{n}}{2}\Big) \geq 1 - 2e^{-n/32}. \tag{12}$$

Thus, with high probability,

$$\|\tilde{V}\tilde{x}\|_2 \approx \sqrt{n}\|\tilde{x}\|_2. \tag{13}$$

Combining (11) and (13), we get $\|\tilde{V}\tilde{x}\|_1 \approx n\|\tilde{x}\|_2$ with high probability. Hence, the minimum values of the following two programs are within fixed constant multiples of each other:

$$\min \|\tilde{V}\tilde{x}\|_1 \text{ such that } \tilde{V}_{i^*,:}\tilde{x} = 1 \quad \approx \quad \min n\|\tilde{x}\|_2 \text{ such that } \tilde{V}_{i^*,:}\tilde{x} = 1. \tag{14}$$

By the Cauchy-Schwarz inequality and concentration estimates of the length of a Gaussian vector, any feasible point in the programs (14) satisfies

$$\|\tilde{x}\|_2 \geq \frac{1}{\|\tilde{V}_{i^*,:}\|_2} \approx \frac{1}{\sqrt{k}}, \tag{15}$$

with failure probability that decays exponentially in k. Considering the $\tilde{x}$ for which the inequality in (15) is achieved, we get that the minimal value to the right hand program in (14) scales like $n/\sqrt{k}$, proving the lemma. □

2.2 Proof of Theorem 1

The proof of Theorem 1 hinges on the following lemma. Let S be a superset of the support of v. Relative to the candidate $x = e_1$, any nonzero $\tilde{x}$ gives components on S^c that can only increase $\|Vx\|_1$. Nonzero $\tilde{x}$ can give components on S that decrease $\|Vx\|_1$. If the ℓ_1 norm of $\tilde{V}\tilde{x}$ on S^c is large enough and the ℓ_1 norm of $\tilde{V}\tilde{x}$ on S is small enough, then the minimizer to (9) must be e_1.

Lemma 5. *Let* $V = [v, \tilde{V}]$ *with* $\|v\|_\infty = 1$, $V_{i^*,:} = [1, \tilde{a}^t]$, *supp*$(v) \subseteq S$, *and* $|S| = s$. *Suppose that* $\|\tilde{V}_S\tilde{x}\|_1 \leq 2s\|\tilde{x}\|_1$ *and* $\|\tilde{V}_{S^c}\tilde{x}\|_1 \geq (2\|\tilde{a}\|_\infty + 2)s\|\tilde{x}\|_1$ *for all* $\tilde{x}$. *Then,* e_1 *is the unique solution to* (9).

Proof. For any x, observe that

$$\begin{aligned} \|Vx\|_1 &= \|vx(1) + \tilde{V}_S\tilde{x}\|_1 + \|\tilde{V}_{S^c}\tilde{x}\|_1 && (16)\\ &\geq \|v\|_1|x(1)| - 2s\|\tilde{x}\|_1 + \|\tilde{V}_{S^c}\tilde{x}\|_1 && (17)\\ &\geq \|v\|_1|x(1)| + 2\|\tilde{a}\|_\infty s\|\tilde{x}\|_1 && (18) \end{aligned}$$

where the first inequality is from the upper bound on $\|\tilde{V}_S\tilde{x}\|_1$ and the second inequality is from the lower bound on $\|\tilde{V}_{S^c}\tilde{x}\|_1$. Note that $x = e_1$ is feasible and has value $\|Ve_1\|_1 = \|v\|_1$. Hence, at a minimizer $\tilde{x}^\#$,

$$\|v\|_1|x^\#(1)| + 2\|\tilde{a}\|_\infty s\|\tilde{x}^\#\|_1 \leq \|v\|_1. \tag{19}$$

Using the constraint $x^\#(1) + \tilde{a}^t\tilde{x}^\# = 1$, a minimizer must satisfy

$$\|v\|_1(1 - \|\tilde{a}\|_\infty\|\tilde{x}^\#\|_1) + 2\|\tilde{a}\|_\infty s\|\tilde{x}^\#\|_1 \leq \|v\|_1. \tag{20}$$

