On Benefits of Selection Diversity via Bilevel Exclusive Sparsity

Haichuan Yang[†], Yijun Huang[†], Lam Tran[†], Ji Liu[†♭], and Shuai Huang[‡]
[†]Department of Computer Science, University of Rochester
[♭] Goergen Institute for Data Science, University of Rochester
[‡]Department of Industrial and Systems Engineering, University of Washington
{hyang1990.cs, huangyj0, lam.c.tran, ji.liu.uwisc, shuai.huang.ie}@gmail.com

Abstract

Sparse feature (dictionary) selection is critical for various tasks in computer vision, machine learning, and pattern recognition to avoid overfitting. While extensive research efforts have been conducted on feature selection using sparsity and group sparsity, we note that there has been a lack of development on applications where there is a particular preference on diversity. That is, the selected features are expected to come from different groups or categories. This diversity preference is motivated from many real-world applications such as advertisement recommendation, privacy image classification, and design of survey.

In this paper, we proposed a general bilevel exclusive sparsity formulation to pursue the diversity by restricting the overall sparsity and the sparsity in each group. To solve the proposed formulation that is NP hard in general, a heuristic procedure is proposed. The main contributions in this paper include: 1) A linear convergence rate is established for the proposed algorithm; 2) The provided theoretical error bound improves the approaches such as L_1 norm and L_0 types methods which only use the overall sparsity and the quantitative benefits of using the diversity sparsity is provided. To the best of our knowledge, this is the first work to show the theoretical benefits of using the diversity sparsity; 3) Extensive empirical studies are provided to validate the proposed formulation, algorithm, and theory.

1. Introduction

Nowadays people can easily extract tons of features in computer vision applications, for example, the well-known open source library VLFeat [34] provides a large collection of feature extraction algorithms, while the number of labeled samples are usually limit. Without restricting the number of active features, the learned model will overfit the training data and dramatically degrades the performance on testing data. Sparse feature (dictionary) selection is one of key steps in various tasks in computer vision, machine learning, and pattern recognition to overcome the overfitting issue (e.g., visual-tracking [47], face recognition [38], sparse coding for image classification [37], joint restoration and recognition [45]). Extensive research efforts have been conducted on sparse feature selection by explicitly or implicitly restricting the number of active features (e.g., LASSO [31]) or the number of active feature group (e.g., group LASSO [42]), where the assumption is that active features (or active groups of features) should be sparse.

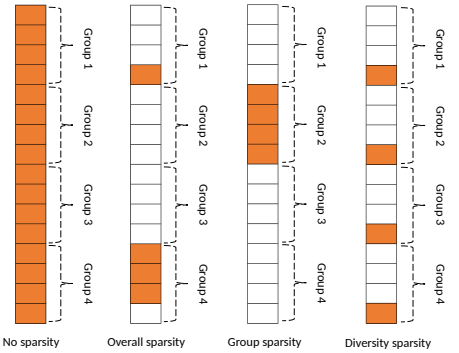


Figure 1. Illustration of different sparsity preferences. The orange color indicates the active (or selected) feature. "No sparsity" potentially selects all features; "overall sparsity" (commonly used sparsity such as L_1 norm and L_0 norm regularization) only restricts the number of selected features without any restriction on the structure; "group sparsity" prefers selects features staying in the same group; and "diversity sparsity" restricts selected features in different groups.

In the paper, we are interested in the sparse feature selection with a particular preference on diversity, that is, the selected features are expected from different groups or categories. Figure 1 provides an illustration for the different sparsity preference that are commonly used and the diversity preference considered in this paper.

This diversity preference is motivated from many real-world applications such as privacy image classification, design of survey, and advertisement recommendation:

- (**Private image detection**) Private image detection is a new emerging task with the thriving activity of posting photos on social media. Sometimes the user uploads photos with the intent to limit the viewership to only friends and family but not to the general public due to personal information in the photos. The reason why an image is considered as private may vary, as shown in Figure 4. Different privacy type is usually better to indicate with different type of features. To better identify the general private image, diversity in feature selection is highly preferred.
- (**Design of survey**) Survey is a very useful data collection technique that has been widely used in many fields such as HCI, education, healthcare, marketing, and etc. On one hand, the total number of questions should be restricted in a certain range. On the other hand, the principle for the design of the survey is to be thorough and comprehensive to cover all aspects.
- (**Advertisement recommendation**) Many recommendation systems are designed to help users to find interesting and relevant items from a large information space. From a user perspective, users will feel frustrated when they are facing a monotonous recommendation list. On one hand, the total number of recommended advertisements is given. On the other hand, users expect the recommended advertisements from diverse categories.

To emphasize the diversity in feature selection, a popular approach is to use $L_{1,2}$ norm regularization [48, 9, 19, 17] such as exclusive LASSO. It is a soft diversity regularization and the selection is seriously affected by the magnitiute of weights of true features. Thus, it does not fit well for applications with strict requirement on diversity such as design of survey and advertisement recommendation. More importantly, so far there is no any theoretical result showing the benefit of emphasizing the selection diversity.

This paper proposes a novel general bilevel exclusive sparsity formulation to pursue a diverse feature selection, which directly restricts the numbers of the total selected features and the features selected from each group:

$$\min_{\mathbf{w}\in\mathbb{R}^p} \quad f(\mathbf{w}) \tag{1a}$$

$$\text{s.t.} \quad \|\mathbf{w}\|_0 \leq s \tag{1b}$$

$$\|\mathbf{w}_g\|_0 \leq \mathbf{t}_g \quad g \in \mathcal{G} \tag{1c}$$

$\mathbf{w}$ is the model we are pursuing. Its nonzero elements indicate the selected features (or items). $f(\mathbf{w})$ is a convex smooth function characterizing the empirical loss on the training samples. For example, $f(\mathbf{w})$ can take the form of the least square loss $\frac{1}{2}\sum_{i=1}^{n}(X_i - \mathbf{y}_i)^2$ or logistic regression loss $\sum_{i=1}^{n}\log(1+\exp(-\mathbf{y}_i X_i^\top \mathbf{w}))$. The first constraint (1b) controls the overall sparsity, that is, the total number of selected features, where the L_0 norm $\|\mathbf{w}\|_0$ is defined as the number of nonzeros in $\mathbf{w}$. The second constraint (1c) maintains the selection diversity by restricting the maximal number of selected features from each group, where $\mathbf{w}_g$ is a sub-vector of $\mathbf{w}$ indexed by $g \subset \{1, 2, \cdots, p\}$ and $\mathcal{G}$ is a super set containing all g's. (Note that $\mathbf{t}_g$'s are positive integers and their sum is greater than s; otherwise the constraint (1b) can be removed.) The group information usually comes from the prior information which is usually decided by the specific applications and domain knowledge. For example, in many computer vision applications, the group is automatically decided by the kind of features such as SIFT features, HOG features, GIST features, and etc. When group information is unavailable, a typical way is to cluster the features using algorithms like K-means [19].

In this paper, we derive an efficient computational algorithm to solve (1) which is NP-hard in general, and further provide theoretical results that show convergence of the algorithm and consistency of the learning formulation. More importantly, our analysis shows the clear advantages of using bilevel exclusive sparsity than using the single overall sparsity in two senses: 1) for the noiseless case, our approach needs fewer samples to find the true solution than the approach only using a single level of sparsity; 2) our approach improves the estimation error bound from $O(n^{-1}s\log p)$ to $O(n^{-1}s\log(p/|\mathcal{G}|))$ under certain assumptions where n is the number of training samples, p is the total number of features, and $|\mathcal{G}|$ is number of groups. To the best of our knowledge, this is the first work to show the benefits of using the diversity sparsity structure, which provides fundamental understanding on diversity preference. Extensive numerical studies have been conducted on both synthetic data and real-world data which show that the proposed method is superior over existing methods such as LASSO [46], L_0 norm approach [43], and exclusive LASSO [19].

Notation and definition

- $\Omega(s, \mathbf{t})$ is used to denote the set $\{\mathbf{w} \in \mathbb{R}^p \mid \|\mathbf{w}\|_0 \leq s,\ \|\mathbf{w}_g\|_0 \leq \mathbf{t}_g\ \forall g \in \mathcal{G}\}$;
- $\mathbf{t} \in \mathbb{R}^{|\mathcal{G}|}$ a vector consisting of all $\mathbf{t}_g$ in $\mathcal{G}$ in a certain order;
- $\bar{\mathbf{w}} \in \Omega(s, \mathbf{t})$ denotes the target solution which takes an arbitrary point in this region $\Omega(s, \mathbf{t})$;

- $\mathrm{P}_{\Omega(s,\mathbf{t})}(\mathbf{u}) = \arg\min_{\mathbf{w}\in\Omega(s,\mathbf{t})} \|\mathbf{w}-\mathbf{u}\|^2$. $\mathrm{P}_{\Omega(\infty,\mathbf{t})}(\mathbf{w})$ removes the overall sparsity restriction and can be simply calculated by keeping the top $\mathbf{t}_g$ largest elements in group g. Similarly, $\mathrm{P}_{\Omega(s,\infty)}(\mathbf{w})$ removes the sparsity restriction for all groups and can be calculated by keeping the top s largest elements from $\mathbf{w}$;
- $\mathrm{supp}(\mathbf{w})$ denotes the support index set, that is, the index set of nonzero elements in $\mathbf{w}$;
- $[\mathbf{w}]_S$ denotes the subvector indexed by the set S;
- $|S|$ returns the cardinality of S if S is a set or the absolute value of S if it is a real number.

