Extracting Salient Features From Less Data via ℓ_1-Minimization

Wotao Yin and Yin Zhang
Department of Computational and Applied Mathematics,
Rice University, Houston, Texas 77005, USA.
(wotao.yin@ and yzhang@rice.edu).

MRI (magnetic resonance imaging) is a widely used medical imaging modality that creates an image from scanned data that are essentially the Fourier coefficients of this image. A typical abdominal scan may take around 90 minutes. Can we reduce this time to 30 minutes by using one third of the Fourier coefficients, while maintaining image quality? In this article, we hope to convince the reader that such reductions are achievable through a new and promising approach called *compressive sensing* (or compressed sensing). The main computational engine that drives compressive sensing is ℓ_1-related minimization algorithms.

1. Introduction

Exploiting sparsity is a common task in computational sciences, as it is in signal processing. Recently, sparsity has been skillfully utilized to increase data acquisition capacity in a new approach called *compressive sensing.* Seminal contributions in this area include Candés and Tao [3] and Donoho [11]. In a nutshell, this technique encodes a sparse signal into a shorter "code" whose length is roughly proportional to the sparsity level rather than the length of the signal. The decoding process, on the other hand, involves solving an optimization problem. This is very different from the traditional paradigm where a full-length code is first acquired, then compressed, and the decoding process is relatively inexpensive. This paradigm shift can potentially bring great benefits to certain applications. However, solving large-scale optimization problems arising from compressive sensing poses real challenges.

1.1 A synthetic example

Let us try to acquire a sparse signal $\bar{x} \in \mathbb{R}^n$ of length $n = 200$ depicted in Figure 1(a). Let $k = \|\bar{x}\|_0$ be the number of nonzeros in $\bar{x}$, which is 10. First, $\bar{x}$ is *encoded* into a "*compressed code*" $b = R\bar{x} \in \mathbb{R}^m$, $m < n$, by a linear transform R. Typically in signal acquisition practice, such encoding is not calculated on a computer but obtained by certain physical or digital means. Notice that since $\bar{x}$ is "unknown" at this time, R can only be constructed independently (non-adaptively) of $\bar{x}$. In this synthetic example, we let $R \in \mathbb{R}^{m\times n}$ be formed from a subset of $m = 80$ rows of the n-dimensional discrete cosine transform (DCT) matrix Φ. The number m is called the *sample size.* $\Phi\bar{x}$ and $R\bar{x}$ are depicted in Figures 1(b) and (c) where those in $\Phi\bar{x}$ but not in $R\bar{x}$, *i.e.*, the missing measurements, are replaced with zeros in (c). After the compressed code $b = R\bar{x}$ is acquired by a sensor and becomes available, we need to *decode* it to recover the original signal. That is where optimization enters the picture. Although the linear equations $Rx = b$ have an infinite number of solutions because $m < n$, one may use the fact that the number of nonzeros in $\bar{x}$, $\|\bar{x}\|_0$, is small and try to recover $\bar{x}$ as the solution to the ℓ_0-problem:

$$\min_{x\in I\!R^n} \{\|x\|_0 : Ax = b\} \tag{1}$$

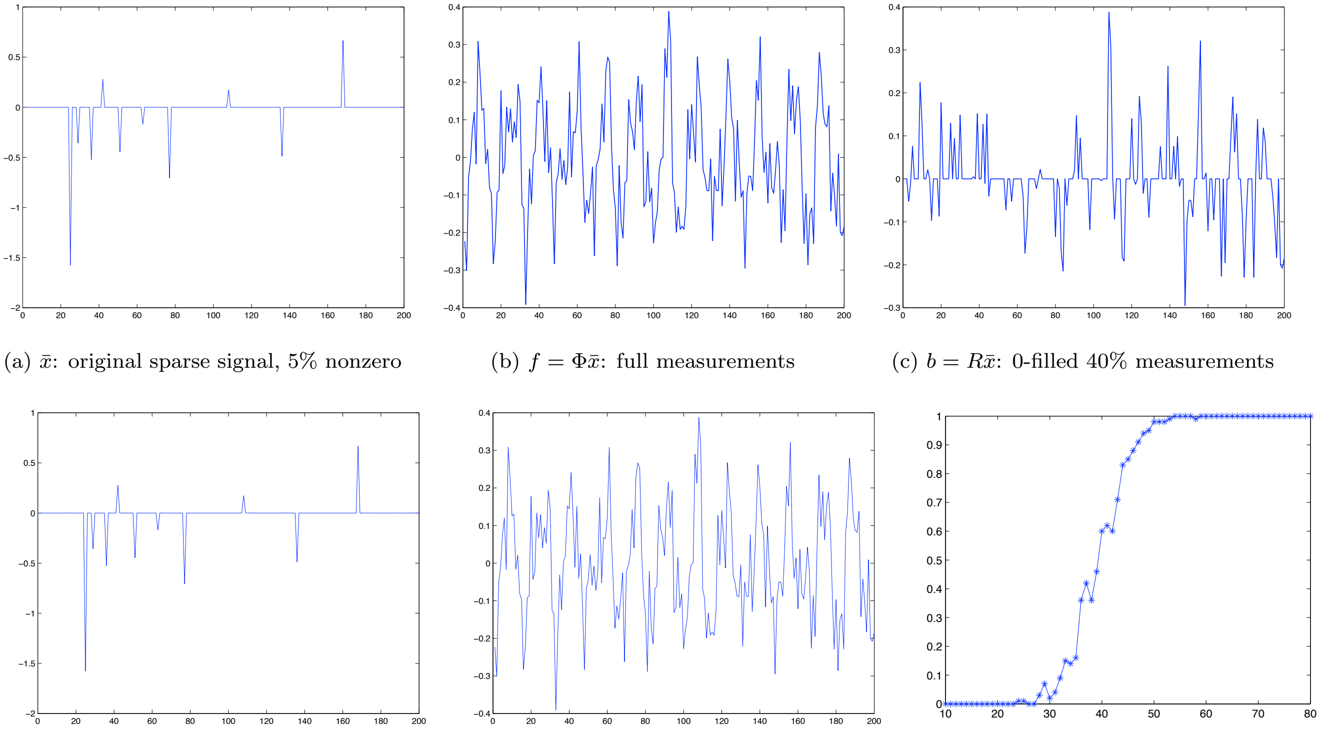
for $A = R$, where the "ℓ_0-norm" of x is the number of nonzeros in x. The solution of (1) will be $\bar{x}$ unless there exists another solution to $Rx = b$ that is equally sparse or sparser than $\bar{x}$ (which does not happen under favorable conditions; see next section). However, the ℓ_0-problem (1) is combinatorial and generally NP-hard [26]. A much more tractable alternative is the ℓ_1-problem (also called *basis pursuit*):

