The Lost Honour of $\ell_2$-Based Regularization

Uri Ascher

Department of Computer Science
University of British Columbia
October 2012
OUTLINE

- Motivation-introduction
- Poor data
- Highly ill-conditioned large problems
- Conclusions
Outline motivation-introduction

- Sparse solutions
- Piecewise smooth surface reconstruction
- $\ell_1$-based and $\ell_2$-based regularizations
JPEG IMAGE COMPRESSION

- Represent raw image using discrete cosine transform (DCT) or wavelet transform.
- Retain only largest coefficients, setting the rest to 0.
- Below, compression factor \( \approx \frac{1}{16} \):
Mathematically, consider the problem

\[
\min_u \|Wu\|_p \\
\text{s.t.} \quad \|Ju - b\|_2 \leq \epsilon
\]

where \( J \) is an \( m \times n \) real matrix, \( m \leq n \), and \( \epsilon \geq 0 \) depends on noise. A related formulation:

\[
\min_u \frac{1}{2} \|Ju - b\|_2^2 + \beta R(u). 
\]

Ideally we want \( p = 0 \), meaning min number of nonzero components, i.e., find sparsest model representation. But this problem is NP-hard.

- Using \( p = 2 \), i.e., \( \ell_2 \) norm, \( z = Wu \) typically is not sparse.
- Using \( p = 1 \), i.e., \( \ell_1 \) norm, \( z = Wu \) often is sparse!
Motivation-Introduction

Sparse solutions

How to obtain sparse representation?

- Mathematically, consider the problem

\[
\min_u \|Wu\|_p \\
\text{s.t.} \quad \|Ju - b\|_2 \leq \epsilon
\]

where \( J \) is an \( m \times n \) real matrix, \( m \leq n \), and \( \epsilon \geq 0 \) depends on noise.

A related formulation:

\[
\min_u \frac{1}{2} \|Ju - b\|_2^2 + \beta R(u).
\]

- Ideally we want \( p = 0 \), meaning min number of nonzero components, i.e., find sparsest model representation. But this problem is NP-hard.

- Using \( p = 2 \), i.e., \( \ell_2 \) norm, \( z = Wu \) typically is not sparse.
- Using \( p = 1 \), i.e., \( \ell_1 \) norm, \( z = Wu \) often is sparse!
Mathematically, consider the problem
\[
\min_u \|Wu\|_p \quad \text{s.t.} \quad \|Ju - b\|_2 \leq \epsilon
\]
where \(J\) is an \(m \times n\) real matrix, \(m \leq n\), and \(\epsilon \geq 0\) depends on noise.

A related formulation:
\[
\min_u \frac{1}{2} \|Ju - b\|_2^2 + \beta R(u).
\]

Ideally we want \(p = 0\), meaning min number of nonzero components, i.e., find sparsest model representation. But this problem is NP-hard.

Using \(p = 2\), i.e., \(\ell_2\) norm, \(z = Wu\) typically is not sparse.

Using \(p = 1\), i.e., \(\ell_1\) norm, \(z = Wu\) often is sparse!
How to obtain sparse representation?

- Mathematically, consider the problem
  \[
  \min_u \|Wu\|_p \\
  \text{s.t.} \quad \|Ju - b\|_2 \leq \epsilon
  \]
  where $J$ is an $m \times n$ real matrix, $m \leq n$, and $\epsilon \geq 0$ depends on noise.

A related formulation:

\[
\min_u \frac{1}{2} \|Ju - b\|_2^2 + \beta R(u).
\]

- Ideally we want $p = 0$, meaning min number of nonzero components, i.e., find sparsest model representation. But this problem is NP-hard.
- Using $p = 2$, i.e., $\ell_2$ norm, $z = Wu$ typically is not sparse.
- Using $p = 1$, i.e., $\ell_1$ norm, $z = Wu$ often is sparse!
Find \( u(x, y) \) that cleans given image \( b(x, y) \).

Cameraman 256 \( \times \) 256. Left: clean image. Right: with noise added.
**Image deblurring in 2D**

Find $u(x, y)$ that sharpens given image $b(x, y)$.

