Journal of Machine Learning Research 14 (2013) 2979-3010 Submitted 10/12; Revised 6/13; Published 10/13

Multi-Stage Multi-Task Feature Learning

Pinghua Gong GPH08@MAILS.TSINGHUA.EDU.CN
State Key Laboratory on Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Department of Automation, Tsinghua University
Beijing 100084, China

Jieping Ye JIEPING.YE@ASU.EDU
Computer Science and Engineering
Center for Evolutionary Medicine and Informatics
The Biodesign Institute
Arizona State University
Tempe, AZ 85287, USA

Changshui Zhang ZCS@MAIL.TSINGHUA.EDU.CN
State Key Laboratory on Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Department of Automation, Tsinghua University
Beijing 100084, China

Editor: Hui Zou

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 non-convex regularizer. To solve the non-convex optimization problem, we propose a Multi-Stage Multi-Task Feature Learning (MSMTFL) algorithm; we also provide intuitive interpretations, detailed convergence and reproducibility analysis for the proposed 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.

Keywords: multi-task learning, multi-stage, non-convex, sparse learning

1. Introduction

Multi-task learning (MTL) (Caruana, 1997) exploits the relationships among multiple related tasks to improve the generalization performance. It has been successfully applied to many applications such as speech classification (Parameswaran and Weinberger, 2010), handwritten character recognition (Obozinski et al., 2006; Quadrianto et al., 2010) and medical diagnosis (Bi et al., 2008). One common assumption in multi-task learning is that all tasks should share some common structures including the prior or parameters of Bayesian models (Schwaighofer et al., 2005; Yu et al., 2005; Zhang et al., 2006), a similarity metric matrix (Parameswaran and Weinberger, 2010), a classification weight vector (Evgeniou and Pontil, 2004), a low rank subspace (Chen et al., 2010; Negahban and Wainwright, 2011) and a common set of shared features (Argyriou et al., 2008; Gong et al., 2012; Kim and Xing, 2009; Kolar et al., 2011; Lounici et al., 2009; Liu et al., 2009; Negahban and Wainwright, 2008; Obozinski et al., 2006; Yang et al., 2009; Zhang et al., 2010).

Multi-task feature learning, which aims to learn a common set of shared features, has received a lot of interests in machine learning recently, due to the popularity of various sparse learning formulations and their successful applications in many problems. In this paper, we focus on a specific multi-task feature learning setting, 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 in the past, many of them require the relevant features to be shared by all tasks. This is too restrictive in real-world applications (Jalali et al., 2010). To overcome this limitation, Jalali et al. (2010) 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 (Donoho et al., 2006; Obozinski et al., 2011) which is more restrictive than RIP (Candes and Tao, 2005; Zhang, 2012). The $\ell_1 + \ell_{1,\infty}$ regularizer is a convex relaxation of an ℓ_0-type one, in which a globally optimal solution can be obtained. However, a convex regularizer is known to be too loose to approximate the ℓ_0-type one and often achieves suboptimal performance (either require restrictive conditions or obtain a suboptimal error bound) (Zou and Li, 2008; Zhang, 2010, 2012; Zhang and Zhang, 2012; Shen et al., 2012; Fan et al., 2012). To remedy the limitation, a non-convex regularizer can be used instead. However, the non-convex formulation is usually difficult to solve and a globally optimal solution can not be obtained in most practical problems. Moreover, the solution of the non-convex formulation heavily depends on the specific optimization algorithms employed. Even with the same optimization algorithm adopted, different initializations usually lead to different solutions. Thus, it is often challenging to analyze the theoretical behavior of a non-convex formulation.

We propose a non-convex formulation, called capped-ℓ_1,ℓ_1 regularized model for multi-task feature learning. The proposed model aims to simultaneously learn the features specific to each task as well as the common features shared among tasks. We propose a Multi-Stage Multi-Task Feature Learning (MSMTFL) algorithm to solve the non-convex optimization problem. We also provide intuitive interpretations of the proposed algorithm from several aspects. In addition, we present a detailed convergence analysis for the proposed algorithm. To address the reproducibility issue of the non-convex formulation, we show that the solution generated by the MSMTFL algorithm is unique (i.e., the solution is reproducible) under a mild condition, which facilitates the theoretical analysis of the MSMTFL algorithm. Although the MSMTFL algorithm may not obtain a globally optimal solution, we 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 used in Jalali et al. (2010), MSMTFL improves the error bound during the multi-stage iteration, that is, 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.

1.1 Notations and Organization

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 the normal distribution with mean μ and variance σ^2. For a $d\times m$ matrix W and sets $I_i \subseteq \mathbb{N}_d \times \{i\}$, $I \subseteq \mathbb{N}_d \times \mathbb{N}_d$, we let $\mathbf{w}_{I_i}$ be the $d\times 1$ vector with the j-th entry being w_{ji}, if $(j,i) \in I_i$, and 0, otherwise. We also let W_I be a $d\times m$ matrix with the (j,i)-th entry being w_{ji}, if $(j,i)\in I$, and 0, otherwise.

In Section 2, we introduce a non-convex formulation and present the corresponding optimization algorithm. In Section 3, we discuss the convergence and reproducibility issues of the MSMTFL algorithm. In Section 4, we present a detailed theoretical analysis on the MSMTFL algorithm, in terms of the parameter estimation error bound. In Section 5, we provide a sketch of the proof of the presented theoretical results (the detailed proof is provided in the Appendix). In Section 6, we report the experimental results and we conclude the paper in Section 7.

2. The Proposed Formulation and Algorithm

In this section, we first propose a non-convex formulation for multi-task feature learning, based on the capped-ℓ_1,ℓ_1 regularization. Then, we show how to solve the corresponding non-convex optimization problem. Finally, we provide intuitive interpretations and discussions for the proposed algorithm.

2.1 A Non-convex 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 $W = [\mathbf{w}_1,\cdots,\mathbf{w}_m] \in \mathbb{R}^{d\times m}$ $(\mathbf{w}_i \in \mathbb{R}^d, i \in \mathbb{N}_m)$ 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 (Zhang, 2010, 2012) on that vector. Formally, we formulate our proposed model as follows:

$$\text{capped}-\ell_1,\ell_1: \quad \min_{W\in\mathbb{R}^{d\times m}} \left\{ l(W) + \lambda \sum_{j=1}^{d} \min\left(\|\mathbf{w}^j\|_1, \theta\right) \right\}, \tag{1}$$

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 following quadratic loss function:

$$l(W) = \sum_{i=1}^{m} \frac{1}{mn_i} \|X_i\mathbf{w}_i - \mathbf{y}_i\|^2. \tag{2}$$

Intuitively, due to the capped-ℓ_1,ℓ_1 penalty, the optimal solution of Equation (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 Equation (1), some features can be shared by some tasks but not all the tasks. Therefore, the proposed formulation can leverage the common features shared among tasks.

2.2 Two Relevant Non-convex Formulations

In this subsection, we discuss two relevant non-convex formulations. The first one is the capped-ℓ_1 feature learning formulation:

$$\text{capped}-\ell_1: \quad \min_{W\in\mathbb{R}^{d\times m}} \left\{ l(W) + \lambda \sum_{j=1}^{d}\sum_{i=1}^{m} \min\left(|w_{ji}|, \theta\right) \right\}. \tag{3}$$

Although the optimal solution of formulation (3) has a similar sparse pattern to that of the proposed capped-ℓ_1,ℓ_1 multi-task feature learning (i.e., the optimal solution can have many zero rows and enable some entries of a non-zero row to be zero), the models for different tasks decouple and thus formulation (3) is equivalent to the single task feature learning. Thus, the existing analysis for the single task setting in Zhang (2010, 2012) can be trivially adapted to this setting. The second one is the capped-ℓ_1,ℓ_2 multi-task feature learning formulation:

$$\text{capped}-\ell_1,\ell_2: \quad \min_{W\in\mathbb{R}^{d\times m}} \left\{ l(W) + \lambda \sum_{j=1}^{d} \min\left(\|\mathbf{w}^j\|, \theta\right) \right\}. \tag{4}$$

Due to the use of the capped-ℓ_1,ℓ_2 penalty, the optimal solution $W^\star$ of formulation (4) has many zero rows. However, any non-zero row of $W^\star$ is less likely to contain zero entries because of the Euclidean norm imposed on the rows of W. In other words, each row of $W^\star$ is either a zero vector or a vector composed of all non-zero entries. Thus, in this setting, some relevant features are required to be shared by all tasks. This is obviously different from the motivation of the proposed capped-ℓ_1,ℓ_1 multi-task feature learning, that is, some features are shared by some tasks but not all the tasks.

2.3 Optimization Algorithm

The formulation in Equation (1) is non-convex and is difficult to solve. In this paper, we propose an algorithm called Multi-Stage Multi-Task Feature Learning (MSMTFL) to solve the optimization problem (see details in Algorithm 1).[1] In this algorithm, a key step is how to efficiently solve Equation (5). Observing that the objective function in Equation (5) can be decomposed into the sum of a differential loss function and a non-differential regularization term, we employ FISTA (Beck and Teboulle, 2009) to solve the sub-problem. In the following, we present some intuitive interpretations of the proposed algorithm from several aspects.

1. We can use MSMTFL-type algorithms to solve the non-convex multi-task feature learning problems in Eqs. (3) and (4). Please refer to Appendix C for details.

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:

$$\min_{W \in \mathbb{R}^{d \times m}} \left\{ l(W) + \sum_{j=1}^{d} \lambda_j^{(\ell-1)} \|\mathbf{w}^j\|_1 \right\}. \tag{5}$$

    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

2.3.1 LOCALLY LINEAR APPROXIMATION

First, we define two auxiliary functions:

$$\mathbf{h} : \mathbb{R}^{d \times m} \mapsto \mathbb{R}_+^d, \; \mathbf{h}(W) = \left[\|\mathbf{w}^1\|_1, \cdots, \|\mathbf{w}^d\|_1 \right]^T,$$

$$g : \mathbb{R}_+^d \mapsto \mathbb{R}_+, \; g(\mathbf{u}) = \sum_{j=1}^{d} \min(u_j, \theta).$$

We note that $g(\cdot)$ is a concave function and we say that a vector $\mathbf{s} \in \mathbb{R}^d$ is a sub-gradient of g at $\mathbf{v} \in \mathbb{R}_+^d$, if for all vector $\mathbf{u} \in \mathbb{R}_+^d$, the following inequality holds:

$$g(\mathbf{u}) \leq g(\mathbf{v}) + \langle \mathbf{s}, \mathbf{u} - \mathbf{v} \rangle,$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product. Using the functions defined above, Equation (1) can be equivalently rewritten as follows:

$$\min_{W \in \mathbb{R}^{d \times m}} \left\{ l(W) + \lambda g(\mathbf{h}(W)) \right\}. \tag{6}$$

Based on the definition of the sub-gradient for a concave function given above, we can obtain an upper bound of $g(\mathbf{h}(W))$ using a locally linear approximation at $\mathbf{h}(\hat{W}^{(\ell)})$:

$$g(\mathbf{h}(W)) \leq g(\mathbf{h}(\hat{W}^{(\ell)})) + \left\langle \mathbf{s}^{(\ell)}, \mathbf{h}(W) - \mathbf{h}(\hat{W}^{(\ell)}) \right\rangle,$$

where $\mathbf{s}^{(\ell)}$ is a sub-gradient of $g(\mathbf{u})$ at $\mathbf{u} = \mathbf{h}(\hat{W}^{(\ell)})$. Furthermore, we can obtain an upper bound of the objective function in Equation (6), if the solution $\hat{W}^{(\ell)}$ at the ℓ-th iteration is available:

$$\forall W \in \mathbb{R}^{d \times m} : l(W) + \lambda g(\mathbf{h}(W)) \leq l(W) + \lambda g(\mathbf{h}(\hat{W}^{(\ell)})) + \lambda \left\langle \mathbf{s}^{(\ell)}, \mathbf{h}(W) - \mathbf{h}(\hat{W}^{(\ell)}) \right\rangle. \tag{7}$$

It can be shown that a sub-gradient of $g(\mathbf{u})$ at $\mathbf{u} = \mathbf{h}(\hat{W}^{(\ell)})$ is

$$\mathbf{s}^{(\ell)} = \left[I(\|(\hat{\mathbf{w}}^{(\ell)})^1\|_1 < \theta), \cdots, I(\|(\hat{\mathbf{w}}^{(\ell)})^d\|_1 < \theta) \right]^T, \tag{8}$$

which is used in Step 4 of Algorithm 1. Since both λ and $\mathbf{h}(\hat{W}^{(\ell)})$ are constant with respect to W, we have

$$\begin{aligned}\hat{W}^{(\ell+1)} &= \arg\min_{W} \left\{ l(W) + \lambda g(\mathbf{h}(\hat{W}^{(\ell)})) + \lambda \left\langle \mathbf{s}^{(\ell)}, \mathbf{h}(W) - \mathbf{h}(\hat{W}^{(\ell)}) \right\rangle \right\} \\ &= \arg\min_{W} \left\{ l(W) + \lambda (\mathbf{s}^{(\ell)})^T \mathbf{h}(W) \right\},\end{aligned}$$

which, as shown in Step 3 of Algorithm 1, obtains the next iterative solution by minimizing the upper bound of the objective function in Equation (6). Thus, in the viewpoint of the locally linear approximation, we can understand Algorithm 1 as follows: The original formulation in Equation (6) is non-convex and is difficult to solve; the proposed algorithm minimizes an upper bound in each step, which is convex and can be solved efficiently. It is closely related to the Concave Convex Procedure (CCCP) (Yuille and Rangarajan, 2003). In addition, we can easily verify that the objective function value decreases monotonically as follows:

$$\begin{aligned}l(\hat{W}^{(\ell+1)}) + \lambda g(\mathbf{h}(\hat{W}^{(\ell+1)})) &\le l(\hat{W}^{(\ell+1)}) + \lambda g(\mathbf{h}(\hat{W}^{(\ell)})) + \lambda \left\langle \mathbf{s}^{(\ell)}, \mathbf{h}(\hat{W}^{(\ell+1)}) - \mathbf{h}(\hat{W}^{(\ell)}) \right\rangle \\ &\le l(\hat{W}^{(\ell)}) + \lambda g(\mathbf{h}(\hat{W}^{(\ell)})) + \lambda \left\langle \mathbf{s}^{(\ell)}, \mathbf{h}(\hat{W}^{(\ell)}) - \mathbf{h}(\hat{W}^{(\ell)}) \right\rangle \\ &= l(\hat{W}^{(\ell)}) + \lambda g(\mathbf{h}(\hat{W}^{(\ell)})),\end{aligned}$$

where the first inequality is due to Equation (7) and the second inequality follows from the fact that $\hat{W}^{(\ell+1)}$ is a minimizer of the right hand side of Equation (7).

