GEOPHYSICS, VOL. 73, NO. 4 (JULY-AUGUST 2008); P. A23–A26, 5 FIGS.
10.1190/1.2944169

New insights into one-norm solvers from the Pareto curve

Gilles Hennenfent[1], Ewout van den Berg[2], Michael P. Friedlander[2], and Felix J. Herrmann[1]

ABSTRACT

Geophysical inverse problems typically involve a trade-off between data misfit and some prior model. Pareto curves trace the optimal trade-off between these two competing aims. These curves are used commonly in problems with two-norm priors in which they are plotted on a log-log scale and are known as L-curves. For other priors, such as the sparsity-promoting one-norm prior, Pareto curves remain relatively unexplored. We show how these curves lead to new insights into one-norm regularization. First, we confirm theoretical properties of smoothness and convexity of these curves from a stylized and a geophysical example. Second, we exploit these crucial properties to approximate the Pareto curve for a large-scale problem. Third, we show how Pareto curves provide an objective criterion to gauge how different one-norm solvers advance toward the solution.

INTRODUCTION

Many geophysical inverse problems are ill posed (Parker, 1994) because their solutions are not unique or are acutely sensitive to changes in data. To solve this kind of problem stably, additional information must be introduced. This technique is called regularization (see, e.g., Phillips, 1962; Tikhonov, 1963).

Specifically, when the solution of an ill-posed problem is known to be (almost) sparse, Oldenburg et al. (1983) and others have observed that a good approximation to the solution can be obtained by using one-norm regularization to promote sparsity. More recently, results in information theory have breathed new life into the idea of promoting sparsity to regularize ill-posed inverse problems. These results establish that under certain conditions, the sparsest solution of a (severely) underdetermined linear system can be recovered exactly by seeking the minimum one-norm solution (Candès et al., 2006; Donoho, 2006; Rauhut, 2007). This has led to tremendous activity in the newly established field of compressed sensing. Several new one-norm solvers have appeared in response (see, e.g., Daubechies et al., 2004; van den Berg and Friedlander, 2008). In the context of geophysical applications, it is a challenge to evaluate and compare these solvers against more standard approaches such as iteratively reweighted least-squares (IRLS) (Gersztenkorn et al., 1986), which uses a quadratic approximation to the one-norm regularization function.

In this paper, we propose an approach to understand the behavior of algorithms for solving one-norm regularized problems. The approach consists of tracking on a graph the data misfit versus the one norm of successive iterates. The Pareto curve traces the optimal trade-off in the space spanned by these two axes and gives a rigorous yardstick for measuring the quality of the solution path generated by an algorithm. In the context of the two-norm (i.e., Tikhonov) regularization, the Pareto curve often is plotted on a log-log scale and is called the L-curve (Lawson and Hanson, 1974). We draw on the work of van den Berg and Friedlander (2008), who examine the theoretical properties of the one-norm Pareto curve. Our goal is to understand the compromises accepted implicitly when an algorithm is given a limited number of iterations.

PROBLEM STATEMENT

Consider the following underdetermined system of linear equations

$$\mathbf{y} = \mathbf{A}\mathbf{x}_0 + \mathbf{n}, \tag{1}$$

where the $n =$ vectors $\mathbf{y}$ and $\mathbf{n}$ represent observations and additive noise, respectively. The n-by-N matrix $\mathbf{A}$ is the modeling operator that links the model $\mathbf{x}_0$ to the noise-free data given by $\mathbf{y} - \mathbf{n}$. We assume that $N \gg n$ and $\mathbf{x}_0$ have few nonzero or significant entries. We use the terms *model* and *observations* in a broad sense so that many linear geophysical problems can be cast in the form shown in equation 1. In the case of wavefield reconstruction, for example, $\mathbf{y}$ is the acquired seismic data with missing traces, and $\mathbf{A}$ can be the restriction operator combined with the curvelet synthesis operator so that

Manuscript received by the Editor 18 December 2007; revised manuscript received 3 March 2008; published online 24 June 2008.
[1]University of British Columbia, Department of Earth and Ocean Science, Seismic Laboratory for Imaging and Modeling, Vancouver, British Columbia, Canada. E-mail: ghennenfent@eos.ubc.ca; mpf@cs.ubc.ca; fherrmann@eos.ubc.ca.
[2]University of British Columbia, Department of Computer Science, Scientific Computing Laboratory, Vancouver, British Columbia, Canada. E-mail: ewout78@cs.ubc.ca; mpf@cs.ubc.ca.

