Sparse Regularization via Convex Analysis

Ivan Selesnick

Electrical and Computer Engineering
Tandon School of Engineering
New York University
Brooklyn, New York, USA

2019
Convex or non-convex: Which is better for inverse problems?

Benefits of convex optimization:

1. Absence of suboptimal local minima
2. Continuity of solution as a function of input data
3. Algorithms guaranteed to converge to a global optimum
4. Regularization parameters easier to set

But convex regularization tends to *under-estimate* signal values.

Non-convex regularization often performs better!

Can we design non-convex sparsity-inducing penalties that maintain the convexity of the cost function to be minimized?
Convex function
Non-Convex function
Goal

Goal: Find a sparse approximate solution to a linear system $y = Ax$.

Minimize a cost function:

$$J(x) = \frac{1}{2} \| y - Ax \|_2^2 + \lambda \| x \|_1$$

or

$$F(x) = \frac{1}{2} \| y - Ax \|_2^2 + \lambda \psi(x)$$

Question: How to define $\psi$?

Let us allow $\psi$ to be non-convex such that $F$ is convex.

This is the \textit{Convex Non-Convex} (CNC) approach.
Linear Filter

Given noisy data \( y \in \mathbb{R}^N \), perform smoothing via:

\[
\hat{x} = \arg \min_{x \in \mathbb{R}^N} \left\{ \sum_{n=0}^{N-1} |y(n) - x(n)|^2 + \lambda \sum_{n=1}^{N-1} |x(n) - x(n-1)|^2 \right\}
\]

which can be written

\[
\hat{x} = \arg \min_{x \in \mathbb{R}^N} \left\{ \|y - x\|_2^2 + \lambda \|Dx\|_2^2 \right\}
\]

where

\[
\|x\|_2^2 := \sum_n |x(n)|^2
\]

\[
D = \begin{bmatrix}
-1 & 1 \\
-1 & 1 & \ddots \\
& & & & & -1 \\
& & & & & 1
\end{bmatrix}
\]

Solution:

\[
\hat{x} = (I + \lambda D^T D)^{-1} y
\]
Total Variation Denoising (Nonlinear Filter)

Given noisy data $y \in \mathbb{R}^N$, perform smoothing via:

$$
\hat{x} = \arg \min_{x \in \mathbb{R}^N} \left\{ \frac{1}{2} \sum_{n=0}^{N-1} |y(n) - x(n)|^2 + \lambda \sum_{n=1}^{N-1} |x(n) - x(n-1)| \right\}
$$

which can be written

$$
\hat{x} = \arg \min_{x \in \mathbb{R}^N} \left\{ \frac{1}{2} \|y - x\|_2^2 + \lambda \|Dx\|_1 \right\}
$$

where

$$
\|x\|_2^2 := \sum_n |x(n)|^2, \quad \|x\|_1 := \sum_n |x(n)|
$$

$$
D = \begin{bmatrix}
-1 & 1 & & & \\
& -1 & 1 & & \\
& & \ddots & \ddots & \\
& & & -1 & 1
\end{bmatrix}
$$

Solution? No closed form solution.

Use iterative algorithm … but note the cost function is not differentiable!
Biosensor Signal

Biosensor data

Total variation denoising
Two Penalties

The $\ell_1$ norm induces sparsity unlike the sum of squares.

\[ f(x) = x \]
\[ f(x) = \sin x \]
\[ f(x) = \frac{1}{20} \exp x \]
Combine Quadratic and Sparse Regularization

\[
\arg\min_{u,v \in \mathbb{R}^N} \left\{ \frac{1}{2} \| y - u - v \|_2^2 + \lambda_1 \| Du \|_1 + \frac{\lambda_2}{2} \| Dv \|_2^2 \right\}
\]

\[\hat{x} = u + v\]
Combine Quadratic and Sparse Regularization

\[
\arg \min_{u,v \in \mathbb{R}^N} \left\{ \frac{1}{2} \| y - u - v \|_2^2 + \lambda_1 \| Du \|_1 + \frac{\lambda_2}{2} \| Dv \|_2^2 \right\}
\]

