On the Power of Compressed Sensing with Generative Models

Akshay Kamath [1] **Sushrut Karmalkar** [1] **Eric Price** [1]

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 (Bora et al., 2017) 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 and strength of this latter task, showing that existing bounds are tight: First, we provide a lower bound matching the (Bora et al., 2017) upper bound for compressed sensing with L-Lipschitz generative models G which holds even for the more relaxed goal of *non-uniform* recovery. Second, we show that generative models generalize sparsity as a representation of structure by constructing a ReLU-based neural network with 2 hidden layers and $O(n)$ activations per layer whose range is precisely the set of 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 floating point numbers 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 $\hat{x}$ of x satisfying

$$\|x - \hat{x}\| \le C \min_{k\text{-sparse } x'} \|x - x'\| \tag{1}$$

for some constant C and norm $\|\cdot\|$. In this paper, we will focus on 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 (Candès et al., 2006) show that (1) is achievable when A consists 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 ℓ_1 or ℓ_2 (Do Ba et al., 2010).

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) (Goodfellow et al., 2014) and variational autoencoders (VAEs) (Kingma & Welling, 2014). 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 (Bora et al., 2017) 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

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

where $r, \delta > 0$ are parameters, $B_k^2(r)$ denotes the radius-r ℓ_2 ball in $\mathbb{R}^k$ and Lipschitzness is defined with respect

*Equal contribution [1]Department of Computer Science, The University of Texas, Austin, Texas. Correspondence to: Akshay Kamath <kamath@cs.utexas.edu>, Sushrut Karmalkar <sushrutk@cs.utexas.edu>, Eric Price <ecprice@cs.utexas.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

[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.

to the ℓ_2-norms, using only $m = O(k + k \log \frac{Lr}{\delta})$ 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 (2) as "generative-model recovery".

Our first theorem is that the (Bora et al., 2017) result is tight: for any setting of parameters n, k, L, r, δ, there exists an L-Lipschitz function $G : \mathbb{R}^k \to \mathbb{R}^n$ such that the (Bora et al., 2017) measurement bound is optimal for achieving (2):

Theorem 1.1. *Consider any n, k, L, r, δ. 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 $\hat{x}$ satisfying (2) with probability at least $3/4$, then $m = \Omega(\min(k + k \log(Lr/\delta), n))$.*

The same result holds if the ℓ_2 norms in (2) are replaced with ℓ_1 norms.

That our lower bound caps out at $m = \Theta(n)$ is of course necessary, since the problem is trivial for $m = n$; thus our bound is tight for the whole range of possible parameters. Notably, and in contrast to sparse recovery, the additive error δ is necessary for Lipschitz generative model recovery. One cannot achieve (2) with $\delta = 0$ and $m = o(n)$.

Our second result directly relates the two notions of structure: sparsity and generative models. We produce a simple ReLU-based neural network $G_{sp} : \mathbb{R}^{2k} \to \mathbb{R}^n$ whose range is precisely the set of all k-sparse vectors.

Theorem 1.2. *There exists a 2-hidden-layer ReLU-based neural network $G_{sp} : \mathbb{R}^{2k} \to \mathbb{R}^n$ with width $O(nk)$ such that $Im(G) = \{x \mid \|x\|_0 \leq k\}$.*

This matches a second result of (Bora et al., 2017), which shows that for ReLU-based neural networks, one can avoid the additive δ term and achieve a different result from (2):

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

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.

Connecting the results. Theorem 1.2 directly implies a weaker form of Theorem 1.1. The network G_{sp} produces all k-sparse binary vectors from seeds of radius $r = n\sqrt{k}$ and with $L = 2$. The standard sparse recovery lower bound shows that recovering these vectors for $\delta = \sqrt{k}$ requires $\Omega(k \log(n/k))$ measurements, which is $\Omega(k \log n)$ for $n > k^{1.1}$. Therefore we immediately see an $\Omega(k \log \frac{Lr}{\delta})$ bound for Lipschitz recovery for these parameters. The advantage of Theorem 1.1 over such an approach is that it applies to *all* values of L, r, and δ, rather than these polynomially-bounded ones; and indeed, such an approach would not show that the additive δ is necessary in (2).

