The lost honour of ℓ_2-based regularization

Kees van den Doel, Uri Ascher and Eldad Haber *

November 20, 2012

Abstract

In the past two decades, regularization methods based on the ℓ_1 norm, including sparse wavelet representations and total variation, have become immensely popular. So much so, that we were led to consider the question whether ℓ_1-based techniques ought to altogether replace the simpler, faster and better known ℓ_2-based alternatives as the default approach to regularization techniques.

The occasionally tremendous advances of ℓ_1-based techniques are not in doubt. However, such techniques also have their limitations. This article explores advantages and disadvantages compared to ℓ_2-based techniques using several practical case studies. Taking into account the considerable added hardship in calculating solutions of the resulting computational problems, ℓ_1-based techniques must offer substantial advantages to be worthwhile. In this light our results suggest that in many applications, though not all, ℓ_2-based recovery may still be preferred.

1 Introduction

Ill-posed problems typically require some regularization in order to compute a credible approximate solution in a stable, well-defined manner. In this article we consider such problems where the objective is to recover a function $u(\mathbf{x})$, with $\mathbf{x} \in \Omega \subset \mathbb{R}^d$ (typically $d = 2$ or $d = 3$), from observed and discrete data b. Given is a forward operator, $F(u)$, which predicts data for any suitable function u, and the challenge is to find u such that the predicted data match the observed data to within a reasonable tolerance.

It is convenient for our discussion at this point to consider a linear forward operator, with u discretized on some mesh in Ω and reshaped as a vector of unknowns $\mathbf{u}$,

*Department of Computer Science, University of British Columbia, Vancouver, Canada kvdoel/ascher/haber@cs.ubc.ca . This work was supported in part by NSERC Discovery Grant 84306.

and with the observed and predicted data likewise written as $\mathbf{b}$ and $F = J\mathbf{u}$, respectively. Here J is an $m \times n$ sensitivity matrix, $m \leq n$, which often has a nontrivial null space. Then we write down the Tikhonov-type regularized problem [60, 25, 61]

$$\min_{\mathbf{u}} \frac{1}{2}\|J\mathbf{u} - \mathbf{b}\|_2^2 + \beta R(\mathbf{u}), \tag{1}$$

where $\|\cdot\|_p$ denotes the usual vector ℓ_p norm, $\beta > 0$ is a parameter, and R is a regularization operator. We focus on the following possibilities for R:

1. Consider

$$R(\mathbf{u}) = \frac{1}{p}\|W\mathbf{u}\|_p^p, \tag{2}$$

for the choices $p = 1$ (referred to as L1) or $p = 2$ (referred to as L2). Here W is an $n \times n$ weight matrix, e.g., some wavelet or curvelet transform, or just the identity [33, 24, 10]. For notational purposes we stipulate that W is not a discretized gradient operator.[1]

2. Recalling that $\mathbf{u}$ represents a discretization of a function $u(\mathbf{x})$ on Ω, choose $R(\mathbf{u})$ to be an appropriate discretization of

$$\mathcal{R}(u) = \frac{1}{p}\int_\Omega |\nabla u|^p, \tag{3}$$

again considering the cases $p = 2$ or $p = 1$. The case $p = 2$ leads to a discretization of the Laplacian operator on Ω when considering necessary conditions for the minimization (1): denote this by L2G. The case $p = 1$ leads to total variation [53, 51]: denote this by L1G.[2]

For many years the almost automatic choices of regularization in (2) and (3) have been based on the ℓ_2-norm, i.e., $p = 2$. This yields a straightforward linear least squares problem that can be effectively solved even when the problem is very large (see, e.g., [55, 32]). Large computational problems are manageable even if F is nonlinear in u and R is more complex but still ℓ_2-based (see, e.g., [29, 16, 17]). Furthermore, the ℓ_2-based regularization enjoys a favorable statistical interpretation for models with a prior that is normally distributed [58, 8, 61, 41].

[1] Of course, wavelet function bases do approximate derivatives, too. For instance, our distinction as such is particularly blurred by tight frame wavelets [7]. And yet, the distinction of L1 from L1G should be intuitively clear. Note also that one can always transform L1 and L2 by a change of variables into a form where W becomes the identity. However, we retain our notational redundancy for convenience.

[2] Note that the gradient magnitude $|\nabla u|$ is the ℓ_2 norm of ∇u. Thus, the L1G expression is one of a discrete ℓ_1 norm only if $d = 1$. Also, a further regularization is required when using L1G upon considering necessary conditions for (1); see, e.g., [1].

In the past two decades, however, regularization methods based on the ℓ_1-norm (i.e., $p = 1$ in (2) and (3)) have become immensely popular; see, e.g., the books [51, 46, 24]. In fact, we have been led to consider the idea that ℓ_1-based techniques ought to altogether replace the simpler, faster and better known ℓ_2-based alternatives. There are two essential motivations for this exciting trend.

- It is natural to choose for the regularization term a penalty function as in (3) thus expressing the a priori information that $u(\mathbf{x})$ ought to be smooth. But if $u(\mathbf{x})$ has jump discontinuities then using L2G essentially smears out such discontinuities because the Dirac δ-function is not square integrable. On the other hand the δ-function is integrable, and thus using L1G better accommodates jump discontinuities.

- Whether the term R is aimed at penalizing the magnitude of the gradient or the solution itself, the ℓ_1-based regularization tends to produce sparse approximations. In the L1G context this is expressed in the observation that the reconstruction tends toward being piecewise constant, so the gradient is mostly zero and thus sparse. In the L1 wavelet (or DCT) approximation context, where $W\mathbf{u}$ in (2) corresponds to coefficients of different wavelet (or cosine) basis functions, a compressed approximation involving only a few basis functions often results (unlike the case when using $p = 2$).

The rather fundamental importance of the above two reasons for using $p = 1$ is not in doubt. Among many other researchers we have ourselves contributed to this volume of work [1, 30, 36]. We have found that for well-conditioned problems with sufficient high quality data,[3] ℓ_1-based regularization can, in many cases, "deliver on its promise". However, for problems with poor data, or ill-conditioned problems typically resulting from discretizations of highly ill-posed problems, we have found that this is often not the case. To demonstrate and motivate the ensuing discussion, let us consider the following example.

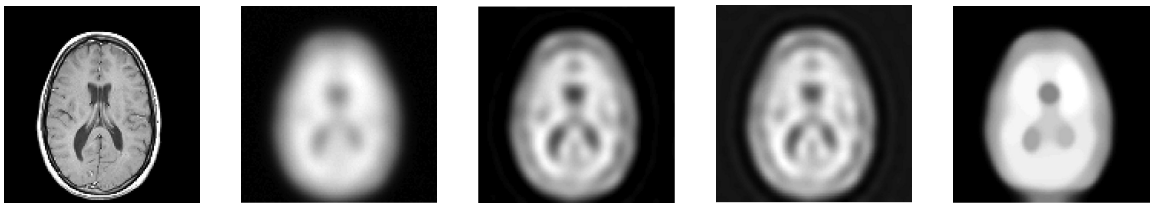
Example 1 Image Deblurring

Let J be a discretization of a known image blurring operator and $\mathbf{u}$ an image reshaped into a vector. Our goal is to recover the clean image given noisy blurred data. For the following numerical experiments we have used three codes: (i) RestoreTools [33], which employs an L2-type recovery strategy (viz. $p = 2$ and $W = I$ in (2)); (ii) the GPSR package [26], which employs a wavelet L1 recovery algorithm; and (iii) a straightforward total variation (L1G) code. The above two packages, in our opinion, are both excellent representations of good software for the problems they aim to solve. However, the L2 code requires, comparatively speaking, only a small

[3] We further explain in Section 3 what we mean by the intuitive terms "high quality" vs "poor" data.

fraction of computational time to terminate successfully, hence it is to be preferred unless the L1 reconstructions are demonstrably better.