Solving for \( v \) gives

\[
v = (I + \lambda_2 D^T D)^{-1} (y - u)
\]

\[
x = v + u = (I + \lambda_2 D^T D)^{-1} (y + \lambda_2 D^T Du)
\]

Substituting \( v \) back in to the cost function:

\[
J(u) = \frac{\lambda_2}{2} (y - u)^T D^T (I + \lambda_2 DD^T)^{-1} D(y - u) + \lambda_1 \| Du \|_1
\]

or

\[
J(u) = \frac{\lambda_2}{2} \| R^{-1} D(y - u) \|_2^2 + \lambda_1 \| Du \|_1
\]

\[
RR^T = I + \lambda_2 DD^T \quad (R \text{ is a banded matrix})
\]

Since \( x \) depends on \( Du \), not \( u \) directly, define \( g = Du \). So we need to minimize

\[
F(g) = \frac{\lambda_2}{2} \| R^{-1} Dy - R^{-1} g \|_2^2 + \lambda_1 \| g \|_1
\]
Scalar case

\[ \hat{x} = \arg \min_x \left\{ \frac{1}{2} (y - x)^2 + \lambda |x| \right\} \]

\[ \Rightarrow \]

\[ \hat{x} \]

\[ \lambda \]

\[ 0 \]

\[ y \]

\[ 0 \]
Non-convex scalar penalty functions (alternatives to $\ell_1$ norm)

<table>
<thead>
<tr>
<th>Penalty</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log</td>
<td>$\phi_a(x) = \frac{1}{a} \log(1 + a</td>
</tr>
<tr>
<td>Rat</td>
<td>$\phi_a(x) = \frac{</td>
</tr>
<tr>
<td>Exp</td>
<td>$\phi_a(x) = \frac{1}{a} \left(1 - e^{-a</td>
</tr>
</tbody>
</table>
| MC      | $\phi_a(x) = \begin{cases} 
|x| - \frac{a}{2}x^2, & |x| \leq 1/a \\
\frac{1}{2a}, & |x| \geq 1/a 
\end{cases}$ |

The penalties are parameterized such that

$$
\phi_a'(0^+) = 1
$$

$$
\phi_a''(0^+) = -a.
$$
Non-convex scalar penalty functions

Penalty functions with $a = 1$. 

Penalty functions

$$|x|$$

Log
Rat
Exp
MC
Logarithmic penalty

Logarithmic penalty (a = 1)

Second derivative

x
MC penalty

MC penalty (a = 1)

Second derivative
$\ell_p$ penalty, $0 < p < 1$ (precluded)
We consider henceforth only the minimax-concave (MC) penalty function

\[ \phi_a(x) = \begin{cases} 
|x| - \frac{a^2}{2}x^2, & |x| \leq 1/a \\
\frac{1}{2a}, & |x| \geq 1/a 
\end{cases} \]
The parameter $a \geq 0$ controls the non-convexity of $\phi_a$. 
Scalar case using MC penalty

\[ \hat{x} = \arg \min_x \left\{ \frac{1}{2} (y - x)^2 + \lambda \phi_a(x) \right\} \]

\( \hat{x} \) is a continuous function of \( y \) when \( a < \lambda \).
Scalar case using MC penalty

Consider

\[ f(x) = \frac{1}{2} (y - x)^2 + \lambda \phi_a(x). \]

For what values ‘a’ is \( f \) a convex function?

\[ f(x) = \frac{1}{2} x^2 + \lambda \phi_a(x) + \left[ \frac{1}{2} y^2 - yx \right]. \]

It is sufficient to consider the convexity of

\[ f_0(x) = \frac{1}{2} x^2 + \lambda \phi_a(x). \]
Scalar case using MC penalty

\[ f_0(x) = \frac{1}{2}x^2 + \lambda \phi_a(x) \]

Is \(f_0\) convex?

