Iteratively Reweighted ℓ_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.

SAHD @ Duke — July 27, 2015

Outline

Introduction

- Goal: Recover signal $\boldsymbol{x} \in \mathbb{C}^N$ from noisy linear measurements

$$\boldsymbol{y} = \boldsymbol{\Phi x} + \boldsymbol{w} \in \mathbb{C}^M$$

  where possibly $M \ll N$.

- Approach: Solve optimization problem

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} R(\boldsymbol{x}) \text{ s.t. } \|\boldsymbol{y} - \boldsymbol{\Phi x}\|_2 \leq \delta$$

  with δ selected based on statistics of $\|\boldsymbol{w}\|_2$.

- Question: How to choose penalty/regularization $R(\boldsymbol{x})$?

Typical Choices of Penalty

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

ℓ_0 penalty: $R(\boldsymbol{x}) = \|\boldsymbol{\Psi x}\|_0$

- Impractical: optimization problem is NP hard

ℓ_1 penalty (generalized LASSO): $R(\boldsymbol{x}) = \|\boldsymbol{\Psi x}\|_1$

- Tightest convex relaxation of ℓ_0 penalty
- Fast algorithms: Douglas-Rachford, NESTA-UP, MFISTA, GAMP ...

Many other penalties, such as $R(\boldsymbol{x}) = \|\boldsymbol{\Psi x}\|_p$ for $p \in (0, 1)$.

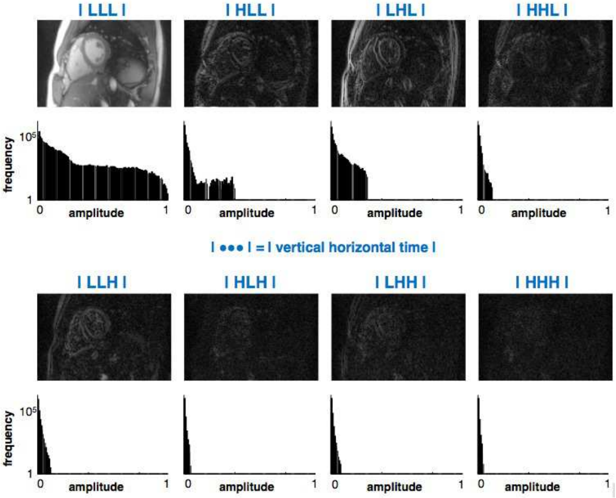
Choice of Analysis Operator

How to choose $\boldsymbol{\Psi}$ in practice?

- Maybe a wavelet dictionary? Which one?
- Maybe a concatenation of several dictionaries $\begin{bmatrix} \boldsymbol{\Psi}_1 \\ \vdots \\ \boldsymbol{\Psi}_D \end{bmatrix}$?
- 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:



Composite ℓ_1 Penalties

We propose to use composite ℓ_1 penalties of the form

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

where

- operators $\boldsymbol{\Psi}_d$ have unit-norm rows (but otherwise arbitrary),
- weights λ_d are learned from the data.

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

1. Composite-ℓ_1 minimization (Co-L1)
2. Iteratively reweighted composite-ℓ_1 minimization (Co-IRW-L1)

The Co-L1 Algorithm

1: input: $\{\boldsymbol{\Psi}_d\}_{d=1}^D$, $\boldsymbol{\Phi}$, $\boldsymbol{y}$, $\delta \geq 0$, $\epsilon \geq 0$
2: initialization: $\lambda_d^{(1)} = 1 \ \forall d$
3: for $t = 1, 2, 3, \dots$
4: $\quad \boldsymbol{x}^{(t)} \leftarrow \arg\min_{\boldsymbol{x}} \sum_{d=1}^{D} \lambda_d^{(t)} \|\boldsymbol{\Psi}_d \boldsymbol{x}\|_1$ s.t. $\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2 \leq \delta$
5: $\quad \lambda_d^{(t+1)} \leftarrow \dfrac{L_d}{\epsilon + \|\boldsymbol{\Psi}_d \boldsymbol{x}^{(t)}\|_1}, \quad d = 1, \dots, D$
6: end
7: output: $\boldsymbol{x}^{(t)}$

