Lecture Notes - Compressive Sensing - INSA

claude.petit@inria.fr

December 2022

These lecture notes supplement the slides of the course and can make references to them.

1 What is compressive sensing ? Why is it useful ?

1.1 Film and digital camera

In the traditional digital signal processing approach, one starts by acquiring (receiving, measuring, calculating, sampling) data before compressing them. The acquisition step is often done from analog data which are transformed into digital data (via sampling and quantization). It is then necessary to compress data in order to be able to transmit them on a network or to store them in digital memories, for efficiency, time sparing and economy of size. Compression can only be done after the acquisition process.

Compressive sensing is a technique for sampling and compressing data at the same time.

Let's start with an example to concretely illustrate the process of sampling, analog / digital conversion and compression: let's consider a film camera, with which we want to take a picture, say, of the Eiffel Tower. The camera lenses are directed towards the object to be photographed and the light passes through them until it hits the photo film. The silver salts deposited on the plastic wrap are impressed by the light and create a physical image on this film. Suppose for simplicity that this image is a square of size 1x1 cm (a typical real size is 24x36 mm). By enlarging the negative, one can obtain photos of several square meters while keeping the same level of details - the same sharpness - as the image contained in these few square millimeters. We can consider that the number of points on the image is practically infinite and mathematically model the image by a continuous function

$$x_R : [01] \times [0,1] \longrightarrow \mathbb{R} \text{ ou } \mathbb{R}^3 \qquad (1)$$

This function is analog because the data are practically continuous. The function takes its values in $\mathbb{R}$, if the filmstrip is in black and white, or $\mathbb{R}^3$ if the filmstrip is in color, in order to obtain, for each point of the image, either the luminosity, either the three values of the components of each color. The value $x(t_1, t_2)$ at point (t_1, t_2) represents the luminosity of the image at this point (we will consider from now on that the image is in black and white). This value varies continuously between 0 (if the point is completely black) and say 1 if it is white.

In a digital camera, the plastic film is replaced by an array of photosensitive sensors, each of which captures the brightness of the rays it receives. There is a fixed number $N_1 \times N_2$ of sensors and the image is then a matrix of size $N_1 \times N_2$ whose number of points (these are pixels) is finite and fixed. Likewise, the value of the luminosity in each pixel must be quantized. The image is then a discrete object that can be mathematically represented by a matrix or a vector:

$$x : \{1, N_1\} \times \{1, N_2\} \longrightarrow \{0, 255\} \text{ ou } \{0, 255\}^3 \qquad (2)$$

The set $\{0.255\}$ is an example of the number of different values that one can assign to the luminuosity ($256 = 2^8$ for a luminosity coded on 8 bits). For current 4K UHD formats, the standard resolution is 3840×2160 pixels, with 12 bits for color and 120 frames per second. The size of the files quickly becomes gigantic and it is not possible to use them without having first compressed them.

The compression step consists of two processes: the first reduces the size, by various methods (we will see one of them) and the other encodes the signal using an entropy encoder. The resulting file (for example a JPEG image) can then be used, transmitted, stored or copied more easily.

When we want to use a compressed file, say an image, it is necessary to decompress it via a new transformation. We then get a file $\hat{x}$ that we want to be identical to the initial signal x.

When $x = \hat{x}$, we say that the compression method is "lossless" and when the two signals are not the same, we say that the method is "lossy".

In conclusion of this introduction, the digital signal processing process requires a sampling step, then a compression step. Compressive sensing tries to perform both operations at the same time.

1.2 Sampling and compression: two fundamental questions

1. Is it possible to perfectly reconstruct a continuous signal (i.e. containing an infinity of values) from a sample of a few measurements of that signal (i.e. a few values of the signal measured at given moments)

2. Is it possible to reduce the size of a discrete image without altering it?

1.2.1 The Shannon-Nyquist sampling theorem

The answer to question 1. is yes, but ... There are two conditions needed to reconstruct perfectly a function from a certain number of its samples: on the one hand, the signal must be restricted to a particular class of signals (signals with limited spectrum width) and on the other hand have enough samples: the sampling rate must be twice the highest frequency.

There exists lots of different Fourier transforms for differents types of signals: Fourier series for periodic functions, usual Fourier transform for functions of class L^2 or L^1 on $\mathbb{R}^n$, FFT or DCT for finite signals, Walsh transform for binary vectors, Fourier transforms on a graph, on finite fields, groups, commutative or not, wavelets, Gelfand transform, etc.

The Fourier transform of a function $x \in L^2(\mathbb{R})$ is

$$\hat{x}(u) = \int_{\mathbb{R}} x(t) e^{-2\pi i u t} dt \tag{3}$$

It is an isometry on $L^2(\mathbb{R})$. The calculus of the inverse Fourier transform makes it possible to recover the initial signal from its transform:

$$x(t) = \int_{\mathbb{R}} \hat{x}(u) e^{2\pi i u t} du \tag{4}$$

Now, if the spectrum is included in $[-B, B]$, then

$$x(t) = \int_{-B}^{B} \hat{x}(u) e^{2\pi i u t} du \tag{5}$$

We will see that if we restrict ourselves to the space $L^2[-B, B]$ of square integrable signals whose spectrum is bounded, of maximum frequency B, then we can reconstruct the initial signal as long as the sampling is carried out with a period less than or equal to $1/2B$.

Recall that the Fourier transform of a gate function is a cardinal sine:

$$\Pi(t) = \mathbb{1}_{\left[-\frac{T}{2}, \frac{T}{2}\right]}(t) \Rightarrow \tag{6}$$

$$\hat{\Pi}(u) = T \frac{\sin(\pi u T)}{\pi u T} = T \mathrm{sinc}(\pi u T) \tag{7}$$

The Dirac mass in 0 is the distribution defined by:

$$\delta(t) = \begin{cases} 0 & \text{if } t \neq 0 \\ +\infty & \text{if } t = 0 \end{cases} \tag{8}$$

and by the property:

$$\int_{\mathbb{R}} \delta(t) dt = 1 \tag{9}$$

Since δ is zero almost everywhere, its integral (Lebesgue or Riemann) should be zero. δ therefore cannot be a function. We can calculate its Fourier transform in the sense of distributions (all the following calculations are done in the sense of distributions):

$$\hat{\delta}(u) = 1 \tag{10}$$

The Dirac mass in 0 is the neutral element for the convolution product:

$$\forall x,\ x \star \delta = x \tag{11}$$

The Dirac mass in n is the translated version of the Dirac mass in 0:

$$\delta_n(t) = \delta(t - n) \tag{12}$$

The convolution of a signal by a Dirac mass in n is the translated signal by a n factor:

$$x \star \delta_n(t) = x(t - n) \tag{13}$$

Let III_T the Dirac comb with period T. It is defined by

$$\mathrm{III}_T(t) = \sum_{n \in \mathbb{Z}} \delta(t - nT) \tag{14}$$

Remember that the Fourier transform (in the sense of distributions) of a Dirac comb III_T is a Dirac comb with period $1/T$ (up to a multiplicative factor):

$$\widehat{\mathrm{III}}_T(u) = \frac{1}{T} \mathrm{III}_{1/T}(u) \tag{15}$$

Multiplying a signal by a Dirac comb si equivalent to sample this signal at period T:

$$x_e(t) = x(t) \times \mathrm{III}_T(t) \tag{16}$$

$$= x(t) \times \sum_{n \in \mathbb{Z}} \delta(t - nT) \tag{17}$$

$$= \sum_{n \in \mathbb{Z}} x(t) \delta(t - nT) \tag{18}$$

$$= \sum_{n \in \mathbb{Z}} x(nT) \delta(t - nT) \tag{19}$$

The passage from the penultimate to the last line comes from the fact that the Dirac comb is zero everywhere except at the points nT. Here again, this

calculation is only valid in the sense of distributions. The result is a zero distribution almost everywhere except at the points defined by the Dirac masses. It is therefore a sequence $(x(nT))_{n\in\mathbb{Z}}$ of real numbers made up of a sample of measures from x. This property is at the heart of the proof of the sampling theorem.

