A FAST AND ACCURATE FIRST-ORDER ALGORITHM FOR COMPRESSED SENSING

J. Bobin and E. J. Candès

Applied and Computational Mathematics (MC 217-50), California Institute of Technology
Pasadena, CA 91125, USA

ABSTRACT

This paper introduces a new, fast and accurate algorithm for solving problems in the area of compressed sensing, and more generally, in the area of signal and image reconstruction from indirect measurements. This algorithm is inspired by recent progress in the development of novel first-order methods in convex optimization, most notably Nesterov's smoothing technique. In particular, there is a crucial property that makes these methods extremely efficient for solving compressed sensing problems. Numerical experiments show the promising performance of our method to solve problems which involve the recovery of signals spanning a large dynamic range.

Index Terms— Compressed sensing, sparsity, ℓ_1 and total-variation minimization, smoothing technique in optimization.

1. INTRODUCTION

Compressed sensing (CS) [3, 2, 4] is a new sampling theory based on the revelation that one can exploit sparsity or compressibility when acquiring signals of general interest, and that one can design nonadaptive sampling techniques that condense the information in a compressible signal into a small amount of data. There are early indications showing that this revelation may change the way engineers think about signal acquisition in areas ranging from analog-to-digital conversion, digital optics, magnetic resonance imaging, and seismics.

In compressed sensing, one acquires a signal $x \in \mathbb{R}^n$ by collecting data of the form $b = Ax + n$: x is the signal of interest (more precisely, its coefficient sequence in a representation where it is assumed to be fairly sparse), A is an $m \times n$ "sampling" matrix, and n is a noise term. A standard approach in CS attempts to reconstruct x by solving

$$\min_x f(x) \text{ s. t. } \|b - Ax\|_{\ell_2} < \epsilon, \tag{1}$$

where ϵ^2 is an estimated upper bound on the noise power. The choice of the regularizing function f depends on prior assumptions about the signal x of interest : if x is (approximately) sparse, an appropriate convex function is the ℓ_1 norm (as advocated by the CS theory), if x is a piecewise constant object, the total-variation (TV) norm provides accurate recovery results, and so on.

Solving large-scale problems such as (1) (think of the unknown as a mega-pixel image) is challenging. Although, one cannot review the vast literature on this subject, the majority of the algorithms that have been proposed are unable to solve these problems *accurately* with low computational complexity. On the one hand, interior-point methods are accurate but problematic because they need to solve large systems of linear equations along the way (the Newton step). On the other hand, newly introduced techniques based upon iterative thresholding, see [5, 1, 6] and the many earlier references therein converge slowly in the sense that they require a very large number of iterations when high accuracy is required. This is important as in many applications of interest, the signals exhibit a wide dynamic range.

In this paper, we build upon the recent work of Nesterov [8], which develops a series of first-order methods with improved convergence rates, by adapting these ideas to develop algorithms for solving signal recovery problems. These novel algorithms show a lot of promise ; they are fast, accurate and seem robust in the sense that their performance does not depend on the fine tuning of various controlling parameters.

2. FAST FIRST-ORDER ALGORITHMS

2.1. General formulation

Consider the saddle point problem

$$\min_{x \in \mathcal{Q}_p} \max_{u \in \mathcal{Q}_d} \langle u, Wx \rangle, \tag{2}$$

where $x \in \mathbb{R}^n$, $u \in \mathbb{R}^p$ and $W \in \mathbb{R}^{p \times n}$. We will refer to $\mathcal{Q}_p$ and $\mathcal{Q}_d$ as the primal and dual feasible sets. Put $f(x) = \max_{u \in Q_d} \langle u, Wx \rangle$; the function f is convex but generally nonsmooth. If $W = I$ and $\mathcal{Q}_d = \{u : \|u\|_{\ell_\infty} \leq 1\}$, then $f(x) = \|x\|_{\ell_1}$ and (2) is the CS recovery problem if we set $\mathcal{Q}_p = \{x : \|b - Ax\|_{\ell_2} < \epsilon\}$.

