

---

# Generalized Block-Diagonal Structure Pursuit: Learning Soft Latent Task Assignment against Negative Transfer

---

Zhiyong Yang<sup>1,2</sup>   Qianqian Xu<sup>3</sup>   Yangbangyan Jiang<sup>1,2</sup>  
Xiaochun Cao<sup>1,2,6</sup>   Qingming Huang<sup>3,4,5,6\*</sup>

<sup>1</sup>State Key Laboratory of Information Security, Institute of Information Engineering, CAS

<sup>2</sup>School of Cyber Security, University of Chinese Academy of Sciences

<sup>3</sup>Key Lab. of Intelligent Information Processing, Institute of Computing Technology, CAS

<sup>4</sup>School of Computer Science and Tech., University of Chinese Academy of Sciences

<sup>5</sup>Key Laboratory of Big Data Mining and Knowledge Management, CAS

<sup>6</sup>Peng Cheng Laboratory

yangzhiyong@iie.ac.cn, xuqianqian@ict.ac.cn, jiangyangbangyan@iie.ac.cn  
caoxiaochun@iie.ac.cn, qmhuang@ucas.ac.cn

## Abstract

In multi-task learning, a major challenge springs from a notorious issue known as *negative transfer*, which refers to the phenomenon that sharing the knowledge with dissimilar and hard tasks often results in a worsened performance. To circumvent this issue, we propose a novel multi-task learning method, which simultaneously learns latent task representations and a block-diagonal Latent Task Assignment Matrix (LTAM). Different from most of the previous work, pursuing the Block-Diagonal structure of LTAM (assigning latent tasks to output tasks) alleviates negative transfer via punishing inter-group knowledge transfer and sharing. This goal is challenging since our notion of Block-Diagonal Property extends the traditional notion for homogeneous and square matrices. In this paper, we propose a spectral regularizer which is proven to leverage the expected structure. Practically, we provide a relaxation scheme which improves the flexibility of the model. With the objective function given, we then propose an alternating optimization method, which reveals an interesting connection between our method and the optimal transport problem. Finally, the method is demonstrated on a simulation dataset, three real-world benchmark datasets and further applied to two personalized attribute learning datasets.

## 1 Introduction

Multi-Task Learning (MTL) is a learning paradigm whose aim is to leverage useful information contained in multiple related tasks to help improve the generalization performance of all the tasks [Caruana, 1997]. Nowadays, MTL has emerged as a fundamental building block for a wide range of applications ranging from scene parsing [Xu et al., 2018], attribute learning [Cao et al., 2018, Yang et al., 2019a, 2018], text classification [Liu et al., 2017], sequence labeling [Lin et al., 2018], to travel time estimation [Li et al., 2018], *etc.*

The fundamental belief of MTL lies in that sharing knowledge among multiple tasks often results in an improvement in generalization performance, which is especially of great significance in the

---

\*Corresponding author.

presence of insufficient data annotations [Heskes, 1998]. Based on the belief, a great number of studies have been carried out to explore the problem of how to share valuable knowledge across different tasks. The early studies of MTL (e.g. [Argyriou et al., 2008a]) hold that all the tasks share common and sparse features. However, [Kang et al., 2011] later points out that if not all the tasks are indeed related, then sharing common features with dissimilar and hard tasks often results in performance degradation, which is termed as *negative transfer*.

To address this issue, recent studies in the odyssey against *negative transfer* fall in two major directions. One line of the researches leverages the grouping effect based on the *latent-task-agnostic* idea which develops structural regularizers where only the original per-task parameters are utilized. [Kang et al., 2011, Kshirsagar et al., 2017] directly formulate the tasking grouping as a mixed integer programming (or a relaxation [Frecon et al., 2018]), which simultaneously learns the group index and the model parameters. [Argyriou et al., 2008b, Zhou et al., 2011a, Jacob et al., 2009, Lee et al., 2016, Liu and Pan, 2017, McDonald et al., 2014] leverage the tasking grouping via enforcing a specific structure, hopefully block-diagonal, on the task correlation matrix. As an extension of this formulation [Zhong and Kwok, 2012] resorts to feature-specific task clustering. The other line of researches formulates the MTL based on the latent task, where the model parameter is represented as a linear combination of latent task basis. [Kumar and III, 2012] gives an early trial of this formulation in search of a more flexible MTL model. Similarly, in the work of [Maurer et al., 2013], a sparse coding model is proposed for MTL, where the dictionary is set as the latent task basis and the code is set as the linear combination coefficients of such basis. Recently, [Lee et al., 2018] also provides an asymmetric learning framework based on the latent task representation where transferring knowledge from unreliable tasks to reliable tasks is explicitly punished.

The two aforementioned directions, i.e., learning grouped model structure and latent task representation provide complementary functions in a sense that the former one avoids inter-group negative transfer, while the latter one focuses on learning a more flexible model. However, the related studies on how to bridge the efforts of these two directions are sparse. To the best of our knowledge, the only two studies along this direction are [Crammer and Mansour, 2012, Barzilai and Crammer, 2015]. However, both studies adopt a strong assumption that each group of tasks is only assigned with one latent task basis.

To leverage a flexible grouping structure with latent task representations, we should allow each task cluster to have multiple latent tasks. Motivated by this, we study the structural learning problem of how to learn a block-diagonal Latent Task Assignment Matrix (LATM). With the block-diagonal structure, tasks within each group share a subset (not necessarily one) of the latent task basis. Since LATM is not a squared matrix and marginal constraints are also necessary to avoid isolated tasks/latent tasks, our notion of block-diagonality generalizes the one adopted in the self-expression scenario [Lu et al., 2019, Lee et al., 2016, Liu and Pan, 2017], which makes traditional structural regularizers not available to solve our problem. Our first contribution then comes as an equivalent spectral condition that realizes our pursuit of the generalized block-diagonal structure. Then we propose a new MTL method named *Generalized Block-Diagonal Structural Pursuit (GBDSP)*, which utilizes the spectral condition as a novel regularizer with a relaxation scheme. In our optimization method, the intermediate solution produced provides new insights into how negative transfer is alleviated in our model. Theoretical studies show how the proposed regularizer guarantees the expected structure. Finally, empirical studies demonstrate the effectiveness of the proposed method.

## 2 Generalized Block-Diagonal Structure Pursuit

**Notations** The notations adopted in this paper are enumerated as follows.  $\mathbb{S}_m$  denotes the set of all symmetric matrices in  $\mathbb{R}^{m \times m}$ . The eigenvalues of a symmetric matrix  $\mathbf{A} \in \mathbb{S}_m$  are denoted as  $\lambda_1(\mathbf{A}), \dots, \lambda_m(\mathbf{A})$  such that  $\lambda_1(\mathbf{A}) \geq \lambda_2(\mathbf{A}) \geq \dots \geq \lambda_m(\mathbf{A})$ .  $\langle \cdot, \cdot \rangle$  denotes the inner product for two matrices or two vectors. Given two vectors  $\mathbf{a}$  and  $\mathbf{b}$ ,  $\mathbf{a} \oplus \mathbf{b}$  denotes the outer sum  $\mathbf{a}\mathbf{1}^\top + \mathbf{1}\mathbf{b}^\top$ . Given two matrices  $\mathbf{A}$  and  $\mathbf{B}$ ,  $\mathbf{A} \oplus \mathbf{B}$  denotes the direct sum of two matrices, i.e.,  $\mathbf{A} \oplus \mathbf{B} = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{bmatrix}$ , and we say  $\mathbf{A} \succeq \mathbf{B}$ , if  $\mathbf{A} - \mathbf{B}$  is positive semi-definite. For distributions,  $\mathcal{N}(\mu, \sigma^2)$  denotes the normal distribution.  $\mathcal{P}_m$  denotes the set of all permutation matrices in  $\mathbb{R}^{m \times m}$ . For two matrices  $\mathbf{A}$  and  $\mathbf{B}$  having the same size,  $d(\mathbf{A}, \mathbf{B}) = \|\mathbf{A} - \mathbf{B}\|_F^2$ . Given an event  $\mathcal{A}$ ,  $\delta(\mathcal{A})$  denotes the corresponding indicator function. *Moreover, let us note two notations in our paper that are prone*

to be confused.  $k$  denotes the dimension of the latent task representation.  $K(K \leq k)$  denotes the number of given groups.

## 2.1 Model Formulation

Before entering our new method, we first provide a brief introduction of the multi-task learning setting we adopted in this paper. Here we adopt the latent task representation framework proposed in [Kumar and III, 2012]. Given  $T$  tasks to be learned simultaneously, we denote the training data as:  $\{(\mathbf{X}^{(1)}, \mathbf{Y}^{(1)}), \dots, (\mathbf{X}^{(T)}, \mathbf{Y}^{(T)})\}$ . Here  $\mathbf{X}^{(i)} = [\mathbf{X}_1^{(i)}, \dots, \mathbf{X}_{n_i}^{(i)}]^\top$ , where  $\mathbf{X}_j^{(i)} \in \mathbb{R}^{d \times 1}$  is the input feature for the  $j$ -th sample of the  $i$ -th task,  $n_i$  denotes the number of instances and  $d$  represents the feature dimension. Similarly  $\mathbf{Y}^{(i)} = [\mathbf{Y}_1^{(i)}, \dots, \mathbf{Y}_{n_i}^{(i)}]^\top \in \mathbb{R}^{n_i \times 1}$ , where  $\mathbf{Y}_j^{(i)}$  is the corresponding response for the  $j$ -th sample of the  $i$ -th task. Following the standard multi-task learning paradigm, we learn a model  $\hat{\mathbf{Y}}^{(i)}(x) = \mathbf{W}^{(i)\top} \mathbf{x}$  to estimate the output response for each task  $i$ . Here we call  $\mathbf{W} = [\mathbf{W}^{(1)}, \dots, \mathbf{W}^{(T)}] \in \mathbb{R}^{d \times T}$  the *per-task parameter matrix*. Furthermore, to model the relationship among the tasks, we assume that the per-task parameters lie in a low dimensional subspace. To this end, we introduce a set of latent task basis  $\mathbf{L} \in \mathbb{R}^{d \times k}$ , where  $k < T$ . For each given task  $i$ , its parameter is then represented as a linear combination of the basis by letting  $\mathbf{W}^{(i)} = \mathbf{L}\mathbf{S}^{(i)}$ , where  $\mathbf{S}^{(i)} \in \mathbb{R}^{k \times 1}$  is the combination coefficients. Given a loss function  $\ell(y, \hat{y})$ , the empirical risk for the  $i$ -th task is defined as  $\mathcal{J}^{(i)} = \sum_{j=1}^{n_i} \ell(\mathbf{Y}_j^{(i)}, \hat{\mathbf{Y}}_j^{(i)})$ . Given proper regularizers  $\Omega(\mathbf{L})$ ,  $\Omega(\mathbf{S})$ , [Kumar and III, 2012] learns  $\mathbf{L}, \mathbf{S}$  from the problem  $\arg\min_{\mathbf{L}, \mathbf{S}} \sum_{i=1}^T \mathcal{J}^{(i)} + \alpha_1 \cdot \Omega(\mathbf{L}) + \alpha_2 \cdot \Omega(\mathbf{S})$ . In this paper, we adopt the  $F$ -norm penalty for  $\mathbf{L}$ , i.e., we set  $\Omega(\mathbf{L}) = \|\mathbf{L}\|_F^2$ . And we seek new solutions against negative transfer from  $\Omega(\mathbf{S})$ . In this setting, we must deal with both the latent task representations and the true tasks. To differentiate the two, we refer the former ones to latent tasks  $(\mathbf{l}_1, \dots, \mathbf{l}_k)$  and the latter ones to output tasks  $(\mathbf{o}_1, \dots, \mathbf{o}_T)$ .

With the latent task formulation  $\mathbf{W} = \mathbf{L}\mathbf{S}$ ,  $\mathbf{S}$  then captures the importance of the latent tasks to the output tasks. In a natural sense, we regard  $S_{i,j}$  as  $\mathbb{P}(\mathbf{l} = i | \mathbf{o} = j)$ , namely the possibility of choosing  $\mathbf{l}_i$  to represent  $\mathbf{o}_j$ . In this probabilistic view,  $\mathbf{L}\mathbf{S}^{(i)}$  now becomes  $\mathbb{E}_{\mathbf{l}|\mathbf{o}=i}(\mathbf{L})$ , i.e., the expectation of the latent tasks representations assigned to task  $\mathbf{o}_i$ . We then call  $\mathbf{S}$  the Latent Task Assignment Matrix (LTAM), since the conditional possibility could be considered as a soft assignment score. Before developing a proper regularizer, we must first answer the question that *can we choose  $\mathbf{S}$  arbitrarily?* Unfortunately, we will immediately see that the answer is negative. Let us denote  $\mathbf{S}^\dagger \in \mathbb{R}^{k \times T}$  by  $S_{i,j}^\dagger = \mathbb{P}(\mathbf{l} = i, \mathbf{o} = j)$ , the joint distribution of  $\mathbf{l}$  and  $\mathbf{o}$ . Note that  $\mathbf{S}^\dagger \mathbf{1}$  and  $\mathbf{S}^{\dagger\top} \mathbf{1}$  are marginal distributions on  $\mathbf{l}$  and  $\mathbf{o}$ , we come to two extreme cases that must be ruled out from consideration. If  $(\mathbf{S}^\dagger \mathbf{1}_T)_i = 0$  then  $\mathbf{l}_i$  becomes an isolated latent task which is irrelevant with all the output task. Similarly, if  $(\mathbf{S}^{\dagger\top} \mathbf{1}_k)_j = 0$  then  $\mathbf{o}_j$  becomes isolated with no latent tasks assigned to it. To remove extreme cases of such kinds, we then pose normalization constraints on  $\mathbf{S}^\dagger$  for each row and column in the form:  $\mathbf{S}^\dagger \mathbf{1}_T = \mathbf{a} > \mathbf{0}_k$ ,  $\mathbf{S}^{\dagger\top} \mathbf{1}_k = \mathbf{b} > \mathbf{0}_T$ . To maintain fairness, we do not expect to introduce extra bias from the choice of marginal distribution. Such a spirit guides us to put out  $\mathbf{a} = \mathbf{1}_k/k$ ,  $\mathbf{b} = \mathbf{1}_T/T$ . Moreover, this also simplifies the relation between  $\mathbf{S}$  and  $\mathbf{S}^\dagger$  with  $\mathbf{S} = T\mathbf{S}^\dagger$ . From all above, we adopt the transportation polytopes  $\Pi(\mathbf{a}, \mathbf{b}) = \{\mathbf{S}^\dagger \in \mathbb{R}_+^{k \times T} : \mathbf{S}^\dagger \mathbf{1} = \mathbf{a}, \mathbf{S}^{\dagger\top} \mathbf{1} = \mathbf{b}\}$  as the feasible set for the parameter  $\mathbf{S}^\dagger$ .

So far we have known that  $\mathbf{S}$  must satisfy the marginal constraints to make the solution non-trivial. Now let us step further to seek out what else we should pose on  $\mathbf{S}$  to suppress inter-group transfer. In this paper, we adopt a basic assumption that the latent tasks and output tasks could be clustered into  $K$  independent groups. In order to avoid negative transfer, we hope the possibility to assign  $\mathbf{l}_i$  to  $\mathbf{o}_j$  is nonzero if and only if  $(i, j)$  belongs to the same group. This leads to a block-diagonal structure of  $\mathbf{S}^\dagger$  up to row and column permutations. Next, we give a formal definition of the desired block-diagonal structure with  $\mathbf{S}^\dagger \in \Pi(\mathbf{a}, \mathbf{b})$ , based on the following simple idea. If the columns and rows of  $\mathbf{S}^\dagger$  could be partitioned into  $K$  groups,  $\mathbf{S}^\dagger$  could then be expressed as a direct sum of  $K$  blocks up to proper column and row permutations. The maximum of such  $K$  then implies the number of groups in the matrix.<sup>1</sup> This motivates the following definition of the grouping structure, which is termed as the *Generalized  $K$  Block Diagonal Property (GKB DP)* in our paper.

<sup>1</sup>If  $K$  is not the maximum of such numbers, we can always find out more disjoint blocks.

**Definition 1** (Generalized  $K$  Block Diagonal Property). Given a matrix  $\mathbf{S}^\dagger \in \Pi(\mathbf{a}, \mathbf{b})$ , if there exists a permutation matrix over rows  $\mathbf{P}_r \in \mathcal{P}_k$  and a permutation matrix over columns  $\mathbf{P}_c \in \mathcal{P}_T$  such that:  $\mathbf{P}_r \mathbf{S}^\dagger \mathbf{P}_c = \bigoplus_{i=1}^K \hat{\mathbf{S}}^{(i)}$  where  $\hat{\mathbf{S}}^{(i)} \neq \mathbf{0}$ ,  $\hat{\mathbf{S}}^{(i)} \in \mathbb{R}^{k_i \times T_i}$ ,  $\sum_i k_i = k$ ,  $\sum_i T_i = T$ , then we define  $\chi_{\mathbf{S}^\dagger}(\mathbf{P}_r, \mathbf{P}_c) = K$ . Moreover, we say  $\mathbf{S}^\dagger$  is Generalized  $K$  Block Diagonal if  $\chi_{\mathbf{S}^\dagger} = \max_{\mathbf{P}_r \in \mathcal{P}_k, \mathbf{P}_c \in \mathcal{P}_T} \chi_{\mathbf{S}^\dagger}(\mathbf{P}_r, \mathbf{P}_c) = K$ .