$$\min_{x\in I\!R^n} \{\|x\|_1 : Ax = b\}, \tag{2}$$

which is a convex program that always has a solution whenever $Ax = b$ is consistent. As we will show in Section 2, problem (2) yields the same solution as the ℓ_0-problem under some mild conditions,

From a different perspective, this is also an example of missing data recovery [35]. Given a portion of data b (Figure 1(c)) that is known, one can recover the complete data f (Figure 1(b)) by exploiting the sparsity of $\bar{x}$ representing f under a basis Φ, *i.e.*, $\Phi\bar{x} = f$. Specifically, solving (2), for A equal to the sub-matrix of Φ corresponding to b, gives the optimal solution $x_{\text{opt}} = \bar{x}$ so that the original signal is reconstructed as $f = \Phi x_{\text{opt}}$ (Figure 1(e)).

Ideally, we would like to take the smallest number of measurements possible, that is, $m = \bar{k} \equiv \|\bar{x}\|_0$.



(a) $\bar{x}$: original sparse signal, 5% nonzero

(b) $f = \Phi\bar{x}$: full measurements

(c) $b = R\bar{x}$: 0-filled 40% measurements

(d) x_{opt}: signal recovered by (2)

(e) Φx_{opt}: recovered measurements

(f) Recovery rate for m =10 to 80

Figure 1: Signal recovery from under-sampled measurement. Φ is a discrete cosine transform.

However, we must pay a price for not knowing the locations of the nonzeros in $\bar{x}$ (there are "n choose $\bar{k}$" possibilities!). It was shown in [3, 4, 30] that, when R is a certain random matrix, then by solving (2) for $m = O(\bar{k}\log(n/\bar{k}))$ one can recover $\bar{x}$ with high probability. Though larger than $\bar{k}$, such an m can be asymptotically much smaller than n when $\bar{k} \ll n$.

To illustrate this point, we performed similar calculations depicted above for $m = 10, 11, \ldots, 80$, each with 100 repetitions of randomly chosen m measurements. The percentages of successful recovery for all m are plotted in Figure 1(e), which shows that it is generally safe to have $m > 6\|\bar{x}\|_0 = 60$ for this combination of $\|\bar{x}\|_0$, n, and A.

1.2 Hidden sparsity

If compressive sensing were only applicable to spatially and temporally sparse signals, it would have few applications. Most images, for example, are not sparse in the pixel domain, but rather have sparse representations in either the Fourier (spectral) or Wavelet (spectral-scale) domain. Let the vector $\bar{u}$ represent such an image. A compression algorithm (*e.g.* JPEG2000) would find an invertible matrix Φ (*e.g.*, a wavelet basis) such that the vector $\bar{x} = \Phi\bar{u}$ has a relatively small number of large-magnitude components. Let $\tilde{x}$ be the sparse vector formed by taking only the large-magnitude components of $\bar{x}$. Then, $\bar{u}$ can be accurately approximated by $\Phi^{-1}\tilde{x}$. This is not surprising because the useful information in most images is relatively sparse compared to pixel values. Since $\Phi\bar{u}$ is sparse, one can recover $\bar{u}$ from $b = R\bar{u}$ by solving

$$\min_u\{\|\Phi u\|_1 : Ru = b\}, \tag{3}$$

which is equivalent to solving (2) with $x = \Phi u$ and $A = R\Phi^{-1}$.