The "true image", or ground truth, is a 128×128 *MRI image from* MATLAB*'s collection. The blurring kernel is* $e^{-\|\mathbf{x}\|_2^2/2\sigma}$ *with* $\sigma = 0.01$ *and the blurred data is further corrupted by* 1% *white noise. In all three methods, the data is fit to an accuracy of* 1% *by tuning the regularization parameter* β *(see, e.g., [61]). The results are presented in Figure 1.*



(a) True (b) Data (c) L2 (d) L1 (e) L1G

Figure 1: The ground truth image (a) is blurred and corrupted by noise to create the data (b). Recovered solutions obtained for this data by RestoreTools (L2), GPSR (L1) and total variation (L1G) are displayed in (c-e), respectively.

It is apparent that, at least for this problem instance, the ℓ_1*-based reconstructions do not yield more pleasing results than the simple* ℓ_2*-based one. The L1G image is typically blocky, and in the present context may be considered the worst of the three: indeed, sparsity of the surface gradient is not a good regularization objective here. The first two recoveries are more comparable in terms of quality. In fact, it may be argued that the L2 result is altogether better than the* ℓ_1*-based ones.*

Image deblurring is a favorite application in the literature for discussing and comparing both L1 and L1G techniques. Indeed, in many such examples ℓ_1-based regularization is to be preferred (see, e.g., [24, 11, 36]). However, Example 1 is by no means esoteric. Furthermore, similar comparative observations arise when working on certain nonlinear ill-posed problems such as electrical impedance tomography (EIT) and direct current (DC) resistivity [1]; we return to this in Section 4.

The goals of this paper are therefore to explore, bearing in mind the occasionally impressive advances of ℓ_1-based regularization techniques, also some of their limitations. Taking into account the often considerable added hardship in calculating solutions of the resulting computational problems, ℓ_1-based techniques must offer substantial advantages to be worthwhile. In this light our results suggest that in many applications, ℓ_2-based recovery may be preferred. To this end we provide the following cautionary notes:

1. Only the left term in the objective function of (1) is really mandated by the stated data fitting problem. The choice of regularization is discretionary: different choices may generally yield different solutions that as such must all be

considered acceptable. The further specification of regularization reflects a prior which depends on additional knowledge that may or may not be truly available.

2. It is not true that one must always seek a sparse approximate solution, especially if an appropriate basis to span the solution is not known.

3. Codes such as those reported in [26, 3, 42, 4], which perform well when applied in the context of using wavelets for denoising or deblurring, may occasionally perform relatively poorly when applied in a wider context.

4. In our experience, if the data is not of sufficiently high quality, in the sense that there is too much noise, then ℓ_1-based methods may occasionally perform worse than the corresponding ℓ_2-based methods.

5. If the data is not of sufficiently high quality, in the sense that it is too sparse or rare, then ℓ_1-based methods may occasionally perform worse than the corresponding ℓ_2-based methods.

6. If the computational problem is highly ill-conditioned then ℓ_1-based methods may occasionally perform worse than the corresponding ℓ_2-based methods.

In this paper we explore examples, or case studies, which demonstrate the claims above and explain when ℓ_2-based methods merit prime consideration. Some analysis is also provided. We group our discussion into two classes: problems with poor data, considered in Section 3, and highly ill-conditioned problems, considered in Section 4. The latter section is far longer and more involved than the others, and Theorem 1, as well as the analysis in Section 4.1, are new. Before these, Section 2 provides a quick review of ℓ_1-based regularization. We review the theory and the requisite assumptions necessary for ℓ_1-based recovery to perform well.

Finally, we summarize the paper in Section 5.

2 ℓ_1-based regularization

Several books, e.g., [56, 51, 11, 46, 24], contain descriptions of ℓ_1-based regularization methods in the context mentioned earlier, and it is not our intention to reproduce them here. We only touch upon a few items. For early efforts in geophysics and data assimilation, see [14, 59]. For advanced uses of such methods in machine learning, see, e.g., [49, 47].

In the context of a discrete cosine or a wavelet-type transform, the problem (1) may be viewed as a noisy version of the problem

$$\min_{\mathbf{u}} \quad R(\mathbf{u}), \tag{4a}$$

$$\text{s.t.} \quad J\mathbf{u} = \mathbf{b}, \tag{4b}$$

where J has a full row rank $m < n$. Note that this can be a well-conditioned problem for both choices of p in (2). For L1 (i.e., $p = 1$ in (2)) problem (4) can be cast as a linear programming problem, and linear programming theory already guarantees that there is an optimal basic feasible solution which is m-sparse (i.e., with only at most m non-zero components) [62, 50]. In contrast, when using L2, all components of the optimal $\mathbf{u}$ are typically non-zero.

This has been well-known since at least the 1960s. Moreover, though, since the above transforms utilize elaborate basis functions it is reasonable to expect that much fewer than m basis functions may suffice, corresponding to a much sparser solution. The discovery [13, 22, 20] that using L1 often yields such a sparse solution, effectively solving a very hard combinatorial problem, is much newer and constitutes a major breakthrough.

However, it is not always the case that the solution of the constrained optimization problem using the ℓ_1 norm yields a sparse solution. Furthermore, for (1) in general it does not automatically follow that if such a sparse solution exists it is an appropriate estimate of the true solution, see [19] and Section 4.1.

Much effort has been devoted to the question, under what conditions the ℓ_1 solution of (4) produces the sparsest possible solution of (4b), referred to as the ℓ_0 solution. Of course, a more practical goal would probably be to seek a "sufficiently sparse" solution, but the quest for optimum in this regard sheds light on what is required more generally. The restricted isometry property (RIP) [9] and the null space property of [21, 15] both provide sufficient conditions, whereas the γ-condition of [40] is both necessary and sufficient for obtaining the sparsest solution by L1.

These conditions are of great value for understanding the design of compressed sensing methods. Unfortunately, though, for realistic instances of the matrix J they are generally intractable (NP-hard) to verify numerically. Moreover, in Section 4.1 we show that such conditions are violated for a specific case of the inverse potential problem, when attempting to recover a pair of point charges by ℓ_1-based methods.

The vector norm function $\|\cdot\|_p$ is well-known to be convex only when $p \geq 1$. Thus, ℓ_1 is marginally convex. Even more sparsity-inducing is the use of a nonconvex norm with $0 < p < 1$ [44, 54, 12]. But there is a price to pay for lack of convexity, in terms of both poorer theory and the necessity of convergent algorithms which typically apply a continuation (homotopy) procedure starting from a convex ℓ_p-norm.

Several famous codes cited earlier for solving (4) use methods that are based on gradient projection with acceleration (see for instance the extended Chapter 6 of [5] and references therein). The advantage of these methods is that they extend directly to problems with non-smooth constraints and require the objective function gradient to be only Lipschitz continuous. However, bear in mind that for solving simple unconstrained convex quadratic problems such methods boil down to accelerated gradient descent without preconditioning, generally thought to be unforgivably slow. These methods seem to work well for compressed sensing problems because the

corresponding problems (4) are well conditioned in an appropriate sense. Unfortunately, other applications involving, for instance, PDE-constrained optimization (as in Section 4), are highly ill-conditioned and therefore, similar numerical optimization methods should not be expected to be robust and efficient in the latter context.

Total variation (L1G) has been discovered and peaked earlier than sparse wavelet basis reconstruction and compressed sensing. The books [61, 51, 11] and many papers develop both theory and algorithms using this approach. In practice, some regularization such as a Huber switching function [56] is often used, and this really gives a mix of ℓ_1 with ℓ_2 elements while still retaining the L1G spirit [1]. See also [6] for another approach to round excessive L1G sharpness. Moreover, one popular iterative scheme to carry out the resulting algorithm is lagged diffusivity, which is a special case of iteratively reweighted least squares (IRLS) [61, 1].