Noting that $\|v\|_1 \leq s$, a minimizer must satisfy

$$2\|\tilde{a}\|_\infty s\|\tilde{x}^\#\|_1 \leq \|\tilde{a}\|_\infty s\|\tilde{x}^\#\|_1. \tag{21}$$

Hence, $\tilde{x}^\# = 0$. The constraint provides $x^\#(1) = 1$, which proves that e_1 is the unique solution to (9). □

To prove Theorem 1 by applying Lemma 5, we need to study the minimum value of $\|\tilde{V}_{S^c}\tilde{x}\|_1/\|\tilde{x}\|_1$ for matrices $\tilde{V}_{S^c}$ with i.i.d. $\mathcal{N}(0,1)$ entries. Precisely, we will show the following lemma.

Lemma 6. *Let A be a $n \times k$ matrix with i.i.d. $\mathcal{N}(0,1)$ entries, with $k \leq n/16$. There is a universal constant $\tilde{c}$, such that with high probability, $\|Ax\|_1/\|x\|_1 \geq \tilde{c}n/\sqrt{k}$ for all $x \neq 0$. This probability is at least $1 - 2e^{-n/32} - \gamma_1 e^{-\gamma_2 n}$.*

Proof of Lemma 6. We are to study the problem

$$\min \|Ax\|_1 \text{ such that } \|x\|_1 = 1, \tag{22}$$

which is equivalent to

$$\min \|Ax\|_1 \text{ such that } \|x\|_1 \geq 1. \tag{23}$$

The minimum value of (23) can be bounded from below by that of

$$\min \|Ax\|_1 \text{ such that } \|x\|_2 \geq 1/\sqrt{k} \tag{24}$$

because the feasible set of (23) is included in the feasible set of (24). We now write both the objective and constraint in terms of Ax. To that end, we apply the lower bound in (12) to get

$$\mathbb{P}(\|x\|_2 \leq 2\frac{\|Ax\|_2}{\sqrt{n}} \text{ for all } x) \geq 1 - 2e^{-n/32} \tag{25}$$

The feasible set of (24) is contained by the set $\{x \mid \|Ax\|_2 \geq \frac{1}{2}\sqrt{\frac{n}{k}}\}$ with high probability. Hence, a lower bound to (24) is with high probability given by

$$\min \|Ax\|_1 \text{ such that } \|Ax\|_2 \geq \frac{1}{2}\sqrt{\frac{n}{k}} \tag{26}$$

In order to find a lower bound on (26), we apply Lemma 4 to the range of A, which is a k-dimensional random subspace of $\mathbb{R}^n$ with $k \leq n/16$. Taking $\eta = 1/16$, we see that with high probability, the minimal value of (26) is bounded from below by $\frac{c_\eta}{2}\frac{n}{\sqrt{k}}$. The minimal value of (26), and hence of (22), is bounded from below by $\tilde{c}n/\sqrt{k}$ for some universal constant $\tilde{c}$ with probability at least $1 - 2e^{-n/32} - \gamma_1 e^{-\gamma_2 n}$. □

To prove the theorem by applying Lemma 5, we also need to study the maximum value of $\|\tilde{V}_S\tilde{x}\|_1/\|\tilde{x}\|_1$ for matrices $\tilde{V}_S$ with i.i.d. $\mathcal{N}(0,1)$ entries.

