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. Pareto curves trace the optimal trade off between these two competing aims. These curves are commonly used in problems with two-norm priors where they are plotted on a log-log scale and are known as L-curves. For other priors, such as the sparsity-promoting one norm, Pareto curves remain relatively unexplored. We show how these curves lead to new insights into one-norm regularization. First, we confirm the 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 towards the solution.

INTRODUCTION

Many geophysical inverse problems are ill posed (Parker, 1994)—their solutions are not unique or are acutely sensitive to changes in the 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 *exactly* recovered 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

[1]Seismic Laboratory for Imaging and Modeling, Department of Earth and Ocean Sciences, The University of British Columbia, 6339 Stores Road, Vancouver, V6T 1Z4, BC, Canada

[2]Scientific Computing Laboratory, Department of Computer Science, The University of British Columbia, 2366 Main Mall, Vancouver, V6K 1Z4, BC, Canada

Accepted 2008 March 7. Received 2008 March 3. In original form 2007 December 18.

have appeared in response (see, e.g., Daubechies et al., 2004; van den Berg and Friedlander, 2008, and references therein). 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 letter, 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 tradeoff 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 is often 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 implicitly accepted 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 that $\mathbf{x}_0$ has 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, $\mathbf{A}$ can be the restriction operator combined with the curvelet synthesis operator so that $\mathbf{x}_0$ is the curvelet representation of the fully-sampled wavefield (Herrmann and Hennenfent, 2008; Hennenfent and Herrmann, 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

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

which is closely related to quadratic programming (QP); the positive parameter λ is the Lagrange multiplier, which balances the tradeoff 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).

Accepted 2008 March 7. Received 2008 March 3. In original form 2007 December 18.

It is generally not clear, however, how to choose the parameter λ such that the solution of QP_λ is, in some sense, optimal. 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

$$\mathrm{BP}_\sigma : \quad \min_{\mathbf{x}} \ \|\mathbf{x}\|_1 \quad \text{s.t.} \quad \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 \leq \sigma.$$

This formulation is often preferred when an estimate of the noise level $\sigma \geq 0$ in the data is available. BP_σ can be solved using ISTc or the spectral projected-gradient algorithm (SPGℓ_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)

$$\mathrm{LS}_\tau : \quad \min_{\mathbf{x}} \ \tfrac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 \quad \text{s.t.} \quad \|\mathbf{x}\|_1 \leq \tau.$$

Because an estimate of the one norm of the solution $\tau \geq 0$ is typically not available for geophysical problems, this formulation is seldom used directly. It is, however, a key internal problem used by SPGℓ_1 in order to solve BP_σ.

To understand the connection between these approaches and 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 tradeoff 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 between the three parameters of QP_λ, BP_σ, and LS_τ. The coordinates of a point on the Pareto curve are (τ, σ) and the slope of the tangent at this point is $-\lambda$. The 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 i) is convex and decreasing, ii) is continuously differentiable, and iii) 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 that it is possible to obtain a good approximation to the curve with very few interpolating points, as illustrated later in this letter.

Accepted 2008 March 7. Received 2008 March 3. In original form 2007 December 18.

