Efficient Methods for Overlapping Group Lasso

Presented by Xingguo Li

CSCI 8980, UMN

Authors: Lei Yuan, Jun Liu, and Jieping Ye

March 31, 2014
Outline

- The Overlapping Group Lasso
- The Proximal Operator and Efficient Computation
  - Key Properties of the Proximal Operator
  - Reformulation as a Smooth Convex Problem
  - Proximal Splitting Methods
- Extensions
  - $\ell_q$ Norm Overlapping Group Lasso
  - Capped Norm Overlapping Group Lasso
- Numerical Experiments
  - Synthetic Data
  - Gene Expression Data
Overlapping Group Lasso

\[
\min_{x \in \mathbb{R}^p} f(x) = l(x) + \phi_{\lambda_2}^{\lambda_1}(x) \tag{1}
\]

- \( l(x) \) is a smooth convex loss function, e.g.
  \[
l(x) = \sum_{i=1}^{n} (y_i - a_i^T x)^2
  \]
- \( \phi_{\lambda_2}^{\lambda_1}(x) = \lambda_1 \|x\|_1 + \lambda_2 \sum_{i=1}^{g} w_i \|x_{G_i}\| \) is the overlapping group Lasso penalty
  - \( \lambda_1 = 0, \lambda_2 > 0 \): group Lasso (Yuan et al., 2006)
  - \( \lambda_1 > 0, \lambda_2 = 0 \): Lasso (Tibshirani, 1996)
“FoGLasso”, Fast overlapping Group Lasso, based on accelerated gradient descent (AGD) (Beck et al., 2009).

- Approximation (Linearization) of $f(x)$ as

$$f_{L,x}(y) = \left[ l(x) + \langle l'(x), y - x \rangle + \frac{L}{2} \| y - x \|^2 \right] + \phi_{\lambda_2}^\lambda(y) \quad (2)$$

- A sequence of approximate solutions $\{x_i\}$ by proximal operator,

$$x_{i+1} = \arg \min_y f_{L_i,s_i}(y) = \pi_{\lambda_2/L_i}^{\lambda_1/L_i} (s_i - \frac{1}{L_i} l'(s_i)),$$  \hspace{2cm} (3)

where $s_i = x_i + \beta_i(x_i - x_{i-1})$ and $L_i$ can be determined by line search.
Algorithm 1: “FoGLasso”

**Input:** $L_0 > 0, x_0, k$

**Output:** $x_{k+1}$

1. Initialize $x_1 = x_0, \alpha_{-1} = 0, \alpha = 1$, and $L = L_0$.
2. for $i = 1$ to $k$ do
   3. Set $\beta_i = \frac{\alpha_{i-2}-1}{\alpha_{i-1}}, s_i = x_i + \beta_i(x_i - x_{i-1})$
   4. Find the smallest $L = 2^j L_{i-1}, j = 0, 1, \cdots$ such that $f(x_{i+1}) \leq f_{L,s_i}(x_{i+1})$ holds, where $x_{i+1} = \pi_{\lambda_1/L_i}(s_i - \frac{1}{L_i} l'(s_i))$
   5. Set $L_i = L$ and $\alpha_{i+1} = \frac{1+\sqrt{1+4\alpha_i^2}}{2}$
6. end for
The proximal operator:

\[ x_{i+1} = \pi_{\lambda_1/L_i} \left( s_i - \frac{1}{L_i} l'(s_i) \right) \]

Definition (recall \( \phi_{\lambda_2}(x) = \lambda_1 \| x \|_1 + \lambda_2 \sum_{i=1}^{g} w_i \| x_{G_i} \| \)):

\[ \pi_{\lambda_1}(v) = \arg \min_{x \in \mathbb{R}^p} \left\{ g_{\lambda_2}(x) \equiv \frac{1}{2} \| x - v \|_2^2 + \phi_{\lambda_2}(x) \right\} \tag{4} \]