Boat $256 \times 256$: clean and blurred.
**Tikhonov regularization: encourage piecewise smoothness**

- Denote by $u$ discretized mesh function of $u(x, y)$, likewise data $b$.
- Tikhonov: $\min \frac{1}{2} \|Ju - b\|^2 + \beta R(u)$.
- Consider

$$R(u) \text{ discretization of } \frac{1}{p} \int_{\Omega} |\nabla u|^p$$

- Using $p = 2$, i.e., $\ell_2$ norm on gradient, then solution is smeared across discontinuities, i.e., where $u(x, y)$ has jumps.
- Using $p = 1$, i.e., $\ell_1$ norm on gradient (total variation), then solution sharper near jumps but often is blocky: sparse gradient!
Tikhonov regularization: encourage piecewise smoothness

- Denote by $u$ discretized mesh function of $u(x, y)$, likewise data $b$.
- Tikhonov: $\min \frac{1}{2} \|Ju - b\|^2 + \beta R(u)$.
- Consider

\[ R(u) \text{ discretization of } \frac{1}{p} \int_{\Omega} |\nabla u|^p \]

- Using $p = 2$, i.e., $\ell_2$ norm on gradient, then solution is smeared across discontinuities, i.e., where $u(x, y)$ has jumps.
- Using $p = 1$, i.e., $\ell_1$ norm on gradient (total variation), then solution sharper near jumps but often is blocky: sparse gradient!
Tikhonov regularization: encourage piecewise smoothness

- Denote by \( u \) discretized mesh function of \( u(x, y) \), likewise data \( b \).
- Tikhonov: \[
\min \frac{1}{2} \| Ju - b \|^2 + \beta R(u) .
\]
- Consider

\[
R(u) \text{ discretization of } \frac{1}{p} \int_{\Omega} |\nabla u|^p
\]

- Using \( p = 2 \), i.e., \( \ell_2 \) norm on gradient, then solution is smeared across discontinuities, i.e., where \( u(x, y) \) has jumps.
- Using \( p = 1 \), i.e., \( \ell_1 \) norm on gradient (total variation), then solution sharper near jumps but often is blocky: \textbf{sparse gradient}!
Huber function regularization

- [Ascher, Haber & Huang, 2006; Haber, Heldman & Ascher, 2007; Huang & Ascher, 2008]: modify total variation.

- Use the Huber function

  \[ \rho(s) = \begin{cases} 
  s, & s \geq \gamma, \\
  s^2/(2\gamma) + \gamma/2, & s < \gamma 
  \end{cases} \]

  \[ \Rightarrow \nabla R(u) \leftarrow -\nabla \cdot \left( \min\left\{ \frac{1}{\gamma}, \frac{1}{|\nabla u|} \right\} \nabla u \right) \]

  - a particular anisotropic diffusion operator.

- Choose \( \gamma \) adaptively, close to total variation:

  \[ \gamma = \frac{h}{|\Omega|} \int_{\Omega} |\nabla u|. \]
Image deblurring in 2D: example

\begin{align*}
\text{TYPE} &= \text{`DISK'}, \quad \text{RADIUS} = 5, \quad \eta = 1, \quad \beta = 10^{-4}.
\end{align*}
SURFACE TRIANGLE MESH DE NOISING

[Huang & Ascher, 2008; 2009]: use an even sharper (non-convex) regularizer.

Left: true. Center: corrupted. Right: reconstructed

Left: corrupted. Right: reconstructed
For sharper reconstruction can consider $\ell_p$, $0 < p < 1$ instead of $p = 1$.

However, convexity is lost, yielding both computational and theoretical difficulties [Saab, Chartrand & Yilmaz, 2008; Chartrand, 2009].

[Levin, Fergus, Durand & Freeman, 2007]: choose $p = 0.8$ for an image deblurring application.

But as it turns out, there is nothing special about $p = 0.8$, and in most graphics applications $p = 1$ is fine (and easier to work with).
For sharper reconstruction can consider $\ell_p$, $0 < p < 1$ instead of $p = 1$.

However, convexity is lost, yielding both computational and theoretical difficulties [Saab, Chartrand & Yilmaz, 2008; Chartrand, 2009].

[Levin, Fergus, Durand & Freeman, 2007]: choose $p = 0.8$ for an image deblurring application.

But as it turns out, there is nothing special about $p = 0.8$, and in most graphics applications $p = 1$ is fine (and easier to work with).
For sharper reconstruction can consider $\ell_p$, $0 < p < 1$ instead of $p = 1$.

However, convexity is lost, yielding both computational and theoretical difficulties [Saab, Chartrand & Yilmaz, 2008; Chartrand, 2009]. [Levin, Fergus, Durand & Freeman, 2007]: choose $p = 0.8$ for an image deblurring application.

But as it turns out, there is nothing special about $p = 0.8$, and in most graphics applications $p = 1$ is fine (and easier to work with).
For sharper reconstruction can consider $\ell_p$, $0 < p < 1$ instead of $p = 1$.

However, convexity is lost, yielding both computational and theoretical difficulties [Saab, Chartrand & Yilmaz, 2008; Chartrand, 2009].

[Levin, Fergus, Durand & Freeman, 2007]: choose $p = 0.8$ for an image deblurring application.

But as it turns out, there is nothing special about $p = 0.8$, and in most graphics applications $p = 1$ is fine (and easier to work with).
Lots of exciting recent work on sparse recovery of signals, including compressive sensing, using $\ell_1$. Texts: [Mallat, 2009; Elad, 2010]. Some key papers: [Donoho & Tanner, 2005; Donoho, 2006; Candes, Wakin & Boyd, 2008; Juditsky & Nemirovski, 2008].