2. Related Work

Extensive research efforts have been conducted on the sparse learning and sparsity regularizer. In this section, we provide a brief overview of the methods that are mostly related to our work, which include algorithms and models restricting the overall sparsity and the structured sparsity.

Overall sparsity. The most popular approaches to enforce the overall sparsity is to use the L_1 norm regularization [31, 7], which leads to a convex problem. L_0 norm approaches directly restrict overall sparsity, but it turns out to be NP hard problem in general. Iteratively hard threshold (IHT) [14, 5, 43] is a commonly used approach to solve it, which is nothing but the projected gradient descent. The model L_0 norm with least square loss is considered in [14, 5], while [43] considers the general convex smooth loss and provides theoretical error bounds and convergence rate. Our work is also based on IHT, but we consider bilevel exclusive sparsity structure simultaneously. In [26], IHT is categorized as greedy pursuits algorithm, together with Compressive Sampling Matching Pursuit (CoSaMP) [24], Orthogonal Matching Pursuit (OMP) [33], Regularized OMP (ROMP) [25], and forward-backward greedy algorithms [22]. All of these approaches essentially admit the same theoretical error bound. One interesting work [18] shows that by considering projection with a relaxed L_0 constraint, some assumptions which are needed for most methods can be released. Although some method (e.g. CoSaMP) can achieve better accuracy by designing more crafted procedures, there is no much theoretical improvements, and it usually brings considerable computational cost. The reason we choose IHT is that it is a straightforward approach to directly control the sparsity at all levels.

Structured sparsity. Besides of the overall sparsity, there is another type of structured sparsity. Group LASSO [4, 42] was proposed to extend LASSO for selecting groups of features. Fused LASSO [32] penalizes the sparsity of features and their difference. Exclusive LASSO [48, 9, 19, 17] also seeks the diversity in selection and uses the convex regularizer $L_{1,2}$ norm to characterize the exclusive sparsity structure. The key difference from this paper lies on two folds: 1) $L_{1,2}$ norm is a soft constraint for diversity preference, which is not suitable for applications with strict requirement on diversity and is not robust when weights of true features are in very different scales; and 2) They do not have any theoretical guarantee or error bound to show the benefit by using the exclusive sparsity. To our best knowledge, this paper is the first work to show the theoretical guarantee and benefits of using the diversity preference. Bach et.al. show the extension of the usage of L_1 norm to the modeling of a series of structured sparsity constraint composing with groups of features [2]. [41] presented a bilevel sparsity model. In addition, there are some work about pursuing sparsity of multiple levels [39, 40, 36]. The difference with our "bilevel" concept is their models restrict the overall sparsity and the group sparsity while our model restricts the overall sparsity and exclusive group sparsity. More importantly, previous works do not show any theoretical benefit by using the bilevel sparsity structure.

3. Algorithm

The problem in (1) is NP hard in general. In this section, we derive the IHT algorithm for solving (1). The IHT algorithm is an iterative algorithm which is able to directly control the sparsity of the solution unlike the convex relaxation based approaches. It consists of two steps per iteration: gradient descent and projection. The gradient descent step aims to decrease the objective and the projection uses to satisfy all constraints:

$$\mathbf{w}^{k+1} = \underbrace{\mathrm{P}_{\Omega(s,\mathbf{t})}(\underbrace{\mathbf{w}^k - \eta\nabla f(\mathbf{w}^k)}_{\text{gradient descent}})}_{\text{projection}} \tag{2}$$

where $\mathbf{w}^k$ are the weights in the k-th iteration. The combination of these two steps per iteration will monotonically decreases the objective function value as long as the step length η is small enough. We will show it converges in a linear rate to a ball of the target solution. It is also worth to mention that the IHT algorithm is essentially the projected gradient descent algorithm from the perspective of optimization. It is simple and efficient as long as the projection step is easy to calculate. In (2), the projection step is to project a point to the set $\Omega(s,\mathbf{t})$ which is the intersection between the sets $\Omega(s,\infty)$ and $\Omega(\infty,\mathbf{t})$. In general, the projection to the intersection between two sets would be difficult to calculate efficiently, but in our case due to the special struc-

ture, this projection can be decomposed into two simple projects.

Lemma 1. *The projection on $P_{\Omega(s,\mathbf{t})}$ defined in* (2) *can be calculated from two simple sequential projections as following*

$$P_{\Omega(s,\mathbf{t})}(\mathbf{w}) = P_{\Omega(s,\infty)}\left(P_{\Omega(\infty,\mathbf{t})}(\mathbf{w})\right).$$

The projection operation $\mathrm{P}_{\Omega(s,\infty)}$ sets all the elements to zero except the s elements which have the largest magnitude, and $\mathrm{P}_{\Omega(\infty,\mathbf{t})}$ makes the same operation to each group.

Note that the lemma above greatly simplifies the computational procedure of (1) by decomposing the procedure into two simpler operations $\mathrm{P}_{\Omega(\infty,\mathbf{t})}(\mathbf{w})$ and $\mathrm{P}_{\Omega(s,\infty)}(\mathbf{w})$. Specifically, $\mathrm{P}_{\Omega(\infty,\mathbf{t})}(\mathbf{w})$ basically removes the total sparsity restriction and it is simply to calculate by keeping the top $\mathbf{t}_g$ largest elements in group g. Similarly, $\mathrm{P}_{\Omega(s,\infty)}(\mathbf{w})$ removes the sparsity restriction for all groups which is also simply to calculate by keeping the top s largest elements of $\mathbf{w}$.

The η is the step length in (2). To guarantee the convergence, η can be dynamically decided by linear search [6]. The idea behinds linear search is to decrease the step length η if $\mathbf{w}^{k+1}$ does not reduce the objective function value from $\mathbf{w}^k$.

4. Main Results

This section discusses the convergence rate of (2) and the advantages of the proposed algorithm and the formation by using two levels of sparsity constraints. All proofs are provided in the supplemental materials.

The main theoretical results and significance can be summarized as follows:

- (Convergence) The general analysis [43] for the IHT update guarantees its convergence, but it is unclear whether it converges to the true solution. We show that the IHT update in (2) converges to a ball around the true solution with a linear rate and the radius of this ball converges to zero when the number of samples goes to infinity. This means that the solution provided by (2) will be driven to the true solution when more and more samples are received.
- (Benefits of using bilevel exclusive sparsity) It is a core problem in machine learning and sparse learning (or compressed sensing) to understand the dependence of the error bound and the sample complexity. Under some commonly used assumptions in sparse learning for the least squares objective function, we prove the error bound of the proposed algorithm is

$$\|\hat{\mathbf{w}} - \mathbf{w}^*\| \leq O\left(\sqrt{s\log(p/|\mathcal{G}|)/n}\right),$$

where $\hat{\mathbf{w}}$ is the output of running the IHT updates for a certain number of iterations. To the best of our knowledge, this work is the first to provide an error bound for the exclusive sparsity motivated approaches. More importantly, our analysis shows the advantages over the approaches (such as L_0 norm approaches and L_1 norm approaches) only using the overall sparsity with the well known error bound

$$O\left(\sqrt{s\log(p)/n}\right).$$

We believe that this analysis provides some fundamental understanding on the exclusive sparsity.

4.1. Convergence Rate

For the convergence property, we need to make some assumptions similar to most papers in compressed sensing and sparse learning. The following two assumptions essentially assume the lower bound $\rho_-(s,\mathbf{t})$ and the upper bound $\rho_+(s,\mathbf{t})$ for the curvature of $f(\cdot)$ in a low dimensional space $\Omega(s,\mathbf{t})$. However, the key difference from existing assumption lies on that our assumptions are less restrictive. While in many existing works such as [43, 22], where ρ_+ and ρ_- need to hold for any pairs $(\mathbf{w},\mathbf{u})$ satisfying $\mathbf{w}-\mathbf{u} \in \Omega(s,\mathbf{t})$. Our assumptions only needs to hold for a subset $(\mathbf{w},\bar{\mathbf{w}})$ with $\mathbf{w}-\bar{\mathbf{w}} \in \Omega(s,\mathbf{t})$, where $\bar{\mathbf{w}}$ is a fixed target model.

Assumption 1. (Target Restricted Strong Convexity.) *There exists $\rho_-(s,\mathbf{t}) > 0$ which satisfies:*

$$\langle \nabla f(\mathbf{w}) - \nabla f(\bar{\mathbf{w}}), \mathbf{w}-\bar{\mathbf{w}}\rangle \geq \rho_-(s,\mathbf{t})\|\mathbf{w}-\bar{\mathbf{w}}\|^2, \quad \forall\, \mathbf{w}-\bar{\mathbf{w}} \in \Omega(s,\mathbf{t})$$

Assumption 2. (Restricted Lipschitz Gradient.) *There exists $\rho_+(s,\mathbf{t}) < +\infty$ which satisfies:*

$$\langle \nabla f(\mathbf{w}) - \nabla f(\bar{\mathbf{w}}), \mathbf{w}-\bar{\mathbf{w}}\rangle \geq \rho_+(s,\mathbf{t})^{-1}\|[\nabla f(\mathbf{w}) - \nabla f(\bar{\mathbf{w}})]_S\|^2, \quad \forall \mathit{supp}(\mathbf{w}-\bar{\mathbf{w}}) \subseteq S, S \in \Omega(s,\mathbf{t}).$$

For readers who are familiar with RIP constant δ defined in Eq. (1.7) [8] for the least square loss function, $\rho_-(s,\mathbf{t})$ and $\rho_+(s,\mathbf{t})$ are actually nothing but $1-\delta$ and $1+\delta$ respectively.