An important issue we should mention is that a monotonic decrease of the objective function value does not guarantee the convergence of the algorithm, even if the objective function is strictly convex and continuously differentiable (see an example in the book (Bertsekas, 1999, Fig 1.2.6)). In Section 3.1, we will formally discuss the convergence issue.

2.3.2 BLOCK COORDINATE DESCENT

Recall that $g(\mathbf{u})$ is a concave function. We can define its conjugate function as (Rockafellar, 1970):

$$g^\star(\mathbf{v}) = \inf_{\mathbf{u}} \{\mathbf{v}^T \mathbf{u} - g(\mathbf{u})\}.$$

Since $g(\mathbf{u})$ is also a closed function (i.e., the epigraph of $g(\mathbf{u})$ is convex), the conjugate function of $g^\star(\mathbf{v})$ is the original function $g(\mathbf{u})$ (Bertsekas, 1999, Chap. 5.4), that is:

$$g(\mathbf{u}) = \inf_{\mathbf{v}} \{\mathbf{u}^T \mathbf{v} - g^\star(\mathbf{v})\}. \tag{9}$$

Substituting Equation (9) with $\mathbf{u} = \mathbf{h}(W)$ into Equation (6), we can reformulate Equation (6) as:

$$\min_{W,\mathbf{v}} \left\{ f(W,\mathbf{v}) = l(W) + \lambda \mathbf{v}^T \mathbf{h}(W) - \lambda g^\star(\mathbf{v}) \right\} \tag{10}$$

A straightforward algorithm for optimizing Equation (10) is the block coordinate descent (Grippo and Sciandrone, 2000; Tseng, 2001) summarized below:

- Fix $W = \hat{W}^{(\ell)}$:

$$\hat{\mathbf{v}}^{(\ell)} = \arg\min_{\mathbf{v}} \left\{ l(\hat{W}^{(\ell)}) + \lambda \mathbf{v}^T \mathbf{h}(\hat{W}^{(\ell)}) - \lambda g^\star(\mathbf{v}) \right\}$$
$$= \arg\min_{\mathbf{v}} \left\{ \mathbf{v}^T \mathbf{h}(\hat{W}^{(\ell)}) - g^\star(\mathbf{v}) \right\}. \quad (11)$$

  Based on Equation (9) and the Danskin's Theorem (Bertsekas, 1999, Proposition B.25), one solution of Equation (11) is given by a sub-gradient of $g(\mathbf{u})$ at $\mathbf{u} = \mathbf{h}(\hat{W}^{(\ell)})$. That is, we can choose $\hat{\mathbf{v}}^{(\ell)} = \mathbf{s}^{(\ell)}$ given in Equation (8). Apparently, Equation (11) is equivalent to Step 4 in Algorithm 1.

- Fix $\mathbf{v} = \hat{\mathbf{v}}^{(\ell)} = \left[I(\|(\hat{\mathbf{w}}^{(\ell)})^1\|_1 < \theta), \cdots, I(\|(\hat{\mathbf{w}}^{(\ell)})^d\|_1 < \theta) \right]^T$:

$$\hat{W}^{(\ell+1)} = \arg\min_{W} \left\{ l(W) + \lambda (\hat{\mathbf{v}}^{(\ell)})^T \mathbf{h}(W) - \lambda g^\star(\hat{\mathbf{v}}^{(\ell)}) \right\}$$
$$= \arg\min_{W} \left\{ l(W) + \lambda (\hat{\mathbf{v}}^{(\ell)})^T \mathbf{h}(W) \right\}, \quad (12)$$

  which corresponds to Step 3 of Algorithm 1.

The block coordinate descent procedure is intuitive, however, it is non-trivial to analyze its convergence behavior. We will present the convergence analysis in Section 3.1.

2.3.3 DISCUSSIONS

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 multi-stage refinement of that of Lasso. Basically, the MSMTFL algorithm solves a sequence of weighted Lasso problems, where the weights λ_j's are set as the product of the parameter λ in Equation (1) and a $\{0,1\}$-valued indicator function. Specifically, a penalty is imposed in the current stage if the ℓ_1-norm of some row of W in the last stage is smaller than the threshold θ; otherwise, no penalty is imposed. In other words, MSMTFL in the current stage tends to shrink the small rows of W and keep the large rows of W in the last stage. However, Lasso (corresponds to $\ell = 1$) penalizes all rows of W in the same way. It may incorrectly keep the irrelevant rows (which should have been zero rows) or shrink the relevant rows (which should have been large rows) to be zero vectors. MSMTFL overcomes this limitation by adaptively penalizing the rows of W according to the solution generated in the last stage. One important question is whether the MSMTFL algorithm can improve the performance during the multi-stage iteration. In Section 4, we will theoretically show that the MSMTFL algorithm indeed achieves the stagewise improvement in terms of the parameter estimation error bound. That is, the error bound in the current stage improves the one in the last stage. Empirical studies in Section 6 also validate the presented theoretical analysis.

3. Convergence and Reproducibility Analysis

In this section, we first present the convergence analysis. Then, we discuss the reproducibility issue for the MSMTFL algorithm.

3.1 Convergence Analysis

The main convergence result is summarized in the following theorem, which is based on the block coordinate descent interpretation.

Theorem 1 *Let $(W^\star, \mathbf{v}^\star)$ be a limit point of the sequence $\{\hat{W}^{(\ell)}, \hat{\mathbf{v}}^{(\ell)}\}$ generated by the block coordinate descent algorithm. Then $W^\star$ is a critical point of Equation (1).*

Proof Based on Equation (11) and Equation (12), we have

$$\begin{aligned} f(\hat{W}^{(\ell)}, \hat{\mathbf{v}}^{(\ell)}) &\le f(\hat{W}^{(\ell)}, \mathbf{v}), \ \forall \mathbf{v} \in \mathbb{R}^d, \\ f(\hat{W}^{(\ell+1)}, \hat{\mathbf{v}}^{(\ell)}) &\le f(W, \hat{\mathbf{v}}^{(\ell)}), \ \forall W \in \mathbb{R}^{d\times m}. \end{aligned} \tag{13}$$

It follows that

$$f(\hat{W}^{(\ell+1)}, \hat{\mathbf{v}}^{(\ell+1)}) \le f(\hat{W}^{(\ell+1)}, \hat{\mathbf{v}}^{(\ell)}) \le f(\hat{W}^{(\ell)}, \hat{\mathbf{v}}^{(\ell)}),$$

which indicates that the sequence $\{f(\hat{W}^{(\ell)}, \hat{\mathbf{v}}^{(\ell)})\}$ is monotonically decreasing. Since $(W^\star, \mathbf{v}^\star)$ is a limit point of $\{\hat{W}^{(\ell)}, \hat{\mathbf{v}}^{(\ell)}\}$, there exists a subsequence $\mathcal{K}$ such that

$$\lim_{\ell\in\mathcal{K}\to\infty} (\hat{W}^{(\ell)}, \hat{\mathbf{v}}^{(\ell)}) = (W^\star, \mathbf{v}^\star).$$

We observe that

$$\begin{aligned} f(W, \mathbf{v}) &= l(W) + \lambda \mathbf{v}^T \mathbf{h}(W) - \lambda g^\star(\mathbf{v}) \\ &\ge l(W) + \lambda g(\mathbf{h}(W)) \ge 0, \end{aligned}$$

where the first inequality above is due to Equation (9). Thus, $\{f(\hat{W}^{(\ell)}, \hat{\mathbf{v}}^{(\ell)})\}_{\ell\in\mathcal{K}}$ is bounded below. Together with the fact that $\{f(\hat{W}^{(\ell)}, \hat{\mathbf{v}}^{(\ell)})\}$ is decreasing, $\lim_{\ell\to\infty} f(\hat{W}^{(\ell)}, \hat{\mathbf{v}}^{(\ell)}) > -\infty$ exists. Since $f(W, \mathbf{v})$ is continuous, we have

$$\lim_{\ell\to\infty} f(\hat{W}^{(\ell)}, \hat{\mathbf{v}}^{(\ell)}) = \lim_{\ell\in\mathcal{K}\to\infty} f(\hat{W}^{(\ell)}, \hat{\mathbf{v}}^{(\ell)}) = f(W^\star, \mathbf{v}^\star).$$

Taking limits on both sides of Equation (13) with $\ell \in \mathcal{K} \to \infty$, we have

$$f(W^\star, \mathbf{v}^\star) \le f(W, \mathbf{v}^\star), \ \forall W \in \mathbb{R}^{d\times m},$$

which implies

$$\begin{aligned} W^\star &\in \arg\min_W f(W, \mathbf{v}^\star) \\ &= \arg\min_W \left\{ l(W) + \lambda (\mathbf{v}^\star)^T \mathbf{h}(W) - \lambda g^\star(\mathbf{v}^\star) \right\} \\ &= \arg\min_W \left\{ l(W) + \lambda (\mathbf{v}^\star)^T \mathbf{h}(W) \right\}. \end{aligned} \tag{14}$$

Therefore, the zero matrix O must be a sub-gradient of the objective function in Equation (14) at $W = W^\star$:

$$O \in \partial l(W^\star) + \lambda \partial \left((\mathbf{v}^\star)^T \mathbf{h}(W^\star) \right) = \partial l(W^\star) + \lambda \sum_{j=1}^d v_j^\star \partial \left(\|(\mathbf{w}^\star)^j\|_1 \right), \tag{15}$$

where $\partial l(W^\star)$ denotes the sub-differential (which is a set composed of all sub-gradients) of $l(W)$ at $W = W^\star$. We observe that

$$\hat{\mathbf{v}}^{(\ell)} \in \partial g(\mathbf{u})|_{\mathbf{u}=\mathbf{h}(\hat{W}^{(\ell)})},$$

which implies that $\forall \mathbf{x} \in \mathbb{R}^d_+$:

$$g(\mathbf{x}) \le g(\mathbf{h}(\hat{W}^{(\ell)})) + \left\langle \hat{\mathbf{v}}^{(\ell)}, \mathbf{x} - \mathbf{h}(\hat{W}^{(\ell)}) \right\rangle.$$

Taking limits on both sides of the above inequality with $\ell \in \mathcal{K} \to \infty$, we have:

$$g(\mathbf{x}) \le g(\mathbf{h}(W^\star)) + \langle \mathbf{v}^\star, \mathbf{x} - \mathbf{h}(W^\star) \rangle,$$

which implies that $\mathbf{v}^\star$ is a sub-gradient of $g(\mathbf{u})$ at $\mathbf{u} = \mathbf{h}(W^\star)$, that is:

$$\mathbf{v}^\star \in \partial g(\mathbf{u})|_{\mathbf{u}=\mathbf{h}(W^\star)}. \tag{16}$$

Note that the objective function of Equation (1) can be written as a difference of two convex functions:

$$l(W) + \lambda \sum_{j=1}^{d} \min\left(\|\mathbf{w}^j\|_1, \theta\right) = (l(W) + \lambda\|W\|_{1,1}) - \lambda \sum_{j=1}^{d} \max\left(\|\mathbf{w}^j\|_1 - \theta, 0\right).$$

Based on Wright et al. (2009); Toland (1979), we know that $W^\star$ is a critical point of Equation (1) if the following holds:

$$O \in (\nabla l(W^\star) + \lambda\partial\|W^\star\|_{1,1}) - \lambda \sum_{j=1}^{d} \partial \max\left(\|(\mathbf{w}^\star)^j\|_1 - \theta, 0\right). \tag{17}$$

Substituting Equation (16) into Equation (15), we can obtain Equation (17). Therefore, $W^\star$ is a critical point of Equation (1). This completes the proof of Theorem 1. ∎

Due to the equivalence between Algorithm 1 and the block coordinate descent algorithm above, Theorem 1 indicates that any limit point of the sequence $\{\hat{W}^{(\ell)}\}$ generated by Algorithm 1 is a critical point of Equation (1). The remaining issue is to analyze the performance of the critical point. In the sequel, we will conduct analysis in two aspects: reproducibility and the parameter estimation performance.