Lots of exciting recent work on total variation-based methods. Texts: [Chan & Shen, 2005; Osher & Fedkiw, 2003]. Some key papers: [Rudin, Osher & Fatemi, 1992].

Perhaps $\ell_1$-based regularization methodology ought to altogether replace the veteran $\ell_2$-based regularization approaches.

A step too far?! – This does not quite agree with our experience in several situations.
VIRTUES OF $\ell_1$-BASED REGULARIZATION

- Lots of exciting recent work on sparse recovery of signals, including compressive sensing, using $\ell_1$. Texts: [Mallat, 2009; Elad, 2010]. Some key papers: [Donoho & Tanner, 2005; Donoho, 2006; Candes, Wakin & Boyd, 2008; Juditsky & Nemirovski, 2008].


- Perhaps $\ell_1$-based regularization methodology ought to altogether replace the veteran $\ell_2$-based regularization approaches.

- A step too far?! – This does not quite agree with our experience in several situations.
VIRTUES OF $\ell_1$-BASED REGULARIZATION

- Lots of exciting recent work on sparse recovery of signals, including compressive sensing, using $\ell_1$. Texts: [Mallat, 2009; Elad, 2010]. Some key papers: [Donoho & Tanner, 2005; Donoho, 2006; Candes, Wakin & Boyd, 2008; Juditsky & Nemirovski, 2008].


- Perhaps $\ell_1$-based regularization methodology ought to altogether replace the veteran $\ell_2$-based regularization approaches.

- A step too far?! – This does not quite agree with our experience in several situations.
Virtues of $\ell_1$-based regularization

- Lots of exciting recent work on sparse recovery of signals, including compressive sensing, using $\ell_1$. Texts: [Mallat, 2009; Elad, 2010]. Some key papers: [Donoho & Tanner, 2005; Donoho, 2006; Candes, Wakin & Boyd, 2008; Juditsky & Nemirovski, 2008].


- Perhaps $\ell_1$-based regularization methodology ought to altogether replace the veteran $\ell_2$-based regularization approaches.

- A step too far?! – This does not quite agree with our experience in several situations.
Notation

We saw four regularization methods with variations:

- $\frac{1}{2} \| W u \|_2^2$ – denote by \textbf{L2}
- $\| W u \|_1$ – denote by \textbf{L1}
- $\frac{1}{2} \int_\Omega |\nabla u|^2$ – denote by \textbf{L2G}
- $\int_\Omega |\nabla u|$ – denote by \textbf{L1G} (TV)
Another image deblurring example

- Goal: recover clean image given noisy, blurred data.
- Blurring kernel: $e^{-\|x\|^2_{2}/2\sigma}$ with $\sigma = 0.01$; the blurred data is further corrupted by 1% white noise.

Left: ground truth. Right: Blurred and noisy image.
Another image deblurring example cont.

Try (i) RestoreTools [Hansen, Nagy & O’Leary, 2006], which is L2-type recovery strategy \((p = 2, W = I)\); (ii) GPSR [Figuerido, Nowak & Wright, 2007], which employs a wavelet L1 recovery algorithm; and (iii) a straightforward L1G code.

Note the L2 recovery costs a tiny fraction compared to the L1 solve, and the result is not worse.
Outline

- Motivation-introduction
- Poor data
- Highly ill-conditioned large problems
- Conclusions
Outline poor data

- Outliers
- Relatively high noise level
- Rare (sparse) data
**Poor data**

- **Outliers:** $\ell_1$-based data fitting is good for outliers! (Known for decades; for point clouds, see LOP [Lipman, Cohen-Or, Levin & Tal-Ezer, 2007].)

  Can see this by considering overdetermined $n \times m$ problem

  \[
  \min_x \| \hat{J}x - y \|_1,
  \]

  $\hat{J}$ “long and skinny”. This can be cast as a linear programming problem: optimal solution $x$ gives at least $m$ zero residual rows. Outliers usually correspond to non-basic residual rows, hence their values matter less.

- **High noise level:** $\ell_1$-based no longer has particular advantage.

- **Rare (sparse) data** given only at a few locations: $\ell_1$-based no longer has particular advantage.

- **Note:** what counts as “poor” depends on the circumstances.
Outliers: $\ell_1$-based data fitting is good for outliers! (Known for decades; for point clouds, see LOP [Lipman, Cohen-Or, Levin & Tal-Ezer, 2007].)

Can see this by considering overdetemined $n \times m$ problem

$$\min_x \| \hat{J}x - y \|_1,$$

$\hat{J}$ “long and skinny”. This can be cast as a linear programming problem: optimal solution $x$ gives at least $m$ zero residual rows. Outliers usually correspond to non-basic residual rows, hence their values matter less.