Eventually, the Fourier transform permutes the convolution product with the classical product:

$$\widehat{x \star y}(u) = \widehat{x}(u) \times \widehat{y}(u) \tag{20}$$

$$\widehat{x \times y}(u) = \widehat{x} \star \widehat{y}(u) \tag{21}$$

All this being stated, we can now prove Shannon's theorem:

Let $x \in L^2(\mathbb{R})$ be a square integrable signal. The sampled T-period signal is :

$$x_e(t) = x(t) \times \mathrm{III}_T(t) \tag{22}$$

Its Fourier transform is

$$\hat{x}_e(u) = \hat{x} \star \frac{1}{T}\,\mathrm{III}_{1/T}(u) \tag{23}$$

$$= \frac{1}{T}\sum_{n\in\mathbb{Z}} \hat{x} \star \delta(u - n/T) \tag{24}$$

$$= \frac{1}{T}\sum_{n\in\mathbb{Z}} \hat{x}(u - n/T) \tag{25}$$

In words, the spectrum of the sampled signal is made of an infinity of copies from the spectrum of x, each translated by a factor $1/T$.

If the signal is bandlimited in $[-B, B]$ and if

$$T \leq \frac{1}{2B} \tag{26}$$

then the different copies are of disjoint supports and by a frequency windowing (ie by multiplying by a gate function), we can recover a single copy of the initial spectrum. Consider for this the function

$$\hat{h}(u) = T\mathbb{1}_{\left[-\frac{1}{2T},\frac{1}{2T}\right]}(u) = T\mathbb{1}_{\left[-\frac{1}{2},\frac{1}{2}\right]}(Tu) \tag{27}$$

it is the Fourier transform of the cardinal sine

$$h(t) = T\frac{1}{T}\mathrm{sinc}(\pi t/T) = \frac{\sin(\pi t/T)}{\pi t/T} \tag{28}$$

To window $\hat{x}(u)$, we multiply it by $\hat{h}(u)$ and the equality below is licit as long as the Shannon condition is satisfied:

$$\hat{x}_e(u) \times \hat{h}(u) = \frac{1}{T}\sum_{n\in\mathbb{Z}} \hat{x}(u - n/T) \times T\mathbb{1}_{\left[-\frac{1}{2T},\frac{1}{2T}\right]}(u) \tag{29}$$

$$= \hat{x}(u) \times \mathbb{1}_{\left[-\frac{1}{2T},\frac{1}{2T}\right]}(u) \tag{30}$$

thus,

$$\hat{x}(u) = \hat{x}_e(u) \times \hat{h}(u) = \widehat{x_e \star h}(u) \tag{31}$$

and by inverse Fourier transform, one has

$$x(t) = x_e \star h(t) \tag{32}$$

Replace x_e by its expression:

$$x(t) = x_e \star h(t) \tag{33}$$

$$= \left(\sum_{n\in\mathbb{Z}} x(nT)\delta_{nT}\right) \star h(t) \tag{34}$$

$$= \sum_{n\in\mathbb{Z}} x(nT)\left[\delta_{nT} \star h(t)\right] \tag{35}$$

$$= \sum_{n\in\mathbb{Z}} x(nT)h(t - nT) \tag{36}$$

$$= \sum_{n\in\mathbb{Z}} x(nT)\mathrm{sinc}\left(\frac{\pi}{T}(t - nT)\right) \tag{37}$$

When $T = 1/2B$ it comes

$$x(t) = \sum_{n\in\mathbb{Z}} x(\frac{n}{2B})\mathrm{sinc}\left(2\pi Bt - n\pi\right) \tag{38}$$

We have succeeded in reconstructing the signal $x(t)$ from the sequence formed of its samples $(x(nT))_n$. The formula also indicates that the family $(\mathrm{sinc}(\pi(t - nT)/T))_{n\in\mathbb{Z}}$ is an orthonormal basis of the space of square integrable signals whose spectrum is included in $[-B, B]$.

Shannon's theorem can be rigorously demonstrated without using distributions (by the Poisson formula), but the above demonstration has the interest of clearly showing the links between the time domain signal, the sampled signal and their Fourier transforms.

When the condition on the sampling frequency is not respected (in case of downsampling, below the optimal rate), the copies of the spectrum of $x(t)$ are no longer disjoint and one speaks of aliasing effect. It is then no longer possible to isolate the spectrum of x and to reconstruct the time signal. The downsampling results in a "moiré" effect in the images: periodic artefacts appear in areas, similar to what can be observed by looking at two grids, one behind the other. There also exists a problem of identifying the initial signal: a too small number of interpolation points can be verified by several different signals. This is called signal ambiguity.

The sampling rate of a compact-disc is 44 kHz with 16 bits of quantization. It has been chosen to correspond to a little bit more than twice the maximal frequency that a human can hear (20 Hz - 20 kHz for most people). Hi-Res audio proposes audio files with

96 or 192 kHz and 24 bits of quantization. Most of the people can't hear any difference between a usual CD file and a Hi-Res audio file.

What should be retained from this paragraph are the two important conditions for the reconstruction: the fact that the signals must be bandlimited and the fact that the sampling must be carried out at a frequency higher than twice the maximum frequency of the signal.

1.2.2 Discrete cosine transform compression method

We now answer question 2. There are many different methods of compressing signals. The one presented here, suitable for video, was used in the JPEG format and uses sparsity of video signals.

DCT compression (discrete cosine transform) in dimension 2 is carried out by splitting the image into blocks of $N_1 \times N_2$ pixels (called tiles) and by expressing each of the blocks as a linear combination of suitably chosen tiles.

Each tile is a matrix of real coefficients of size $N_1 \times N_2$. We can therefore consider it as a real vector of size $N_1 \times N_2$ and choose a basis of the vector space $\mathbb{R}^{N_1 \times N_2}$ to represent them. In what follows, we will therefore keep in mind that each image block is an element of a real vector space.

