

SOLUTION OF ℓ_1 MINIMIZATION PROBLEMS BY LARS/HOMOTOPY METHODS

Iddo Drori, David L. Donoho

Stanford University, Department of Statistics
Sequoia Hall, 390 Serra Mall, Stanford, CA 94305-4065

ABSTRACT

Many applications in signal processing lead to the optimization problems

$$\min \|x\|_1 \text{ subject to } y = Ax,$$

and

$$\min \|x\|_1 \text{ subject to } \|y - Ax\| \leq \varepsilon,$$

where A is a given d times n matrix, $d < n$, and y is a given $n \times 1$ vector.

In this work we consider ℓ_1 minimization by using LARS, Lasso, and homotopy methods [1, 2, 3] (Efron et al., Tibshirani, Osborne et al.). While these methods were first proposed for use in statistical model selection, we show that under certain conditions these methods find the sparsest solution rapidly, as opposed to conventional general purpose optimizers which are prohibitively slow.

We define a phase transition diagram which shows how algorithms behave for random problems, as the ratio of unknowns to equations and the ratio of the sparsity to equations varies. We find that whenever the number k of nonzeros in the sparsest solution is less than $d/2\log(n)$ then LARS/homotopy obtains the sparsest solution in k steps each of complexity $O(d^2)$.

1. INTRODUCTION

The problem we wish to solve is finding the sparsest solution to an underdetermined system of equations:

$$(P_0) \quad \min \|x\|_0 \text{ subject to } y = Ax.$$

where y is observed data, A is a known, $d \times n$ matrix, $d < n$, and x is an unknown vector in R^n , and $\|x\|_0$ represents the number of non-zeros. This is a non-convex combinatorial optimization problem, and in general, finding the sparsest solution is NP hard. Therefore we solve the problem for the ℓ_1 norm [4]:

$$(P_1) \quad \min \|x\|_1 \text{ subject to } y = Ax.$$

Since $d < n$, in such settings, the system of equations $y = Ax$ is underdetermined, and the ℓ^1 -norm minimization is a way

to regularize the solution. This problem can be cast as a standard linear program which is convex and tractable, and solved using general purpose solvers such as simplex and interior point methods [5] in order $O(n^3)$ which is slow for large scale problems. When the solution is sufficiently sparse there exists equivalence between the ℓ_1 and sparsest solutions [6].

There has been much interest in the statistical community in fitting regression models while imposing a sparsity constraint on the regression variables. This leads to the formulation of a minimization problem related to (P_1) ,

$$(L_q) \quad \min \|y - Ax\|_2^2 \text{ subject to } \|x\|_1 \leq q.$$

The matrix $A_{d \times n}$ is implicitly assumed to have $d > n$, i.e. representing an overdetermined linear system. Thus, the problem considered is a least-squares fit subject to an ℓ^1 -norm constraint on the variables; it is named Lasso by Tibshirani [2]. In the signal processing community, it is known in its augmented formulation

$$(D_\lambda) \quad \min \|y - Ax\|_2^2/2 + \lambda \|x\|_1.$$

Problem (D_λ) is named Basis Pursuit Denoising (BPDN) by Chen *et al.* [4]. It is equivalent to (L_q) under an appropriate correspondence of parameters. If \tilde{x}_λ is a solution to (D_λ) for some $\lambda \geq 0$, it also solves (L_q) for $q = \|\tilde{x}_\lambda\|_1$. One important distinction, perhaps, is that (D_λ) is studied in the underdetermined setting, i.e. $d < n$.

In the $d > n$ setting, Osborne, Presnell and Turlach (2000) [3] and later Efron, Hastie, Johnstone, and Tibshirani (2004) [1] developed an algorithm for solving (D_λ) for all $\lambda \geq 0$ or (L_q) for all $q \geq 0$. In detail, associate to each problem $(D_\lambda) : \lambda \in [0, \infty)$ a solution \tilde{x}_λ , then this identifies a polygonal solution path $\{\tilde{x}_\lambda : \lambda \in [0, \infty)\}$, with $x_\lambda = 0$ for λ large and, as $\lambda \rightarrow 0$, \tilde{x}_λ converging to the solution of (P_1) . The homotopy method of Osborne *et al.* [3], a.k.a the LARS/Lasso algorithm of Efron *et al.* [1], follows the solution path by jumping from vertex to vertex; it starts at $\tilde{x}_\lambda = 0$ for λ large, and then, in a sequence of steps, successively obtains the solutions \tilde{x}_{λ_ℓ} at a special problem-dependent sequence λ_ℓ associated to vertices of the polygonal path. The name homotopy refers to the fact that the objective function for (D_λ) is undergoing a homotopy from the ℓ^2 to the ℓ^1 objective as t decreases.

