# Lower Bounds for Compressed Sensing with Generative Models

**Akshay Kamath**
The University of Texas at Austin
kamath@cs.utexas.edu

**Sushrut Karmalkar**
The University of Texas at Austin
sushrutk@cs.utexas.edu

**Eric Price**
The University of Texas at Austin
ecprice@cs.utexas.edu

## Abstract

The goal of compressed sensing is to learn a structured signal $x$ from a limited number of noisy linear measurements $y \approx Ax$. In traditional compressed sensing, "structure" is represented by sparsity in some known basis. Inspired by the success of deep learning in modeling images, recent work starting with [BJPD17] has instead considered structure to come from a generative model $G : \mathbb{R}^k \to \mathbb{R}^n$. We present two results establishing the difficulty of this latter task, showing that existing bounds are tight.

First, we provide a lower bound matching the [BJPD17] upper bound for compressed sensing from $L$-Lipschitz generative models $G$. In particular, there exists such a function that requires roughly $\Omega(k \log L)$ linear measurements for sparse recovery to be possible. This holds even for the more relaxed goal of *nonuniform* recovery.

Second, we show that generative models generalize sparsity as a representation of structure. In particular, we construct a ReLU-based neural network $G : \mathbb{R}^{2k} \to \mathbb{R}^n$ with $O(1)$ layers and $O(kn)$ activations per layer, such that the range of $G$ contains all $k$-sparse vectors.

## 1 Introduction

In compressed sensing, one would like to learn a structured signal $x \in \mathbb{R}^n$ from a limited number of linear measurements $y \approx Ax$. This is motivated by two observations: first, there are many situations where linear measurements are easy, in settings as varied as streaming algorithms, single-pixel cameras, genetic testing, and MRIs. Second, the unknown signals $x$ being observed are structured or "compressible": although $x$ lies in $\mathbb{R}^n$, it would take far fewer than $n$ words to describe $x$. In such a situation, one can hope to estimate $x$ well from a number of linear measurements that is closer to the size of the *compressed representation* of $x$ than to its ambient dimension $n$.

In order to do compressed sensing, you need a formal notion of how signals are expected to be structured. The classic answer is to use *sparsity*. Given linear measurements[1] $y = Ax$ of an arbitrary vector $x \in \mathbb{R}^n$, one can hope to recover an estimate $x^*$ of $x$ satisfying

$$\|x - x^*\| \leq C \min_{k\text{-sparse } x'} \|x - x'\| \tag{1}$$

for some constant $C$ and norm $\|\cdot\|$. In this paper, we will focus on the $\ell_2$ norm and achieving the guarantee with $3/4$ probability. Thus, if $x$ is well-approximated by a $k$-sparse vector $x'$, it should be accurately recovered. Classic results such as [CRT06] show that (1) is achievable when $A$ consists

[1] The algorithms we discuss can also handle post-measurement noise, where $y = Ax + \eta$. We remove this term for simplicity: this paper focuses on lower bounds, and handling this term could only make things harder.

of $m = O(k \log \frac{n}{k})$ independent Gaussian linear measurements. This bound is tight, and in fact no distribution of matrices with fewer rows can achieve this guarantee in either $\ell_1$ or $\ell_2$ [DIPW10].

Although compressed sensing has had success, sparsity is a limited notion of structure. Can we learn a richer model of signal structure from data, and use this to perform recovery? In recent years, deep convolutional neural networks have had great success in producing rich models for representing the manifold of images, notably with generative adversarial networks (GANs) [GPAM+14] and variational autoencoders (VAEs) [KW14]. These methods produce generative models $G : \mathbb{R}^k \to \mathbb{R}^n$ that allow approximate sampling from the distribution of images. So a natural question is whether these generative models can be used for compressed sensing.

In [BJPD17] it was shown how to use generative models to achieve a guarantee analogous to (1): for any $L$-Lipschitz $G : \mathbb{R}^k \to \mathbb{R}^n$, one can achieve

$$\|x - x^*\|_2 \leq C \min_{z' \in B_k(r)} \|x - G(z')\|_2 + \delta, \tag{2}$$

where $r, \delta > 0$ are parameters, $B_k(r)$ denotes the radius-$r$ $\ell_2$ ball in $\mathbb{R}^k$ and Lipschitzness is defined with respect to the $\ell_2$-norms, using only $m = O(k \log \frac{Lr}{\delta})$ measurements. Thus, the recovered vector is almost as good as the nearest point in the *range of the generative model*, rather than in the set of $k$-sparse vectors. We will refer to the problem of achieving the guarantee in (2) as "function-sparse recovery".

Our main theorem is that the [BJPD17] result is tight: for any setting of parameters $n, k, L, r, \delta$, there exists an $L$-Lipschitz function $G : \mathbb{R}^k \to \mathbb{R}^n$ such that any algorithm achieving (2) with $3/4$ probability must have $\Omega(\min(k \log \frac{Lr}{\delta}, n))$ linear measurements. Notably, the additive error $\delta$ that was unnecessary in sparse recovery is necessary for general Lipschitz generative model recovery.

A concurrent paper [LS19] proves a lower bound for a restricted version of (2). They show a lower bound when the vector that $x$ lies in the image of $G$ and for a particular value of $\delta$. Our results, in comparison, apply to the most general version of the problem and are proven using a simpler communication complexity technique.