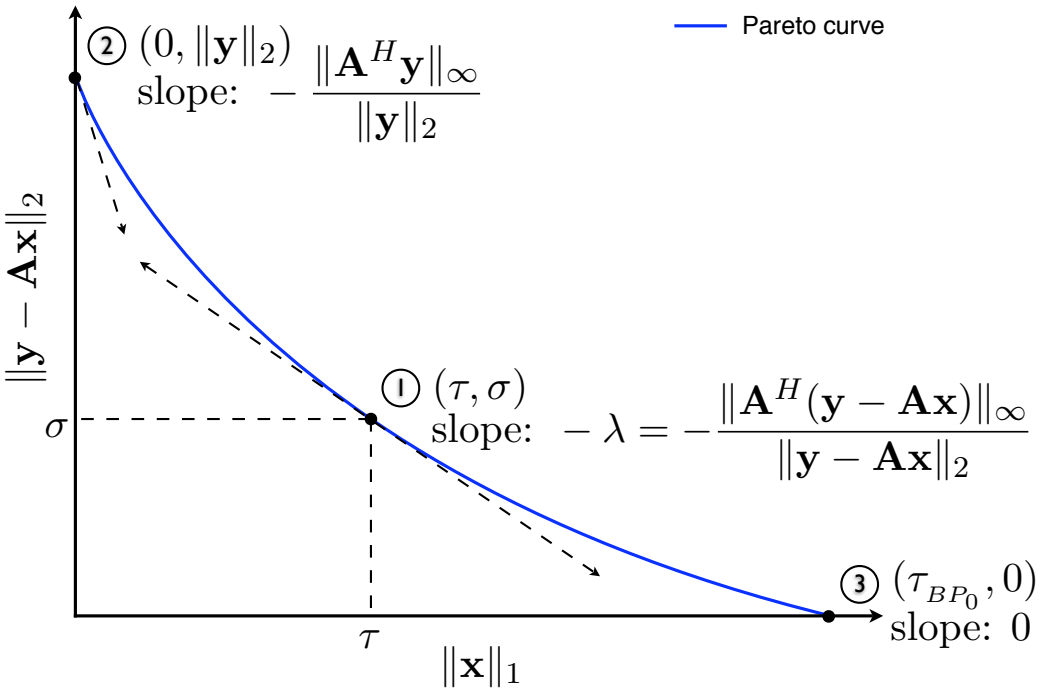


Figure 1: Schematic illustration of a Pareto curve. Point ① exposes the connection between the three parameters of QP_λ, BP_σ, and LS_τ. Point ③ corresponds to a solution of BP_σ with $\sigma = 0$.

COMPARISON OF ONE-NORM SOLVERS

To illustrate the usefulness of the Pareto curve, we compare IST, ISTc, SPGℓ_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 only be attained in the limit as $\lambda \to 0$.

We construct a benchmark problem that is typically 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 randomly generated, 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, and hence the paths start at $(0, \|\mathbf{y}\|_2)$—point ② in Figure 1. The number of iterations is large enough for each solver to converge, and therefore the solution paths end at $(\tau_{BP_0}, 0)$—point ③ in Figure 1.

Accepted 2008 March 7. Received 2008 March 3. In original form 2007 December 18.

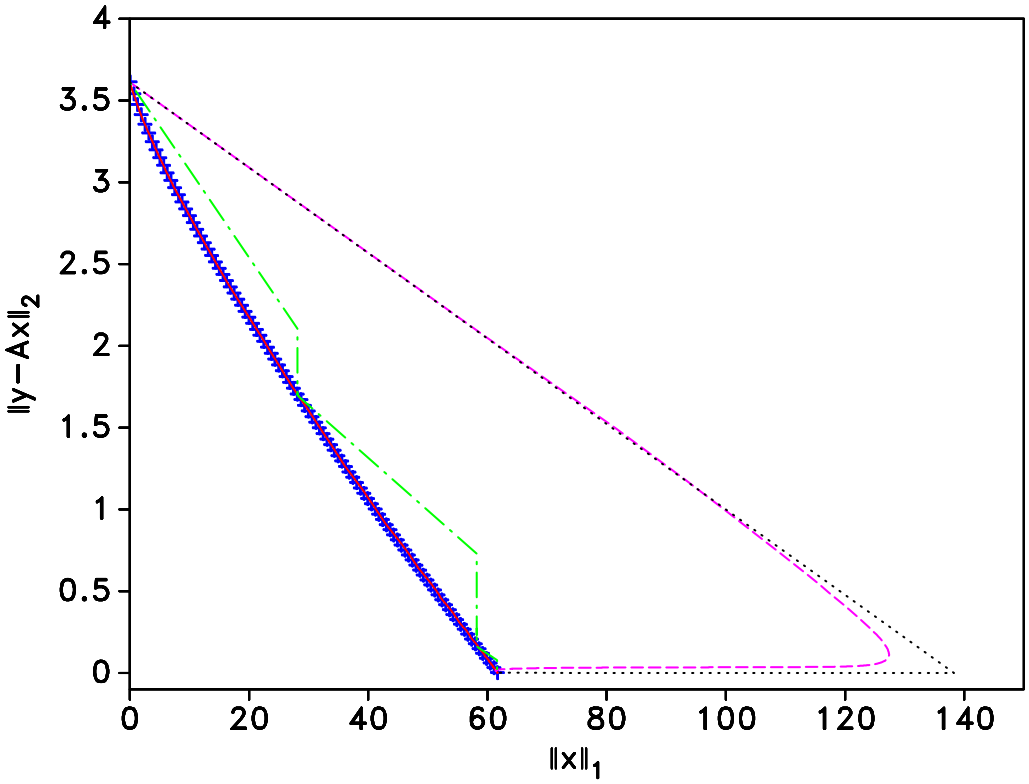


Figure 2: Pareto curve and solution paths (large enough number of iterations) of four solvers for a BP_0 problem. The symbols + represent a sampling of the Pareto curve. The solid (—) line, obscured by the Pareto curve, is the solution path of ISTc, the chain (– · –) line the path of SPGLℓ_1, the dashed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

