Sparse Recovery via ℓ_1 and L_1 Optimization

by Stanley Osher* and Wotao Yin†

1. Introduction

A sparse signal is a signal that has very few nonzero elements or one that becomes so under a basis change or through a certain transform. Exploiting sparsity has become a common task in data sciences. Compressed sensing [1, 2], regularized regression (e.g., LASSO [3]), and regularized inverse problems (e.g., total variation image reconstruction [4]) have made ℓ_1 optimization a central tool in data processing problems. As the name suggests, ℓ_1 optimization problems recover sparse solutions by solving an optimization problem involving an ℓ_1–norm.

Today, the scope of ℓ_1 optimization is quickly expanding. The size, complexity, and diversity of instances have grown significantly. Beyond 1D signals and 2D images, high-dimensional quantities such as video, 4D CT, and multi-way tensors have become the data or unknown variables in models. New applications have motivated structured solutions to optimization problems that significantly generalize our notion of sparsity. Such applications look for low-rank matrices or tensors, sparse graphs, tree structured data representations, and sparse representations involving only a few dictionary atoms.

This article gives self–contained introductions to ℓ_1 optimization for sparse vectors (Section 2), L_1 optimization for finding functions with compact support (Section 3), and computing sparse solutions from measurements that are corrupted by unknown noisy (Section 4).

2. Can We Trust ℓ_1 Optimization?

Let A be an $m \times n$ matrix and let $b \in \mathbb{R}^m$ be a vector. Suppose that we wish to find the sparsest solution to the linear equations $Ax = b$. Mathematically, this problem is equivalent to minimizing the ℓ_0–"norm" of x (denoted by $\|x\|_0$), which counts the number of nonzero entries of x, subject to $Ax = b$. However, this is a combinatorial problem and is NP–hard [5]. A computationally tractable alternative is to perform the ℓ_1–norm minimization in place of ℓ_0–"norm" minimization. We call this problem the basis pursuit problem:

$$\text{(1)} \qquad \underset{x \in \mathbb{R}^n}{\text{minimize}} \, \|x\|_1 \ \text{subject to} \ Ax = b.$$

Note that the ℓ_1–norm is a convex function. Altogether, problem (1) is a *convex optimization problem*. Before discussing the numerical solution to (1), we should question the quality of the ℓ_0–to–ℓ_1 relaxation: if $\bar{x} \in \mathbb{R}^n$ is the unknown sparse vector and $b = A\bar{x}$, can we trust (1) to recover $\bar{x}$?

First, whenever the linear system $Ax = b$ has a unique solution $\bar{x}$, $\bar{x}$ is the unique solution to (1). In this case, minimizing the ℓ_1–norm is unnecessary. Therefore, it is more interesting to consider the under–determined case: when $Ax = b$ has infinitely many solutions. How can we find the needle $\bar{x}$ in the haystack $\{x : Ax = b\}$?

Without loss of generality, let us assume that the matrix A has full row rank, or otherwise some rows of $Ax = b$ can be removed without affecting the solution of (1). In this case, A is an m–by–n matrix, where $m < n$. We now introduce some notation for the conditions that will guarantee several properties of $\bar{x}$.

* Department of Mathematics, University of California at Los Angeles
E-mail: sjo@math.ucla.edu
† Department of Mathematics, University of California at Los Angeles
E-mail: wotaoyin@math.ucla.edu

Let $S = \{i : \bar{x}_i \neq 0\}$ denote the set of nonzero elements of $\bar{x}$ (a.k.a., the support of $\bar{x}$), let A_i denote the ith column of A, and let A_S denote the submatrix of A formed by the columns A_i for $i \in S$.

When S is fixed, the necessary and sufficient condition for the model (1) to return $\bar{x}$ uniquely is

1a. A_S has full column rank, and
1b. there exists a "*dual certificate*" denoted by $y \in \mathbb{R}^m$ that obeys

   i) $\langle y, A_i \rangle = \mathrm{sign}(\bar{x}_i)$, for $i \in S$, and

   ii) $-1 < \langle y, A_j \rangle < 1$, for $j \notin S$.

Condition 1a basically says that if an oracle tells us that all nonzero elements of $\bar{x}$ fall within S, then we can solely rely on the linear subsystem $A_S x_S = b$ to recover $\bar{x}_S$, which is the nonzero part of $\bar{x}$. Minimizing ℓ_1–norm cannot help here because it is a locally linear function. If condition 1a is not satisfied, then $\bar{x}$ cannot be the unique solution. Indeed, in this case there exists a nonzero vector $t \in \mathbb{R}^n$ such that $At = 0$ and $\mathrm{supp}(t) \subseteq S$. Consider $x_\alpha = \bar{x} + \alpha t$. Restricting α to the interval $(-\epsilon, \epsilon)$ for some sufficiently small ϵ, we have $\mathrm{sign}(\bar{x}) = \mathrm{sign}(\bar{x} + \alpha t)$ and thus $\|x_\alpha\|_1 = \langle \mathrm{sign}(\bar{x}), x_\alpha \rangle = \langle \mathrm{sign}(\bar{x}), \bar{x} + \alpha t \rangle = \|\bar{x}\|_1 + \alpha \langle \mathrm{sign}(\bar{x}), t \rangle$. Therefore, there exists some $\alpha \neq 0$ such that $\|x_\alpha\|_1 \leq \|\bar{x}\|_1$. Together with $Ax_\alpha = A\bar{x} = b$, $\bar{x}$ cannot be the unique solution to (1).