Unlike the case for wavelet-type solutions, where a sparse representation is sought for the same high-quality surface or image approximation, here the regularization is applied directly to the surface variables to be recovered. Along with the advantage in directly penalizing piecewise smoothness, the tendency of the L1G regularization to give sparse gradients, translating into a "blocky image", is not always what one necessarily wants (see, e.g., Figure 1(e)). L1G penalizes large jumps in the solution more than small jumps, and this may introduce distortion in the reconstructed surface. Various nonconvex alternatives to L1G are listed in [56], for instance, and these occasionally yield sharper results for some applications. However, the nonconvex nature of these regularizations again leads to both theoretical and practical additional difficulties.

Our focus in this article is on exploring situations where use of the L1 or L1G regularization ($p = 1$ in (2), (3)) may reasonably be compared to use of L2 or L2G ($p = 2$ in (2), (3)). Therefore, employing any of the even sharper non-convex options mentioned above is not under further consideration.

The above synopsis has been restricted to linear problems. There is very little ℓ_1 theory for nonlinear problems. Moreover, it is easy to see that some of the basic sparsity arguments fail for this case. Consider the problem

$$\begin{aligned} \min_{\mathbf{u}} \quad & \|\mathbf{u}\|_1 \\ \text{s.t.} \quad & F(\mathbf{u}) = \mathbf{b}, \end{aligned}$$

where the forward mapping function $F : \mathbb{R}^n \to \mathbb{R}^m$ is smooth and has significant curvature (see Figure 2). In such a case the problem need not even have m-sparse solutions; indeed the optimal solution may have n non-zero entries. Thus, the justification of using L1 for nonlinear problems is far from obvious. On the other hand, L1G remains of interest, because of its sharpening property. In Section 4.3 we explore the use of L1G for a particular popular nonlinear case study.

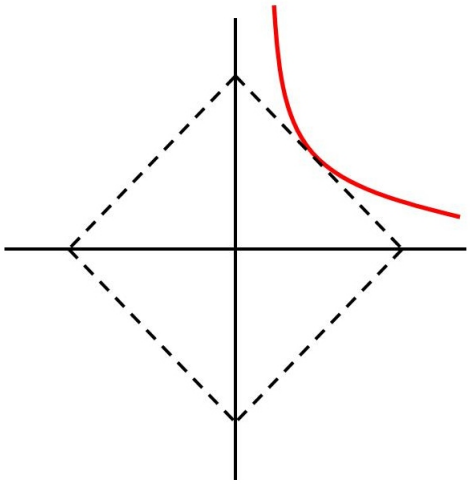


Figure 2: When the constraint (solid red) is nonlinear, it does not need to intersect the level set of $\|\mathbf{u}\|_1$ (dashed black) at a vertex, so the solution is not necessarily sparse.

3 Poor data

The perceived quality of a given data set depends on several factors, not simply on some idealized noise level. One of these is the inverse problem operator. For instance, in Example 1 the deblurring operation, which is essentially to improve contrast and sharpness of the image, counters an image smoothing operation which aims to remove noise. Thus, a noise level in the data which may otherwise be considered benign (say, in a pure denoising application) can be an important obstruction here.

In the context of data fitting, it has been known for decades that ℓ_1 data fitting is more robust than ℓ_2 against outliers in the data. See for instance [50] and also [45] for a recent use in the context of 3D graphics. However, such a comparative statement does not necessarily hold true for other types of noise such as white noise.

In general, bearing in mind the additional complications in carrying out ℓ_1-based regularization, the data must be of sufficiently high quality to allow its favourable properties (when relevant) to be expressed. A common situation yielding lack of sufficiently good data is when the data is relatively rare, being given only at relatively few locations in Ω. Let us next discuss a simple such example where the data locations are rare (or sparse) in the domain of definition.

Example 2 Rare data reconstruction of piecewise smooth functions

Consider the recovery of a (real) signal $u^*(t)$ on $[0,1]$ from m noisy samples $u_i \approx u^*(t_i)$, and assume we know that u^* is piecewise smooth but may have jump discontinuities. We discretize the interval $[0,1]$ with a uniform grid of $n = 512$ points, and use in a given experiment a subset of $m \ll n$ samples taken at random points t_j

from this grid. The integral appearing in (3) is discretized using a piecewise linear function $u(t)$ on all n grid points. Thus, the recovery problem is formulated as in (1), with J being the $m \times n$ matrix consisting of m columns forming an identity matrix interspersed with $n - m$ zero columns. In the limit case of no noise the formulation (4) yields interpolation through the data points (t_i, u_i) of the sample.

We compare L2G and L1G regularizations. It is easy to verify that in the L2G case these data points are connected by straight lines, whereas with L1G (total variation) regularization the behavior is indeterminate, only restricting u to be monotone.

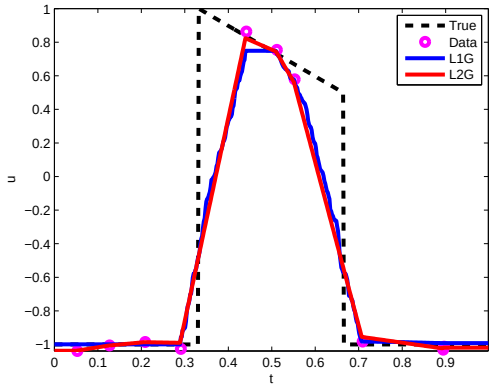
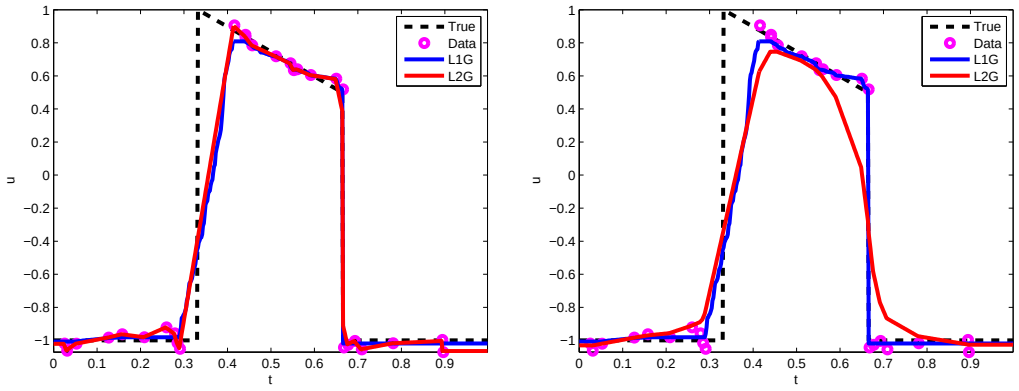


Figure 3: Reconstructions of a piecewise smooth function from a few noisy samples: using L2G and L1G for m=9 data pairs, with β_{L2G}=.04, β_{L1G}=.08.



(a) β_{L2G}=.002, β_{L1G}=.08 (b) β_{L2G}=.02, β_{L1G}=.08

Figure 4: Reconstructions of a piecewise smooth function from a few noisy samples: using L2G and L1G for m=28 data pairs.

Figures 3 and 4 depict reconstruction results for $m = 9$ and $m = 28$ samples. The ground truth signal $u^*(t)$ contains two jumps, and we added 5% Gaussian noise to the selected values to form the corresponding data sets.

Figure 3 shows the result for 9 *samples, with the regularization parameters tuned by the discrepancy principle to obtain a data misfit of* $5 \pm 0.1\%$. *There is little difference between the L1G and L2G reconstructions.*

The reconstruction in Figure 4(a) using 28 *samples starts to show the advantages of L1G. Because the data contains two samples across the right discontinuity, the regularization parameter* β_{L2G} *now had to be decreased to* $\beta_{L2G} = .002$ *in order to obtain the desired misfit of roughly* 5%. *As a result the L2G reconstruction exhibits considerable oscillation in the flat sections, although note that the second jump is reproduced as well as by the L1G method. In Figure 4(b) we increased* β_{L2G} *until the flat sections became reasonably smooth according to the "eyeball norm". Observe that the oscillation has disappeared, but the second jump is now completely blurred as well.*

Example 2 illustrates that L1G regularization performs well when there is enough quality data to require the reconstructed model to have discontinuities. But when the data is "too sparse", L2G regularization performs as well as L1G even in the presence of discontinuities in the underlying ground truth function. This lesson seems perhaps obvious in hindsight. However, it extends to more complex situations where the insight is no longer so obvious. For instance, the problems considered in Section 4 have data specified only at the boundary of a given physical domain Ω, which is a lower-dimensional manifold; several examples can be found in the literature where some L1G variant is applied to such problems. For another instance, consider a point cloud in 3D, obtained as a set of somewhat noisy and not very dense 3D laser scan measurements of a body with edges, such as a desk corner. In order to obtain a good surface reconstruction we need at each point the normal to the surface that the (cleaned) point cloud represents [38]. Since the curvature across an edge is infinite the data can be effectively very sparse there, and indeed a global ℓ_1-reconstruction approach [2] might not work well then. See Figure 10 in [39] for such an example. Poor data are often encountered in ocean and atmospheric data assimilation, as well as in other time-dependent geophysical applications [27, 23].