Accepted 2008 March 7. Received 2008 March 3. In original form 2007 December 18.

The two solvers SPGℓ_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 towards 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 $QP_{\lambda_{i-1}}$, which is by definition on the Pareto curve. This explains why ISTc so closely follows the curve. SPGℓ_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 SPGℓ_1 solution path. IST solves QP_{0^+}. Because there is hardly any regularization, IST first works towards minimizing the data misfit. When the data misfit is sufficiently small, the effect of the one-norm penalization starts, yielding a change of direction towards 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 the 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 the various solvers where 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 effect of the one-norm regularization kicks in; 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.

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 4(a), corresponds to a shot record with 35% of the traces missing. The interpolated result, shown in Figure 4(b), is obtained by solving BP_0 using SPGℓ_1. This problem has more than half a million unknowns and forty-two thousand data points.

The 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

Accepted 2008 March 7. Received 2008 March 3. In original form 2007 December 18.

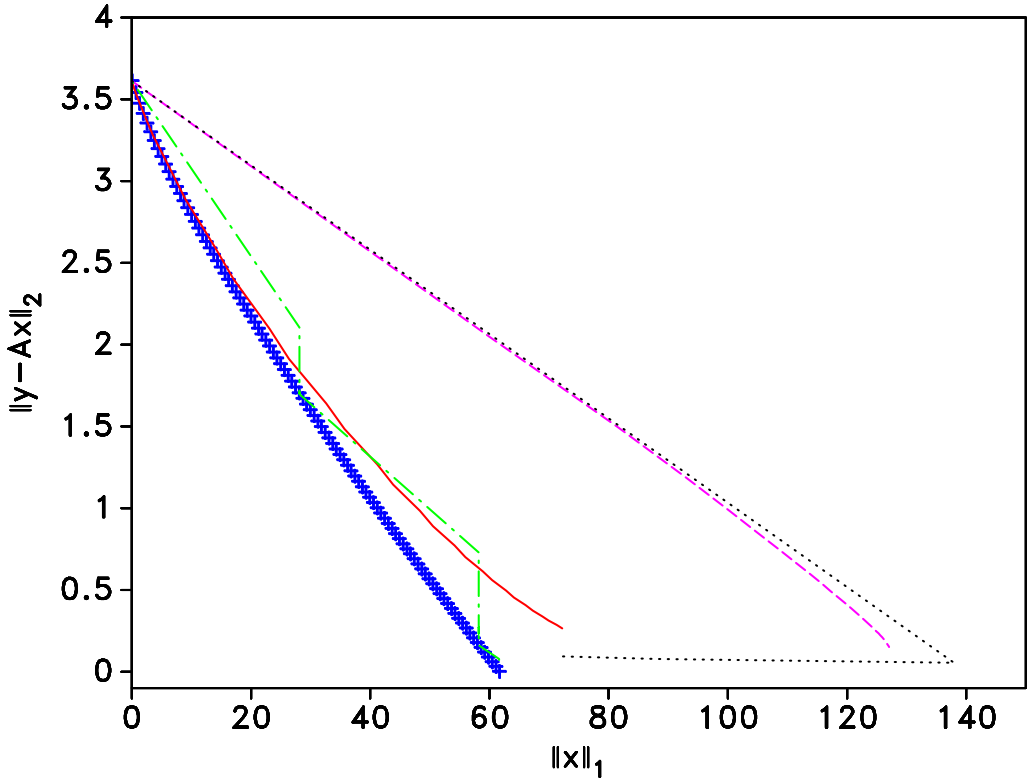
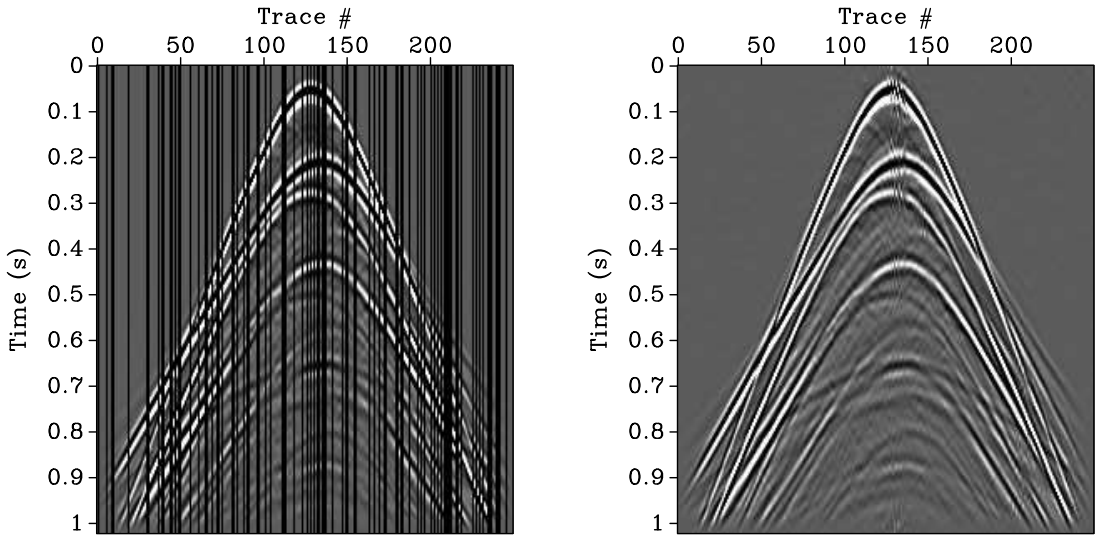


