Tutorial on compressive sensing and random projections

Kamesh Krishnamurthy

September 26, 2016

Janelia Theoretical Neuroscience Workshop

1 Solving $\mathbf{y} = A\mathbf{x}$

Consider the problem of taking linear measurements about some signal $\mathbf{x}$ – for e.g., $\mathbf{x}$ could be an image or a time-series signal and A could be the Fourier transform operator, in which case, we make frequency domain measurements about our signal. Now, suppose that our signal $\mathbf{x}$ resides in some high-dimensional space of dimension N and we only take M measurements. Can we recover $\mathbf{x}$ from the measurements $\mathbf{y}$?

In general, when $M > N$, and A is full-rank, we have a overdetermined system of equations and we can use methods like least-squares to give us a solution. What about the case when $M < N$? In this case, we have an underdetermined system of equations, and in general, we don't have a unique solution for $\mathbf{x}$. In certain scenarios, of all the possible solutions to $\mathbf{y} = A\mathbf{x}$, the one you care about is "small" in some sense. One notion of small is the solution having the least "energy" or the minimum $L_2 -$ norm. In this case we get the pseudo-inverse solution:

$$\begin{aligned} \hat{x} &= \arg\min \|\mathbf{x}\|_2 \quad \text{s.t. } \mathbf{y} = A\mathbf{x} \\ &= A^* \left(AA^*\right)^{-1} \mathbf{y} \end{aligned}$$

This amounts to "growing" the L_2 "ball" till you satisfy the constraint as shown in the schematic in Fig. 1.1

A classic example of this is finding the minimum energy reconstruction of a signal $f \in \mathbb{R}^N$ from a limited number of $M < N$ Fourier measurements : $\tilde{f}(\omega_1), \tilde{f}(\omega_2) \cdots \tilde{f}(\omega_M)$. In this case, the solution is simple : it is simply the reconstructed signal with the measured Fourier coefficients.

$$\hat{f}(t) = \sum_{i=1}^{M} \tilde{f}(\omega_i) e^{i2\pi\omega_i t/N}$$

However, in many cases the minimum energy solution is way off and not what we are looking for. For instance, consider the Logan-Shepp phantom[6] image and its corresponding minimum L_2 solution reconstructed from a limited number of (2D) Fourier measurements (Fig.[1.2]). The minimum energy solution has a number of artifacts which obscure the structure in the image.

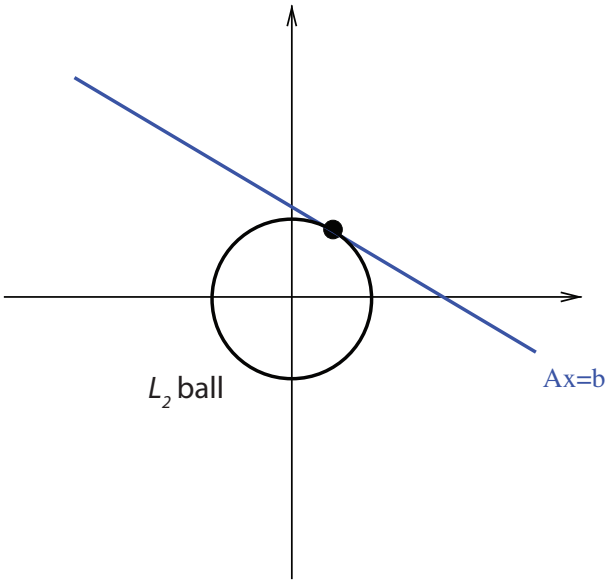

Figure 1.1: Geometric illustration of the pseudo-inverse solution in 2D

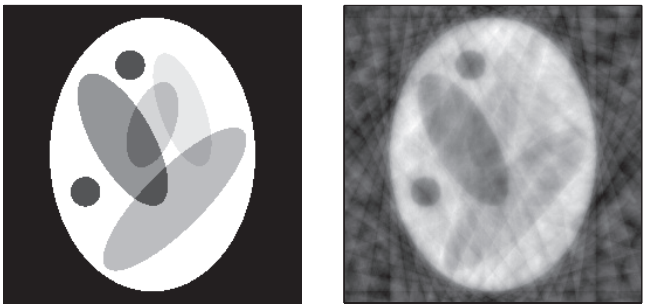

Figure 1.2: minimum L_2 reconstruction (right) performs poorly when reconstructing the image(left) from incomplete 2D Fourier measurements

2 Solving $\mathbf{y} = A\mathbf{x}$ for sparse signals

In many scenarios, the signal $\mathbf{x}$ has a sparse or "compressible" structure – i.e., only a few elements of $\mathbf{x}$ are significant. A signal is called $K-$sparse if it has at most K non-zero entries; the location of these non-zero entries can, however, be arbitrary. Given such a structure for the signal, can we exploit this information to recover the signal from incomplete measurements?

Consider this interesting empirical observation : if you find a solution which minimized the L_1 norm instead of the L_2 norm, then in many cases you can *exactly* reconstruct the signal from highly incomplete measurements, provided the signal has a sparsity structure.

