Sparse Extreme Multi-label Learning with Oracle Property

Weiwei Liu[1] **Xiaobo Shen**[2]

Abstract

The pioneering work of sparse local embeddings for extreme classification (SLEEC) (Bhatia et al., 2015) has shown great promise in multi-label learning. Unfortunately, the statistical rate of convergence and oracle property of SLEEC are still not well understood. To fill this gap, we present a unified framework for SLEEC with nonconvex penalty. Theoretically, we rigorously prove that our proposed estimator enjoys oracle property (i.e., performs as well as if the underlying model were known beforehand), and obtains a desirable statistical convergence rate. Moreover, we show that under a mild condition on the magnitude of the entries in the underlying model, we are able to obtain an improved convergence rate. Extensive numerical experiments verify our theoretical findings and the superiority of our proposed estimator.

1. Introduction

Extreme multi-label learning (Yu et al., 2014; Prabhu & Varma, 2014; Bhatia et al., 2015; Yen et al., 2016; Liu & Tsang, 2017; Babbar & Schölkopf, 2017) refers to learn a classifier that is able to automatically annotate a data point with the most relevant subset of labels from an extremely large number of labels, which has opened up a new research frontier in data mining and machine learning. A wide range of challenging applications, such as product categorization for e-commerce (Shen et al., 2011) and document, video and image annotation, can benefit from being formulated as multi-label learning (Dembczynski et al., 2010; Tsoumakas et al., 2012; Liu & Tsang, 2015b; Gibaja & Ventura, 2015; Du et al., 2017; Liu et al., 2017; Shen et al., 2018a;b) tasks with hundreds of thousands or even millions of labels.

Due to the simplicity and ease of implementation, embedding approaches (Hsu et al., 2009; Yu et al., 2014; Prabhu & Varma, 2014; Liu & Tsang, 2015a; Bhatia et al., 2015; Babbar & Schölkopf, 2017; Liu et al., 2019) have been proved to be the most popular methods for addressing extreme multi-label learning tasks. Specifically, based on an assumption that the label matrix is low-rank, embedding approaches project label vectors into a lower dimensional compressed label space. A regression is then learned for each compressed label and a decompression matrix is used to lift the embedded label vectors back to the original label space. Because the low rank assumption is violated in most real world applications, leading embedding approaches can not obtain high prediction accuracies, and scale to large-scale data sets.

To break the low-rank assumption and boost classification accuracy, the pioneering work of SLEEC (Bhatia et al., 2015) is developed to learn a small ensemble of local distance preserving embeddings. Extensive empirical studies in (Bhatia et al., 2015) show that SLEEC significantly outperforms the state-of-the-art embedding and tree-based methods. Although SLEEC has achieved great success in extreme multi-label classification, the statistical rate of convergence and oracle property of SLEEC remain less explored.

To bridge this gap, we propose a unified framework for SLEEC with nonconvex penalty. Theoretically, we show that our proposed estimator enjoys oracle property, which performs as well as if the underlying model were known beforehand, as well as attains a desirable statistical convergence rate of $\mathcal{O}(\frac{\sigma\sqrt{\varpi}+\sqrt{s^*}}{\mu\sqrt{n}})$, where σ, ϖ, μ are positive constants, n is the sample size and s^* denotes the cardinality of the true support of underlying model. Considering the magnitude of the entries in the underlying model, we are able to achieve a refined convergence rate of $\mathcal{O}(\frac{\sqrt{s^*}}{\mu\sqrt{n}})$ under suitable conditions. Moreover, we adapt an accelerated proximal gradient method with soft-thresholding to solve the proposed estimator. Empirical results on various data sets validate our theoretical results and the superiority of our proposed estimator.

We organize this paper as follows. §2 presents some preliminaries of SLEEC. §3 introduces our proposed estimator. §4 analyzes the statistical properties of our proposed estimator. §5 presents an optimization algorithm and experimental results are presented in §6. The last section provides our

[1] School of Computer Science, Wuhan University, Wuhan, China [2] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China. Correspondence to: Weiwei Liu <liuweiwei863@gmail.com>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

conclusions.

2. Preliminaries

In this section, we briefly review some preliminaries of SLEEC. SLEEC learns low dimensional embeddings which non-linearly capture label correlations by preserving the pairwise distances between only the closest (rather than all) label vectors. Regressors are then trained in the embedding space. SLEEC uses a k-nearest neighbour (kNN) classifier in the embedding space for prediction.

Assume $x_i \in \mathbb{R}^{d\times 1}$ is a real vector representing an input or instance (feature), $y_i \in \{0,1\}^{L\times 1}$ is the corresponding output or label vector ($i \in \{1,\ldots,n\}$). n denotes the number of training data. The input matrix is $X = [x_1,\ldots,x_n] \in \mathbb{R}^{d\times n}$ and the output matrix is $Y = [y_1,\ldots,y_n] \in \{0,1\}^{L\times n}$. SLEEC maps the label vector y_i to ϖ-dimensional vector $z_i \in \mathbb{R}^{\varpi\times 1}$ ($\varpi < L$ is a small constant) and learns a set of regressors $V \in \mathbb{R}^{\varpi\times d}$ s.t. $z_i \approx Vx_i, \forall i \in \{1,\ldots,n\}$. During the prediction, for a testing instance x, SLEEC first computes its embedding Vx and then perform kNN over the set $[Vx_1,\ldots,Vx_n]$. We denote the transpose of the vector/matrix by the superscript ^{T} and the logarithms to base 2 by $\log$. Let $||\cdot||_F$ represent the Frobenius norm. Given a matrix A, $||A||_1$ denotes the sum of absolute elements of A.

SLEEC aims to learn a embedding matrix $Z = [z_1,\ldots,z_n] \in \mathbb{R}^{\varpi\times n}$ through the following formula:

$$\min_{Z\in\mathbb{R}^{\varpi\times n}} ||P_\Omega(Y^TY) - P_\Omega(Z^TZ)||_F^2 \tag{1}$$

where the index set Ω denotes the set of neighbors: $(i,j) \in \Omega$ iff $j \in \mathcal{N}_i$. $\mathcal{N}_i$ denotes a set of nearest neighbors of i. $P_\Omega(\cdot)$ is defined as:

$$\big(P_\Omega(Y^TY)\big)_{(i,j)} = \begin{cases} y_i^Ty_j, & \text{if } (i,j)\in\Omega \\ 0, & \text{otherwise.}\end{cases}$$

Based on embedding matrix Z, SLEEC minimizes the following objective with l_1 and l_2 regularization to find regressors V, which is able to reduce the prediction time and the model size, and avoid overfitting.

$$\min_{V\in\mathbb{R}^{\varpi\times d}} ||Z - VX||_F^2 + \mu||V||_F^2 + \lambda||VX||_1 \tag{2}$$

where $\mu > 0$ and $\lambda > 0$ are the regularization parameters.