Theorem 2. *Let $\bar{\mathbf{w}}$ be an arbitrary target model and*

$$\alpha = 2(1 - 2\eta\rho_-(3s,3\mathbf{t}) + \eta^2\rho_+(3s,3\mathbf{t})\rho_-(3s,3\mathbf{t}))^{1/2}. \tag{3}$$

If the step length η in (2) *can appropriately set to a value such that α is less than* 1 *and $\mathbf{w}^0$ is initialized as a feasible point in $\Omega(s,\mathbf{t})$, we have the following results*

- *The k-th iteration satisfies*

$$\|\mathbf{w}^k - \bar{\mathbf{w}}\| \leq \alpha^k \|\mathbf{w}^0 - \bar{\mathbf{w}}\| + \frac{2}{(1-\alpha)\rho_+(3s, 3\mathbf{t})}\Delta;$$

- *If* $k \geq \left\lceil \log \frac{2\Delta}{(1-\alpha)\rho_+(3s,3t)\|\mathbf{w}^0 - \bar{\mathbf{w}}\|} / \log \alpha \right\rceil$, *then the target features in* $\bar{\mathbf{w}}$ *can be identified by* $\mathbf{w}^k$ *is at least*

$$|supp(\mathbf{w}^k) \cap supp(\bar{\mathbf{w}})| \geq \left| \left\{ j \mid |\bar{\mathbf{w}}_j| > \frac{4\Delta}{(1-\alpha)\rho_+(3s,3t)} \right\} \right|$$

where $\Delta = \left\| P_{\Omega(2s,2\mathbf{t})}(\nabla f(\bar{\mathbf{w}})) \right\|$.

This theorem suggests a few important observations: 1) $\mathbf{w}^k$ in (2) converges to the ball $\mathcal{B}_{\bar{\mathbf{w}}}\left(\frac{2}{(1-\alpha)\rho_+(3s,3\mathbf{t})}\Delta\right)$ in a linear rate; 2) The number of correctly selected features by $\mathbf{w}^k$ (i.e., $|\text{supp}(\mathbf{w}^k) \cap \text{supp}(\bar{\mathbf{w}})|$), should be more than $|\{j \mid |\bar{\mathbf{w}}_j| > \frac{4\Delta}{(1-\alpha)\rho_+(3s,3t)}\}|$; 3) If all channels (that is, all nonzeros) of the target solution $\bar{\mathbf{w}}$ are strong enough, all true features can be identified correctly after a few iterations.

Acute readers may notice that the implicit assumption in our theory is $\alpha < 1$, which essentially requires the ratio $\rho_-/\rho_+ > 3/4$ by choosing the optimal steplength. This assumption is consistent with former studies [24, 8]. There is a recent analysis for $L0$ minimization [18] which can relax this requirement ($\rho_-/\rho_+ > 0$) by enlarging the scope (i.e. larger Ω) of the Assumptions 1 and 2. By applying their strategy, our result can be potentially relaxed.

4.2. Theoretical Benefits of Using Bilevel Exclusive Sparsity

We explain the benefits of using the bilevel sparsity structures in this subsection. In particular, we show the advantage over the model only using the overall sparsity such as Dantzig Selector [7], LASSO [46] and L_0 minimization [43].

To simplify the following analysis, we consider the least square loss $f(\mathbf{w}) = \frac{1}{2n}\|X\mathbf{w} - \mathbf{y}\|^2$. Assume that $\mathbf{w}^* \in \Omega(s, \mathbf{t}) \subset \mathbb{R}^p$ is a sparse vector, and y is generated from $y = X\mathbf{w}^* + \boldsymbol{\epsilon} \in \mathbb{R}^n$ where $\boldsymbol{\epsilon}_i$ are i.i.d Guassian noise $\mathcal{N}(0, \sigma^2)$[1]. The data matrix X has i.i.d samples (or rows) and each sample follows the sub-Guassian distribution.

We will consider two cases to show the advantage of the proposed model: noiseless case $\boldsymbol{\epsilon} = 0$ and noisy case $\boldsymbol{\epsilon} \neq 0$. First for the noiseless case, we have the following result:

[1]The Gaussian noise can be relaxed to sub-Gaussian noise; for example, any bounded random variable is sub-Gaussian

Theorem 3. (Noiseless linear regression) *For the least square loss without observation noise (that is,* $\boldsymbol{\epsilon} = 0$*), assume that matrix* X *is sub-Guassian and has independent rows or columns. If the number of samples* n *is more than*

$$O\left(\min\left\{ s\log p,\ \log(\max_{g\in\mathcal{G}}|g|)\sum_{g\in\mathcal{G}}\mathbf{t}_g \right\}\right) \quad (4)$$

then by appropriately choosing η *(for example,* $\eta = 1/\rho_+(3s, 3\mathbf{t})$*) such that* α *defined in* (3) *is less than* 1*, we have with high probability*[2] *that the sequence* $\{\mathbf{w}^k\}$ *generated from* (2) *converges to identify all features in the true solution* $\mathbf{w}^*$.

This theorem basically suggests that the true solution $\mathbf{w}^*$ can be exactly recovered when the number of samples is more than the quantity in (4). For the models which only consider the overall sparsity, for example, Dantzig selector [7], LASSO [46], and L_0 minimization [43], they have a similar exact recovery condition when the number of samples is more than $O(s \log p)$. Apparently, the required number of samples for our model is fewer in the order sense. Particularly, when $\sum_{g\in\mathcal{G}} \mathbf{t}_g = O(s)$ and $\log(\max_{g\in\mathcal{G}}|g|) \ll \log p$, the required number of samples in (4) is much less than $s \log p$ required by Dantzig selector, LASSO, and L_0 minimization. For example, taking $s = p/\log p$ and $|g| = \log p$, we have

$$s\log p = p \quad \text{and} \quad (4) = p\frac{\log\log p}{\log p}$$

which suggest that the proposed model constantly improved the sample complexity. In other words, it implies the quantitative benefits of using the additional exclusive sparsity. Next, we will observe similar benefits in the noisy case.

Theorem 4. (Noisy linear regression) *Under the same setting as in Theorem 3 except that the noise allows to be nonzero, we have with high probability*

- *There exists a number* k'*, such that*

$$\|\mathbf{w}^k - \mathbf{w}^*\| \leq O\left(\min\left\{\sqrt{\frac{s\log p}{n}}, \sqrt{\frac{\log(\max_{g\in\mathcal{G}}|g|)\sum_{g\in\mathcal{G}}\mathbf{t}_g}{n}}\right\}\right), \quad (5)$$

for all $k > k'$*, that is, all true features are identified after a certain number of iterations.*

[2]"With high probability" is a typical statement to simply the complicated probability definition and explanation. It basically says that the probability will converge to 1 when the problem dimension goes to infinity.

- *If* $|\mathbf{w}_j^*| > O\left(\min\left\{s\log p, \log(\max_{g\in\mathcal{G}}|g|)\sum_{g\in\mathcal{G}}\mathbf{t}_g\right\}\right)$, *there exists a number* k'', *such that supp*$(\mathbf{w}^k) =$ *supp*$(\mathbf{w}^*)$ *for all* $k > k''$, *that is, all true features are identified after a certain number of iterations.*

To see the benefits, recall the error bound for Dantzig selector [7], LASSO [46], and L_0 minimization [43] is $O(\sqrt{n^{-1}s\log p})$. Our result actually improves it to Eq. (5). To see a clear quantitative improvement, we can assume that all groups have comparable sizes and $\mathbf{t}_g$ is chosen appropriately ($\sum_{g\in\mathcal{G}}\mathbf{t}_g = O(s)$), then the error bound in (5) becomes $O(\sqrt{n^{-1}\log(p/|\mathcal{G}|)})$. Considering the same scenario as in the noiseless case, we can obtain the constant improvement as well.

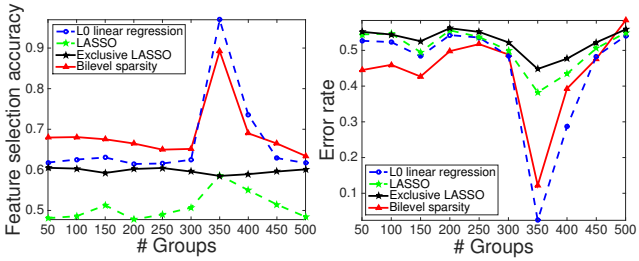
5. Experiments

Although this paper mainly focuses on the theoretical side, we provide an empirical study to validate the proposed model and algorithm with L_0 norm based approach, L_1 norm based approach, and $L_{1,2}$ norm based approach. Experiments are conducted on synthetic data, real data, and a real application problem.

