Multi-Stage Multi-Task Feature Learning*

[†]**Pinghua Gong,** [‡]**Jieping Ye,** [†]**Changshui Zhang**
[†]State Key Laboratory on Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology (TNList)
Department of Automation, Tsinghua University, Beijing 100084, China
[‡]Computer Science and Engineering, Center for Evolutionary Medicine and Informatics
The Biodesign Institute, Arizona State University, Tempe, AZ 85287, USA
[†]{gph08@mails, zcs@mail}.tsinghua.edu.cn, [‡]jieping.ye@asu.edu

Abstract

Multi-task sparse feature learning aims to improve the generalization performance by exploiting the shared features among tasks. It has been successfully applied to many applications including computer vision and biomedical informatics. Most of the existing multi-task sparse feature learning algorithms are formulated as a convex sparse regularization problem, which is usually suboptimal, due to its looseness for approximating an ℓ_0-type regularizer. In this paper, we propose a non-convex formulation for multi-task sparse feature learning based on a novel regularizer. To solve the non-convex optimization problem, we propose a Multi-Stage Multi-Task Feature Learning (MSMTFL) algorithm. Moreover, we present a detailed theoretical analysis showing that MSMTFL achieves a better parameter estimation error bound than the convex formulation. Empirical studies on both synthetic and real-world data sets demonstrate the effectiveness of MSMTFL in comparison with the state of the art multi-task sparse feature learning algorithms.

1 Introduction

Multi-task learning (MTL) exploits the relationships among multiple related tasks to improve the generalization performance. It has been applied successfully to many applications such as speech classification [16], handwritten character recognition [14, 17] and medical diagnosis [2]. One common assumption in multi-task learning is that all tasks should share some common structures including the prior or parameters of Bayesian models [18, 21, 24], a similarity metric matrix [16], a classification weight vector [6], a low rank subspace [4, 13] and a common set of shared features [1, 8, 10, 11, 12, 14, 20].

In this paper, we focus on multi-task feature learning, in which we learn the features specific to each task as well as the common features shared among tasks. Although many multi-task feature learning algorithms have been proposed, most of them assume that the relevant features are shared by all tasks. This is too restrictive in real-world applications [9]. To overcome this limitation, Jalali et al. (2010) [9] proposed an $\ell_1 + \ell_{1,\infty}$ regularized formulation, called dirty model, to leverage the common features shared among tasks. The dirty model allows a certain feature to be shared by some tasks but not all tasks. Jalali et al. (2010) also presented a theoretical analysis under the incoherence condition [5, 15] which is more restrictive than RIP [3, 27]. The $\ell_1 + \ell_{1,\infty}$ regularizer is a convex relaxation for the ℓ_0-type one, which, however, is too loose to well approximate the ℓ_0-type regularizer and usually achieves suboptimal performance (requiring restrictive conditions or obtaining a suboptimal error bound) [23, 26, 27]. To remedy the shortcoming, we propose to use a non-convex regularizer for multi-task feature learning in this paper.

[*]This work was completed when the first author visited Arizona State University.

Contributions: We propose to employ a capped-ℓ_1,ℓ_1 regularized formulation (non-convex) to learn the features specific to each task as well as the common features shared among tasks. To solve the non-convex optimization problem, we propose a Multi-Stage Multi-Task Feature Learning (MSMTFL) algorithm, using the concave duality [26]. Although the MSMTFL algorithm may not obtain a globally optimal solution, we theoretically show that this solution achieves good performance. Specifically, we present a detailed theoretical analysis on the parameter estimation error bound for the MSMTFL algorithm. Our analysis shows that, under the sparse eigenvalue condition which is *weaker* than the incoherence condition in Jalali et al. (2010) [9], MSMTFL improves the error bound during the multi-stage iteration, i.e., the error bound at the current iteration improves the one at the last iteration. Empirical studies on both synthetic and real-world data sets demonstrate the effectiveness of the MSMTFL algorithm in comparison with the state of the art algorithms.

Notations: Scalars and vectors are denoted by lower case letters and bold face lower case letters, respectively. Matrices and sets are denoted by capital letters and calligraphic capital letters, respectively. The ℓ_1 norm, Euclidean norm, ℓ_∞ norm and Frobenius norm are denoted by $\|\cdot\|_1$, $\|\cdot\|$, $\|\cdot\|_\infty$ and $\|\cdot\|_F$, respectively. $|\cdot|$ denotes the absolute value of a scalar or the number of elements in a set, depending on the context. We define the $\ell_{p,q}$ norm of a matrix X as $\|X\|_{p,q} = \left(\sum_i \left((\sum_j |x_{ij}|^q)^{1/q}\right)^p\right)^{1/p}$. We define $\mathbb{N}_n$ as $\{1,\cdots,n\}$ and $N(\mu,\sigma^2)$ as a normal distribution with mean μ and variance σ^2. For a $d\times m$ matrix W and sets $\mathcal{I}_i \subseteq \mathbb{N}_d \times \{i\}, \mathcal{I} \subseteq \mathbb{N}_d \times \mathbb{N}_d$, we let $\mathbf{w}_{\mathcal{I}_i}$ be a $d\times 1$ vector with the j-th entry being w_{ji}, if $(j,i)\in\mathcal{I}_i$, and 0, otherwise. We also let $W_{\mathcal{I}}$ be a $d\times m$ matrix with the (j,i)-th entry being w_{ji}, if $(j,i)\in\mathcal{I}$, and 0, otherwise.

