Exclusive Feature Learning on Arbitrary Structures via $\ell_{1,2}$-norm

Deguang Kong\(^1\), Ryohei Fujimaki\(^2\), Ji Liu\(^3\), Feiping Nie\(^1\), Chris Ding\(^1\)

\(^1\)Dept. of Computer Science, University of Texas Arlington, TX, 76019; \(^2\)NEC Laboratories America, Cupertino, CA, 95014; \(^3\)Dept. of Computer Science, University of Rochester, Rochester, NY, 14627

Email: doogkong@gmail.com, rfujimaki@nec-labs.com, jliu@cs.rochester.edu, feipingnie@gmail.com, chqding@uta.edu

Abstract

Group LASSO is widely used to enforce the structural sparsity, which achieves the sparsity at the inter-group level. In this paper, we propose a new formulation called “exclusive group LASSO”, which brings out sparsity at intra-group level in the context of feature selection. The proposed exclusive group LASSO is applicable on any feature structures, regardless of their overlapping or non-overlapping structures. We provide analysis on the properties of exclusive group LASSO, and propose an effective iteratively re-weighted algorithm to solve the corresponding optimization problem with rigorous convergence analysis. We show applications of exclusive group LASSO for uncorrelated feature selection. Extensive experiments on both synthetic and real-world datasets validate the proposed method.

1 Introduction

Structure sparsity induced regularization terms [1, 8] have been widely used recently for feature learning purpose, due to the inherent sparse structures of the real world data. Both theoretical and empirical studies have suggested the powerfulness of structure sparsity for feature learning, e.g., Lasso [24], group LASSO [29], exclusive LASSO [31], fused LASSO [25], and generalized LASSO [22]. To make a compromise between the regularization term and the loss function, the sparse-induced optimization problem is expected to fit the data with better statistical properties. Moreover, the results obtained from sparse learning are easier for interpretation, which give insights for many practical applications, such as gene-expression analysis [9], human activity recognition [14], electronic medical records analysis [30], etc.

Motivation Of all the above sparse learning methods, group LASSO [29] is known to enforce the sparsity on variables at an inter-group level, where variables from different groups are competing to survive. Our work is motivated from a simple observation: in practice, not only features from different groups are competing to survive (i.e., group LASSO), but also features in a seemingly cohesive group are competing to each other. The winner features in a group are set to large values, while the loser features are set to zeros. Therefore, it leads to sparsity at the intra-group level. In order to make a distinction with standard LASSO and group LASSO, we called it “exclusive group LASSO” regularizer. In “exclusive group LASSO” regularizer, intra-group sparsity is achieved via $\ell_1$ norm, while inter-group non-sparsity is achieved via $\ell_2$ norm. Essentially, standard group LASSO achieves sparsity via $\ell_2$ norm, while the proposed exclusive group LASSO achieves sparsity via $\ell_{1,2}$ norm. An example of exclusive group LASSO is shown in Fig. (1) via Eq. (2). The significant difference from the standard LASSO is to encourage similar features in different groups to co-exist (Lasso usually allows only one of them surviving). Overall, the exclusive group LASSO regularization encourages intra-group competition but discourages inter-group competition.
Figure 1: Explanation of differences between group LASSO and exclusive group LASSO. Group setting: \(G_1 = \{1, 2\}, G_2 = \{3, 4\}, G_3 = \{5, 6, 7\}\). Group LASSO solution of Eq.\((3)\) at \(\lambda = 2\) using least square loss is: \(w = [0.0337; 0.0891; 0; -0.2532; 0.043; 0.011]\). Exclusive group LASSO solution of Eq.\((2)\) at \(\lambda = 10\) is: \(w = [0.0749; 0; -0.0713; -0.1888; 0; 0]\). Clearly, group LASSO introduces sparsity at an inter-group level, whereas exclusive LASSO enforces sparsity at an intra-group level.