Specifically, if we solve the following problem :

$$\begin{aligned} \arg\min \|\mathbf{x}\|_1 \quad \text{s.t.} \quad & \mathbf{y} = A\mathbf{x} \\ \text{where } \|\mathbf{x}\|_1 \quad = \quad & \sum_{i=1}^{N} |x_i| \end{aligned} \tag{2.1}$$

then we can *exactly* recover the true (sparse) $\mathbf{x}$ even when the number of measurements $\mathbf{y}$ ($M-$dimensional) are substantially fewer than the dimension, N, of $\mathbf{x}$. This fact was well known to researchers studying seismology data, where the reflected signals naturally had a sparse structure due to discrete transitions in the earth's crust . Fig.[2.1] shows that the minimum L_1 reconstruction of Logan-Shepp phantom from incomplete ($\backsim 1\%$) Fourier measurements gives back the exact image[6]. Note, that the sparsity structure exploited in this case is the sparsity in the *gradient* of the image and not the pixels themselves.

Another early example of the success of L_1 success, with more theoretical backing, is the basis pursuit problem. Consider a measured signal $f \in \mathbb{R}^n$ which is made up of the superposition of two signals, one which is sparse in an ortho-basis Φ_1 and the other which is sparse in another ortho-basis Φ_2 – i.e. $f = \Phi\alpha$ where $\Phi = [\Phi_1\Phi_2]$ and α is sparse. We would like to know α so that we can split the signal into its components. A practical example comes from astronomy where telescope images often contain elements which look "texture" like and elements that look like lines or rods. We would like to separate the texture "background" from the more linear features. A theorem due to [10], showed that if the "coherence" μ between the bases defined as

$$\mu(\Phi_1, \Phi_2) \quad = \quad \sqrt{n} \max_{1 \le i,j \le n} \left| \left\langle \varphi_i^{(1)}, \varphi_j^{(2)} \right\rangle \right| \tag{2.2}$$

is small, then solving the following problem gives back the exact α :

$$\arg\min \|\hat{\alpha}\|_1 \quad \text{s.t.} \quad f = \Phi\hat{\alpha}$$

provided

$$\|\alpha\|_0 \quad \le \quad \frac{0.9}{\mu(\Phi_1, \Phi_2)} \sqrt{n}$$

The coherence between two bases is a measure of how "different" do bases function in one basis look compared to the other. And, for any two orthobases, $\mu \ge 1$, and for good reconstruction we want the coherence to be low. Time and frequency bases would be examples of bases which have small coherence. We will return later to the notion of coherence. But, note that the restriction here about the sparsity of f : $(O\sqrt{n})$ is rather restrictive; it requires f to be quite sparse. There are more powerful theorems that guarantee that L_1 will do well under much more general conditions!

Let us now consider one example of a family of more general theorems[6, 7, 10] which guarantee exact reconstruction :

- **Theorem** [Candes & Tao 2006]: Let A be a $M \times N$ matrix with entires A_{ij} drawn i.i.d from $\mathcal{N}(0,1)$. Let $\mathbf{x} \in \mathbb{R}^N$ be an unknown, but *fixed* $K-$sparse vector, and furthermore assume we have access to the M measurements $\mathbf{y} = A\mathbf{x}$. Then we can reconstruct $\mathbf{x}$ exactly with overwhelming probability by solving

$$\begin{aligned} \arg\min \|\hat{\mathbf{x}}\|_1 \quad \text{s.t.} \quad & \mathbf{y} = A\hat{\mathbf{x}} \\ \text{provided } M \quad \gtrsim \quad & K \log\left(\frac{N}{K}\right) \end{aligned} \tag{2.3}$$

Figure 2.1: minimum L_1 reconstruction (right) returns the exact original image (left) from incomplete 2D Fourier measurements

> Moreover, no other method can reconstruct $\mathbf{x}$ with "fewer" (in order) measurements – even if you use some other (even *adaptive*) sensing and any reconstruction method, you cannot do better (in order) than Gaussian sensing followed by L_1 reconstruction.

This remarkable theorem states that if you know that $\mathbf{x}$ is $K-$sparse you only need to take slightly more than $\mathcal{O}(K)$ **non-adaptive** measurements to get back $\mathbf{x}$ by convex optimisation, and this scheme is universally optimal in some sense. For Gaussian matrices, something like $\sim 4K$ measurements will suffice. In the what follows we will review some theory for why this works, and in particular the following important practical questions:

- What are the requirements on the linear measurement operator A ?
- For some A, how many measurements M are required to guarantee exact reconstruction of a N dimensional signal with K non-zero entries?
- How strict is the sparsity requirement? In particular, will the results hold for compressible signals, where only K entries are significant, but other entries are small and non-zero?
- What happens if we have measurement noise?

3 Why does the L_1 solution work?

To understand why the L_1 solution gives the exact result for the underdetermined system, let us first consider the related question: When can we recover any $K-$sparse vector $\mathbf{x}$ from M measurements $\mathbf{y} = A\mathbf{x}$ using any method whatsoever? It turns out that if *any* of $2K$ columns of A are linearly independent (so necessarily $M \geq 2K$), then there is a unique solution to $\mathbf{y} = A\mathbf{x}$ for a $K-$sparse vector $\mathbf{x}$. To see this, assume there are two solutions $\mathbf{x}, \widetilde{\mathbf{x}}$ then $\mathbf{x} - \widetilde{\mathbf{x}}$ is at most $2K-$sparse, and $A(\mathbf{x} - \widetilde{\mathbf{x}}) = 0$ – this cannot be true unless $\mathbf{x} = \widetilde{\mathbf{x}}$ because any $2K$ columns of A are linearly independent. This argument also suggests an algorithm to find the unique solution for the case $M \geq 2K$: choose every subset of K columns and try to solve $\mathbf{y} = A_K \mathbf{x}_K$ where A_K is the submatrix of A with the K selected columns and $\mathbf{x}_K$ is the vector with non-zero entries of $\mathbf{x}$. This problem can be equivalently formulated as minimising the "L_0"norm

$$\begin{aligned} \arg\min \|\mathbf{x}\|_o \quad \text{s.t.} \quad & \mathbf{y} = A\mathbf{x} \\ \text{where } \|\mathbf{x}\|_o \quad = \quad & \text{no. of non-zero entries in } \mathbf{x} \end{aligned} \tag{3.1}$$

Unfortunately, this problem which we call the L_0 problem contains within it the subset-sum problem which is known to be NP complete. So it's hopeless to use this for even moderate size problems.