2 The Proposed Formulation

Assume we are given m learning tasks associated with training data $\{(X_1,\mathbf{y}_1),\cdots,(X_m,\mathbf{y}_m)\}$, where $X_i \in \mathbb{R}^{n_i\times d}$ is the data matrix of the i-th task with each row as a sample; $\mathbf{y}_i \in \mathbb{R}^{n_i}$ is the response of the i-th task; d is the data dimensionality; n_i is the number of samples for the i-th task. We consider learning a weight matrix $W = [\mathbf{w}_1,\cdots,\mathbf{w}_m] \in \mathbb{R}^{d\times m}$ consisting of the weight vectors for m linear predictive models: $\mathbf{y}_i \approx \mathbf{f}_i(X_i) = X_i\mathbf{w}_i,\ i\in\mathbb{N}_m$. In this paper, we propose a non-convex multi-task feature learning formulation to learn these m models simultaneously, based on the capped-ℓ_1,ℓ_1 regularization. Specifically, we first impose the ℓ_1 penalty on each row of W, obtaining a column vector. Then, we impose the capped-ℓ_1 penalty [26, 27] on that vector. Formally, we formulate our proposed model as follows:

$$\min_W \left\{ l(W) + \lambda \sum_{j=1}^{d} \min\left(\|\mathbf{w}^j\|_1, \theta\right) \right\}, \tag{1}$$

where $l(W)$ is an empirical loss function of W; λ (>0) is a parameter balancing the empirical loss and the regularization; θ (>0) is a thresholding parameter; $\mathbf{w}^j$ is the j-th row of the matrix W. In this paper, we focus on the quadratic loss function: $l(W) = \sum_{i=1}^m \frac{1}{mn_i}\|X_i\mathbf{w}_i - \mathbf{y}_i\|^2$.

Algorithm 1: MSMTFL: Multi-Stage Multi-Task Feature Learning

```
Initialize $\lambda_j^{(0)} = \lambda$;
for $\ell = 1, 2, \cdots$ do
    Let $\hat{W}^{(\ell)}$ be a solution of the following problem:
        $\min_{W\in\mathbb{R}^{d\times m}} \left\{ l(W) + \sum_{j=1}^{d} \lambda_j^{(\ell-1)} \|\mathbf{w}^j\|_1 \right\}.$   (2)
    Let $\lambda_j^{(\ell)} = \lambda I(\|(\hat{\mathbf{w}}^{(\ell)})^j\|_1 < \theta)$ $(j = 1,\cdots,d)$, where $(\hat{\mathbf{w}}^{(\ell)})^j$ is the $j$-th row of $\hat{W}^{(\ell)}$ and $I(\cdot)$
    denotes the $\{0,1\}$ valued indicator function.
end
```

Intuitively, due to the capped-ℓ_1, ℓ_1 penalty, the optimal solution of Eq. (1) denoted as $W^\star$ has many zero rows. For a nonzero row $(\mathbf{w}^\star)^k$, some entries may be zero, due to the ℓ_1-norm imposed on each row of W. Thus, under the formulation in Eq. (1), a certain feature can be shared by some tasks but not all the tasks. Therefore, the proposed formulation can leverage the common features shared among tasks.

The formulation in Eq. (1) is non-convex and is difficult to solve. To this end, we propose a Multi-Stage Multi-Task Feature Learning (MSMTFL) algorithm (see Algorithm 1). Note that if we terminate the algorithm with $\ell = 1$, the MSMTFL algorithm is equivalent to the ℓ_1 regularized multi-task feature learning algorithm (Lasso). Thus, the solution obtained by MSMTFL can be considered as a refinement of that of Lasso. Although Algorithm 1 may not find a globally optimal solution, the solution has good performance. Specifically, we will theoretically show that the solution obtained by Algorithm 1 improves the performance of the parameter estimation error bound during the multi-stage iteration. Moreover, empirical studies also demonstrate the effectiveness of our proposed MSMTFL algorithm. We provide more details about intuitive interpretations, convergence analysis and reproducibility discussions of the proposed algorithm in the full version [7].

3 Theoretical Analysis

In this section, we theoretically analyze the parameter estimation performance of the solution obtained by the MSMTFL algorithm. To simplify the notations in the theoretical analysis, we assume that the number of samples for all the tasks are the same. However, our theoretical analysis can be easily extended to the case where the tasks have different sample sizes.

We first present a sub-Gaussian noise assumption which is very common in the analysis of sparse regularization literature [23, 25, 26, 27].

Assumption 1 *Let $\bar{W} = [\bar{\mathbf{w}}_1, \cdots, \bar{\mathbf{w}}_m] \in \mathbb{R}^{d \times m}$ be the underlying sparse weight matrix and $\mathbf{y}_i = X_i \bar{\mathbf{w}}_i + \boldsymbol{\delta}_i$, $\mathbb{E}\mathbf{y}_i = X_i \bar{\mathbf{w}}_i$, where $\boldsymbol{\delta}_i \in \mathbb{R}^n$ is a random vector with all entries δ_{ji} $(j \in \mathbb{N}_n, i \in \mathbb{N}_m)$ being independent sub-Gaussians: there exists $\sigma > 0$ such that $\forall j \in \mathbb{N}_n, i \in \mathbb{N}_m, t \in \mathbb{R}$: $\mathbb{E}_{\delta_{ji}} \exp(t\delta_{ji}) \leq \exp\left(\sigma^2 t^2/2\right)$.*

Remark 1 *We call the random variable satisfying the condition in Assumption 1 sub-Gaussian, since its moment generating function is upper bounded by that of the zero mean Gaussian random variable. That is, if a normal random variable $x \sim N(0, \sigma^2)$, then we have $\mathbb{E}\exp(tx) = \int_{-\infty}^{\infty} \exp(tx) \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx = \exp(\sigma^2 t^2/2) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-(x - \sigma^2 t)^2/(2\sigma^2)\right) dx = \exp(\sigma^2 t^2/2) \geq \mathbb{E}_{\delta_{ji}} \exp(t\delta_{ji})$.*

Remark 2 *Based on the Hoeffding's Lemma, for any random variable $x \in [a, b]$ and $\mathbb{E}x = 0$, we have $\mathbb{E}(\exp(tx)) \leq \exp\left(\frac{t^2(b-a)^2}{8}\right)$. Therefore, both zero mean Gaussian and zero mean bounded random variables are sub-Guassians. Thus, the sub-Gaussian noise assumption is more general than the Gaussian noise assumption which is commonly used in the literature [9, 11].*

We next introduce the following sparse eigenvalue concept which is also common in the analysis of sparse regularization literature [22, 23, 25, 26, 27].