However, the l_1 penalty used in Eq.(2) introduces a bias into the resulting estimator (Zou, 2006; Zhang & Huang, 2008; Zhang, 2010), which compromises the estimation accuracy. Moreover, Fan *et al.* (Fan & Li, 2001) has argued that the oracle property does not hold for l_1 penalty. The following section introduces a novel estimator to address the issues.

3. The Proposed Estimator

This section proposes a unified framework for SLEEC with nonconvex penalty. We follow SLEEC to learn Z and perform prediction. Given embedding matrix Z, we consider a multiple regression model as follows:

$$Z = V^*X + W \tag{3}$$

where $V^* \in \mathbb{R}^{\varpi\times d}$ represents the unknown sparse regression coefficient matrix and $W \in \mathbb{R}^{\varpi\times n}$ denotes a noise matrix with independent and identically distributed (i.i.d) zero mean Gaussian entries with variance σ^2 ($\sigma > 0$). We propose to estimate V^* by minimizing the following objective:

$$\hat{V} = \mathop{\arg\min}_{V\in\mathbb{R}^{\varpi\times d}} ||Z - VX||_F^2 + \mu/2||V||_F^2 + \mathscr{P}_\lambda(V) \tag{4}$$

where $\mathscr{P}_\lambda(V)$ is a decomposable nonconvex regularization: $\mathscr{P}_\lambda(V) = \sum_{(i,j)} p_\lambda(V_{(i,j)})$ and $p_\lambda(\cdot)$ is a univariate nonconvex function. Nonconvex penalty functions, such as smoothly clipped absolute deviation (SCAD) penalty (Fan & Li, 2001) and minimax concave penalty (MCP) (Zhang, 2010), have recently attracted much attention because they can eliminate the estimation bias and attain attractive statistical properties. This work takes SCAD and MCP penalties as the example. Let $\mathbb{I}(\cdot)$ be the indicator function. $p_\lambda(\cdot)$ in SCAD is defined as

$$p_\lambda(t) = \lambda\int_0^{|t|}(\mathbb{I}(a \le \lambda) + \frac{(b\lambda - a)_+}{(b-1)\lambda}\mathbb{I}(a > \lambda))da \tag{5}$$

where $(b\lambda - a)_+ = \max(0, b\lambda - a)$, $b > 2$ and $\lambda > 0$. For MCP, we have

$$p_\lambda(t) = \lambda\int_0^{|t|}(1 - \frac{a}{\lambda b})_+da \tag{6}$$

where $b > 0$ is a fix parameter. These nonconvex penalties can be further decomposed as an l_1 penalty plus a concave part: $p_\lambda(t) = \lambda|t| + q_\lambda(t)$. For SCAD, the concave component is

$$q_\lambda(t) = \frac{-(|t|+\lambda)^2}{2(b-1)}\mathbb{I}(\lambda < |t| \le b\lambda) + (\frac{(b+1)\lambda^2}{2} - \lambda|t|)\mathbb{I}(|t| > b\lambda)$$

Regarding MCP, we have

$$q_\lambda(t) = -\frac{t^2}{2b}\mathbb{I}(|t| \le b\lambda) + (\frac{b\lambda^2}{2} - \lambda|t|)\mathbb{I}(|t| > b\lambda)$$

The decomposability of $p_\lambda(t)$ is equivalent to the decomposability of $\mathscr{P}_\lambda(V)$ as: $\mathscr{P}_\lambda(V) = \lambda||V||_1 + \mathscr{Q}_\lambda(V)$, where $\mathscr{Q}_\lambda(V) = \sum_{(i,j)} q_\lambda(V_{(i,j)})$. This paper relies on the following regularity conditions on $p_\lambda(t)$ and $q_\lambda(t)$:

(i) Both function $q_\lambda(t)$ and its derivative $q'_\lambda(t)$ pass through the origin: $q_\lambda(0) = q'_\lambda(0) = 0$.

(ii) There exits a constant ν such that the derivative $p'_\lambda(t)$ satisfies $p'_\lambda(t) = 0$, for $|t| \ge \nu > 0$.

(iii) $q'_\lambda(t)$ is monotone and Lipschitz continuous: for $\ddot{t} \ge t$, there exists a constant $\zeta \ge 0$ such that $q'_\lambda(\ddot{t}) - q'_\lambda(t) \ge -\zeta(\ddot{t} - t)$.

(iv) $|q'_\lambda(t)|$ is upper bounded by λ: $|q'_\lambda(t)| \le \lambda$ for any t.

A variety of nonconvex penalty functions satisfy the above conditions. For example, SCAD penalty satisfies the conditions with $\nu = b\lambda$ and $\zeta = 1/(b-1)$. Regarding MCP, we have $\nu = b\lambda$ and $\zeta = 1/b$.

4. Main Theory

In this section, we show that the estimator in Eq.(4) enjoys the oracle properties, namely, our proposed estimator performs as well as if the underlying model were known beforehand. For matrices A and B with compatible dimension, $\langle A, B\rangle$ denotes the trace inner product on matrix space that $\langle A, B\rangle = trace(A^TB)$. Given a matrix $A \in \mathbb{R}^{n_1 \times n_2}$, we define $||A||_\infty = \max_{(i,j)\in S}\{|A_{(i,j)}|\}$, $||A||_0 = |supp(A)|$, where S denotes the set of index notation for matrix A, $supp(A)$ represents the support of A: $supp(A) = \{(i,j) : A_{(i,j)} \neq 0\}$, and $|supp(A)|$ is the cardinality of $supp(A)$. Before we present the following theorem, we introduce the definition of an oracle estimator, denoted by $\hat{V}_O$. Let $S^* = supp(V^*)$. Its complement and cardinality are denoted by $\bar{S}^*$ and $s^* = |S^*|$, respectively. The oracle estimator $\hat{V}_O$ is defined as

$$\hat{V}_O = \underset{supp(V)\subseteq S^*}{\arg\min}\ \mathcal{L}(V) \tag{7}$$

where $\mathcal{L}(V) = ||Z - VX||_F^2 + \mu/2||V||_F^2$. Because we do not know the true support of V^* in practice, the oracle estimator defined above is not a practical estimator. We define $\widetilde{\mathcal{L}}_\lambda(V) = \mathcal{L}(V) + \mathscr{Q}_\lambda(V)$. The following theorem shows that our proposed estimator enjoys oracle property.

Theorem 1. *Suppose the nonconvex penalty $\mathscr{P}_\lambda(V) = \sum_{(i,j)} p_\lambda(V_{(i,j)})$ satisfies regularity conditions (i), (ii), (iii). We assume the oracle estimator $\hat{V}_O$ defined in Eq.(7) satisfies $\min_{(i,j)\in S^*} |(\hat{V}_O)_{(i,j)}| \ge \nu$. If $\mu > \zeta$, $||X||_F \le 1/n$, and V^* satisfies $||V^*||_\infty \le 1/(\mu\sqrt{n})$, we have*

(i) $\hat{V} = \hat{V}_O$.

(ii) $||\hat{V} - V^*||_F \le \frac{4\sigma\sqrt{\varpi}+2\sqrt{s^*}}{\mu\sqrt{n}}$.