Like images, many signals by their nature are sparse in certain domains. The principle of compressive sensing is that such a signal can be recovered from a relatively small number of measurements provided its sparsity is appropriately exploited. However, a good sparse representation for a given signal is not always obvious. Recently, some progress has been made on signals arising from low-light imaging, medical imaging (MRI and CT), infra-red sensing, bio-sensing, radar signal processing, multi-sensor networks and distributive sensing, and analog-to-information conversion. The interested reader can visit the Rice compressive sensing website [6] for a list of recent papers.

2. When are the ℓ_0- and ℓ_1-problems equivalent?

We give an informal proof of the fact that whenever A is random, $\bar{x}$ is sufficiently sparse and $b = A\bar{x}$, then with high probability $\bar{x}$ will solve the "basis pursuit" problem (2). Following the proof in [34], we will use a classic result developed by Kashin [21], and Garnaev and Gluskin [19].

2.1 A sufficient condition for recovery

We first derive a sufficient condition for $\bar{x}$ to be the unique solution of (2) assuming that $A \in \mathbb{R}^{m\times n}$ has rank m and $m < n$. Let $\bar{x}$ satisfy $A\bar{x} = b$ and denote the null space of A by Null(A). Since

$$\{x : Ax = b\} \equiv \{\bar{x} + v : v \in \mathrm{Null}(A)\},$$

$\bar{x}$ uniquely solves (2) if and only if

$$\|\bar{x} + v\|_1 > \|\bar{x}\|_1,\ \forall v \in \mathrm{Null}(A) \setminus \{0\}. \tag{4}$$

Let S be the support of $\bar{x}$ and Z be its complement, *i.e.*,

$$S = \{i : \bar{x}_i \neq 0\}, \quad Z = \{i : \bar{x}_i = 0\},$$

and v_S be the sub-vector of v corresponding to the index set S (we apply similar notation for other vectors). Then we calculate

$$\begin{aligned}\|\bar{x} + v\|_1 &= \|\bar{x}_S + v_S\|_1 + \|0 + v_Z\|_1 \\ &= \|\bar{x}\|_1 + (\|v_Z\|_1 - \|v_S\|_1) + \\ &\quad (\|\bar{x}_S + v_S\|_1 - \|\bar{x}_S\|_1 + \|v_S\|_1),\end{aligned}$$

where in the right-hand side we have added and subtracted the terms $\|\bar{x}\|_1$ and $\|v_S\|_1$ (noting that $\|\bar{x}\|_1 = \|\bar{x}_S\|_1$ given that $\bar{x}_Z = 0$).

In the above identity, the last term in parentheses is nonnegative by the triangle inequality; hence, $\|\bar{x}+v\|_1 > \|\bar{x}\|_1$ if $\|v_Z\|_1 > \|v_S\|_1$. Therefore, a sufficient condition for $\bar{x}$ to be the unique solution of (2) is that $\|v_Z\|_1 > \|v_S\|_1$, or equivalently $\|v\|_1 > 2\|v_S\|_1$, for all nonzero v in the null space of A. In view of the inequality

$$\|v_S\|_1 \leq \sqrt{|S|}\|v_S\|_2 \leq \sqrt{\|\bar{x}\|_0}\|v\|_2,$$

where we used the facts that (i) the length of v_S is $|S|$ (the cardinality of the set S) which equals $\|\bar{x}\|_0$, and (ii) v_S is a sub-vector of v, we derive another sufficient condition that $\bar{x}$ uniquely solves (2) if

$$\sqrt{\|\bar{x}\|_0} < \frac{1}{2}\frac{\|v\|_1}{\|v\|_2},\ \ \forall v \in \mathrm{Null}(A) \setminus \{0\}. \tag{5}$$

This condition requires nothing but sparsity of $\bar{x}$ for it to solve (2) uniquely. This uniqueness implies that there can exist at most one vector $\bar{x} \in \{x : Ax = b\}$ whose sparsity meets the condition (5); otherwise, it would not be the unique solution of (2). Such an $\bar{x}$, whenever it exists, must be the sparsest solution to $Ax = b$. In other words, *the ℓ_1- and ℓ_0-problems are equivalent* in the sense

$$\begin{aligned}\bar{x} = &\ \arg\min\{\|x\|_1 : Ax = b\} \\ = &\ \arg\min\{\|x\|_0 : Ax = b\}.\end{aligned} \tag{6}$$

The remaining question is how restrictive the condition (5) is? More precisely, how big can the bound on the right-hand side of (5) be? The answer will, of course, depend on the properties of matrix A.

2.2 Kashin-Garnaev-Gluskin result

We will make use of a classic result established in the late 1970's and early 1980's by Russian mathematicians. In our context, this result has to do with the ratio of the ℓ_1-norm to the ℓ_2-norm restricted to a subspace. We know that in the entire space $I\!R^n$, the ratio can vary from 1 to $\sqrt{n}$, namely,

$$1 \leq \frac{\|v\|_1}{\|v\|_2} \leq \sqrt{n}, \quad \forall v \in I\!R^n \setminus \{0\}.$$

Here we will only concern ourselves with the lower bound. Roughly, this ratio is small for sparse vectors that have many zero or near-zero elements. However, it turns out that in many subspaces this ratio can have much larger lower bounds than 1.