For the DCT, the $N_1 \times N_2$ chosen basis vectors are cosine products. In the FFT (fast Fourier transform) the functions of the basis are complex exponentials, which can be broken down into sines and cosines; then sines are transformed into cosines. It is thus possible to construct an orthonormal basis of the space of discrete signals of size $N_1 \times N_2$ only using cosines. We can prove that the $N_1 \times N_2$ functions

$$\phi_{n_1,n_2}(k_1,k_2) = ... \tag{39}$$

$$... \cos\left[\frac{(2n_1+1)\pi}{2N_1}k_1\right]\cos\left[\frac{(2n_2+1)\pi}{2N_2}k_2\right] \tag{40}$$

form an orthogonal basis of $\mathbb{R}^{N_1 \times N_2}$. By multiplying by an adequate constant, this basis can be made orthonormal (this is what we will assume, while forgetting the constants). If the pixels of a digital image are given by the vector $x = (x_{n_1,n_2})$, the coefficients of its DCT transform are

$$c_{k_1,k_2} = \sum_{n_1=0}^{N_1-1}\sum_{n_2=0}^{N_2-1} x_{n_1,n_2} \times \phi_{n_1,n_2}(k_1,k_2) \tag{41}$$

Low frequencies (corresponding to areas of small variations of luminuosity) are stored in the upper left corner of the DCT image and high frequencies (corresponding to high changes of luminuosity like near edges or textures) are stored in the lower right corner. The image is read in diagonal from low to high frequencies. It must be noticed that most of the real world images are sparse in some way.

DCT compression is better than discrete Fourier transform, because the DCT is the mirrorerd function (symetrized) of the DFT. This eliminates the Gibbs effect visible on some DFT, near the discontinuity points of the signal. The Gibbs effect arises when developping a periodic function which is not continuous in every point. It causes high contrast around the edges in digital images or echoes in MP3 files.

There exists other methods of compression like Zipf's algorithm (used in ZIP files), which are less efficient because they do not exploit sparsity of the initial signals.

For audio files, the FLAC (Free Lossless Audio Codec) specification format describes a lossless compression algorithm that is mostly still used more than twenty years after its discovery.

2 How to formalize compressive sensing ?

2.1 Sparse signals

A vector x is k-sparse if it has at most k non-zero coordinates (and we obviously assume that k is very small compared to n). The degree of sparsity of x is measured with the function

$$||x||_0 = \text{card}(\text{supp}(x)) \tag{42}$$

$||x||_0$ is equal to the number of non-zero coordinates of x. It is not a mathematical norm (why?). The set of signals k-Sparse is denoted by Σ_k. It is not a vector space (why?).

Sparsity depends on the basis in which the vector is given (cf. Lena's DCT).

Most real-world signals are not exactly sparse, but have coefficients close to zero. We have to consider compressible vectors, whose coordinates tend rapidly towards 0 and which can be approached by sparse vectors. This leads to the following definitions:

The best s-sparse approximation of a vector x relatively to the l_p-norm is the vector $\sigma_k(x)_p$ given by

$$\sigma_k(x)_p = \min_{y \in \Sigma_k} ||x - y||_p \tag{43}$$

A simple way to get this approximation is to threshold to 0 all the $n-k$ smallest coefficients (in absolute value) of x.

A signal is said to be compressible if its coordinates x_i ranged in decreasing order satisfy

$$\exists\ C, q > 0 :\ |x_i| \leq \frac{C}{i^q}\ \forall i = 1, ..., n \tag{44}$$

We denote by x_S a vector x whose coordinates indexed elsewhere than in S have been set to 0. We will denote by M_S a matrix whose columns indexed elsewhere than in S have been fixed at 0. We have in particular

$$M = M_S + M_{\overline{S}} \tag{45}$$

and

$$M_S x = M_S x_S = M x_S \tag{46}$$

Forall $x \in \mathbb{R}^n$.

2.2 Mathematical model for compressive sensing

In the traditional approach of digital signal processing, a lot of data are sampled and then discarded at the compression step. Is it possible to sample only the data that will be useful ? This is the main objective of compressive sensing.

Mathematically the problem is very simple: we have a vector of measurements

$$y = Mx \tag{47}$$

formed by linear combinations of a vector x by a known sensing matrix M. We would like to find x given y and M, by exploiting the fact that x is s-sparse.

There are many phenomena and operations in signal processing that can be modelized by a matrix product: all classical transformations on finite signals (Fourier, wavelets, etc.) are linear. The same goes for a lot of filtering, coding operations, etc.

Solving (47) is in fact the same as solving a simple linear system. We can summarize this in the form of the program P_0:

$$P_0 :\ \hat{x} = \underset{z:Mz=y}{\operatorname{argmin}} ||z||_0 \tag{48}$$

In fact, there are three (big) issues:

1. The above program is not convex because of the l_0 norm. It cannot therefore be solved by conventional optimization techniques.

2. It is a NP-hard problem: solving it amounts to testing all the vectors whose support is of cardinality k. There are $\binom{n}{k}$ of them, which grows exponentialy with n.

3. The linear system is underdetermined and therefore admits an infinity of solutions.

There are roughly three different philosophies for tackling the problem in a practical way, which results in algorithms of three different types: convex relaxation algorithms, greedy or thresholding algorithms.

A (false) problem to consider is that of signals which are not sparse in the current basis, but in another basis to be determined. The study of the algorithm shows that the problem does not change and that it is not necessary to know the basis in which the signal is sparse, in order to to find the solution.

Compressive sensing is often presented as part of the sparse representation methods, which is true, but there are key differences between classic sparse methods and compressing sensing. In the former, the encoding is non-linear (it is used to determine a dictionary, depending on the signal, in which the latter is sparse) and the decoding is linear. In compressing sensing, the encoding is linear (Mx is calculated) while the decoding is non-linear (one of the families of algorithms mentioned above is applied).

Both families seek to minimize the difference between the initial signal and its estimate with equivalent decomposition, but different criteria. In a way, sparsity representations try to determine a local sparse solution, while compressive sensing determines the sparsiest solution.

At this stage we have just posed the mathematical problem that interests us: to solve an underdetermined system by exploiting the sparsity of the solution. We must now answer two questions:

1. Which matrices should be used as sensing matrices?

2. What are the methods and algorithms that concretely make it possible to find x from y?

2.3 Which sensing matrix to choose and the RIP property

This section presents an interesting type of matrix to solve the P_0 problem and a simple method to obtain such matrices.

Let $\epsilon > 0$, and let k a given integer. A matrix M satisfy the (ϵ, k)-RIP property if

$$\forall x \in \Sigma_k,\ (1-\epsilon)||x||_2^2 \leq ||Mx||_2^2 \leq (1+\epsilon)||x||_2^2 \tag{49}$$

In words, M doesn't change two much the norm of each k-sparse vector x. It should be clear that

$$(\epsilon, k) - \text{RIP} \iff \forall x \in \Sigma_k,\ \left|\frac{||Mx||_2^2 - ||x||_2^2}{||x||_2^2}\right| \leq \epsilon \tag{50}$$

We recall that the spectral norm of a matrix is defined par any of the following formula:

$$||M||_2 = \sup_{x\neq 0} \frac{||Mx||_2}{||x||_2} \quad (51)$$

$$= \sup_{||x||_2=1} ||Mx||_2 \quad (52)$$

$$= \sup_{||x||_2\leq 1} ||Mx||_2 \quad (53)$$

$$= \sqrt{\rho(M'M)} \quad (54)$$

$$= \sigma_{\max}(M) \quad (55)$$

where ρ is the spectral radius (absolute value of the largest eigenvalue) and $\sigma_{\max}$ the largest singular value. In the case of a hermitian or symmetric matrix, the spectral norm can be defined by

$$||A||_2 = \sup_{x\neq 0} \frac{\langle Ax, x\rangle}{||x||^2} = \rho(A) \quad (56)$$

This is the expression of the Rayleigh ratio, which is maximised by the eigenvector relative to the largest eigenvalue and in that case, the Rayleigh ratio is equal to the spectral radius.