We note that “exclusive LASSO” was first used in \([31]\) for multi-task learning. Our “exclusive group LASSO” work, however, has clear difference from \([31]\): (1) we give a clear physical intuition of “exclusive group LASSO”, which leads to sparsity at an intra-group level (Eq.\((2)\)), whereas \([31]\) focuses on “Exclusive LASSO” problem in a multi-task setting; (2) we target a general “group” setting which allows arbitrary group structure, which can be easily extended to multi-task/multi-label learning. The main contributions of this paper include: (1) we propose a new formulation of “exclusive group LASSO” with clear physical meaning, which allow any arbitrary structure on feature space; (2) we propose an effective iteratively re-weighted algorithm to tackle non-smooth “exclusive group LASSO” term with rigorous convergence guarantees. Moreover, an effective algorithm is proposed to handle both non-smooth \(\ell_1\) and exclusive group LASSO term (Lemma \([4.1]\)); (3) The proposed approach is validated via experiments on both synthetic and real data sets, specifically for uncorrelated feature selection problems.

Notation Throughout the paper, matrices are written as boldface uppercase, vectors are written as boldface lowercase, and scalars are denoted by lower-case letters \((a, b)\). \(n\) is the number of data points, \(p\) is the dimension of data, \(K\) is the number of class in a dataset. For any vector \(w \in \mathbb{R}^p, \ell_q\) norm of \(w\) is: \(\|w\|_q = \left(\sum_{i=1}^{p} |w_i|^q\right)^{\frac{q}{2}}\) for \(q \in (0, \infty)\). A group of variables is a subset \(g \subset \{1, 2, \cdots, p\}\). Thus, the set of possible groups is the power set of \(\{1, 2, \cdots, p\}\): \(\mathcal{G} \in \mathcal{P}(\{1, 2, \cdots, p\})\). \(G_g \in \mathcal{P}(\{1, 2, \cdots, p\})\) denotes a set of group \(g\), which is known in advance depending on applications. If two groups have one or more overlapped variables, we say that they are overlapped. For any group variable \(w_{G_g} \in \mathbb{R}^p\), only the entries in the group \(g\) are preserved which are the same as those in \(w\), while the other entries are set to zeros. For example, if \(G_g = \{1, 2, 4\}\), \(w_{G_g} = [w_1, w_2, 0, w_4, 0, 0, \cdots, 0]\), then \(\|w_{G_g}\|_2 = \sqrt{w_1^2 + w_2^2 + w_4^2}\). Let \(supp(w) \subset \{1, 2, \cdots, p\}\) be a set which \(w_i \neq 0\), and \(zero(w) \subset \{1, 2, \cdots, p\}\) be a set which \(w_i = 0\). Clearly, \(zero(w) = \{1, 2, \cdots, p\} \setminus supp(w)\). Let \(\nabla f(w)\) be gradient of \(f\) at \(w \in \mathbb{R}^p\), for any differentiable function \(f: \mathbb{R}^p \rightarrow \mathbb{R}\).

2 Exclusive group LASSO

Let \(G\) be a group set, the exclusive group LASSO penalty is defined as:

\[
\forall w \in \mathbb{R}^p, \quad \Omega_{Eg}^\mathcal{G}(w) = \sum_{g \in \mathcal{G}} \|w_{G_g}\|_1.
\]

When the groups of \(g\) form different partitions of the set of variables, \(\Omega_{Eg}^\mathcal{G}\) is a \(\ell_1/\ell_2\) norm penalty. A \(\ell_2\) norm is enforced on different groups, while in each group, \(\ell_1\) norm is used to make a sum over each intra-group variable. Minimizing such a convex risk function often leads to a solution that some entries in a group are zeros. For example, for a group \(G_g = \{1, 2, 4\}\), there exists a solution \(w\), such that \(w_1 = 0, w_2 \neq 0, w_4 \neq 0\). A concrete example is shown in Fig\([1]\) in which we solve:

\[
\min_{w \in \mathbb{R}^p} J_1(w), \quad J_1(w) = f(w) + \lambda \Omega_{Eg}^\mathcal{G}(w),
\]

using least square loss function \(f(w) = \|y - X^Tw\|_2^2\). As compared to standard group LASSO \([29]\) solution of Eq.\((3)\),
In the following, we derive a very efficient yet simple algorithm. Moreover, the proposed algorithm is a generic algorithm, which allows arbitrary structure on feature space, irrespective of specific feature structures [10], e.g., linear structure [28], tree structure [15], graph structure [7], etc.

\[
f(w) + \lambda \sum_g \|w_{G_g}\|_2. \tag{3}
\]

We observe that group LASSO introduces sparsity at an inter-group level, whereas exclusive LASSO enforces sparsity at an intra-group level.