The second result in this paper is to directly relate the two notions of structure: sparsity and generative models. We produce a simple Lipschitz neural network $G_{sp} : \mathbb{R}^{2k} \to \mathbb{R}^n$, with ReLU activations, 2 hidden layers, and maximum width $O(kn)$, so that the range of $G$ contains all $k$-sparse vectors.

A second result of [BJPD17] is that for ReLU-based neural networks, one can avoid the additive $\delta$ term and achieve a different result from (2):

$$\|x - x^*\|_2 \leq C \min_{z' \in \mathbb{R}^k} \|x - G(z')\|_2 \tag{3}$$

using $O(kd \log W)$ measurements, if $d$ is the depth and $W$ is the maximum number of activations per layer. Applying this result to our sparsity-producing network $G_{sp}$ implies, with $O(k \log n)$ measurements, recovery achieving the standard sparsity guarantee (1). So the generative-model representation of structure really is more powerful than sparsity.

## 2 Proof overview

As described above, this paper contains two results: an $\Omega(\min(k \log \frac{Lr}{\delta}, n))$ lower bound for compressed sensing relative to a Lipschitz generative model, and an $O(1)$-layer generative model whose range contains all sparse vectors. These results are orthogonal, and we outline each in turn.

### 2.1 Lower bound for Lipschitz generative recovery.

Over the last decade, lower bounds for sparse recovery have been studied extensively. The techniques in this paper are most closely related to the techniques used in [DIPW10].

Similar to [DIPW10], our proof is based on communication complexity. We will exhibit an $L$-Lipschitz function $G$ and a large finite set $Z \subset Im(G) \subset B_n(R)$ of points that are well-separated. Then, given a point $x$ that is picked uniformly at random from $Z$, we show how to identify it from $Ax$ using the function-sparse recovery algorithm. This implies $Ax$ also contains a lot of information, so $m$ must be fairly large.

Formally, we produce a generative model whose range includes a large, well-separated set:

**Theorem 2.1.** *Given $R > 0$ satisfying $R > 2Lr$, there exists an $O(L)-$Lipschitz function $G : \mathbb{R}^k \to \mathbb{R}^n$, and $X \subseteq B_k(r)$ such that*

*(1) for all $x \in X$, $G(x) \in \{\pm\frac{R}{\sqrt{n}}\}^n$*

*(2) for all $x \in X$, $\|G(x)\|_2 \leq R$*

*(3) for all $x, y \in X$, $\|G(x) - G(y)\|_2 \geq \frac{R}{\sqrt{6}}$*

*(4) $\log(|X|) = \Omega\left(\min(k\log(\frac{Lr}{R})), n\right)$*

Now, suppose we have an algorithm that can perform function-sparse recovery with respect to $G$ from Theorem 2.1, with approximation factor $C$, and error $\delta < R/8$ within the radius $r$ ball in $k$-dimensions. Set $t = \Theta(\log n)$, and for any $z_1, z_2, \ldots, z_t \in Z = G(X)$ take

$$z = \epsilon^t z_1 + \epsilon^{t-1} z_2 + \epsilon^{t-1} z_3 + \ldots + z_t$$

for $\epsilon = \frac{1}{8(C+1)}$ a small constant. The idea of the proof is the following: given $y = Az$, we can recover $\widehat{z}$ such that

$$\|\widehat{z} - z_1\|_2 \leq \|z - z_1\|_2 + \|\widehat{z} - z\|_2 + \delta \leq (C+1)\|z - z_1\|_2 + \delta < R/8 + R/8 = R/4$$

and so, because $Z$ has minimum distance $R/\sqrt{6}$, we can exactly recover $z_t$ by rounding $\widehat{z}$ to the nearest element of $Z$. But then we can repeat the process on $(Az - Az_t)$ to find $z_{t-1}$, then $z_{t-2}$, up to $z_1$, and learn $t\lg|Z| = \Omega(tk\log(Lr/R))$ bits total. Thus $Az$ must contain this many bits of information; but if the entries of $A$ are rational numbers with poly$(n)$ bounded numerators and (the same) poly$(n)$ bounded denominator, then each entry of $Az$ can be described in $O(t + \log n)$ bits, so

$$m \cdot O(t + \log n) \geq \Omega(tk\log(Lr/R))$$

or $m \geq \Omega(k\log(Lr/R))$.