- Many groups are zero (identify \( x_{G_i} = 0 \))
- \( g_{\lambda_2}(x) \) is nonsmooth (smooth reformulation)
- More proximal operator solver (Dykstra-like, ADMM)
Lemma 1

Suppose $\lambda_1, \lambda_2 \geq 0$ and $w_i > 0$, $i = 1, 2, \cdots, g$. Let $\mathbf{x}^* = \pi_{\lambda}^{\lambda_1}(\mathbf{v})$ and $\odot$ be point-wise product, then the following holds:

1. if $v_i > 0$, then $0 \leq x_i^* \leq v_i$;
2. if $v_i < 0$, then $v_i \leq x_i^* \leq 0$;
3. if $v_i = 0$, then $x_i^* = 0$;
4. $\text{SGN}(\mathbf{v}) \subseteq \text{SGN}(\mathbf{x}^*)$; and
5. $\pi_{\lambda_2}^{\lambda_1}(\mathbf{v}) = \text{sgn}(\mathbf{v}) \odot \pi_{\lambda_2}^{\lambda_1}(|\mathbf{v}|)$.

\[
\text{SGN}(t) = \begin{cases} 
\{1\}, & t > 0 \\
\{-1\}, & t < 0 \\
[-1, 1], & t = 0 
\end{cases}
\]

\[
\text{sgn}(t) = \begin{cases} 
1, & t > 0 \\
-1, & t < 0 \\
0, & t = 0 
\end{cases}
\]
Key Properties of the Proximal Operator

**Theorem 1**

Let \( \mathbf{u} = \text{sgn}(\mathbf{v}) \odot \max(|\mathbf{v}| - \lambda_1, 0) \), and

\[
\pi^0_{\lambda_2}(\mathbf{u}) = \arg \min_{\mathbf{x} \in \mathbb{R}^p} \left\{ h_{\lambda_2}(\mathbf{x}) \equiv \frac{1}{2} \| \mathbf{x} - \mathbf{u} \|^2 + \lambda_2 \sum_{i=1}^{g} w_i \| \mathbf{x}_{G_i} \| \right\}. \tag{5}
\]

Then, the following holds: \( \pi^1_{\lambda_2}(\mathbf{v}) = \pi^0_{\lambda_2}(\mathbf{u}) \).

- Nice! \( \pi^1_{\lambda_2}(\mathbf{v}) \) reduces to (5).
- Difficulty: groups may overlap.
- Many groups are zero (sparse solution solution desired), how to identify?
Key Properties of the Proximal Operator

- Sufficient condition for a group to be zero:

**Lemma 2**

Let \( \mathbf{x}^* = \arg \min_{\mathbf{x} \in \mathbb{R}^p} h_{\lambda_2}(\mathbf{x}) \). If the \( i \)-th group satisfies 
\[ \| \mathbf{u}_{G_i} \| \leq \lambda_2 w_i, \text{ then } \mathbf{x}^*_G = \mathbf{0}, \text{ i.e. the } i \text{-th group is zero.} \]

- Given \( S_i = \bigcup_{j \neq i} \mathbf{x}^*_G = \mathbf{0}(G_j \cap G_i) \), a much weaker condition
  (much more zero groups can be identified):

**Lemma 3**

Let \( \mathbf{x}^* = \arg \min_{\mathbf{x} \in \mathbb{R}^p} h_{\lambda_2}(\mathbf{x}) \). If the \( i \)-th group satisfies 
\[ \| \mathbf{u}_{G_i - S_i} \| \leq \lambda_2 w_i, \text{ then } \mathbf{x}^*_G = \mathbf{0}, \text{ i.e. the } i \text{-th group is zero.} \]