TV minimization can also be cast as (2). In [8], Nesterov proposed to substitute f by the smooth approximation

$$f_\mu(x) = \max_{u \in Q_d} \langle u, Wx \rangle - \mu p_d(u), \tag{3}$$

where $p_d(u)$ is a *prox-function* for $\mathcal{Q}_d$; that is, $p_d(u)$ is continuous and strongly convex on Q_d, with convexity parameter σ_d (we will assume that p_d vanishes on $\mathcal{Q}_d$). Nesterov proved that f_μ is continuously differentiable, that $\nabla f_\mu(x) = W^* u_\mu(x)$ where $u_\mu(x)$ is the optimal solution of (3), and that ∇f_μ is Lipschitz with constant $L_\mu = \frac{1}{\mu\sigma_d}\|W\|^2$ ($\|W\|$ is the operator norm of W). Nesterov's algorithm minimizes f_μ over $\mathcal{Q}_p$ by iteratively estimating three sequences $\{x_k\}$, $\{y_k\}$ and $\{z_k\}$ while smoothing the feasible set $\mathcal{Q}_p$. The algorithm depends on two scalar sequences $\{\alpha_k\}$ and $\{\tau_k\}$ discussed below, and takes the following form :

Initialize x_0. For $k \geq 0$,
1. Compute $\nabla f_\mu(x_k)$.
2. Compute y_k :
   $y_k = \text{Argmin}_{x \in Q_p} \frac{L_\mu}{2}\|x - x_k\|_{\ell_2}^2 + \langle \nabla f_\mu(x_k), x - x_k \rangle$.
3. Compute z_k :
   $z_k = \text{Argmin}_{x \in Q_p} \frac{L_\mu}{2\sigma_p} p_p(x) + \sum_{i=0}^k \alpha_i \langle \nabla f_\mu(x_i), x - x_i \rangle$.
4. Update x_k :
   $x_k = \tau_k z_k + (1 - \tau_k) y_k$.

Stop when a given criterion is valid.

In this algorithm the function $p_p(x)$ is a prox-function for the primal feasible set Q_p with strong convexity parameter σ_p. At step k, y_k is the current guess at the optimal solution. If we only performed the second step of the algorithm with y_{k-1} instead of x_k, we would obtain a standard first-order technique. The novelty is that the sequence z_k "keeps in mind" the previous iterations, and the point x_k at which the gradient of f_μ is evaluated happens to be a subtle average between z_k and y_k. It has been shown in [8] that if $\alpha_k = 1/2(k+1)$ and $\tau_k = 2/(k+3)$, then the algorithm converges to $x^\star = \text{Argmin}_{x \in Q_p} f_\mu(x)$ with the convergence rate

$$f_\mu(y_k) - f_\mu(x^\star) \leq \frac{4L_\mu p_p(x^\star)}{(k+1)^2 \sigma_p}. \tag{4}$$

The decay is far better than what is achievable via standard gradient-based optimization techniques (k^{-2} vs. k^{-1}).

2.2. The choice of the primal set prox-function

A good primal prox-function is a smooth function that is likely to have some positive effect near the solution. In the setting of (1), a suitable smoothing prox-function may be

$$p_p(x) = \frac{1}{2}||b - Ax||_{\ell_2}^2 + \frac{\rho}{2}\|x - x_0\|_{\ell_2}^2 \tag{5}$$

for some $x_0 \in \mathbb{R}^n$, e.g. an initial guess of the solution. Notice that the bound on the error at iteration k in (4) is proportional to $p_p(x^\star)$; choosing x_0 wisely (a good first guess) can make $p_p(x^\star)$ small ; when nothing is known about the solution as in our later experiments, a natural choice may be $x_0 = A^* b$.

2.3. Projecting onto $\mathcal{Q}_p$

When $m < n$, the feasible set is unbounded which is a departure from Nesterov's algorithm. Now applying the algorithm above requires computing the projection onto $\mathcal{Q}_p$ to estimate y_k and z_k. After an appropriate change of variables, the minimization problems in steps 2 and 3 can be recast as

$$\min_{x \in \mathcal{Q}_p} \frac{\gamma}{2}\|b - Ax\|_{\ell_2}^2 + \frac{1}{2}\|x\|_{\ell_2}^2 + \langle c, x \rangle. \tag{6}$$