$\mathbf{x}_0$ is the curvelet representation of the fully sampled wavefield (Hennenfent and Herrmann, 2008; Herrmann and Hennenfent, 2008).

Because $\mathbf{x}_0$ is assumed to be (almost) sparse, one can promote sparsity as a prior via one-norm regularization to overcome the singular nature of $\mathbf{A}$ when estimating $\mathbf{x}_0$ from $\mathbf{y}$. A common approach is to solve the convex optimization problem

$$\mathrm{QP}_\lambda\colon \min_{\mathbf{x}} \tfrac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1, \tag{2}$$

which is related closely to quadratic programming (QP). The positive parameter λ is the Lagrange multiplier, which balances the trade-off between the two norm of the data misfit and the one norm of the solution. Many algorithms are available for solving QP_λ, including IRLS; iterative soft thresholding (IST), introduced by Daubechies et al. (2004); and the IST extension to include cooling (ISTc) (Figueiredo and Nowak, 2003), which was tailored to geophysical applications by Herrmann and Hennenfent (2008).

Generally it is not clear; however, how to choose the parameter λ so that the solution of QP_λ is optimal in some sense. A directly related optimization problem, the basis-pursuit (BP) denoise problem (Chen et al., 1998), minimizes the one norm of the solution given a maximum misfit and is given by

$$\mathrm{BP}_\sigma\colon \min_{\mathbf{x}} \|\mathbf{x}\|_1 \text{ s.t. } \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 \le \sigma. \tag{3}$$

This formulation often is preferred when an estimate of the noise level $\sigma \ge 0$ in the data is available. BP_σ can be solved using ISTc or the spectral projected-gradient algorithm ($\mathrm{SPG}\ell_1$) introduced by van den Berg and Friedlander (2008).

For interest, a third optimization problem, connected to QP_λ and BP_σ, minimizes the misfit given a maximum one norm of the solution and is given by the lasso (LS) problem (Tibshirani, 1996)

$$\mathrm{LS}_\tau\colon \min_{\mathbf{x}} \tfrac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 \text{ s.t. } \|\mathbf{x}\|_1 \le \tau. \tag{4}$$

Because an estimate of the one norm of the solution $\tau \ge 0$ is not typically available for geophysical problems, this formulation seldom is used directly. However, it is a key internal problem used by $\mathrm{SPG}\ell_1$ to solve BP_σ.

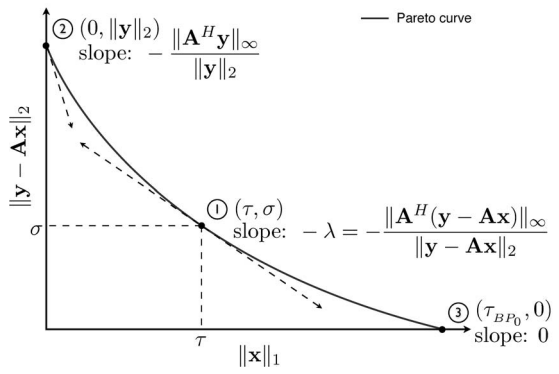

Figure 1. Schematic illustration of a Pareto curve. Point 1 exposes the connection among parameters QP_λ, BP_σ, and LS_τ. Point 3 corresponds to a solution of BP_σ with $\sigma = 0$.