- High noise level: $\ell_1$-based no longer has particular advantage.
- Rare (sparse) data given only at a few locations: $\ell_1$-based no longer has particular advantage.
- Note: what counts as “poor” depends on the circumstances.
Poor data

- **Outliers**: $\ell_1$-based data fitting is good for outliers! (Known for decades; for point clouds, see LOP [Lipman, Cohen-Or, Levin & Tal-Ezer, 2007].)
  
  Can see this by considering over-determined $n \times m$ problem

  \[
  \min_x \| \hat{J} x - y \|_1,
  \]

  $\hat{J}$ “long and skinny”. This can be cast as a linear programming problem: optimal solution $x$ gives at least $m$ zero residual rows. Outliers usually correspond to non-basic residual rows, hence their values matter less.

- **High noise level**: $\ell_1$-based no longer has particular advantage.

- **Rare (sparse) data** given only at a few locations: $\ell_1$-based no longer has particular advantage.

- Note: what counts as “poor” depends on the circumstances.
Poor data

- **Outliers**: $\ell_1$-based data fitting is good for outliers! (Known for decades; for point clouds, see LOP [Lipman, Cohen-Or, Levin & Tal-Ezer, 2007].)

  Can see this by considering overdetermined $n \times m$ problem

  \[
  \min_x \| \hat{J} x - y \|_1,
  \]

  $\hat{J}$ “long and skinny”. This can be cast as a linear programming problem: optimal solution $x$ gives at least $m$ zero residual rows. Outliers usually correspond to non-basic residual rows, hence their values matter less.

- **High noise level**: $\ell_1$-based no longer has particular advantage.

- **Rare (sparse) data** given only at a few locations: $\ell_1$-based no longer has particular advantage.

- Note: what counts as “poor” depends on the circumstances.
**Poor data**

- **Outliers**: $\ell_1$-based data fitting is good for outliers! (Known for decades; for point clouds, see LOP [Lipman, Cohen-Or, Levin & Tal-Ezer, 2007].)
  Can see this by considering **overdetermined** $n \times m$ problem

\[
\min_x \| \hat{J}x - y \|_1,
\]

$\hat{J}$ “long and skinny”. This can be cast as a linear programming problem: optimal solution $x$ gives at least $m$ zero residual rows. Outliers usually correspond to non-basic residual rows, hence their values matter less.

- **High noise level**: $\ell_1$-based no longer has particular advantage.

- **Rare (sparse) data** given only at a few locations: $\ell_1$-based no longer has particular advantage.

- **Note**: what counts as “poor” depends on the circumstances.
Rare (sparse) data: a simple example

- Recover signal $u^*(t)$ on $[0, 1]$ discretized at $n = 512$ equidistant points, from $m \ll n$ data pairs $(t_i, u_i)_{i=1}^m$.
- Draw locations $t_i$ at random from the $n$-mesh. Calculate $u_i$ as $u^*(t_i)$ plus 5% white noise.
- Tune $\beta$ in $\min \frac{1}{2} \| Ju - b \|^2 + \beta R(u)$ by discrepancy principle.

Using $m=9$ data pairs, with $\beta_{L1G}=.08$, $\beta_{L2G}=.04$

For $m=9$, no significant quality difference between $\ell_1$-based and $\ell_2$-based
Rare (Sparse) data: a simple example

- Recover signal $u^*(t)$ on $[0, 1]$ discretized at $n = 512$ equidistant points, from $m \ll n$ data pairs $(t_i, u_i)_{i=1}^m$.
- Draw locations $t_i$ at random from the $n$-mesh. Calculate $u_i$ as $u^*(t_i)$ plus 5% white noise.
- Tune $\beta$ in $\min \frac{1}{2} \|Ju - b\|^2 + \beta R(u)$ by discrepancy principle.

Using $m=9$ data pairs, with $\beta_{L1G} = .08$, $\beta_{L2G} = .04$

For $m=9$, no significant quality difference between $\ell_1$-based and $\ell_2$-based.
**Rare (Sparse) Data: A Simple Example Cont.**

Using $m=28$ data pairs, with $\beta_{L1G}=.08$. **Left:** $\beta_{L2G}=.002$, **Right:** $\beta_{L2G}=.02$

For $m=28$, the $\ell_1$-based method performs better than the $\ell_2$-based one.
Example: Edge aware resampling (EAR)

[Huang, Wu, Gong, Cohen-Or, Ascher & Zhang, 2012; Huang, Li, Zhang, Ascher & Cohen-Or, 2009]

Left: Point cloud for fin shape 150° angle at edge; 1.0% noise. Right: Profile of normals using classical PCA.
Alternative: $\ell_1$ for normals [Avron, Sharf, Greif & Cohen-Or, 2010].