Note that **GKBDP** extends the notion of block-diagonal property deployed in self-expression [Lee et al., 2016, Liu and Pan, 2017, Lu et al., 2019, Jiang et al., 2018, 2019, Yang et al., 2019b] which is only available for square matrices. Furthermore, the traditional self-expression-based block-diagonal property requires  $\mathbf{P}_r = \mathbf{P}_c^\top$  [Lu et al., 2019], i.e., a simultaneous permutation on columns and rows (the  $i$ -th row and the  $i$ -th column represent the same object). However, this is not the case in our notion of **GKBDP** since the columns and rows here represent heterogeneous concepts (the  $i$ -th row is for  $l_i$ ,  $i$ -th column is for  $o_i$ ). Moreover we also consider the marginal constraints  $\mathbf{S}^\dagger \mathbf{1} = \mathbf{a}$ , and  $\mathbf{S}^{\dagger\top} \mathbf{1} = \mathbf{b}$  to avoid isolated  $l$  or  $o$ . Based on the aforementioned facts, traditional regularization schemes are thus not directly applicable to leverage the **GKBDP**.

Next, we derive an equivalent condition for **GKBDP**, which directly leads to the formulation of our method. First, we define an auxiliary bipartite graph  $\mathcal{G}_{l \cup o} = (\mathcal{V}_{l \cup o}, \mathcal{E}_{l \cup o}, \mathbf{A}_{l \cup o})$ . The vertices of  $\mathcal{G}_{l \cup o}$  include all  $l$  and  $o$ . Denote  $\mathcal{V}_o$  as the set of all output tasks and  $\mathcal{V}_l$  as the set of all latent tasks, the vertex set  $\mathcal{V}_{l \cup o}$  is then defined as  $\mathcal{V}_{l \cup o} = \mathcal{V}_l \cup \mathcal{V}_o$ . To define the edge set, we first define an affinity matrix  $\mathbf{A}_{l \cup o}$  in the form  $\mathbf{A}_{l \cup o} = \begin{bmatrix} \mathbf{0} & \mathbf{S}^\dagger \\ \mathbf{S}^{\dagger\top} & \mathbf{0} \end{bmatrix}$ , where the  $(i, j) \in \mathcal{E}_{l \cup o}$  if and only if  $\mathbf{A}_{l \cup o} ij \neq 0$ . Then the well-known graph Laplacian follows as  $\Delta(\mathbf{S}^\dagger) = \text{diag}(\mathbf{A}_{l \cup o} \mathbf{1}) - \mathbf{A}_{l \cup o}$ . With the definition of  $\mathcal{G}_{l \cup o}$ , we could derive the following theorem.

**Theorem 1.** If  $\mathbf{S}^\dagger \in \Pi(\mathbf{a}, \mathbf{b})$ ,  $\chi_{\mathbf{S}} = K$  holds if and only if  $\dim(\text{Null}(\Delta(\mathbf{S}^\dagger))) = K$ , i.e., the 0 eigenvalue of  $\Delta(\mathbf{S}^\dagger)$  has multiplicity  $K$ . Moreover, denote  $\mathcal{A}^{(i)}$  as the set of latent and output tasks belonging to the  $i$ -th block of  $\mathbf{S}$ , the eigenspace of 0 is spanned by  $\iota_{\mathcal{A}^{(1)}}, \iota_{\mathcal{A}^{(2)}}, \dots, \iota_{\mathcal{A}^{(K)}}$ , where  $\iota_{\mathcal{A}^{(i)}} \in \mathbb{R}^{(k+T) \times 1}$ ,  $[\iota_{\mathcal{A}^{(i)}}]_j = 1$  if  $j \in \mathcal{A}^{(i)}$ , otherwise  $[\iota_{\mathcal{A}^{(i)}}]_j = 0$ .

The proof can be found in Appendix B.1. With the theorem, we can now step further to seek a suitable regularizer realizing **GKBDP**. It becomes straightforward that leveraging **GKBDP** requires the sum of bottom  $K$  eigenvalues to be as small as possible. Let  $N = k + T$  denote the total number of nodes in  $\mathcal{G}_{l \cup o}$ , we then need to minimize  $\sum_{N-K+1}^N \lambda_i(\Delta(\mathbf{S}^\dagger))$  with the constraint  $\mathbf{S}^\dagger \in \Pi(\mathbf{a}, \mathbf{b})$ . Following the variational characterization of eigenvalues, we could reformulate eigenvalue calculation as an optimization problem with the following theorem.

**Theorem 2.** Let  $\mathcal{M} = \{\mathbf{U} : \mathbf{U} \in \mathbb{S}_N, \mathbf{I} \succeq \mathbf{U} \succeq \mathbf{0}, \text{tr}(\mathbf{U}) = K\}$ , then  $\forall \mathbf{A} \in \mathbb{S}^N$ :  $\sum_{N-K+1}^N \lambda_i(\mathbf{A}) = \min_{\mathbf{U} \in \mathcal{M}} \langle \mathbf{A}, \mathbf{U} \rangle$ , with an optimal value reached at  $\mathbf{U} = \mathbf{V}_K \mathbf{V}_K^\top$ , where  $\mathbf{V}_K$  represents the eigenvectors of the smallest  $K$  eigenvalues of  $\mathbf{A}$ .

The theorem slightly extends the results in [Overton and Womersley, 1992b], which considers top- $K$  eigenvalues. A proof for the theorem could be found in Appendix B.2. Back to our practical problem, Thm.2 provides a regularizer as  $\Omega(\mathbf{S}^\dagger) = \inf \{\langle \Delta(\mathbf{S}^\dagger), \mathbf{U} \rangle : \mathbf{U} \in \mathcal{M}\}$ . Denote  $\tilde{\mathcal{J}} = \sum_{i=1}^T \mathcal{J}^{(i)}$ ,  $\Omega_1 = \alpha_1 \cdot \|\mathbf{L}\|_F^2/2$ ,  $\Omega_2 = \alpha_3 \cdot \langle \Delta(\mathbf{S}^\dagger), \mathbf{U} \rangle$ , we then reach an MTL model based on the latent task framework:

$$\min_{\mathbf{L}, \mathbf{S}, \mathbf{S}^\dagger, \mathbf{U}} \tilde{\mathcal{J}} + \Omega_1 + \Omega_2, \quad \text{s.t. } \mathbf{S}^\dagger \in \Pi(\mathbf{a}, \mathbf{b}), \mathbf{U} \in \mathcal{M}, \mathbf{S} = \mathbf{T} \mathbf{S}^\dagger. \quad (\text{Obj}_0)$$

This model exactly realizes **GKBDP**. However, this exact model is impractical in the following sense. First, it is hard to solve (Obj<sub>0</sub>) directly since multiple constraints are wrapped together on  $\mathbf{S}$ . Moreover, it encourages a strict structural control of  $\mathbf{S}$ , prohibiting overlapped subspaces even when it benefits the performance. These problems lead us to a relaxed implementation of (Obj<sub>0</sub>), bringing us possibilities to embrace a practical and a more flexible solution.

To avoid directly controlling the structure of  $\mathbf{S}$ , we relax the constraint  $\mathbf{S} = \mathbf{T} \mathbf{S}^\dagger$  as a distance penalty<sup>2</sup>  $\Omega_3 = \alpha_2 \cdot d(\mathbf{S}, \mathbf{T} \mathbf{S}^\dagger)/2$ . This brings us to the final optimization problem:

<sup>2</sup>Let us note that, in the rest of the paper,  $\mathbf{S} = \mathbf{T} \mathbf{S}^\dagger$  no longer holds

$$\min_{\mathbf{L}, \mathbf{S}, \mathbf{S}^\dagger, \mathbf{U}} \mathcal{J} = \tilde{\mathcal{J}} + \Omega_1 + \Omega_2 + \Omega_3, \quad s.t. \mathbf{S}^\dagger \in \Pi(\mathbf{a}, \mathbf{b}), \mathbf{U} \in \mathcal{M}. \quad (\text{Obj})$$

The relaxation scheme improves the flexibility of our model via leveraging a partial structural control, which decomposes  $\mathbf{S}$  into a structural component  $T\mathbf{S}^\dagger$  and a dense component as the residual  $\mathbf{S} - T\mathbf{S}^\dagger$ . The new dense component allows  $\mathbf{S}$  to slightly (controlled by the magnitude of  $\alpha_2$ ) violate the structural constraint, in search of a better performance.

Alternatively, we could also interpret (Obj) as a way to leverage hierarchical priors on the model. This is specified by the generative model in the following:

$$\begin{aligned} [\mathbf{Y}^{(i)} | \mathbf{X}^{(i)}, \mathbf{L}, \mathbf{S}^{(i)}] &\sim \mathcal{N}(\mathbf{X}^{(i)} \mathbf{L} \mathbf{S}^{(i)}, \mathbf{I}), & [\text{vec}(\mathbf{L}) | \alpha_1] &\sim \mathcal{N}(\mathbf{0}, \frac{1}{\alpha_1} \mathbf{I}), \\ [\mathbf{S}^{(i)} | T\mathbf{S}^{\dagger(i)}, \alpha_2] &\sim \mathcal{N}(T\mathbf{S}^{\dagger(i)}, \frac{1}{\alpha_2} \mathbf{I}), & [\mathbf{S}^\dagger | \mathbf{U}, \alpha_3] &\sim g, \quad \mathbf{U} \sim h, \end{aligned}$$

where

$$g \propto \exp(-\alpha_3 \cdot \langle \Delta(\mathbf{S}^\dagger), \mathbf{U} \rangle) \cdot \delta(\mathbf{S}^\dagger \in \Pi(\mathbf{a}, \mathbf{b})), \quad h \propto \delta(\mathbf{U} \in \mathcal{M}).$$

Here  $g$  specifies an exponential distribution restricted on the set  $\Pi(\mathbf{a}, \mathbf{b})$ , and  $h$  specifies a uniform distribution on the set  $\mathcal{M}$ . With this process, our objective function is equivalent to a Maximum A Posterior (MAP) formulation in the following sense:

$$\mathcal{J} = -\log(\mathbb{P}(\mathbf{L}, \mathbf{S}, \mathbf{S}^\dagger, \mathbf{U} | \mathbf{X}, \mathbf{Y}, \alpha_1, \alpha_2, \alpha_3)) + \text{const.}$$

This fact gives us an alternative perspective on the relationship between  $\mathbf{S}$  and  $\mathbf{S}^\dagger$ . With the relaxation scheme, the constraints are moved to the mean of the prior distribution of  $\mathbf{S}$ . This provides  $\mathbf{S}$  with a possibility to activate the overlapping off-diagonal block elements with a moderate variance  $\alpha_2$ .

## 2.2 Optimization

It is easy to see that  $\mathcal{J}$  in (Obj) is not a jointly convex function. But fortunately, it is easy to show that the four subproblems with respect to  $\mathbf{L}, \mathbf{S}, \mathbf{S}^\dagger, \mathbf{U}$  are all convex. Instead of directly solving the overall non-convex problem, this fact motivates us to adopt an alternating optimization scheme where only one of the four parameters is updated each time and the others are fixed as constants. Now we elaborate the four subproblems, respectively.

**L and S subroutine:** Theoretically, both subroutines solve a strongly convex unconstrained quadratic programming and enjoy a closed-form solution. However, calculating the closed form of the  $\mathbf{L}$  subproblem suffers from a heavy computational complexity. Instead of adopting the closed form directly for the  $\mathbf{L}$  subproblem, we adopt a gradient-based optimizer in our paper. Please see Appendix C for more details.

**U subroutine:** According to Thm.2,  $\mathbf{U}$  could be solved from:  $\mathbf{U} = \mathbf{V}_K \mathbf{V}_K^\top$ , where  $\mathbf{V}_K$  denotes eigenvectors associated with the smallest  $K$  eigenvalues of  $\Delta(\mathbf{S}^\dagger)$ . Denote  $\mathbf{V}_K = [\mathbf{f}_1, \dots, \mathbf{f}_{k+T}]^\top$ , according to Thm.1, when  $\chi_{\mathbf{S}^\dagger} = K$ , up to some orthogonal transformation,  $\mathbf{f}_i \in \mathbb{R}^{K \times 1}$  becomes an indicator vector with only one non-zero entry where  $[\mathbf{f}_i]_j = 1$  only if the corresponding latent/output task  $i$  is in group  $\iota_{\mathcal{A}(j)}$ . In this way, we see that  $\mathbf{f}_i$  is a strong group indicator. Consequently, we name  $\mathbf{f}_i$  as the *embedding vector* for the latent (output) task.

**S<sup>†</sup> subroutine:** With  $\mathbf{U}$  updated with  $\mathbf{U} = \mathbf{V}_K \mathbf{V}_K^\top$ , the following proposition shows a way to solve this subproblem:

**Proposition 1.** *The  $\mathbf{S}^\dagger$  subproblem could be reformulated as:*

$$\min_{\mathbf{S}^\dagger \in \Pi(\mathbf{a}, \mathbf{b})} \frac{\vartheta}{2} \|\mathbf{S}^\dagger - \bar{\mathbf{S}}\|_F^2 + \langle \mathcal{D}, \mathbf{S}^\dagger \rangle, \quad (\text{Primal})$$

where  $\vartheta = \frac{\alpha_2 \cdot T^2}{\alpha_3}$ ,  $\bar{\mathbf{S}} = \frac{\mathbf{S}}{T}$  and  $\mathcal{D}_{ij} = \|\mathbf{f}_i - \mathbf{f}_{k+j}\|^2$ .

The proof could be found in Appendix A.1. From Prop.1, we see that the subproblem recovers a smoothed Optimal Transport (OT) [Peyré et al., 2019] problem. More specifically, the calculation of the Wasserstein-2 distance between  $\mathbf{l}$  and  $\mathbf{o}$ , based on the spectral embedding.

Similar to the recent results [Blondel et al., 2018, Peyré et al., 2019], we can show that the regularized OT problem has a close connection with the original OT problem.

**Proposition 2.** *The following properties hold true:*

(a) Denote  $\mathbf{S}^\dagger_{\vartheta_0}$  as the solution of problem (Primal) when  $\vartheta = \vartheta_0$ . Then we have :

$$\mathbf{S}^\dagger_{\vartheta_0} \xrightarrow{\vartheta_0 \rightarrow 0} \operatorname{argmin}_{\mathbf{S}^\dagger} \left\{ d(\mathbf{S}^\dagger, \bar{\mathbf{S}}) : \mathbf{S}^\dagger \in \operatorname{argmin}_{\mathbf{S}^\dagger \in \Pi(\mathbf{a}, \mathbf{b})} \langle \mathcal{D}, \mathbf{S}^\dagger \rangle \right\}.$$

(b) Denote

$$\mathcal{J}_{OT} = \min_{\mathbf{S}^\dagger \in \Pi(\mathbf{a}, \mathbf{b})} \langle \mathcal{D}, \mathbf{S}^\dagger \rangle, \quad \mathcal{J}_{REG} = \min_{\mathbf{S}^\dagger \in \Pi(\mathbf{a}, \mathbf{b})} (\vartheta/2) \cdot d(\mathbf{S}, \bar{\mathbf{S}}) + \langle \mathcal{D}, \mathbf{S}^\dagger \rangle,$$

we have:

$$\vartheta \cdot \max \{ d(\bar{\mathbf{S}}\mathbf{1} - \mathbf{a})/T, d(\bar{\mathbf{S}}^\top \mathbf{1}, \mathbf{b})/k \} \leq \mathcal{J}_{REG} - \mathcal{J}_{OT} \leq \vartheta \cdot (\min\{\|\mathbf{a}\|_2^2, \|\mathbf{b}\|_2^2\} + \|\bar{\mathbf{S}}\|_F^2).$$

The proof can be found in Appendix A.2. Prop.2-(a) shows that asymptotically, when  $\vartheta \rightarrow 0$ , the solution of the regularized OT problem approaches a specific solution of the original OT problem. More specifically, it will pick out an optimal coupling from the OT solution set with the smallest regularization term  $d(\mathbf{S}^\dagger, \bar{\mathbf{S}})$ . From a non-asymptotic perspective, Prop.2-(b) shows how fast this approximation will take place. Consequently, we will get a reasonable approximation of the original OT problem with a small  $\vartheta$ . Moreover, if the regularizer  $\langle \Delta(\mathbf{S}^\dagger), \mathbf{U} \rangle$  is sufficiently small,  $\mathbf{f}$  approaches the grouping indicator of a K-connected bipartite graph. At the same time,  $\mathcal{D}_{ij}$ , the distance between the embedding vectors, approaches zero when  $\mathbf{l}_i$  and  $\mathbf{o}_j$  belong to the same group indicated by  $\mathbf{f}$ . Under this circumstance, the transportation cost  $\mathcal{D}_{i,j}$  is small only if  $\mathbf{l}_i$  and  $\mathbf{o}_j$  belong to the same group. By contrast, the inter-group negative transfer is suppressed with a large transportation cost. Moreover, with  $\alpha_2 \rightarrow +\infty$ ,  $\mathbf{L}\mathbf{S}^{(i)} \rightarrow \mathbb{E}_{\mathbf{l}|\mathbf{o}=i}(\mathbf{L})$ . This indicates that the conditional expectation also embraces the idea of barycenter projection mapping [Seguy et al., 2018] in the sense  $\mathbb{E}_{\mathbf{l}|\mathbf{o}=i}(\mathbf{L}) = \operatorname{argmin}_{\mathbf{z}} \mathbb{E}_{\mathbf{l}|\mathbf{o}=i}(d(\mathbf{L}^{(i)}, \mathbf{z}))$ . Under this condition, the task parameter of  $\mathbf{o}_i$  becomes a barycenter of the latent task embeddings. Finally, we show that this subproblem could be solved efficiently from the dual formulation.

**Proposition 3.** *The dual problem of (Primal) could be solved from:*

$$(\mathbf{h}^*, \mathbf{g}^*) = \operatorname{argmin}_{\mathbf{h}, \mathbf{g}} \frac{1}{2\vartheta} \cdot \|(\mathbf{h} \oplus \mathbf{g} - \mathcal{D} + \vartheta \bar{\mathbf{S}})_+\|_F^2 - \langle \mathbf{h}, \mathbf{a} \rangle - \langle \mathbf{g}, \mathbf{b} \rangle, \quad (\text{Dual})$$

and the primal solution is given by  $\mathbf{S}^{\dagger*} = \left[ \frac{\mathbf{h}^* \oplus \mathbf{g}^* - \mathcal{D}}{\vartheta} + \bar{\mathbf{S}} \right]_+$ .