Lemma 7. *Let A be a $s\times k$ matrix with i.i.d. $\mathcal{N}(0,1)$ entries. Then $\sup_{x\neq 0}\|Ax\|_1/\|x\|_1 \leq 2s$ with probability at least $1-ke^{-s}$.*

Proof. Note that elementary matrix theory gives that the $\ell_1\to\ell_1$ operator norm of A is

$$\max_{x\neq 0}\frac{\|Ax\|_1}{\|x\|_1} = \max_{1\leq i\leq k}\|Ae_i\|_1 \tag{27}$$

As Ae_i is an $s\times 1$ vector of i.i.d. standard normals, we have

$$\mathbb{P}(\|Ae_i\|_1 > t) \leq 2^s e^{-t^2/2s} \tag{28}$$

Hence,

$$\mathbb{P}(\max_i \|Ae_i\|_1 > t) \leq k2^s e^{-t^2/2s} \tag{29}$$

Taking $t=2s$, we conclude

$$\mathbb{P}(\max_i \|Ae_i\|_1 > 2s) \leq k2^s e^{-2s} \leq ke^{-s} \tag{30}$$

□

We can now combine Lemmas 5, 6, and 7 to prove Theorem 1.

Proof of Theorem 1. Let $\tilde{c}$ be the universal constant given by Lemma 6 and let $c=\tilde{c}/5$. We will show that for $\|v\|_0 \leq c\frac{n/\sqrt{\log n}}{\sqrt{k}}$, the minimizer to (1) is v with at least the stated probability.

Let S be any superset of $\mathrm{supp}(v)$ with cardinality $s=\lfloor c\frac{n/\sqrt{\log n}}{\sqrt{k}}\rfloor$. As per Lemma 5, e_1 is the solution to (9), and hence v is the unique solution to (1), if the following events occur simultaneously:

$$\|\tilde{V}_S\tilde{x}\|_1 \leq 2s\|\tilde{x}\|_1 \text{ for all } \tilde{x} \tag{31}$$

$$\|\tilde{a}\|_\infty \leq 2\sqrt{\log n} \tag{32}$$

$$\|\tilde{V}_{S^c}\tilde{x}\|_1 \geq 5\sqrt{\log n}s\|\tilde{x}\|_1 \text{ for all } \tilde{x} \tag{33}$$

Applying Lemma 7 to the $s\times k$ matrix $\tilde{V}_S$, we get that (31) holds with probability at least $1-ke^{-s} = 1-ke^{-\lfloor c\sqrt{n/\log n}\rfloor}$. Classical results on the maximum of a gaussian vector establishes that (32) holds with probability at least $1-k/n^2$. Because $s\leq n/2$ and $k\leq n/32$, we have that $\tilde{V}_{S^c}$ has height at least $n/2$ and width at most $n/32$. Hence, Lemma 6 gives that $\|\tilde{V}_{S^c}\tilde{x}\|_1/\|\tilde{x}\|_1 \geq \tilde{c}n/\sqrt{k}$ for all $\tilde{x}\neq 0$ with probability at least $1-2e^{-n/64}-\gamma_1 e^{-\gamma_2 n/2}$. Because $s\leq \frac{\tilde{c}}{5}\frac{n/\sqrt{\log n}}{\sqrt{k}}$, we conclude (33), allowing us to apply Lemma 5. Hence, successful recovery occurs with probability at least $1-2e^{-n/64}-\gamma_1 e^{-\gamma_2 n/2}-ke^{-\lfloor c\sqrt{n/\log n}\rfloor}-k/n^2$.

□

2.3 Proof of Theorem 2

We will prove the following lemma, of which Theorem 2 is a special case.

Lemma 8. *Let $k \le n/32$. There exists universal constants c, C such that for sufficiently large n, for all $s \le c\frac{n/\sqrt{\log n}}{\sqrt{k}}$, and for $i = i^*$, any minimizer $z^\#$ of (1) satisfies*

$$\left\| z^\# - \frac{v}{v(i^*)} \right\|_2 \le C \frac{\sqrt{n}}{s} \frac{\|v - v_s\|_1}{\|v\|_\infty} \tag{34}$$

with probability at least $1 - 2e^{-n/64} - 2e^{-n/32} - \gamma_1 e^{-\gamma_2 n/2} - ke^{-s} - k/n^2$.