Condition 1a also implies that $|S|$, the number of nonzero components in $\bar{x}$, must obey $|S| \leq m$. In general, we clearly need to take at least $|S|$ linear measurements in order to recover a signal with $|S|$ nonzero elements. Later we will discuss how large m needs to be.

Condition 1b reveals the power of ℓ_1–norm minimization: when a dual certificate y exists, it determines S. To see this, suppose $x \in \mathbb{R}^n$ satisfies $Ax = b$ but $x_j \neq 0$ for some $j \notin S$. Given a dual certificate y, we will show

$$\|x\|_1 > \langle y, Ax \rangle = \langle y, A\bar{x} \rangle = \|\bar{x}\|_1 \tag{2}$$

and thus x cannot be a solution to (1).

Proof of (2). For $i \in S$, by condition 1b (i), we have $|\bar{x}_i| = \langle y, A_i \rangle \bar{x}_i$ and $|x_i| \geq \langle y, A_i \rangle x_i$. For any $j \notin S$ and $x_j \neq 0$, by condition 1b (ii), we have $|x_j| \geq \langle y, A_j \rangle x_j$. Therefore, $\|\bar{x}\|_1 = \sum_{i \in S} \langle y, A_i \rangle \bar{x}_i = \langle y, A\bar{x} \rangle$ and $\|x\|_1 > \sum_{i \in S} \langle y, A_i \rangle x_i + \sum_{j \notin S} \langle y, A_j \rangle x_j = \langle y, Ax \rangle$. □

The two conditions have a nice geometric interpretation. Consider the hyperplane $\mathcal{H} = \{x \in \mathbb{R}^n : p^T x = \alpha\}$, where $p = A^T y$ and $\alpha = y^T b$, and the ℓ_1–"ball" $\mathcal{B} = \{x \in \mathbb{R}^n : \|x\|_1 \leq \beta\}$, where $\beta = \|\bar{x}_S\|_1$. Condition 1b ensures that $\mathcal{H} \cap \mathcal{B}$ is the face of $\mathcal{B}$ where $\mathrm{sign}(x) = \mathrm{sign}(\bar{x})$. Condition 1a further ensures that this face intersects $\{x \in \mathbb{R}^n : Ax = b\}$ at exactly one point, $\bar{x}_S$. Altogether, they ensure that $\bar{x}_S$ is uniquely recovered by (1).

We already saw that Condition 1 is sufficient for (1) to uniquely recover $\bar{x}$. In fact, it is also necessary. In addition, it is both necessary and sufficient for the following relaxed problems to have a unique solution:

$$\underset{x \in \mathbb{R}^n}{\text{minimize}}\, \lambda \|x\|_1 + \frac{1}{2}\|Ax - b\|_2^2, \tag{3}$$

$$\underset{x \in \mathbb{R}^n}{\text{minimize}}\, \|x\|_1 \text{ subject to } \|Ax - b\|_2 \leq \delta; \tag{4}$$

see [6] for proofs. The same condition also guarantees that, when b contains noise and/or when $\bar{x}$ are not exactly but approximately sparse, the solutions to (3) and (4) remain close to $\bar{x}$ in certain norms (assuming appropriate parameters λ and δ, respectively); a generalized condition gives similar properties for the analysis–ℓ_1 model [7], where the signal is (approximately) sparse under a linear transform Ψ and thus $\|\Psi x\|_1$, instead of $\|x\|_1$, is minimized; see [8] and the references therein. The total variation model [4] is a well–known example.

The existence of a dual certificate is the key to the success of ℓ_1 optimization. Given $|S|$ (the number of nonzero elements in the signal), the set of vectors y obeying Condition 1b part (i) is larger if m is larger. Now fixing m, the set of vectors y obeying part (ii) is larger if n is smaller. Therefore, recovery by ℓ_1 optimization is, in general, more likely to succeed if there are a lot of linear measurements and just a few nonzero components.