**Analysis of exclusive group LASSO** For each group \( g \), feature index \( u \in \text{supp}(g) \) will be non-zero. Let \( v_g \in \mathbb{R}^p \) be a variable which preserves the values of non-zero index for group \( g \). Consider all groups, for optimization goal \( w \), we have \( \text{supp}(w) = \bigcup_g \text{supp}(v_g). \) (1) For non-overlapping case, different groups form a partition of feature set \( \{1, 2, \ldots, p\} \), and there exists a unique decomposition of \( w = \sum_g v_g. \) Since there is not any common elements for any two different groups \( G_u \) and \( G_v \), i.e., \( \text{supp}(w_{G_u}) \cap \text{supp}(w_{G_v}) = \phi \) thus it is easy to see: \( v_g = w_{G_g} \), \( \forall g \in G. \) (2) However, for overlapping groups, there could be element sets \( I \subseteq (G_u \cap G_v) \), and therefore, different groups \( G_u \) and \( G_v \) may have opposite effects to optimize the features in set \( I. \) For feature \( i \in I \), it is prone to give different values if optimized separately, i.e., \( \{w_{G_u}\}_i \neq \{w_{G_v}\}_i \). For example, \( G_u = [1, 2], G_v = [2, 3] \), whereas group \( u \) may require \( w_2 = 0 \) and group \( v \) may require \( w_2 \neq 0 \). Thus, there will be many possible combinations of feature values, and it leads to: \( \Omega^G_{E_g} = \inf_{v_g} \sum_g \|v_g\|_1. \) Further, if some groups are overlapped, the final zeros sets will be a subset of unions of all different groups, \( \text{zero}(w) \subseteq \bigcap_g \text{zero}(v_g). \)

**Illustration of Geometric shape of exclusive LASSO** Figure 2 shows the geometric shape for both norms in \( \mathbb{R}^3 \) with different group settings, where in (a): \( G_1 = [1, 2], G_2 = [3] \); and in (b): \( G_1 = [1, 2], G_2 = [2, 3] \). For the non-overlapping case, variables \( w_1, w_2 \) usually can not be zero simultaneously. In contrast, for the overlapping case, variable \( w_2 \) cannot be zero unless both groups \( G_1 \) and \( G_2 \) require \( w_2 = 0 \).

**Properties of exclusive LASSO** The regularization term of Eq.1 is a convex formulation. If \( \bigcup_{g \in G} \{1, 2, \ldots, p\} \), then \( \Omega^G_{E_g} := \sqrt{\Omega^G_{E_g}} \) is a norm. See Appendix for proofs.

**3 An effective algorithm for solving \( \Omega^G_{E_g} \) regularizer**

The challenge of solving Eq.1 is to tackle the exclusive group LASSO term, where \( f(w) \) can be any convex loss function w.r.t \( w \). It is generally felt that exclusive group LASSO term is much more difficult to solve than the standard LASSO term (shrinkage thresholding). Existing algorithm can formulate it as a quadratic programming problem [19], which can be solved by interior point method or active set method. However, the computational cost is expensive, which limits its use in practice. Recently, a primal-dual algorithm [27] is proposed to solve the similar problem, which casts the non-smooth problem into a min-max problem. However, the algorithm is a gradient descent type method and converges slowly. Moreover, the algorithm is designed for multi-task learning problem, and cannot be applied directly for exclusive group LASSO problem with arbitrary structures. In the following, we first derive a very efficient yet simple algorithm. Moreover, the proposed algorithm is a generic algorithm, which allows arbitrary structure on feature space, irrespective of specific feature structures [10], e.g., linear structure [28], tree structure [15], graph structure [7], etc.
Theoretical analysis guarantees the convergence of algorithm. Moreover, the algorithm is easy to implement and ready to use in practice.

