Compressive Sensing
– Mathematical Principles and Practical Implications –

Y. M. Cho (Seoul National University)

I. Abstract

The mathematical foundations of the compressive sensing which goes against the common wisdom of data acquisition (the Nyquist–Shannon theorem) is reviewed. The compressive sensing asserts that one can reconstruct images or signals of interest accurately from a number of samples far smaller than the desired resolution of the the image (e.g., the number of pixels in the image). The compressive sensing has far reaching implications. It suggests the new data acquisition protocols that translates analog information to digital form with fewer sensors considered necessary.

II. Introduction

The convenience and comfort in modern civilization would be unthinkable without the high resolution imaging devices digital cameras, televisions, magnetic resonance imagings (MRIs), etc. But the complexity of modern civilization demands better devices which have to deal with extremely large amount (billions or even trillions of pixels) of data, which creates problems to store and transmit the data. This necessitates the image compression algorithms which can reduce large amount of data sets by orders of magnitudes.

The basic concept of the data compression in imaging devices is simple: we transform the image to an appropriate basis and code only the important parts to convert high–resolution images into a relatively small bits, keeping the essential features intact. This in effect turns a large digital data set into a substantially smaller one. A most notable example of this image compression was the change of the sinusoid–basis to the wavelet–basis in JPEG [1], which replaced the classical JPEG to the modern JPEG–2000 standards [2]. But this compression is an "adaptive" compression of existing data which has already been acquired.

The compressive sensing is about the compression of data at the acquisition level, at the sampling level, without damaging the capability to reproduce the full image. The advantage of the compressive sensing is obvious: if we can compress the data at the sampling level, we can avoid the large digital data set to begin with. But this is against the common wisdom in data acquisition.

The celebrated Nyquist–Shannon theorem tells that the sampling rate must be at least twice the maximum frequency present in the signal (the Nyquist rate). According to this theorem the number of samples needed to reconstruct a signal without error is dictated by the Nyquist rate. In other words one must sample the signal at or above the Nyquist rate. And this principle underlies all existing signal acquisition protocols used in consumer audio and visual electronics, medical imaging devices, radio receivers, and so on. But in many applications the Nyquist rate is so high that too many samplings are needed, which makes compression of data prior to storage and transmission a necessity. This requires the compressive sensing, an imaging protocol which reproduces high-resolution images with much fewer samplings than required by the Nyquist-Shannon rate.

There is another motivation for the compressive sensing. In many imaging devices we may have to deal with the undersampled situations where the number of available measurements is much smaller than the dimension of the signal. This is because sampling gets very expensive and/or very difficult for various reasons. There are countless examples of this. In radiology and biomedical imaging, for example, the sensing process becomes very slow that one can only measure the object a few times (as in MRI). And in gene expression studies one may have to infer the gene expression level of thousands of genes with a low number (typically tens) of observations. Moreover, in the applications which use high-speed analog to digital converters (medical scanners and radars), increasing sampling rate gets very expensive because of the current limitation of technology. Also, in imaging devices which use the neutron scattering, the measurements gets very expensive. In other cases the number of sensors may be limited, and so on. In these cases one is forced to reproduce the images with the undersamplings (i.e., the compressed sensing).

A first indication that the compressive sensing is possible appeared in 1970s, when seismologists reconstructed images of reflective layers within the earth based on the data that did not seem to satisfy the Nyquist-Shannon criterion [3]. The theory of compressive sensing developed later around 2004 by Candes, Donoho, Romberg, and Tao (and others) asserts that one can recover certain signals and images from far fewer samples or measurements than the traditional methods require [4-7]. The theory is based on a wide ranges of fields in applied sciences and engineering probability theory, information theory, mathematical optimization theory, theoretical computer science, and others. The purpose of this article is to review the key mathematical ideas and some important results, and to discuss the implications of the compressive sensing. Doing this we will show why the compressive sensing is a concrete protocol for sensing and compressing data simultaneously.

The compressive sensing is about reproducing the full signal with undersampled data. It relies on two principles: the sparsity which has to do with the signals of interest and the incoherence which has to do with the sensing method. In addition to this, of course, it relies on the mathematicaltheorems based on two principles which assures that we can reproduce the full signal with the undersampled data.

The sparsity is based on the idea that for many signals the information rate are often much smaller than suggested by its bandwidth. In otherwords in many cases the useful informations are sparsely distributed that, when expressed in a proper basis, they have concise representations so that we can compress them to make them much smaller. This is because each signal has its own unique structurewhich is distributed sparsely when we measure it. It is this sparsity which allows us efficient compression of data and efficient data acquisition process.

The incoherence, on the other hand, is based on the idea that signals having a sparse representation must be spread out in the domain (the dual space) in which they are acquired, just like the Dirac's delta function in space or time is spread out in the dual space, momentum or energy (or wave vector or frequency) space. This tells that the sampling waveforms, unlike the signal of interest, must have an extremely dense representation in the basis in which the signal has concise representation. So, when we encode the signal, we should exploit this incoherence. In other words the incoherence tells us what should be the most effective sensing methods.

Of course, it has been well known that typical signals have structures, so that they can be compressed without much loss. For example, JPEG 2000exploits the fact that many signals have sparse representation in a fixed basis, so that one can store only a small number of adaptively chosen largest coefficients, discarding all insignificant ones (typically more than 90 to 95 percent). Obviously this type of data compression (discarding most of the data after the massive data acquisition) is extremely wasteful. To be economic we need data compression at the acquisition level, so that we do not have to measure the unnecessary informations only to discard them later.