Figure 3: Pareto curve and optimization paths (same, limited number of iterations) of four solvers for a BP_0 problem (see Figure 2 for legend).

Accepted 2008 March 7. Received 2008 March 3. In original form 2007 December 18.

be convex—is smooth and well behaved, and empirically supports our earlier claim. However problems of practical interest are often significantly larger, and it may 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 the samples that were not included in the interpolation. The figure also plots the iterates taken by SPGℓ_1 in order to obtain the reconstruction shown in Figure 4(b). The plot shows that the iterates remain to the Pareto curve and that they convergence towards the BP_0 solution.



(a) (b)

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

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 in order to make an informed decision on how to best truncate the solution process. The Pareto curve serves as the optimal reference, which makes an unbiased comparison between different one-norm solvers possible.

Of course, in practice it is prohibitively expensive to compute the entire Pareto curve exactly. We observe, however, that the Pareto curves for many of the one-norm regularized problems are regular, as confirmed by the theoretical Result 1. This

Accepted 2008 March 7. Received 2008 March 3. In original form 2007 December 18.

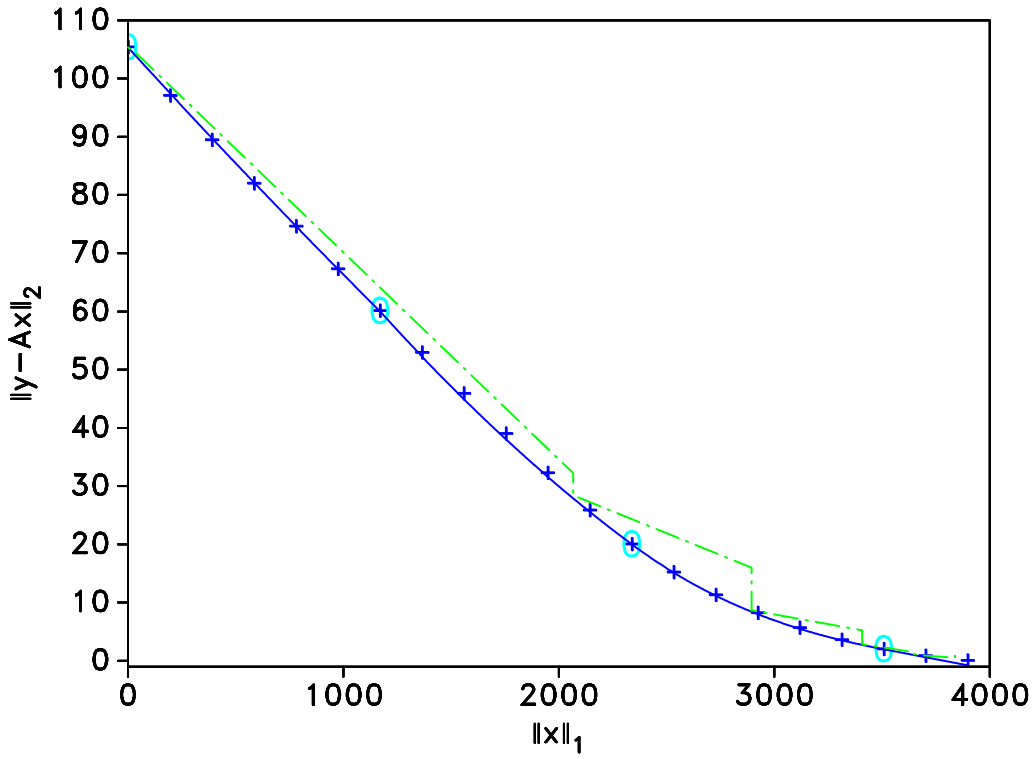


Figure 5: Pareto curve and SPGℓ_1 solution path for a CRSI problem. The symbols + represent a fine, accurate sampling of the Pareto curve. The solid (—) line is an approximation to the Pareto curve using the few, circled points, the chain (– · –) line the solution path of SPGℓ_1.

Accepted 2008 March 7. Received 2008 March 3. In original form 2007 December 18.

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, the insights from the Pareto curve can be leveraged to large-scale one-norm regularized problems, as we illustrate on a geophysical example. This prospect is particularly exciting given the current resurgence of this type of regularization in many different areas of research.

ACKNOWLEDGMENTS