In Theorem 2.2, we also show how to improve Theorem 1.2 to have width $O(n)$, at the cost of exponential Lipschitzness.

Concurrent work. A concurrent paper (Liu & Scarlett, 2019) proves a very similar lower bound to our Theorem 1.1. However, the (Liu & Scarlett, 2019) result is weaker in an important way, analogous to the implication from Theorem 1.2: it requires n to equal Lr/δ, so the lower bound is equal to $\Theta(k \log n)$. As a result, it neither applies to superpolynomial L, nor does it imply that any dependence on δ is necessary.

Our result is also stronger than (Liu & Scarlett, 2019) in a couple other ways. Our bound applies to *non-uniform* algorithms where each matrix A only works for $3/4$ of possible inputs x, rather than requiring A to work for all x, and our bound applies to the ℓ_1 as well as the ℓ_2 guarantee. The (Liu & Scarlett, 2019) approach likely can be extended to non-uniform algorithms, but extending their techniques to ℓ_1 seems quite challenging. Even in the standard sparse-recovery setting, our communication-complexity–based techniques extend to the ℓ_1 guarantee, while (to our knowledge) the information-theory techniques used in (Liu & Scarlett, 2019) do not.

2. Proof overview

As described above, this paper contains two results: a tight lower bound for compressed sensing relative to a Lipschitz generative model, and an $O(1)$-layer generative model whose range contains all sparse vectors. The techniques are independent, 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 (Do Ba et al., 2010).

Similar to (Do Ba et al., 2010), our proof is based on communication complexity. We will exhibit an L-Lipschitz function G and a large finite set $Z \subset Im(G) \subset B_n^p(R)$ 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 generative model recovery algorithm. This implies Ax also contains a lot of informa-

tion, 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 $R > 2Lr$, $p \in \{1, 2\}$, there exists an $O(L)-$Lipschitz function $G : \mathbb{R}^k \to \mathbb{R}^n$, and $X \subseteq B_k^p(r)$ such that*

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

(2) hence for all $x \in X$, $\|G(x)\|_p = R$

(3) for all $x, y \in X$, $\|G(x) - G(y)\|_p \geq \frac{R}{6^{1/p}}$

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

Now, suppose we have an algorithm that can perform generative model recovery with respect to G from Theorem 2.1, with approximation factor C, and error $\delta < R/24$ 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 $\epsilon = \frac{1}{48(C+1)}$ a small constant. The idea of the proof is the following: given $y = Az$, we can recover $\widehat{z}$ such that

$$\begin{aligned} \|\widehat{z} - z_t\| &\leq \|z - z_t\| + \|\widehat{z} - z\| + \delta \\ &\leq (C+1)\|z - z_t\| + \delta \\ &\leq (C+1)\frac{\epsilon R}{1-\epsilon} + \delta \\ &< R/24 + R/24 = R/12 \end{aligned}$$

where the first inequality comes from the generative model recovery guarantee for z_t when treating $z - z_t$ as noise. Now, because Z has minimum distance $R/6^{1/p}$, 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 $i \in [t]$ 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.

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

We construct this set in two stages. First, we consider the $k = 1$ case, producing a Lipschitz map from $\mathbb{R}$ to $\mathbb{R}^n$ with Lr/R points of appropriate distance. We do this by linearly interpolating between elements of a high-distance code over $\{\pm R/n^{1/p}\}^n$; because codewords are $\Theta(R)$ apart, an L-Lipschitz function from $[-r, r]$ can reach Lr/R such elements (as long as this is less than the $2^{\Omega(n)}$ total number of codewords).