The compressive sensing allows us to directly acquire just the "minimum" important information of the signal (about four times the number of the sparsity) from just a few sensors as in random sensing, and "decompresses" the measured (undersampled) data to reconstruct the signal with accuracy at least as good as what one has with typical compression after the massive data acquisition. This means that the compressive sensing can practically reduce the number of necessary samplings by a factor of roughly ten or twenty percent. This, of course, is an impressive gain. And here the sparsity of the signal and the incoherence of the sensing method play the central role.

This means that one can design an efficient sensing protocol that captures useful information from a sparse signal and condense it into a small amount of data. All we need is correlating the signal with a small number of fixed waveforms that are incoherent with the sparsifying basis. What is really remarkable about this protocol is the followings: First it allows a sensor which captures the information in a sparse signal non-adaptively, without trying to comprehend it. Second, it allows numerical optimization procedures with which we can reconstruct the full signal from the small amount of collected data. In other words, we can have a very effiient and simple signal acquisition protocol which allows us to sample the signal at a very low rate yet reconstruct it fully, in a signal independent fashion. This is the essence of the compressive sensing.

Although the compressive sensing is a relatively new and young field which has yet to be matured, it clearly provides us great opportunity to advance modern civilization a step further. In Korea the first international workshop on compressive sensing was organized last November by School of Electrical and Computer Engineering, Ulsan Institute of Science and Technology. But clearly we need more scientists and scientific activities in this field, domestically as well as internationally. This article is intended to provide a tutorial introduction to the subject. But there already exist many excellent review articles on the subject by the early proponents. We recommend the articles in the special section of *IEEE Signal Processing Magazine*, Volume 25, March 2008, in particular the articles by Romberg, Candes and Watkin, and Baraniuk, to the interested readers [8,9].

Ⅲ. Problem of Compressive Sensing

Consider the general problem of reconstructing an image or a signal, which we choose to be a real-valued function $f(x)$ of space (or time) for simplicity. We can discretize x to n points x_i $(i=1,2,...,n)$ and $f(x)$ to n numbers $f_i=f(x_i)$ and treat it as an n-dimensional vector $\vec{f}=(f_1,f_2,f_3,...,f_n)$ in $\mathbb{R}^n$, where $(f_1,f_2,f_3,...,f_n)$ are identified as the coordinates (pixels) of $\vec{f}$ in the canonical basis. It is assumed that n is a large number. Now, introduce another complete set of orthonormal vectors (the measuring vectors) $\hat{a}_i=(a_{i1},a_{i2},a_{i3},...,a_{in})$ with $\hat{a}_i \cdot \hat{a}_j=\delta_{ij}$ in R^n, and measure $\vec{f}$ by $\hat{a}_i$ to reproduce it

$$g_i=\hat{a}_i \cdot \vec{f}=a_{ij}\ f_j, \qquad \vec{f}=\Sigma_{i=1}^n\ g_i \hat{a}_i. \tag{1}$$

Here we have assumed that $\hat{a}_i$ are orthonormal, but this assumption is not essential. All we need is the linear independence. The measurement generates another vector $\vec{g}=(g_1,g_2,...,g_n)$ in $\mathbb{R}^n$, and clearly all information about $\vec{f}$ is stored in $\vec{g}$ (and vise versa).

The choice of $\hat{a}_i$ determines the type of the information we collect, so that it tells what type of coded imaging systems we use. For example, if $\hat{a}_i$ are the sinusoids with frequencies ω_i, we are essentially collecting the Fourier coefficients of the signal (as in magnetic resonance imaging). If they are the delta functions $\delta(x-x_i)$, we are measuring the (one-dimensional) pixels f_i. For two-dimensional images, if they are the delta ridges, we are measuring the line integrals of the image (as in tomography). And if they are the indicator functions on square, we are again collecting the pixels (as in standard digital cameras).

Since $\vec{f}$ and $\vec{g}$ are equivalent to each other, we can certainly reproduce the signal $\vec{f}$ with $\vec{g}$,

$$\vec{f} = A\vec{g}\,, \quad f_i = A_{ij}\, g_j, \quad A_{ij} = (a^{-1})_{ij}\,. \qquad (2)$$

But in many cases we find that most of the n coefficients g_i are very small, so that actually only a small number k of them become important. In this case the signal f is called *k-sparse*, and is deemed *compressible*. Of course, here we are interested in the case when $k \ll n$.

So, if the signal is k-sparse, we can approximate $\vec{f}$ with only k-coefficients (neglecting $(n-k)$ of them), and express it by a linear combination of k basis vectors of $\hat{a}_i$. In other words, we only need a much smaller data (k coefficients) to reproduce the n-dimensional vector $\vec{f}$. This implies that we could reproduce the signal with only k measurements, if we are clever enough.

This raises the interesting question: Can we construct an imaging device which can reproduce an arbitrary k-sparse n-dimensional signal $\vec{f}$ with only m ($m \simeq k$) samplings which is much smaller than n? From the above discussion one might think that the answer is definitely yes. Indeed the above argument implies that, choosing $\hat{a}_1 = \vec{f}/|\vec{f}|$, we could always approximate the signal $\vec{f}$ with only one measurement g_1 and one vector $\hat{a}_1$, making all other g_i with $i = 2,3,\ldots,n$ zero. But this is totally misleading. The reason why we could easily approximate the signal $\vec{f}$ with k coefficients was simply because we already had all information about $\vec{f}$, namely all g_i and $\hat{a}_i$. In other words we could do this because we knew everything and knew which g_i could safely be neglected. But what we want to know is the answer in the absence of a *priori* knowledge of g_i, for arbitrary k-sparse signal. In this case the answer to this question is not simple.