Definition 1 *Given $1 \leq k \leq d$, we define*

$$\rho_i^+(k) = \sup_{\mathbf{w}} \left\{ \frac{\|X_i \mathbf{w}\|^2}{n\|\mathbf{w}\|^2} : \|\mathbf{w}\|_0 \leq k \right\},\ \rho_{max}^+(k) = \max_{i \in \mathbb{N}_m} \rho_i^+(k),$$

$$\rho_i^-(k) = \inf_{\mathbf{w}} \left\{ \frac{\|X_i \mathbf{w}\|^2}{n\|\mathbf{w}\|^2} : \|\mathbf{w}\|_0 \leq k \right\},\ \rho_{min}^-(k) = \min_{i \in \mathbb{N}_m} \rho_i^-(k).$$

Remark 3 *$\rho_i^+(k)$ $(\rho_i^-(k))$ is in fact the maximum (minimum) eigenvalue of $(X_i)_{\mathcal{S}}^T (X_i)_{\mathcal{S}}/n$, where $\mathcal{S}$ is a set satisfying $|\mathcal{S}| \leq k$ and $(X_i)_{\mathcal{S}}$ is a submatrix composed of the columns of X_i indexed by $\mathcal{S}$. In the MTL setting, we need to exploit the relations of $\rho_i^+(k)$ $(\rho_i^-(k))$ among multiple tasks.*

We present our parameter estimation error bound on MSMTFL in the following theorem:

Theorem 1 *Let Assumption 1 hold. Define $\bar{\mathcal{F}}_i = \{(j,i) : \bar{w}_{ji} \neq 0\}$ and $\bar{\mathcal{F}} = \cup_{i \in \mathbb{N}_m} \bar{\mathcal{F}}_i$. Denote $\bar{r}$ as the number of nonzero rows of $\bar{W}$. We assume that*

$$\forall (j,i) \in \bar{\mathcal{F}}, \|\bar{\mathbf{w}}^j\|_1 \geq 2\theta \tag{3}$$

$$\text{and } \frac{\rho_i^+(s)}{\rho_i^-(2\bar{r}+2s)} \leq 1 + \frac{s}{2\bar{r}}, \tag{4}$$

where s is some integer satisfying $s \geq \bar{r}$. If we choose λ and θ such that for some $s \geq \bar{r}$:

$$\lambda \geq 12\sigma\sqrt{\frac{2\rho_{max}^+(1)\ln(2dm/\eta)}{n}}, \tag{5}$$

$$\theta \geq \frac{11m\lambda}{\rho_{min}^-(2\bar{r}+s)}, \tag{6}$$

then the following parameter estimation error bound holds with probability larger than $1-\eta$:

$$\|\hat{W}^{(\ell)} - \bar{W}\|_{2,1} \leq 0.8^{\ell/2}\frac{9.1m\lambda\sqrt{\bar{r}}}{\rho_{min}^-(2\bar{r}+s)} + \frac{39.5m\sigma\sqrt{\rho_{max}^+(\bar{r})(7.4\bar{r}+2.7\ln(2/\eta))/n}}{\rho_{min}^-(2\bar{r}+s)}, \tag{7}$$

where $\hat{W}^{(\ell)}$ is a solution of Eq. (2).

Remark 4 *Eq. (3) assumes that the ℓ_1-norm of each nonzero row of $\bar{W}$ is away from zero. This requires the true nonzero coefficients should be large enough, in order to distinguish them from the noise. Eq. (4) is called the sparse eigenvalue condition [27], which requires the eigenvalue ratio $\rho_i^+(s)/\rho_i^-(s)$ to grow sub-linearly with respect to s. Such a condition is very common in the analysis of sparse regularization [22, 25] and it is slightly weaker than the RIP condition [3, 27].*

Remark 5 *When $\ell = 1$ (corresponds to Lasso), the first term of the right-hand side of Eq. (7) dominates the error bound in the order of*

$$\|\hat{W}^{Lasso} - \bar{W}\|_{2,1} = O\left(m\sqrt{\bar{r}\ln(dm/\eta)/n}\right), \tag{8}$$

since λ satisfies the condition in Eq. (5). Note that the first term of the right-hand side of Eq. (7) shrinks exponentially as ℓ increases. When ℓ is sufficiently large in the order of $O(\ln(m\sqrt{\bar{r}/n}) + \ln\ln(dm))$, this term tends to zero and we obtain the following parameter estimation error bound:

$$\|\hat{W}^{(\ell)} - \bar{W}\|_{2,1} = O\left(m\sqrt{\bar{r}/n + \ln(1/\eta)/n}\right). \tag{9}$$

Jalali et al. (2010) [9] gave an $\ell_{\infty,\infty}$-norm error bound $\|\hat{W}^{Dirty} - \bar{W}\|_{\infty,\infty} = O\left(\sqrt{\ln(dm/\eta)/n}\right)$ as well as a sign consistency result between $\hat{W}$ and $\bar{W}$. A direct comparison between these two bounds is difficult due to the use of different norms. On the other hand, the worst-case estimate of the $\ell_{2,1}$-norm error bound of the algorithm in Jalali et al. (2010) [9] is in the same order with Eq. (8), that is: $\|\hat{W}^{Dirty} - \bar{W}\|_{2,1} = O\left(m\sqrt{\bar{r}\ln(dm/\eta)/n}\right)$. When dm is large and the ground truth has a large number of sparse rows (i.e., $\bar{r}$ is a small constant), the bound in Eq. (9) is significantly better than the ones for the Lasso and Dirty model.