The proof can be found in Appendix A.3. From Prop.3, we can recover the primal solution from (Dual), which only involves  $\mathcal{O}(k + T)$  parameters instead of  $\mathcal{O}(kT)$ . In this spirit, we first solve  $\mathbf{h}^*, \mathbf{g}^*$  from (Dual) with the L-BFGS [Zhu et al., 1997] method as the optimizer, and then recover  $\mathbf{S}^{\dagger*}$  from the dual parameters.

**Summary** Our optimization procedure then alternatively solves the four subproblems until a convergence condition is reached, with irrelevant variables fixed as their latest version. Moreover, since all the subproblems are convex, it is easy to see that the iteration over subproblems then keeps the overall loss function non-increasing.

### 2.3 Theoretical Analysis

In this section, we present theoretical analysis shedding light on how the hyperparameters  $\alpha_1, \alpha_2, \alpha_3$  affect our proposed model. Let us start with defining a proper hypothesis space  $\mathcal{H}$  that covers the solution returned by the optimization algorithm. Recall that (Obj) is non-increasing during the



optimization procedure. This means that if we choose a feasible candidate  $\mathbf{L}_0, \mathbf{S}_0, \mathbf{S}_0^\dagger, \mathbf{U}_0$  as the initialization of the algorithm, denote by  $\mathcal{J}_0$  the corresponding objective function value, we will have  $\|\mathbf{L}\|_F^2 \leq 2\mathcal{J}_0/\alpha_1, d(\mathbf{S}, T\mathbf{S}^\dagger) \leq 2\mathcal{J}_0/\alpha_2, \langle \Delta(\mathbf{S}^\dagger), \mathbf{U} \rangle \leq 2\mathcal{J}_0/\alpha_3$ , for all outputs from the optimization algorithm. This naturally defines a hypothesis class  $\mathcal{H} = \mathcal{H}(\mathbf{L}, \mathbf{S}, \mathbf{S}^\dagger, \mathbf{U})$ :

$$\mathcal{H}(\mathbf{L}, \mathbf{S}, \mathbf{S}^\dagger, \mathbf{U}) = \left\{ \left\{ \hat{Y}^{(i)}(\mathbf{X}_i^{(t)}) = (\mathbf{L}\mathbf{S}^{(i)})^\top \mathbf{X}_i^{(t)} \right\}_{ti} : \|\mathbf{L}\|_F^2 \leq \xi_1, \right. \\ \left. d(\mathbf{S}, T\mathbf{S}^\dagger) \leq \xi_2, \langle \Delta(\mathbf{S}^\dagger), \mathbf{U} \rangle \leq \xi_3, \mathbf{S}^\dagger \in \Pi(\mathbf{a}, \mathbf{b}), \mathbf{U} \in \mathcal{M} \right\},$$

where  $\xi_1 = 2\mathcal{J}_0/\alpha_1, \xi_2 = 2\mathcal{J}_0/\alpha_2, \xi_3 = 2\mathcal{J}_0/\alpha_3$ . Now we are ready to represent the theoretical results based on  $\mathcal{H}$ .

As the first step, we explore how well a model learned from  $\mathcal{H}$  generalizes to the overall population. The empirical risk  $\hat{\mathcal{R}}(\mathbf{L}, \mathbf{S})$  over the observed dataset and the task-averaged risk  $\mathcal{R}(\mathbf{L}, \mathbf{S})$  are given as:  $\hat{\mathcal{R}}(\mathbf{L}, \mathbf{S}) = \sum_{i=1}^T \mathcal{J}^{(i)}/T, \mathcal{R}(\mathbf{L}, \mathbf{S}) = \sum_{i=1}^T \mathbb{E}_{\mu_i} [\mathcal{J}^{(i)}]/T$ , where the training data  $\mathbf{X}_j^{(i)}, \mathbf{Y}_j^{(i)}$  are sampled from  $\mu_i$ . We then bound the term  $\Delta = \mathcal{R}(\mathbf{L}, \mathbf{S}) - \hat{\mathcal{R}}(\mathbf{L}, \mathbf{S})$ . Following the spirit of [Maurer et al., 2016], we have the following bound for the hypothesis space:

**Theorem 3.** Suppose that  $n_1 = n_2 \cdots = n_T = n$ , the loss function  $\ell(y, \cdot) : \hat{y} \mapsto [0, M]$ ,  $\ell(y, \cdot)$  is  $M\phi$ -Lipschitz continuous, and  $\forall (\mathbf{L}, \mathbf{S})$  are chosen from  $\mathcal{H}$ , the following bound holds with possibility at least  $1 - \delta$ :

$$\frac{\Delta}{M} \leq \kappa_1 \phi \aleph \left( \frac{\xi_1 k \|\mathbf{COV}(\mathbf{X})\|_1}{nT} \right)^{1/2} + 2\kappa_2 \phi \aleph \cdot \left( \frac{\xi_1 \|\mathbf{COV}(\mathbf{X})\|_\infty}{n} \right)^{1/2} + \left( \frac{9 \ln(2/\delta)}{2nT} \right)^{1/2},$$

where  $\kappa_1$  and  $\kappa_2$  are two universal constants,  $\aleph = \sqrt{\xi_2} + 1$ .  $\mathbf{COV}(\mathbf{X})$  is the covariance operator defined as  $\langle \mathbf{COV}(\mathbf{X})\mathbf{u}, \mathbf{v} \rangle = (1/nT) \cdot \sum_{ti} \langle \mathbf{u}, \mathbf{X}_i^{(t)} \rangle \langle \mathbf{X}_i^{(t)}, \mathbf{v} \rangle$ .

The proof can be found in Appendix B.3. With the sample complexity given, we narrow our focus to the problem that how  $\xi_2, \xi_3$  benefit the hypothesis space. The following theorem shows that  $\xi_2, \xi_3$  control the spectral properties of  $\mathbf{S}$  and  $\mathbf{S}^\dagger$ .

**Theorem 4** (Spectral Properties of  $\mathbf{S}$ ). Let  $k \leq T$ , define the SVD of  $\mathbf{S}$  as  $\mathbf{S} = \mathbf{P}\mathbf{\Lambda}\mathbf{Q}^\top$ , where  $\mathbf{P} = [\mathbf{p}_1, \cdots, \mathbf{p}_k], \mathbf{Q} = [\mathbf{q}_1, \cdots, \mathbf{q}_k]$  are left and right singular vectors respectively,  $\mathbf{\Lambda} = \text{diag}(\sigma_i(\mathbf{S}))$  with  $\sigma_1(\mathbf{S}) \geq \sigma_2(\mathbf{S}) \geq \cdots \geq \sigma_k(\mathbf{S}) \geq 0$ . The following properties hold for all  $\mathbf{S} \in \mathcal{H}$ :

- (a) The bottom  $K$  eigenvalues of the graph Laplacian induced by  $\mathbf{S}$  is bounded by :  $\sum_{i=N-K+1}^N \lambda_i(\Delta(\mathbf{S})) \leq T\xi_3 + \sqrt{\xi_2 K}(\sqrt{2} + \sqrt{k} + \sqrt{T})$ , where  $\Delta(\mathbf{S})$  is obtained from replacing  $\mathbf{S}^\dagger$  in  $\Delta(\mathbf{S}^\dagger)$  with  $\mathbf{S}$ .
- (b) Define  $\mathbb{M}(\mathbf{P}) = \text{Span}\{\mathbf{p}_1, \cdots, \mathbf{p}_K\}$  and  $\mathbb{M}(\mathbf{Q}) = \text{Span}\{\mathbf{q}_1, \cdots, \mathbf{q}_K\}$ , if  $\xi_3 + \sqrt{\xi_2}/T < 1/T$  and  $\text{rank}(\mathbf{S}) \geq K$ , then we have:

$$\frac{\sigma_1(\mathbf{S})}{\sigma_K(\mathbf{S})} = \max_{\substack{\mathbf{x}, \mathbf{y} \in \mathbb{M}(\mathbf{P}) \\ \|\mathbf{x}\|_2=1, \|\mathbf{y}\|_2=1}} \frac{\|\mathbf{S}\mathbf{x}\|_2}{\|\mathbf{S}\mathbf{y}\|_2} = \max_{\substack{\mathbf{x}, \mathbf{y} \in \mathbb{M}(\mathbf{Q}) \\ \|\mathbf{x}\|_2=1, \|\mathbf{y}\|_2=1}} \frac{\|\mathbf{S}^\top \mathbf{x}\|_2}{\|\mathbf{S}^\top \mathbf{y}\|_2} \leq \frac{1}{k} \cdot \frac{T + k\sqrt{\xi_2}}{1 - T\xi_3 - \sqrt{\xi_2}}.$$

The proof can be found in Appendix B.4. Thm.4.(a) implies that, with a small  $\xi_2$  and  $\xi_3$ , the grouping structure of  $\mathbf{S}$  could also be controlled, even though the structural penalty is not directly exhibited on  $\mathbf{S}$ . More specifically, if we pick  $\xi_3 = O(T^{-3/2})$  and  $\xi_2 = O(1/T^2)$ , we can reach a small  $\sum_{i=N-K+1}^N \lambda_i(\Delta(\mathbf{S}))$  with  $O(T^{-1/2})$  if  $T \gg k$ . Thm.4.(b) states that shrinking  $\xi_2, \xi_3$  helps to remain a smaller numerical perturbation of  $\mathbf{S}\mathbf{x}$  ( $\mathbf{S}^\top \mathbf{x}$ ) over the principle subspaces, i.e., the subspaces spanned by principle left/right singular vectors. Besides numerical benefits, the following theorem shows that  $\xi_3$  guarantees good structure recovery in a non-asymptotic manner.

**Theorem 5.** Assume that  $k \leq T$ , and that the ground-truth grouping is indicated by  $\mathcal{G} = \{(i, j) : 1 \leq i \leq k, 1 \leq j \leq T, \mathbf{l}_i \text{ and } \mathbf{o}_j \text{ are in the same group}\}$  with  $K$  disjoint groups. Moreover, for a matrix  $\mathbf{W}$ , denote  $\text{supp}(\mathbf{W})$  as  $\{(i, j) : W_{i,j} \neq 0\}$ . For all  $\mathbf{S}^\dagger$  obtained from the space  $\mathcal{H}$  such

that  $\lambda_{K+1}(\Delta(\mathbf{S}^\dagger)) > \lambda_K(\Delta(\mathbf{S}^\dagger)) > 0$  and  $\inf_{\mathbf{S}^* \in \Pi(\mathbf{a}, \mathbf{b}), \text{Supp}(\mathbf{S}^*) = \mathcal{G}} \|\Delta(\mathbf{S}^\dagger) - \Delta(\mathbf{S}^*)\|_F \leq \epsilon$ , we have:

$$\|\mathbf{S}^{\dagger \text{supp}^c}\|_1 \leq \frac{1}{2} \cdot \left( \xi_3 + \frac{\sqrt{\frac{2}{k} + \frac{6}{T}} \cdot \epsilon}{\lambda_{K+1}(\Delta(\mathbf{S}^\dagger))} \right) \leq \frac{1}{2} \cdot \xi_3 + \frac{4}{\sqrt{kT} \lambda_{K+1}(\Delta(\mathbf{S}^\dagger))}$$

$\mathbf{S}^{\dagger \text{supp}^c}$  denotes the projection of  $\mathbf{S}^\dagger$  onto the complement of the support set of the expected block-diagonal structure in the sense that  $\mathbf{S}_{i,j}^{\dagger \text{supp}^c} = 0$  if  $i$  and  $j$  belong to the same group,  $\mathbf{S}_{i,j}^{\dagger \text{supp}^c} = \mathbf{S}_{i,j}^\dagger$  otherwise.

The proof can be found in Appendix B.5. Under the assumptions of Thm.5, a smaller  $\xi_3$  embraces a better recovery of the true block-diagonal structure. More specifically, it shrinks  $\|\mathbf{S}^{\dagger \text{supp}^c}\|_1$ , i.e., the overall magnitude of the elements that violate the true grouping structure. Picking  $\xi_3 = O(T^{-1/4})$  and  $k = O(\sqrt{T})$ , if  $\lambda_{K+1}(\Delta(\mathbf{S}^\dagger)) = O(1/k)$ , under the worst case, we have  $\|\mathbf{S}^{\dagger \text{supp}^c}\|_1 = O(T^{-1/4})$ .

### 3 Empirical Study

#### 3.1 Experiment Settings

For all the experiments, hyper-parameters are tuned based on the training and validation set, and the results on the test set are recorded. The experiments are done with 5 repetitions for each involved algorithm. Except for the Simulated Dataset, the train/valid/test ratio is fixed as 70%/15%/15%. For regression datasets (Simulated dataset and School), we adopt the overall rmse on all samples as the evaluation metric. For classification datasets, we adopt the average of task-wise AUC as the evaluation metric. For regression problem,  $\mathcal{J}(\cdot)$  in GBDSP is chosen as the square-loss. For classification problem,  $\mathcal{J}(\cdot)$  in GBDSP is chosen as the squared surrogate loss for AUC [Gao et al., 2016]. All the experiments are run with MATLAB 2016b and a Ubuntu 16.04 system. In the next subsection, we show our experimental results on a simulated dataset. More experiments for real-world datasets could be found in Appendix D.

#### 3.2 Simulated Dataset

To test the effectiveness of GBDSP we generate a simple simulated annotation dataset with  $T = 150$  simulated tasks, where the dataset is produced according to the assumption in our model. For each task, 500 samples are generated with  $d = 300$  features such that  $\mathbf{X}^{(i)} \in \mathbb{R}^{500 \times 300}$  and  $\mathbf{x}_k^{(i)} \sim \mathcal{N}(0, \mathbf{I}_{80})$ . Specifically, we generate latent task representations with  $k = 100$  basis. This yields an  $\mathbf{L} \in \mathbb{R}^{300 \times 100}$  and an  $\mathbf{S} \in \mathbb{R}^{100 \times 150}$ . To leverage the group structure, we split the latent tasks and output tasks into 5 groups, in a way that  $\mathbf{L} = [\mathbf{L}_1, \dots, \mathbf{L}_5]$ , where  $\mathbf{L}_1 \in \mathbb{R}^{300 \times 20}$ ,  $\mathbf{L}_2 \in \mathbb{R}^{300 \times 20}$ ,  $\mathbf{L}_3 \in \mathbb{R}^{300 \times 10}$ ,  $\mathbf{L}_4 \in \mathbb{R}^{300 \times 30}$ ,  $\mathbf{L}_5 \in \mathbb{R}^{300 \times 20}$ , and that  $\mathbf{S} = \bigoplus_{i=1}^5 \mathbf{S}_i$  where  $\mathbf{S}_1 \in \mathbb{R}^{20 \times 30}$ ,  $\mathbf{S}_2 \in \mathbb{R}^{20 \times 30}$ ,  $\mathbf{S}_3 \in \mathbb{R}^{10 \times 15}$ ,  $\mathbf{S}_4 \in \mathbb{R}^{30 \times 45}$ ,  $\mathbf{S}_5 \in \mathbb{R}^{20 \times 30}$ . For the  $i$ -th group, the elements in  $\mathbf{L}_i$  is sampled i.i.d from  $\mathcal{N}(m_i, 0.01)$ , where  $m_i = 5i$ .  $\mathbf{S}_i$  is generated as  $\mathbf{S}_i = s_i \mathbf{1}$ , i.e., every element in  $\mathbf{S}_i$  shares the same value. Moreover,  $s_i$  is calculated from the constraint that  $\mathbf{S} \in \Pi(\mathbf{a}, \mathbf{b})$ . Then the task parameter is generated as  $\mathbf{W} = \mathbf{L}\mathbf{S}$ . For each task, the outputs are generated as  $\mathbf{Y}^{(i)} = \mathbf{X}^{(i)}(\mathbf{W}^{(i)} + \epsilon^{(i)})$ , where  $\epsilon^{(i)} \in \mathbb{R}^{200 \times 1}$ , and  $\epsilon^{(i)} \sim \mathcal{N}(0, 0.1^2 \mathbf{I}_{500})$ . Based on this setting, we compare GBDSP with GOMTL in the simulation dataset to see how the block-diagonal structure benefits the latent task representation based MTL.

First, we show how well could GOMTL and GBDSP recover the block-diagonal structure. We compare  $\mathbf{S}$  obtained from GOMTL and  $\mathbf{S}^\dagger$  obtained from GBDSP, with the initial value of  $\hat{\mathbf{L}}$  set as  $\hat{\mathbf{L}} = \mathbf{L} + \mathcal{N}(0, 0.05 \mathbf{I})$ . As shown in Fig.1(a)- Fig.1(c), GBDSP recovers a much clearer structure than GOMTL. Moreover, we provide a closer look at the embedding vectors in GBDSP. To do this, we visualize the spectral embeddings  $\mathbf{f}$  in a 3d space with t-SNE [Maaten and Hinton, 2008], which is shown in Fig.2(c). In this figure, the points with different colors represent latent/output tasks in different groups. Clearly, we see that the clusters are well-separated in the spectral embedding space, which again verifies the grouping power of the proposed method.

Next, we check whether GBDSP could improve the performance with a structural LATM. In Fig.2, we plot the performance of GOMTL and GBDSP with different training set ratio



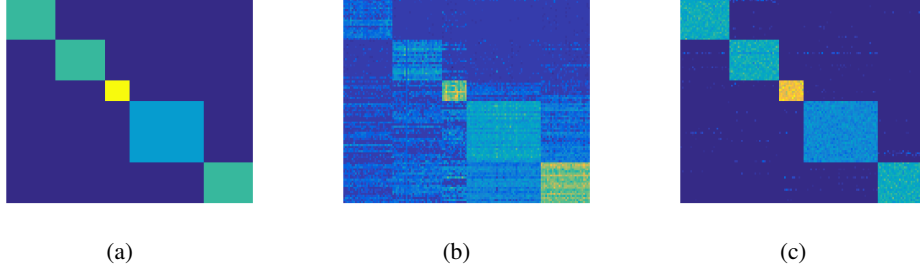


Figure 1: Visualizations over the Simulated Dataset. (a)-(c) provide structural comparisons over the LATM: (a) shows The true LATM; (b) shows the LATM recovered by GOMTL (c) shows the LATM recovered by **GBDSP**.