4 Large, highly ill-conditioned problems

In this section we consider applying ℓ_1-based techniques to large, highly ill-conditioned problems such as typically arise in applications involving PDE-constrained optimization. Using first an example we show in §4.1 that L1 techniques may not only be expensive to carry out but also have difficulty in producing solutions which are as sparse as a given ground truth. In §4.2 we then supply some analytical evidence supporting this observation. Finally, in §4.3 we show by another example that while L1G is not nearly as severely afflicted as L1, its advantage over L2G in recovering surface discontinuities requires favourable conditions to shine through.

4.1 Inverse potential problem

In the inverse potential problem one seeks to recover an electrical source distribution in a given domain Ω from measurements of the potential on the domain's boundary. This problem arises in EEG source modeling [48] and in electromyography [18, 19]. In [19] the sought source is a combination of discrete tripoles corresponding to muscle fibers, and as such invites a sparse reconstruction. However, in 3D the computational problem using L1 indeed became much too large and difficult to work with, and our eventual success in solving the research problem stated in [19] followed a further realization that, given the specific goals of those computations, the sparse view was not anyway the most effective. This has still left open the question of what an L1 reconstruction can do for such a problem (regardless of cost), a question that we now proceed to explore in a more manageable 2D context with Ω being the unit square.

The forward model

$$-\Delta v = u(\mathbf{x}), \quad \mathbf{x} \in \Omega, \tag{5}$$

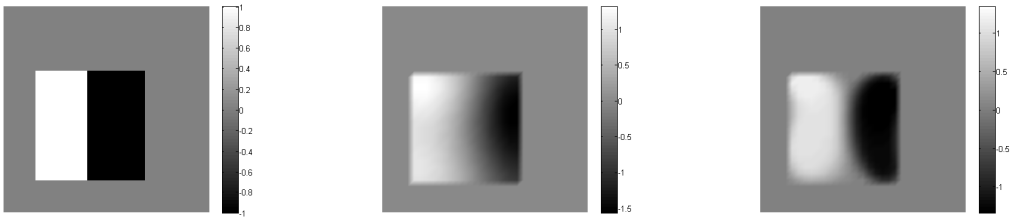
with Neumann boundary conditions on v predicts the potential v for given electrical source u. The total charge must be zero due to these boundary conditions. Note that $v(\mathbf{x})$ is only determined up to an overall additive constant, reflecting the physical principle that only a potential difference is physically meaningful.

The inverse problem of finding u from values of v on the boundary does not have a unique solution, even under idealized conditions [35]. The best one can do is construct an "equivalent source" u that explains the data. Such a reconstruction gives incomplete but still useful information about the actual source. Hence the role of regularization is to provide additional information leading to a distribution u that conforms to prior expectations, a rather fundamental difference from sparse signal reconstruction. Denoting the discretized Poisson operator of (5) by A and the data projection operator by Q (see [19] for details), we obtain a problem in the form (1) with

$$J = QA^{-1}. \tag{6}$$

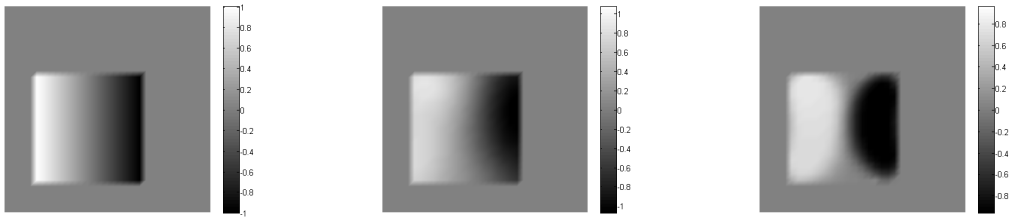
Example 3 Inverse Potential Problem

In this numerical experiment the support of the source u is restricted to the offset inner square (assumed known in the reconstruction) as depicted in Figures 5–7. The potential is measured on the boundary, taking the average boundary potential as the ground level, i.e., we subtract the average boundary potential from each datum. This is necessary as only potential differences are measurable. Figures 5–7 depict results for three different source distributions in the region. In each case, synthetic data is computed on a 64^2 grid, to which we add 1% Gaussian noise. The reconstruction is done with our various regularizations (3) and (2) on a 32^2 grid. The regularization constant β is tuned to obtain a resulting misfit of $1 \pm 0.1\%$ (see, e.g., [61]).



(a) True (b) L2G (c) L1G

Figure 5: Reconstructions of a piecewise constant charge distribution from boundary data.



(a) True (b) L2G (c) L1G

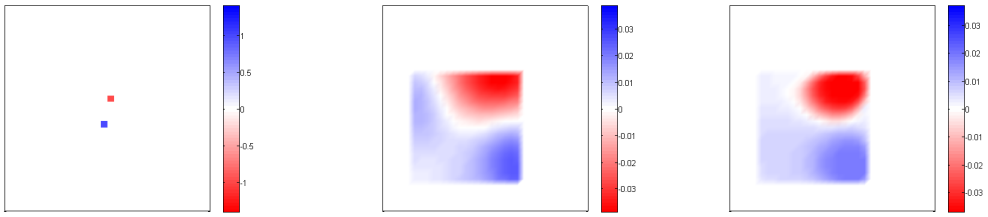
Figure 6: Reconstructions of a smoothed-step charge distribution from boundary data.

Figures 5 and 6 serve as an appetizer. We consider, respectively, piecewise constant and smoothed-step dipole distributions. Observe that the L1G reconstruction results in a well-defined interface between the positively and negatively charged regions, whereas the L2G reconstruction is smooth, irrespective of the true model. As such, the use of L1G is especially recommended if we know a priori that u is piecewise smooth. However, it is not possible to determine from the reconstructions whether u has a jump or not: notice the similarity between Figures 5(b) and 6(b) and that between Figures 5(c) and 6(c).

Next, we explore the main theme of this section by considering a point charge pair. The true model (ground truth) depicted in Figure 7(a) is now very sparse.

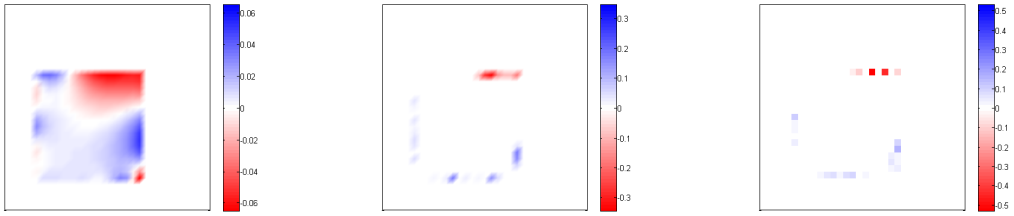
The results shown in Figure 7 are in a similar vein as those in Figures 5 and 6. The dipole structure is apparent from the L2G and L1G reconstructions, but not much more is. The L1G reconstruction hints at a dipole pair but may mislead one to infer an incorrect orientation.