Recently, ℓ^1 -norm minimization problems have attracted attention [4, 7, 8, 9], with an eye to a range of important practical applications, particularly in conjunction with sparse representation. Applications of (P_1) have been proposed in the context of time-frequency representation [4], overcomplete signal representation [7], compressed sensing (CS) [10, 8], and error-correcting codes (ECC) [9]. In such applications, the underlying problem is to obtain a solution to $y = Ax$ which is as sparse as possible, in the sense of having few nonzero entries. The above-cited literature shows that, when the solution x_0 of (P_0) is sufficiently sparse, then the solution of (P_1) is either x_0 or an approximation to it.

2. THE PHASE PLANE, BREAKDOWN CURVE, AND SPARSITY PHASE TRANSITIONS

In this paper, we to bring to the fore the interplay between indeterminacy and sparsity. First, a *problem suite*, is a collection \mathcal{S} of (y, A) pairs obeying some common conditions, in particular, all of the same problem size (d, n) and all having a solution with at most k nonzeros. Second, for a given suite \mathcal{S} and algorithm \mathcal{A} , we say that the breakdown curve exceeds k if, for all members of the suite, the algorithm \mathcal{A} succeeds. The definition of success is that the algorithm correctly solves (P_1) and finds the sparsest solution. If the suite \mathcal{S} is a collection of random problems, we apply the same terminology if the algorithm is successful for a large fraction (rather than all) members of the suite. Finally, given a collection of problem suites, indexed by k, d, n , algorithm success/failure defines regions in a three-dimensional space indexed by k, d, n . Experience shows that it is convenient to consider slices $n = \text{constant}$, k and d varying. The two parameters k and d determine the sparsity of x_0 and the indeterminacy of the $d \times n$ system of equations; it is convenient to re-express these in the sparsity/indeterminacy ‘phase plane’ (δ, ρ) , where $\delta = d/n$ and $\rho = k/d$, and the interesting range is $0 \leq \delta, \rho \leq 1$. This plane shows how algorithms behave, as the ratio of unknowns to equations varies, but also the ratio of the sparsity to equations varies. In this paper, we describe algorithm performance by identifying regions of this plane where a certain algorithm is above or below a breakdown curve.

Theoretical results show that such a phase plane picture makes sense, from the following viewpoint. Suppose as an algorithm we consider a generic solver (P_1) , say, Michael Saunders’ interior point PDGO [5]. As matrix ensemble, we consider random matrices whose columns are uniformly distributed on the unit sphere. As problem suite, we consider such $d \times n$ random matrices together with vectors $y = Ax_0$ generated by vectors x_0 with k randomly-sited nonzeros. Figure 1 (a) shows a phase plane defined by the relative error $\|x - x_0\|_2 / \|x_0\|_2$ with $n = 500$ and varying d and k . Evidently, PDGO typically succeeds in recovering x_0 (despite the underdetermined system), provided the sparsity level is below the superimposed curve. This curve is derived theoretically in

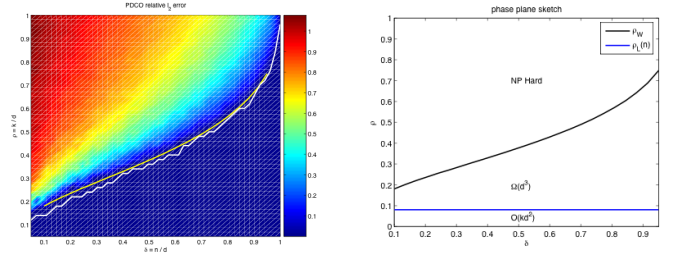


Fig. 1. (a) Empirical phase transition in the (ρ, δ) plane defined by the relative error $\|x - x_0\|_2 / \|x_0\|_2$ with $n = 500$. The solution x_0 is recovered provided the sparsity level is below the theoretical curve ρ_W . (b) Theoretical phase transition in the (ρ, δ) plane. The solution x_0 is recovered provided the sparsity level is below the theoretical curve ρ_W .

[6, 11], which shows that for $k < \rho(d/n)d(1 + o(1))$, the ℓ^1 solution in such an ensemble also gives the sparsest possible solution.