Remark. Theorem 1 shows that our proposed estimator in Eq.(4) is identical to the oracle estimator under suitable conditions. This is a very strong result because we do not even have any oracle knowledge on the true support. Moreover, our proposed estimator is able to achieve the desirable statistical convergence rate of $\mathcal{O}(\frac{\sigma\sqrt{\varpi}+\sqrt{s^*}}{\mu\sqrt{n}})$ for estimating V^*.

Before proving Theorem 1, we first present the following lemmas.

Lemma 1. *Under regularity conditions (iii), we have*

$$\widetilde{\mathcal{L}}_\lambda(\ddot{V}) \ge \widetilde{\mathcal{L}}_\lambda(V) + \langle \nabla\widetilde{\mathcal{L}}_\lambda(V), \ddot{V} - V\rangle + \frac{\mu-\zeta}{2}||\ddot{V} - V||_F^2$$

Proof. Recall that $\mathscr{Q}_\lambda(V)$ is the concave part of the nonconvex penalty $\mathscr{P}_\lambda(V)$, which implies $-\mathscr{Q}_\lambda(V)$ is convex. Because $\mathscr{Q}_\lambda(V) = \sum_{(i,j)} q_\lambda(V_{(i,j)})$, where $q_\lambda(V_{(i,j)})$ satisfies regularity condition (iii), we have

$$\begin{aligned}&\big(q'_\lambda(\ddot{V}_{(i,j)}) - q'_\lambda(V_{(i,j)})\big)(\ddot{V}_{(i,j)} - V_{(i,j)})\\ \ge& -\zeta(\ddot{V}_{(i,j)} - V_{(i,j)})^2\end{aligned}$$

This implies the convex function $-\mathscr{Q}_\lambda(V)$ satisfies

$$\langle \nabla(-\mathscr{Q}_\lambda(\ddot{V})) - \nabla(-\mathscr{Q}_\lambda(V)), \ddot{V} - V\rangle \le \zeta||\ddot{V} - V||_F^2 \tag{8}$$

Following (Nesterov, 2014), Eq.(8) is equivalent to the definition of strong convexity, and $-\mathscr{Q}_\lambda(V)$ satisfies

$$-\mathscr{Q}_\lambda(\ddot{V}) \le -\mathscr{Q}_\lambda(V) - \langle \nabla\mathscr{Q}_\lambda(V), \ddot{V} - V\rangle + \frac{\zeta}{2}||\ddot{V} - V||_F^2 \tag{9}$$

Because $\mathcal{L}(V) = ||Z - VX||_F^2 + \mu/2||V||_F^2$ is strongly convex with modulus μ, we have

$$\mathcal{L}(\ddot{V}) \ge \mathcal{L}(V) + \langle \nabla\mathcal{L}(V), \ddot{V} - V\rangle + \frac{\mu}{2}||\ddot{V} - V||_F^2 \tag{10}$$

By subtracting Eq.(9) from Eq.(10), we obtain the result. □

Lemma 2. *If $||X||_F \le 1/n$, and V^* satisfies $||V^*||_\infty \le 1/(\mu\sqrt{n})$, we have*

$$||\hat{V}_O - V^*||_F \le \frac{4\sigma\sqrt{\varpi} + 2\sqrt{s^*}}{\mu\sqrt{n}}$$

Proof. Let $\Psi = \hat{V}_O - V^*$. According to Eq.(3), we have

$$\begin{aligned}&\mathcal{L}(\hat{V}_O) - \mathcal{L}(V^*)\\ =&||Z - \hat{V}_OX||_F^2 + \mu/2||\hat{V}_O||_F^2 - ||Z - V^*X||_F^2 - \mu/2||V^*||_F^2\\ =&||Z - V^*X - \Psi X||_F^2 + \mu/2||\Psi + V^*||_F^2\\ &- ||Z - V^*X||_F^2 - \mu/2||V^*||_F^2\\ =&||\Psi X||_F^2 - 2\langle W, \Psi X\rangle + \mu/2||\Psi||_F^2 + \mu\langle\Psi, V^*\rangle\end{aligned}$$

(11)

Because the oracle estimator $\hat{V}_O$ minimizes the objective in Eq.(7), we have $\mathcal{L}(\hat{V}_O) \leq \mathcal{L}(V^*)$. Using Hölder's inequality and the assumption, we obtain

$$\begin{aligned} & ||\Psi||_F^2 \\ \leq & 2/\mu(2\langle W, \Psi X\rangle - \mu\langle \Psi, V^*\rangle) \\ \leq & 2/\mu(2||W||_F||\Psi||_F||X||_F + \mu||\Psi||_F||V^*||_F) \\ \leq & \frac{4\sigma\sqrt{\varpi} + 2\sqrt{s^*}}{\mu\sqrt{n}}||\Psi||_F \end{aligned} \tag{12}$$

Therefore, we derive the result. □

Proof. (of Theorem 1). Let $M \in \partial||\hat{V}||_1$. Since $\hat{V}$ satisfies the optimality condition, we have

$$\max_{\ddot{V} \in \mathbb{R}^{\varpi \times d}} \langle \nabla\widetilde{\mathcal{L}}_\lambda(\hat{V}) + \lambda M, \hat{V} - \ddot{V}\rangle \leq 0 \tag{13}$$

Next, we prove that there exist some $M_O \in \partial||\hat{V}_O||_1$ such that $\hat{V}_O$ satisfies the optimality condition

$$\max_{\ddot{V} \in \mathbb{R}^{\varpi \times d}} \langle \nabla\widetilde{\mathcal{L}}_\lambda(\hat{V}_O) + \lambda M_O, \hat{V}_O - \ddot{V}\rangle \leq 0 \tag{14}$$

By the definition of $\widetilde{\mathcal{L}}_\lambda(\cdot)$, we obtain

$$\begin{aligned} & \langle \nabla\widetilde{\mathcal{L}}_\lambda(\hat{V}_O) + \lambda M_O, \hat{V}_O - \ddot{V}\rangle \\ = & \underbrace{\langle \nabla\mathscr{Q}_\lambda(\hat{V}_O) + \lambda M_O, \hat{V}_O - \ddot{V}\rangle}_{(i)} + \underbrace{\langle \nabla\mathcal{L}(\hat{V}_O), \hat{V}_O - \ddot{V}\rangle}_{(ii)} \end{aligned} \tag{15}$$

Regarding term (i) in Eq.(15), we consider two cases: $(i,j) \in S^*$ and $(i,j) \in \bar{S^*}$.

For $(i,j) \in \bar{S^*}$, since $(\hat{V}_O)_{(i,j)} = 0$ and by regularity conditions (i), we obtain $\left(\nabla\mathscr{Q}_\lambda(\hat{V}_O)\right)_{(i,j)} = 0$. As $M_O \in \partial||\hat{V}_O||_1$, by setting $(M_O)_{(i,j)} = 0$ for $(i,j) \in \bar{S^*}$, then we have $(\nabla\mathscr{Q}_\lambda(\hat{V}_O) + \lambda M_O)_{(i,j)\in\bar{S^*}} = 0$.