The main condition listed above is based on a fixed support set S. It is numerically verifiable only when S is either given (by an oracle) or known to have moderately many possibilities. However, there are, in general, exponentially many possible support sets S, and the correct set is hard to guess in advance. How can we ensure the existence of a dual certificate for every possible sparse signal $\bar{x}$? Such a setting is commonly referred to as "uniform sparse recovery." There has been very encouraging answers based on conditions such as the mutual incoherence condition [9, 10], the null-space property (NSP) [2, 11], the restricted isometry principle (RIP) [1], the spherical section property [12], and so on.

A surprising result in uniform sparse recovery is that as long as the entries of A are sampled from subgaussian distributions, then with overwhelming probability, all sparse signals with no more than $k = O(m/\log(n/m))$ nonzero elements can be uniquely recovered by (1). Furthermore, even if b contains noise and/or $\bar{x}$ is approximately k sparse, the recovery remains stable. The constant in the big–O is very mild. The result says that we shall trust ℓ_1 optimization for recovering any sparse signal from a number of qualified linear observations that are merely a few times more in number than the nonzero elements in the signal. We pay a mild price for not knowing the locations of the nonzero elements because of the help of ℓ_1 minimization.

3. L_1 Optimization for Compactly Support Solutions

We have seen the power of ℓ_1 optimization for inducing sparsity in finite dimensional problems. In this section, we consider extensions of ℓ_1 optimization to infinite dimensional calculus of variations type problems arising, for example, in physics and elsewhere.

We can think continuously and start with a very simple, but canonical example. Let $u, f : \mathbb{R}^1 \to \mathbb{R},\ u \in H^1,\ f \in L^2$, and consider the toy problem:

$$\operatorname*{minimize}_{u \in H^1} \frac{1}{2}\int |u_x|^2 - \int fu + \frac{1}{\mu}\int |u|.$$

(We will abuse notation below and let $\|\cdot\|_2, \|\cdot\|_1$ now be the continuous L_2 and L_1 norms and $\langle\ ,\ \rangle$ be the continuous L^2 inner product.)

When minimizing this problem, we are led to the Euler–Lagrange equation

$$u_{xx} + f = \frac{1}{\mu}p(u), \tag{5}$$

where $p(u)$ is a subgradient of $\|u\|_1$. We have $\|u\|_1 = \langle u, p(u)\rangle$ and, for any v,

$$\|v\|_1 - \|u\|_1 \geq \langle v - u, p(u)\rangle,$$
$$\|v\|_1 \geq \langle v, p(u)\rangle.$$

We might also consider gradient descent on this toy problem, with t being the descent direction, obtaining

$$u_t = u_{xx} + f - \frac{1}{\mu}p(u) \tag{6}$$

as an evolution equation.

We can hope that these L_1 regularized (or perturbed) problems will have solutions that vanish a lot, e.g., have compact support. The theoretical framework for this was developed by H. Brezis in some important papers from the early 1970's [13, 14], without any connection to compressive sensing and without any suggested numerical implementation. He considered a wide class of second–order elliptic equations and, with Friedman [14], an extension to parabolic equations. In [15, 16] we showed that many interesting problems of physics can be rewritten in this L^1 form and demonstrated advantages, both numerically and in physical understanding that arises from this approach.

It is instructive to give the following formal argument, which helps explain why the measure of the support shrinks as $\mu \downarrow 0$. Consider a solution to (5) on an interval $x_1 \leq x \leq x_2$ with $u(x_1) = u(x_2) = 0$ and $u(x) > 0$ for $x_1 < x < x_2$. Hence, $p(u(x)) = 1$ for $x_1 < x < x_2$.

Integrating (5) from x_1 to x_2 gives us

$$\mu\left(u_x(x_2) - u_x(x_1) + \int_{x_1}^{x_2} f\right) = x_2 - x_1,$$

but $u_x(x_2) \leq 0 \leq u_x(x_1)$ so:

$$x_2 - x_1 \leq \mu \int_{x_1}^{x_2} f(x)\mathrm{d}x.$$

This gives a bound on the interval in terms of f, which diminishes with μ. Similarly, if instead $u(x_1) = u(x_2) = 0$ and $u(x) < 0$ for $x_1 < x < x_2$, then we have $p(u(x)) = -1$ for $x_1 < x < x_2$ and

$$-\mu\left(u_x(x_2) - u_x(x_1) + \int_{x_1}^{x_2} f\right) = x_2 - x_1.$$

This time

$$-u_x(x_2) \leq 0 < -u_x(x_1), \quad \text{so}$$
$$x_2 - x_1 \leq -\mu \int_{x_1}^{x_2} f(x)\mathrm{d}x$$

and we get the same kind of estimate. This formal argument can be generalized to a wide class of elliptic problems.