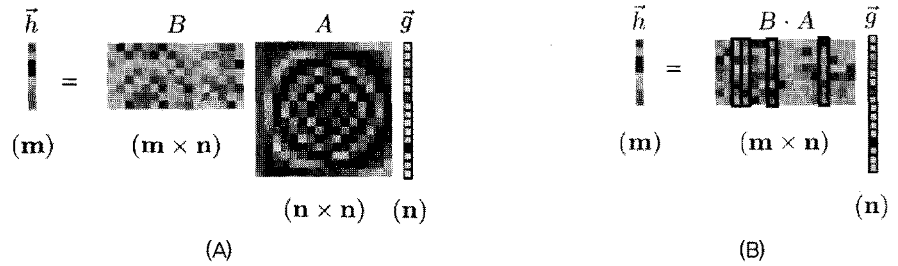
To understand the meaning of this question, remember how the existing imaging devices work. Here the full n samples f_i of the signal are acquired, the complete set of transform coefficients g_i is computed, k largest coefficients are identified and all remaining small coefficients are discarded, and k largest coefficients are encoded for the reproduction (and transmission) of the signal. But this process has three inherent drawbacks. First, the initial number of measurements n is too many. Second, all coefficients g_i must be computed even though only k of them is needed. Third, the locations of large coefficients must be encoded for the reproduction of the image.

The compressive sensing addresses these inefficiencies by directly acquiring the compressed signal. Consider a general linear measurement process which uses a different sampling method based on m orthonormal measuring (encoding) vectors $\hat{b}_i$ $(i = 1,2,\ldots,m)$ in stead of n vectors $\hat{a}_i$. Now, compute m inner products $(h_1, h_2, \ldots, h_m)$ $(m \ll n)$ between $\vec{f}$ and $\hat{b}_i = (b_{i1}, b_{i2}, b_{i3}, \ldots, b_{im})$ $(i = 1,2,\ldots,m)$ to encode $\vec{f}$,

$$h_i = \hat{b}_i \cdot \vec{f} = b_{ij}\, f_j = b_{ij}\, a_{jk}\, g_k = B_{ij}\, A_{jk}\, g_k, \quad B_{ij} = b_{ij} \qquad (3)$$
$$\vec{h} = B\vec{f} = B \cdot A\vec{g}\,.$$

where B and $B{\cdot}A$ are $m \times n$ matrices. Here again the orthonormality of $\hat{b}_i$ is not essential (only the linear independence is), but we assume that for simplicity. Clearly $\vec{h} = (h_1, h_2, \ldots, h_m)$ can be viewed as an m-dimensional vector in $\mathbb{R}^m$. We can interpret that the basis $\hat{a}_i$ (or the matrix A) is used for measuring the signal, but the basis $\hat{b}_i$ (or the matrix B) is used to represent or encode it.

But an important point here is that, unlike the complete set $\hat{a}_i$, $\hat{b}_i$ form an incomplete set of basis in $\mathbb{R}^n$ because $m \ll n$. Of course we can extend the $\hat{b}_i$ basis to make it complete, adding $n-m$ orthonormal vectors. But what is crucial is that we use only m number of $\hat{b}_i$ to encode the signal. In other words the dimension of the encoding space $\mathbb{R}^m$ is much smaller than the dimension of the measuring space $\mathbb{R}^n$, and this reduction of the dimensionality is the crux of the



⟨Fig. 1⟩ The undersampling of a 4-sparse n-pixel image $\vec{f}=A\vec{g}$ encoded to an m-dimensional vector $\vec{h}$ in $\mathbb{R}^m$ by the matrix B. Here A and B represent an $(n\times n)$ discrete cosine transform (DCT) matrix and $(m\times n)$ random Gaussian encoding matrix. (A) shows how $\vec{f}$ is transformed to $\vec{g}$ and encoded to $\vec{h}$. And (B) shows the resulting dimensionality reduction of $\vec{g}$ to $\vec{h}$. There are 4 columns in $B{\cdot}A$ that correspond to 4 non-zero coe ffiients of $\vec{g}$, and $\vec{h}$ becomes a linear combination of these columns.

compressive sensing.

Now, the problem that we face in compressive sensing is designing a stable matrix B such that

1. The salient information in any k-sparse signal should not be damaged by the dimensionality reduction from $\vec{f}\in\mathbb{R}^n$ to $\vec{h}\in\mathbb{R}^m$.
2. There must be a practical and reliable reconstruction algorithm to recover $\vec{f}$ from only $m(m \simeq k)$ encoding vectors $\hat{b}_i$.
3. The encoding matrix B must be non-adaptive, so that it should not depend on the signal. In other words B must be applicable to any k-sparse signal.

If we do, than we can reconstruct the signal $\vec{f}$ with only m $(m \ll n)$ measurements, without going through the large n measurements. But the problem appears to be ill-conditioned, because in this case the encoding matrix B must allow the reconstruction of the n-dimensional vector $\vec{f}$ from only m measurements.

This dimensionality reduction of the encoding space is schematically shown in ⟨Fig. 1.⟩ Here the n-dimensional signal $\vec{f}$ is encoded by an m-dimensional vector $\vec{h}$ with the $m\times n$ matrix B. And the problem here is to reproduce the full signal $\vec{f}$ with the undersampled $\vec{h}$. Under the normal circumstance this would be hopeless to solve, because this is an underdetermined matrix equation. But we can solve this. What allows us to solve this problem is the k-sparsity of $\vec{f}$, that the signal has only k non-zero coefficients.

Ⅳ. Incoherent Sampling: Mathematical Principles

To solve the problem we have to do two things. First, we have to identify the important (nonzero) coefficients of g_i and register them. Second, we have to invent an algorithm to reconstruct the signal with the registered data. The first problem is the problem of undersampling to figure out what is the best strategy to minimize the measurements to encode (to identify the important coefficients) a sparse signal. Mathematically this translates to the problem of finding a best choice of $\hat{b}_i$ which can minimize the measurements and at the same time register the important coefficients (the sparse spikes) of $\vec{f}$, or equivalently $\vec{g}$.