3.2 Reproducibility of The Algorithm

In general, it is difficult to analyze the performance of a non-convex formulation, as different solutions can be obtained due to different initializations. One natural question is whether the solution generated by Algorithm 1 (based on the initialization of $\lambda_j^{(0)} = \lambda$ in Step 1) is reproducible. In other words, is the solution of Algorithm 1 unique? If we can guarantee that, for any $\ell \ge 1$, the solution $\hat{W}^{(\ell)}$ of Equation (5) is unique, then the solution generated by Algorithm 1 is unique. That is, the solution is reproducible. The main result is summarized in the following theorem:

Theorem 2 *If $X_i \in \mathbb{R}^{n_i \times d}$ $(i \in \mathbb{N}_m)$ has entries drawn from a continuous probability distribution on $R^{n_i d}$, then, for any $\ell \geq 1$, the optimization problem in Equation (5) has a unique solution with probability one.*

Proof Equation (5) can be decomposed into m independent smaller minimization problems:

$$\hat{\mathbf{w}}_i^{(\ell)} = \arg\min_{\mathbf{w}_i \in \mathbb{R}^d} \frac{1}{mn_i} \|X_i \mathbf{w}_i - \mathbf{y}_i\|^2 + \sum_{j=1}^{d} \lambda_j^{(\ell-1)} |w_{ji}|.$$

Next, we only need to prove that the solution of the above optimization problem is unique. To simplify the notations, we unclutter the above equation (by ignoring some superscripts and subscripts) as follows:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{mn} \|X\mathbf{w} - \mathbf{y}\|^2 + \sum_{j=1}^{d} \lambda_j |w_j|, \tag{18}$$

The first order optimal condition is $\forall j \in \mathbb{N}_d$:

$$\frac{2}{mn} \mathbf{x}_j^T (\mathbf{y} - X\hat{\mathbf{w}}) = \lambda_j \mathrm{sign}(\hat{w}_j),$$

where $\mathrm{sign}(\hat{w}_j) = 1$, if $\hat{w}_j > 0$; $\mathrm{sign}(\hat{w}_j) = -1$, if $\hat{w}_j < 0$; and $\mathrm{sign}(\hat{w}_j) \in [-1, 1]$, otherwise. We define

$$\mathcal{E} = \left\{ j \in \mathbb{N}_d : \frac{2}{mn} |\mathbf{x}_j^T (\mathbf{y} - X\hat{\mathbf{w}})| = \lambda_j \right\},$$
$$\mathbf{s} = \mathrm{sign}\left(\frac{2}{mn} X_{\mathcal{E}}^T (\mathbf{y} - X\hat{\mathbf{w}}) \right),$$

where $X_{\mathcal{E}}$ denotes the matrix composed of the columns of X indexed by $\mathcal{E}$. Then, the optimal solution $\hat{\mathbf{w}}$ of Equation (18) satisfies

$$\hat{\mathbf{w}}_{\mathbb{N}_d \setminus \mathcal{E}} = \mathbf{0},$$
$$\hat{\mathbf{w}}_{\mathcal{E}} = \arg\min_{\mathbf{w}_{\mathcal{E}} \in \mathbb{R}^{|\mathcal{E}|}} \frac{1}{mn} \|X_{\mathcal{E}} \mathbf{w}_{\mathcal{E}} - \mathbf{y}\|^2 + \sum_{j \in \mathcal{E}} \lambda_j |w_j|, \tag{19}$$

where $\mathbf{w}_{\mathcal{E}}$ denotes the vector composed of entries of $\mathbf{w}$ indexed by $\mathcal{E}$. Since $X \in \mathbb{R}^{n_i \times d}$ is drawn from the continuous probability distribution, X has columns in general positions with probability one and hence $\mathrm{rank}(X_{\mathcal{E}}) = |\mathcal{E}|$ (or equivalently $\mathrm{Null}(X_{\mathcal{E}}) = \{\mathbf{0}\}$), due to Lemma 3, Lemma 4 and their discussions in Tibshirani (2013). Therefore, the objective function in Equation (19) is strictly convex. Noticing that $X\hat{\mathbf{w}}$ is unique (Tibshirani, 2013), thus $\mathcal{E}$ is unique. This implies that $\hat{\mathbf{w}}_{\mathcal{E}}$ is unique. Thus, the optimal solution $\hat{\mathbf{w}}$ of Equation (18) is also unique and so is the optimization problem in Equation (5) for any $\ell \geq 1$. This completes the proof of Theorem 2. ∎

Theorem 2 is important in the sense that it makes the theoretical analysis for the parameter estimation performance of Algorithm 1 possible. Although the solution may not be globally optimal, we show in the next section that the solution has good performance in terms of the parameter estimation error bound.

Remark 3 *Zhang (2010, 2012) study the capped-ℓ_1 regularized formulation for the single task setting and propose the multi-stage algorithm for such formulation. However, Zhang (2010, 2012) neither provide detailed convergence analysis nor discuss the reproducibility issues. The presented analysis is applicable to the multi-stage algorithm proposed in Zhang (2010, 2012), as it is a special case of the proposed algorithm with $m = 1$. To our best knowledge, this is the first work that discusses the reproducibility issue for multi-stage optimization algorithms.*

4. Parameter Estimation Error Bound

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 learning literature (Zhang and Zhang, 2012; Zhang, 2008, 2009, 2010, 2012).

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 $\forall j \in \mathbb{N}_n, i \in \mathbb{N}_m, t \in \mathbb{R}$:*

$$\mathbb{E}_{\delta_{ji}} \exp(t\delta_{ji}) \leq \exp\left(\frac{\sigma^2 t^2}{2}\right).$$

Remark 4 *We call the random variable satisfying the condition in Assumption 1 sub-Gaussian, since its moment generating function is bounded by that of a zero mean Gaussian random variable. That is, if a normal random variable $x \sim N(0, \sigma^2)$, then we have:*

$$\begin{aligned}\mathbb{E}\exp(tx) &= \int_{-\infty}^{\infty} \exp(tx) \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx \\ &= \exp(\sigma^2 t^2/2) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\sigma^2 t)^2}{2\sigma^2}\right) dx \\ &= \exp(\sigma^2 t^2/2).\end{aligned}$$

Remark 5 *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-Gaussians. Thus, the sub-Gaussian noise assumption is more general than the Gaussian noise assumption which is commonly used in the multi-task learning literature (Jalali et al., 2010; Lounici et al., 2009).*

We next introduce the following sparse eigenvalue concept which is also common in the analysis of sparse learning literature (Zhang and Huang, 2008; Zhang and Zhang, 2012; Zhang, 2009, 2010, 2012).

Definition 6 *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 7 $\rho_i^+(k)$ $(\rho_i^-(k))$ *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 8 *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*

$$\forall (j,i) \in \bar{\mathcal{F}}, \|\bar{\mathbf{w}}^j\|_1 \geq 2\theta \tag{20}$$

$$\text{and } \frac{\rho_i^+(s)}{\rho_i^-(2\bar{r}+2s)} \leq 1 + \frac{s}{2\bar{r}}, \tag{21}$$

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

$$\lambda \geq 12\sigma\sqrt{\frac{2\rho_{max}^+(1)\ln(2dm/\eta)}{n}}, \tag{22}$$

$$\theta \geq \frac{11m\lambda}{\rho_{min}^-(2\bar{r}+s)}, \tag{23}$$

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

$$\|\hat{W}^{(\ell)} - \bar{W}\|_{2,1} \leq 0.8^{\ell/2}\frac{9.1m\lambda\sqrt{\bar{r}}}{\rho_{min}^-(2\bar{r}+s)} + \frac{39.5m\sigma\sqrt{\rho_{max}^+(\bar{r})(7.4\bar{r}+2.7\ln(2/\eta))/n}}{\rho_{min}^-(2\bar{r}+s)}, \tag{24}$$

where $\hat{W}^{(\ell)}$ *is a solution of Equation (5).*

Remark 9 *Equation (20) 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. Equation (21) is called the sparse eigenvalue condition (Zhang, 2012), 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 (Zhang and Huang, 2008; Zhang, 2009) and it is slightly weaker than the RIP condition (Candes and Tao, 2005; Huang and Zhang, 2010; Zhang, 2012).*

Remark 10 *When* $\ell = 1$ *(corresponds to Lasso), the first term of the right-hand side of Equation (24) dominates the error bound in the order of*

$$\|\hat{W}^{Lasso} - \bar{W}\|_{2,1} = O\left(m\sqrt{\bar{r}\ln(dm/\eta)/n}\right), \tag{25}$$

since λ *satisfies the condition in Equation (22). Note that the first term of the right-hand side of Equation (24) 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:*

$$\|\hat{W}^{(\ell)} - \bar{W}\|_{2,1} = O\left(m\sqrt{\bar{r}/n + \ln(1/\eta)/n}\right). \tag{26}$$

Jalali et al. (2010) 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) is in the same order with Equation (25), 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 Equation (26) is significantly better than the ones for the Lasso and Dirty model.

Remark 11 *Jalali et al. (2010) 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 Equation (21). From the viewpoint of the parameter estimation error, our proposed algorithm can achieve a better bound under weaker conditions. Please refer to (Van De Geer and Bühlmann, 2009; Zhang, 2009, 2012) for more details about the incoherence condition, the RIP condition, the sparse eigenvalue condition and their relationships.*

Remark 12 *The capped-ℓ_1 regularized formulation in Zhang (2010) 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 (Zhang, 2010, 2012), 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.*

In addition, we want to emphasize that the support recovery analysis in Zhang (2012) can not be easily adapted to the proposed capped-ℓ_1,ℓ_1 multi-task feature learning setting. The key difficulty is that, in order to achieve a similar support recovery result for the formulation in Equation (1), we need to assume that each row of the underlying sparse weight matrix $\bar{W}$ is either a zero vector or a vector composed of all nonzero entries. However, this is not the case in the proposed multi-task formulation. Although this assumption holds for the capped-ℓ_1,ℓ_2 multi-task feature learning problem in Equation (4), each subproblem involved for solving Equation (4) is a reweighed ℓ_2 regularized problem and its first-order optimality condition is quite different from that of the reweighed ℓ_1 regularized problem. Thus, it is also challenging to extend the analysis in Zhang (2012) to the capped-ℓ_1,ℓ_2 multi-task feature learning setting.

5. Proof Sketch of Theorem 8

In this section, we present a proof sketch of Theorem 8. We first provide several important lemmas (detailed proofs are available in the Appendix A) and then complete the proof of Theorem 8 based on these lemmas.

Lemma 13 *Let $\bar{\Upsilon} = [\bar{\epsilon}_1, \cdots, \bar{\epsilon}_m]$ with $\bar{\epsilon}_i = [\bar{\varepsilon}_{1i}, \cdots, \bar{\varepsilon}_{di}]^T = \frac{1}{n}X_i^T(X_i\bar{\mathbf{w}}_i - \mathbf{y}_i)$ ($i \in \mathbb{N}_m$). 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 8, the followings hold with probability larger than $1-\eta$*:*

$$\|\bar{\Upsilon}\|_{\infty,\infty} \leq \sigma\sqrt{\frac{2\rho^{+}_{max}(1)\ln(2dm/\eta)}{n}}, \tag{27}$$

$$\|\bar{\Upsilon}_{\bar{\mathcal{H}}}\|_F^2 \leq m\sigma^2\rho^{+}_{max}(\bar{r})(7.4\bar{r}+2.7\ln(2/\eta))/n. \tag{28}$$

Lemma 13 gives bounds on the residual correlation ($\bar{\Upsilon}$) with respect to $\bar{W}$. We note that Equation (27) and Equation (28) are closely related to the assumption on λ in Equation (22) and the second term of the right-hand side of Equation (24) (error bound), respectively. This lemma provides a fundamental basis for the proof of Theorem 8.