5.1. Synthetic Data

In this section, we evaluate our method with synthetic data. Our sparsity method can be used in many algorithms which pursue a sparse solution. In the experiments, we conduct two classical models in machine learning and artificial intelligence, i.e., the least square and logistic regression models. These two algorithms are both linear models which allow us to use a $n \times p$ measurement matrix X to measure the data for n dimensional observations. For both algorithms, the measurement matrix X is generated with i.i.d. standard Gaussian distribution. For generating the true sparse vector $\mathbf{w}^*$, we first generate a dense p dimensional vector with i.i.d. standard Gaussian distribution, then we randomly set its elements to zeros until it satisfies our requirement (i.e., the bilevel sparse constraint). For the sparsity parameters correspond to s and $\mathbf{t}$, we use random integers from 1 to $\lfloor p/|\mathcal{G}|\rfloor$ to set elements of $\mathbf{t}$, where $\mathcal{G}$ is the set of all groups. Elements are uniformly assigned to $|\mathcal{G}|$ groups. The overall sparsity s is set as $0.75 \times \text{sum}(\mathbf{t})$, where the function sum$(\cdot)$ means the sum all of elements in the vector. The settings of group structure and sparsity s are the same for all the methods. All the synthetic experiments are repeated 30 times and we report the averaged performance.

Linear regression Firstly, we demonstrate the experiment using linear regression to recover the sparse vector $\mathbf{w}^*$. The dimension of the vector $\mathbf{w}^*$ is set as $p = 1000$ and the number of measurements is set as $n = 600$. After generating the true sparse vector $\mathbf{w}^*$, we can get the observations $\mathbf{y} \in \mathbb{R}^n$ by $\mathbf{y} = X\mathbf{w}^* + \epsilon$, where ϵ is a n dimensional noise vector generated from i.i.d. Gaussian distribution with mean 0 and variance 0.01. The cost function of least square is $f(\mathbf{w}) = \frac{1}{2n}\|X\mathbf{w} - y\|^2$. We compare our methods with three methods, i.e., L_0 norm minimization [43], L_1 norm regularization (LASSO [46]), and $L_{1,2}$ norm regularization (Exclusive LASSO [19]). While LASSO and L_0 minimization only considers the overall sparsity, exclusive LASSO only restrict the exclusive sparsity. For fair comparison, methods with regularization will be truncated and we only keep the largest s elements. Figure 2 shows the accuracy of feature selection (i.e. $|\text{supp}(\mathbf{w}^*) \cap \text{supp}(\mathbf{w})|/|\text{supp}(\mathbf{w}^*)|$) and the relative error rate which is $\|\mathbf{w} - \mathbf{w}^*\|/\|\mathbf{w}^*\|$. The step size η of projected gradient descent is chosen by line search method. For all methods, the relative error rate decreases just with the improvement of feature selection accuracy. Within these four methods, the proposed bilevel method is superior when compared to other methods.



(a) Feature selection accuracy. (b) Relative error rate.

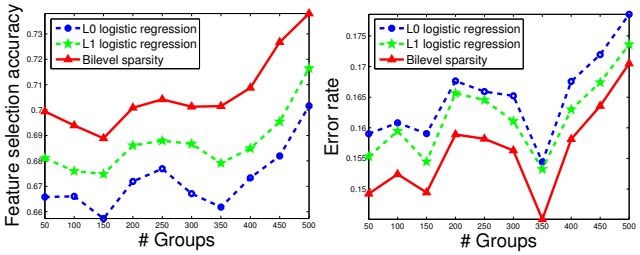
Figure 2. Feature selection accuracy and relative error rate for linear regression loss.

Logistic Regression Next we move to the logistic regression model which is a classical classification model. The model learns a linear classifier by maximizing the likelihood of sample-label pairs in the training set. The ith sample X_i is generated from Gaussian distribution and its label $\mathbf{y}_i$ is decided by sign$(X_i\mathbf{w}^*)$. We set the feature dimension p to 1000 and generate the true sparse model $\mathbf{w}^*$ in a similar fashion to the linear regression experiment. The number of training samples is $n = 2000$. Besides the comparison with true sparse vector $\mathbf{w}^*$, we need to estimate the classification error of our learned classifier. In addition, we generate a matrix $X^{test} \in \mathbb{R}^{n\times p}$ by using the mean of the generated matrix X, then its label vector is sign$(X^{test}\mathbf{w}^*)$. Figure 3 shows the feature selection accuracy and classification error rate of our method, and its counterparts, L_0 logistic regression and L_1 logistic regression. The result of L_1 norm regularization method only keep the s largest

Table 1. Testing accuracy comparison among L1 logistic regression, L0 logistic regression, and the proposed Bilevel exclusive sparsity model on real datasets: Computer (ISOLET [3]), Handwritten Digits (GISETTE and MNIST [21]), Cancer (LEU [15], ALLAML [13], Colorectal [1] and Prostate-GE [30]) and Social Media (TwitterHealthData [29]).

Data	#Samples	#Features	#Selected features	L_1 logistic regression	L_0 logistic regression	Exclusive LASSO	Bilevel sparsity(#Groups)
ISOLET	1560	617	18	74.74%±0.1127%	80.63%±0.1013%	77.96%±0.3559%	**81.14%±0.0729 %** (10)
MNIST	3119	784	8	96.40%±0.0670%	97.13%±0.1011%	95.03%±0.0957%	**97.29%±0.1138 %** (10)
ALLAML	72	7129	4	83.33%±0.0728%	85.28%±0.1785%	79.44%±0.1642%	**85.28%±0.0561 %** (4)
Colorectal	112	16331	6	92.14%±0.0275%	92.14%±0.2203%	80.71%±0.2168%	**93.75%±0.0060 %** (8)
GISETTE	6000	5000	20	90.82%±0.0109%	93.78%±0.1528%	91.92%±0.0743%	**94.17%±0.1360 %** (20)
Prostate-GE	102	5966	7	84.12%±0.0814%	88.24%±0.1720%	86.67%±0.0792%	**88.63%±0.0118 %** (8)
LEU	72	571	16	93.06%±0.0384%	93.61%±0.3812%	83.13%±0.0114%	**95.56%±0.0673 %** (8)
TwitterHealthData	6873	3846	36	81.96%±0.2071%	83.39%±0.0601%	83.33%±0.1957%	**83.50%±0.0241%** (20)

elements, and all of the other methods are conducted in a standard L_2 norm regularized logistic regression with the selected features. The L_1 norm and L_2 norm regularized logistic regression models are implemented with LIBLINEAR [12]. From Figure 3, we can see that our method consistently outperforms other methods.



(a) Feature selection accuracy. (b) Classification error rate.

Figure 3. Feature selection accuracy and classification error rate for logistic regression loss.

5.2. Real Data Sets

We compare the proposed method to exclusive LASSO, L_1 norm and L_0 norm logistic regression on real datasets. For each dataset, we randomly sampled 50% of the data for training and the remaining data for testing. We repeat all experiments 10 times and reported the averaged performance for three algorithms in Table 1. For our approach, s is set to $\text{sum}(\mathbf{t})$ roughly for all experiments. We observed that the proposed bilevel sparsity algorithm is the best performer compared to L_1 norm and L_0 norm in logistic regression.

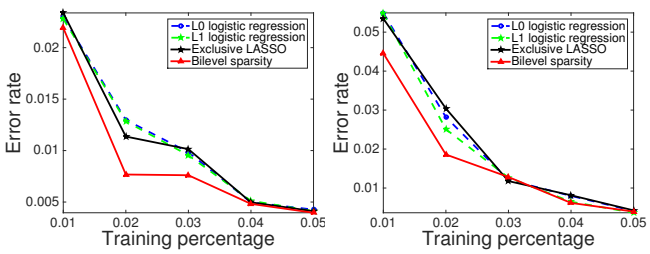
5.3. Application on Privacy Image Classification

Photo privacy is an important problem in the digital age. Many photos are shared on social media websites such as Facebook, Twitter, Flickr, and etc. Many time, the user uploaded photos without noticing privacy sensitive information is leaked in their photos. This can lead on identify theft, stolen private photos, or cause embarrassment. Solving the photo privacy problem helps to alleviate the aforementioned problems and alerts the user before they post their photos online.

Figure 4. Samples of private images data set.

This experiment use two datasets for photo privacy detection. The first data set is a public dataset from [44] which consists of with roughly 4K private photos and 33.5K public photos from Flickr. Due to the imbalance number of photos between the two classes, we randomly sampled 4K photos from the public photos set. For the second dataset, we collected photos [20] that are considered as private risk in general which come from 6 classes, i.e., nude photos, identification card, wedding photos, documents, people smoking, and family (group) photos. A sample of photos in our data set is shown in Figure 4. This data set consists of roughly 3.4K private photos. We randomly sampled 3.4K public photos from the public photos of dataset from [44] and used them as public photos.

We extracted 5 types of features: color histogram, histogram of oriented gradients (HOG) [11], linear binary pattern (LBP) [27], GIST [28], and bag of visual words (BoVW) with SIFT points [23, 10]. For each image, we generate 5 sub feature vectors which are 768-dimensional color histogram, 1984-dimensional HOG feature vector, 3712-dimensional LBP feature vector, 512-dimensional GIST feature vector, and 512-dimensional BoVW vector. These features capture the different type of image statistics such as colors, textures, and other patterns.



(a) Zerr et.al.'s [44] privacy image dataset. (b) Our privacy image dataset.