The L_1 problem is however admits a polynomial time solution – in fact, it is a linear program. To see this, note that solving the problem

$$\begin{aligned} \arg\min \|\mathbf{x}\|_1 \quad \text{s.t.} \quad & \mathbf{y} = A\mathbf{x} \\ \text{where} \quad \|\mathbf{x}\|_1 \quad = \quad & \sum_{i=1}^{N} |x_i| \end{aligned}$$

can be recast as the following equivalent problem

$$\begin{aligned} \text{minimise} \quad & \textstyle\sum_i t_i \\ \text{subject to} \quad -t_i \quad & \leq x_i \leq \quad t_i \\ \text{and} \quad \mathbf{y} \quad & = \quad A\mathbf{x} \end{aligned}$$

which is a well known linear program that can be solved using, for e.g., the simplex method.

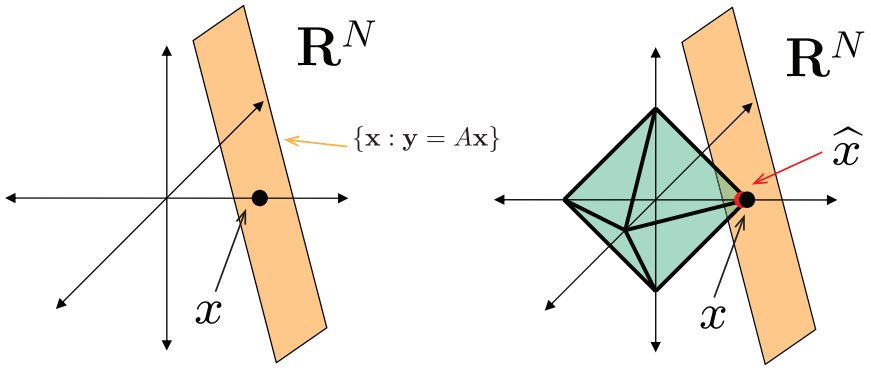

Figure 3.1: The L_1 ball intersects the constraint surface at points which are sparse. In high dimensions the L_1 ball is even more "pointed" and looks a lot like the L_0 ball. . Schematic adapted from [2]

To get an intuition for why, with sufficient measurements, the L_1 problem returns the same solution as the L_0 problem[11] (which is the best you can do), it's useful to view the problem geometrically (Fig. [3.1]). Sparse signals reside in union of planes (the L_0 ball); in particular 1-sparse signals are the union of the axes in $\mathbb{R}^N$. So when searching for sparse solutions to $\mathbf{y} = A\mathbf{x}$ we look for places where the hyperplane $\mathbf{y} = A\mathbf{x}$ intersects the L_0 ball as shown in the schematic Fig.[3.1] . It turns out that the L_1 ball, because of its pointed vertices, also intersects the hyperplane at places where the solutions are sparse, and with sufficient measurements the solution to the L_0 and the L_1 problem are exactly identical. In very high dimensions the L_0 ball looks a lot like the L_1 ball and nothing like the L_2 ball. Of course, you can't always get back the L_0 solution by solving the L_1 problem, otherwise P=NP! But, as we will see CS theory tells us that the two give the same solution by taking slightly more measurements than the minimal amount required.

4 Non-sparse signals : the best $K-$term approximation

The problem used to motivate compressive sensing assumed that the signals of interest were $K-$ sparse, however in most practical situations, the signals are not exactly sparse, but they are "compressible" : only a few entries are significant and the remaining entries decay rapidly but are non-zero. How do the compressive sensing results hold for compressible signals? What about signals with no known structure *a priori* ? To understand the extension of the classical compressive sensing results to non-sparse signals it is useful to ask the following question: *what's the best you can hope to do if you only got to make K measurements of a non-sparse signal?* As discussed above, of course, we can't expect to get back a general signal from incomplete measurements, but can we attempt to get back the "best reconstruction possible from K measurements" or will the compressive sensing paradigm fall apart for general signals?

If you had access to an oracle who told you where the K most significant entries in $\mathbf{x}$ were, then of course you'd want to measure those with your K measurements, and we would end with a best $K-$term approximation of $\mathbf{x}$. Let's call this $K-$term approximation as $\mathbf{x}_K$, which can also be written as

$$\mathbf{x}_K \quad = \quad \underset{\mathbf{y}:\text{k-sparse}}{\arg\min} \left\|\mathbf{y}-\mathbf{x}\right\|_2$$

This is the benchmark we would like to compare the reconstructed vector from the L_1 recovery. CS theory essentially tells you that with just a few more samples than K for **any** $\mathbf{x}$ – not necessarily sparse or compressible, we will recover the best $K-$term approximation $\mathbf{x}_K$. This recovered signal will be a good approximation of the original signal only for sparse or compressible signals, but the CS theory essentially tells you that the "best $K-$term approximation" results hold for any $\mathbf{x}$. So, sensing in a non-adaptive way followed by L_1 reconstruction gives us a performance close to an oracle with perfect knowledge of the largest entries in the signal!