**Key idea** The idea of the proposed algorithm is to find an auxiliary function for Eq. (1) which can be easily solved. Then the updating rules for \( w \) is derived. Finally, we prove the solution is exactly the optimal solution we are seeking for the original problem. Since it is a convex problem, the optimal solution is the global optimal solution.

**Procedure** Instead of directly optimizing Eq. (1), we propose to optimize the following objective (the reasons will be seen immediately below), i.e.,

\[
J_2(w) = f(w) + \lambda w^T F w,
\]

where \( F \in \mathbb{R}^{p \times p} \) is a diagonal matrices which encodes the exclusive group information, and its diagonal element is given by \( F_{ii} = \left( \sum_g (I_{G_g})_i \Vert w_{G_g} \Vert_1 \right) / \Vert w_i \Vert \).

Let \( I_{G_g} \in \{0, 1\}^{p \times 1} \) be group index indicator for group \( g \in G \). For example, group \( G_1 \) is \( \{1, 2, \cdots, p\} \), then \( I_{G_1} = [1, 1, 0, \cdots, 0] \). Thus the group variable \( w_{G_g} \) can be explicitly expressed as \( w_{G_g} = \text{diag}(I_{G_g}) \times w \).

Note computation of \( F \) depends on \( w \), thus minimization of \( w \) depends on both \( F \). In the following, we propose an efficient iteratively re-weighted algorithm to find out the optimal global solution for \( w \), where in each iteration, \( w \) is updated along the gradient descent direction. This process is iterated until the algorithm converges. Taking the derivative of Eq.(4) w.r.t \( w \) and set \( \partial J_2 / \partial w = 0 \). We have

\[
\nabla_w f(w) + 2\lambda F w = 0.
\]

Then the complete algorithm is:

1. Updating \( w^t \) via Eq.(6);
2. Updating \( F^t \) via Eq.(5).

The above two steps are iterated until the algorithm converges. The convergence is implied by the monotonic decreasing property of these two steps shown in the following.

**3.1 Convergence Analysis**

In the following, we prove the convergence of algorithm.

**Theorem 3.1.** Under the updating rule of Eq. (6), \( J_1(w^{t+1}) - J_1(w^t) \leq 0 \).

The proof is provided in Appendix.

**Discussion** We note reweighted strategy [26] was also used in solving problems like zero-norm of the parameters of linear models. However, it cannot be directly used to solve “exclusive group LASSO” problem proposed in this paper, and cannot handle arbitrary structures on feature space.

**4 Uncorrelated feature learning via exclusive group LASSO**

**Motivation** It is known that in Lasso-type (including elastic net) [24, 32] variable selection, variable correlations are not taken into account. Therefore, some strongly correlated variables tend to be in or out of the model together. However, in practice, feature variables are often correlated. See an
Lemma 4.1. The optimal solution to Eq.(10) is the optimal solution to Eq.(11), where
\[ u = \text{arg min}_x \frac{1}{2} ||x - a||^2 + \alpha ||x||_1 = \text{sgn}(a)(a - \alpha)_+ , \]
and \( \text{sgn}(\cdot), \text{SGN}(\cdot) \) are the operators defined in the component fashion: if \( v > 0 \), \( \text{sgn}(v) = 1, \text{SGN}(v) = \{1\} \); else if \( v = 0 \), \( \text{sgn}(v) = 0, \text{SGN}(v) = [-1,1] \); else if \( v < 0 \), \( \text{sgn}(v) = -1, \text{SGN}(v) = \{-1\} \).

The proof is provided in Appendix.
5 Experiment Results

To validate the effectiveness of our method, we first conduct experiment using Eq.(7) on two synthetic datasets, and then show experiments on real-world datasets.