We can borrow computational techniques from ℓ_1 optimization to devise efficient and novel numerical methods for these and a wide variety of classical problems. The key tool from a numerical point of view is the simple "soft thresholding" or "shrink" operator. Recall:

$$\begin{aligned}\operatorname{shrink}(x, u) &= \arg\min_y \ \mu|y| + \frac{1}{2}|x - y|^2 \\ &= \begin{cases} x - \mu, & x \geq \mu, \\ 0, & |x| \leq \mu, \\ x + \mu, & x \leq -\mu. \end{cases}\end{aligned}$$

In [16] we applied this approach to PDEs that come from a variational problem, either by minimization, obtaining an elliptic PDE, or by gradient descent to obtain a parabolic PDE. Additionally, some PDEs can be rewritten using the L^1 subgradient such as the divisible sandpile problem and the signum-Gordon equation [15]. Given a linear second order elliptic operator $\mathcal{L}(u)$, we would like to solve numerically

$$0 \in -\mathcal{L}(u) - f + \mu p(u),$$
$$0 \in u_t - \mathcal{L}(u) - f + \mu p(u).$$

Let $Au = -\mathcal{L}(u) - f$, $Bu = \mu p(u)$, and τ be the time step for the time dependent problem. A very convenient implicit and unconditionally stable method is known as the Douglas–Rachford splitting algorithm [17]. Let

$$u^k \approx u(k\Delta t)$$

then update:

$$u^{k+1} = (1+\tau B)^{-1}\left((1+\tau A)^{-1}(1-\tau B)+\tau B\right)u^k,$$

which can be written as

$$\begin{aligned} u^{k+1} &= (I+\tau B)^{-1}\tilde{u}^k, \\ \tilde{u}^{k+1} &= \tilde{u}^k + (1+\tau A)^{-1}(2u^{k+1}-\tilde{u}^k) - u^{k+1}. \end{aligned}$$

Note that computing $(I+\tau B)^{-1}g = v$ means that we are solving for v in

$$g = v + \tau\mu p(v)$$

or solving

$$\underset{v}{\text{minimize}}\, \|v\|_1 + \frac{1}{2\tau\mu}\|v-g\|^2.$$

The solution is

$$v = \text{shrink}(g, \tau\mu),$$

which is simple to implement. See [17] for a convergence proof of this method. This is unconditionally stable and the possible multi–valuedness of $p(u)$ gives no difficulties.

In [16] we constructed an efficient numerical scheme for solving obstacle problems in the divergence form. We reformulated the problem in terms of an L^1 like penalty on the variational problem. This is an exact regularizer. The technique also applies to classical obstacle problems as well as some related free boundary problems, e.g., Hele–Shaw and two phase membrane. The resulting methods are quite simple, again involving the shrink operator, and seem to outperform classical approaches.

Perhaps the most significant application in this set of ideas involves obtaining compactly supported approximations to eigenfunctions of the Schrodinger equation [18, 19]. These have long been sought [20] and are called Wannier functions. These were developed in solid state physics and quantum chemistry. In [18, 19, 21] we developed and analyzed a natural and easy to implement method to do this.

Consider the Hamiltonian

$$\hat{H} = -\frac{1}{2}\Delta + V(x),$$

where Δ is the Laplacian and V is a potential with eigenvalues $\lambda_1 < \lambda_2 \cdots$.

We obtain compactly supported approximations to eigenfunctions by solving the variational problem

$$E_0 = \min_{\varphi_1,\varphi_2,\dots,\varphi_N} \sum_{j=1}^{N} \langle \varphi_j, \hat{H}\varphi_j \rangle$$

subject to $\langle \varphi_i, \varphi_j \rangle = \delta_{jk}$, for $i,j = 1,\dots,N$.

We get densely supported φ_j, (think of sines and cosines when $V = 0$). Physicists and chemists want short–ranged interaction. The original Wannier functions (1937) involve a subspace rotation of the φ_j following by a cut–off to get compactly supported approximate eigenfunctions.

We just add an L^1 regularization in the previous variational problem, obtaining

$$E = \min_{\psi_1,\psi_2,\dots,\psi_N} \sum_{i=1}^{N} \frac{1}{\mu}\|\psi_j\|_1 + \langle \psi_j, \hat{H}\psi_j \rangle \tag{7}$$

subject to $\langle \psi_i, \psi_j \rangle = \delta_{jk}$, for $i,j = 1,\dots,N$.

It turns out that this can be solved rapidly using the split Bregman algorithm with an extra (nonconvex) projection step [22]. The L^1 term actually often speeds up the optimization!

We have a fairly complete approximation theory [23]. We can also impose shift invariance, i.e orthogonality to the translations of the eigenfunctions by lattice vectors [24]. The only nonlinear steps in the algorithm are very simple scalar operations. The resulting approximate eigenfunctions resemble Meyer wavelets [25], but have compact support and are intimately connected to the Schrodinger equation.