Figure 5. Classification error rate on privacy image datasets.

Different types (groups) of features depict different aspects of the visual information and emphasize different attributes of the images. As shown in Figure 4, the six private images classified into the "private" image due to different reasons look very different visually. For example, nude photos and documents photos have very different visual information. Therefore, we need to use different kinds of features to classify all of them into the general private image framework. Although traditional overall sparsity can avoid this, but the selected features may concentrate in just a few groups due to the biased dataset. For example, if the training set is dominanted by the number of nude images as private images, the selected features are also dominanted by nude images as well while important features from other types of private images could be absent in the selection procedure. This motivates us to emphasize the diversity in feature selection.

We demonstrate the performance of learning models with a relatively small training data (i.e., number of data is much smaller than the number of features). In our experiment, we randomly chose $1\% \sim 5\%$ samples from all the samples as the training set and leave the rest as the testing set. We construct the features into 5 groups based on the 5 different feature types mentioned above. By tuning the parameters on training set, we set the sparsity $\mathbf{t}_g$ as half of the number of total features in the group g and the overall sparsity $s = 0.75 \times \text{sum}(\mathbf{t})$. The classification error rate is shown in Figure 5. Exclusive Lasso is also compared here by setting the observations $\mathbf{y}$ as 1 and -1 for positive and negative samples respectively. We can see that our bilevel sparse model consistently outperforms the other sparse learning models. We also notice that all the error rates decrease significantly as the size of training set grows from 1% to 2%.

5.4. Application on a Complex Survey Dataset

We implemented our method on a survey dataset that was used for understanding which business practices drive firm's performance. This is a complex problem since many factors and practices could affect a firm performance. Thus, the original survey was quite comprehensive that consists of more than 20 categories (each category can be a group in the context of this paper) such as human resources, supply chain, information technology, to name a few. To measure each category, several related items (each item is a variable) were used to measure the same latent construct. The survey was assigned to 197 companies while these companies were classified as two classes, the high-performing company and low-performing company. The dataset comprised 1701 variables and 197 observations. The primary objective of the experiment is show how the proposed method can help to improve the design of complex survey. For our approach, s is set to $0.5 \times \text{sum}(\mathbf{t})$. Thus, as shown in Figure 6, the proposed method is more powerful to capture the inherent structure in the survey data. For a given number of feature that will be selected, the proposed method provides better approximation of the original dataset. Therefore, it implies that our method could help optimize the design of the survey, minimize the redundancy, and maximize the information collection power.

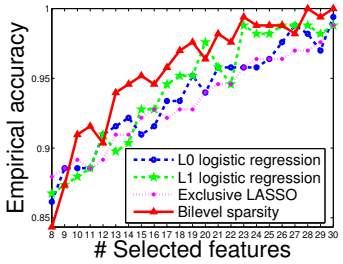


Figure 6. Empirical accuracy vs. # of selected features.

6. Conclusion

In this paper, we propose a novel bilevel exclusive sparsity formulation to emphasize the diversity in feature or item in general selection, motivated by diversity preference in many real problems such as private image classification, advertisement recommendation, and design of survey. The proposed formulation can be specified by many common tasks such as regression and classification. We propose to use IHT to solve this formulation which is NP-hard in general, and prove a linear convergence rate of IHT and the error bound between the output of IHT and the true solution. Our theoretical analysis shows the improvement of using diversity sparsity over the commonly used sparse learning approaches such as LASSO and L_0 norm approaches. To the best of our knowledge, this is the first work to show such error bound and theoretical benefit of using the diversity sparsity. Experiments on synthetic data and real-world data demonstrate the effectiveness of our method and validate and the correctness of our theory.

Acknowledgement

The authors from University of Rochester are supported by the NSF grant CNS-1548078 and the NEC fellowship. Shuai Huang is supported by NSF grant CMMI-1505260.

References

[1] T. Alexandrov, J. Decker, B. Mertens, A. M. Deelder, R. A. T. P. Maass, and H. Thiele. Biomarker discovery in maldi-tof serum protein profiles using discrete wavelet transformation. *Bioinformatics*, 25(5):643–649, 2009.

[2] F. Bach, R. Jenatton, J. Mairal, G. Obozinski, et al. Structured sparsity through convex optimization. *Statistical Science*, 27(4):450–468, 2012.

[3] K. Bache and M. Lichman. UCI machine learning repository. 2013.

[4] S. Bakin et al. Adaptive regression and model selection in data mining problems. 1999.

[5] T. Blumensath and M. E. Davies. Iterative thresholding for sparse approximations. *Journal of Fourier Analysis and Applications*, 14(5-6):629–654, 2008.

[6] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[7] E. Candes and T. Tao. The dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2313–2351, 2007.

[8] E. J. Candes and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

[9] X. Chen, X.-T. Yuan, Q. Chen, S. Yan, and T.-S. Chua. Multi-label visual classification with label exclusive context. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 834–841. IEEE, 2011.

[10] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *In Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

[11] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pages 886–893, 2005.

[12] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[13] S. P. Fodor. DNA SEQUENCING: Massively Parallel Genomics. *Science*, 277:393–395, 1997.

[14] S. Foucart. Hard thresholding pursuit: an algorithm for compressive sensing. *SIAM Journal on Numerical Analysis*, 49(6):2543–2563, 2011.

[15] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, 286(5439):531–537, 1999.

[16] D. Hsu, S. Kakade, and T. Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab*, 17(52), 2012.

[17] Y. Huang and J. Liu. Exclusive sparsity norm minimization with random groups via cone projection. *arXiv preprint arXiv:1510.07925*, 2015.

[18] P. Jain, A. Tewari, and P. Kar. On iterative hard thresholding methods for high-dimensional m-estimation. In *Advances in Neural Information Processing Systems*, pages 685–693, 2014.

[19] D. Kong, R. Fujimaki, J. Liu, F. Nie, and C. Ding. Exclusive feature learning on arbitrary structures via $\ell_{1,2}$-norm. In *NIPS*, pages 1655–1663, 2014.

[20] T. Lam, D. Kong, H. X. Jin, and J. Liu. Privacy-cnh: A framework to detect photo privacy with convolutional neural network using hierarchical features. *AAAI*, 2016.

[21] Y. LeCun and C. Cortes. The mnist database of handwritten digits. 1998.

[22] J. Liu, R. Fujimaki, and J. Ye. Forward-backward greedy algorithms for general convex smooth functions over a cardinality constraint. *ICML*, 2014.

[23] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, Nov. 2004.

[24] D. Needell and J. A. Tropp. Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, 2009.

[25] D. Needell and R. Vershynin. Signal recovery from incomplete and inaccurate measurements via regularized orthogonal matching pursuit. *IEEE Journal of Selected Topics in Signal Processing*, 4(2):310–316, 2010.

[26] N. Nguyen, D. Needell, and T. Woolf. Linear convergence of stochastic iterative greedy algorithms with sparse constraints. *arXiv preprint arXiv:1407.0088*, 2014.

[27] T. Ojala, M. Pietikinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51 – 59, 1996.

[28] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vision*, 42(3):145–175, May 2001.

[29] A. Sadilek, S. Brennan, H. Kautz, and V. Silenzio. nemesis: Which restaurants should you avoid today? *First AAAI Conference on Human Computation and Crowdsourcing*, 2013.

[30] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, E. S. Lander, M. Loda, P. W. Kantoff, T. R. Golub, and W. R. Sellers. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1(2):203 – 209, 2002.

[31] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[32] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

[33] J. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004.

[34] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms, 2008.

[35] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv:1011.3027*, 2010.

[36] J. Wang, M. Wang, P. Li, L. Liu, Z. Zhao, X. Hu, and X. Wu. Online feature selection with group structure analysis. *IEEE Transactions on Knowledge and Data Engineering*.

[37] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3360–3367. IEEE, 2010.

[38] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(2):210–227, 2009.

[39] S. Xiang, X. Tong, and J. Ye. Efficient sparse group feature selection via nonconvex optimization. In *ICML*, pages 284–292, 2013.

[40] S. Xiang, T. Yang, and J. Ye. Simultaneous feature and feature group selection through hard thresholding. In *SIGKDD*, pages 532–541. ACM, 2014.

[41] S. Xiang, L. Yuan, W. Fan, Y. Wang, P. M. Thompson, J. Ye, A. D. N. Initiative, et al. Bi-level multi-source learning for heterogeneous block-wise missing data. *NeuroImage*, 102:192–206, 2014.

[42] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[43] X.-T. Yuan, P. Li, and T. Zhang. Gradient hard thresholding pursuit for sparsity-constrained optimization. *ICML*, 2014.

[44] S. Zerr, S. Siersdorfer, J. Hare, and E. Demidova. Privacy-aware image classification and search. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 35–44, New York, NY, USA, 2012. ACM.

[45] H. Zhang, J. Yang, Y. Zhang, N. M. Nasrabadi, and T. S. Huang. Close the loop: Joint blind image restoration and recognition with sparse representation prior. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 770–777. IEEE, 2011.

[46] T. Zhang et al. Some sharp performance bounds for least squares regression with l1 regularization. *The Annals of Statistics*, 37(5A):2109–2144, 2009.

[47] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja. Robust visual tracking via multi-task sparse learning. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2042–2049. IEEE, 2012.