There are two issues that make the above outline not totally satisfactory, which we only briefly address how to resolve here. First, the theorem statement makes no supposition on the entries of $A$ being polynomially bounded. To resolve this, we perturb $z$ with a tiny (polynomially small) amount of additive Gaussian noise, after which discretizing $Az$ at an even tinier (but still polynomial) precision has negligible effect on the failure probability. The second issue is that the above outline requires the algorithm to recover all $t$ vectors, so it only applies if the algorithm succeeds with $1 - 1/t$ probability rather than constant probability. This is resolved by using a reduction from the *augmented indexing* problem, which is a one-way communication problem where Alice has $z_1, z_2, \ldots, z_t \in Z$, Bob has $i \in [Z]$ and $z_{i+1}, \cdots, z_n$, and Alice must send Bob a message so that Bob can output $z_i$ with $2/3$ probability. This still requires $\Omega(t\log|Z|)$ bits of communication, and can be solved in $O(m(t + \log n))$ bits of communication by sending $Az$ as above. Formally, our lower bound states:

**Theorem 2.2.** *Consider any $L, r, \delta$ with $\delta \leq Lr/4$. There exists an L-Lipschitz function $G^* : \mathbb{R}^k \to \mathbb{R}^n$ such that, if $\mathcal{A}$ is an algorithm which picks a matrix $A \in \mathbb{R}^{m\times n}$, and given $Ax$ returns an $x^*$ satisfying* (2) *with probability $\geq 3/4$, then $m = \Omega(\min(k\log(Lr/\delta), n))$.*

**Constructing the set.** The above lower bound approach, relies on finding a large, well-separated set $Z$ as in Theorem 2.1.

We construct this aforementioned set $Z$ within the $n$-dimensional $\ell_2$ ball of radius $R$ such that any two points in the set are at least $\Omega(R)$ apart. Furthermore, since we wish to use a function-sparse recovery algorithm, we describe a function $G : \mathbb{R}^k \to \mathbb{R}^n$ and set the radius $R$ such that $G$ is $L$-Lipschitz. In order to get the desired lower bound, the image of $G$ needs to contain a subset of at least $(Lr)^{\Omega(k)}$ points.

First, we construct a mapping as described above from $\mathbb{R}$ to $\mathbb{R}^{n/k}$ i.e we need to find $(Lr)^{\Omega(k)}$ points in $B_{n/k}(R)$ that are mutually far apart. We show that certain binary linear codes over the alphabet $\{\pm R/\sqrt{n}\}$ yield such points that are mutually $R/\sqrt{3k}$ apart. We construct a $O(L)$-Lipschitz mapping of $O(\sqrt{Lr})$ points in the interval $[0, r/\sqrt{k}]$ to a subset of these points.

In order to extend this construction to a mapping from $\mathbb{R}^k$ to $\mathbb{R}^n$, we apply the above function in a coordinate-wise manner. This would result in a mapping with the same Lipschitz parameter. The points in $\mathbb{R}^n$ that are images of these points lie in a ball of radius $R$ but could potentially be $R/\sqrt{3k}$ close. To get around this, we use an error correcting code over a large alphabet to choose a subset of these points that is large enough and such that they are still mutually $R/\sqrt{6}$ far apart.

### 2.2 Sparsity-producing generative model.

To produce a generative model whose range consists of all $k$-sparse vectors, we start by mapping $\mathbb{R}^2$ to the set of positive 1-sparse vectors. For any pair of angles $\theta_1, \theta_2$, we can use a constant number of unbiased ReLUs to produce a neuron that is only active at points whose representation $(r, \theta)$ in polar coordinates has $\theta \in (\theta_1, \theta_2)$. Moreover, because unbiased ReLUs behave linearly, the activation can be made an arbitrary positive real by scaling $r$ appropriately. By applying this $n$ times in parallel, we can produce $n$ neurons with disjoint activation ranges, making a network $\mathbb{R}^2 \to \mathbb{R}^n$ whose range contains all 1-sparse vectors with nonnegative coordinates.

By doing this $k$ times and adding up the results, we produce a network $\mathbb{R}^{2k} \to \mathbb{R}^n$ whose range contains all $k$-sparse vectors with nonnegative coordinates. To support negative coordinates, we just extend the $k = 1$ solution to have two ranges within which it is non-zero: for one range of $\theta$ the output is positive, and for another the output is negative.

This results in the following theorem:

**Theorem 2.3.** *There exists a 2 layer neural network* $G_{sp} : \mathbb{R}^{2k} \to \mathbb{R}^n$ *with width* $O(nk)$ *such that* $\{x \mid \|x\|_0 = k\} \subseteq Im(G)$

## 3 Lower bound proof

In this section, we prove a lower bound for the sample complexity of function-sparse recovery by a reduction from a communication game. We show that the communication game can be won by sending a vector $Ax$ and then performing function-sparse recovery. A lower bound on the communication complexity of the game implies a lower bound on the number of bits used to represent $Ax$ if $Ax$ is discretized. We can then use this to lower bound the number of measurements in $A$.

Since we are dealing in bits in the communication game and the entries of a sparse recovery matrix can be arbitrary reals, we will need to discretize each measurement. We show first that discretizing the measurement matrix by rounding does not change the resulting measurement too much and will allow for our reduction to proceed.

**Notation.** We use $B_k(r) = \{x \in \mathbb{R}^k \mid \|x\|_2 \leq r\}$ to denote the $k$-dimensional ball of radius $r$. Given a function $g : \mathbb{R}^a \to \mathbb{R}^b$, $g^{\otimes k} : \mathbb{R}^{ak} \to \mathbb{R}^{bk}$ denotes a function that the maps a point $(x_1, \ldots, x_{ak})$ to $(g(x_1, \ldots, x_a), g(x_{a+1}, \ldots, x_{2a}), \ldots, g(x_{a(k-1)+1}, \ldots, x_{ak}))$. For any function $G : A \to B$, we use $Im(G)$ to denote $\{G(x) \mid x \in A\}$.