In several of the theorems for non-sparse signals, guarantees are provided that the accuracy of the reconstructed vector $\hat{\mathbf{x}}$ from, say K, measurements is close to $\mathbf{x}_K$. These guarantees usually bound the reconstruction error ($\left\|\hat{\mathbf{x}}-\mathbf{x}\right\|_1$ and $\left\|\hat{\mathbf{x}}-\mathbf{x}\right\|_2$) by the benchmark $\left\|\mathbf{x}_K-\mathbf{x}\right\|_1$ and are often referred to as "**oracle bounds**". If indeed $\left\|\mathbf{x}_K-\mathbf{x}\right\|_1$ is small, then these bounds say that $\hat{\mathbf{x}}$ will be close to the true signal; otherwise, they say that you can expect to do as well as the case where an oracle tells you the location of the K largest elements in $\mathbf{x}$. It doesn't get better than this!

[Optional] Oracle bounds often compare a $2-$ norm ($\left\|\hat{\mathbf{x}}-\mathbf{x}\right\|_2$) to a $1-$norm ($\left\|\mathbf{x}_K-\mathbf{x}\right\|_1$). To know if and when these bounds are tight/good, let us look at the behavior $1-$norm and $2-$norm of compressible signals. A common model for compressible signals is a power-law decay of rank-ordered entries in $\mathbf{x}$. Let $\mathbf{x}$ be a rank-ordered compressible signal s.t.

$$|\mathbf{x}_i| \quad \geq \quad |\mathbf{x}_{i+1}| \ \forall i$$

and

$$|\mathbf{x}_i| \quad \leq \quad \frac{R}{i^p} \quad \forall i > K;\, p > 1 \tag{4.1}$$

for some constant R. Let us consider the signal $\mathbf{x}_K$ which is constructed from $\mathbf{x}$ by retaining the K largest components

$$(\mathbf{x}_K)_i \quad = \quad \begin{cases} \mathbf{x}_i & i \leq K \\ 0 & i > K \end{cases}$$

We are interested in observing how the 1-norm and 2-norm of the residual $(\mathbf{x} - \mathbf{x}_K)$ behave for compressible signals in the following limit :

$$N \gg K \gg 1 \tag{4.2}$$

This is the relevant regime for real-life signals such as images. The norms of the residuals are given by

$$\begin{aligned} \|\mathbf{x} - \mathbf{x}_K\|_1 &= \sum_{i=K+1}^{N} |\mathbf{x}_i| \\ \|\mathbf{x} - \mathbf{x}_K\|_2 &= \left(\sum_{i=K+1}^{N} |\mathbf{x}_i|^2 \right)^{1/2} \end{aligned} \tag{4.3}$$

For compressible signals we have $|\mathbf{x}_i| \leq R/i^p$ for $i > K$. Let us first consider the case when $\mathbf{x}_i \sim R/i^p$ for $i > K$ (the case when the fall-off is faster than R/i^p will turn out to better for the CS theorems); i.e. the bound $\mathbf{x}_i \leq R/i^p$ is tight. In this case, for the limit in (4.2) we can approximate the sums by integrals

$$\begin{aligned} \|\mathbf{x} - \mathbf{x}_K\|_1 \approx R \int_K^N \frac{1}{t^p} \mathrm{dt} &= \frac{R}{p-1} \left[\frac{1}{K^{p-1}} - \frac{1}{N^{p-1}} \right] \\ &\approx \left(\frac{R}{p-1} \right) K^{1-p} \\ \|\mathbf{x} - \mathbf{x}_K\|_2 \approx R \left(\int_K^N \frac{1}{t^{2p}} \mathrm{dt} \right)^{1/2} &= R \left[\frac{1}{2p-1} \left(\frac{1}{K^{2p-1}} - \frac{1}{N^{2p-1}} \right) \right]^{1/2} \\ &\approx \left(\frac{R}{\sqrt{2p-1}} \right) K^{(1/2-p)} \end{aligned} \tag{4.4}$$

so

$$\frac{\|\mathbf{x} - \mathbf{x}_K\|_1}{\sqrt{K}} \approx \text{constant} \times \|\mathbf{x} - \mathbf{x}_K\|_2 \tag{4.5}$$

Note that these relations will only hold for $p > 1$ otherwise we can no longer neglect the dependence on N(which, for e.g., will enter as a logarithmic term for $p = 1$).

The take away from all this is the following: if you see a bound comparing $\|\hat{\mathbf{x}} - \mathbf{x}\|_2$ to $\|\mathbf{x} - \mathbf{x}_K\|_1 / \sqrt{K}$, then this bound is appropriate for compressible signals as described in (4.1) , but if the smaller entries in the signal fall off much slower, then this bound is not assured to be tight. In any case, it always makes sense to compare $\|\hat{\mathbf{x}} - \mathbf{x}\|_1$ to $\|\mathbf{x} - \mathbf{x}_K\|_1$.

5 Requirements for the sensing procedure

5.1 Incoherence and random sampling

Let $\mathbf{x} \in \mathbb{R}^N$ be $K-$sparse, and suppose we make M linear measurements of $\mathbf{x}$ using sensing vectors chosen *uniformly randomly* from Φ_2 (for this result, we needn't restrict Φ_2 to be an orthobasis). Then [5] has shown the following:

- If the number of measurements

$$M \gtrsim K \cdot \mu^2(\mathbb{I}, \Phi_2) \log N \tag{5.1}$$

then solving L_1 problem (2.1) exactly recovers the sparse component of the signal. Here $\mathbb{I}$ is the canonical (e.g. time) basis.