As we have pointed out, if we know the locations of the non-vanishing coefficients, we only need k measurements (with k important $\hat{a}_i$). In fact with $\hat{b}_1=\vec{f}/|\vec{f}|$ (we can always make f to be 1-sparse so that) we need only one measurement to reproduce the signal. But this is an adaptive choice, which we can do only when we already know the signal $\vec{f}$. Obviously it does not work for other signals.

Clearly the answer to the problem is deeply related to the the signal reconstruction: the small set of measurements should be able to reproduce the full signal. For this to be possible, of course, the small set of the measurements should record all important characters of the signal. So we have to find out a most efficient way to register the important characters of any sparse signal, without trying to comprehend the nature of the signal in advance.

At this point the concept of the incoherence, the other pillar of the compressive sensing, plays the important role. To explain this consider two bases $\hat{a}_i$ and $\hat{b}_i$ (or two matrices A and B) we have introduced before. The coherence between the two bases is defined by

$$\mu(A,B) = \sqrt{n} \cdot max\,|\hat{a}_i \cdot \hat{b}_j|, \quad (1 \le i,j \le n) \qquad (4)$$
$$1 \le \mu(A,B) \le \sqrt{n}\ .$$

In other words the coherence measures the largest correlation between any two elements of A and B. The linear algebra assures that $\mu(A,B)$ ranges from 1 to $\sqrt{n}$. When $\mu=1$, we call A and B are maximally incoherent.

A good example of maximally incoherent bases is the delta-function basis and the sinusoidal basis. Another example of maximally incoherent bases is the delta-function basis and *the noiselet basis* [10]. The noiselet basis is made of $\hat{b}_i = (b_{i1}, b_{i2}, \ldots, b_{in})$ where b_{ik} are given by the noiselets, the unit-normed and binary valued step functions taking values $+1/\sqrt{n}$ or $-1/\sqrt{n}$ with equal probability. The noiselet basis is also known to be highly incoherent with wavelet bases (It is well known that $\mu \simeq 2$ between the noiselet basis and a typical wavelet basis). And the random matrices are largely incoherent with any fixed basis.

The importance of the incoherence in compressive sensing comes from the following observation: When we do not know *a priori* the location of the important coefficients of a sparse signal, the best strategy is to choose the encoding basis $\hat{b}_i$ to be maximally incoherent with $\hat{a}_i$. In other words the best strategy to encode a sparse signal is a random sensing.

The justification for this random sensing strategy is that each sparse signal has its own unique set of measurement in the maximally incoherent basis $\hat{b}_i$. To understand this it is important to realize that the mathematical basis for this incoherence principle is very much like the Heisenberg's uncertainty principle in quantum mechanics. In physics it is well known that a sparse peak (the Dirac's delta-function) in time or space is evenly distributed in the dual space (energy or momentum space, or equivalently frequency and wave vector space) after the Fourier transformation. Here the k-sparse signal (k delta-function peaks) in the measurement domain is evenly distributed, so that a best strategy to encode them is to adopt the random sensing based on noiselet or random basis.

Remarkably, this random sensing strategy has other important advantages. First, it is nonadaptive. In other words it works for any sparse signal, so that we do not have to taylor the random matrix to each signal. Second, the noiselet basis (or any random basis) can easily be generated so that it need not be stored to be applied. This is crucial for numerically efficient implementation of compressive sensing, which is very important for practical applications.

To make the incoherent sampling mathematically precise, let us try to reconstruct the image with m measurements $\vec{h}$. Clearly there are huge number of n-pixel images which can provide exactly the same measurements (the same encoding vector $\vec{h}$), simply because m is much smaller than n. So we have to find out the sparsest n-pixel image, namely the unique $\vec{f}$, that is consistent with the m measurements. Mathematically this amounts to finding out the n-dimensional vector $\vec{x}$ which satisfies the following convex optimization problem.

Problem I: Find $\vec{x}$ which minimizes $\|A\vec{x}\|_0$, subject to $B\vec{x} = \vec{h}$.

Here $\|A\vec{x}\|_0$ denotes the l_0-norm (i.e., the number of non-zero coefficients) of the vector $A\vec{x}$, and the constraint min $\|A\vec{x}\|_0$ assures that $\vec{x}$ indeed has the minimum number of non-zero coefficients (i.e., k-sparse) as we wish. Unfortunately solving the above problem is NP-hard (i.e., non-deterministic and polynomial-time hard). This means that solving the problem is computationally infeasible, requiring the exhaustive enumeration of all ${}_nC_k$ possible location of the k non-zero entries in $\vec{x}$ [11].

Surprisingly there is an optimization problem that works almost as well, the optimization based on the l_1-norm in stead of the l_0-norm

$$min \, \|A\vec{x}\|1 \quad subject \ to \quad B\vec{x}=\vec{h}. \qquad (5)$$

This problem is far easier to solve, because it can be recast as a linear program which can be solved by a number of modern techniques [12]. And the solution can exactly recover the k-sparse signal $\vec{f}$. This is based on the following central theorem in compressive sensing [13].

Theorem I: If

$$m \geq c \cdot \mu^2(A,B) \cdot k \log n \qquad (6)$$

for some positive constant c, the solution to (5) is exact with overwhelming probability.

Now, the following comments are in order:

1. The importance of the incoherence in encoding is obvious in the theorem. The smaller the coherence, the fewer samples are necessary. This is why the incoherent sampling is crucial in the compressive sensing.
2. With the maximal incoherence the number of necessary samplings becomes of the order of $k \log n$, in stead of n. This is a huge reduction in sampling.
3. This is independent of the signal, which does not require any a *priori* knowledge of the number of non-zero coefficients and their locations. All we need is a decoder which can "decompress" the undersampled data.