**Matrix conditioning.** We first show that, without loss of generality, we may assume that the measurement matrix $A$ is well-conditioned. In particular, we may assume that the rows of $A$ are orthonormal.

We can multiply $A$ on the left by any invertible matrix to get another measurement matrix with the same recovery characteristics. If we consider the singular value decomposition $A = U\Sigma V^*$, where $U$ and $V$ are orthonormal and $\Sigma$ is 0 off the diagonal, this means that we can eliminate $U$ and make the entries of $\Sigma$ be either 0 or 1. The result is a matrix consisting of $m$ orthonormal rows.

**Discretization.** For well-conditioned matrices $A$, we use the following lemma (similar to one from [DIPW10]) to show that we can discretize the entries without changing the behavior by much:

**Lemma 3.1.** *Let* $A \in \mathbb{R}^{m \times n}$ *be a matrix with orthonormal rows. Let* $A'$ *be the result of rounding* $A$ *to* $b$ *bits per entry. Then for any* $v \in \mathbb{R}^n$ *there exists an* $s \in \mathbb{R}^n$ *with* $A'v = A(v - s)$ *and* $\|s\|_2 < n2^{-b} \|v\|_2$.

*Proof.* Let $A'' = A - A'$ be the error when discretizing $A$ to $b$ bits, so each entry of $A''$ is less than $2^{-b}$. Then for any $v$ and $s = A^T A'' v$, we have $As = A'' v$ and

$$\begin{aligned} \|s\|_2 &= \left\|A^T A'' v\right\|_2 \leq \|A'' v\|_2 \\ &\leq m2^{-b} \|v\|_2 \leq n2^{-b} \|v\|_2 . \end{aligned}$$

□

**The Augmented Indexing problem.** As in [DIPW10], we use the Augmented Indexing communication game which is defined as follows: There are two parties, Alice and Bob. Alice is given a string $y \in \{0,1\}^d$. Bob is given an index $i \in [d]$, together with $y_{i+1}, y_{i+2}, \ldots, y_d$. The parties also share an arbitrarily long common random string $r$. Alice sends a single message $M(y,r)$ to Bob, who must output $y_i$ with probability at least $2/3$, where the probability is taken over $r$. We refer to this problem as Augmented Indexing. The communication cost of Augmented Indexing is the minimum, over all correct protocols, of length $|M(y,r)|$ on the worst-case choice of $r$ and $y$.

The following theorem is well-known and follows from Lemma 13 of [MNSW98] (see, for example, an explicit proof in [DIPW10])

**Theorem 3.2.** *The communication cost of* Augmented Indexing *is* $\Omega(d)$.

**A well-separated set of points.** We would like to prove Theorem 2.1, getting a large set of well-separated points in the image of a Lipschitz generative model. Before we do this, though, we prove a $k = 1$ analog:

**Lemma 3.3.** *There is a set of points $P$ in $B_n(1) \subset \mathbb{R}^n$ of size $2^{\Omega(n)}$ such that for each pair of points $x, y \in P$*

$$\|x - y\| \in \left[\sqrt{\frac{1}{3}}, \sqrt{\frac{2}{3}}\right]$$

*Proof.* Consider a $\tau$-balanced linear code over the alphabet $\{\pm\frac{1}{\sqrt{n}}\}$ with message length $M$. It is known that such codes exist with block length $O(M/\tau^2)$ [BATS09]. Setting the block length to be $n$ and $\tau = 1/6$, we get that there is a set of $2^{\Omega(n)}$ points in $\mathbb{R}^n$ such that the pairwise hamming distance is between $\left[\frac{n}{3}, \frac{2n}{3}\right]$, i.e. the pairwise $\ell_2$ distance is between $\left[\sqrt{\frac{1}{3}}, \sqrt{\frac{2}{3}}\right]$. □

Now we wish to extend this result to arbitrary $k$ while achieving the parameters in Theorem 2.1.

*Proof of Theorem 2.1.* We first define an $O(L)$-Lipschitz map $g : \mathbb{R} \to \mathbb{R}^{n/k}$ that goes through a set of points that are pairwise $\Theta\left(\frac{R}{\sqrt{k}}\right)$ apart. Consider the set of points $P$ from Lemma 3.3 scaled to $B_{n/k}(\frac{R}{\sqrt{k}})$. Observe that $|P| \geq \exp\left(\Omega\left(n/k\right)\right) \geq \min\left(\exp\left(\Omega\left(n/k\right)\right), Lr/R\right)$. Choose subset $P'$ that such that it contains exactly $\min\left(Lr/R, \exp(\Omega(n/k))\right)$ points and let $g_1 : [0, r/\sqrt{k}] \to P'$ be a piecewise linear function that goes through all the points in $P'$ in order. Then, we define $g : \mathbb{R} \to \mathbb{R}^{n/k}$ as:

$$g(x) = \begin{cases} g_1(0) & \text{if } x < 0 \\ g_1(x) & \text{if } 0 \leq x \leq r/\sqrt{k} \\ g_1(\frac{R}{\sqrt{k}}) & \text{if } x \geq r/\sqrt{k} \end{cases}$$

Let $I = \{\frac{r}{\sqrt{k}|P'|}, \ldots, \frac{r}{\sqrt{k}}\}$ be the points that are pre-images of elements of $P'$. Observe that $g$ is $O(L)$-Lipschitz since within the interval $[0, r/\sqrt{k}]$, since it maps each interval of length $\frac{r}{\sqrt{k}|P'|} \geq \frac{rR}{\sqrt{k}Lr} = \frac{R}{L\sqrt{k}}$ to an interval of length at most $O(R/\sqrt{k})$.