5.1 Synthetic datasets

(1) Linear-correlated features. Let data $X^1 = [x_1^1, x_2^1, \cdots, x_n^1] \in \mathbb{R}^{p \times n}$, where each data $x_i^1 \sim \mathcal{N}(0_{p \times 1}, I_{p \times p})$, $X^2 = [x_1^2, x_2^2, \cdots, x_n^2] \in \mathbb{R}^{p \times n}$, where each data $x_i^2 \sim \mathcal{N}(0_{p \times 1}, I_{p \times p})$, $I$ is identity matrix. We generate a group of $p$-features, which is a linear combination of features in $X_1$ and $X_2$, i.e., $X_j^i = 0.5(X_1^j + X_2^j) + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma)$, $\sigma = 0.1$. Construct data matrix $X = [X_1; X_2; X_3]$, clearly, $X \in \mathbb{R}^{3p \times n}$. Features in the dimension $[2p + 1, 3p]$ are highly correlated with features in dimension $[1, p]$ and $[p + 1, 2p]$. Let $w^1 \in \mathbb{R}^p$, where each $w_i^1 \sim \text{Uniform}(-0.5, 0.5)$, and $w^2 \in \mathbb{R}^p$, where each $w_i^2 \sim \text{Uniform}(-0.5, 0.5)$. Let $\hat{w} = [w_1^1; w_2^1; 0_{p \times 1}]$. We generate predicator $y \in \mathbb{R}^n$, and $y = \hat{w}^T X + \epsilon_g$, where $(\epsilon_g)_i \sim \mathcal{N}(0, 0.1)$.

We solve Eq.(7) using current $y$ and $X$ with least square loss. The group settings are: $(i, p+i, 2p+i)$, for $1 \leq i \leq p$. We compare the computed $w^*$ against ground truth solution $\hat{w}$ and plain LASSO solution (i.e., $\beta = 0$ in Eq.(7)). We use the root mean square error (RMSE) and mean absolute error (MAE) error to evaluate the differences of values predicted by a model and the values actually observed. We generate $n = 1000$ data, with $p = [120, 140, \cdots, 220, 240]$ and do 5-fold cross validation. Generalization error of RMSE and MAE are shown in Figures 3(a) and 3(b). Clearly, our approach outperforms standard LASSO solution and exactly recovers the true features.

(2) Correlated features on Hub structure. Let data $X = [X^1; X_2^2; \cdots; X^B] \in \mathbb{R}^{n \times qB}$, where each block $X^b = [X_{11}^b; X_{12}^b; \cdots; X_{n1}^b] \in \mathbb{R}^{n \times p}$, $1 \leq b \leq B$, $q = p \times B$. In each block, for each data point $1 \leq i \leq n$, $X_{1i}^b = \frac{1}{B} \sum_{2 \leq j \leq B} X_{ji}^b + \frac{1}{B} z_i + \epsilon_i$, where $X_{ji}^b \sim \mathcal{N}(0, 1)$, $z_i \sim \mathcal{N}(0, 1)$ and $\epsilon_i \sim \mathcal{N}(0, 0.1)$. Let $w^1, w^2, \cdots, w^B \in \mathbb{R}^p$, where $w^b = [w_{1i}^b; 0_{p \times 1}]^T$, where $w_{1i}^b \sim \mathcal{N}(0, 0.5)$. Let $\hat{w} = [w_1^1; w_2^1; \cdots; w_B^1]$, we generate predicator $y \in \mathbb{R}^n$, and $y = \hat{w}^T X + \epsilon_g$, where $(\epsilon_g)_i \sim \mathcal{N}(0, 0.1)$.

The group settings are: $(b - 1) \times p + 1, \cdots, b \times p)$, for $1 \leq b \leq B$. We generate $n = 1000$ data, $B = 10$, with varied $p = [20, 21, \cdots, 24, 25]$ and do 5-fold cross validation. Generalization error of RMSE and MAE are shown in Figs 3(c)-(d). Clearly, our approach outperforms standard LASSO solution, and recovers the exact features.
5.2 Real-world datasets

To validate the effectiveness of proposed method, we perform feature selection via proposed uncorrelated feature learning framework of Eq.(7) on 4 datasets (shown in Table 1), including 2 UCI datasets: isolet [6], ionosphere [5], 1 image dataset: mnist with only figure “0” and “1” [16], and 1 biology dataset: Leuml [13].