**Left:** Normals after applying $\ell_1$-minimization. **Right:** Normals after applying EAR.
Motivation-introduction
Poor data
Highly ill-conditioned large problems
Conclusions
Highly ill-conditioned large problems

Outline highly ill-conditioned large problems

- Computational difficulty with L1
- Computed myography and inverse potential
- Nonlinearity, EIT and DC resistivity
- Assessment
Highly ill-conditioned large problems

Computational difficulties with $\ell_1$

**Computational difficulty with L1**

$$\min_u \frac{1}{2} \| J u - b \|^2 + \beta R(u)$$

- Several famous codes for L1 employ gradient projection with acceleration (because such methods extend well for constrained non-smooth problems).
- However, such methods have poor convergence properties for large, highly ill-conditioned problems.
- The situation is better for L1G (total variation and variants), although still L2G is faster.
Computational difficulty with $L_1$

$$\min_u \frac{1}{2} \| J u - b \|^2 + \beta R(u)$$

- Several famous codes for L1 employ gradient projection with acceleration (because such methods extend well for constrained non-smooth problems).
- However, such methods have poor convergence properties for large, highly ill-conditioned problems.
- The situation is better for L1G (total variation and variants), although still L2G is faster.
Computationally difficulty with L1

\[ \min_u \frac{1}{2} \| J u - b \|^2 + \beta R(u) \]

- Several famous codes for L1 employ gradient projection with acceleration (because such methods extend well for constrained non-smooth problems).
- However, such methods have poor convergence properties for large, highly ill-conditioned problems.
- The situation is better for L1G (total variation and variants), although still L2G is faster.

- Determine electric activity of individual muscles in human limb using sEMG.
- Applications: prosthetic control, muscle function assessment.

---

**Source localization in electromyography**

Preprocessing

- Morphometric MRI data
- MRI scan
- FEM model

Inverse equations

- Solve inverse equations
- 3D current activations

Visualization and analysis

sEMG imaging

- sEMG data

---
**Example: MRI data segmented into geometry model**

3D model of the upper arm constructed from MRI data of a human subject segmented into different anatomical regions: brachialis, biceps, triceps, fat and skin, and bone.
Highly ill-conditioned large problems

Computed myography

Reconstructions

\( \ell_1 \) and \( \ell_2 \)
CMG problem

Potential problem

\[-\nabla \cdot (\sigma \nabla v) = u \quad \text{in} \quad \Omega,\]

subject to Neumann boundary conditions.

- Inverse problem: recover source $u$ from measurements of $v$ on boundary.
- In previous notation, $J = QA^{-1}$, with $A$ the discretization matrix of the potential problem and $Q$ the data projector operator.
- Highly ill-posed: many different sources explain data. Regularization should select solution that agrees with a priori information.
- Model consists of set of discrete tripole s: use L1 regularization?
- Typical 3D calculation: 50,000 finite elements. L1 codes based on gradient projection are hopeless!
**CMG problem**

Potential problem

\[-\nabla \cdot (\sigma \nabla v) = u \quad \text{in} \quad \Omega,\]

subject to **Neumann boundary conditions**.

- **Inverse problem**: recover source $u$ from measurements of $v$ on boundary.
- In previous notation, $J = QA^{-1}$, with $A$ the discretization matrix of the potential problem and $Q$ the data projector operator.
- Highly ill-posed: many different sources explain data. Regularization should select solution that agrees with a priori information.
- Model consists of set of discrete **tripoles**: use L1 regularization?
- Typical 3D calculation: 50,000 finite elements. L1 codes based on gradient projection are hopeless!
CMG PROBLEM

Potential problem

\[-\nabla \cdot (\sigma \nabla \nu) = u \quad \text{in} \quad \Omega,\]

subject to **Neumann boundary conditions**.

- **Inverse problem**: recover source \(u\) from measurements of \(\nu\) on boundary.
- In previous notation, \(J = QA^{-1}\), with \(A\) the discretization matrix of the potential problem and \(Q\) the data projector operator.
- Highly ill-posed: many different sources explain data. Regularization should select solution that agrees with a priori information.
- Model consists of set of discrete *tripoles*: use \(L_1\) regularization?
- Typical 3D calculation: 50,000 finite elements. \(L_1\) codes based on gradient projection are hopeless!
CMG PROBLEM

Potential problem

\[-\nabla \cdot (\sigma \nabla \upsilon) = u \quad \text{in} \quad \Omega,\]

subject to Neumann boundary conditions.