4. Computing Paths of Sparse Solutions

When the vector b is corrupted due to noise at an unknown level, it is not straightforward to calculate the correct value of λ in (3). Certain methods, such as cross validation, exist to solve this problem, but they need the solutions to (3) corresponding to all (or largely many) parameter values $\lambda \geq 0$. While solving a single ℓ_1 problem is inexpensive, solving (3) for the entire path of solutions x_λ for all $\lambda \geq 0$ can be time–consuming.

In addition, an *unpleasant* by–product of minimizing ℓ_1–norm in model (3) is the loss of signal magnitude. Consider a toy problem $b = ax + \epsilon$, where ϵ is noise and a, b, x are strictly positive scalars. The solution to (3) is

$$x_\lambda = \begin{cases} 0, & \lambda > ab, \\ \frac{b}{a} - \frac{\lambda}{a^2}, & \lambda \in (0, ab]. \end{cases}$$

Unless $\lambda = 0$, we always get $x_\lambda < b/a$. Roughly speaking, model (3) returns a sparse x_λ by reducing the magnitudes of its components; otherwise, the solution will have many nonzero elements since the noise ϵ cannot be sparsely represented. However, the magnitudes of the true nonzero components are also reduced, causing the solution to be *biased*.

This section describes a simple solution to resolve these issues. We restrict our discussion to the Euclidean space $\mathbb{R}^n$. The optimality condition of (3) is:

$$0 = \lambda p + A^T(Ax_\lambda - b), \quad p \in \partial\|x_\lambda\|_1. \tag{8}$$

Introducing $\lambda = 1/t$ and then replacing $\lambda p = \frac{p}{t}$ in (8) by $\frac{\mathrm{d}p}{\mathrm{d}t}$ so that we can evolve p over time t, we arrive at the new system, known as inverse-scale space (ISS) [26, 27]:

$$\dot{p}(t) = -A^T(Ax(t) - b), \quad p \in \partial\|x(t)\|_1. \tag{9}$$

This is an ordinary differential inclusion, for which we set initial solution $p(0) = x(0) = 0$. For well-definedness, we let x to be right continuous, let p be right continuously differentiable, and let $\dot{p}$ denote the right time derivative of p.

It is easy to evolve the system (9) because at each time $t \geq 0$, either $p_i(t)$ is changing value or $x_i(t)$ is so, but not both. This is because $p_i(t)$ is a subgradient of $|x_i(t)|$, so $x_i(t)$ must stay 0 whenever $p_i(t)$ is changing value between $(-1,1)$, and once $x_i(t)$ becomes strictly positive or strictly negative, $p_i(t)$ must stay 1 or -1, respectively. We can construct a solution path to (9) by keeping x fixed and evolving p, at all but a set of time points where some $p_i(t)$ reaches either 1 or -1. At those times, x is updated as follows. Let $S_1 = \{i : p_i(t) = 1\}$, $S_2 = \{i : p_i(t) = -1\}$, and $T = (S_1 \cup S_2)^c$. Following (9), $x(t)$ is a solution to the system:

$$x_{S_1} \geq 0,\ x_{S_2} \leq 0,\ x_T = 0, \tag{10a}$$

$$0 = A^T_{S_1 \cup A_2}(Ax - b). \tag{10b}$$

Equation (10b) prevents (9) from evolving $p_i(t)$ above 1 or below -1. The system (10) is equivalent to the problem

$$\underset{x \in \mathbb{R}^n}{\text{minimize}}\ \frac{1}{2}\|Ax - b\|^2 \text{ subject to (10a).} \tag{11}$$

The entire path $\{p(t), x(t)\}_{t \geq 0}$ can be obtained by alternating between evolving $p(t)$ by (9) and, upon some $p_i(t)$ hitting either 1 or -1, updating $x(t)$ by solving (10) or (11). Note that $Ax(t)$ is always unique even if $x(t)$ is not, so the path $\{p(t), Ax(t)\}_{t \geq 0}$ is unique. When $Ax = b$ is consistent, there exists $T > 0$ such that $x(t) \equiv x(T)$, for $t \geq T$, and $x(T)$ is a solution to (1).

Applying model (9) to the toy example gives

$$x(t) = \begin{cases} 0, & \frac{1}{t} > ab, \\ \frac{b}{a}, & \frac{1}{t} \in (0, ab]. \end{cases}$$

Notice that the bias $-\frac{\lambda}{a^2}$ is gone!

Of course, one can manually add $-\frac{\lambda}{a^2}$ back to the solution of (3) and, in the general case, manually update x_λ by solving an additional *de-biasing* problem, for example, minimize $\frac{1}{2}\|Ax - b\|^2$ subject to x on the same support as x_λ. However, this will *not* recover the solution path $x(t)$ of ISS. In general, $x(t)$ and x_λ, for $\lambda = 1/t$, do not have the same support. This is because bias not only reduces magnitude but also affects the support of x_λ. Debiasing only changes the values of x_λ, not its support. Therefore, introducing bias and then removing it are not as effective as avoiding bias at the beginning.