- leverages existing ℓ_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: $\{\boldsymbol{\Psi}_d\}_{d=1}^D$, $\boldsymbol{\Phi}$, $\boldsymbol{y}$, $\delta \geq 0$,
2: if $\boldsymbol{x} \in \mathbb{R}^N$, use $\Lambda = (1, \infty)$ and the real version of $\log p(\boldsymbol{x}; \boldsymbol{\lambda}, \boldsymbol{\epsilon})$;
if $\boldsymbol{x} \in \mathbb{C}^N$, use $\Lambda = (2, \infty)$ and the complex version of $\log p(\boldsymbol{x}; \boldsymbol{\lambda}, \boldsymbol{\epsilon})$.
3: initialization: $\lambda_d^{(1)} = 1 \ \forall d$, $\boldsymbol{W}_d^{(1)} = \boldsymbol{I} \ \forall d$
4: for $t = 1, 2, 3, \ldots$
5: $\quad \boldsymbol{x}^{(t)} \leftarrow \arg\min_{\boldsymbol{x}} \sum_{d=1}^{D} \lambda_d^{(t)} \|\boldsymbol{W}_d^{(t)} \boldsymbol{\Psi}_d \boldsymbol{x}\|_1$ s.t. $\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2 \leq \delta$
6: $\quad (\lambda_d^{(t+1)}, \epsilon_d^{(t+1)}) \leftarrow \arg \max_{\lambda_d \in \Lambda, \epsilon_d > 0} \log p(\boldsymbol{x}^{(t)}; \boldsymbol{\lambda}, \boldsymbol{\epsilon}), \ d = 1, ..., D$
7: $\quad \boldsymbol{W}_d^{(t+1)} \leftarrow \operatorname{diag}\left\{ \frac{1}{\epsilon_d^{(t+1)} + |\boldsymbol{\psi}_{d,1}^{\mathsf{T}} \boldsymbol{x}^{(t)}|}, \cdots, \frac{1}{\epsilon_d^{(t+1)} + |\boldsymbol{\psi}_{d,L_d}^{\mathsf{T}} \boldsymbol{x}^{(t)}|} \right\}, \ d = 1, ..., D$
8: end
9: output: $\boldsymbol{x}^{(t)}$

- IRW version of Co-L1: tunes both λ_d and $\boldsymbol{W}_d$ for all d.
- also tunes regularization parameters ϵ_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 ℓ_0 minimization,
3. Bayesian estimation according to a particular hierarchical prior,
4. variational EM algorithm under a particular prior.

Outline

Optimization Interpretations of Co-L1

Co-L1 is an MM approach to the weighted log-sum optimization problem

$$\arg\min_{\boldsymbol{x}} \sum_{d=1}^{D} L_d \log(\epsilon + \|\boldsymbol{\Psi}_d \boldsymbol{x}\|_1) \text{ s.t. } \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2 \leq \delta.$$

and

As $\epsilon \to 0$, Co-L1 aims to solve the weighted $\ell_{1,0}$ problem

$$\arg\min_{\boldsymbol{x}} \sum_{d=1}^{D} L_d \, 1_{\|\boldsymbol{\Psi}_d \boldsymbol{x}\|_1 > 0} \text{ s.t. } \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2 \leq \delta.$$

Note: L_d is the size of dictionary $\boldsymbol{\Psi}_d$, and $1_{\square}$ 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

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

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

and

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

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

Outline

A Stepping Stone

The IRW version of real-valued Co-L1: tunes both inter-dictionary weights λ_d and intra-dictionary weights $\boldsymbol{W}_d$ for given parameters ϵ_d.