- **Inverse problem**: recover source \( u \) from measurements of \( \upsilon \) on boundary.
- In previous notation, \( J = QA^{-1} \), with \( A \) the discretization matrix of the potential problem and \( Q \) the data projector operator.
- Highly ill-posed: many different sources explain data. Regularization should select solution that agrees with a priori information.
- Model consists of set of discrete **tripoles**: use L1 regularization?
- Typical 3D calculation: 50,000 finite elements. L1 codes based on gradient projection are hopeless!
CMG PROBLEM

Potential problem

\[-\nabla \cdot (\sigma \nabla \nu) = u \quad \text{in} \quad \Omega,\]

subject to Neumann boundary conditions.

- Inverse problem: recover source \( u \) from measurements of \( \nu \) on boundary.
- In previous notation, \( J = QA^{-1} \), with \( A \) the discretization matrix of the potential problem and \( Q \) the data projector operator.
- Highly ill-posed: many different sources explain data. Regularization should select solution that agrees with a priori information.
- Model consists of set of discrete tripoles: use L1 regularization?
- Typical 3D calculation: 50,000 finite elements. L1 codes based on gradient projection are hopeless!
Inverse potential problem

- Study the simpler problem with $\sigma \equiv 1$ on a square (2D):

  $$-\Delta v = u(x), \quad x \in \Omega.$$ 

- Generate data using $64 \times 64$ grid, pollute by 1% white noise. Ground truth for sought source $u$ are various tripole distributions.

**Left: ground truth. Center: L2G. Right: L1G.**
More example instances

Left: ground truth. Center: L2G. Right: L1G.

Next, consider a point charge pair for a source:

Left: ground truth. Center: L2G. Right: L1G.
More example instances

Left: ground truth. Center: L2G. Right: L1G.

Next, consider a point charge pair for a source:

Left: ground truth. Center: L2G. Right: L1G.
**L1 and L2 for a point charge pair**


Indeed, the L1 methods give sparse solutions, but not very sparse and not very good ones.
Simple analysis

\[
\min_u \frac{1}{2} \|Ju - b\|^2 + \beta \|Wu\|_1
\]

- **Singular value decomposition:**
  \[J = U\Sigma V^T\]
  with \(\Sigma = \{\text{diag } \sigma_i\}, \sigma_1 \geq \cdots \geq \sigma_m \geq 0\), and \(U, V\) orthogonal.

- **Special case:** \(W = V^T\). Then for \(z = Wu, c = U^Tb\), consider
  \[
  \min_z \frac{1}{2} \|\Sigma z - c\|^2 + \beta \|z\|_1
  \]

- **Representation for true model** \(z^* = Wu^*\) satisfies \(\sigma_i z_i^* = c_i + \varepsilon_i\).
  Consider a sparse representation
  \[
  z_i^* = 1 \text{ if } i \in T, \quad = 0 \text{ if } i \notin T.
  \]

Under what conditions can we expect to **stably** calculate \(z\) with the same sparsity (i.e. \(z_i = 0 \text{ iff } i \notin T\))?
Simple analysis

\[
\min_u \frac{1}{2} \|Ju - b\|^2 + \beta \|Wu\|_1
\]

- Singular value decomposition:
  \[
  J = U\Sigma V^T
  \]
  with \(\Sigma = \{\text{diag } \sigma_i\}\), \(\sigma_1 \geq \cdots \geq \sigma_m \geq 0\), and \(U, V\) orthogonal.

- Special case: \(W = V^T\). Then for \(z = Wu\), \(c = U^Tb\), consider
  \[
  \min_z \frac{1}{2} \|\Sigma z - c\|^2 + \beta \|z\|_1
  \]

- Representation for true model \(z^* = Wu^*\) satisfies \(\sigma_i z_i^* = c_i + \varepsilon_i\). Consider a sparse representation
  \[
  z_i^* = 1 \text{ if } i \in T, \quad = 0 \text{ if } i \notin T.
  \]

Under what conditions can we expect to stably calculate \(z\) with the same sparsity (i.e. \(z_i = 0 \text{ iff } i \notin T\))?
**Simple analysis for L1**

- Assume noise with mean $0$ and covariance $\rho^2 I$, and define

\[
\sigma_+ = \max_{i \notin T} \sigma_i, \quad \sigma_- = \min_{i \in T} \sigma_i.
\]

- **Theorem:**
  Using L1, the true and reconstructed models, $z^*$ and $z$, are expected to have the same zero structure only if either $\sigma_+ \leq \sigma_-$ or

\[
\rho \leq \frac{\sigma_-^2}{\sqrt{\sigma_+^2 - \sigma_-^2}}.
\]

  (Can’t get the sparse computed $z$ stably for just any sparse $z^*$.)

- Further difficulties arise in the latter case in determining the regularization parameter $\beta$. 
Simple analysis for L1

- Assume noise with mean 0 and covariance $\rho^2 I$, and define

$$
\sigma_+ = \max_{i \notin T} \sigma_i, \quad \sigma_- = \min_{i \in T} \sigma_i.
$$

- **Theorem:**

Using L1, the true and reconstructed models, $z^*$ and $z$, are expected to have the same zero structure only if either $\sigma_+ \leq \sigma_-$ or

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

(Can’t get the sparse computed $z$ stably for just any sparse $z^*$.)