For $(i,j) \in S^*$, using the assumption $|(\hat{V}_O)_{(i,j)\in S^*}| \geq \nu$, the definition of $\mathscr{P}_\lambda(V)$ and regularity conditions (ii), we obtain $(\nabla\mathscr{Q}_\lambda(\hat{V}_O) + \lambda M_O)_{(i,j)\in S^*} = (\nabla\mathscr{P}_\lambda(\hat{V}_O))_{(i,j)\in S^*} = p'_\lambda((\hat{V}_O)_{(i,j)\in S^*}) = 0$. Therefore, term (i) in Eq.(15) is always zero: $\langle \nabla\mathscr{Q}_\lambda(\hat{V}_O) + \lambda M_O, \hat{V}_O - \ddot{V}\rangle = 0$.

Because $\hat{V}_O$ is the global solution to the minimization problem in Eq.(7), $\hat{V}_O$ in term (ii) of Eq.(15) satisfies the optimality condition: $\max_{\ddot{V} \in \mathbb{R}^{\varpi \times d}} \langle \nabla\mathcal{L}(\hat{V}_O), \hat{V}_O - \ddot{V}\rangle \leq 0$. By taking the maximum over $\ddot{V} \in \mathbb{R}^{\varpi \times d}$ on both sides of Eq.(15), we obtain Eq.(14). Now, we are going to prove that $\hat{V} = \hat{V}_O$.

Applying Lemma 1, we have

$$\begin{aligned} & \widetilde{\mathcal{L}}_\lambda(\hat{V}) \\ \geq & \widetilde{\mathcal{L}}_\lambda(\hat{V}_O) + \langle \nabla\widetilde{\mathcal{L}}_\lambda(\hat{V}_O), \hat{V} - \hat{V}_O\rangle + \frac{\mu - \zeta}{2}||\hat{V} - \hat{V}_O||_F^2 \\ & \widetilde{\mathcal{L}}_\lambda(\hat{V}_O) \\ \geq & \widetilde{\mathcal{L}}_\lambda(\hat{V}) + \langle \nabla\widetilde{\mathcal{L}}_\lambda(\hat{V}), \hat{V}_O - \hat{V}\rangle + \frac{\mu - \zeta}{2}||\hat{V}_O - \hat{V}||_F^2 \end{aligned} \tag{16}$$

Using the convexity of l_1 norm, we obtain

$$\begin{aligned} \lambda||\hat{V}||_1 \geq \lambda||\hat{V}_O||_1 + \lambda\langle M_O, \hat{V} - \hat{V}_O\rangle \\ \lambda||\hat{V}_O||_1 \geq \lambda||\hat{V}||_1 + \lambda\langle M, \hat{V}_O - \hat{V}\rangle \end{aligned} \tag{17}$$

Adding Eq.(16) to Eq.(17), we have

$$\begin{aligned} 0 \geq & \langle \nabla\widetilde{\mathcal{L}}_\lambda(\hat{V}) + \lambda M, \hat{V}_O - \hat{V}\rangle \\ & + \langle \nabla\widetilde{\mathcal{L}}_\lambda(\hat{V}_O) + \lambda M_O, \hat{V} - \hat{V}_O\rangle + (\mu - \zeta)||\hat{V}_O - \hat{V}||_F^2 \end{aligned} \tag{18}$$

According to Eq.(13) to Eq.(14), we obtain

$$\begin{aligned} & \langle \nabla\widetilde{\mathcal{L}}_\lambda(\hat{V}) + \lambda M, \hat{V} - \hat{V}_O\rangle \\ \leq & \max_{\ddot{V} \in \mathbb{R}^{\varpi \times d}} \langle \nabla\widetilde{\mathcal{L}}_\lambda(\hat{V}) + \lambda M, \hat{V} - \ddot{V}\rangle \leq 0 \\ & \langle \nabla\widetilde{\mathcal{L}}_\lambda(\hat{V}_O) + \lambda M_O, \hat{V}_O - \hat{V}\rangle \\ \leq & \max_{\ddot{V} \in \mathbb{R}^{\varpi \times d}} \langle \nabla\widetilde{\mathcal{L}}_\lambda(\hat{V}_O) + \lambda M_O, \hat{V}_O - \ddot{V}\rangle \leq 0 \end{aligned} \tag{19}$$

Therefore $(\mu - \zeta)||\hat{V}_O - \hat{V}||_F^2 \leq 0$. As $\mu > \zeta$, we derive $\hat{V} = \hat{V}_O$.

By Lemma 2 and the first result, we derive the second result and complete the proof. □

Considering the magnitude of the entries in V^*, the following theorem provides a refined statistical rate of convergence. Let $S_1^* \cup S_2^* = S^* = supp(V^*)$, $s_1^* = |S_1^*|$, $s_2^* = |S_2^*|$ and $|S^*| = s^* = s_1^* + s_2^*$, $m_1 = \min\{\varpi, d\}$ and $m_2 = \max\{\varpi, d\}$.

Theorem 2. *We assume that $|V^*_{(i,j)\in S_1^*}| \geq \nu$, while $|V^*_{(i,j)\in S_2^*}| < \nu$. Suppose the nonconvex penalty $\mathscr{P}_\lambda(V) = \sum_{(i,j)} p_\lambda(V_{(i,j)})$ satisfies regularity conditions (i), (ii), (iii) and (iv). Given $\mu > \zeta$, for the estimator defined in Eq.(4) with regularization parameter $\lambda = C\sqrt{\log m_1/nm_2}$ ($C > 0$), and $\max_{(i,j)\in S^* \cup \bar{S^*}} |\nabla\mathcal{L}(V^*)_{(i,j)}| \leq \lambda$, we have*

$$||\hat{V} - V^*||_F \leq \underbrace{\frac{C\sqrt{s_1^* \log m_1}}{(\mu - \zeta)\sqrt{nm_2}}}_{\Xi_1 : |V^*_{(i,j)}| \geq \nu} + \underbrace{\frac{3C\sqrt{s_2^* \log m_1}}{(\mu - \zeta)\sqrt{nm_2}}}_{\Xi_2 : |V^*_{(i,j)}| < \nu}$$

Proof. Let $M \in \partial||\hat{V}||_1$ and $M^* \in \partial||V^*||_1$. Since $\hat{V}$ satisfies the optimality condition, we have

$$\max_{\ddot{V}\in\mathbb{R}^{\varpi\times d}} \langle \nabla\widetilde{\mathcal{L}}_\lambda(\hat{V}) + \lambda M, \hat{V} - \ddot{V}\rangle \leq 0 \tag{20}$$

Applying Lemma 1, we have

$$\begin{aligned}
&\widetilde{\mathcal{L}}_\lambda(\hat{V})\\
\geq&\widetilde{\mathcal{L}}_\lambda(V^*) + \langle \nabla\widetilde{\mathcal{L}}_\lambda(V^*), \hat{V} - V^*\rangle + \frac{\mu-\zeta}{2}||\hat{V} - V^*||_F^2\\
&\widetilde{\mathcal{L}}_\lambda(V^*)\\
\geq&\widetilde{\mathcal{L}}_\lambda(\hat{V}) + \langle \nabla\widetilde{\mathcal{L}}_\lambda(\hat{V}), V^* - \hat{V}\rangle + \frac{\mu-\zeta}{2}||V^* - \hat{V}||_F^2
\end{aligned} \tag{21}$$

