1 Sparse Recovery

Suppose there is some hidden vector $x \in \mathbb{R}^n$ that we can only access using linear measurements. That is, for any query vector $v \in \mathbb{R}^n$ and we can request the inner product $v \cdot x = \sum_{j=1}^{n} v_j x_j$. We would like to recover $x$, while making as few queries $a_1, \ldots, a_m$ as possible.

If $A$ is a matrix $m \times n$ with rows as the query vectors $a_1, \ldots, a_m$, then is same as saying that we want to recover $x$ given the product $Ax = b$.

Of course as stated, this problem is trivial, as we need to make $n$ queries to recover any arbitrary $x$ in general, and $n$ queries also suffices as we can just query each of the $n$ coordinates, using $a_i = e_i$ for $i = 1, \ldots, n$ where $e_i$ are the standard coordinate basis vectors. In general any $n$ linearly independent query vectors $a_1, \ldots, a_n$ would suffice.

The question gets much more interesting when $x$ is $s$-sparse, i.e. it has only $s$ non-zero entries. As we shall see in this can one can recover $x$ uniquely using just $m = O(s \log n/s)$ (or even $2s$ queries).

So the general question we will be interested is the following: Given some measurements $b = Ax$ where $A$ is an $m \times n$ matrix with $m \ll n$ of some hidden $s$-sparse solution $x$, find $x$.

Note that in general the linear system $Ax = b$ is very under-determined, and the solutions form an at least $n - m$ dimensional subspace. So, we wish to find a sparse vector in this space of solutions. This is called the sparse recover, or $\ell_0$ minimization problem.

Applications: This problem has several applications, and related problems are extensive studied in a popular area called compressed sensing. The reason is that interesting data is often quite sparse. For example, images often become very sparse when measured in the wavelet or Fourier basis. This allows significant compression of images. The standard approach is to first capture the entire image, then compress by discarding non-useful frequencies. One can ask why not just measure a few pixels to begin with. (For more on this and compressed sensing, search for the term “single pixel camera”.)

Similarly, if we are trying to find parameters in a machine learning model that fit our data, then among the various options, we would like to choose the one with fewest non-zero parameters (by Occam’s razor principle, this is likely to be a good model.)

In many settings measurement are expensive to begin with. For example, in oil exploration, one gets linear measurements by exploding dynamite.

While we only consider exactly $s$-sparse setting here, these techniques extend very robustly to almost sparse vectors. This has made compressed sensing a huge success area in recent years.

2 Formalizing the problem

We want to design an $m \times n$ matrix $A$, with $m$ as small as possible, so that any $s$-sparse vector $x$ can be recovered uniquely from $b = Ax$. Later we will also focus on how to recover $x$ efficiently from $b$. 
It turns out that $m = 2s$ will suffice, and there is a clean characterization of such $A$’s.

**Theorem 1**  A can uniquely encode $s$-sparse vectors iff every $2s$ columns of $A$ are linearly independent.

**Proof:** If every subset of $2s$ columns of $A$ is linearly independent this means that $Ay \neq 0$, for any non-zero $2s$-sparse vector $y$.

Suppose there are two $s$-sparse solutions $x$ and $x'$ such that $Ax = Ax' = b$, then $A(x - x') = 0$. But as $x - x'$ is $2s$-sparse, this must mean that $x - x' = 0$.

Conversely, if $A$ has some subset $S$ of $2s$ columns that are linearly dependent, then $Ay = 0$, $y \neq 0$ and $y$ is $2s$-sparse. Writing $y$ arbitrarily as $x - x'$ for two $s$-sparse vectors $x, x'$ with disjoint supports (and at least one of them non-zero), gives $Ax = Ax'$, and hence unique recovery fails (we cannot distinguish between $x$ and $x'$).

Such a $2s \times n$ matrix $A$ can be designed, say, by choose entry iid as $N(0,1)$. The probability that any $2s \times 2s$ sub-matrix does not have full-rank will be essentially 0.

**The Decoding Algorithm** While we designed an $A$ with only $2s$ rows, given $b = Ax$, how to we actually find the corresponding $s$-sparse $x$?

Note that we want to solve the $\ell_0$-minimization problem

$$\min \|x\|_0, \quad \text{such that } Ax = b \quad x \in \mathbb{R}^n,$$

which is NP-hard in general. The best we know is the following naive algorithm.

**Algorithm:** For every possible $\binom{n}{s}$ subsets $S$ of $s$ coordinates, consider the linear system $A_{|S}\big|x_{|S} = b$, where $A_{|S}$ is the $m \times |S|$ matrix with columns restricted to $S$. (i.e. if we can guess the $s$ coordinates of $x$, then we can determine the value of $x$ on these coordinates). Note that the running time here is $n^{O(s)}$, which is terrible as $s$ grows.