To extend this construction to a mapping from $\mathbb{R}^k$ to $\mathbb{R}^n$, we take the product distribution of k such functions, each run with $n' = n/k$. This results in a Lipschitz generative model with the desired radius and number of elements; unfortunately, the minimum distance would be too small. We fix this by concatenating the code: we use an error correcting code over $[n/k]^k$ to choose a subset of these points that is still large enough but has the desired distance.

2.2. Sparsity-producing generative model.

For our second result, 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 θ_1, θ_2, we can use a constant number of unbiased ReLUs to produce a neuron that is only active at points whose representation (r, θ) 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 θ the

output is positive, and for another the output is negative. This results in Theorem 1.2.

2.3. Low width sparsity producing generative model

We also show how to improve this width to $O(n)$. This construction is randomized as opposed to the explicit construction in Theorem 1.2.

We pick n random vectors in $v_1, \ldots, v_n \in \mathbb{R}^{k+1}$. We associated every set $S \subset [n]$ such that $|S| = k$ with a point x_S on the radius 1 ball such that $\langle v_i, x_S \rangle = 0 \forall i \in S$ (such a point must exist because these are k linear equations in x_S, which has $k+1$ variables). We then construct a neural network G such that for $r_i := G(x_S)_i > 0 \forall i \in S$ and $G(x_S)_i = 0 \forall i \notin S$. Further, we show that for values $r'_i \in [0, r_i]$ there exists a pre-image x' within a small region around x_S such that $G(x')_i = r'_i \forall i \in S$ and $G(x')_i = 0 \forall i \notin S$. Scaling such x' up can produce any vector with support S.

The crux of the proof lies in showing that the regions which we identify around the points x_S are non-overlapping. We do this using an ϵ-net argument. The result is as follows:

Theorem 2.2. *There exists a 2-hidden-layer neural network $G_{rand} : \mathbb{R}^{k+1} \to \mathbb{R}^n$ with width $O(n)$ such that $Im(G) = \{x \mid \|x\|_0 \leq k\}$.*

3. Proof of lower bound

In this section, we prove a lower bound for the sample complexity of generative model 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 generative model 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 $B_k^p(r) = \{x \in \mathbb{R}^k \mid \|x\|_p \leq r\}$ to denote the k-dimensional ℓ_p 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 Σ is 0 off the diagonal, this means that we can eliminate U and make the entries of Σ 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 (Do Ba et al., 2010)) 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 $\|s\|_p < n^2 2^{-b} \|v\|_p$ for $p \in \{1, 2\}$.*

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 $As = A''v$. For $p = 2$, we have:

$$\begin{aligned}\|s\|_2 &= \left\|A^T A'' v\right\|_2 \leq \|A'' v\|_2 \\ &\leq m 2^{-b} \|v\|_2 \leq n 2^{-b} \|v\|_2 .\end{aligned}$$

and for $p = 1$,

$$\begin{aligned}\|s\|_1 &= \left\|A^T A'' v\right\|_1 \leq \sqrt{n} \|A'' v\|_1 \\ &\leq m\sqrt{n} 2^{-b} \|v\|_1 \leq n^2 2^{-b} \|v\|_1 .\end{aligned}$$

□

The Augmented Indexing problem. As in (Do Ba et al., 2010), 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 (Miltersen et al., 1998) (see, for example, an explicit proof in (Do Ba et al., 2010))

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

$$\|x - y\| \in \left[\left(\frac{1}{3}\right)^{1/p}, \left(\frac{2}{3}\right)^{1/p} \right]$$

Proof. Consider a τ-balanced linear code over the alphabet $\{\pm\frac{1}{n^{1/p}}\}$ with message length M. It is known that such codes exist with block length $O(M/\tau^2)$ (Ben-Aroya & Ta-Shma, 2009). 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 ℓ_p distance is between $\left[\left(\frac{1}{3}\right)^{1/p}, \left(\frac{2}{3}\right)^{1/p}\right]$. □