As an improvement to an earlier result by Kashin [21], Garnaev and Gluskin [19] established that for any natural number $p < n$, there exist p-dimensional subspaces $V_p \subset I\!R^n$ in which

$$\frac{\|v\|_1}{\|v\|_2} \geq \frac{C\sqrt{n-p}}{\sqrt{\log(n/(n-p))}}, \forall v \in V_p \setminus \{0\}, \quad (7)$$

where C is an absolute constant independent of the dimensions. In other words, these subspaces do not contain excessively sparse vectors. Moreover, such subspaces are abundant because *every p-dimensional subspace spanned by iid (independently identically distributed) random vectors of the standard Gaussian distribution will satisfy inequality (7) with high probability.* (This is an instance of a mathematical phenomenon commonly referred to as *concentration of measure*; see [25], for example.)

2.3 How sparse is enough?

If A is an m by n random matrix with iid standard-Gaussian entries, then it is known that the null space of A can be spanned by iid random vectors. In particular, vectors in the null space of A will satisfy, with high probability, the Garnaev and Gluskin inequality (7) for $V_p = \mathrm{Null}(A)$ and $p = n - m$. Combining the sufficient condition (5) with the Garnaev and Gluskin inequality (7), we have the result that for a random Gaussian matrix A, $\bar{x}$ will uniquely solve (2) with high probability whenever

$$\|\bar{x}\|_0 < \frac{C^2}{4}\frac{m}{\log(n/m)}. \quad (8)$$

(The constant C above is the same one from (7)). This result can be interpreted as follows. As long as the sparsity of a signal $\bar{x}$ is less than a certain fraction of the number of random measurements m, where the value of the fraction only logarithmically deteriorates as the signal dimension n increases, with high probability this signal can be recovered from the random measurements by solving the basis-pursuit problem (2).

The sparsity bound given in (8) is the best order currently available, first established in [3] for Gaussian random matrices, which is a significant improvement upon previously existing bounds. The same order has been extended to some other random matrices such as Bernoulli matrices whose entries are ±1 [4]. For certain partial orthonormal (for example, partial DCT) matrices, a slightly weaker bound has been proved [30]. Moreover, an in-depth study on the constant in (8), $C^2/4$, can be found in [12].

3. Imaging and other applications

To demonstrate the potential benefit of compressive sensing in practical applications, let us simulate a compressed MRI (Magnetic Resonance Imaging) experiment using under-sampled measurements (see [24] for a more realistic work).

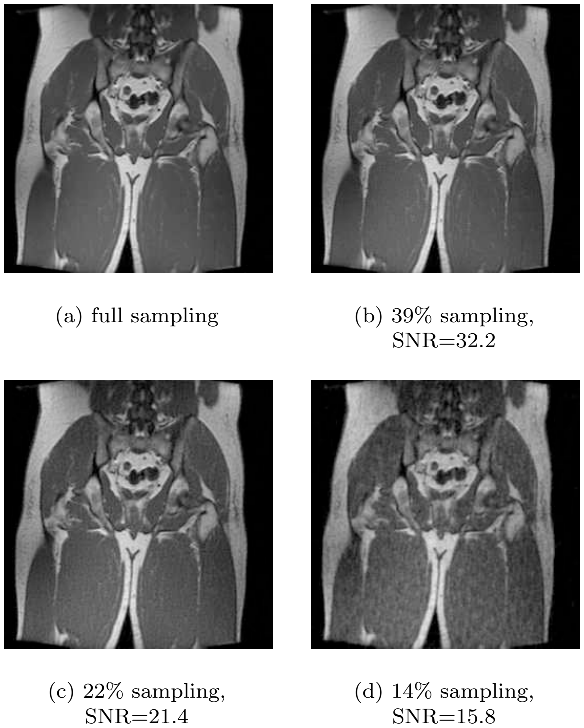
3.1 Compressed MRI simulation

First, we need an abridged overview of MRI — a non-invasive and safe medical imaging technique. In MRI, images are obtained in the form of the frequency response of tissues. First, a strong magnetic field and an RF (radio frequency) pulse are directed to a section of the anatomy, causing the protons in that area to be "excited": they get aligned along the magnetic direction and spin with a certain frequency. Next, on turning off the RF pulse, the protons return to their natural, rather chaotic, state while releasing RF signals that are captured by external coils in the form of phases and magnitudes at selected frequencies. In other words, the image of spatial energy (or density), denoted by $\bar{u}$, is constructed from data acquired in the frequency domain, the so-called k–space. Roughly, at a given resolution, a complete set of sampled frequencies is $f = \mathcal{F}\bar{u}$ where $\mathcal{F}$ is a discrete Fourier transform. Therefore, an image can be constructed through a Fourier inversion $\bar{u} = \mathcal{F}^{-1}f$.

An MRI scan can be a long and uncomfortable process. For example, a patient must repeatedly hold his/her breath during an abdominal scan while strictly immobilized throughout the process, which can last 1 to 2 hours. Potentially, compressive sensing can help construct $\bar{u}$ with a much smaller number of sample frequencies. This would mean that the MRI scan duration could be significantly reduced.