---

input: $\{\boldsymbol{\Psi}_d\}_{d=1}^{D}$, $\boldsymbol{\Phi}$, $\boldsymbol{y}$, $\delta \geq 0$, $\epsilon_d > 0 \ \forall d$,

initialization: $\lambda_d^{(1)} = 1 \ \forall d$, $\boldsymbol{W}_d^{(1)} = \boldsymbol{I} \ \forall d$

for $t = 1, 2, 3, \ldots$

$$\boldsymbol{x}^{(t)} \leftarrow \arg\min_{\boldsymbol{x}} \sum_{d=1}^{D} \lambda_d^{(t)} \|\boldsymbol{W}_d^{(t)} \boldsymbol{\Psi}_d \boldsymbol{x}\|_1 \text{ s.t. } \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2 \leq \delta$$

$$\lambda_d^{(t+1)} \leftarrow \left[\frac{1}{L_d} \sum_{l=1}^{L_d} \log\left(1 + \frac{|\boldsymbol{\psi}_{d,l}^{\mathsf{T}} \boldsymbol{x}^{(t)}|}{\epsilon_d}\right)\right]^{-1} + 1, \quad d = 1, ..., D$$

$$\boldsymbol{W}_d^{(t+1)} \leftarrow \operatorname{diag}\left\{\frac{1}{\epsilon_d + |\boldsymbol{\psi}_{d,1}^{\mathsf{T}} \boldsymbol{x}^{(t)}|}, \cdots, \frac{1}{\epsilon_d + |\boldsymbol{\psi}_{d,L_d}^{\mathsf{T}} \boldsymbol{x}^{(t)}|}\right\}, \ d = 1, ..., D$$

end

output: $\boldsymbol{x}^{(t)}$

Optimization Interpretations of real-Co-IRW-L1-ϵ

Real-Co-IRW-L1-ϵ is an MM approach to the non-convex optimization

$$\arg\min_{\boldsymbol{x}} \sum_{d=1}^{D} \sum_{l=1}^{L_d} \log\left[\left(\epsilon_d + |\boldsymbol{\psi}_{d,l}^{\mathsf{T}} \boldsymbol{x}|\right) \sum_{i=1}^{L_d} \log\left(1 + \frac{|\boldsymbol{\psi}_{d,i}^{\mathsf{T}} \boldsymbol{x}|}{\epsilon_d}\right)\right] \text{ s.t. } \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2 \leq \delta$$

and

As $\epsilon \to 0$, real-Co-IRW-L1-ϵ aims to solve the ℓ_0+ weighted $\ell_{0,0}$ problem

$$\arg\min_{\boldsymbol{x}} \left[\|\boldsymbol{\Psi}\boldsymbol{x}\|_0 + \sum_{d=1}^{D} L_d \, 1_{\|\boldsymbol{\Psi}_d \boldsymbol{x}\|_0 > 0}\right] \text{ s.t. } \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2 \leq \delta.$$

Note: L_d is the size of dictionary $\boldsymbol{\Psi}_d$, and $1_{\square}$ is the indicator function.

Bayesian Interpretations of real-Co-IRW-L1-ϵ

Real-Co-IRW-L1 is an MM approach to Bayesian MAP estimation under an AWGN likelihood and the hierarchical prior

$$p(\boldsymbol{x}|\boldsymbol{\lambda}) = \prod_{d=1}^{D}\prod_{l=1}^{L_d} \frac{\lambda_d}{2\epsilon_d}\left(1 + \frac{|\boldsymbol{\psi}_{d,l}^{\mathsf{T}}\boldsymbol{x}|}{\epsilon_d}\right)^{-(\lambda_d+1)} \quad \text{i.i.d. generalized-Pareto}$$

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

and

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

$$p(\boldsymbol{x};\boldsymbol{\lambda}) = \prod_{d=1}^{D}\prod_{l=1}^{L_d} \frac{\lambda_d - 1}{2\epsilon_d}\left(1 + \frac{|\boldsymbol{\psi}_{d,l}^{\mathsf{T}}\boldsymbol{x}|}{\epsilon_d}\right)^{-\lambda_d} \quad \text{i.i.d. generalized-Pareto}$$

The Co-IRW-L1 Algorithm

Finally, we self-tune ϵ_d and allow for real or complex quantities:

1: input: $\{\boldsymbol{\Psi}_d\}_{d=1}^D$, $\boldsymbol{\Phi}$, $\boldsymbol{y}$, $\delta \geq 0$,

2: if $\boldsymbol{x} \in \mathbb{R}^N$, use $\Lambda = (1, \infty)$ and the real version of $\log p(\boldsymbol{x}; \boldsymbol{\lambda}, \boldsymbol{\epsilon})$;
if $\boldsymbol{x} \in \mathbb{C}^N$, use $\Lambda = (2, \infty)$ and the complex version of $\log p(\boldsymbol{x}; \boldsymbol{\lambda}, \boldsymbol{\epsilon})$.