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 $\Theta\left(\frac{R}{k^{1/p}}\right)$ apart. Consider the set of points P from Lemma 3.3 scaled to $B^p_{n/k}(\frac{R}{k^{1/p}})$. 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' such that it contains exactly $\min\left(Lr/R, \exp(\Omega(n/k))\right)$ points and let $g_1 : [0, r/k^{1/p}] \to P'$ be a piecewise linear function that goes through all the points in P' in any order. Then, we define $g : \mathbb{R} \to \mathbb{R}^{n/k}$ as:

$$g(x) = \begin{cases} g_1(0) & \text{if } x < 0 \\ g_1(x) & \text{if } 0 \leq x \leq r/k^{1/p} \\ g_1(\frac{R}{k^{1/p}}) & \text{if } x \geq r/k^{1/p} \end{cases}$$

Let $I = \{\frac{r}{k^{1/p}|P'|}, \ldots, \frac{r}{k^{1/p}}\}$ be the points that are pre-images of elements of P'. Observe that g is $O(L)$-Lipschitz since within the interval $[0, r/k^{1/p}]$, since it maps each interval of length $\frac{r}{k^{1/p}|P'|} \geq \frac{rR}{k^{1/p}Lr} = \frac{R}{Lk^{1/p}}$ to an interval of length at most $O(R/k^{1/p})$.

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

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

Also, for every point $(x_1, \ldots, x_k) \in I^k$, $\|G(x_1, \ldots, x_k)\|_p = (\sum_{i \in [k]} \|g(x_i)\|_p^p)^{1/p} \leq R$. However, there still exist distinct points $x, y \in I^k$(for instance points that differ at exactly one coordinate) such that $\|G(x) - G(y)\|_p \leq O(\frac{R}{k^{1/p}})$.

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 $\frac{R}{6^{1/p}}$ in p-norm. So, if we set $G = g^{\otimes k}$ and $X \subset I^k$ to $G^{-1}(X')$, we get a set of points of cardinality $(|P'|/2)^{k/2} \geq (\min(\exp(\Omega(n/k)), Lr/R))^{k/2}$ with minimum distance $\frac{R}{6^{1/p}}$ in p-norm that lie within the ℓ_p ball of radius R. □

Lower bound. We now prove the lower bound for generative model recovery.

Proof of Theorem 1.1. 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 $x \in Z$, $\|x\| \leq \sqrt{Lr\delta}$ and for all $x, x' \in Z$, $\|x - x'\| \geq \sqrt{Lr\delta}/6$. Let $d = \lfloor \log |X| \rfloor \log n$, and let $D = 48(C + 1)$.

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 $y^1, y^2, \ldots, y^{\log n}$:

$$\underbrace{y_1, \ldots, y_{\log |X|}}_{y^1}, \underbrace{y_{\log |X|+1}, \ldots, y_{2 \log |X|}}_{y^2}, \ldots, \underbrace{y_{d - \log |X|}, \ldots, y_d}_{y^{\log n}}$$

where each chunk contains $\lfloor \log |X| \rfloor$ bits and represents an index into X.

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 on a recovery 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 $x \in \left\{\pm\frac{1}{n^{1/p}}\right\}$ the x's need not be discretized.

From Bob's input i, he can compute the chunk $j = j(i)$ 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

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

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

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

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 $B_n^p(R/D^j)$ 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 $\|u\| \leq \frac{R}{D^j}(1 - 1/n^2) \leq \frac{R}{D^j} - \|s\|$ with probability $\frac{\mathrm{Vol}(B_n^p(\frac{R}{D^j}(1-1/n^2)))}{\mathrm{Vol}(B_n^p(\frac{R}{D^j}))} = (1 - 1/n^2)^n > 1 - 1/n$. But $\{w - u \mid \|u\| \leq \frac{R}{D^j} - \|s\|\} \subseteq \{w - s - u \mid \|u\| \leq \frac{R}{D^j}\}$, 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