At first glance, this lemma appears to have poor error bounds for large n and poor probabilistic guarantees for small s. On further inspection, the bounds can be improved by simply considering a larger s, possibly even larger than the size of the support of v. Larger values of s simultaneously increase the denominator and decrease the s-term approximation error in the numerator of (34). Taking the largest permissible value $s = \lfloor c\frac{n/\sqrt{\log n}}{\sqrt{k}} \rfloor$, we arrive at Theorem 2.

Lemma 8 hinges on the following analog of Lemma 5.

Lemma 9. *Fix $1 \le s < n$ and $\alpha > 0$. Let $V = [v, \tilde{V}]$ with $\|v\|_\infty = 1$, $V_{i^*,:} = [1, \tilde{a}^t]$, $\delta = \|v - v_s\|_1$, supp$(v) \subseteq S$, and $|S| = s$. If $\|\tilde{V}_S \tilde{x}\|_1 \le 2s\|\tilde{x}\|_1$ and $\|\tilde{V}_{S^c}\tilde{x}\|_1 \ge (2\|\tilde{a}\|_\infty + 2 + \alpha)s\|\tilde{x}\|_1$ for all $\tilde{x} \in \mathbb{R}^k$, then any $x^\#$ minimizing (9) satisfies*

$$|x_1^\# - 1| \le \frac{2\delta}{s}, \qquad \text{and} \qquad \|\tilde{x}^\#\|_1 \le \frac{2\delta}{s(\|\tilde{a}\|_\infty + \alpha)}. \tag{35}$$

Proof. For any x, observe that

$$\begin{aligned}
\|Vx\|_1 &= \|v \cdot x(1) + \tilde{V}_S \tilde{x}\|_1 + \|\tilde{V}_{S^c}\tilde{x}\|_1 && (36)\\
&\ge \|v\|_1 |x(1)| - 2s\|\tilde{x}\|_1 + \|\tilde{V}_{S^c}\tilde{x}\|_1 && (37)\\
&\ge \|v\|_1 |x(1)| + (2\|\tilde{a}\|_\infty + \alpha)s\|\tilde{x}\|_1 && (38)\\
&\ge (\|v_s\|_1 - \delta)|x(1)| + (2\|\tilde{a}\|_\infty + \alpha)s\|\tilde{x}\|_1 && (39)
\end{aligned}$$

where the first inequality is from the upper bound on $\|\tilde{V}_S\tilde{x}\|_1$ and the second inequality is from the lower bound on $\|\tilde{V}_{S^c}\tilde{x}\|_1$. Note that $x = e_1$ is feasible and has value $\|Ve_1\|_1 = \|v\|_1 \le \|v_s\|_1 + \delta$. Hence, at a minimizer $\tilde{x}^\#$,

$$(\|v_s\|_1 - \delta)|x^\#(1)| + (2\|\tilde{a}\|_\infty + \alpha)s\|\tilde{x}^\#\|_1 \le \|v_s\|_1 + \delta. \tag{40}$$

Using the constraint $x^\#(1) + \tilde{a}\tilde{x}^\# = 1$, a minimizer must satisfy

$$(\|v_s\|_1 - \delta)(1 - \|\tilde{a}\|_\infty \|\tilde{x}^\#\|_1) + (2\|\tilde{a}\|_\infty + \alpha)s\|\tilde{x}^\#\|_1 \le \|v_s\|_1 + \delta. \tag{41}$$

Noting that $\|v_s\|_1 \le s$, a minimizer must satisfy

$$\|\tilde{x}^\#\|_1 \le \frac{2\delta}{(\|\tilde{a}\|_\infty + \alpha)s}. \tag{42}$$

Applying the constraint again, we get

$$|x^\#(1) - 1| \le \frac{2\delta}{s}. \tag{43}$$

□

We now complete the proof of Theorem 2 by proving Lemma 8.

Proof of Lemma 8. Let $\tilde{c}$ be the universal constant given by Lemma 6 and let $c = \tilde{c}/6$. We will show that for any $s \leq c\frac{n/\sqrt{\log n}}{\sqrt{k}}$, the minimizer to (1) is near v with at least the stated probability.