\(\phi_a\) is not differentiable. So we can not simply check that the second derivative of \(f_0\) is positive . . .
Scalar case using MC penalty

Let us write

$$\phi_a(x) = |x| - s_a(x)$$

We see the Huber function:

$$s_a(x) = \begin{cases} \frac{a^2 x^2}{2}, & |x| \leq 1/a \\ |x| - \frac{1}{2a}, & |x| \geq 1/a. \end{cases}$$
Scalar case using MC penalty

Writing $\phi_a$ as

$$\phi_a(x) = |x| - s_a(x),$$

we have

$$f(x) = \frac{1}{2} (y - x)^2 + \lambda \phi_a(x)$$

$$= \frac{1}{2} (y - x)^2 + \lambda \left[|x| - s_a(x)\right]$$

$$= \frac{1}{2} x^2 - \lambda s_a(x) + \left[\lambda |x| + \frac{1}{2} y^2 - xy\right]$$

$$g(x) \text{ convex in } x$$

$g$ convex $\implies f$ convex

Note that $g$ is differentiable unlike $f$.

Is $g$ convex? It depends on $a$ and $\lambda$. 
Scalar case using MC penalty

\[ g(x) = \frac{1}{2} x^2 - \lambda s_a(x) \]

\[ g(x) = 0.5 x^2 - \lambda s_a(x) \]

\[ a = 1, \lambda = 0.8, g \text{ is convex.} \]
Scalar case using MC penalty

\[ g(x) = \frac{1}{2}x^2 - \lambda s_a(x) \]

\[ g(x) = 0.5x^2 - \lambda s_a(x) \]

\[ a = 1, \lambda = 1.3, \text{ } g \text{ is not convex.} \]
Scalar case using MC penalty

The Huber function is differentiable. But not twice differentiable.

\[ s(x) \]

\[ s'(x) \]

\[ s''(x) \]
Scalar case using MC penalty

\[ g(x) = \frac{1}{2}x^2 - \lambda s_a(x) \]

When is \( g \) convex?

We can not check the second derivative of \( g \) because it is not twice differentiable (see previous page).

How can we ensure \( g \) (and hence \( f \)) is convex?
The Huber function can be written as

$$s_a(x) = \min_v \left\{ \frac{a}{2}(x - v)^2 + |v| \right\}.$$  

As infimal convolution

$$s_a(x) = \left( \frac{a}{2} (\cdot)^2 \square |\cdot| \right)(x)$$  

where infimal convolution (Moreau-Yosida regularization) is defined as

$$(f \square g)(x) := \min_v \{ f(v) + g(x - v) \}$$
Huber function as an infimal convolution

\[ \left( \frac{a}{2} (\cdot)^2 \square |\cdot| \right)(x) \]

\[ g(x) = |x| \]

\[ 0.5x^2 \]
The epigraph of a function is a set comprising points on and above the graph.

The epigraph of

\[ epi\{f \Box g\} = epi\{f\} + epi\{g\} \]
Scalar case using MC penalty

The Huber function can be written as

\[ s_a(x) = \min_v \left\{ \frac{a}{2} (x - v)^2 + |v| \right\} . \]

When is \( g \) convex?

\[
g(x) = \frac{1}{2} x^2 - \lambda s_a(x)
\]

\[
= \frac{1}{2} x^2 - \lambda \min_v \left\{ \frac{a}{2} (x - v)^2 + |v| \right\}
\]

\[
= \frac{1}{2} x^2 - \lambda \min_v \left\{ \frac{a}{2} (x^2 - 2xv + v^2) + |v| \right\}
\]

\[
= \frac{1}{2} x^2 - \lambda \frac{a}{2} x^2 + \lambda \max_v \left\{ \frac{a}{2} (2xv - v^2) - |v| \right\}
\]

affine in \( x \)  

convex in \( x \)

\[
= \frac{1}{2} (1 - a\lambda) x^2 + \text{convex function}
\]

The function \( g \) is \textit{convex} if \( 1 - a\lambda \) is non-negative, i.e.,

\[ a \leq 1/\lambda \]

We do \textit{not} need derivatives!
Scalar MC penalty

The MC penalty can be written as

\[ \phi_a(x) = |x| - s_a(x) \]

\[ = |x| - \min_v \left\{ \frac{a}{2} (x - v)^2 + |v| \right\} \]

\[ = |x| - \left( \frac{a}{2} (\cdot)^2 \boxplus |\cdot| \right)(x) \]
Multivariate case

\[ F(x) = \frac{1}{2} \| y - Ax \|_2^2 + \lambda \psi(x) \]

How can we set \( \psi \) so that \( F \) is convex and promotes sparsity of \( x \)?