To understand the connection between these approaches and to compare their related solvers in different scenarios, we propose to follow Daubechies et al. (2007) and van den Berg and Friedlander (2008) and look at the Pareto curve.

PARETO CURVE

Figure 1 gives a schematic illustration of a Pareto curve. The curve traces the optimal trade-off between $\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2$ and $\|\mathbf{x}\|_1$ for a specific pair of $\mathbf{A}$ and $\mathbf{y}$ in equation 1. Point 1 clarifies the connection among the three parameters of QP_λ, BP_σ, and LS_τ. Coordinates of a point on the Pareto curve are (τ, σ), and the slope of the tangent at this point is $-\lambda$. End points of the curve, points 2 and 3, are two special cases. When $\tau = 0$, the solution of LS_τ is $\mathbf{x} = 0$ (point 2). It coincides with the solutions of BP_σ with $\sigma = \|\mathbf{y}\|_2$ and QP_λ with $\lambda = \|\mathbf{A}^H\mathbf{y}\|_\infty/\|\mathbf{y}\|_2$. (The infinity norm $\|\cdot\|_\infty$ is given by $\max(|\cdot|)$.) When $\sigma = 0$, the solution of BP_σ (point 3) coincides with the solutions of LS_τ, where τ is the one norm of the solution, and QP_λ, where $\lambda = 0^+$, i.e., λ infinitely close to zero from above.

These relations are formalized as follows in van den Berg and Friedlander (2008):

Result 1. The Pareto curve (1) is convex and decreasing, (2) is continuously differentiable, and (3) has a negative slope $\lambda = \|\mathbf{A}^H\mathbf{r}\|_\infty/\|\mathbf{r}\|_2$ with the residual $\mathbf{r}$ given by $\mathbf{y} - \mathbf{A}\mathbf{x}$.

For large-scale geophysical applications, it is not practical or even feasible to sample the entire Pareto curve. However, its regularity, as implied by this result, means it is possible to obtain a good approximation to the curve with very few interpolating points, as illustrated later in this paper.

COMPARISON OF ONE-NORM SOLVERS

To illustrate the usefulness of the Pareto curve, we compare IST, ISTc, $\mathrm{SPG}\ell_1$, and IRLS on a noise-free problem and compute a solution of BP_σ for $\sigma = 0$, i.e., BP_0. This case is especially challenging for solvers that attack QP_λ, e.g., IST, ISTc, and IRLS, because the corresponding solution can be attained only in the limit as $\lambda \to 0$.

We construct a benchmark problem that typically is used in the compressed-sensing literature (Donoho et al., 2006). The matrix $\mathbf{A}$ is taken to have Gaussian-independent and identically distributed entries. A sparse solution $\mathbf{x}_0$ is generated randomly, and the observations $\mathbf{y}$ are computed according to equation 1.

Solution paths

Figure 2 shows the solution paths of the four solvers as they converge to the BP_0 solution. The starting vector provided to each solver is the zero vector; hence, the paths start at $(0, \|\mathbf{y}\|_2)$, point 2 in Figure 1. The number of iterations are large enough for each solver to converge; therefore, the solution paths end at $(\tau_{BP_0}, 0)$, point 3 in Figure 1.

The two solvers $\mathrm{SPG}\ell_1$ and ISTc approach the BP_0 solution from the left and remain close to the Pareto curve. In contrast, IST and IRLS aim at a least-squares solution before turning back toward the BP_0 solution. ISTc solves QP_λ for a decreasing sequence $\lambda_i \to 0$. The starting vector for QP_{λ_i} is the solution of $\mathrm{QP}_{\lambda_{i-1}}$, which by definition is on the Pareto curve. This explains why ISTc follows the curve so closely. $\mathrm{SPG}\ell_1$ solves a sequence of LS_τ problems for an increasing sequence of $\tau_i \to \tau_{BP_0}$, hence the vertical segments along the $\mathrm{SPG}\ell_1$ solution path. IST solves QP_{0^+}. Because there is hardly any regular-

ization, IST first works toward minimizing the data misfit. When the misfit is sufficiently small, the effect of the one-norm penalization starts, yielding a change of direction toward the BP_0 solution. IRLS solves a sequence of weighted, damped least-squares problems. Because the weights are initialized to ones, IRLS first reaches the standard least-squares solution. The estimates obtained from subsequent reweightings have a smaller one norm while maintaining the residual (close) to zero. Eventually, IRLS gets to the BP_0 solution.