- Iterative procedure to identify the zero groups.
Focus on reduced problem $\mathbf{u} \succ 0$. Rewrite $\pi_{\lambda_2}^0(\mathbf{u})$ as:

$$
\pi_{\lambda_2}^0(\mathbf{u}) = \arg \min_{\mathbf{x} \in \mathbb{R}^p} \max_{\mathbf{0} \preceq \mathbf{x} \preceq \mathbf{u}} \left\{ h_{\lambda_2}(\mathbf{x}) \equiv \frac{1}{2} \| \mathbf{x} - \mathbf{u} \|^2 + \lambda_2 \sum_{i=1}^{g} w_i \| \mathbf{x}_{G_i} \| \right\}.
$$

Use dual norm of $\| \cdot \|$, rewrite $h_{\lambda_2}(\mathbf{x})$ as:

$$
h_{\lambda_2}(\mathbf{x}) = \max_{\mathbf{Y} \in \Omega} \frac{1}{2} \| \mathbf{x} - \mathbf{u} \|^2 + \sum_{i=1}^{g} \langle \mathbf{x}, \mathbf{Y}^i \rangle,
$$

where $\Omega = \left\{ \mathbf{Y} \in \mathbb{R}^{p \times g} : \mathbf{Y}^i_{G_i} = \mathbf{0}, \| \mathbf{Y}^i \| \leq \lambda_2 w_i, i = 1, \ldots, g \right\}$.

Reformulation as a min-max problem:

$$
\pi_{\lambda_2}^0(\mathbf{u}) = \arg \min_{\mathbf{x} \in \mathbb{R}^p} \max_{\mathbf{Y} \in \Omega} \left\{ \psi(\mathbf{x}, \mathbf{Y}) \equiv \frac{1}{2} \| \mathbf{x} - \mathbf{u} \|^2 + \langle \mathbf{x}, \mathbf{Y} \mathbf{e} \rangle \right\}
$$
Reformulation as a Smooth Convex Problem (continue....)

\[ \psi(x, Y) \text{ is convex in } x, \text{ concave in } Y. \text{ Methodology for } \min h_{\lambda_2}(\cdot): \]

- w.r.t. \( Y \), \( \arg\min_{Y \in \Omega} \{ w(Y) = -\psi(\max(u - Ye, 0), Y) \} \)
- w.r.t. \( x \), \( x = \max(u - Ye, 0) \Rightarrow \) construct solution to \( h_{\lambda_2}(\cdot) \)

**Theorem 2**

The function \( w(Y) \) is convex and continuously differentiable with

\[
 w'(Y) = -\max(u - Ye, 0)e^T \quad (8)
\]

In addition, \( w'(Y) \) is Lipschitz continuous with constant \( g \), i.e.,

\[
 \| w'(Y_1) - w'(Y_2) \|_F \leq g \| Y_1 - Y_2 \|_F, \forall Y_1, Y_2 \in \mathbb{R}^{p \times g}. \quad (9)
\]

Use accelerated gradient descent (AGD) method to solve \( \psi(x, Y) \).
Theorem 3

Let \( \text{gap} \tilde{Y} = \max_{Y \in \Omega} \psi(\tilde{x}, Y) - \min_{x \in \mathbb{R}^p, 0 \leq x \leq u} \psi(x, \tilde{Y}) \) be the duality gap. Then, the following holds:

\[
\text{gap}(\tilde{Y}) = \sum_{i=1}^{g} (\lambda_2 w_i \|\tilde{x}_{G_i}\| - \langle \tilde{x}_{G_i}, \tilde{Y}_{G_i} \rangle).
\]  

(10)

In addition, we have

\[
w(\tilde{Y}) - w(Y^*) \leq \text{gap}(\tilde{Y}),
\]

(11)

\[
h(\tilde{x}) - h(x^*) \leq \text{gap}(\tilde{Y}).
\]

(12)

Serve as the stopping criteria (e.g. \( < 10^{-10} \)).
Proximal Splitting Methods

- Dykstra-like Proximal Splitting Method (Combettes et al., 2009)
- ADMM (Boyd et al., 2011)