Remember that for every non-zero vector x and for every subset S with less than k elements,

$$M_S x = M_S x_S = M x_S \quad (57)$$

Therefore, for all non-zero x of $\mathbb{R}^n$ (sparse or not),

$$||M_S x||_2^2 - ||x||_2^2 = \langle M_S x, M_S x\rangle - \langle x, x\rangle \quad (58)$$

$$= \langle M_S' M_S x, x\rangle - \langle x, x\rangle \quad (59)$$

$$= \langle (M_S' M_S - I)x, x\rangle \quad (60)$$

Thus,

$$(\epsilon, k) - \text{RIP} \Rightarrow \forall x \neq 0, \forall S, \ \frac{\langle (M_S' M_S - I)x, x\rangle}{||x||_2^2} \leq \epsilon \quad (61)$$

$$\Rightarrow \max_{x\neq 0} \frac{\langle (M_S' M_S - I)x, x\rangle}{||x||_2^2} \leq \epsilon \quad (62)$$

Now, since $M_S' M_S - I$ is symmetric, the above expression is exactly the definition of the spectral norm.

$$\Rightarrow ||M_S' M_S - I||_2 \leq \epsilon \quad (63)$$

In words, for any subset S of cardinality smaller than the sparsity level k, the extracted matrix M_S behaves like an orthogonal isometry when it is applied to the support vectors S. Beware of the term "extracted" because M_S has the same size as M.

We can notice that it is necessary to impose a RIP property with parameters $(\epsilon, 2k)$ if we want two k-sparse vectors to have distinct images by M (see Theorem 2.13 in Foucart & Rauhut regarding the NSP property (null space property). In fact, we then have

$$||M(x-y)||_2^2 > 0 \quad (64)$$

Remember that if $x, y \in \Sigma_k$ then $(x-y) \in \Sigma_{2k}$ (take two disjoint supports).

Why is a sensing matrix with the RIP property of interest in compressive sensing ? Recall that our goal is to solve the following problem:

$$P_0 : \ \underset{z:Mz=y}{\operatorname{argmin}} ||z||_0 \quad (65)$$

In the previous lines, we proved in fact that

If $M \in \mathbb{R}^{m\times n}$ and M is $(\epsilon, 2k)$-RIP, then if $x \in \Sigma_k$ is solution, it is the only solution of P_0. The NSP property imposes anyway that the minimum number of measurements m to retrieve all the k-sparse signals is $m \geq 2k$.

2.4 IHT: a sparse signal recovery algorithm (at last)

The IHT algorithm (for Iterative Hard Tresholding) iteratively calculates a solution to the P_0 problem. At each iteration t, it performs a hard thresholding on a sparse candidate vector x^t, i.e. it sets to 0 all the coordinates below a threshold value, keeping only the highest s coordinates in absolute value.

The function H_k is define as a function from $\mathbb{R}^n$ onto $\mathbb{R}^n$ which changes a vector x into a k-sparse x_S where all the coordinates except the k largest have been set to zero.

$$\text{IHT} : \begin{cases} x(0) = 0 \\ x^{t+1} = H_k\left(x^t + M'(y - Mx^t)\right) \end{cases} \quad (66)$$

The exit of the algorithm is

$$\hat{x} = \lim_{t\to+\infty} x^t \quad (67)$$

x^t is equal to $H_k\left((I - M'M)x^t + M'y\right)$ and $y = Mx$,

$$y - Mx^t = M(x - x^t) \quad (68)$$

The thresholding step if of the form

$$x^{t+1} = H_k\left(x^t + \text{erreur}(x - x^t)\right) \quad (69)$$

The following Theorem shows the importance of the RIP property for the success of IHT:

Theorem 1 *Let $M \in \mathbb{R}^{m\times n}$ with $m << n$ and let $\epsilon > 0$. If M satisfies $(\epsilon, 3k)$-RIP, then*

$$||x^{t+1} - x|| \leq 2\epsilon ||x^t - x|| \quad (70)$$

Specifically, if $\epsilon < 1/2$ then $\hat{x} = x$

Proof: let

$$u^t = x^t + M'(y - Mx^t) \tag{71}$$
$$= x^t + M'M(x - x^t) \tag{72}$$

so that

$$x^{t+1} = H_k(u^t) \tag{73}$$

For all $x \in \Sigma_k$, since x^{t+1} is by construction the best k-sparse approximation of u^t,

$$x^{t+1} = \arg\min_{x \in \Sigma_k} ||u^t - x||^2 \tag{74}$$

so that, for all x,

$$||u^t - x^{t+1}||^2 \leq ||u^t - x||^2 \tag{75}$$
$$||u^t - x^{t+1}||^2 \leq ||u^t - x^t||^2 \tag{76}$$

Now,

$$||u^t - x^{t+1}||^2 = ||u^t - x||^2 + ||x^{t+1} - x||^2 ... \tag{77}$$
$$... - 2\left\langle u^t - x, x^{t+1} - x \right\rangle \tag{78}$$
$$\leq ||u^t - x||^2 \tag{79}$$

By reordering both members of the above inequality,

$$||x^{t+1} - x||^2 \leq 2\left\langle u^t - x, x^{t+1} - x \right\rangle \tag{80}$$
$$= 2\left\langle (I - M'M)(x^t - x), x^{t+1} - x \right\rangle \tag{81}$$
$$= 2\ \langle \bullet \rangle \tag{82}$$

We have now to prove that

$$\langle \bullet \rangle \leq \epsilon . ||x^t - x|| . ||x^{t+1} - x|| \tag{83}$$

Let

$$u = x^t - x \tag{84}$$
$$v = x^{t+1} - x \tag{85}$$
$$T = \text{supp}(u) \cup \text{supp}(v) \tag{86}$$

so that

$$T \subset \left(\text{supp}(x^t) \cup \text{supp}(x) \cup \text{supp}(x^{t+1})\right) \tag{87}$$

This proves that $\text{card}(T) \leq 3k$

$$\langle (I - M'M)u, v \rangle = u'_T(I - M'_T M_T)v_T \tag{88}$$
$$\leq ||(I - M'_T M_T)u_T||_2 ||v_T||_2 \tag{89}$$
$$\leq ||(I - M'_T M_T)||_2 ||u_T||_2 ||v_T||_2 \tag{90}$$
$$\leq \epsilon ||u_T||_2 ||v_T||_2 \tag{91}$$

By successively using the Cauchy-Schwarz inequality and the properties of the operator norm, one deduces

$$||x^{t+1} - x||^2 \leq (2\epsilon).||x^t - x||.||x^{t+1} - x|| \tag{92}$$
$$\iff ||x^{t+1} - x|| \leq (2\epsilon)||x^t - x|| \tag{93}$$
$$\Rightarrow ||x^t - x|| \leq (2\epsilon)^t ||x|| \tag{94}$$

The last inequality, for $\epsilon < 1/2$ proves that the application is a contractant one and that

$$\lim_{t \to +\infty} x^t = x \tag{95}$$

2.5 Gaussian concentration, RIP and Johnson-Lindenstrauss Lemma

2.5.1 Concentration inequality

The question that comes immediately at the end of the previous section is how to build a sensing matrix that has the RIP property. The answer is very simple: at random!