The authors are grateful to Sergey Fomel and Tamas Nemeth for their valuable input, and to Eric Verschuur for the synthetic data. The authors also thank the anonymous reviewers and associate editor for their comments that certainly helped improve this letter. This publication was prepared using Madagascar (`rsf.sf.net`), a package for reproducible computational experiments, SPGℓ_1 (`cs.ubc.ca/labs/scl/spgl1`), and Sparco (`cs.ubc.ca/labs/scl/sparco`), a suite of linear operators and problems for testing algorithms for sparse signal reconstruction. This research was in part financially supported by NSERC Discovery Grant 22R81254 of F.J.H. and by CRD Grant DNOISE 334810-05 of F.J.H. and M.P.F., and was carried out as part of the SINBAD project with support, secured through ITF, from the following organizations: BG Group, BP, Chevron, ExxonMobil, and Shell.

REFERENCES

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. (`http://www.optimization-online.org/DB_HTML/2008/01/1889.html`).

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**, no. 2, 489–509.

Chen, S. S., D. L. Donoho, and M. A. Saunders, 1998, Atomic decomposition by basis pursuit: SIAM Journal on Scientific Computing, **20**, no. 1, 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, **LVII**, 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. (`http://adsabs.harvard.edu/abs/2007arXiv0706.4297D`).

Donoho, D. L., 2006, Compressed sensing: IEEE Transactions on Information Theory, **52**, no. 4, 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. (`http://stat.stanford.edu/~idrori/StOMP.pdf`).

Accepted 2008 March 7. Received 2008 March 3. In original form 2007 December 18.

Figueiredo, M. and R. Nowak, 2003, An EM algorithm for wavelet-based image restoration: IEEE Transactions on Image Processing, **12**, no. 8, 906–916.

Gersztenkorn, A., J. B. Bednar, and L. Lines, 1986, Robust iterative inversion for the one-dimensional acoustic wave equation: Geophysics, **51**, no. 2, 357–369.

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

Herrmann, F. J. and G. Hennenfent, 2008, Non-parametric seismic data recovery with curvelet frames: Geophysical Journal International, **173**, 233–248.

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

Oldenburg, D., T. Scheuer, and S. Levy, 1983, Recovery of the acoustic impedance from reflection seismograms: Geophysics, **48**, no. 10, 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**, no. 1, 84–97.

Rauhut, H., 2007, Random sampling of sparse trigonometric polynomials: Applied and Computational Harmonic Analysis, **22**, no. 1, 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 Royal Statistics, **58**, no. 1, 267–288.

Accepted 2008 March 7. Received 2008 March 3. In original form 2007 December 18.