Now, consider the function $G := g^{\otimes k} : \mathbb{R}^k \to \mathbb{R}^n$. Observe that $G$ is also $O(L)$ Lipschitz,

$$\begin{aligned}\|G(x_1, \ldots, x_k) - G(y_1, \ldots, y_k)\|_2^2 &= \sum_{i \in [k]} \|g(x_i) - g(y_i)\|_2^2 \\ &\leq \sum_{i \in [k]} O(L^2) \|x_i - y_i\|_2^2 \\ &= O(L^2) \|x - y\|_2^2\end{aligned}$$

Also, for every point $(x_1, \ldots, x_k) \in I^k$, $\|G(x_1, \ldots, x_k)\|_2 = \sqrt{\sum_{i \in [k]} \|g(x_i)\|_2^2} \leq R$. However, there still exist distinct points $x, y \in I^k$(for instance points that differ at exactly one coordinate) such that $\|G(x) - G(y)\|_2 \leq O(\frac{R}{\sqrt{k}})$.

We construct a large subset of the points in $I^k$ such that any two points in this subset are far apart using error correcting codes. Consider the $A \subset P'$ s.t. $|A| > |P'|/2$ is a prime. For any integer $z > 0$, there is a prime between $z$ and $2z$, so such a set $A$ exists. Consider a Reed-Solomon code of block length $k$, message length $k/2$, distance $k/2$ and alphabet $A$. The existence of such a code implies that there is a subset $X'$ of $(P')^k$ of size at least $(|P'|/2)^{k/2}$ such that every pair of distinct elements from this set disagree in $k/2$ coordinates.

This translates into a distance of $\frac{R}{\sqrt{6}}$ in 2-norm. So, if we set $G = g^{\otimes k}$ and $X \subset I^k$ to $G^{-1}(X')$, we get have a set of $(|P'|/2)^{k/2} \geq (\min(\exp(\Omega(n/k)), Lr/R))^{k/2}$ points which are $\frac{R}{\sqrt{6}}$ apart in 2-norm, lie within the $\ell_2$ ball of radius $R$. □

**Lower bound.** We now prove the lower bound for function-sparse recovery.

*Proof of Theorem 2.2.* An application of Theorem 2.1 with $R = \sqrt{Lr\delta}$ gives us a set of points $Z$ and $G$ such that $Z = G(X) \subseteq \mathbb{R}^n$ such that $\log(|Z|) = \Omega(\min(k \log(\frac{Lr}{\delta}), n))$, and for all $x \in Z$, $\|x\|_2 \leq \sqrt{Lr\delta}$ and for all $x, x' \in Z$, $\|x - x'\|_2 \geq \sqrt{Lr\delta}/\sqrt{6}$. Let $d = \lfloor \log |X| \rfloor \log n$, and let $D = 16\sqrt{3}(C+1)$.

We will show how to solve the Augmented Indexing problem on instances of size $d = \log(|Z|) \cdot \log(n) = \Omega(k \log(Lr) \log n)$ with communication cost $O(m \log n)$. The theorem will then follow by Theorem 3.2.

Alice is given a string $y \in \{0,1\}^d$, and Bob is given $i \in [d]$ together with $y_{i+1}, y_{i+2}, \ldots, y_d$, as in the setup for Augmented Indexing.

Alice splits her string $y$ into $\log n$ contiguous chunks $y^1, y^2, \ldots, y^{\log n}$, each containing $\lfloor \log |X| \rfloor$ bits. She uses $y^j$ as an index into the set $X$ to choose $x_j$. Alice defines

$$x = D^1 x_1 + D^2 x_2 + \cdots + D^{\log n} x_{\log n}.$$

Alice and Bob use the common randomness $\mathcal{R}$ to agree upon a random matrix $A$ with orthonormal rows. Both Alice and Bob round $A$ to form $A'$ with $b = \Theta(\log(n))$ bits per entry. Alice computes $A'x$ and transmits it to Bob. Note that, since $x \in \left\{\pm \frac{1}{\sqrt{n}}\right\}$ the $x$'s need not be discretized.

From Bob's input $i$, he can compute the value $j = j(i)$ for which the bit $y_i$ occurs in $y^j$. Bob's input also contains $y_{i+1}, \ldots, y_n$, from which he can reconstruct $x_{j+1}, \ldots, x_{\log n}$, and in particular can compute

$$z = D^{j+1} x_{j+1} + D^{j+2} x_{j+2} + \cdots + D^{\log n} x_{\log n}.$$

Set $w = \frac{1}{D^j}(x - z) = \frac{1}{D^j} \sum_{i=1}^{j} D^i x_i$. Bob then computes $A'z$, and using $A'x$ and linearity, he can compute $\frac{1}{D^j} \cdot A'(x - z) = A'w$. Then

$$\|w\|_2 \leq \frac{1}{D^j} \sum_{i=1}^{j} R \cdot D^i < R.$$

So from Lemma 3.1, there exists some $s$ with $A'w = A(w - s)$ and

$$\|s\|_2 < n^2 2^{-b} \|w\|_2 < \frac{R}{D^j n^2}.$$

Ideally, Bob would perform recovery on the vector $A(w-s)$ and show that the correct point $x_j$ is recovered. However, since $s$ is correlated with $A$ and $w$, Bob needs to use a slightly more complicated technique.