Let S be any superset of $\mathrm{supp}(v_s)$ with cardinality s. Applying Lemma 9 with $\alpha = \sqrt{\log n}$, we observe that a minimizer $x^\#$ to (9) satisfies $|x^\#(1) - 1| \leq 2\delta/s$ and $\|\tilde{x}^\#\|_1 \leq 2\delta/(s\sqrt{\log n})$ if the following events occur simultaneously:

$$\|\tilde{V}_S\tilde{x}\|_1 \leq 2s\|\tilde{x}\|_1 \text{ for all } \tilde{x} \tag{44}$$

$$\|\tilde{a}\|_\infty \leq 2\sqrt{\log n} \tag{45}$$

$$\|\tilde{V}_{S^c}\tilde{x}\|_1 \geq 6\sqrt{\log n}s\|\tilde{x}\|_1 \text{ for all } \tilde{x} \tag{46}$$

Applying Lemma 7 to the $s \times k$ matrix $\tilde{V}_S$, we get that (44) holds with probability at least $1 - ke^{-s}$. Classical results on the maximum of a gaussian vector establishes that (45) holds with probability at least $1 - k/n^2$. Because $s \leq n/2$ and $k \leq n/32$, we have that $\tilde{V}_{S^c}$ has height at least $n/2$ and width at most $n/32$. Hence, Lemma 6 gives that $\|\tilde{V}_{S^c}\tilde{x}\|_1/\|\tilde{x}\|_1 \geq \tilde{c}n/\sqrt{k}$ for all $\tilde{x} \neq 0$ with probability at least $1 - 2e^{-n/64} - \gamma_1 e^{-\gamma_2 n/2}$. If $s \leq \frac{\tilde{c}}{6}\frac{n/\sqrt{\log n}}{\sqrt{k}}$, we conclude (46), allowing us to apply Lemma 5.

It remains to show that $Vx^\#$ is near v. Observe that

$$\begin{aligned}
\|Vx^\# - v\|_2 &= \|Vx^\# - Ve_1\|_2 && (47)\\
&\leq \|v\|_2|x^\#(1) - 1| + \|\tilde{V}\tilde{x}^\#\|_2 && (48)\\
&\leq \|v\|_2|x^\#(1) - 1| + \sigma_{\max}(\tilde{V})\|\tilde{x}^\#\|_2 && (49)\\
&\leq \sqrt{n}|x^\#(1) - 1| + \frac{3}{2}\sqrt{n}\|\tilde{x}^\#\|_1 && (50)\\
&\leq \sqrt{n}\frac{2\delta}{s} + \frac{3}{2}\sqrt{n}\frac{2\delta}{s\sqrt{\log n}} && (51)\\
&\leq C\frac{\sqrt{n}}{s}\delta && (52)
\end{aligned}$$

The the third inequality uses the fact that $\|v\|_\infty = 1$ and $\sigma_{\max}(\tilde{V}) \leq \frac{3}{2}\sqrt{n}$, which occurs with probability at least $1 - 2e^{-n/32}$ due to the upper bound in (12). Hence, approximate recovery occurs with probability at least $1 - 2e^{-n/64} - 2e^{-n/32} - \gamma_1 e^{-\gamma_2 n/2} - ke^{-s} - k/n^2$.

□

3 Simulations

In this section, we present computer simulations that demonstrate when solving the n linear programs (1) can find an approximately sparse vector $v \in \mathbb{R}^n$ from a subspace spanned by it and k random vectors. We demonstrate that recovery performance depends significantly on the precise sense of sparsity used when selecting the sparsest of the n outputs.