MR images often have sparse representations under some wavelet transform Φ. By solving $\min_u\{\|\Phi u\|_1 : Ru = b\}$ or its variants, we can obtain a given image $\bar{u}$ from an under-sampled frequency set $b = R\bar{u}$, where R represents a partial discrete Fourier transform.

Let us simulate this approach to see how much compressive sensing could help. Figure 2(a) depicts



(a) full sampling

(b) 39% sampling, SNR=32.2

(c) 22% sampling, SNR=21.4

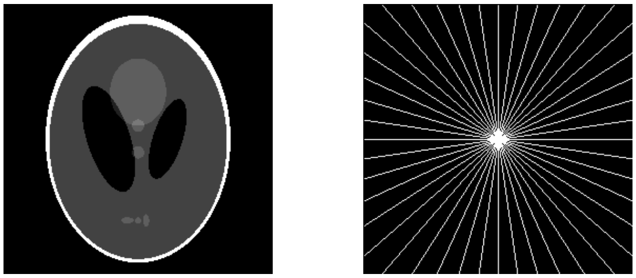
(d) 14% sampling, SNR=15.8

Figure 2: (a) Original Image; (b)-(d) Pelvis MR images recovered from incomplete measurements using the wavelet-based model (9) (where the higher the SNR (signal-noise ratio) is, the better the image quality is).

a clean 256×256 pelvis MR image that is our $\bar{u}$. We tried the use of 39%, 22%, and 14% of its Fourier coefficients to reconstruct the image, simulating three different levels of under-sampling. Since in MR imaging, one has some freedom in selecting sample frequencies (however, practical constraints do exist), the reconstruction results were obtained by random sampling with a strong bias towards lower frequencies. The images in Figure 2(b)–(d) were obtained by solving the model

$$\min_u \|\Phi u\|_1 + \frac{\mu}{2}\|Ru - b\|^2, \qquad (9)$$

with a Haar-wavelet transform Φ and a large μ. From a visual examination, it seems that using one third of measurements, if properly chosen, would be quite sufficient for obtaining a high-quality image for this case.



(a) Shepp-Logan phantom

(b) 22 radial lines

Figure 3: Fourier samples taken at the frequencies on the 22 radial lines (b) are sufficient to exactly recover Shepp-Logan phantom (a) using total variation.

3.2 Total variation

Many natural images possess a “blocky” structure. For such images, minimizing the *total variation* (*cf.* [36]) yields a better image quality [31]. For a 2D digital image u, the total variation of u, $TV(u)$, is defined as the sum of Euclidean norms of local finite differences, *i.e.*, $TV(u) \equiv \sum_{i,j} \|(Du)_{ij}\|_2$, where $(Du)_{ij}$ represents a first-order finite difference vector of u at pixel (i, j). Since $TV(u)$ is the ℓ_1-norm of “gradient magnitude”, minimizing $TV(u)$ tends to yield a solution with sparse finite differences, namely an image with constant-intensity blocks. Therefore, total-variation regularization has been widely used in image processing tasks such as noise removal, deblurring, edge detection, *etc.*. A similar argument can be used to justify the use of higher-order finite differences to regularize images of appropriate characteristics.

In [2], Candés and Romberg demonstrated that, by minimizing the total-variation, the Shepp-Logan phantom in Figure 3(a) can be almost exactly recovered from Fourier samples taken on 22 radial lines depicted in Figure 3(b).

Since finite difference operators are not invertible, minimizing total variation cannot be directly transformed into a problem of the form (2). This poses a major algorithmic challenge.

3.3 Broad applications

There are other potential applications of compressive sensing besides MRI, especially in areas where signal acquisitions are relatively expensive and time-consuming. For example, an infrared sensor is over a hundred times more expensive than an image sensor of the same resolution in a consumer digital camera. In a CT (computed tomography) scan, a series of two-dimensional X-rays are used to construct a three-dimensional image, but a long exposure to the radiation from X-rays can be dangerous. In wireless sensor networks for collecting physical or environmental measurements, a large number of spatially distributed sensors acquire and transmit a deluge of data, relying on low capacity batteries. In all of these examples, physical hardware capacities are being stressed, and improving their sensing resolution or speed is expensive. Compressive sensing offers an invaluable alternative to expensive physical improvements by using much cheaper computing power after data collection. Some recent explorations and applications of compressive sensing can be found at the Rice compressive sensing website [6].

4. Algorithms

Let us turn to optimization, the ultimate tool for obtaining a sparse signal (or its sparse representation) from under-sampled measurements.