The measure concentration phenomenom for Gaussian vectors makes it possible to demonstrate the following theorem:

Theorem 2 (CI: concentration inequality) *Let $M \in \mathbb{R}^{m \times n}$ a matrix whose coefficients are i.i.d. Gaussian random variables $\mathcal{N}(0, 1/m)$. Then, $\forall x \in \mathbb{R}^n,\ \forall \epsilon \in]0, 1[$,*

$$\mathbb{P}\left[\left|||Mx||_2^2 - ||x||_2^2\right| > \epsilon ||x||_2^2\right] \leq 2e^{-m\epsilon^2/12} \tag{96}$$

Let $\phi_X(t) = \mathbb{E}[e^{tX}]$ the moment generating function of a random variable X.

We recall the expression of $\phi_X(t)$ for a centered Gaussian random variable $X \sim \mathcal{N}(0, \sigma^2)$ and for a chi square distribution:

$$\phi_X(t) = e^{\sigma^2 t^2/2} \tag{97}$$
$$\phi_{X^2}(t) = \frac{1}{\sqrt{1 - 2t\sigma^2}} \tag{98}$$
$$\tag{99}$$

The following proof will make use (without saying it) of the Cramer transform. It is easy to see that

$$\mathbb{E}\left[||Mx||_2^2\right] = ||x||_2^2 \tag{100}$$

Let $y = Mx$, $\gamma = ||x||_2^2/m$, and $z_i = y_i^2 - \gamma$ so that $y_i \sim \mathcal{N}(0, \gamma)$ and

$$\gamma = \mathbb{V}(y_i) = \mathbb{E}\left[y_i^2\right] = ||x||_2^2/m \tag{101}$$

then

$$||Mx||_2^2 - ||x||_2^2 = \sum_{i=1}^{m} z_i = S_m \tag{102}$$

Let $A = A(M, x, \epsilon)$ the event inside the probability

$$\mathbb{P}\left[\left|||Mx||_2^2 - ||x||_2^2\right| > \epsilon||x||_2^2\right] \tag{103}$$
$$= \mathbb{P}(A) \tag{104}$$
$$= \mathbb{P}\left[S_m > \epsilon m\gamma\right] + \mathbb{P}\left[S_m < -\epsilon m\gamma\right] \tag{105}$$

Using Markov inequality,

$$\mathbb{P}\left[S_m > \epsilon m\gamma\right] \leq \frac{\mathbb{E}\left[e^{tS_m}\right]}{e^{t\epsilon m\gamma}} \tag{106}$$

and by i.i.d. nature of the z_i

$$\mathbb{P}\left[S_m > \epsilon m\gamma\right] \leq \phi_{z_i}(t)^m e^{-mt\epsilon\gamma} = p(t)^m \tag{107}$$

with $p(t) = \phi_{z_1}(t)\exp(-t\epsilon\gamma)$.

Since y_i is Gaussian, $y_i^2/\gamma \sim \chi_1^2$ and $z_i/\gamma + 1$ is a chi square distribution, so that

$$p(t) = \mathbb{E}\left[\exp\left(ty_i^2 - t\gamma\right)\right] e^{-t\epsilon\gamma} \tag{108}$$
$$= \frac{e^{-t(1+\epsilon)\gamma}}{\sqrt{1-2\gamma t}} \tag{109}$$

An easy calculus (do it !) shows that the minimum of $p(t)$ in $\mathbb{R}$ is reached in

$$t^* = \frac{1}{2\gamma}\frac{\epsilon}{1+\epsilon} \tag{110}$$

Then

$$\ln p(t^*) = \frac{1}{2}\ln(1+\epsilon) - \epsilon/2 \tag{111}$$
$$\leq \frac{1}{2}(\epsilon - \epsilon^2/2 + \epsilon^3/3) - \epsilon/2 \tag{112}$$
$$\leq -\epsilon^2/4 + \epsilon^3/6 \tag{113}$$

and

$$\mathbb{P}\left[S_m > \epsilon m\gamma\right] \leq \exp\left(-m\left(\epsilon^2/4 - \epsilon^3/6\right)\right) \tag{114}$$

$$\mathbb{P}\left[S_m < -\epsilon m\gamma\right] \leq \exp\left(-m\left(\epsilon^2/4 - \epsilon^3/6\right)\right) \tag{115}$$

Eventually,

$$\mathbb{P}(A) \leq 2\exp\left(-m\left(\epsilon^2/4 - \epsilon^3/6\right)\right) \tag{116}$$

If $0 < \epsilon < 1$, it is clear that $\epsilon^2/4 - \epsilon^3/6 < \epsilon^2/12$ which gives a less optimal constant but a simpler formula. □

$\mathbb{P}(A)$ tends rapidly to zero when m tends to infinity.

2.5.2 The Johnson-Lindenstrauss Lemma

(96) says that, for any fixed $x \in \mathbb{R}^n$, when choosing a Gaussian matrix M at random, the probability of the event $A = A(M, x, \epsilon)$ is close to 1 for m high enough. But M depends on x.

The RIP-(ϵ, k) property for M says that for any $x \in \Sigma_k$, $A(M, x, \epsilon)$ is realized. For the RIP, a unique matrix M should satisfies the inequality for all sparse vectors, so that M shouldn't depend on x.

The link between the two properties is given by the Johnson-Lindenstrauss Lemma, a geometric theorem which says that a small number of points relatively to the dimension of the space can be projected in a smaller space without altering the distances between the points.

One possible and convenient form of the Lemma is the following:

Theorem 3 (J.L. Lemma) *Let $M \in \mathbb{R}^{m\times n}$ a matrix whose coefficients are i.i.d. Gaussian random variables $\mathcal{N}(0, 1/m)$. Let $\epsilon \in]0, 1[$, let $\delta > 0$, let Q a finite set of vectors of cardinality $|Q|$. If m satisfies:*

$$m > \frac{12}{\epsilon^2}\ln\left(\frac{2|Q|}{\delta}\right) \tag{117}$$

Then

$$\mathbb{P}\left[\sup_{x\in Q}\left|\frac{||Mx||_2^2}{||x||_2^2} - 1\right| \leq \epsilon\right] \geq 1 - \delta \tag{118}$$

Let us note RIP(ϵ, k) the event « the matrix M satisfies the RIP proprerty of order (ϵ, k) ». Then,

$$\text{RIP}(\epsilon, k) = \bigcap_{x\in\Sigma_k} A(M, x, \epsilon) \tag{119}$$

But this intersection is infinite and uncountable.

$$\mathbb{P}_M\left(\text{RIP}(\epsilon, k)\right) = \mathbb{P}_M\Big[\forall x \in \Sigma_k : A(M, x, \epsilon)\Big] \tag{120}$$

