Efficient Methods for Overlapping Group Lasso: Supplemental Material

A. Properties of the Function $\omega(\cdot)$ in (15)

Theorem 3. *The function $\omega(Y)$ is convex and continuously differentiable with*

$$\omega'(Y) = -\max(\mathbf{u} - Y\mathbf{e}, \mathbf{0})\mathbf{e}^{\mathrm{T}}. \tag{24}$$

In addition, $\omega'(Y)$ is Lipschitz continuous with constant g^2, i.e.,

$$\|\omega'(Y_1) - \omega'(Y_2)\|_F \leq g^2\|Y_1 - Y_2\|_F, \quad \forall\, Y_1, Y_2 \in \mathbb{R}^{p\times g}. \tag{25}$$

To prove Theorem 3, we first present two technical lemmas. The first lemma is related to the optimal value function [4, 9], and it was used in a recent study [27] on infinite kernel learning.

Lemma 4. *[4] Let X be a metric space and U be a normed space. Suppose that for all $\mathbf{x} \in X$, the function $\psi(\mathbf{x}, \cdot)$ is differentiable and that $\psi(\mathbf{x}, Y)$ and $D_Y\psi(\mathbf{x}, Y)$ (the partial derivative of $\psi(\mathbf{x}, Y)$ with respect to Y) are continuous on $X \times U$. Let Φ be a compact subset of X. Define the optimal value function as $\varphi(Y) = \inf_{\mathbf{x}\in\Phi} \psi(\mathbf{x}, Y)$. The optimal value function $\varphi(Y)$ is directionally differentiable. In addition, if $\forall Y \in U$, $\psi(\cdot, Y)$ has a unique minimizer $\mathbf{x}(Y)$ over Φ, then $\varphi(Y)$ is differentiable at Y and the gradient of $\varphi(Y)$ is given by $\varphi'(Y) = D_Y\psi(\mathbf{x}(Y), Y)$.*

The second lemma shows that the operator $\mathbf{y} = \max(\mathbf{x}, \mathbf{0})$ is non-expansive.

Lemma 5. $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^p$, *we have* $\|\max(\mathbf{x}, \mathbf{0}) - \max(\mathbf{y}, \mathbf{0})\| \leq \|\mathbf{x} - \mathbf{y}\|$.

Proof. The results follows since $|\max(x, 0) - \max(y, 0)| \leq |x - y|$, $\forall x, y \in \mathbb{R}$. □

Proof of Theorem 3: To prove the differentiability of $\omega(Y)$, we apply Lemma 4 with $X = \mathbb{R}^p$, $U = \mathbb{R}^{p\times g}$ and $\Phi = \{\mathbf{x} \in X : \mathbf{u} + \lambda_2 \sum w_i\mathbf{e} \geq \mathbf{x} \geq \mathbf{0}\}$. It is easy to verify that 1) $\psi(\mathbf{x}, \cdot)$ is differentiable; 2) $\psi(\mathbf{x}, Y)$ and $D_Y\psi(\mathbf{x}, Y) = \mathbf{x}\mathbf{e}^{\mathrm{T}}$ are continuous on $X \times U$; 3) Φ be a compact subset of X; and 4) $\forall Y \in U$, $\psi(\mathbf{x}, Y)$ has a unique minimizer $\mathbf{x}(Y) = \max(\mathbf{u} - Y\mathbf{e}, \mathbf{0})$ over Φ. Note that, the last result follows from $\mathbf{u} > 0$ and $\mathbf{u} - Y\mathbf{e} \leq \mathbf{u} + \lambda_2 \sum w_i\mathbf{e}$, where the latter inequality utilizes $\|Y^i\| \leq \lambda_2 w_i$; and this indicates that $\mathbf{x}(Y) = \max(\mathbf{u} - Y\mathbf{e}, \mathbf{0}) = \arg\min_{\mathbf{x}} \psi(\mathbf{x}, Y) = \arg\min_{\mathbf{x}\in\Phi} \psi(\mathbf{x}, Y)$. It follows from Lemma 4 that

$$\varphi(Y) = \inf_{\mathbf{x}\in\Phi} \psi(\mathbf{x}, Y) = \psi(\max(\mathbf{u} - Y\mathbf{e}, \mathbf{0}), Y)$$

is differentiable with $\varphi'(Y) = \max(\mathbf{u} - Y\mathbf{e}, \mathbf{0})\mathbf{e}^{\mathrm{T}}$.

