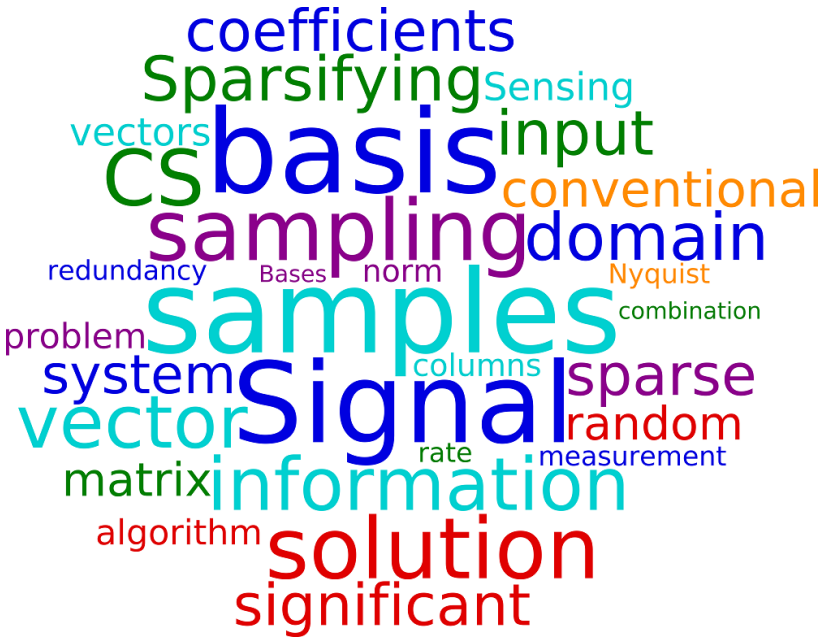
Accepted version of: H. E. A. Laue, “Demystifying Compressive Sensing [Lecture Notes],” in *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 171–176, July 2017. DOI:10.1109/MSP.2017.2693649



Demystifying Compressive Sensing

Heinrich E. A. Laue

The conventional Nyquist-Shannon sampling theorem has been fundamental to the acquisition of signals for decades, relating a uniform sampling rate to the bandwidth of a signal. However, many signals can be compressed after sampling, implying a high level of redundancy. The theory of compressive sensing/sampling (CS) presents a sampling framework based on the 'rate of information' of a signal and not the bandwidth, thereby minimising redundancy during sampling. This means that a signal can be recovered from far fewer samples than conventionally required.

RELEVANCE

When first exposed to a field and especially one which challenges existing paradigms, it is often useful to start by gaining a high-level understanding of the principles underlying that field. This article employs a new set of analogies, illustrations and numerical examples to provide intuitive explanations for the fundamental principles in CS. Armed with a feel for *why* CS makes sense, the interested reader may then proceed to more technical introductions to the field, such as [1]–[3].

PREREQUISITES

The body of this article requires knowledge of linear algebra, conventional sampling theory, basic probability, and basic optimisation.

ANALOGY 1—LISTENING WITH HALF AN EAR

Three journalists are taking notes at a press conference. The first is inexperienced and quite naïve; so afraid of missing something important, he frantically writes down every word being said. The second journalist is more experienced and while also listening attentively to every word, he interprets what is said and summarises the facts concisely in his notes. The third journalist is quite lazy; not listening attentively at all, he only picks up every second or third word being said. The experienced journalist sees the lazy one daydreaming and is greatly surprised afterwards to find their notes almost identical. 'How did you get all the facts, when you were clearly only listening with half an ear?', he asks. 'Did you not know', the lazy one replies, 'that the speaker is known to waffle, using ten words to convey a single concept? I am not likely to miss anything important when I know he could not be saying very much, though his words may be many.'

The naïve journalist represents conventional Nyquist-rate sampling, with words analogous to samples. This scheme makes no effort to interpret what is being sampled; the sampling rate is based purely on the signal bandwidth. The experienced journalist represents conventional signal compression. Sampling is still done at the Nyquist rate, but the system interprets the sampled signal and expresses it more concisely for storage or transmission. The lazy, or rather, the efficient journalist represents compressive sampling. This scheme samples well below the Nyquist rate by assuming that the unique concepts or information being conveyed is little, and that this information is distributed over the many conventional samples so that missing a particular sample is unlikely to lead to significant loss of information.