3: initialization: $\lambda_d^{(1)} = 1 \ \forall d$, $\boldsymbol{W}_d^{(1)} = \boldsymbol{I} \ \forall d$

4: for $t = 1, 2, 3, \ldots$

5: $$\boldsymbol{x}^{(t)} \leftarrow \arg\min_{\boldsymbol{x}} \sum_{d=1}^D \lambda_d^{(t)} \|\boldsymbol{W}_d^{(t)} \boldsymbol{\Psi}_d \boldsymbol{x}\|_1 \text{ s.t. } \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2 \leq \delta$$

6: $$(\lambda_d^{(t+1)}, \epsilon_d^{(t+1)}) \leftarrow \arg \max_{\lambda_d \in \Lambda, \epsilon_d > 0} \log p(\boldsymbol{x}^{(t)}; \boldsymbol{\lambda}, \boldsymbol{\epsilon}), \ d = 1, \ldots, D$$

7: $$\boldsymbol{W}_d^{(t+1)} \leftarrow \operatorname{diag}\left\{ \frac{1}{\epsilon_d^{(t+1)} + |\boldsymbol{\psi}_{d,1}^\mathsf{T} \boldsymbol{x}^{(t)}|}, \cdots, \frac{1}{\epsilon_d^{(t+1)} + |\boldsymbol{\psi}_{d,L_d}^\mathsf{T} \boldsymbol{x}^{(t)}|} \right\}, \ d = 1, \ldots, D$$

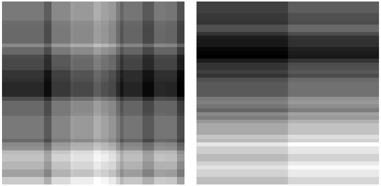
8: end

9: output: $\boldsymbol{x}^{(t)}$

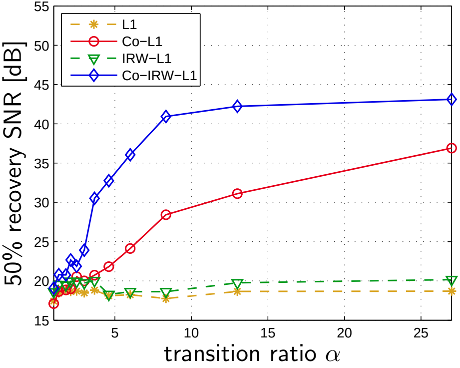
Outline

Experiment: Synthetic finite difference image

$\alpha = 1$ $\alpha = 27$



- 48×48 image with a total of 28 horiz & vert transitions.
- $\alpha \triangleq \frac{\#\text{ vertical transitions}}{\#\text{ horizontal transitions}}$
- "spread-spectrum" $\mathbf{\Phi}$
- sampling ratio $\frac{M}{N} = 0.3$
- AWGN @ 30 dB SNR
- $\mathbf{\Psi}_1 =$ vertical finite difference, $\mathbf{\Psi}_2 =$ horizon. finite difference

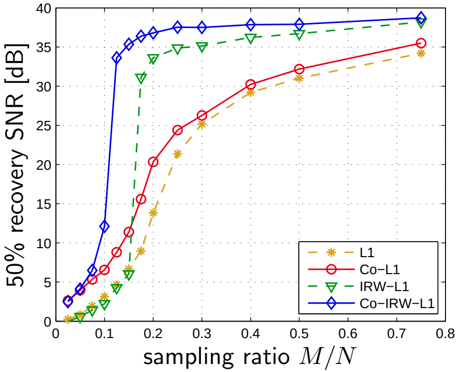


$\Rightarrow$ The composite algorithms significantly outperform the non-composite ones

$\Rightarrow$ Performance improves as sparsities become more disparate!