Lemma 14 *Use the notations of Lemma 13 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 Equation (5) 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* $2\|\bar{\epsilon}_i\|_\infty < \hat{\lambda}_{\mathcal{G}_i}$*, 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 $\bar{\mathcal{F}} \cap \mathcal{G} = \emptyset \Rightarrow \Delta\hat{W}_{\mathcal{G}}^{(\ell)} = \hat{W}_{\mathcal{G}}^{(\ell)}$. Lemma 14 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 15 *Using the notations of Lemma 14, 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 13 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* $\hat{\mathbf{w}}_{\mathcal{G}_i}$*,* $I_i = \mathcal{G}_i^c \cup \mathcal{J}_i$*,* $I = \cup_{i\in\mathbb{N}_m}I_i$ *and* $\bar{\mathcal{F}} = \cup_{i\in\mathbb{N}_m}\bar{\mathcal{F}}_i$*. Then, the following inequalities hold at any stage* $\ell \geq 1$*:*

$$\|\Delta\hat{W}^{(\ell)}\|_{2,1} \leq \frac{\left(1+1.5\sqrt{\frac{2\bar{r}}{s}}\right)\sqrt{8m\left(4\|\bar{\Upsilon}_{\mathcal{G}_{(\ell)}^c}\|_F^2+\sum_{(j,i)\in\bar{\mathcal{F}}}(\hat{\lambda}_{ji}^{(\ell-1)})^2\right)}}{\rho^{-}_{min}(2\bar{r}+s)}, \tag{29}$$

$$\|\Delta\hat{W}^{(\ell)}\|_{2,1} \leq \frac{9.1m\lambda\sqrt{\bar{r}}}{\rho^{-}_{min}(2\bar{r}+s)}. \tag{30}$$

Lemma 15 is established based on Lemma 14, by considering the relationship between Equation (22) and Equation (27), and the specific definition of $\mathcal{G} = \mathcal{G}_{(\ell)}$. Equation (29) provides a parameter estimation error bound in terms of $\ell_{2,1}$-norm by $\|\bar{\Upsilon}_{\mathcal{G}_{(\ell)}^c}\|_F^2$ and the regularization parameters $\hat{\lambda}_{ji}^{(\ell-1)}$ (see the definition of $\hat{\lambda}_{ji}$ ($\hat{\lambda}_{ji}^{(\ell-1)}$) in Lemma 14). This is the result directly used in the proof of Theorem 8. Equation (30) 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 Equation (24).

Lemma 16 *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 13. Then under the condition of Equation (20), 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 16 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 8. This recursive relation is crucial for the shrinkage part of the error bound in Equation (24).

5.1 Proof of Theorem 8

We now complete the proof of Theorem 8 based on the lemmas above.

Proof For notational simplicity, we denote the right-hand side of Equation (28) as:

$$u = m\sigma^2\rho_{max}^+(\bar{r})(7.4\bar{r}+2.7\ln(2/\eta))/n. \tag{31}$$

Based on $\bar{\mathcal{H}}\subseteq\mathcal{G}_{(\ell)}^c$, Lemma 13 and Equation (22), the followings hold with probability larger than $1-\eta$:

$$\begin{aligned}
\|\bar{\Upsilon}_{\mathcal{G}_{(\ell)}^c}\|_F^2 &= \|\bar{\Upsilon}_{\bar{\mathcal{H}}}\|_F^2+\|\bar{\Upsilon}_{\mathcal{G}_{(\ell)}^c\setminus\bar{\mathcal{H}}}\|_F^2\\
&\le u+|\mathcal{G}_{(\ell)}^c\setminus\bar{\mathcal{H}}|\|\bar{\Upsilon}\|_{\infty,\infty}^2\\
&\le u+\lambda^2|\mathcal{G}_{(\ell)}^c\setminus\bar{\mathcal{H}}|/144\\
&\le u+(1/144)m\lambda^2\theta^{-2}\|\hat{W}^{(\ell-1)}_{\mathcal{G}_{(\ell)}^c\setminus\bar{\mathcal{H}}}-\bar{W}_{\mathcal{G}_{(\ell)}^c\setminus\bar{\mathcal{H}}}\|_{2,1}^2,
\end{aligned}\tag{32}$$

where the last inequality follows from

$$\begin{aligned}
&\forall(j,i)\in\mathcal{G}_{(\ell)}^c\setminus\bar{\mathcal{H}}, \|(\hat{\mathbf{w}}^{(\ell-1)})^j\|_1^2/\theta^2=\|(\hat{\mathbf{w}}^{(\ell-1)})^j-\bar{\mathbf{w}}^j\|_1^2/\theta^2\ge 1\\
\Rightarrow &|\mathcal{G}_{(\ell)}^c\setminus\bar{\mathcal{H}}|\le m\theta^{-2}\|\hat{W}^{(\ell-1)}_{\mathcal{G}_{(\ell)}^c\setminus\bar{\mathcal{H}}}-\bar{W}_{\mathcal{G}_{(\ell)}^c\setminus\bar{\mathcal{H}}}\|_{2,1}^2.
\end{aligned}$$

According to Equation (29), we have:

$$\begin{aligned}
\|\hat{W}^{(\ell)}-\bar{W}\|_{2,1}^2 &= \|\Delta\hat{W}^{(\ell)}\|_{2,1}^2\\
&\le \frac{8m\left(1+1.5\sqrt{\frac{2\bar{r}}{s}}\right)^2\left(4\|\bar{\Upsilon}_{\mathcal{G}_{(\ell)}^c}\|_F^2+\sum_{(j,i)\in\bar{\mathcal{F}}}(\hat{\lambda}_{ji}^{(\ell-1)})^2\right)}{(\rho_{min}^-(2\bar{r}+s))^2}\\
&\le \frac{78m\left(4u+(37/36)m\lambda^2\theta^{-2}\left\|\hat{W}^{(\ell-1)}-\bar{W}\right\|_{2,1}^2\right)}{(\rho_{min}^-(2\bar{r}+s))^2}\\
&\le \frac{312mu}{(\rho_{min}^-(2\bar{r}+s))^2}+0.8\left\|\hat{W}^{(\ell-1)}-\bar{W}\right\|_{2,1}^2\\
&\le\cdots\le 0.8^\ell\left\|\hat{W}^{(0)}-\bar{W}\right\|_{2,1}^2+\frac{312mu}{(\rho_{min}^-(2\bar{r}+s))^2}\frac{1-0.8^\ell}{1-0.8}\\
&\le 0.8^\ell\frac{9.1^2m^2\lambda^2\bar{r}}{(\rho_{min}^-(2\bar{r}+s))^2}+\frac{1560mu}{(\rho_{min}^-(2\bar{r}+s))^2}.
\end{aligned}$$

In the above derivation, the first inequality is due to Equation (29); the second inequality is due to the assumption $s\ge\bar{r}$ in Theorem 8, Equation (32) and Lemma 16; the third inequality is due to Equation (23); the last inequality follows from Equation (30) 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 Equation (31) into the above inequality, we verify Theorem 8. ∎

Remark 17 *The assumption $s \ge \bar{r}$ used in the above proof indicates that at each stage, the zero entries of $\hat{W}^{(\ell)}$ should be greater than $m\bar{r}$ (see definition of s in Lemma 15). This requires the solution obtained by Algorithm 1 at each stage is sparse, which is consistent with the sparsity of $\bar{W}$ in Assumption 1.*

6. Experiments

In this section, we present empirical studies on both synthetic and real-world data sets. In the synthetic data experiments, we present the performance of the MSMTFL algorithm in terms of the parameter estimation error. In the real-world data experiments, we show the performance of the MSMTFL algorithm in terms of the prediction error.

6.1 Competing Algorithms

We present the empirical studies by comparing the following six algorithms:

- Lasso: ℓ_1-norm regularized feature learning algorithm with $\lambda\|W\|_{1,1}$ as the regularizer
- L1,2: $\ell_{1,2}$-norm regularized multi-task feature learning algorithm with $\lambda\|W\|_{1,2}$ as the regularizer (Obozinski et al., 2006)
- DirtyMTL: dirty model multi-task feature learning algorithm with $\lambda_s\|P\|_{1,1}+\lambda_b\|Q\|_{1,\infty}$ ($W = P+Q$) as the regularizer (Jalali et al., 2010)
- CapL1,L1: our proposed multi-task feature learning algorithm with $\lambda\sum_{j=1}^d \min(\|\mathbf{w}^j\|_1,\theta)$ as the regularizer
- CapL1: capped-ℓ_1 regularized feature learning algorithm with $\lambda\sum_{j=1}^d\sum_{i=1}^m \min(|w_{ji}|,\theta)$ as the regularizer
- CapL1,L2: capped-ℓ_1,ℓ_2 regularized multi-task feature learning algorithm with $\lambda\sum_{j=1}^d \min(\|\mathbf{w}^j\|,\theta)$ as the regularizer

In the experiments, we employ the quadratic loss function in Equation (2) for all the compared algorithms. We use MSMTFL-type algorithms (similar to Algorithm 1) to sovle capped-ℓ_1 and capped-ℓ_1,ℓ_2 regularized feature learning problems (details are provided in Appendix C).

6.2 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 8. 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 six algorithms in different parameter settings (Figure 2 and Figure 3). For a fair comparison, we compare the smallest estimation errors of the six algorithms in all the parameter settings as done in (Zhang, 2009, 2010). We observe that the parameter estimation errors of the capped-ℓ_1,ℓ_1, capped-ℓ_1 and capped-ℓ_1,ℓ_2 regularized feature learning formulations solved by MSMTFL-type algorithms are the smallest among all algorithms. In most cases, CapL1,L1 achieves a slightly smaller error than CapL1 and CapL1,L2. This empirical result demonstrates the effectiveness of the MSMTFL algorithm. We also have the following observations: (a) When λ is large enough, all six algorithms tend to have the same parameter estimation error. This is reasonable, because the solutions $\hat{W}$'s obtained by the six 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.

6.3 Real-World Data Experiments

We conduct experiments on two real-world data sets: MRI and Isolet data sets.

The MRI data set is collected from the ANDI database, which contains 675 patients' MRI data preprocessed using FreeSurfer.[2]. 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.

The Isolet data set[3] 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 one subset. The 5 tasks respectively have 1560, 1560, 1560, 1558, and 1559 samples,[4] 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

2. FreeSurfer can be found at `www.loni.ucla.edu/ADNI/`.
3. The data set can be found at `www.zjucadcg.cn/dengcai/Data/data.html`.
4. Three samples are historically missing.

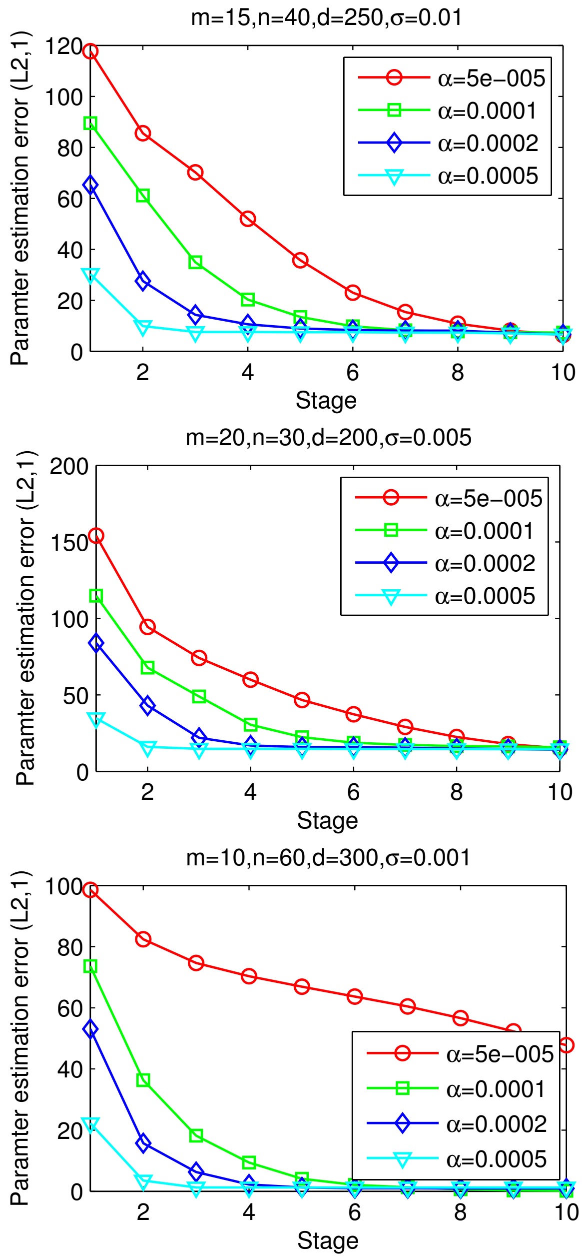

Figure 1: Averaged parameter estimation error $\|\hat{W}-\bar{W}\|_{2,1}$ vs. Stage (ℓ) plots for MSMTFL on the synthetic data set (averaged over 10 runs). Here we set $\lambda = \alpha\sqrt{\ln(dm)/n}$, $\theta = 50m\lambda$. Note that $\ell = 1$ corresponds to Lasso; the results show the stage-wise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

form the test set. We evaluate the six algorithms in terms of the normalized mean squared error (nMSE) and the averaged means squared error (aMSE), which are commonly used in multi-task learning problems (Zhang and Yeung, 2010; Zhou et al., 2011; Gong et al., 2012). For each training