To further fix ideas, Figure 1 (b) displays a phase plane, with several curves bounding different regions. The suites being considered again involve random matrices A and vectors y , and the phase plane is partitioned into several regions. The highest region is marked Hard; in this region, the sparsest solution to $y = Ax$ is not very sparse, and finding the sparsest solution requires solving (P_0) . This problem is NP-hard in general. The Hard region is bounded below by the curve ρ_W mentioned above. Below, the indicated curve, a general solver for (P_1) is a good procedure for finding the sparsest solution. The second highest region is marked $\Omega(n^3)$, which means that in this regime, algorithms we discuss for solving (P_1) correctly find the sparsest possible solution require at least $O(d^3)$ time. The next bounding curve is at $d/(2\log(n))$, and sits on top of a region labelled $O(kd^2)$. In this regime there are algorithms which stop in k steps each one of which costs $O(d^2)$.

In short, there are algorithms successful for finding the sparsest solution in the lower two regions. Note that n is fixed in this 2-D picture; as n increases, the upper curve ρ_W will stay fixed, but the bottom curve is lower.

Borrowing terminology from statistical physics, we call the boundaries between the indicated regions of such a diagram *phase transitions*. In overview then, in this paper we present algorithms which in certain conditions rapidly solve (P_1) , and explain conditions under which this is possible using terminology associated with phase plane, breakdown curves, and phase transitions.

3. PROPOSAL FOR SOLVING (P_1)

The LARS, Lasso and homotopy methods described can be implemented using already-published software [12]. A key point here is that this software is not proposed for solving

linear programs as presented here; it is instead proposed for solving problems of approximate modeling of noisy data. Our contribution is to propose its use in a different setting.

Malioutov *et al.* [13] apply the homotopy method to the formulation (D_λ) in the underdetermined setting, when the data is noisy. We follow their ideas and suggest the following scheme for solving (P_1) . We apply the homotopy method to the noiseless data $y = Ax$. Follow the solution path from $0 = x_{t_0}$ to \tilde{x}_0 . When the algorithm reaches the $t = 0$ limit, (P_1) is solved.

If the homotopy method stops in k steps, the work required is kd^2 ; which is substantially less than the $O(n^3)$ work required to solve a generic system of equations. So early stopping of the homotopy method implies a fast solution. The homotopy algorithm has the k -step solution property at a given problem instance (y, A) if that instance $y = Ax_0$ for some k -sparse vector x_0 , and if the homotopy algorithm stops in at most k steps.

The mutual coherence $M(A)$ of a matrix A whose columns are normalized to length 1 is the maximal off-diagonal entry of the Gram matrix $A^T A$. A matrix is incoherent if $M(A)$ is small, and the smallest it can be is $1/\sqrt{d}$, [7, 14]. Such matrices are somewhat like orthogonal matrices, but they can be very non-square; we can have $M(A) = 1/\sqrt{d}$ for matrices which are $d \times d^2$. Consider the suite $\mathcal{S}_{inc}(d, n, \mu, k)$ of problems (y, A) involving matrices with $M(A) \leq \mu$ and involving left-hand sides y admitting sparse representation $y = Ax_0$ with $\|x_0\|_0 \leq k$. If $k < (\mu^{-1} + 1)/2$, then the homotopy algorithm has the k -step property. For example, picking $\mu = 1/\sqrt{d}$, we find that homotopy has the k -step property for $k < \sqrt{d}/2$.

Suppose that A is a random matrix from the uniform spherical ensemble. That is, the columns of A are uniformly distributed points on the unit sphere. As indicated above, such matrices A are incoherent, but much more is true. To see this, consider the setting where n and d are both large, tending to infinity together in a proportional way: $d = \delta n$, $0 < \delta < 1$. The parameter δ gives the limiting shape of the matrix A . Since the matrix A has exchangeable columns, we may focus attention on the situation where $y = Ax_0$ and x_0 has nonzeros in the k positions $1, \dots, k$. Fix $\varepsilon > 0$. Suppose that $y = Ax_0$ where x_0 has k nonzeros in the positions $1, \dots, k$ say, where

$$k \leq \frac{d}{2 \log(n)} (1 - \varepsilon).$$

With overwhelming probability for large n the solution of the minimum ℓ^1 problem (P_1) is unique and is precisely x_0 , and the homotopy algorithm runs k steps and stops, delivering the solution x_0 . This $\frac{d}{2 \log(n)}$ ceiling is much stronger than the result $k \leq \sqrt{d/8 \log(n)}$ implied by incoherence alone. Most importantly for applications it is nearly proportional to d .