Without minimizing the ℓ_1-norm in (9), is $x(t)$ still sparse? The answer is interesting: $x(t)$ and x_λ are both sparse for the same reason: $p \in \mathrm{Range}(A^T)$, which holds for both (8) and (9). In our work, we deal with the underdetermined case where A has more columns than rows, so p stays in a small m-dimensional subspace in a large n-dimensional space. On the other hand, from the subgradient relation between p and x, x is sparse if few components of p equal 1 or -1. The ℓ_1 subgradient p takes value in the hyperbox $[-1,1]^n$. The faces of the hyperbox are precisely the vectors which have 1 or -1 in some component Having more components equal to 1 or -1 means that p is on a smaller dimensional face. For example, if all components are equal to 1, the hyperbox face is just a single point. Therefore, Therefore, when the dimension of $\mathrm{range}(A^T)$ is small, it is unlikely for p to have many 1 or -1 components, so x is likely sparse. More formal analysis can be found in [28]. The point is that x is sparse because p, the ℓ_1 subgradient at x, is in the range of A^T, not because of the usual properties of the ℓ_1-norm.

One of the main advantages of (9) is how quickly and easily it computes the solution path. As argued above, it can be computed piece-wise, and every piece is a sign-constrained least-squares problem (11) that is similar to the previous one, so one can warm-start and solve it very quickly using QR updates. There are also other methods to obtain an approximate solution path even faster: for example, (discrete-time) Bregman iteration [29, 30], (continuous-time) linearized Bregman ISS [27], and (discrete-time) linearized Bregman iteration [30, 31].

Bregman iteration is the forward Euler iteration of (9):

$$p^{k+1} = p^k - \frac{\Delta t}{m}A^T(Ax^k - b), \quad p^k \in \partial\|x^k\|_1,$$

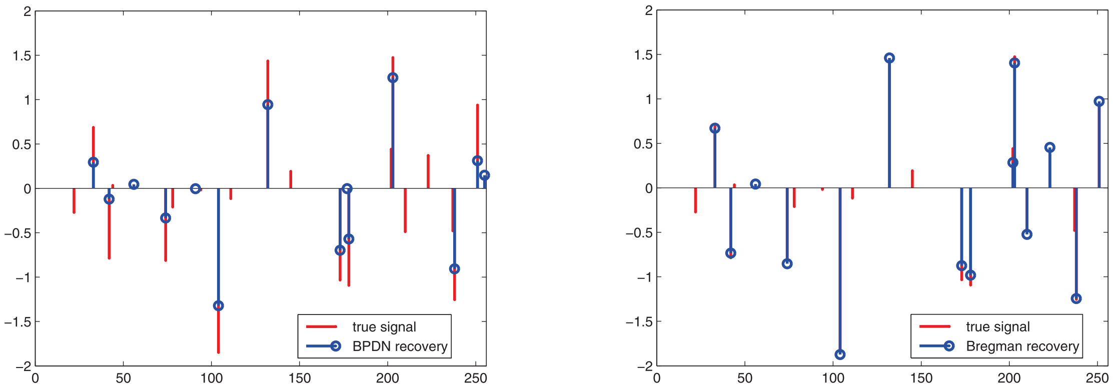
which is the optimality condition to

$$x^{k+1} = \underset{x \in \mathbb{R}^n}{\arg\min}\, D(x; x^k) + \frac{\Delta t}{2m}\|Ax - b\|^2, \tag{12}$$

where $D(x; x^k) := \|x\|_1 - \|x^k\|_1 - \langle p^k, x - x^k\rangle$ is the Bregman distance induced by ℓ_1-norm. Interestingly, after a change of variable, (12) reduces to the equivalent "add-back-the-residual" iteration:

$$x^{k+1} = \underset{x \in \mathbb{R}^n}{\arg\min}\, \|x\|_1 + \frac{\Delta t}{2m}\|Ax - b^k\|^2, \tag{13a}$$

$$b^{k+1} = b^k + (b - Ax^k). \tag{13b}$$



Model (3) with hand tuned λ

5th Bregman iteration of (12) or (13)

Figure 1. Recover x (red) from Gaussian noisy measurements $b = Ax + \epsilon$. Left: the solution (blue) of model (3) *(also known as BPDN or* basis pursuit denoising*) with hand picked $\lambda = 49$ for best sparsity-noise tradeoff. Right: the 5th iterate (blue) of the Bregman iteration* (12) *or* (13)*. Conclusion: while true x (blue) cannot be recovered due to unknown noise, the Bregman solution (right blue) recovers the lost magnitude in the ℓ_1 solution (left blue) and three additional large elements, near the right end.*

Each iteration of (13a) requires minimizing a problem similar to (3) except that residual $(b - Ax^k)$ is *added back* to the measurement. This is better than solving (3) and tuning its λ because this restores lost magnitude in the signal. See Figure 1.