4.1 Formulations and challenges

Let $J(x)$ be a convex, sparsity-promoting function, such as the ℓ_1-norm or the total variation. To recover a sparse signal representation $\bar{x}$ from measurements $b \approx A\bar{x}$, we can either solve

$$\min_x \{J(x) : Ax = b\}, \tag{10}$$

when b is relatively accurate, or solve

$$\min_x \{J(x) : H(Ax, b) \le \epsilon\} \tag{11}$$

when b is more noisy, where H is a measure of the closeness of Ax to b. For an appropriate penalty parameter μ (which can be found by a noise-statistics computation, cross validation, or simply trial and error), (11) is equivalent to

$$\min_x J(x) + \mu H(Ax, b) \tag{12}$$

for some $\mu > 0$. The most common choices of J and H are, respectively, $J(x) = \|x\|_1$ and $H(Ax, b) = \frac{1}{2}\|Ax - b\|_2^2$. In Statistics, minimizing this H subject to $\|x\|_1 \le \delta$ is the so-called LASSO problem. More generally, the regularization term $J(x)$ can be a mixture of multiple terms representing multiple features of a sparse solution. For example, a signal may possess a piece-wise constant feature and have a sparse representation under a certain transform Φ at the same time. In this case, we may use a mixed regularization term:

$$J(x) = TV(u) + \lambda\|\Phi x\|_1.$$

Similarly, the fidelity-measure function $H(x)$ could also consist of multiple terms.

All these problems are non-smooth convex optimization problems that can be easily transformed into smooth problems with convex constraints. However, algorithmic challenges arise from the facts that (i) real-world application problems are invariably large-scale (an image of 1024×1024 resolution leads to over a million variables); (ii) the data matrices involved are generally dense; and (iii) real-time or near real-time processing is often required (as in MRI). For these problems, conventional algorithms requiring matrix factorizations are generally not effective or even applicable.

On the other hand, when A is a partial transform matrix, fast matrix-vector multiplications are often available. Moreover, the sparsity in solutions presents unusual opportunities to achieve relatively fast convergence with first-order methods. These features make the development of efficient optimization algorithms for compressive sensing applications an interesting research area.

4.2 Some recent algorithms

We mention a few algorithms recently developed for solving large-scale compressive sensing problems, fully realizing that any such list would be unavoidably incomplete. In addition to those briefly reviewed below, there are many other algorithms based upon ideas such as minimizing a non-convex ℓ_p-"quasi-norm" for $p < 1$, iteratively weighted least squares, group testing, homotopy methods in statistics, combinatorial methods, and ℓ_1-Bregman iterations. We again refer the reader to the Rice CS resource website [6] for more comprehensive lists of algorithmic papers and software.

Orthogonal Matching Pursuit (OMP) based methods (*e.g.*, [32, 13, 7]) do not solve (2) *per se*, but use an iterative greedy approach to identify nonzero (or large-magnitude) components of x so that the residual $b - Ax$ is minimized in some sense while keeping other components of x at zero. The recent algorithm StOMP [13] is a good representative of such greedy algorithms that can perform well on problems with highly sparse solutions and noiseless measurements.

A recent code called ℓ_1_ℓ_s [22] is based on an interior-point algorithm that uses a preconditioned conjugate gradient (PCG) method to approximately solve linear systems in a truncated-Newton framework. The authors exploit the structure of the Hessian to construct their preconditioner. Their computational results show that about a hundred PCG steps are sufficient for obtaining accurate MRI images in the compressive sensing framework. Though generally slower than first-order methods, this algorithm may offer a certain advantage on problems of less sparsity where first-order methods could potentially encounter slow convergence.

The recent method GPSR [18], which stands for gradient projection for sparse reconstruction, reformulates the unconstrained version (12) of (2) into a quadratic program with nonnegativity constraints and applies a projected gradient algorithm, with optional Barzilai-Borwein steps and a non-monotone line search. Although motivated from very different viewpoints, this algorithm has a certain similarity with shrinkage methods introduced below; however, their performance can be quite different on some problems.

SPGL1 [33] is a recent code for solving a sequence of problems of the form

$$\min_x \|Ax - b\|, \quad \text{s.t. } \|x\|_1 \leq \lambda, \tag{13}$$

for $\lambda = \lambda_1, \lambda_2, \ldots, \lambda_j = \bar{\lambda}$ until reaching the desired value $\bar{\lambda}$. The choice of λ is based on a root finding algorithm (e.g., Newton's method) using two results: (i) the curve formed by the minimizers $x_{\text{opt}}(\lambda)$ is convex and continuously differentiable in λ, (ii) the dual solution of (13) gives the gradient of the curve at λ.

Recently, a general method was proposed in [27] for minimizing $J(x) + H(x)$, where J is non-smooth, H is smooth, and both are convex. It is required that J be "simple" so that there exists a closed-form solution to minimizing J plus some auxiliary functions. The ℓ_1-norm is such a "simple" function since the problem $\min_x \lambda\|x\|_1 + \frac{1}{2}\|x - y\|_2^2$ has the closed-form solution shrink(y, λ), which is defined in (16) below. When H has Lipschitz continuous gradients, the objective value in this method converges at a rate $O(k^{-2})$, where k is the iteration number. This result shows that in general, minimizing the sum of J and H is not harder than minimizing the smooth function H alone as long as J is "simple".