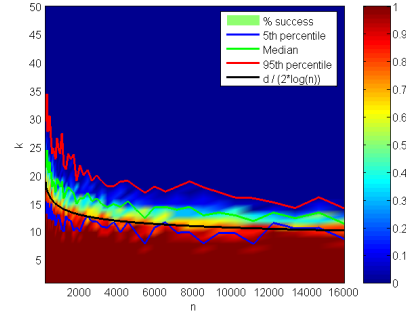


Fig. 2. (k, n) plane: The probability in which the number of iterations until convergence is k for $d = 200$, $k = 1 \dots 50$, and $n = 200 \dots 16000$.

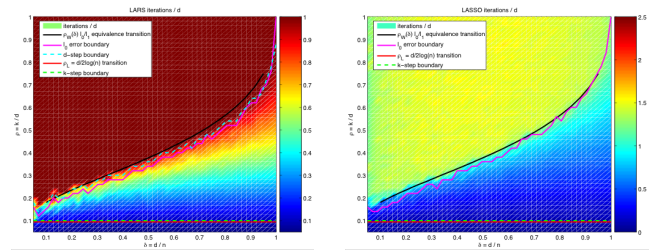


Fig. 3. LARS (left) and LASSO (right) iterations/ d for vectors x_0 with k nonzeros drawn from random uniform distribution and matrices $A_{d \times n}$ from random signs ensembles, for $n = 200$.

4. EMPIRICAL EVIDENCE

We validate the property described above by exploring the (k, n) space. We generate matrices $A_{d \times n}$ from the uniform spherical ensemble for varying values of n and fixed d , and vectors x with k non-zeros from a random uniform distribution. We then compute the observation $y = Ax$ and solve the problem given (A, y) using the homotopy method. We repeat the experiment for each point in the (k, n) plane and compute the number of times the property occurs. Figure 2 shows the the probability of success for each value in the (k, n) plane for $d = 200$, and $k = 1 \dots 50$ and $n = 200 \dots 16000$.

Figure 3 shows the number of iterations divided by d in the ρ, δ plane, for a vector x_0 from random uniform distribution with k non-zeros, and matrices $A_{d \times n}$ from random signs ensembles. The curves overlaid show the derived theoretical ℓ_0/ℓ_1 phase transition $\rho_W(\delta)$ (in black); empirical ℓ_0 error boundary (in magenta); derived theoretical k -step phase transition $\rho_L = d/(2 \log(n))$ (in red); and empirical k -step boundary (dashed green). For LARS the Figure shows the empirical d -step boundary (dashed cyan).

The number of LARS iterations to solution in the (k, n) plane has a phase transition at the curve $\rho_L = d/(2 \log(n))$. Below ρ_L the number of iterations to convergence is k .

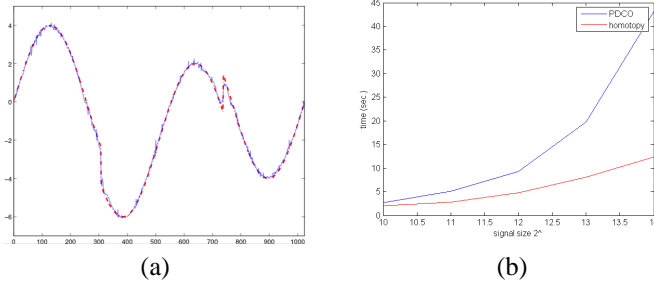


Fig. 4. (a) Compressed sensing and (b) computation times.

5. APPLICATIONS

Finally, we present some practical results of important applications involving ℓ_1 minimization solved by LARS/homotopy methods.

5.1. Compressed Sensing

The notion of Compressed Sensing [10, 15, 8], is that a signal, compressible in a known basis, (e.g. wavelet or Fourier), is reconstructed by ℓ_1 minimization from fewer measurements than the nominal sampling density, provided that the samples are made on a specially transformed version of the signal. Roughly speaking, the samples measure linear functionals which look like random linear combinations of the basis functions. Figure 4 (a) shows a signal of length 2^{10} (blue) reconstructed (red) by compressed sensing from $d = 180$ sample using the LARS/homotopy method. Compressed sensing computation time is 2.5 seconds by 258 iterations with 0.06 RMSE. Figure 4 (b) compares actual computation time of PDCO and homotopy for compressed sensing with signal lengths 2^n for $n = 10 \dots 15$ on a standard PC running matlab.

