Iteratively Reweighted $\ell_1$ Approaches to Sparse Composite Regularization

Phil Schniter

THE OHIO STATE UNIVERSITY

Joint work with Rizwan Ahmad (OSU)

Supported in part by NSF grant CCF-1018368.

Outline

1. Introduction and Motivation for Composite Penalties
2. Co-L1 and its Interpretations
3. Co-IRW-L1 and its Interpretations
4. Numerical Experiments
Introduction

- **Goal**: Recover signal $\mathbf{x} \in \mathbb{C}^N$ from noisy linear measurements
  
  $$\mathbf{y} = \Phi \mathbf{x} + \mathbf{w} \in \mathbb{C}^M$$

  where possibly $M \ll N$.

- **Approach**: Solve optimization problem
  
  $$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} R(\mathbf{x}) \text{ s.t. } \|\mathbf{y} - \Phi \mathbf{x}\|_2 \leq \delta$$

  with $\delta$ selected based on statistics of $\|\mathbf{w}\|_2$.

- **Question**: How to choose penalty/regularization $R(\mathbf{x})$?
Introduction and Motivation for Composite Penalties

Typical Choices of Penalty

Suppose $\Psi x$ is (approximately) sparse for analysis operator $\Psi \in \mathbb{C}^{L \times N}$:

\[ R(x) = \| \Psi x \|_0 \]

- **$\ell_0$ penalty:** Impractical: optimization problem is NP hard

\[ R(x) = \| \Psi x \|_1 \]

- Tightest convex relaxation of $\ell_0$ penalty
- Fast algorithms: Douglas-Rachford, NESTA-UP, MFISTA, GAMP . . .

Many other penalties, such as $R(x) = \| \Psi x \|_p$ for $p \in (0, 1)$. 
Choice of Analysis Operator

How to choose $\Psi$ in practice?

- Maybe a wavelet dictionary? Which one?
- Maybe a concatenation of several dictionaries $\begin{bmatrix} \Psi_1 \\ \vdots \\ \Psi_D \end{bmatrix}$?
- What if signal is more sparse in one dictionary than another? Can we use this to our advantage?
Example: Undecimated Wavelet Transform of MRI Cine

Note different sparsity rate in each subband of 1-level UWT:
We propose to use composite $\ell_1$ penalties of the form

$$R(x; \lambda) \triangleq \sum_{d=1}^{D} \lambda_d \| \Psi_d x \|_1, \quad \lambda_d \geq 0$$

where

- operators $\Psi_d$ have unit-norm rows (but otherwise arbitrary),
- weights $\lambda_d$ are learned from the data.

We propose two algorithms to jointly estimate $x$ and $\lambda = [\lambda_1, \ldots, \lambda_D]^T$:

1. Composite-$\ell_1$ minimization (Co-L1)
2. Iteratively reweighted composite-$\ell_1$ minimization (Co-IRW-L1)
The Co-L1 Algorithm

1: input: $\{\Psi_d\}_{d=1}^D, \Phi, y, \delta \geq 0, \epsilon \geq 0$
2: initialization: $\lambda_d^{(1)} = 1 \ \forall d$
3: for $t = 1, 2, 3, \ldots$

4: $x^{(t)} \leftarrow \arg \min_x \sum_{d=1}^D \lambda_d^{(t)} \|\Psi_d x\|_1$ s.t. $\|y - \Phi x\|_2 \leq \delta$

5: $\lambda_d^{(t+1)} \leftarrow \frac{L_d}{\epsilon + \|\Psi_d x^{(t)}\|_1}$, $d = 1, \ldots, D$

6: end
7: output: $x^{(t)}$

- leverages existing $\ell_1$ solvers,
- applies to both real- and complex-valued cases,
- reduces to IRW-L1 algorithm [Candes,Wakin,Boyd’08] when $L_d = 1 \ \forall d$
  (single-atom dictionaries).
The Co-IRW-L1 Algorithm