Remark 6 *Jalali et al. (2010) [9] presented an $\ell_{\infty,\infty}$-norm parameter estimation error bound and hence a sign consistency result can be obtained. The results are derived under the incoherence condition which is more restrictive than the RIP condition and hence more restrictive than the sparse eigenvalue condition in Eq. (4). From the viewpoint of the parameter estimation error, our proposed algorithm can achieve a better bound under weaker conditions. Please refer to [19, 25, 27] for more details about the incoherence condition, the RIP condition, the sparse eigenvalue condition and their relationships.*

Remark 7 *The capped-ℓ_1 regularized formulation in Zhang (2010) [26] is a special case of our formulation when $m = 1$. However, extending the analysis from the single task to the multi-task setting is nontrivial. Different from previous work on multi-stage sparse learning which focuses on a single task [26, 27], we study a more general multi-stage framework in the multi-task setting. We need to exploit the relationship among tasks, by using the relations of sparse eigenvalues $\rho_i^+(k)$ ($\rho_i^-(k)$) and treating the ℓ_1-norm on each row of the weight matrix as a whole for consideration. Moreover, we simultaneously exploit the relations of each column and each row of the matrix.*

4 Proof Sketch

We first provide several important lemmas (please refer to the full version [7] or supplementary materials for detailed proofs) and then complete the proof of Theorem 1 based on these lemmas.

Lemma 1 *Let $\bar{\Upsilon} = [\bar{\boldsymbol{\epsilon}}_1, \cdots, \bar{\boldsymbol{\epsilon}}_m]$ with $\bar{\boldsymbol{\epsilon}}_i = [\bar{\epsilon}_{1i}, \cdots, \bar{\epsilon}_{di}]^T = \frac{1}{n} X_i^T (X_i \bar{\mathbf{w}}_i - \mathbf{y}_i)$ $(i \in \mathbb{N}_m)$. Define $\bar{\mathcal{H}} \supseteq \bar{\mathcal{F}}$ such that $(j,i) \in \bar{\mathcal{H}}$ $(\forall i \in \mathbb{N}_m)$, provided there exists $(j,g) \in \bar{\mathcal{F}}$ ($\bar{\mathcal{H}}$ is a set consisting of the indices of all entries in the nonzero rows of $\bar{W}$). Under the conditions of Assumption 1 and the notations of Theorem 1, the followings hold with probability larger than $1-\eta$:*

$$\|\bar{\Upsilon}\|_{\infty,\infty} \leq \sigma \sqrt{\frac{2\rho^+_{max}(1)\ln(2dm/\eta)}{n}}, \tag{10}$$

$$\|\bar{\Upsilon}_{\bar{\mathcal{H}}}\|_F^2 \leq m\sigma^2 \rho^+_{max}(\bar{r})(7.4\bar{r} + 2.7\ln(2/\eta))/n. \tag{11}$$

Lemma 1 gives bounds on the residual correlation ($\bar{\Upsilon}$) with respect to $\bar{W}$. We note that Eq. (10) and Eq. (11) are closely related to the assumption on λ in Eq. (5) and the second term of the right-hand side of Eq. (7) (error bound), respectively. This lemma provides a fundamental basis for the proof of Theorem 1.

Lemma 2 *Use the notations of Lemma 1 and consider $\mathcal{G}_i \subseteq \mathbb{N}_d \times \{i\}$ such that $\bar{\mathcal{F}}_i \cap \mathcal{G}_i = \emptyset$ $(i \in \mathbb{N}_m)$. Let $\hat{W} = \hat{W}^{(\ell)}$ be a solution of Eq. (2) and $\Delta\hat{W} = \hat{W} - \bar{W}$. Denote $\hat{\boldsymbol{\lambda}}_i = \hat{\boldsymbol{\lambda}}_i^{(\ell-1)} = [\lambda_1^{(\ell-1)}, \cdots, \lambda_d^{(\ell-1)}]^T$. Let $\hat{\lambda}_{\mathcal{G}_i} = \min_{(j,i)\in\mathcal{G}_i} \hat{\lambda}_{ji}$, $\hat{\lambda}_{\mathcal{G}} = \min_{i\in\mathcal{G}_i} \hat{\lambda}_{\mathcal{G}_i}$ and $\hat{\lambda}_{0i} = \max_j \hat{\lambda}_{ji}$, $\hat{\lambda}_0 = \max_i \hat{\lambda}_{0i}$. If $2\|\bar{\boldsymbol{\epsilon}}_i\|_\infty < \hat{\lambda}_{\mathcal{G}_i}$, then the following inequality holds at any stage $\ell \geq 1$:*

$$\sum_{i=1}^m \sum_{(j,i)\in\mathcal{G}_i} |\hat{w}_{ji}^{(\ell)}| \leq \frac{2\|\bar{\Upsilon}\|_{\infty,\infty} + \hat{\lambda}_0}{\hat{\lambda}_{\mathcal{G}} - 2\|\bar{\Upsilon}\|_{\infty,\infty}} \sum_{i=1}^m \sum_{(j,i)\in\mathcal{G}_i^c} |\Delta\hat{w}_{ji}^{(\ell)}|.$$

Denote $\mathcal{G} = \cup_{i\in\mathbb{N}_m}\mathcal{G}_i$, $\bar{\mathcal{F}} = \cup_{i\in\mathbb{N}_m}\bar{\mathcal{F}}_i$ and notice that $\bar{\mathcal{F}} \cap \mathcal{G} = \emptyset \Rightarrow \Delta\hat{W}^{(\ell)} = \hat{W}^{(\ell)}$. Lemma 2 says that $\|\Delta\hat{W}_{\mathcal{G}}^{(\ell)}\|_{1,1} = \|\hat{W}_{\mathcal{G}}^{(\ell)}\|_{1,1}$ is upper bounded in terms of $\|\Delta\hat{W}_{\mathcal{G}^c}^{(\ell)}\|_{1,1}$, which indicates that the error of the estimated coefficients locating outside of $\bar{\mathcal{F}}$ should be small enough. This provides an intuitive explanation why the parameter estimation error of our algorithm can be small.