In addition, one can apply an existing code for (3) to solver the subproblem (13a). Furthermore, (13) has another interesting property of *error forgetting* [32]: the subproblem (13a) can be solved inexactly with error, but the errors do not accumulate; instead, they cancel each other so that x^k still converges quickly.

We can get even faster linearized Bregman algorithms by smoothing. Simply add $\frac{1}{\kappa}\dot{x}$ to (9) and obtain

$$\text{(14)} \qquad \dot{p}(t) + \frac{1}{\kappa}\dot{x} = -A^T(Ax(t) - b), \quad p \in \partial\|x(t)\|_1.$$

It has a piece-wise smooth solution, which converges to the unsmoothed solution exponentially fast as κ increases. By introducing $z = p + \frac{1}{\kappa}x$, (14) reduces to an ordinary differential equation:

$$\text{(15)} \qquad \dot{z}(t) = -A^T(\kappa A\,\mathrm{shrink}(z(t), 1) - b).$$

There is no inclusion anymore. This is because the mapping between z and (p, x) is one-one. Given z, we uniquely recover $x = \kappa\,\mathrm{shrink}(z, 1)$ and $p = z - \frac{1}{\kappa}x$. The forward Euler iteration of (15) is known as the linearized Bregman iteration, which evolves quickly [33] and can be easily parallelized for problems with massive amounts of data [34].

All we have discussed in this section generalizes naturally to other regularization function in place of the ℓ_1 norm. If one is using the minimization model:

$$\mathrm{minimize}\, r(x) + t f(x)$$

where r enforces a solution structure and f is a differentiable data fidelity function, we encourage trying the ISS system

$$\dot{p}(t) = -f'(x), \quad p(t) \in \partial r(x(t)),$$

which will likely reduce bias and compute a solution path quickly while still keeping the desired structure for the solution.

Acknowledgments

We would like to thank Damek Davis and Hayden Schaeffer for valuable suggestions for improving this article. S. Osher was supported in part by ONR grants N00014140444, N000141410683, and N000141110719. W. Yin was supported in part by NSF grant DMS-1317602.

References

[1] E. J. Candes and T. Tao. “Near-optimal signal recovery from random projections: universal encoding strategies?” In: *IEEE Transactions on Information Theory* 52.12 (Dec. 2006), pp. 5406–5425.

[2] D. Donoho. “Compressed sensing”. In: *IEEE Transactions on Information Theory* 52.4 (Apr. 2006), pp. 1289–1306.

[3] R. Tibshirani. “Regression shrinkage and selection via the lasso: a retrospective”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.3 (June 2011), pp. 273–282.

[4] L. I. Rudin, S. Osher, and E. Fatemi. “Nonlinear total variation based noise removal algorithms”. In: *Physica D: Nonlinear Phenomena* 60.1-4 (Nov. 1992), pp. 259–268.

[5] B. K. Natarajan. “Sparse approximate solutions to linear systems”. In: *SIAM Journal on Computing* 24.2 (Apr. 1995), pp. 227–234.

[6] H. Zhang, W. Yin, and L. Cheng. “Necessary and sufficient conditions of solution uniqueness in 1-norm minimization”. In: *Journal of Optimization Theory and Applications* (Aug. 2014).

[7] M. Elad, P. Milanfar, and R. Rubinstein. “Analysis versus synthesis in signal priors”. In: *Inverse Problems* 23.3 (June 2007), pp. 947–968.

[8] H. Zhang, M. Yan, and W. Yin. “One condition for solution uniqueness and robustness of l1-synthesis and l1-analysis minimizations”. In: *arXiv preprint arXiv:1304.5038* (2013), pp. 1–13.

[9] D. Donoho and X. Huo. “Uncertainty principles and ideal atomic decomposition”. In: *IEEE Transactions on Information Theory* 47.7 (2001), pp. 2845–2862.

[10] M. Elad and A. Bruckstein. “A generalized uncertainty principle and sparse representation in pairs of bases”. In: *IEEE Transactions on Information Theory* 48.9 (Sept. 2002), pp. 2558–2567.

[11] A. Cohen, W. Dahmen, and R. DeVore. “Compressed sensing and best k-term approximation”. In: *Journal of the American Mathematical Society* 22.1 (July 2008), pp. 211–231.

[12] Y. Zhang. “Theory of compressive sensing via $\ell 1$-minimization: a non-RIP analysis and extensions”. In: *Journal of the Operations Research Society of China* 1.1 (Mar. 2013), pp. 79–105.

[13] H. Brezis. “Solutions with compact support of variational inequalities”. In: *Russian Mathematical Surveys* 29.2 (Apr. 30, 1974), pp. 103–108.