### 3 The miracle of $\ell_1$ minimization

Suppose $A$ is a generic matrix with $m$ large enough so that unique recovery of $s$-sparse vectors is possible. The algorithm people use in practice to determine the sparse $x$ from $A$ is to solve the following:

$$\min \|x\|_1, \quad \text{such that } Ax = b \quad x \in \mathbb{R}^n.$$

Note that we are solving the wrong problem: minimizing $\ell_1$ instead of $\ell_0$. So it is a miracle that this should work!

Let us first note that this is just a linear program, which can be solved very efficiently.

$$\min \sum_i z_i, \quad \text{such that } Ax = b, z_i \geq x_i \text{ and } z_i \geq -x_i \quad \forall i \in [m], \quad x \in \mathbb{R}^n.$$

Our goal will be to understand why this heuristic of $\ell_1$ minimization works! In fact, this has been very widely used heuristic in machine learning and various other fields for several years: If you want to find a sparse solution that fits some data, just minimize $\ell_1$ instead. This is often referred to as $\ell_1$ regularization. However, until recently it was a big mystery as to why this works so well.
We will prove the following result.
Call a matrix $A$ good, if it can solve sparse recovery via $\ell_1$ minimization for every $s$-sparse vector $x$.

**Theorem 2** Let $A$ be a random $m \times n$ matrix with $m = O(s \log(n/s))$, where each entry is say, $\pm 1$ with probability 1/2 each (there is nothing special about $\pm 1$, and any reasonable model of randomness works). Then $A$ is good with high probability.

In fact, we will do something stronger and identify deterministic condition on $A$ called the restricted isometry property that will ensure that it is good.

**Definition 3 (Restricted Isometry)** For an integer $k$ and $\delta > 0$, we say that $A$ satisfies the $(k,\delta)$-isometry property, if

$$(1 - \delta)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \delta)\|x\|_2^2$$

holds for all $k$-sparse vectors $x$.

So, to prove theorem 2, we show the following two facts.

**Theorem 4** If $A$ satisfies the $(3s,1/2)$-isometry property, then $A$ is good.

**Theorem 5** For any $\delta > 0$, a random matrix $A$ with $m = O((s \log(n/s))/\delta^2)$ satisfies the $(3s,\delta)$-isometry property.

Remark: Observe that the property that each subset of $2s$-columns has full rank is the same as saying that $A$ has $(2s,\delta)$-isometry for $\delta < 1$.

Remark: All these results above holds in the more general setting where the vector is almost $s$-sparse. The proofs of these more general results follows the same spirit as the proof of Theorem 2 below.

4 Some basic facts

We recall some very basic facts about norms, vectors and matrices that we will need in the next few lectures. We also some intuition on what isometry and restricted isometry means.

For a vector $v \in \mathbb{R}^n$, its $\ell_p$ norm, for $p > 0$ is defined as $\|v\|_p = (\sum_{i=1}^n |v_i|^p)^{1/p}$. For $p = 2$, this is the standard euclidean length. For $p = 1$, $\|v\|_1 = \sum_i |v_i|$. As $p$ approaches infinity, $\|v\|_\infty = \max_i |v_i|$.

It is also common to refer to the number of non-zero coordinates, or the sparsity of $v$, as $\|v\|_0$, the $\ell_0$ norm of $v$. Even though this does not formally make sense, the idea is that as $a^0 = 1$ if $a \neq 0$ and $0^0 = 0$, $\sum_i v_i^0$ can be thought of as counting the non-zero coordinates of $v$ (but note that we are not taking the $1/p$-th power here).

For two vectors $u, v \in \mathbb{R}^n$, the inner product $\langle u, v \rangle := u^T v = \sum_i u_i v_i$. Here $u^T$ denotes the transpose of $u$. Sometimes we will also the notation $u \cdot v$ to denote $\langle u, v \rangle$. So, $\|u\|_2 = (\langle u, u \rangle)^{1/2}$. The Cauchy-Schwarz inequality states that $|\langle u, v \rangle| \leq \|u\|_2 \|v\|_2$. In fact, $\langle u, v \rangle = \|u\|_2 \|v\|_2 \cos \theta$, where $\theta$ is the angle between $u$ and $v$.