Lemma 3 *Using the notations of Lemma 2, we denote $\mathcal{G} = \mathcal{G}_{(\ell)} = \bar{\mathcal{H}}^c \cap \{(j,i) : \hat{\lambda}_{ji}^{(\ell-1)} = \lambda\} = \cup_{i\in\mathbb{N}_m}\mathcal{G}_i$ with $\bar{\mathcal{H}}$ being defined as in Lemma 1 and $\mathcal{G}_i \subseteq \mathbb{N}_d \times \{i\}$. Let $\mathcal{J}_i$ be the indices of the largest s coefficients (in absolute value) of $\hat{\mathbf{w}}_{\mathcal{G}_i}$, $\mathcal{I}_i = \mathcal{G}_i^c \cup \mathcal{J}_i$, $\mathcal{I} = \cup_{i\in\mathbb{N}_m}\mathcal{I}_i$ and $\bar{\mathcal{F}} = \cup_{i\in\mathbb{N}_m}\bar{\mathcal{F}}_i$. Then, the following inequalities hold at any stage $\ell \geq 1$:*

$$\|\Delta\hat{W}^{(\ell)}\|_{2,1} \leq \frac{\left(1 + 1.5\sqrt{\frac{2\bar{r}}{s}}\right)\sqrt{8m\left(4\|\bar{\Upsilon}_{\mathcal{G}^c_{(\ell)}}\|_F^2 + \sum_{(j,i)\in\bar{\mathcal{F}}}(\hat{\lambda}_{ji}^{(\ell-1)})^2\right)}}{\rho^-_{min}(2\bar{r}+s)}, \tag{12}$$

$$\|\Delta\hat{W}^{(\ell)}\|_{2,1} \leq \frac{9.1m\lambda\sqrt{\bar{r}}}{\rho^-_{min}(2\bar{r}+s)}. \tag{13}$$

Lemma 3 is established based on Lemma 2, by considering the relationship between Eq. (5) and Eq. (10), and the specific definition of $\mathcal{G} = \mathcal{G}_{(\ell)}$. Eq. (12) provides a parameter estimation error bound in terms of $\ell_{2,1}$-norm by $\|\bar{\Upsilon}_{\mathcal{G}^c_{(\ell)}}\|_F^2$ and the regularization parameters $\hat{\lambda}_{ji}^{(\ell-1)}$ (see the definition of $\hat{\lambda}_{ji}$ ($\hat{\lambda}_{ji}^{(\ell-1)}$) in Lemma 2). This is the result directly used in the proof of Theorem 1. Eq. (13) states that the error bound is upper bounded in terms of λ, the right-hand side of which constitutes the shrinkage part of the error bound in Eq. (7).

Lemma 4 *Let $\hat{\lambda}_{ji} = \lambda I\left(\|\hat{\mathbf{w}}^j\|_1 < \theta, j \in \mathbb{N}_d\right), \forall i \in \mathbb{N}_m$ with some $\hat{W} \in \mathbb{R}^{d\times m}$. $\bar{\mathcal{H}} \supseteq \bar{\mathcal{F}}$ is defined in Lemma 1. Then under the condition of Eq. (3), we have:*

$$\sum_{(j,i)\in\bar{\mathcal{F}}} \hat{\lambda}_{ji}^2 \leq \sum_{(j,i)\in\bar{\mathcal{H}}} \hat{\lambda}_{ji}^2 \leq m\lambda^2\|\bar{W}_{\bar{\mathcal{H}}} - \hat{W}_{\bar{\mathcal{H}}}\|_{2,1}^2/\theta^2.$$

Lemma 4 establishes an upper bound of $\sum_{(j,i)\in\bar{\mathcal{F}}} \hat{\lambda}_{ji}^2$ by $\|\bar{W}_{\bar{\mathcal{H}}} - \hat{W}_{\bar{\mathcal{H}}}\|_{2,1}^2$, which is critical for building the recursive relationship between $\|\hat{W}^{(\ell)} - \bar{W}\|_{2,1}$ and $\|\hat{W}^{(\ell-1)} - \bar{W}\|_{2,1}$ in the proof of Theorem 1. This recursive relation is crucial for the shrinkage part of the error bound in Eq. (7).

4.1 Proof of Theorem 1

Proof For notational simplicity, we denote the right-hand side of Eq. (11) as:

$$u = m\sigma^2\rho_{max}^+(\bar{r})(7.4\bar{r} + 2.7\ln(2/\eta))/n. \tag{14}$$

Based on $\bar{\mathcal{H}} \subseteq \mathcal{G}_{(\ell)}^c$, Lemma 1 and Eq. (5), the followings hold with probability larger than $1-\eta$:

$$\begin{aligned}\|\bar{\Upsilon}_{\mathcal{G}_{(\ell)}^c}\|_F^2 &= \|\bar{\Upsilon}_{\bar{\mathcal{H}}}\|_F^2 + \|\bar{\Upsilon}_{\mathcal{G}_{(\ell)}^c\setminus\bar{\mathcal{H}}}\|_F^2 \le u + |\mathcal{G}_{(\ell)}^c \setminus \bar{\mathcal{H}}|\|\bar{\Upsilon}\|_{\infty,\infty}^2 \\ &\le u + \lambda^2|\mathcal{G}_{(\ell)}^c \setminus \bar{\mathcal{H}}|/144 \le u + (1/144)m\lambda^2\theta^{-2}\|\hat{W}_{\mathcal{G}_{(\ell)}^c\setminus\bar{\mathcal{H}}}^{(\ell-1)} - \bar{W}_{\mathcal{G}_{(\ell)}^c\setminus\bar{\mathcal{H}}}\|_{2,1}^2,\end{aligned} \tag{15}$$