We perform classification tasks on these different datasets. The compared methods include: proposed method of Eq.(7) (shown as Exclusive), plain LASSO, ReliefF [21], F-statistics [3]. We use logistic regression as the loss function in our method and plain LASSO method. In our method, parameter $\alpha$, $\beta$ are tuned to select different numbers of features. Exclusive LASSO groups are set according to feature correlations (i.e., threshold $\theta$ is set to 0.90 in Eq.8). After the specific number of features are selected, we feed them into support vector machine (SVM) with linear kernel, and classification results with different number of selected features are shown in Fig.(3).

A first glance at the experimental results indicates the better performance of our method as compared to plain LASSO. Moreover, our method is also generally better than the other two popularly used feature selection methods, such as ReliefF and F-statistics. The experiment result also further confirms our intuition: elimination of correlated features is really helpful for feature selection and thus improves the classification performance. Because $\ell_1$, $\ell_2$, $\ell_\infty$, or non-convex feature learning via $\ell_p$, $\ell_\infty$([20], $\ell_{2,1}$ [12, 18]) are designed for multi-task or multi-label feature learning, thus we do not compare against these methods.

We show the running time of plain exclusive LASSO and exclusive group LASSO ($\alpha > 0$ in Eq.7) in the following table. We run different algorithms on a Intel i5-3317 CPU, 1.70GHz, 8GRAM desktop.

We show the running time of plain exclusive LASSO and exclusive group LASSO ($\alpha > 0$ in Eq.7) in the following table. We run different algorithms on a Intel i5-3317 CPU, 1.70GHz, 8GRAM desktop.

Further, we list the mean and variance of classification accuracy of different algorithms in the following table, using 50% of all the features. Compared methods include (1) Lasso (L1); (2) Plain exclusive group LASSO ($\alpha = 0$ in Eq.7); (3) Exclusive group LASSO ($\alpha > 0$ in Eq.7).

The above experiment results indicate that the advantage of our method (exclusive group LASSO) over plain LASSO comes from the exclusive LASSO term. The experiment results also suggest that the plain exclusive LASSO performs very similar to LASSO. However, the exclusive group LASSO ($\alpha > 0$ in Eq.7) performs definitely better than both standard LASSO and plain exclusive LASSO (1%-4% performance improvement). The exclusive LASSO regularizer eliminates the correlated and redundant features.

We show the running time of plain exclusive LASSO and exclusive group LASSO ($\alpha > 0$ in Eq.7) in the following table. We run different algorithms on a Intel i5-3317 CPU, 1.70GHz, 8GRAM desktop.

The above experiment results indicate that the computational cost of exclusive group LASSO is slightly higher than that of plain exclusive LASSO. The reason is that, the solution to exclusive group LASSO is given by simple thresholding on the plain exclusive LASSO result. This further confirms our theoretical analysis results shown in Lemma 4.1.

6 Conclusion

In this paper, we propose a new formulation called “exclusive group LASSO” to enforce the sparsity for features at an intra-group level. We investigate its properties and propose an effective algorithm with rigorous convergence analysis. We show applications for uncorrelated feature selection, which indicate the good performance of proposed method. Our work can be easily extended for multi-task or multi-label learning.

Acknowledgement The majority of the work was done during the internship of the first author at NEC Laboratories America, Cupertino, CA.
Appendix

Proof of a valid norm of $\Omega_E^G$: Note if $\Omega_E^G(w) = 0$, then $w = 0$. For any scalar $a$, $\Omega_E^G(a)w = |a|\Omega_E^G(w)$. This proves absolute homogeneity and zero property hold. Next we consider triangle inequality. Consider $w, \tilde{w} \in \mathbb{R}^p$. Let $v_g$ and $\tilde{v}_g$ be optimal decomposition of $w, \tilde{w}$ such that $\Omega_E^G(w) = \sqrt{\sum_g \|v_g\|^2}$ and $\Omega_E^G(\tilde{w}) = \sqrt{\sum_g \|\tilde{v}_g\|^2}$. Since $v_g + \tilde{v}_g$ is a decomposition of $w + \tilde{w}$, we have:

$$\Omega_E^G(w + \tilde{w}) \leq \sqrt{\sum_g \|v_g + \tilde{v}_g\|^2} \leq \sqrt{\sum_g \|v_g\|^2} + \sqrt{\sum_g \|\tilde{v}_g\|^2} = \Omega_E^G(w) + \Omega_E^G(\tilde{w}).$$

To prove Theorem 3.1 we need two lemmas.