Dykstra-like Proximal Splitting Method: *convex feasibility problem*

\[
\text{find } x \in \{ \bigcap_{i=1}^{m} C_i \mid C_i \text{ is a convex set} \}
\]

- Iterative scheme by cycling through all convex sets
- Convergence guarantee under certain conditions

Consider \( \pi_{\lambda_2}^0(u) = \arg\min_{x \in \mathbb{R}^p} \frac{1}{2} \| x - u \|^2 + \lambda_2 \sum_{i=1}^{g} w_i \| x_{G_i} \| \) as the projection of \( u \) onto a collection of convex sets \( \{ w_i \| x_{G_i} \| \} \).
Algorithm 2: Dykstra-like Proximal Splitting Methods

1: Set $x_0 = u, q_{1,0}, \ldots, q_{g,0} = x_0, n = 0$
2: repeat $n = n + 1$
3: for $i = 1$ to $g$ do
4: \hspace{1em} $p_{i,n} = \text{prox}_{\lambda \|x_{G_i}\|} q_{i,n}$
5: \hspace{1em} $x_{n+1} = \sum_{i=1}^{g} w_i q_{i,n}$
6: for $i = 1$ to $g$ do
7: \hspace{1em} $q_{i,n+1} = x_{n+1} + q_{i,n} - p_{i,n}$
8: until Convergence

\[ p = \text{prox}_{\lambda \|x_{G_i}\|} q = \arg\min_{x \in \mathbb{R}^p} \|x - q\|^2/2 + \lambda \|x_{G_i}\| \]

$\Rightarrow p_{G_i} = \frac{\max(\|q_{G_i}\| - \lambda, 0)}{\|q_{G_i}\|} q_{G_i}$ (closed form)
ADMM

- Reformulation with auxiliary variables:
  \[
  \min_{x, z} \frac{1}{2} \| x - u \|^2 + \lambda \sum_{i=1}^{g} w_i \| z_i \|
  \text{ s.t. } z_i = x_{G_i}, \ i = 1, \cdots, g.
  \]

- Augmented Lagrangian:
  \[
  L_\rho(x, z, y) = \frac{1}{2} \| x - u \|^2 + \lambda \sum_{i=1}^{g} w_i \| z_i \|
  + \sum_{i=1}^{g} y_i^T (z_i - x_{G_i}) + \frac{\rho}{2} \sum_{i=1}^{g} \| z_i - x_{G_i} \|^2
  \]

- ADMM iterations:
  \[
  x^{k+1} := \arg\min_x L_\rho(x, z^k, y^k)
  
  z^{k+1} := \arg\min_x L_\rho(x, z^k, y^k)
  
  y^{k+1} := \arg\min_x L_\rho(x, z^k, y^k)
  \]
ADMM

For $x$, \( \frac{\partial}{\partial x} L_\rho(x, z^k, y^k) = x - u - \sum_{i=1}^{g} \tilde{y}^k_i + \rho \sum_{i=1}^{g} \tilde{e}_i \odot x - \rho \sum_{i=1}^{g} \tilde{z}^k_i \)

\[ \Rightarrow x^{k+1} = (u + \sum_{i=1}^{g} \tilde{y}^k_i + \rho \sum_{i=1}^{g} \tilde{z}^k_i) \odot (e + \rho \sum_{i=1}^{g} \tilde{e}_i) \]