In (13), $\psi(\mathbf{x}, Y)$ is convex in $\mathbf{x}$ and concave in Y, and the constraint sets are closed convex for both $\mathbf{x}$ and Y, thus the existence of the saddle point is guaranteed by the well-known von Neumann Lemma [21]. As a result,

$$\varphi(Y) = \inf_{\mathbf{x}\in\Phi} \psi(\mathbf{x}, Y) = \psi(\max(\mathbf{u} - Y\mathbf{e}, \mathbf{0}), Y)$$

is concave, and $\omega(Y) = -\varphi(Y)$ is convex. For any Y_1, Y_2, we have

$$\begin{aligned}
&\|\omega'(Y_1) - \omega'(Y_2)\|_F \\
=&\|\max(\mathbf{u} - Y_1\mathbf{e}, \mathbf{0})\mathbf{e}^{\mathrm{T}} - \max(\mathbf{u} - Y_2\mathbf{e}, \mathbf{0})\mathbf{e}^{\mathrm{T}}\|_F \\
\leq&\|\mathbf{e}\| \times \|\max(\mathbf{u} - Y_1\mathbf{e}, \mathbf{0}) - \max(\mathbf{u} - Y_2\mathbf{e}, \mathbf{0})\| \\
\leq&\|\mathbf{e}\| \times \|(Y_1 - Y_2)\mathbf{e}\| \\
\leq&g^2\|Y_1 - Y_2\|_F,
\end{aligned} \tag{26}$$

where the second inequality follows from Lemma 5. We prove (25). □

B. Dykstra-like Proximal Splitting Method for Computing the Proximal Operator in (5)

In the field of signal processing, one classical problem is the *convex feasibility problem*:

$$\text{find } x \in \bigcap_{i=1}^{m} C_i, \tag{27}$$

where C_i's are convex sets. Efficient methods have been designed for (27) where at each iteration, only one convex set is considered and the solution is updated iteratively by cycling through all convex sets. Under certain conditions, convergence is guaranteed. For our problem, since (5) can be considered as the projection of a vector $\mathbf{u}$ onto a collection of convex sets induced by the regularization components $w_i\|\mathbf{x}_{G_i}\|$, the proximal splitting ideas can be applied.

We define $f_i = \lambda\|\mathbf{x}_{G_i}\|$, the proximal operator in (5) can be rewritten as:

$$\min_{\mathbf{x}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{x}-\mathbf{u}\|^2 + \sum_{i=1}^{g} w_i f_i \tag{28}$$

Then, the Dykstra-like proximal algorithm can be summarized in Algorithm 2.

Algorithm 2 Dykstra-like Proximal Splitting Method

Set $\mathbf{x}_0 = \mathbf{u}$, $\mathbf{q}_{1,0}, \ldots, \mathbf{q}_{g,0} = \mathbf{x}_0$, $n = 0$
repeat
  for $i = 1, \ldots, g$ **do**
    $\mathbf{p}_{i,n} = \text{prox}_{f_i}\mathbf{q}_{i,n}$
  end for
  $\mathbf{x}_{n+1} = \sum_{i=1}^{g} w_i \mathbf{p}_{i,n}$
  for $i = 1, \ldots, g$ **do**
    $\mathbf{q}_{i,n+1} = \mathbf{x}_{n+1} + \mathbf{q}_{i,n} - \mathbf{p}_{i,n}$
  end for
  $n = n + 1$
until Convergence