Bob first chooses another vector $u$ uniformly from $B_n(R/D^j)$ and computes $A(w-s-u) = A'w - Au$. He then runs the estimation algorithm $\mathcal{A}$ on $A$ and $A(w-s-u)$, obtaining $\hat{w}$. We have that $u$ is independent of $w$ and $s$, and that $\|u\|_2 \leq \frac{R}{D^j}(1-1/n^2) \leq \frac{R}{D^j} - \|s\|_2$ with probability $\frac{\text{Vol}(B_n(\frac{R}{D^j}(1-1/n^2)))}{\text{Vol}(B_n(\frac{R}{D^j}))} = (1-1/n^2)^n > 1-1/n$. But $\{w-u \mid \|u\|_2 \leq \frac{R}{D^j} - \|s\|_2\} \subseteq \{w-s-u \mid \|u\|_2 \leq \frac{R}{D^j}\}$, so as a distribution over $u$, the ranges of the random variables $w-s-u$ and $w-u$ overlap in at least a $1-1/n$ fraction of their volumes. Therefore $w-s-u$ and $w-u$ have statistical distance at most $1/n$. The distribution of $w-u$ is independent of $A$, so running the recovery algorithm on $A(w-u)$ would work with probability at least $3/4$. Hence with probability at least $3/4 - 1/n \geq 2/3$ (for $n$ large enough), $\hat{w}$ satisfies the recovery criterion for $w-u$, meaning

$$\|w-u-\hat{w}\|_2 \leq C \min_{w' \in Im(G)} \|w-u-w'\|_2 + \delta$$

Now,

$$\begin{aligned}
\|x_j - \hat{w}\|_2 &\leq \|w-u-x_j\|_2 + \|w-u-\hat{w}\|_2 \\
&\leq (1+C)\|w-u-x_j\|_2 + \delta \\
&\leq (1+C)\left(\|u\|_2 + \frac{1}{D^j} \cdot \sum_{i=1}^{j-1} \left\|D^i x_i\right\|_2\right) + \delta \\
&\leq 2(1+C)R/D + \delta \\
&< R \cdot \frac{2(1+C)}{D} + \delta \\
&= \frac{1}{8\sqrt{3}} \cdot R + \delta.
\end{aligned}$$

Since $\delta < Lr/4$, this distance is strictly bounded by $R/2\sqrt{6}$. Since the minimum distance in $X$ is $R/\sqrt{6}$, this means $\left\|D^j x_j - \hat{w}\right\|_2 < \left\|D^j x' - \hat{w}\right\|_2$ for all $x' \in X, x' \neq x_j$. So Bob can correctly identify $x_j$ with probability at least $2/3$. From $x_j$ he can recover $y^j$, and hence the bit $y_i$ that occurs in $y^j$.

Hence, Bob solves Augmented Indexing with probability at least $2/3$ given the message $A'x$. Each entry of $A'x$ takes $O(\log n)$ bits to describe because $A'$ is discretized to up to $\log(n)$ bits and $x \in \{\pm\frac{1}{\sqrt{n}}\}^n$. Hence, the communication cost of this protocol is $O(m \cdot \log n)$. By Theorem 3.2, $m \log n = \Omega(\min(k \log(Lr/\delta), n) \cdot \log n)$, or $m = \Omega(\min(k \log(Lr/\delta), n))$. □

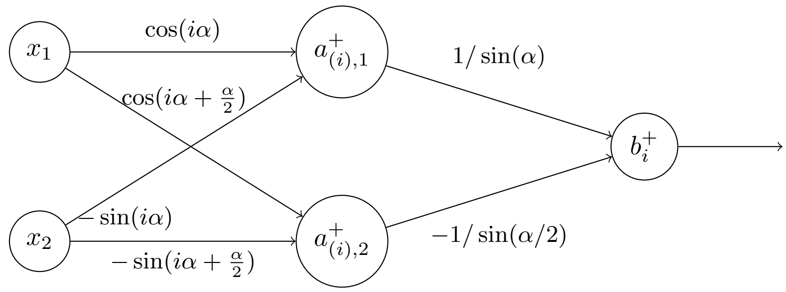
# 4 Reduction from $k$-sparse recovery

We show that the set of all $k$-sparse vectors in $\mathbb{R}^n$ is contained in the image of a 2 layer neural network. This shows that function-sparse recovery is a generalization of sparse recovery.