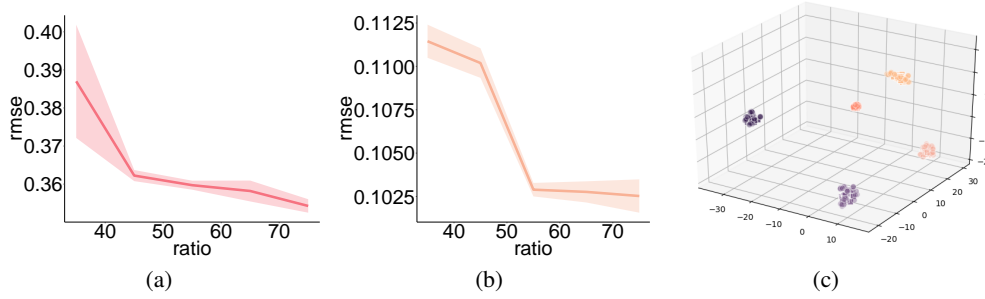


Figure 2: (a-b) Performance curve with different training data ratio: (a) GOMTL (b) **GBDSP** (c) shows the spectral group embedding of  $\mathbf{o}$  and  $\mathbf{l}$  in **GBDSP**.

(0.35, 0.45, 0.55, 0.65, 0.75). The corresponding results show that **GBDSP** consistently provides a better performance with a smaller variance, which supports the idea that learning a block-diagonal LATM structure improves the performance.

## 4 Conclusion

To simultaneously leverage a latent task representation and alleviate the inter-group negative transfer issue, we develop a novel MTL method **GBDSP**, which simultaneously separates the latent tasks and out tasks into a given number of groups. Moreover, we adopt an optimization method to solve the model parameters, which gives an alternative update scheme for our multi-convex objective function. The solution produced by the optimization method shows a close connection between our method and the optimal transport problem, which brings new insight into how negative transfer could be prevented across latent tasks and output tasks. Furthermore, we provide theoretical analysis on the spectral properties of the model parameters. Empirical results on the simulated dataset show that **GBDSP** could roughly recover the correct grouping structure with good performance, and results on the real-world datasets further verify the effectiveness of our proposed model on the problem of personalized attribute prediction.

## 5 Acknowledgements

This work was supported in part by National Natural Science Foundation of China: 61620106009, U1636214, 61836002, 61861166002, 61672514 and 61976202, in part by National Basic Research Program of China (973 Program): 2015CB351800, in part by Key Research Program of Frontier Sciences, CAS: QYZDJ-SSW-SYS013, in part by the Strategic Priority Research Program of Chinese Academy of Sciences, Grant No. XDB28000000, in part by the Science and Technology Development Fund of Macau SAR (File no. 0001/2018/AFJ) Joint Scientific Research Project, in part by Beijing Natural Science Foundation (No. 61971016, L182057, and 4182079), in part by Peng Cheng

Laboratory Project of Guangdong Province PCL2018KP004, and in part by Youth Innovation Promotion Association CAS.

## References

- F. Alizadeh. Interior point methods in semidefinite programming with applications to combinatorial optimization. *SIAM journal on Optimization*, 5(1):13–51, 1995.
- G. P. and James Hays. SUN attribute database: Discovering, annotating, and recognizing scene attributes. In *CVPR*, pages 2751–2758, 2012.
- A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008a.
- A. Argyriou, M. Pontil, Y. Ying, and C. A. Micchelli. A spectral regularization framework for multi-task structure learning. In *Neurips*, pages 25–32, 2008b.
- A. Barzilai and K. Crammer. Convex multi-task learning by clustering. In *Artificial Intelligence and Statistics*, pages 65–73, 2015.
- M. Blondel, V. Seguy, and A. Rolet. Smooth and sparse optimal transport. In *AISTATS*, pages 880–889, 2018.
- J. Cao, Y. Li, and Z. Zhang. Partially shared multi-task convolutional neural network with local constraint for face attribute learning. In *CVPR*, pages 4290–4299, 2018.
- R. Caruana. Multitask learnings. *Machine Learning*, 28(1):41–75, 1997.
- K. Crammer and Y. Mansour. Learning multiple tasks using shared hypotheses. In *Neurips*, pages 1475–1483, 2012.
- J. Frecon, S. Salzo, and M. Pontil. Bilevel learning of the group lasso structure. In *Neurips*, pages 8301–8311, 2018.
- W. Gao, L. Wang, R. Jin, S. Zhu, and Z. Zhou. One-pass AUC optimization. *AI*, 236:1–29, 2016.
- G. H. Golub and C. F. Van Loan. *Matrix computations*, volume 3. JHU press, 2012.
- L. Han and Y. Zhang. Multi-stage multi-task learning with reduced rank. In *AAAI*, pages 1638–1644, 2016.
- T. Heskes. Solving a huge number of similar tasks: A combination of multi-task learning and a hierarchical bayesian approach. In *ICML*, pages 233–241, 1998.
- R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge university press, 2012.
- L. Jacob, J.-p. Vert, and F. R. Bach. Clustered multi-task learning: A convex formulation. In *Neurips*, pages 745–752, 2009.
- J.-Y. Jeong and C.-H. Jun. Variable selection and task grouping for multi-task learning. In *KDD*, pages 1589–1598, 2018.
- Y. Jiang, Z. Yang, Q. Xu, X. Cao, and Q. Huang. When to learn what: Deep cognitive subspace clustering. In *ACM MM*, pages 718–726, 2018.
- Y. Jiang, Q. Xu, Z. Yang, X. Cao, and Q. Huang. Duet robust deep subspace clustering. In *ACM MM*, 2019.
- Z. Kang, K. Grauman, and F. Sha. Learning with whom to share in multi-task feature learning. In *ICML*, pages 521–528, 2011.
- A. Kovashka and K. Grauman. Discovering attribute shades of meaning with the crowd. *IJCV*, 114(1):56–73, 2015.

- A. Kovashka, D. Parikh, and K. Grauman. Whittlesearch: Image search with relative attribute feedback. In *CVPR*, pages 2973–2980, 2012.
- M. Kshirsagar, E. Yang, and A. C. Lozano. Learning task clusters via sparsity grouped multitask learning. In *ECML PKDD*, pages 673–689, 2017.
- A. Kumar and H. D. III. Learning task grouping and overlap in multi-task learning. In *ICML*, pages 1723–1730, 2012.
- G. Lee, E. Yang, and S. Hwang. Asymmetric multi-task learning based on task relatedness and loss. In *ICML*, pages 230–238, 2016.
- H. Lee, E. Yang, and S. J. Hwang. Deep asymmetric multi-task feature learning. In *ICML*, pages 2962–2970, 2018.
- Y. Li, K. Fu, Z. Wang, C. Shahabi, J. Ye, and Y. Liu. Multi-task representation learning for travel time estimation. In *KDD*, pages 1695–1704, 2018.
- Y. Lin, S. Yang, V. Stoyanov, and H. Ji. A multi-lingual multi-task architecture for low-resource sequence labeling. In *ACL*, pages 799–809, 2018.
- P. Liu, X. Qiu, and X. Huang. Adversarial multi-task learning for text classification. In *ACL*, pages 1–10, 2017.
- S. Liu and S. J. Pan. Adaptive group sparse multi-task learning via trace lasso. In *IJCAI*, pages 2358–2364, 2017.
- C. Lu, J. Feng, Z. Lin, T. Mei, and S. Yan. Subspace clustering by block diagonal representation. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):487–501, 2019.
- L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (Nov):2579–2605, 2008.
- A. Maurer, M. Pontil, and B. Romera-Paredes. Sparse coding for multitask and transfer learning. In *ICML*, pages 343–351, 2013.
- A. Maurer, M. Pontil, and B. Romera-Paredes. The benefit of multitask representation learning. *The Journal of Machine Learning Research*, 17(1):2853–2884, 2016.
- A. M. McDonald, M. Pontil, and D. Stamos. Spectral k-support norm regularization. In *Neurips*, pages 3644–3652, 2014.
- F. Nie, Z. Hu, and X. Li. Calibrated multi-task learning. In *KDD*, pages 2012–2021, 2018.
- M. L. Overton and R. S. Womersley. On the sum of the largest eigenvalues of a symmetric matrix. *SIAM Journal on Matrix Analysis and Applications*, 13(1):41–45, 1992a.
- M. L. Overton and R. S. Womersley. On the sum of the largest eigenvalues of a symmetric matrix. *SIAM Journal on Matrix Analysis and Applications*, 13(1):41–45, 1992b.
- G. Peyré, M. Cuturi, et al. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.
- V. Seguy, B. B. Damodaran, R. Flamary, N. Courty, A. Rolet, and M. Blondel. Large-scale optimal transport and mapping estimation. In *(ICLR)*, 2018.
- C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016.
- U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.
- Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

- D. Xu, W. Ouyang, X. Wang, and N. Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *CVPR*, pages 675–684, 2018.
- L. Xu, A. Huang, J. Chen, and E. Chen. Exploiting task-feature co-clusters in multi-task learning. In *AAAI*, pages 1931–1937, 2015.
- Z. Yang, Q. Xu, X. Cao, and Q. Huang. From common to special: When multi-attribute learning meets personalized opinions. In *AAAI*, pages 515–522, 2018.
- Z. Yang, Q. Xu, X. Cao, and Q. Huang. Learning personalized attribute preference via multi-task AUC optimization. In *AAAI*, pages 5660–5667, 2019a.
- Z. Yang, Q. Xu, W. Zhang, X. Cao, and Q. Huang. Split multiplicative multi-view subspace clustering. *IEEE Transactions on Image Processing*, 2019b.
- Z. Yin and Y. Shen. On the dimensionality of word embedding. In *NIPS*, pages 887–898, 2018.
- Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2014.
- W. Zhong and J. T. Kwok. Convex multitask learning with flexible task clusters. In *ICML*, pages 483–490, 2012.
- J. Zhou, J. Chen, and J. Ye. Clustered multi-task learning via alternating structure optimization. In *Neurips*, pages 702–710, 2011a.
- J. Zhou, J. Chen, and J. Ye. Malsar: Multi-task learning via structural regularization. *Arizona State University*, 21, 2011b.
- C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, 23(4):550–560, 1997.

## Appendix

### A Proofs of the Propositions

#### A.1 Proof of Proposition 1

*Proof.* Given the solution of the  $\mathbf{U}$  subproblem, the  $\mathbf{S}^\dagger$  subproblem could be formulated as:

$$\min_{\mathbf{S}^\dagger \in \Pi(\mathbf{a}, \mathbf{b})} \frac{\alpha_2}{2} \|\mathbf{S} - \mathbf{S}_c^\dagger\|_F^2 + \alpha_3 \cdot \left\langle \text{diag} \left( \begin{bmatrix} 0 & \mathbf{S}^\dagger \\ \mathbf{S}^{\dagger\top} & 0 \end{bmatrix} \mathbf{1} \right) - \begin{bmatrix} 0 & \mathbf{S}^\dagger \\ \mathbf{S}^{\dagger\top} & 0 \end{bmatrix}, \mathbf{U} \right\rangle.$$

With the fact that

$$\left\langle \text{diag} \left( \begin{bmatrix} 0 & \mathbf{S}^\dagger \\ \mathbf{S}^{\dagger\top} & 0 \end{bmatrix} \mathbf{1} \right) - \begin{bmatrix} 0 & \mathbf{S}^\dagger \\ \mathbf{S}^{\dagger\top} & 0 \end{bmatrix}, \mathbf{U} \right\rangle = \left\langle \text{diag}(\mathbf{U}) \mathbf{1}^\top - \mathbf{U}, \begin{bmatrix} 0 & \mathbf{S}^\dagger \\ \mathbf{S}^{\dagger\top} & 0 \end{bmatrix} \right\rangle,$$

and simple scaling of the constants, we could reformulate the problem as:

$$\min_{\mathbf{S}^\dagger \in \Pi(\mathbf{a}, \mathbf{b})} \frac{\vartheta}{2} \|\bar{\mathbf{S}} - \mathbf{S}^\dagger\|_F^2 + \left\langle \Delta^{(1)} + \Delta^{(2)\top}, \mathbf{S}^\dagger \right\rangle.$$

where  $\Delta = \text{diag}(\mathbf{U}) \mathbf{1}^\top - \mathbf{U}$ ,  $\Delta^{(1)} = \Delta(1 : k, (k+1) : \text{end})$ ,  $\Delta^{(2)} = \Delta((k+1) : \text{end}, 1 : k)$ .

Then the proof directly follows the fact that

$$\Delta_{ij}^{(1)} + \Delta_{ji}^{(2)} = U_{ii} + U_{k+j, k+j} - U_{i, k+j} - U_{k+j, i} = \|\mathbf{f}_i - \mathbf{f}_{k+j}\|_2^2.$$

□

#### A.2 Proof of Proposition 2

*Proof. (a):* Since  $\Pi(\mathbf{a}, \mathbf{b})$  is bounded, the sequence  $\{\mathbf{S}^\dagger_{\vartheta}\}_{\vartheta \rightarrow 0}$  must admit at least one convergent subsequence. Pick any such subsequence with a limit point  $\mathbf{S}^{\dagger*}$ , and pick any  $\mathbf{S}^\dagger \in \text{argmin}_{\mathbf{S}^\dagger \in \Pi(\mathbf{a}, \mathbf{b})} \langle \mathbf{D}, \mathbf{S}^\dagger \rangle$ , from the optimality of  $\mathbf{S}^\dagger_{\vartheta}$  in this subsequence and the optimality of  $\mathbf{S}^\dagger$ , we have:

$$0 < \langle \mathbf{D}, \mathbf{S}^\dagger_{\vartheta} - \mathbf{S}^\dagger \rangle \leq \vartheta \cdot (d(\mathbf{S}^\dagger_{\vartheta}, \bar{\mathbf{S}}) - d(\mathbf{S}^\dagger_{\vartheta}, \mathbf{S}_c)).$$

Now we prove that  $\mathbf{S}^{\dagger*}$  is a feasible solution. Obviously since  $\Pi(\mathbf{a}, \mathbf{b})$  is closed, we have  $\mathbf{S}^{\dagger*} \in \Pi(\mathbf{a}, \mathbf{b})$ . Moreover, taking the limit  $\vartheta \rightarrow 0$  in the inequality above, we have  $\langle \mathbf{D}, \mathbf{S}^{\dagger*} - \mathbf{S}^\dagger \rangle = 0$ .

This implies that  $\mathbf{S}^{\dagger*} \in \text{argmin}_{\mathbf{S}^\dagger \in \Pi(\mathbf{a}, \mathbf{b})} \langle \mathbf{D}, \mathbf{S}^\dagger \rangle$ . Above all, we know  $\mathbf{S}^{\dagger*}$  is a feasible solution. Now we continue to show its optimality. Again, by dividing  $\vartheta$  on both sides of the inequality above and taking the limit  $\vartheta \rightarrow 0$ , we have:  $d(\mathbf{S}^\dagger, \bar{\mathbf{S}}) - d(\mathbf{S}^{\dagger*}, \bar{\mathbf{S}}) \geq 0$ . This implies that  $d(\mathbf{S}^\dagger, \bar{\mathbf{S}})$  reaches a minimal value in the feasible set at  $\mathbf{S}^\dagger = \mathbf{S}^{\dagger*}$ . Moreover, since the optimization problem is strongly convex, we know such  $\mathbf{S}^{\dagger*}$  must be unique. Since  $\Pi(\mathbf{a}, \mathbf{b})$  is closed and every convergent subsequence of  $\{\mathbf{S}^\dagger_{\vartheta}\}$  converges to the optimal solution of the problem,  $\{\mathbf{S}^\dagger_{\vartheta}\}$  converges to the optimal solution of the problem. This ends our proof.