The last piece of puzzle in Algorithm 2 is to solve $\mathbf{p} = \text{prox}_{f_i}\mathbf{q}$, defined as:

$$\mathbf{p} = \arg\min_{\mathbf{x}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{x}-\mathbf{q}\|^2 + \lambda\|\mathbf{x}_{G_i}\|$$

Clearly, we have $\mathbf{p}_{\overline{G}_i} = \mathbf{q}_{\overline{G}_i}$. For index set G_i, a close form solution is known to exist:

$$\mathbf{p}_{G_i} = \frac{\max(\|\mathbf{q}_{G_i}\| - \lambda, 0)}{\|\mathbf{q}_{G_i}\|}\mathbf{q}_{G_i}.$$

Thus, at each iteration, we have a closed-form solution.

C. Alternating Direction Method of Multipliers for Computing the Proximal Operator in (5)

Besides splitting the proximal operators, we can also bypass the difficulty brought by overlapping groups by introducing auxiliary variables, and reformulate (5) as:

$$\begin{aligned} \min_{\mathbf{x},\mathbf{z}} \quad & \frac{1}{2}\|\mathbf{x}-\mathbf{u}\|^2 + \lambda\sum_{i=1}^{g} w_i\|\mathbf{z}_i\| \\ \text{s.t.} \quad & \mathbf{z}_i = \mathbf{x}_{G_i}, \quad i = 1, \ldots, g \end{aligned} \tag{29}$$

We can therefore form the augmented Lagrangian as follows:

$$L_\rho(\mathbf{x},\mathbf{z},\mathbf{y}) = \frac{1}{2}\|\mathbf{x}-\mathbf{u}\|^2 + \lambda\sum_{i=1}^{g} w_i\|\mathbf{z}_i\| + \sum_{i=1}^{g}\mathbf{y}_i^T(\mathbf{z}_i - \mathbf{x}_{G_i}) + (\rho/2)\sum_{i=1}^{g}\|\mathbf{z}_i - \mathbf{x}_{G_i}\|^2.$$

The Alternating Direction Method of Multipliers (ADMM) consists of the following iterations:

$$\begin{aligned}
\mathbf{x}^{k+1} &:= \arg\min_{\mathbf{x}} L_\rho(\mathbf{x}, \mathbf{z}^k, \mathbf{y}^k) \\
\mathbf{z}^{k+1} &:= \arg\min_{\mathbf{z}} L_\rho(\mathbf{x}^{k+1}, \mathbf{z}, \mathbf{y}^k) \\
\mathbf{y}_i^{k+1} &:= \mathbf{y}_i^k + \rho(\mathbf{z}_i^{k+1} - \mathbf{x}_{G_i}^{k+1})
\end{aligned} \tag{30}$$

One nice property of ADMM is, each iterative step admits a closed-form solution. We define $\otimes$ as the point-wise product, $\odot$ as the point-wise division, $\mathbf{e}$ the p-dimensional vector with all ones, and the indicator vector $\tilde{\mathbf{e}}_i$ such that $\tilde{\mathbf{e}}_i(j) = 1$ if $j \in G_i$ and 0 otherwise. We further define $\tilde{\mathbf{y}}_i, \tilde{\mathbf{z}}_i \in \mathbb{R}^p$ such that $\tilde{\mathbf{y}}_i(G_i) = \mathbf{y}_i, \tilde{\mathbf{y}}_i(G_i^C) = 0$ and $\tilde{\mathbf{z}}_i(G_i) = \mathbf{z}_i, \tilde{\mathbf{z}}_i(G_i^C) = 0$. For updating $\mathbf{x}$, we have:

$$\frac{\partial}{\partial \mathbf{x}} L_\rho(\mathbf{x}, \mathbf{z}^k, \mathbf{y}^k) = \mathbf{x} - \mathbf{u} - \sum_{i=1}^g \tilde{\mathbf{y}}_i^k + \rho \left(\sum_{i=1}^g \tilde{\mathbf{e}}_i \right) \otimes \mathbf{x} - \rho \left(\sum_{i=1}^g \tilde{\mathbf{z}}_i^k \right)$$

and therefore,

$$\mathbf{x}^{k+1} = \left(\mathbf{u} + \sum_{i=1}^g \tilde{\mathbf{y}}_i^k + \rho \sum_{i=1}^g \tilde{\mathbf{z}}_i^k \right) \odot \left(\mathbf{e} + \rho \sum_{i=1}^g \tilde{\mathbf{e}}_i \right).$$