ANALOGY 2—FILLING IN THE GAPS

Imagine playing a game where a friend chooses a word and reveals the letters and their positions one by one, at random, until you correctly identify all the letters. You have gotten as far as 'co–pr–ss'. You see that the answer must be 'compress', and that the missing letters are 'm' and 'e'. How did you identify these two letters? You did not see the letters as unrelated parts to be identified on their own, but saw them as collectively conveying a single concept—a word.

Compressive sampling is analogous to such a game, where letters represent samples. Instead of seeing samples as unrelated to each other, CS identifies the 'concepts' the samples are collectively conveying. These concepts are few compared to the number of samples, like many letters convey a single word. Some samples required by the Nyquist theorem are missing, but by identifying the concepts being conveyed, the CS algorithm is able to fill in the missing samples.

On the other hand, imagine playing the game without the assumption that the letters form a word. This significantly increases the complexity of the problem, since any combination of letters is a possible solution. Without any vocabulary to draw from, there is no way to identify the missing letters and you will have to wait until all the letters have been revealed.

Conventional sampling is like the latter scenario. It does not interpret what the samples are conveying, but treats each sample as an individual 'concept' to be identified correctly. All the Nyquist samples must be taken; there is no way of filling in any gaps.

PROBLEM STATEMENT

SPARSIFYING BASES

Consider a discrete signal vector $\mathbf{x} = [x_1, x_2, \ldots, x_N]^{\mathrm{T}}$ of N samples taken at the conventional

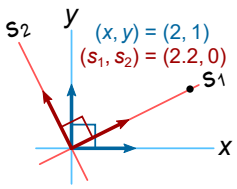


Fig. 1: Illustration of a two-dimensional data point expressed in standard and sparsifying bases.

Nyquist rate of twice the signal bandwidth. A signal vector in a conventional sampling domain can be expressed in a different domain/basis for analysis or processing. For example, time domain signals may be transformed to the Fourier domain to analyse their frequency content.

Consider a set of orthonormal basis vectors placed as columns in the *transform matrix* $\mathbf{\Psi} = [\boldsymbol{\psi}_1\ \boldsymbol{\psi}_2\ \cdots\ \boldsymbol{\psi}_N]$. The signal vector $\mathbf{x}$ can then be expressed as a weighted sum of basis vectors [1]

$$\mathbf{x} = \sum_{n=1}^{N} s_n \boldsymbol{\psi}_n = \mathbf{\Psi}\mathbf{s} \tag{1}$$

where the $N \times 1$ vector $\mathbf{s}$ contains the coefficients of the signal in its new basis, found as the projection of $\mathbf{x}$ onto each of the basis vectors by the dot product $s_n = \langle \boldsymbol{\psi}_n, \mathbf{x} \rangle$, or $\mathbf{s} = \mathbf{\Psi}^{\mathrm{T}}\mathbf{x}$. Each coefficient is represented by its own basis vector which is separable from all others.

Sampling at the Nyquist rate guarantees perfect recovery of the original signal, suggesting that no fewer than N coefficients are required to fully describe the signal. However, a signal vector can often be expressed in a different basis where many coefficients are zero (or close to zero) [2]. The remaining K non-zero, or *significant*, coefficients are sufficient to fully describe the signal. When $\mathbf{x}$ is expressed in a *sparsifying basis*, it results in a K*-sparse* vector $\mathbf{s}$ with only $K \ll N$ significant coefficients. Sparsity in $\mathbf{s}$ implies redundancy in $\mathbf{x}$, since N samples represent a signal with effectively only K degrees of freedom. Here, the significant coefficients in the sparsifying basis can be seen as the 'concepts' or 'information' being conveyed by $\mathbf{x}$.

Figure 1 illustrates a two-dimensional data point at $(2, 1)$ in the standard basis. However, in the sparsifying basis shown, the data point is $(2.2, 0)$, which has only one significant coefficient.