**(b)** Let  $\mathbf{S}_1^*$  and  $\mathbf{S}_2^*$  be the (an) optimal solution of  $\mathcal{J}_{REG}$  and  $\mathcal{J}_{OT}$ , respectively. From the optimality of  $\mathbf{S}_1^*$ , we have:

$$\langle \mathbf{D}, \mathbf{S}_1^* \rangle + \frac{\vartheta}{2} \|\mathbf{S}_1^* - \bar{\mathbf{S}}\|_F^2 \leq \langle \mathbf{D}, \mathbf{S}_2^* \rangle + \frac{\vartheta}{2} \|\mathbf{S}_2^* - \bar{\mathbf{S}}\|_F^2.$$

Furthermore, according to the optimality of  $\mathbf{S}_2^*$ , we have:

$$\langle \mathbf{D}, \mathbf{S}_2^* \rangle + \frac{\vartheta}{2} \|\mathbf{S}_1^* - \bar{\mathbf{S}}\|_F^2 \leq \langle \mathbf{D}, \mathbf{S}_1^* \rangle + \frac{\vartheta}{2} \|\mathbf{S}_1^* - \bar{\mathbf{S}}\|_F^2.$$

Above all we have:

$$\frac{\vartheta}{2} \|\mathbf{S}_1^* - \bar{\mathbf{S}}\|_F^2 \leq \mathcal{J}_{REG} - \mathcal{J}_{OT} \leq \frac{\vartheta}{2} \|\mathbf{S}_2^* - \bar{\mathbf{S}}\|_F^2.$$



Since  $\mathbf{S}_1^* \in \Pi(\mathbf{a}, \mathbf{b})$ , we must have:

$$\min_{\mathbf{S} \in \Pi(\mathbf{a}, \mathbf{b})} \|\mathbf{S} - \bar{\mathbf{S}}\|_F^2 \leq \frac{\vartheta}{2} \|\mathbf{S}_1^* - \bar{\mathbf{S}}\|_F^2.$$

Moreover, we have:

$$\min_{\mathbf{S}_1 = \mathbf{a}} \|\mathbf{S} - \bar{\mathbf{S}}\|_F^2 \leq \min_{\mathbf{S} \in \Pi(\mathbf{a}, \mathbf{b})} \|\mathbf{S} - \bar{\mathbf{S}}\|_F^2, \quad \min_{\mathbf{S}^\top \mathbf{1} = \mathbf{b}} \|\mathbf{S} - \bar{\mathbf{S}}\|_F^2 \leq \min_{\mathbf{S} \in \Pi(\mathbf{a}, \mathbf{b})} \|\mathbf{S} - \bar{\mathbf{S}}\|_F^2.$$

It is easy to see that

$$\min_{\mathbf{S}_1 = \mathbf{a}} \frac{1}{2} \|\mathbf{S} - \bar{\mathbf{S}}\|_F^2 = \frac{\|\bar{\mathbf{S}}\mathbf{1} - \mathbf{a}\|_2^2}{T}, \quad \min_{\mathbf{S}^\top \mathbf{1} = \mathbf{b}} \frac{1}{2} \|\mathbf{S} - \bar{\mathbf{S}}\|_F^2 = \frac{\|\bar{\mathbf{S}}^\top \mathbf{1} - \mathbf{b}\|_2^2}{k}.$$

All these lead to:

$$\vartheta \cdot \max \left\{ \frac{\|\bar{\mathbf{S}}\mathbf{1} - \mathbf{a}\|_2^2}{T}, \frac{\|\bar{\mathbf{S}}^\top \mathbf{1} - \mathbf{b}\|_2^2}{k} \right\} \leq \mathcal{J}_{REG} - \mathcal{J}_{OT}.$$

Moreover, we have  $\|\mathbf{S} - \bar{\mathbf{S}}\|_F^2 \leq 2\|\mathbf{S}\|_F^2 + 2\|\bar{\mathbf{S}}\|_F^2$ . For all  $\mathbf{S} \in \Pi(\mathbf{a}, \mathbf{b})$ , we have:

$$\|\mathbf{S}\|_F^2 = \sum_{i,j} S_{i,j} = \sum_i a_i^2 \sum_j \left( \frac{S_{i,j}}{a_i} \right)^2 \leq \sum_i a_i^2 \sum_j \left( \frac{S_{i,j}}{a_i} \right) \leq \|\mathbf{a}\|_2^2.$$

Similarly, we have  $\|\mathbf{S}\|_F^2 \leq \|\mathbf{b}\|_2^2$ . These lead to :

$$\mathcal{J}_{REG} - \mathcal{J}_{OT} \leq \vartheta \cdot (\|\bar{\mathbf{S}}\|_F^2 + \min\{\|\mathbf{a}\|_2^2, \|\mathbf{b}\|_2^2\}).$$

□

### A.3 Proof of Proposition 3

*Proof.* The lagrangian dual of the problem could be written as:

$$\max_{\mathbf{f}, \mathbf{g}, \Lambda \geq 0} \min_{\mathbf{S}^\dagger} \frac{\vartheta}{2} \|\bar{\mathbf{S}} - \mathbf{S}^\dagger\|_F^2 + \langle \mathcal{D}, \mathbf{S}^\dagger \rangle - \langle \mathbf{S}^\dagger \mathbf{1} - \mathbf{a}, \mathbf{f} \rangle - \langle \mathbf{S}^{\dagger \top} \mathbf{1} - \mathbf{b}, \mathbf{g} \rangle - \langle \Lambda, \mathbf{S}^\dagger \rangle. \quad (1)$$

Since (*Primal*) is strongly convex, the strong duality reduces to the Slater condition. It is easy to see that (*Primal*) satisfies this condition, since  $\mathbf{S}_\dagger = \frac{\mathbf{1}_{k \times T}}{kT} \in \Pi(\mathbf{a}, \mathbf{b})$  and obviously  $\mathbf{S}_\dagger > 0$ . The primal problem could then be solved from its dual problem Eq.(1). Now we show that it is equivalent to solving (*Dual*). Solving the inner minimization problem of Eq.(1), we have:

$$\mathbf{S}^\dagger = \frac{\mathbf{h} \oplus \mathbf{g} + \Lambda - \mathcal{D}}{\vartheta} + \bar{\mathbf{S}}. \quad (2)$$

Note that since the inner minimization problem is strongly convex toward  $\mathbf{S}^\dagger$ , the solution is unique. Plugging the solution into the outer maximization problem yields:

$$\max_{\mathbf{f}, \mathbf{g}, \Lambda \geq 0} \frac{1}{2\vartheta} \|\mathbf{h} \oplus \mathbf{g} + \Lambda - \mathcal{D}\|_F^2 + \langle \mathbf{f}, \mathbf{a} \rangle + \langle \mathbf{g}, \mathbf{b} \rangle + \left\langle \mathcal{D} - \mathbf{h} \oplus \mathbf{g} - \Lambda, \frac{\mathbf{h} \oplus \mathbf{g} + \Lambda - \mathcal{D}}{\vartheta} + \bar{\mathbf{S}} \right\rangle.$$

Fixing  $\mathbf{f}, \mathbf{g}$ , we see the maximization problem is strongly concave toward  $\Lambda$ . This means that  $\Lambda$  can be uniquely determined by  $\mathbf{f}, \mathbf{g}$  with the solution of the partial maximization problem:

$$\max_{\Lambda \geq 0} -\frac{1}{2\vartheta} \|\mathbf{h} \oplus \mathbf{g} + \Lambda - \mathcal{D}\|_F^2 + \langle \mathbf{f}, \mathbf{a} \rangle + \langle \mathbf{g}, \mathbf{b} \rangle + \text{const}. \quad (3)$$

This yields:

$$\Lambda = [\mathbf{h} \oplus \mathbf{g} + \Lambda - \mathcal{D} + \bar{\mathbf{S}}]_- . \quad (4)$$

Plugging Eq.(4) into Eq.(2) and (3) then complete the proof. □

## B Proofs of the Theorems

### B.1 Proof of Theorem 1

*Proof.*

(a) We first show that if  $\chi_{S^\dagger} = K$  then  $\dim(\text{Null}(\Delta(S^\dagger))) = K$ . Since  $\chi_{S^\dagger} = K$ ,  $\exists P_r \in \mathcal{P}_k$ ,  $\exists P_c \in \mathcal{P}_T$ , such that:  $P_r S^\dagger P_c = \bigoplus_{i=1}^K \hat{S}^{(i)}$ , where  $\hat{S}^{(i)} \in \mathbb{R}^{k_i \times T_i}$ ,  $\hat{S}^{(i)} \neq \mathbf{0}$ ,  $\forall i = 1, 2, \dots, K$ . Let  $\tilde{P}_0 = \begin{bmatrix} P_r & 0 \\ 0 & P_c^\top \end{bmatrix}$ , we have  $\tilde{P}_0 \in \mathcal{P}_{k+T}$  and :

$$\tilde{A} = \tilde{P}_0 A_{l \cup o} \tilde{P}_0^\top = \begin{bmatrix} 0 & K_1 \\ K_2 & 0 \end{bmatrix},$$

where  $K_1 = \bigoplus_{i=1}^K \hat{S}^{(i)}$ ,  $K_2 = \bigoplus_{i=1}^K \hat{S}^{(i)\top}$ . In the following, we show that  $\tilde{A}$  could be rearranged to a direct sum of  $K$  submatrices, following the same permutation on its rows and columns. We denote  $e_i = [e_{i,1} \dots e_{i,k+T}]^\top$ , where  $e_{i,j} = 1$ , if  $i = j$ , otherwise we have  $e_{i,j} = 0$ . Furthermore, we define  $e_{a:b} = [e_a, e_{a+1}, \dots, e_b]$ . Define a permutation matrix

$$\tilde{P}_1 = [e_{1:k_1}, e_{(k_1+1):(k_1+T_1)}, e_{(k_1+1):d_2}, e_{(k_1+T_1+1):(k_1+T_2)}, \dots, e_{(d-k_K+1):d}, e_{(k+T-t_K+1):(k+T)}]$$

Then we have:

$$\tilde{A}_1 = \tilde{P}_1^\top \tilde{A} \tilde{P}_1 = \bigoplus_{i=1}^K E_i, \quad E_i = \begin{bmatrix} \mathbf{0} & \hat{S}^{(i)} \\ \hat{S}^{(i)\top} & \mathbf{0} \end{bmatrix}$$

Since  $\chi_{S^\dagger}(P_r, P_c)$  reaches its maximum at  $K$ , the corresponding Graph  $\tilde{\mathcal{G}}$  with  $\tilde{A}_1$  being its affinity matrix has  $K$  connected components. Meanwhile,  $\tilde{A}_1$  differs from  $A_{l \cup o}$  only by row and column permutations. This means that  $\tilde{\mathcal{G}}$  only rearranges the nodes in  $\mathcal{G}_{l \cup o}$ . Hence,  $\mathcal{G}_{l \cup o}$  also has  $K$  connected bipartite components. From the spectral property of a graph Laplacian [Von Luxburg, 2007], we have  $\dim(\text{Null}(\Delta(S))) = K$ . This ends the proof of (a).

(b) Next, we show that if  $\dim(\text{Null}(\Delta(S))) = K$  then  $\chi_{S^\dagger} = K$ . Since  $\dim(\text{Null}(\Delta(S))) = K$ , we know that  $\mathcal{G}_{l \cup o}$  has  $K$  connected components  $g_1, \dots, g_k$ . Denote  $g(l_i)$  as the corresponding group index that  $l_i$  belongs to and  $g(o_j)$  be the group index that  $o_j$  belongs to. Furthermore let  $\mathcal{G}_l$  be a rearrangement of the indexes of latent tasks such that if  $g(l_{i_1}) < g(l_{i_2})$  then  $\mathcal{G}_l(i_1) < \mathcal{G}_l(i_2)$ , if  $g(l_{i_1}) = g(l_{i_2})$ , then  $\mathcal{G}_l(i_1) < \mathcal{G}_l(i_2)$  only if  $i_1 < i_2$ . Meanwhile, let  $\mathcal{G}_o$  be the same rearrangement of the indexes of output tasks. Then define two permutation matrix  $P_r = [e_{\mathcal{G}_l(1)}^{k_1^\top}, e_{\mathcal{G}_l(2)}^{k_2^\top}, \dots, e_{\mathcal{G}_l(k)}^{k_k^\top}]$  and  $P_c = [e_{\mathcal{G}_o(1)}^T, e_{\mathcal{G}_o(2)}^T, \dots, e_{\mathcal{G}_o(k)}^T]$ , where  $e_j^i \in \mathbb{R}^i$ ,  $e_{j,k}^i = 0$  if  $j \neq k$ ,  $e_{j,j}^i = 1$ . Then we have  $P_r S^\dagger P_c = \bigoplus \hat{S}^{(i)}$ . Now we prove that  $\hat{S}^{(i)} \neq \mathbf{0}, \forall i$  by contradiction. Without loss of generality, we assume that  $\hat{S}^{(K)} = \mathbf{0}$ , then all the elements in the last  $k_i$  rows and the last  $T_i$  columns of  $P_r S^\dagger P_c$  must be 0. This contradicts with the fact that  $P_r S^\dagger P_c \mathbf{1} = P_r S^\dagger \mathbf{1} > \mathbf{0}$  and  $P_c^\top S^\dagger P_r^\top \mathbf{1} = P_c^\top S^\dagger \mathbf{1} > \mathbf{0}$ . This implies that  $\chi_{S^\dagger} \geq K$ . Then we show that  $\chi_{S^\dagger}$  could not exceed  $K$  by contradiction. If  $\chi_{S^\dagger} > K$ , then at least one  $\hat{S}^{(i)}$  could be written as a direct sum of two nonzero sub-matrices. This contradicts with the fact that  $\begin{bmatrix} \mathbf{0} & \hat{S}^{(i)} \\ \hat{S}^{(i)\top} & \mathbf{0} \end{bmatrix}$  is a connected bipartite component. This shows that  $\chi_{S^\dagger} \leq K$ . Above all, we must have  $\chi_{S^\dagger} = K$ . □

### B.2 Proof of Theorem 2

We note that the proof is similar to the literature [Overton and Womersley, 1992a, Alizadeh, 1995], where the largest  $K$  eigenvalues instead of the smallest eigenvalues are analyzed. We provide a proof here to make our paper self-contained.

*Proof.* Denote the eigenvalue decomposition of  $A$  as

$$A = Q \Lambda Q^\top, \quad \Lambda = \text{diag}(\lambda_1(A), \dots, \lambda_N(A)).$$

For any element  $\mathbf{U}$  in the feasible set  $\Gamma$ , we have:  $\langle \mathbf{A}, \mathbf{U} \rangle = \sum_i C_{ii} \lambda_i(\mathbf{A})$ , where  $\mathbf{C} = \mathbf{Q}^\top \mathbf{U} \mathbf{Q}$ . Since  $\mathbf{C}$  has the same eigenvalues as  $\mathbf{U}$ , we have  $\mathbf{C} \in \Gamma$  if and only if  $\mathbf{U} \in \Gamma$ . Then we have:

$$\min_{\mathbf{U} \in \Gamma} \langle \mathbf{A}, \mathbf{U} \rangle \iff \min_{\mathbf{C} \in \Gamma} \sum_i C_{ii} \lambda_i(\mathbf{A}). \quad (5)$$

Define  $\mathbf{e}^i \in \mathbb{R}^{N \times 1}$ ,  $\mathbf{e}_i^i = 1$  and  $\mathbf{e}_s^i = 0$ , if  $s \neq i$ , then we reach the fact that:  $C_{ii} = \frac{\mathbf{e}^{i\top} \mathbf{C} \mathbf{e}^i}{\mathbf{e}^{i\top} \mathbf{e}^i}$ . We could then attain the following inequality based on the extremal property of the top/bottom eigenvalue of  $\mathbf{C}$ :

$$0 \leq \lambda_N(\mathbf{C}) = \min_{\mathbf{x}} \frac{\mathbf{x}^\top \mathbf{C} \mathbf{x}}{\mathbf{x}^\top \mathbf{x}} \leq C_{ii} \leq \max_{\mathbf{x}} \frac{\mathbf{x}^\top \mathbf{C} \mathbf{x}}{\mathbf{x}^\top \mathbf{x}} = \lambda_1(\mathbf{C}) \leq 1.$$

Since  $\lambda_1(\mathbf{A}) \geq \lambda_2(\mathbf{A}) \cdots \geq \lambda_N(\mathbf{A})$ , the minimum of (5) is reached at  $\sum_{i=N-K+1}^N \lambda_i(\mathbf{A})$  when  $C_{ii} = 0, i \leq N - K$ ,  $C_{ii} = 1, i \geq N - K + 1$ . This directly shows that  $\sum_{i=N-K+1}^N \lambda_i(\mathbf{A}) = \min_{\mathbf{U} \in \Gamma} \langle \mathbf{A}, \mathbf{U} \rangle$ .

Now it only remains to prove that  $\mathbf{U} = \mathbf{V}_K \mathbf{V}_K^\top$  is an optimal solution. Since  $\mathbf{V}_K$  is the eigenvectors associated with the smallest  $K$  eigenvalues of  $\mathbf{A}$ , we have  $\mathbf{Q} = [\mathbf{V}_K^\perp, \mathbf{V}_K]$ , where  $\mathbf{V}_K^\perp$  denotes the eigenvectors associated with the largest  $N - K$  eigenvalues, and we have  $\mathbf{V}_K^\top \mathbf{V}_K^\perp = \mathbf{0}$  and  $\mathbf{V}_K^\perp{}^\top \mathbf{V}_K = \mathbf{0}$ . In this sense, we obtain:

$$\mathbf{Q}^\top \mathbf{U} \mathbf{Q} = \begin{bmatrix} \mathbf{V}_K^\perp{}^\top \\ \mathbf{V}_K^\top \end{bmatrix} \mathbf{V}_K \mathbf{V}_K^\top \begin{bmatrix} \mathbf{V}_K^\perp \\ \mathbf{V}_K \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{I}_K \end{bmatrix} [\mathbf{0}, \mathbf{I}_K] = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_K \end{bmatrix}.$$

Then the proof follows that  $\mathbf{C} = \mathbf{Q}^\top \mathbf{U} \mathbf{Q}$  satisfies the optimal condition analysed above.  $\square$

### B.3 Proof of Theorem 3

In the proof of Thm.3, we adopt the Gaussian Average as the complexity measure of the hypothesis space, which is defined as follows.

**Definition 2** (Gaussian Average). *Given a set  $\mathcal{C} \subset \mathbb{R}^n$ , the Gaussian Average is defined as*

$$\mathbb{G}(\mathcal{C}) = \mathbb{E} \left[ \sup_{\mathbf{c} \in \mathcal{C}} \langle \boldsymbol{\gamma}, \mathbf{c} \rangle \right],$$

where  $\boldsymbol{\gamma} = [\gamma_1 \cdots \gamma_n]$  and  $\gamma_1 \cdots \gamma_n \stackrel{i.i.d}{\sim} \mathcal{N}(0, 1)$ .

Define the multi-task dataset over  $T$  tasks  $\mathcal{X} = \{\mathbf{X}^{(i)}\}_{i=1}^T$ , where  $\mathbf{X}^{(i)} = \{\mathbf{X}_1^{(i)}, \dots, \mathbf{X}_n^{(i)}\}$ . Given an  $M\phi$ -Lipschitz continuous loss function  $\ell$  with range  $[0, M]$ , we are interested in the following classes.