1: input: \( \{\Psi_d\}_{d=1}^D, \Phi, y, \delta \geq 0, \)
2: if \( x \in \mathbb{R}^N \), use \( \Lambda = (1, \infty) \) and the real version of \( \log p(x; \lambda, \epsilon) \);  
   if \( x \in \mathbb{C}^N \), use \( \Lambda = (2, \infty) \) and the complex version of \( \log p(x; \lambda, \epsilon) \).
3: initialization: \( \lambda_d^{(1)} = 1 \ \forall d, \ W_d^{(1)} = I \ \forall d \)
4: for \( t = 1, 2, 3, \ldots \)
5: \( x^{(t)} \leftarrow \arg\min_x \sum_{d=1}^D \lambda_d^{(t)} \| W_d^{(t)} \Psi_d x \|_1 \) s.t. \( \| y - \Phi x \|_2 \leq \delta \)
6: \( (\lambda_d^{(t+1)}, \epsilon_d^{(t+1)}) \leftarrow \arg\max_{\lambda_d \in \Lambda, \epsilon_d > 0} \log p(x^{(t)}; \lambda, \epsilon), \ d = 1, \ldots, D \)
7: \( W_d^{(t+1)} \leftarrow \text{diag} \left\{ \frac{1}{\epsilon_d^{(t+1)} + |\psi_{d,1}^{T} x^{(t)}|}, \ldots, \frac{1}{\epsilon_d^{(t+1)} + |\psi_{d,L_d}^{T} x^{(t)}|} \right\}, \ d = 1, \ldots, D \)
8: end
9: output: \( x^{(t)} \)

- IRW version of Co-L1: tunes both \( \lambda_d \) and \( W_d \) for all \( d \).
- also tunes regularization parameters \( \epsilon_d \) for all \( d \).
Understanding Co-L1 and Co-IRW-L1

In the sequel, we provide four interpretations of each algorithm:

1. MM optimization of a particular non-convex penalty,
2. a particular approximation of $\ell_0$ minimization,
3. Bayesian estimation according to a particular hierarchical prior,
4. variational EM algorithm under a particular prior.
Outline

1. Introduction and Motivation for Composite Penalties
2. Co-L1 and its Interpretations
3. Co-IRW-L1 and its Interpretations
4. Numerical Experiments
Co-L1 is an MM approach to the **weighted log-sum** optimization problem

\[
\arg \min_x \sum_{d=1}^{D} L_d \log(\epsilon + \|\Psi_d x\|_1) \quad \text{s.t.} \quad \|y - \Phi x\|_2 \leq \delta.
\]

and

As \(\epsilon \to 0\), Co-L1 aims to solve the **weighted \(l_{1,0}\) problem**

\[
\arg \min_x \sum_{d=1}^{D} L_d \mathbb{1}_{\|\Psi_d x\|_1 > 0} \quad \text{s.t.} \quad \|y - \Phi x\|_2 \leq \delta.
\]

Note: \(L_d\) is the size of dictionary \(\Psi_d\), and \(\mathbb{1}_{\square}\) is the indicator function.
Bayesian Interpretations of Co-L1

As $\epsilon \to 0$, Co-L1 is an MM approach to Bayesian MAP estimation under an AWGN likelihood and the hierarchical prior

$$p(x|\lambda) = \prod_{d=1}^{D} \left( \frac{\lambda_d}{2} \right)^{L_d} \exp \left( -\lambda_d \| \Psi_d x \|_1 \right) \quad \text{i.i.d. Laplacian}$$

$$p(\lambda) = \prod_{d=1}^{D} p(\lambda_d), \quad p(\lambda_d) \propto \begin{cases} \frac{1}{\lambda_d} & \lambda_d > 0 \\ 0 & \text{else} \end{cases}, \quad \text{Jeffrey’s non-informative}$$

and

As $\epsilon \to 0$, Co-L1 is a variational EM approach to estimating (deterministic) $\lambda$ under an AWGN likelihood and the prior

$$p(x; \lambda) = \prod_{d=1}^{D} \left( \frac{\lambda_d}{2} \right)^{L_d} \exp \left( -\lambda_d \| \Psi_d x \|_1 \right) \quad \text{i.i.d. Laplacian}$$
Outline

1. Introduction and Motivation for Composite Penalties
2. Co-L1 and its Interpretations
3. Co-IRW-L1 and its Interpretations
4. Numerical Experiments
A Stepping Stone

The IRW version of real-valued Co-L1: tunes both inter-dictionary weights \( \lambda_d \) and intra-dictionary weights \( W_d \) for given parameters \( \epsilon_d \).