Using the convexity of l_1 norm, we obtain

$$\begin{aligned}
\lambda||\hat{V}||_1 &\geq \lambda||V^*||_1 + \lambda\langle M^*, \hat{V} - V^*\rangle\\
\lambda||V^*||_1 &\geq \lambda||\hat{V}||_1 + \lambda\langle M, V^* - \hat{V}\rangle
\end{aligned} \tag{22}$$

Adding Eq.(21) to Eq.(22), we have

$$\begin{aligned}
&0 \geq \langle \nabla\widetilde{\mathcal{L}}_\lambda(\hat{V}) + \lambda M, V^* - \hat{V}\rangle\\
&+\langle \nabla\widetilde{\mathcal{L}}_\lambda(V^*) + \lambda M^*, \hat{V} - V^*\rangle + (\mu-\zeta)||V^* - \hat{V}||_F^2
\end{aligned} \tag{23}$$

Using Eq.(20), we obtain

$$\begin{aligned}
&\langle \nabla\widetilde{\mathcal{L}}_\lambda(\hat{V}) + \lambda M, \hat{V} - V^*\rangle\\
\leq&\max_{\ddot{V}\in\mathbb{R}^{\varpi\times d}} \langle \nabla\widetilde{\mathcal{L}}_\lambda(\hat{V}) + \lambda M, \hat{V} - \ddot{V}\rangle \leq 0
\end{aligned} \tag{24}$$

which implies

$$\begin{aligned}
&(\mu-\zeta)||V^* - \hat{V}||_F^2\\
\leq&\langle \nabla\widetilde{\mathcal{L}}_\lambda(V^*) + \lambda M^*, \hat{V} - V^*\rangle\\
\leq&\sum_{(i,j)\in S^*\cup\bar{S^*}} |(\nabla\mathcal{L}(V^*)+\nabla\mathscr{Q}_\lambda(V^*)+\lambda M^*)_{(i,j)}||(\hat{V}-V^*)_{(i,j)}|
\end{aligned} \tag{25}$$

We divide the summation in Eq.(25) into three parts: $(i,j) \in \bar{S^*}$, $(i,j) \in S_1^*$ and $(i,j) \in S_2^*$.

For $(i,j) \in \bar{S^*}$, since $(V^*)_{(i,j)} = 0$ and by regularity conditions (i), we obtain $\big(\nabla\mathscr{Q}_\lambda(V^*)\big)_{(i,j)} = q'_\lambda(0) = 0$. Using the assumption of Theorem 2, we have $\max_{(i,j)\in\bar{S^*}} |\nabla\mathcal{L}(V^*)_{(i,j)}| \leq \max_{(i,j)\in S^*\cup\bar{S^*}} |\nabla\mathcal{L}(V^*)_{(i,j)}| \leq \lambda$, thus we obtain $\max_{(i,j)\in\bar{S^*}} |(\nabla\mathcal{L}(V^*) + \nabla\mathscr{Q}_\lambda(V^*))_{(i,j)}| \leq \lambda$. Since $M^* \in \partial||V^*||_1$, we have $-\lambda \leq \lambda M^*_{(i,j)} \leq \lambda$. Thus, there always exist some $M^*_{(i,j)\in\bar{S^*}}$ such that $|(\nabla\mathcal{L}(V^*) + \nabla\mathscr{Q}_\lambda(V^*) + \lambda M^*)_{(i,j)\in\bar{S^*}}| = 0$, and we obtain

$$\begin{aligned}
&\sum_{(i,j)\in\bar{S^*}} |(\nabla\mathcal{L}(V^*)+\nabla\mathscr{Q}_\lambda(V^*)+\lambda M^*)_{(i,j)}|\times\\
&|(\hat{V}-V^*)_{(i,j)}| = 0
\end{aligned} \tag{26}$$

For $(i,j) \in S_1^*$, using the assumption $|(V^*)_{(i,j)\in S_1^*}| \geq \nu$, the definition of $\mathscr{P}_\lambda(V)$ and regularity conditions (ii), we obtain $(\nabla\mathscr{Q}_\lambda(V^*) + \lambda M^*)_{(i,j)\in S_1^*} = (\nabla\mathscr{P}_\lambda(V^*))_{(i,j)\in S_1^*} = p'_\lambda((V^*)_{(i,j)\in S_1^*}) = 0$. Using Hölder's inequality and the assumption, we obtain

$$\begin{aligned}
&\sum_{(i,j)\in S_1^*} |(\nabla\mathcal{L}(V^*)+\nabla\mathscr{Q}_\lambda(V^*)+\lambda M^*)_{(i,j)}||(\hat{V}-V^*)_{(i,j)}|\\
=&\sum_{(i,j)\in S_1^*} |(\nabla\mathcal{L}(V^*))_{(i,j)}||(\hat{V} - V^*)_{(i,j)}|\\
\leq&\lambda\sqrt{s_1^*}||V^* - \hat{V}||_F
\end{aligned} \tag{27}$$

For $(i,j) \in S_2^*$, we have the assumption $|(V^*)_{(i,j)\in S_2^*}| < \nu$. Using the regularity conditions (iv), we obtain $\max_{(i,j)\in S_2^*} |(\nabla\mathscr{Q}_\lambda(V^*))_{(i,j)}| = \max_{(i,j)\in S_2^*} |q'_\lambda((V^*)_{(i,j)})| \leq \lambda$. By the assumption of Theorem 2, we have $\max_{(i,j)\in S_2^*} |\nabla\mathcal{L}(V^*)_{(i,j)}| \leq \max_{(i,j)\in S^*\cup\bar{S^*}} |\nabla\mathcal{L}(V^*)_{(i,j)}| \leq \lambda$. Since $M^* \in \partial||V^*||_1$, we have $|M^*_{(i,j)}| \leq 1$. Thus, we obtain

$$\begin{aligned}
&|(\nabla\mathcal{L}(V^*) + \nabla\mathscr{Q}_\lambda(V^*) + \lambda M^*)_{(i,j)\in S_2^*}|\\
\leq&\max_{(i,j)\in S_2^*}|(\nabla\mathcal{L}(V^*)| + \max_{(i,j)\in S_2^*}|\nabla\mathscr{Q}_\lambda(V^*)| + \max_{(i,j)\in S_2^*}|\lambda M^*|\\
\leq&3\lambda
\end{aligned} \tag{28}$$

which implies

$$\begin{aligned}
&\sum_{(i,j)\in S_2^*} |(\nabla\mathcal{L}(V^*)+\nabla\mathscr{Q}_\lambda(V^*)+\lambda M^*)_{(i,j)}||(\hat{V}-V^*)_{(i,j)}|\\
\leq&3\lambda\sum_{(i,j)\in S_2^*} |(\hat{V} - V^*)_{(i,j)}|\\
\leq&3\lambda\sqrt{s_2^*}\sqrt{\sum_{(i,j)\in S_2^*} |(\hat{V} - V^*)_{(i,j)}|^2}\\
\leq&3\lambda\sqrt{s_2^*}||\hat{V} - V^*||_F
\end{aligned} \tag{29}$$

Combining Eq.(25), Eq.(26), Eq.(27) and Eq.(29), we complete the proof. □

Remark. The upper bound in Theorem 2 includes two parts corresponding to different magnitudes of the entries in V^*: (1) Ξ_1 corresponds to the set of entries with larger

Table 1. Statistics of six real-world data sets.