This theorem indeed assures that there exists a concrete acquisition protocol for the compressive sensing encode the signal non-adaptively in a compressed form with random sensing, and decode the compressed data with the l_1-minimization.

A closely related concept to the incoherence is *the uniform uncertainty principle*, which has to do with the uniqueness of the solution of the above convex optimization problem. To explain this consider the $m \times n$ matrix B made of m random test vectors $\hat{b}_i$. If we have

$$m \geq c \cdot k \log n, \qquad (7)$$

where C is a known constant, then for any k-sparse vector $\vec{f}$ the energy of the measurement $B\vec{f}$ will be comparable to the energy of $\vec{f}$ itself [7],

$$\frac{1}{2}\frac{m}{n} \leq \frac{\|B\vec{f}\|_2^2}{\|\vec{f}\|_2^2} \leq \frac{3}{2}\frac{m}{n}. \qquad (8)$$

Here $\|\vec{f}\|_2^2$ denotes the standard l_2-norm (the energy) of the signal $\vec{f}$. In other words the proportion of energy of $\vec{f}$ that appears in the measurements is roughly the same as the undersampling ratio m/n. This is known as the uniform uncertainty principle, and in this case the matrix B is said to obey the uniform uncertainty principle.

The importance of the uniform uncertainty principle is that it assures the solution is unique. To understand this let (8) holds for sets of size $2k$, and measure the k-sparse vector $\vec{h}=B\vec{f}$ as above. Now, we ask if there is any other k-sparse (or sparser) vector $\vec{f'}$ that yields the same measurement $\vec{h}$. The answer is no. If there were such $\vec{f'}$, then $\vec{u}=\vec{f'}-\vec{f}$ would be $2k$-sparse and have $B\vec{u}=0$. This is incompatible with (8).

V. Geometry of l_1 Minimization

To solve the compressive sensing problem it was crucial (and very ingenious) that we could replace the l_0-minimization condition by the l_1-minimization, even though they are fundamentally different. This is of utmost importance, because this makes the NP-hard problem to a solvable one and thus allows us to have realistic signal reproduction algorithm. Without this the compressive sensing would have been practically impossible.

One might wonder what is so special about this l_1 minimization, and why can't we impose the more familiar (and standard) l_2-minimization which minimizes the energy of the signal. The truth is that there is good reason why the l_2-minimization does not work and why the l_1 minimization works so well. In fact the l_2

minimization almost never find the k-sparse solution, producing instead a non-sparse solution with many non-zero coefficients. This has to do with the difference of the geometry between two minimizations [14,15].

To understand this let $n=3$ for simplicity and consider the l_1 and l_2 balls in $\mathbb{R}^3$ as shown in ⟨Fig. 2.⟩ Here (A) shows the subspaces containing 2-sparse vectors in $\mathbb{R}^3$ which lie close to the axes.

And (B) illustrates the l_1 ball of a given radius R, which is made of the points $|x_1|+|x_2|+|x_3| \le R$. Clearly it is anisotropic. More importantly it is "pointy" along the three axes. Now let the point $\vec{x}$ be a sparse vector which we want to reproduce with the encoding $\vec{h}$, and let the green hyperplane to represent the set of all signals $\vec{x}$ which provides the same encoded value $\vec{h}$. The task in (5) is to find the point on this hyperplane which has the minimum l_1 norm.

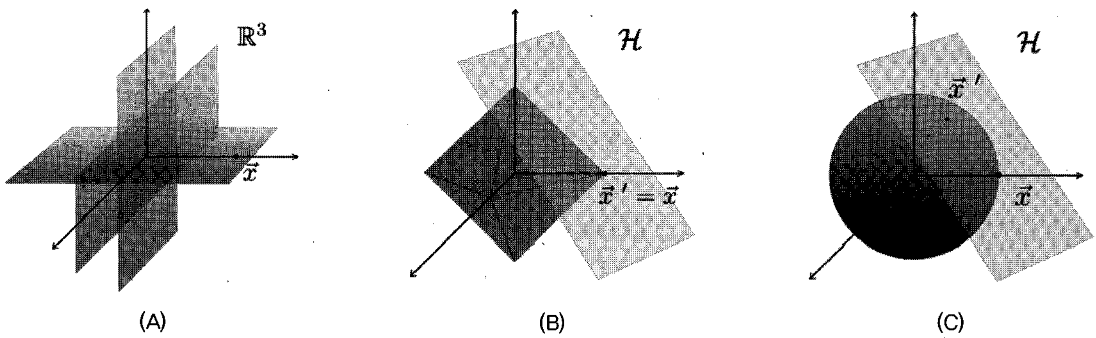
To visualize how the recovery program accomplishes this, take an l_1 ball with a small radius and enlarge it gradually till it hits the hyperplane. By definition the first point of contact is the vector that solves the problem. Now, it is intuitively clear that the pointiness of the l_1 ball together with the linearity (i.e., the flatness) of the hyperplane tells that the intersection (the contact point) occurs at precisely where the sparse signals are located, on the axes.

In comparison, consider the l_2 minimization shown in (C). Obviously the l_2 ball is spherical and isotropic. In this case the first encounter of the expanding l_2 ball with the hyperplane almost always occurs off the axes, where the sparse vectors are not located. So with the the l_2 minimization it is very difficult to find the k-sparse signal which satisfies the encoded data. And this difference becomes more dramatic in higher dimension when n becomes very large. This is why the l_2 minimization becomes completely useless in the signal recovery.