5.2. Decoding Error Correcting Codes

Efficiently recovering a signal despite malicious errors [16] can be formulated as solving a minimum ℓ_1 problem. For large n which is divisible by 4 we generate a random orthogonal matrix $U_{n \times n}$. Then form a matrix $A_{d \times n}$ with $\lfloor \frac{3n}{4} \rfloor$ by the first d rows of U . Then generate its $m = n - d$ by n orthocomplement B from the last m rows. In order to communicate a block x of m pieces of information to a receiver we transmit the noiseless signal $S = B^T x$. The receiver gets the corrupted signal $r = B^T x + z$, and solves the minimum ℓ_1 problem: $\min \|r - B^T x\|_1$. Since $AB^T = 0$, this is equivalent to solving the minimum ℓ_1 problem: $\min \|z\|_1$ subject to $y = Az$, and then recovering the signal $\hat{x} = B(r - \hat{z})$. Figure 5 shows the recovered noise \hat{z} and reconstructed signal \hat{x} (red) of length $n = 2^{10}$ by the LARS/homotopy method. Computation time is 108.4 seconds by 269 iterations with RMSE $5.8e-8$.

- ## 6. REFERENCES
- [1] Bradley Efron, Trevor Hastie, Iain M. Johnstone, and Robert Tibshirani, "Least angle regression," *The Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.
 - [2] Robert Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society*, vol. 58, no. 1, pp. 267–288, 1996.
 - [3] Michael R. Osborne, Brett Presnell, and Berwin A. Turlach, "A new approach to variable selection in least squares problems," *IMA J. Numerical Analysis*, vol. 20, pp. 389–403, 2000.
 - [4] Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comp.*, vol. 20, no. 1, pp. 33–61, 1999.
 - [5] Michael A. Saunders and Byunggyoo Kim, "Pdco: Primal-dual interior method for convex objectives," <http://www.stanford.edu/group/SOL/software/pdco.html>.
 - [6] David L. Donoho, "For most underdetermined systems of linear equations, the minimal ℓ^1 -norm near-solution approximates the sparsest near-solution," *Comm. Pure and Appl. Math.*, 2004.
 - [7] David L. Donoho and Xiaoming Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Trans. Info. Theory*, vol. 47, no. 7, pp. 2845–2862, 2001.
 - [8] Emmanuel J. Candès, Justin Romberg, and Terence Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," Tech. Rep., California Institute of Technology, 2004.
 - [9] Emmanuel Candès and Terence Tao, "Decoding by linear programming," Tech. Rep., California Institute of Technology, 2004.
 - [10] David L. Donoho, "Compressed sensing," Tech. Rep., Dept. of Statistics, Stanford University, 2004.
 - [11] David L. Donoho, "High-dimensional centrosymmetric polytopes with neighborliness proportional to dimension," Tech. Rep., Dept. of Statistics, Stanford University, 2005.
 - [12] Bradley Efron and Trevor Hastie, "Lars software web site," <http://www-stat.stanford.edu/hastie/Papers/LARS/>.
 - [13] Dmitry M. Malioutov, Müjdat Çetin, and Alan S. Willsky, "Optimal sparse representations in general overcomplete bases," in *IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2004.
 - [14] David L. Donoho and Michael Elad, "Optimally sparse representation from overcomplete dictionaries via ℓ^1 norm minimization," *Proc. Natl. Acad. Sci. USA*, vol. 100, no. 5, pp. 2197–2002, March 2002.
 - [15] Anna C. Gilbert, S. Guha, P. Indyk, S. Muthukrishnan, and M. Strauss, "Near-optimal sparse fourier representations via sampling," 2002.
 - [16] David L. Donoho, "Neighborly polytopes and sparse solution of underdetermined linear equations," Tech. Rep., Dept. of Statistics, Stanford University, 2004.

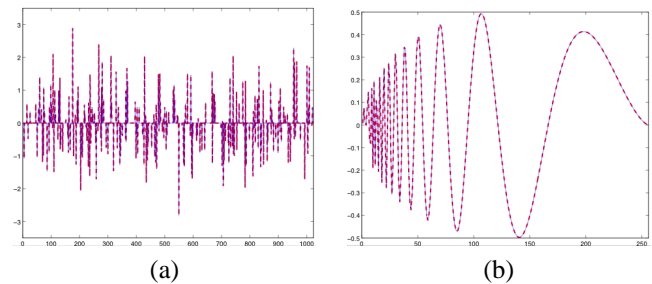


Fig. 5. Decoding error correcting codes: (a) recovered noise \hat{z} and (b) reconstructed signal \hat{x} .