If $v$ is an $s$-sparse vector, then the norms are related as

$$\|v\|_1 \leq s^{1/2} \|v\|_2 \leq s \|v\|_\infty$$
If $A \in \mathbb{R}^{m \times n}$ is an $m \times n$ matrix with the $(i, j)$-th entry $a_{ij}$, and $x \in \mathbb{R}^n$, then $b = Ax \in \mathbb{R}^m$ is an $m$ dimensional vector with entries $b_i = \sum_{j=1}^n a_{ij}x_j$. So, $b_i$ is the inner product of the $i$-th row with $x$. Another useful view of $b$ is as combination of columns of $A$, weighted by coefficients of $x$, i.e., if $c_j$ denotes the $c_j$-th column of $A$, then $b = \sum_{j=1}^n x_jc_j$.

**Some intuition on Isometry.** A matrix $A$ is called an isometry if it preserves the lengths for every $x \in \mathbb{R}^n$. That is $\|Ax\|_2 = \|x\|_2$.

**Lemma 6** $A$ is an isometry if the columns of $A$ are an orthonormal set of vectors (i.e., each has length 1 are any two are orthogonal).

**Proof:** If the columns of $A$ are orthonormal, then $A^TA = I$. So, we have

$$\|Ay\|^2 = \langle Ay, Ay \rangle = (Ay)^TAy = y^TA^TAy = y^Ty = \|y\|^2$$

Let spell this out so that we understand what really happened under the hood. $Ay = \sum_i y_ia_i$. So,

$$\langle Ay, Ay \rangle = \sum_{ij}y_iy_j\langle a_i, a_j \rangle = \sum_i y_i^2\langle a_i, a_i \rangle + \sum_{i \neq j} y_iy_j \langle a_i, a_j \rangle = 0$$

Now, for an isometry one would need $m \geq n$ (do you see why?). So the $(k, \delta)$-isometry property can be seen as a significant relaxation of isometry, where only require the lengths to be preserved approximately up to $\delta$ relative error, and moreover, we only need the property of vectors $x$ with support at most $k$.

A useful intuition for $(k, \delta)$ might be to consider the case when the columns of $A$ are unit length and satisfy $|\langle c, c' \rangle| \leq \delta/k$.

**Lemma 7** If the columns of $A$ have length 1 and satisfy $|\langle c, c' \rangle| \leq \delta/k$, then $A$ satisfies $(k, 2\delta)$-isometry.

**Proof:** For any vector $x$ with support $k$,

$$\|\sum_i x_ic_i\|^2 = \sum_i x_i^2\|c_i\|^2 + \sum_{i<j} 2x_ix_j\langle c_i, c_j \rangle$$

$$\leq \sum_i x_i^2 + \sum_{i<j} \frac{\delta}{k}(x_i^2 + x_j^2) \leq (1 + 2\delta)\|x\|^2$$

The same proof shows that $\|Ax\|^2 \geq (1 - 2\delta)\|x\|^2$.

**Remark:** This condition is called *decoherence* in the literature, and is more restrictive that $(k, \delta)$-isometry. In particular, If we want a random matrix $A$ to satisfy the stronger condition $|\langle c, c' \rangle| \leq \delta/k$, then we can only ensure that $m = O_\delta(k^2 \log n)$ (as opposed to $O(k \log n)$ in Theorem 5). (this is a fun and easy exercise in probabilistic method to try, once you have seen the proof of Theorem 5).
5 Proof

We now prove Theorem 5. You will prove Theorem 5 in the exercises.

Proof: Let \( x \) be the (hidden) \( s \)-sparse vector, and let \( b = Ax \). Let \( x^* \) be the solution obtained by solving the \( \ell_1 \)-minimization LP. Clearly, \( \|x^*\|_1 \leq \|x\|_1 \). Let \( x^* = x + h \).

Our goal will be to show that if \( A \) has the \((3s, 1/2)\)-isometry property, then \( h = 0 \).

Let \( T_0 \subseteq [n] \) denote the support of \( x \). For a vector \( y \) and a subset of coordinates \( T \subseteq [n] \), let \( y_T \) denote the vector restricted to coordinates in \( T \). Let \( T_0^c \) denote the remaining coordinates \( [n] \setminus T_0 \).

The first crucial observation is that as \( \|x^*\|_1 = \|x + h\|_1 = \|x^* + h\|_1 \), the \( \ell_1 \) mass of \( h \) outside \( T_0 \) can be at most inside \( T_0 \) (this is much easier to see by a picture).

Lemma 8 \( \|h_{T_0}\|_1 \geq \|h_{T_0^c}\|_1 \).

Proof: We use that \( \|x\|_1 \geq \|x^*\|_1 \) and expand out the \( \ell_1 \) norms.

\[
|x|_1 \geq |x + h|_1 = \sum_{i \in T_0} |x_i + h_i| + \sum_{i \in T_0^c} |x_i + h_i| \geq \sum_{i \in T_0} (|x_i| - |h_i|) + \sum_{i \in T_0^c} |h_i| = \|x\|_1 - \|h_{T_0}\|_1 + \|h_{T_0^c}\|_1
\]

where the second inequality uses hat \( |a + b| \leq |a| - |b| \) for any real numbers \( a, b \).