Although the above argument is very intuitive, it does tell that we can have a practical recovery program based on the l_1 minimization. But actual computation to find the solution, however, is non-trivial. This is becausethis type of problem obviously has to deal with hundreds of thousands of encoding data to construct millions of pixels signals. Clearly this will involve a good deal of computation. Fortunately there have been great advances recently in convex optimization which can deal with this type of problems [16]. The computational complexity of this type of problem (after the linearization of the problem) becomes about the order of $O(n^3)$, but it is manageable. And a good rule of thumb is that solving the l_1 minimization program is about thirty to fifty times as expensive as solving the l_2 minimization problem.

Ⅵ. Stability Of Reconstruction Algorithm And Error Estimate

The above theorem assures that a compressive sensing is possible, with overwhelming probability of success. This is very nice. But this is not enough,



⟨Fig. 2⟩ The geometry of l_1 and l_2 minimizations in R^3. Here (A) shows the subspaces containing 2-sparse vectors in $\mathbb{R}^3$. (B) shows the l_1 ball $|x_1|+|x_2|+|x_3|=R$ and the hyperplane $\mathcal{H}$ defined by $\vec{h}=B\vec{x}$. The encounter of the hyperplane with the l_1 ball occurs exactly where the sparse vectors are located. (C) shows the l_2 balls with $|x_1|^2+|x_2|^2+|x_3|^2=R$, but here the encounter in general occurs at the point where there is no sparse vectors because the real world is not so ideal. First of all, in reality most of the signals are not exactly sparse, but only approximately so. Moreover, they always have noise, especially because sensing devices do not have infinite precision. To make things worse, the above theorem assures that the solution "almost always" exists with overwhelming probability, but not always. In other words, the theorem admits that there is a possibility of failure, even though this possibility is extremely small. Therefore the compressive sensing must be able to deal with these non-ideal realities. At the very least we must have assurance of the stability in the reconstruction algorithm, that a small perturbations in data should not cause a large perturbation in the reconstruction. And of course, we need assurance of no failure. In other words we have to improve the above theorem to be deterministic, not probabilistic.

To deal with these problems, we have to consider the problem of recovering the signal with a contaminated data, with

$$\vec{h} = B\vec{f} \rightarrow B\vec{f} + \vec{z}, \tag{9}$$

where $\vec{z}$ represents a small but finite stochastic noise or unknown error term. This means that we have to modify the above problem to a new problem which can take care of the noise.

Problem II: Find $\vec{x}$ which minimizes $\|A\vec{x}\|_1$, subject to $\|B\vec{x} - \vec{h}\|_2 \leq \epsilon$, where ϵ is the upper bound of the noise in the data.

A key notion to resolve this problem is the so-called *the restricted isometry property*, which is a refinement of the uniform uncertainty principle [4]. Consider the encoding matrix B again. Define the isometry constant δ_k for each integer $k = 1,2,...$ as the smallest number such that

$$1 - \delta_k \leq \frac{\|B\vec{x}\|_2^2}{\|\vec{x}\|_2^2} \leq 1 + \delta_k \tag{10}$$

holds for all k-sparse vectors $\vec{x}$. If δ_k is not too close to one, the matrix B is said to have the restricted isometry property. When this property holds, B approximately preserves the Euclidean length of k-sparse signals. In this case the k-sparse vectors can not be in the null space of B. This is important, because otherwise there would be no hope to reconstruct these signals.

Another way to describe the restricted isometry property is that all subsets of k columns of B are in fact nearly orthogonal. This is because in this case B approximately preserves the Euclidean length of k-sparse signals. But notice that the columns of B can not be exactly orthogonal because the matrix has more columns than rows.

The importance of this restricted isometry property comes from the following observation. Suppose that δ_{2k} is sufficiently smaller than one. Then the distances between any two k-sparse vectors $\vec{x}_1$ and $\vec{x}_2$ must satisfy

$$1 - \delta_{2k} \leq \frac{\|B(\vec{x}_1 - \vec{x}_2)\|_2^2}{\|\vec{x}_1 - \vec{x}_2\|_2^2} \leq 1 + \delta_{2k}. \tag{11}$$

In other words all pairwise distances between two k-sparse signals must be well preserved in the measurement space. This fact guarantees the existence of efficient and stable reconstruction algorithm for discriminating k-sparse signals, as the following theorem assures [13].

Theorem II: Assume the matrix B has the restricted isometry property and let $\delta_{2k} < \sqrt{2} - 1$. Then the solution $\vec{x}_0$ to the Problem II obeys

$$\begin{aligned} \|\vec{x} - \vec{x}_0\|1 &\leq C_0 \cdot \|\vec{x} - \vec{x}_k\|_1, \\ \|\vec{x} - \vec{x}_0\|_2 &\leq C_0 \cdot \frac{\|\vec{x} - \vec{x}_k\|_1}{\sqrt{k}} + C_1 \cdot \epsilon, \end{aligned} \tag{12}$$

for some constants C_0 and C_1, where $\vec{x}_k$ is the $\vec{x}$ with all but the largest k components set to zero.

This theorem tells that minimizing the l_1–norm recovers the k–largest entries of an n–dimensional unknown signal with only m measurements. Clearly this is stronger than the previous theorem. If $\vec{x}$ is k–sparse, then $\vec{x}=\vec{x}_k$, and thus the recovery is exact. And this new theorem deals with all signals. If $\vec{x}$ is not k–sparse, then the theorem asserts that the quality of the recovered signal is as good as if one knew ahead of time the location of k largest values of $\vec{x}$ and decided to measure those directly. So, the reconstruct is as good as the one provided by an oracle which, with the full and perfect knowledge about $\vec{x}$, extracts the k most significant coefficients for us.