Practical considerations

In geophysical applications, problem sizes are large, and there is a severe computational constraint. We can use the technique outlined above to understand the robustness of a given solver that is limited by a maximum number of iterations or matrix-vector products that can be performed.

Figure 3 shows the Pareto curve and the solution paths of various solvers in which the maximum number of iterations is fixed. This roughly equates to using the same number of matrix-vector products for each solver. Whereas, SPGℓ_1 continues to provide a fairly accurate approximation to the BP_0 solution, those computed by IST, ISTc, and IRLS suffer from larger errors. IST stops before the one-norm regularization goes into effect; hence, the data misfit at the candidate solution is small, but the one norm is completely incorrect. ISTc and IRLS accumulate small errors along their paths because there are not enough iterations to solve each subproblem to sufficient accuracy. Note that both solvers accumulate errors along both axes.

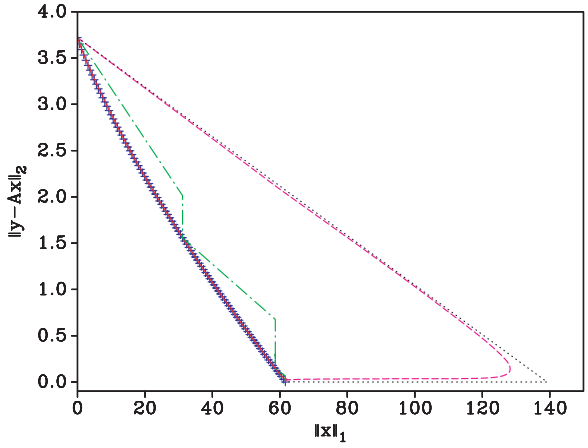

Figure 2. Pareto curve and solution paths (large enough number of iterations) of four solvers for a BP_0 problem. The symbols + represent a sampling of the Pareto curve. The solid line (—), obscured by the Pareto curve, is the solution path of ISTc; the chain line (—·—) is the path of SPGLℓ_1; the dashed line (– –) is the path of IST; and the dotted line (···) is the path of IRLS.

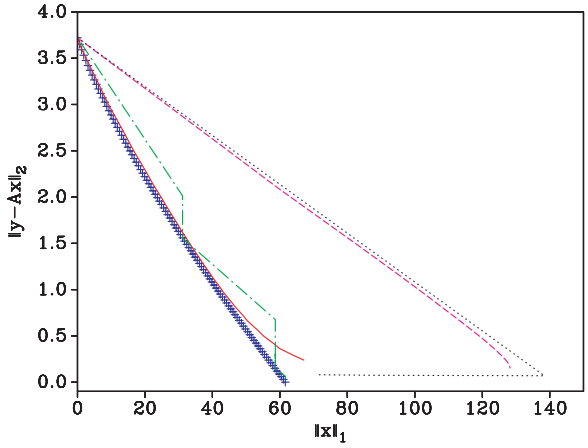

Figure 3. Pareto curve and optimization paths (same, limited number of iterations) of four solvers for a BP_0 problem (see Figure 2 for legend).

GEOPHYSICAL EXAMPLE

As a concrete example of the use of the Pareto curve in the geophysical context, we study the problem of wavefield reconstruction with sparsity-promoting inversion in the curvelet domain (CRSI) (Herrmann and Hennenfent, 2008). The simulated acquired data, shown in Figure 4a, corresponds to a shot record with 35% of traces missing. The interpolated result, shown in Figure 4b, is obtained by solving BP_0 using SPGℓ_1. This problem has more than half a million unknowns and 42,000 data points.