where the last inequality follows from $\forall(j,i) \in \mathcal{G}_{(\ell)}^c \setminus \bar{\mathcal{H}}, \|(\hat{\mathbf{w}}^{(\ell-1)})^j\|_1^2/\theta^2 = \|(\hat{\mathbf{w}}^{(\ell-1)})^j - \bar{\mathbf{w}}^j\|_1^2/\theta^2 \ge 1 \Rightarrow |\mathcal{G}_{(\ell)}^c \setminus \bar{\mathcal{H}}| \le m\theta^{-2}\|\hat{W}_{\mathcal{G}_{(\ell)}^c\setminus\bar{\mathcal{H}}}^{(\ell-1)} - \bar{W}_{\mathcal{G}_{(\ell)}^c\setminus\bar{\mathcal{H}}}\|_{2,1}^2$. According to Eq. (12), we have:

$$\|\hat{W}^{(\ell)} - \bar{W}\|_{2,1}^2 = \|\Delta\hat{W}^{(\ell)}\|_{2,1}^2 \le \frac{8m\left(1+1.5\sqrt{\frac{2\bar{r}}{s}}\right)^2\left(4\|\bar{\Upsilon}_{\mathcal{G}_{(\ell)}^c}\|_F^2 + \sum_{(j,i)\in\bar{\mathcal{F}}}(\hat{\lambda}_{ji}^{(\ell-1)})^2\right)}{(\rho_{min}^-(2\bar{r}+s))^2}$$

$$\le \frac{78m\left(4u + (37/36)m\lambda^2\theta^{-2}\left\|\hat{W}^{(\ell-1)} - \bar{W}\right\|_{2,1}^2\right)}{(\rho_{min}^-(2\bar{r}+s))^2} \le \frac{312mu}{(\rho_{min}^-(2\bar{r}+s))^2} + 0.8\left\|\hat{W}^{(\ell-1)} - \bar{W}\right\|_{2,1}^2$$

$$\le 0.8^\ell\left\|\hat{W}^{(0)} - \bar{W}\right\|_{2,1}^2 + \frac{312mu}{(\rho_{min}^-(2\bar{r}+s))^2}\frac{1-0.8^\ell}{1-0.8} \le 0.8^\ell\frac{9.1^2m^2\lambda^2\bar{r}}{(\rho_{min}^-(2\bar{r}+s))^2} + \frac{1560mu}{(\rho_{min}^-(2\bar{r}+s))^2}.$$

In the above derivation, the first inequality is due to Eq. (12); the second inequality is due to the assumption $s \ge \bar{r}$ in Theorem 1, Eq. (15) and Lemma 4; the third inequality is due to Eq. (6); the last inequality follows from Eq. (13) and $1 - 0.8^\ell \le 1$ ($\ell \ge 1$). Thus, following the inequality $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ ($\forall a, b \ge 0$), we obtain:

$$\|\hat{W}^{(\ell)} - \bar{W}\|_{2,1} \le 0.8^{\ell/2}\frac{9.1m\lambda\sqrt{\bar{r}}}{\rho_{min}^-(2\bar{r}+s)} + \frac{39.5\sqrt{mu}}{\rho_{min}^-(2\bar{r}+s)}.$$

Substituting Eq. (14) into the above inequality, we verify Theorem 1. □

5 Experiments

We compare our proposed MSMTFL algorithm with three competing multi-task feature learning algorithms: ℓ_1-norm multi-task feature learning algorithm (Lasso), $\ell_{1,2}$-norm multi-task feature learning algorithm (L1,2) [14] and dirty model multi-task feature learning algorithm (DirtyMTL) [9]. In our experiments, we employ the quadratic loss function for all the compared algorithms.

5.1 Synthetic Data Experiments

We generate synthetic data by setting the number of tasks as m and each task has n samples which are of dimensionality d; each element of the data matrix $X_i \in \mathbb{R}^{n\times d}$ ($i \in \mathbb{N}_m$) for the i-th task is sampled i.i.d. from the Gaussian distribution $N(0,1)$ and we then normalize all columns to length 1; each entry of the underlying true weight $\bar{W} \in \mathbb{R}^{d\times m}$ is sampled i.i.d. from the uniform distribution in the interval $[-10, 10]$; we randomly set 90% rows of $\bar{W}$ as zero vectors and 80% elements of the remaining nonzero entries as zeros; each entry of the noise $\boldsymbol{\delta}_i \in \mathbb{R}^n$ is sampled i.i.d. from the Gaussian distribution $N(0,\sigma^2)$; the responses are computed as $\mathbf{y}_i = X_i\bar{\mathbf{w}}_i + \boldsymbol{\delta}_i$ ($i \in \mathbb{N}_m$).

We first report the averaged parameter estimation error $\|\hat{W} - \bar{W}\|_{2,1}$ vs. Stage (ℓ) plots for MSMTFL (Figure 1). We observe that the error decreases as ℓ increases, which shows the advantage of our proposed algorithm over Lasso. This is consistent with the theoretical result in Theorem 1. Moreover, the parameter estimation error decreases quickly and converges in a few stages.

We then report the averaged parameter estimation error $\|\hat{W} - \bar{W}\|_{2,1}$ in comparison with four algorithms in different parameter settings (Figure 2). For a fair comparison, we compare the smallest estimation errors of the four algorithms in all the parameter settings [25, 26]. As expected, the parameter estimation error of the MSMTFL algorithm is the smallest among the four algorithms. This empirical result demonstrates the effectiveness of the MSMTFL algorithm. We also have the following observations: (a) When λ is large enough, all four algorithms tend to have the same parameter estimation error. This is reasonable, because the solutions $\hat{W}$'s obtained by the four algorithms are all zero matrices, when λ is very large. (b) The performance of the MSMTFL algorithm is similar for different θ's, when λ exceeds a certain value.