Many signals have sparsifying bases, a fact well known in conventional compression where a signal may be expressed in a sparsifying basis so only the largest coefficients can be retained [1].

Images are typically sparse in the discrete cosine transform (DCT) or wavelet bases [4], audio signals in the modified discrete cosine transform (MDCT) basis [5], magnetic resonance (MR) images in the spatial, spatial finite differences or wavelet domains [6], and sensor array data in direction of arrival (DoA) [7].

INCOHERENCE AND COMPRESSIVE SAMPLING

CS exploits redundancy to reduce the number of samples that must be taken to fully describe a signal. We have seen that redundancy can be quantified in terms of the number of significant coefficients in a sparsifying basis. CS aims to reduce the required number of samples without any prior knowledge of the signal, only the assumption that some sparsifying basis exists.

Imagine as a first approach that you decide to simply neglect some of the conventional Nyquist samples at random. The problem here is that some samples that you neglect may contain crucial information. For example, what if the signal contains spikes or sharp discontinuities? Neglecting samples in these areas would lead to significant loss of information.

In order for this approach to work, it would be necessary for the information to be distributed evenly over all the conventional samples, so that no one sample conveys significantly more information than the rest. While the signal should be sparse in some other basis, it should certainly *not* be sparse in the domain in which it is sampled.

The *incoherence* between two domains expresses the idea that a vector which is sparse in one domain will be non-sparse in the other, and occurs when the basis vectors between the domains are dissimilar [2]. Consider the time and frequency domains, where the time domain is represented by the standard basis and the frequency domain by the Fourier basis. A single frequency component will result in a sine wave with most time samples being significant. Similarly, a time-domain impulse can only be represented by a multitude of frequency components. The bottom of Figure 2 shows an example where $\mathbf{\Psi}$ is the inverse DCT matrix. The blue arrow represents a matrix multiplication with $\mathbf{\Psi}$ or its inverse, depending on the direction. With only two significant coefficients in $\mathbf{s}$, almost all conventional samples in $\mathbf{x}$ are significant.

If the bases of $\mathbf{x}$ and $\mathbf{s}$ are incoherent, and if $\mathbf{s}$ is sparse, then the information of interest (the K significant coefficients in $\mathbf{s}$) will be distributed over all N samples in $\mathbf{x}$. Since this information is comparatively little ($K \ll N$), neglecting some of the samples in $\mathbf{x}$ is unlikely to lead to a significant loss of information. Also, the choice of which samples to neglect becomes almost arbitrary, as long as enough are kept [2]. Aliasing is not a problem when the samples that are kept are not spaced uniformly; therefore, samples are typically neglected at random [2].

So far we require sparsity in $\mathbf{s}$, and non-sparsity in $\mathbf{x}$, which is met when there is incoherence between the conventional and sparsifying bases. But what happens if $\mathbf{x}$ itself is sparse? Redundancy is still present, but neglecting samples at random will result in a loss of information. The