The constraint $x \in \mathcal{Q}_p$ can be relaxed as follows with an appropriate choice of λ : $\min_x \frac{\gamma+\lambda}{2}\|b - Ax\|_{\ell_2}^2 + \frac{1}{2}\|x\|_{\ell_2}^2 + \langle c, x \rangle$. The solution is then given by

$$x = (I + \beta A^* A)^{-1}(\beta A^* b - c), \tag{7}$$

where $\beta = \gamma + \lambda$. Hence, we need to solve a positive definite system of the form $(I + \beta A^* A)\Delta x = \Delta b$. Note that the eigenvalues of this system are of the form $1 + \beta\lambda(A^* A)$, where $\lambda(A^* A)$ are the eigenvalues of $A^* A$.

Random matrices. When $A \in \mathbb{R}^{m \times n}$, $m < n$, is a random matrix with i.i.d. entries with mean zero and variance 1, it is well known that the nonzero eigenvalues of $A^* A/n$ are all very near the interval $[(1 - \sqrt{m/n})^2, (1 + \sqrt{m/n})^2]$ [7]. As a consequence, apart from many eigenvalues at 1, the eigenvalues of $I + \beta A^* A$ are highly clustered. In this case, choosing $\beta = \gamma + \max\{0, \|A^* A(b - Ac)\|_{\ell_2}/\epsilon - \gamma - 1\}$ is close to the optimal value and, hence, steps 2 and 3 can be performed with only a few Conjugate Gradients (CG) iterations. To drive this point home, Table 1 reports the results of a simple experiment in which A is a random Gaussian matrix with $n = 256$ and a varying value of m. The value of β is set to $0.01/\|A^* A\|$, the vectors b and c are random, and CG iterations are applied to compute an approximate solution x_{CG} to (7). Each x_{CG} is compared to the true solution x obtained by solving the system with a direct solver. The results demonstrate that for random measurement matrices, steps 2 and 3 can be computed with a handful of CG iterations with excellent accuracy.

Projections. In a wide range of CS applications, the matrix A is a projection onto a subspace of $\mathbb{R}^n$ as when the Fourier, the Hadamard or the noiselet transforms are

m/n	# iterations	$\|(x_{\text{CG}} - x)/|x|\|_{\ell_\infty}$
0.01	3	$2\,10^{-4}$
0.04	4	$5\,10^{-5}$
0.08	4	$7\,10^{-5}$
0.16	4	$4\,10^{-4}$

Table 1. Conjugate gradients and projection onto $\mathcal{Q}_p$. Averaged numbers of CG steps and averaged maximum (entry-wise) relative errors. These averages are over 100 trials for each value of m/n.

subsampled. Steps 2 and 3 become trivial since one has an explicit formula

$$x = \big(I - \frac{\beta}{1+\beta} A^*A\big)(\beta A^*b - c).$$

where $\beta = \gamma + \max\{0, \|A^*A(b - Ac)\|_{\ell_2}/\epsilon - \gamma - 1\}$. In practice, the computational cost of this step is just the cost of applying A^*A ; e.g. two FFTs when A is a partial DFT.

2.4. Nesterov and Continuation

In the spirit of the fixed continuation technique introduced in [6], the nature of the algorithm makes it amenable to some sort of continuation. We discuss two possibilities.

In Step 2 and 3 of the proposed algorithm, updating y_k and z_k is similar to a projected gradient step. For instance, y_k is of the form

$$y_k = \mathcal{P}_{\mathcal{Q}_p}\left(x_k - \frac{1}{L_\mu}\nabla f_\mu(x_k)\right), \tag{8}$$

where $\mathcal{P}_{\mathcal{Q}_p}$ is the projector onto $\mathcal{Q}_p$. As $L_\mu \propto \mu^{-1}$, the step size scales like μ. Thus the algorithm converges faster when μ is larger. For ℓ_1 minimization, $\nabla f_\mu = 1/\mu W^*\mathcal{P}_{\{x | \|x\|_\infty \le \mu\}}(Wx_k)$. The parameter μ is thus equivalent to a threshold, and the accuracy on each entry of the final estimate will also scale like μ. High accuracy requires a small value of μ. To overcome this trade-off between speed and accuracy, we propose a kind of continuation scheme. The idea is to apply Nesterov's algorithm sequentially with a decreasing sequence $\{\mu_t\}$.