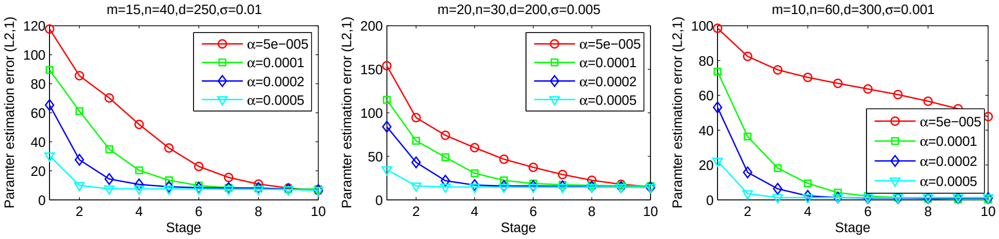

Figure 1: Averaged parameter estimation error $\|\hat{W} - \bar{W}\|_{2,1}$ vs. Stage (ℓ) plots for MSMTFL on the synthetic data set (averaged over 10 runs). Here we set $\lambda = \alpha\sqrt{\ln(dm)/n}$, $\theta = 50m\lambda$. Note that $\ell = 1$ corresponds to Lasso; the results show the stage-wise improvement over Lasso.

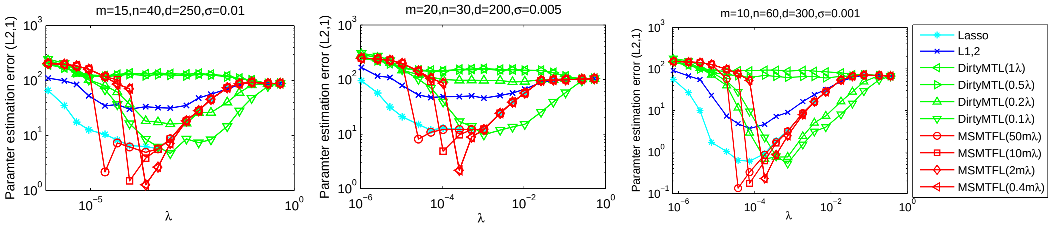

Figure 2: Averaged parameter estimation error $\|\hat{W} - \bar{W}\|_{2,1}$ vs. λ plots on the synthetic data set (averaged over 10 runs). MSMTFL has the smallest parameter estimation error among the four algorithms. Both DirtyMTL and MSMTFL have two parameters; we set $\lambda_s/\lambda_b = 1, 0.5, 0.2, 0.1$ for DirtyMTL ($1/m \leq \lambda_s/\lambda_b \leq 1$ was adopted in Jalali et al. (2010) [9]) and $\theta/\lambda = 50m, 10m, 2m, 0.4m$ for MSMTFL.

5.2 Real-World Data Experiments

We conduct experiments on two real-world data sets: MRI and Isolet data sets. (1) The MRI data set is collected from the ANDI database, which contains 675 patients' MRI data preprocessed using FreeSurfer[1]. The MRI data include 306 features and the response (target) is the Mini Mental State Examination (MMSE) score coming from 6 different time points: M06, M12, M18, M24, M36, and M48. We remove the samples which fail the MRI quality controls and have missing entries. Thus, we have 6 tasks with each task corresponding to a time point and the sample sizes corresponding to 6 tasks are 648, 642, 293, 569, 389 and 87, respectively. (2) The Isolet data set[2] is collected from 150 speakers who speak the name of each English letter of the alphabet twice. Thus, there are 52 samples from each speaker. The speakers are grouped into 5 subsets which respectively include 30 similar speakers, and the subsets are named Isolet1, Isolet2, Isolet3, Isolet4, and Isolet5. Thus, we naturally have 5 tasks with each task corresponding to a subset. The 5 tasks respectively have 1560, 1560, 1560, 1558, and 1559 samples (Three samples are historically missing), where each sample includes 617 features and the response is the English letter label (1-26).

In the experiments, we treat the MMSE and letter labels as the regression values for the MRI data set and the Isolet data set, respectively. For both data sets, we randomly extract the training samples from each task with different training ratios (15%, 20% and 25%) and use the rest of samples to form the test set. We evaluate the four multi-task feature learning algorithms in terms of normalized mean squared error (nMSE) and averaged means squared error (aMSE), which are commonly used in

[1] www.loni.ucla.edu/ADNI/

[2] www.zjucadcg.cn/dengcai/Data/data.html

Table 1: Comparison of four multi-task feature learning algorithms on the MRI data set in terms of averaged nMSE and aMSE (standard deviation), which are averaged over 10 random splittings.

measure	traning ratio	Lasso	L1,2	DirtyMTL	MSMTFL
nMSE	0.15	0.6651(0.0280)	0.6633(0.0470)	0.6224(0.0265)	**0.5539(0.0154)**
	0.20	0.6254(0.0212)	0.6489(0.0275)	0.6140(0.0185)	**0.5542(0.0139)**
	0.25	0.6105(0.0186)	0.6577(0.0194)	0.6136(0.0180)	**0.5507(0.0142)**
aMSE	0.15	0.0189(0.0008)	0.0187(0.0010)	0.0172(0.0006)	**0.0159(0.0004)**
	0.20	0.0179(0.0006)	0.0184(0.0005)	0.0171(0.0005)	**0.0161(0.0004)**
	0.25	0.0172(0.0009)	0.0183(0.0006)	0.0167(0.0008)	**0.0157(0.0006)**

multi-task learning problems [28, 29]. For each training ratio, both nMSE and aMSE are averaged over 10 random splittings of training and test sets and the standard deviation is also shown. All parameters of the four algorithms are tuned via 3-fold cross validation.

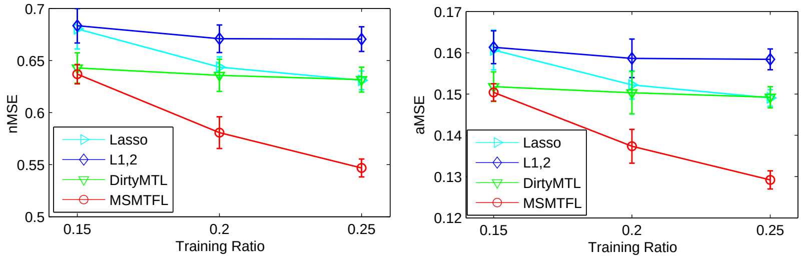

Figure 3: Averaged test error (nMSE and aMSE) vs. training ratio plots on the Isolet data set. The results are averaged over 10 random splittings.