Points in Figure 5 are samples of the corresponding Pareto curve. The regularity of these points strongly indicates that the underlying curve, which we know to be convex, is smooth and well behaved and empirically supports our earlier claim. However, problems of practi-

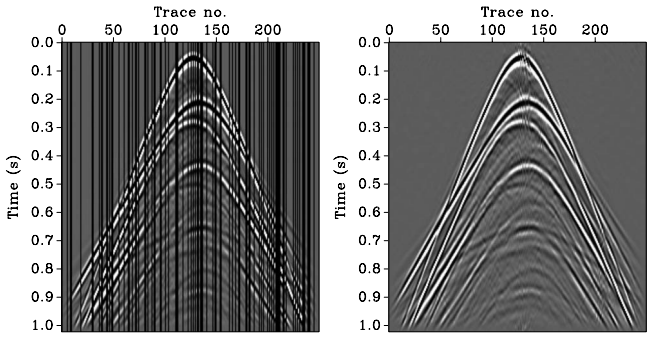

Figure 4. CRSI on synthetic data. (a) Input and (b) interpolated data using CRSI with SPGℓ_1.

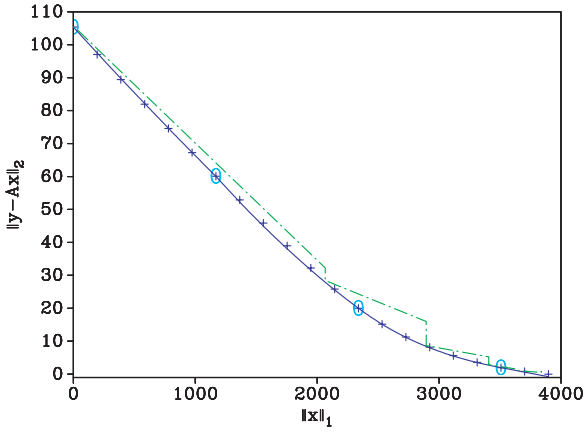

Figure 5. Pareto curve and SPGℓ_1 solution path for a CRSI problem. The symbols + represent a fine accurate sampling of the Pareto curve. The solid line (—) is an approximation to the Pareto curve using the few circled points, and the chain line (—·—) is the solution path of SPGℓ_1.

cal interest are often significantly larger, and it can be prohibitively expensive to compute a similarly fine sampling of the curve.

Because the curve is well behaved, we can leverage its smoothness and use a small set of samples to obtain a good interpolation. The solid line in Figure 5 shows an interpolation based only on information from the circled samples. The interpolated curve closely matches samples that were not included in the interpolation. Figure 5 also plots the iterates taken by SPGℓ_1 to obtain the reconstruction shown in Figure 4b. The plot shows that the iterates remain close to the Pareto curve and converge toward the BP_0 solution.

CONCLUSIONS

The sheer size of seismic problems makes it a certainty that there will be significant constraints on the amount of computation that can be done when solving an inverse problem. Hence, it is especially important to explore the nature of a solver's iterations to make an informed decision on how best to truncate the solution process. The Pareto curve serves as the optimal reference, which makes an unbiased comparison among different one-norm solvers possible.

Of course, it is prohibitively expensive in practice to compute the entire Pareto curve exactly. However, we observe that the Pareto curves for many of the one-norm regularized problems are regular, as confirmed by the theoretical result 1. This suggests that it is possible to approximate the Pareto curve by fitting a curve to a small set of sample points, taking into account derivative information at these points. As such, insights from the Pareto curve can be leveraged to large-scale one-norm regularized problems, as we illustrate in a geophysical example. This prospect is particularly exciting given the current resurgence of this type of regularization in many areas of research.