- Moreover, if M is less that $O\left(K \cdot \mu^2(\mathbb{I}, \Phi_2) \log N\right)$, then no algorithm (even combinatorially hard ones) can recover $\mathbf{x}$ from such measurements.

where $\mu(\Phi_1, \Phi_2)$ is the coherence between the two bases as defined in [2.2]. There are a few points worth mentioning about this result:

1. The result holds for a random set of M measurements – i.e. any typical set of M measurements is equally good as long as M satisfies [5.1].
2. The coherence between the bases $\mu(\mathbb{I}, \Phi_2)$ plays a critical role in deciding how many measurements are required for recovery. So, we ideally want bases which have very low ($O(1)$) coherence. An example of incoherent bases is: Φ_1-time and Φ_2- Fourier. As a trivial example, if $\mathbb{I} = \Phi_2$ then the bases are maximally coherent$\left(\mu = \sqrt{N}\right)$, and we need around $N \log N$ samples. So, taking time samples of a signal which is $1-$sparse in the time domain, will require us to collect $\sim N \log N$ before we recover the signal $\mathbf{x}$(why the $\log N$ term? – hint c.f. point above!).
3. If Φ_2 is a random basis, for e.g. if each element of the vector $\varphi \in \Phi_2$ is sampled i.i.d from $\mathcal{N}\left(0, 1/\sqrt{N}\right)$ then with very high probability it is incoherent with any orthobasis Φ_1.

This result can be extended to non-sparse signals. If the number of measurements satisfies (5.1), then the solution to

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \left[\|\mathbf{y} - A\mathbf{x}\|^2 + \lambda \sum_{i=1}^{N} |\mathbf{x}_i| \right] \tag{5.2}$$

will satisfy (with very large probability)

$$\begin{aligned} \|\hat{\mathbf{x}} - \mathbf{x}\|_2 &\lesssim \frac{\|\mathbf{x} - \mathbf{x}_K\|_1}{\sqrt{K}} \\ \|\hat{\mathbf{x}} - \mathbf{x}\|_1 &\lesssim \|\mathbf{x} - \mathbf{x}_K\| \end{aligned}$$

So, with slightly more than K measurements we are close to the best $K-$ term approximation for any $\mathbf{x}$. The problem (5.2) is also referred to as LASSO in the literature, and the problem tries to find the best (in the L_2 sense) solution to a linear system while at the same time penalising non-sparse solution. The parameter λ controls this trade-off and can be chosen appropriately (see [5]).

5.2 Why do we need at least $K \cdot \mu^2 \log N$ measurements?

To get some intuition about why we need $K \cdot \mu^2(\Phi_1, \Phi_2) \log N$ samples using our random sampling scheme, consider the simple case when Φ_1 is the Fourier domain and Φ_2 is the time

domain. In this case, $\mu = 1$ – the bases are maximally incoherent. Further, let's assume that the signal $\mathbf{x} \in \mathbb{R}^N$ is a "Dirac comb" which is $K-$ sparse :

$$x[t] = \sum_{j=0}^{K-1} \delta[t - \tau j]$$

where $N = \tau K$ and the spacing between the spikes is τ. The Fourier transform of $\mathbf{x}$ will be $\tau = N/K$ sparse with the spacing between the spikes K:

$$\tilde{x}[f] = K \sum_{j=0}^{\tau-1} \delta[f - jK]$$

So the Fourier transform of a Dirac comb is a Dirac comb, and the spacing between the spikes of the combs in the time and frequency domain are inversely related. This is the classic time-frequency duality of the Fourier transform.

Now let's say we take $\sim K$ Fourier samples of $\mathbf{x}$. The probability we will sample a zero element is $(1 - \tau/N)$. So the probability that all of our M random Fourier measurements are zero is $(1 - 1/K)^M$. Therefore, *any* method would fail with probability at least $1/N$, if

$$\begin{aligned} \left(1 - \frac{1}{K}\right)^M &\geq \frac{1}{N} \\ \text{i.e. } M &\leq K \log N \end{aligned}$$

So we need *at least* $K \log N$ random non-adaptive samples for any (even combinatorially hard) method to work. And, it turns out that a linear program will work fine if M satisfies (5.1)! Extensions to the case when $\mu > 1$ are simple (see [4]).

5.3 Restricted isometries

There are two parallel theories of compressive sensing: i) a set of results based on incoherence as discussed above and ii) a complimentary set of results based on a property of the sensing matrix A called the restricted isometry property or RIP. The theorems based on incoherence, discussed above, rely on taking measurements with randomly selected measurement vectors from a basis which is incoherent with the basis in which the signal is sparse or compressible. However, the theorems which make use of RIP are deterministic and their guarantees hold as long as A has the requisite RIP property.