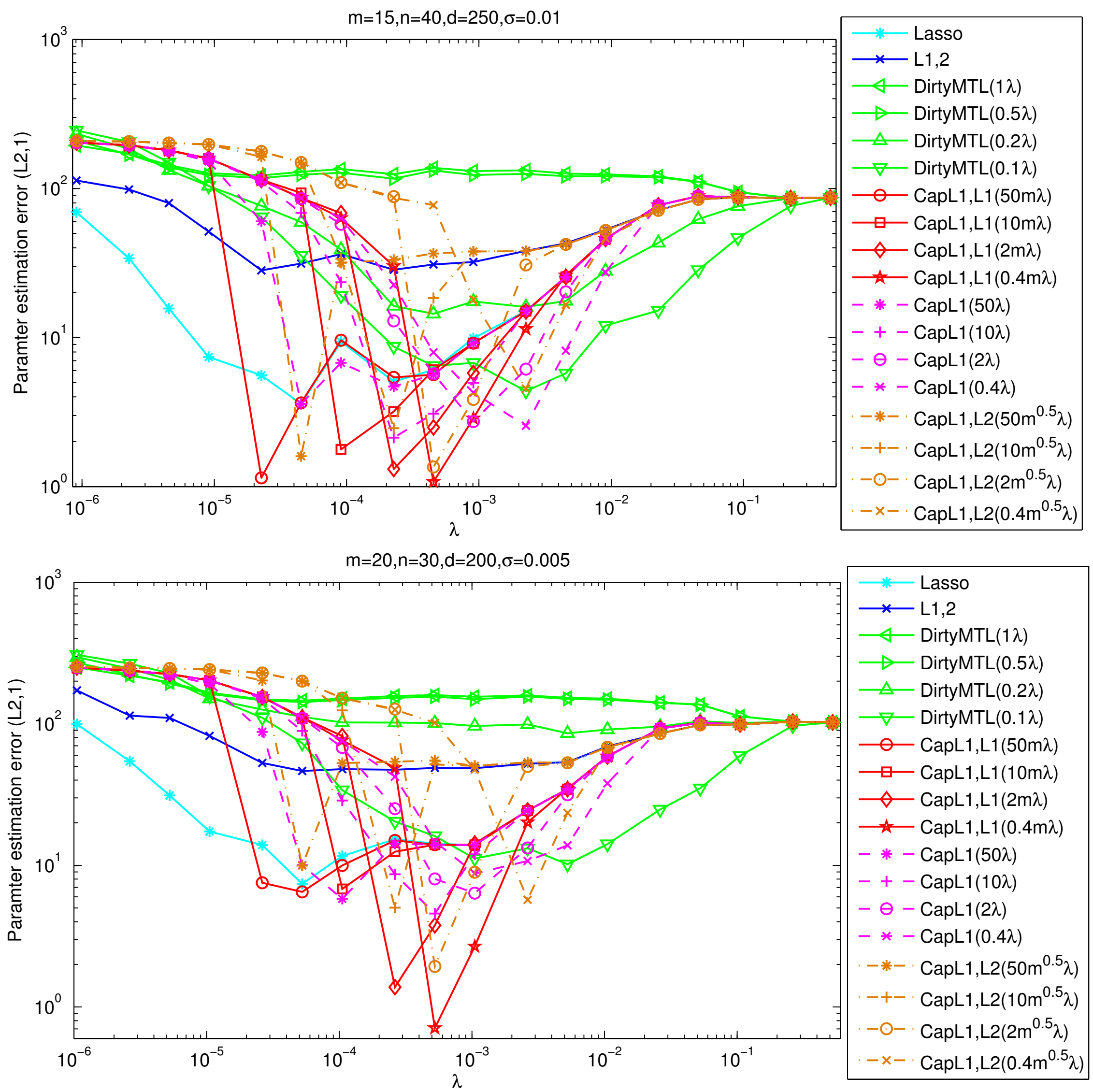

Figure 2: Averaged parameter estimation error $\|\hat{W}-\bar{W}\|_{2,1}$ vs. λ plots on the synthetic data set (averaged over 10 runs). DirtyMTL, CapL1,L1, CapL1, CapL1,L2 have two parameters; we set $\lambda_s/\lambda_b = 1, 0.5, 0.2, 0.1$ for DirtyMTL ($1/m \le \lambda_s/\lambda_b \le 1$ was adopted in Jalali et al. (2010)), $\theta/\lambda = 50m, 10m, 2m, 0.4m$ for CapL1,L1, $\theta/\lambda = 50, 10, 2, 0.4$ for CapL1 and $\theta/\lambda = 50m^{0.5}, 10m^{0.5}, 2m^{0.5}, 0.4m^{0.5}$ for CapL1,L2 (The settings of θ/λ for CapL1,L1, CapL1 and CapL1,L2 are based on the relationships of $\|\mathbf{w}^j\|_1, |w_{ji}|$ and $\|\mathbf{w}^j\|$, where $\mathbf{w}^j \in \mathbb{R}^{1\times m}$ and w_{ji} are the j-th row and the (j,i)-th entry of W, respectively).

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 six algorithms are tuned via 3-fold cross validation.

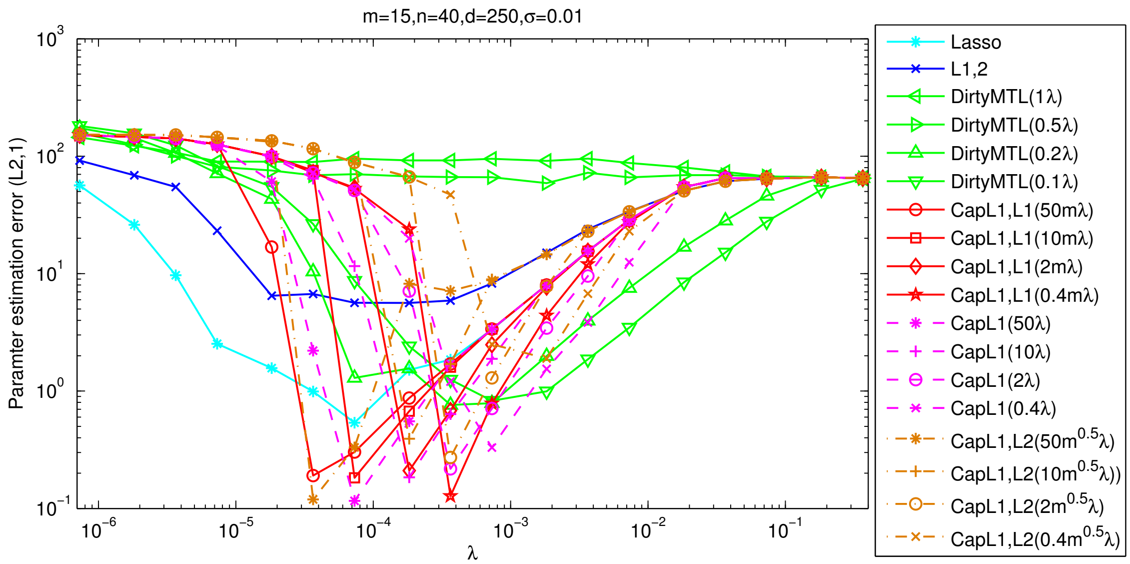

Figure 3: (continued) Averaged parameter estimation error $\|\hat{W} - \bar{W}\|_{2,1}$ vs. λ plots on the synthetic data set (averaged over 10 runs). DirtyMTL, CapL1,L1, CapL1, CapL1,L2 have two parameters; we set $\lambda_s/\lambda_b = 1, 0.5, 0.2, 0.1$ for DirtyMTL ($1/m \leq \lambda_s/\lambda_b \leq 1$ was adopted in Jalali et al. (2010)), $\theta/\lambda = 50m, 10m, 2m, 0.4m$ for CapL1,L1, $\theta/\lambda = 50, 10, 2, 0.4$ for CapL1 and $\theta/\lambda = 50m^{0.5}, 10m^{0.5}, 2m^{0.5}, 0.4m^{0.5}$ for CapL1,L2 (The settings of θ/λ for CapL1,L1, CapL1 and CapL1,L2 are based on the relationships of $\|\mathbf{w}^j\|_1, |w_{ji}|$ and $\|\mathbf{w}^j\|$, where $\mathbf{w}^j \in \mathbb{R}^{1\times m}$ and w_{ji} are the j-th row and the (j,i)-th entry of W, respectively).

Table 1 and Table 2 show the experimental results in terms of the averaged nMSE (aMSE) and the standard deviation. From these results, we observe that CapL1,L1 and CapL1,L2 outperform all the other competing feature learning algorithms on both data sets in terms of the regression errors (nMSE and aMSE). On the MRI data set, CapL1,L1 achieves slightly better performance than CapL1,L2 and on the Isolet data set, CapL1,L2 achieves slightly better performance than CapL1,L1. These empirical results demonstrate the effectiveness of the proposed MSMTFL (-type) algorithms.

7. Conclusions

In this paper, we propose a non-convex formulation for multi-task feature learning, which learns the specific features of each task as well as the common features shared among tasks. The non-convex formulation adopts the capped-ℓ_1,ℓ_1 regularizer to better approximate the ℓ_0-type one than the commonly used convex regularizer. To solve the non-convex optimization problem, we propose a Multi-Stage Multi-Task Feature Learning (MSMTFL) algorithm and provide intuitive interpretations from several aspects. We also present a detailed convergence analysis and discuss the reproducibility issue for the proposed algorithm. Specifically, we show that, under a mild condition, the solution generated by MSMTFL is unique. Although the solution may not be globally optimal, we theoretically show that it has good performance in terms of the parameter estimation error bound. Experimental results on both synthetic and real-world data sets demonstrate the effectiveness of our

measure	traning ratio	Lasso	L1,2	DirtyMTL
nMSE	0.15	0.6577(0.0193)	0.6443(0.0326)	0.6150(0.0160)
	0.20	0.6294(0.0255)	0.6541(0.0182)	0.6110(0.0122)
	0.25	0.6007(0.0120)	0.6407(0.0310)	0.5997(0.0218)
aMSE	0.15	0.0190(0.0008)	0.0184(0.0006)	0.0173(0.0006)
	0.20	0.0178(0.0009)	0.0184(0.0005)	0.0170(0.0007)
	0.25	0.0173(0.0007)	0.0183(0.0004)	0.0169(0.0007)
measure	traning ratio	CapL1,L1	CapL1	CapL1,L2
nMSE	0.15	**0.5551(0.0082)**	0.6448(0.0238)	**0.5591(0.0082)**
	0.20	**0.5539(0.0094)**	0.6245(0.0396)	**0.5612(0.0086)**
	0.25	**0.5513(0.0097)**	0.5899(0.0203)	**0.5595(0.0063)**
aMSE	0.15	**0.0163(0.0007)**	0.0187(0.0009)	**0.0165(0.0007)**
	0.20	**0.0161(0.0006)**	0.0177(0.0010)	**0.0163(0.0006)**
	0.25	**0.0162(0.0007)**	0.0171(0.0009)	**0.0164(0.0007)**

Table 1: Comparison of six feature learning algorithms on the MRI data set in terms of the averaged nMSE and aMSE (standard deviation), which are averaged over 10 random splittings. The two best results are in bold.

measure	traning ratio	Lasso	L1,2	DirtyMTL
nMSE	0.15	0.6798(0.0120)	0.6788(0.0149)	0.6427(0.0172)
	0.2	0.6465(0.0105)	0.6778(0.0104)	0.6371(0.0111)
	0.25	0.6279(0.0099)	0.6666(0.0110)	0.6304(0.0093)
aMSE	0.15	0.1605(0.0028)	0.1602(0.0033)	0.1517(0.0039)
	0.2	0.1522(0.0022)	0.1596(0.0021)	0.1500(0.0023)
	0.25	0.1477(0.0024)	0.1568(0.0025)	0.1482(0.0019)
measure	traning ratio	CapL1,L1	CapL1	CapL1,L2
nMSE	0.15	**0.6421(0.0153)**	0.6541(0.0122)	**0.5819(0.0125)**
	0.2	**0.5847(0.0081)**	0.5962(0.0051)	**0.5589(0.0056)**
	0.25	**0.5496(0.0106)**	0.5569(0.0158)	**0.5422(0.0063)**
aMSE	0.15	**0.1516(0.0035)**	0.1544(0.0028)	**0.1373(0.0030)**
	0.2	**0.1376(0.0020)**	0.1404(0.0012)	**0.1316(0.0014)**
	0.25	**0.1293(0.0028)**	0.1310(0.0042)	**0.1275(0.0013)**

Table 2: Comparison of six feature learning algorithms on the Isolet data set in terms of the averaged nMSE and aMSE (standard deviation), which are averaged over 10 random splittings. The two best results are in bold.

proposed MSMTFL algorithm in comparison with the state of the art multi-task feature learning algorithms.

There are several interesting issues that need to be addressed in the future. First, we will explore the conditions under which a globally optimal solution of the proposed formulation can be obtained by the MSMTFL algorithm. Second, we plan to explore general theoretical bounds for multi-task learning settings (involving different loss functions and non-convex regularization terms) using multi-stage algorithms. Third, we will adapt the GIST algorithm (Gong et al., 2013a,b) to solve the non-convex multi-task feature learning problem and derive theoretical bounds.