ACKNOWLEDGMENTS

The authors are grateful to Sergey Fomel and Tamas Nemeth for their valuable comments and to Eric Verschuur for the synthetic data. The authors also thank the anonymous reviewers and associate editor for their comments, which certainly helped to improve this paper. This publication was prepared using Madagascar, a package for reproducible computational experiments; SPGℓ_1; and Sparco, a suite of linear operators and problems for testing algorithms for sparse signal reconstruction. This research was supported financially in part by NSERC Discovery Grant 22R81254, of F.J.H., and by CRD Grant DNOISE 334810-05, of F.J.H. and M.P.F. Research was carried out as part of the SINBAD project with support secured through Industry Technology Facilitator from the following organizations: BG Group, BP, Chevron, ExxonMobil, and Shell.

REFERENCES

Candès, E. J., J. Romberg, and T. Tao, 2006, Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information: IEEE Transactions on Information Theory, **52**, 489–509.

Chen, S. S., D. L. Donoho, and M. A. Saunders, 1998, Atomic decomposition by basis pursuit: SIAM Journal on Scientific Computing, **20**, 33–61.

Daubechies, I., M. Defrise, and C. De Mol, 2004, An iterative thresholding algorithm for linear inverse problems with a sparsity constraint: Communications on Pure and Applied Mathematics, **57**, 1413–1457.

Daubechies, I., M. Fornasier, and I. Loris, 2007, Accelerated projected gradient method for linear inverse problems with sparsity constraints: ArXiv e-prints, 706, no. 0706.4297, accessed April 23, 2008; http://adsabs.harvard.edu/abs/2007arXiv0706.4297D.

Donoho, D. L., 2006, Compressed sensing: IEEE Transactions on Information Theory, **52**, 1289–1306.

Donoho, D. L., Y. Tsaig, I. Drori, and J. L. Starck, 2006, Sparse solution of underdetermined linear equations by stagewise orthogonal matching pursuit: Technical Report TR-2006-2, Stanford Statistics Department, accessed April 23, 2008; http://stat.stanford.edu/~idrori/StOMP.pdf.

Figueiredo, M., and R. Nowak, 2003, An EM algorithm for wavelet-based image restoration: IEEE Transactions on Image Processing, **12**, 906–916.

Gersztenkorn, A., J. B. Bednar, and L. Lines, 1986, Robust iterative inversion for the one-dimensional acoustic wave equation: Geophysics, **51**, 357–369.

Hennenfent, G., and F. J. Herrmann, 2008, Simply denoise: Wavefield reconstruction via jittered undersampling: Geophysics, **73**, no. 3, V19–V28.

Herrmann, F. J., and G. Hennenfent, 2008, Non-parametric seismic data recovery with curvelet frames: Geophysical Journal International, **173**, no. 1, 233–248.

Lawson, C. L., and R. J. Hanson, 1974, Solving least squares problems: Prentice-Hall, Inc.

Oldenburg, D., T. Scheuer, and S. Levy, 1983, Recovery of the acoustic impedance from reflection seismograms: Geophysics, **48**, 1318–1337.

Parker, R. L., 1994, Geophysical inverse theory: Princeton University Press.

Phillips, D. L., 1962, A technique for the numerical solution of certain integral equations of the first kind: Journal of the Association for Computing Machinery, **9**, 84–97.

Rauhut, H., 2007, Random sampling of sparse trigonometric polynomials: Applied and Computational Harmonic Analysis, **22**, 16–42.

Tikhonov, A. N., 1963, Solution of incorrectly formulated problems and regularization method: Soviet mathematics - Doklady, **4**, 1035–1038.

Tibshirani, R., 1996, Regression shrinkage and selection via the LASSO: Journal of the Royal Statistical Society, Series B, **58**, 267–288.

van den Berg, E., and M. P. Friedlander, 2008, Probing the Pareto frontier for basis pursuit solutions: Technical Report TR-2008-01, UBC Computer Science Department, accessed April 23, 2008; http://www.optimization-online.org/DB_HTML/2008/01/1889.html.