Datasets	#Training	#Testing	#Features	#Labels	#Card-Features	#Card-Labels
Bibtex	4,880	2,515	1,836	159	68.47	2.40
Delicious	12,920	3,185	500	983	18.17	19.03
Mediamill	30,993	12,914	120	101	120.00	4.38
Wiki10	14,146	6,616	101,938	30.938	673.45	18.64
Delicious-L	196,606	100,095	782,585	205,443	301.17	75.54
Amazon	490,449	153,025	135,909	670,091	75.68	5.45

Table 2. Precision@k (k=1,3,5) comparisons on three medium-sized data sets. The best results are in bold.

Datasets		CS	CPLST	ML-CSSP	1-vs-All	REML	FastXML	LEML	SLEEC	SML-SCAD	SML-MCP
Bibtex	P@1	58.87	62.38	44.98	62.62	65.13	63.42	62.54	65.08	66.43	**67.39**
	P@3	33.53	37.84	30.43	39.09	41.45	39.23	38.41	39.64	41.18	**42.86**
	P@5	23.72	27.62	23.53	28.79	30.12	28.86	28.21	28.87	30.25	**31.56**
Delicious	P@1	61.36	65.31	63.04	65.02	66.30	**69.61**	65.67	67.59	67.83	68.79
	P@3	56.46	59.95	56.26	58.88	61.73	64.12	60.55	61.36	63.45	**65.49**
	P@5	52.07	55.31	50.16	53.28	56.89	59.27	56.08	56.56	58.39	**60.56**
Mediamill	P@1	83.82	83.35	78.95	83.57	86.37	84.22	84.01	87.82	**89.56**	88.32
	P@3	67.32	66.18	60.93	65.50	73.97	67.33	67.20	73.45	**74.46**	73.89
	P@5	52.80	51.46	44.27	48.57	59.53	53.04	52.80	59.17	**60.53**	59.86

magnitudes; and (2) Ξ_2 corresponds to the set of entries with smaller magnitudes. By setting $\zeta = \mu/2$, when $s_1^* = s^*$ or $s_2^* = s^*$, we are able to achieve the convergence rate of $\mathcal{O}(\frac{\sqrt{s^*}}{\mu\sqrt{n}})$, which is sharper than the rate in Theorem 1.

5. Optimization Algorithm

In this section, we present an optimization algorithm to solve Eq.(4), which can be reformulated as

$$\min_{V\in\mathbb{R}^{\varpi\times d}} \mathcal{F}(V) := \widetilde{\mathcal{L}}_\lambda(V) + \mathcal{G}_\lambda(V) \tag{30}$$

where $\mathcal{G}_\lambda(V) = \lambda||V||_1$. Given $\mu > \zeta$, Lemma 1 shows that $\widetilde{\mathcal{L}}_\lambda(V)$ is strongly convex under regularity conditions (iii). $\mathcal{G}_\lambda(V)$ is a convex and non-smooth function. Thus, we adapt an accelerated proximal gradient (APG) method (Beck & Teboulle, 2009; chuan Toh & Yun, 2009) to iteratively minimize a quadratic approximation to $\mathcal{F}(V)$ at $\ddot{V} \in \mathbb{R}^{\varpi\times d}$ by

$$\begin{aligned}&\Phi_\tau(V,\ddot{V})\\ =&\widetilde{\mathcal{L}}_\lambda(\ddot{V}) + <\nabla\widetilde{\mathcal{L}}_\lambda(\ddot{V}), V-\ddot{V}> + \frac{\tau}{2}\|V-\ddot{V}\|_F^2 + \mathcal{G}_\lambda(V)\\ =&\frac{\tau}{2}\|V-B\|_F^2 + \mathcal{G}_\lambda(V) + \widetilde{\mathcal{L}}_\lambda(\ddot{V}) - \frac{1}{2\tau}\|\nabla\widetilde{\mathcal{L}}_\lambda(\ddot{V})\|_F^2\end{aligned} \tag{31}$$

where $\tau > 0$ is a constant and $B = \ddot{V} - \frac{1}{\tau}\nabla\widetilde{\mathcal{L}}_\lambda(\ddot{V})$. To minimize $\Phi_\tau(V,\ddot{V})$ w.r.t. V, it is reduced to solve the following Moreau projection problem (Wright et al., 2009):

$$\Upsilon_{\tau,\lambda}(B) = \arg\min_{V\in\mathbb{R}^{\varpi\times d}} \frac{\tau}{2}\|V-B\|_F^2 + \mathcal{G}_\lambda(V) \tag{32}$$

Note that the objective of problem (32) is separable w.r.t each entry in V. Wright *et al.* (Wright et al., 2009) have shown that the optimization problem (32) w.r.t each entry in V can be solved by the soft-thresholding operator:

$$\big(\Upsilon_{\tau,\lambda}(B)\big)_{(i,j)} = sgn(B_{(i,j)})\max\{0, |B_{(i,j)}| - \lambda/\tau\} \tag{33}$$

where $sgn(\cdot)$ denotes the sign function.

6. Experiment

In this section, we evaluate the performance of the proposed methods for extreme multi-label classification (MLC). All the computations are performed on a Red Hat Enterprise 64-Bit Linux workstation with 18-core Intel Xeon CPU E5-2680 2.80 GHz processor and 256 GB memory.

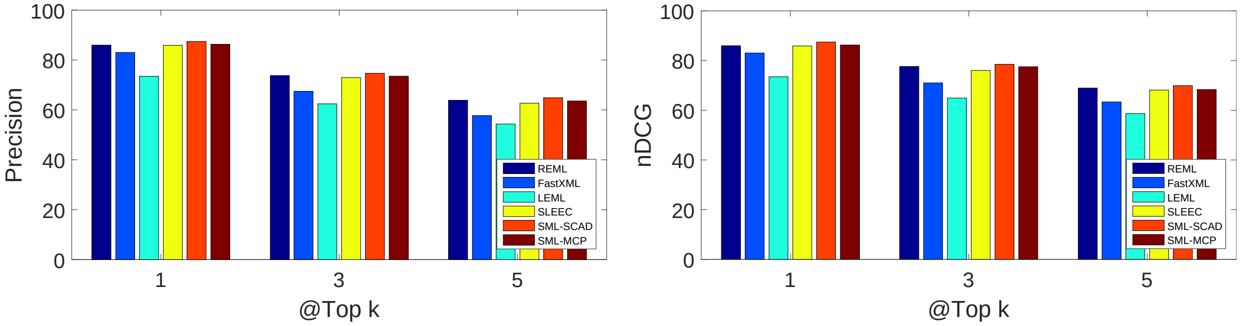
6.1. Experimental Setup

Datasets The experiments are conducted on a variety of real world multi-label data sets [1], which fall into two categories. The first category contains three medium-sized data sets, i.e., Bibtex, Delicious and Mediamill. The second category contains three large-scale data sets with more than

[1] http://manikvarma.org/downloads/XC/XMLRepository.html

Table 3. nDCG@k (k=1,3,5) comparisons on three medium-sized data sets. The best results are in bold.