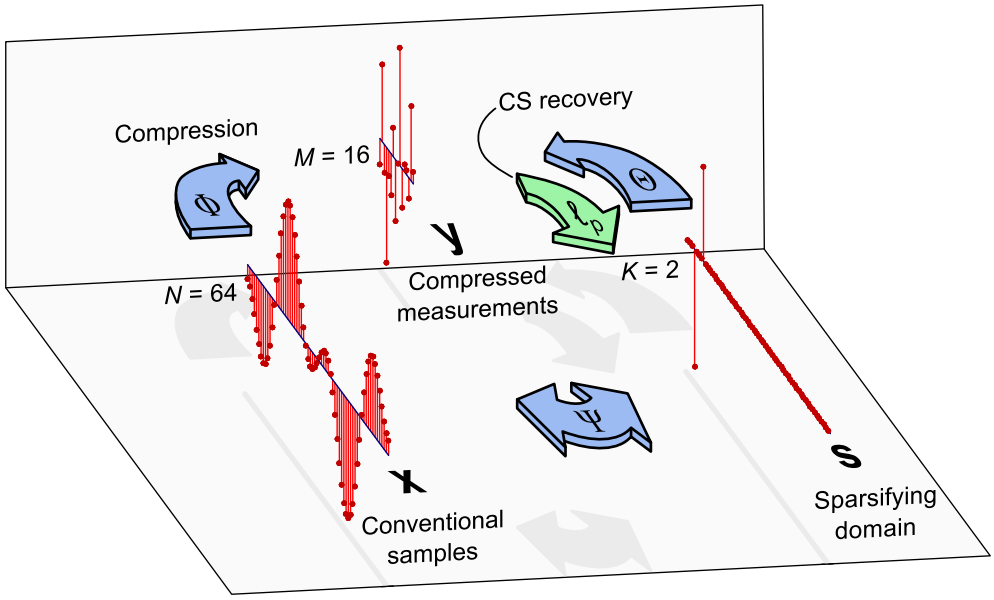


Fig. 2: Example of a CS system with $N = 64$, $M = 16$, $K = 2$, random Gaussian $\mathbf{\Phi}$, and $\mathbf{\Psi}$ the inverse DCT matrix. Blue arrows represent matrix multiplications. Amplitudes are not to scale.

solution is to transform the signal $\mathbf{x}$ to an *intermediate domain* before sampling—a domain in which the signal is non-sparse, and which is incoherent with the sparsifying domain.

How will we select such an intermediate domain? An interesting fact is that a random basis is highly likely to be incoherent with almost any other basis [2]. This means that without prior knowledge of the sparsifying domain, we may transform the signal to a random domain and be sure that the incoherence requirement will be met. In this way we can design *universal* sampling schemes requiring only the assumption that *some* sparsifying basis exists. No knowledge of the sparsifying basis is required during sampling, only during signal reconstruction.

How is compressive sampling implemented? The simultaneous process of transforming to an intermediate domain and neglecting transformed samples may be described by the system $\mathbf{y} = \mathbf{\Phi x}$, where $\mathbf{y}$ is the $M \times 1$ vector of compressed samples (or *measurements*), and $\mathbf{\Phi}$ is the rectangular $M \times N$ *sensing matrix*, with $K < M < N$. The rows $\{\boldsymbol{\phi}_m\}_{m=1}^{M}$ describe the basis vectors of the intermediate domain, where only M out of N basis vectors are used. Each measurement is found as the projection of $\mathbf{x}$ onto the corresponding basis vector as $y_m = \boldsymbol{\phi}_m \mathbf{x}$. Figure 2 illustrates how a random $\mathbf{\Phi}$ is used to compress $\mathbf{x}$ into the measurement vector $\mathbf{y}$.

Each measurement y_m is a unique weighted combination of all the elements in $\mathbf{x}$, or $y_m = \sum_{n=1}^{N} \phi_{m,n} x_n$. Notice how this ensures that the information in $\mathbf{x}$ is distributed over all the

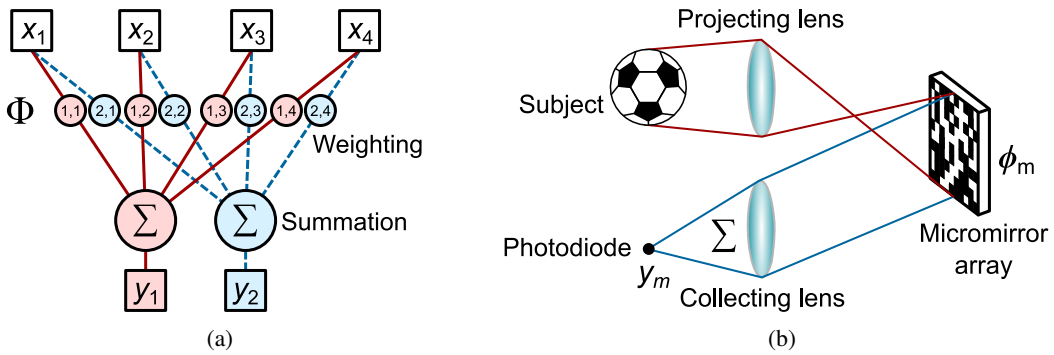


Fig. 3: (a) General illustration of a sensing matrix implementation ($N = 4$, $M = 2$). (b) Single-pixel CS camera (adapted from [1]).

measurements in $\mathbf{y}$. The processes of *weighting* and *summation* are used for hardware implementations of the sensing matrix. This is illustrated in Figure 3(a) for $N = 4$ and $M = 2$.