For $z$, use subdifferential,

\[ 0 \in z_i^{k+1} - x^{k+1}_G + \frac{1}{\rho} y^k_i + \frac{\lambda w_i}{\rho} \partial \| z_i^{k+1} \|, \]
where \( \partial \| z_i^{k+1} \| = \begin{cases} \frac{z_i^{k+1}}{\| z_i^{k+1} \|} & \| z_i^{k+1} \| \neq 0 \\ \{ t \mid t \in \mathbb{R}^{|G_i|}, \| t \| < 1 \} & \| z_i^{k+1} \| = 0 \end{cases} \)

\[ \Rightarrow z_i^{k+1} = \max \{ \| x^{k+1}_G \| - \tilde{\lambda}_i, 0 \} \frac{x^{k+1}_G}{\| x^{k+1}_G \|}, \]
where \( \tilde{x}^{k+1}_G = x^{k+1}_G - \frac{1}{\rho} y^k_i, \tilde{\lambda}_i = \frac{\lambda w_i}{\rho} \)
\( \psi_{q, \lambda_2}^\lambda (x) = \lambda_1 \| x \|_1 + \lambda_2 \sum_{i=1}^{g} w_i \| x_{G_i} \|_q \) \hfill (13)

\[ \pi_{q, \lambda_2}^\lambda (v) = \arg \min_{x \in \mathbb{R}^p} \left\{ g_{q, \lambda_2}^\lambda (x) \equiv \frac{1}{2} \| x - v \|_2^2 + \phi_{q, \lambda_2}^\lambda (x) \right\} \] \hfill (14)

- Generalize \( \psi_{\lambda_2}^\lambda (x) \) and \( \pi_{\lambda_2}^\lambda (v) \) to

- Same properties hold for \( \ell_q \) proximal operator: \( 1/q + 1/\bar{q} = 1 \),
  Necessary condition: If \( \| u_{G_i} \|_{\bar{q}} \leq \lambda_2 w_i \), then \( x_{G_i}^* = 0 \).
  A weaker condition: If \( \| u_{G_i - S_i} \|_{\bar{q}} \leq \lambda_2 w_i \), then \( x_{G_i}^* = 0 \).
Same result holds for the duality gap for smooth reformulation:

$$\text{gap}(\tilde{Y}) = \sum_{i=1}^{g} (\lambda_2 w_i \|\tilde{x}_{G_i}\|_q - \langle \tilde{x}_{G_i}, \tilde{Y}_{G_i}^i \rangle).$$

Feasible region of the dual variable $Y$:

$$\Omega = \left\{ Y \in \mathbb{R}^{p \times g} : Y_{G_i^c}^i = 0, \|Y_i\|_q \leq \lambda_2 w_i, i = 1, \cdots, g \right\}$$

Efficient bisection root-finding based $\ell_q$-norm projection (Liu et al., 2010)
Consider the problem:

$$\min_{x \in \mathbb{R}^p} l(x) + \lambda_1 \|x\|_0 + \lambda_2 \sum_{i=1}^{g} w_i l(\|x_{G_i}\| \neq 0) \quad (15)$$

- $\ell_1$-norm regularization introduces bias.
- Nonconvex capped norms: closer to $\ell_0$-norm than $\ell_1$-norm (Zhang 2011, Shen et al. 2012): for some small $\theta_1, \theta_2 > 0$,

  $$\|x\|_0 \approx \sum_{j=1}^{p} \min \left( 1, \frac{|x_j|}{\theta_1} \right)$$

  $$\sum_{i=1}^{g} w_i l(\|x_{G_i}\| \neq 0) \approx \sum_{i=1}^{g} w_i \min \left( 1, \frac{|x_{G_i}|}{\theta_2} \right)$$
Capped Norm Overlapping Group Lasso

- Decompose $\sum_{j=1}^{p} \min \left(1, \frac{|x_j|}{\theta_1} \right)$ and $\sum_{i=1}^{g} w_i \min \left(1, \frac{|x_{G_i}|}{\theta_2} \right)$, approximate the problem 15 as:

$$\min_{x \in \mathbb{R}^p} l(x) + \frac{\lambda_1}{\theta_1} \|x\|_1 + \frac{\lambda_2}{\theta_2} \sum_{i=1}^{p} \|x_{G_i}\| - P(x) - D(x)$$

$$P(x) = \frac{\lambda_1}{\theta_1} \sum_{i=1}^{p} \max(|x_j| - \theta_1, 0) \text{ convex in } x$$

$$D(x) = \frac{\lambda_2}{\theta_2} \sum_{i=1}^{p} w_i \max(\|x_{G_i}\| - \theta_2, 0) \text{ convex in } x$$