Datasets		CS	CPLST	ML-CSSP	1-vs-All	REML	FastXML	LEML	SLEEC	SML-SCAD	SML-MCP
Bibtex	nDCG@1	58.87	62.38	44.98	62.62	65.13	63.42	62.54	65.08	66.43	**67.39**
	nDCG@3	52.19	57.63	44.67	59.13	60.01	59.51	58.22	60.47	61.02	**61.23**
	nDCG@5	53.25	59.71	47.97	61.58	62.46	61.70	60.53	62.64	62.89	**63.04**
Delicious	nDCG@1	61.36	65.31	63.04	65.02	66.30	**69.61**	65.67	67.59	67.83	68.79
	nDCG@3	57.66	61.16	57.91	60.43	62.65	65.47	61.77	62.87	63.95	**66.76**
	nDCG@5	54.44	57.80	53.36	56.28	59.10	61.90	58.47	59.28	60.12	**62.13**
Mediamill	nDCG@1	83.82	83.35	78.95	83.57	86.73	84.22	84.01	87.82	**89.56**	88.32
	nDCG@3	75.29	74.21	68.97	73.84	82.67	75.41	75.23	81.50	**83.84**	82.35
	nDCG@5	71.92	70.55	62.88	68.18	78.32	72.37	71.96	79.22	**81.32**	80.63



(a) Precision

(b) nDCG

Figure 1. (a) Top P@k and (b) nDCG@k of large-scale MLC on the Wiki10 data set.

hundreds of thousands of labels, i.e., Wiki10, Delicious-L and Amazon. The split of training and testing sets in the data sets is publicly available in (Bhatia et al., 2015). The statistics of the six real-world data sets are summarized in Table 1.

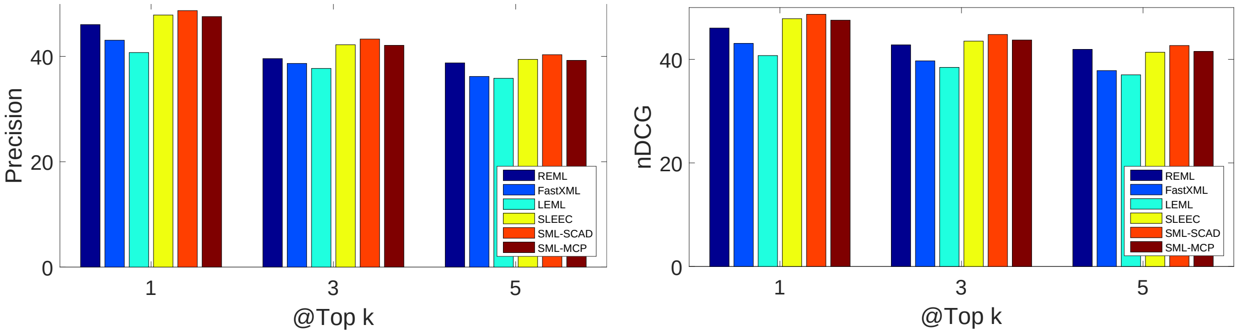
Baselines and Parameters We abbreviate our proposed methods with SCAD penalty and MCP penalty to SML-SCAD and SML-MCP, respectively. We compare the proposed methods with the state-of-the-art embedding methods including SLEEC (Bhatia et al., 2015), LEML (Yu et al., 2014) and tree-based FastXML (Prabhu & Varma, 2014), which can scale to the large-scale data sets. We further compare our methods with CS (Hsu et al., 2009), CPLST (Chen & Lin, 2012), ML-CSSP (Bi & Kwok, 2013), and 1-vs-All (Hariharan et al., 2012) on the medium-sized data sets. The codes of baseline methods are provided by the respective authors.

For the proposed methods, as $\mu > \zeta$ in the estimator, μ is set to $\mu = 2/(b-1)$ for SML-SCAD and $\mu = 2/b$ for SML-MCP, respectively. In addition, b is empirically set as 3 and 2 for SML-SCAD and SML-MCP, respectively. The parameter λ is tuned by validation on a small validation set. The dimensions of embedding are set as 100 and 50 for the medium-sized and large-scale data sets, respectively. Following (Bhatia et al., 2015), the number of nearest neighbours are selected via cross validation. The parameters for all the other baseline algorithms are set using fine grained validation on each data set so as to achieve the highest possible prediction accuracy for each method.

Evaluation Metrics We use two widely-used metrics to evaluate the multi-label classification performance. Precision at k (P@k) is the fraction of true positive predictions in the top k scoring labels. nDCG at k (nDCG@k) measures the usefulness or gain of a label based on its position in the predicted label list. The details of the metrics can be referred in (Prabhu & Varma, 2014; Bhatia et al., 2015; Zhang et al., 2015; Peng et al., 2018a;b; Zhou et al., 2018).

6.2. Results

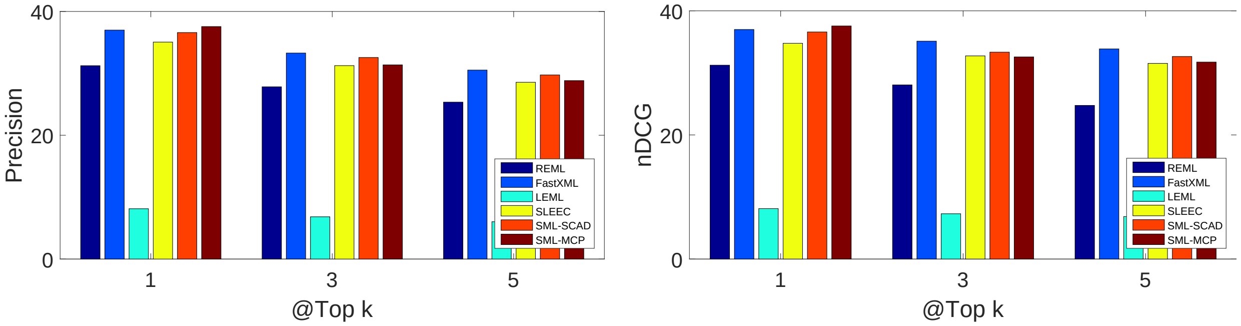
Results on medium-sized data sets We compare the classification performance on three medium-sized data sets, i.e., Bibtex, Delicious, Mediamill. The Precision@k and nDCG@k ($k = 1, 3, 5$) of all the methods are reported in Tables 2 and 3, respectively. From Tables 2 and 3, we can see that 1) SLEEC and its variant REML achieve better classification results than other baseline methods, which is consistent with the empirical results in (Bhatia et al., 2015). 2) Our proposed methods, SML-SCAD and SML-MCP, out-



(a) Precision

(b) nDCG

Figure 2. (a) Top P@k and (b) nDCG@k of large-scale MLC on the Delicious-L data set.