For this last source distribution a sparse reconstruction seems natural, and one such, obtained using an L1 regularization, is depicted in Figure 8(b). The L2 reconstruction is depicted in Figure 8(a) for comparison. We see that the L1 reconstruction is somewhat sparse, but all the reconstructed sources are on the boundary of the sup-



(a) True (b) L2G (c) L1G

Figure 7: Reconstructions of a point charge pair from boundary data using gradient regularization.



(a) L2 (b) L1 (c) Weighted L1

Figure 8: Reconstructions of a point charge pair from boundary data using regularizations (2).

port of $u(\mathbf{x})$*, and the* $L1$ *solution is not as sparse as the true model.*

The reason for the observed source distribution is that sources near the detector affect the data more and are therefore favoured [31]. This effect can be reduced by a location dependent re-weighting of the regularization function as suggested in [28, 43], which amounts to normalizing the columns of J *to unit 2-norm. Letting*

$$a_j = \Big(\sum_{i=1}^{m}(J_{ij})^2\Big)^{1/2} \quad \hat{J}_{ij} = J_{ij}/a_j \quad \hat{u}_i = a_i u_i,$$

we can write $J\mathbf{u} = \mathbf{b}$ *as* $\hat{J}\hat{\mathbf{u}} = \mathbf{b}$ *and apply the* $L1$ *regularization to* $\hat{\mathbf{u}}$*. (Note though that computing* a_i *for large scale problems may not be practical.) The resulting reconstruction is depicted in Figure 8(c). The sparsity has improved a little, but we are still far from the* ℓ_0 *solution.*

For this example, since $\hat{J}$ has normalized columns, the famous RIP condition defined and analyzed in [9] applies. This condition requires that there be a $\delta \leq \sqrt{2}-1$ such that for all 4-sparse $\mathbf{u}$ we have

$$(1-\delta)\|\hat{\mathbf{u}}\|_2^2 \leq \|J\mathbf{u}\|_2^2 \leq (1+\delta)\|\hat{\mathbf{u}}\|_2^2. \tag{7}$$

However, here it can be shown to be violated on physical grounds. Let $\mathbf{u}$ be a 4-sparse source, i.e., non-zero only for indices i in a set $\mathcal{T}$ with $|\mathcal{T}| = 4$, and further let it have values ± 1, so

$$\|\hat{\mathbf{u}}\|_2^2 = \sum_{i \in T} a_i^2 \geq 4 \min_k (a_k^2) > 0.$$

(The value a_i is just the 2-norm of the boundary potential when a unit source is placed at location i.) Note that $\|J\mathbf{u}\|_2^2$ is the ℓ_2 norm of the boundary potential. By placing the positive and negative charges very close together, so that they almost cancel each other, we can make the boundary potential and thereby $\|J\mathbf{u}\|_2^2$ arbitrary small, and thus δ becomes arbitrarily close to 1. Hence the RIP condition is violated. Note that this does not prove that the sparsest solution cannot be obtained, as the RIP is a sufficient but not necessary condition.

The necessary and sufficient γ-condition of [40] for obtaining the ℓ_0 solution from the ℓ_1 solution relies on properties of the solution $\mathbf{y}$ to the equation

$$(J^T\mathbf{y})_i = z_i, \tag{8}$$

for selected indices i such that $z_i \neq 0$. In our case, to determine if it is possible to recover a 2-sparse source, the n-vector $\mathbf{z}$ should be 2-sparse with entries ± 1, so (8) has just two equations. Further, $J^T\mathbf{y} = A^{-1}Q^T\mathbf{y}$, and we can interpret $\mathbf{y}$ as describing electrical sources on the boundary only, such that the generated potential equals 1 at point $\mathbf{p}_1$ and -1 at point $\mathbf{p}_2$. These correspond to the location of the point charges described by $\mathbf{z}$. The γ-condition then implies that we can find a $\mathbf{y}$ such that the potential $J^T\mathbf{y}$ is between -1 and 1 everywhere else. Unfortunately, however, on physical grounds we can see that this is not possible. To see this note that if we place $\mathbf{p}_1$ and $\mathbf{p}_2$ very close together then a very large electrical field will exist between the points, which must be caused by very large boundary sources, which in turn will generate close to those sources an even larger (> 1) field. Analytically we observe that in the continuum limit, since $\mathbf{z}$ is a harmonic function it must take its extreme values on the boundary. Since it takes on values ± 1 inside, it must take on larger values on the boundary, and hence the γ-condition is violated.

4.2 The effect of ill-conditioning on L1 regularization

In this subsection we consider the regularized L1 problem

$$\min_{\mathbf{u}} \frac{1}{2}\|J\mathbf{u} - \mathbf{b}\|_2^2 + \beta\|W\mathbf{u}\|_1, \tag{9}$$

and show, for a special choice of W which in a sense favours sparsity, that in the highly ill-conditioned case and in the presence of noise the correct sparsity of a ground truth model can be recovered only if the singular values of J and the sparsity structure combine in a beneficial manner. This helps explain the negative results of Example 3.

Let the singular value decomposition (SVD) of the $m \times n$ matrix J be given by

$$J = U\Sigma V^\top,$$

where U and V are orthogonal matrices and $\Sigma = \text{diag}\,\{\sigma_1, \ldots, \sigma_m\}$ is $m \times n$ with the singular values ordered so that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_m$. Further, consider a true model $\mathbf{u}^*$ such that $\mathbf{z}^* = V^T\mathbf{u}^*$ satisfies

$$z_i^* = \begin{cases} 1 & i \in \mathcal{T} \\ 0 & i \notin T \end{cases}. \tag{10}$$

The emphasis in (10) is on the nature of $\mathcal{T}$, i.e., the sparsity: setting the nonzero values to 1 is just for convenience. For notational simplicity let us also assume, without loss of generality, that $U = I$, the identity. Then it also makes sense to consider the case where $z_i^* = 0,\ i > m$. Suppose further that the data $\mathbf{b}$ is contaminated by Gaussian noise $\boldsymbol{\epsilon}$ with mean 0 and covariance $\rho^2 I$. We have

$$\mathbf{b} = \Sigma\mathbf{z}^* + \boldsymbol{\epsilon}.$$

Thus, for $i \in \mathcal{T}$, $z_i^* = (b_i - \epsilon_i)/\sigma_i = 1$.

Turning to approximate solutions and setting $\mathbf{z} = V^T\mathbf{u}$, recall first the truncated SVD method, even though it has nothing to do with L1 methods. Thus, we set $\beta = 0$ in (9), obtaining the least squares problem

$$\min_{\mathbf{z}} \frac{1}{2}\|\Sigma\mathbf{z} - \mathbf{b}\|_2^2, \tag{11}$$

and then, since the noise ϵ_i is obviously magnified by σ_i^{-1}, we set

$$z_i = \begin{cases} b_i/\sigma_i & i \leq r \\ 0 & i > r \end{cases}, \tag{12}$$

where the effective rank $r,\ r \leq m$, is such that the error term depending on σ_r^{-1} has tolerable size. Using this regularization method, it is obvious that a necessary and sufficient condition for obtaining the same sparsity for $\mathbf{z}$ and $\mathbf{z}^*$ is that $\mathcal{T} = \{1, 2, \ldots, r\}$. Indeed, no (very) small singular value index can be tolerated in the set $\mathcal{T}$ of the given true model. In particular, we cannot stably obtain the sparse approximate solution for just any true model. This requirement becomes rather restrictive in the highly ill-conditioned case, where $r \ll m$.

Of course, the truncated SVD method not only does not have L1 magic, it also requires carrying out the SVD, something we wish to avoid for the large problems considered in this section. Let us now return to the Tikhonov-type method (9) with $\beta > 0$, and consider the special case of the L1 approach with $W = V^\top$. This special

case is in a sense the most favourable for the sparsity-inducing algorithm to work well. This is so because the subspace defined by $\Sigma\mathbf{z} = \mathbf{b}$ has the best possible orientation, with respect to the faces of the polyhedron $\|\mathbf{z}\|_1 = \text{constant}$, to cause intersection at a face corresponding to the correct sparsity. See for example Figure 1 in [10] for a sparsity spoiling orientation that cannot occur in our case. So, if we encounter difficulties caused by ill-conditioning in this special case then they will persist upon using a more general W.