Let $n = 100$, $1 \leq s \leq n$, and $S = \{1, \ldots, s\}$. Let 1_S be the vector that is 1 on S and 0 on S^c. We attempt to recover the approximately s-sparse vector $v = 1_S + \epsilon u$, where $\epsilon = 0.01$ and u has i.i.d. Gaussian entries and is normalized such that $\|u\|_1 = 1$. Let $i^* = \mathrm{argmax}_i\,|v(i)|$. We solve (1)

for $1 \leq i \leq n$ using YALMIP [12] with the SDPT3 solver [15, 16]. Among these n outputs, we let $z^\#$ be given by the one that (a) corresponds to $i = i^*$; (b) is sparsest in the sense of ℓ_1/ℓ_∞; (c) is sparsest in the sense of ℓ_1/ℓ_2; or (d) is sparsest in the sense of thresholded-ℓ_0 at the level ϵ. We call a recovery successful if $\left\| z^\# - v/v(i^*) \right\|_2 \leq \epsilon$. Figure 1 shows the probability of successful recovery, as computed over 10 independent trials, for many values of k and the approximate sparsity s. Near and below the phase transitions, simulations were performed for all even values of k and s. In the large regions to the top-right of the phase transitions, simulations were performed only for values of k and s that are multiples of 5. In this region, the probably of recovery was always zero.

We also compute

$$\min \|z\|_1 \text{ such that } z \in \text{span}\{\tilde{v}_1, \ldots, \tilde{v}_k\}, z(i) = 1, \tag{53}$$

for i.i.d. $\tilde{v}_j \sim \mathcal{N}(0, I_n)$ and for all $1 \leq i \leq n$. The curve in Figure 1a shows the dependence on k of the minimal value of (53) for a single fixed i, as found by the median over 50 independent trials. The curve in Figure 1b shows the solution to (53) that is sparsest in the ℓ_1/ℓ_∞ sense, also found by the median of 50 trials. We note that that the least ℓ_1/ℓ_∞ solution to (53) is also the minimal value of the problem

$$\min \frac{\|z\|_1}{\|z\|_\infty} \text{ such that } z \in \text{span}\{\tilde{v}_1, \ldots, \tilde{v}_k\}, z \neq 0. \tag{54}$$

These curves represent upper bounds for the sparsity of a recoverable signal. For sparsities of v above this curve, there will be linear combinations of random vectors that are considered sparser than v in the provided sense. Hence, recovering v would be impossible. Figures 1c and 1d do not have corresponding lines because we do not have a tractable method for directly optimizing sparsity in the ℓ_1/ℓ_2 or thresholded-ℓ_0 senses.

Our primary observation is that the selection process for the sparsest output among the n programs (1) can greatly effect recovery performance. Comparing Figure 1a and 1b, we see that v can be recovered by the $i = i^*$ program yet discarded as not sparse enough in the ℓ_1/ℓ_∞ sense. This effect is most significant for small k. In the parameter regime we simulated, the ℓ_1/ℓ_∞ selector exhibits the worst overall recovery performance. The thresholded-ℓ_0 selector gives the best performance, though we remark it involves the choice of a threshold parameter.

Our second observation is that recovery can succeed even when the $i = i^*$ instance of (1) fails. That is, solving all n programs of form (1) can outperform the result of a single program, even when an oracle tells us the index of the largest coefficient. For example, compare Figures 1a to 1c and 1d. This effect is more likely when v has many large components and k is small. To see why, note that successful recovery is expected when $\|\tilde{V}_{i,:}\|_\infty$ is small. If k is small, a large deviation of $\|\tilde{V}_{i,:}\|_\infty$ is fairly likely. If there are many indices i where v is large, it is likely that $\|\tilde{V}_{i,:}\|_\infty$ will be small enough for successful recovery with some other value of i. On the other hand, if k is big, a large deviation of $\|\tilde{V}_{i,:}\|_\infty$ is extremely unlikely. Hence, considering many programs would have little benefit.

Our third observation is that in the case where an oracle tells us i^*, and in the case where we minimize sparsity in an ℓ_1/ℓ_∞ sense, the recoverable sparsity is quite close to the upper bound provided by a considering random space with no planted sparse vector. With the ℓ_1/ℓ_2 and thresholded-ℓ_0 selectors, we expect a discrepancy between the observed recovery and the sparsest element of W in the respective sense.