**1) Target hypothesis Space.** Recall we define the target hypothesis space as  $\mathcal{H}(\mathbf{L}, \mathbf{S}, \mathbf{S}^\dagger, \mathbf{U})$ . For the case of simplicity, we rename it here as  $\mathcal{F}(\mathcal{X})$ .

$$\mathcal{F}(\mathcal{X}) = \mathcal{H}(\mathbf{L}, \mathbf{S}, \mathbf{S}^\dagger, \mathbf{U}) = \left\{ \left\{ \hat{Y}^{(i)}(\mathbf{X}_i^{(t)}) = (\mathbf{L} \mathbf{S}^{(i)})^\top \mathbf{X}_i^{(t)} \right\}_{ti} : \right. \\ \left. \|\mathbf{L}\|_F^2 \leq \xi_1, d(\mathbf{S}, \mathbf{T} \mathbf{S}^\dagger) \leq \xi_2, \langle \Delta(\mathbf{S}^\dagger), \mathbf{U} \rangle \leq \xi_3, \mathbf{S}^\dagger \in \Pi(\mathbf{a}, \mathbf{b}), \mathbf{U} \in \mathcal{M} \right\}.$$

To bound  $\Delta$ , we need to analysis  $\ell \circ \mathcal{F}(\mathcal{X})$ , which is formed as the composition of the loss function and the hypothesis space:

$$\ell \circ \mathcal{F}(\mathcal{X}) = \left\{ \left\{ \ell(f_t(\mathbf{X}_i^{(t)}), \mathbf{Y}_i^{(t)}) \right\}_{i=1}^n \right\}_{t=1}^T : \left\{ f_t(\mathbf{X}_i^{(t)}) \right\}_{ti} \in \mathcal{F}(\mathcal{X}) \right\}$$

**2) Latent response space.** Moreover, given the latent task embedding  $\mathbf{L} = [\mathbf{L}^{(1)}, \dots, \mathbf{L}^{(k)}]$ , then  $\mathbf{L}^{(j)} \mathbf{X}_i^{(t)}$  gives a response for  $\mathbf{X}_i^{(t)}$  on the  $j$ -th latent task. With the choice of  $\mathcal{F}(\mathcal{X})$ , one could only choose latent responses from the following space.

$$\mathcal{L}_c(\mathcal{X}) = \left\{ \left\{ \mathbf{L}^{(k)\top} \mathbf{X}_i^{(t)} \right\}_{kti} : \mathbf{L} = [\mathbf{L}^{(1)}, \dots, \mathbf{L}^{(k)}], \|\mathbf{L}\|_F \leq c \right\}.$$

**3) Output response space.** We define the output response class for a single task as :

$$\mathcal{I}_c = \left\{ \mathbf{y} \in \mathbb{R}^k \mapsto \sum_{i=1}^k y_i \cdot s_i : \mathbf{s} = [s_1 \cdots s_k]^\top, \|\mathbf{s}\|_2 \leq c \right\}.$$

For  $\mathcal{I}_c$ , we define two quantities that are essential for the proof:

$$Q(\mathcal{I}_c) = \sup_{\mathbf{y} \neq \mathbf{y}' \in \mathbb{R}^{kn}} \frac{1}{\|\mathbf{y} - \mathbf{y}'\|} \mathbb{E} \left[ \sup_{s \in \mathcal{I}_c} \sum_{i=1}^n \gamma_i (s(\mathbf{y}_i) - s(\mathbf{y}'_i)) \right]$$

where  $\mathbf{y}$  concatenates all the  $k$ -dimensional latent responses for all samples ( $n$ ) in a single task, i.e.,  $\mathbf{y} = [\mathbf{y}_1; \cdots; \mathbf{y}_n]$ ,  $\mathbf{y}' = [\mathbf{y}'_1; \cdots; \mathbf{y}'_n]$ ,  $\mathbf{y}_i, \mathbf{y}'_i \in \mathbb{R}^{k \times 1}$  are responses for the  $i$ -th instance. Moreover, we define the supremum of the Lipschitz constants over  $\mathcal{I}_c$  as

$$\text{Lip}(\mathcal{I}_c) = \sup \{B : \|s(\mathbf{x}) - s(\mathbf{x}')\| \leq B\|\mathbf{x} - \mathbf{x}'\|, s \in \mathcal{I}_c\}$$

Given  $\mathcal{I}_c$ , we define the representation space as:

$$\mathcal{S}_c = \left\{ \mathbf{y} \in \mathbb{R}^{kTn} \mapsto \{f^{(t)}(\mathbf{y}_{ti})\}_{ti} : f^{(t)}(\cdot) \in \mathcal{I}_c \right\}$$

Here  $\mathbf{y}, \mathbf{y}'$  concatenate all latent responses for all tasks and all samples, i.e.,  $\mathbf{y} = [\mathbf{y}_{11}; \cdots; \mathbf{y}_{1n}; \cdots; \mathbf{y}_{T1}; \cdots; \mathbf{y}_{Tn}]$ , where  $\mathbf{y}_{ti}, \mathbf{y}'_{ti}$  could be regarded as latent responses for the  $i$ -th instance of the  $t$ -th task.

The following Lemma shows that the generalization ability could be controlled by  $\mathbb{G}(\ell \circ \mathcal{F}(\mathcal{X}))$ .

**Lemma 1.** [Thm.9 of [Maurer et al., 2016](#)] Suppose that  $n_1 = n_2 = \cdots = n_T = n$ , the loss function  $l(y, \cdot) : \hat{y} \mapsto [0, M]$ ,  $\forall (\mathbf{L}, \mathbf{S})$  chosen from  $\mathcal{F}(\mathcal{X}) = \mathcal{H}(\mathbf{L}, \mathbf{S}, \mathbf{S}^\dagger, \mathbf{U})$  the following bound holds with possibility at least  $1 - \delta$  :

$$\frac{\Delta}{M} \leq \frac{\sqrt{2\pi} \mathbb{G} \left( \frac{\ell \circ \mathcal{F}(\mathcal{X})}{M} \right)}{nT} + \sqrt{\frac{9 \ln(2/\delta)}{2nT}}.$$

Then we proceed the derivation by providing an explicit bound for  $\mathbb{G}(\ell \circ \mathcal{F}(\mathcal{X}))$ . The following Lemma shows that  $\mathbb{G}(\ell \circ \mathcal{F}(\mathcal{X}))$  could be bounded above by the Gaussian Average of a simpler class.

**Lemma 2.** If  $\ell(\cdot, y)$  is  $M\phi$ -Lipschitz continuous, we have:

$$\mathbb{G} \left( \frac{\ell \circ \mathcal{F}(\mathcal{X})}{M} \right) \leq \phi \cdot \mathbb{G}(\mathcal{S}_\aleph \circ \mathcal{L}_{\xi_1^{1/2}}(\mathcal{X})),$$

where  $\aleph = \sqrt{\xi_2} + 1$ .

*Proof.* Since  $\ell$  is  $\phi$ -Lipschitz continuous, we have:

$$\mathbb{G} \left( \frac{\ell \circ \mathcal{F}(\mathcal{X})}{M} \right) \leq \phi \cdot \mathbb{G}(\mathcal{F}(\mathcal{X})).$$

For every  $\left\{ \hat{Y}^{(t)}(\mathbf{x}_i^{(t)}) \right\}_{ti} \in \mathcal{H}(\mathbf{L}, \mathbf{S}, \mathbf{S}^\dagger, \mathbf{U})$ , the corresponding  $\mathbf{L}$  and  $\mathbf{S}^{(i)}$  satisfy  $\|\mathbf{L}\|_F \leq \sqrt{\xi_1}$  and

$$\|\mathbf{S}^{(i)}\|_2 \leq \sqrt{\xi_2} + \|T\mathbf{S}^{\dagger(i)}\|_2 \leq \sqrt{\xi_2} + \|T\mathbf{S}^{\dagger(i)}\|_2 \leq \sqrt{\xi_2} + \frac{T}{T} = \sqrt{\xi_2} + 1.$$

This implies that  $\mathcal{F}(\mathcal{X}) \subset \mathcal{S}_\aleph \circ \mathcal{L}_{\xi_1^{1/2}}(\mathcal{X})$ . According to the definition of Gaussian Average, we then reach:

$$\mathbb{G}(\mathcal{F}(\mathcal{X})) \leq \mathbb{G}(\mathcal{S}_\aleph \circ \mathcal{L}_{\xi_1^{1/2}}(\mathcal{X})).$$

□

**Lemma 3.**

$$\mathbb{G}(\mathcal{S}_\aleph \circ \mathcal{L}_{\xi_1^{1/2}}(\mathcal{X})) \leq \kappa_1 \aleph \cdot (\xi_1 n k T \|\mathbf{COV}(\mathbf{X})\|_1)^{1/2} + 2\kappa_2 \aleph \cdot T (\xi_1 n k \|\mathbf{COV}(\mathbf{X})\|_\infty)^{1/2}.$$

*Proof.* From the composition rule of Gaussian Average [Maurer et al., 2016], we have:

$$\mathbb{G}(\mathcal{S}_{\aleph} \circ \mathcal{L}_{\xi_1^{1/2}}(\mathcal{X})) \leq \kappa_1 \cdot \text{Lip}(\mathcal{I}_{\aleph}) \cdot \mathbb{G}(\mathcal{L}(\mathcal{X})) + \kappa_2 \sqrt{T} \cdot \text{Diam}[\mathcal{L}(\mathcal{X})] \cdot Q(\mathcal{I}_{\aleph}),$$

where  $\text{Diam}[\mathcal{L}(\mathcal{X})] = \sup_{\mathbf{y}, \mathbf{y}' \in \mathcal{L}(\mathcal{X})} \|\mathbf{y} - \mathbf{y}'\|$ . Then we bound  $\text{Lip}(\mathcal{I}_{\aleph})$ ,  $\mathbb{G}(\mathcal{L}(\mathcal{X}))$ ,  $\text{Diam}(\mathcal{L}(\mathcal{X}))$ , and  $Q(\mathcal{I}_{\aleph})$ , respectively.

- $\forall f \in \mathcal{I}_c$ , uniformly we have:

$$|f(\mathbf{x}_1) - f(\mathbf{x}_2)| \leq \aleph \|\mathbf{x}_1 - \mathbf{x}_2\|, \quad (6)$$

since  $f$  is a linear functional induced by a vector  $\mathbf{s}$  with  $\|\mathbf{s}\| \leq \aleph$ . This suggests that  $\text{Lip}(\mathcal{I}_{\aleph}) \leq \aleph$ .

- For  $\mathbb{G}(\mathcal{L}(\mathcal{X}))$ , we have:

$$\begin{aligned} \mathbb{G}(\mathcal{L}(\mathcal{X})) &= \mathbb{E} \left[ \sup_{\mathcal{L}(\mathcal{X})} \sum_{jtn} \gamma_{jtn} \cdot (\mathbf{L}^{(j)\top} \mathbf{X}_i^{(t)}) \right] \\ &= \mathbb{E} \left[ \sup_{\mathcal{L}(\mathcal{X})} \sum_j \mathbf{L}^{(j)\top} \left( \sum_{ti} \gamma_{jtn} \cdot \mathbf{X}_i^{(t)} \right) \right] \\ &\leq \sqrt{\xi_1} \mathbb{E} \left[ \left( \sum_j \left\| \sum_{ti} \gamma_{jtn} \mathbf{x}_i^{(t)} \right\|^2 \right)^{1/2} \right] \\ &\leq \sqrt{\xi_1} \left( \sum_j \mathbb{E} \left[ \left\| \sum_{ti} \gamma_{jtn} \mathbf{x}_i^{(t)} \right\|^2 \right] \right)^{1/2} \\ &\leq \left( \xi_1 k \sum_{ti} \|\mathbf{x}_i^{(t)}\|^2 \right)^{1/2} \leq (\xi_1 knT \|\mathbf{COV}(\mathbf{X})\|_1)^{1/2}. \end{aligned}$$

- Since  $\text{Diam}(\mathcal{L}(\mathcal{X})) \leq 2 \sup_{\mathbf{y} \in \mathcal{L}(\mathcal{X})} \|\mathbf{y}\|$ , we have:

$$\begin{aligned} \text{Diam}[\mathcal{L}(\mathcal{X})] &\leq 2 \sup_{\|\mathbf{L}\|_F \leq \xi_1^{1/2}} \left( \sum_{jti} (\mathbf{L}^{(j)\top} \mathbf{x}_i^{(t)})^2 \right)^{1/2} \\ &\leq 2 \sup_{\|\mathbf{L}\|_F \leq \xi_1^{1/2}} \left( \sum_j \|\mathbf{L}^{(j)}\|_2^2 \cdot \sum_{ti} \left( \left( \frac{\mathbf{L}^{(j)}}{\|\mathbf{L}^{(j)}\|_2} \right)^\top \mathbf{x}_i^{(t)} \right)^2 \right)^{1/2} \\ &\leq 2\xi_1^{1/2} \cdot \left( k \cdot \sum_{ti} \sup_{\|\mathbf{y}\|_2=1} (\mathbf{y}^\top \mathbf{x}_i^{(t)})^2 \right)^{1/2} \\ &\leq 2(\xi_1 nkT \cdot \|\mathbf{COV}(\mathbf{X})\|_\infty)^{1/2}. \end{aligned}$$



- For  $Q(\mathcal{I}_{\aleph})$ , we have:

$$\begin{aligned}
& \mathbb{E} \left[ \sup_{f \in \mathcal{I}_{\aleph}} \sum_{i=1}^n \gamma_i (f(\mathbf{y}_i) - f(\mathbf{y}'_i)) \right] \\
&= \mathbb{E} \left[ \sup_{\|\mathbf{s}\| \leq \aleph} \sum_{i=1}^n \gamma_i (\mathbf{s}^\top (\mathbf{y}_i - \mathbf{y}'_i)) \right] \\
&\leq \aleph \cdot \mathbb{E} \left[ \left\| \sum_i \gamma_i (\mathbf{y}_i - \mathbf{y}'_i) \right\| \right] \\
&\leq \aleph \left( \mathbb{E} \left[ \left\| \sum_i \gamma_i (\mathbf{y}_i - \mathbf{y}'_i) \right\|^2 \right] \right)^{1/2} \\
&= \aleph \left( \sum_i \left\| (\mathbf{y}_i - \mathbf{y}'_i) \right\|^2 \right)^{1/2} \\
&= \aleph \cdot \|\mathbf{y} - \mathbf{y}'\|.
\end{aligned}$$

This suggests that  $Q(\mathcal{I}_{\aleph}) \leq \aleph$ .

□

**Proof of Theorem 3:** Thm.3 directly follows Lem.1-Lem.3.

#### B.4 Proof of Theorem 4

Our proof requires the following three lemmas from [Golub and Van Loan, 2012].

**Lemma 4.** Let  $\mathbf{A} \in \mathbb{R}^{m \times n}$  with  $m \leq n$ . Moreover, denote the singular values of  $\mathbf{A}$  as  $\sigma_1(\mathbf{A}) \geq \sigma_2(\mathbf{A}) \geq \dots \geq \sigma_m(\mathbf{A})$ . Let  $\mathbf{B} = \begin{bmatrix} \mathbf{0} & \mathbf{A} \\ \mathbf{A}^\top & \mathbf{0} \end{bmatrix}$ . The eigenvalues of  $\mathbf{B}$  are  $\lambda_1(\mathbf{B}) \geq \lambda_2(\mathbf{B}) \geq \dots \geq \lambda_{m+n}(\mathbf{B})$ . We have :  $\lambda_1(\mathbf{B}) = \sigma_1(\mathbf{A}), \dots, \lambda_m(\mathbf{B}) = \sigma_m(\mathbf{A}), \lambda_{m+1}(\mathbf{B}) = \lambda_{m+2}(\mathbf{B}) = \dots = \lambda_n(\mathbf{B}) = 0, \lambda_{n+1}(\mathbf{B}) = -\sigma_m(\mathbf{A}), \lambda_{n+2}(\mathbf{B}) = -\sigma_{m-1}(\mathbf{A}), \dots, \lambda_{n+m}(\mathbf{B}) = -\sigma_1(\mathbf{A})$ .