While a form of CS may be implemented in software to compress Nyquist-rate signals, the real power of CS lies in developing new hardware-based sub-Nyquist sampling schemes.

Figure 3(b) illustrates a CS application, the single-pixel camera, where an N-pixel image is reconstructed from $M < N$ measurements [1], [4]. An image is projected onto a digital micromirror device (DMD) which is an array of tiny mirrors, each representing a pixel. For each measurement y_m, the micromirrors are randomly set to either reflect light toward a collecting lens, or away from it. This is the process of weighting (by zero or one), with the weights for each measurement taken from the corresponding row in $\mathbf{\Phi}$. The lens then collects the rays from the DMD and concentrates it onto a single photodiode—the process of summation. After the first measurement, new weights are programmed into the DMD for the next measurement.

SOLUTION

EQUIVALENT SYSTEMS AND DESIGN REQUIREMENTS

So far we have seen that CS is possible when a signal has a sparse representation in some domain which is incoherent with the sampling domain. To understand the problem in more detail it is useful to consider it from a slightly different angle. Consider the effect of $\mathbf{\Phi}$ and $\mathbf{\Psi}$ not separately, but together in the following equivalent formulation of the CS problem:

$$\mathbf{y} = \mathbf{\Phi x} = \mathbf{\Phi \Psi s} = \mathbf{\Theta s} \tag{2}$$

where $\mathbf{\Theta}$ is the $M \times N$ *compressed transform matrix*. The columns $\{\boldsymbol{\theta}_n\}_{n=1}^{N}$ are vectors in an overcomplete 'basis', with more vectors than dimensions. Each vector still represents a coefficient in $\mathbf{s}$, but the set cannot be orthogonal and there will be some similarity between the vectors.

The goal is to recover $\mathbf{s}$ from $\mathbf{y} = \mathbf{\Theta s}$. Notice that the system is underdetermined since there are less equations than unknowns. It would generally be impossible to determine $\mathbf{s}$ uniquely since we do not have sufficient information about it from $\mathbf{y}$. However, the assumption of sparsity in $\mathbf{s}$ is the additional 'information' we require to obtain a unique solution. Out of the infinite number of possible solutions, we choose to consider only the *sparsest solution*, i.e. the sparsest vector $\mathbf{s}$ that satisfies $\mathbf{y} = \mathbf{\Theta s}$. We are applying a form of Occam's razor: out of all the possible explanations for the measured data, we assume that the simplest (sparsest) one must be correct.

Figure 2 shows the relationship between $\mathbf{\Phi}$, $\mathbf{\Psi}$ and $\mathbf{\Theta}$. The arrow labelled 'CS recovery' indicates that using a CS algorithm, we are able to go 'upstream' by finding the length-N vector $\mathbf{s}$ from the smaller length-M vector $\mathbf{y}$. Having $\mathbf{s}$, we can find the original vector $\mathbf{x}$ through $\mathbf{\Psi}$.

Consider the following example. An underdetermined system in the form $\mathbf{y} = \mathbf{\Theta s}$ is given by

$$\begin{bmatrix} -0.357 \\ 0.612 \\ 0.137 \end{bmatrix} = \begin{bmatrix} -1.036 & -0.293 & -0.127 & -0.503 & -0.127 \\ -0.844 & 0.088 & -0.385 & 0.105 & 1.048 \\ 0.241 & -0.069 & -0.444 & 0.733 & 0.412 \end{bmatrix} \begin{bmatrix} 0 \\ 1.0 \\ 0 \\ 0 \\ 0.5 \end{bmatrix} \tag{3}$$

where a random Gaussian sensing matrix $\mathbf{\Phi}$ with zero mean and variance $1/N$ and the inverse DCT transform matrix $\mathbf{\Psi}$ have been used to obtain $\mathbf{\Theta} = \mathbf{\Phi\Psi}$.