- Further difficulties arise in the latter case in determining the regularization parameter $\beta$. 

---

Kees van den Doel & Eldad Haber (UBC)
Simple analysis for L1

- Assume noise with mean 0 and covariance $\rho^2 I$, and define
  
  $$
  \sigma_+ = \max_{i \notin T} \sigma_i, \quad \sigma_- = \min_{i \in T} \sigma_i.
  $$

- **Theorem:**
  Using L1, the true and reconstructed models, $z^*$ and $z$, are expected to have the same zero structure only if either $\sigma_+ \leq \sigma_-$ or
  
  $$
  \rho \leq \frac{\sigma_-^2}{\sqrt{\sigma_+^2 - \sigma_-^2}}.
  $$

  (Can’t get the sparse computed $z$ stably for just any sparse $z^*$.)

- Further difficulties arise in the latter case in determining the regularization parameter $\beta$. 

Nonlinear Inverse Problems

Nonlinear objective function and $\ell_1$ constraints.

Note: when the constraint (solid red) is non-linear, it does not necessarily intersect the level set of $\|u\|_1$ at a vertex, so the solution is not necessarily sparse.

Still, L1G can be very useful due to less smearing across jumps!
Nonlinear objective function and $\ell_1$ constraints.

- Note: when the constraint (solid red) is non-linear, it does not necessarily intersect the level set of $\|u\|_1$ at a vertex, so the solution is not necessarily sparse.
- Still, $\ell_1$G can be very useful due to less smearing across jumps!
Nonlinear inverse problems

Nonlinear objective function and $\ell_1$ constraints.

- Note: when the constraint (solid red) is non-linear, it does not necessarily intersect the level set of $\|u\|_1$ at a vertex, so the solution is not necessarily sparse.
- Still, L1G can be very useful due to less smearing across jumps!
EIT and DC resistivity

[Haber, Heldman & Ascher, 2007; van den Doel & Ascher, 2011; Haber, Chung & Herrmann, 2011; Roosta-Khorasani, van den Doel & Ascher, 2012]

- **Forward problem**:

\[ \nabla \cdot (\sigma \nabla v^i) = q^i, \quad x \in \Omega, \quad i = 1, \ldots, s, \]
\[ \frac{\partial v^i}{\partial \nu}|_{\partial \Omega} = 0. \]

- Take \( \Omega \) to be a square. Construct \( s \) data sets, choosing different current patterns:

\[ q^i(x) = \delta_{x, p^i_L} - \delta_{x, p^i_R}, \]

where \( p^i_L \) and \( p^i_R \) are located on the left and right boundaries resp. Place each at \( \sqrt{s} \) different, equidistant locations, \( 1 \leq i_L, i_R \leq \sqrt{s} \).

- Predict data by measuring field at specified locations:

\[ F(u) = (F_1(u), \ldots, F_s(u))^T. \]
Often we have *a priori information* that $\sigma_{\text{min}} \leq \sigma(x) \leq \sigma_{\text{max}}$.

Predict data by measuring field at specified locations:

$$F(u) = (F_1(u), \ldots, F_s(u))^T,$$

where

$$u(x) = \psi^{-1}(\sigma(x)),$$

with transfer function

$$\psi(t) = .5(\sigma_{\text{max}} - \sigma_{\text{min}}) \tanh(t) + .5(\sigma_{\text{max}} + \sigma_{\text{min}}).$$

In the following experiments use

$$\sigma_{\text{min}} = 1, \quad \sigma_{\text{max}} = 10,$$

and calculate on a $64 \times 64$ grid.
Often we have a priori information that $\sigma_{\text{min}} \leq \sigma(x) \leq \sigma_{\text{max}}$.

Predict data by measuring field at specified locations:

$$F(u) = (F_1(u), \ldots, F_s(u))^T,$$
where

$$u(x) = \psi^{-1}(\sigma(x)),$$
with transfer function

$$\psi(t) = .5(\sigma_{\text{max}} - \sigma_{\text{min}}) \tanh(t) + .5(\sigma_{\text{max}} + \sigma_{\text{min}}).$$

In the following experiments use

$$\sigma_{\text{min}} = 1, \quad \sigma_{\text{max}} = 10,$$
and calculate on a $64 \times 64$ grid.
Note: having many experiments here helps, both theoretically and practically! Thus, may want $s$ large, although this brings major computational difficulties.