A widely used method for solving ℓ_1-minimization problems of the form

$$\min_u \ \mu\|u\|_1 + H(u), \tag{14}$$

for a convex and differentiable H, is an iterative procedure based on shrinkage (also called soft thresholding; see (16) below). In the context of solving (14) with a quadratic H, this method was independently proposed and analyzed in [17, 28, 10, 1], and then further studied or extended in [14, 15, 9, 5, 20, 8]. It turns out that this algorithm can be directly derived from the classic forward-backward operator splitting technique (*c.f.* [23]). The basic shrinkage algorithm can be written as the fixed-point iteration: for $i = 1, \ldots, n$,

$$u_i^{k+1} = \text{shrink}((u^k - \tau\nabla H(u^k))_i, \mu\tau), \tag{15}$$

where $\tau > 0$ serves as a step-length for gradient descent (which may vary with k) and

$$\text{shrink}(t, \alpha) = t - \text{Proj}_{[-\alpha,\alpha]}(t) \tag{16}$$

for any $t \in \mathbb{R}$ and $\alpha > 0$. It is easy to see that the larger μ is, the larger the allowable distance between u^{k+1} and u^k.

A new result in [20] is the finite convergence of the support and the signs of u^k under a non-degeneracy condition. That is, $\text{sign}(u^k) \equiv \text{sign}(u_{\text{opt}})$ (assuming $\text{sign}(0) = 0$) for all $k \geq K$, where u_{opt} denotes the solution of (14) (however, an estimate or bound for K is still unknown). It was also proved in [20] that the rate of convergence is q-linear under suitable conditions on τ and H, and the rate depends on the condition of a sub-Hessian, rather than the entire Hessian, of H at u_{opt}. These results provide explanations why sparsity in solutions can help accelerate convergence of first-order methods.

Various modifications and enhancements have been proposed to improve the efficiency of the basic iteration (15), including [16, 18]. In our view, the basic iteration (15) would not be practically effective without a continuation (or path-following) strategy [20, 33] in which a gradually decreasing sequence of μ-values is used to guide the iterates towards the final optimal solution. In [20], the performance of a fixed-point continuation (FPC) algorithm was compared with those of StOMP [13], GPSR [18] and ℓ_1_ℓ_s [22].

In addition, a general block-coordinate gradient descent method for linearly constrained separable problems [29] can be applied to solving (14).

5. Concluding remarks

Compressive sensing is a new, application-driven, interdisciplinary area where optimization can have a great impact. Given the diversity of applications, successful algorithms should be able to take full advantage of problem structure. We have just seen the beginning of activities in this direction.

Taking advantage of sparsity has always been one of the central tasks in computational algorithms. However, it is fair to say that most previous efforts have been concentrated on sparsity in problem data rather than sparsity in solutions. How to optimally exploit solution sparsity certainly deserves closer examinations in algorithmic studies.

Noise and errors naturally appear in measurements in practical applications. A good algorithm for compressive sensing should be robust with respect to noise and errors under normal conditions. Comprehensive and in-depth research in this direction has yet to be conducted.

Unlike for most other problems, algorithm designers for compressive sensing have some freedom in selecting problem data. For example, which measurement matrix should we use for a given problem, a random Gaussian or a partial DCT matrix? Which frequencies should we sample in MRI? This interaction between problem data, sparse solution and algorithms presents a rich and unique set of research opportunities. Moreover, if data are acquired over a period of time, can we develop a "warm-start" algorithm that produces approximate solutions whose accuracy progressively improves with the increase in available measurements?

The past few years have seen a burst of activities using ℓ_1-related optimization in areas such as statistics, machine learning, signal processing, imaging, and computer vision. While the gradient-descent method is probably the most well-known and widely used tool, researchers in these areas have developed rich analytical results and efficient computational tools for solving various ℓ_1-related optimization problems. Historically, research in optimization has been stimulated by the demand of engineering applications, and subsequently contributed to the practice of these applications. We believe that today we are witnessing the same phenomenon repeating itself in the area of ℓ_1-related optimization.

Acknowledgments

We would like to thank Donald Goldfarb, Richard Tapia and Luis N. Vicente for reading a draft of this article and their valuable suggestions for improving this article.

REFERENCES

[1] J. Bect, L. Blanc-Feraud, G. Aubert, and A. Chambolle, *A ℓ_1-unified variational framework for image restoration*, European Conference on Computer Vision, Prague, Lecture Notes in Computer Science, 3024 (2004), pp. 1–13.

[2] E. Candès and J. Romberg, *Practical signal recovery from random projections*, Wavelet applications in signal and image processing XI, SPIE, (2005) pp. 5914.

[3] E. Candès and T. Tao, *Near optimal signal recovery from random projections: Universal encoding strategies*, IEEE Transactions on Information Theory, 52 (2006), pp. 5406–5425.

[4] A. Cohen, W. Dahmen, and R. A. DeVore, *Compressed sensing and best k-term approximation*, submitted, 2006.

[5] P. L. Combettes and J.-C. Pesquet, *Proximal thresholding algorithm for minimization over orthonormal bases*, SIAM J. Optim., to appear.

[6] Compressive Sensing Resources, `http://www.dsp.ece.rice.edu/cs`, 2008.

[7] W. Dai and O. Milenkovic, *Subspace pursuit for compressive sensing: Closing the gap between performance and complexity*, arXiv:0803.0811, 2008.

[8] J. Darbon and S. Osher, *Fast discrete optimization for sparse approximations and deconvolutions*, preprint, 2007.