Assuming for now that we know which coefficients in $\mathbf{s}$ are significant, notice how the system may be reduced to an equivalent *overdetermined* sub-system

$$\begin{bmatrix} -0.357 \\ 0.612 \\ 0.137 \end{bmatrix} = \begin{bmatrix} -0.293 & -0.127 \\ 0.088 & 1.048 \\ -0.069 & 0.412 \end{bmatrix} \begin{bmatrix} 1.0 \\ 0.5 \end{bmatrix} \tag{4}$$

in the form $\mathbf{y} = \mathbf{\Theta}'\mathbf{s}'$, where the columns in $\mathbf{\Theta}$ corresponding to the non-zero elements in $\mathbf{s}$ have been extracted to give $\mathbf{\Theta}'$, and $\mathbf{s}'$ contains the non-zero coefficients in $\mathbf{s}$.

Under which conditions can $\mathbf{s}$ be recovered from $\mathbf{y}$? Notice firstly that the columns in $\mathbf{\Theta}'$

should be linearly independent, since the coefficients in $\mathbf{s}$ will not be separable if they can be expressed as a linear combination of the vectors corresponding to the other coefficients. This condition would have sufficed if we knew the K locations beforehand. However, we do not, and must in addition ensure that there is only one K-sparse solution to choose—a *unique solution*.

Suppose for the sake of contradiction that two distinct K-sparse solutions $\mathbf{s}$ and $\hat{\mathbf{s}}$ exist, such that $\mathbf{y} = \mathbf{\Theta s} = \hat{\mathbf{y}} = \mathbf{\Theta}\hat{\mathbf{s}}$, or $\mathbf{\Theta}(\mathbf{s} - \hat{\mathbf{s}}) = \mathbf{0}$. Then the difference vector $\boldsymbol{\delta} = \mathbf{s} - \hat{\mathbf{s}}$ is at most $2K$-sparse [8]. By definition, the equivalent sub-system $\mathbf{\Theta}'\boldsymbol{\delta}' = \mathbf{0}$ has a non-trivial solution if and only if the $2K$ or less columns in $\mathbf{\Theta}'$ are linearly dependent. Conversely, if these columns are linearly independent, $\mathbf{\Theta}(\mathbf{s} - \hat{\mathbf{s}}) = \mathbf{0}$ cannot be satisfied and no more than one K-sparse solution can exist. To guarantee a unique solution for *any* combination of K or less significant coefficients in $\mathbf{s}$, all subsets of $2K$ columns in $\mathbf{\Theta}$ must be linearly independent [8].

The restricted isometry property (RIP) goes a step further and considers whether sub-matrices with $2K$ columns are *nearly* linearly dependent, by placing bounds on the conditioning of these sub-matrices [9]. This ensures robustness in the presence of noise, since small perturbations may produce large errors when solving a nearly linearly dependent system.

Compressed transform matrices $\mathbf{\Theta}$ with independent and identically distributed (i.i.d.) random entries have been shown to meet the RIP criterion with high probability. Examples include Gaussian matrices with zero mean and variance $1/M$, and Bernoulli matrices with equiprobable $\pm 1/\sqrt{M}$ entries [2]. For practical implementations $\mathbf{\Phi}$ may be chosen in the same way as $\mathbf{\Theta}$ (replacing M with N), and the resulting $\mathbf{\Theta}$ will still be able to meet the criteria for arbitrary choice of $\mathbf{\Psi}$ [2]. As a result, the incoherence requirement between $\mathbf{\Phi}$ and $\mathbf{\Psi}$ will also be satisfied.

MINIMISING SPARSITY—PERFECT RECOVERY ALGORITHMS

We have seen how it is possible for a CS system to preserve the information in $\mathbf{x}$ from only the measurements $\mathbf{y}$. But how will one go about recovering $\mathbf{x}$ from $\mathbf{y}$?