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

Now,

$$\begin{aligned}
\|x_j - \hat{w}\| &\leq \|w - u - x_j\| + \|w - u - \hat{w}\| \\
&\leq (1 + C)\|w - u - x_j\| + \delta \\
&\leq (1 + C)\left(\|u\| + \frac{1}{D^j} \cdot \sum_{i=1}^{j-1} \left\|D^i x_i\right\|\right) + \delta \\
&\leq 2(1 + C)R/D + \delta \\
&< R \cdot \frac{2(1 + C)}{D} + \delta \\
&= \frac{1}{24} \cdot R + \delta.
\end{aligned}$$

Since $\delta < Lr/24$, this distance is strictly bounded by $R/12$. Since the minimum distance in X is $R/6$, this means $\left\|D^j x_j - \hat{w}\right\| < \left\|D^j x' - \hat{w}\right\|$ 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 $x \in \{\pm\frac{1}{n^{1/p}}\}^n$. 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. Generator for k-sparse vectors

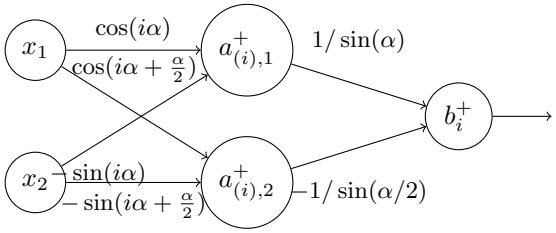
4.1. Explicit Construction

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 generative model 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^{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:

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

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

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

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

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

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 $Im(G) = \{x \mid \|x\|_0 \leq 1\}$. □

Proof of Theorem 1.2. 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 ℓ_2/ℓ_2 (k, C)-sparse recovery.

4.2. Randomized Low Width Construction

We describe the neural network $G_{rand} : \mathbb{R}^{k+1} \to \mathbb{R}^n$ here and we prove in the appendix that given any k-sparse vector y, we show that there exists a vector in $\hat{x} \in \mathbb{R}^{k+1}$ such that $G(\hat{x}) = y$.

For every $i \in [2n]$, pick v_i uniformly at random from $\mathbb{S}^k$. Then pick $v_i^{'} \perp v_i$ uniformly at random from $\mathbb{S}^k$.

We define for every $i \in [2n]$ and $b \in \{0, 1\}$

$$F_i(x) = \Big[\big[\langle x, \tan(\alpha)v_i^{'} + v_i\rangle\big]_+ - 2\big[\langle x, v_i\rangle\big]_+\Big]_+ + \Big[\big[\langle x, \tan(\alpha)v_i^{'} - v_i\rangle\big]_+ - 2\big[\langle x, -v_i\rangle\big]_+\Big]_+$$

where we set $\alpha = 1/n^{O(k)}$. As illustrated in Figure 1, F_i has a non-zero value only when the projection of x on the plane formed by v_i and v_i' is at an angle smaller than α from v_i' or $-v_i'$.

Finally, we define the output of the neural network for each $i \in [n]$:

$$G(x)_i = F_{2i}(x) - F_{2i+1}(x)$$

For negative output values, y_i we use F_{2i+1} and for positive we use F_{2i}. As is clear from the construction of F_i, we use 2 layers and in each layer, we have at most 4 nodes for each $i \in [n]$. So, the width of G is $O(n)$.

Fact 4.2. *Let $u, w \in \mathbb{S}^k$ be two orthogonal vectors chosen uniformly . Define $F(x) = \big[[\langle x, v_i + \tan(\alpha)v_i^{'}\rangle]_+ - 2\big[\langle x, v_i^{'}\rangle]_+\big]_+$.*