- Use stochastic methods for model reduction in case that $s$ is large.
- Synthesize experiment’s data by using “true” $u(x)$, calculating data on twice-as-fine grid, adding 3% Gaussian noise.
- Generalized Tikhonov:

$$\min_u \frac{1}{2} \|F(u) - b\|^2 + \beta R(u)$$
L2G vs L1G: results

Left: true. Center: $s = 4$, L2G. Right: $s = 4$, L1G.

Left: true. Center: $s = 64$, L2G. Right: $s = 64$, L1G.
Decrease noise to 1%, increase $\sqrt{s}$.

Left: $s = 1024$, L2G. Right: $s = 1024$, L1G.
Indeed, better results for larger $s$.

For $s = 64$ the L1G results are somewhat better than L2G.

Upon increasing $s$ further, and using finer grids, and having less noise, L1G becomes significantly better than L2G.

Expensive computing may be encountered when trying to obtain more accurate solutions. Even more so in 3D. Adaptive grids (meshes) and random sampling of right hand side combinations help decrease the pain significantly!
L2G vs L1G results: observations

- Indeed, better results for larger $s$.
- For $s = 64$ the L1G results are somewhat better than L2G.
- Upon increasing $s$ further, and using finer grids, and having less noise, L1G becomes significantly better than L2G.
- Expensive computing may be encountered when trying to obtain more accurate solutions. Even more so in 3D. Adaptive grids (meshes) and random sampling of right hand side combinations help decrease the pain significantly!
L2G vs L1G results: observations

- Indeed, better results for larger $s$.
- For $s = 64$ the L1G results are somewhat better than L2G.
- Upon increasing $s$ further, and using finer grids, and having less noise, L1G becomes significantly better than L2G.
- Expensive computing may be encountered when trying to obtain more accurate solutions. Even more so in 3D. Adaptive grids (meshes) and random sampling of right hand side combinations help decrease the pain significantly!
Yes, L1G, aka total variation, is worth considering!

although, it does not always deliver and it’s often not cheap to work with.

Bigger doubts linger regarding L1 (when relevant), not only because of the possibly prohibitive cost:

For the inverse potential problem, can show based on physical grounds that the $\gamma$-criterion of Juditsky & Nemirovski is violated.

Furthermore, our numerical results for the inverse potential problem do not show positive evidence.

Can show that in some situations, using L1 can be highly restrictive and ineffective, regardless of cost.
Yes, $L1G$, aka total variation, is worth considering!

although, it does not always deliver and it’s often not cheap to work with.

Bigger doubts linger regarding $L1$ (when relevant), not only because of the possibly prohibitive cost:

- For the inverse potential problem, can show based on physical grounds that the $\gamma$-criterion of Juditsky & Nemirovski is violated.
- Furthermore, our numerical results for the inverse potential problem do not show positive evidence.
- Can show that in some situations, using $L1$ can be highly restrictive and ineffective, regardless of cost.
**$\ell_1$-BASED REGULARIZATION FOR LARGE, HIGHLY ILL-CONDITIONED PROBLEMS?**

- Yes, **L1G**, aka total variation, is worth considering!
- although, it does not always deliver and it’s often not cheap to work with.
- Bigger doubts linger regarding **L1** (when relevant), not only because of the possibly prohibitive cost:
  - For the inverse potential problem, can show based on physical grounds that the $\gamma$-criterion of Juditsky & Nemirovski is violated.
  - Furthermore, our numerical results for the inverse potential problem do not show positive evidence.
  - Can show that in some situations, using **L1** can be highly restrictive and ineffective, regardless of cost.
Conclusions

- In many situations, $\ell_1$-based regularization is well-worth using. Such techniques can provide exciting advances (e.g. in model reduction).

- However, such techniques are not suitable for all problems, and it is dangerous (and may consume many student-years) to apply them blindly.

- In practice, always consider first using $\ell_2$-based regularization, because they are simpler and more robust. Only upon deciding that these are not sufficiently good for the given application, proceed to examine $\ell_1$-based alternatives (when this makes sense).
Conclusions

- In many situations, $\ell_1$-based regularization is well-worth using. Such techniques can provide exciting advances (e.g. in model reduction).
- However, such techniques are not suitable for all problems, and it is dangerous (and may consume many student-years) to apply them blindly.
- In practice, always consider first using $\ell_2$-based regularization, because they are simpler and more robust. Only upon deciding that these are not sufficiently good for the given application, proceed to examine $\ell_1$-based alternatives (when this makes sense).
Conclusions

- In many situations, $\ell_1$-based regularization is well-worth using. Such techniques can provide exciting advances (e.g. in model reduction).
- However, such techniques are not suitable for all problems, and it is dangerous (and may consume many student-years) to apply them blindly.
- In practice, always consider first using $\ell_2$-based regularization, because they are simpler and more robust. Only upon deciding that these are not sufficiently good for the given application, proceed to examine $\ell_1$-based alternatives (when this makes sense).