Proof of J.L. Lemma: let

$$B(M, \epsilon) = \left[\sup_{x\in Q}\left|\frac{||Mx||_2^2}{||x||_2^2} - 1\right| \leq \epsilon\right] \tag{121}$$

$$\begin{aligned}\overline{B(M,\epsilon)} &= \left[\exists\ x \in Q :\ \left|||Mx||_2^2 - ||x||_2^2\right| > \epsilon||x||_2^2\right] \\ &= \cup_{x\in Q}\Big[\left|||Mx||_2^2 - ||x||_2^2\right| > \epsilon||x||_2^2\Big]\end{aligned}$$

so by the union bound,

$$\mathbb{P}(\overline{B(M,\epsilon)}) \leq \sum_{x\in Q}\mathbb{P}\left(A(M, x, \epsilon)\right) \tag{122}$$
$$\leq 2|Q|\exp\left(-m\epsilon^2/12\right) \tag{123}$$

Now the condition on m is equivalent to

$$2|Q|\exp(-m\epsilon^2/12) < \delta \tag{124}$$

and the proof is complete. □

The last step between the concentration inequality, the J-L Lemma and the RIP property is a covering argument that will let us replace Q by Σ_k thanks to the following result:

Theorem 4 *Let $\rho \in]0, 1/2[$. There exists a finite subset $\mathcal{U}$ of the unit sphere $\mathcal{S}_S = \{x \in \mathbb{R}^n : supp(x) \subset S, |S| = k, ||x||_2 = 1\}$ such that for all $x \in \mathcal{S}_S$*

$$\min_{u\in\mathcal{U}} ||x - u||_2 \leq \rho \tag{125}$$

and the cardinality of $\mathcal{U}$ satisfies

$$|\mathcal{U}| \leq \left(1 + \frac{2}{\rho}\right)^k \tag{126}$$

This means that we can choose a finite number of points in the unit ball such that all points of the ball are not far away from these points. We do not prove this result (see Foucart).

Now, the CI gives, for $\epsilon \in]0, 1[$ (depending on ρ and δ to be determined later),

$$\mathbb{P}\left(\bigcup_{u\in\mathcal{U}} A(M, u, \epsilon)\right) \leq 2|\mathcal{U}|e^{-mc\epsilon^2} \tag{127}$$

$$\leq 2\left(1 + \frac{2}{\rho}\right)^k e^{-mc\epsilon^2} \tag{128}$$

where c is a constant. In the last step, we have to replace the finite set $\mathcal{U}$ by the infinite uncountable Σ_k. Let $B = A'_S A_S - I$. We recall that the RIP(ϵ, k) property is equivalent to:

$$\forall S \subset [\![1..n]\!] \text{ s.t. } |S| = k,\ ||M'_S M_S - I||_2 \leq \epsilon \tag{129}$$

The CI applied to $u \in \mathcal{U}$ gives

$$|\langle Bu, u\rangle| \leq \epsilon,\ \forall u \in \mathcal{U} \tag{130}$$

Let $x \in \mathcal{S}_S$ and $u \in \mathcal{U}$ such that $||x - u||_2 \leq \rho < 1/2$.Then,

$$|\langle Bx, x\rangle| = |\langle Bu, u\rangle + \langle B(x + u), x - u\rangle| \tag{131}$$

$$\leq |\langle Bu, u\rangle| + |\langle B(x + u), x - u\rangle| \tag{132}$$

$$< \epsilon + ||B||_2||x + u||_2||x - u||_2 \tag{133}$$

$$\leq \epsilon + 2\rho||B||_2 \tag{134}$$

Taking the maximum over $x \in \mathcal{S}_S$ gives

$$||B||_2 < \epsilon + 2\rho||B||_2 \tag{135}$$

so that

$$||B||_2 < \frac{\epsilon}{1 - 2\rho} \tag{136}$$

Now if we put $\epsilon = (1 - 2\rho)\delta$, then $||B||_2 < \delta$ so that

$$\mathbb{P}\Big[||A'_S A_S - I||_2 \geq \delta\Big] \leq ...$$

$$...2\left(1 + \frac{2}{\rho}\right)^k \exp\left(-mc(1 - 2\rho)^2\delta^2\right)$$

We have change a condition on the finite set $\mathcal{U}$ with precision ϵ into a condition on the infinite set $\mathcal{S}_S$ with precision δ. Moreover, the sparsity k appears now in the inequality. The condition can also be expressed as follows:

$$\mathbb{P}\Big[||B||_2 \leq \delta\Big] = \mathbb{P}\left(\bigcap_{x\in\mathcal{S}_S} A(M, x, \delta)\right) \tag{137}$$

which is equivalent to

$$\mathbb{P}\Big[||B||_2 \geq \delta\Big] = \mathbb{P}\left(\bigcup_{x\in\mathcal{S}_S} \overline{A(M, x, \delta)}\right) \tag{138}$$

$$\leq \mathbb{P}\left(\bigcup_{u\in\mathcal{U}} \overline{A(M, u, \epsilon)}\right) \tag{139}$$

$$\leq 2\left(1 + \frac{2}{\rho}\right)^k \exp\left(-mc(1 - 2\rho)^2\delta^2\right) \tag{140}$$

Now, there exists exactly

$$\left(\begin{array}{c} n \\ k \end{array}\right) \leq \left(\frac{en}{k}\right)^k$$

possible sets $\mathcal{S}_S$ of cardinality k (why do we not write of cardinality $\leq k$?). In words,

$$\Sigma_k^1 = \bigcup_{S:|S|=k} \mathcal{S}_S \tag{141}$$

where Σ_k^1 is the set of all vectors of norm 1 with a support of cardinal k (and the union is finite). We can drop the restriction to vectors of norm 1, because the linearity of the transformation M makes the inequalities equivalent (we can divide both side by $||x||_2$). A last use of the union bound gives eventually

$$\mathbb{P}_M\left(\overline{\text{RIP}}(\delta, k)\right) \leq \sum_{S:|S|=k} \mathbb{P}\left(\bigcup_{x\in\mathcal{S}_S} \overline{A(M, x, \delta)}\right) \tag{142}$$

$$\leq 2\left(\frac{en}{k}\right)^k \left(1 + \frac{2}{\rho}\right)^k \exp\left(-mc(1 - 2\rho)^2\delta^2\right) \tag{143}$$

We can know prove that the RIP property is always satisfied by a random Gaussian matrix with a

probability arbitrary close to 1 (up to a factor τ), as soon as the number m of measurements is greater than a function of n and k:

$$2\left(\frac{en}{k}\right)^k \left(1+\frac{2}{\rho}\right)^k \exp\left(-mc(1-2\rho)^2\delta^2\right) \leq \tau$$
$$\iff m \geq \frac{k\ln(en/k) + k\ln(1+2/\rho) + \ln(2/\tau)}{c\delta^2(1-2\rho)^2}$$
$$\iff m \geq Ck\ln(en/k) + C'k + C''\ln(2/\tau)$$

Theorem 5 *Let $M \in \mathbb{R}^{m\times n}$ a matrix whose coefficients are i.i.d. Gaussian random variables $\mathcal{N}(0,1/m)$. For all $\epsilon \in]0,1[$, M satisfies the (ϵ,k)-RIP property as soon as*

$$m \geq C \times k\ln(en/k) \tag{144}$$

with a probability arbitrary close to 1.

Which means that the number of measurements needed to recover a k-sparse vector is roughly proportional to k up to a logarithm factor in n/k.