Then,

$$\Pr_{x\sim\mathbb{S}^k}[F(x) > 0] = \frac{2\alpha}{\pi}$$

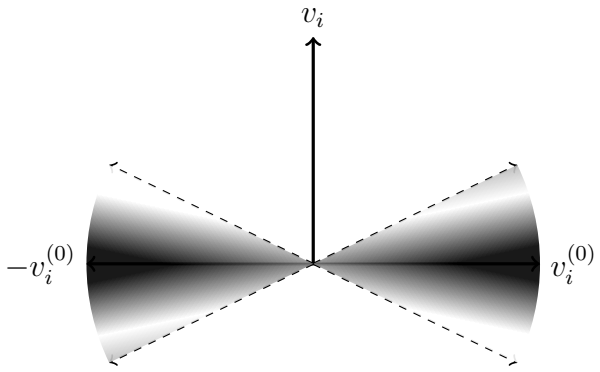


Figure 1. The region where $F_i(x)$ is non-zero

We define $W_S = \{x \in \mathbb{S}^k \mid F_i(x) > 0 \forall i \in S\}$. Let w_S denote the point such that $\|w_S\|_2 = 1$ and $\langle v_i, w_S\rangle = 0$ for all $i \in S$.

Fact 4.3. *The point w_S lies within W_S almost surely.*

4.3. Proof Sketch

We prove that the construction presented here works by proving the following statements:

w_S and $w_{S'}$ are far apart We show that for every S and S' of cardinality k, w_S and $w_{S'}$ are not too close. As an example, consider S and S' such that $S \cap S' = \emptyset$. The resulting w_S and $w_{S'}$ are randomly distributed and independent. So, we know that with high probability they are more than $1/poly(n)$ apart with probability $1/n^k$. When you now consider S and S' that intersect, the w_S and $w_{S'}$ are not independent. However, we may use the fact that that conditioned on being in the subspace $W_{S\cap S'}$, these points are randomly distributed and independent. Since we need to take a union bound over all $\binom{n}{k}$ pairs of such sets (some of which intersect at all but $o(k)$ elements), we show that for any $S, S' \subset [n]$ with $|S| = |S'| = k$, w_S and $w_{S'}$ are at least $1/n^k$ apart. We formally prove this in Lemma B.1.

W_S and $W_{S'}$ are disjoint We show that the regions W_S and $W_{S'}$ around w_S and w_S' do not intersect for every $S \neq S' \subseteq [n]$ with $|S| = |S'| = k$. Intuitively, this would hold because we define $\alpha = 1/n^{8k}$ and the individual regions are likely to have very small volume and since there are only $\binom{n}{k}$ many such regions, they are unlikely to intersect. This statement is true if w_S and $w_{S'}$ are randomly distributed or even if the constraints that define W_S and $W_{S'}$ are independent. Since neither of those statements is true, we use a technique involving ϵ-nets in Lemma B.2 to show that these sets are indeed non-intersecting.

Every k-sparse vector has a pre-image Now, that we have disjoint regions within which G is non-zero at exactly the coordinates in S, we show that given a desired k-sparse output vector y that is non-zero at coordinates $S \subseteq [n]$, there exists a point $\hat{x}$ in W_S such that $G(\hat{x}) = y$. We describe a set of linear constraints such that satisfying those constraints yields such a $\hat{x}$. If y has large ℓ_2 weight, though, such a point might not exist within W_S. However, we may recover a point in W_S that is correct up to scaling and then scale the norm of that point to get output y.

Theorem 2.2 is formally proved in the appendix in Section B.

References

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

Bora, A., Jalal, A., Price, E., and Dimakis, A. G. Compressed sensing using generative models. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 537–546. JMLR.org, 2017. URL http://dl.acm.org/citation.cfm?id=3305381.3305437.

Candès, E. J., Romberg, J., and Tao, T. Stable signal re-

covery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59(8):1208–1223, 2006.

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

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6114.

Liu, Z. and Scarlett, J. Information-theoretic lower bounds for compressive sensing with generative models, 2019.

Miltersen, P. B., Nisan, N., Safra, S., and Wigderson, A. On data structures and asymmetric communication complexity. *J. Comput. Syst. Sci.*, 57(1):37–49, 1998. doi: 10.1006/jcss.1998.1577. URL https://doi.org/10.1006/jcss.1998.1577.