Initialize μ_0 and $x_0 = x_{\mu_0}$. For $t \ge 1$,
1. Apply Nesterov's algorithm with $\mu = \mu_t$ and $x_0 = x_{\mu_{t-1}}$
2. Decrease the value of μ : $\mu_{t+1} = \rho\mu_t$ with $\rho < 1$
Stop when the desired value of μ is reached.

Above, x_0 is the point defining the prox-function, see (5), and x_{μ_t} is the computed solution for $\mu = \mu_t$. As μ_t decreases, this algorithm sequentially computes tighter approximations x_{μ_t} to x. Numerical results are given in Section 3.2.

3. NUMERICAL EXPERIMENTS

We apply the algorithm discussed above to the classical ℓ_1 norm and 2D total-variation norm minimization problems. Both these examples assume incomplete data.

3.1. Recovery of signals with wide dynamic range

We assume that x is sparse and that A is a partial Fourier transform (this models the problem of recovering a signal with a sparse spectrum from data sampled at a rate much lower than the Nyquist rate) and use ℓ_1 minimization to recover x from b. The number of entries of x is $n = 65,536$, the number of measurements is $m = n/4$ and only $n/100 \simeq 655$ entries of x are nonzero. The amplitudes of the nonzero components are selected at random ; the lowest value is equal to 1 and the highest is equal to 10^d where $d = 1, \cdots, 4$. Note that when $d = 3, 4$, the signals exhibit a large dynamic range (DR), which is challenging for most numerical methods. The stopping criterion is $1 - f_\mu(x_{k+1})/f_\mu(x_k) < 10^{-6}$.

As discussed earlier, the parameter μ is of paramount importance as it fixes the accuracy of the smooth approximation. The lower this parameter, the closer $x_\mu^\star$ to $x^\star$. When minimizing the ℓ_1 norm, one can check that the entries of $x_\mu^\star$ will differ from those of $x^\star$ by an amount which is on the order of μ. In our experiment, the std. of the additive Gaussian noise is $\sigma = 0.01$ and we set $\mu = 30\sigma$ as this gives a good balance between speed and accuracy (the lowest nonzero entry of x has an amplitude equal to 1). The ℓ_2 norm error on the residual has been set to $\epsilon = \sigma\sqrt{m^2 + 2\sqrt{2m}}$.

Figure 1 presents the results of an experiment when the DR is 10^4 (40dB) and Table 2 reports on experiments corresponding to various values of the DR. Each value in Table 2 is an average computed from 25 random trials.

The results highlight that whatever the DR, the error on the nonzero entries is at most on the order of μ. Hence, the algorithm achieves very high accuracy whenever needed, as in the case of high DR. Since μ is smaller than the smallest entry, the consequence is that all the spikes (all the nonzero components) are correctly detected. As expected, the higher the DR, the higher the number of iterations. However, the number of iterations is still remarkably small. For a DR of 40dB, the algorithm returns a highly accurate solution in just about 1,200 FFTs.

3.2. TV Minimization

The total-variation norm of a 2D digital object x_{ij} is given by

$$\|x\|_{TV} = \sum_{i,j}\sqrt{[D_1x]_{ij}^2 + [D_2x]_{ij}^2},$$

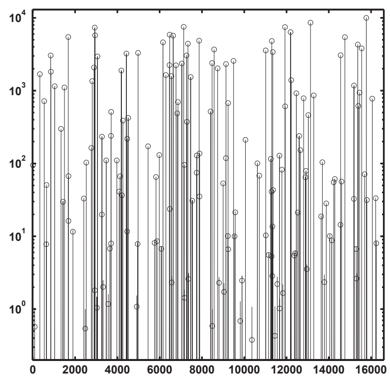

Fig. 1. Solid line : original signal. Dots - ○ : recovered signal.

d	# iterations	ℓ_∞ error off Support(x)	Detection rate
1	47	0.11	100$\%$
2	64	0.15	100$\%$
3	109	0.16	100$\%$
4	305	0.16	100$\%$