For updating $\mathbf{z}_i$, we use the sub-differential method: $\mathbf{z}^*$ is the optimal solution if and only if 0 belongs to the sub-differential set $\partial L_\rho(\mathbf{x}^{k+1}, \mathbf{z}^*, \mathbf{y}^k)$. Decouple the problem with respect to groups, we have:

$$0 \in \mathbf{z}_i^{k+1} - \mathbf{x}_{G_i}^{k+1} + \frac{1}{\rho}\mathbf{y}_i^k + \frac{\lambda w_i}{\rho} \partial \|\mathbf{z}_i^{k+1}\|$$

where

$$\partial \|\mathbf{z}_i^{k+1}\| = \begin{cases} \frac{\mathbf{z}_i^{k+1}}{\|\mathbf{z}_i^{k+1}\|} & \|\mathbf{z}_i^{k+1}\| \neq 0 \\ \{\mathbf{t} | \mathbf{t} \in \mathbb{R}^{|G_i|}, \|\mathbf{t}\| \leq 1\} & \|\mathbf{z}_i^{k+1}\| = 0. \end{cases}$$

Thus, we have:

$$\mathbf{z}_i^{k+1} = \frac{\max\{\|\tilde{\mathbf{x}}_{G_i}^{k+1}\| - \tilde{\lambda}_i, 0\}}{\|\tilde{\mathbf{x}}_{G_i}^{k+1}\|} \tilde{\mathbf{x}}_{G_i}^{k+1}$$

where

$$\tilde{\mathbf{x}}_{G_i}^{k+1} = \mathbf{x}_{G_i}^{k+1} - \frac{1}{\rho}\mathbf{y}_i^k, \quad \tilde{\lambda}_i = \frac{\lambda w_i}{\rho}.$$

Optimality conditions and stopping criterion The KKT conditions for (29) are primal feasibility:

$$\mathbf{z}_i^* - \mathbf{x}_{G_i}^* = 0 \tag{31}$$

and the dual feasibility:

$$\begin{aligned}
0 &= \mathbf{x}^* - \mathbf{u} - \sum_{i=1}^g \tilde{\mathbf{y}}_i^* \\
0 &\in \lambda w_i \partial \|\mathbf{z}_i^*\| + \mathbf{y}_i^*
\end{aligned} \tag{32}$$

Since $\mathbf{z}^{k+1}$ minimizes $L_\rho(\mathbf{x}^{k+1}, \mathbf{z}, \mathbf{y}^k)$, we have

$$\begin{aligned}
0 &\in \mathbf{z}_i^{k+1} - \mathbf{x}_{G_i}^{k+1} + \frac{1}{\rho}\mathbf{y}_i^k + \frac{\lambda w_i}{\rho} \partial \|\mathbf{z}_i^{k+1}\| \\
&= \frac{1}{\rho}\mathbf{y}_i^{k+1} + \frac{\lambda w_i}{\rho} \partial \|\mathbf{z}_i^{k+1}\|
\end{aligned}$$

Therefore, the second condition in the dual feasibility is always satisfied, and the optimization comes down to attaining the primal and the first dual feasibility.