Experiment: Shepp-Logan Phantom

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

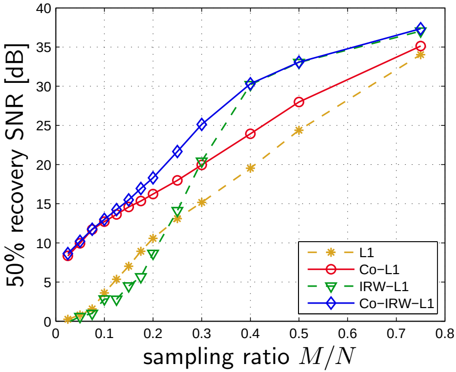


$\Rightarrow$ The composite algorithms significantly outperform the non-composite ones

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

Experiment: Cameraman

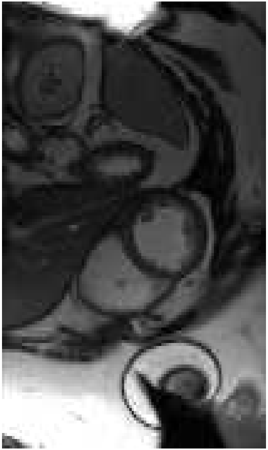
- 96×104 image
- "spread-spectrum" $\mathbf{\Phi}$
- AWGN @ 40 dB SNR
- $\mathbf{\Psi} \in \mathbb{R}^{7N \times N}$ = 2D UWT-db1, $\mathbf{\Psi}_d \in \mathbb{R}^{N \times N}$ $\forall d$



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

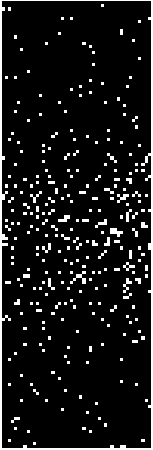
$\Rightarrow$ Performance gap is larger for small M/N

Experiment: 1D Dynamic MRI



x-y profile

x-t profile

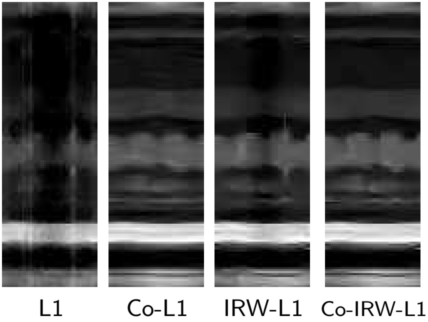


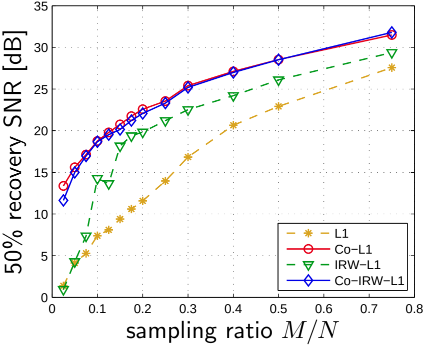
k-t sampling

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

Experiment: 1D Dynamic MRI (cont.)

sampling ratio $M/N = 0.3$





- 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

	Shepp-Logan	Cameraman	dMRI
L1	20.8s	23.1s	29.3s
Co-L1	32.7s	34.2s	86.4s
IRW-L1	45.9s	48.4s	54.1s
Co-IRW-L1	72.1s	96.4s	131s

The composite algs run 1.5–$3\times$ 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 ℓ_0 methods, Bayesian methods, and variational Bayesian methods.

Conclusions

Thanks!

Iteratively Reweighted ℓ_1 (IRW-L1)

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

1: input: $\boldsymbol{\Psi} = [\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_L]^\mathsf{T}$, $\boldsymbol{\Phi}$, $\boldsymbol{y}$, $\delta \geq 0$, $\epsilon \geq 0$

2: initialization: $\boldsymbol{W}^{(1)} = \boldsymbol{I}$

3: for $t = 1, 2, 3, \ldots$

4: $\quad \boldsymbol{x}^{(t)} \leftarrow \arg\min_{\boldsymbol{x}} \|\boldsymbol{W}^{(t)}\boldsymbol{\Psi}\boldsymbol{x}\|_1$ s.t. $\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2 \leq \delta$

5: $\quad \boldsymbol{W}^{(t+1)} \leftarrow \operatorname{diag}\left\{ \frac{1}{\epsilon + |\boldsymbol{\psi}_1^\mathsf{T}\boldsymbol{x}^{(t)}|}, \cdots, \frac{1}{\epsilon + |\boldsymbol{\psi}_L^\mathsf{T}\boldsymbol{x}^{(t)}|} \right\}$

6: end

7: output: $\boldsymbol{x}^{(t)}$

- behaves more like ℓ_0 minimization than ℓ_1 minimization alone,
- leverages existing ℓ_1 solvers.