(a) Precision

(b) nDCG

Figure 3. (a) Top P@k and (b) nDCG@k of large-scale MLC on the Amazon data set.

perform SLEEC and REML, and are most successful on all data sets, which back up our theoretical analysis.

Results on large-scale data sets We compare the classification performance on three large-scale data sets. The classification performances on Wiki10, Delicious-L, Amazon are shown in Figures 1, 2, 3, respectively. From Figures 1, 2, 3, we can observe that the proposed method with nonconvex penalties, including SCAD and MCP penalty, achieve the best results.

7. Conclusion

SLEEC has been one of the most successful methods in extreme multi-label classification. However, the statistical rate of convergence and oracle property of SLEEC remain less explored. In this paper, we present a unified framework for SLEEC with nonconvex penalty. Our theoretical results show that our proposed estimator enjoys oracle property, and achieves an attractive statistical convergence rate. In addition, we can obtain a sharper convergence rate when a certain condition on the magnitude of the entries in the underlying model is imposed. Numerical experiments support our theoretical results and demonstrate the effectiveness of the proposed method.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grants 61822113 and 41801285, Fundamental Research Funds for the Central Universities under Grant 30919011229, Fundamental Research Founds for National University and China University of Geosciences (Wuhan) under Grant G1323541864.

References

Babbar, R. and Schölkopf, B. Dismec: Distributed sparse machines for extreme multi-label classification. In *WSDM*, pp. 721–729, 2017.

Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sciences*, 2(1):183–202, 2009.

Bhatia, K., Jain, H., Kar, P., Varma, M., and Jain, P. Sparse local embeddings for extreme multi-label classification. In *NIPS*, pp. 730–738, 2015.

Bi, W. and Kwok, J. T. Efficient multi-label classification with many labels. In *ICML*, pp. 405–413, 2013.

Chen, Y. and Lin, H. Feature-aware label space dimension reduction for multi-label classification. In *NIPS*, pp. 1538–1546, 2012.

chuan Toh, K. and Yun, S. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. Technical report, National University of Singapore, November 2009.

Dembczynski, K., Cheng, W., and Hüllermeier, E. Bayes optimal multilabel classification via probabilistic classifier chains. In *ICML*, pp. 279–286, 2010.

Du, B., Wang, Z., Zhang, L., Zhang, L., and Tao, D. Robust and discriminative labeling for multi-label active learning based on maximum correntropy criterion. *IEEE Trans. Image Processing*, 26(4):1694–1707, 2017.

Fan, J. and Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

Gibaja, E. and Ventura, S. A tutorial on multilabel learning. *ACM Computing Surveys*, 47(3):52:1–52:38, 2015.

Hariharan, B., Vishwanathan, S. V. N., and Varma, M. Efficient max-margin multi-label classification with applications to zero-shot learning. *Machine Learning*, 88(1-2): 127–155, 2012.

Hsu, D., Kakade, S., Langford, J., and Zhang, T. Multi-label prediction via compressed sensing. In *NIPS*, pp. 772–780, 2009.

Liu, W. and Tsang, I. W. Large margin metric learning for multi-label prediction. In *AAAI*, pp. 2800–2806, 2015a.

Liu, W. and Tsang, I. W. On the optimality of classifier chain for multi-label classification. In *NIPS*, pp. 712–720, 2015b.

Liu, W. and Tsang, I. W. Making decision trees feasible in ultrahigh feature and label dimensions. *Journal of Machine Learning Research*, 18(81):1–36, 2017.

Liu, W., Tsang, I. W., and Müller, K. An easy-to-hard learning paradigm for multiple classes and multiple labels. *Journal of Machine Learning Research*, 18(94):1–38, 2017.

Liu, W., Xu, D., Tsang, I. W., and Zhang, W. Metric learning for multi-output tasks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2):408–422, 2019.

Nesterov, Y. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer, 2014.

Peng, X., Feng, J., Xiao, S., Yau, W., Zhou, J. T., and Yang, S. Structured autoencoders for subspace clustering. *IEEE Trans. Image Processing*, 27(10):5076–5086, 2018a.

Peng, X., Lu, C., Yi, Z., and Tang, H. Connections between nuclear-norm and frobenius-norm-based representations. *IEEE Trans. Neural Netw. Learning Syst.*, 29(1):218–224, 2018b.

Prabhu, Y. and Varma, M. FastXML: A fast, accurate and stable tree-classifier for extreme multi-label learning. In *SIGKDD*, pp. 263–272, 2014.

Shen, D., Ruvini, J., Somaiya, M., and Sundaresan, N. Item categorization in the e-commerce domain. In *CIKM*, pp. 1921–1924, 2011.

Shen, X., Liu, W., Tsang, I. W., Sun, Q., and Ong, Y. Compact multi-label learning. In *AAAI*, pp. 4066–4073, 2018a.

Shen, X., Liu, W., Tsang, I. W., Sun, Q., and Ong, Y. Multilabel prediction via cross-view search. *IEEE Trans. Neural Netw. Learning Syst.*, 29(9):4324–4338, 2018b.

Tsoumakas, G., Zhang, M., and Zhou, Z. Introduction to the special issue on learning from multi-label data. *Machine Learning*, 88(1-2):1–4, 2012.

Wright, S. J., Nowak, R. D., and Figueiredo, M. A. T. Sparse reconstruction by separable approximation. *IEEE Trans. Signal Processing*, 57(7):2479–2493, 2009.

Yen, I. E., Huang, X., Ravikumar, P., Zhong, K., and D-hillon, I. S. PD-Sparse : A primal and dual sparse approach to extreme multiclass and multilabel classification. In *ICML*, pp. 3069–3077, 2016.

Yu, H., Jain, P., Kar, P., and Dhillon, I. S. Large-scale multi-label learning with missing labels. In *ICML*, pp. 593–601, 2014.

Zhang, C.-H. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38(2): 894–942, 2010.

Zhang, C.-H. and Huang, J. The sparsity and bias of the lasso selection in high-dimensional linear regression. *Annals of Statistics*, 36(4):1567–1594, 2008.

Zhang, L., Zhang, Q., Zhang, L., Tao, D., Huang, X., and Du, B. Ensemble manifold regularized sparse low-rank approximation for multiview feature embedding. *Pattern Recognition*, 48(10):3102–3112, 2015.

Zhou, J. T., Zhao, H., Peng, X., Fang, M., Qin, Z., and Goh, R. S. M. Transfer hashing: From shallow to deep. *IEEE Trans. Neural Netw. Learning Syst.*, 29(12):61916201, 2018.

Zou, H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.