In developing a CS recovery algorithm, our aim is to obtain a unique solution to an underdetermined system which is the sparsest of all solutions. This can be formulated as

$$\min \|\tilde{\mathbf{s}}\|_0 \quad \text{subject to} \quad \mathbf{y} = \mathbf{\Theta}\tilde{\mathbf{s}} \tag{5}$$

where $\|\cdot\|_0$ is the ℓ_0-'norm' defined as the number of non-zero elements in a vector. The problem reads: 'Minimise the number of non-zero elements in $\tilde{\mathbf{s}}$, subject to $\tilde{\mathbf{s}}$ being a possible solution to the system'. The general ℓ_p-norm is defined as $\|\mathbf{s}\|_p = \sqrt[p]{\sum_{n=1}^{N} |s_n|^p}$.

To illustrate, consider the following combinatorial ℓ_0 minimisation algorithm:

for $k \leftarrow 1$ **to** $M-1$

  for all combinations of k out of N coefficient locations in $\tilde{\mathbf{s}}$

    Find the least-squares solution $\hat{\mathbf{s}}'$ to $\mathbf{y} = \mathbf{\Theta}'\tilde{\mathbf{s}}'$

    if $\hat{\mathbf{y}} = \mathbf{\Theta}'\hat{\mathbf{s}}' = \mathbf{y} \longrightarrow$ **break**

This algorithm tries all possible combinations of sparse coefficient locations, starting with a single one, until it finds an exact least-squares solution to the sub-system.

Consider the system in (3) with $\mathbf{s}$ unknown. The first try would assume that only s_1 is significant and solve the resulting sub-system, giving the least-squares solution $s_1 = -0.062$. This gives $\hat{\mathbf{y}} = [0.064,\, 0.052,\, -0.015]^{\mathrm{T}} \neq \mathbf{y}$, which is not an exact solution.

Eventually we reach the combination (s_2, s_5), which leads to the sub-system in (4) with $\mathbf{s}'$ unknown, for which the least-squares solution is $(s_2, s_5) = (1.0, 0.5)$, giving $\hat{\mathbf{y}} = [-0.357,\, 0.612,\, 0.137]^{\mathrm{T}} = \mathbf{y}$, an exact solution. This also happens to be the first combination which gives an exact solution, as expected. While there are many more solutions, the algorithm accepts this one as correct since it is sparsest. If desired, we can now calculate $\hat{\mathbf{x}} = \mathbf{\Psi}\hat{\mathbf{s}}$.

Unfortunately, a combinatorial ℓ_0 algorithm is computationally infeasible for problems of practical sizes [1]. For a practically feasible recovery algorithm we use the ℓ_p-norm formulation

$$\min \|\tilde{\mathbf{s}}\|_p \quad \text{subject to} \quad \mathbf{y} = \mathbf{\Theta}\tilde{\mathbf{s}}, \quad 0 < p \leq 1 \tag{6}$$

which can be solved using a variety of efficient optimisation algorithms [10]. The ℓ_1 problem can be recast as a linear program [3], [8], and requires around $O(K \log(N/K))$ measurements when using a random sensing matrix [2]. For $0 < p < 1$ the problem is non-convex with multiple minima, but local optimisers nevertheless perform well, particularly when $p = 0.5$ [10].

To visualise why the ℓ_p-norm favours sparsity when $0 < p \leq 1$, consider the graphs in Figure 4 for $p = 0.5$, $p = 1$, and $p = 2$. The ℓ_p balls shown in blue represent the points at which $\tilde{\mathbf{s}}$ has constant ℓ_p-norm. Since the norms must be minimised, imagine the balls being inflated until they first touch the lines of possible solutions shown in red. For $0 < p \leq 1$, the ℓ_p-norms favour sparse solutions that lie on the axes. By comparison, minimising the common Euclidean norm ($p = 2$) is not useful since it does not obtain a solution which is necessarily sparse.