- “Difference of two convex functions” (DC) programming
Algorithm 3: DC Programming for Overlapping Group Lasso with the Capped Norm

\[
\frac{\partial}{\partial x_j} P(x) \begin{cases} 
\frac{\lambda_1}{\theta_1} \text{sgn}(x_j) & |x_j| > \theta_1 \\
0 & |x_j| \leq \theta_1
\end{cases} \quad \frac{\partial}{\partial x_{G_i}} D(x_{G_i}) \begin{cases} 
\frac{x_{G_i}}{\|x_{G_i}\|} & \|x_{G_i}\| > \theta_2 \\
0 & \|x_{G_i}\| \leq \theta_2
\end{cases}
\]

**Input:** $\theta_0, \theta_1 > 0, x_0, k$

**Output:** $x_{k+1}$

1. Initialize $x_1 = x_0$

2. for $i = 1$ to $k$

3. Choose $U^i \in \partial P(x^i)$ and $V^i \in \partial D(x^i)$

4. Solve $x^{i+1} = \arg\min_{x \in \mathbb{R}^p} l(x) + \frac{\lambda_1}{\theta_1} \|x\|_1 + \frac{\lambda_2}{\theta_2} \sum_{i=1}^p \|x_{G_i}\| - \langle U^k + V^k, x \rangle$ (via “FoG Lasso”)

5. end for
Experiments: Efficiency of Calculating the Proximal Operator

Figure 1: Time comparison for computing the proximal operators. The group number $g$ is fixed in the left figure and the problem size $p$ is fixed in the middle figure. The right figure illustrates the effectiveness of the preprocessing.
Figure 2: Results of the convex overlapping group Lasso formulation (top row) and the nonconvex overlapping group Lasso with the capped norm (bottom row).
Sparse Pattern Recovery

TABLE 1
Cross-Validation Performance of Sparse Pattern Recovery of the Convex Overlapping Group Lasso Formulation and the Nonconvex Overlapping Group Lasso Formulation Based on the Capped Norm on Synthetic Data with Different Problem Sizes

<table>
<thead>
<tr>
<th>n</th>
<th>Entry Rate</th>
<th>Group Rate</th>
<th>Entry Rate</th>
<th>Group Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>300</td>
<td>0.71</td>
<td>0.60</td>
<td>0.77</td>
<td>0.71</td>
</tr>
<tr>
<td>400</td>
<td>0.80</td>
<td>0.61</td>
<td>0.82</td>
<td>0.70</td>
</tr>
</tbody>
</table>

- Nonconvex formulation outperforms convex formulation.
Comparison with SLasso, Prox-Grad, and ADMM

Figure 3: Comparison of SLasso (Jenatton et al. 2009), ADMM (Boyd et al. 2010), Prox-Grad (Chen et al. 2012), and “FoGLasso” in terms of computational time (in seconds and in the log scale).
Comparison with Picard-Nesterov

For each particular method, the first row denotes the number of outer iterations required for convergence, while the second row represents the total number of inner iterations.

- Same complexity of $O(pg)$ for inner iteration.
Computation of the Proximal Operator

**Figure 4:** Performance of the computation of the proximal operator in FoGlasso. The left plot shows the objective function value during the FoGlasso iteration. The middle plot shows the percentage of the identified zero groups. The right plot shows the number of inner iterations for achieving the duality gap less than $10^{-10}$ when one solves the proximal operator via the dual reformulation.

- Most zero groups are identified after $\sim 100$ steps.
Convergence with Inexact Proximal Operator

**Figure 5** : Illustration of the objective function values of the first 50 iterations with different stopping criteria used for computing the proximal operator.

- No dramatic change w.r.t different termination conditions.
Thank you!