We can generalize the scalar case . . .
Separable penalty (*precluded*)

Conventional penalty: additive (separable)

\[ \phi(x) = \phi(x_1) + \phi(x_2) \]
Generalized Huber function

Let $B \in \mathbb{R}^{M \times N}$. We define the \textit{generalized Huber function}

$$S_B(x) := \min_{v \in \mathbb{R}^N} \left\{ \frac{1}{2} \| B(x - v) \|_2^2 + \| v \|_1 \right\}.$$ 

In the notation of infimal convolution, we have

$$S_B(x) := \left( \frac{1}{2} \| B \cdot \|_2^2 \square \| \cdot \|_1 \right)(x).$$
Example 1. Generalized Huber function

\[ B = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{bmatrix} \]

\[ S_B(x) := \min_v \left\{ \frac{1}{2} \| B(x - v) \|_2^2 + \| v \|_1 \right\} \]
Example 2. Generalized Huber function

\[ B = \begin{bmatrix} 1 & 0.5 \end{bmatrix} \]

\[ S_B(x) := \min_v \left\{ \frac{1}{2} \| B(x - v) \|_2^2 + \| v \|_1 \right\} \]
Example 3. Generalized Huber function

\[ B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \]

\[ S_B(x) := \min_v \left\{ \frac{1}{2} \| B(x - v) \|_2^2 + \| v \|_1 \right\} \]

If \( B \) is diagonal, then \( S_B \) is separable.
The generalized Huber function is differentiable.

Its gradient is given by

\[ \nabla S_B(x) = B^T B \left( x - \arg \min_{v \in \mathbb{R}^N} \left\{ \frac{1}{2} \| B(x - v) \|_2^2 + \| v \|_1 \right\} \right) \]

Neither the generalized Huber function nor its gradient have simple closed form expressions. But we will still be able to use them . . .

When \( B = I \) we recover a well known identity

\[ \nabla S_I(x) = x - \arg \min_{v \in \mathbb{R}^N} \left\{ \frac{1}{2} \| x - v \|_2^2 + \| v \|_1 \right\} \]
We define the \textit{generalized MC (GMC)} penalty

\[
\psi_B(x) := \|x\|_1 - S_B(x) \\
:= \|x\|_1 - \min_{v \in \mathbb{R}^N} \left\{ \frac{1}{2} \|B(x - v)\|_2^2 + \|v\|_1 \right\}.
\]
Example 1. Generalized MC penalty

\[ \psi_B(x) = \|x\|_1 - S_B(x) \]
Example 2. Generalized MC penalty

\[ \psi_B(x) = \|x\|_1 - S_B(x) \]
Example 3. Generalized MC penalty

\[ \psi_B(x) = ||x||_1 - S_B(x) \]
Theorem. The function

\[ F(x) = \frac{1}{2} \|y - Ax\|^2_2 + \lambda \psi_B(x) \]

\[ = \frac{1}{2} \|y - Ax\|^2_2 + \lambda \left[ \|x\|_1 - S_B(x) \right] \]

\[ = \frac{1}{2} \|y - Ax\|^2_2 - \lambda \|x\|_1 - \lambda \min_{v \in \mathbb{R}^N} \left\{ \frac{1}{2} \|B(x - v)\|^2_2 + \|v\|_1 \right\}. \]

is *convex* if

\[
B^T B \preceq \frac{1}{\lambda} A^T A
\]

even when \( \psi_B \) is *non-convex.*
Convexity condition – proof

Write $F$ as

$$F(x) = \frac{1}{2} \| y - Ax \|_2^2 + \lambda \psi_B(x)$$

$$= \frac{1}{2} \| y - Ax \|_2^2 + \lambda [\| x \|_1 - S_B(x)]$$

$$= \left[ \frac{1}{2} \| Ax \|_2^2 + \frac{1}{2} \| y \|_2^2 - y^T Ax \right] + \lambda \| x \|_1 - \lambda S_B(x)$$

$$= \frac{1}{2} \| Ax \|_2^2 - \lambda S_B(x) + \frac{1}{2} \| y \|_2^2 - y^T Ax + \lambda \| x \|_1$$

$$\underbrace{G(x)}_{\text{convex in } x} + \underbrace{\frac{1}{2} \| y \|_2^2 - y^T Ax + \lambda \| x \|_1}_{\text{convex in } x}$$