Majorize-Minimization (MM) Interpretation of IRW-L1

IRW-L1 is an MM approach to the log-sum optimization problem

$$\arg\min_{\boldsymbol{x}} \; \textstyle\sum_{l=1}^{L} \log(\epsilon + |\boldsymbol{\psi}_l^{\mathsf{T}} \boldsymbol{x}|) \text{ s.t. } \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2 \leq \delta.$$

How to see this? Reformulate as

$$\arg\min_{\boldsymbol{x},\boldsymbol{u}} \; \sum_l \log(\epsilon + u_l) \text{ s.t. } \begin{cases} \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2 \leq \delta \\ |\boldsymbol{\psi}_l^{\mathsf{T}} \boldsymbol{x}| \leq u_l \;\forall l, \end{cases}$$

$$\Leftrightarrow \quad \arg\min_{\boldsymbol{v}} g(\boldsymbol{v}) \text{ s.t. } \boldsymbol{v} \in \mathcal{C}$$

for $\boldsymbol{v} = \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{x} \end{bmatrix}$, convex $\mathcal{C}$, and concave g.

MM procedure: Iterate for $t = 1, 2, 3, \ldots$

1. create surrogate $g(\boldsymbol{v}; \boldsymbol{v}^{(t)})$ that majorizes $g(\boldsymbol{v})$ at $\boldsymbol{v}^{(t)}$,
2. minimize the surrogate over $\boldsymbol{v} \in \mathcal{C}$, producing $\boldsymbol{v}^{(t+1)}$.

MM Interpretation of IRW-L1 (cont.)

Our concave $g(\boldsymbol{v})$ is majorized by the tangent at $\boldsymbol{v}^{(t)}$. So MM becomes

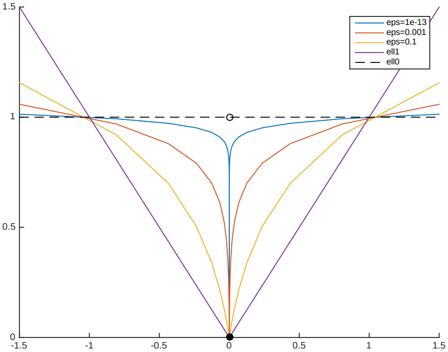
$$\begin{aligned}
\boldsymbol{v}^{(t+1)} &= \arg\min_{\boldsymbol{v}\in\mathcal{C}} g(\boldsymbol{v}^{(t)}) + \nabla g(\boldsymbol{v}^{(t)})^{\mathsf{T}}[\boldsymbol{v} - \boldsymbol{v}^{(t)}] \\
&= \arg\min_{\boldsymbol{v}\in\mathcal{C}} \nabla g(\boldsymbol{v}^{(t)})^{\mathsf{T}}\boldsymbol{v} \\
\Leftrightarrow \quad \boldsymbol{x}^{(t+1)} &= \arg\min_{\boldsymbol{x}} \underbrace{\sum_l \frac{1}{\epsilon + |\boldsymbol{\psi}_l^{\mathsf{T}}\boldsymbol{x}^{(t)}|} |\boldsymbol{\psi}_l^{\mathsf{T}}\boldsymbol{x}|}_{\|\boldsymbol{W}^{(t)}\boldsymbol{\Psi}\boldsymbol{x}\|_1} \text{ s.t. } \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2 \le \delta
\end{aligned}$$

Implications of MM:

- IRW-L1 convergence is guaranteed
- but possibly to a suboptimal local minimum (since non-convex).

Approximate-ℓ_0 Interpretation of IRW-L1

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



$\Rightarrow$ As $\epsilon \to 0$, the log-sum penalty becomes a scaled and shifted version of the ℓ_0 penalty.