1: input: \( \{ \Psi_d \}_{d=1}^{D}, \Phi, y, \delta \geq 0, \epsilon_d > 0 \ \forall d \),
2: initialization: \( \lambda_d^{(1)} = 1 \ \forall d, \ W_d^{(1)} = I \ \forall d \)
3: for \( t = 1, 2, 3, \ldots \)
4: \( x^{(t)} \leftarrow \arg \min_{x} \sum_{d=1}^{D} \lambda_d^{(t)} \| W_d^{(t)} \Psi_d x \|_1 \) s.t. \( \| y - \Phi x \|_2 \leq \delta \)
5: \( \lambda_d^{(t+1)} \leftarrow \left[ \frac{1}{L_d} \sum_{l=1}^{L_d} \log \left( 1 + \frac{\| \psi_{d,l} x^{(t)} \|}{\epsilon_d} \right) \right]^{-1} + 1, \ d = 1, ..., D \)
6: \( W_d^{(t+1)} \leftarrow \text{diag} \left\{ \frac{1}{\epsilon_d + \| \psi_{d,1} x^{(t)} \|}, \ldots, \frac{1}{\epsilon_d + \| \psi_{d,L_d} x^{(t)} \|} \right\}, \ d = 1, ..., D \)
7: end
8: output: \( x^{(t)} \)
Real-Co-IRW-L1-\(\epsilon\) is an MM approach to the non-convex optimization
\[
\arg\min_{x} \sum_{d=1}^{D} \sum_{l=1}^{L_d} \log \left( (\epsilon_d + |\psi_{d,l}^T x|) \right) \sum_{i=1}^{L_d} \log \left( 1 + \frac{|\psi_{d,i}^T x|}{\epsilon_d} \right) \quad \text{s.t.} \quad \|y - \Phi x\|_2 \leq \delta
\]

and

As \(\epsilon \to 0\), real-Co-IRW-L1-\(\epsilon\) aims to solve the \(\ell_0+\) weighted \(\ell_{0,0}\) problem
\[
\arg\min_{x} \left[ \|\Psi x\|_0 + \sum_{d=1}^{D} L_d 1_{\|\Psi_d x\|_0 > 0} \right] \quad \text{s.t.} \quad \|y - \Phi x\|_2 \leq \delta.
\]

Note: \(L_d\) is the size of dictionary \(\Psi_d\), and \(1_{\square}\) is the indicator function.
Real-Co-IRW-L1 is an MM approach to Bayesian MAP estimation under an AWGN likelihood and the hierarchical prior

$$p(\mathbf{x} | \lambda) = \prod_{d=1}^{D} \prod_{l=1}^{L_d} \frac{\lambda_d}{2\epsilon_d} \left(1 + \frac{|\psi_{d,l}^\top \mathbf{x}|}{\epsilon_d}\right)^{-(\lambda_d+1)}$$

i.i.d. generalized-Pareto

$$p(\lambda) = \prod_{d=1}^{D} p(\lambda_d), \quad p(\lambda_d) \propto \begin{cases} \frac{1}{\lambda_d} & \lambda_d > 0 \\ 0 & \text{else} \end{cases}, \quad \text{Jeffrey's non-informative}$$

and

Real-Co-IRW-L1 is a variational EM approach to estimating (deterministic) $\lambda$ under an AWGN likelihood and the prior

$$p(\mathbf{x}; \lambda) = \prod_{d=1}^{D} \prod_{l=1}^{L_d} \frac{\lambda_d - 1}{2\epsilon_d} \left(1 + \frac{|\psi_{d,l}^\top \mathbf{x}|}{\epsilon_d}\right)^{-\lambda_d}$$

i.i.d. generalized-Pareto
Co-IRW-L1 Algorithm

Finally, we self-tune $\epsilon_d$ and allow for real or complex quantities:

1: input: $\{\Psi_d\}_{d=1}^D$, $\Phi$, $y$, $\delta \geq 0$,
2: if $x \in \mathbb{R}^N$, use $\Lambda = (1, \infty)$ and the real version of $\log p(x; \lambda, \epsilon)$;
   if $x \in \mathbb{C}^N$, use $\Lambda = (2, \infty)$ and the complex version of $\log p(x; \lambda, \epsilon)$.
3: initialization: $\lambda_d^{(1)} = 1 \forall d$, $W_d^{(1)} = I \forall d$
4: for $t = 1, 2, 3, \ldots$
5: $x^{(t)} \leftarrow \arg\min_x \sum_{d=1}^D \lambda_d^{(t)} \| W_d^{(t)} \Psi_d x \|_1$ s.t. $\| y - \Phi x \|_2 \leq \delta$
6: $(\lambda_d^{(t+1)}, \epsilon_d^{(t+1)}) \leftarrow \arg\max_{\lambda_d \in \Lambda, \epsilon_d > 0} \log p(x^{(t)}; \lambda, \epsilon)$, $d = 1, \ldots, D$
7: $W_d^{(t+1)} \leftarrow \text{diag} \left\{ \frac{1}{\epsilon_d^{(t+1)} + |\psi_{d,1}^{(t)}|^2}, \ldots, \frac{1}{\epsilon_d^{(t+1)} + |\psi_{d,L_d}^{(t)}|^2} \right\}$, $d = 1, \ldots, D$
8: end
9: output: $x^{(t)}$
Outline

1. Introduction and Motivation for Composite Penalties
2. Co-L1 and its Interpretations
3. Co-IRW-L1 and its Interpretations
4. Numerical Experiments
Numerical Experiments


declaration
declaration

Experiment: Synthetic finite difference image

\( \alpha = 1 \quad \alpha = 27 \)

- 48 x 48 image with a total of 28 horiz & vert transitions.
- \( \alpha \triangleq \frac{\text{# vertical transitions}}{\text{# horizontal transitions}} \)
- “spread-spectrum” \( \Phi \)
- sampling ratio \( \frac{M}{N} = 0.3 \)
- AWGN @ 30 dB SNR
- \( \Psi_1 = \text{vertical finite difference} \), \( \Psi_2 = \text{horizon. finite difference} \)

\Rightarrow The composite algorithms significantly outperform the non-composite ones

\Rightarrow Performance improves as sparsities become more disparate!
Numerical Experiments

Experiment: Shepp-Logan Phantom

- 96 × 96 image
- “spread-spectrum” $\Phi$
- AWGN @ 30 dB SNR
- $\Psi \in \mathbb{R}^{7N \times N} = 2\text{D UWT-db1}$, $\Psi_d \in \mathbb{R}^{N \times N} \forall d$

$\Rightarrow$ The composite algorithms significantly outperform the non-composite ones

$\Rightarrow$ Performance gap is larger for small $M/N$
Experiment: Cameraman

- 96 × 104 image
- “spread-spectrum” \( \Phi \)
- AWGN @ 40 dB SNR
- \( \Psi \in \mathbb{R}^{7N \times N} = 2D \) UWT-db1,
  \( \Psi_d \in \mathbb{R}^{N \times N} \ \forall d \)

\[ \Rightarrow \] The composite algorithms significantly outperform the non-composite ones

\[ \Rightarrow \] Performance gap is larger for small \( M/N \)
Experiment: 1D Dynamic MRI

- $144 \times 48$ spatiotemporal profile extracted from MRI cine
- $\Phi$: variable density random Fourier
- AWGN @ 30 dB SNR
- $\Psi \in \mathbb{R}^{3N \times N}$: 2D $[\text{db1;db2;db3}]$ DWT
Numerical Experiments

Experiment: 1D Dynamic MRI (cont.)

The composite algs significantly outperform the non-composite ones.

- Performance gap is larger for small $M/N$
- No advantage to Co-IRW-L1 over Co-L1 in this experiment
### Runtimes for Previous Experiments

<table>
<thead>
<tr>
<th></th>
<th>Shepp-Logan</th>
<th>Cameraman</th>
<th>dMRI</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>20.8s</td>
<td>23.1s</td>
<td>29.3s</td>
</tr>
<tr>
<td>Co-L1</td>
<td>32.7s</td>
<td>34.2s</td>
<td>86.4s</td>
</tr>
<tr>
<td>IRW-L1</td>
<td>45.9s</td>
<td>48.4s</td>
<td>54.1s</td>
</tr>
<tr>
<td>Co-IRW-L1</td>
<td>72.1s</td>
<td>96.4s</td>
<td>131s</td>
</tr>
</tbody>
</table>

The composite algs run 1.5–3× slower than the non-composite ones.
Conclusions

- We proposed a new “composite-L1” approach to L2-constrained signal reconstruction that learns and exploits differences in sparsity across sub-dictionaries.