The second key fact is that \( Ah = 0 \) as \( Ax = Ax^* = b \).

The key idea of the proof will be that if \( Ah = 0 \), then \((3s, \delta)\)-isometry property of \( A \) will force \( h \) to have a lot of \((\ell_2)\) mass outside \( T_0 \) (for instance if \( h \) was \( 3s \)-sparse vector, then \( h \) would have to be 0 as by \((3s, \delta)\)-isometry \( \|Ah\|_2^2 \geq (1 - \delta)\|h\|_2^2 \)). We now give the details.

Lemma 9 Let \( x \) and \( x' \) be supported on disjoint subsets \( T, T' \) of \([n]\) with \(|T| + |T'| \leq 3s \). Then \(|\langle Ax, Ax' \rangle| \leq \delta\|x\|_2\|x'\|_2 \).

Proof: As both sides are homogenous, we can assume that \( \|x\|_2 = \|x'\|_2 = 1 \). As \( x + x' \) and \( x - x' \) are \( 3s \)-sparse vectors, and moreover as \( x, x' \) have disjoint supports, \( \|x + x'\|_2^2 = \|x\|_2^2 + \|x'\|_2^2 = 2 \). Similarly, \( \|x - x'\|_2 = \sqrt{2} \). So, by the \((3s, \delta)\)-isometry property of \( A \),

\[
2(1 - \delta) \leq \|A(x \pm x')\|_2^2 \leq 2(1 + \delta)
\]

By the parallelogram trick to relate the inner product to lengths,

\[
\langle Ax, Ax' \rangle = \frac{1}{4} (\|Ax + Ax'\|^2 - \|Ax - Ax'\|^2) \leq \frac{1}{4} ((1 + 2\delta) - (1 - 2\delta)) = \delta
\]

Let us sort the coordinates \( i \) in \( T_0^c \) according non-increasing value of \( h(i) \), and group them into blocks \( T_1, T_2, \ldots \) each of size \( s \). Then Lemma 8 can be written as

\[
\|h_{T_0}\|_1 \geq \sum_{i \geq 1} \|h_{T_i}\|_1
\]

By Cauchy Schwarz, and as every coordinate of \( h_{T_i} \) is smaller than every coordinate in \( h_{T_{i-1}} \), we have

\[
\frac{1}{\sqrt{s}} \|h_{T_i}\|_1 \leq \|h_{T_i}\|_2 \leq \frac{1}{\sqrt{s}} \|h_{T_{i-1}}\|_2
\]
Let us denote $T_{01} = T_0 \cup T_1$. Then $h_{T_{01}} + \sum_{i \geq 2} h_{T_i} = h$, and so

$$Ah_{T_{01}} + \sum_{i \geq 2} Ah_{T_i} = Ah = 0$$

Taking inner product with $Ah_{T_{01}}$ on both the sides we get

$$\langle Ah_{T_{01}}, Ah_{T_0} \rangle = -\sum_{i \geq 2} \langle Ah_{T_i}, Ah_{T_0} \rangle$$

By the isometry property the lhs is at least $(1 - \delta) \|h_{T_{01}}\|_2^2$. Similarly, by lemma ?? and noting that $h_{T_{01}}$ has disjoint support from $h_{T_i}$ for each $i \geq 2$, $|\langle Ah_{T_i}, Ah_{T_{01}} \rangle| \leq \delta \|h_{T_i}\|_2 \|h_{T_{01}}\|_2$

This gives

$$(1 - \delta) \|h_{T_{01}}\|_2^2 \leq \delta \sum_{i \geq 2} \|h_{T_i}\|_2^2$$

Now, we relate the $\ell_2$-norms to the $\ell_1$ norms to obtain a contraction to (??). First, by Cauchy Schwarz, $|h_{T_0}|_1 \leq s^{1/2} |h_{T_0}|_2 \leq s^{1/2} |h_{T_0}|_2$. Next, the key point (and is this where we use that we form $T_1, T_2, \ldots$, by sorting the values of entries of $h$) is that for each $i \geq 2$

$$\|h_{T_i}\|_2 \leq \|h_{T_{i-1}}\|_1$$

Together this gives

$$(1 - \delta) \|h_{T_0}\|_1 \leq \delta \sum_{i \geq 1} \|h_{T_i}\|_1$$

which contradicts (??) when $\delta < 1/2$. \qed