Thus, we are considering the problem

$$\min_{\mathbf{z}} \ \frac{1}{2}\|\Sigma\mathbf{z} - \mathbf{b}\|_2^2 + \beta\|\mathbf{z}\|_1. \tag{13}$$

Because (13) is just a sum of decoupled terms, we can solve it explicitly for each component of $\mathbf{z}$. The solution has $z_i = 0$ where the gradient of the data fitting term is bounded by the gradient of the regularization term, which gives

$$\beta \geq |\sigma_i(\sigma_i z_i^* + \epsilon_i)|.$$

Otherwise

$$z_i = ((\sigma_i z_i^* + \epsilon_i) \pm \beta/\sigma_i)/\sigma_i,$$

where the sign in front of β is not needed for our purposes.

In order for $\mathbf{z}$ to have the same sparsity as $\mathbf{z}^*$ we therefore must have

$$\begin{aligned} \beta &\leq |\sigma_i(\sigma_i + \epsilon_i)| \quad \text{for } i \in \mathcal{T}, \\ \beta &\geq |\sigma_i \epsilon_i| \quad \text{for } i \notin \mathcal{T}. \end{aligned}$$

Squaring these inequalities and replacing ϵ_i^2 by its expected value ρ^2 gives the condition

$$\max_{i\notin\mathcal{T}} \rho^2\sigma_i^2 \leq \beta^2 \leq \min_{i\in\mathcal{T}} \sigma_i^2(\sigma_i^2 + \rho^2).$$

Thus, the regularization parameter β must satisfy

$$\rho\sigma_+ \leq \beta \leq \sigma_-\sqrt{\sigma_-^2 + \rho^2}, \tag{14a}$$

with

$$\sigma_+ = \max_{i\notin\mathcal{T}} \sigma_i, \qquad \sigma_- = \min_{i\in\mathcal{T}} \sigma_i. \tag{14b}$$

From (14) it follows that the correct sparsity pattern can be comfortably recovered if $\sigma_+ \leq \sigma_-$, i.e., if all small singular values are not in $\mathcal{T}$ and all others are in $\mathcal{T}$, just as for the truncated SVD method.

The case where L1 may offer potential advantage over truncated SVD is when $\sigma_+ > \sigma_-$. In this case, (14a) yields the requirement

$$\rho \leq \frac{\sigma_-^2}{\sqrt{\sigma_+^2 - \sigma_-^2}}. \tag{15}$$

We summarize this as follows:

Theorem 1 *Consider the L1 regularization problem* (9).

For the specific case defined above using (10), (13) *and* (14b), *the true and reconstructed models,* $\mathbf{z}^*$ *and* $\mathbf{z}$, *are expected to have the same zero structure only if either* $\sigma_+ \leq \sigma_-$ *or* (15) *holds.*

Unfortunately, if $\sigma_- \ll 1$ and/or $\sigma_+ \gg \sigma_-$ then the condition (15) may be too restrictive in practice, possibly holding only for an unrealistically small noise level.

Further difficulties arise upon considering the usual practical process of selecting the regularization parameter β by the discrepancy principle (see, e.g., [61]), i.e., such that the total misfit μ satisfies

$$\mu^2 = \frac{1}{m}\sum_i (\sigma_i(z_i - z_i^*) - \epsilon_i)^2 \approx \rho^2.$$

Let us next compute the misfit for β satisfying (14a), assuming ρ is such that this is possible, i.e., one of the conditions of Theorem 1 holds, and show that the misfit can easily be much too large in the ill-conditioned case. Conversely, this means that if β was selected by the discrepancy principle, condition (14a) would be violated.

Let us choose $\beta = \rho\sigma_+$, i.e., the smallest possible β satisfying (14). Replacing ϵ_i^2 by its expected value the expected misfit squared becomes

$$\mu^2 = \frac{1}{m}\left(\sum_{i\notin\mathcal{T}, i\leq m} \rho^2 + \sum_{i\in\mathcal{T}} \rho^2\sigma_+^2/\sigma_i^2\right).$$

The discrepancy principle requirement $\mu \approx \rho$ can now be written as

$$\frac{1}{|\mathcal{T}|}\sum_{i\in\mathcal{T}} \sigma_+^2/\sigma_i^2 \approx 1.$$

However if J is ill-conditioned the mean value of σ_+^2/σ_i^2 over the set $\mathcal{T}$ could be very large, implying that β (chosen to recover the correct sparsity) is too large to satisfy the discrepancy principle. Conversely, the value of β selected by the discrepancy principle will be too small to recover the correct sparsity of $\mathbf{z}^*$.

It is important to emphasize that we do not claim that L1 variants *cannot* work for highly ill-conditioned problems. Rather, they *may not necessarily* work. It all depends on how the sparsity of the true solution $\mathcal{T}$ and the singular values of J relate. Moreover, we do not know of a method that does better than L1 in the present sense. But then, our expectations regarding sparsity are lower for most other methods in the first place.

4.3 Nonlinear, highly ill-posed examples

In this subsection we study the DC resistivity problem on the unit square. The forward problem for v, given by

$$-\nabla \cdot (\sigma(u)(\mathbf{x})\nabla v) = q(\mathbf{x}), \quad \mathbf{x} \in \Omega, \tag{16}$$

subject to Neumann boundary conditions, predicts the potential v for given external source q and conductivity σ (parameterized in terms of u). The inverse problem is to recover the conductivity $\sigma(u)$ from partial measurements of the potential v^i, when different current patterns q^i, $i = 1, \ldots, s$, are injected into the region.

For experiment i, q^i consists of a positive point source on the left boundary and an opposite point source on the right boundary, so

$$q^i(\mathbf{x}) = \delta_{\mathbf{x},\mathbf{p}_\mathbf{L}^{\mathbf{i_L}}} - \delta_{\mathbf{x},\mathbf{p}_\mathbf{R}^{\mathbf{i_R}}},$$

where $\mathbf{p}_\mathbf{L}^{\mathbf{i_L}}$ and $\mathbf{p}_\mathbf{R}^{\mathbf{i_R}}$ are located on the left and right boundaries. Different data sets are obtained by varying the positions $\mathbf{p}_\mathbf{L}^{\mathbf{i_R}}$ and $\mathbf{p}_\mathbf{R}^{\mathbf{i_R}}$ of the two opposing point sources. We place each at $\sqrt{s}$ equidistant points including the corners, in all possible combinations, which gives a total of s data sets for a perfect square. Voltage is measured on the boundary, so the number of point in each data set is the number of boundary points of the discretization mesh. See [17, 52] and references therein for details of the problem setup such as the discretization of (16) and the solution of the resulting optimization problem.

For this nonlinear inverse problem it is well-known that, unlike for the inverse potential problem, increasing the number of data sets s allows a more accurate recovery of the resistivity $1/\sigma$. There is no reason to apply L1 here, and the purpose of the following experiments is to determine, for a piecewise continuous surface recovery, roughly at what point of such computational refinement the L1G regularization becomes worthwhile.

Example 4 EIT and DC-resistivity

We have chosen to recover a grid approximation $\mathbf{u}$ *of*

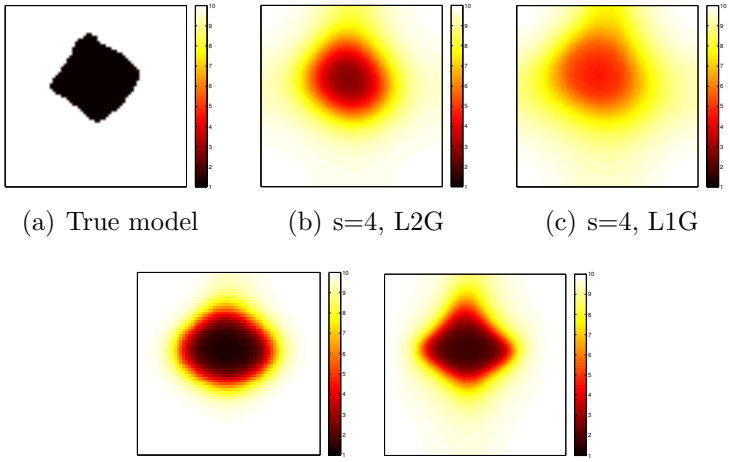
$$u(\mathbf{x}) = P^{-1}(\sigma(\mathbf{x})), \tag{17a}$$

where the transfer function

$$P(t) = \frac{1}{2}(\sigma_{max} - \sigma_{min})\tanh(t) + \frac{1}{2}(\sigma_{max} + \sigma_{min}) \tag{17b}$$

enforces a priori known upper and lower bounds on the possible conductivity.