**Lemma 5.** Let  $\mathbf{A}, \mathbf{B} \in \mathbb{S}^N$ ,  $\lambda_1(\mathbf{A}) \geq \lambda_2(\mathbf{A}) \geq \dots \geq \lambda_N(\mathbf{A})$ ;  $\lambda_1(\mathbf{B}) \geq \lambda_2(\mathbf{B}) \geq \dots \geq \lambda_N(\mathbf{B})$ . Then:

$$\lambda_N(\mathbf{A}) + \lambda_j(\mathbf{B}) \leq \lambda_j(\mathbf{A} + \mathbf{B}) \leq \lambda_1(\mathbf{A}) + \lambda_j(\mathbf{B}), \text{ for } j = 1, 2, \dots, i.$$

**Lemma 6** (Wielandt-Hoffman). For  $\mathbf{A} \in \mathbb{S}^N$ , and  $\mathbf{E} \in \mathbb{S}^N$ , we have:

$$\sum_{i=1}^N (\lambda_i(\mathbf{A} + \mathbf{E}) - \lambda_i(\mathbf{A}))^2 \leq \|\mathbf{E}\|_F^2.$$

Moreover, for  $\tilde{\mathbf{A}} \in \mathbb{R}^{m \times n}$ , and  $\tilde{\mathbf{E}} \in \mathbb{R}^{m \times n}$ , we have:

$$\sum_{i=1}^N \left( \sigma_i(\tilde{\mathbf{A}} + \tilde{\mathbf{E}}) - \sigma_i(\tilde{\mathbf{A}}) \right)^2 \leq \|\tilde{\mathbf{E}}\|_F^2.$$

*Proof.*

**proof of (a)** According to the definition of  $\Delta(\mathbf{S}^\dagger)$  and  $\Delta(\mathbf{S})$ , we have :

$$\begin{aligned}
\left\| T\Delta(\mathbf{S}^\dagger) - \Delta(\mathbf{S}) \right\|_F &\leq \left\| \begin{bmatrix} \mathbf{0} & T\mathbf{S}^\dagger \\ T\mathbf{S}^{\dagger\top} & \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{0} & \mathbf{S} \\ \mathbf{S}^\top & \mathbf{0} \end{bmatrix} \right\|_F + \left\| T\mathbf{S}^\dagger \mathbf{1}_k - \mathbf{S} \mathbf{1}_k \right\|_2 + \left\| T\mathbf{S}^{\dagger\top} \mathbf{1}_T - \mathbf{S}^\top \mathbf{1}_T \right\|_2 \\
&= \sqrt{2} \left\| T\mathbf{S}^\dagger - \mathbf{S} \right\|_F + \left\| T\mathbf{S}^\dagger \mathbf{1}_k - \mathbf{S} \mathbf{1}_k \right\|_2 + \left\| T\mathbf{S}^{\dagger\top} \mathbf{1}_T - \mathbf{S}^\top \mathbf{1}_T \right\|_2 \\
&\leq \sqrt{2\xi_2} + \left\| \mathbf{S} - T\mathbf{S}^\dagger \right\|_F \cdot (\|\mathbf{1}_k\|_2 + \|\mathbf{1}_T\|_2) \\
&\leq \sqrt{2\xi_2} + \sqrt{\xi_2}(\sqrt{k} + \sqrt{T}).
\end{aligned}$$

According to Lem.6, we have:

$$\begin{aligned} \sum_{i=N-K+1}^N \lambda_i(\Delta(\mathbf{S})) &\leq \sum_{i=N-K+1}^N T\lambda_i(\Delta(\mathbf{S}^\dagger)) + \sqrt{2\xi_2 K} + \sqrt{\xi_2 K}(\sqrt{k} + \sqrt{T}) \\ &= T\xi_3 + \sqrt{2\xi_2 K} + \sqrt{\xi_2 K}(\sqrt{k} + \sqrt{T}). \end{aligned}$$

**proof of (b)** Let  $\mathbf{A} = \text{diag}(\mathbf{S}^\dagger \mathbf{1})$ ,  $\mathbf{B} = -\Delta(\mathbf{S}^\dagger)$ , we have  $\mathbf{A} + \mathbf{B} = \mathbf{A}_{\mathcal{L} \cup \mathcal{O}}$ . Applying Lem. 5 for  $j = 1, 2, \dots, K$  we have :

$$\lambda_N(\text{diag}(\mathbf{S}^\dagger \mathbf{1})) - \lambda_{N-j+1}(\Delta(\mathbf{S}^\dagger)) \leq \lambda_j(\mathbf{A}_{\mathcal{L} \cup \mathcal{O}}) \stackrel{(\star)}{=} \sigma_j(\mathbf{S}^\dagger) \leq \lambda_1(\text{diag}(\mathbf{S}^\dagger \mathbf{1})) - \lambda_{N-j+1}(\Delta(\mathbf{S}^\dagger)), \quad (7)$$

where  $(\star)$  is due to Lem.4. According to Thm.2 and the definition of  $\xi_3$ , we have:

$$\sum_{i=1}^K \lambda_{N-i+1}(\Delta(\mathbf{S}^\dagger)) \leq \langle \Delta(\mathbf{S}^\dagger), \mathbf{U} \rangle \leq \xi_3. \quad (8)$$

With Lem.4 and Eq.(7)-(8), we have :

$$\frac{1}{T} - \xi_3 \leq \sigma_i(\mathbf{S}^\dagger) \leq \frac{1}{k}, \text{ for } i = 1, 2, \dots, K.$$

Since  $d(\mathbf{S}, T\mathbf{S}^\dagger) \leq \xi_2$ , according to Lem.6, we have:

$$T\sigma_i(\mathbf{S}^\dagger) - \sqrt{\xi_2} \leq \sigma_i(\mathbf{S}) \leq T\sigma_i(\mathbf{S}^\dagger) + \sqrt{\xi_2}, \quad i = 1, 2, \dots, K.$$

The theorem follows from Thm.4.2.3 in [Horn and Johnson, 2012] and the fact that  $\sigma_i(\mathbf{S}) = \sqrt{\lambda_i(\mathbf{S}^\top \mathbf{S})}$ . □

## B.5 Proof of Theorem 5

First, we need the following two lemmas about the sine  $\Theta$  theorem, the proof of which could be found in [Yin and Shen, 2018, Lem.1] and [Yu et al., 2014, Thm.1] respectively.

**Lemma 7.** Let  $\mathbf{X}$  and  $\mathbf{Y}$  be two orthogonal matrices of  $\mathbb{R}^{n \times n}$ . Let  $\mathbf{X} = [\mathbf{X}_0, \mathbf{X}_1]$  and  $\mathbf{Y} = [\mathbf{Y}_0, \mathbf{Y}_1]$ , where  $\mathbf{X}_0$  and  $\mathbf{Y}_0$  are the first  $K$  columns of  $\mathbf{X}$  and  $\mathbf{Y}$ , respectively. Then, we have:

$$\|\mathbf{X}_0 \mathbf{X}_0^\top - \mathbf{Y}_0 \mathbf{Y}_0^\top\|_F \leq \sqrt{2} \|\mathbf{X}_0^\top \mathbf{Y}_1\|_F.$$

**Lemma 8** (sine  $\Theta$ ). Let  $\Sigma, \hat{\Sigma}$  be symmetric with eigenvalues  $\lambda_1 \geq \dots, \lambda_p$  and  $\hat{\lambda}_1 \geq \dots, \hat{\lambda}_p$ , respectively. Fix  $1 \leq K \leq p$ , and let  $\mathbf{X}_0 = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_K] \in \mathbb{R}^{p \times K}$  and  $\hat{\mathbf{Y}}_0 = [\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2, \dots, \hat{\mathbf{v}}_K]$  and let  $\mathbf{X}_1 = [\mathbf{v}_{K+1}, \dots, \mathbf{v}_p]$  and  $\mathbf{Y}_1 = [\hat{\mathbf{v}}_{K+1}, \dots, \hat{\mathbf{v}}_p]$ . For  $1 \leq j \leq p$ , we have  $\Sigma \mathbf{v}_j = \lambda_j \mathbf{v}_j$  and  $\hat{\Sigma} \hat{\mathbf{v}}_j = \hat{\lambda}_j \hat{\mathbf{v}}_j$ . If  $\delta = |\hat{\lambda}_{K+1} - \lambda_K| > 0$ , we have:

$$\|\mathbf{X}_0^\top \mathbf{Y}_1\|_F \leq \frac{\|\Sigma - \hat{\Sigma}\|_F}{\delta}.$$

Now we are ready to prove the theorem.

*Proof.* For all  $\mathbf{S}^* \in \Pi(\mathbf{a}, \mathbf{b})$ , such that  $\text{Supp}(\mathbf{S}^*) = \mathcal{G}$ , according to Thm.2, we have that  $V_K^* = [\mathbf{f}_1^*, \dots, \mathbf{f}_{k+T}^*]^\top$  the eigenvectors associated with the bottom  $K$  eigenvalues of the Laplacian  $\Delta(\mathbf{S}^*)$  much satisfy that  $j \leq k$ ,  $[\mathbf{f}_j^*]_g = 1$  only if  $\mathbf{l}_j$  belongs to group  $g$  and that for  $j > k$ ,  $[\mathbf{f}_j^*]_g = 1$  only if  $\mathbf{o}_j$  belongs to group  $g$ . Moreover, we define  $V_K = [\mathbf{f}_1, \dots, \mathbf{f}_{k+T}]^\top$  as the eigenvectors associated with the bottom  $K$  eigenvalues of  $\Delta(\mathbf{S}^\dagger)$  and define

$$\mathcal{D}_{i,j}^* = \|\mathbf{f}_i^* - \mathbf{f}_{k+j}^*\|_2^2, \quad \mathcal{D}_{i,j} = \|\mathbf{f}_i - \mathbf{f}_{k+j}\|_2^2.$$

Since  $\mathbf{S}^\dagger$  is selected from  $\mathcal{H}$ , according to Thm.2, we have:

$$\langle \mathcal{D}, \mathbf{S}^\dagger \rangle = \sum_{i=N-K+1}^N \lambda_i (\Delta(\mathbf{S}^\dagger)) \leq \langle \mathbf{U}, \Delta(\mathbf{S}^\dagger) \rangle, \quad \forall \mathbf{U} \in \mathcal{M}.$$

This implies  $\langle \mathcal{D}, \mathbf{S}^\dagger \rangle \leq \xi_3$ .

Since for all  $i$ ,  $\mathbf{f}_i^*$  is a 0-1 group indicator, we have

$$\mathcal{D}_{ij}^* = \begin{cases} 0, & \mathbf{l}_i \text{ and } \mathbf{o}_j \text{ belong to the same group} \\ 2, & \text{otherwise} \end{cases}$$

and thus  $\langle \mathcal{D}^*, \mathbf{S}^\dagger \rangle = 2 \|\mathbf{S}^{\dagger \text{supp}^c}\|_1$ . Then, we have :

$$\|\mathbf{S}^{\dagger \text{supp}^c}\|_1 \leq \frac{1}{2} \cdot (\xi_3 + \langle \mathcal{D}^* - \mathcal{D}, \mathbf{S}^\dagger \rangle).$$

It now only remains to give an upper bound for  $\langle \mathcal{D}^* - \mathcal{D}, \mathbf{S}^\dagger \rangle$ . In fact:

$$\langle \mathcal{D}^* - \mathcal{D}, \mathbf{S}^\dagger \rangle = \langle \mathbf{U}^* - \mathbf{U}, \Delta(\mathbf{S}^\dagger) \rangle \leq \|\Delta(\mathbf{S}^\dagger)\|_F \cdot \|\mathbf{U} - \mathbf{U}^*\|_F \leq \frac{\sqrt{\frac{2}{k} + \frac{6}{T}} \cdot \epsilon}{\lambda_{K+1}(\Delta(\mathbf{S}^\dagger))},$$

where  $\mathbf{U}^* = \mathbf{V}_K^* \mathbf{V}_K^{*\top}$ ,  $\mathbf{U} = \mathbf{V}_K \mathbf{V}_K^\top$ , and the last inequality follows from that

$$\begin{aligned} \|\Delta(\mathbf{S}^\dagger)\|_F &= \sqrt{\left\| \text{diag}\left(\frac{\mathbf{1}_k}{k} + \frac{\mathbf{1}_T}{T}\right) - \begin{bmatrix} \mathbf{0} & \mathbf{S}^\dagger \\ \mathbf{S}^{\dagger\top} & \mathbf{0} \end{bmatrix} \right\|_F^2} \\ &= \sqrt{\left\| \frac{\mathbf{1}_k}{k} \right\|_F^2 + \left\| \frac{\mathbf{1}_T}{T} \right\|_F^2 + 2 \left\| \mathbf{S}^\dagger \right\|_F^2} \\ &\leq \sqrt{\frac{1}{k} + \frac{1}{T} + \frac{2}{T^2}}. \end{aligned}$$

and Lem.7 and Lem.8, with  $\hat{\Sigma} = \Delta(\mathbf{S}^\dagger)$ ,  $\Sigma = \Delta_{\mathbf{S}^*}$ , and  $|\hat{\lambda}_{K+1} - \lambda_K| = \lambda_{K+1}(\Delta(\mathbf{S}^\dagger))$ . This leads to an upper bound of  $\|\mathbf{S}^{\dagger \text{supp}^c}\|_1$ :

$$\|\mathbf{S}^{\dagger \text{supp}^c}\|_1 \leq \frac{1}{2} \cdot \left( \xi_3 + \frac{\sqrt{\frac{2}{k} + \frac{6}{T}} \cdot \epsilon}{\lambda_{K+1}(\Delta(\mathbf{S}^\dagger))} \right).$$

The final upper bound for  $\|\mathbf{S}^{\dagger \text{supp}^c}\|_1$  follows that:

$$\epsilon \leq \sup_{\mathbf{S}, \mathbf{S}' \in \Pi(\mathbf{a}, \mathbf{b})} \|\Delta(\mathbf{S}) - \Delta(\mathbf{S}')\|_F \leq 2\sqrt{2} \cdot \sup_{\mathbf{S} \in \Pi(\mathbf{a}, \mathbf{b})} \|\mathbf{S}\|_F \leq 2\sqrt{\frac{2}{T}}.$$

□

## C Details of the Optimization Algorithm

### Solving the $L$ subproblem.

Recall the  $L$  subproblem of (Obj), where we fix all the other variable and solve  $L$ . This subproblem could be formulated as:

$$\underset{\mathbf{L}}{\text{argmin}} \sum_{i=1}^T \sum_{j=1}^{n_j} \mathcal{J}(\mathbf{Y}^{(i)}, \mathbf{X}^{(i)} \mathbf{L} \mathbf{S}^{(i)}) + \frac{\alpha_1}{2} \cdot \|\mathbf{L}\|_F^2.$$

Recall our setting in the experiments, we set  $\mathcal{J}(\mathbf{Y}^{(i)}, \mathbf{X}^{(i)} \mathbf{L} \mathbf{S}^{(i)}) = \frac{1}{N_i} \|\mathbf{Y}^{(i)} - \mathbf{X}^{(i)} \mathbf{L} \mathbf{S}^{(i)}\|_F^2$  for regression problems, where  $N_i$  is the number of instances for the  $i$ -th task. And we adopt the squared surrogate loss for AUC:

$$\mathcal{J}(\mathbf{Y}^{(i)}, \mathbf{X}^{(i)} \mathbf{L} \mathbf{S}^{(i)})_{LS} = \sum_{x_p \in \mathcal{S}_{+,i}} \sum_{x_q \in \mathcal{S}_{-,i}} \frac{s(\mathbf{g}^{(i)}(x_p) - \mathbf{g}^{(i)}(x_q))}{n_{+,i} n_{-,i}}$$

for classification problem, where  $g^{(i)}(\mathbf{x}) = \langle \mathbf{L}\mathbf{S}^{(i)}, \mathbf{x} \rangle$ ,  $\mathcal{S}_{+,i}$  is the set of positive instances for the  $i$ -th task,  $\mathcal{S}_{-,i}$  is the set of negative instances for the  $i$ -th task. To simplify the squared surrogate AUC loss, we build an AUC graph, where the vertexes are the instances and the edges are only activated across different classes. Specifically, for each task  $i$ , we define the graph as  $\mathcal{G}_{AUC}^{(i)} = (\mathcal{V}^{(i)}, \mathcal{E}^{(i)}, \mathcal{W}^{(i)})$ . The vertex set  $\mathcal{V}^{(i)}$  is the set of all the instances in  $(\mathbf{X}^{(i)}, \mathbf{y}^{(i)})$ . There exists an edge  $(k, m) \in \mathcal{E}^{(i)}$  with weight  $\mathcal{W}_{km}^{(i)} = \frac{1}{n_{+,i}n_{-,i}}$  if and only if  $y_k^{(i)} \neq y_m^{(i)}$ . Given  $\mathcal{W}^{(i)}$ , the Laplacian matrix  $\Delta_{AUC}^{(i)}$  of  $\mathcal{G}_{AUC}^{(i)}$  could be expressed as:  $\Delta_{AUC}^{(i)} = \text{diag}(\mathcal{W}^{(i)}\mathbf{1}) - \mathcal{W}^{(i)}$ . With the definition of  $\Delta_{AUC}^{(i)}$ , we could reformulate the empirical loss  $\mathcal{J}(\mathbf{Y}^{(i)}, \mathbf{X}^{(i)}\mathbf{L}\mathbf{S}^{(i)})_{AUC}$  as :

$$\mathcal{J}(\mathbf{Y}^{(i)}, \mathbf{X}^{(i)}\mathbf{L}\mathbf{S}^{(i)})_{AUC} = (\mathbf{Y}^{(i)} - \mathbf{X}^{(i)}\mathbf{L}\mathbf{S}^{(i)})^\top \Delta_{AUC}^{(i)} (\mathbf{Y}^{(i)} - \mathbf{X}^{(i)}\mathbf{L}\mathbf{S}^{(i)}).$$

The gradient of  $\mathbf{L}$  with these two loss functions are then given as:

$$\begin{aligned} \nabla_{\mathbf{L}}(\mathcal{J}_{LS}) &= \sum_{i=1}^T \frac{1}{N_i} \left( \mathbf{X}^{(i)\top} \mathbf{X}^{(i)} \mathbf{L} \mathbf{S}^{(i)} \mathbf{S}^{(i)\top} - \mathbf{X}^{(i)\top} \mathbf{Y}^{(i)} \mathbf{S}^{(i)\top} \right) + \alpha_1 \mathbf{L}, \\ \nabla_{\mathbf{L}}(\mathcal{J}_{AUC}) &= \sum_{i=1}^T \left( \mathbf{X}^{(i)\top} \Delta_{AUC}^{(i)} \mathbf{X}^{(i)} \mathbf{L} \mathbf{S}^{(i)} \mathbf{S}^{(i)\top} - \mathbf{X}^{(i)\top} \Delta_{AUC}^{(i)} \mathbf{Y}^{(i)} \mathbf{S}^{(i)\top} \right) + \alpha_1 \mathbf{L}. \end{aligned}$$

Though both losses enjoy a closed-form solution, it comes with an extremely high time complexity of  $O(k^6 T^6)$ . We adopt L-BFGS [Zhu et al., 1997] as our optimizer, which only requires loss and gradient evaluations. If  $\mathbf{X}^{(i)\top} \mathbf{X}^{(i)}$ ,  $\mathbf{X}^{(i)\top} \mathbf{Y}^{(i)}$ ,  $\mathbf{X}^{(i)\top} \Delta_{AUC}^{(i)} \mathbf{X}^{(i)}$  and  $\mathbf{X}^{(i)\top} \Delta_{AUC}^{(i)} \mathbf{Y}^{(i)}$  are precomputed and cached to the memory, we come to a complexity of  $O(d^2 T + kdT)$  per gradient evaluation.