Let $\mathbf{x}_K$ be a $K-$ sparse vector, then the $M \times N$ matrix A is said to have a **restricted isometry constant** δ_K of order K provided that δ_K is the smallest scalar which satisfies

$$(1 - \delta_K) \|\mathbf{x}_K\|_2^2 \leq \|A\mathbf{x}_K\|_2^2 \leq (1 + \delta_K) \|\mathbf{x}_K\|_2^2$$

for *all* $K-$sparse vectors $\mathbf{x}_K$. The matrix A is said to satisfy the RIP of order K provided that the constant δ_K is sufficiently small. The RIP is essentially a requirement that all subset of K columns of A are *approximately* orthogonal – of course, they can't all be perfectly orthogonal since $M < N$. If this condition is satisfied, then the lengths of the sparse vectors are preserved when they are projected on the column space of A. A simple extension is that if A satisfies the RIP of order $2K$ then the distances between $K-$ sparse vectors are preserved by A. Another way of stating the RIP is that any submatrix of A formed by choosing K columns is well-conditioned (actually the condition number is $(1 - \delta_K) / (1 + \delta_K)$).

A theorem by [6, 8] shows that if A satisfies $\delta_{2K} < \sqrt{2} - 1$, then we essentially get back the guarantees in the previous section with incoherent sampling. More precisely :

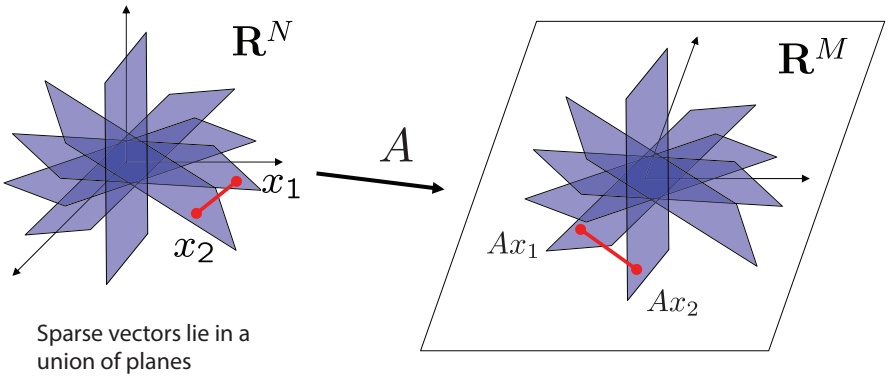

Figure 5.1: A matrix A satisfying the RIP of order $2K$ (δ_{2K} is sufficiently small) will approximately preserve the distance between all $K-$sparse vectors $\mathbf{x}_1$ and $\mathbf{x}_2$. Schematic adapted from [2]

- **Theorem** [Candes, Romber, Tao 06]: If A satisfies $\delta_{2K} < \sqrt{2} - 1$ then the solution $\hat{\mathbf{x}}$ to

$$\arg\min \|\tilde{\mathbf{x}}\|_1 \quad \text{s.t.} \quad \mathbf{y} = A\tilde{\mathbf{x}}$$

guarantees that

$$\begin{aligned} \|\hat{\mathbf{x}} - \mathbf{x}\|_2 &\lesssim \frac{\|\mathbf{x} - \mathbf{x}_K\|_1}{\sqrt{K}} \\ \|\hat{\mathbf{x}} - \mathbf{x}\|_1 &\lesssim \|\mathbf{x} - \mathbf{x}_K\| \end{aligned} \tag{5.3}$$

Note that

- The theorem is valid for *all* **x**unlike the incoherent sampling theorems which applied to a fixed **x**and random measurements. If $\mathbf{x}$ happens to be $K-$sparse, then we get exact recovery otherwise we get the best $K-$term approximation.
- $\delta_{2K} < 1$ will guarantee that there is a unique $K-$sparse solution to $\mathbf{y} = A\mathbf{x}$, but you have to solve a NP-hard problem to find it! However, the above theorem states that more stringent requirement $\delta_{2K} < \sqrt{2} - 1$ will not only guarantee that the solution is unique, but the L_1 problem will find it!

The way we have stated the theorem (5.3) makes no mention of the number of measurements or randomness! All it says is that if A satisfies the RIP of order $2K$ then we get back the best $K-$sparse approximation by solving L_1. So we are left with the task of constructing matrices which satisfy the RIP for which M is close to K. Calculating the restricted isometry constant of a matrix is actually a NP-hard problem [13].

This is where randomness enters the picture: random matrices satisfy the RIP (for δ_{2K}) with very high probability. More precisely, let A be an $M \times N$ random matrix with the entries A_{ij} sampled i.i.d from a distribution F. Then A will satisfy the RIP ($\delta_{2K} < \sqrt{2} - 1$) with very high probability provided

$$\begin{aligned} M &\gtrsim K \log\left(\frac{N}{K}\right) \quad \text{for } F \equiv \mathcal{N}\left(0, \frac{1}{\sqrt{M}}\right) \\ M &\gtrsim K \log\left(\frac{N}{K}\right) \quad \text{for } F \equiv \text{Bernoulli}\left(\pm\frac{1}{\sqrt{M}}\right) \end{aligned}$$

There are other distributions F for which the RIP holds with M being slightly more than above [1]. So, there's a kind of universality in sensing with random matrices. Also, the RIP hold for a Fourier (DFT) matrix provided $M \gtrsim K (\log N)^4$[6]. An incoherent sampling matrix satisfies RIP with high probability provided $M \gtrsim K\mu^2 (\log N)^4$. Another remarkable result due to [5, 11] is that no other sensing mechanism – adaptive or non-adaptive – or any other reconstruction algorithm can do better with substantially fewer samples, provided that the signal is sparse or compressible in a power law sense (4.1).