[48] Y. Zhou, R. Jin, and S. Hoi. Exclusive lasso for multi-task feature selection. In *ICAIS*, pages 988–995, 2010.

Supplemental Materials

Proof to Lemma 1

Proof. Denote $\text{supp}(\text{P}_{\Omega(\infty,\mathbf{t})}(\mathbf{w}))$ and $\text{supp}(\text{P}_{\Omega(s,\mathbf{t})}(\mathbf{w}))$ by A and B respectively for short. We first prove $B \subseteq A$.

Suppose $B \nsubseteq A$, then we can find an element $b \in B$ but $b \notin A$. Without the loss of generality, we assume that b is in a certain group g. Since $A \cap g$ contains the indices of the $\mathbf{t}_g$ largest (magnitude) elements of group g, there exists at least one element $a \in A \cap g$ and $a \notin B \cap g$ (otherwise $|B \cap g| \geq \mathbf{t}_g + 1$). Replacing b by a in B, the constraints are still satisfied, but we can get a better solution since $|\mathbf{w}_a| > |\mathbf{w}_b|$. This contradicts $B = \text{supp}(\text{P}_{\Omega(s,\mathbf{t})}(\mathbf{w}))$.

Because we already know $B \subseteq A$, we can construct B by selecting the A's elements corresponding to the largest s (magnitude) elements. Therefore, $\text{supp}(\text{P}_{\Omega(s,\mathbf{t})}(\mathbf{w})) = \text{supp}(\text{P}_{\Omega(s,\infty)}(\text{P}_{\Omega(\infty,\mathbf{t})}(\mathbf{w})))$, which proves Lemma 1. □

Lemma 5. *$\forall supp(\mathbf{w} - \bar{\mathbf{w}}) \subseteq S, S \in \Omega(s,\mathbf{t})$, if $2\eta - \eta^2\rho_+(s,\mathbf{t})) > 0$, then*

$$\|\mathbf{w} - \bar{\mathbf{w}} - \eta[\nabla f(\mathbf{w}) - \nabla f(\bar{\mathbf{w}})]_S\|^2 \leq (1 - 2\eta\rho_-(s,\mathbf{t}) + \eta^2\rho_-(s,\mathbf{t})\rho_+(s,\mathbf{t}))\|\mathbf{w} - \bar{\mathbf{w}}\|^2. \tag{6}$$

Proof.

$$\begin{aligned}
&\|\mathbf{w} - \bar{\mathbf{w}} - \eta[\nabla f(\mathbf{w}) - \nabla f(\bar{\mathbf{w}})]_S\|^2 \\
=&\|\mathbf{w} - \bar{\mathbf{w}}\|^2 + \eta^2\|[\nabla f(\mathbf{w}) - \nabla f(\bar{\mathbf{w}})]_S\|^2 - 2\eta\langle \mathbf{w} - \bar{\mathbf{w}}, [\nabla f(\mathbf{w}) - \nabla f(\bar{\mathbf{w}})]_S\rangle \\
\leq&\|\mathbf{w} - \bar{\mathbf{w}}\|^2 + (\eta^2\rho_+(s,\mathbf{t}) - 2\eta)\langle \mathbf{w} - \bar{\mathbf{w}}, [\nabla f(\mathbf{w}) - \nabla f(\bar{\mathbf{w}})]_S\rangle \\
\leq&\|\mathbf{w} - \bar{\mathbf{w}}\|^2 - (2\eta - \eta^2\rho_+(s,\mathbf{t}))\rho_-(s,\mathbf{t})\|\mathbf{w} - \bar{\mathbf{w}}\|^2 \\
=&(1 - 2\eta\rho_-(s,\mathbf{t}) + \eta^2\rho_+(s,\mathbf{t})\rho_-(s,\mathbf{t}))\|\mathbf{w} - \bar{\mathbf{w}}\|^2.
\end{aligned}$$

It completes the proof. □

Proof to Theorem 2

Proof. Let us prove the first claim.

$$\begin{aligned}
&\|\mathbf{w}^{k+1} - (\mathbf{w}^k - \eta\nabla f(\mathbf{w}^k))\|^2 \\
=&\|\mathbf{w}^{k+1} - \bar{\mathbf{w}}\|^2 + \|\bar{\mathbf{w}} - (\mathbf{w}^k - \eta\nabla f(\mathbf{w}^k))\|^2 + 2\langle \mathbf{w}^{k+1} - \bar{\mathbf{w}}, \bar{\mathbf{w}} - (\mathbf{w}^k - \eta\nabla f(\mathbf{w}^k))\rangle ?
\end{aligned}$$

Define $\bar{\Omega} = \text{supp}(\bar{\mathbf{w}}), \Omega_{k+1} = \text{supp}(\mathbf{w}^{k+1})$, and $\bar{\Omega}_{k+1} = \bar{\Omega} \cup \Omega_{k+1}$. From $\|\mathbf{w}^{k+1} - (\mathbf{w}^k - \eta\nabla f(\mathbf{w}^k))\|^2 \leq \|\bar{\mathbf{w}} - (\mathbf{w}^k - \eta\nabla f(\mathbf{w}^k))\|^2$, we have

$$\begin{aligned}
\|\mathbf{w}^{k+1} - \bar{\mathbf{w}}\|^2 &\leq 2\langle \mathbf{w}^{k+1} - \bar{\mathbf{w}}, \mathbf{w}^k - \eta\nabla f(\mathbf{w}^k) - \bar{\mathbf{w}}\rangle \\
&= 2\langle \mathbf{w}^{k+1} - \bar{\mathbf{w}}, [\mathbf{w}^k - \eta\nabla f(\mathbf{w}^k) - \bar{\mathbf{w}}]_{\bar{\Omega}_{k+1}}\rangle \\
&\leq 2\|\mathbf{w}^{k+1} - \bar{\mathbf{w}}\|\|[\mathbf{w}^k - \eta\nabla f(\mathbf{w}^k) - \bar{\mathbf{w}}]_{\bar{\Omega}_{k+1}}\|.
\end{aligned}$$

It follows

$$\begin{aligned}
\|\mathbf{w}^{k+1} - \bar{\mathbf{w}}\| &\leq 2\|[\mathbf{w}^k - \eta\nabla f(\mathbf{w}^k) - \bar{\mathbf{w}}]_{\bar{\Omega}_{k+1}}\| \\
&= 2\|[\mathbf{w}^k - \eta\nabla f(\mathbf{w}^k) - \bar{\mathbf{w}} + \eta\nabla f(\bar{\mathbf{w}}) - \eta\nabla f(\bar{\mathbf{w}})]_{\bar{\Omega}_{k+1}}\| \\
&\leq 2\|[\mathbf{w}^k - \eta\nabla f(\mathbf{w}^k) - \bar{\mathbf{w}} + \eta\nabla f(\bar{\mathbf{w}})]_{\bar{\Omega}_{k+1}}\| + 2\eta\|[\nabla f(\bar{\mathbf{w}})]_{\bar{\Omega}_{k+1}}\| \\
&\leq 2\|[\mathbf{w}^k - \eta\nabla f(\mathbf{w}^k) - \bar{\mathbf{w}} + \eta\nabla f(\bar{\mathbf{w}})]_{\bar{\Omega}_{k+1}\cup\Omega_k}\| + 2\eta\|[\nabla f(\bar{\mathbf{w}})]_{\bar{\Omega}_{k+1}}\| \\
&= 2\|\mathbf{w}^k - \bar{\mathbf{w}} - \eta[\nabla f(\mathbf{w}^k) - \nabla f(\bar{\mathbf{w}})]_{\bar{\Omega}_{k+1}\cup\Omega_k}\| + 2\eta\|[\nabla f(\bar{\mathbf{w}})]_{\bar{\Omega}_{k+1}}\|.
\end{aligned}$$

From the inequality of Lemma 5, we have

$$\begin{aligned}\|\mathbf{w}^{k+1}-\bar{\mathbf{w}}\| &\leq \alpha\|\mathbf{w}^{k}-\bar{\mathbf{w}}\|+2\eta\|[\nabla f(\bar{\mathbf{w}})]_{\bar{\Omega}_{k+1}}\| \\ &\leq \alpha\|\mathbf{w}^{k}-\bar{\mathbf{w}}\|+2\eta\max_{j}\|[\nabla f(\bar{\mathbf{w}})]_{\bar{\Omega}_{j+1}}\| \\ &\leq \alpha\|\mathbf{w}^{k}-\bar{\mathbf{w}}\|+2\eta\Delta. \end{aligned} \tag{7}$$

Since Δ is constant, using the recursive relation of (7), we have

$$\begin{aligned}\|\mathbf{w}^{k}-\bar{\mathbf{w}}\| &\leq \alpha^{k}\|\mathbf{w}^{0}-\bar{\mathbf{w}}\|+2\eta\Delta\sum_{i=0}^{k}\alpha^{i} \\ &= \alpha^{k}\|\mathbf{w}^{0}-\bar{\mathbf{w}}\|+2\eta\Delta\frac{1-\alpha^{k}}{1-\alpha} \\ &\leq \alpha^{k}\|\mathbf{w}^{0}-\bar{\mathbf{w}}\|+2\eta\Delta\frac{1}{1-\alpha}. \end{aligned} \tag{8}$$