A synthetic conductivity model is used to compute the data $\mathbf{b}$*, which is calculated on a grid that is twice as fine as the grid used for the reconstruction, and either* 3% *or* 1% *Gaussian noise is added to it.*



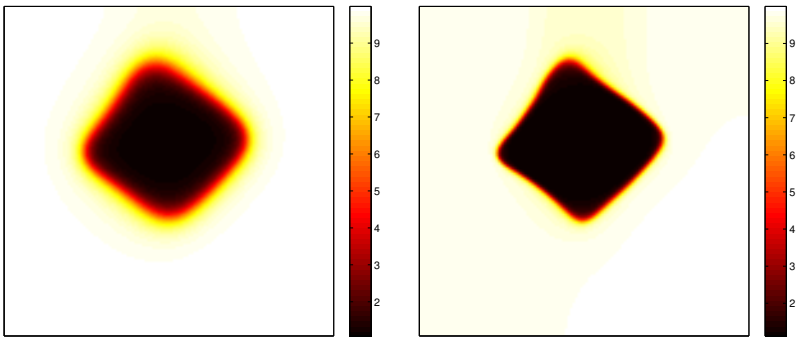
(a) True model (b) s=4, L2G (c) s=4, L1G

(d) s=64, L2G (e) s=64, L1G

Figure 9: Conductivity reconstructions for different numbers s of data sets with noise level 3%.

The ground truth model used to synthesize data consists of an object with conductivity $\sigma = 1$ (black) placed in a background of conductivity $\sigma = 10$ (white); see Figure 9(a). In (17b) we set $\sigma_{min} = 1$ and $\sigma_{max} = 10$. The inverse problem involves minimizing expressions of the form (1), (3). We compare $p = 1$ (total variation, or L1G) with $p = 2$ (L2G). A 128^2 uniform grid is used in these calculations.

Figure 9 shows the obtained reconstructions using $s = 4$ and $s = 64$ current configurations at a noise level of 3%. The regularization parameter β was tuned to result in a misfit of $3 \pm 0.1\%$. Observe that in the case of rare data $s = 4$ there appears to be no advantage to using the L1G regularization, whereas with 64 data sets the L1G reconstruction is only marginally better than L2G.



(a) s=1024, L2G (b) s=1024, L1G

Figure 10: Reconstructions for a larger number of data sets $s = 1024$ and with the noise level at only 1%. Here L1G clearly outshines L2G.

Next, we use $s = 1024$ data sets at a noise level of 1%, with β correspondingly tuned. In order to accommodate so many right hand sides we employ the stochastic adaptive algorithm described in [17]. The results are depicted in Figure 10. At this increased model accuracy and resolution the result obtained using L1G is clearly better than that obtained using L2G.

The situation described in Example 4 is not uncommon in practice. Often in geophysical experiments results of the sort depicted in Figure 9(d,e) are of sufficient quality and the lower noise level and larger number of experiments s required for obtaining the result in Figure 10(b) is a sort of luxury that is not always attained. Moreover, the forward problem considered in this section is often indicative of what is observed numerically also for more complex problems such as low frequency electromagnetic and seismic data inversions. Finally, weighted L2G variants that are routinely used in geophysical applications may further improve reconstructions without resorting to ℓ_1-based regularization. In view of the occasionally significantly higher cost of computing with L1G it cannot be automatically concluded that the latter is worthwhile for this application, although it is a viable option that we always entertain in the course of our research.

5 Summary

In this paper we have investigated the relative performance of ℓ_1-based regularization techniques on several examples and case studies. We have shown cases where such methods are worse than ℓ_2-based ones in the sense of costing more without delivering more (Examples 1 and 4), and other cases where such methods produce better results (see Figures 4b and 10). Further, we have shown cases where the ℓ_1-based results appear to be more misleading than corresponding ℓ_2-based results (Example 3).

In Section 4.2 we have analyzed the effect of ill-conditioning on the ability of an L1 method to correctly recover solution sparsity. Theorem 1 and the arguments following it suggest severe limitations in case of extreme ill-conditioning such as arises in certain inverse problems.

The results in Section 4.3 demonstrate how and when L1G becomes favoured as the quality of the data improves. This in itself is intuitively expected, but less clear is where the cross-over point occurs in realistic situations. Unfortunately, we had to tweak the problem beyond what may be expected in many geophysical situations in order to observe the L1G takeover.

Let us again stress our overall conviction that the swing of the pendulum in recent years towards ℓ_1-based techniques is rather important and not merely refreshing. Our purpose here, far from opposing this trend, is to simply suggest that this virtual pendulum should not swing too far and away, to realms beyond reason. To this end we note the following.

- In many situations, ℓ_1-based regularization is well-worth using. Such techniques can provide exciting advances (e.g., in model reduction, computer graphics, image processing and reconstruction of surfaces with discontinuities).
- However, such techniques are not good for all problems, and it is dangerous (and may consume many student-years) to apply them blindly.
- In practice, we recommend to always consider first using ℓ_2-based regularization techniques, because they are simpler, more easy to compute with, and do not introduce nonlinearities or lower smoothness. Only upon deciding that these are not sufficiently good for the given application, it is highly advisable to proceed to examine ℓ_1-based alternatives (when this makes sense).
- Last but not least, the possibility of combining ℓ_1- and ℓ_2-based techniques suggests itself. We have already commented on using the Huber switching function as well as IRLS techniques [56, 61, 33, 1, 30] for this purpose in the L1G–L2G context, but these ideas are also very popular in the image processing and computer vision literature in mixing the L1 and L2 approaches [34]. Another popular approach is to employ an empirical Bayesian framework in order to learn an appropriate mix [37, 57].

References

[1] U. Ascher, E. Haber, and H. Huang. On effective methods for implicit piecewise smooth surface recovery. *SIAM J. Sci. Comput.*, 28:339–358, 2006.

[2] C. Avron, A. Sharf, C. Greif, and D. Cohen-Or. ℓ_1-sparse reconstruction of sharp point set surfaces. *ACM trans. on Graphics*, 29(5):135:1–12, 2010.

[3] S. Becker, J. Bobin, and E.J. Candes. Nesta: a fast and accurate first-order method for sparse recovery. *SIAM J. on Imaging Sciences*, 4:1–39, 2010.

[4] E. van den Berg and M. Friedlander. Sparse optimization with least-squares constraints. *SIAM J. Optimization*, 21:1201–1229, 2011.

[5] D. Bertsekas. *Convex Optimization Theory*. Athena Scientific, 2009.

[6] K. Bredies, K. Kunisch, and T. Pock. Total generalized variation. *SIAM J. Imaging Sciences*, 3:492–526, 2010.

[7] J.-F. Cai, S. Osher, and Z. Shen. Split bregman methods and frame based image restoration. *SIAM J. multiscale modeling and simulation*, 8(2):337–369, 2009.

[8] D. Calvetti and E. Somersalo. *Introduction to Bayesian Scientific Computing*. Springer, 2007.

[9] E. Candes, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Pure Appl. Math.*, 59(8):1207–1223, 2006.

[10] E. J. Candes, M. B. Wakin, and S. Boyd. Enhancing sparsity by reweighted l1 minimization. *Journal of Fourier Analysis and Applications*, 14(5):877–1905, 2008.