An important feature of this theorem is that, unlike the previous one, it involves no probability at all. It is deterministic. It guarantees to recover all sparse k vectors exactly, and essentially the k largest entries of all vectors. In other words there is no chance of failure.

Furthermore, it handles the the noise gracefully. The reconstruction error is bounded by two terms, the first term which occurs without any noise and the second term which is proportional to the noise level.

In conclusion, this theorem assures that a realistic compressive sensing mechanism which works all kind of not necessarily sparse signals and handles noise nicely is indeed possible. That is, if we can design efficient sensing matrices which obeys the restricted isometry property. Can we?

Ⅶ. Random Sensing Matrices

So the remaining problem is to design sensing matrices which has the restricted isometry property, with the property that column vectors taken from arbitrary subsets are nearly orthogonal. And the larger these subsets, the better. This is where the randomness re–enters in the picture.

Fortunately we do not have to worry about this problem either. There already exist several ways to construct such matrix. We list just some of them:

A. Sample n column vectors uniformly at random on the unit sphere of $\mathbb{R}^m$.

B. Sample independently and identically distributed entries from the normal distribution with mean 0 and variance $1/m$.

C. Sample random projection P and normalize $B=\sqrt{n/m}\,P$.

D. Sample independently and identically distributed entries from a symmetric Bernoulli distribution or other sub–Gaussian distribution.

All these matrices are known to obey the restricted isometry property with overwhelming probability, provided that [5,6]

$$m \geq C \cdot k \log \frac{n}{k}, \qquad (13)$$

where C is another constant which depends on each case. And, of course, there are other ways to construct such metrix. So we do not have to worry about devising an encoding matrix which satisfies the restricted isometry property.

A remarkable feature of these random matrices is that they are universal the sparsity basis need not even be known designing the encoding matrix. Moreover, they are almost optimal. In other words there are no measurement matrices and no reconstruction algorithm whatsoever which can reproduce the signal with substantially fewer samples. This means that the random sensing with the l_1 minimization is an optimal compressive sensing strategy.

Ⅷ. Discussions and Implications

In this article we have described the underlying logics of the compressive sensing the motivation, the basic concepts, the mathematical principles, the undersampling mechanism, the conditions for signal reproduction, and the stability of the reproduction algorithm. For the clarity of the logic we have skipped the details of the arguments. But our discussion assures that there exists a robust and stable compressive sensing mechanism which can reproduce practically all signals with far fewer samplings than required by the conventional imaging technique (the Nyquist–Shannon rate).

Theoretically the above discussion tells that one can in principle compress the n data to the order of $k\log(n/k)$. But a practical question at the end of the day is how much can we actually compress the data in reality. As we have already pointed out, the rough estimate is four times the sparsity. Now, assuming that k/n is a few percent (which is generally true in many applications), one may conclude that about ten or twenty percent of n samplings would suffice in average. In other words one may expect eighty to ninety percent compression of data. And in real applications the compressive sensing often works better than this.

An important feature of the compressive sensing which can be very useful is that it is asymmetric in data acquisition and signal reconstruction. In many imaging devices, the data acquisition process is complicated and expensive. The compressive sensing trades off this complicated data acquisition process with the data reconstruction process, which can be performed on a digital computer which need not even be collocated with the sensor.

The compressive sensing has huge applications, in almost all areas that use imaging devices. As a general rule of thumb, any two stage techniques or indirect imaging involving the use of a computer for the reconstruction of the signal is bound to find an application of compressive sensing technique.

Actual status of implementation of compressive sensing varies very much. In some cases one may only need new softwares which can be used with the existing devices, but in other cases one may have to develop new hardwares as well. But it is still developing field and there are many unanswered questions and challenges. And, of course, there are certain disadvantages and tradeoffs to make in compressive sensing. Here we discuss two examples, the compressive sensing in digital camera and MRI, for simplicity.

The compressive sensing allows us to build a simpler, smaller, and cheaper digital camera that can operate efficiently across a much broader spectral range than the conventional silicon–based cameras the new "single–pixel" camera based on an optical computer comprised of a digital microwave device (DMD), a single photon detector, and an analog–digital (A/D) converter [17]. Here the optical computer computes random linear measurements of the image under view, and the image is then reconstructed by a digital computer.

The single–pixel design reduces the required size, complexity, and cost of the photon detector array down to a single unit, which enables the use of exotic detectors that would be impossible in conventional digital cameras. These include a photo–multiplier tube or an avalanche photo–diode for low–light (photon–limited) imaging, a sandwich of several photo–diodes sensitive to different light wavelengths for multimodal sensing, a spectrometer for hyperspectral imaging, and others.

In addition to this sensing flexibilities, a practical advantage of the single–pixel design is that the quantum efficiency of a photodiode is higher than that of the pixel sensor in a typical CCD or CMOS array, since the fill factor of a DMD can reach 90 percent whereas that of a CCD/CMOS array is only about 50 percent. Another advantage is that each DMD measurement receives about $n/2$ times more photons than an average pixel sensor, which signifocantly reduces the image distortion from dark noise and read–out noise. But of course, there are challenges. One challenge is to extend the single–pixel concept to wavelengths where the DMD fails as a modulator, such as THz and X rays.

MRI is a medical imaging device which has an inherently slow data acquisition process, so that the compressive sensing offers potentially significant scan time reduction, with benefits for patients and health care budget [18]. It obeys two key requirements for the compressive sensing: MRI imagery is naturally sparse and compressible and MRI scanners naturally acquire the encoded samples (rather than the direct pixel samples).