Acknowledgments

We would like to thank the action editor and anonymous reviewers for their constructive comments. This work is partly supported by 973 Program (2013CB329503), NSFC (Grant No. 91120301, 61075004 and 61021063), NIH (R01 LM010730) and NSF (IIS-0953662, CCF-1025177).

Appendix A. Proofs of Lemmas 13 to 16

In this appendix, we provide detailed proofs for Lemmas 13 to 16. In our proofs, we use several lemmas (summarized in Appendix B) from Zhang (2010).

We first introduce some notations used in the proof. Define

$$\pi_i(k_i, s_i) = \sup_{\mathbf{v}\in\mathbb{R}^{k_i}, \mathbf{u}\in\mathbb{R}^{s_i}, I_i, \mathcal{J}_i} \frac{\mathbf{v}^T A^{(i)}_{I_i,\mathcal{J}_i}\mathbf{u}\|\mathbf{v}\|}{\mathbf{v}^T A^{(i)}_{I_i,I_i}\mathbf{v}\|\mathbf{u}\|_\infty},$$

where $s_i + k_i \leq d$ with $s_i, k_i \geq 1$; I_i and $\mathcal{J}_i$ are *disjoint* subsets of $\mathbb{N}_d$ with k_i and s_i elements respectively (with some abuse of notation, we also let I_i be a subset of $\mathbb{N}_d \times \{i\}$, depending on the context.); $A^{(i)}_{I_i,\mathcal{J}_i}$ is a sub-matrix of $A_i = n^{-1}X_i^T X_i \in \mathbb{R}^{d\times d}$ with rows indexed by I_i and columns indexed by $\mathcal{J}_i$.

We let $\mathbf{w}_{I_i}$ be a $d \times 1$ vector with the j-th entry being w_{ji}, if $(j,i) \in I_i$, and 0, otherwise. We also let W_I be a $d \times m$ matrix with (j,i)-th entry being w_{ji}, if $(j,i) \in I$, and 0, otherwise.

Proof of Lemma 13 For the j-th entry of $\bar{\epsilon}_i$ $(j \in \mathbb{N}_d)$:

$$|\bar{\varepsilon}_{ji}| = \frac{1}{n}\left|\left(\mathbf{x}^{(i)}_j\right)^T (X_i\bar{\mathbf{w}}_i - \mathbf{y}_i)\right| = \frac{1}{n}\left|\left(\mathbf{x}^{(i)}_j\right)^T \boldsymbol{\delta}_i\right|,$$

where $\mathbf{x}^{(i)}_j$ is the j-th column of X_i. We know that the entries of $\boldsymbol{\delta}_i$ are independent sub-Gaussian random variables, and $\|1/n\mathbf{x}^{(i)}_j\|^2 = \|\mathbf{x}^{(i)}_j\|^2/n^2 \leq \rho^+_i(1)/n$. According to Lemma 18, we have $\forall t > 0$:

$$\Pr(|\bar{\varepsilon}_{ji}| \geq t) \leq 2\exp(-nt^2/(2\sigma^2\rho^+_i(1))) \leq 2\exp(-nt^2/(2\sigma^2\rho^+_{max}(1))).$$

Thus we obtain:

$$\Pr(\|\bar{\Upsilon}\|_{\infty,\infty} \leq t) \geq 1 - 2dm\exp(-nt^2/(2\sigma^2\rho^+_{max}(1))).$$

Let $\eta = 2dm\exp(-nt^2/(2\sigma^2\rho^+_{max}(1)))$ and we can obtain Equation (27). Equation (28) directly follows from Lemma 21 and the following fact:

$$\|\mathbf{x}_i\|^2 \leq ay_i \Rightarrow \|X\|_F^2 = \sum_{i=1}^m \|\mathbf{x}_i\|^2 \leq ma\max_{i\in\mathbb{N}_m} y_i.$$

■

Proof of Lemma 14 The optimality condition of Equation (5) implies that

$$\frac{2}{n}X_i^T(X_i\hat{\mathbf{w}}_i - \mathbf{y}_i) + \hat{\boldsymbol{\lambda}}_i \odot \text{sign}(\hat{\mathbf{w}}_i) = \mathbf{0},$$

where $\odot$ denotes the element-wise product; $\text{sign}(\mathbf{w}) = [\text{sign}(w_1), \cdots, \text{sign}(w_d)]^T$, where $\text{sign}(w_i) = 1$, if $w_i > 0$; $\text{sign}(w_i) = -1$, if $w_i < 0$; and $\text{sign}(w_i) \in [-1, 1]$, otherwise. We note that $X_i\hat{\mathbf{w}}_i - \mathbf{y}_i = X_i\hat{\mathbf{w}}_i - X_i\bar{\mathbf{w}}_i + X_i\bar{\mathbf{w}}_i - \mathbf{y}_i$ and we can rewrite the above equation into the following form:

$$2A_i\Delta\hat{\mathbf{w}}_i = -2\bar{\boldsymbol{\epsilon}}_i - \hat{\boldsymbol{\lambda}}_i \odot \text{sign}(\hat{\mathbf{w}}_i).$$

Thus, for all $\mathbf{v} \in \mathbb{R}^d$, we have

$$2\mathbf{v}^T A_i \Delta\hat{\mathbf{w}}_i = -2\mathbf{v}^T\bar{\boldsymbol{\epsilon}}_i - \sum_{j=1}^{d} \hat{\lambda}_{ji} v_j \text{sign}(\hat{w}_{ji}). \tag{33}$$

Letting $\mathbf{v} = \Delta\hat{\mathbf{w}}_i$ and noticing that $\Delta\hat{w}_{ji} = \hat{w}_{ji}$ for $(j,i) \notin \bar{\mathcal{F}}_i, i \in \mathbb{N}_m$, we obtain

$$\begin{aligned}
0 &\le 2\Delta\hat{\mathbf{w}}_i^T A_i \Delta\hat{\mathbf{w}}_i = -2\Delta\hat{\mathbf{w}}_i^T\bar{\boldsymbol{\epsilon}}_i - \sum_{j=1}^{d} \hat{\lambda}_{ji}\Delta\hat{w}_{ji}\text{sign}(\hat{w}_{ji}) \\
&\le 2\|\Delta\hat{\mathbf{w}}_i\|_1\|\bar{\boldsymbol{\epsilon}}_i\|_\infty - \sum_{(j,i)\in\bar{\mathcal{F}}_i} \hat{\lambda}_{ji}\Delta\hat{w}_{ji}\text{sign}(\hat{w}_{ji}) - \sum_{(j,i)\notin\bar{\mathcal{F}}_i} \hat{\lambda}_{ji}\Delta\hat{w}_{ji}\text{sign}(\hat{w}_{ji}) \\
&\le 2\|\Delta\hat{\mathbf{w}}_i\|_1\|\bar{\boldsymbol{\epsilon}}_i\|_\infty + \sum_{(j,i)\in\bar{\mathcal{F}}_i} \hat{\lambda}_{ji}|\Delta\hat{w}_{ji}| - \sum_{(j,i)\notin\bar{\mathcal{F}}_i} \hat{\lambda}_{ji}|\hat{w}_{ji}| \\
&\le 2\|\Delta\hat{\mathbf{w}}_i\|_1\|\bar{\boldsymbol{\epsilon}}_i\|_\infty + \sum_{(j,i)\in\bar{\mathcal{F}}_i} \hat{\lambda}_{ji}|\Delta\hat{w}_{ji}| - \sum_{(j,i)\in\mathcal{G}_i} \hat{\lambda}_{ji}|\hat{w}_{ji}| \\
&\le 2\|\Delta\hat{\mathbf{w}}_i\|_1\|\bar{\boldsymbol{\epsilon}}_i\|_\infty + \sum_{(j,i)\in\bar{\mathcal{F}}_i} \hat{\lambda}_{0i}|\Delta\hat{w}_{ji}| - \sum_{(j,i)\in\mathcal{G}_i} \hat{\lambda}_{\mathcal{G}_i}|\hat{w}_{ji}| \\
&= \sum_{(j,i)\in\mathcal{G}_i} (2\|\bar{\boldsymbol{\epsilon}}_i\|_\infty - \hat{\lambda}_{\mathcal{G}_i})|\hat{w}_{ji}| + \sum_{(j,i)\notin\bar{\mathcal{F}}_i\cup\mathcal{G}_i} 2\|\bar{\boldsymbol{\epsilon}}_i\|_\infty|\hat{w}_{ji}| + \sum_{(j,i)\in\bar{\mathcal{F}}_i} (2\|\bar{\boldsymbol{\epsilon}}_i\|_\infty + \hat{\lambda}_{0i})|\Delta\hat{w}_{ji}|.
\end{aligned}$$

The last equality above is due to $\mathbb{N}_d \times \{i\} = \mathcal{G}_i \cup (\bar{\mathcal{F}}_i \cup \mathcal{G}_i)^c \cup \bar{\mathcal{F}}_i$ and $\Delta\hat{w}_{ji} = \hat{w}_{ji}, \forall (j,i) \notin \bar{\mathcal{F}}_i \supseteq \mathcal{G}_i$. Rearranging the above inequality and noticing that $2\|\bar{\boldsymbol{\epsilon}}_i\|_\infty < \hat{\lambda}_{\mathcal{G}_i} \le \hat{\lambda}_{0i}$, we obtain:

$$\begin{aligned}
\sum_{(j,i)\in\mathcal{G}_i} |\hat{w}_{ji}| &\le \frac{2\|\bar{\boldsymbol{\epsilon}}_i\|_\infty}{\hat{\lambda}_{\mathcal{G}_i} - 2\|\bar{\boldsymbol{\epsilon}}_i\|_\infty} \sum_{(j,i)\notin\bar{\mathcal{F}}_i\cup\mathcal{G}_i} |\hat{w}_{ji}| + \frac{2\|\bar{\boldsymbol{\epsilon}}_i\|_\infty + \hat{\lambda}_{0i}}{\hat{\lambda}_{\mathcal{G}_i} - 2\|\bar{\boldsymbol{\epsilon}}_i\|_\infty} \sum_{(j,i)\in\bar{\mathcal{F}}_i} |\Delta\hat{w}_{ji}| \\
&\le \frac{2\|\bar{\boldsymbol{\epsilon}}_i\|_\infty + \hat{\lambda}_{0i}}{\hat{\lambda}_{\mathcal{G}_i} - 2\|\bar{\boldsymbol{\epsilon}}_i\|_\infty} \|\Delta\hat{\mathbf{w}}_{\mathcal{G}_i^c}\|_1.
\end{aligned} \tag{34}$$

Then Lemma 14 can be obtained from the above inequality and the following two inequalities.

$$\max_{i\in\mathbb{N}_m} \frac{2\|\bar{\boldsymbol{\epsilon}}_i\|_\infty + \hat{\lambda}_{0i}}{\hat{\lambda}_{\mathcal{G}_i} - 2\|\bar{\boldsymbol{\epsilon}}_i\|_\infty} \le \frac{2\|\bar{\Upsilon}\|_{\infty,\infty} + \hat{\lambda}_0}{\hat{\lambda}_{\mathcal{G}} - 2\|\bar{\Upsilon}\|_{\infty,\infty}} \text{ and } \sum_{i=1}^{m} x_i y_i \le \|\mathbf{x}\|_\infty \|\mathbf{y}\|_1.$$

■

Proof of Lemma 15 According to the definition of $\mathcal{G}$ ($\mathcal{G}_{(\ell)}$), we know that $\bar{\mathcal{F}}_i \cap \mathcal{G}_i = \emptyset$ ($i \in \mathbb{N}_m$) and $\forall (j,i) \in \mathcal{G}$ ($\mathcal{G}_{(\ell)}$), $\hat{\lambda}_{ji}^{(\ell-1)} = \lambda$. Thus, all conditions of Lemma 14 are satisfied, by noticing the relationship between Equation (22) and Equation (27). Based on the definition of $\mathcal{G}$ ($\mathcal{G}_{(\ell)}$), we easily obtain $\forall j \in \mathbb{N}_d$:

$$(j,i) \in \mathcal{G}_i, \forall i \in \mathbb{N}_m \text{ or } (j,i) \notin \mathcal{G}_i, \forall i \in \mathbb{N}_m.$$

and hence $k_\ell = |\mathcal{G}_1^c| = \cdots = |\mathcal{G}_m^c|$ (k_ℓ is some integer). Now, we assume that at stage $\ell \geq 1$:

$$k_\ell = |\mathcal{G}_1^c| = \cdots = |\mathcal{G}_m^c| \leq 2\bar{r}. \tag{35}$$