$G$ convex $\implies$ $F$ convex
Convexity condition – proof

Write $G$ as

$$G(x) = \frac{1}{2} \|Ax\|_2^2 - \lambda S_B(x)$$

$$= \frac{1}{2} \|Ax\|_2^2 - \lambda \min_v \left\{ \frac{1}{2} \|B(x - v)\|_2^2 + \|v\|_1 \right\}$$

$$= \frac{1}{2} \|Ax\|_2^2 - \lambda \min_v \left\{ \frac{1}{2} \|Bx\|_2^2 + \frac{1}{2} \|Bv\|_2^2 - v^T B^T B x + \|v\|_1 \right\}$$

$$= \frac{1}{2} \|Ax\|_2^2 - \frac{1}{2} \|Bx\|_2^2 + \lambda \max_v \left\{ v^T B^T B x - \frac{1}{2} \|Bv\|_2^2 - \|v\|_1 \right\}$$

affine in $x$

convex in $x$

$$= \frac{1}{2} x^T (A^T A - \lambda B^T B) x + \text{ convex function}$$

The function $G$ (hence $F$) is *convex* if $A^T A - \lambda B^T B$ is positive semidefinite, i.e.

$$B^T B \preceq \frac{1}{\lambda} A^T A.$$
Convexity condition

A straightforward choice of $B$ to satisfy

$$B^T B \preceq \frac{1}{\lambda} A^T A.$$  

is

$$B = \sqrt{\frac{\gamma}{\lambda}} A$$

for some $\gamma$ with $0 \leq \gamma \leq 1$. 
Optimization Algorithm

We can use the Forward-Backward Splitting (FBS) algorithm

\[
F(x) = \frac{1}{2} \|y - Ax\|^2 + \lambda \psi_B(x)
\]

\[
= \frac{1}{2} \|y - Ax\|^2 + \lambda \|x\|_1 - \lambda S_B(x)
\]

\[
= \frac{1}{2} \|y - Ax\|^2 - \lambda S_B(x) + \lambda \|x\|_1
\]

\[
= f_1(x) + f_2(x)
\]

The FBS algorithm is given by:

\[
w^{(i)} = x^{(i)} - \mu [\nabla f_1(x^{(i)})]
\]

\[
x^{(i+1)} = \arg\min_x \left\{ \frac{1}{2} \|w^{(i)} - x\|^2 + \mu f_2(x) \right\}
\]

\[
= \text{prox}_{\mu f_2}(w^{(i)})
\]
Optimization Algorithm

Let $0 \leq \gamma \leq 1$ and $B = \sqrt{\gamma/\lambda} A$. Then a minimizer of the objective function $F$ is obtained by the iteration:

$v^{(i)} = \arg \min_{v \in \mathbb{R}^N} \left\{ \frac{\gamma}{2} \| A(x^{(i)} - v) \|_2^2 + \lambda \| v \|_1 \right\}$

$z^{(i)} = \gamma A(x^{(i)} - v^{(i)})$

$x^{(i+1)} = \arg \min_{x \in \mathbb{R}^N} \left\{ \frac{1}{2} \| y + z^{(i)} - Ax \|_2^2 + \lambda \| x \|_1 \right\}$

*Interpretation:* iteratively adjusted additive data perturbation of $\ell_1$ norm regularized problem ...
Biosensor Signal

Biosensor data

Time (n)

0 100 200 300 400 500 600 700 800 900 1000
0
50
100
150

SIPS Denoising

Time (n)

0 100 200 300 400 500 600 700 800 900 1000
0
50
100
150
ECG Signal
ECG Signal

ECG data

SIPS Denoising
ECG Signal

ECG data

SIPS Denoising (non-separable non-convex penalty)
Consider the function

\[
J(x) := \frac{1}{2} \|Ax - b\|^2_2 + \lambda R(x)
\]

where \( R \) is a convex regularizer of the form

\[
R(x) = \Phi(G(Lx))
\]

where \( L \) is a linear operator, \( G \) is possibly nonlinear, and \( \Phi \) promotes sparsity.