Then we move to (2), when $k \geq \lceil \log \frac{2\Delta}{(1-\alpha)\rho_{+}(3s,3t)\|\mathbf{w}^{0}-\bar{\mathbf{w}}\|} / \log\alpha \rceil$, from the conclusion of (1), we have

$$\|\mathbf{w}^{k}-\bar{\mathbf{w}}\|_{\infty} \leq \|\mathbf{w}^{k}-\bar{\mathbf{w}}\| \leq \frac{4\Delta}{(1-\alpha)\rho_{+}(3s,3t)}. \tag{9}$$

For any $j \in \bar{\Omega}$,

$$\begin{aligned}\|\mathbf{w}^{k}-\bar{\mathbf{w}}\|_{\infty} &\geq |[\mathbf{w}^{k}-\bar{\mathbf{w}}]_{j}| \\ &\geq -|[\mathbf{w}^{k}]_{j}|+|[\bar{\mathbf{w}}]_{j}|. \end{aligned}$$

So

$$\begin{aligned}|[\mathbf{w}^{k}]_{j}| &\geq |[\bar{\mathbf{w}}]_{j}|-\|\mathbf{w}^{k}-\bar{\mathbf{w}}\|_{\infty} \\ &\geq |[\bar{\mathbf{w}}]_{j}|-\frac{4\Delta}{(1-\alpha)\rho_{+}(3s,3t)}. \end{aligned}$$

Therefore, $[\mathbf{w}^{k}]_{j}$ is non-zero if $|[\bar{\mathbf{w}}]_{j}| > \frac{4\Delta}{(1-\alpha)\rho_{+}(3s,3t)}$, and (2) is proved. □

Lemma 6. *The value of Δ is bounded by*

$$\Delta \leq \min\left(O\left(\sqrt{\frac{s\log p+\log 1/\eta'}{n}}\right),\ O\left(\sqrt{\frac{\max_{g\in\mathcal{G}}\log|g|\sum_{g\in\mathcal{G}}\mathbf{t}_{g}+\log 1/\eta'}{n}}\right)\right), \tag{10}$$

with high probability $1-\eta'$.

Proof. We introduce the following notation for matrix and it is different from the vector notation. For a matrix X in $\mathbb{R}^{n\times p}$, X_h will be a $\mathbb{R}^{n\times|h|}$ matrix that only keep the columns corresponding to the index set h. Here we restrict h by $\mathbf{w}_h \in \Omega(s,\mathbf{t})$ for any $\mathbf{w} \in \mathbb{R}^p$. We denote $\Sigma_h = X_h^{\top}X_g$, For the theorem, we can first show that $\|X_h^{\top}\epsilon\| \leqslant \sqrt{n}\left(\sqrt{|h|}+\sqrt{2\rho_{+}(2s,2\mathbf{t})\log(\frac{1}{\eta})}\right)$ with probability $1-\eta$. To this end, we have to point out that our columns of X are normalized to $\sqrt{n}$ and hence $X_h^{\top}\epsilon$ will be a $\frac{p}{m}$-variate Gaussian random variable with n on the diagonal of covariance matrix. We further use λ_i as the eigenvalues of Σ_h with decreasing order, i.e., λ_1 being the largest, or equivalently, $\lambda_1 = \|\Sigma_h\|_{spec}$.

Also, using the trick that $tr(\Sigma_h^2) = \lambda_1^2 + \lambda_2^2 + \cdots + \lambda_{|h|}^2$ and Proposition 1.1 from [16], we have

$$\begin{aligned} e^{-t} &\geqslant \mathbf{Pr}\left(||X_h^\top \epsilon||^2 > \sum_{i=1}^{|h|} \lambda_i + 2\sqrt{\sum_{i=1}^{|h|} \lambda_i^2 t} + 2\lambda_1 t \right) \\ &\geqslant \mathbf{Pr}\left(||X_h^\top \epsilon||^2 > \sum_{i=1}^{|h|} \lambda_i + 2\sqrt{2\sum_{i=1}^{|h|} \lambda_i \lambda_1 t} + 2\lambda_1 t \right) \\ &\geqslant \mathbf{Pr}\left(||X_h^\top \epsilon|| \geqslant \sqrt{\sum_{i=1}^{|h|} \lambda_i} + \sqrt{2\lambda_1 t} \right). \end{aligned}$$

Substitute t with $\log(\frac{1}{\eta})$ and the facts that $\sum_{i=1}^{|h|} \lambda_i = |h|n$ and $\lambda_1 = \|\Sigma\|_{spec} \leqslant n\rho_+(2s, 2\mathbf{t})$, we have

$$\|X_h^\top \epsilon\| \leqslant \sqrt{n}\left(\sqrt{|h|} + \sqrt{2\rho_+(2s, 2\mathbf{t})\log(1/\eta)}\right)$$

with probability $1 - \eta$.

For the least square loss, we have $\nabla f(\bar{\mathbf{w}}) = \frac{1}{n} X^\top (X\bar{\mathbf{w}} - y) = \frac{1}{n} X^\top \epsilon$. To estimate the upper bound of $\|\mathrm{P}_{\Omega(2s,2\mathbf{t})}(\nabla f(\mathbf{w}))\|$, we use the following fact

$$\|\mathrm{P}_{\Omega(2s,2\mathbf{t})}(\nabla f(\bar{\mathbf{w}}))\| = \|\mathrm{P}_{\Omega(2s,2\mathbf{t})}(X^\top \epsilon)\| \leq \min\left(\|\mathrm{P}_{\Omega(2s,\infty)}(X^\top \epsilon)\|,\ \|\mathrm{P}_{\Omega(\infty,2\mathbf{t})}(X^\top \epsilon)\|\right).$$

We consider the upper bounds of $\|\mathrm{P}_{\Omega(2s,\infty)}(X^\top \epsilon)\|$ and $\|\mathrm{P}_{\Omega(\infty,2\mathbf{t})}(X^\top \epsilon)\|$ respectively:

$$\begin{aligned} &\mathbf{Pr}\left(\|\mathrm{P}_{\Omega(2s,\infty)}(X^\top \epsilon)\| \geq n^{-1/2}\left(\sqrt{2s} + \sqrt{2\rho_+(2s, 2\mathbf{t})\log(1/\eta)}\right)\right) \\ =&\mathbf{Pr}\left(\max_{|h|=2s} \|X_h^\top \epsilon\| \geq n^{-1/2}\left(\sqrt{2s} + \sqrt{2\rho_+(2s, 2\mathbf{t})\log(1/\eta)}\right)\right) \\ \leq &\sum_{|h|=2s} \mathbf{Pr}\left(\|X_h^\top \epsilon\| \geq n^{-1/2}\left(\sqrt{2s} + \sqrt{2\rho_+(2s, 2\mathbf{t})\log(1/\eta)}\right)\right) \\ \leq &\binom{p}{2s}\eta. \end{aligned}$$

By taking $\eta' = \eta\binom{p}{2s}$, we obtain

$$\begin{aligned} \eta' \geq &\mathbf{Pr}\left(\|\mathrm{P}_{\Omega(2s,\infty)}(X^\top \epsilon)\| \geq n^{-1/2}\left(\sqrt{2s} + \sqrt{2\rho_+(2s, 2\mathbf{t})\log\left(\binom{p}{2s}/\eta'\right)}\right)\right) \\ \geq &\mathbf{Pr}\left(\|\mathrm{P}_{\Omega(2s,\infty)}(X^\top \epsilon)\| \geq O\left(\sqrt{\frac{s\log(p) + \log 1/\eta'}{n}}\right)\right), \end{aligned}$$

where the last inequality uses the fact that $\rho_+(2s, 2\mathbf{t})$ is bounded by a constant with high probability.

Next we consider the upper bound of $\|\mathrm{P}_{\Omega(\infty,2\mathbf{t})}(X^\top\epsilon)\|$. Similarly, we have

$$\begin{aligned}
&\mathbf{Pr}\left(\|\mathrm{P}_{\Omega(\infty,2\mathbf{t})}(X^\top\epsilon)\|\geq n^{-1/2}\left(\sqrt{2\sum_{g\in\mathcal{G}}g}+\sqrt{2\rho_+(2s,2\mathbf{t})\log(1/\eta)}\right)\right)\\
=&\mathbf{Pr}\left(\max_{|h\cap g|=2\mathbf{t}_g\ \forall g\in\mathcal{G}}\|X_h^\top\epsilon\|\geq n^{-1/2}\left(\sqrt{2\sum_{g\in\mathcal{G}}g}+\sqrt{2\rho_+(2s,2\mathbf{t})\log(1/\eta)}\right)\right)\\
\leq&\sum_{|h\cap g|\leq 2\mathbf{t}_g\ \forall g\in\mathcal{G}}\mathbf{Pr}\left(\|X_h^\top\epsilon\|\geq n^{-1/2}\left(\sqrt{2\sum_{g\in\mathcal{G}}g}+\sqrt{2\rho_+(2s,2\mathbf{t})\log(1/\eta)}\right)\right)\\
\leq&\eta\prod_{g\in\mathcal{G}}\binom{|g|}{2\mathbf{t}_g}.
\end{aligned}$$