We will show in the second part of this proof that Equation (35) holds for all ℓ. Based on Lemma 19 and Equation (21), we have:

$$\begin{aligned}
\pi_i(2\bar{r}+s,s) &\leq \frac{s^{1/2}}{2}\sqrt{\rho_i^+(s)/\rho_i^-(2\bar{r}+2s)-1} \\
&\leq \frac{s^{1/2}}{2}\sqrt{1+s/(2\bar{r})-1} \\
&= 0.5s(2\bar{r})^{-1/2},
\end{aligned}$$

which indicates that

$$0.5 \leq t_i = 1 - \pi_i(2\bar{r}+s,s)(2\bar{r})^{1/2}s^{-1} \leq 1.$$

For all $t_i \in [0.5, 1]$, under the conditions of Equation (22) and Equation (27), we have

$$\frac{2\|\bar{\epsilon}_i\|_\infty + \lambda}{\lambda - 2\|\bar{\epsilon}_i\|_\infty} \leq \frac{2\|\bar{\Upsilon}\|_{\infty,\infty} + \lambda}{\lambda - 2\|\bar{\Upsilon}\|_{\infty,\infty}} \leq \frac{7}{5} \leq \frac{4-t_i}{4-3t_i} \leq 3.$$

Following Lemma 14, we have

$$\|\hat{W}_{\mathcal{G}}\|_{1,1} \leq 3\|\Delta\hat{W}_{\mathcal{G}^c}\|_{1,1} = 3\|\Delta\hat{W} - \Delta\hat{W}_{\mathcal{G}}\|_{1,1} = 3\|\Delta\hat{W} - \hat{W}_{\mathcal{G}}\|_{1,1}.$$

Therefore

$$\begin{aligned}
&\|\Delta\hat{W} - \Delta\hat{W}_I\|_{\infty,1} = \|\Delta\hat{W}_{\mathcal{G}} - \Delta\hat{W}_{\mathcal{J}}\|_{\infty,1} \\
&\leq \|\Delta\hat{W}_{\mathcal{J}}\|_{1,1}/s = (\|\Delta\hat{W}_{\mathcal{G}}\|_{1,1} - \|\Delta\hat{W} - \Delta\hat{W}_I\|_{1,1})/s \\
&\leq s^{-1}(3\|\Delta\hat{W} - \hat{W}_{\mathcal{G}}\|_{1,1} - \|\Delta\hat{W} - \Delta\hat{W}_I\|_{1,1}),
\end{aligned}$$

which implies that

$$\begin{aligned}
&\|\Delta\hat{W}\|_{2,1} - \|\Delta\hat{W}_I\|_{2,1} \leq \|\Delta\hat{W} - \Delta\hat{W}_I\|_{2,1} \\
&\leq (\|\Delta\hat{W} - \Delta\hat{W}_I\|_{1,1}\|\Delta\hat{W} - \Delta\hat{W}_I\|_{\infty,1})^{1/2} \\
&\leq \left(\|\Delta\hat{W} - \Delta\hat{W}_I\|_{1,1}\right)^{1/2}\left(s^{-1}(3\|\Delta\hat{W} - \hat{W}_{\mathcal{G}}\|_{1,1} - \|\Delta\hat{W} - \Delta\hat{W}_I\|_{1,1})\right)^{1/2} \\
&\leq \left(\left(3\|\Delta\hat{W} - \hat{W}_{\mathcal{G}}\|_{1,1}/2\right)^2\right)^{1/2} s^{-1/2} \\
&\leq (3/2)s^{-1/2}(2\bar{r})^{1/2}\|\Delta\hat{W} - \hat{W}_{\mathcal{G}}\|_{2,1} \\
&\leq (3/2)(2\bar{r}/s)^{1/2}\|\Delta\hat{W}_I\|_{2,1}.
\end{aligned}$$

In the above derivation, the third inequality is due to $a(3b-a) \leq (3b/2)^2$, and the fourth inequality follows from Equation (35) and $\bar{\mathcal{F}} \cap \mathcal{G} = \emptyset \Rightarrow \Delta\hat{W}_{\mathcal{G}} = \hat{W}_{\mathcal{G}}$. Rearranging the above inequality, we obtain at stage ℓ:

$$\|\Delta\hat{W}\|_{2,1} \leq \left(1 + 1.5\sqrt{\frac{2\bar{r}}{s}}\right)\|\Delta\hat{W}_I\|_{2,1}. \tag{36}$$

From Lemma 20, we have:

$$\begin{aligned}
&\max(0, \Delta\hat{\mathbf{w}}_{I_i}^T A_i \Delta\hat{\mathbf{w}}_i) \\
&\geq \rho_i^-(k_\ell + s)(\|\Delta\hat{\mathbf{w}}_{I_i}\| - \pi_i(k_\ell + s, s)\|\hat{\mathbf{w}}_{\mathcal{G}_i}\|_1/s)\|\Delta\hat{\mathbf{w}}_{I_i}\| \\
&\geq \rho_i^-(k_\ell + s)[1 - (1-t_i)(4-t_i)/(4-3t_i)]\|\Delta\hat{\mathbf{w}}_{I_i}\|^2 \\
&\geq 0.5 t_i \rho_i^-(k_\ell + s)\|\Delta\hat{\mathbf{w}}_{I_i}\|^2 \\
&\geq 0.25\rho_i^-(2\bar{r} + s)\|\Delta\hat{\mathbf{w}}_{I_i}\|^2 \\
&\geq 0.25\rho_{min}^-(2\bar{r} + s)\|\Delta\hat{\mathbf{w}}_{I_i}\|^2,
\end{aligned}$$

where the second inequality is due to Equation (34), that is

$$\begin{aligned}
\|\hat{\mathbf{w}}_{\mathcal{G}_i}\|_1 &\leq \frac{2\|\bar{\epsilon}_i\|_\infty + \hat{\lambda}_{0i}}{\hat{\lambda}_{\mathcal{G}_i} - 2\|\bar{\epsilon}_i\|_\infty}\|\Delta\hat{\mathbf{w}}_{\mathcal{G}_i^c}\|_1 \\
&\leq \frac{(2\|\bar{\epsilon}_i\|_\infty + \hat{\lambda}_{0i})\sqrt{k_\ell}}{\hat{\lambda}_{\mathcal{G}_i} - 2\|\bar{\epsilon}_i\|_\infty}\|\Delta\hat{\mathbf{w}}_{\mathcal{G}_i^c}\| \\
&\leq \frac{(2\|\bar{\epsilon}_i\|_\infty + \hat{\lambda}_{0i})\sqrt{k_\ell}}{\hat{\lambda}_{\mathcal{G}_i} - 2\|\bar{\epsilon}_i\|_\infty}\|\Delta\hat{\mathbf{w}}_{I_i}\| \\
&\leq \frac{(4-t_i)\sqrt{k_\ell}}{4-3t_i}\|\Delta\hat{\mathbf{w}}_{I_i}\|;
\end{aligned}$$

the third inequality follows from $1-(1-t_i)(4-t_i)/(4-3t_i) \geq 0.5t_i$ for $t_i \in [0.5, 1]$ and the fourth inequality follows from the assumption in Equation (35) and $t_i \geq 0.5$.

If $\Delta\hat{\mathbf{w}}_{I_i}^T A_i \Delta\hat{\mathbf{w}}_i \leq 0$, then $\|\Delta\hat{\mathbf{w}}_{I_i}\| = 0$. If $\Delta\hat{\mathbf{w}}_{I_i}^T A_i \Delta\hat{\mathbf{w}}_i > 0$, then we have

$$\Delta\hat{\mathbf{w}}_{I_i}^T A_i \Delta\hat{\mathbf{w}}_i \geq 0.25\rho_{min}^-(2\bar{r} + s)\|\Delta\hat{\mathbf{w}}_{I_i}\|^2. \tag{37}$$

By letting $\mathbf{v} = \Delta\hat{\mathbf{w}}_{I_i}$, we obtain the following from Equation (33):

$$\begin{aligned}
2\Delta\hat{\mathbf{w}}_{I_i}^T A_i \Delta\hat{\mathbf{w}}_i &= -2\Delta\hat{\mathbf{w}}_{I_i}^T \bar{\boldsymbol{\epsilon}}_i - \sum_{(j,i)\in I_i} \hat{\lambda}_{ji}\Delta\hat{w}_{ji}\mathrm{sign}(\hat{w}_{ji}) \\
&= -2\Delta\hat{\mathbf{w}}_{I_i}^T \bar{\boldsymbol{\epsilon}}_{\mathcal{G}_i^c} - 2\Delta\hat{\mathbf{w}}_{I_i}^T \bar{\boldsymbol{\epsilon}}_{\mathcal{G}_i} - \sum_{(j,i)\in\bar{\mathcal{F}}_i} \hat{\lambda}_{ji}\Delta\hat{w}_{ji}\mathrm{sign}(\hat{w}_{ji}) \\
&\quad - \sum_{(j,i)\in\mathcal{J}_i} \hat{\lambda}_{ji}|\Delta\hat{w}_{ji}| - \sum_{(j,i)\in\bar{\mathcal{F}}_i^c\cap\mathcal{G}_i^c} \hat{\lambda}_{ji}|\Delta\hat{w}_{ji}| \\
&= -2\Delta\hat{\mathbf{w}}_{I_i}^T \bar{\boldsymbol{\epsilon}}_{\mathcal{G}_i^c} - 2\Delta\hat{\mathbf{w}}_{\mathcal{J}_i}^T \bar{\boldsymbol{\epsilon}}_{\mathcal{J}_i} - \sum_{(j,i)\in\bar{\mathcal{F}}_i} \hat{\lambda}_{ji}\Delta\hat{w}_{ji}\mathrm{sign}(\hat{w}_{ji}) \\
&\quad - \sum_{(j,i)\in\mathcal{J}_i} \hat{\lambda}_{ji}|\Delta\hat{w}_{ji}| - \sum_{(j,i)\in\bar{\mathcal{F}}_i^c\cap\mathcal{G}_i^c} \hat{\lambda}_{ji}|\Delta\hat{w}_{ji}| \\
&\leq 2\|\Delta\hat{\mathbf{w}}_{I_i}\|\|\bar{\boldsymbol{\epsilon}}_{\mathcal{G}_i^c}\| + 2\|\bar{\boldsymbol{\epsilon}}_{\mathcal{J}_i}\|_\infty \sum_{(j,i)\in\mathcal{J}_i} |\Delta\hat{w}_{ji}| + \sum_{(j,i)\in\bar{\mathcal{F}}_i} \hat{\lambda}_{ji}|\Delta\hat{w}_{ji}| - \sum_{(j,i)\in\mathcal{J}_i} \hat{\lambda}_{ji}|\Delta\hat{w}_{ji}| \\
&\leq 2\|\Delta\hat{\mathbf{w}}_{I_i}\|\|\bar{\boldsymbol{\epsilon}}_{\mathcal{G}_i^c}\| + \left(\sum_{(j,i)\in\bar{\mathcal{F}}_i} \hat{\lambda}_{ji}^2\right)^{1/2} \|\Delta\hat{\mathbf{w}}_{\bar{\mathcal{F}}_i}\| \\
&\leq 2\|\Delta\hat{\mathbf{w}}_{I_i}\|\|\bar{\boldsymbol{\epsilon}}_{\mathcal{G}_i^c}\| + \left(\sum_{(j,i)\in\bar{\mathcal{F}}_i} \hat{\lambda}_{ji}^2\right)^{1/2} \|\Delta\hat{\mathbf{w}}_{I_i}\|. \qquad (38)
\end{aligned}$$

In the above derivation, the second equality is due to $I_i = \mathcal{J}_i \cup \bar{\mathcal{F}}_i \cup (\bar{\mathcal{F}}_i^c \cap \mathcal{G}_i^c)$; the third equality is due to $I_i \cap \mathcal{G}_i = \mathcal{J}_i$; the second inequality follows from $\forall (j,i) \in \mathcal{J}_i, \hat{\lambda}_{ji} = \lambda \geq 2\|\bar{\boldsymbol{\epsilon}}_i\|_\infty \geq 2\|\bar{\boldsymbol{\epsilon}}_{\mathcal{J}_i}\|_\infty$ and the last inequality follows from $\bar{\mathcal{F}}_i \subseteq \mathcal{G}_i^c \subseteq I_i$. Combining Equation (37) and Equation (38), we have

$$\|\Delta\hat{\mathbf{w}}_{I_i}\| \leq \frac{2}{\rho_{min}^-(2\bar{r}+s)} \left[2\|\bar{\boldsymbol{\epsilon}}_{\mathcal{G}_i^c}\| + \left(\sum_{(j,i)\in\bar{\mathcal{F}}_i} \hat{\lambda}_{ji}^2\right)^{1/2} \right].$$

Notice that

$$\|\mathbf{x}_i\| \leq a(\|\mathbf{y}_i\| + \|\mathbf{z}_i\|) \Rightarrow \|X\|_{2,1}^2 \leq m\|X\|_F^2 = m\sum_i \|\mathbf{x}_i\|^2 \leq 2ma^2(\|Y\|_F^2 + \|Z\|_F^2).$$