**Solving the  $\mathbf{S}$  subproblem** With the other parameters fixed,  $\mathbf{S}$  could be solved from the following problem:

$$\underset{\mathbf{S}}{\text{argmin}} \sum_{i=1}^T \sum_{j=1}^{n_j} \mathcal{J}(\mathbf{Y}^{(i)}, \mathbf{X}^{(i)}\mathbf{L}\mathbf{S}^{(i)}) + \frac{\alpha_2}{2} \|\mathbf{S} - T\mathbf{S}^\dagger\|_F^2.$$

With the squared loss for regression, the closed-form solution reads

$$\mathbf{S}^{(i)\star} = \left( \frac{\mathbf{A}_i^\top \mathbf{A}_i}{N_i} + \alpha_2 \mathbf{I} \right)^{-1} \left( \frac{\mathbf{A}_i^\top \mathbf{Y}^{(i)}}{N_i} + \alpha_2 T \mathbf{S}^\dagger \right),$$

where  $\mathbf{A}_i = \mathbf{X}^{(i)}\mathbf{L}$ . Similarly, for the squared surrogate loss for AUC, the closed-form solution becomes:

$$\mathbf{S}^{(i)\star} = \left( \frac{\mathbf{A}_i^\top \Delta_{AUC}^{(i)} \mathbf{A}_i}{N_i} + \alpha_2 \mathbf{I} \right)^{-1} \left( \frac{\mathbf{A}_i^\top \Delta_{AUC}^{(i)} \mathbf{Y}^{(i)}}{N_i} + \alpha_2 T \mathbf{S}^\dagger \right).$$

## D Experiments

### D.1 Real-world Dataset

**Competitors** Now we briefly introduce competitors adopted in this paper. To show the improvement toward different types of methods, our method is compared with the following methods:

- RAMUSA [Han and Zhang, 2016] adopts a capped trace norm regularizer to minimize only the singular values smaller than an adaptively tuned threshold.
- GOMTL [Kumar and III, 2012] decomposes per-task parameters as linear combinations of latent task basis, where  $\Omega(\mathbf{S})$  is set to  $\ell_1$  penalty. In this way, it learns an arbitrary sparse LATM.
- CoCMTL [Xu et al., 2015] realizes the task-specific co-clustering via minimizing the truncated sum-of-squares of the singular values of the task matrix.
- CMTL [Jacob et al., 2009] assumes that the per-tasks parameters are clustered into a given number of groups. Specifically, it leverages a clustered model parameter via simultaneously encouraging a large between-cluster variance and a small within-cluster variance.
- NC-CMTL [Nie et al., 2018] explores shared information among different tasks with a non-convex low-rank spectral regularizer and a robust re-weighting scheme.

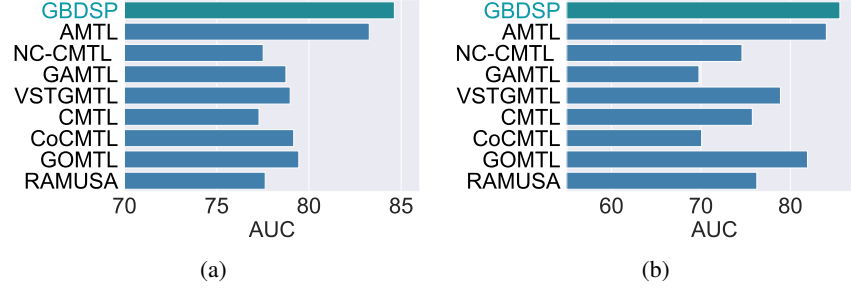


Figure 3: (a) Performance Comparison Curve over the Simulation Dataset, with varying training data ratio. (b) Performance Comparison over the Sun Dataset.

- VSTGMTL [Jeong and Jun, 2018] implements simultaneous variable selection and learning with a low-rank decomposition.
- AMTL [Lee et al., 2016] assumes that each task parameter is a linear combination of other tasks and leverages asymmetric transfer between tasks with a sparse selection on the asymmetric transfer matrix.
- GAMTL [Liu and Pan, 2017] Similar to AMTL, GAMTL also assumes that each task parameter is a linear combination of other tasks. Moreover, it also adopts the trace lasso penalty to maintain a better grouping structure.

**AWA2-Attribute:** Based on the AWA-2 dataset [Xian et al., 2018], we construct an MTL dataset containing 85 tasks. Each task in this dataset is a binary classification to recognize whether a given attribute is presented in a given instance. To construct this dataset, for each attribute we sampled 50 positive instances and 150 negative instances. This yields to a total volume of 17,000 samples.

**AWA2-Class:** Similar to AWA2-Attribute, we perform another MTL dataset concerning the animal class recognition. The  $i$ -th task in this dataset is a binary classification to recognize whether a given sample belongs to the  $i$ -th class in AWA2. Similarly, we sampled 50 positive instances and 150 negative instances for each task. This yields to a total volume of 10,000 samples.

**School Dataset :** This dataset is collected from the Inner London Education Authority, which consists of examination scores of 15362 students from 139 schools in London [Kumar and III, 2012]. Here, the score prediction for each school corresponds to a task, thus giving a total of 139 tasks.

**Shoes Dataset.** The Shoes Dataset [Kovashka and Grauman, 2015] is a popular attribute prediction benchmark, which consists of 14,658 online shopping shoe images with 7 attributes (BR: brown, CM: comfortable, FA: fashionable, FM: formal, OP: open, ON: ornate, PT: pointy). In this dataset, annotators with various knowledge are invited to judge whether a specific attribute is present in an image. Here, the tasks are then consisted of predicting the attribute annotations for different users. Specifically, each user is randomly assigned with 50 images, and there are at least 190 users for each attribute who take part in the process, which results in a total volume of 90,000 annotations.

**Sun Dataset.** The Sun Dataset [Kovashka and Grauman, 2015] contains 14,340 scene images from SUN Attribute Database [andJames Hays, 2012], with personalized annotations over 5 attributes (CL: Cluttered, MO: Modern, OP: Opening Area, RU: Rustic, SO: Soothe). With a similar annotating procedure, 64,900 annotations are obtained in this dataset. The tasks are defined in the same way as Shoes dataset.

**Pre-processing.** For AWA2-Attribute and AWA2-Attribute, we adopt the ILSVRC-pretrained ResNet101 feature which is used in [Xian et al., 2018]. For School Dataset, we use the features provided in the MALSAR package [Zhou et al., 2011b]. For Shoes dataset, we simply adopt the GIST and color histogram provided in [Kovashka et al., 2012] as input features, whereas we deploy the 2048-dim feature vectors extracted by the Inception-V3 [Szegedy et al., 2016] network for Sun’s data. The reason leading us to two different feature extraction strategies lies in that the images in Shoes dataset are photographed on a white background, while images in Sun dataset usually suffer from much more complicated backgrounds. For the classification datasets, we perform PCA to reduce the redundancy of these features before training. For sun and shoes dataset, we notice that users who extremely prefer to provide merely one class of labels may lead to large biases. To e-



Table 1: Performance Comparison over General MTL Datasets (mean  $\pm$  std)

Algorithms	AWA2-Attr( $\uparrow$ )	AWA2-Cls ( $\uparrow$ )	School ( $\downarrow$ )
RAMUSA [Han and Zhang, 2016]	88.60 $\pm$ 0.66	93.06 $\pm$ 0.53	10.52 $\pm$ 0.09
GOMTL [Kumar and III, 2012]	89.56 $\pm$ 0.33	88.22 $\pm$ 1.18	<b>10.26<math>\pm</math>0.11</b>
CoCMTL [Xu et al., 2015]	92.29 $\pm$ 0.35	94.69 $\pm$ 0.73	12.06 $\pm$ 0.09
CMTL [Zhong and Kwok, 2012]	<b>92.95<math>\pm</math>0.35</b>	94.81 $\pm$ 0.70	12.06 $\pm$ 0.09
VSTGMTL [Jeong and Jun, 2018]	89.31 $\pm$ 0.39	92.03 $\pm$ 0.94	10.17 $\pm$ 0.08
GAMTL [Liu and Pan, 2017]	89.39 $\pm$ 0.42	92.55 $\pm$ 0.53	10.50 $\pm$ 0.12
NC-CMTL [Nie et al., 2018]	<b>92.99<math>\pm</math>0.32</b>	95.10 $\pm$ 0.60	10.53 $\pm$ 0.12
AMTL [Lee et al., 2016]	92.15 $\pm$ 0.34	<b>95.76<math>\pm</math>0.44</b>	12.15 $\pm$ 0.09
<b>GBDSP</b>	92.73 $\pm$ 0.29	<b>97.86<math>\pm</math>0.22</b>	<b>10.10<math>\pm</math>0.08</b>

liminate such effect, we manually remove users who give less than 8 annotations for the minority class.

**Performance comparison** The performance results over general MTL datasets are shown in Tab. 1, and the results for personalized attribute prediction datasets are shown in Fig. 3(a) and Fig. 3(b). Then we could make the following observations: 1) Our proposed algorithm consistently outperforms all the competitors in all datasets except AWA2-Attr. Moreover, on AWA2-Attr, **GBDSP** shows competitive performance with CMTL and NC-CMTL. 2) Comparing with GOMTL, **GBDSP** always shows significant performance improvements. This implies that leveraging a block-diagonal rather than arbitrarily sparse LATM helps to improve the performance. 3) In most cases, we find AMTL outperforms the other low-rank constrained methods on all the datasets, as it explicitly models and reduces the influence of negative transfer via asymmetric learning. 4) Our proposed method outperforms AMTL on most results. One possible reason is that AMTL avoids negative transfer via differentiates the hard tasks with the easy tasks, while **GBDSP** provides a finer-grained against negative transfer via suppressing inter-group transfer with the latent task representation.

**Task Correlation study on AWA-Attr and AWA-Cls** To see how learning a block-diagonal  $\tilde{S}$  helps to recover a valuable task correlation matrix, we compare the predicted correlation matrix with the semantic correlation matrix obtained from the class-attribute relations. Here the predicted task correlation matrix is calculated from  $\tilde{R} = |\tilde{S}|^T |\tilde{S}|$ , with  $\tilde{R}_{ij}$  measures the similarity between the latent task assignment between  $\mathbf{o}_i$  and  $\mathbf{o}_j$ . The semantic correlation is obtained based on the predicate-matrix  $P$  provided by the AW2 dataset<sup>3</sup>, where  $P \in \{0, 1\}^{50 \times 85}$  and  $P_{ij} = 1$  only if attribute  $j$  is relevant to class  $i$ . For AWA-Attr, we formulate the semantic task correlation matrix  $\bar{R}$  as  $\bar{R}_{ij} = \sum_{k=1}^{50} \delta(P_{ki} = P_{kj})$ . Here  $\bar{R}_{ij}$  counts how many times attribute  $i$  and  $j$  are simultaneously relevant/irrelevant to the same class. For AWA-Cls, we have a similar construction of  $\bar{R}$  as  $\bar{R}_{ij} = \sum_{k=1}^{85} \delta(P_{ik} = P_{jk})$ , where  $\bar{R}_{ij}$  counts how many times class  $i$  and  $j$  are simultaneously relevant/irrelevant to the same attribute. Moreover, the matrices are normalized such that  $\max_j \{\tilde{R}_{ij}\} = 1$  and  $\max_j \{\bar{R}_{ij}\} = 1, \forall i$ .

In Fig. 5 and Fig. 6, we visualize  $\tilde{R}$  generated by GOMTL and **GBDSP** respectively together with  $\bar{R}$ . Moreover, as shown in Fig. 4, we also calculate the cosine similarity between the  $\tilde{R}$  and  $\bar{R}$ . The results indicate that learning a block-diagonal instead of an arbitrary sparse  $S$  matrix helps to preserve the semantic similarity among tasks.

**Fine-grained Comparisons on Shoes and Sun Dataset** Fig. 7 shows attribute-wise comparison on Shoes and Sun Dataset.

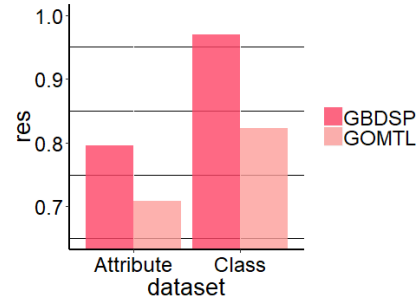


Figure 4: Task Correlation Comparison

<sup>3</sup><https://cvml.ist.ac.at/AwA2/>

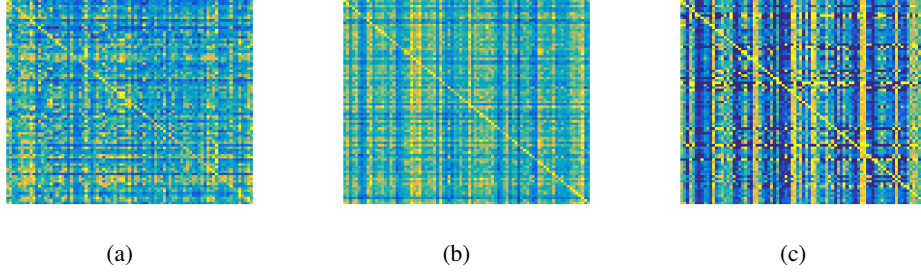


Figure 5: Comparison over the correlation structure for AWA2-Attribute: (a)  $\tilde{\mathbf{R}}$  obtained from GOMTL, (b)  $\tilde{\mathbf{R}}$  obtained from **GBDSP** (c)  $\bar{\mathbf{R}}$

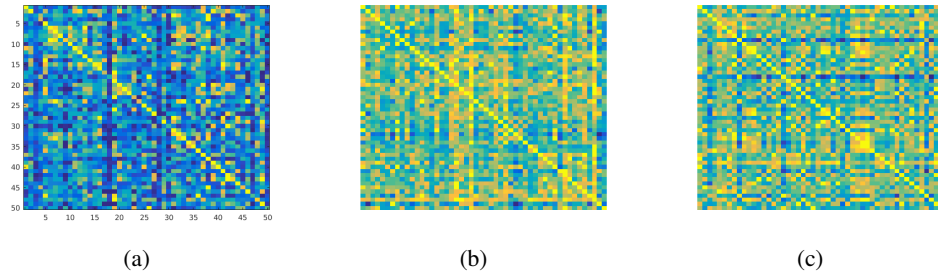


Figure 6: Comparison over the correlation structure for AWA2-Class: (a)  $\tilde{\mathbf{R}}$  obtained from GOMTL, (b)  $\tilde{\mathbf{R}}$  obtained from **GBDSP** (c)  $\bar{\mathbf{R}}$

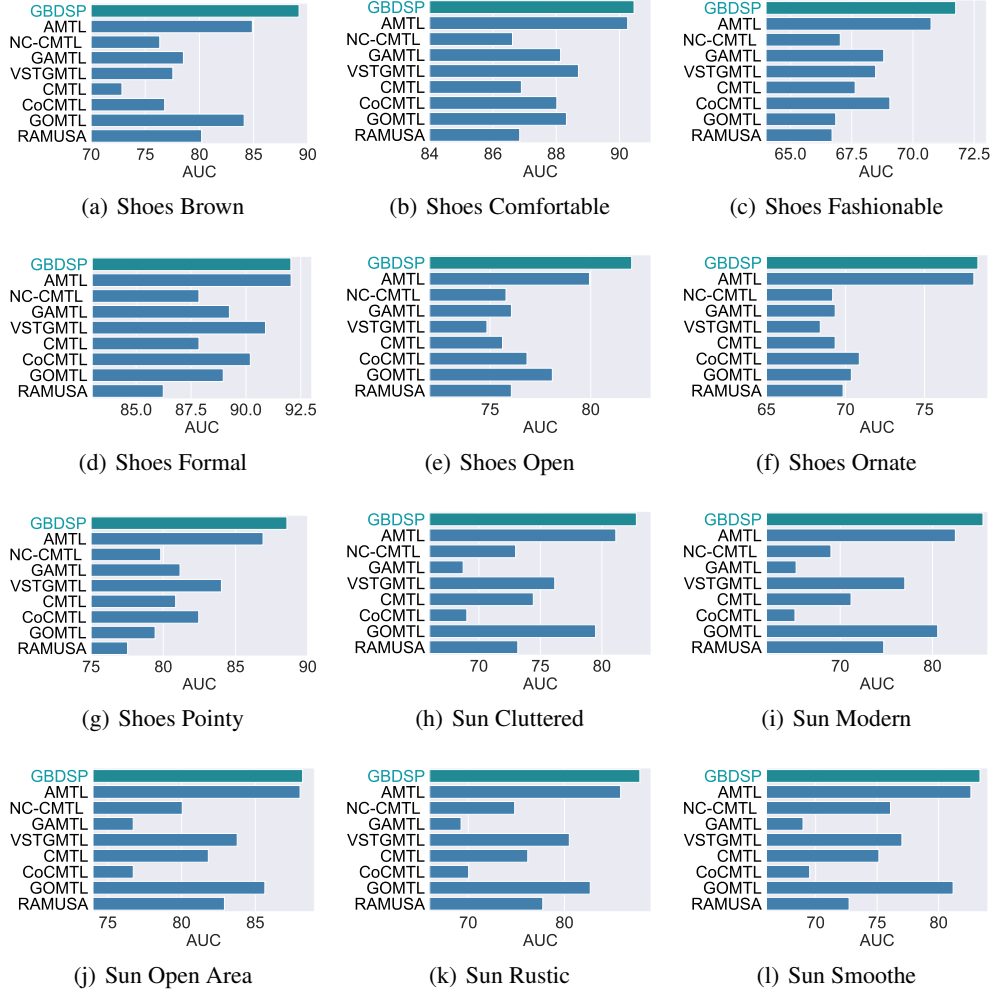


Figure 7: Attribute-wise Performance Comparisons on Shoes and Sun