Potentially, the recovery region could be improved significantly beyond that indicated by Figure 1. For example, one could consider even more than n linear programs, each with a normalization

against a different random direction in $\mathbb{R}^n$. Such an approach immediately gives rise to a natural tradeoff: recovery performance may be improved at the expense of more linear programs that need to be solved. We leave this relationship for future study.

References

[1] S. Artstein-Avidan, V. Milman. Logarithmic reduction of the level of randomness in some probabilistic geometric constructions. *J. Functional Analysis* 235, 297-329, 2006.

[2] F. Bach, J. Mairal, J. Ponce. Convex Sparse Matrix Factorizations *arXiv preprint 0812.1869*, 2008.

[3] G. Beylkin, L. Monzón. On approximation of functions by exponential sums. *Applied and Computational Harmonic Analysis*, 19(1), 17-48, 2005.

[4] E. J. Candès, Y. C. Eldar, T. Strohmer, V. Voroninski. Phase Retrieval via Matrix Completion. *SIAM J. on Imaging Sciences* 6(1), 199–225, 2011.

[5] E. J. Candès, J. Romberg, T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59 1207-1223, 2005.

[6] T.F. Coleman, A. Pothen. The null space problem I. Complexity. *SIAM J. Algebraic and Discrete Methods*, 7(4):527-537, 1986.

[7] T. Figiel, J. Lindenstrauss, V. Milman. The dimension of almost spherical sections of convex bodies. *Acta Math.*, 139(1-2):53-94, 1977.

[8] L-A. Gottlieb, T. Neylon. Matrix sparsification and the sparse null space problem. *APPROX and RANDOM*, 6302:205-218, 2010.

[9] V. Guruswami, J. Lee, A. Razborov. Almost Euclidean subspaces of l_1^n via expander codes. *Combinatorica* 30(1): 47-68, 2010.

[10] B. Kashin. The widths of certain finite-dimensional sets and classes of smooth functions. *Izv. Akad. Nauk SSSR Ser. Mat.*, 41(2):334-351, 478, 1977.

[11] J. Lee. Kernels of Random Sign Matrices. Tcs math blog post. Available: `tcsmath.wordpress.com/2008/05/08/kernels-of-random-sign-matrices/`. 2008.

[12] J. Löfberg. YALMIP : A Toolbox for Modeling and Optimization in MATLAB. *Proceedings of the CACSD Conference*, Taipei, Taiwan, 2004.

[13] S. Lovett, S. Sodin. Almost Euclidean sections of the N-dimensional cross-polytope using O(N) random bits. *Electronic Colloquium on Computational Complexity, Report* 7-12. 2007.

[14] D. Spielman, H. Wang, J. Wright. Exact Recovery of Sparsely-Used Dictionaries. *J. Machine Learning Research - Proceedings Track 23* 37.1-37.18, 2012

[15] K.C. Toh, M.J. Todd, R.H. Tutuncu. SDPT3 — a Matlab software package for semidefinite programming, *Optimization Methods and Software*, 11, 545-581, 1999

[16] R.H Tutuncu, K.C. Toh, M.J. Todd. Solving semidefinite-quadratic-linear programs using SDPT3, *Mathematical Programming Ser. B*, 95, 189-217, 2003.

[17] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. C. Eldar and G. Kutyniok, editors, *Compressed Sensing: Theory and Applications*. Cambridge University Press, 2010.

[18] M. Zibulevsky and B. Pearlmutter. Blind source separation by sparse decomposition. *Neural Computation*, 13(4), 2001.