Define $r_i = \mathbf{z}_i - \mathbf{x}_{G_i}$. We have $\mathbf{y}_i^{k+1} = \mathbf{y}_i^k + \rho r^{k+1}$. Since $\mathbf{x}^{k+1}$ minimizes $L_\rho(\mathbf{x}, \mathbf{z}^k, \mathbf{y}^k)$, we have

$$\begin{aligned} 0 &= \mathbf{x}^{k+1} - \mathbf{u} - \sum_{i=1}^{g} \tilde{\mathbf{y}}_i^k + \rho \left(\sum_{i=1}^{g} \tilde{\mathbf{e}}_i \right) \otimes \mathbf{x}^{k+1} - \rho \left(\sum_{i=1}^{g} \tilde{\mathbf{z}}_i^k \right) \\ &= \mathbf{x}^{k+1} - \mathbf{u} - \sum_{i=1}^{g} \tilde{\mathbf{y}}_i^{k+1} + \rho \left(\sum_{i=1}^{g} (\tilde{\mathbf{z}}_i^{k+1} - \tilde{\mathbf{z}}_i^k) \right) \end{aligned}$$

or equivalently,

$$\rho \left(\sum_{i=1}^{g} (\tilde{\mathbf{z}}_i^k - \tilde{\mathbf{z}}_i^{k+1}) \right) = \mathbf{x}^{k+1} - \mathbf{u} - \sum_{i=1}^{g} \tilde{\mathbf{y}}_i^{k+1}.$$

This means that the quantity

$$s^{k+1} = \rho \left(\sum_{i=1}^{g} (\tilde{\mathbf{z}}_i^{k+1} - \tilde{\mathbf{z}}_i^k) \right)$$

can be viewed as the residual for the first dual feasibility. Paired with the primal residual r^{k+1}, we can terminate the algorithm by checking whether they are small enough.

D. Alternating Direction Method of Multipliers for Solving Overlapping Group Lasso

Using the least squared loss and observing that ℓ_1 norm is a special case of (2), we can rewrite the overlapping group lasso problem (1) as:

$$\begin{aligned} \min_{\mathbf{x},\mathbf{z}} \quad & \frac{1}{2}\|A\mathbf{x} - \mathbf{u}\|^2 + \lambda \sum_{i=1}^{g} w_i \|\mathbf{z}_i\| \\ \text{s.t.} \quad & \mathbf{z}_i = \mathbf{x}_{G_i} \end{aligned}$$

We can therefore form the augmented Lagrangian as follows:

$$L_\rho(A\mathbf{x}, \mathbf{z}, \mathbf{y}) = \frac{1}{2}\|A\mathbf{x} - \mathbf{u}\|^2 + \lambda \sum_{i=1}^{g} w_i \|\mathbf{z}_i\| + \sum_{i=1}^{g} \mathbf{y}_i^T (\mathbf{z}_i - \mathbf{x}_{G_i}) + (\rho/2) \sum_{i=1}^{g} \|\mathbf{z}_i - \mathbf{x}_{G_i}\|^2$$

The Alternating Direction Method of Multipliers (ADMM) consists of the iterations:

$$\begin{aligned} \mathbf{x}^{k+1} &:= \arg\min_{\mathbf{x}} L_\rho(\mathbf{x}, \mathbf{z}^k, \mathbf{y}^k) \\ \mathbf{z}^{k+1} &:= \arg\min_{\mathbf{z}} L_\rho(\mathbf{x}^{k+1}, \mathbf{z}, \mathbf{y}^k) \\ \mathbf{y}_i^{k+1} &:= \mathbf{y}_i^k + \rho(\mathbf{z}_i^{k+1} - \mathbf{x}_{G_i}^{k+1}) \end{aligned}$$

We define $\mathbf{e}$ the p-dimensional vector with all ones, and the indicator vector $\tilde{\mathbf{e}}_i$ such that $\tilde{\mathbf{e}}_i(j) = 1$ if $j \in G_i$ and 0 otherwise. We further define $\tilde{\mathbf{y}}_i, \tilde{\mathbf{z}}_i \in \mathbb{R}^p$ such that $\tilde{\mathbf{y}}_i(G_i) = \mathbf{y}_i, \tilde{\mathbf{y}}_i(G_i^C) = 0$ and $\tilde{\mathbf{z}}_i(G_i) = \mathbf{z}_i, \tilde{\mathbf{z}}_i(G_i^C) = 0$. For updating $\mathbf{x}$, we have:

$$\frac{\partial}{\partial \mathbf{x}} L_\rho(\mathbf{x}, \mathbf{z}^k, \mathbf{y}^k) = A^T A\mathbf{x} - A^T\mathbf{u} - \sum_{i=1}^{g} \tilde{\mathbf{y}}_i^k + \rho \left(\sum_{i=1}^{g} \tilde{\mathbf{e}}_i \right) \otimes \mathbf{x} - \rho \left(\sum_{i=1}^{g} \tilde{\mathbf{z}}_i^k \right)$$

and therefore, the update for $\mathbf{x}^{k+1}$ involves solving the following linear system:

$$\tilde{A}\mathbf{x} = \tilde{b},$$

where

$$\begin{aligned} \tilde{A} &= A^T A + \text{diag}\left(\rho \sum_{i=1}^{g} \tilde{\mathbf{e}}_i \right) \\ \tilde{b} &= A^T\mathbf{u} + \sum_{i=1}^{g} \tilde{\mathbf{y}}_i^k + \rho \sum_{i=1}^{g} \tilde{\mathbf{z}}_i^k \end{aligned}$$

Please note that, for a given problem, $\tilde{A}$ is fixed. Therefore, for moderate size problems, we can save the Cholesky decomposition of $\tilde{A}$ such that the linear system can be solved very fast in each iteration. For large (high dimensional) problems, the storage of $\tilde{A}$ might not be practical. However, since we can calculate $\tilde{A}\mathbf{x}$ without having to calculate $\tilde{A}$, methods such as Preconditioned Conjugate Gradient (PCG) or BB method can be applied.

For updating $\mathbf{z}_i$, we use the sub-differential method: $\mathbf{z}^*$ is the optimal solution if and only if 0 belongs to the sub-differential set $\partial L_\rho(\mathbf{x}^{k+1}, \mathbf{z}^*, \mathbf{y}^k)$. Decouple the problem with respect to groups, we have:

$$0 \in \mathbf{z}_i^{k+1} - \mathbf{x}_{G_i}^{k+1} + \frac{1}{\rho}\mathbf{y}_i^k + \frac{\lambda w_i}{\rho}\partial\|\mathbf{z}_i^{k+1}\|$$

where

$$\partial\|\mathbf{z}_i^{k+1}\| = \begin{cases} \frac{\mathbf{z}_i^{k+1}}{\|\mathbf{z}_i^{k+1}\|} & \|\mathbf{z}_i^{k+1}\| \neq 0 \\ \{\mathbf{t} | \mathbf{t} \in \mathbb{R}^{|G_i|}, \|\mathbf{t}\| \leq 1\} & \|\mathbf{z}_i^{k+1}\| = 0. \end{cases}$$

Thus, we have:

$$\mathbf{z}_i^{k+1} = \frac{\max\{\|\tilde{\mathbf{x}}_{G_i}^{k+1}\| - \tilde{\lambda}_i, 0\}}{\|\tilde{\mathbf{x}}_{G_i}^{k+1}\|}\tilde{\mathbf{x}}_{G_i}^{k+1}$$

where

$$\tilde{\mathbf{x}}_{G_i}^{k+1} = \mathbf{x}_{G_i}^{k+1} - \frac{1}{\rho}\mathbf{y}_i^k, \quad \tilde{\lambda}_i = \frac{\lambda w_i}{\rho}.$$

E. Additional Experiments

To illustrate the scalability of our proposed method, we also evaluate our method using numbers (p) of genes larger than 2000. The results are summarized in Table 2.

Table 2: Scalability study of the proposed FoGLasso algorithm under different numbers (p) of genes involved. The reported results are the total computational time (seconds) including all nine regularization parameter values.

p	3000	4000	5000	6000	7000	8141
pathways	37.6	48.3	62.5	68.7	86.2	99.7
edges	58.8	84.8	102.7	140.8	173.3	247.8