[11] T. Chan and J. Shen. *Image Processing and Analysis: Variational, PDE, Wavelet and Stochastic Methods*. SIAM, 2005.

[12] R. Chartrand. Fast algorithms for nonconvex compressive sensing: MRI reconstruction from very few data. *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2009.

[13] S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43:129–159, 2001.

[14] J. Claerbout and F. Muir. Robust modeling with erratic data. *Geophysics*, 38:826–844, 1973.

[15] A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best k-term approximation. *Journal of the American Mathematical Society*, 22(1):211–231, 2008.

[16] K. van den Doel and U. Ascher. Dynamic level set regularization for large distributed parameter estimation problems. *Inverse Problems*, 23:1271–1288, 2007.

[17] K. van den Doel and U. Ascher. Adaptive and stochastic algorithms for EIT and DC resistivity problems with piecewise constant solutions and many measurements. *SIAM J. Scient. Comput.*, 2012. DOI: 10.1137/110826692.

[18] K. van den Doel, U. Ascher, and D. Pai. Computed myography: three dimensional reconstruction of motor functions from surface EMG data. *Inverse Problems*, 24:065010, 2008.

[19] K. van den Doel, U. Ascher, and D. Pai. Source localization in electromyography using the inverse potential problem. *Inverse Problems*, 27:025008, 2011.

[20] D. Donoho. For most large underdetermined systems of linear equations, the minimal l1 solution is also the sparsest solution. *Comm. Pure Applied Math.*, 7:907–934, 2006.

[21] D. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 47(7):2845–2862, 2001.

[22] D. Donoho and J. Tanner. Sparse nonnegative solutions of underdetermined linear equations by linear programming. *Proc. Nat. Acad. Sciences*, 102:9446–9451, 2005.

[23] A. Ebtehaj, E. Foufoula-Georgiou, and G. Lerman. Sparse regularization for precipitation downscaling. *J. Geophys. Res.*, 117:D08107 doi:10.1029/2011JD017057, 2012.

[24] M. Elad. *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer, 2010.

[25] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Kluwer, 1996.

[26] M. Figueiredo, R. Nowak, and S. Wright. Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems. *IEEE J. Special Topics on Signal Processing*, 1:586–598, 2007.

[27] M. Freitag, N. Nichols, and C. Budd. Resolution of sharp fronts in the presence of model error in variational data assimilation. *Q.J.R. Meteorol. Soc.*, page doi: 10.1002/qj.2002, 2012.

[28] R. E. Greenblatt. Probabilistic reconstruction of multiple sources in the neuroelectromagnetic inverse problem. *Inverse problems*, 9:271–284, 1993.

[29] E. Haber, U. Ascher, and D. Oldenburg. Inversion of 3D electromagnetic data in frequency and time domain using an inexact all-at-once approach. *Geophysics*, 69:1216–1228, 2004.

[30] E. Haber, S. Heldmann, and U. Ascher. Adaptive finite volume method for distributed non-smooth parameter identification. *Inverse Problems*, 23:1659–1676, 2007.

[31] M. S. Hämäläinen and R. J. Ilmoniemi. Interpreting magnetic fields of the brain-minimum norm estimates. *Med. Biol. Eng. Comput.*, 32:35–42, 1994.

[32] P. C. Hansen. *Rank Deficient and Ill-Posed Problems*. SIAM, Philadelphia, 1998.

[33] P.-C. Hansen, J. Nagy, and D. O'Leary. *Deblurring Images: Matrices, Spectra and Filtering*. SIAM, 2006.

[34] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[35] H. L. F. Helmholtz. Ueber einige gesetze der verheilung elektrischer ströme in körperlicher leitern mit anwendung auf die thierisch-elektrischen versuche. *Ann. Physik und Chemie*, 9:211–233, 1853.

[36] H. Huang. *Efficient Reconstruction of 2D Images and 3D Surfaces*. PhD thesis, University of BC, Vancouver, 2008.

[37] H. Huang, E. Haber, and L. Horesh. Optimal estimation of l1 regularization prior from a regularized empirical bayesian risk standpoint. *Inverse Problems and Imaging*, 2013.

[38] H. Huang, D. Li, R. Zhang, U. Ascher, and D. Cohen-Or. Consolidation of unorganized point clouds for surface reconstruction. *ACM Trans. Graphics (SIGGRAPH Asia)*, 29(5), 2009.

[39] H. Huang, S. Wu, M. Gong, D. Cohen-Or, U. Ascher, and H. Zhang. Edge-aware point set resampling. *ACM trans. on Graphics*, 2013.

[40] A. Juditsky and A. Nemirovski. On verifiable sufficient conditions for sparse signal recovery via l1-minimization. *Mathematical Programming Ser. B*, 127:57–88, 2008.

[41] J. Kaipo and E. Somersalo. *Statistical and Computational Inverse Problems*. Springer, 2005.

[42] S. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinvesky. A method for large-scale l1-regularized least squares problems with applications in signal processing and statistics. *IEEE J. Select. Topics Signal Process*, 2007.

[43] T. K ohler, M. Wagner, M. Fuchs, H. A. Wischmann, R. Denkckhahn, and A. Theissen. Depth Normalization in MEG/EEG Current Density Imaging. In *18th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Amsterdam*, pages 812–813, 2006.

[44] A. Levin, R. Fergus, F. Durand, and W. Freeman. Image and depth from a conventional camera with a coded aperture. *ACM trans. on Graphics (SIGGRAPH)*, 26(3):70, 2007.

[45] Y. Lipman, D. Cohen-Or, D. Levin, and H. Tal-Ezer. Parameterization-free projection for geometry reconstruction. *ACM trans. on Graphics (SIGGRAPH)*, 26(3):22, 2007.

[46] S. Mallat. *A Wavelet Tour of Signal Processing: the Sparse Way*. Academic Press, 2009. 3rd Ed.

[47] N. Meinshausen and P. Buehlmann. Stability selection. *J. Royal Stat. Soc.*, B72:417–473, 2010.

[48] C. M. Michel, M. M. Murray, G. Lantz, S Gonzalez, L. Spinelli, and R. G. de Peralta. EEG source imaging. *Clinical neurophysiology*, 115:2195–2222, 2004.

[49] K. Murphy. *Machine Learning: a Probabilistic Perspective*. MIT Press, 2012.

[50] M. Osborne. *Finite Algorithms in Optimization and Data Analysis*. Wiley, 1985.

[51] S. Osher and R. Fedkiw. *Level Set Methods and Dynamic Implicit Surfaces*. Springer, 2003.

[52] F. Roosta-Khorasani, K. van den Doel, and U. Ascher. Stochastic algorithms for inverse problems involving PDEs and many measurements. *Submitted*, 2012.

[53] L. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60:259–268, 1992.

[54] R. Saab, R. Chartrand, and O. Yilmaz. Stable sparse approximations via nonconvex optimization. *33rd IEEE International Conference Acoustics, Speech and Signal Proc. (ICASSP)*, 2008.

[55] Y. Saad. *Iterartive Methods for Sparse Linear Systems*. PWS Publishing Company, 1996.

[56] G. Sapiro. *Geometric Partial Differential Equations and Image Analysis*. Cambridge, 2001.

[57] K. Swersky, M. Ranzato, D. Buchman, B. Marlin, and N. de Freitas. On autoencoders and score matching for energy based models. In *Proc. 28th Intl. Conf. Machine Learning, Bellevue, WA, USA*, 2011.

[58] A. Tarantola. *Inverse problem theory*. Elsevier, Amsterdam, 1987.

[59] H. Taylor, S. Banks, and J. McCoy. Deconvolution with the l1 norm. *Geophysics*, 44:39–52, 1979.

[60] A. N. Tikhonov and V. Ya. Arsenin. *Methods for Solving Ill-posed Problems*. John Wiley and Sons, Inc., 1977.

[61] C. Vogel. *Computational methods for inverse problem*. SIAM, Philadelphia, 2002.

[62] H. Wagner. Linear programming techniques for regression analysis. *Proc.*, 54:206–212, 1959.