**Lemma 4.1.** *There exists a 2 layer neural network $G : \mathbb{R}^2 \to \mathbb{R}^n$ with width $O(n)$ such that $\{x \mid \|x\|_0 = 1\} \subseteq Im(G)$*

Our construction is intuitively very simple. We define two gadgets $G_i^+$ and $G_i^-$. $G_i^+ \geq 0$ and $G_i^+(x_1, x_2) \neq 0$ iff $\arctan(x_2/x_1) \in [i \cdot \frac{2\pi}{n}, (i+1) \cdot \frac{2\pi}{n}]$. Similarly $G_i^-(x_1, x_2) \leq 0$ and $G_i^-(x_1, x_2) \neq 0$ iff $\arctan(x_2/x_1) \in [\pi + i \cdot \frac{2\pi}{n}, \pi + (i+1) \cdot \frac{2\pi}{n}]$. Then, we set the $i^{\text{th}}$ output node $(G(x_1, x_2))_i = G_i^+(x_1, x_2) + G_i^-(x_1, x_2)$. Varying the distance of $(x_1, x_2)$ from the origin will allow us to get the desired value at the output node $i$.

*Proof.* Let $\alpha = \frac{\pi}{n+1}$. Let $[x]_+ = x \cdot \mathbb{I}(x \geq 0)$ denote the unbiased ReLU function that preserves positive values and $[x]_- = x \cdot \mathbb{I}(x \leq 0)$ denote the unbiased ReLU function that preserves negative values. We define $G_i^+ : \mathbb{R}^2 \to \mathbb{R}$ as follows:




$G_i^+$ is a 2 layer neural network gadget that produces positive values at output node $i$ of $G$. We define each of the hidden nodes of the neural network $G_i^+$ as follows:

$$\begin{aligned}
a^+_{(i),1} &= \Big[\cos(i\alpha)x_1 - \sin(i\alpha)x_2\Big]_+ \\
a^+_{(i),2} &= \Big[\cos\big(i\alpha + \frac{\alpha}{2}\big)x_1 - \sin\big(i\alpha + \frac{\alpha}{2}\big)x_2\Big]_+ \\
b^+_{(i)} &= \Big[\frac{a^+_{(i),1}}{\sin(\alpha)} - \frac{a^+_{(i),2}}{\sin(\alpha/2)}\Big]_+
\end{aligned}$$

In a similar manner, $G_i^-$ which produces negative values at output node $i$ of $G$ with the internal nodes defined as:

$$\begin{aligned}
a^-_{(i),1} &= \Big[\cos(\pi + i\alpha)x_1 - \sin(\pi + i\alpha)x_2\Big]_+ \\
a^-_{(i),2} &= \Big[\cos\big(\pi + i\alpha + \frac{\alpha}{2}\big)x_1 - \sin\big(\pi + i\alpha + \frac{\alpha}{2}\big)x_2\Big]_+ \\
b^-_{(i)} &= \Big[\frac{a^-_{(i),2}}{\sin(\alpha/2)} - \frac{a^-_{(i),1}}{\sin(\alpha)}\Big]_-
\end{aligned}$$

The last ReLU activation preserves only negative values. Since $G_i^+$ and $G_i^-$ are identical up to signs in the second hidden layer, we only analyze $G_i^+$'s.

Consider $i \in [n]$. Let $\beta = i\alpha$ and $(x_1, x_2) = (t\sin(\theta), t\cos(\theta))$. Then using the identity $\sin(A)\cos(B) - \cos(A)\sin(B) = \sin(A - B)$,

$$\begin{aligned}
\cos(\beta)x_1 - \sin(\beta)x_2 &= t\big(\cos(\beta)\sin(\theta) - \sin(\beta)\cos(\theta)\big) \\
&= t\sin(\theta - \beta)
\end{aligned}$$

This is positive only when $\theta \in (\beta, \pi + \beta)$. Similarly, $\cos(\beta + \alpha/2)x_1 - \sin(\beta + \alpha/2)x_2 = t\sin(\theta - (\beta + \alpha/2))$ and is positive only when $\theta \in (\beta + \alpha/2, \pi + \beta + \alpha/2)$. So, $a^+_{(i),1}$ and $a^+_{(i),2}$ are both non-zero when $\theta \in (\beta + \alpha/2, \pi + \beta)$. Using some elementary trigonometry, we may see that:

$$\begin{aligned}
\frac{a_1^{(i)}}{\sin(\alpha)} - \frac{a_2^{(i)}}{\sin(\alpha/2)} &= t\Big(\frac{\sin(\theta - \beta)}{\sin(\alpha)} - \frac{\sin(\theta - (\beta + \frac{\alpha}{2}))}{\sin(\alpha/2)}\Big) \\
&= \frac{t\sin(\beta - \theta + \alpha)}{\sin(\alpha/2)}
\end{aligned}$$

In Fact A.1, we show a proof of the above identity. Observe that when $\theta > \beta + \alpha$, this term is negative and hence $b^i = 0$. So, we may conclude that $G_i^+((x_1, x_2)) \neq 0$ if and only if $(x_1, x_2) = (t\sin(\theta), t\cos(\theta))$ with $\theta \in ((i-1)\alpha, i\alpha)$. Also, observe that $G_i^+(t\sin(\beta + \alpha/2), t\cos(\beta + \alpha/2)) = t$. Similarly $G_i^-$ is non-zero only if and only if $\theta \in [\pi + i\alpha, \pi + (i+1)\alpha]$ and $G_i^-(t\sin(\pi + i\alpha + \alpha/2), t\cos(\pi + i\alpha + \alpha/2)) = -t$. Since $\alpha = \frac{\pi}{n+1}$, the intervals within which each of $G_1^+, \ldots, G_n^+$ ,$G_1^-, \ldots, G_n^-$ are non-zero do not intersect.