Table 1 and Figure 3 show the experimental results in terms of averaged nMSE (aMSE) and the standard deviation. From these results, we observe that: (a) Our proposed MSMTFL algorithm outperforms all the competing feature learning algorithms on both data sets, with the smallest regression errors (nMSE and aMSE) as well as the smallest standard deviations. (b) On the MRI data set, the MSMTFL algorithm performs well even in the case of a small training ratio. The performance for the 15% training ratio is comparable to that for the 25% training ratio. (c) On the Isolet data set, when the training ratio increases from 15% to 25%, the performance of the MSMTFL algorithm increases and the superiority of the MSMTFL algorithm over the other three algorithms is more significant. Our results demonstrate the effectiveness of the proposed algorithm.

6 Conclusions

In this paper, we propose a non-convex multi-task feature learning formulation based on the capped-ℓ_1,ℓ_1 regularization. The proposed formulation learns the specific features of each task as well as the common features shared among tasks. We propose to solve the non-convex optimization problem by employing a Multi-Stage Multi-Task Feature Learning (MSMTFL) algorithm, using concave duality. We also present a detailed theoretical analysis in terms of the parameter estimation error bound for the MSMTFL algorithm. The analysis shows that our MSMTFL algorithm achieves good performance under the sparse eigenvalue condition, which is weaker than the incoherence condition. Experimental results on both synthetic and real-world data sets demonstrate the effectiveness of our proposed MSMTFL algorithm in comparison with the state of the art multi-task feature learning algorithms. In our future work, we will focus on a general non-convex regularization framework for multi-task feature learning settings (involving different loss functions and non-convex regularization terms) and derive theoretical bounds.

Acknowledgements

This work is supported in part by 973 Program (2013CB329503), NSFC (Grant No. 91120301, 60835002 and 61075004), NIH (R01 LM010730) and NSF (IIS-0953662, CCF-1025177).

References

[1] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[2] J. Bi, T. Xiong, S. Yu, M. Dundar, and R. Rao. An improved multi-task learning approach with applications in medical diagnosis. *Machine Learning and Knowledge Discovery in Databases*, pages 117–132, 2008.

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

[4] J. Chen, J. Liu, and J. Ye. Learning incoherent sparse and low-rank patterns from multiple tasks. In *SIGKDD*, pages 1179–1188, 2010.

[5] D. Donoho, M. Elad, and V. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory*, 52(1):6–18, 2006.

[6] T. Evgeniou and M. Pontil. Regularized multi–task learning. In *SIGKDD*, pages 109–117, 2004.

[7] P. Gong, J. Ye, and C. Zhang. Multi-stage multi-task feature learning. *arXiv:1210.5806*, 2012.

[8] P. Gong, J. Ye, and C. Zhang. Robust multi-task feature learning. In *SIGKDD*, pages 895–903, 2012.

[9] A. Jalali, P. Ravikumar, S. Sanghavi, and C. Ruan. A dirty model for multi-task learning. In *NIPS*, pages 964–972, 2010.

[10] S. Kim and E. Xing. Tree-guided group lasso for multi-task regression with structured sparsity. In *ICML*, pages 543–550, 2009.

[11] K. Lounici, M. Pontil, A. Tsybakov, and S. Van De Geer. Taking advantage of sparsity in multi-task learning. In *COLT*, pages 73–82, 2009.

[12] S. Negahban and M. Wainwright. Joint support recovery under high-dimensional scaling: Benefits and perils of $\ell_{1,\infty}$-regularization. In *NIPS*, pages 1161–1168, 2008.

[13] S. Negahban and M. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2):1069–1097, 2011.

[14] G. Obozinski, B. Taskar, and M. Jordan. Multi-task feature selection. *Statistics Department, UC Berkeley, Tech. Rep*, 2006.

[15] G. Obozinski, M. Wainwright, and M. Jordan. Support union recovery in high-dimensional multivariate regression. *Annals of statistics*, 39(1):1–47, 2011.

[16] S. Parameswaran and K. Weinberger. Large margin multi-task metric learning. In *NIPS*, pages 1867–1875, 2010.

[17] N. Quadrianto, A. Smola, T. Caetano, S. Vishwanathan, and J. Petterson. Multitask learning without label correspondences. In *NIPS*, pages 1957–1965, 2010.

[18] A. Schwaighofer, V. Tresp, and K. Yu. Learning gaussian process kernels via hierarchical bayes. In *NIPS*, pages 1209–1216, 2005.

[19] S. Van De Geer and P. Bühlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[20] X. Yang, S. Kim, and E. Xing. Heterogeneous multitask learning with joint sparsity constraints. In *NIPS*, pages 2151–2159, 2009.

[21] K. Yu, V. Tresp, and A. Schwaighofer. Learning gaussian processes from multiple tasks. In *ICML*, pages 1012–1019, 2005.

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

[23] C. Zhang and T. Zhang. A general theory of concave regularization for high dimensional sparse estimation problems. *Statistical Science*, 2012.

[24] J. Zhang, Z. Ghahramani, and Y. Yang. Learning multiple related tasks using latent independent component analysis. In *NIPS*, pages 1585–1592, 2006.

[25] T. Zhang. Some sharp performance bounds for least squares regression with ℓ_1 regularization. *The Annals of Statistics*, 37:2109–2144, 2009.

[26] T. Zhang. Analysis of multi-stage convex relaxation for sparse regularization. *JMLR*, 11:1081–1107, 2010.

[27] T. Zhang. Multi-stage convex relaxation for feature selection. *Bernoulli*, 2012.

[28] Y. Zhang and D. Yeung. Multi-task learning using generalized t process. In *AISTATS*, 2010.

[29] J. Zhou, J. Chen, and J. Ye. Clustered multi-task learning via alternating structure optimization. In *NIPS*, pages 702–710, 2011.