[14] H. Brezis and A. Friedman. “Estimates on the support of solutions of parabolic variational inequalities”. In: *Illinois Journal of Mathematics* 20.1 (1976), pp. 82–97.

[15] R. E. Caflisch, S. J. Osher, H. Schaeffer, and G. Tran. “PDEs with compressed solutions”. In: *UCLA CAM Report 13-67, to appear in Communications in Mathematical Sciences* (Nov. 22, 2013).

[16] G. Tran, H. Schaeffer, W. M. Feldman, and S. J. Osher. “An ℓ_1 penalty method for general obstacle problems”. In: *UCLA CAM Report 14-27* (Apr. 4, 2014), *to appear in SIAM J. Applied Math.*

[17] P. L. Lions and B. Mercier. “Splitting algorithms for the sum of two nonlinear operators”. In: *SIAM Journal on Numerical Analysis* 16.6 (Dec. 1979), pp. 964–979.

[18] V. Ozolins, R. Lai, R. Caflisch, and S. Osher. “Compressed modes for variational problems in mathematics and physics”. In: *Proceedings of the National Academy of Sciences* 110.46 (Nov. 12, 2013), pp. 18368–18373.

[19] V. Ozolins, R. Lai, R. Caflisch, and S. Osher. “Compressed plane waves yield a compactly supported multiresolution basis for the laplace operator”. In: *Proceedings of the National Academy of Sciences* 111.5 (Feb. 4, 2014), pp. 1691–1696.

[20] G. H. Wannier. “The structure of electronic excitation levels in insulating crystals”. In: *Physical Review* 52.3 (Aug. 1, 1937), pp. 191–197.

[21] F. Barekat, K. Yin, R. E. Caflisch, S. J. Osher, R. Lai, and V. Ozolins. “Compressed Wannier modes found from an L_1 regularized energy functional”. In: *UCLA CAM Report 14-23* (Mar. 26, 2014).

[22] R. Lai and S. Osher. “A splitting method for orthogonality constrained problems”. In: *Journal of Scientific Computing* 58.2 (Feb. 1, 2014), pp. 431–449.

[23] O. Tekin, K. Yin, and F. Barekat. “Spectral results for perturbed variational eigenvalue problems and their applications to compressed PDEs”. In: *UCLA CAM Report 14-47* (June 2014).

[24] F. Barekat, R. Lai, K. Yin, S. Osher, R. Caflisch, and V. Ozolins. “Projection to the set of shift orthogonal functions”. In: *UCLA CAM Report 14-18* (Feb. 20, 2014).

[25] Y. Meyer. *Ondelettes et opérateurs*. Hermann, 1990.

[26] M. Burger, S. Osher, J. Xu, and G. Gilboa. “Nonlinear inverse scale space methods for image restoration”. In: *Variational, Geometric, and Level Set Methods in Computer Vision Lecture Notes in Computer Science*. Ed. by N. Paragios, O. Faugeras, T. Chan, and C. Schnörr. Vol. 3752. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, Berlin, Heidelberg, 2005, pp. 25–36.

[27] S. Osher, F. Ruan, J. Xiong, Y. Yao, and W. Yin. “Sparse recovery via differential inclusions”. In: *UCLA CAM Report 14-61* (June 2014).

[28] D. L. Donoho and J. Tanner. “Neighborliness of randomly projected simplices in high dimensions”. In: *Proceedings of the National Academy of Sciences of the United States of America* 102.27 (July 2005), pp. 9452–9457.

[29] S. Osher, M. Burger, D. Goldfarb, J. Xu, and W. Yin. “An iterative regularization method for total variation-based image restoration”. In: *SIAM Journal on Multiscale Modeling and Simulation* 4.2 (2005), pp. 460–489.

[30] W. Yin, S. Osher, D. Goldfarb, and J. Darbon. “Bregman iterative algorithms for L1-minimization with applications to compressed sensing”. In: *SIAM Journal on Imaging Sciences* 1.1 (2008), pp. 143–168.

[31] S. Osher, Y. Mao, B. Dong, and W. Yin. “Fast linearized Bregman iteration for compressive sensing and sparse denoising”. In: *Communications in Mathematical Sciences* 8.1 (Mar. 2010), pp. 93–111.

[32] W. Yin and S. Osher. “Error forgetting of Bregman iteration”. In: *Journal of Scientific Computing* 54.2 (2012), pp. 684–698.

[33] M.-J. Lai and W. Yin. “Augmented ℓ_1 and nuclear-norm models with a globally linearly convergent algorithm”. In: *SIAM Journal on Imaging Sciences* 6.2 (June 2013), pp. 1059–1091.

[34] Z. Peng, M. Yan, and W. Yin. “Parallel and distributed sparse optimization”. In: *2013 Asilomar Conference on Signals, Systems and Computers*. IEEE, Nov. 2013, pp. 659–646.