So, given a vector $z'$ such that $\|z\|_0 = 1$ with $z_{i'} \neq 0$, if $z_{i'} > 0$, set

$$\begin{aligned}
x_1 &= |z_{i'}|\sin(i'\alpha + \alpha/2) \\
x_2 &= |z_{i'}|\cos(i'\alpha + \alpha/2)
\end{aligned}$$

and if $z_{i'} < 0$, set

$$x_1 = |z_{i'}| \sin(\pi + i'\alpha + \alpha/2)$$
$$x_2 = |z_{i'}| \cos(\pi + i'\alpha + \alpha/2)$$

Observe that:

$$G^+_{i'}((x_1, x_2)) + G^-_{i'}((x_1, x_2)) = z_{i'}$$

and for all $j \neq i'$

$$G^+_j((x_1, x_2)) + G^-_j((x_1, x_2)) = 0$$

So, if $G(x) = (G^+_1(x) + G^-_1(x), \ldots, G^+_n(x) + G^-_n(x))$, $G$ is a 2-layer neural network with width $O(n)$ such that $\{x \mid \|x\|_0 = 1\} \subseteq Im(G)$. □

*Proof of Theorem 2.3.* Given a vector $z$ that is non-zero at $k$ coordinates, let $i_1 < i_2 < \cdots < i_k$ be the indices at which $z$ is non-zero. We may use copies of $G$ from Lemma 4.1 to generate 1-sparse vectors $v_1, \ldots, v_k$ such that $(v_j)_{i_j} = z_{i_j}$. Then, we add these vectors to obtain $z$. It is clear that we only used $k$ copies of $G$ to create $G_{sp}$. So, $G_{sp}$ can be represented by a neural network with 2 layers. □

Theorem 1 provides a reduction which uses only 2 layers. Then, using the algorithm from Theorem 3, we can recover the correct $k$-sparse vector using $O(kd\log(nk))$ measurements. Since $d = 4$ and $\leq n$, this requires only $O(k \log n)$ linear measurements to perform $\ell_2/\ell_2$ $(k, C)$-sparse recovery.

## References


[BATS09] Avraham Ben-Aroya and Amnon Ta-Shma. Constructing small-bias sets from algebraic-geometric codes. In *2009 50th Annual IEEE Symposium on Foundations of Computer Science*, pages 191–197. IEEE, 2009.

[BJPD17] Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G. Dimakis. Compressed sensing using generative models. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 537–546. JMLR.org, 2017.

[CRT06] E. J. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59(8):1208–1223, 2006.

[DIPW10] K. Do Ba, P. Indyk, E. Price, and D. Woodruff. Lower bounds for sparse recovery. *SODA*, 2010.

[GPAM+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[KW14] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[LS19] Zhaoqiang Liu and Jonathan Scarlett. Information-theoretic lower bounds for compressive sensing with generative models, 2019.

[MNSW98] Peter Bro Miltersen, Noam Nisan, Shmuel Safra, and Avi Wigderson. On data structures and asymmetric communication complexity. *J. Comput. Syst. Sci.*, 57(1):37–49, 1998.

## A Trigonometric identity

**Fact A.1.**

$$\frac{\sin(\beta + \frac{\alpha}{2} - \theta)}{\sin(\alpha/2)} - \frac{\sin(\beta - \theta)}{\sin(\alpha)} = \frac{\sin(\beta - \theta + \alpha)}{\sin(\alpha/2)}$$

*Proof.*

$$\begin{aligned}
\frac{\sin(\beta + \frac{\alpha}{2} - \theta)}{\sin(\alpha/2)} - \frac{\sin(\beta - \theta)}{\sin(\alpha)} &= \frac{\sin(\beta + \frac{\alpha}{2} - \theta)\sin(\alpha) - \sin(\beta - \theta)\sin(\alpha/2)}{\sin(\alpha)\sin(\alpha/2)} \\
&= \frac{\frac{1}{2}\big(\cos(\beta - \theta - \frac{\alpha}{2}) - \cos(\beta - \theta + \frac{3\alpha}{2}) - \cos(\beta - \theta - \frac{\alpha}{2}) + \cos(\beta - \theta + \frac{\alpha}{2})\big)}{\sin(\alpha)\sin(\alpha/2)} \\
&= \frac{\cos(\beta - \theta + \frac{\alpha}{2}) - \cos(\beta - \theta + \frac{3\alpha}{2})}{2\sin(\alpha)\sin(\alpha/2)} \\
&= \frac{\sin(\beta - \theta + \alpha)\sin(\alpha)}{\sin(\alpha)\sin(\alpha/2)} \\
&= \frac{\sin(\beta - \theta + \alpha)}{\sin(\alpha/2)}
\end{aligned}$$

where we use the identity that $\sin(A)\sin(B) = \frac{1}{2}[\cos(A - B) - \cos(A + B)]$ $\square$