Table 2. ℓ_1-recovery results as a function of dynamic range. Average number of iterations required to reach convergence, average maximum size off the support of x, and average detection rate in $\%$.

where D_1 and D_2 are the horizontal and vertical differences : e. g. $[D_1x]_{ij} = x_{i+1,j} - x_{i,j}$. Set W to be the linear map defined via $[Wx]_{ij} = ([D_1x]_{ij}, [D_2x]_{ij})$. Then minimizing the TV norm can be cast as a saddle point problem since $\|x\|_{TV} = \max_{u \in \mathcal{Q}_d} \langle Wx, u \rangle$, where $\mathcal{Q}_d$ is the family $u = \{u_{ij}\}$ with $u_{ij} \in \mathbb{R}^2$ and $\|u_{i,j}\|_{\ell_2} \leq 1$ for each pair (i, j).

We apply the algorithm to solve a TV minimization problem from incomplete frequency data. Here, we collect noisy Fourier coefficients about the 256×256 classical Logan-Shepp phantom ; these $m = 5357$ samples lie on radial lines just as in [3]. The level of the additive noise is $\sigma = 0.01$; $\epsilon = \sqrt{m^2 + 2\sqrt{2m}}\sigma$.

With $\mu = 10^{-7}$, it takes 1092 iterations to reach convergence[1]. Applying a continuation technique with a sequence of μ's equal to $\{\mu_0, \mu_0/2, \cdots, 10^{-7}\}$, with $\mu_0 = \|A^*b\|_{TV}/n \simeq 10^{-4}$, lowers the iteration count ; convergence is reached in 512 steps. The SNR is equal to 58.2 dB without continuation and equal to 66.4 dB with continuation. These preliminary results show that continuation with Nesterov's algorithm can definitely improve the overall speed of convergence.

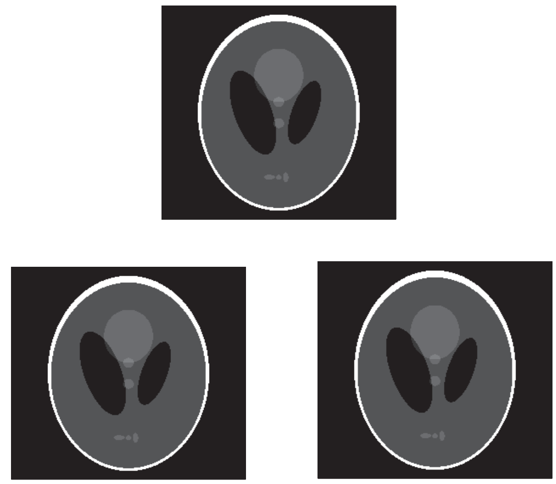

Fig. 2. Top : Original 256×256 image. Bottom-left : Recovery without continuation. Bottom-right : Recovery with continuation.

4. REFERENCES

[1] J.-F. Cai, S. Osher, and Z. Shen. Linearized Bregman iterations for compressed sensing. *Math. Comp. (to appear)*, 2008.

[2] E. Candès and T. Tao. Near optimal signal recovery from random projections : Universal encoding strategies ? *IEEE Trans. on Information Theory*, 52(12) :5406–5425, 2006.

[3] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles : Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. on Information Theory*, 52(2) :489–509, 2006.

[4] D. L. Donoho. Compressed sensing. *IEEE Trans. on Information Theory*, 52(4) :1289–1306, April 2006.

[5] M.A. Figueiredo, R. Nowak, and S.J. Wright. Gradient projection for sparse reconstruction : Application to compressed sensing and other inverse problems. *IEEE Journal of Selected Topics in Signal Processing*, 1(4) :586 – 597, Dec. 2007.

[6] E. T. Hale, W. Yin, and Y. Zhang. A fixed-point continuation method for l1-regularized minimization with applications to compressed sensing. *Technical Report - Rice University*, 2007.

[7] Mehta M.L. *Random Matrices*. Academic Press, 3rd edition, 1991.

[8] Y. Nesterov. Smooth minimization of non-smooth functions. *Math. Program., Serie A*, 103 :127–152, 2005.

1. convergence is reached when $1 - f_\mu(x_{k+1})/f_\mu(x_k) < 10^{-6}$