Thus, we have

$$\|\Delta\hat{W}_I\|_{2,1} \leq \frac{\sqrt{8m\left(4\|\bar{\Upsilon}_{\mathcal{G}_{(\ell)}^c}\|_F^2 + \sum_{(j,i)\in\bar{\mathcal{F}}} (\hat{\lambda}_{ji}^{(\ell-1)})^2\right)}}{\rho_{min}^-(2\bar{r}+s)}. \qquad (39)$$

Therefore, at stage ℓ, Equation (29) in Lemma 15 directly follows from Equation (36) and Equation (39). Following Equation (29), we have:

$$\begin{aligned}
\|\hat{W}^{(\ell)} - \bar{W}\|_{2,1} &= \|\Delta\hat{W}^{(\ell)}\|_{2,1} \\
&\leq \frac{\left(1+1.5\sqrt{\frac{2\bar{r}}{s}}\right)\sqrt{8m\left(4\|\bar{\Upsilon}_{\mathcal{G}^c_{(\ell)}}\|_F^2 + \sum_{(j,i)\in\bar{\mathcal{F}}}(\hat{\lambda}^{(\ell-1)}_{ji})^2\right)}}{\rho^-_{min}(2\bar{r}+s)} \\
&\leq \frac{8.83\sqrt{m}\sqrt{4\|\Upsilon\|^2_{\infty,\infty}|\mathcal{G}^c_{(\ell)}| + \bar{r}m\lambda^2}}{\rho^-_{min}(2\bar{r}+s)} \\
&\leq \frac{8.83\sqrt{m}\lambda\sqrt{\frac{8}{144}\bar{r}m + \bar{r}m}}{\rho^-_{min}(2\bar{r}+s)} \\
&\leq \frac{9.1m\lambda\sqrt{\bar{r}}}{\rho^-_{min}(2\bar{r}+s)},
\end{aligned}$$

where the first inequality is due to Equation (39); the second inequality is due to $s \geq \bar{r}$ (assumption in Theorem 8), $\hat{\lambda}_{ji} \leq \lambda$, $\bar{r}m = |\bar{\mathcal{H}}| \geq |\bar{\mathcal{F}}|$ and the third inequality follows from Equation (35) and $\|\bar{\Upsilon}\|^2_{\infty,\infty} \leq (1/144)\lambda^2$. Therefore, Equation (30) in Lemma 15 holds at stage ℓ.

Notice that we obtain Lemma 15 at stage ℓ, by assuming that Equation (35) is satisfied. To prove that Lemma 15 holds for all stages, we next need to prove by induction that Equation (35) holds at all stages.

When $\ell = 1$, we have $\mathcal{G}^c_{(1)} = \bar{\mathcal{H}}$, which implies that Equation (35) holds. Now, we assume that Equation (35) holds at stage ℓ. Thus, by hypothesis induction, we have:

$$\begin{aligned}
\sqrt{|\mathcal{G}^c_{(\ell+1)} \setminus \bar{\mathcal{H}}|} &\leq \sqrt{m\theta^{-2}\|\hat{W}^{(\ell)}_{\mathcal{G}^c_{(\ell+1)}\setminus\bar{\mathcal{H}}} - \bar{W}_{\mathcal{G}^c_{(\ell+1)}\setminus\bar{\mathcal{H}}}\|^2_{2,1}} \\
&\leq \sqrt{m}\theta^{-1}\left\|\hat{W}^{(\ell)} - \bar{W}\right\|_{2,1} \\
&\leq \frac{9.1m^{3/2}\lambda\sqrt{\bar{r}}\theta^{-1}}{\rho^-_{min}(2\bar{r}+s)} \\
&\leq \sqrt{\bar{r}m},
\end{aligned}$$

where θ is the thresholding parameter in Equation (1); the first inequality above follows from the definition of $\mathcal{G}_{(\ell)}$ in Lemma 15:

$$\begin{aligned}
&\forall (j,i) \in \mathcal{G}^c_{(\ell+1)} \setminus \bar{\mathcal{H}}, \|(\hat{\mathbf{w}}^{(\ell)})^j\|_1^2/\theta^2 = \|(\hat{\mathbf{w}}^{(\ell)})^j - \bar{\mathbf{w}}^j\|_1^2/\theta^2 \geq 1 \\
\Rightarrow &|\mathcal{G}^c_{(\ell+1)} \setminus \bar{\mathcal{H}}| \leq m\theta^{-2}\|\hat{W}^{(\ell)}_{\mathcal{G}^c_{(\ell+1)}\setminus\bar{\mathcal{H}}} - \bar{W}_{\mathcal{G}^c_{(\ell+1)}\setminus\bar{\mathcal{H}}}\|^2_{2,1};
\end{aligned}$$

the last inequality is due to Equation (23). Thus, we have:

$$|\mathcal{G}^c_{(\ell+1)} \setminus \bar{\mathcal{H}}| \leq \bar{r}m \Rightarrow |\mathcal{G}^c_{(\ell+1)}| \leq 2\bar{r}m \Rightarrow k_{\ell+1} \leq 2\bar{r}.$$

Therefore, Equation (35) holds at all stages. Thus the two inequalities in Lemma 15 hold at all stages. This completes the proof of the lemma. ■

Proof of Lemma 16 The first inequality directly follows from $\bar{\mathcal{H}} \supseteq \bar{\mathcal{F}}$. Next, we focus on the second inequality. For each $(j,i) \in \bar{\mathcal{F}}$ $(\bar{\mathcal{H}})$, if $\|\hat{\mathbf{w}}^j\|_1 < \theta$, by considering Equation (20), we have

$$\|\bar{\mathbf{w}}^j - \hat{\mathbf{w}}^j\|_1 \geq \|\bar{\mathbf{w}}^j\|_1 - \|\hat{\mathbf{w}}^j\|_1 \geq 2\theta - \theta = \theta.$$

Therefore, we have for each $(j,i) \in \bar{\mathcal{F}}$ $(\bar{\mathcal{H}})$:

$$I\left(\|\hat{\mathbf{w}}^j\|_1 < \theta\right) \leq \|\bar{\mathbf{w}}^j - \hat{\mathbf{w}}^j\|_1/\theta.$$

Thus, the second inequality of Lemma 16 directly follows from the above inequality. ■

Appendix B. Lemmas from Zhang (2010)

Lemma 18 *Let $\mathbf{a} \in \mathbb{R}^n$ be a fixed vector and $\mathbf{x} \in \mathbb{R}^n$ be a random vector which is composed of independent sub-Gaussian components with parameter σ. Then we have:*

$$\Pr(|\mathbf{a}^T\mathbf{x}| \geq t) \leq 2\exp\left(-t^2/(2\sigma^2\|\mathbf{a}\|^2)\right), \forall t > 0.$$

Lemma 19 *The following inequality holds:*

$$\pi_i(k_i, s_i) \leq \frac{s_i^{1/2}}{2}\sqrt{\rho_i^+(s_i)/\rho_i^-(k_i+s_i) - 1}.$$

Lemma 20 *Let $\mathcal{G}_i \subseteq \mathbb{N}_d \times \{i\}$ such that $|\mathcal{G}_i^c| = k_i$, and let $\mathcal{J}_i$ be indices of the s_i largest components (in absolute values) of $\mathbf{w}_{\mathcal{G}_i}$ and $I_i = \mathcal{G}_i^c \cup \mathcal{J}_i$. Then for any $\mathbf{w}_i \in \mathbb{R}^d$, we have*

$$\max(0, \mathbf{w}_{I_i}^T A_i \mathbf{w}_i) \geq \rho_i^-(k_i+s_i)(\|\mathbf{w}_{I_i}\| - \pi_i(k_i+s_i, s_i)\|\mathbf{w}_{\mathcal{G}_i}\|_1/s_i)\|\mathbf{w}_{I_i}\|.$$

Lemma 21 *Let $\bar{\epsilon}_i = [\bar{\varepsilon}_{1i}, \cdots, \bar{\varepsilon}_{di}] = \frac{1}{n}X_i^T(X_i\bar{\mathbf{w}}_i - \mathbf{y}_i)$ $(i \in \mathbb{N}_m)$, and $\bar{\mathcal{H}}_i \subseteq \mathbb{N}_d \times \{i\}$. Under the conditions of Assumption 1, the followings hold with probability larger than $1-\eta$:*

$$\|\bar{\epsilon}_{\bar{\mathcal{H}}_i}\|^2 \leq \sigma^2\rho_i^+(|\bar{\mathcal{H}}_i|)(7.4|\bar{\mathcal{H}}_i| + 2.7\ln(2/\eta))/n.$$

Appendix C. MSMTFL-type Algorithms

We present the multi-stage (-type) algorithms for the formulations in Equation (3) and Equation (4) below.

References

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

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

Algorithm 2: MSMTFL-CapL1: Multi-Stage Multi-Task Feature Learning for solving the capped-ℓ_1 regularized feature learning problem in Equation (3)

Initialize $\lambda_j^{(0)} = \lambda$;
for $\ell = 1, 2, \cdots$ **do**
    Let $\hat{W}^{(\ell)}$ be a solution of the following problem:

$$\min_{W \in \mathbb{R}^{d \times m}} \left\{ l(W) + \sum_{j=1}^{d} \lambda_{ji}^{(\ell-1)} |w_{ji}| \right\}.$$

    Let $\lambda_{ji}^{(\ell)} = \lambda I(|\hat{w}_{ji}^{(\ell)}| < \theta)$ $(j = 1, \cdots, d, i = 1, \cdots, m)$, where $\hat{w}_{ji}^{(\ell)}$ is the (j,i)-th entry of $\hat{W}^{(\ell)}$ and $I(\cdot)$ denotes the $\{0,1\}$-valued indicator function.
end

Algorithm 3: MSMTFL-CapL1,L2: Multi-Stage Multi-Task Feature Learning for solving the capped-ℓ_1, ℓ_2 regularized multi-task feature learning problem in Equation (4)

Initialize $\lambda_j^{(0)} = \lambda$;
for $\ell = 1, 2, \cdots$ **do**
    Let $\hat{W}^{(\ell)}$ be a solution of the following problem:

$$\min_{W \in \mathbb{R}^{d \times m}} \left\{ l(W) + \sum_{j=1}^{d} \lambda_{j}^{(\ell-1)} \|\mathbf{w}^j\| \right\}.$$

    Let $\lambda_j^{(\ell)} = \lambda I(\|(\hat{\mathbf{w}}^{(\ell)})^j\| < \theta)$ $(j = 1, \cdots, d)$, 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

D.P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.

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.

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

R. Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

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

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

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

J. Fan, L. Xue, and H. Zou. Strong oracle optimality of folded concave penalized estimation. *ArXiv Preprint ArXiv:1210.5992*, 2012.

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

P. Gong, C. Zhang, Z. Lu, J. Huang, and J. Ye. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. In *ICML*, 2013a.

P. Gong, C. Zhang, Z. Lu, J. Huang, and J. Ye. *GIST: General Iterative Shrinkage and Thresholding for Non-convex Sparse Learning*. Tsinghua University, 2013b. URL `http://www.public.asu.edu/~jye02/Software/GIST`.

L. Grippo and M. Sciandrone. On the convergence of the block nonlinear gauss-seidel method under convex constraints. *Operations Research Letters*, 26(3):127–136, 2000.

J. Huang and T. Zhang. The benefit of group sparsity. *The Annals of Statistics*, 38(4):1978–2004, 2010.

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

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

M. Kolar, J. Lafferty, and L. Wasserman. Union support recovery in multi-task learning. *Journal of Machine Learning Research*, 12:2415–2435, 2011.

J. Liu, S. Ji, and J. Ye. Multi-task feature learning via efficient $\ell_{2,1}$-norm minimization. In *UAI*, pages 339–348, 2009.

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

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

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

G. Obozinski, B. Taskar, and M.I. Jordan. Multi-task feature selection. Technical report, Statistics Department, UC Berkeley, 2006.

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

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

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

R.T. Rockafellar. *Convex Analysis*. Princeton University Press (Princeton, NJ), 1970.

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

X. Shen, W. Pan, and Y. Zhu. Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*, 107(497):223–232, 2012.

R. Tibshirani. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490, 2013.

J.F. Toland. A duality principle for non-convex optimisation and the calculus of variations. *Archive for Rational Mechanics and Analysis*, 71(1):41–61, 1979.

P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494, 2001.

S.A. 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.

S. Wright, R. Nowak, and M. Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(7):2479–2493, 2009.

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

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

A.L. Yuille and A. Rangarajan. The concave-convex procedure. *Neural Computation*, 15(4):915–936, 2003.

C.H. 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.

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

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

T. Zhang. Adaptive forward-backward greedy algorithm for sparse learning with linear models. In *NIPS*, pages 1921–1928, 2008.

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

T. Zhang. Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research*, 11:1081–1107, 2010.

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

Y. Zhang and D.Y. Yeung. Multi-task learning using generalized t process. *Journal of Machine Learning Research - Proceedings Track*, 9:964–971, 2010.

Y. Zhang, D. Yeung, and Q. Xu. Probabilistic multi-task feature selection. In *NIPS*, pages 2559–2567, 2010.

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

H. Zou and R. Li. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36(4):1509, 2008.