Thus, by taking $\eta'=\eta\prod_{g\in\mathcal{G}}\binom{|g|}{2\mathbf{t}_g}$, we have

$$\begin{aligned}
\eta'\geq&\mathbf{Pr}\left(\|\mathrm{P}_{\Omega(\infty,2\mathbf{t})}(X^\top\epsilon)\|\geq n^{-1/2}\left(\sqrt{2\sum_{g\in\mathcal{G}}\mathbf{t}_g}+\sqrt{2\rho_+(2s,2\mathbf{t})\log\left(\prod_{g\in\mathcal{G}}\binom{|g|}{2\mathbf{t}_g}/\eta'\right)}\right)\right)\\
\geq&\mathbf{Pr}\left(\|\mathrm{P}_{\Omega(\infty,2\mathbf{t})}(X^\top\epsilon)\|\geq n^{-1/2}\left(\sqrt{2\sum_{g\in\mathcal{G}}\mathbf{t}_g}+\sqrt{4\rho_+(2s,2\mathbf{t})\sum_{g\in\mathcal{G}}\mathbf{t}_g\log|g|+2\varphi_+(1)\log 1/\eta'}\right)\right)\\
\geq&\mathbf{Pr}\left(\|\mathrm{P}_{\Omega(\infty,2\mathbf{t})}(X^\top\epsilon)\|\geq n^{-1/2}\left(\sqrt{2\sum_{g\in\mathcal{G}}\mathbf{t}_g}+\sqrt{4\rho_+(2s,2\mathbf{t})\max_{g\in\mathcal{G}}\log|g|\sum_{g\in\mathcal{G}}\mathbf{t}_g+2\varphi_+(1)\log 1/\eta'}\right)\right)\\
\geq&\mathbf{Pr}\left(\|\mathrm{P}_{\Omega(\infty,2\mathbf{t})}(X^\top\epsilon)\|\geq O\left(\sqrt{\frac{\max_{g\in\mathcal{G}}\log|g|\sum_{g\in\mathcal{G}}\mathbf{t}_g+\log 1/\eta'}{n}}\right)\right).
\end{aligned}$$

Summarizing two upper bounds, we have with high probability $(1-2\eta')$

$$\|\mathrm{P}_{\Omega(2s,2\mathbf{t})}(\nabla f(\bar{\mathbf{w}}))\|\leq\min\left(O\left(\sqrt{\frac{s\log p+\log 1/\eta'}{n}}\right),\ O\left(\sqrt{\frac{\max_{g\in\mathcal{G}}\log|g|\sum_{g\in\mathcal{G}}\mathbf{t}_g+\log 1/\eta'}{n}}\right)\right).$$

□

Lemma 7. *For the least square loss, assume that matrix X to be sub-Gaussian with zero mean and has independent rows or columns. If the number of samples n is more than*

$$O\left(\min\left\{s\log p,\ \log(\max_{g\in\mathcal{G}}|g|)\sum_{g\in\mathcal{G}}\mathbf{t}_g\right\}\right),$$

then with high probability, we have with high probability

$$\rho_+(3s,3\mathbf{t})\leq\frac{3}{2} \tag{11}$$

$$\rho_-(3s,3\mathbf{t})\geq\frac{1}{2}. \tag{12}$$

Thus, α defined in (3) *is less than* 1 *by appropriately choosing η (for example, $\eta=1/\rho_+(3s,3\mathbf{t})$).*

Proof. For the linear regression loss, we have

$$\rho_+^{1/2}(3s,3\mathbf{t}) \le \frac{1}{\sqrt{n}} \max_{\mathbf{w}\in\Omega(3s,3\mathbf{t})} \frac{\|X\mathbf{w}\|}{\|\mathbf{w}\|} = \max_{|h|\le 3s, |h\cap g|\le \mathbf{t}_g} \|X_h\|$$
$$\rho_-^{1/2}(3s,3\mathbf{t}) \ge \frac{1}{\sqrt{n}} \min_{\mathbf{w}\in\Omega(3s,3\mathbf{t})} \frac{\|X\mathbf{w}\|}{\|\mathbf{w}\|} = \min_{1\le|h|\le 3s, |h\cap g|\le \mathbf{t}_g} \|X_h\|$$

From the random matrix theory [35, Theorem 5.39], we have

$$\mathbf{Pr}\left(\|X_h\| \ge \sqrt{n} + O(\sqrt{3s}) + O(\sqrt{\log\frac{1}{\eta}}\right) \le O(\eta)$$

Then we have

$$\begin{aligned}
&\mathbf{Pr}\left(\sqrt{n}\rho_+^{1/2}(3s,3\mathbf{t}) \ge \sqrt{n} + O(\sqrt{s}) + O(\sqrt{\log\frac{1}{\eta}})\right)\\
\le&\mathbf{Pr}\left(\max_{|h|\le 3s, |h\cap g|\le \mathbf{t}_g} \|X_h\| \ge \sqrt{n} + O(\sqrt{s}) + O(\sqrt{\log\frac{1}{\eta}})\right)\\
\le&|\{h \mid |h| = 3s\}|\mathbf{Pr}\left(\|X_h\| \ge \sqrt{n} + O(\sqrt{s}) + O(\sqrt{\log\frac{1}{\eta}})\right)\\
=&\binom{p}{3s}\mathbf{Pr}\left(\|X_h\| \ge \sqrt{n} + O(\sqrt{s}) + O(\sqrt{\log\frac{1}{\eta}})\right) \le O\left(\binom{p}{3s}\eta\right)
\end{aligned}$$

which implies (by taking $\eta' = \binom{p}{3s}\eta$):

$$\mathbf{Pr}\left(\sqrt{n}\rho_+^{1/2}(3s,3\mathbf{t}) \ge \sqrt{n} + O\left(\sqrt{s\log p}\right) + O\left(\sqrt{\log\frac{1}{\eta'}}\right)\right) \le \eta'$$

Taking $n = O(s\log p)$, we have $\rho_+^{1/2}(3s,3\mathbf{t}) \le \sqrt{\frac{3}{2}}$ with high probability. Next, we consider it from a different perspective.

$$\begin{aligned}
&\mathbf{Pr}\left(\sqrt{n}\rho_+^{1/2}(3s,3\mathbf{t}) \ge \sqrt{n} + O\left(\sqrt{\sum_{g\in\mathcal{G}}\mathbf{t}_g}\right) + O\left(\sqrt{\log\frac{1}{\eta}}\right)\right)\\
\le&\mathbf{Pr}\left(\sqrt{n}\rho_+^{1/2}(+\infty,3\mathbf{t}) \ge \sqrt{n} + O\left(\sqrt{\sum_{g\in\mathcal{G}}\mathbf{t}_g}\right) + O\left(\sqrt{\log\frac{1}{\eta}}\right)\right)\\
=&\mathbf{Pr}\left(\max_{|h\cap g|\le\mathbf{t}_g, g\in\mathcal{G}} \|X_h\| \ge \sqrt{n} + O\left(\sqrt{\sum_{g\in\mathcal{G}}\mathbf{t}_g}\right) + O\left(\sqrt{\log\frac{1}{\eta}}\right)\right)\\
\le&\prod_{g\in\mathcal{G}}\binom{|g|}{\mathbf{t}_g}\mathbf{Pr}\left(\|X_h\| \ge \sqrt{n} + O\left(\sqrt{\sum_{g\in\mathcal{G}}\mathbf{t}_g}\right) + O\left(\sqrt{\log\frac{1}{\eta}}\right)\right)\\
\le&\eta\prod_{g\in\mathcal{G}}\binom{|g|}{\mathbf{t}_g} \le \eta\log\max_{g\in\mathcal{G}}|g|\sum_{g\in\mathcal{G}}\mathbf{t}_g\\
\Rightarrow&\\
&\mathbf{Pr}\left(\sqrt{n}\rho_+^{1/2}(3s,3\mathbf{t}) \ge \sqrt{n} + O\left(\sqrt{\sum_{g\in\mathcal{G}}\mathbf{t}_g}\log\max_{g\in\mathcal{G}}|g|\right) + O\left(\log\frac{1}{\eta'}\right)\right) \le \eta'
\end{aligned}$$

It indicates that if $n \geq O(\sum_{g\in\mathcal{G}} \mathbf{t}_g \max_{g\in\mathcal{G}}|g|)$, then we have $\rho_+^{1/2}(3s, 3\mathbf{t}) \leq \sqrt{\frac{3}{2}}$ with high probability as well. Similarly, we can prove $\rho_-^{1/2}(3s, 3\mathbf{t}) \leq \sqrt{\frac{1}{2}}$ with high probability. □

Proof to Theorem 3

Proof. Since n is large enough as shown in (4), from Lemma 7, we have $\alpha < 1$ and are allowed to apply Theorem 2. Since $\Delta = 0$ for the noiseless case, we prove the theorem by letting $\bar{\mathbf{w}}$ be $\mathbf{w}^*$. □

Proof to Theorem 4

Proof. Since n is large enough as shown in (4), from Lemma 7, we have $\alpha < 1$ and are allowed to apply Theorem 2. From Lemma 6, we obtain the upper bound for Δ. When the number of iterations k is large enough such that $\alpha^k\|\mathbf{w}^0 - \bar{\mathbf{w}}\|$ reduces the magnitude of Δ, we can easily prove the error bound of $\mathbf{w}^k$ letting $\bar{\mathbf{w}}$ be $\mathbf{w}^*$. The second claim can be similarly proven by applying the second claim in Theorem 2. □