For example, total variation regularization is expressed as

\[
R(x) = \|G(Lx)\|_1 = \sum_i |g_i(Lx)|
\]

with

\[
g_i(Lx) = \begin{cases} 
(D_h x)_i & \text{anisotropic TV} \\
\sqrt{(D_h x)_i^2 + (D_v x)_i^2} & \text{isotropic TV}
\end{cases}
\]
Generalizations (TV, nuclear norm, etc)

Suppose \( R \) satisfies

A1) \( R(\cdot) = \Phi(G(L \cdot)) \) is convex and bounded from below by zero;

A2) \( \Phi(G(\cdot)) \) is a proper, lower semicontinuous and coercive function.

Consider the function

\[
J_B(x) := \frac{1}{2} \|Ax - b\|_2^2 + \lambda R_B(x)
\]  

(2)

with non-convex regularizer

\[
R_B(x) := R(x) - \min_{v \in \mathbb{R}^N} \left\{ \frac{1}{2} \|B(x - v)\|_2^2 + R(v) \right\}.
\]

The function \( J_B(x) \) is convex (strongly convex) if

\[
A^T A - \lambda B^T B \succeq 0 \quad (> 0).
\]

Numerical Examples – 1D Total Variation denoising

Summary

▶ We show how to construct non-convex regularizers that preserve the convexity of functionals for sparse-regularized linear least-squares.
▶ Generalizes the $\ell_1$ norm.
▶ Can be used in conjunction with other convex non-smooth regularizers (TV, nuclear norm, etc).
▶ Resulting regularizers are non-separable.
▶ Optimization implementable using proximal algorithms (as for $\ell_1$ norm).


Optimization Algorithm (saddle point algorithm)

With

\[ J_B(x) = \frac{1}{2} \| Ax - b \|^2_2 + \lambda R_B(x) \]
\[ R_B(x) = R(x) - \min_v \left\{ \frac{1}{2} \| B(x - v) \|^2_2 + R(v) \right\} \]

we have

\[ J_B(x) = \frac{1}{2} \| Ax - b \|^2_2 + \lambda R(x) - \lambda \min_v \left\{ \frac{1}{2} \| B(x - v) \|^2_2 + R(v) \right\} \]
\[ = \frac{1}{2} \| Ax - b \|^2_2 + \lambda R(x) + \lambda \max_v \left\{ -\frac{1}{2} \| B(x - v) \|^2_2 - R(v) \right\} \]

and

\[ \hat{x} = \arg \min_x J_B(x) \]
\[ = \arg \min_x \left\{ \frac{1}{2} \| Ax - b \|^2_2 + \lambda R(x) + \lambda \max_v \left\{ -\frac{1}{2} \| B(x - v) \|^2_2 - R(v) \right\} \right\} \]

or a saddle point problem

\[ (\hat{x}, \hat{v}) = \arg \min_x \max_v \left\{ \frac{1}{2} \| Ax - b \|^2_2 + \lambda R(x) - \frac{1}{2} \| B(x - v) \|^2_2 - \lambda R(v) \right\} \]
Optimization Algorithm (saddle point algorithm)

This saddle-point problem is an instance of a monotone inclusion problem. Hence, the solution can be obtained using the forward-backward (FB) algorithm.

\begin{align*}
\text{for } k = 0, 1, 2, \ldots \\
  w_k &= x_k - \mu \left[ A^T (Ax_k - b) + \lambda B^T B (v_k - x_k) \right] \\
  u_k &= v_k - \mu \lambda B^T B (v_k - x_k) \\
  x_{k+1} &= \arg \min_{x \in \mathbb{R}^n} \left\{ \mathcal{R}(x) + \frac{1}{2\mu \lambda} \|x - w_k\|_2^2 \right\} \\
  v_{k+1} &= \arg \min_{v \in \mathbb{R}^n} \left\{ \mathcal{R}(v) + \frac{1}{2\mu \lambda} \|v - u_k\|_2^2 \right\}
\end{align*}

end

This algorithm reduces to ISTA for $B = 0$ and $\mathcal{R}(x) = \|x\|_1$. 