Let's summarize the main lines of the proof:

- Gaussian matrices satisfies CI: $\forall x \in \mathbb{R}^n$, $\mathbb{P}\left(A(M,x,\epsilon)\right) \sim 1$
- if $x \in Q$ finite , the J.L. Lemma gives: $\mathbb{P}\left(\forall x \in Q : A(M,x,\epsilon)\right) \sim 1$
- if x on the sphere, with support S and $|S| = k$, covering property gives $\mathbb{P}\left(\forall x \in \mathcal{S}_S : A(M,x,\epsilon)\right) \sim 1$
- if $x \in \Sigma_k$, union bound gives $\mathbb{P}\left(\forall x \in \Sigma_k : A(M,x,\epsilon)\right) = \mathbb{P}\left(\mathrm{RIP}(\epsilon,k)\right) \sim 1$

3 Recovery guarantees

3.1 Introduction and the RIP as a first example

A guarantee of perfect recovery is a condition on the sensing matrix M and the parameters n,k,m to ensure that the estimated sparse signal $\widehat{x}$, solution of a reconstruction algorithm, is indeed equal to the initial signal x:

$$x = \widehat{x} \tag{145}$$

The condition can be deterministic or probably approximatively correct (PAC or one can also say "with overwhelming probability") in which case a parameter τ measures how far the probability of perfect recovery is from 1.

The guarantee can also be uniform if one single matrix M allows the reconstruction of all sparse signal x (M doesn't depend on x) or non-uniform if, for any fixed signal x, there exists a matrix M (depending on x) that allows the reconstruction.

A guarantee can be sufficient, necessary, or both.

Note that a recovery guarantee depends on the nature of the problem, but also on the algorithm used to solve it. Let's summarize:

- Deterministic framework:
  - Uniform: $\exists\ M$ s.t. $\forall x, \widehat{x} = x$ $\Rightarrow m = 2k$ necessary, but in fact, $m = 2k$ sufficient (in theory).
  - Non uniform: $\forall x, \exists\ M$ s.t. $\widehat{x} = x$ $\Rightarrow m = k+1$ necessary, and in fact, $m = k+1$ often sufficient (in theory).

  The problems are solved with algebraic methods of recovery using interpolation Vandermonde or FFT matrices. These methods are not stable, not robust to noise or approximate sparsity.
- Random framework
  - Uniform: $\mathbb{P}_M\left(\forall x, \widehat{x} = x\right) \geq 1-\tau$
  - Non uniform:$\forall x, \mathbb{P}_M\left(\widehat{x} = x\right) \geq 1-\tau$

  For example, the RIP property obtained with random Gaussian matrix is uniform, necessary, sufficient and probabilist (there exists deterministic matrix with RIP, but difficult to build and less efficient than random ones). If M satisfies the (ϵ, kp)-RIP condition, then with high probability, we can reconstruct every k- sparse vector by using BP (if $p=2$), IHT (if $p=3$) or OMP (if $p=12$).

3.2 Spark and NSP

Let M be a $m \times n$ matrix. The spark of M is the minimum number of columns that are linearly dependent.

$$\text{spark } M = \min_{x\neq 0} ||x||_0 \text{ subject to } Mx = 0 \tag{146}$$

The spark is the minimum weight of the non-trivial vectors in the nullspace:

$$\text{spark } M = \min\left\{k : \ker M \cap \Sigma_k \neq 0\right\} \tag{147}$$

The spark is clearly connected to the rank of the sensing matrix: $\exists$ a set of spark M columns that are dependent, so any set of spark M-1 columns are free. The rank is the maximum number of linearly

independent columns of M: $\exists$ at least a set of rank M columns that are free, so any set of rank $M+1$ columns are dependent. In words,

$$\text{spark } M - 1 \leq \text{rank} \tag{148}$$

Let M be a $m \times n$ matrix with $m \leq n$

- $2 \leq \text{spark } M \leq m + 1$
- in general spark $M \neq$rank $M + 1$
- if M is a random matrix with i.i.d. entries and continuous density, then spark $M =$ rank $M + 1$ with probability one.
- calculating spark M is complex.
- rank M can be computed via Gaussian elimination.

The spark gives a guarantee of unicity for the problem P_0, but is also a necessary and sufficient solution for recovery of any k-sparse vector under problem P_0. But solving this problem, as already said, is NP-hard.

Theorem 6 *Let $M \in \mathbb{R}^{m \times n}$ and $k \leq m$. The following are equivalent:*

- *Every $x \in \Sigma_k$ is the unique k-sparse solution of $Mz = Mx$*
- $\ker M \cap \Sigma_{2k} = \{0\}$
- $\forall S \subset [\![1, n]\!]$ *with* $|S| \leq 2k$, M_S *is injective*
- *Every subset of $2k$ columns of M are linearly independent*
- *spark $M > 2k$*

The first sentence has to be understood as follows: forall $x \in \Sigma_k$, there is a unique solution to the equation $Mz = Mx$ where the unknow is z. In words, x is the only k-sparse solution of $Mx = Mz$; note that this is exactly the problem P_0. As a consequence, exact recovery of every k-sparse vector needs

$$m \geq 2k \tag{149}$$

and we have already said that in fact, $m = 2k$ is sufficient for perfect recovery, but with unstable methods not usable anymore in high dimension.

There is a strong link between spark and the theory of error correcting codes (which is one possible application for compressive sensing). If M is the generator matrix of a linear error correcting code, then the spark of M is exactly the minimum distance of the code.

So the spark gives a unicity recovery guarantee for P_0. The next guarantee, called null space property (NSP), gives the same guarantee for P_1.

Let M be an $m \times n$ matrix. Then M has the null space property (NSP) of order k if, for all $v \neq 0 \in \ker M$ and for all index sets S s.t. $|S| \leq k$,

$$||v_S||_1 < ||v_{\overline{S}}||_1 \tag{150}$$

Where $\overline{S}$ is the complementary of S; equivalently,

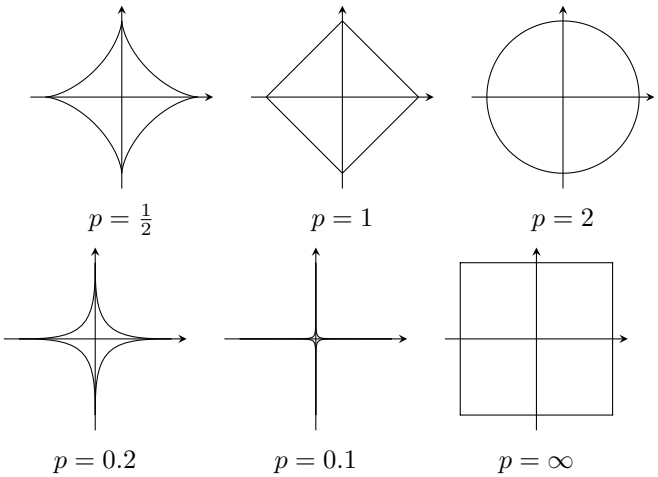
$$||v_S||_1 < \frac{1}{2}||v||_1 \tag{151}$$

The algebraic interpretation is that the vectors of the kernel must not be too concentrate on small subsets. The "weight" of the kernel vectors must be diluted within all their coordinates.