- Relative to standard L1 methods, our composite L1 methods give significant improvements in reconstruction SNR at low sampling rates, at the cost of $1.5-3 \times$ slower runtimes.

- Our algorithms can be interpreted as MM approaches to non-convex optimization, approximate $\ell_0$ methods, Bayesian methods, and variational Bayesian methods.
Thanks!
Iteratively Reweighted $\ell_1$ (IRW-L1)

From [Candes, Wakin, Boyd, JFA’08] . . .

1: input: $\Psi = [\psi_1, \ldots, \psi_L]^T$, $\Phi$, $y$, $\delta \geq 0$, $\epsilon \geq 0$
2: initialization: $W^{(1)} = I$
3: for $t = 1, 2, 3, \ldots$
4: $x^{(t)} \leftarrow \arg\min_x \|W^{(t)}\Psi x\|_1$ s.t. $\|y - \Phi x\|_2 \leq \delta$
5: $W^{(t+1)} \leftarrow \text{diag} \left\{ \frac{1}{\epsilon + |\psi_1^T x^{(t)}|}, \ldots, \frac{1}{\epsilon + |\psi_L^T x^{(t)}|} \right\}$
6: end
7: output: $x^{(t)}$

- behaves more like $\ell_0$ minimization than $\ell_1$ minimization alone,
- leverages existing $\ell_1$ solvers.
IRW-L1 is an MM approach to the log-sum optimization problem
\[
\arg \min_x \sum_{l=1}^L \log(\epsilon + |\psi_l^T x|) \text{ s.t. } \|y - \Phi x\|_2 \leq \delta.
\]

How to see this? Reformulate as
\[
\arg \min_{x,u} \sum_l \log(\epsilon + u_l) \text{ s.t. } \begin{cases} 
\|y - \Phi x\|_2 \leq \delta \\
|\psi_l^T x| \leq u_l \forall l,
\end{cases}
\]
\[
\iff \arg \min_v g(v) \text{ s.t. } v \in C
\]
for \(v = \begin{bmatrix} u \\ x \end{bmatrix}\), convex \(C\), and concave \(g\).

**MM procedure**: Iterate for \(t = 1, 2, 3, \ldots\)

1. create surrogate \(g(v; \nu^{(t)})\) that majorizes \(g(v)\) at \(\nu^{(t)}\),
2. minimize the surrogate over \(v \in C\), producing \(\nu^{(t+1)}\).
Our concave $g(v)$ is majorized by the tangent at $v^{(t)}$. So MM becomes

$$v^{(t+1)} = \arg \min_{v \in \mathcal{C}} g(v^{(t)}) + \nabla g(v^{(t)})^T [v - v^{(t)}]$$

$$= \arg \min_{v \in \mathcal{C}} \nabla g(v^{(t)})^T v$$

$$\Leftrightarrow x^{(t+1)} = \arg \min_{x} \sum_{l} \frac{1}{\epsilon + |\psi_l^T x^{(t)}|} |\psi_l^T x| \text{ s.t. } ||y - \Phi x||_2 \leq \delta$$

Implications of MM:

- IRW-L1 convergence is guaranteed
- but possibly to a suboptimal local minimum (since non-convex).
Approximate-$\ell_0$ Interpretation of IRW-L1

\[ \sum_l \log(\epsilon + |u_l|) \]
\[ = \sum_l \log(1 + |u_l|/\epsilon) + \text{const} \]
\[ \propto \sum_l \frac{\log(1 + |u_l|/\epsilon)}{\log(1 + 1/\epsilon)} + \text{const} \]
\[ = \sum_l \lim_{p \to 0} \frac{1}{p} \left( (1 + |u_l|/\epsilon)^p - 1 \right) \]
\[ = \lim_{p \to 0} \sum_l \frac{(1 + |u_l|/\epsilon)^p - 1}{(1 + 1/\epsilon)^p - 1} + \text{const} \]
\[ \approx \lim_{p \to 0} \sum_l |u_l|^p + \text{const} \quad \text{(for } \epsilon \ll 1) \]
\[ = \|u\|_0 + \text{const} \]

$\Rightarrow$ As $\epsilon \to 0$, the log-sum penalty becomes a scaled and shifted version of the $\ell_0$ penalty.