Random matrices also have the desirable property that they are universal sensing matrices in some sense. If $\mathbf{x}$ is sparse in a basis $B : \mathbf{x} = B\alpha$, and we take measurements using a random matrix $A : \mathbf{y} = A\mathbf{x} = AB\alpha$, then if the matrix A satisfies the RIP, then so will AB. Thus, we are guaranteed to recover the sparse coefficients α even if we don't know the sparsity basis *a priori* – the measurements can be completely non-adaptive.

To see this, we show that if $[A]_{ij} \sim i.i.d\mathcal{N}\left(0, \sigma^2\right)$ satisfies RIP (δ_{2K} is small) with high probability, then $A \cdot B$ also satisfy this property for an orthonormal matrix B. Let us look at the statistics of the element of $A \cdot B$. Each entry of the matrix is a linear combination of Gaussians, and hence will be a Gaussian itself.

$$\begin{aligned}
[AB]_{ij} &= \sum_{k=1}^{N} A_{ik}B_{kj} \qquad (5.4)\\
\left\langle [AB]_{ij} \right\rangle &= \sum_{k=1}^{N} \langle A_{ik} \rangle B_{kj} = 0\\
\left\langle [AB]_{ij}^2 \right\rangle &= \sum_{k=1}^{N}\sum_{l=1}^{N} \langle A_{ik}A_{il} \rangle B_{kj}B_{lj}\\
&= \sum_{k=1}^{N}\sum_{l=1}^{N} \delta_{kl}\sigma^2 B_{kj}B_{lj}\\
&= \sigma^2 \sum_{k=1}^{N} B_{kj}^2 = \sigma^2
\end{aligned}$$

where (5.4) follows because rows and columns of B have unit norm. So, the elements of the matrix AB are also Gaussians with zero mean and variance σ^2. Let us see if they are independent

$$\begin{aligned}
\left\langle [AB]_{ij} [AB]_{rs} \right\rangle &= \left\langle \sum_{k=1}^{N}\sum_{l=1}^{N} A_{ik}A_{rl}B_{kj}B_{ls} \right\rangle\\
&= \sum_{k=1}^{N}\sum_{l=1}^{N} \delta_{ir}\delta_{kl}\sigma^2 B_{kj}B_{ls}\\
&= \delta_{ir}\sigma^2 \sum_{k=1}^{N} B_{kj}B_{ks}\\
&= \delta_{ir}\delta_{js}\sigma^2
\end{aligned}$$

where, the first dirac delta comes from the fact that entries of A are uncorrelated, and the second one comes from the fact that the rows of B are orthonormal. Therefore entries of AB have the same joint distribution as entries A and it will also satisfy RIP with high probability!

6 Compressive sensing with noise

Now we consider the case when there is noise in the measurements.

It turns out that we can still get good reconstruction and the performance degradation is graceful. Let the matrix A satisfy RIP ($\delta_{2K} < \sqrt{2}-1$)– the distances between projections of K-sparse signals will be preserved– and let the measurements be noisy

$$\mathbf{y} \quad = \quad A\mathbf{x} + \mathbf{z} \quad \text{where } \left\langle \mathbf{z}_i^2 \right\rangle = \sigma^2$$

Then [REF Candes Noosy] showed that we can get a good reconstruction by solving a different optimization problem(LASSO)

$$\hat{\mathbf{x}} \quad = \quad \arg\min_{\mathbf{x}} \left[\|\mathbf{y} - A\mathbf{x}\|^2 + \lambda \sum_{i=1}^{N} |\mathbf{x}_i| \right]$$

For properly tuned λ the theorem states that

$$\|\hat{\mathbf{x}} - \mathbf{x}\|_2 \quad \leq \quad C_1 K\sigma + C_2 \frac{\|\mathbf{x} - \mathbf{x}_K\|_1}{\sqrt{K}} \tag{6.1}$$

The result cannot be any better. It states that the reconstruction error in the noisy case is bounded by the reconstruction error of the best possible reconstruction plus a term that scales linearly with the noise.

7 Random projections and stable embeddings

We already saw that random $M \times N$ matrices obey the RIP for sparse signals with high probability provided M is sufficiently large. Now we briefly discuss another surprising property of random projections : random projections also provide "stable" low dimensional embeddings [3]. If the signals $\mathbf{x}$ resides in a high dimensional space of ambient dimension N, but has some low-dimensional structure (like a manifold) of effective dimensionality K, then a random projection of the signal to a space with dimension $M \sim K$ will preserve local distance between the points in the high dimensional space. Thus, random projections can be used for dimensionality reduction in a *non-adaptive* way.

Let us look at one particular result with this flavour: the **Johnson-Lindenstrauss** lemma [9, 3]. The lemma shows that any K point set in an Euclidean space (say, $\mathbb{R}^N$) can be *linearly* embedded in a space of dimension $O\left(\log K/\epsilon^2\right)$ without distorting pairwise distances by more than $(1 \pm \epsilon)$. Specifically, consider a set of K points in $\mathbb{R}^N$, then the lemma shows that there exists a linear map $A : \mathbb{R}^N \to \mathbb{R}^M$ which preserves the pair-wise distances up to $(1 \pm \epsilon)$ for all the points u, v for $\epsilon \in (0, 1/2)$ and $M \sim \log K/\epsilon^2$

$$(1-\epsilon)\|u-v\|_2^2 \quad \leq \|Au - Av\|_2^2 \leq \quad (1+\epsilon)\|u-v\|_2^2$$