Before giving the geometric interpretation of NSP, this is the right place to recall the geometric interpretation of the problem P_p. Let's recall that:

$$P_p : \ \hat{x} = \underset{z:Mz=y}{\text{argmin}} ||z||_p \tag{152}$$

Solving P_p is equivalent to find the vector(s) of minimum p-norm, solution of the system $Mz = y$ (z is the unknown and y is fixed). The solution of $Mz = y$ formed a linear subspace of $\mathbb{R}^n$. Solving P_p is finding all vectors in the intersection of the smallest ball $||z||_p$ and the subspace $Mz = y$. The existence and the unicity of the solution(s) depends on p, because the geometry of the unit balls depends on p:

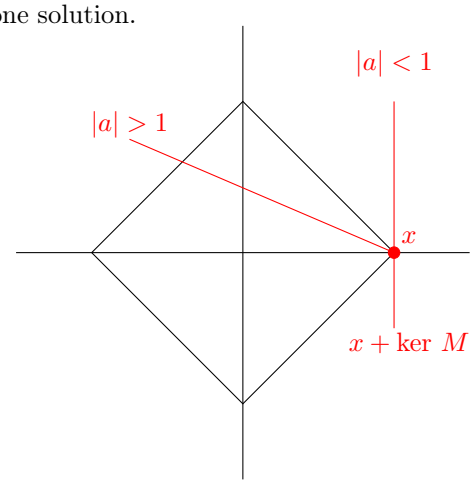


$\forall v \in \ker M$, $x + v$ is solution because $Mx + Mv = Mx$. In fact, the subspace of solutions can be written $x + \ker M = \{x + v; v \in \ker M\}$.

Let's come back to the NSP and let's take an example in dimension $n = 2$. Suppose that $x = (1, 0)'$ is the 1-sparse vector to recover. The support of x is $S = \{1\}$ (x is colinear to the x axis). Let $M = (1, a) \in \mathbb{R}^{1 \times 2}$ the sensing matrix (i.e. we make

one measure to recover the two dimensional vector x). It is easy to see that $||x||_0 = ||x||_1 = 1$ and that $\ker M \propto (-a, 1)'$ is the linear space of all vectors colinear to $v = (-a, 1)'$. $\ker M$ is of dimension 1 and for all $v \in \ker M$, $||v_S||_1 = |a|$ and $||v_{\overline{S}}||_1 = 1$, so that M verifies the NSP of order 1 if, and only if, $|a| < 1$.

As we can see in the figure below, the number of intersections between the line $x + \ker M$ and the ball $||x||_1$ depends on the slope $-1/a$ of the line. When NSP is verified, i.e. when $|a| < 1$, there is one unique solution, but when $|a| \geq 1$, there exists more than one solution.



Theorem 7 *Let M be an $m \times n$ matrix, and let $k \leq m$. Then, the following are equivalent:*

- *If a solution x of P_1 satisfies $||x||_0 \leq k$, it is the unique solution.*
- *M satisfies NSP of order k.*

NSP is a necessary and sufficient condition that guaranties to find the unique solution of P_1 by using $||.||_1$ minimization algorithms.

3.3 ERC for OMP

At each iteration t, OMP:

- Selects most correlated atom with residue r_{t-1} ($r_0 = y$):

$$\lambda_t = \underset{\lambda \notin \Lambda^{t-1}}{\operatorname{argmax}} |\langle r_{t-1}, M_\lambda \rangle| \tag{153}$$

- Update estimated support $\Lambda^t = \Lambda^{t-1} \cup \{\lambda_t\}$
- Computes estimate x_t and residue r_t.

$$x_t = \underset{z \in \mathbb{R}^n : \text{supp}(z) \subset \Lambda^t}{\operatorname{argmin}} ||y - Mz||_2$$
$$= \begin{cases} x_t|_{\Lambda^t} = M_{\Lambda^t}^\dagger y = \left(M'_{\Lambda^t} M_{\Lambda^t}\right)^{-1} M'_{\Lambda^t} y \\ x_t|_{\overline{\Lambda}^t} = 0 \end{cases}$$

$$r_t = y - Mx_t = y - M_{\Lambda^t} . x_t|_{\Lambda^t} = y - P_t y = Q_t y$$
$$P_t = M_{\Lambda^t} M_{\Lambda^t}^\dagger = M_{\Lambda^t} \left(M'_{\Lambda^t} M_{\Lambda^t}\right)^{-1} M'_{\Lambda^t}$$

and the scalar products are the

$$W_i(t) = |< r_{t-1}, M_i >| \tag{154}$$

The iteration t is a success if, and only if,

$$\max_{j \in \Lambda^t} |W_j(t)| \geq \max_{j \notin \Lambda^t} |W_j(t)| \tag{155}$$

Theorem 8 (ERC) *OMP recovers any x of support S s.t. $|S| = k$ if, and only if, M_S is injective and*

$$\max_{i \in S^c} ||M_S^\dagger M_i|| < 1 \tag{156}$$

Success for OMP $\Rightarrow$ success for P_1 (and BP) and ERC for $S \Rightarrow$ NSP for S with $||.||_1$.

3.4 Mutual coherence

The mutual coherence of a matrix M is the positive real number

$$\mu(M) = \max_{i \neq j} \left| \left\langle \frac{M_i}{||M_i||}, \frac{M_j}{||M_j||} \right\rangle \right| \tag{157}$$

If the matrix M is normalized (i.e. $||M_j||_2 = 1$, which we suppose from now on), $\mu(M) = \max_{i \neq j} |\langle M_i, M_j \rangle|$.

$$\mu = \max_{i \neq j} |\langle M_i, M_j \rangle| \tag{158}$$

$G = M'M = (\langle M_i, M_j \rangle)_{i,j}$ is the Gram matrix of M. Each coefficient gives a measures of the angle between any pair of columns. It is easy to see that $\mu \in [0, 1]$, $\mu = 0 \iff$ columns of M are orthonormal and $\mu = 1 \iff$ some columns are colinear.

Welsh bound: $\mu \geq \sqrt{\frac{n-m}{m(n-1)}}$

Theorem 9

$$\mu(M) < \frac{1}{2k-1} \tag{159}$$

is sufficient to perfectly recover any $x \in \Sigma_k$ with OMP, BP, P_1. If x solution of P_0, then x also solution of P_1.

Theorem 10

$$\mu(M) < \frac{1}{2k-1} \frac{|x_{min}|}{|x_{max}|} \tag{160}$$

is sufficient to perfectly recover any $x \in \Sigma_k$ with IHT.

The condition is tight: if $\mu(M) = (2k-1)^{-1}, \exists\ x$ impossible to recover.

3.5 Connections between different guarantees

Let $M \in \mathbb{R}^{m \times n}$ with normalized columns.

- spark $M \geq 1 + 1/\mu$
- M is (ϵ, k)-RIP with $\epsilon = k\mu$, for all $k < 1/\mu$.
- if M is $(2k, \epsilon)$-RIP, then spark $M > 2k$.
- if M is $(2k, \epsilon)$-RIP with $\epsilon < \sqrt{2} - 1$ and

  $$\frac{\sqrt{2}\epsilon}{1 - (1 + \sqrt{2})\epsilon} < \sqrt{\frac{k}{n}}$$

  then M is NSP-$2k$
- If M has coherence μ, the minimum number of measurements to recover $x \in \Sigma_k$ is of order

  $$m \geq C\mu^2 k \ln n$$