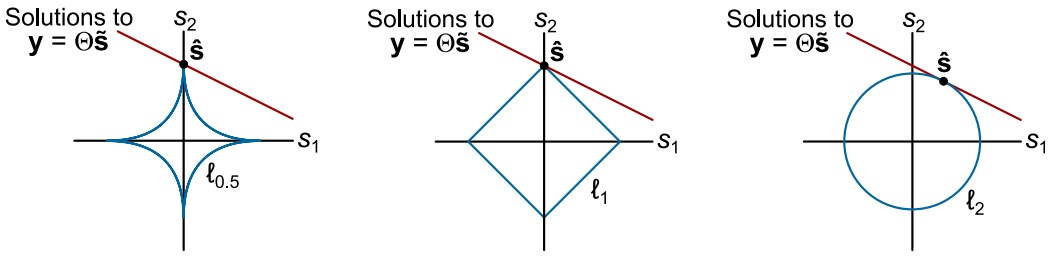


Fig. 4: Illustration of ℓ_p-norms for (a) $p = 0.5$, (b) $p = 1$, and $p = 2$ (Euclidean norm). Blue lines occur where the vector $\mathbf{s}$ has constant ℓ_p-norm. Any point on a red line is a possible solution.

CONCLUSION

In this article, we learned how CS exploits redundancy to reduce sampling rates, and saw under which conditions the original signal is preserved despite the system being underdetermined. We saw how universal sampling schemes requiring the existence but not knowledge of a sparsifying basis may be developed by using random sensing matrices. Lastly, we saw how an ℓ_0 minimisation algorithm can recover a signal with knowledge of the sparsifying basis, and considered the rationale behind ℓ_p minimisation algorithms for $0 < p \leq 1$.

ACKNOWLEDGEMENTS

The author would like to thank Prof. Warren du Plessis, Prof. Pieter Jacobs, Marius-Corné Meijer, Wade Smith and Winston Jimu for many insightful discussions. The author would also like to thank the anonymous reviewers for their input which helped explain certain concepts better.

AUTHOR

Heinrich E. A. Laue (laueheinrich@gmail.com) is a postgraduate student with the Department of Electrical, Electronic and Computer Engineering at the University of Pretoria, South Africa. His research focuses on developing compressive feed networks for antenna arrays.

REFERENCES

[1] R. G. Baraniuk, "Compressive sensing," *IEEE Signal Process. Mag.*, vol. 24, no. 4, pp. 118–121, 2007.

[2] E. J. Candes and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

[3] M. Fornasier and H. Rauhut, "Compressive sensing," in *Handbook of Mathematical Methods in Imaging*, O. Scherzer, Ed. New York, NY: Springer New York, 2011, pp. 187–228.

[4] M. F. Duarte, M. A. Davenport, D. Takhar, J. N. Laska, T. Sun, K. E. Kelly, and R. G. Baraniuk, "Single-pixel imaging via compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 83–91, Mar. 2008.

[5] M. D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M. E. Davies, "Sparse representations in audio and music: From coding to source separation," *Proc. IEEE*, vol. 98, no. 6, pp. 995–1005, Jun. 2010.

[6] M. Lustig, D. Donoho, and J. M. Pauly, "Sparse MRI: The application of compressed sensing for rapid MR imaging," *Magnetic Resonance in Medicine*, vol. 58, no. 6, pp. 1182–1195, Dec. 2007.

[7] A. Massa, P. Rocca, and G. Oliveri, "Compressive sensing in electromagnetics—a review," *IEEE Antennas Propag. Mag.*, vol. 57, no. 1, pp. 224–238, Feb. 2015.

[8] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization," *Proc. Nat. Acad. of Sci.*, vol. 100, no. 5, pp. 2197–2202, Mar. 2003.

[9] S. Foucart, "A note on guaranteed sparse recovery via ℓ_1-minimization," *Appl. and Computational Harmonic Anal.*, vol. 29, no. 1, pp. 97–103, Jul. 2010.

[10] Z. Zhang, Y. Xu, J. Yang, X. Li, and D. Zhang, "A survey of sparse representation: Algorithms and applications," *IEEE Access*, vol. 3, pp. 490–530, May 2015.