This result is tight – you cannot do the embedding into a substantially lower dimensional space without distorting the distances a lot [9]. Moreover, the linear map A can be constructed by populating the $M \times N$ matrix by i.i.d entries from the same distributions F that were suitable for RIP : $F \equiv \mathcal{N}\left(0, 1/\sqrt{M}\right)$ and $F \equiv \text{Bernoulli}\left(\pm 1/\sqrt{M}\right)$! The proof is not complicate [9] and uses the following ideas:

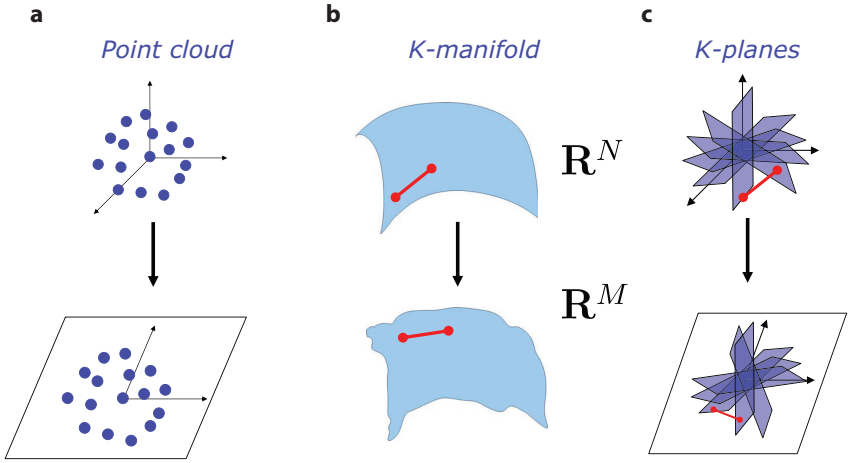

Figure 7.1: Random projections provide stable embeddings from $\mathbb{R}^N \to \mathbb{R}^M$ for a) point clouds b) $K-$manifold and c) $K-$planes (i.e $K-$sparse signals) provided M is comparable to K or $\log$ (# of points in cloud). Schematic adapted from [3]

- If $A_{ij} \sim \mathcal{N}(0,1)$, then using the **Hoeffding** concentration inequality[12] it's easy to show that the lengths of vectors are concentrated around the mean

$$\mathrm{P}\left(\frac{1}{\sqrt{M}}\|A\mathbf{x}\|_2^2 \geq (1+\epsilon)\|\mathbf{x}\|_2^2\right) \quad \leq \quad \exp\left(-\frac{M}{4}\left(\epsilon^2-\epsilon^3\right)\right)$$

- From which we see that probability of *one* pair of distances getting distorted is exponentially small in M

$$\mathrm{P}\left((1-\epsilon)\|\mathbf{x}\|_2^2 \leq \|A\mathbf{x}\|_2^2 \leq (1+\epsilon)\|\mathbf{x}\|_2^2\right) \quad \geq \quad 1-2e^{-M\left(\epsilon^2-\epsilon^3\right)/4}$$

- There are $O\left(K^2\right)$ pairs, so use the union bound to show that the RHS holds for all pairs with non-zero probability provided $M \sim O\left(\log K/\epsilon^2\right)$.

A similar covering argument like the one used to show that points of clouds can be stably embedded in a lower dimension, can be extended to show that signals lying on other low-dimensional structures like manifolds or union of planes (sparsity) can be stably embedded by a random projection in a space with dimensionality M comparable to the effective dimensionality of the low dimensional structure [3]. This is illustrated in the schematic in Fig. []. Thus, random projections give us a way to do proximity-preserving dimensionality reduction in a non-adaptive way! This is very useful for practical applications, because a lot of natural signals like images or sound have a sparsity or smooth manifold structure, so we can first project them randomly (and non-adaptively) to a lower dimension and then perform computational tasks such as clustering or learning in the lower dimensional representation. This paradigm suggests a provocative and counterintuitive strategy for the brain – represent structured stimuli by means of random receptive fields!

References

[1] Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, 2008.

[2] Richard G Baraniuk. Compressive sensing. *IEEE signal processing magazine*, 24(4), 2007.

[3] Richard G Baraniuk, Volkan Cevher, and Michael B Wakin. Low-dimensional models for dimensionality reduction and signal recovery: A geometric perspective. *Proceedings of the IEEE*, 98(6):959–971, 2010.

[4] Emmanuel J Candes and Yaniv Plan. A probabilistic and ripless theory of compressed sensing. *IEEE Transactions on Information Theory*, 57(11):7235–7254, 2011.

[5] Emmanuel J Candès, Yaniv Plan, et al. Near-ideal model selection by â1 minimization. *The Annals of Statistics*, 37(5A):2145–2177, 2009.

[6] Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.

[7] Emmanuel J Candes and Terence Tao. Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215, 2005.

[8] Emmanuel J Candes and Terence Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE transactions on information theory*, 52(12):5406–5425, 2006.

[9] Sanjoy Dasgupta and Anupam Gupta. An elementary proof of the johnson-lindenstrauss lemma. *International Computer Science Institute, Technical Report*, pages 99–006, 1999.

[10] David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

[11] David L Donoho and Jared Tanner. Sparse nonnegative solution of underdetermined linear equations by linear programming. *Proceedings of the National Academy of Sciences of the United States of America*, 102(27):9446–9451, 2005.

[12] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

[13] Andreas M Tillmann and Marc E Pfetsch. The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing. *IEEE Transactions on Information Theory*, 60(2):1248–1259, 2014.