The potential applications of the compressive sensing in MRI are the followings rapid 3–d angiography, whole–heart coronary imaging, brain imaging, dynamic heart imaging. And different applications bring us different constraints imposed by MRI scanning hardware or by patient considerations. The compressive sensing can significantly accelerate the magnetic resonance angiography, enabling better temporalresolution or alternatively improving the resolution of current imagery without compromising the scan time. In whole-heart coronary imaging, it can accelerate the data acquisition, allowing the entire heart to be imaged in a single breath. Similarly in brain imaging it reduces the data collection time while improving the resolution of the image. And in dynamic heart imaging it can recover the dynamic sequence at much higher rate and at the same time reduce the image artifact significantly.

But here again, many crucial issues remains unsettled. These include optimizing sampling trajectories, developing improved encoding mechanism, improving the speed of the reconstruction, and improving the quality of the reconstructed image in terms of clinical significance.

Our main concern in compressive sensing in this paper has been the reproduction of the signal with undersamplings, to reduce n samplings to k undersamplings. But in principle the compressive sensing can also be used to enhance resolution of the images with the n samplings, to turn the low resolution datasets to high resolution samples.

Ultimately, one could think of a compressive sensing for quantum information. In all current imaging devices, the signal processing means the classical signal processing (just like all present computers, no matter how complicated, are performing classical operations). But suppose we have quantum signal. In this case one might ask whether a compressive sensing of the quantum signal is possible. Of course, at this point it is not even clear to figure out what this question means, or whether this question has any meaning at all. But certainly this is a mind provoking question.

Acknowledgments

The work is supported in part by Ulsan National Institute of Science and Technology and by the National Research Foundation (Grant 2010-002-1564) of Korea.

References

[1] S. Mallat, *A Wavelet Tour of Signal Processing* (2nd edition), Academic (San Diego), 1999.

[2] A. Skodras, C. Christopoulos, and T. Ebrahimi, "The JPEG2000 still image compression standard", *IEEE Signal Processing Mag.* 18, 36 (2001); D. Taubman and M. Marcellin, *JPEG 2000: Image Compression Fundamentals, Standards, and Practice*, Kluwer (Norwell), 2001.

[3] J. Claerbout and F. Muir, "Robust modeling with erratic data", *Geophys. Mag.* 38, 826 (1973); F. Santosa and W. Symes, "Linear inversion of band-limited reflction seismograms", *SIAM J. Sci. Statist. Comput.* 7, 1307 (1986). See also B. Hayes, *The best bits*, American Scientist, July, 2009.

[4] E. Candes and T. Tao, "Decoding by linear programming", *IEEE Trans. Inform. Theory* 51, 4203 (2005).

[5] E. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information", *IEEE Trans. Inform. Theory* 52, 489 (2006).

[6] D. Donoho, "Compressed sensing", *IEEE Trans. Inform. Theory* 52, 1289 (2006).

[7] E. Candes and T. Tao, "Near optimal signal recovery from random projections: Universal encoding strategies?", *IEEE Trans. Inform. Theory* 52, 5406 (2006).

[8] J. Romberg, "Imaging via Compressive Sampling", IEEE Signal Processing Mag. 25, 14 (2008) E. Candes and M. Watkin, "An Introduction to Compressive Sampling", IEEE Signal Processing Mag. 25, 21 (2008) R. Baraniuk, "Compressive Sensing", *IEEE Signal Processing Mag.* 25, 118 (2008).

[9] See also, E. Candes, "Compressive Sampling", *Proceedings of the International Congress of Mathematicians*, 1, Madrid (2006).

[10] R. Coifman,F. Geshwind, and Y. Meyer, "Noiselets", *Appl. Comp. Harmon. Anal.* 10, 27 (2001).

[11] B. Natarajan, "Sparse approximation solutions to linear systems", *SIAM J. Comput.* 24, 227 (1995).

[12] S.S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit", *SIAM J. Sci. Comput.* 20, 33 (1999).

[13] E. Candes and J. Romberg, "Sparsity and incoherence in compressive sampling", *Inverse Prob.* 23, 969 (2007).

[14] E. Candes, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements", *Commun. Pure Appl. Math.* 59, 1207 (2006).

[15] D. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition", *IEEE Trans. Inform. Theory* 47, 2845 (2001) J. Tropp, "Just relax: Convex programming methods for identifying sparse signals in noise", *IEEE Trans. Inform. Theory* 52, 1030 (2006).

[16] S. Boyd and L. Vandenburghe, *Convex Optimization*, Cambridge University Press, 2004.

[17] D. Tackhar, J. Laska, M. Wakin, M. Duarte, D. Baron, S. Sarvotham, K. Kelly, and R. Baraniuk, "A new compressive imaging camera architecture using potical-domain compression" *Proceedings of VIth Computational Imaging*, 43, San Jose (2006).

[18] See, for example, M. Lustig, D. Donoho, J. Santos, J. Pauly, "Compressed sensing MRI", *IEEE Signal Processing Mag.* 25, 14 (2008).

조 용 민

1962년 4월~1966년 3월 Department of Physics, Seoul National University (BS)
1966년 7월~1970년 4월 Instructor, Korea Air Force Academy (Physics)
1970년 9월~1975년 6월 Department of Physics, University of Chicago (PhD)
1975년 7월~1978년 6월 Faculty Staff, New York University
1978년 7월~1980년 6월 Guest Scientist, Max Planck Institute
1980년 7월~1982년 11월 Invited Professor, University of Paris IV
1982년 12월~2009년 8월 Professor, Seoul National University
1966년 4월~1999년 3월 Executive Director, Asia Pacific Center for Theoretical Physics
2009년 9월~Present Chair Professor, Ulsan National Institute of Science and Technology
2009년 9월~Present Professor Emeritus, Seoul National University
2009년 9월~Present Professor Emeritus, Korea Air Force Academy