Lemma 6.1. Under the updating rule of Eq.(6), $J_2(w^{t+1}) < J_2(w^t)$.

Lemma 6.2. Under the updating rule of Eq.(6),

$$J_1(w^{t+1}) = J_1(w^t).$$

Proof of Theorem 3.1 From Lemma 6.1 and Lemma 6.2, it is easy to see $J_1(w^{t+1}) - J_1(w^t) \leq 0$. This completes the proof.

Proof of Lemma 6.1 Eq.(13) is a convex function, and optimal solution of Eq.(6) is obtained by taking derivative $\frac{\partial J_2}{\partial w} = 0$, thus obtained $w^*$ is global optimal solution, $J_2(w^{t+1}) < J_2(w^t)$.

Before proof of Lemma 6.2, we need the following Proposition.

Proposition 6.3. $w^TFw = \sum_{g=1}^{G} (\|w_g\|_2^2)^2$.

Proof of Lemma 6.2 Let $\Delta = \text{LHS} - \text{RHS}$ of Eq.(13). We have $\Delta$, where

$$\Delta = \sum_g \|w_g^{t+1}\|^2 - \sum_{i,g} \frac{(1-g_i)\|w_g^*\|^2}{|w_i|} (w_i^{t+1})^2 + \sum_{i,g} \frac{(1-g_i)\|w_g^*\|^2}{|w_i|} (w_i^t)^2 - \sum_g \|w_g^t\|^2,$$

$$= \sum_g \|w_g^{t+1}\|^2 - \sum_{i,g} \frac{(1-g_i)\|w_g^*\|^2}{|w_i|} (w_i^{t+1})^2 = \sum_g \left[ \left( \sum_{i \in G_g} |w_i^{t+1}| \right)^2 - \left( \sum_{i \in G_g} |w_i^t| \right)^2 \right].$$

$$= \sum_g \left[ (\sum_{i \in G_g} a_i b_i)^2 - (\sum_{i \in G_g} a_i^2)(\sum_{i \in G_g} b_i^2) \right] \leq 0,$$

where $a_i = \frac{|w_i^{t+1}|}{|w_i^t|}, b_i = \sqrt{|w_i^t|}$. Due to proposition 6.3, Eq.(14) is equivalent to Eq.(15). Eq.(16) holds due to Cauchy inequality [23]; for any scalar $a_i, b_i, (\sum_i a_i^2)(\sum_i b_i^2) \leq (\sum_i a_i^2)(\sum_i b_i^2)$.

Proof of Lemma 6.1 For notation simplicity, let $\Omega_{Eg}^G(w) = \sum_g \|w_g\|_2^2$. Let $w^*$ be the optimal solution to Eq.(11), then we have

$$0 \in w^* - u + \beta \partial \Omega_{Eg}^G(w^*).$$

In order to prove that $w^*$ is also the global optimal solution to Eq.(10), i.e.,

$$0 \in w^* - u + \alpha \text{SGN}(w^*) + \beta \partial \Omega_{Eg}^G(w^*).$$

First, from Eq.(12), we have $0 \in u - a + \alpha \text{SGN}(u)$, and this leads to $u \in a - \alpha \text{SGN}(u)$. According to the definition of $\Omega_{Eg}^G(w)$, from Eq.(11), it is easy to verify that (1) if $u_i = 0$, then $w_i = 0$; (2) if $u_i \neq 0$, then $\text{sign}(w_i) = \text{sign}(u_i)$ and $|w_i| \leq |u_i|$. This indicates that $\text{SGN}(u) \subset \text{SGN}(w)$, and thus

$$u \in a - \alpha \text{SGN}(w).$$

Put Eqs.(17)[19] together, and this exactly recovers Eq.(18), which completes the proof.

Note the derivation needs Cauchy inequality [23], where for any scalar $a_i, b_i$, $(\sum_a a_i b_i)^2 \leq (\sum_a a_i^2)(\sum b_i^2)$. Let $a_g = \|v_g\|_1, b_g = \|\tilde{v}_g\|_1$, then we can get the inequality.
References