[9] I. Daubechies, M. Defrise, and C. De Mol, *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint*, Commun. Pure Appl. Anal., 57 (2004), pp. 1413–1457.

[10] C. De Mol and M. Defrise, *A note on wavelet-based inversion algorithms*, Commun. Pure Appl. Anal., 313 (2002), pp. 85–96.

[11] D. Donoho, *Compressed sensing*, IEEE Transactions on Information Theory, 52 (2006), pp. 1289–1306.

[12] D. Donoho and J. Tanner, *Counting faces of randomly-projected polytopes when the projection radically lowers dimension*, submitted to J. AMS, 2005.

[13] D. Donoho, Y. Tsaig, I. Drori, and J.-C. Starck, *Sparse solution of underdetermined linear equations by stagewise orthogonal matching pursuit*, submitted to IEEE Transactions on Information Theory, 2006.

[14] M. Elad, *Why simple shrinkage is still relevant for redundant representations?*, IEEE Transactions on Information Theory, 52 (2006), pp. 5559–5569.

[15] M. Elad, B. Matalon, J. Shtok, and M. Zibulevsky, *A wide-angle view at iterated shrinkage algorithms*, SPIE (Wavelet XII), San-Diego CA, August (2007), pp. 26–29.

[16] M. Elad, B. Matalon, and M. Zibulevsky, *Coordinate and subspace optimization methods for linear least squares with non-quadratic regularization*, Appl. Comput. Harmon. Anal., 23 (2007), pp. 346–367.

[17] M. Figueiredo and R. Nowak, *EM algorithm for wavelet-based image restoration*, IEEE Transactions on Image Processing, 12 (2003), pp. 906–916.

[18] M. Figueiredo, R. Nowak, and S. Wright, *Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems* IEEE Journal on Selected Topics in Signal Processing, 1 (2007), pp. 586–597.

[19] A. Garnaev and E. D. Gluskin, *The widths of a Euclidean ball*, Dokl. Akad. Nauk SSSR, 277 (1984), pp. 1048–1052.

[20] E. Hale, W. Yin, and Y. Zhang, *A fixed-point continuation method for ℓ_1-regularization with application to compressed sensing*, Tech. Report TR07-07, Dept. of Computational and Applied Mathematics, Rice University, 2007.

[21] B. S. Kashin, *Diameters of certain finite-dimensional sets in classes of smooth functions*, Izv. Akad. Nauk SSSR, Ser. Mat., 41 (1977), pp. 334–351.

[22] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, *A method for large-scale ℓ_1-regularized least squares*, IEEE Journal on Selected Topics in Signal Processing, 1 (2007), pp. 606–617.

[23] P. L. Lions and B. Mercier, *Splitting algorithms for the sum of two nonlinear operators*, SIAM J. Numer. Anal., 16 (1979), pp. 964–979.

[24] M. Lustig, D. Donoho, and J. Pauly, *Sparse MRI: The application of compressed sensing for rapid MR imaging*, Magnetic Resonance in Medicine, to appear.

[25] V. D. Milman and G. Schechtman, *Asymptotic Theory of Finite Dimensional Normed Spaces, With an Appendix by M. Gromov*, Lecture Notes in Mathematics, Vol. 1200, Springer, Berlin, 2001.

[26] B. K. Natarajan, *Sparse approximate solutions to linear systems*, SIAM J. Comput., 24 (1995), pp. 227–234.

[27] Y. Nesterov, *Gradient methods for minimizing composite objective function*, `http://www.optimization-online.org`, Discussion Paper 2007/76, CORE, 2007.

[28] R. Nowak and M. Figueiredo, *Fast wavelet-based image deconvolution using the EM algorithm*, Proceedings of the 35th Asilomar Conference on Signals, Systems, and Computers, Monterey, CA, (2001).

[29] P. Tseng and S. Yun, *A block-coordinate gradient descent method for linearly constrained nonsmooth separable optimization*, preprint, 2008.

[30] M. Rudelson and R. Vershynin, *Geometric approach to error correcting codes and reconstruction of signals*, Int. Math. Res. Not., 64 (2005), pp. 4019–4041.

[31] L. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Physica D, 60 (1992), pp. 259–268.

[32] J. Tropp and A. Gilbert, *Signal recovery from partial information via orthogonal matching pursuit*, preprint, 2005.

[33] E. Van den Berg and M. P. Friedlander, *SPGL1: A MATLAB Solver for Large-Scale Sparse Reconstruction*, `http://www.cs.ubc.ca/labs/scl/spgl1`, 2007.

[34] Y. Zhang, *A simple proof for recoverability of ℓ_1-minimization: Go over or under?*, Tech. Report TR05-19, Dept. of Computational and Applied Mathematics, Rice University, 2005.

[35] Y. Zhang, *When is missing data recoverable?*, Tech. Report TR06-15, Dept. of Computational and Applied Mathematics, Rice University, 2006.

[36] W. P. Ziemer, *Weakly Differentiable Functions: Sobolev Spaces and Functions of Bounded Variation*, Graduate Texts in Mathematics, Springer, Berlin, 1989.