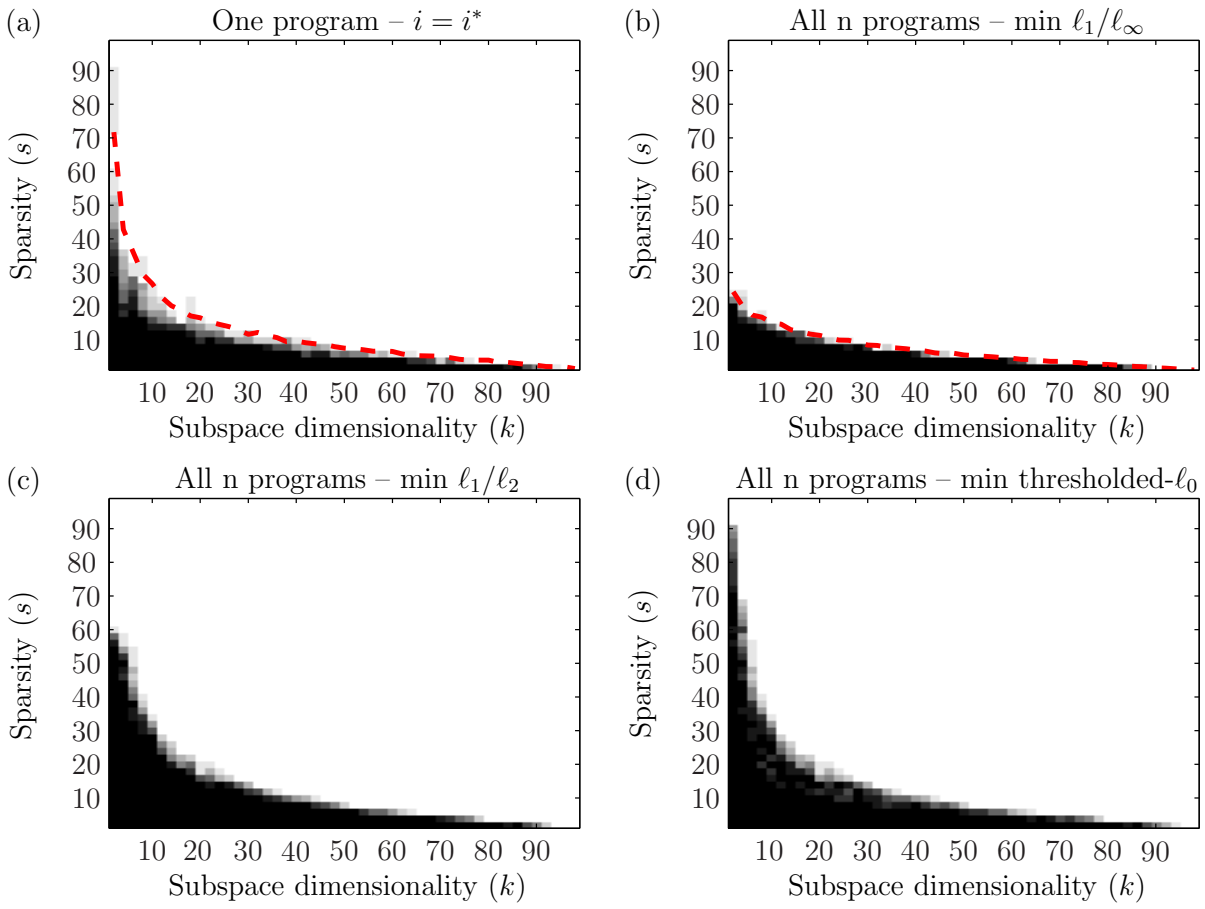

Figure 1: Empirical probability of recovery versus approximate sparsity and subspace dimensionality k. For a given s, the vector to be recovered is a noisy version of 1_S, where $|S| = s$. For all values of k and s, we solve n programs of form (1), one for each $1 \leq i \leq n$. In panel (a), the output corresponding to $i = i^*$ was selected. In panels (b), (c), (d), the output was selected to minimize sparsity in the ℓ_1/ℓ_∞, ℓ_1/ℓ_2, and thresholded-ℓ_0 senses, respectively. The diagram shows the probability of recovery, as measured by 10 independent trials. White represents a recovery with probability zero. Black represents recovery with probability 1. The dashed lines in panels (a) and (b) represent upper bounds for the maximal recoverable sparsity based on the worst-case linear combination of only the random vectors.